UnifabriX

Leader in CXL Smart Memory Technology

The Quest for Bandwidth and Capacity: Memory Edition

Ronen Hyatt

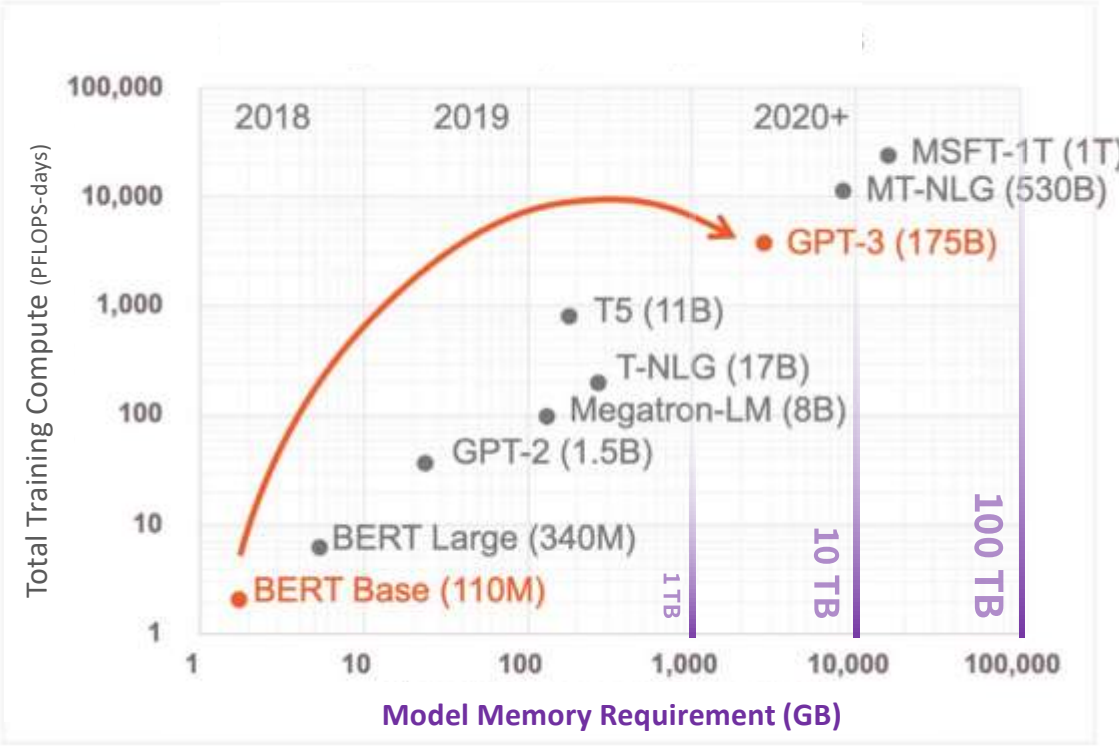
CEO & Founder, UnifabriX

Memory Capacity Matters

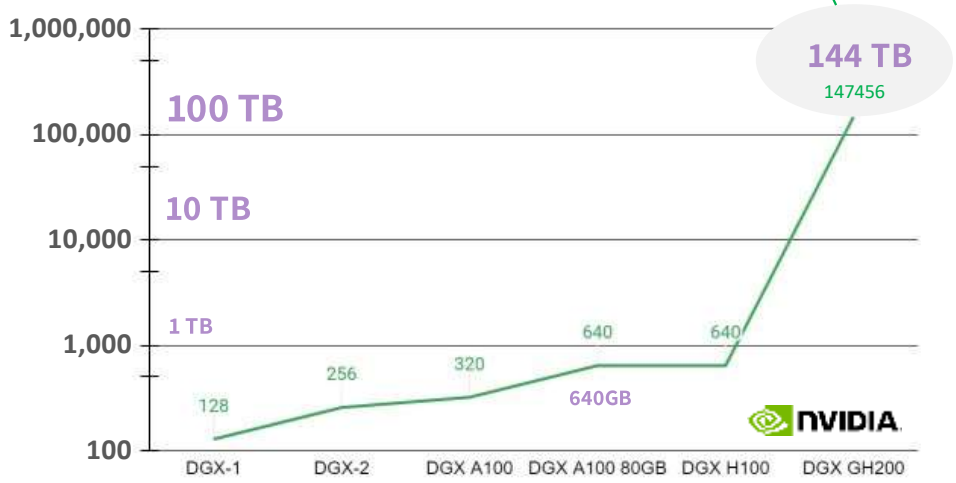
ML models grow exponentially

*The emergence of **Terabyte-Class Models** pushes the limits of the infrastructure towards **memory fabrics***

Memory and Compute Requirements



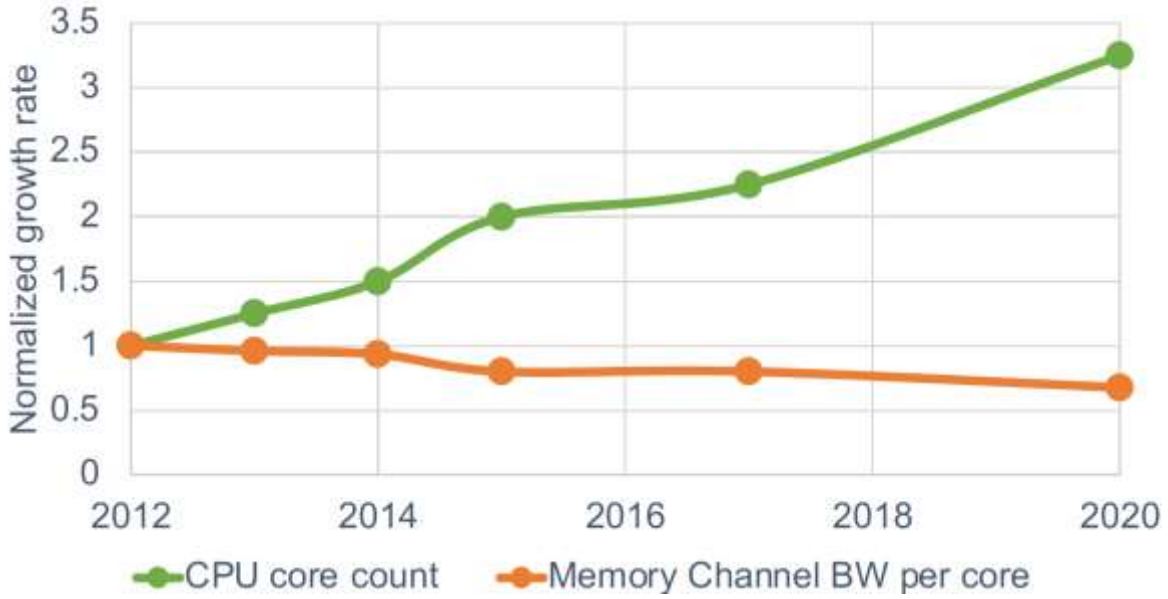
Proprietary Memory Fabric

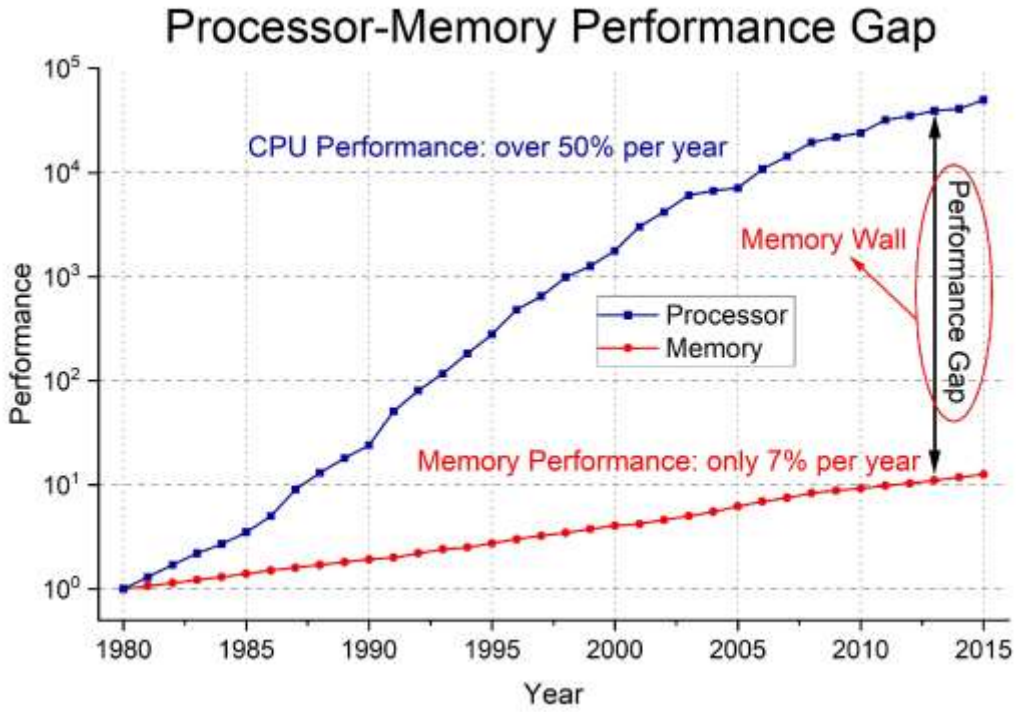


GPU Memory (GB) over NVIDIA DGX Generations

The Ever-Worsening Compute-Memory Gap

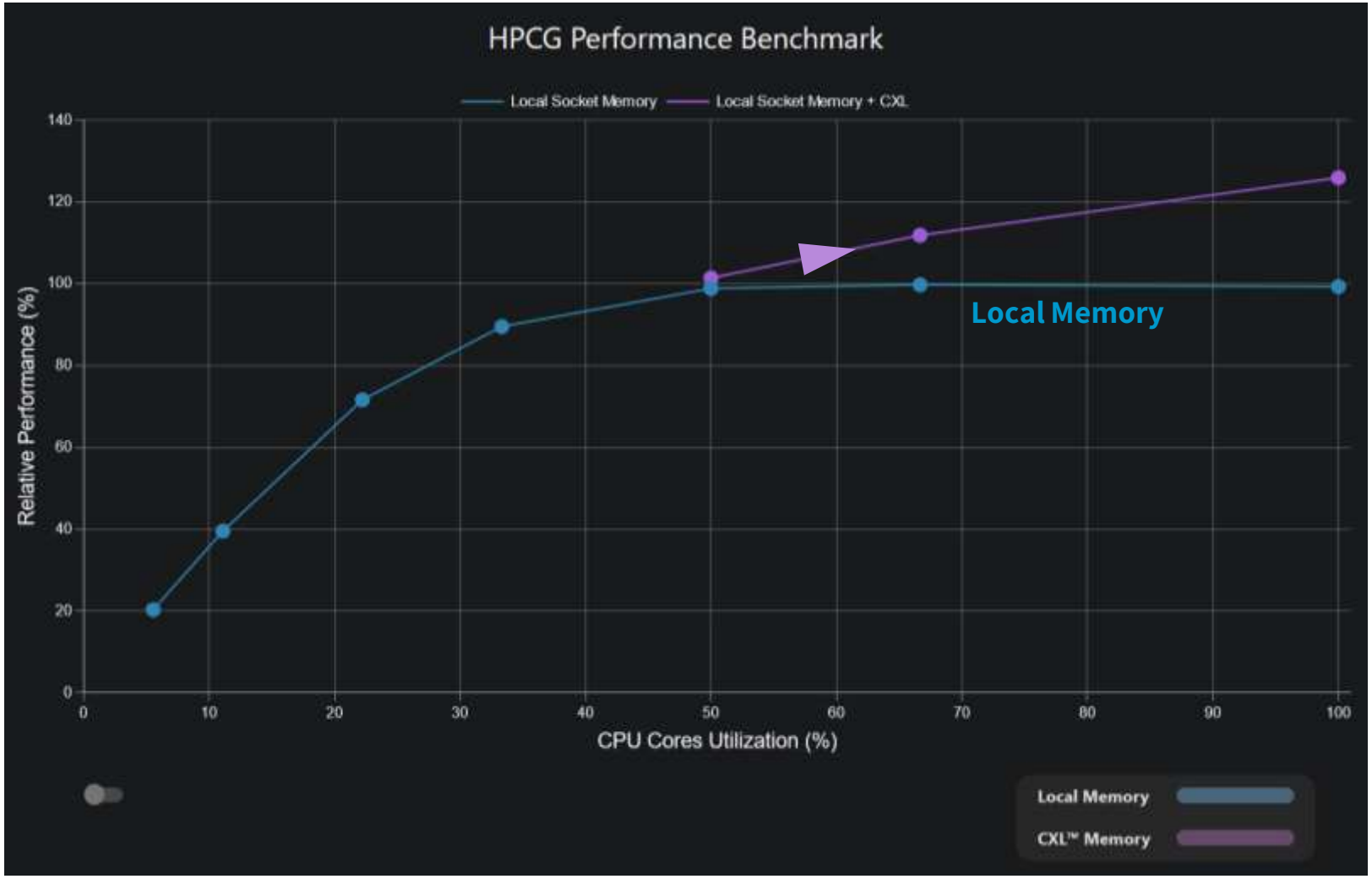
Memory performance advances at a slower pace than compute performance.
Consequently, the performance gap creates a "Memory Wall" effect.





When Memory Bandwidth is Exhausted

Compute cores become stranded



TOP500 The List.

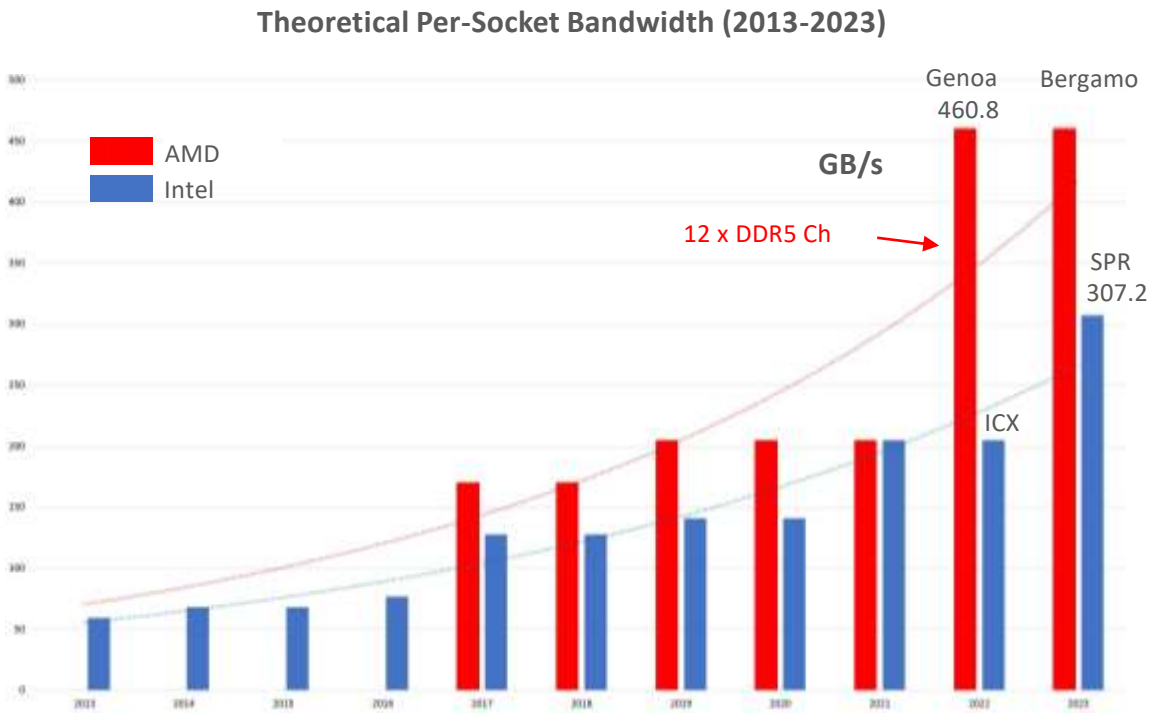
System	Cores	Rmax (PFlop/s)	HPCG (TFlop/s)
Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	16004.50
Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	14054.00
LUMI - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	3408.47
Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	2925.75
Leonardo - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	2566.75

The Simple Truth Behind Memory Bandwidth

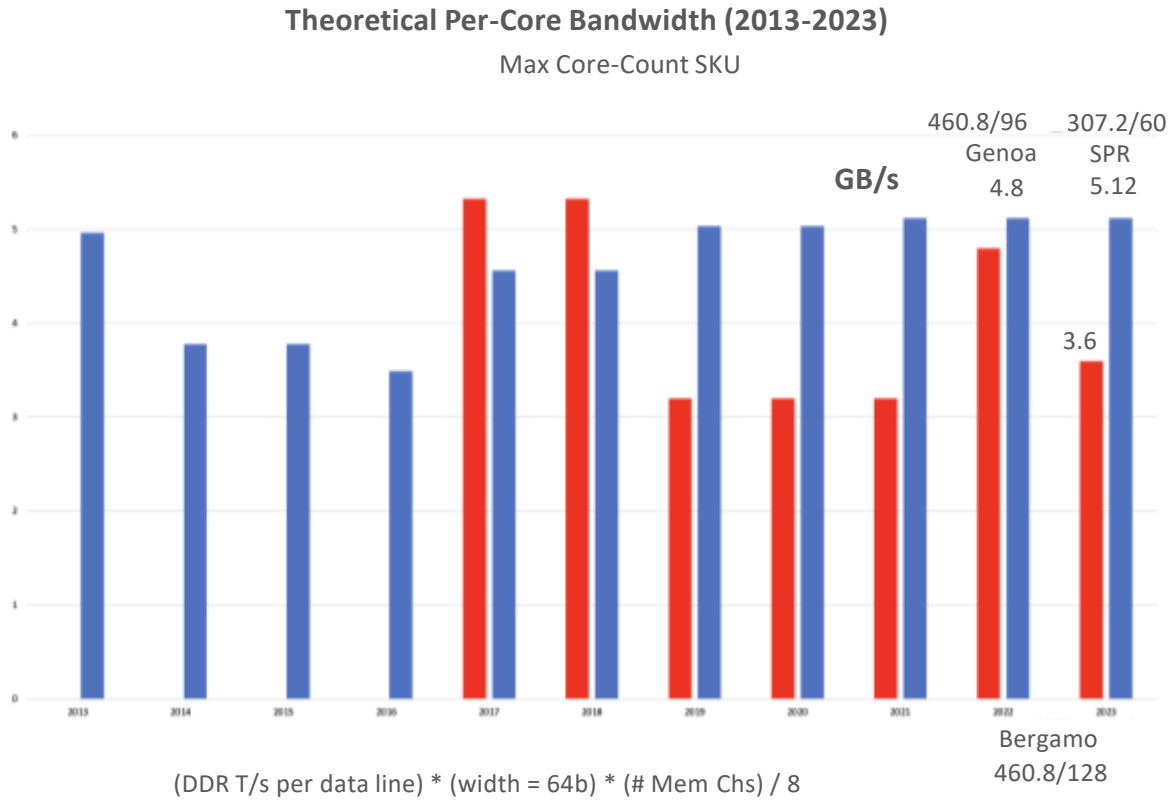
Techniques: Move up the ladder of / DDR rate / DDR generation / # of memory channels

Maximal Memory [Capacity/Core] and [BW/Core] is FIXED on a per CPU-SKU basis

The (Compute:Memory) Ratio is LOCKED on system build



(DDR T/s per data line) * (width = 64b) * (# Mem Chs) / 8



(DDR T/s per data line) * (width = 64b) * (# Mem Chs) / 8
/ (#Cores Max SKU)

The solution

UnifabriX is redefining memory composability

UnifabriX MAX is the world's first **Software-Defined Memory Pool**
to provide **memory Bandwidth and Memory Capacity on-demand**,
using the standard-based OPEN ecosystem of CXL

Full flexibility with setting memory [capacity/core] and [BW/core] independently of CPU SKU

Meet MAX: World's first Software-Defined Memory Pool

- Inventory Management
- Orchestration API
- Performance Telemetry
- Autonomous Tiering
- HeatMap
- Adaptive Memory Sharing
- Smart Interleaving
- Memory-aaS
- Workload SLA
- FlexMemory
- RAS
- Memory Health PFA
- Security
- Virtualization

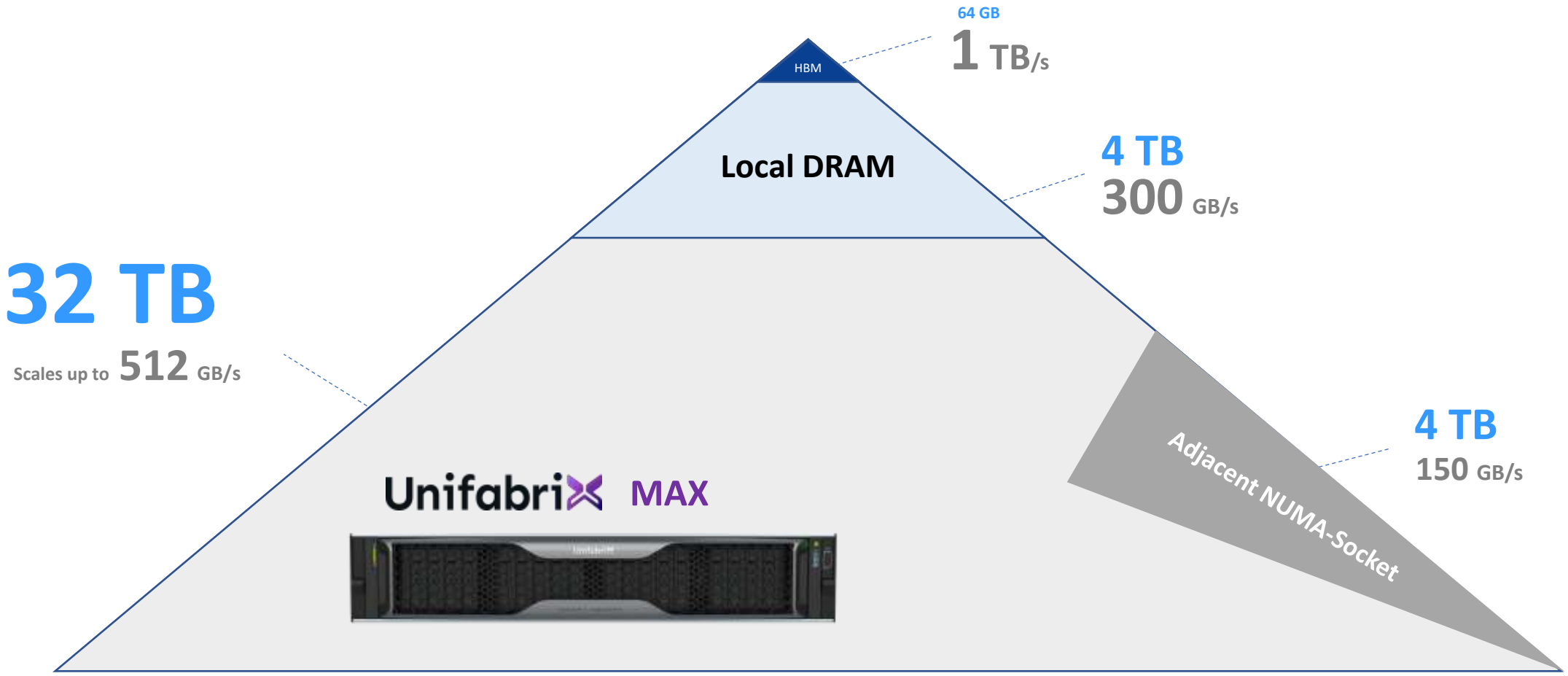
- Standard 2U FF
- 4-32 TB Memory
- CPU-Agnostic

- 8 x CXL 1.1/2.0 FE Ports
- Type-3 / Type-2
- CDFP Gen5/Gen6 x16
- SSDc-oM (NVMe-oCXL)
- EoCXL (Ethernet-oCXL)

- 2 x 400GE QSFP-DD
- CXLoE (CXL-oEthernet)

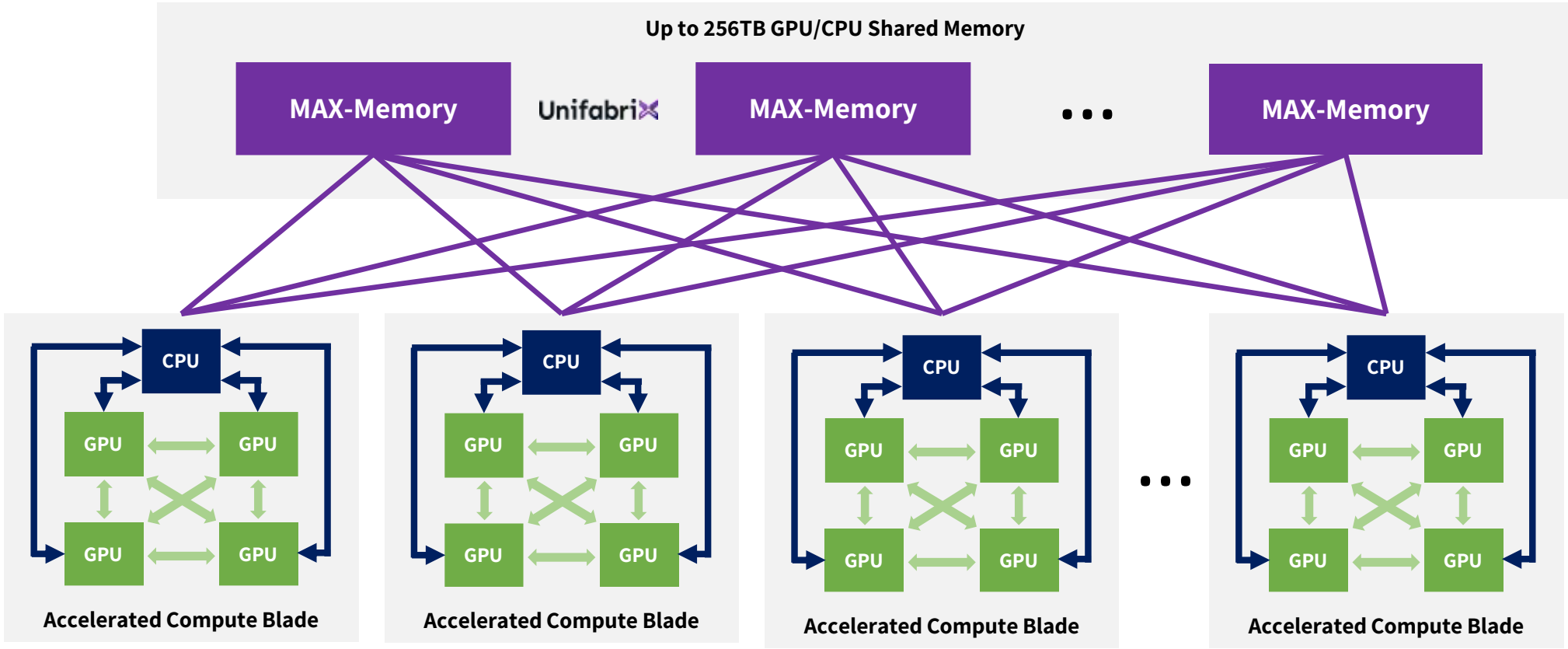
- 2 x CXL 3.0 Fabric Ports

MAX Memory Hierarchy: More Capacity, More Bandwidth



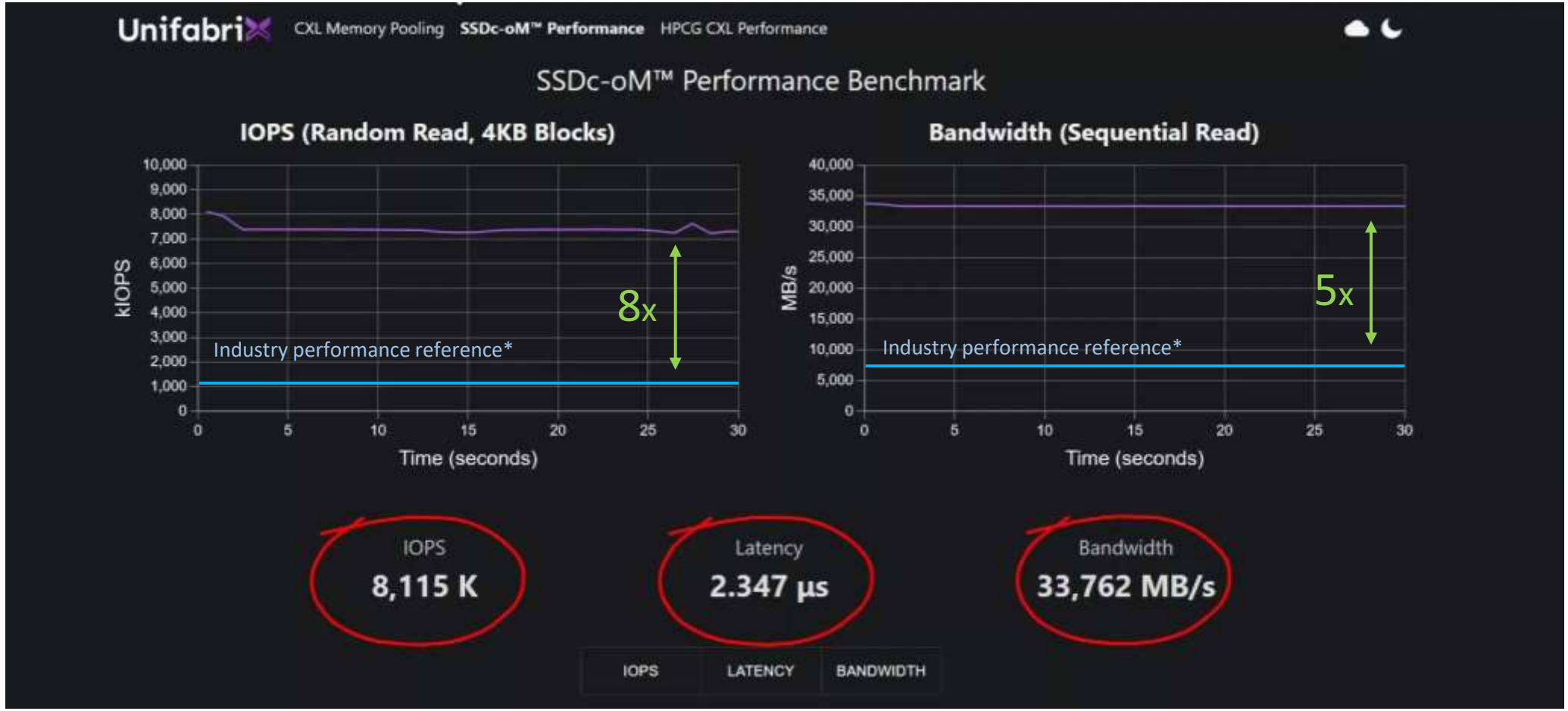
SuperScaling HPC & AI with MAX-Memory

UnifabriX MAX provides up to 256TB of GPU/CPU shared memory for the most demanding HPC & Generative AI workloads



MAX: Dual-Personality Memory-aaS and Storage-aaS

Abstracts the memory media via load-store semantics or via NVMe block semantics.
World's fastest NVMeTM Technology over CXL 2.0/1.1

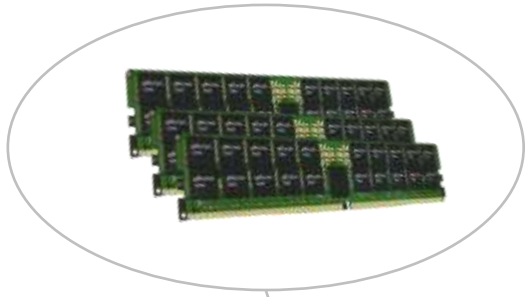


*Compared to state-of-the-art high-performance PCIe Gen4 NVMe Drive

MAXimizing Memory Versatility

Leverages industry standard memory media and emerging datacenter form factors

Standard Memory DIMM Form-Factor



Standard Memory EDSFF Form-Factor

Standard Memory Semantics NAND/SCM