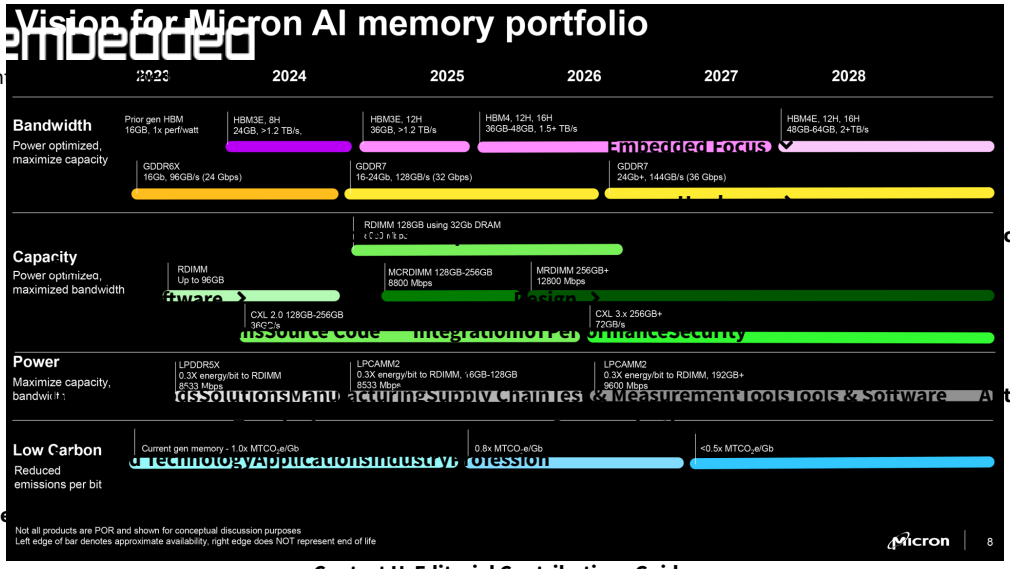
Technical Article (https://www.embedded.com/category/technical-article/)

High-bandwidth memory (HBM) options for demanding compute

January 15, 2024 Anton Shilov (https://www.embedded.com/author/nitin-dahad/)

While unbeatable in terms of performance, HBM is expensive and power hungry for many applications. We review the many memory different options for demanding compute applications.

Explosive growth of generative artificial intelligence (AI) applications in recent quarters has spurred demand for AI servers and skyrocketing demand for AI processors. Most of these processors — including compute GPUs from AMD and Nvidia, specialized processors like Intel's Gaudi or AWS's Inferentia and Trainium and FPGAs — use high-bandwidth memory (HBM) as it provides the highest memory bandwidth possible today. As a result, memory makers Micron, Samsung, and SK Hynix were set to double bit output of HBM in 2023 and increase it further in 2024, according to TrendForce (https://www.trendforce.com/presscenter/news/20230809-11785.html), a pledge that is set to become a challenge for the industry.



(https://www.embedded.com/?attachment_id=4481289)

Micron roadmap as of November 2023. (Source: Micron)

But there are a lot of AI processors, particularly those that are designed to run inference workloads, as well as HPC processors that take advantage of GDDR6/GDDR6X or even LPDDR5/LPDDR5X. Furthermore, general purpose CPUs, which can also run AI workloads (optimized for particular instructions), are poised to use commodity memory, which is why in the coming years we are going to see MCRDIMMs and MRDIMMs memory modules that will significantly increase capacity and bandwidth to new levels. But HBM is set to remain bandwidth king.

Memory Bandwidth and Capacity Comparison								
	HBM4	HBM3E	HBM3	HBM2E	GDDR7	GDDR6	LPDDR5X	DDR5
Max Capacity per Stack, Chip, or Module	36 GB - 64 GB	36 GB	24 GB	16 GB	3 GB	2 GB	16 GB	2 TB per module
Data Transfer Rate	?	9.2 GT/s	6.4 GT/s	3.6 GT/s	32 GT/s	24 GT/s	9.6 GT/s	8.4 GT/s
DRAM ICs per Stack	16			8	?	?	8	8 for 64 Gb, 16 for <64 Gb ICs
Interface Width	2048-bit	1024-bit			32-bit		32-bit, 64-bit	64-bit per module
Signaling	?	NRZ			PAM-3	NRZ	?	NRZ
Voltage	?	1.1 V		1.2 V	1.2 V	1.2 V	1.01 - 1.12V	1.1 V
Bandwidth per Stack, Chip, or Module	1.5 TB/s - 2+ TB/s	1.2 TB/s	819.2 GB/s	460.8 GB/s	128 GB/s	96 GB/s	76.8 GB/s	67.2 GB/s

All numbers are for a stack, chip, or module

(https://www.embedded.com/?attachment_id=4481290)

(Source: embedded.com)

HBM: bandwidth at all costs

Considering performance specifications and capabilities of modern types of memory, it is evident why HBM is so popular among bandwidth-hungry applications. At around 1.2 TB/s per stack, no conventional memory can beat HBM3E in terms of bandwidth. But that bandwidth comes at a cost and with some constraints, when it comes to capacity and costs.

"HBM has not just superior bandwidth, but also power consumption because the distances are small," said David Kanter, executive director of MLCommons (https://mlcommons.org/), an artificial intelligence engineering consortium that, among other things, develops industry benchmarks for AI hardware. "The central weakness is that it requires advanced packaging which is currently limiting supply and also adding cost. "But HBM will almost certainly always have a place."

SK Hynix's HBM3E memory stacks. (Source: SK Hynix)

These peculiarities of HBM are why DDR, GDDR, and LPDDR types of memory are also used for many bandwidth-hungry applications, including AI, HPC, graphics, and workstations. Development of these capacity-optimized and bandwidth-optimized types of memory is proceeding rapidly, so there is clear demand for them among developers of AI hardware, according to Micron.

“HBM is a very promising technology with a market that has a lot of potential for future growth,” said Krishna Yalamanchi, senior manager in the compute and networking business unit at Micron. “Currently, the applications are in artificial intelligence, high-performance computing, and other fields that require high bandwidth, high density, and low power consumption. The market is expected to grow rapidly as more processors and platforms adopt it.”

Meanwhile, there is a clear need for both bandwidth and capacity, according to Rambus, which develops licenses and sells to implementers memory controllers for a wide variety of applications, including processors used for AI workloads.

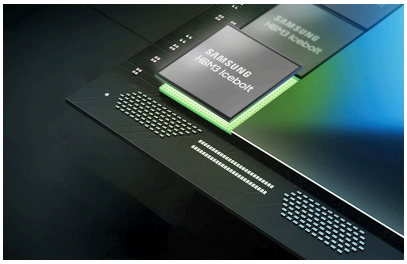
“What we are continuing to see in the AI market is that the datasets are getting larger and larger,” said Joe Salvador, vice president of product marketing, Interface IP at Rambus. “The performance needs and the memory bandwidth and memory sizes are just growing exponentially. One of the interesting things that I saw was that the training models have increased at a rate of 10x per year since 2012 and it looks like [the growth] is not slowing down.”

What is particularly interesting is that those companies who need HBM tend to adopt the latest iteration of the standard practically overnight. Nowadays nobody starts designs with HBM2E, according to Rambus.

“Today, we do not see any new HBM2 or HBM2E design starts,” Salvador said. “The market really has transitioned.” Most new chip designs use either HBM3, or the all-new HBM3E for which Rambus has a memory controller rated for an up to 9.6 GT/s data transfer rate. Integrating a 9.6 GT/s capable HBM3E memory controller should not really increase power consumption significantly, but certainly HBM3E PHY and 9.6 GT/s HBM3E stacks consume more than typical HBM3 PHY and HBM3 stacks, according to Rambus.

HBM production is hard

TrendForce is not alone in its optimistic predictions about the future of HBM memory. Gartner (https://www.gartner.com/en/documents/4641399?_its=JTdCJTIydmlkJTIyJTNBJTIyMjc2NWZkOGMtNjQ4OC00ZTEyLTk0OTItMGI0YmFlNjhjOTYxJTIyJTJDJTIyc3RhdGUlMjIlM0ElMjJybHR%2BMTY5NDc(
projects demand for high-bandwidth memory is projected to surge from 123 million GBs in 2022 to a staggering 972 million GBs by 2027, which means that HBM bit demand is expected to increase from 0.5% of overall DRAM in 2022 to 1.6% in 2027. This spike is attributed to the escalating need for HBM in standard AI and generative AI applications.



Samsung’s HBM3 stacks. (Source: Samsung)

Analysts from Gartner believe that HBM earnings will rise from $1.1 billion in 2022 to $5.2 billion by 2027, whereas HBM prices will drop by 40% relative to 2022 levels. Due to technology advancements and increasing commitment by memory makers, the density of HBM stacks is also going to increase, moving from 16 GB in 2022 to 48 GB by 2027, Gartner believes. Meanwhile, Micron seems to be more optimistic and expects 64 GB HBMNext (HBM4) stacks around 2026. HBM3 and HBM4 specifications allow to build 16-Hi stacks, so it is possible to use 16 32-Gb devices to build 64 GB HBM modules, but this will require memory makers to reduce the distance between memory ICs, which encompasses usage of new production technologies.

Given the fact that Nvidia commands the lion's share of compute GPU market, it is likely that the company is the largest consumer of HBM memory in the industry and will be for some time. The company's A30 is equipped with 24 GB of HBM2, A100 comes with 80 GB of HBM2E, H100 features 80 GB of usable HBM2E (PCIe) or HBM3 (SXM), H200 offers 141 GB of HBM3E, and GH200 is the first product to feature either 96 GB of HBM3 or 141 GB of HBM3E.

But producing HBM known good stack dies (KGSD) is fundamentally more complicated than producing traditional DRAM chips. For starters, DRAM devices used for HBM are completely different than typical DRAM ICs used for commodity memory (e.g., DDR4, DDR5). Memory producers have to make eight or 12 DRAM devices, test them, then package them together on top of a pre-tested high-speed logic layer, then test the whole package. This process is both expensive and lengthy.

"HBM stacks are based on 3D-stacked DRAM architecture that uses through-silicon vias (TSVs) to connect multiple dies vertically which is fundamentally different than commodity DRAM," Yalamanchi said. "This stacked architecture with TSVs allows for a very wide memory interface (1024 bits), a memory capacity of up to 36 GB, and enables high bandwidth operation of over 1 TB/s. The DRAM bank and data architectures are fundamentally redesigned to support such parallel wide interfaces."

But while HBM architecture is complicated, it should not be terribly costly if general production methods are well known, believes Michael Schuette, CTO at DataSecure and CTO/chief scientist at Boolean Labs, who has multiple patents in the field of memory.

"These are not a terrible cost adder, the tools and methods are established from 3D NAND, you do the through silicon vias for the connections and all it takes is porting over the existing TSV methodology [from 3D NAND]," Schuette said.

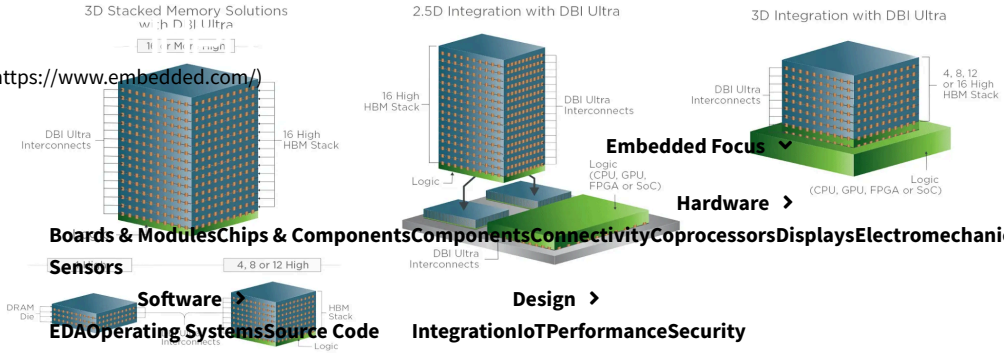
But DRAM devices used for HBM have to feature a wide interface, so they are physically bigger and therefore costlier than regular DRAM ICs. Therefore, the ramp up of HBM memory production to meet demand for AI servers will affect bit supply of all DRAM types, according to Sanjay Mehrotra, chief executive of Micron.

"High-bandwidth memory (HBM) production will be a headwind to industry bit supply growth," Mehrotra said at a conference call recently. "HBM3E die is roughly twice the size of equivalent-capacity DDR5. The HBM product includes a logic interface die and has a substantially more complex packaging stack that impacts yields. As a result, HBM3 and 3E demand will absorb an outsized portion of industry wafer supply. The ramp of HBM3 and 3E production will reduce overall DRAM bit supply growth industrywide — with a particular supply impact on non-HBM products as more capacity is diverted to addressing HBM opportunities. Micron is experiencing a similar impact of our planned HBM3E ramp on our bit supply capability."

HBM3E is essentially HBM3 with a significant speed bump, so while DRAM makers will have to ensure decent yields and then adjust their production methods to build 8-Hi 24 GB and 12-Hi 36 GB HBM3E KGSDs more efficiently, the new type of memory will not represent a significant shift in HBM production. By contrast, its successor will.

HBM4: going wider, going 3D

HBM4 is set to widen memory stack interface to 2048 bits, which will be one of the most significant changes of HBM specifications since the introduction of this memory type eight years ago. Increasing the number of I/O pins by a factor of two while keeping similar physical footprint is extremely challenging for memory makers, SoC developers, foundries, and outsourced assembly and test (OSAT) companies. Samsung indicates (https://www.anandtech.com/show/21104/samsung-announces-shinebolt-hbm3e-memory-hbm-hits-36gb-stacks-at-98-gbps) that HBM4 will require transition from micro-bump bonding used for HBM today (which is hard and expensive already) to direct copper-to-copper bonding, a state-of-the-art technology that is set to be used for integration of multi-chiplet designs in the coming years.



(https://www.embedded.com/?attachment_id=4481296)

3D stacked memory solutions featuring Xperi's DBI Ultra technology. (Source: Xperi)

"If I look at the [upcoming HBM4 specification] and the 2048-bit wide interface, that brings the pin count to 3000 pins, which is in the same league as most server CPUs out there [in terms of pin count]," Schuette said. "If you are trying to route the traces in a small footprint design, you end up with something like a 20-layer redistribution layer/interposer, if you go for a larger footprint, fewer layers, you end up exceeding the max allowed trace length."

SK Hynix even envisions that HBM4 will have to be 3D integrated on system-on-chips for maximum efficiency, but that will increase costs even further.

"In the next few years, I think we may get superior performance and efficiency through tighter integration (e.g., 3D stacking), but that is likely to be more expensive," Kanter said.

Schuette believes that it could be extremely hard to connect HBM4 stacks with a 2048-bit interface to a host processor using conventional methods featuring an interposer and a redistribution layer due to HBM4's extreme pin count.

"The tiniest warp and you get a bad connection," explained Schuette. "If it is just one ground pin, you may not notice, but if it is a signal pin you are toast.

But a 3D packaging technique will require even more sophisticated equipment, so it is highly likely that at least initially only foundries themselves will offer HBM4 integration sometime in 2025 – 2026.

To keep shrinking DRAM cell sizes and keep memory power consumption in check, Samsung reportedly intends to use FinFET transistors for HBM4. The incorporation of FinFETs is expected to optimize performance, power, and area scaling for upcoming HBM devices. However, the impact of this technology on costs is still uncertain. Additionally, the timeline for when Samsung will implement FinFETs in standard DRAM ICs is yet to be determined. For now, Samsung has only confirmed that FinFETs are coming to HBM4.

"There are still going to be cost issues and then there are going to be implementation issues on HBM4 that may give HBM3/HBM3E a longer life, particularly in places that are more cost sensitive," Salvador said.

"It is not an accurate assumption to make that people will want to adopt the fastest version of HBM, as many factors influence the choice of memory technology, such as cost, supply constraints, platform readiness, and performance requirements," Yalamanchi said.

Due to fundamentally different architecture and packaging costs, HBM will remain an expensive type of memory serving growing niche markets. This opinion is partly shared by Michael Schuette. He believes that while HBM serves its target markets quite well, it will hardly be able to address the broader market.

"HBM still appears to be a niche product and will most likely remain one," Schuette said.

Will HBM ever get cost competitive with commodity or specialty memory?

"I do not want to say never, because that is a really long time," Kanter said. "But for HBM to become cost competitive would require radically lower packaging costs and/or significant increase in cost for GDDR. Or perhaps a fundamental technology shift — for instance, if GDDR switches from high-

speed copper signaling to optical. But I am not sure if that would be GDDR at that point."

LPDDR: a lower-power option

While unbeatable in terms of performance, HBM is expensive and power hungry for many applications, so there are developers that opt to use LPDDR5X for their bandwidth-demanding applications as this type of memory offers them the right balance of price, performance, and power consumption.

For example, Apple has been using LPDDR memory in its PCs for years before it became a trend. By now, the company has polished off its LPDDR5-based memory subsystems so well that they deliver performance that is unrivalled by competing solutions. Apple's high-end desktops — Mac Studio and Mac Pro powered by the M2 Ultra SoC — now come with a whopping 800 GB/s of bandwidth using two 512-bit memory interfaces. To put this number into context: AMD's latest Ryzen Threadripper Pro with a 12-channel DDR5-4800 memory subsystem can boast with a peak bandwidth of 460.8 GB/s.

Micron's LPDDR5X memory. (Source: Micron)

Using LPDDR5 across the whole range of its devices, like Apple does, has some additional benefits, such as LPDDR5 controller IP and PHY re-use in different SoCs as well as procuring such memory in high volumes, which gives a leverage to negotiation for better pricing. Apple is certainly not the only company to use LPDDR memory for bandwidth-hungry processors. Tenstorrent uses this memory for its Grayskull AI processors.

"Today they seem to serve different niches and there are broad trends in differences," said Kanter. "HBM is more datacenter oriented, LPDDR more edge oriented. That being said, there are absolutely folks targeting similar markets using different memory types. Take inference in the datacenter – some designs use HBM, some GDDR, some regular DDR, and some LPDDR."

Among the distinct advantages of LPDDR memory chips is their relatively wide interface and rather fast operation. Typical LPDDR5 and LPDDR5X/LPDDR6T ICs feature a 32 or 64-bit interface and support data transfer rates of up to 9.6 GT/s, which is considerably wider and faster than data rates supported by mass produced DDR5 data rates (8 or 16 bits, up to 7.2 GT/s as of October, 2023). Furthermore, mobile memory naturally consumes less power than mainstream DDR memory for client PCs and servers.

For applications that are developed by Tenstorrent, memory bandwidth is crucial, but so is power consumption, which is why usage of LPDDR spans well beyond smartphones and client PCs these days.

GDDR: a balance between price and performance

Tenstorrent brings us to another type of memory, which the company is going to use for its upcoming Wormhole and Blackhole AI processors. Meanwhile, Nvidia uses GDDR6 and GDDR6X for a wide variety of GPUs used for AI inference.

"GDDR memory is used in AI and other applications and would be a good choice for AI inference applications because GDDR still offers higher bandwidth and lower latency than DDR," Yalamanchi said. "GDDR will also have a lower cost and is less complex than HBM. For example, GDDR6 can be found in Nvidia's Tesla T4 GPU which is used for AI inference as well as L40S for AI inference and graphics applications."

Samsung's GDDR6 memory. (Source: Samsung)

GDDR6 typically consumes more power than LPDDR and contemporary GDDR6/GDDR6X chips come with a 32-bit interface (i.e., narrower than some LPDDR5X), but GDDR6/GDDR6X/GDDR7 memory runs considerably faster.

In fact, GDDR7 promises to run at up to 36 GT/s and at a data transfer rate of 32 GT/s memory subsystems based on it are going to be substantially faster than those featuring LPDDR5X, especially keeping in mind that we are talking about potentially wide memory interfaces, such as 384 or 512-bit. Even at 32 GT/s data transfer rate a 384-bit LPDDR7 memory subsystem is set to offer a 1,536 TB/s peak bandwidth, which is way higher than a 512-bit LPDDR5X-9600 memory subsystem (614.4 GB/s). Yet, we can guess that an LPDDR7 memory subsystem will also be significantly more power hungry than a one using LPDDR5X, but given its performance we would consider it a fair trade-off.

MCR-DIMMs and MR-DIMMs

A story about high-performance memory solutions would not be complete without MCR-DIMMs and MR-DIMMs, two types of dual-rank DDR5-based memory modules designed primarily for servers, which are currently in development. The idea behind these technologies is to further improve efficiency of memory modules and increase their peak bandwidth beyond speeds supported by DDR5 amid consistently growing number of cores per CPU.

(https://www.embedded.com/?attachment_id=4481300)

SK Hynix's MCR-DIMM module. (Source: SK Hynix)

On a high level, a Multiplexer Combined Ranks DIMM (MCR-DIMM) is a dual-rank buffered memory module equipped with a multiplexer buffer. This buffer can retrieve 128 bytes of data from both ranks at the same time, and it is designed to work with a memory controller at high speeds of around 8800 MT/s (based on a recently published Micron roadmap), which is 400 MT/s higher than the maximum data rate specified by the original DDR5 specification. These modules are aimed at enhancing performance while also simplifying construction of high-capacity dual-rank modules. MCR-DIMM is backed by Intel and SK Hynix and is set to be supported by Intel's 6th Generation Xeon Scalable 'Granite Rapids' platforms, whereas Micron plans to ship MCR-DIMMs in early 2025.

Multi-Ranked Buffered DIMM (MR DIMM) are very similar conceptually: they are dual-rank modules with a multiplexer buffer that interacts with both ranks simultaneously and operate with a memory controller at speeds beyond those specified for DDR5. This standard is set to begin with a speed of 8,800 MT/s for its first generation, advance to 12,800 MT/s in the second generation, and ultimately surge to 17,600 MT/s by its third generation. This technology is backed by JEDEC, AMD, Google, and Microsoft. Micron intends to start shipments of MR-DIMMs with a 12,800 MT/s speed in 2026. Such modules will offer massive bandwidth and capacity, something that is needed due to rising number of cores inside datacenter CPUs as well as their demand for bandwidth.

"It would be foolish not to embrace new form factors for disaggregated memory," said Schuette. "The server requirements are different from the client's and you will always need ECC on server, while you do not need it in client PCs."

Exotic and hybrid memory subsystems

While using a specific type of memory is perhaps the most obvious course of action for chip and system developers, there are also those who opt for hybrid memory subsystems that use different types of memory.

For example, Intel has Xeon Max CPUs that carry 64 GB of on-package HBM2e and support up to 6 TB of six-channel DDR5 memory using up to 16 DIMMS per socket. These CPUs are aimed primarily at high-performance computing (HPC) environments and can work in HBM Only mode, HBM Flat mode (providing fast and slow memory tiers), and HBM Caching mode.

(https://www.embedded.com/?attachment_id=4481301)

Different modes of Sapphire Rapids HBM's operation. (Source: Lenovo)

Another example is D-Matrix's AI processors that come with 256 MB of SRAM (at 150 TB/s) inside and support up to 32 GB of LPDDR5 memory with a fairly limited bandwidth. These chips are aimed primarily at inference and their architecture is tailored for such workloads.

"Generally, caching or on-die SRAM can reduce some external bandwidth needs," Kanter said. "So, for inference, if we could live with say a neural network that is <100MB, [caching would help]. Similarly, we could integrate memory even closer to reduce off-package bandwidth. But a lot of cutting-edge work for really big training systems, like training the next generation of LLMs, will always need more bandwidth."

While historically hybrid and exotic memory sub-systems comprising of different types of memory have been used by a wide variety of applications, such as ATI's Xenos GPU with eDRAM-based 'daughter die' for the Xbox 360 game console or Intel's Xeon Phi 7200-series co-processors that used both MCDRAM and DDR4 memory, Schuette believes that such memory subsystems are not exactly efficient.

"My opinion is that you get the worst of both worlds," he said. "A huge overhead in design, lots of complexity and I do not even want to get into troubleshooting.

On the other hand, all systems with CPUs and accelerators by definition use hybrid memory subsystems and they have proven to be very efficient..

"Many AI systems are hybrids today," Kanter said. "For example, many training systems favor HBM for the accelerator, but use DDR for the host processor — and the host processor actually does real work here. Similar for data center inference systems."

Anton Shilov *is a veteran technology writer who has covered many aspects of the electronics and embedded systems industry, including semiconductors, computing, displays, and consumer electronics.*

Related Contents:

- Renesas' Balaji Kanigicherla: industry needs to bridge compute & memory (https://www.embedded.com/renesas-balaji-kanigicherla-industry-needs-to-bridge-compute-memory/)
- Flash Memory Summit: exploring memory innovation for the AI era (https://www.embedded.com/flash-memory-summit-exploring-memory-innovation-for-the-ai-era/)
- AI acceleration will need HBM3 to overcome memory bottlenecks (https://www.embedded.com/ai-acceleration-will-need-hbm3-to-overcome-memory-bottlenecks/)
- Interview: multi-die addresses chip complexity but power is a challenge (https://www.embedded.com/interview-multi-die-addresses-chip-complexity-but-power-is-a-challenge/)
- Eliyan eliminates silicon interposer to advance D2D chiplet connect for HPC (https://www.embedded.com/eliyan-eliminates-silicon-interposer-to-advance-d2d-chiplet-connect-for-hpc/)

Tags: Memory/Storage (https://www.embedded.com/tag/memory-storage/)

Previous
Mastering motor control: motor control 101 (https://www.embedded.com/mastering-motor-control-motor-control-101/)

Next
Mastering motor control: measurements sensors and parameters (https://www.embedded.com/mastering-motor-control-measurements-sensors-and-parameters/)

(https://www.embedded.com/)

Leave a Reply

You must **Sign in (https://login.aspencore.com/d045a02d-767d-4f3f-80d7-695dce25d142/login/authorize?response_type=code&scope=profile%20email%20openid&client_id=216002e0-47e9-4d0c-8915-4dfcbabd2e56&state=106f7e21459bcdb0419a48fd1e7712f0&redirect_uri=https%3A%2F%2Fwww.embedded.com%2Fopenid-connect-authorize)** or **Register (https://login.aspencore.com/d045a02d-767d-4f3f-80d7-695dce25d142/login/authorize?response_type=code&scope=profile%20email%20openid&client_id=216002e0-47e9-4d0c-8915-4dfcbabd2e56&state=f6dfeb7838205cce81c656c080dd5ec2&redirect_uri=https%3A%2F%2Fwww.embedded.com%2Fopenid-connect-authorize)** to post a comment.

You may have missed

Products (https://www.embedded.com/category/products/)

Cadence adds radar accelerator and automotive-optimized DSPs (https://www.embedded.com/cadence-adds-radar-accelerator-and-automotive-optimized-dsps/)

March 4, 2024 **Nitin Dahad (https://www.embedded.com/author/nitin-dahad/)**

(https://www.embedded.com/cadence-adds-radar-accelerator-and-automotive-optimized-dsps/)

News (https://www.embedded.com/category/news/)

Infineon integrates Qt GUI into its MCUs (https://www.embedded.com/infineon-integrates-qt-gui-into-its-mcus/)

March 4, 2024 **Nitin Dahad (https://www.embedded.com/author/nitin-dahad/)**

(https://www.embedded.com/infineon-integrates-qt-gui-into-its-mcus/)

Products (https://www.embedded.com/category/products/)

Renesas RZ MPU targets robotics with vision AI and real-time control (https://www.embedded.com/renesas-rz-mpu-targets-robotics-with-vision-ai-and-real-time-control/)

March 4, 2024 **Nitin Dahad (https://www.embedded.com/author/nitin-dahad/)**

(https://www.embedded.com/renesas-rz-mpu-targets-robotics-with-vision-ai-and-real-time-control/)

Products (https://www.embedded.com/category/products/)

Synopsys 1.6T Ethernet IP reduces interconnect power use by 50% (https://www.embedded.com/synopsys-1-6t-ethernet-ip-reduces-interconnect-power-use-by-50/)

March 4, 2024 **Nitin Dahad (https://www.embedded.com/author/nitin-dahad/)**

(https://www.embedded.com/synopsys-1-6t-ethernet-ip-reduces-interconnect-power-use-by-50/)

Products (https://www.embedded.com/category/products/)

Micron's new UFS 4.0 modules shrink dimensions and add functionality (https://www.embedded.com/microns-new-ufs-4-0-modules-shrink-dimensions-and-add-functionality/)

March 1, 2024 **Anton Shilov (https://www.embedded.com/author/nitin-dahad/)**

(https://www.embedded.com/microns-new-ufs-4-0-modules-shrink-dimensions-and-add-functionality/)