Micron's Perspective on Impact of CXL on DRAM Bit Growth Rate

Introduction

*CXL (*Compute Express Link™*) is a high-speed interconnect, industry-standard interface for communications between processors, accelerators, memory, storage, and other IO devices. CXL increases efficiency by allowing composability, scalability, and flexibility for heterogeneous and distributed compute architectures. The key advantage of CXL is the expansion of the memory for compute nodes, filling the gap for data-intensive applications that require high bandwidth, capacity, and low latency.*

In this paper we will present Micron's view: The memory market offers robust growth prospects and Compute Express Link™ (CXL)[1] will be net positive for DRAM bit demand growth and total addressable market (TAM) growth.

We'll start by discussing two challenges in today's IT systems, and then discuss how CXL addresses those problems. Last, we'll explain the impact we believe CXL will have on the memory market.

The Memory Wall Problem

Modern parallel computer architectures are prone to system level bottlenecks that can limit performance for application processing. Historically, this phenomenon has been known as the "memory wall", where the rate of improvement in microprocessor performance far exceeds the rate of improvement in DRAM memory speed. Throughout the past decade the rate of growth of CPU core counts has created an increasing gap between CPU and memory performance (Figure 1) that hinders complex computing challenges.

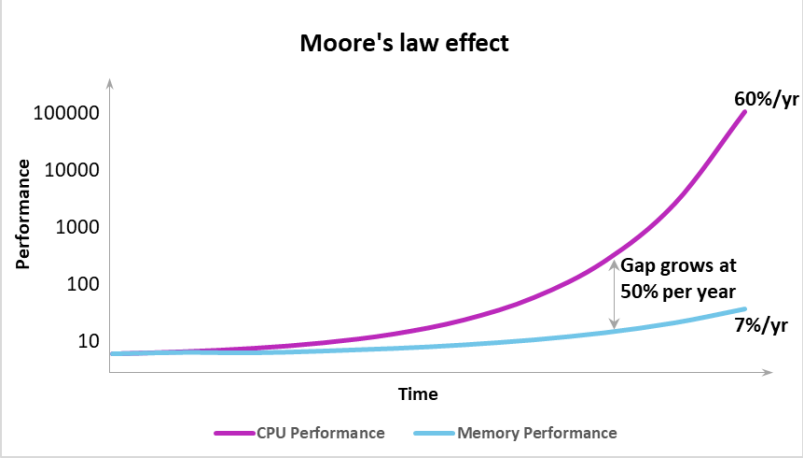

Figure 1 Historical performance gap between CPU and memory performance (Source: Synopsys)

[1] Compute Express Link is a trademark of the Compute Express Link Consortium

Adding processor cores is only a part of solving many application computing challenges. In most cases, it is vital to have the memory bandwidth to supply these processor cores with data. CPU vendors have attempted to mitigate the scaling gap issue through incremental advancements by adding more memory channels and increasing the data rate of those channels in new generation CPUs. New generations of DRAM technologies provide temporary relief with memory data rate evolution.

Table 1 shows the progression over the past decade of CPU core counts and DDR DRAM data rate increases along with the addition of more memory channels in years 2011, 2017, 2021 and 2023. However, even with theoretical memory data rates and more memory channels, it is challenging for memory bandwidth to keep pace with CPU core count growth and maintain 4 GB/s per core over time.

Table 1 Progression of CPU and memory and effect on bandwidth per core

Year	CPU core count	Theoretical DDR memory data rate (GB/s)	Memory channels	System memory bandwidth (GB/s)	Memory bandwidth per core (GB/s)
2010	8	10.7	2	21.3	2.7
2011 2012	10	10.7	4	42.7	4.3
2013	12	14.9	4	59.7	5.0
2014	18	14.9	4	59.7	3.3
2015	18	17.1	4	68.3	3.8
2016	22	19.2	4	76.8	3.5
2017 2018	28	21.3	6	128.0	4.6
2019	32	23.5	6	140.8	4.4
2020	48	25.6	6	153.6	3.2
2021 2022	64	25.6	8	204.8	3.2
2023	96	38.4	12	460.8	4.8
Future	256*	70.4*	12	844.8	3.3

(CPU vendors added more memory channels in years 2011, 2017, 2021 and 2023)
(Projected increases in core count and memory speeds from public statements and JEDEC specifications)*

The relationship between platform processing capabilities, as measured in CPU core count, and available memory capacity scaling is similarly challenged. As shown by the historical trend data in Figure 2, processor core counts have scaled quite rapidly while system memory capacity per core growth has steadily declined. The integration of the memory controller into CPUs has generally led to more direct and constrained processor to memory capacity ratios. Capacity may be increased by adding more DIMMs per channel. However, due to increased channel loading, adding more DIMMs per channel often necessitates a reduction in memory clock speed which reduces memory bandwidth, exacerbating the memory wall issue previously discussed.

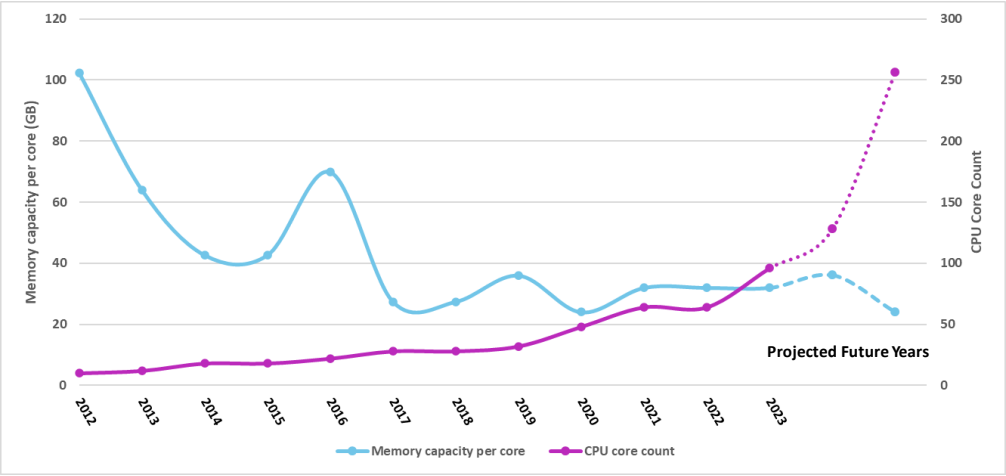

Figure 2 Historical trend lines for typical system memory capacity and CPU core count
(Source: Based on capacity and core counts from publicly available AMD and Intel datasheets, and public statements.)

The Problem of IT Resource Efficiency Maximization

Applications and services are being disaggregated into microservices to optimize available resources as workload demands ebb and flow. One of the key limiters of IT efficiency is that no one combination of infrastructure resources is ideal for all workloads. Workloads have a dynamic need for compute horsepower, memory, storage, latency, and IO bandwidth. As the nature and complexity of algorithms changes, workloads and services are optimized to be delivered over installed and unchanging public and private cloud hardware infrastructures.

IT workloads have historically been provisioned for peak demand. Architects and service planners project maximum resource requirements needed to provide a given level of service for some set period of time—and then ensuring that the proper peak-level (and some extra buffer) of compute, memory, storage and network resources are provided for that workload on a given server or rack of servers, including power to meet peak level demands. However, this often means that there is significant over-provisioning of resources as workload needs infrequently operate at peak levels. For much of recent history, the industry saw incredibly low (below 50% and often well below that[2]) overall data center resource utilization rates.

Over time, virtualization and cloud infrastructure provided significant capabilities to help conserve and even reclaim resources lost to overprovisioning and underutilization—through enhanced automation, workload migration and placement, and other techniques. Many believed that this would dramatically reduce the TAM for server infrastructure in the data center. This had quite the opposite effect and created demand for higher density CPU compute platforms as efficiency allowed savings in other areas such as power and operational management. Thus, reinforcing Jevons paradox “an increase in efficiency in resource will generate an increase in resource consumption rather than a decrease”.

The desire for increased flexibility and improved efficiency has never been higher, and there is constant industry discussion on enabling the composable “data center of the future.” One of the ideals sought from this

[2] TechTarget What is server virtualization? The ultimate guide

next-generation of data centers is more fine-grained control over the use of resources—including a re-thinking of how resources can be shared not only at a data center level, but also across racks and even within servers.

Evolution of the CXL Architected Data Center

CXL has emerged as a cost-effective, flexible, and scalable architectural solution that will shape the data center of the future. CXL will change how traditional rack and stack architecture of servers and fabric switches are deployed in the data center. Purpose-built servers that have dedicated fixed resources comprised of CPU, memory, network and storage components will give way to more flexible and scalable architectures. Servers in the rack, once interconnected to fixed resources for network, storage and compute to function as a solution, will be dynamically composed through software management infrastructure to meet the demands of modern and emerging workloads such as AI and deep learning.

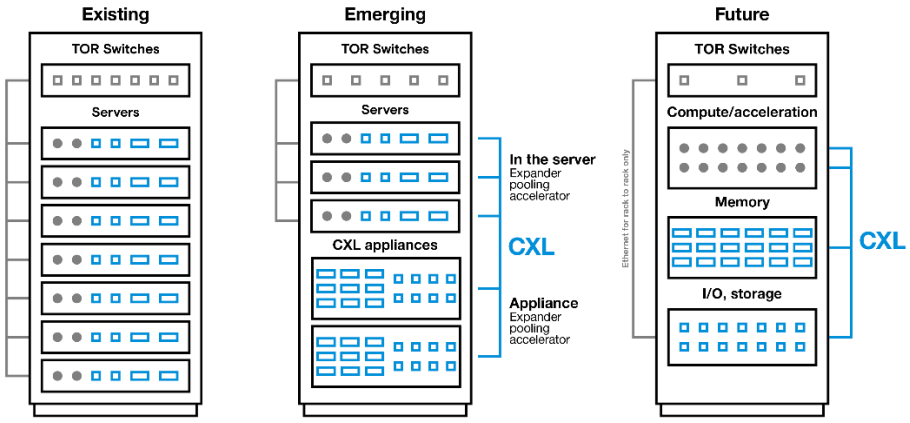

Figure 3 Data center evolution from traditional rack and stack to full composability (Source: Marvell)

The industry has focused on the potential that can be unlocked through CXL devices with memory access, as CXL-attached memory devices. Memory attached points provide high-capacity memory expansion and can be utilized for intensive server workloads with increased memory bandwidth, low latency and memory coherency for heterogeneous compute/processing and achieve tiering in memory infrastructure. Memory tiering will be introduced in much in the same way that tiering was introduced to storage over the past several decades and will eventually include direct-attached memory expansion, memory pooling and memory sharing.

The data center will progress to be more memory-centric with the ability to dynamically compose servers with high Terabyte (TB)-plus memory pools, enabling more applications to run in memory. Storage-class memory becomes the new primary active data storage tier, with NAND and disk drives being used for warm and inactive data to be shared among multiple hosts.

Eventually the data center will migrate towards complete disaggregation of all server elements including compute, memory, network and storage. Container and microservices that are massively deployed will drive dynamic configuration of the underlying resources they need for an optimized solution with balanced compute and memory ratios and no performance penalty. With CXL, the deployment of services and the underlying hardware used in on-demand configurations will appear seamless and quick with the advent of composability management software creating greater efficiencies in the as-a-service model in a heterogeneous environment.

How CXL Addresses the Memory Wall Problem

CXL protocol properties for memory-device cohesion and coherency will address the “memory wall” problem by enabling expansion of memory beyond server DIMM slots. CXL memory expansion serves as a two-prong approach by adding bandwidth to overcome the “memory wall” problem and adding capacity for data-intensive workloads for CXL enabled servers.

For typical workloads it is important to maintain bandwidth per CPU core for ideal efficiency. As the core count rapidly increases, bandwidth falls short (see Table 1). Direct attach CXL memory expansion allows server platforms to scale-up and close the gap for additional bandwidth to maintain balance.

Another factor to consider is the diminishing capacity per core as core counts increase. The workload demands of applications keep growing in the amount of capacity needed to analyze collected data fast, and present results in useful business insights. These high value workloads (i.e., machine learning, NLP, computer vision, recommender systems, in-memory databases, etc.) can be economically addressed with greater levels of memory per system. CXL memory modules can be plugged directly into the server, providing the processors more bandwidth and more capacity beyond the direct-attach memory channels, and do so at a latency comparable to the NUMA link between processors in a dual socket server.

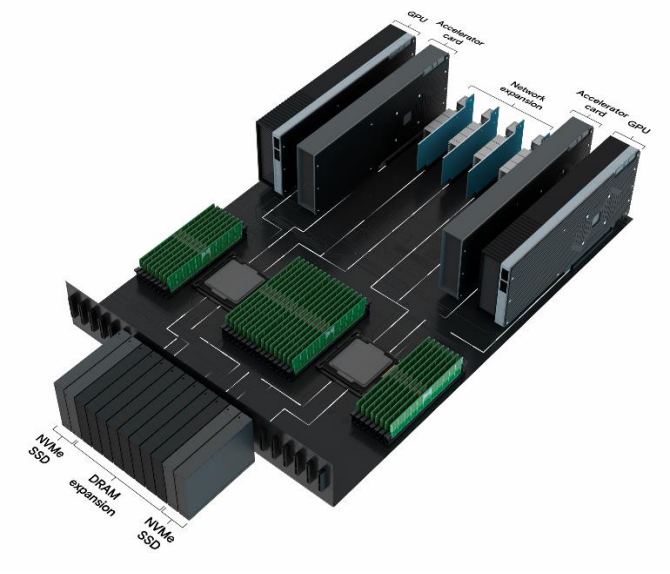

Figure 4 CXL enabled server with CXL memory modules

How CXL Addresses IT Efficiency and Sustainability Challenges

A variety of workloads across application verticals are highly sensitive to computing operations, memory capacity, bandwidth, and latency. Applications that run in the cloud, enterprise, or edge data center on traditional rack servers must meet service level agreements (SLAs). A common approach is to distribute these types of application workloads over multiple systems. Building an IT infrastructure does not always follow a simple rule of thumb to achieve a system balance between compute and device resources. Balancing these resources is dependent upon workloads, which can be compute bound, memory bound, or IO bound.

Initial deployments of CXL based systems provide expansion options for performance and capacity to match the scaling up of compute resources based on workload demands. Memory, storage, network, and

accelerators become interchangeable modules as form factors and connectivity are standardized and servers can be composed as necessary for the workload demands. This approach allows server manufacturers, including cloud providers, to reduce the number of server SKUs they need to develop and maintain to address the myriad applications of their customer base. It also helps IT administrators by right sizing a server with adequate resources to reduce the quantity of servers a single workload must be distributed across to drive improved efficiency and performance.

Over time, the value of CXL architectures will be expanded to the rack, enabling composability. Composability is the ability to more flexibly provision the ratio of memory to compute resources in one or more servers supporting one or more workloads. Balancing resources can be achieved through memory expansion, memory pooling or memory sharing. At the rack, a scale-out approach allows pools of resources (compute, networking, memory, storage, and IO) to be dynamically allocated with seamless integration as required by the application. Compute, memory, network, and storage are assigned to an application or microservice as the instance is brought online through composition management software using native device level discovery within the rack. During peak demands an application may be assigned additional resources on-the-fly to meet SLAs. When the application workload demand moderates the additional resources can be freed and reassigned to other services. Resource sharing or pooling provides higher utilization without the need to overprovision systems, it also means higher performance, reduced software stack complexity, and lower overall system cost.

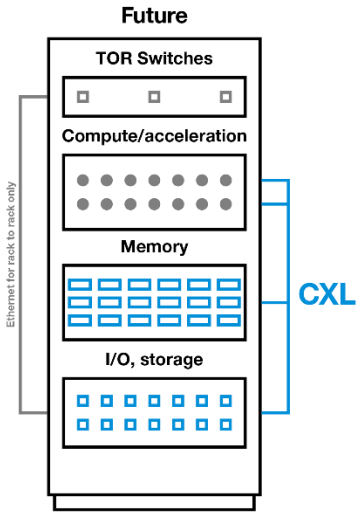

Figure 5 Composable architecture extended to the rack

Of course, there will always be limits to how much shared (and therefore composable) resources will be aggregated to any given workload as customers have security and fault tolerance factors to consider along with workload efficiency and utilization concerns. Although memory pools create the resources necessary to account for peak levels, 85% of organizations require uptime of 99.99% to meet SLAs[3], which must be accounted for in memory pools within the rack, driving some amount of oversubscription even when pooled CXL attached memory is adopted. Furthermore, while memory pooling may alleviate near term memory overprovisioning concerns, careful consideration must be given to memory pool expansion failures to avoid server failures throughout the rack, driving redundancy to avoid downtime. One approach gaining favor is the

[3] DataCenter Knowledge: Ten Ways To Ensure Maximum Data Center Uptime

creation of resource zones, or pods, that provide a balance of effective use of shared resources with the need to minimize the impact of disruptions in service, and to provide the proper security and compliance capabilities.

One of the biggest initiatives for data centers is the push to net-zero emissions. Efficiency is a crucial variable in the sustainability equation for data centers. Just like server virtualization, expanding device sharing and pooling of resources will reduce over provisioning in general, but on a larger scale, in the data center. Converting dedicated device resources to shared pool resources and stage them to be dynamically allocated reduces power at the compute node. Not only a reduction in power consumed per compute node, but also improving airflow and thermals for more efficient cooling in the rack, reducing the demand on HVAC systems to further reduce the power consumption of the data center.

CXL's Impact on DRAM Bit Demand Growth

Now let us address how we expect CXL will impact DRAM bit demand growth.

The net impact of CXL enabled pooling and CXL enabled memory bandwidth expansion on bit demand growth will be positive. All in all, we expect CXL to help sustain high 20s% data center bit growth in the short to medium term.

The near term CXL memory market is dependent upon how quickly CXL enabled server platforms become available to the broad industry. Since CXL memory is an emerging market, the memory growth on CXL will be very fast but won't have a huge impact on the total DRAM market until 2026. The Yole Intelligence market research group predicts that DRAM bit demand on CXL will grow to approach 100 exabits by 2028. Yole Intelligence forecasts that CXL bits will make up 31% of total DRAM bits in servers in 2028[4].

A recent Carnegie Mellon / Microsoft paper discusses how pooling affects CXL TCO savings. The paper proposes a CXL based pooling solution that results in a TCO savings of 4-5% by reducing the memory requirements for a given set of hyperscale workloads by 9-10%. Data center DRAM bit growth compound annual grown rate (CAGR) remains in the 20s% range, which includes impact from CXL. Even with the addition of memory pooling this will have a small impact on overall data center DRAM bit growth. Rough math to simply calculate this effect is achieved by multiplying the 9-10% reduction by the expected CAGR of 20-30%. The theoretical worst-case calculation would suggest a reduction of 2-3 percentage points due to pooling. Of course, this theoretical scenario is infeasible, because pooling comes with latency tradeoffs and software optimization requirements, and pooling is not applicable to all workloads. Secondly, the applicability and extension of pooling is limited by the need for fault tolerance with some level of redundancy, and the risk of a memory pool failure cascading across multiple hosted servers. Finally, any impact would be muted by a gradual CXL adoption period. CXL enabled pooling will not be able to address the current non-CXL data center install base.

While the CXL enabled pooling opportunity may reduce the peak memory required for a given set of workloads, CXL enabled memory bandwidth will increase the peak memory capable of addressing the growing set of memory-intensive and AI workloads that are presently economically infeasible with alternatives to CXL. These CXL alternatives, such as Through Silicon Via (TSV)-based solutions are expensive and face thermal density and signal integrity challenges. **The aggregate of opportunities from memory expansion and memory pooling result in a net-positive for DRAM bit growth.**

[4] Yole Intelligence "DRAM Market Monitor Q1 2023"

CXL Impact on Industry Revenue TAM and Micron's Financial Model

Revenue TAM growth for memory depends on bits and price, and price depends on supply and demand balance. CXL is an interconnect solution, and its technology adoption does not itself add supply into the market. CXL itself should not be a disruptive factor for industry supply and demand and pricing is expected to enable TAM growth. In some configurations, memory connected to CXL interfaces versus the standard memory slots is more cost efficient, enabling server systems to be built and deployed at a scale that would otherwise exceed budget targets. The first use cases for CXL revolve around memory expansion for single-host configurations. Memory expansion restores balance between compute and memory for memory bound workloads that would otherwise be distributed among multiple servers and consolidates memory from those servers to CXL expansion slots. New servers capable of supporting CXL 1.1+ will show up in the market this year in 2023 but primarily used as proof-of-concept for CXL emerging memory solutions. Real deployments will start in late 2024 when CXL 2.0 enabled servers are available with more memory expansion options and marks the beginning of higher amount of average DRAM content in the server. We expect this will be the start of ramping revenue behind the CXL interface and project it to be an estimated $2 billion market in 2025.

Resource expansion is the first step in the CXL evolution before moving to complete composability and memory pooling, which we currently expect to begin to ramp in 2026. In 2026 many new servers will support CXL 3.0, and the server market is forecasted to grow to around 21 million units[5], providing the necessary support for disaggregation. Factors that will influence the adoption rate for memory pooling include CXL switches, and software that can handle tiered memory pools and allocation across multiple hosts to minimize latency. Hyperscalers will be early adopters for memory pool expansion in the near term. It is likely that they will have an even split of growth between single-host memory expansion and memory pooling within the rack. We, as well as industry analyst Yole Intelligence, expect CXL attached memory will be a greater than $20 billion market by 2030 within an estimated $100 billion data center market for memory, with most of the growth after 2025.

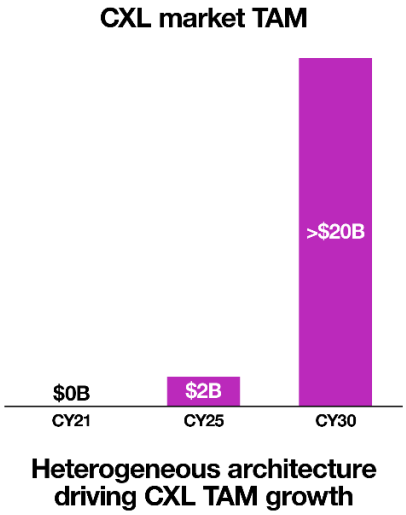

Figure 6 Micron - CXL market TAM projection

[5] Wheeler's Network: CXL Chip Market Poised for Rapid Growth

Our view and expectations of CXL's impact are incorporated into our long-term modeling and our cross-cycle financial model, so the anticipation of CXL technology adoption should not change investor expectations about our financial performance.

Conclusion

CXL provides the necessary architecture to bring balance to the “memory wall” problem and provides a new vector for achieving economical memory solutions through memory expansion. Additionally, CXL flexible and scalable architecture provides higher utilization and operational efficiency of compute and memory resources to scale-up or scale-out resources based on workload demands. CXL attached memory provides tremendous opportunity for growth in new areas for tiered memory storage and enabling memory scaling independent of CPU cores. CXL will help sustain a higher rate of DRAM bit growth than we would see without it. In other words, we don’t expect CXL to cause an acceleration in DRAM bit growth – but it is a net positive for DRAM growth.

Micron’s commitment to CXL technology enables customers and suppliers to drive the ecosystem for memory innovation solutions. To learn more on how Micron is enabling next-generation data center innovation, visit micron.com/solutions/server.

micron.com