2025 EMBEDDED VISION SUMMIT

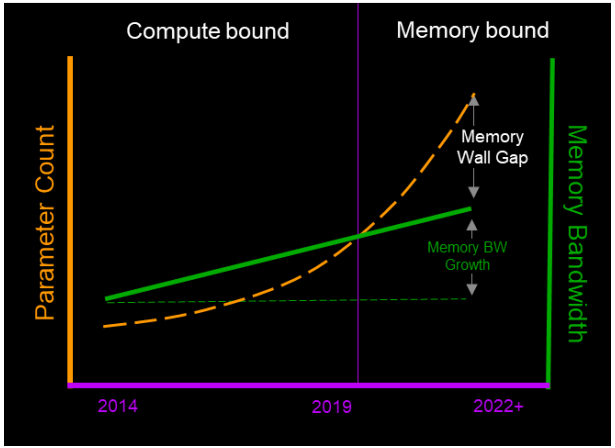
Addressing Evolving AI Model Challenges Through Memory and Storage

Wil Florentino
Senior Segment Marketing Manager
Automotive and Embedded Business Unit

micron

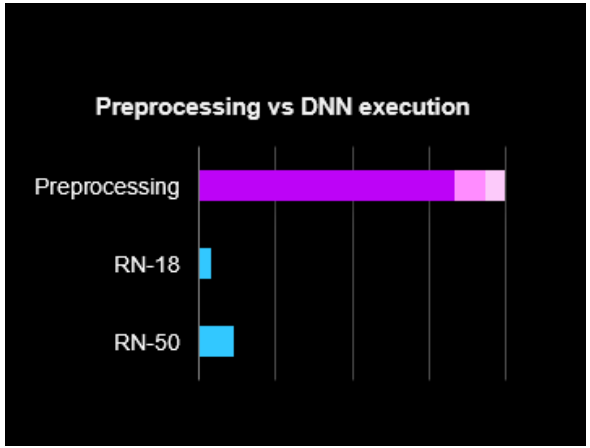
Embedded AI Reveals Memory As Critical Consideration

Model Complexity vs Memory Bandwidth



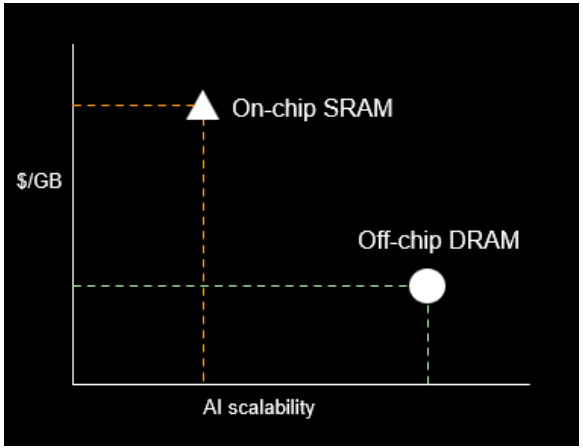
Transformer Size growth 410x / 2 years
AI HW Memory bandwidth 2x / 2 years[1]

Preprocessing Latency vs NN execution



Video image preprocessing overhead impacts latency and DNN execution[2]

$/GB vs Scalability



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

1) "AI and Memory Wall", Medium, 2021
2) "Accelerating Queries over Unstructured Data with ML", Stanford Dawn, 2020
3) "SRAM vs DRAM: Difference Between SRAM & DRAM Explained", Enterprise Storage Forum, 2023

Edge deployments and Memory densities

Embedded AI will be dominated by ASIC accelerators and NPUs

Model size

	Model name	Params	Tasks	Minimum Memory	Algorithms	Hardware
Near-edge deployment Training and inference happens close to the device	Ex. Llama 3.1, Gemma 2, etc.	**1B+**	Scene understanding Inventory tracking Classification Others	**6 GB +**	Multimodal, Transformer, LLM, VLM, CNN, RNN, FNN, CLIP	FPGA, ASIC, SoC, NPU MCU, GPU, ULP-CPU Edge Server, Gateways, IPC, PLC
On device Inferencing happens on the edge device	DEIT_Base	**86.5M**	Classification	**400MB – 4 GB**	LLMs, VLM (2025), CNN, RNN, ML	FPGA ASIC SoC NPU MCU GPU
	YoLov7e6	**97.2M**	Object detection			
	CLIP_ResNet_50x4	**87M**	Zero-shot classification			
TinyML on device Training and inference happens on the edge device	SSD_MobilNet_v2	**4.46M**	Object detection	**320 kB – 64 MB**	CNN, ML	FPGA TML, MCU NPU, ULP-CPU
	SqueezeNet v1.1	**1.24M**	Classification			
	EfficientNet_Lite1	**4.73M**	Object detection			

Table 1: model name, number of parameters, task it can perform, memory required for deployment type, and possible algorithms

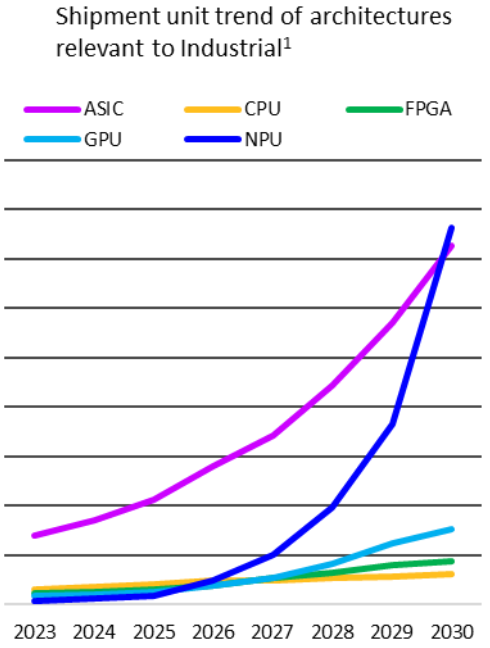


Fig 1: Shipment unit trend and CAGR 2024-2030 of edge AI and TinyML architectures

Omdia, 2024. ASIC-Application Specific Integrated Circuit, CPU-Central Processing Unit, FPGA-Field Programmable Gate Array, TML-TinyML, GPU-Graphics Processing Unit, MCU-Micro Controller Unit, NPU-Neural Processing Unit, ULP-Ultra Low Power, IPC-Industrial Personal Computer, PLC-Programmable Logic Control, CNN-Convolution Neural Network, RNN-Recurrent Neural Network, SDK-Software Development Kit, VLM-Visual Language Models, ViT-Visual Transformers,

Numerous AI Tasks deployed in industrial applications

AI Tasks in Factory Automation

Vision
Object Detection
Image Classification
Semantic Segmentation
Image Captioning
Image Denoising

Numeric
Pattern Recognition
Multistage Reinforcement learning

Multi-Modal
Scene Understanding
Vision Language Action
Multi-Step Planning

AI Tasks in Transportation

Vision
Zero Shot Object Detection
Zero Shot Image Classification
Semantic Segmentation
Facial Expression Recognition

Audio
Automatic Speech Recognition

Multi-Modal
Scene Understanding
Multi-Step Planning
Semantic Reasoning

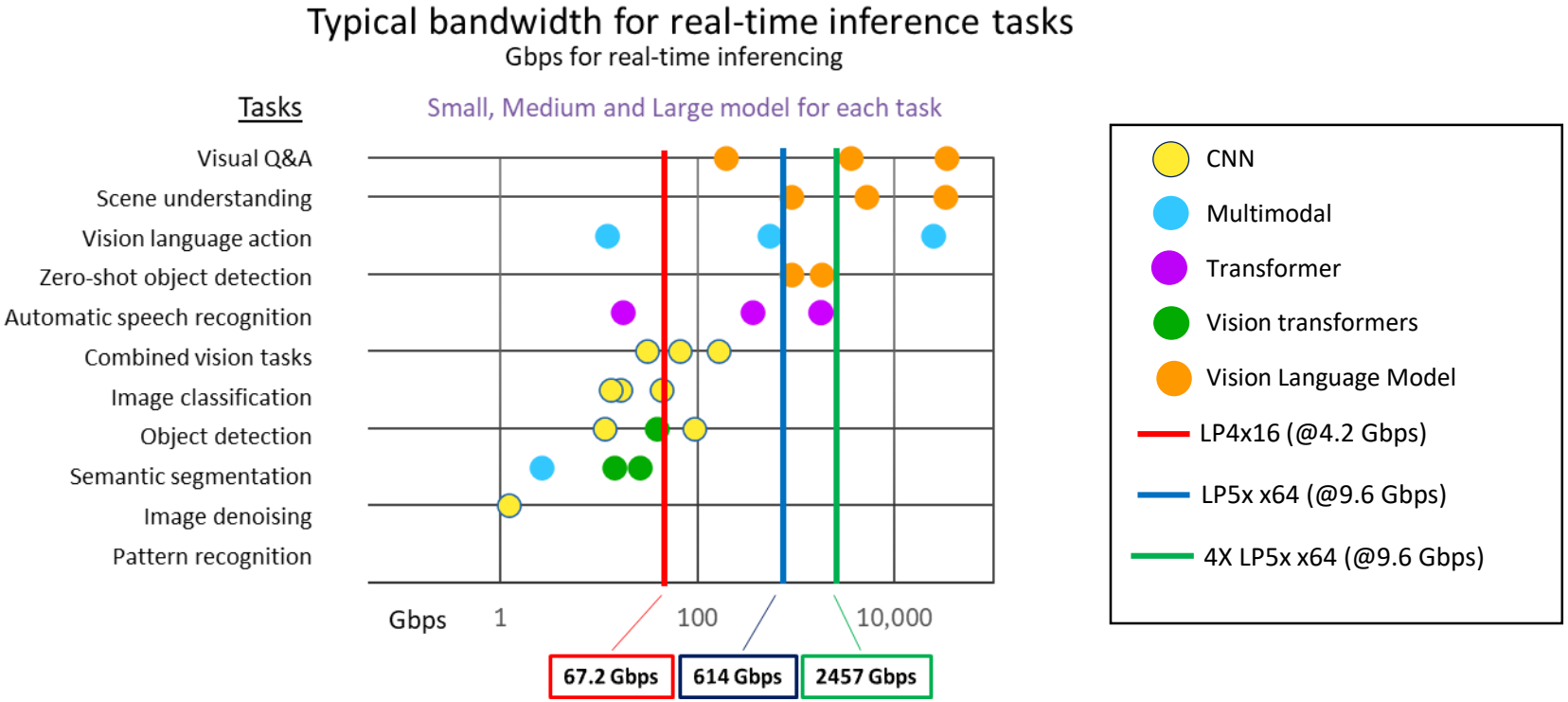
AI Tasks in A&D

Vision
Image Denoising
Object Detection
Object Tracking
Zero Shot Image Classification
Semantic Segmentation

Audio
Automatic Speech Recognition

Multi-Modal
Scene Understanding
Multi-Step Planning
Semantic Reasoning
Adaptive Skill Coordination

Support for more complex models and tasks require higher memory performance

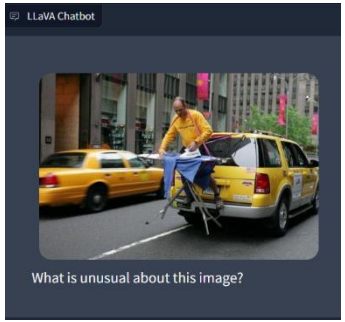


Source: Micron Marketing internal analysis; assumptions on % of system memory bandwidth utilization, inference precision, real-time compute time scale – offered as representation only

Memory bandwidth is critical for generative language

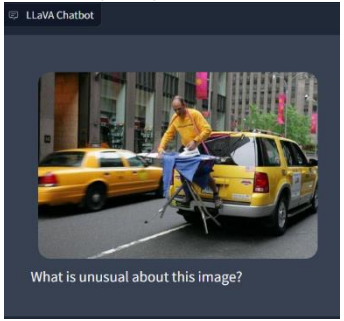
- Models are very large and often need to fit in DRAM
- Bandwidth is critical to quality of service
  - Tokens/sec is highly correlated with DRAM bandwidth

LP4 4.2 (x32): 17 GB/s



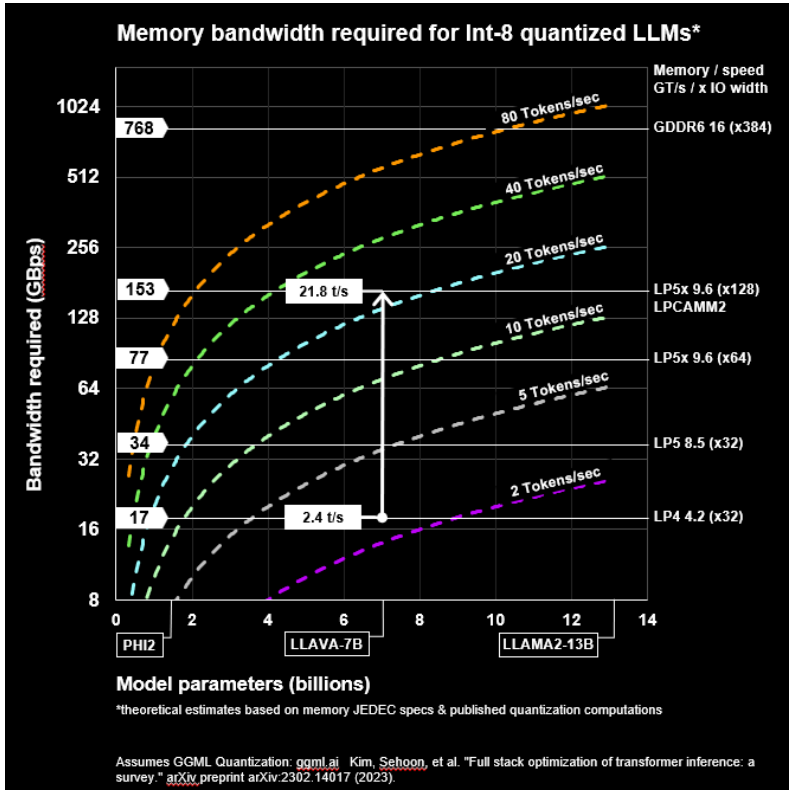
The image shows a person ironing clothes on a …

LP5x 9.6 (x128): 153 GB/s



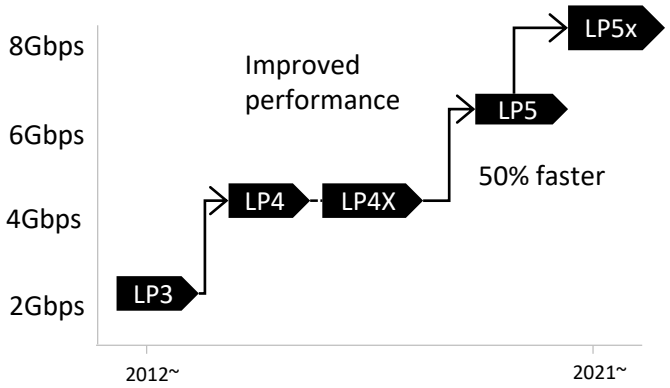
The image depicts an unusual scene where a man is ironing clothes on an ironing board placed on the back of a moving vehicle, specifically a yellow SUV. This is not a typical activity one would expect to see on a city street, as ironing is usually done indoors in a stationary position to ensure safety 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

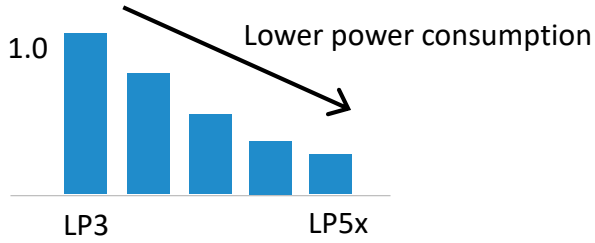


LPDDR5X offers a leap in performance and possibilities

Data rate

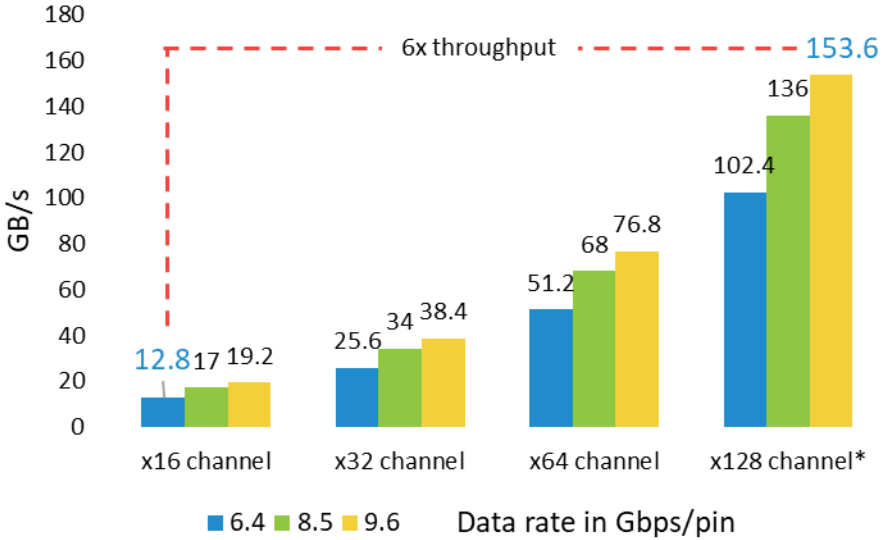


Reduced power consumption
[mW/GBps index]



Based on Jedec standards and performance comparison

LPDDR5X bandwidth for different channel size and pin speed performance



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

*Based on LPCAMM2

LPCAMM2 for AI-equipped systems

Performance

- Speed capability of up to 9600Mbps utilizing LPDDR5X technology
- Full 128-bit, dual-channel, low-power modular memory solution
- AI inference systems

Power efficiency

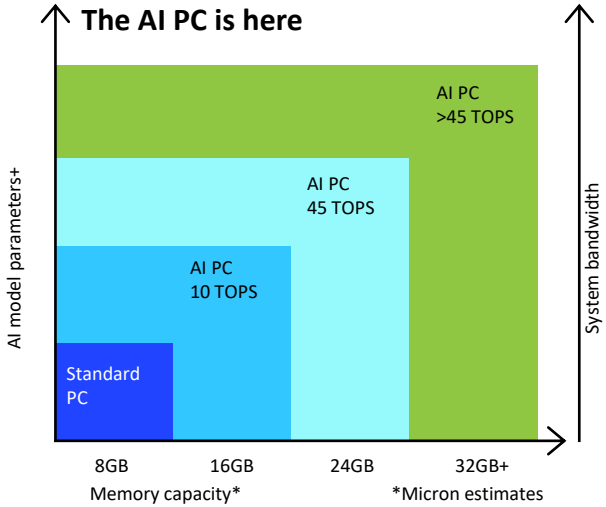
- Consumes 57-61%[1] less active power and up to 80%[1] less system standby power compared to DDR5 SODIMM
- Thermal efficiency, fanless computers

Modularity

- LPCAMM2 provides users with the flexibility to upgrade system memory capacity
- Provides OEMs with a standard PCB for all memory configurations

Form factor

- Creates up to 64%[2] space savings, providing flexible design options (e.g., bigger battery or smaller footprint)
- Space savings for IPC and SBC

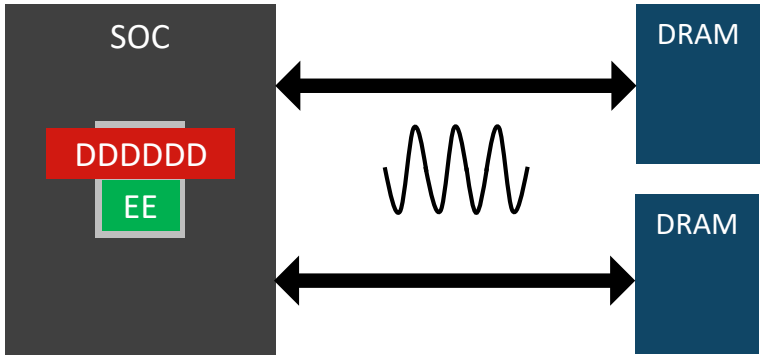


- High speed
- Energy efficient
- Modular and serviceable
- Space savings

[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 LPLPCAMM2 module (11,934 mm3).

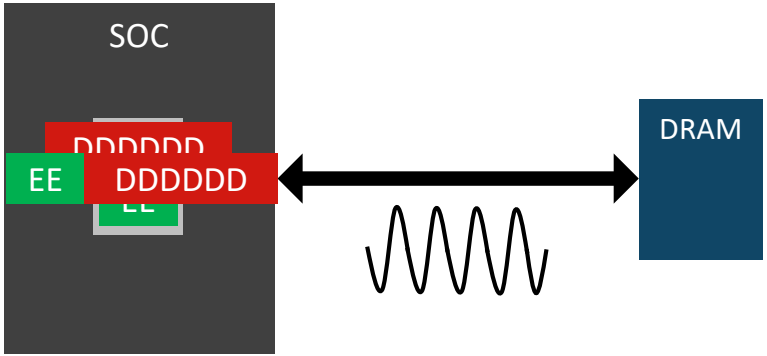
Options for system ECC

Traditional approaches have limitations



Sideband ECC

- Parity bits are generated by the SOC and transmitted in parallel with mission data to a separate component
- No mission data bandwidth/capacity impact
- Higher cost (extra component, board space, routing, SOC PHY area, power, etc.)



Inline ECC

- Parity bits are generated by the SOC and transmitted in serial with mission data (more clocks to transmit same number of bits)
- Reduction in mission data bandwidth and capacity
- Lower cost (no extra component and associated system impacts)

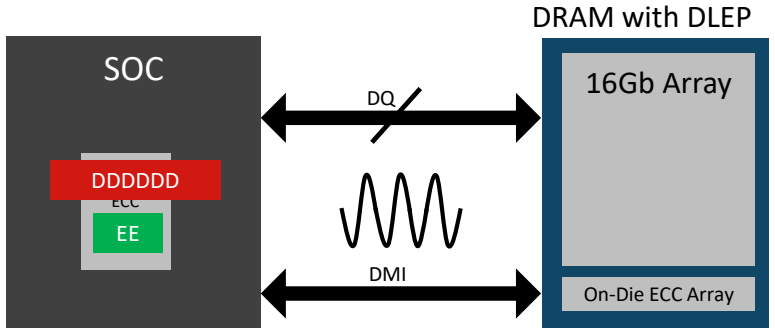
Direct Link ECC Protocol (DLEP)

Higher performance advantage for safety critical systems

EMBEDDED VISION SUMMIT

Highlights

- **Increased effective memory bandwidth**
  - Recovers 15% to 25% of memory BW consumed by inline ECC
- **Increased memory capacity**
  - Up to 11% increase in available memory capacity per device vs standard LP5[3]
- **Reduces power consumption**
  - Approximately 10% lower power (pJ/bit)
- **Significant system BOM savings**
  - $200M+ potential TCO reduction over platform lifetime[1,2]
- **Enables Functional Safety**
  - Compatible with more robust ECC schemes critical for ASIL- D[4]
- **Full Portfolio**
  - Available on all Micron LP5X products



DDDDDD = mission data, EE = ECC parity (ratio of data to parity shown is for conceptual purpose only)

LP5x has a Direct Mask Inversion line

1. $40+ system cost reduction (assumes 2 SoC) from DRAM, SOC and PCB savings in ADAS & IVI for 500GB/s effective memory bandwidth, $200M TCO savings based on 5M car sales over platform lifetime
2. Assumes 20% system bandwidth loss of comparable system using inline ECC
3. 11% capacity increase assumes comparable system using 64+8 inline ECC
4. 256 +16 SECDED transported/stored with DLEP: SBE 100% detect/correct, DBE 100% detect, MBE 99.6% detect

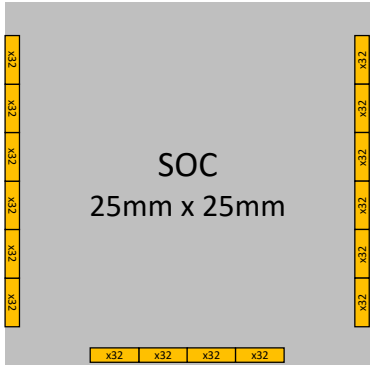
DLEP improvements on 32 channel LP5x bus

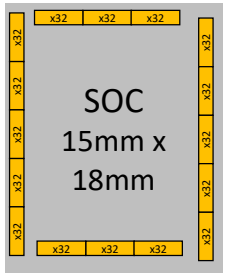
JEDEC System (No DLEP)

Speed	9.6 Gbps
Total CH	32
Density/CH	16 Gb
Total IO	512
Total BW	614 GB/s
Effective BW*	**491 GB/s**
Total Power	19.7 W
Total Capacity	64 GB
Usable Capacity*	**57 GB**
PHY shoreline	56.8 mm
PHY area	46.4 mm^2
Effective BW/mm	8.6 GB/s

DLEP System

Speed	9.6 Gbps
Total CH	32
Density/CH	16 Gb
Total IO	512
Total BW	614 GB/s
Effective BW	**614 GB/s**
Total Power	19.7 W
Total Capacity	64 GB
Usable Capacity	**64 GB**
PHY shoreline	56.8 mm
PHY area	46.4 mm^2
Effective BW/mm	9.6 GB/s





*Assumes system using 64 + 8 SECDED Inline ECC

Micron AI memory and storage portfolio

Leadership products to enable AI workloads

High-bandwidth in-package memory

HBM3E

High-performance graphics memory

GDDR7

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 9550

High-capacity data center NVMe™ SSD

Micron 6500 ION

Summary

Micron memory enables all forms of AI embedded solutions

Smart factory and robotics

Industrial AR/VR

AI-enabled video security and analytics

Low earth orbit (LEO) communication

Smart grid and clean energy

Drones and industrial transport

Inference tasks and model requirements point to memory as a bottleneck

- 200x growth in transformer size vs. memory bandwidth
- Pre- / post-processing is estimated 25x latency vs DNN
- On-chip SRAM is cost prohibitive vs. external DRAM

Memory technology influences AI model execution performance

- Multiple Tasks are used in each industrial application
- Edge deployments will require higher memory density to support AI models
- Memory bandwidth is critical to support real-time task operations
- Generative language memory bandwidth is required for quality of service

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

- LPDDR5 for neural network compute
- LPCAMM2 leverages LPDDR5X performance with DIMM module modularity
- DLEP feature recovers up to 25% memory bandwidth

Resources

- Useful Resources:
  - Micron Technology | Global Leaders in Semiconductors | Micron Technology Inc.
  - AI and Machine Learning | Micron Technology Inc.
- Please visit Micron's booth #303.

Thank You!