2024 embedded VISION SUMMIT®

The Importance of Memory for Breaking the Edge AI Performance Bottleneck

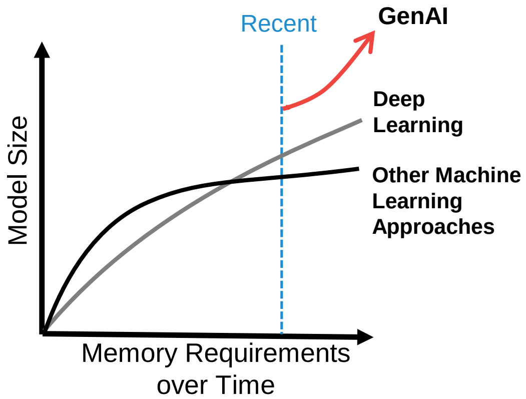
Wil Florentino

Sr. Segment Marketing Manager
Micron Technology

Micron®

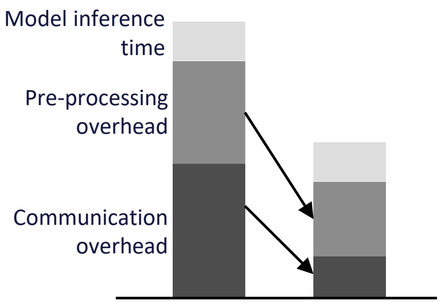
Edge AI reveals memory as the bottleneck

Trend toward memory-bound applications



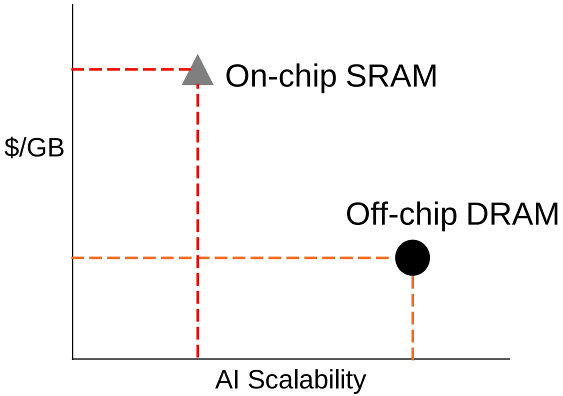
Model complexity vs. memory bandwidth

- Transformer size growth 410x / 2 years
- AI HW memory bandwidth 2x / 2 years[1]



Pre-processing latency in AI execution

- Data pre-preprocessing overhead[2] impacts latency



$/GB vs. scalability

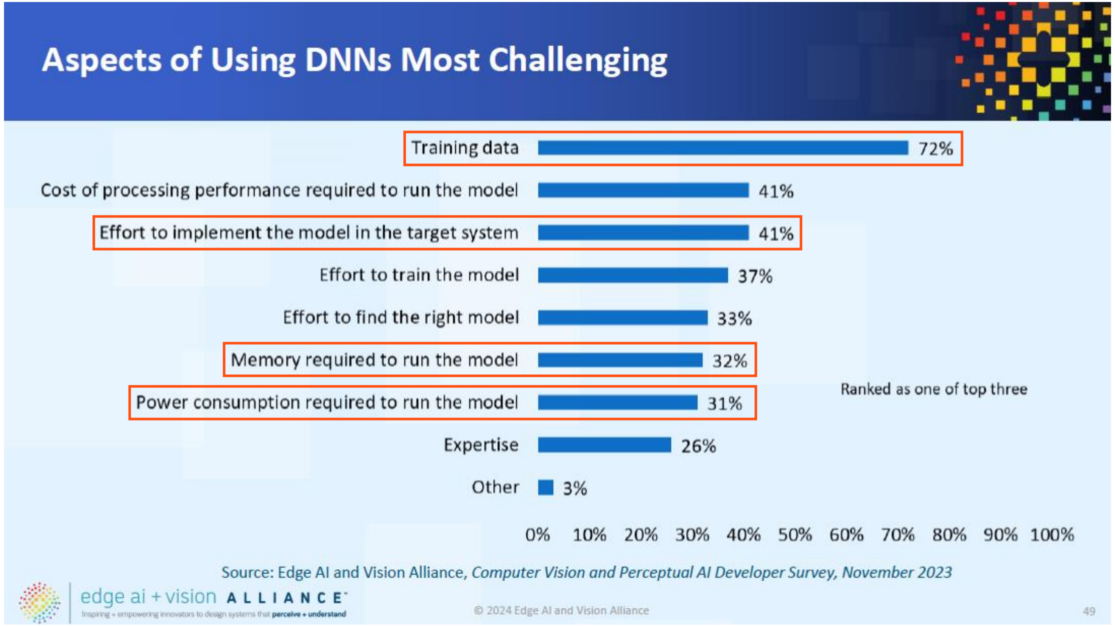
- SRAM: $5,000/GB
- DRAM: $50/GB[3]

[1] "AI and memory wall," Medium, 2021 [2] "Rapid Data Pre-Processing with NVIDIA DALI" NVIDIA Technical Blog, 2021 [3] "SRAM vs. DRAM: Difference between SRAM & DRAM explained," Enterprise Storage Forum, 2023

DNN challenges relate back to memory and storage

Edge AI and Vision Alliance report on DNN implementation challenges

- Training data trade-offs between cost of storage on-premise vs. cloud
- Complexity of on-device implementation in target
- Type of and memory performance influence the efficiency of running the model
- Power consumption



Aspects of Using DNNs Most Challenging

Ranked as one of top three

Source: Edge AI and Vision Alliance, *Computer Vision and Perceptual AI Developer Survey, November 2023*

edge ai + vision ALLIANCE™

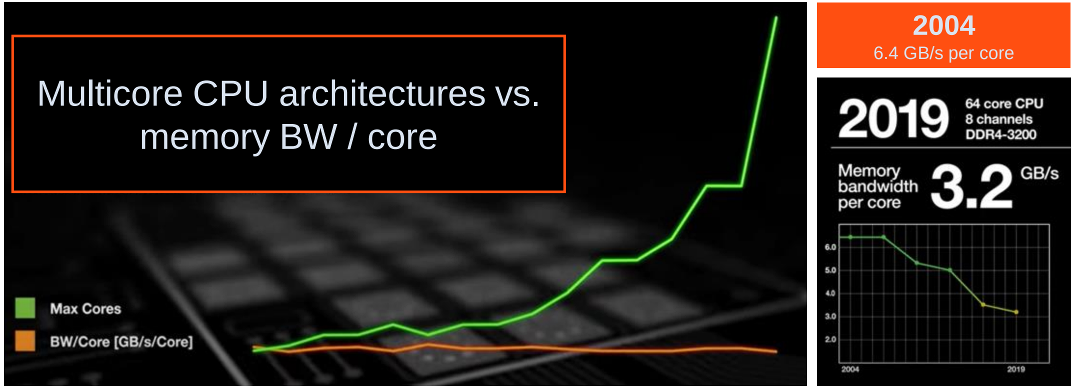
Inspiring + empowering innovators to design systems that perceive + understand

© 2024 Edge AI and Vision Alliance

49

DRAM memory bandwidth per core has been declining

- CPU core counts are increasing at a rate that minimizes available memory bandwidth per core
- New memory technologies are required to meet next-generation bandwidth-per-core requirements in multi-core CPUs
- Edge AI inference compute requires additional memory consideration



Source: Micron. bandwidth normalized to x64 interface, 64Byte random accesses, 66% reads, dual-rank x4 simulation, 16Gb. Best estimates; subject to change.

The many levers of a memory device

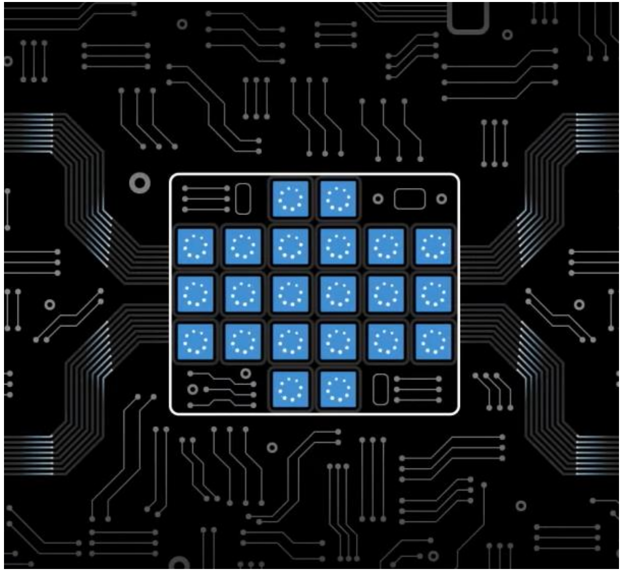
Complex design considerations for memory improve performance and lower costs

Configuration

- Density per die
- Die per package
- I/O width
- Bank groups
- Technology node

Performance

- Speed/pin
- Number of channels
- Prefetch size
- Burst length
- Read latency



Operational

- On-die Error Correction
- Thermal profile
- Refresh management
- Power reduction modes
- Active vs. standby power (picojoule/bit)

Application focus

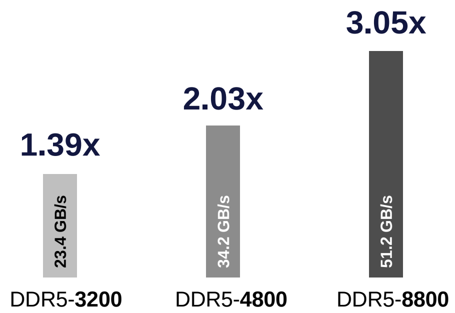
- Functional safety
- Reliability/Availability/Serviceability
- Extended temperature
- Validation and testing
- Product lifecycle
- Industrial rated
- Auto validated

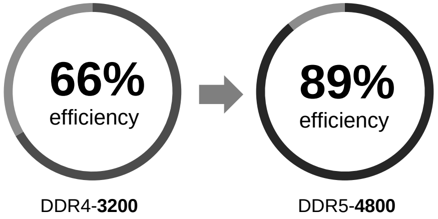
DDR5 for data-intensive training workloads

2x capability

- Burst length
- Bank groups
- Banks

DDR5 memory comparisons
Increased bandwidth **more than 3x**[1]





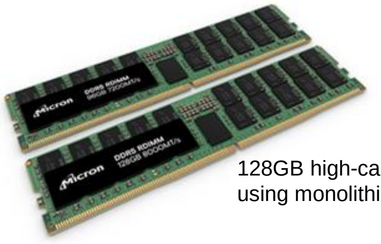
Higher bus efficiency **up to 90%**[1]
Faster transfer speed **up to 8800 MT*/s**[2]

Improved overall workload performance[3]

Cloud
Virtualization 40%

Data center
Business apps 45%

High-performance computing
HPC modeling >200%



128GB high-capacity RDIMM using monolithic 32Gb DRAM

[1] Benchmark simulation comparison of DDR5 vs. DDR4-3200 [2] Based on defined JEDEC specification [3] Results based on internal testing, third-party testing and/or industry workload benchmark testing

*Mega-transfers per second

Compute bandwidth requirements by edge solution

AI TOPs* vs. number of LPDDR4 devices scenarios

🔵 x16 🟢 x32 🟠 x64

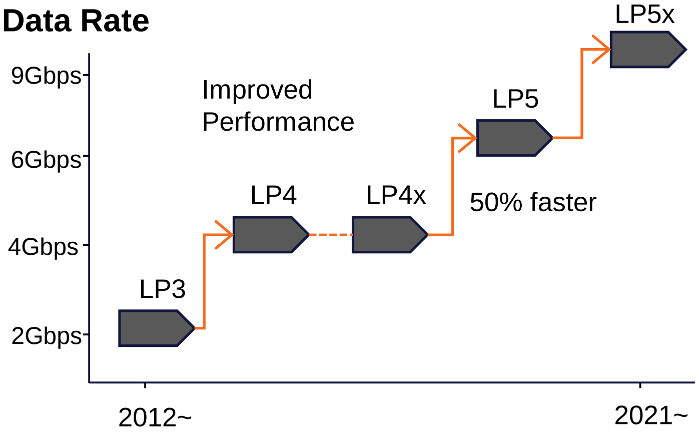
	Sensor edge IoT sensors and ultra low power devices (TinyML)	**Device edge** Cameras, machines and industrial/SFF PC/server	**Network edge** Industrial PC/server, network equipment, NVR/VMS appliances	**Compute edge** Server/NVR/VMS appliances
Power	<1W	2W <= 15W	15W <=75W	15W <= 75W+
SoC/ASIC IO width (typical)	x16	x32	x64	x128
DLA INT 8 TOPS	<4	4–20	20–50	50–100
Est. bandwidth to full utilization of accelerator [saturate accelerator**]	18 GB/s	90 GB/s	225 GB/s	451 GB/s
BW of LP4 @ 4.2Gbps/pin IO per device (x16/x32/x64)	8 GB/s 🔵	17 GB/s 🟢	33 GB/s 🟠	33 GB/s 🟠
Number of LP4 packaged devices	**3**	**6**	**7**	**14!**

[1] *relative reference models only, actuals will vary. [2] **Device Level Accelerator bandwidth assumed roofline modeling (Resnet 50) [3] "V. Sze, Y. -H. Chen, T. -J. Yang and J. S. Emer, "How to Evaluate Deep Neural Network Processors: TOPS/W (Alone) Considered Harmful," in IEEE Solid-State Circuits Magazine, vol. 12, no. 3, pp. 28-41"

LPDDR5 offers a leap in performance and possibilities

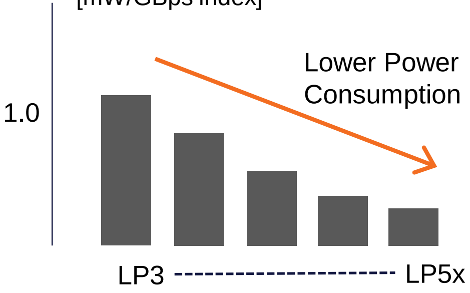
embedded VISION SUMMIT

Data Rate

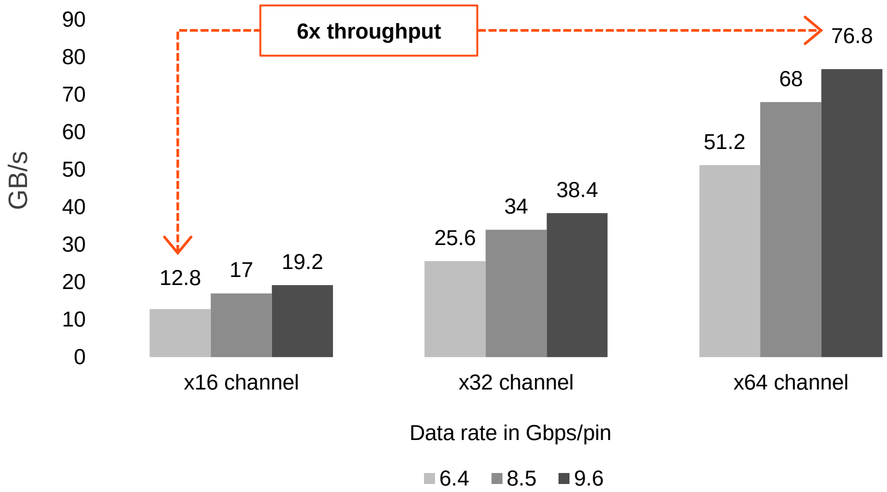


Improved Power Savings Features

[mW/GBps index]



LPDDR5X bandwidth at different channel and pin speed



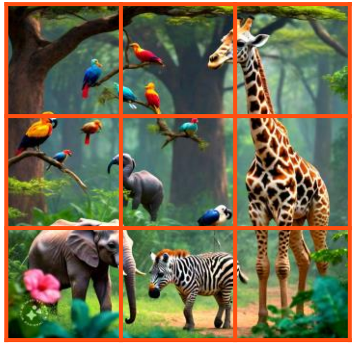
- Reduces number of components to get to same bandwidth
- Improved architecture
- Lower power [pj/bit]

Micron

Memory footprint as a function of batch size

Tiling for small object detection in high-resolution vision

Meta AI-generated image
(Imagine Platform)

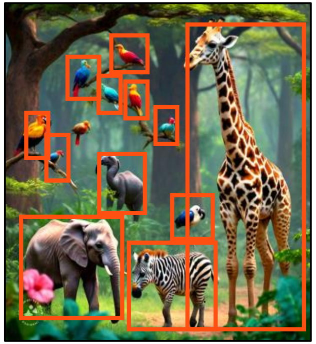


Tiling high-resolution images

Stacked
inputs

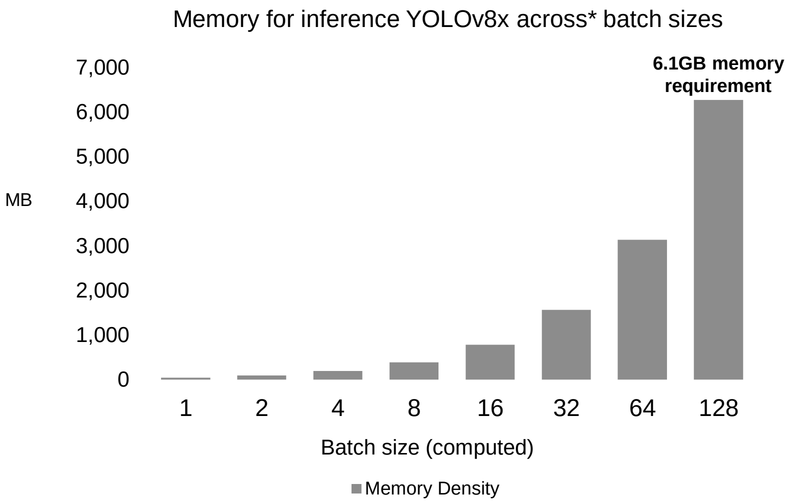
Example: Batch size: 9 x N

Convolutional model



Higher batch size improves results

Batch size impacts the memory footprint



*Parameter size: 273MB

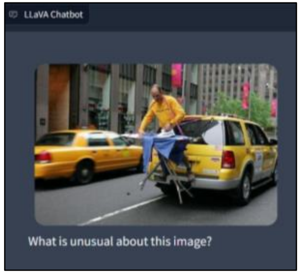
[1] Small object detection: An image tiling based approach, Medium, 2021 [Link]
[2] S. Ngyuyen, et al., "Dynamic tiling: A model-agnostic, adaptive, scalable, and inference-data-centric approach for efficient and accurate small object detection," arXiv:2309.11069v1, 2023
[3] F. Akyon, et al., "SAHI: Slicing aided hyper inference and fine-tuning for small object detection," IEEE ICIP, 2022
[4] F. Unel, et al., "The power of tiling for small object detection," CVPR, 2019
[5] Training vs. inference – Memory consumption by neural networks [Link]
[6] GitHub: TorchInfo [Link]
[7] Model not quantized (fp32). Memory footprint of two largest consecutive layers.

Why memory is important for generative language

- Models are very large and often need to fit in DRAM
- Bandwidth is critical to quality of service
  - Tokens/sec is highly correlated with DRAM bandwidth

LLAVA 7B with 8-bit quantization* ~5 seconds

LP4 4.2 (x32): 17 GB/s

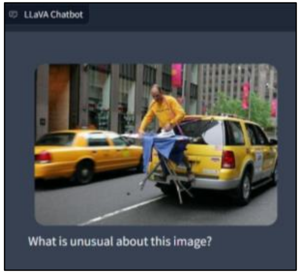


LLaVA Chatbot

What is unusual about this image?

The image shows a person ironing clothes on a…

LP5X 9.6 (x128): 153 GB/s

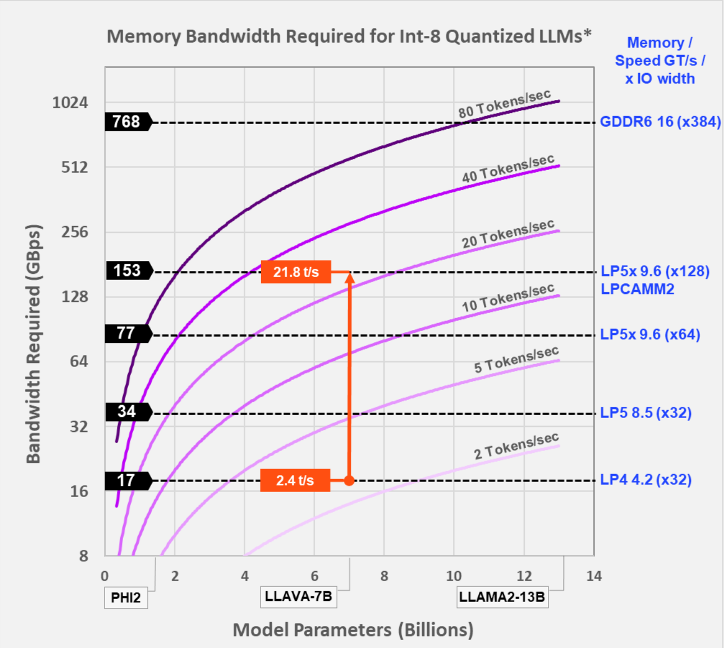


LLaVA Chatbot

What is unusual about this image?

The image depicts an unusual scene where a man is ironing clothes on an ironing board placed on the back of a moving vehicle, specifically a yellow SUV. This is not a typical activity one would expect to see on a city street, as ironing is usually done indoors in a stationary position to ensure safety and to prevent accidents. The man's actions are not only unconventional but also potentially dangerous due to the risk of falling or being hit by other vehicles or pedestrians. Additionally, the presence of a taxicab in the background adds to the urban environment, which makes the scene even more out of the ordinary.

* LLAVA (llava-vl.github.io) | Assume 1 token/word | Excluding time to first token



*theoretical estimates based on memory JEDEC specs & published quantization computations

[1] Assumes GGML Quantization: ggml.ai. [2] Kim, Sehoon, et al. "Full stack optimization of transformer inference: a survey." arXiv preprint arXiv:2302.14017 (2023)

LPCAMM2 for AI-equipped systems

Performance

- LPDDR5x speed of up to **9.6Gbps**
- **Full 128-bit**, dual-channel, low-power modular memory solution

Modularity

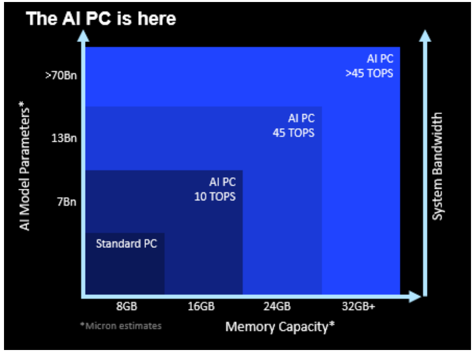
- Flexibility to **upgrade system memory** capacity
- Single PCB for all memory configurations

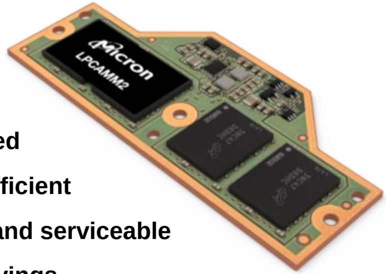
Power efficiency

- Consumes **57%-61%**[1] less active power and up to **80%**[1] less system standby power compared to DDR5 SODIMM
- Thermal efficiency, fanless computers

Form factor

- Up to **64%**[2] space savings
- Space savings for industrial PCs, embedded single-board computers, AIoT systems





High speed

Energy efficient

Modular and serviceable

Space savings

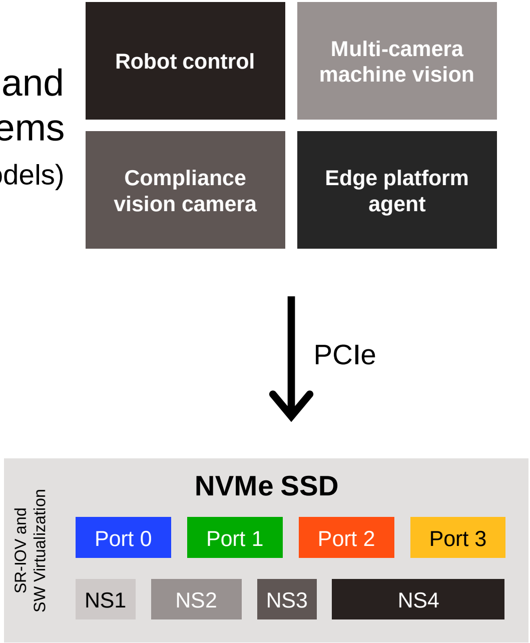
[1] Power measurements in mW per 64-bit bus at the same LPDDR5X speed compared to SODIMM [2] Calculation based on comparison of the total volume of commercially available dual-stacked DDR5 SODIMM module (32,808 mm3) to LPCAMM2 module (11,934 mm3)

Multiport SSD as centralized storage

Supporting multiple subsystems in a single storage device

Multiple HW and SW subsystems

(different AI models)



Robot control

Multi-camera machine vision

Compliance vision camera

Edge platform agent

PCIe

NVMe SSD

SR-IOV and SW Virtualization

Port 0

Port 1

Port 2

Port 3

NS1

NS2

NS3

NS4

Single multiport centralized storage

4150AT product highlights

- **Configurable multiport** (single, dual, triple and quad)
- **SR-IOV** allowing for shared and private namespaces
- **Design flexibility** to match system usage models with TLC, SLC and HE-SLC endurance modes
- Up to **600K read and 100K write** IOPS performance
- **-40 C to 115 C Tc** operating temperature range
- **Fast boot** with TTR <100ms

Legend: SR-IOV = single root I/O virtualization, NS = namespace, PF = physical function, VF = virtual function, Tc = case temperature, TTR = time to ready, TLC = triple-level cell, SLC = single-level cell, HE-SLC = high endurance SLC, IOPS = input/output operations per second

Micron AI memory and storage portfolio

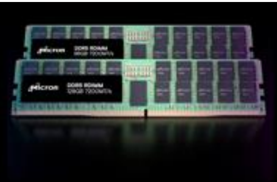
Leadership products to enable AI workloads

High-bandwidth in-package memory

HBM3E

High-performance graphics memory

GDDR6/X



High-capacity DRAM
128GB DDR5 using monolithic 32Gb DRAM

Compute DRAM

DDR5

Low-power memory

LPCAMM2

Low-power memory

LPDDR5X

Universal flash storage

UFS 4.0

Memory expansion with CXL™

CZ120

High-performance data center NVMe™ SSD

Micron 9400

High-capacity data center NVMe™ SSD

Micron 6500 ION

Summary

Micron memory enables all forms of AI embedded solutions

AI at the edge (outside the data center) reveals memory as a bottleneck

- Disproportionate growth between transformer size vs. memory bandwidth
- Data pre-preprocessing overhead impacts latency
- On-chip SRAM is cost prohibitive vs. external DRAM

Memory technology influences AI model execution performance

- Edge AI devices TOPS showcase memory bandwidth gap
- Tiling activation requires in-line memory density resources
- In generative language, bandwidth is required for quality of service

Leading memory technologies offer the best mix of solutions for edge AI applications

- DDR5 for AI training workloads
- LPDDR4 and LPDDR5 for neural network compute
- LPCAMM2 to leverage LPDDR5X performance with DIMM modularity
- Multiport SSD to support different AI models and compute in a single storage

Smart factory and robotics

Industrial AR/VR

Smart grid and clean energy

AI-enabled video security and analytics

Low earth orbit (LEO) communication

Drones and industrial transport

Visit us at **Booth #105**

Thank You