Fabric Technology Required for Composable Memory

Compute Express Link (CXL) 3.0 is a fabric solution to tackle memory and GPU/accelerator inefficiency. The solution was introduced by the CXL Consortium, an open industry standard group formed to develop technical specifications that facilitate breakthrough performance for emerging use models. The Consortium has introduced specifications at a regular cadence with the CXL 3.0 specification being the latest. The specification expands fabric capabilities and management and allows for the creation of a composable and disaggregated memory fabric. The latest CXL 3.0 specification remains a point-to-point connection for memory expansion and memory pooling over short distances but fails to address the need for larger pools of memory over longer distances.

As data continues to grow, database and AI applications are being constrained on memory bandwidth and capacity. At the same time billions of dollars are being wasted on stranded and unutilized memory. According to a recent Carnegie Mellon / Microsoft report [1], Google stated that average DRAM utilization in its datacenters is 40%, and Microsoft Azure said that 25% of its server DRAM is stranded. Although the inherent requirements of data centers for larger memory pools and rack-scale or pod-scale applications are of great importance, the following issues remain:

- Latency
- Security (end-to-end encryption)
- Reliability, addressability, and serviceability (RAS)
- Fabric Management
- Peer-to-peer memory sharing and dataset sharing

IntelliProp introduced the Omega Memory Fabric solution for CXL devices to create a Network Attached Memory (NAM) system. The Omega Memory Fabric chips allow for dynamic allocation and sharing of memory across compute domains - both in and out of the server. This document compares IntelliProp Omega Memory Fabric chips to CXL 3.0 and reviews the applications and use cases for Network Attached Memory (NAM.)

IntelliProp Omega Memory Fabric Chips

IntelliProp's Omega Memory Fabric chips are built from hardware and software components. The hardware consists of three main components:

- A host adaptor card
- An endpoint adaptor
- A discrete switch

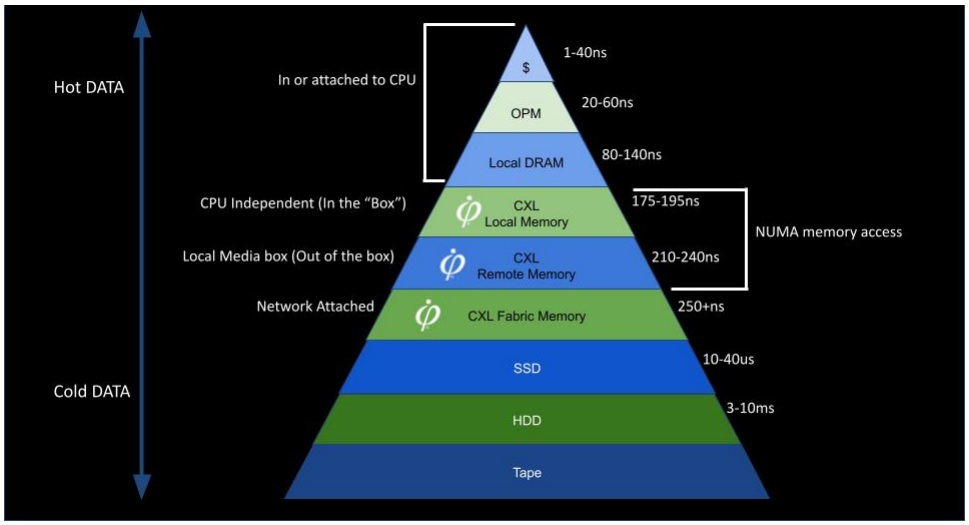
The host adaptor card is used for CXL memory expansion and pooling within the server box and includes an additional fabric port. The endpoint adaptors connect existing CXL memory components, GPUs, or direct fabric-attached memory components onto the fabric. A discrete switch provides additional routing between the server and external memory. The host adaptors and endpoint adaptors have built-in 8-port switches to allow for mesh-like topologies with redundancy and multi-pathing. The discrete switch provides for additional routing between server boxes and additional memory or media boxes.

The Omega Memory Fabric management software is composed of a fabric and resource manager. The software subsystem uses a standardized interface to Redfish and the OpenFabric Management Framework from the Open Fabrics Alliance. Omega's Memory Fabric Manager includes dynamic multi-pathing, congestion management, topology discovery, security with hardware isolation, asset allocation/deallocation, and automatic discovery. By utilizing a standard API interface, IntelliProp's Memory Fabric users have the flexibility to build their own fabric management software or resource management software.

IntelliProp provides customers with three new tiers of memory beyond local memory. Each tier creates a new latency domain.

- The first is CXL local memory that can be pooled and shared with the CPU or cores within a server.
- The second is remote memory that may be located in a different chassis, like a media box. The media box may be in the same rack or nearby racks.
- The third tier is a true memory fabric with memory that could be several switch "hops" away.

The diagram below shows typical memory and storage tiers and latencies. The Omega Memory Fabric management software can access data from hot to cold with increasing latency for colder data. When local memory is insufficient, pages of data are loaded from Non-Volatile Memory (NVM) storage such as SSD and HDD into local memory. IntelliProp's Omega Memory Fabric extends CXL's connection beyond simple memory expansion and provides access to more memory, whether in the box, out of the box, or network attached.

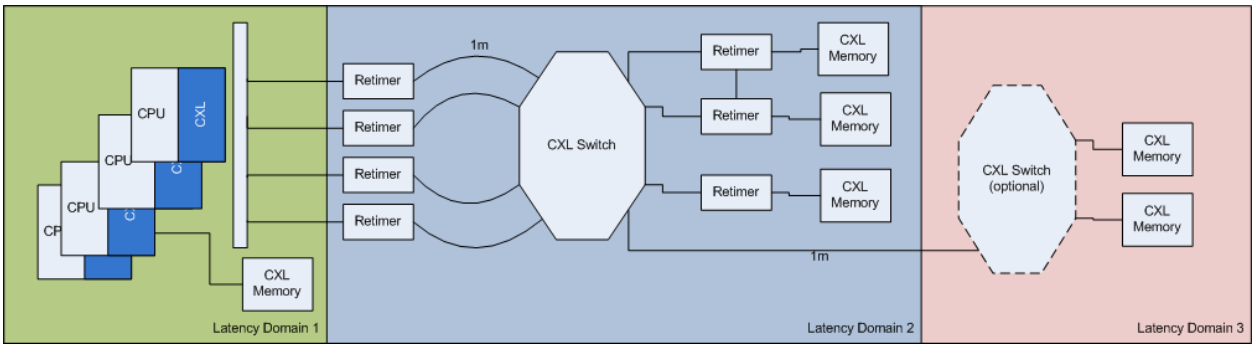


This paper will explore how the Omega Memory Fabric provide features beyond today's CXL solutions and brings future CXL advantages to the data center including latency, security, RAS (reliability, accessibility, serviceability), fabric management, and peer-to-peer access.

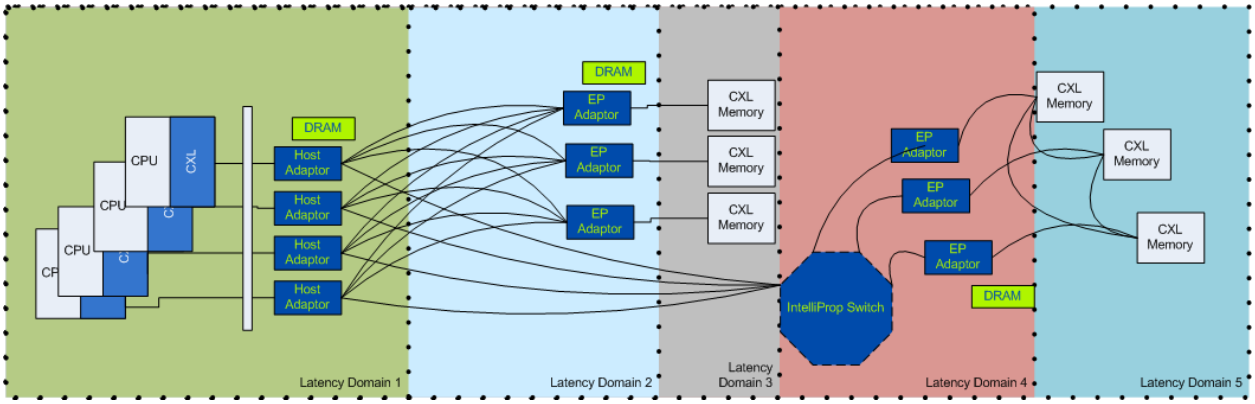
Latency

Latency is very important to existing applications. Because many applications are aware of Non-Uniform Memory Access (NUMA) latencies and have been designed with this in mind, the applications can handle the new tiers of memory with little to no software modification.

The diagram below shows a typical CXL fabric composed of various latency domains. CXL is designed using PCIe electrical levels. Because of that PCIe distances over 6" require re-timers. The diagram shows re-timer circuitry, which can add 10ns of latency in each direction. Smart cables help lower the re-timer requirement for more extended reach.

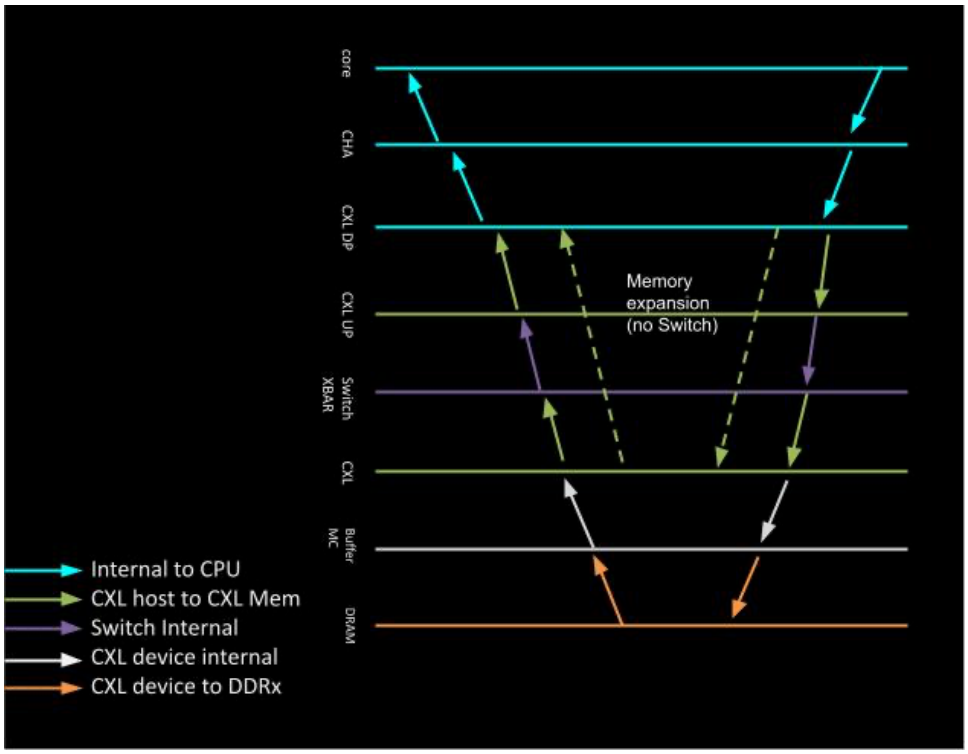


All the components of the Omega Memory Fabric solution are illustrated in the diagram below, which also represents the multiple latency domains created by using a memory semantic fabric. Each domain allows software and resource managers to allocate the appropriate latency profile memory to the applications that need it.



In the Omega Memory Fabric solution, Latency Domain 1 is similar to the latency of a CXL memory expansion solution. This approach creates direct-attached memory expansion. Since the endpoint adapters have built-in switches, a new domain, Latency Domain 2, is added, allowing memory to be disaggregated and shared across the processing elements. Latency Domain 3 allows standard CXL memory components to be added to the memory pool with some additional latency. Latency Domain 4 incorporates IntelliProp Switches to create a larger memory pool and accelerators. Finally, Latency Domain 5 allows CXL memory to be pooled in larger quantities.

The diagram below shows an example of round-trip latency from the CPU to the memory and back. The first segment is for local memory expansion, and the second segment shows memory further out through a CXL switch or fabric.



The table below shows the round-trip latency from a CPU core to the memory and compares the latency between a modeled CXL fabric and the IntelliProp Omega Fabric solution. This comparison doesn't consider any re-timers or cable delays. The IntelliProp latency is measured and scaled from an existing FPGA solution. The CXL latency is derived from target numbers shared by CXL member companies.

Memory "Type"	CXL Reference Latency	IntelliProp CXL Fabric Latency
Cache	1-40ns	
On Package Memory	60-80ns	
Local Memory	80-140ns	
CXL Memory (Direct Attached Memory)	176-191ns (Latency Domain 1)	
Memory Shared (IntelliProp only)		216-231ns (Latency Domain 2)
CXL Memory behind a switch	276-291ns (target) (Latency Domain 2)	266-281ns (Latency Domain 3)
Memory behind a cascade Switch (IntelliProp only)		252-267ns (Latency Domain 4)
Cascaded Switches	376-391ns (target) (Latency Domain 3)	302-317ns (Latency Domain 5)

Security

While CXL fabrics have built-in security, the IntelliProp Omega Memory Fabric solution provides a low latency method for security. IntelliProp does not need to unencrypt packets at the switch level because header information is not encrypted. In CXL 3.0 switches, packets have to be unencrypted and re-encrypted for switching transfers using host-based routing (HBR) or packet-based routing (PBR). The decryption and encryption at every switch hop can increase latency by 28-40 clock cycles.

IntelliProp's Omega Memory Fabric solution utilizes hardware-enforced isolation thereby adding another key security feature.

Reliability Accessibility and Serviceability (RAS)

CXL3.0 is an excellent enabler for creating large amounts of memory pooling and sharing. Since data stored in the memory will be further from the CPU and shared by multiple CPUs, certain data storage management practices have to be enabled to keep data reliable, available, and serviceable (also known as RAS). While CXL 3.0 introduces the industry to memory fabrics, it misses some key points that other protocols have defined in the fabric space previously.

Omega Memory Fabric along with IntelliProp's innovative Fabric Management Software and Network Attached Memory (NAM) system address RAS features not available in CXL3.0:

1. End to End re-try. CXL uses link level reliability (LLR) which only sends acknowledgements from a point to point and retries packets, or flits, at the lowest layer. If a link goes down later in the chain, the originator is not aware of this loss of packet. The IntelliProp NAM fabric does retries from the originator to the destination with acknowledgments from the destination.
2. Dynamic Multi-pathing. When CXL links go down the entire path and switches have to be reset and links re-established. The IntelliProp NAM fabric can dynamically re-route packets between CXL hosts and CXL memory to increase reliability.
3. Congestion Routing. CXL paths are hard defined and all traffic between points flow through the same path. The IntelliProp NAM fabric takes advantage of congestion management built into the switches to help transfer packets onto the least congested routes.

IntelliProp has demonstrated these RAS features using an FPGA based network of CXL host adaptors connected to existing CXL 1.1 Sapphire Rapids hosts, switches, and memory modules.

Fabric Management

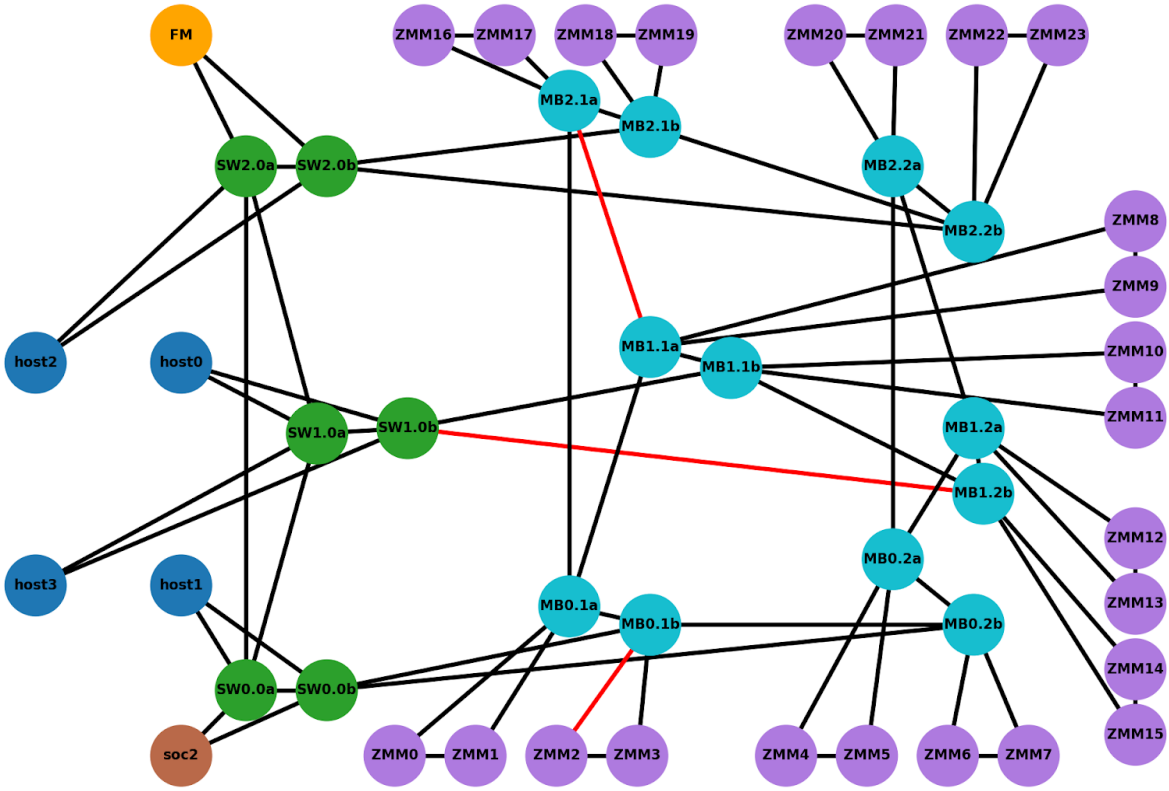
The CXL Consortium has been focused on the CXL protocol interface and on memory expansion. As the focus shifts to scaling out the CXL interface to a fabric, Consortium workgroups are beginning to be devoted to a fabric management specification. Considering that the Gen-Z Consortium spent 4+ years

developing the fabric management specification, the CXL consortium will either leverage heavily from the Gen-Z management specification (most likely) or spend 4+ years developing from scratch.

Meanwhile, IntelliProp has already developed a fabric manager based on the Gen-Z specification that is host agnostic. The fabric manager can be hosted on a CXL capable server attached to the fabric, as well as on other fabric attached devices including FPGA SOC's, PCIe servers, a BMC connected to a switch, or even a CXL memory module.

The IntelliProp Omega Memory Fabric has demonstrated key features including fabric attached component discovery and configuration, link status monitoring, and secondary/distributed fabric manager features for management failover. More impressive is the fabric manager's ability to dynamically allocate and deallocate fabric attached memory to servers for use as memory sematic memory, or block devices. The Omega Memory Fabric manager includes a well-defined API that has been utilized by both the Open Fabric Alliance (OFA) and a commercial hardware manager (Liqid Command Center). IntelliProp and Liqid have posted videos dynamically adding and removing (composing) memory resources for use by a CXL connected server.

Below is a graphical representation of a medium scale fabric with 4 CXL hosts, and two FPGA SOC hosts. One of the FPGA hosts is running the primary IntelliProp fabric manager. The other FPGA host runs the secondary or backup IntelliProp Omega Memory Fabric manager. Also shown are 24 memory modules, 18 fabric switches and all links (including 3 links that were unplugged as represented by the red link lines). This drawing was dynamically generated by the IntelliProp Omega Memory Fabric manager.



Peer to Peer Access

Looking beyond memory, there is customer demand for accelerators and GPUs to connect to the memory fabric to share datasets, memory, and other resources. In order to properly share memory, CXL end points (.mem or .cache) need the ability to have bidirectional control. IntelliProp Omega Memory Fabric allows all connected devices to be utilized as both requesters and responders. For example, GPUs can request additional memory as needed. GPUs can also access the same memory as the CPUs. This allows CPUs to place datasets in memory locations that GPUs can access. The dataset doesn't have to move from CPU memory to GPU memory.

Summary

Taking advantage of previous specifications, the IntelliProp Omega Memory Fabric solution creates a NAM fabric that is low latency, comes with a fabric manager, allows peer-to-peer traffic, and embodies RAS features. By composing and sharing memory as needed, NAM increases memory efficiency and lowers CAPEX costs.

IntelliProp's Omega Memory Fabric eliminates memory bottleneck and allows for dynamic allocation and sharing of memory across compute domains both in and out of the server, delivering on the promise of Composable Disaggregated Infrastructure (CDI) and rack scale architecture, an industry first. Omega Memory Fabric's features go beyond the current CXL 3.0 specifications enabling next-generation performance, scale and efficiency for database and AI applications. For the first time, high-bandwidth, petabyte-level memory can be deployed for vast in-memory datasets, minimizing data movement, speeding computation and greatly improve utilization – allowing for composable memory to become a reality.

[1] Source: Carnegie Mellon University, Microsoft Research and Microsoft Azure report, *First-generation Memory Disaggregation for Cloud Platforms*, March 2022.