



AI LEADERSHIP FORUM

CXL: From Standard to Industry Transformation

- Anil Godbole, Intel/CXL Consortium



Contents

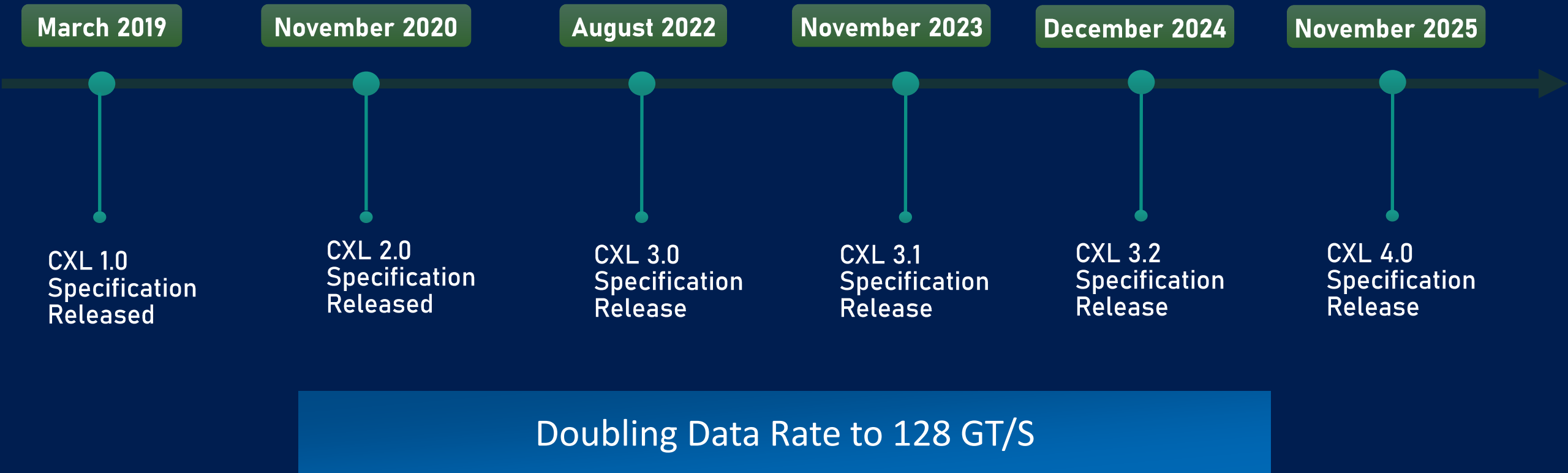
CXL Consortium Update

Popular CXL Use cases

Mem Pooling Developments

Summary & Call to Action

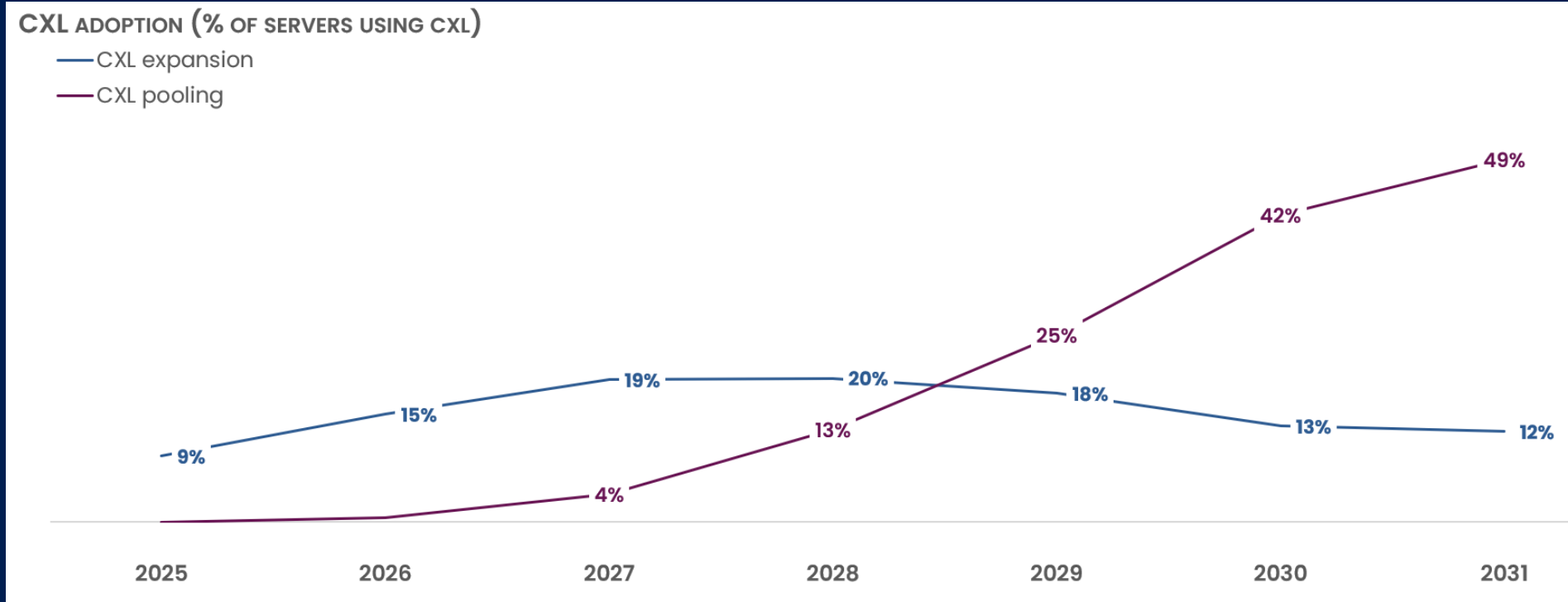
Specification Release Calendar



Key Feature Enhancements for CXL 4.0

- Doubles the bandwidth to **128GTs with zero added latency**
 - Enables rapid data movement between CXL devices, directly improving system performance
 - Maintains previously enabled CXL 3.x protocol enhancements with the 256B Flit format
- CXL bundled port
 - Ability to combine device ports between Host and CXL accelerators (Type 1/2 devices) to increase bandwidth of the connection
- Memory RAS enhancements
 - Advanced CVME (Corrected Volatile Memory Error) controls for granular event reporting
 - Mechanisms for host-initiated Post Package Repair (PPR)
 - Flexible memory sparing operations at boot, collectively improving reliability, error visibility, and maintenance efficiency

Memory Pooling Next Wave of Adoption



- Shipments of high-volume Datacenter CPUs (Intel 6500/6700) have only been recently launched
- Recent announcements from Hyperscalers on use of CXL expansion memory, incl. DDR4 reuse
- AI market is consuming all available memory → Rise in memory prices → Interest in using/reusing DDR4
- CXL-based Memory Pooling looks promising due to interest from AI workloads

Popular CXL Use Cases

Value Proposition of CXL Memory

Higher Memory Capacity

- Improve processor performance
 - Reduced SSD accesses
- Benefiting workloads
 - Virtualized Servers
 - In-memory Databases
 - AI / ML
 - HPC
 - Media (CDN, Video 8K)
 - Medical (Genomics)

Higher Memory Bandwidth

- Improve bandwidth with processor address interleaving
- Benefiting workloads
 - AI/ML
 - HPC
 - Non-relational Databases
 - Virtualized servers

Lower Memory TCO

- Avoid expensive 3DS DIMMs
 - Use standard DIMM capacities for native & CXL memory
 - Use lower-cost memory media on CXL
 - DDR4
- Memory Pooling
 - For optimal provision of local DRAM on servers

Microsoft Azure Adopts CXL Memory

RUNNING SAP APPLICATIONS ON THE MICROSOFT PLATFORM 2 MIN READ

Azure delivers the first cloud VM with Intel Xeon 6 and CXL memory - now in Private Preview



Elyse Ge Hylander

Nov 18, 2025



Intel released their new Intel Xeon 6 6500/6700 series processor with P-cores this year. Intel Xeon 6 processors provide performance and scalability by delivering outstanding performance for transactional and analytical workloads and provide scale-up capacities of up to 64TB of memory.

In addition, Intel Xeon 6 supports the new [Compute Express Link \(CXL\)](#) standard that enables memory expansion to **accommodate larger data sets in a cost-effective manner**. CXL Flat Memory Mode is a unique Intel Xeon 6 capability that enhances the ability to right-size the compute-to-memory ratio and improve scalability without sacrificing performance. This enhanced ability can help run SAP S/4HANA more efficiently and help enable greater flexibility for configurations so they can better align with business needs and improve the total cost of ownership.

In collaboration with SAP and Intel, Microsoft is delighted to announce private preview of CXL technology on Azure M-series family of VMs. We believe that, when combined with advancements in the new Intel Xeon 6 processors, it can tackle the challenges of managing the growing volume of data in SAP software, meet the increased demand for faster compute performance and reduce overall TCO.

Stefan Bäuerle, SVP, Head of BTP, HANA & Persistency at SAP noted:

"Intel Xeon 6 helps deliver system scalability to support the growing demand for high-performance computing and growing database capacity among SAP customers."

Elyse Ge Hylander, Senior Director, Azure SAP Compute stated:

"At Microsoft, we are continually exploring new technological innovations to improve our customer experience. We are thrilled about the potential of Intel's new Xeon 6 processors with CXL and Flat Memory Mode. This is a big step forward to deliver the next-level performance, reliability, and scalability to meet the growing demands of our customers."

Bill Pearson, Vice President of Data Center and Artificial Intelligence at Intel states:

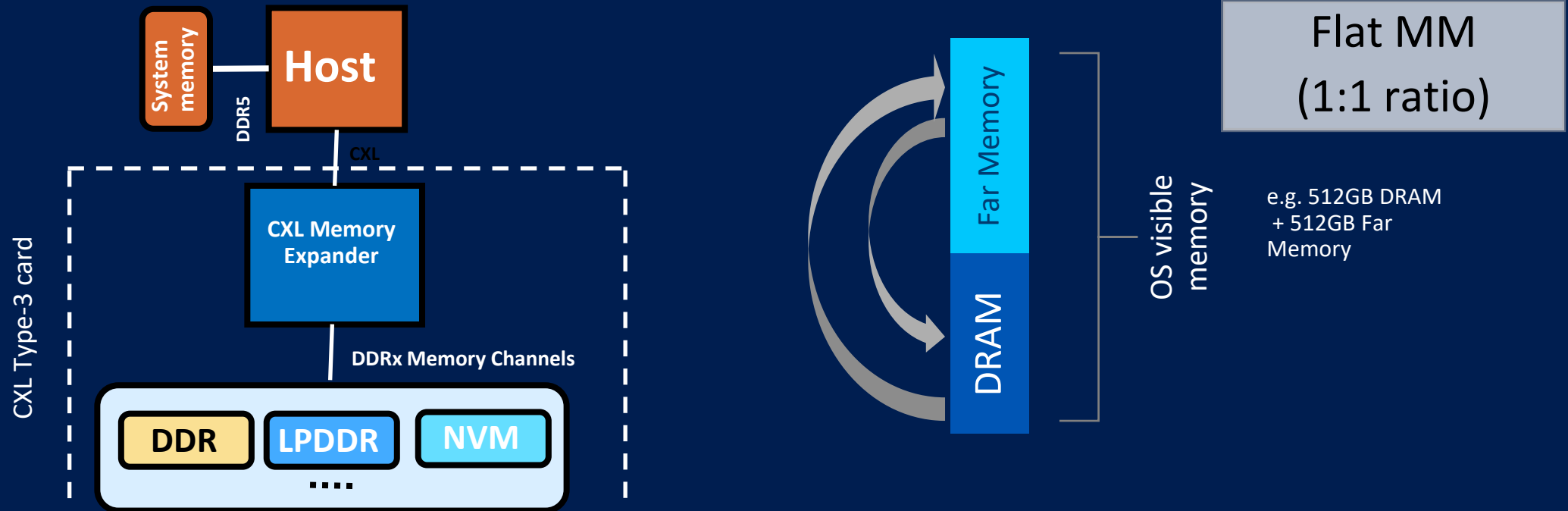
"Intel Xeon 6 represents a significant advancement for Intel, opening up exciting business opportunities to strengthen our collaboration with Microsoft Azure and SAP. The innovative instance architecture featuring CXL Flat Memory Mode is designed to enhance cost efficiency and performance optimization for SAP software and SAP customers."

Blog from Microsoft Azure

[\(Link\)](#)

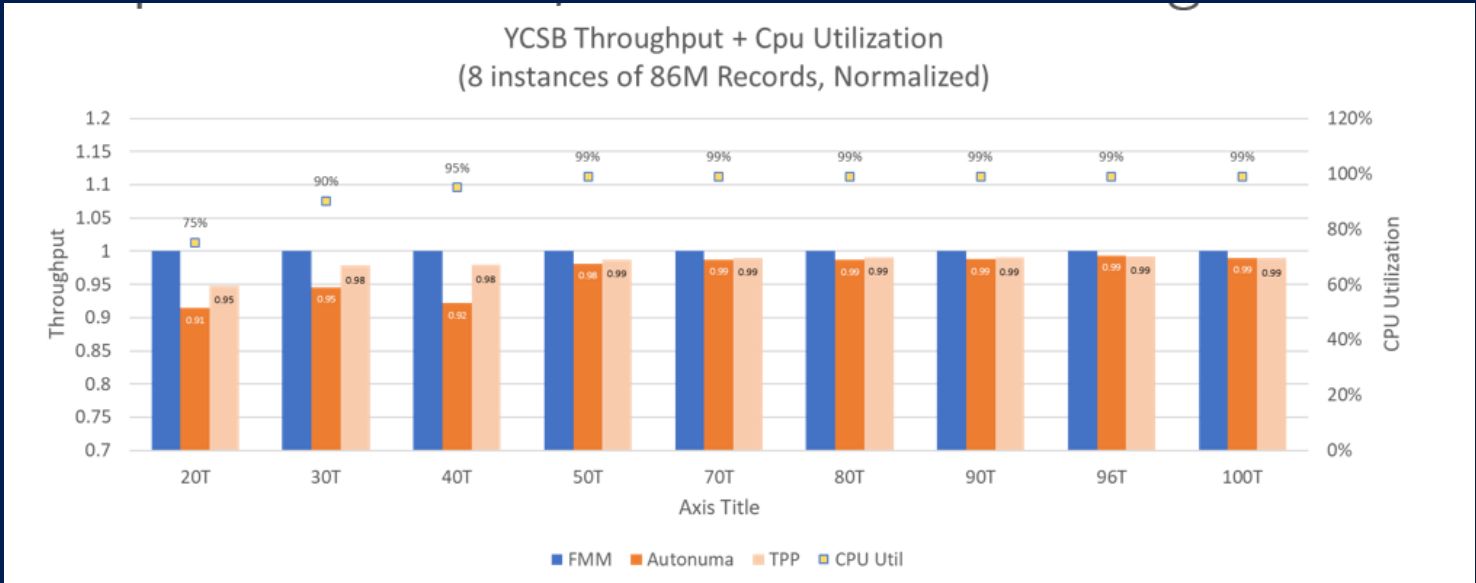
H/W-based CXL Memory Expansion

Intel Flat Memory Mode



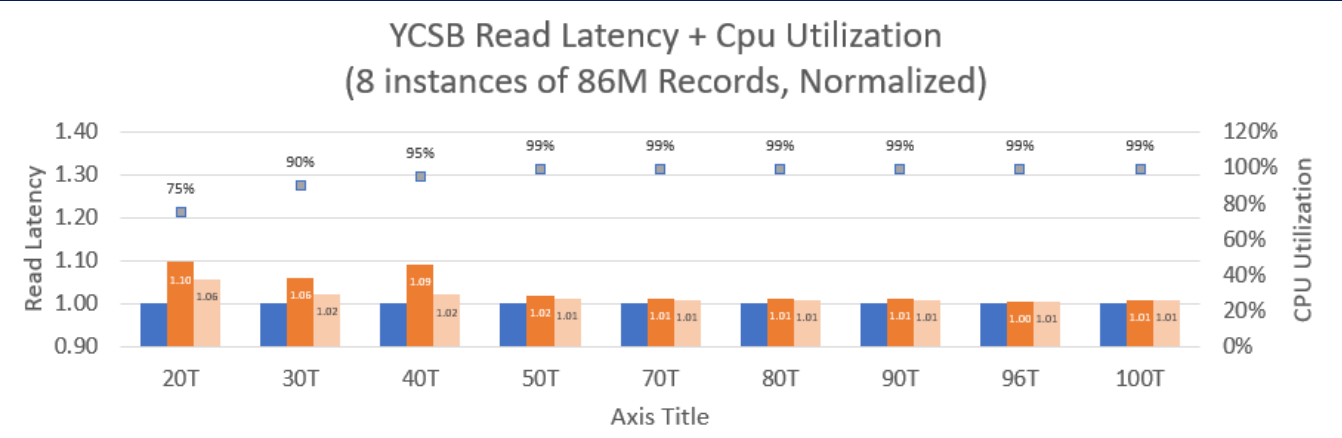
- Both DRAM and far memory exposed to OS as combined physical memory – **One memory tier**
- No Hot/cold page telemetry required and hence no dedicated CPU resources
- Hot data is swapped into DRAM – **one cacheline at a time, not a whole 4KB page**
- Performance very good due to 1:1 Near/Far memory ratios

Intel Flat Memory Mode Performance: (MongoDB) Comparison with S/W-controlled Tiering



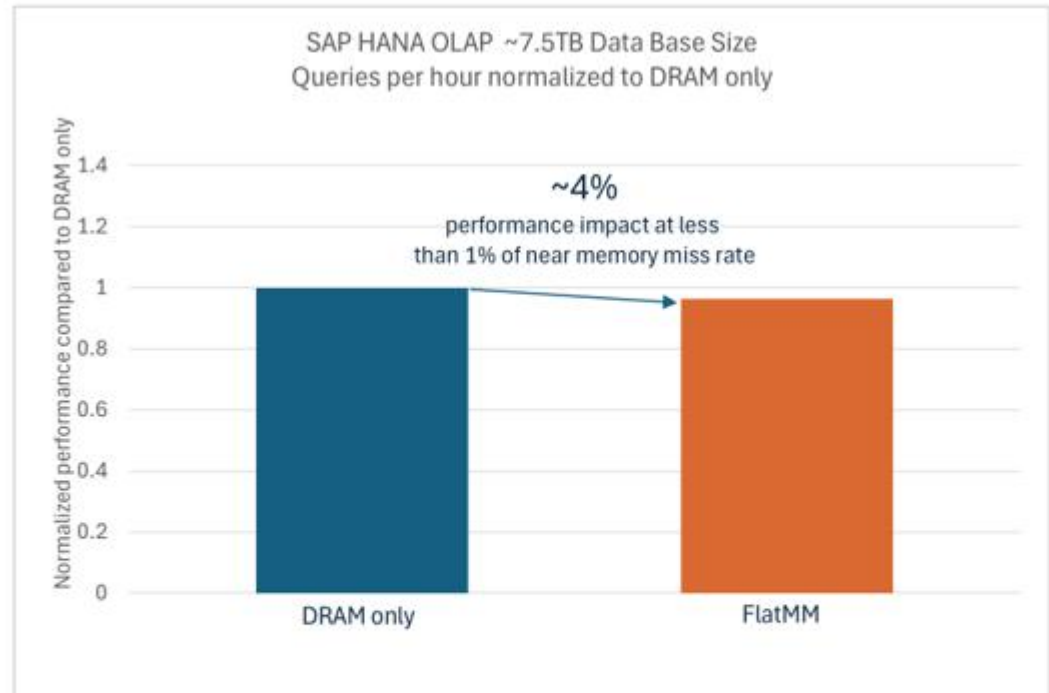
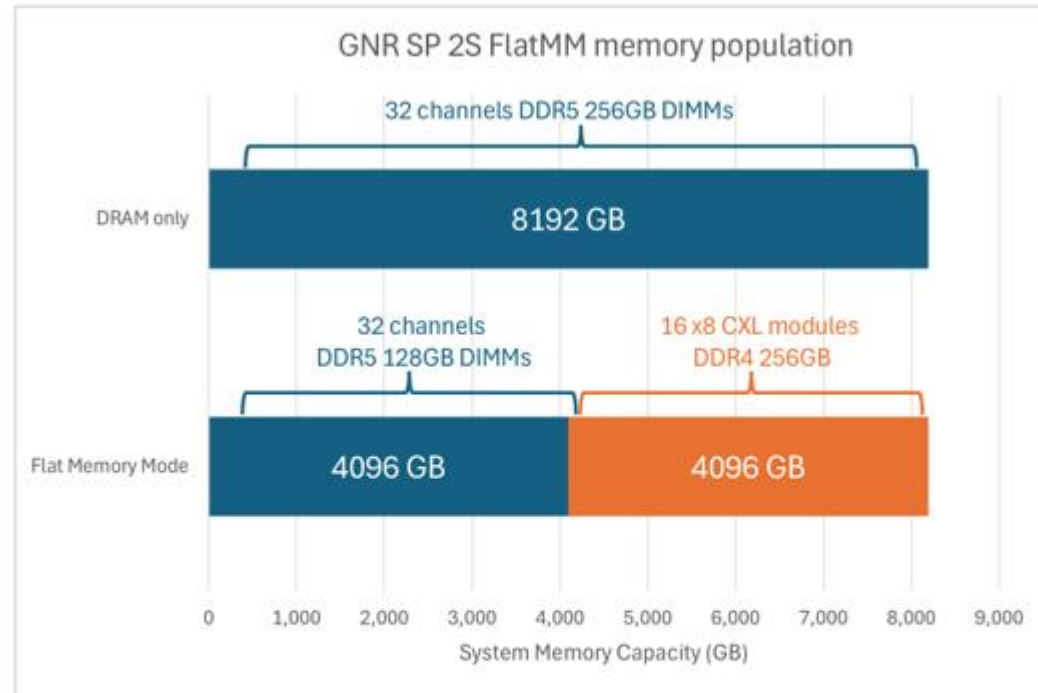
Up to 10% Better Query Throughput
under optimal CPU utilization

Up to 10% Lower Latency under optimal
CPU utilization



Intel Flat Memory Mode Performance Example: SAP HANA running OLAP

40%+ TCO Savings with DDR4 RDIMM Reuse

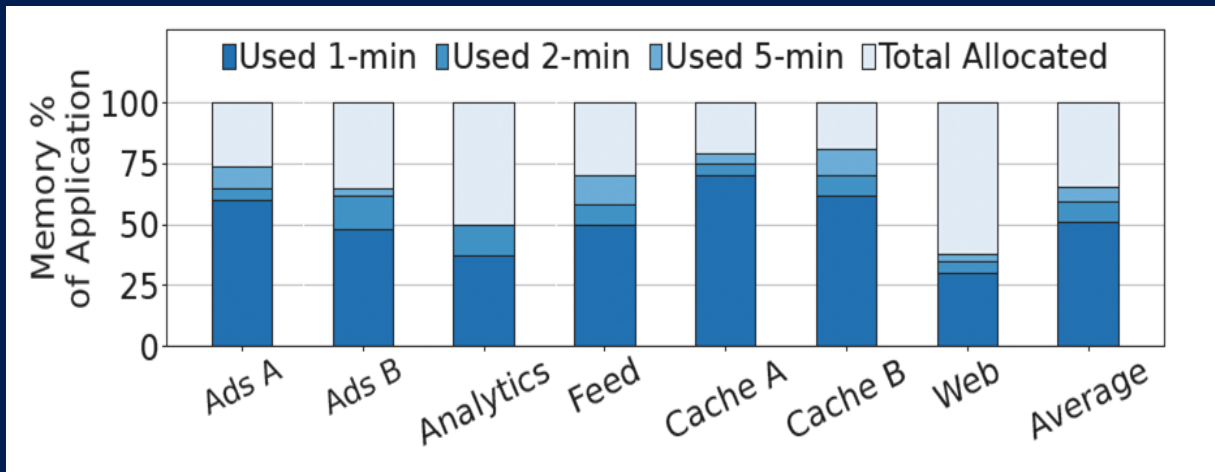


Endorsement by SAP Inc. / Microsoft Azure

<https://sapinsider.org/how-intel-xeon-6-and-cxl-are-the-new-secret-weapons-to-tame-sap-data>

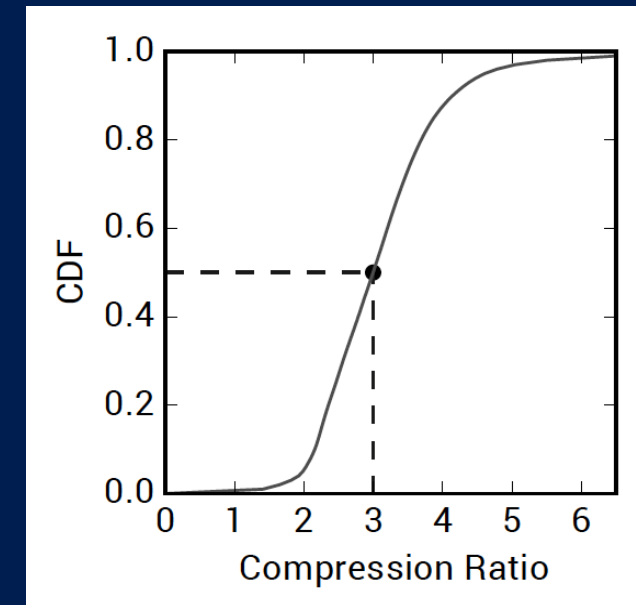
Paper at ACM: <https://dl.acm.org/doi/10.1145/3736227.3736230>

Google (GCP) / Meta: CXL-based Swap



Half of memory hasn't been used in the past minute*

↑
**Lots of 'Cold pages' in main memory –
Make room by demoting to CXL tier
'(Page)Faultless' Swap action**



**Cold data can compress at a 3:1 ratio;
More like 2:1 once we account for
incompressible pages**

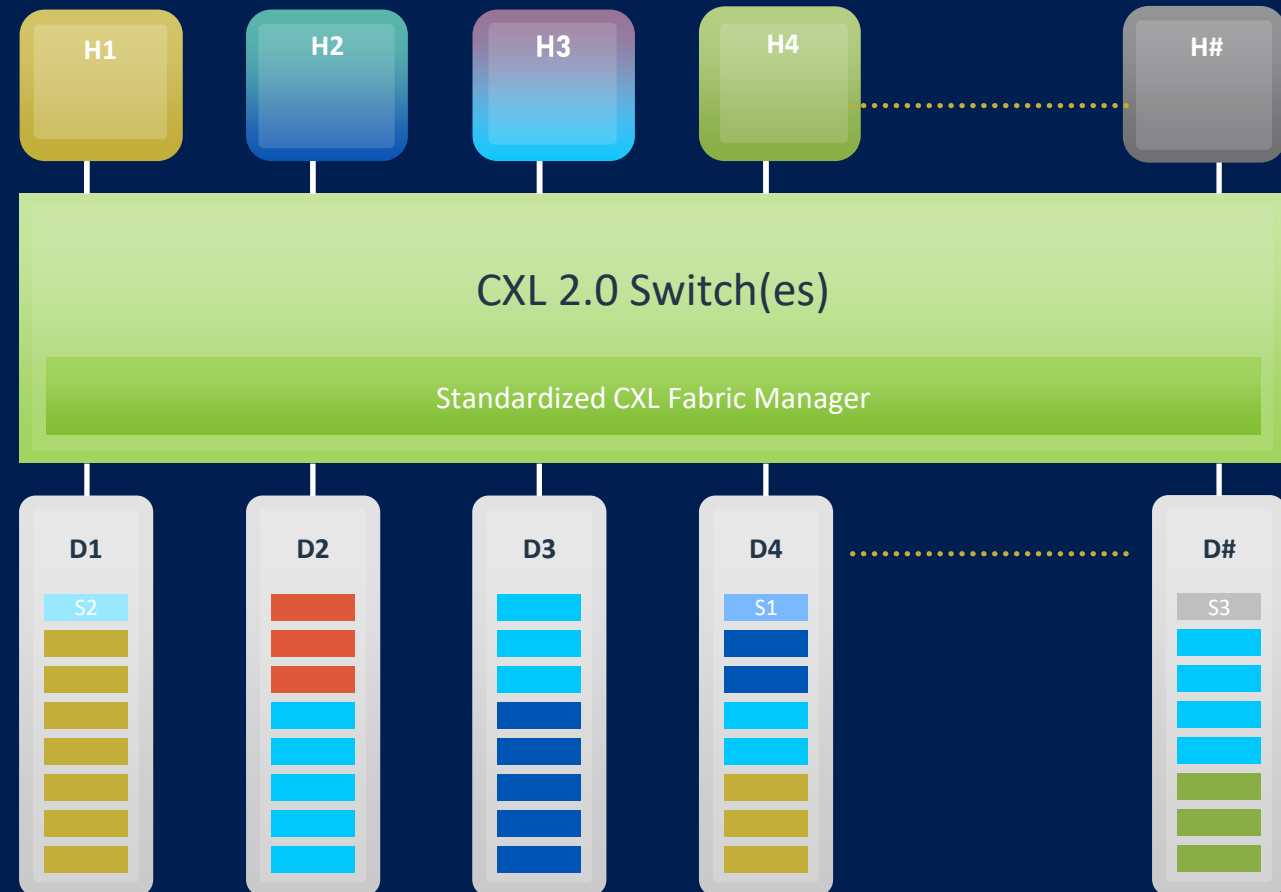
↑
**Compress 'cold' pages to
double capacity in CXL tier**

*Source: Making Memories at HyperScale with CXL; Webinar presented by Google/Meta

Memory Pooling Developments

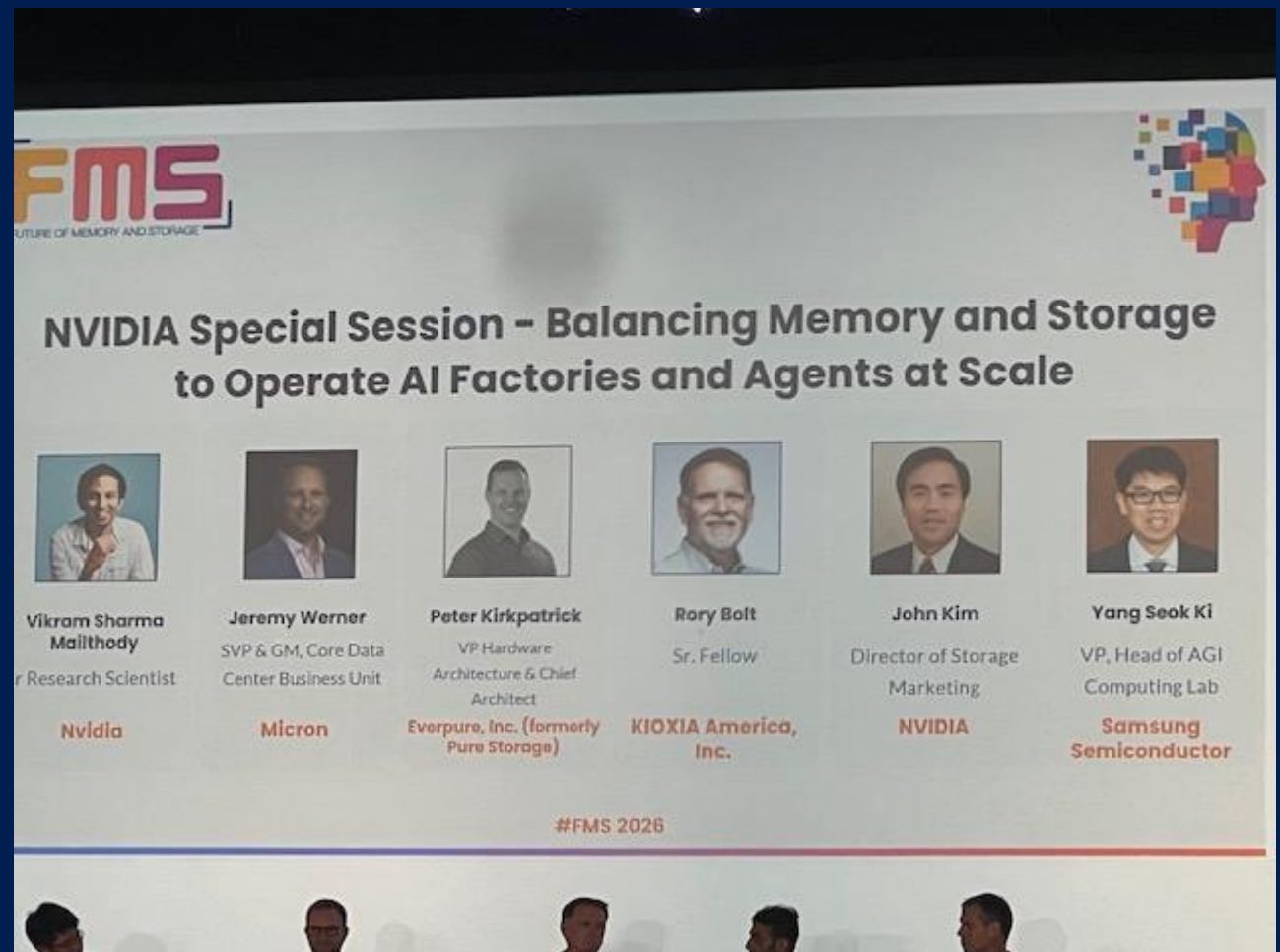
Industry Interest in CXL Memory Pooling

- Introduced in CXL 2.0 / CXL 3.0
- Allows data sharing by 'reference'
 - Lower latency compared to shuffling over networks
- Popular with AI, Distributed data-bases, Big-data & HPC industry & Academia
- Recent Memory pooling Demonstrations
 - Alibaba: Distributed Database acceleration
 - Micron: Vectorized DB for RAG (Retrieval Augmented Generation)
 - KV Cache repository with Nvidia Dynamo



CXL MemPooling for AI

- Gaining traction as a big performance enabler for AI w/l
- Agentic AI: Next big use case
 - Multiple GPUs sharding big LLM (Model wts inTBs)
 - Hyper increases in KV\$ sizes
 - Needs big data sizes & locality
 - Use SSD pools / fast connectivity
 - **New alternative: CXL MemPool**
 - Allows for big data footprints
 - Naturally co-located data
 - Better from latency/power pov



Endorsements by Experts from AI leader companies

Several demos by Hynix/Samsung in their booths

Summary & Call to Action

- CXL Adoption continues to grow
 - Big recent announcements from major CSPs in US, PRC
- Memory Pooling/Sharing use cases in AI expected to be a major demand driver
 - Several demos with significant perf improvements over traditional data-shuffling approaches
 - Big endorsement from Nvidia at FMS'26 acknowledging CXL Memory Pool will be the best solution for Agentic AI era

Call to Action:

- Urging all CXL ecosystem partners to continue evangelizing CXL use cases with their customers/partners
- Work towards a seamless Memory Pooling infrastructure