5 Ways CXL Will Transform Computing

Compute Express Link (CXL) devices can take advantage of available PCIe interfaces to open an additional conduit that complements the overtaxed memory bus. More lanes, more data movement, more performance.

Some technologies experience stunning breakthroughs every year. In memory, it can be decades between major milestones. Burroughs invented magnetic memory in 1952 so ENIAC wouldn't lose time pulling data from punch cards.[1] In the 1970s DRAM replaced magnetic memory while in the 2010s, HBM arrived.

Compute Express Link (CXL) represents the next big step forward. CXL devices essentially take advantage of available PCIe interfaces to open an additional conduit that complements the overtaxed memory bus. That means more lanes, more data movement, and more performance.

Additionally, and arguably more importantly, CXL will change how data centers are built, operate, and work. It's a technology that will have a ripple effect. Here are a few scenarios on how it can potentially impact infrastructure:

1. DLRM Gets Faster and More Efficient
2. Data-Center Real Estate and Facility Costs Drop
3. Free Memory is There for the Taking
4. Higher Utilization of CXL Assets
5. Better Data-Center Operations Through Telemetry

DLRM Gets Faster and More Efficient

Memory bandwidth—the amount of memory that can be transmitted from memory to a processor per second—has chronically been a bottleneck because processor performance increases far faster and more predictably than bus speed or bus capacity. To help contain that gap, designers added more lanes or added co-processors.

For example, the Marvell Structera A does both. It's an intelligent memory accelerator that sports 16 Arm Neoverse N2 cores, 200 Gb/s of memory bandwidth, up to 4 TB of memory, and consumes under 100 W along with processing fabric. It's essentially a server-within-a-server with outsized memory bandwidth for bandwidth-intensive tasks like inference or deep-learning recommendation models (DRLMs). Cloud providers need to program their software to offload tasks, but doing so brings a number of benefits (**Fig. 1**).

Take a typical high-end x86 processor. Today it might
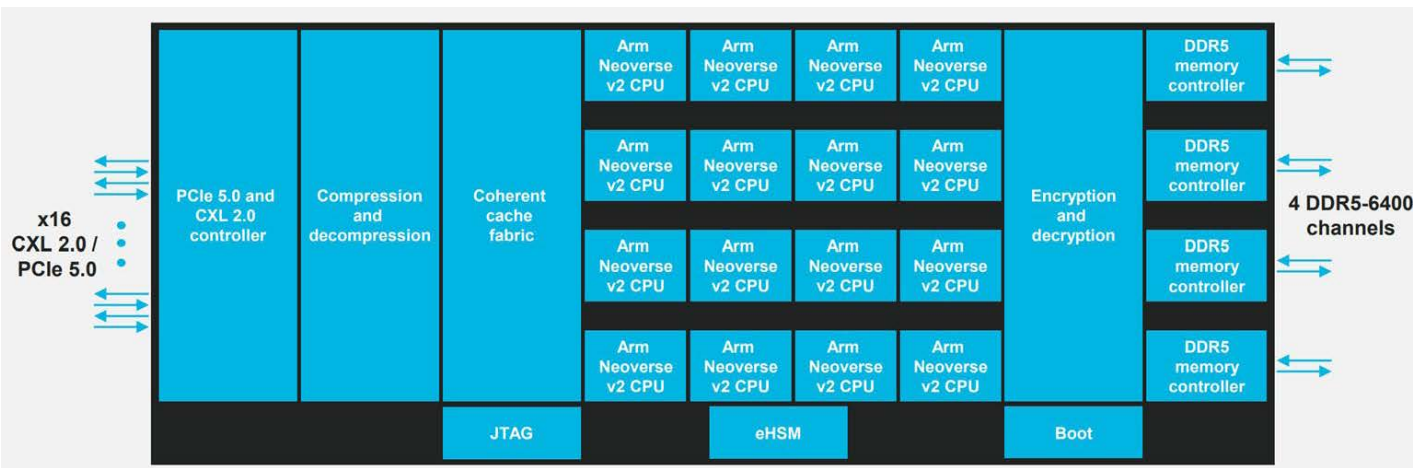
1. Shown is a block diagram of the Marvell Structera A memory accelerator.
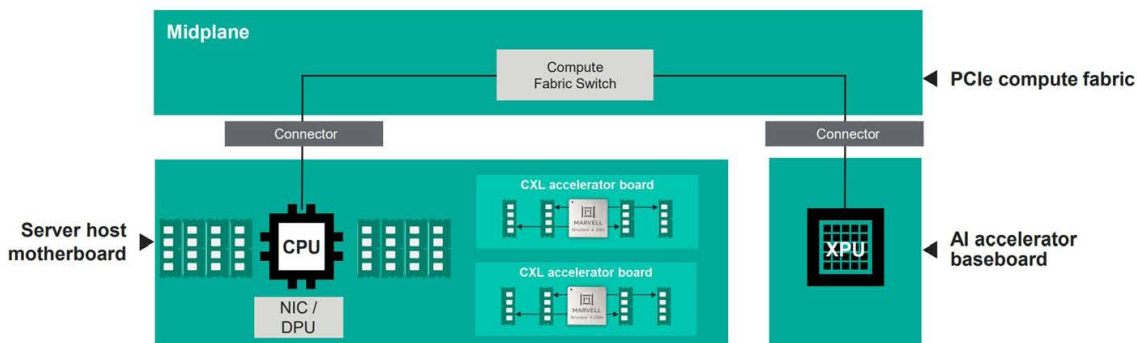
Server has **96 cores** with **800 GB/s of DRAM bandwidth** and consumes **600W power**
8.33 GB/s per CPU core and **0.75W per GB/s**

Increases number of cores by 50% and <u>doubles</u> memory bandwidth

2. The addition of memory accelerators increases the number of cores and expands memory bandwidth.

sport 64 cores, 400 Gb/s of memory bandwidth, up to 2 TB of memory (i.e., four top-of-the-line 512-GB DIMMs), and consume a maximum 400 W for a data-transmission power rate of 1 W per GB/s.

Adding one Structera A increases core count by 25% (64 to 80 processor cores) and memory bandwidth by 50% (400 to 600 Gb/s). Memory capacity, meanwhile, increases by 4 TB. One of the chip's optimized design features is that it sports four memory channels rather than the usual two to give the server an overall capacity of 6 TB, or 300% more than normal.

Total power goes up by 100 W, but, because of the outsized increase in bandwidth, transmission power per GB per second drops 17% to 0.83 W.

Adding two channels increases core count by 50% (to 96 cores), doubles memory bandwidth from the 400 Gb/s base to 800 Gb/s, and boosts memory capacity 6X to up to 12 TB. Power per bit transmitted, meanwhile, drops by 25% to 0.75 W per GB/s (**Fig. 2**).

Now multiply that by 40, roughly the number of servers in a rack. That's 3,840 more processing cores—or the equivalent of 60 additional servers—in the same finite space and 24 Tb/s more memory bandwidth. Total potential memory capacity increases to 400 TB.

**Data-Center Real Estate and Facility Costs Drop**

Now multiply the equipment in the single rack by 10, the number of racks per row. That's 600 extra servers, or equivalent to adding an edge-metro-sized data center (250 to 1,000 servers) without having to actually build one. And it scales. A real mid-sized data center with 14 rows would have the processing headroom of a small hyperscaler facility with 8,400 servers without the super-sized cooling and physical infrastructure.

Will your mileage vary? Absolutely. But the direction remains the same. CXL effectively lets data-center operators squeeze more capacity within the same roofline. With CXL, the footprint of physical infrastructure doesn't need to grow as fast to accommodate demand.

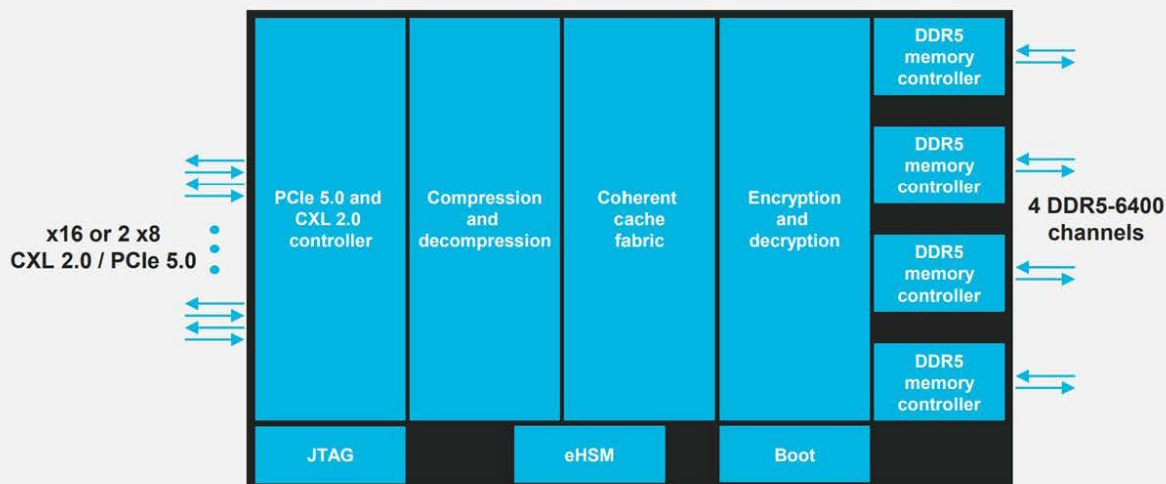**Free Memory is There for the Taking**

Because of its unusual design, DRAM lasts far longer than other components. It can operate for over a decade, although
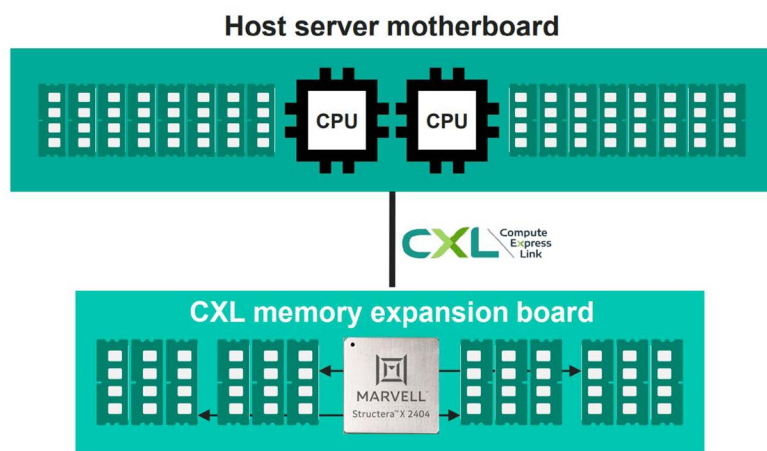
Recycle up to **12 DDR4 DIMMs** per expander (up to 6TB)

3. This a block diagram of Marvell Structera X memory accelerator.

Structera X 2504 high-level **block diagram**

4. With Marvell Structera X, two processors can take advantage of its additional memory.

the servers that contain it might be decommissioned after five years. Exabytes of DDR4 DRAM have been shipped to data-center operators since it debuted in the last decade. Servers based on next-generation processors coming to market in the near future are only compatible with DDR 5.

That means there will be vast stockpiles of useful, but unused, DDR 4 DIMMs that will grow as existing servers get replaced. A CXL memory expander like the Marvell Structera X can take advantage of this surplus. It has four DDR4 memory controller channels that support three DDR4 DIMMs per channel.

In other words, it lets you add 12 DIMMs to complement the 8 DIMM slots natively supported by the processor for 6 TB of capacity with 512-GB DIMMs. It also features LZ4 compression that can double the amount of memory accessed to 12 TB.

That's 12 TB of memory in one server. The NEC Earth Simulator, the last great vector computer ever made and the top supercomputer in the world between 2002 and 2004, only had 10 TB2 of memory (**Fig. 3**).

Let's say you wanted to fully max out the memory, but your CXL system was only compatible with DDR5. 256-GB DDR5 DIMMs currently sell for $2,488 in the spot market.[3] 12X is $29,856. 24X is $59,712. Populating an entire rack would cost $2,388,480.

Using DDR 4 wouldn't be free. There would be testing and integration costs. But the cost would be relatively minor.

But let's say you only wanted 64-GB DIMMs. DDR5 64-GB DIMMs sell for roughly $280 in the spot market.[4] 12X comes to $3,360 more per server, or $134,400 per rack

or $1,344,000 per row. It adds up. At the same time, massive amounts of memory avoid becoming part of the 62 million tons of e-waste generated per year.[5]

Higher Utilization of CXL Assets

Two processors can take advantage of the additional memory the Structera X provides—think of this as an early form of memory pooling, another benefit of CXL (**Fig. 4**).

Microsoft estimates that 25% of memory can become stranded[6] during certain functions because they are only accessible by one or two processors. In 2028, DRAM sales for servers will hit around $32-40 billion, not counting high bandwidth memory. (This is an average of different forecasts, which vary.[7])

Even with the low figure, that means up to $8 billion of memory shipped in one year alone will be idle for a good part of the time. With servers living around 7.6 years[8] these days, that one year of shipments represents only a fraction of the idling memory bought.

Or look at it this way. Memory systems consume between 25-40% of data-center power.[9] For the sake of argument, assume than only 5%, or one-fifth, of it is stranded at any given time. That's 1.25% to 2% of data-center power delivered to stranded assets (25 and 40 × .05).

Data centers consumed 460 terawatt hours (TWh) of power in 2022 and are on track potentially for 1,000 TWh in 2026.[10] That's 5.75 to 9.2 TWh, or billion watt hours, going to memory stuck in idle in 2022 and 12.5 to 20 TWh in 2026. A typical U.S. home consumes 10,500 kWh a year.[11] That's enough power for 1,190 to 2,000 U.S. homes in four years, or

nearly 6,500 homes in Germany.[12]

And, of course, there will be more ideas. Memory appliances will pave the way for truly disaggregated systems and data centers. Near-memory accelerators, potentially, will also be used to leverage machine learning to monitor ongoing system health or enhance security.

Better Data-Center Operations Through Telemetry

Telemetry, the art of monitoring real-time operational data for upping utilization or engaging in predictive maintenance, has been around for years. Advances in embedded processors and analytics, however, are making it far more useful.

CXL telemetry will be particularly useful with memory pooling and disaggregated systems. Instead of being tied to a single server or processor, memory, along with processors and networking, will be shared across a broad array of assets in an ad hoc fashion. Telemetry data will play a big role in developing the "magically delicious" formulas for running these systems.

References

1. ACM Digital Library, 1952
2. Top 500 Org running archive.
3. Memory Net, September 2023.
4. Memory Net, September 2023
5. International Telecommunications Union.
6. Microsoft. Pond: CXL-Based Memory Pooling Systems for Cloud Platforms, March 2023.
7. Yole Group, February 2023.
8. Omdia, March 2024.
8. DGIST, October 2021.
9. University of Illinois, Understanding and Optimizing Power Consumption in Memory Networks
10. International Energy Agency, Electricity 2024 report.
11. US Energy Information Agency.
12. Destatis.De (Statistics agency of the Bundestadt)