Rambus

INTERFACE IP
White Paper

HBM2E and GDDR6: Memory Solutions for AI

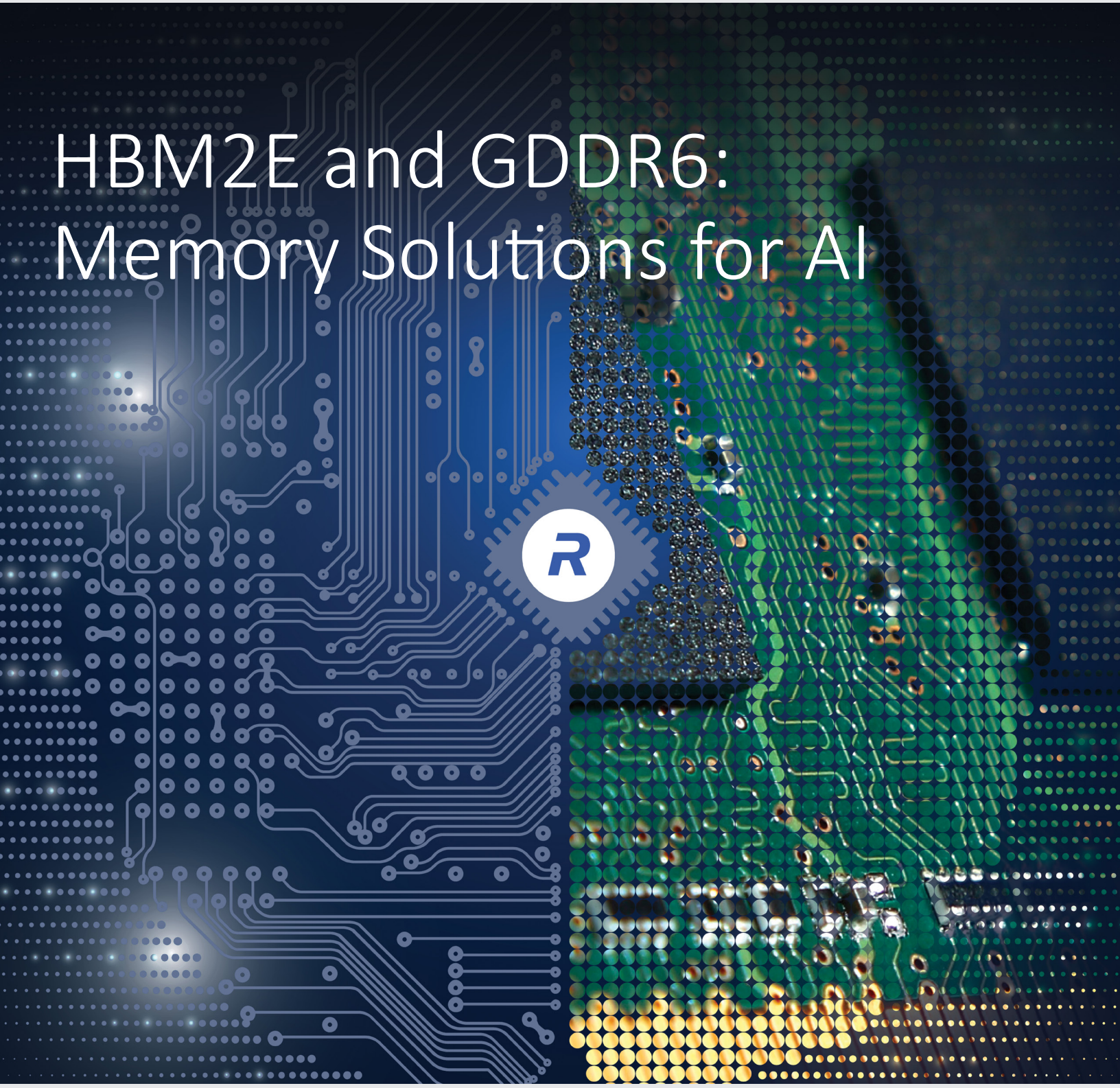
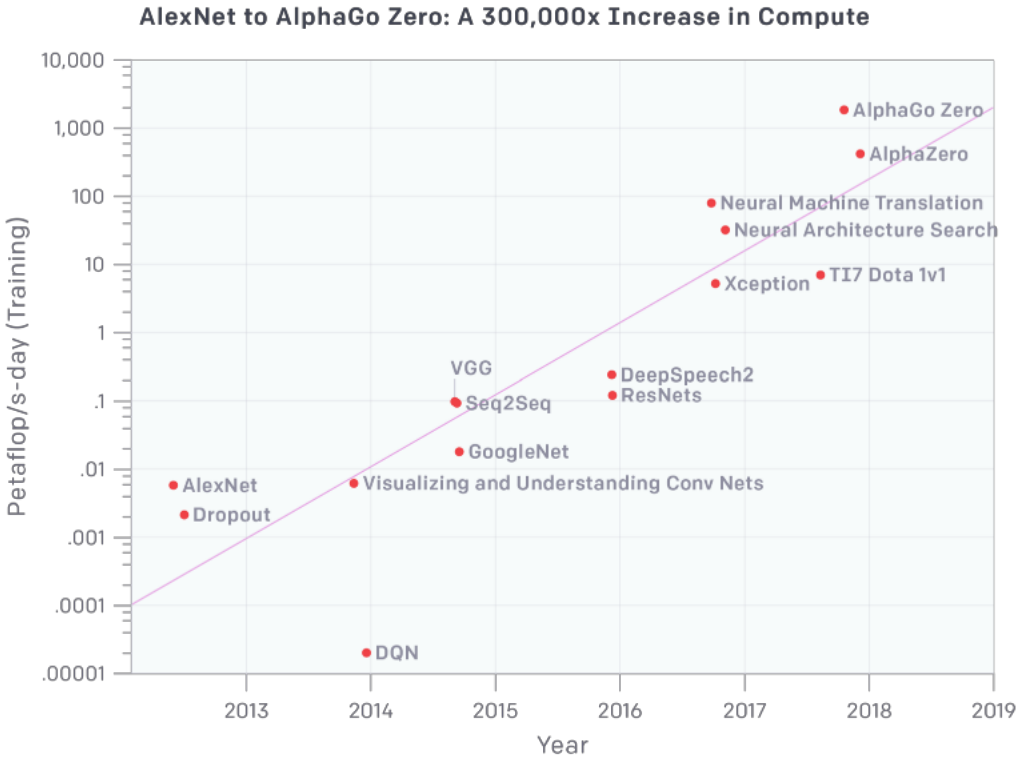


Table of Contents

Introduction

Artificial intelligence/machine learning (AI/ML) changes everything, impacting every industry and touching the lives of everyone. AI is catalyzing breathtaking growth across a broad spectrum of technology markets, from 5G to IoT. This is powerfully illustrated in the growth of AI training sets which have increased by a factor of 300,000 from 2012 to 2019, a doubling every 3.43 months. Supporting this pace requires far more than the improvements that can be realized through Moore's Law, which is slowing in any case, necessitating rapid ongoing improvements in every aspect of AI computer hardware and software.

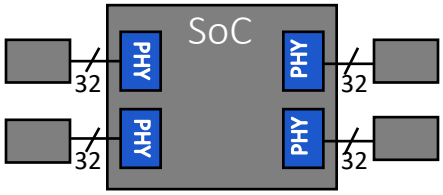


Training capability has grown by a factor of 300,000 from 2012 to the present

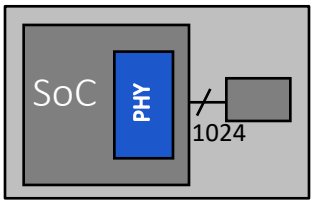
Source: https://openai.com/blog/ai-and-compute/ May 16, 2018

Memory bandwidth will be one such critical area of focus enabling the continued growth of AI. Take as an example advanced driver-assistance systems (ADAS). Complex data processing at Level 3 and higher systems requires memory bandwidths exceeding 200 GB/s. These high bandwidths are essential for the complex AI/ML algorithms needed to rapidly execute massive calculations and safely implement real-time decisions on the road. At Level 5, or full autonomy, vehicles capable of independently reacting to a dynamic environment of traffic signs and signals, as well as accurately predicting the movements of cars, trucks, bikes and pedestrians, will take enormous memory bandwidth. Concurrent with the rapid development of a new generation of AI/ML accelerators and specialized silicon has seen the adoption of new memory solutions such as High Bandwidth Memory (HBM, HBM2, HBM2E) and GDDR6 SDRAM (GDDR6) to deliver this necessary bandwidth.

GDDR6 Memory System
Four 16Gbps x32 GDDR6 DRAMs



HBM2E Memory System
Single 2Gbps HBM2E Device



Memory Solution	GDDR6	HBM2E	
Total Bandwidth	256 GB/s	256 GB/s	
Per-pin data rate	16 Gbps	2 Gbps	
Relative Controller PHY Area	1.5-1.75	1.0	Area advantage for HBM2
Relative Controller PHY Power	3.5-4.5	1.0	Power advantage for HBM2
Interposer	None	Added cost	Cost and complexity advantage for GDDR6
Memory	Similar to GDDR5, DDR4	Stacked, adds cost	Cost advantage for GDDR6

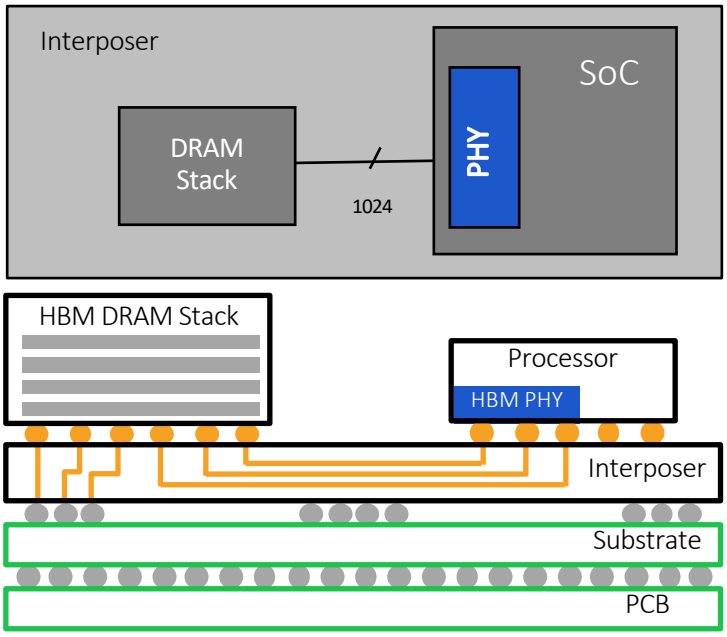
GDDR6 and HBM2E offer differing benefits and design trade-offs

When choosing between HBM2E (the latest generation of HBM) and GDDR6 for AI/ML applications, designers must consider a number of trade-offs and key metrics including cost, power, capacity and implementation complexity. In this white paper, we'll explore the benefits and design considerations for HBM2E and GDDR6. We'll also highlight the applicability of each memory in the overall AI/ML architecture. Lastly, we'll discuss Rambus' HBM2E and GDDR6 interface solutions which can be used to implement a complete memory subsystem.

Part 1: HBM2E Memory

Introduced in 2013, High Bandwidth Memory (HBM) is a high-performance 3D-stacked SDRAM architecture. Like its predecessor, HBM2 specifies up to 8 memory die per stack, while doubling pin transfer rates to 2 Gbps. HBM2 achieves 256 GB/s of memory bandwidth per package (DRAM stack), with the HBM2 specification supporting up to 8 GB of capacity per package.

In late 2018, JEDEC announced the HBM2E specification to support increased bandwidth and capacity. With transfer rates rising to 3.6 Gbps per pin, HBM2E can achieve 461 GB/s of memory bandwidth per stack. In addition, HBM2E supports 12-high stacks with memory capacities of up to 24 GB per stack.



HBM2E Memory System with Single DRAM Stack

All versions of HBM run at a relatively low data rate, but achieve very high bandwidth through the use of an extremely wide interface. Specifically, each HBM2E stack running at up to 3.6 Gbps connects to its associated processor through an interface of 1,024 data “wires.” With command and address, the number of wires grows to about 1,700. This is far more than can be supported on a standard PCB. Therefore, a silicon interposer is used as an intermediary to connect memory stack(s) and processor. Like with an SoC, finely spaced data traces can be etched in the silicon interposer to achieve the desired number of wires needed for the HBM interface.

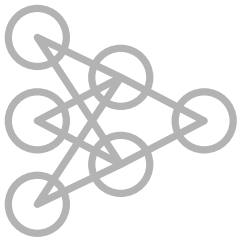
HBM2E offers the capability to achieve tremendous memory bandwidth. Four HBM2E stacks connected to a processor will deliver over 1.8 TB/s of bandwidth. And with 3D stacking of memory, high bandwidth and high capacity can be achieved in an exceptionally small footprint. Further, by keeping data rates relatively low, and the memory close to the processor, overall system power is kept low.

The design trade-off with HBM is increased complexity and costs. The interposer is an additional element that must be designed, characterized and manufactured. 3D stacked memory shipments pale in comparison to the enormous volume and manufacturing experience built up making traditional DDR-type memories (including GDDR). The net is that implementation and manufacturing costs are higher for HBM2E than for GDDR6.

Outstanding bandwidth, capacity and latency in a power-efficient, compact footprint make HBM2E memory a superior choice for AI training hardware.

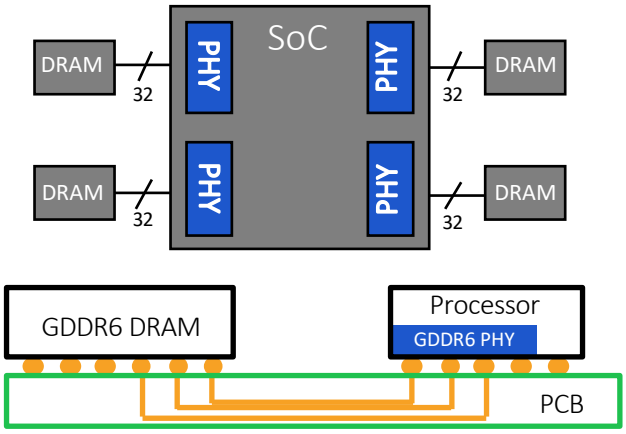
However, for AI training applications, the benefits of HBM2E make it the superior choice. The performance is outstanding, and higher implementation and manufacturing costs can be traded off against savings of board space and power. In data center environments, where physical space is increasingly constrained, HBM2E's compact architecture offers tangible benefits. Its lower power translates to lower heat loads for an environment where cooling is often one of the top operating costs.

In summary, HBM2E offers system designers extremely high-bandwidth capabilities and optimal power efficiency. While implementation of HBM2E systems can be challenging due to greater design complexity and manufacturing costs, savings in board space and cooling can be compelling. For AI training, HBM2E is an ideal solution. It builds on a strong track record of success with HBM2 which was implemented in AI processors such as NVIDIA's Tesla A100 and the second-generation Google TPU.



Part 2: GDDR6 Memory

Graphics DDR SDRAM (GDDR SDRAM) was originally designed for the gaming and graphics market over two decades ago. GDDR has undergone several major evolutions in that time, with the latest generation, GDDR6 delivering data rates of 16 Gbps. GDDR6 offers an impressive combination of bandwidth, capacity, latency and power. It lowers the operating voltage to 1.35V from 1.5V for greater power efficiency, and doubles the data rate (16 vs. 8 Gbps) and capacity (16 vs. 8 GB) of GDDR5 memory. Rambus has already demonstrated a GDDR6 interface running at 18 Gbps showing there's additional headroom for this memory architecture.



GDDR6 Memory System with Four DRAM

Unlike HBM2E, GDDR6 DRAM relies on the same high-volume manufacturing and assembly techniques used to produce standard DDR-type DRAM. More specifically, GDDR6 employs the traditional approach of connecting packaged and tested DRAMs together with an SoC through a standard PCB. Leveraging existing infrastructure and processes offers system designers a familiarity that reduces cost and implementation complexities.

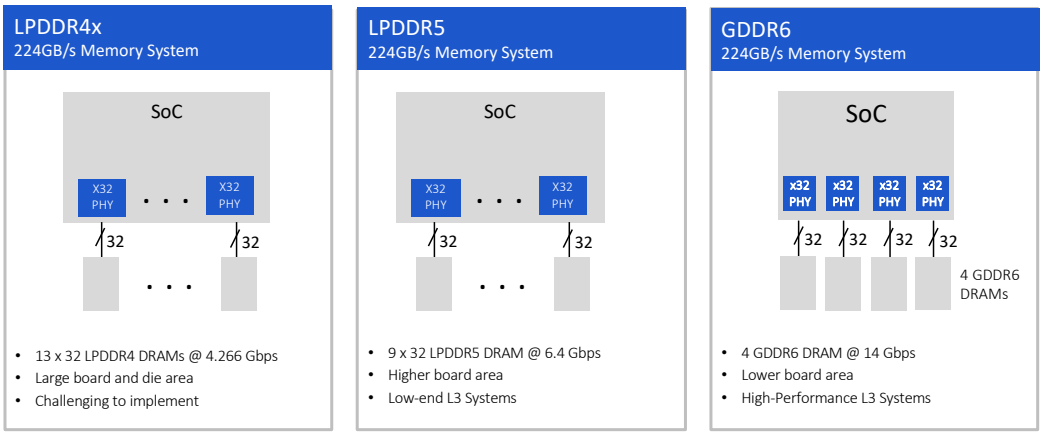
The excellent price-performance of GDDR6 memory, built on time-tested manufacturing processes, make it a great choice for AI inference applications.

As opposed to HBM2E's wide and slow memory interface, the GDDR6 interface is narrow and fast. Two 16-bit wide channels (32 data wires) connect a GDDR6 PHY to an associated SDRAM. Running at 16 Gbps per pin, a GDDR6 interface can deliver a bandwidth of 64 GB/s. Returning to our earlier L3 automotive example, a GDDR6 memory system could hit >200 GB/s bandwidth with four DRAM devices.

The principal design challenge for GDDR6 implementations derives from one of its strongest features: its speed. Maintaining signal integrity (SI) at speeds of 16 Gbps, particularly at lower voltages, requires significant expertise. Designers face tighter timing and voltage margins while the number of sources of loss and their effects all rise rapidly. Interdependencies between the behavior of the interface, package and board require a method of co-design of these components in order to maintain the SI of the system.

On balance, the excellent performance characteristics of GDDR6 memory, built on tried-and-true manufacturing processes, is an ideal memory solution for AI inference. Its price-performance characteristics make it suitable to volume deployment across a broad array of edge network and IoT end-point devices.

L3 ADAS Memory System Implementation Examples



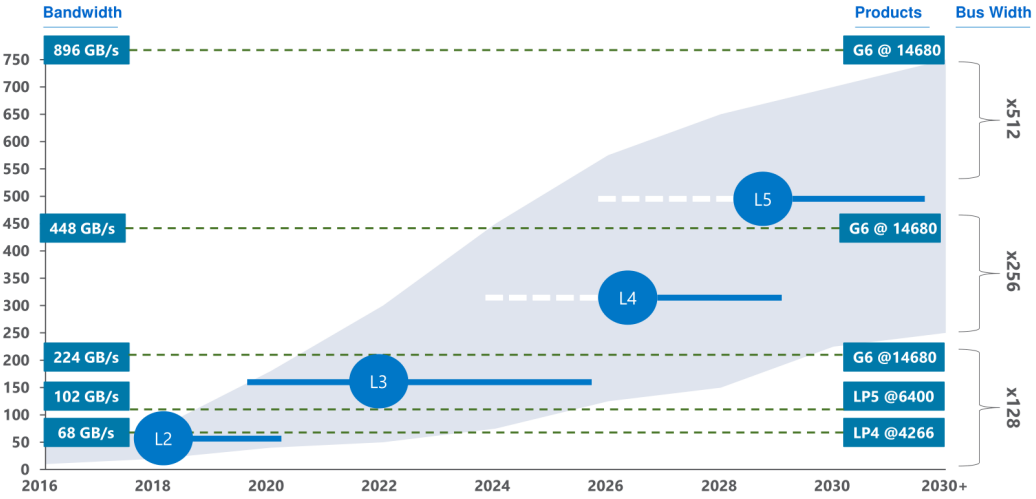
GDDR6 Provides the Best Memory Design and Utilization Efficiency

There may be no more demanding "IoT" AI-inference application than ADAS. Qualification standards in a system responsible for protecting life and property are necessarily high. The net result is that road-tested memory architectures such as LPDDR (with billions of mobile phone deployments) and GDDR6 have seen implementation in early ADAS systems.

As the chart above illustrates, LPDDR4/5 memory architectures can achieve the 200GB/s bandwidth threshold of L3 ADAS systems, but it requires a large number of DRAM devices to do so. GDDR6 is far more efficient from a design and utilization standpoint requiring fewer than half the number of devices to achieve the desired system bandwidth. As bandwidth requirements rise to meet the requirements of L4 and L5 ADAS, GDDR6 becomes the only viable alternative.

As shown in the figure below, at L4 ADAS, bandwidth requirements rise to 300 GB/s. With an LPDDR5 interface running at 6.4 Gbps, 12 DRAM devices would be required to hit that mark. The beachfront of the SoC would be dominated by memory interfaces which would be impractical and complicate the logic layout of the SoC. GDDR6 running at 16 Gbps can provide over 300 GB/s of bandwidth with only five devices and hit the greater than 500 GB/s for L5 ADAS with eight.

ADAS Memory Bandwidth Requirements



Memory Bandwidth Requirements Grow Rapidly with Higher Level ADAS

Source : https://www.anandtech.com/show/12362/micron-rambus-others-team-up-to-spur-gddr6-adoption

In summary, GDDR6 offers a great combination of bandwidth, capacity, power efficiency, reliability and price-performance. With a trusted partner like Rambus, SoC designers can realize all these benefits while tackling the SI challenges presented by operation at speeds of 16 Gbps and beyond.

Part 3: HBM2E and GDDR6 – Partners in AI

Given the bifurcated nature of AI/ML, the choice of memory depends on the application: training or inference. Rather than a question of "or," it's one of "and" as both of these high-bandwidth memories, HBM2E and GDDR6, can play a vital role.

For training, bandwidth and capacity are critical requirements. This is particularly so given that training sets are on a pace to double in size every 3.43 months, as we discussed earlier. Training workloads now run over multiple servers to provide the needed processing power, flipping virtualization on its head. Given the value created through training, there is a powerful "time-to-market" incentive to complete training runs as quickly as possible. Furthermore, training applications run in data centers increasingly constrained for power and space, so there's a premium for solutions that offer power efficiency and smaller size.

Given all these requirements, HBM2E is an ideal memory solution for AI training hardware. It provides excellent bandwidth and capacity capabilities: 461 GB/s of memory bandwidth with 24 GB of capacity for a single 12-high HBM2E stack. Its 3D structure provides these features in a very compact form factor and at a lower power thanks to a low interface speed and proximity between memory and processor.

Training	Inference
Bandwidth Capacity Power Efficiency Compact Size	Bandwidth Latency Price-Performance
HBM2E Memory	GDDR6 Memory

In the case of inference, bandwidth and latency are critical driven by the need to act in real time. With inference being deployed across a broad spectrum of edge and IoT end point devices, implementations will be more cost sensitive than those at the heart of the data center. In addition, for ADAS the memory will need to be built on road-tested technologies and manufacturing processes in order to meet the rigorous qualification requirements. With the ongoing roll out of 5G, there will be a growing number of AI-powered, untethered devices performing complicated inference.

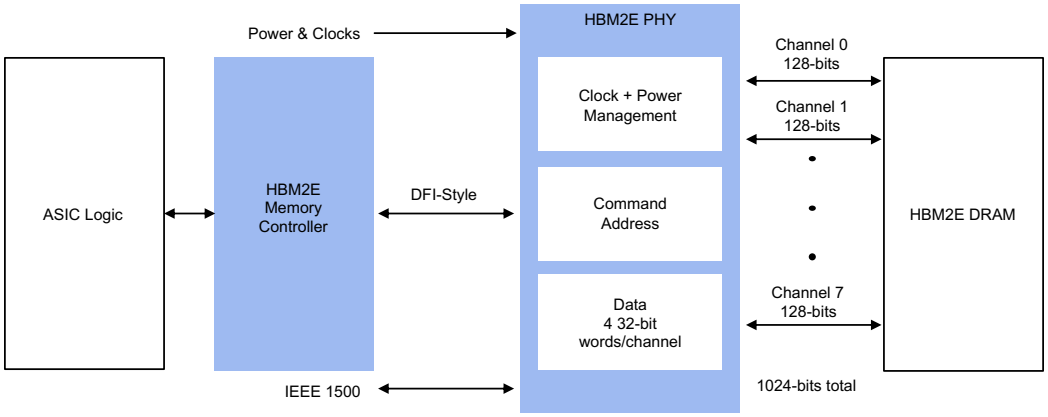
For this increasingly challenging landscape for AI inference, GDDR6 is an ideal solution. It can provide excellent bandwidth with a single or small number of DRAM devices: 64GB/s of memory bandwidth per device at 16 Gbps data rate. Built on mature manufacturing processes, it delivers the price-performance characteristics suitable to volume deployments.

The upshot is that AI/ML is not monolithic, and both training and inference require memory solutions tailored to their specific requirements. HBM2E and GDDR6 meet the needs of training and inference respectively delivering a strong set of benefits that these applications demand. As discussed earlier, both HBM2E and GDDR6 present design challenges to implementation. But with solutions from a trusted partner such as Rambus, the benefits of these memories can be readily achieved. In the next sections, we'll review the HBM2E and GDDR6 interface solutions available from Rambus.

Part 4: The Rambus HBM2E Memory Interface Solution

Optimized for high bandwidth and low latency, the Rambus HBM2E interface delivers maximum performance and flexibility in compact form factor and power-efficient envelope. The interface consists of a co-verified PHY and digital controller comprising a complete HBM2E memory subsystem.

The Rambus HBM2E interface is fully compliant with the JEDEC JESD235B standard. It supports data rates up to 3.6 Gbps per data pin. The interface features 8 independent channels, each containing 128 bits for a total data width of 1024 bits. The resulting bandwidth is 461 GB/s per stack, with the stack consisting of 2, 4, 8 or 12 DRAMs.



HBM2E Memory Interface Subsystem Example

The interface is designed for a 2.5D system with an interposer used for routing signals between the 3D DRAM stack and the PHY on the SoC. This combination of signal density and stacked form factor requires special design consideration. In order to enable easy implementation and improved flexibility of design, Rambus performs complete signal and power integrity analysis on the entire 2.5D system to ensure that all signal, power, and thermal requirements are met.

Additional key features include:

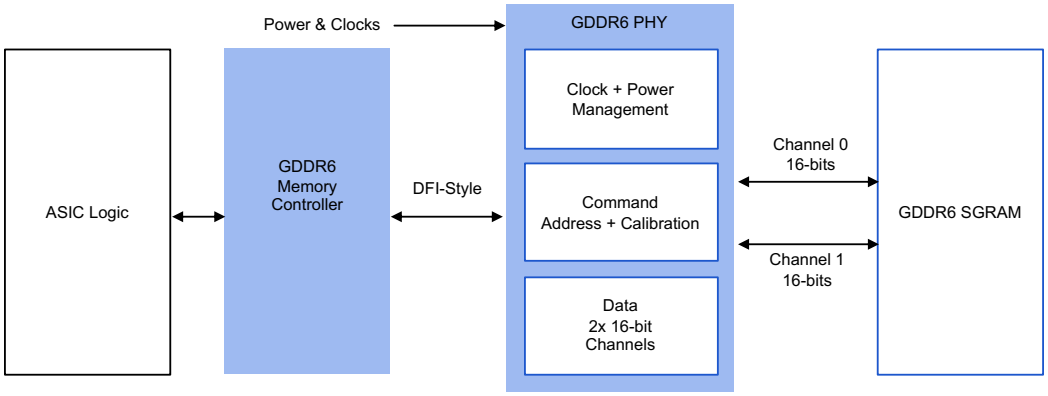
- Co-verified PHY and digital controller
- Speed bins: 0.5, 1.0, 1.5, 1.6, 1.8, 2.0, 2.4, 3.0, 3.2, 3.6 Gbps
- 8 channels and 16 pseudo-channels
- Support for 2, 4, 8 or 12 DRAM stacks
- Supports all standard HBM2E channel densities (4, 6, 8, 12, 16, 24 Gb)
- Memory controller or PHY can be ASIC interface master (PHY independent mode)
- Selectable low-power operating states
- Programmable output impedance
- Pin programmable support for lane repair
- ZQ calibration of output impedance
- IEEE 1500 test support
- Autonomous test support
- SSO noise reduction
- Micro-bump pitch matched to the DRAM pitch
- Utilizes 13 or 15-layer metal stack
- East-West orientation (PHY can be placed in corner of die)
- Register interface for state observation
- LabStation™ software environment for system level bring-up, characterization, and validation

Part 5: The Rambus GDDR6 Memory Interface Subsystem

Designed for performance and power efficiency, the Rambus GDDR6 interface supports the high-bandwidth, low-latency requirements of AI/ML and ADAS inference. It consists of a co-verified PHY and digital controller providing a complete GDDR6 memory subsystem.

The Rambus GDDR6 interface is fully compliant with the JEDEC GDDR6 JESD250 standard, supporting up to 16 Gbps per pin. The GDDR6 interface supports 2 channels, each with 16 bits for a total data width of 32 bits. At 16 Gbps per pin, the Rambus GDDR6 interface offers a bandwidth of 64 GB/s.

Rambus works directly with customers to provide full-system signal and power integrity (SI/PI) analysis, creating an optimized chip layout. Customers receive a hard macro solution with a full suite of test software for quick turn-on, characterization and debug.



GDDR6 Memory Interface Subsystem Example

Additional key features include:

- Co-verified PHY and digital controller
- Flexible delivery of IP core (works with ASIC/SoC layout requirements)
- Speed bins: 12, 14, and 16 Gbps. Rambus has demonstrated 18 Gbps for future scalability.
- Two 16-bit Channels
- Support for GDDR6 SGRAM
- Memory controller or PHY can be ASIC interface master (PHY independent mode)
- Selectable low-power operating states
- Programmable driver/termination impedance value
- Driver/termination impedance calibration
- In-built test support
- Utilizes 13-layer metal stack
- Register interface for state observation
- LabStation™ software environment for system level bring-up, characterization, and validation

Conclusion

AI/ML's evolution proceeds at a lightning pace. Training capabilities are growing at a rate of 10X per year driving rapid improvements in every aspect of computing hardware and software. Meanwhile, AI inference is being deployed across the network edge and in a broad spectrum of IoT devices including in automotive/ADAS. Training and inference have unique application requirements that can be served by tailored memory solutions with HBM2E being ideal for the former and GDDR6 for the latter. Designers can realize the benefits of these high-performance memories by working with Rambus to overcome the design challenges inherent in these architectures. Rambus offers comprehensive HBM2E and GDDR6 memory interface solutions ready for integration into AI/ML training and inference SoCs.

For more information, visit
https://www.rambus.com/interface-ip