

Memory Fabric Technology Landscape

Charles Fan
CEO and Co-founder
MemVerge

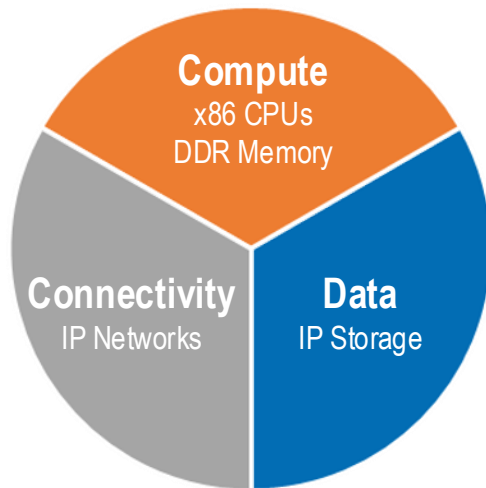


OCT 15-17, 2024
SAN JOSE, CA

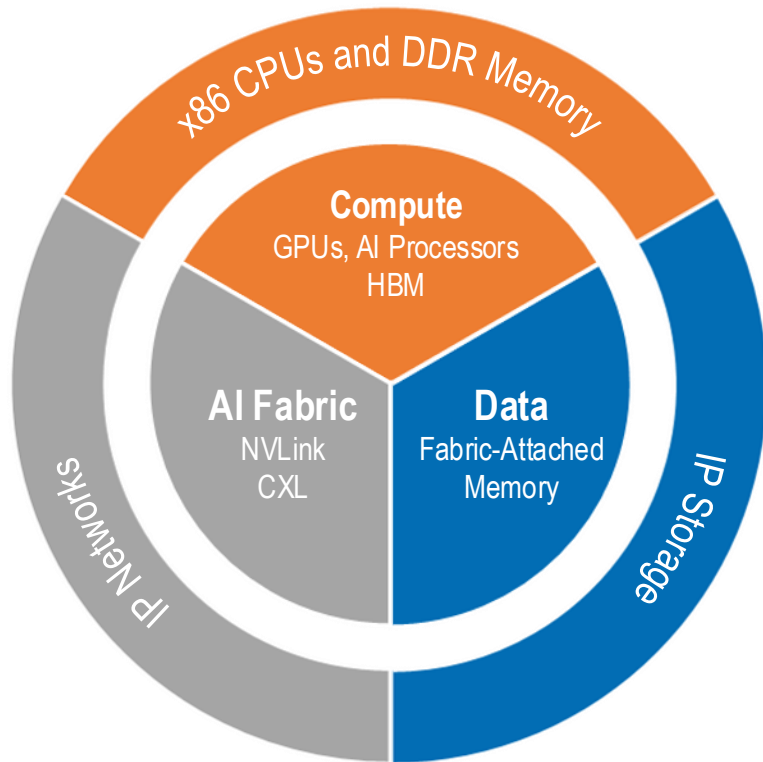


The Emergence of the AI Computer

x86 Era



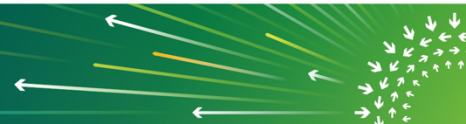
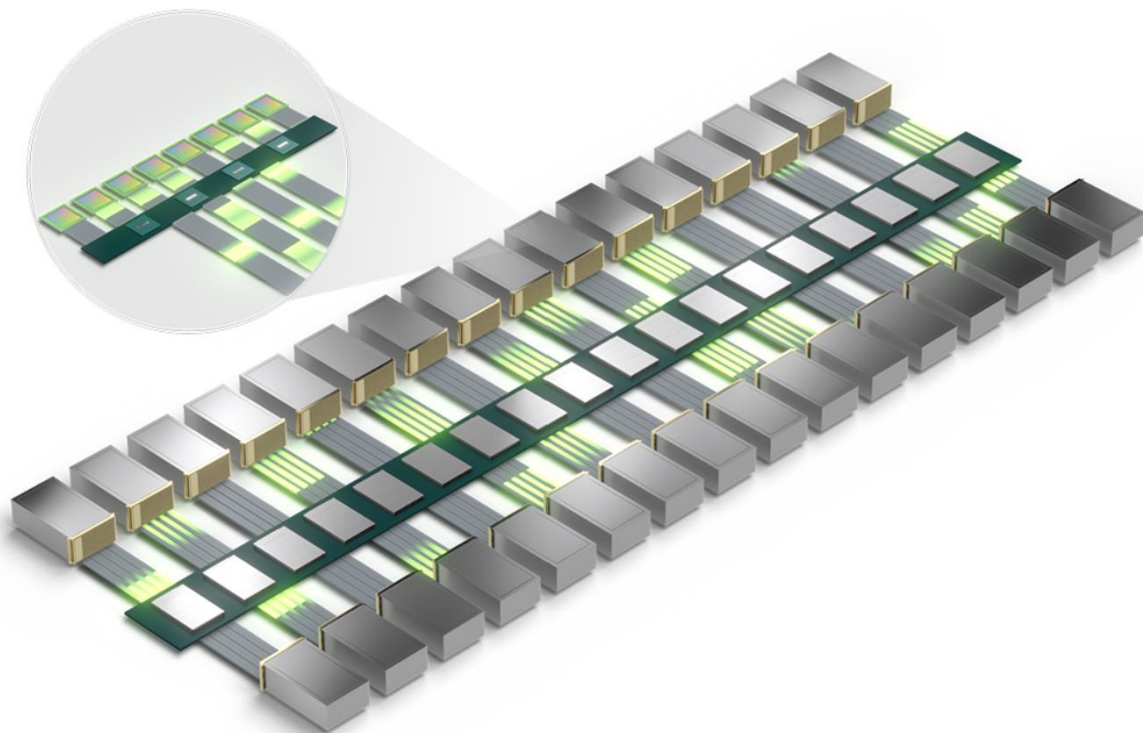
AI Era



NVIDIA NVLink

GPU interconnection within and between AI servers

- H100 @900 GB/s, B200 @1.8 TB/s
- With NVLink Switch System, GB200 NVL72 connects 36 Grace CPUs and 72 Blackwell GPUs and provides total bandwidth of 130 TB/s
- Proprietary to NVIDIA

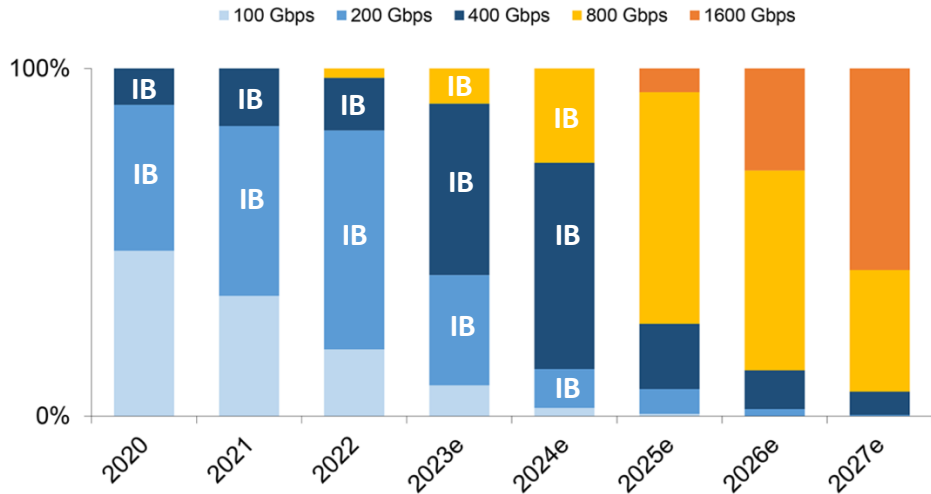


InfiniBand

Backend AI Network

- AI workloads require a new backend infrastructure buildout
- Backend spending forecast doubling to almost \$80B over the next five years
- InfiniBand is currently dominating
- Significant improvements on the Ethernet technology side

Migration to High Speeds in AI Clusters (AI Backend Networks)



*Includes both Ethernet and InfiniBand

* Source: Dell'Oro Group AI Networks Report December 2023



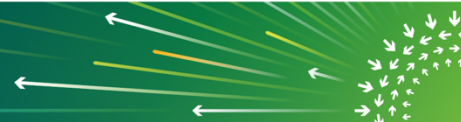
<https://www.delloro.com/news/ai-back-end-networks-to-drive-80-b-of-data-center-switch-spending-over-the-next-five-years/>

<https://www.delloro.com/exploring-the-data-center-switch-and-ai-networks-markets-landscape-in-2024/>



2024

FROM IDEAS TO IMPACT



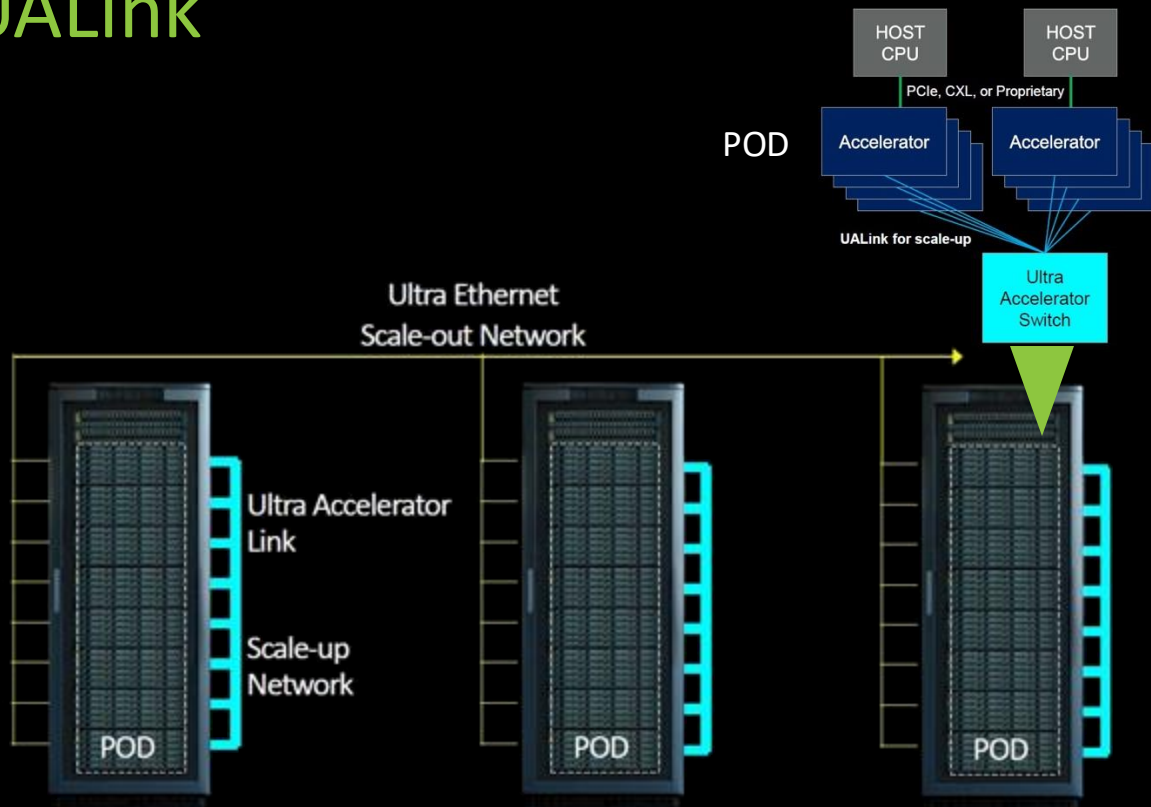
Ultra Ethernet & UALink

Ultra Ethernet

- High bandwidth multi-pathing
- 800 Gbps & 1.6 Tbps

UALink

- Industry group proposed alternative to NVLink

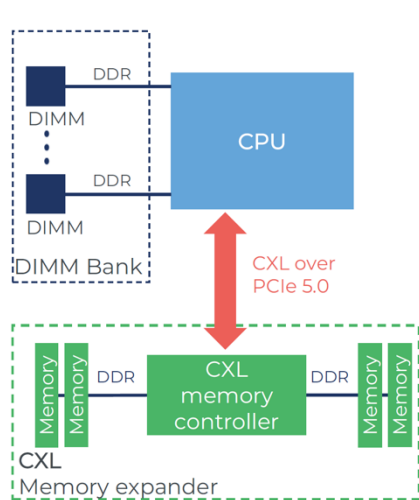


Compute Express Link[®] (CXL[®])

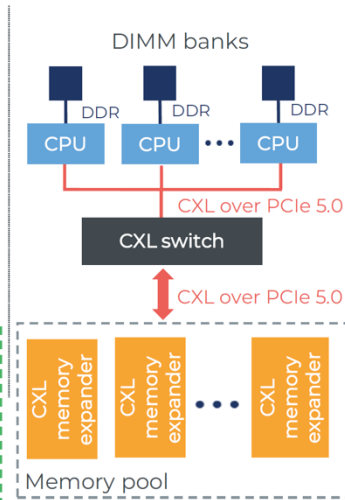
- Industry standard memory fabric on top of PCIe
- In-server memory expansion production deployments start in 2025
- Composable and shareable Fabric-attached memory forthcoming

Memory expansion, Pooling, Disaggregation using CXL

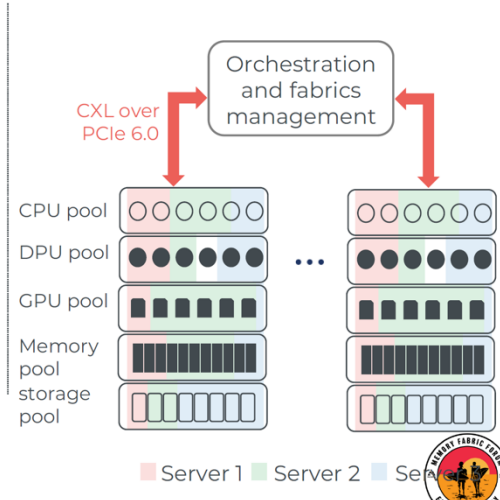
CXL 1.1 In-server memory expansion (server level)



CXL 2.0 Memory pooling (rack level)



CXL 3.1 Fully disaggregation and composability of resources (rack-to-rack)



Source: CXL Consortium
CXL: Compute Express Link

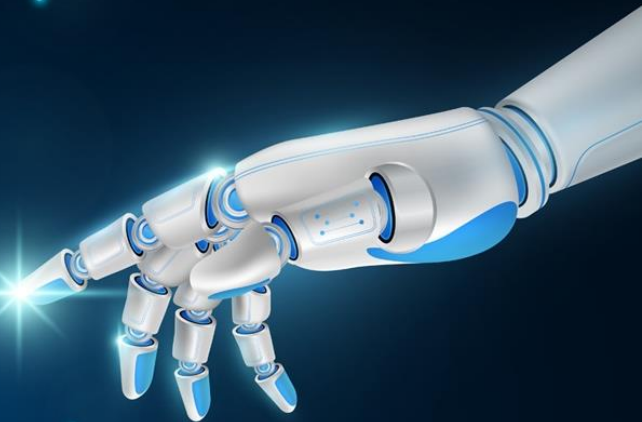
Memory Fabric Forum | www.yolegroup.com | 9





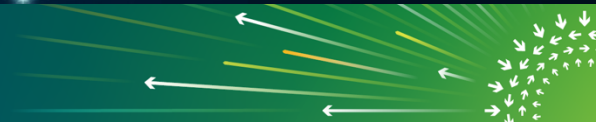
Software for Memory Fabric

in the era of Generative AI



2024

FROM IDEAS TO IMPACT



CXL Memory Expansion

Intelligent Tiering

- **Latency Policy** intelligently manages data placement across heterogeneous memory devices to optimize performance based on the “temperature” of memory pages, or how frequently they are accessed
- **Bandwidth Policy** utilizes the available bandwidth from all DRAM and CXL memory devices with a user-selectable ratio of DRAM to CXL to maintain a balance between bandwidth and latency

Latency QoS Policy

Active QoS Policy: No Policy

QoS Settings Enabled

Policies

Latency Bandwidth

Latency policy activated.
The Latency policy moves frequently used memory pages (Hot Pages) to low-latency NUMA Nodes and least frequently used pages (Cold Pages) to high-latency NUMA Nodes.

Trigger: Process Size Threshold
User processes that use more memory than the threshold become candidates for the active QoS policy.

8096 MIB

Cancel Save

Bandwidth QoS Policy

Active QoS Policy: No Policy

QoS Settings Enabled

Policies

Latency Bandwidth

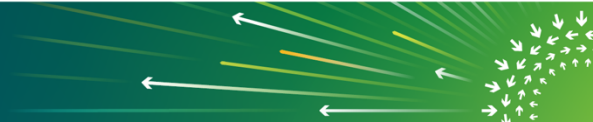
Bandwidth policy activated.
When user processes exceed the memory threshold, a percentage (ratio) of pages are moved to other devices to boost available memory bandwidth.

Trigger: Process Size Threshold
User processes that use more memory than the threshold become candidates for the active QoS policy.

8096 MIB

DRAM:CXL Ratio
QoS-enabled processes will keep a fixed DRAM and CXL NUMA Node distribution (ratio) of memory pages. If multiple CXL Nodes exist, the allocation is evenly distributed among them.

0 50 75 80 90 100
- DRAM - CXL

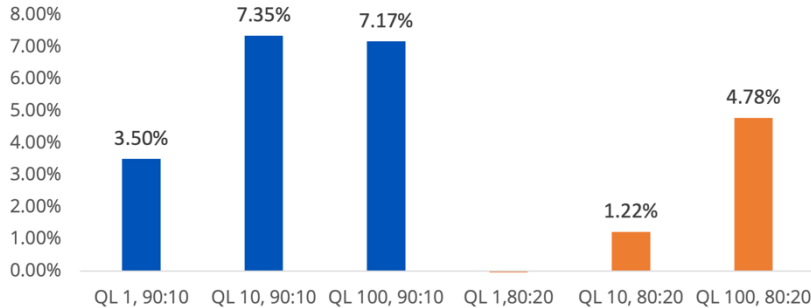


CXL Memory Expansion

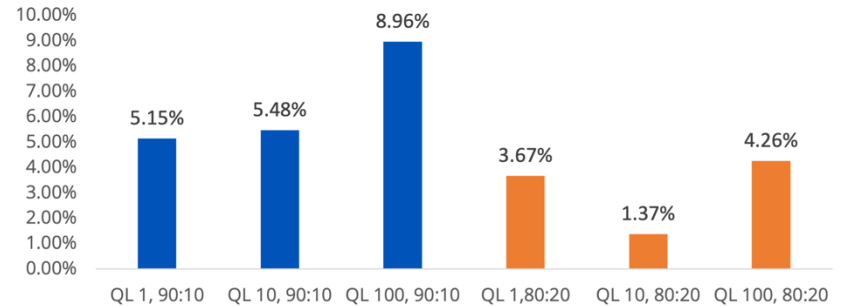
Accelerating Weaviate Vector Databases with Bandwidth Policy

- Using 10% CXL and 20% CXL memory across different query limits (QL)
- Weaviate delivered up to 7.35% more queries per second and up to 8.96% lower latency

Weaviate queries per second with Memory Machine X
(gist-960-Euclidian-128-32 – Queries per Second – EF512)



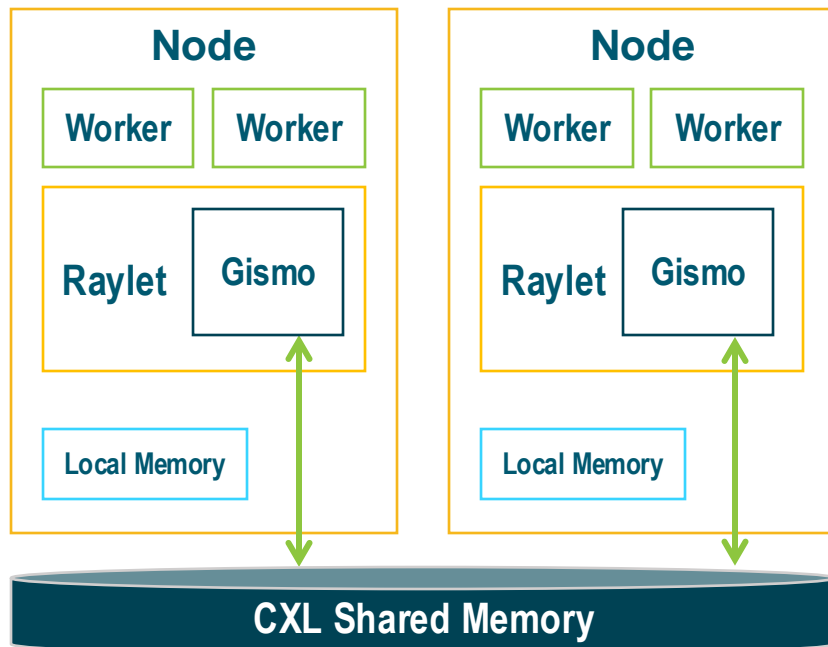
Weaviate latency with Memory Machine X
(gist-960-Euclidian-128-32 – P95 Latency (ms) – EF512)



CXL Memory Sharing

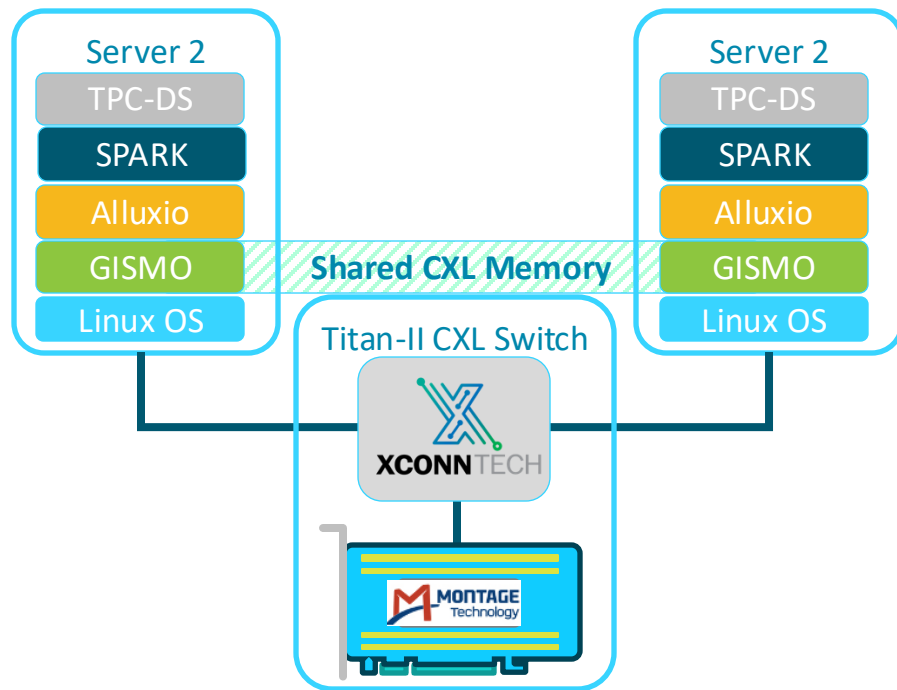
Global IO-Free Shared Memory (GISMO)

- **Efficient Shuffle:** Change from passing the data to passing by reference to the data on the shared memory
- **Faster Data Loading:** Using shared memory to load data to all nodes
- **Reduce Skew and Spilling:** Reduces object spilling and data skewing, replacing local memory with shared memory



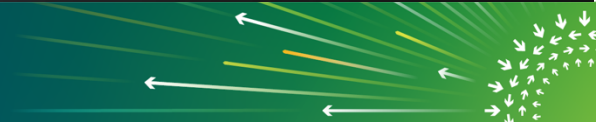
Shared CXL Memory Accelerating TPC & Spark

- Alluxio uses a shared CXL cache (GISMO) vs local node caching, reducing DRAM requirements
- Significantly reduces Disk and Network I/O (HDFS).
 - Data is accessed over the CXL memory bus instead.
- TPC-DS Lower Query Latency
- TPC-DS Higher Queries/Requests per Second (QPS) TPC-DS Reduced time to result



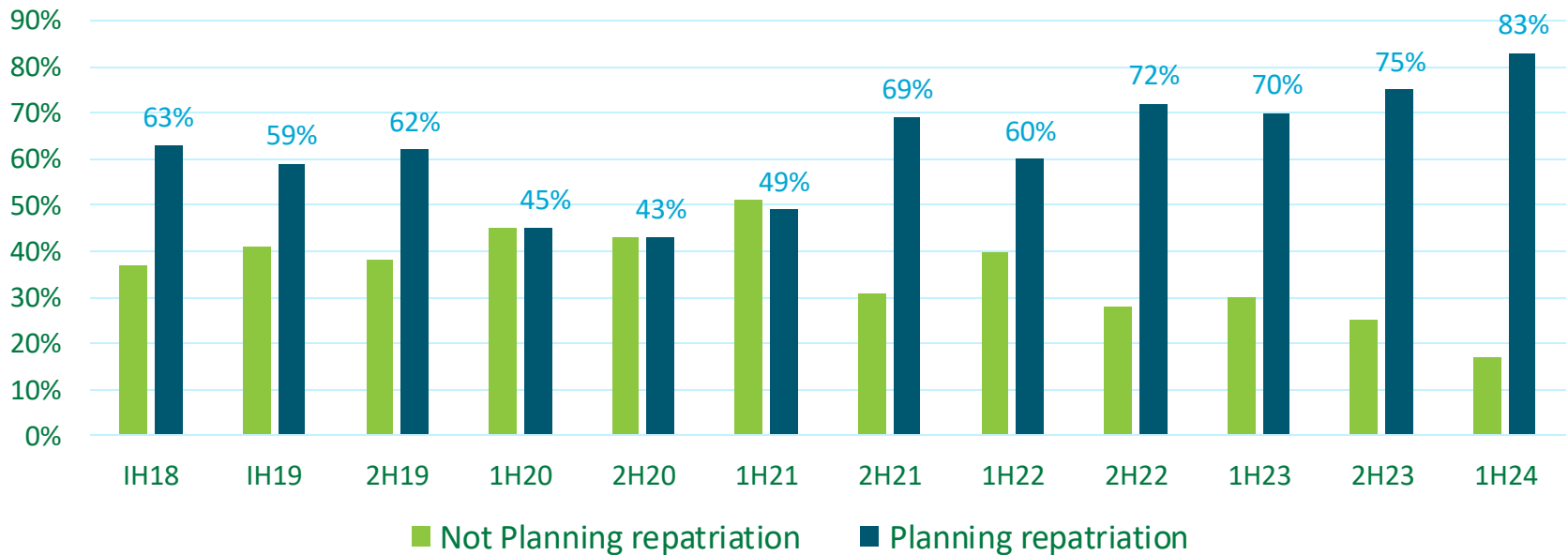


Predictions



Prediction #1: GPU Workloads Will Be Hybrid

Barclays CIO Survey – Percentage of Respondents Planning to Move Workloads Back to Private Cloud /On-Prem from Public Cloud



<https://x.com/MchaelDeI/status/1780672823167742135?prefetchTimest=amp=1728836936140>

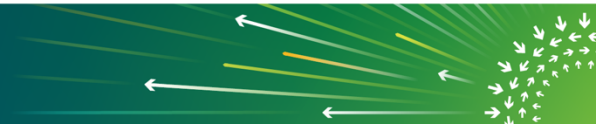


Prediction #2: Co-existence of NVLink and an industry-standard AI Fabric



Prediction #3: Fabric-attached Memory for AI will emerge

- Form the lowest tier in the GPU-centric memory hierarchy
 - HBM – Main DRAM – Fabric-attached Memory
- Inter-node shared memory
 - KV cache
- Replaces today's performance tier of storage
 - Checkpointing store
 - Faster data loading



Thank you!



**MEMORY FABRIC
FORUM**



**OCP
GLOBAL
SUMMIT**

**OCT 15-17, 2024
SAN JOSE, CA**

