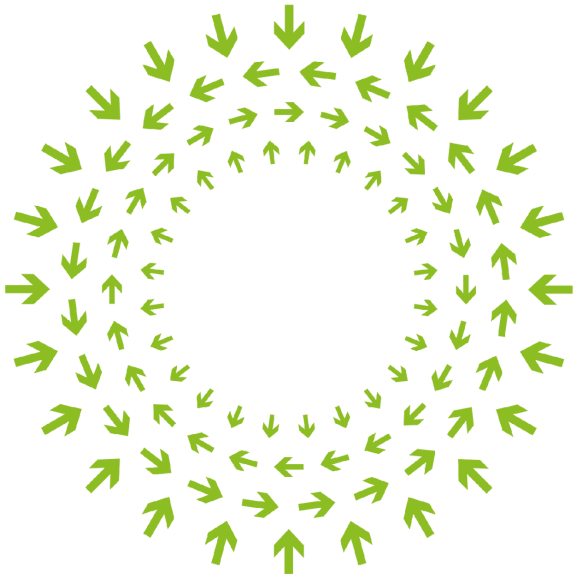


OPEN

Compute Project

Hyperscale Tiered Memory Expander Specification – for Compute Express Link® (CXL®)

Revision 1

Version 1.0

Base Specification Template v1.2

Effective October 27, 2023

Author: Prakash Chauhan and Chris Petersen, Meta
Author: Brian Morris and Jerome Glisse, Google

Compute Express Link® and CXL® are trademarks of the Compute Express Link Consortium

Table of Contents

1. License 2
1.1. Open Web Foundation (OWF) CLA 2
2. Change Log 4
3. Scope 5
4. Overview 6
5. Product Requirements 6
6. Environmental Regulatory Compliance and Requirements 11
7. Physical Specifications 11
8. Electrical Requirements 12
9. Thermal Design Requirements 13
10. Interfaces 13
11. Onboard Power System 15
12. Other/ Miscellaneous 16
13. System Firmware 17
14. Hardware Management 17
15. Security 18
16. Software Support 20
17. Arm SystemReady (only for Arm-based Systems) 21
18. References 21

1. License

Open Web Foundation (OWF) CLA

1.1. Open Web Foundation (OWF) CLA

Contributions to this Specification are made under the terms and conditions set forth in Modified OWF-CLA-1.0.2 (As of June 1, 2023) (“Contribution License”) by:

Google and Meta

Usage of this Specification is governed by the terms and conditions set forth in **Modified OWFa1.0.2 Final Specification Agreement (FSA) (As of June 1, 2023) ("Specification License").**

You can review the applicable Specification License(s) referenced above by the contributors to this Specification on the OCP website at http://www.opencompute.org/participate/legal-documents/. For actual executed copies of either agreement, please contact OCP directly.

Notes:

1) The above license does not apply to the Appendix or Appendices. The information in the Appendix or Appendices is for reference only and non-normative in nature.

NOTWITHSTANDING THE FOREGOING LICENSES, THIS SPECIFICATION IS PROVIDED BY OCP "AS IS" AND OCP EXPRESSLY DISCLAIMS ANY WARRANTIES (EXPRESS, IMPLIED, OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THE SPECIFICATION. NOTICE IS HEREBY GIVEN, THAT OTHER RIGHTS NOT GRANTED AS SET FORTH ABOVE, INCLUDING WITHOUT LIMITATION, RIGHTS OF THIRD PARTIES WHO DID NOT EXECUTE THE ABOVE LICENSES, MAY BE IMPLICATED BY THE IMPLEMENTATION OF OR COMPLIANCE WITH THIS SPECIFICATION. OCP IS NOT RESPONSIBLE FOR IDENTIFYING RIGHTS FOR WHICH A LICENSE MAY BE REQUIRED IN ORDER TO IMPLEMENT THIS SPECIFICATION. THE ENTIRE RISK AS TO IMPLEMENTING OR OTHERWISE USING THE SPECIFICATION IS ASSUMED BY YOU. IN NO EVENT WILL OCP BE LIABLE TO YOU FOR ANY MONETARY DAMAGES WITH RESPECT TO ANY CLAIMS RELATED TO, OR ARISING OUT OF YOUR USE OF THIS SPECIFICATION, INCLUDING BUT NOT LIMITED TO ANY LIABILITY FOR LOST PROFITS OR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR PUNITIVE DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND EVEN IF OCP HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

2. Change Log

Date	Version #	Author	Description
31OCT23	1.0	Prakash Chauhan Brian Morris	Defines requirements for CXL memory expander.

3. Scope

This specification describes the requirements of a CXL memory expansion chip including speeds of the interfaces and landing zones of the chip power, performance, management and board level requirements. This includes hardware, firmware, test, validation, and software requirements for deployment at scale.

4. Overview

This Product Requirement Design Specification is focused on the technical requirements for a CXL Memory Expander chip. At a high level, the CXL Expander consists of a CXL 3.x (PCIe Gen6 data rate capable) x16 interface with 4x DDR4 channels / 8x DDR5 sub-channels (using combo controller and Phy) with support for caching and compression. The device shall have a management sub-system that provides configuration, monitoring, and error reporting (both inband and out-of-band). The solution needs to focus on minimizing power consumption and memory access latency. Please refer to the JEDEC standard JESD319 CXL Memory Controller Specification which is in development and, when published, may be referenced for baseline requirements.

5. Product Requirements

Requirement	Details
Latency	The maximum unloaded latencies listed below are pin-to-pin latencies on the CXL device, i.e. the time from when the read request arrives at the CXL pin of the device until the data response is delivered on the CXL pin of the device. ● Block cache (SRAM) hit latency: 30ns ● Access to cache line in uncompressed region: 90ns ● Access to a cache line in an uncompressed block: 150ns ● Access to a cache line in a compressed block: 250ns ● Tail latency for accessing a cache line in an compressed block of < 1us including worst case lookup latency, decompression, power-state transitions etc
Bandwidth	● The CXL interface should be at least 80% efficient for read-only requests (e.g. a 32GT/s x16 interface should be capable of at least 51GB/s per direction) after accounting for all overheads. ● The CXL interface should be at least 75% efficient for write-only requests (e.g. a 32GT/s x16 interface should be capable of at least 48GB/s per direction) after accounting for all overheads.
Memory controller	● Ability to interleave across channels and control the granularity ● Support for Post Package Repair (PPR) commands and address translation (Hard PPR and Soft PPR)

	● Ability to send custom MR commands to support vendor specific MR options
	● Patrol and demand scrubbing should be implemented and configurable to run periodically
	● Ability to read and report DIMM SPD information including temperature
	● Ability to set target memory speed (e.g. DIMM may be capable of 3200MT/s but we force the bus to 2667MT/s)
	● Open, closed, and adaptive close paging policies
	● Ability to control refresh rates, ZQ calibration, Rcomp, and other periodic training
	● Ability to configure DFE if available
	● Support for Address/control parity protection
	● Enable reading of correctable and uncorrectable error counts
	● MBIST engine to test and initialize DRAM aligned to the requirements in the JEDEC CXL01 Memory Controller Specification
	● Counters
	  ● Read count
	  ● Write count
	  ● Activate count
	  ● Refresh count
	  ● Precharge (due to page conflict) count

Compression and Block cache	● Bandwidth
	  ○ Decompress 46GB/s matched to 4 Channels at 1867MT/s compressed data
	● Block size
	  ○ 4KB required
	  ○ 1KB required
	● Algorithm
	  ○ LZ4 required
	  ○ Zstd optional
	  ○ Configurable compression effort level to optimize for latency
	  ○ Incompressible blocks should be stored uncompressed but placed in the compressed region
	● Regions - Two independent regions
	  ○ User configurable at boot time: capacity allocated for compressed vs uncompressed memory

	● Block Cache ○ Device shall include a block cache to hold recently decompressed blocks or prefetched cache lines from uncompressed regions to take advantage of spatial and temporal locality. ○ 16MB or larger size is desirable ● Performance features ○ Defragmentation of compressed memory ○ Fast table lookups ○ Compression bypass for incompressible blocks ○ Back pressure to host when capacity thresholds exceeded ● Reporting/error handling ○ Out of capacity error ○ Page poisoning recorded in case of over-capacity write ■ Returns poison in case of read request to page ■ Clearable via vendor specific mailbox commands ● Counters/trackers ○ Block cache hit count ○ Block cache miss count ○ Read count ○ Write count ○ At least X entry LRU tracker for hot page addresses (hits in page cache) ○ At least Y entry LRU tracker for compressed page addresses ○ At least Z entry LRU tracker for incompressible page addresses ○ Compression ratio histogram ○ Defragmentation events ○ Capacity of used memory in compressed region ■ Capacity used by compressed blocks ■ Capacity used by incompressible blocks
ECC	DDR4: ● Support for a Reed Solomon based ECC solution ● Must support x4 chipkill ● Must support at least 2x parallel symbol corrections per beat including across 2x or more DRAMs and per codeword DDR5: ● Support for a Reed Solomon based ECC solution with bounded fault alignment

	● Must support x4 chipkill and x8 bounded fault ● Must support at least 2x parallel symbol corrections per beat including across 2x or more DRAMs and per codeword
Other RAS	DRAM ● Patrol and demand scrubbing with hardware based immediate writeback of corrected data in addition to logging address/syndrome of error ● CA Parity retry via ALERT# ● Post Package Repair Memory error logging shall capture the following attributes for the first, next, and last memory error for each channel: ● DIMM ● Rank ● Bank ● Row ● Column ● Device ● Error syndrome ● FRU information CXL Link RAS ● The device shall support CXL Link RAS features aligned to the requirements in the JEDEC CXL01 Memory Controller Specification.
Debug, tracing and monitoring	● PCIe and DDR eye capture and reporting mechanisms need to be provided ● Memory performance counters including event counters for reads, writes, refreshes, bank utilization, bus turnarounds, activations etc ● Ability to track latency of CXL.mem requests and responses and collect statistics such as min, max, average latencies inside the CXL Expander ● Memory command/addressing trace buffer of up to 1ms depth with programmable triggers (no data is captured)

	● Ability to report utilized read, write, “unused”, and total memory bandwidth ● CXL command/addressing trace buffer of up to 512 commands deep with programmable triggers (no data is captured) ● Deep LTSSM state tracking that includes a history of state transitions and reasons for transitions which can be used for link debug ● CXL .mem counters to track counts of opcodes ● CXL counters/events for unexpected/unsupported operations ● CXL counters to link level utilization and debug low utilization/starvation cases ● Ability to track memory retries and the source of those retries ● Ability to report current memory controller configuration and DDR training results ● Ability to retrain DDR interface repeatedly (non-mission mode) to gather interface margins at various operating conditions ● USB/UART debug port to dump logs and boot info for FW (boot logging) ● JTAG support in system for manufacturing 1. SCAN snapshot support ● Please refer to CXL Debug and Mgmt Spec (provided separately) for additional details on expectations for device debug and management
Error handling and injection	● PCIe AER correctable and uncorrectable error handling is required and correctable/uncorrectable error counters need to be implemented in addition to standard AER registers.Internal and external memory should be protected by SECDED mechanisms. Internal memory errors need to be counted and reported. ● Both system ECC and on-die DRAM correctable and uncorrectable errors need to be tracked and counted via the CXL memory device interface and should include syndromes ● Ability to isolate an ECC error down to rank/bank/column/address/DQ level for correctable errors, and down to rank level for uncorrectable errors ● Ability to report errors with FRU (Field Replaceable Unit) identifiers

	• Ability to inject DRAM ECC errors (for desired count and rate of errors) • Support for Linux Kernel Mechanism (EINJ) to inject memory errors through the tool. • Provide clear indication to the system on uncorrectable errors, and provide a trigger to the CPU to follow the crash-dump mechanism (similar to memory UCEs) • Any detected uncorrectable errors should cause the errored address locations to be explicitly poisoned. Any transactions where data is known to be corrupted should have the poison bit set.
Resets	• Support for cold, warm, hot, and CXL resets • Support for FW activation without impacting datapath
Manufacturing features	• JTAG scan support at the PCB level • Support MEMBIST with the ability to load custom patterns, set loop count, adjust numerous timing/voltage parameters (e.g. Vref, RxDq, TxDq, etc), and control whether it is run during POST
Interoperability	• CXL Expander interop and compatibility to work with CXL 2.0 for initial enabling and/or targeted deployment and CXL 3.1 for broader deployment.

6. Environmental Regulatory Compliance and Requirements

Requirement	**Details**
Operating temperature range support	• Maximum Tcase: >= 85C (i.e. higher is better) • Maximum Tj: based on Tcase and Theta-(jc)

7. Physical Specifications

Requirement	**Details**
Package	Package dimensions

	• Width <= 25mm (side with CXL interface) • Length <= 50mm package (side with DDR interface) • Minimum BGA pitch 0.8mm. • Total thickness (from BGA ball to top of package) shall be less than 2.5mm
Ball out	• Ballmap should be designed for easy fan-out with no more than 16 layers of PCB • Standard PCBA process is preferred for low cost.
Board design	Maximum layout count is 16 (14 preferred). No HDI. Back drilling is OK. • VLL(Very Low Loss) preferred material for PCB • 6 DIMMs on two channels on each side of the package • CXL Expander IO orientation preference: CXL to North and memory channels on East or West. Power on South • Airflow is South to North.

8. Electrical Requirements

Below are the DIMM configurations must be validated w/ all major memory suppliers. Silicon and board designs need to meet the speeds as specified below.

DDR4:

DIMM	**1DPC**	**2DPC**	**3DPC**
16GB (2Rx8) – 8Gb	Up to 3200MT/s	Up to 2933MT/s	Up to 1867MT/s
32GB (2Rx4) – 8Gb	Up to 3200MT/s	Up to 2933MT/s	Up to 1867MT/s
64GB (2Rx4) – 16Gb	Up to 3200MT/s	Up to 2933MT/s	Up to 1867MT/s

DDR5:

DIMM	**1DPC**	**2DPC**
32GB (2Rx8) – 16Gb	Up to 5200MT/s	Up to 4800MT/s

48GB (2Rx8) – 24Gb	Up to 5200MT/s	Up to 4800MT/s
64GB (2Rx4) – 16Gb	Up to 5200MT/s	Up to 4800MT/s
96GB (2Rx4) – 24Gb	Up to 5200MT/s	Up to 4800MT/s

PCI Express/CXL:
Links need to be 64GT/s spec compliant on BER with board, substrate and Silicon losses properly accounted for and simulated.

9. Thermal Design Requirements

Requirement	Details
Temperature monitoring	● On-die thermal diode is required. Within +/-2% accuracy. ● Temperature reporting via both in-band and out of band CXL management interface is required. On-die sensor and DIMM temperature sensors need to be monitored ● Configurable thermal thresholds that generate error logs and interrupts, and throttle the chip in case of severe excursion are required.

10. Interfaces

Requirement	Details
CXL Interface	● CXL 3.1 compliant and backwards compatible to CXL 2.0 ● Support for CXL 2.0 device register and command interface as specified in the management section ● Support for CXL performance monitoring interface with low host/device performance overhead and low latency. This interface should be used for all performance telemetry. ● Up to 1x x16 CXL width, but ability to reduce configuration to x8 and x4 to save power consumption ● Support for 2-way multi-home to a single device (2x8 CXL)

	• Support for link width (down to x4) and speed degradation (down to gen3) • Support for 64GT/s link speed • Common clock with SSC support • PCIe PHY settings must be tunable to improve adaptive convergence for different channel designs • BER target of 1e-15 • Support for 2, 3, and 4-way interleaving across multiple controllers

DDR4 requirements	Support for:
	• 4Gb, 8Gb, 16Gb dies
	• 1DPC, 2DPC, 3DPC, and 1, 2 or 3 DIMMs populated in 3DPC
	• Standard JEDEC RDIMMs
	• Up to 6 ranks per channel
	• Support x4, x8 DRAMs
	• Support 2400-3200MT/s capable devices
	• Support bus speeds as follows:
	  ○ 3DPC at 1867MT/s
	  ○ 2DPC at 2933MT/s
	  ○ 1DPC at 3200MT/s
	• 4x 72 bit channels
	• Support for mixed DIMM types and capacities per channel (e.g. 2R x8 and 2R x4)

DDR5 requirements	Support for: • 16Gb, 24Gb, 32Gb dies • 1DPC, 2DPC and 1 or 2 DIMMs populated in 2DPC • Standard JEDEC RDIMMs • Up to 4 ranks per channel

	● Support x4, x8 DRAMs with Bounded fault ● Support 4800-6400MT/s capable devices ● Support bus speeds as follows: ○ 2DPC at 4800MT/s ○ 1DPC at 5200MT/s ● 8x 40/36 bit channels ● Support for mixed DIMM types and capacities per channel (e.g. 2R x8 and 2R x4) ● RFM and ECS ● Read and report DRAM temperature sensors
Sideband Interfaces	● Support for I3C sideband interface. Sideband interface shall provide the ability to debug and restore the primary PCIe interface if it fails for any reason ● Support for QSPI flash to store FW, configuration and logs ● At least 2x GPIOs to support FW-controllable functions (e.g. LED control or a “FW has booted and is live signal”) ● Support for UART to observe firmware debug and informational logs

11. Onboard Power System

Memory expander power envelope is specified below

Requirement	**Details**
CXL Expander Power	● The workload assumptions for the power requirements includes: ● Random 67% reads with 33% writes ● 10% page hits ● For a 75C Tj, typical process, and nominal voltage, the max RMS power of the CXL Expander should not exceed: 1. 15W: CXL 32GT/s x16, DDR4-3200

	2. 20W: CXL 64GT/s x16, DDR5-5200 ● Clock gating and other power optimizations should be used to turn off unused blocks. ● AVS should be supported if possible. ● The CXL Expander power shall not exceed: 1. 20W for peak electrical transient conditions: CXL 32GT/s x16, DDR4-3200 2. 25W for peak electrical transient conditions: CXL 64GT/s x16, DDR5-5200

12. Other/ Miscellaneous

Requirement	**Details**
Provide HW IP information	● PCIe/CXL PHY/controller ● DDR4/5 PHY/controller
Validation plan	● Component level validation, for example high-speed interface test; functional test; power sequencing; reset flow etc ● System integration validation: test full system solution with CXL Expander working in a platform with all compatible DIMMS/DRAMs (customer to provide a validation matrix)
Deliverables to Customer	● PCIe PHY and DDR PHY models and package models ● Flowtherm thermal models ● Ballout and pin definitions ● Package dimensions and drawings ● Datasheet ● Board design guidelines ● Test reports ● FW releases and FW update tools including secure FW flows

13. System Firmware

Requirement	Details
Firmware	● Firmware/software running on the SOC shall be as lightweight as possible and as small as possible. It shall fit in a SPI EEPROM/NOR Flash and not require NAND Flash. ● Firmware updates and activations must be supported via both in band and out of band ● FW should generate FW logs and crash dumps to debug FW ● FW should not introduce dependencies in debug, management and firmware update as much as possible (this is to allow debug and recovery when firmware crashes or gets corrupted)

14. Hardware Management

Inband and out of band management will be supported using CCI commands defined in the CXL 3.1 specification with vendor specific enhancements as needed. More detailed management specifications will be covered by customer specific documentation, as these requirements are hardware agnostic and customer specific.

Requirement	Details
CCI mailbox commands	● Memory Device Registers ● Get/Clear Event Records 1. Common events 2. DRAM events 3. Memory Module events ● Get/Set Event Interrupt Policy ● Get FW Info ● Transfer/Activate FW

	• Get/Set Timestamp • Get Supported Logs • Get Log • Identify • Background Operation Status • Identify Memory Device • Get Health Info • Get/Set Alert Configuration • Maintenance commands • Get Poison List • Inject Poison • Clear Poison • Component state dump log • Get/Set Features

15. Security

Requirement	**Details**
Security	Required: • Support secure boot to check signed FW on boot 1. Hashing of SHA256 or better 2. Signed using HSM-based Ed25519, NIST P384, or RSA4k • Check signed FW before flashing and report at least the following failures: 1. Signature mismatch

	2. Anti-rollback version error 3. Key revocation error ● Support recovery mode in case of bricking attacks or incorrectly signed FW ● Ability to disable debug interfaces (e.g. JTAG and UART), and securely re-enable them if needed in production environments ● Support SPDM-based Attestation ● Support key revocation ● Support rollback protection ● iROT with Caliptra ● Support for CXL IDE is strongly desired but not required. ● We expect encryption is required for DRAM contents. 1. Support for at least AES-XTS 256 is required. 2. Support for one key is required. 3. Support for up to 256 keys with dynamic capacity is desired but not required. More detailed requirements will be provided in detailed discussions

16. Software Support

Feature	Details
Driver	● Support to use open-source CXL driver upstreamed to Linux Redhat kernel distro
Userspace Linux tools	● Support strict version control ● Sources available on Github and contributed to by different CXL vendors (akin to nvme-cli) ● Should not access any user data ● Should support human readable and JSON output formats ● Querying CXL.io or mem counters should not adversely impact user workloads ● Ability to gather diagnostic logs (binary or human readable) to be used for supplier FA ● Validate support for EDAC kernel driver and EINJ.
Validation tools	● PCIe eye/margin tool ● DDR4/DDR5 eye/margin tool ● Power Stress tool ● Error injection tool - A tool which can inject correctable / uncorrectable errors at a desired rate and support interleaving of multiple errors on both CXL and memory interfaces. ● Diagnostic tool – A tool which can perform validation of observed correctable / uncorrectable errors with synthetic test workloads for the addressable memory space.(from the OS) (A repro tool for CEs/UCEs) ● Manufacturing tests: 1. A tool which can perform validation of the entire memory space for any marginal failures in DRAM (from both the OS & BIOS space). 2. A tool which can run manufacturing test patterns with user space control for executing and consuming the tests.

17. Arm SystemReady (only for Arm-based Systems)

Not specified

18. References

1. JEDEC standard JESD319: Compute Express Link (CXL®) Memory Controller Specification