Rambus

INTERFACE IP
White Paper

HBM3E Memory: Break Through to Greater Bandwidth

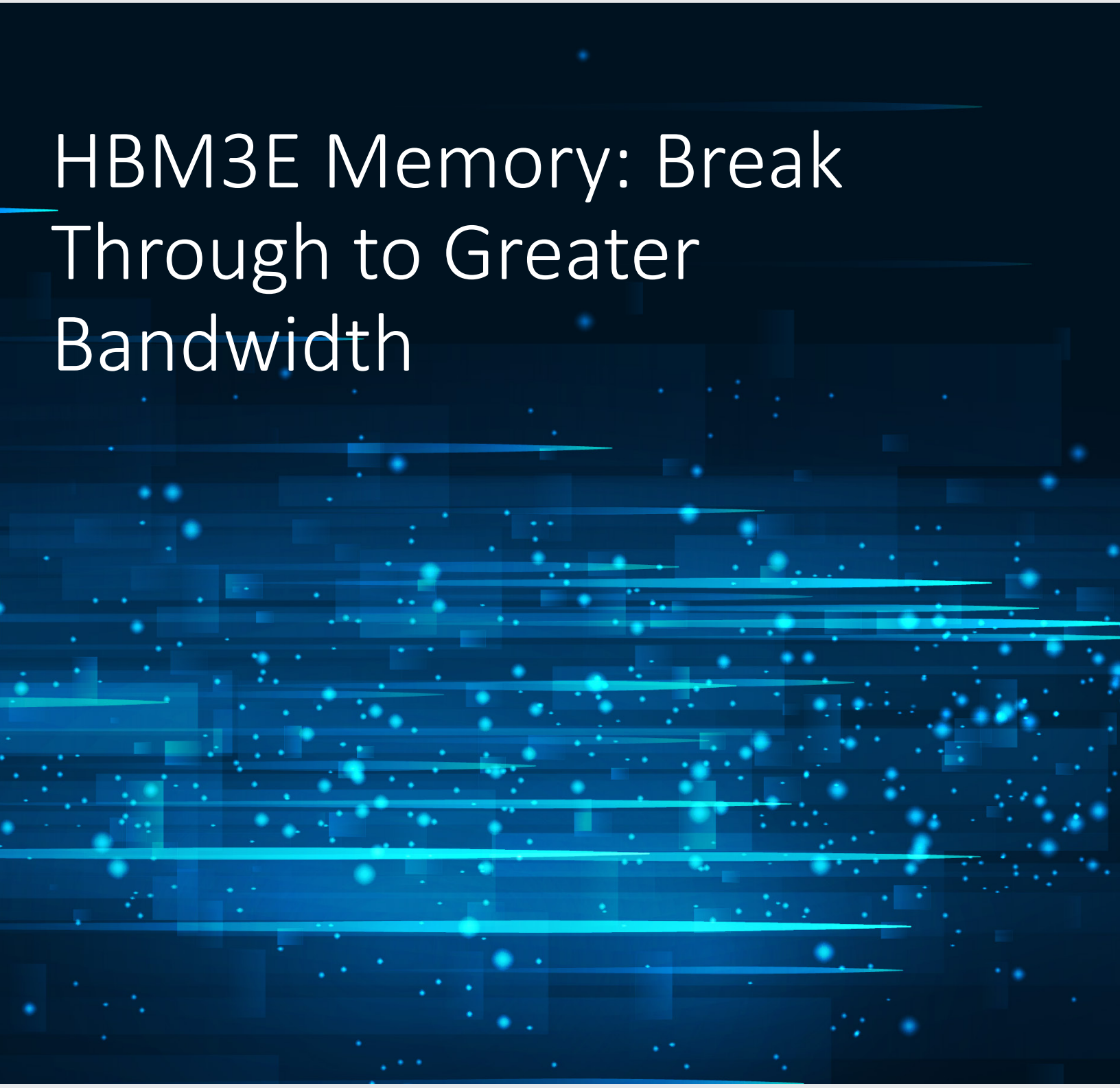
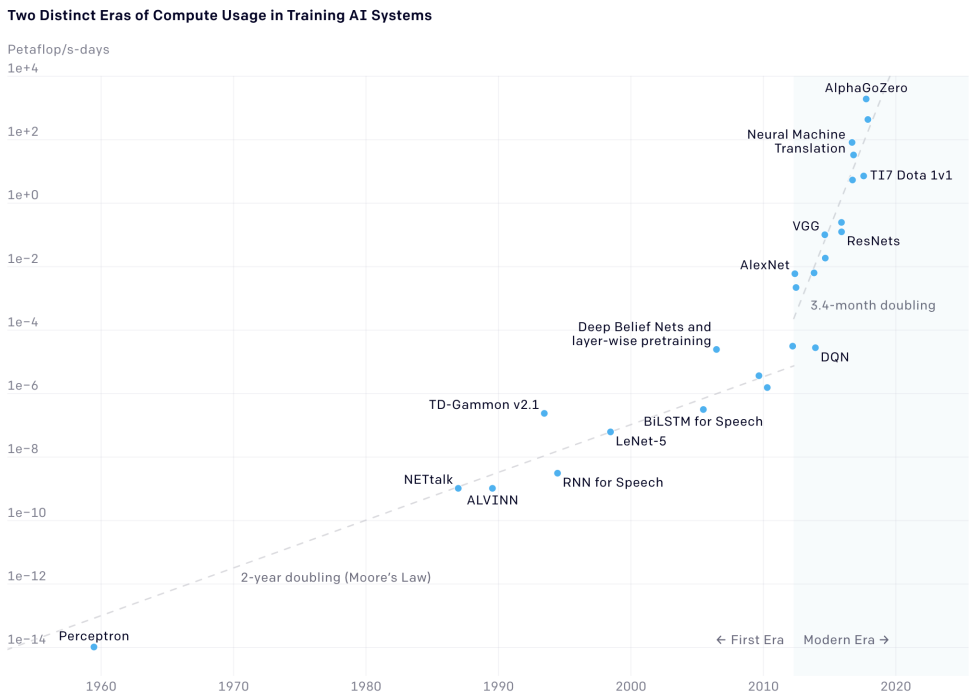


Table of Contents

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. This is powerfully illustrated in the growth of AI training sets which have been growing at a pace of 10X per year and are set to continue to grow ever further over this decade.



Source: www.openai.com/research/ai-and-compute

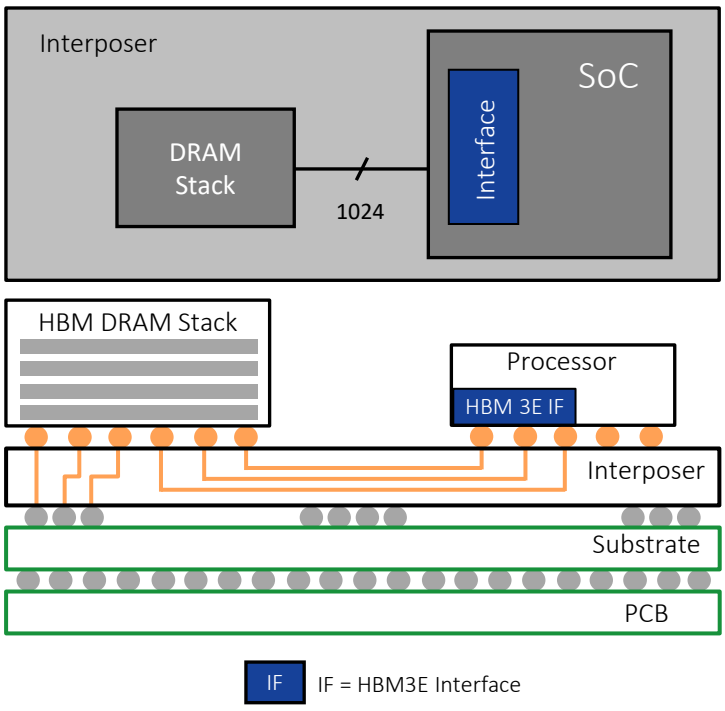
Supporting this pace requires far more than the improvements that can be realized through Moore's Law, which is slowing in any case, necessitating rapid ongoing improvements in every aspect of AI computer hardware and software.

Part 1: HBM Memory Architecture

Introduced in 2013, High Bandwidth Memory (HBM) is a high-performance 2.5D/3D memory architecture. The concept behind HBM is to use a wide data path (1024 bit) that can be run at "slow" data rates, thus delivering high bandwidth at low power. The latest generation HBM3 introduced in January 2022 takes memory performance to new heights. Given its outstanding bandwidth and compact footprint, it has become the memory solution of choice for advanced AI workloads.

The "3D" part is easy to see. HBM memory is a 3D stack of DRAM in a packaged device. The "2.5D" refers to the way the HBM memory devices connect to the processing chip, be it a GPU or AI accelerator. The data path between each HBM memory device and the processor requires 1024 "wires" or traces. With the addition of command and address, clocks, etc. the number of traces necessary grows to about 1,700.

A thousand plus traces is far more than can be supported on a standard PCB. Therefore, a silicon interposer is used as an intermediary to connect memory device(s) and processor. As with an integrated circuit, finely spaced traces can be etched in the silicon interposer allowing us to achieve the desired number of wires needed for the HBM interface. The HBM device(s) and the processor are mounted atop the interposer in what is referred to as a 2.5D architecture.



HBM uses a 2.5D/3D architecture

Part 2: How is HBM3E Different?

HBM3 is the third major generation of the HBM standard, with HBM3E offering an extended data rate and the same feature set. In each generation, we've seen an upward trend of greater bandwidth, 3D-stack height, and DRAM chip density. That translates to higher performance and greater device capacity with each upgrade of the specification.

HBM launched with a 1 Gb/s data rate, and a maximum of 8-high 3D stacks of 16 Gb devices. With HBM3, the data rate scales up to 6.4 Gb/s, and the devices can support 16-high stacks of 32 Gb capacity DRAM. The major DRAM manufacturers have introduced HBM3E devices which push data rates to 9.6 Gb/s. The table below gives a summary of each of the generations of HBM.

Generation	Data Rate (Gb/s)	Bandwidth per Device (GB/s)	Stack Height	Max. DRAM Capacity (Gb)	Max. Device Capacity (GB)
HBM	1.0	128	8	16	16
HBM2	2.0	256	8	16	16
HBM2E	3.6	461	12	24	36
HBM3	6.4	819	16	32	64
HBM3E	9.6	1229	16	32	64

Each HBM generation offers greater bandwidth and capacity

But that's not all. HBM3E/3 also introduces enhancements in power, memory access and RAS over HBM2E.

- **Power**: HBM3E/3 decrease core voltage to 1.1V from HBM2E's 1.2V. Also, HBM3E/3 reduce IO signaling to 400mV from the 1.2V used in HBM2E. Lower voltages translate to lower power. These changes help offset the higher power consumption inherent in moving to higher data rates.

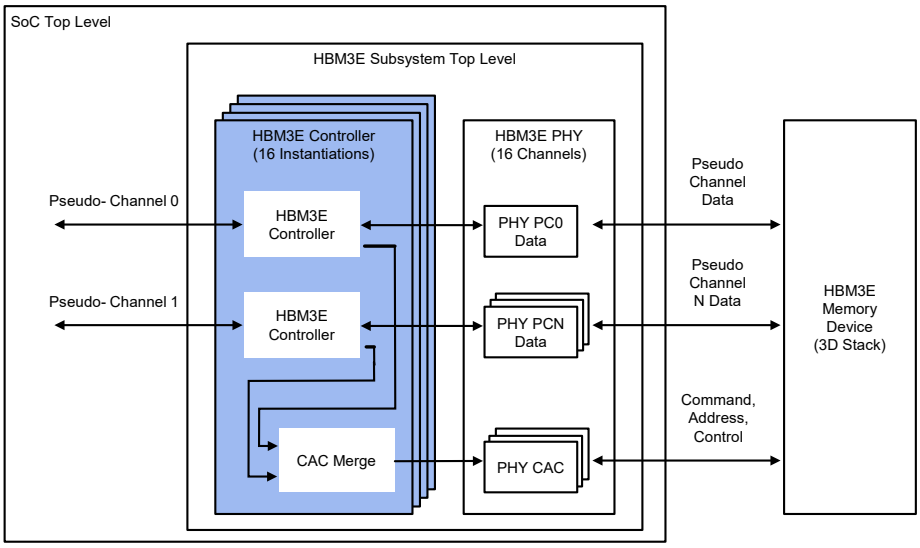
- **Channel Architecture**: HBM3E/3 divide the 1024-bit wide data channel into 16 64-bit channels or 32 32-bit pseudo-channels. This doubles the number of memory channels, and increases performance, over HBM2E's eight 128-bit channels and 16 64-bit pseudo channels.

- **Reliability, Availability, Serviceability (RAS)**: HBM3E/3 introduce additional host-side and device-side ECC as well as support for Refresh Management (RFM) and Adaptive Refresh Management (ARFM).

Part 3: The Rambus HBM3E/3 Memory Controller

Optimized for high bandwidth and low latency, the Rambus HBM3E/3 Memory Controller delivers maximum performance and flexibility for AI training in a compact form factor and power-efficient envelope.

The Rambus HBM3E/3 Memory Controller more than doubles maximum HBM2E signaling speed raising data rates to a market-leading 9.6 Gb/s per data pin (well above the standard speed of 6.4 Gb/s). The interface features 16 independent channels, each containing 64 bits for a total data width of 1024 bits. At maximum data rate, this provides a total interface bandwidth of 1228.8 GB/s or 1.23 Terabytes per second (TB/s) of throughput for every attached HBM3E/3 memory device.



The Rambus HBM3E/3 Controller offers performance up to 9.6 Gb/s

The core accepts commands using a simple local interface and translates them to the command sequences required by HBM3E/3 devices. The core also performs all initialization, refresh and power-down functions. The core queues up multiple commands in the command queue. This enables optimal bandwidth utilization for both short transfers to highly random address locations as well as longer transfers to contiguous address space. The command queue is also used to opportunistically perform look-ahead activates, precharges and auto-precharges, further improving overall throughput. The Reorder functionality is fully integrated into the controller command queue increasing throughput and minimizing gate count.

Additional key features include:

- Supports HBM3E/3 memory devices
- Supports all standard HBM3E/3 channel densities (up to 32 Gb)
- Supports up to 9.6 Gb/s/pin (HBM3E) or 8.4 Gb/s/pin (HBM3)
- Refresh Management (RFM) support
- Maximizes memory bandwidth and minimizes latency via Look-Ahead command processing
- Integrated Reorder functionality
- Achieves high clock rates with minimal routing constraints
- Self-refresh and Power-down Low Power Modes
- Support for HBM3E/3 RAS features
- Built-in hardware-level performance Activity Monitor
- DFI compatible
- End-to-end data parity
- Supports AXI or native interface to user logic
- Full set of Add-On cores available including in-line ECC core
- Delivered fully integrated and verified with target HBM3E/3 PHY

Conclusion

Delivering unrivaled memory bandwidth in a compact, high-capacity footprint, has made HBM the memory of choice for AI/ML and other high-performance computing workloads. HBM3 as the latest generation of the standard raises data rates to 6.4 Gb/s and promises to scale even higher.

The Rambus HBM3E/3 Controller provides industry-leading support of the extended roadmap for HBM3 with performance to 9.6 Gb/s. With this solution, designers can achieve as much as 1.23 TB/s of throughput for every attached HBM3E/3 memory device.

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

rev_20240213