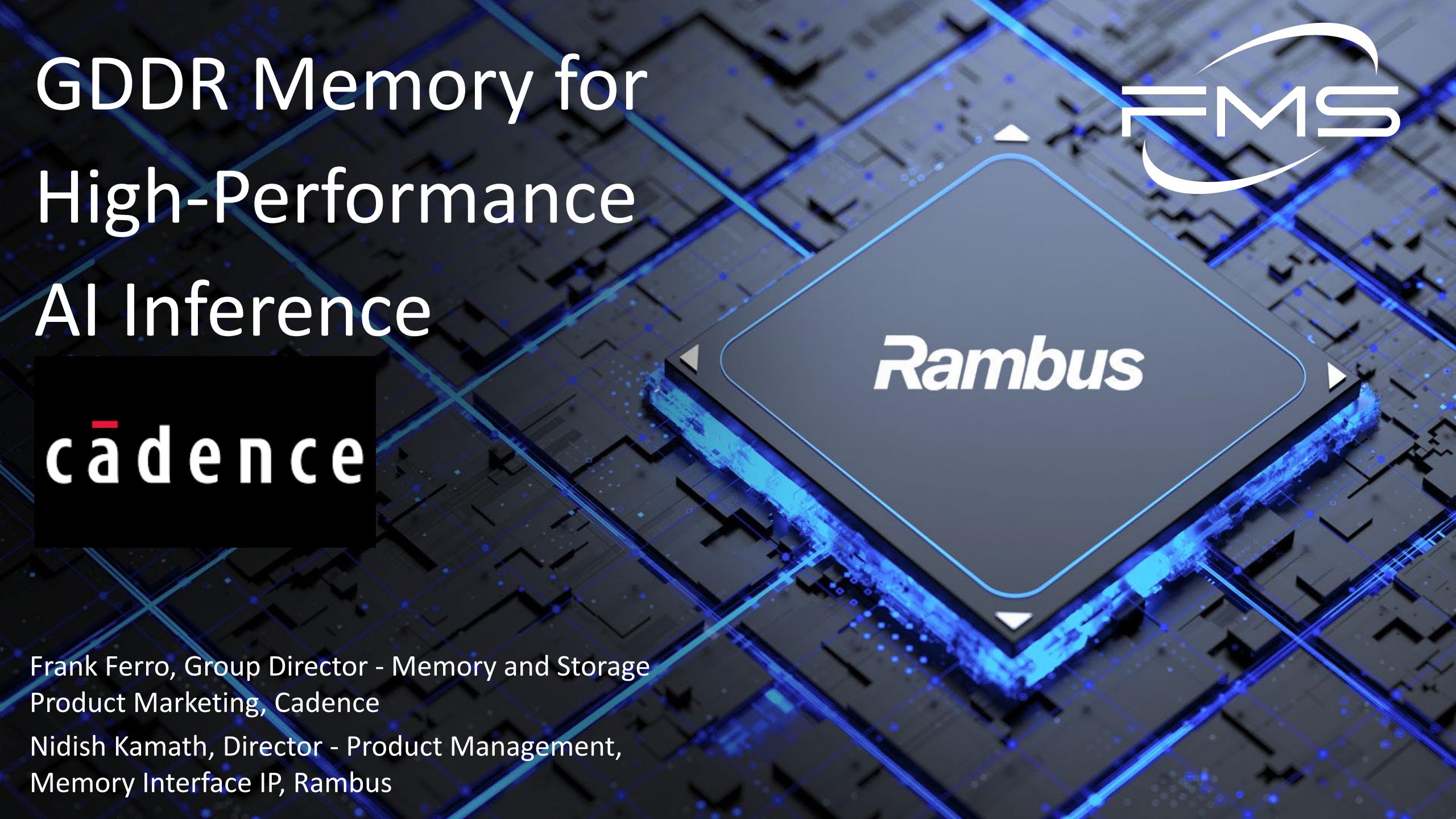
GDDR Memory for High-Performance AI Inference

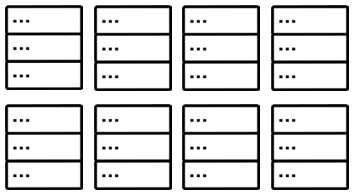


Frank Ferro, Group Director - Memory and Storage Product Marketing, Cadence

Nidish Kamath, Director - Product Management, Memory Interface IP, Rambus

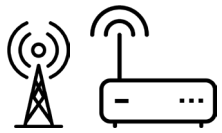
AI Infrastructure Memory Requirements

Cloud



- Analyze broad behaviors, train neural networks for cloud, edge, endpoints
- Need high performance AI training and inference
- Requires the highest performance memories (>1 TB/s) HBM4, HBM3, DDR5

Edge

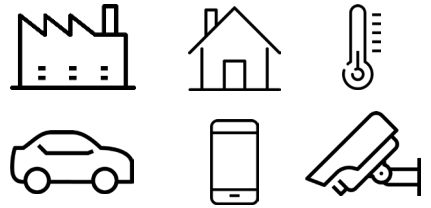


IoT gateways, base stations, etc

On-premise processing

- 5G brings new opportunities for training and inference
- Analyze endpoint data
- Need lower performance training, and cost-efficient AI inference
- Range of memories needed (300-500 GB/s) GDDR7, DDR5, LPDDR5X, HBM3

Endpoints



- Local processing improves security, allows communication of a smaller amount of processed data instead of raw data
- Need cost and power efficient AI inference
- Typically use on-chip memory or low-power mobile memory (50-100 GB/s) LPDDR5x, GDDR6

Infrastructure and data security

Edge Computing Advantages

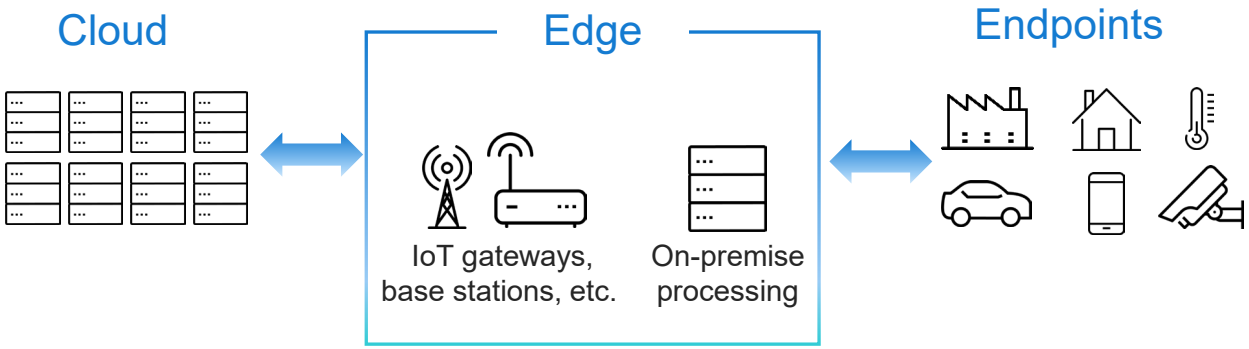
Applications

Sensor Processing:
- Medical imaging
- Weather forecasting
- Security
- Exploration
- Data Recorders

Compute:
- Augmented reality
- Voice recognition
- Image recognition
- Genomics
- High-speed trading

Storage:
- Database Acceleration
- Compression
- Search

Networking:
- Network monitoring
- Load Balancing
- Virtual Machines



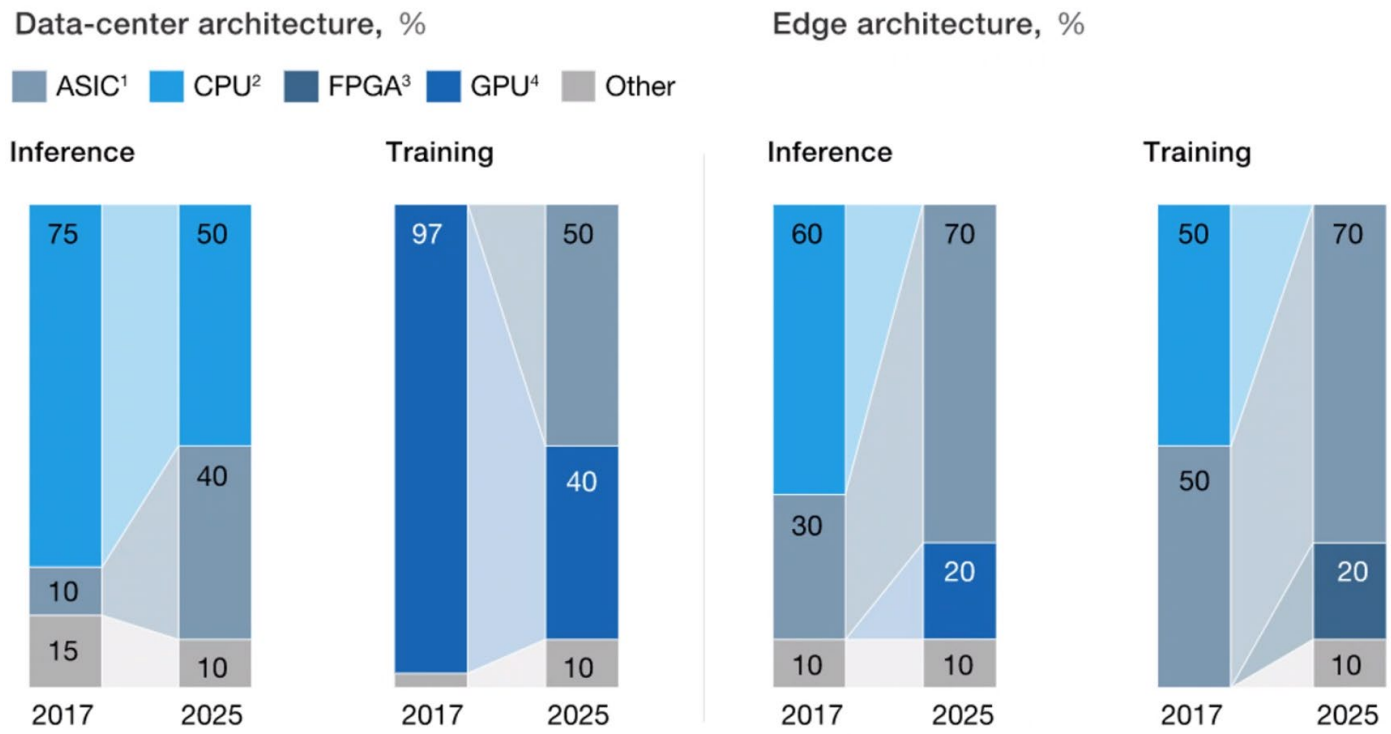
Key Benefits

- Improves energy consumption by minimizing data movement
- Improves security
- AI training performed in the data center then run on the edge
- Reduce transactional latency
- Improve throughput
- Reduces pressure on long-haul links – move smaller amounts of data

AI/ML Driving New Architectures

Advances in computing have pushed bottleneck to memory

- General purpose CPU and GPU cannot reach the performance efficiency of ASIC architectures optimized for specific neural networks
- CPU and GPU hitting the memory wall as the need for bandwidth increases
- The memory subsystem can be optimized with ASIC architectures to achieve the highest bandwidth
- Demand for bandwidth driving innovative memory solutions (2.5D/3D)



Compute moving from GPU to custom ASICs

GDDR7 Standard Overview

JEDEC Publishes GDDR7 Graphics Memory Standard

Cutting-Edge Standard to Enhance Memory Performance in Graphics, Gaming and AI

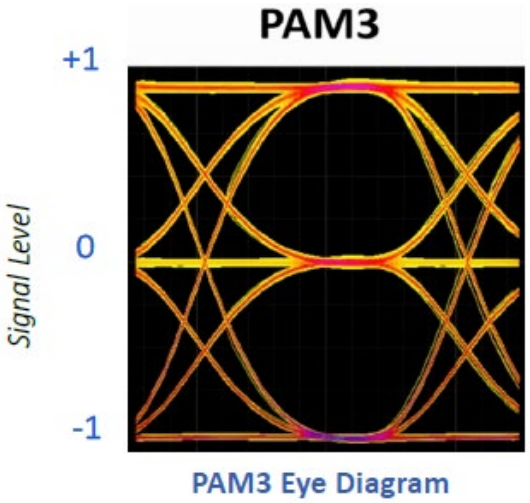
ARLINGTON, Va., USA – MARCH 5, 2024 –JEDEC Solid State Technology Association, the global leader in the development of standards for the microelectronics industry, is pleased to announce the publication of JESD239 Graphics Double Data Rate (GDDR7) SGRAM. This groundbreaking new memory standard is available for free download from the JEDEC website. JESD239 GDDR7 offers double the bandwidth over GDDR6, reaching up to 192 GB/s per device, and is poised to meet the escalating demand for more memory bandwidth in graphics, gaming, compute, networking and AI applications.

JESD239 GDDR7 is the first JEDEC standard DRAM to use the Pulse Amplitude Modulation (PAM) interface for high frequency operations. Its PAM3 interface improves the signal to noise ratio (SNR) for high frequency operation while enhancing energy efficiency. By using 3 levels (+1, 0, -1) to transmit 3 bits over 2-cycles versus the traditional NRZ (non-return-to-zero) interface transmitting 2 bits over 2-cycles, PAM3 offers higher data transmission rate per cycle resulting in improved performance.

Specification	GDDR6	GDDR7
Per bit Bandwidth	24 Gbps	36 Gbps (48 Gbps spec max)
Total Bandwidth (32bit)	768 Gbps	1.15 Tbps
Chip Density (max)	32 Gb	32 Gb
Signaling	NRZ	PAM3/NRZ
Channels/data	2 x 16bit	4 x10bit (8bit data, 2bit error)

GDDR7 Features (vs. GDDR6)

- Three-level pulse amplitude modulation (PAM3) - 3 bits of information in 2 cycles
- 50% increase in data transmission compared at same clock rate (48G max spec)
- RAS features: on die ECC, error check and scrub, command address parity with command blocking



PAM3 Eye Diagram

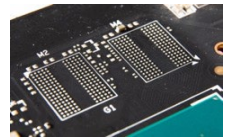
Memory Choices for AI at the Edge

- Needs AI Inference and low-end training
- High performance at the lowest cost
- GDDR6 (768 Gb/s) and GDDR7 (1152 Gb/s) offer the highest performance at a good price point
- Lower system cost – standard PCB and package

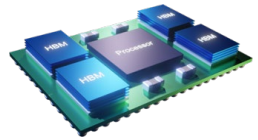
Memory Protocol	Density	Cost/ Performance	Bandwidth	Power Consumption
DDR4/5				
LPDDR5X				
GDDR6/7				
HBM2E/3E				

Edge AI Memory Performance

Traditional DDR DIMM or on PCB



GDDR on PCB (no 2.5D needed)

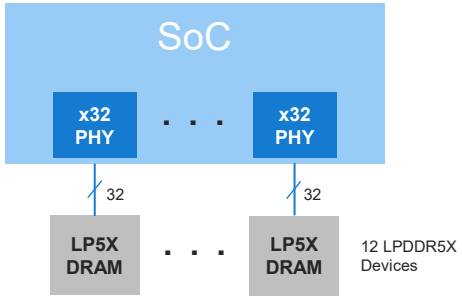


HBM Stacked DRAM – 2.5/3D

Key: Best Good

AI Inference example: Target Memory Bandwidth 500GB/s

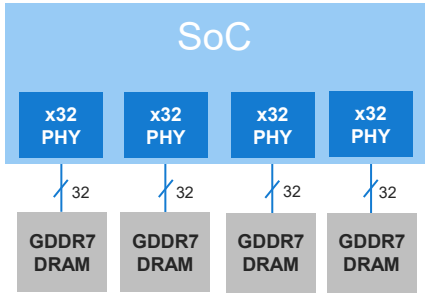
LPDDR5X (9.6 Gbps)



System Bandwidth = 461 GB/s

- 12 x 32 LPDDR DRAMs @ 9.6 Gbps (or 6 x 64)
- Standard PCB manufacturing (384 data pins)
- Good memory density
- Larger board and die area (cost = 2x)

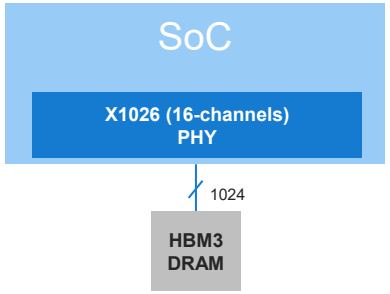
GDDR7 (32 Gbps)



System Bandwidth = 512 GB/s

- 4 GDDR7 DRAM @ 32 Gbps
- Standard PCB manufacturing (128 data pins)
- Lower board area
- Lower overall system cost (cost = 1x)

HBM3 (6.4 Gbps)



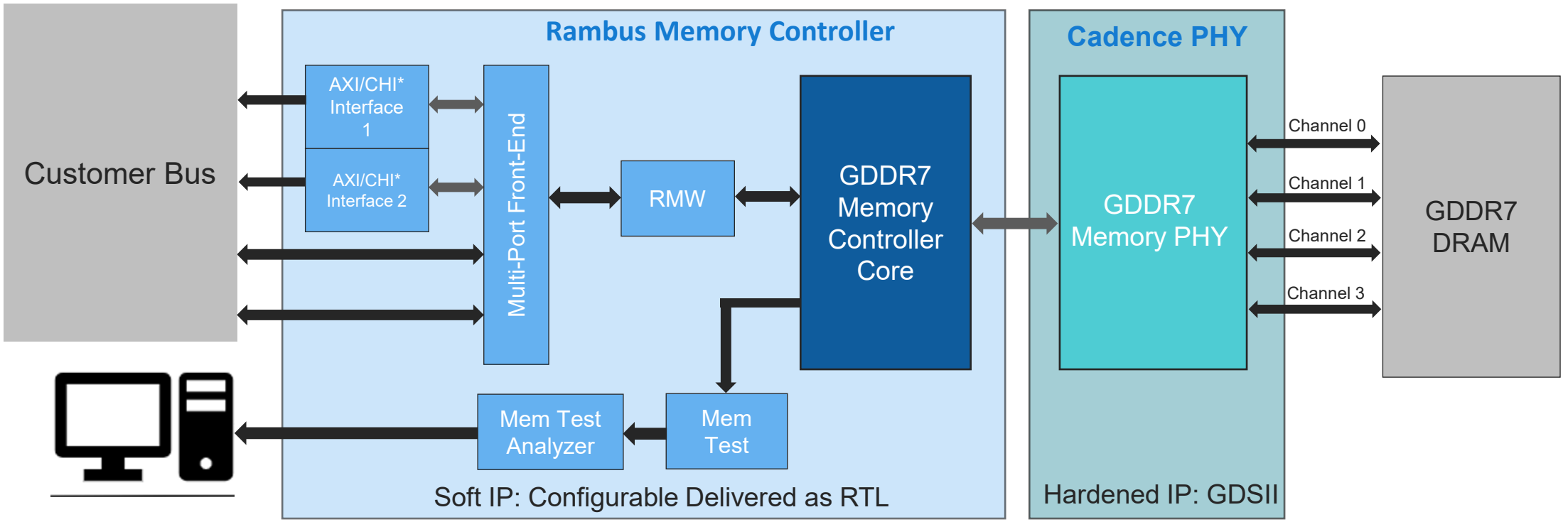
System Bandwidth = 820 GB/s

- 1 HBM3 PHYs/DRAM @ 6.4 Gbps
- 2.5D manufacturing (1024 data pins)
- Low area and power
- Higher cost (cost = 4x)

GDDR7 Memory Subsystem from Cadence and Rambus

- Configurable solution to customer specification
- Fully validated PHY and Controller
- Controller delivered as soft RTL
- Hardened timing closed PHY
- Silicon validated
- Support for design and bring-up

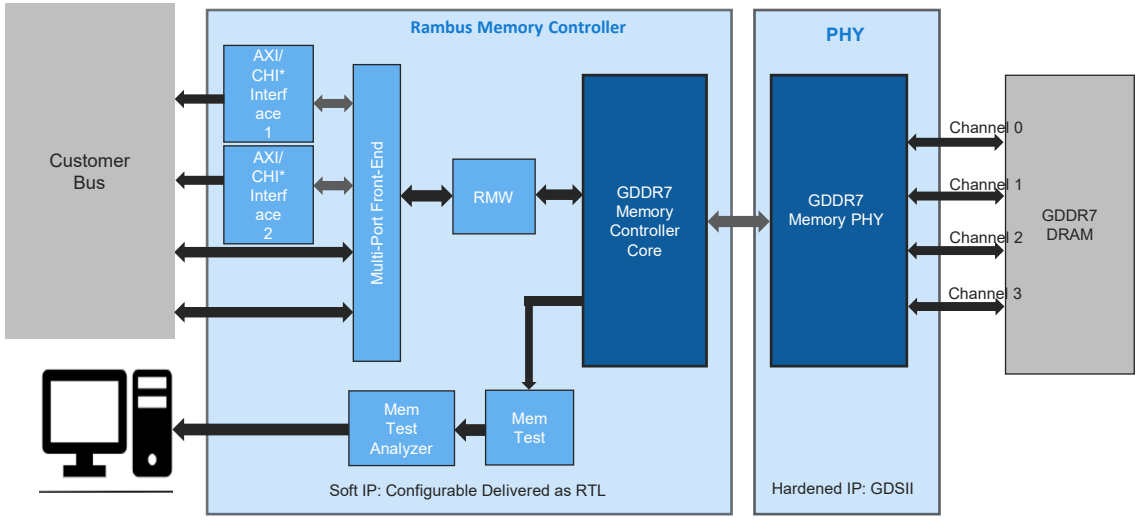
Integrated Memory Subsystem



*For CHI interface options, contact Rambus Sales

GDDR7 Controller Overview

- Supports all standard GDDR7 features
  - All speed grades up to 40 Gbps
  - All bank and per bank refresh
  - EDC, low power modes (self-refresh, power-down)
  - Supports x16 or x8 clamshell modes
- Quad-Controller design
  - One controller per channel
  - Half-rate operation
    - 1.25 GHz controller clock for 40 Gbps operation
  - User interface data width is 32x memory width (i.e., 256 bits for 8-bit memory)
- Optional Add-On Cores – AXI/CHI*, Multi-Port, Memory Test (offers comprehensive Memory Sub-System test support), etc.
- Rambus performs complete regression of each customer's Memory Controller solution + PHY during delivery



*For CHI interface options, contact Rambus Sales

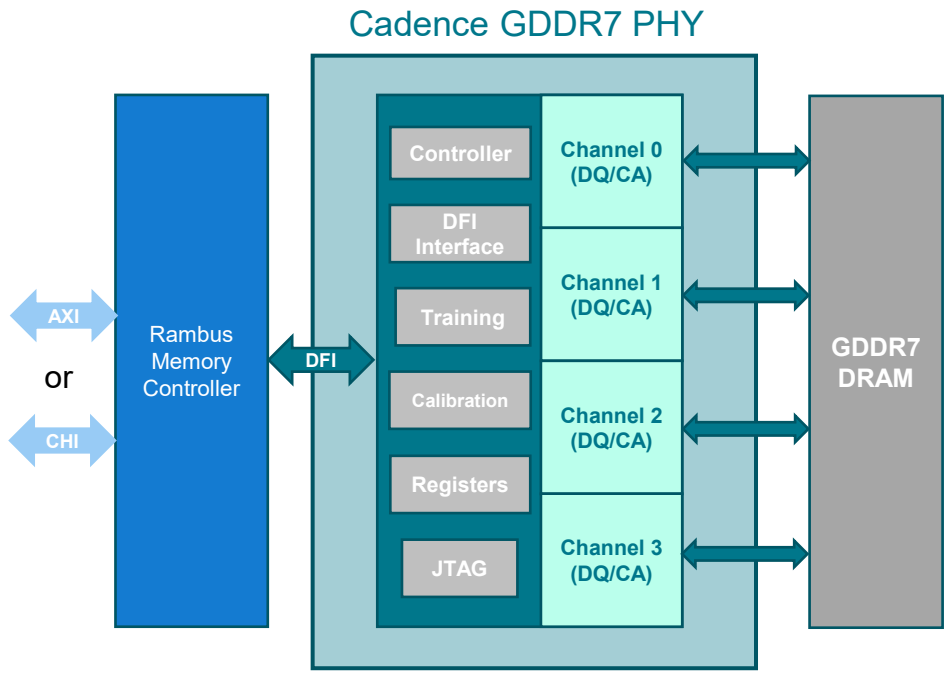
Cadence GDDR7 PHY Features

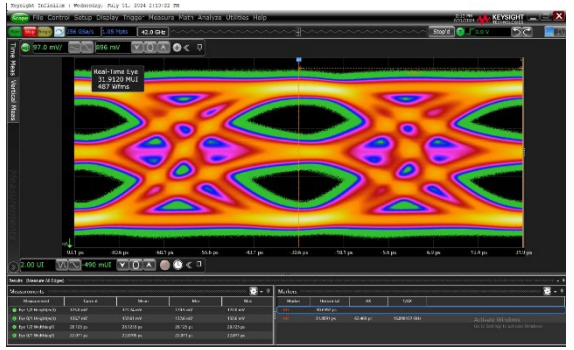
Key Advantages:

- Highest performance at 36Gb/s
- Fully Hardened timing closed PHY
- System design support: PCB and package reference designs
- SI/PI expertise
- Available in advanced nodes at multiple foundries

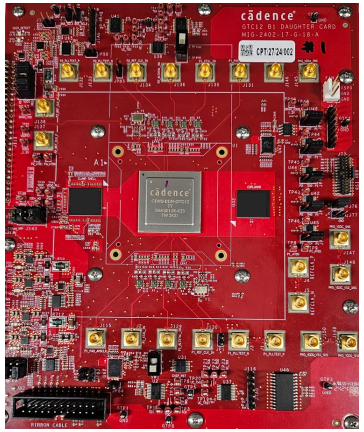
Key Features:

- 1.2Tb/s bandwidth per PHY
- PAM3 signaling or NRZ
- 4 independent channels
- PHY independent mode
- Microcontroller or state machine training
- Low power clock gating
- Advanced equalization





PAM3 TX eye @32Gbps (PRBS)

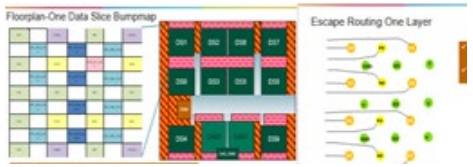


GDDR7 Development Hardware

Cadence Complete System Design Flow

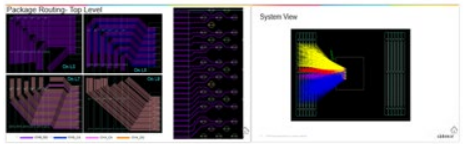
Support for Signal Integrity Challenges

Floor Plan/Bump map



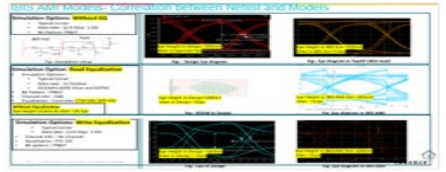
Achieve balance between area, pitch, escape routing, decap

PKG/Board Routing



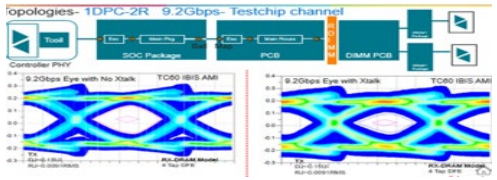
Optimal package board routing solution and stack-up

I/O Models



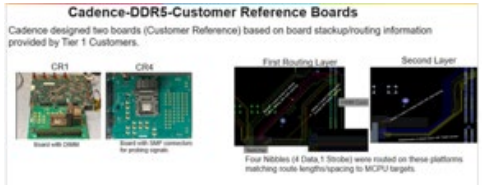
Best in class IBIS and PDN Models

SI/PI Analysis



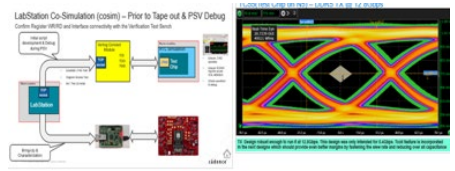
Signal and Power Integrity analysis in Cadence Tools

Reference Platform



Enable customers with proven reference designs

Post Silicon Validation



Fast silicon bring-up and validation

Summary

- Advanced AI applications are demanding far greater memory bandwidth
- AI inference is a growing market segment with many applications at the edge and endpoints of the network
- GDDR6 and GDDR7 offer the best trade-off of bandwidth, cost and density for AI inference
- Cadence and Rambus working in collaboration for test chip and system validation to ensure best performance for GDDR6 and GDDR7 systems
- Cadence and Rambus offer a complete validated GDDR6 and GDDR7 PHY/Controller memory subsystem
- Cadence provides PCB and package design support