OBJECTIVE ANALYSIS
Semiconductor Market Research
Coughlin Associates
Data Storage Consulting

JOINTLY-AUTHORED WHITE PAPER

THE FUTURE OF LOW-LATENCY MEMORY

Why Near Memory Requires a New Interface

The growth of data processed and stored, and the new ways that computers are being used, are causing memory capacity and performance requirements to balloon. Near Memory, the memory connected directly to a processor's pins, must grow larger and faster to keep pace.

But DDR bus speeds decrease as memory is added to a processor's pins. This means that a DDR-based memory system must trade off between speed and size, or that an increasing number of DDR channels must be added to a processor to achieve both ends.

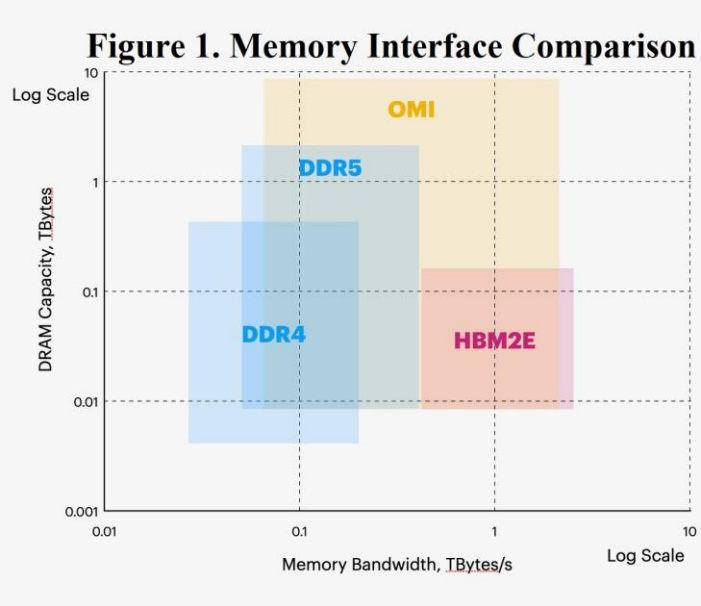


Figure 1. Memory Interface Comparison

This state of affairs is driving changes in the processor-memory interface. These changes will be discussed and evaluated in this white paper. The three established DRAM interfaces seen in Figure 1. will be reviewed, resulting in recommendations and an outlook for the future path of these interfaces.

The DDR Conundrum

For Near Memory, the memory that attaches directly to a processor's pins, the current DDR parallel memory bus has been tweaked and adjusted over the years.

Although its performance has improved impressively for more than two decades, DDR is failing to keep pace with the increasing bandwidth requirements of processor chips. Processor core counts are rising quickly, and clock speeds continue to creep higher, driving a thirst for bandwidth and capacity that runs in direct opposition to the way the DDR bus operates.

To achieve the highest DDR speeds, the bus's capacitive loading must decrease as the bus speed increases. Because of this, the memory channels that previously managed four DIMM slots have shrunk to three, then two, and now the highest-speed channels can only support a single slot.

As a result, the amount of memory per channel is declining.

One response is to add more memory channels to the processor, but this consumes more real estate on the most expensive chip in a computer system,

while also adding to the chip's power budget.

Some processors, notably GPUs, use HBM (High Bandwidth Memory) to get past this issue. HBMs are stacks of DRAM that present 1,000-2,000 parallel signal paths to the processor. This can improve performance but the processor and the HBM must be intimately connected.

Although HBM is a help, it's considerably more expensive than standard DRAM and is limited to stacks of no more than twelve chips, limiting its use to lower-capacity memory arrays. HBM is also complex and inflexible. There's no way to upgrade an HBM-based memory in the field. As a consequence, HBM memory is only adopted where no other solution will work

Todays and tomorrow's computing systems need a growing amount of both Near and Far Memory, that provide as much bandwidth as possible and the processor needs to use the smallest possible die area to communicate with these memories.

The Industry's Response

There have been numerous approaches to manage the different requirements of low latency Near Memory. As mentioned above, Near Memory bandwidth needs have been addressed by adding memory channels to the processor or by harnessing HBMs, but these changes do nothing to help increase the size of the memory, and can even work against it.

Intel introduced very high density DDR modules based on its 3D XPoint memory technology, but this technology is slow, and this requires the processor to always need some DRAM to work around its speed issue. That's difficult to do when each memory channel is restricted to a single slot!

Another approach is to add memory on the other side of a buffer in a Far Memory space. This adds considerable latency, and the DDR bus, which is the only way that today's mainstream processors communicate with Near Memory, is unable to communicate with two different memory speeds.

Since Far Memory has higher latency than the Near Memory on the DDR bus, and since the DDR bus can only communicate with a single memory speed, buffered memory pools were originally connected to the processor as I/O devices. This gave the memory pool the unfortunate distinction of being a memory that was a couple of orders of magnitude higher latency than DRAM at a cost that was slightly higher than DRAM's, thanks to the required support circuitry.

The "orders of magnitude" latency difference came from the way the data was managed, through a hardware and software I/O protocol.

This conundrum drew the attention of various teams of system architects, who produced a number of similar solutions: CAPI, in 2014, OpenCAPI, CCIX, and Gen-Z in 2016, and CXL in 2019. These new protocols allowed the processor to access the slower Far Memory as memory, rather than through an I/O channel, and even to manage coherency between the various caches and memories sprinkled throughout the system.

This solved the problem of increasing overall memory size, but it did not satisfy the need for larger Near Memories with high bandwidths and low latencies.

One of these protocols, CAPI, which originally was layered on top of the PCIe protocol, later developed a new underlying protocol to improve performance. This new protocol became the OpenCAPI standard, and subsequently was developed into a new way to communicate with memory that brought bandwidth close to that of HBM without in-

curring HBM's high cost and capacity restrictions.

This approach, dubbed the "Open Memory Interface" or OMI, uses existing high-speed serial signaling PHYs, with a custom protocol, to connect standard low-cost DDR DRAMs to the processor.
OMI is a latency-optimized subset of OpenCAPI.
This approach allows large arrays of inexpensive DRAM to be connected at high speeds to a processor without burdening the processor with a lot of additional I/O pins.

The relationship between DDR, HBM, and OMI is illustrated in Figure 1. DDR can support larger memory capacities than HBM, and HBM supports higher bandwidth than DDR. OMI provides near-HBM bandwidth at larger capacities than are supported by DDR.

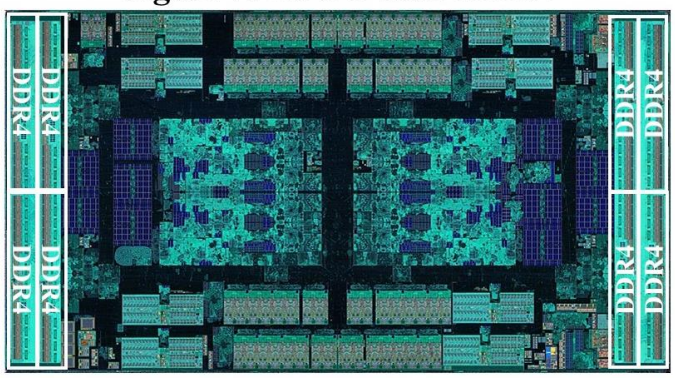
Why I/O Pins are Troublesome

Figure 2 shows a die photo of the I/O Die (IOD) for AMD's EPYC Rome processor, which packages four or eight CCDs (Core Complex Dice) with a single IOD.

The I/O Die supports eight DDR4 3200 channels for an aggregate bandwidth of 204 gigabytes per second (GB/s), or 25.6GB/s per channel.

DDR is the most mature of the technologies listed in this document, and has been the basis for all DRAM for two decades.

Figure 2. EPYC with DDR4



The annotation on the photo shows how much die area is consumed by the DDR I/O on the left and right sides of the chip. Each of the eight channels uses about 7.8mm² out of the chip's total area of 416mm².

One way to estimate the relative economy of the I/O's die area, that is, how well it's being used, is to calculate, on average, how much memory bandwidth is supported by one square millimeter of silicon. In the case of this chip it's (25.6GB/s)/7.8mm², or 3.3GB/s/mm². Shortly we will compare that against similar figures for other memory interfaces.

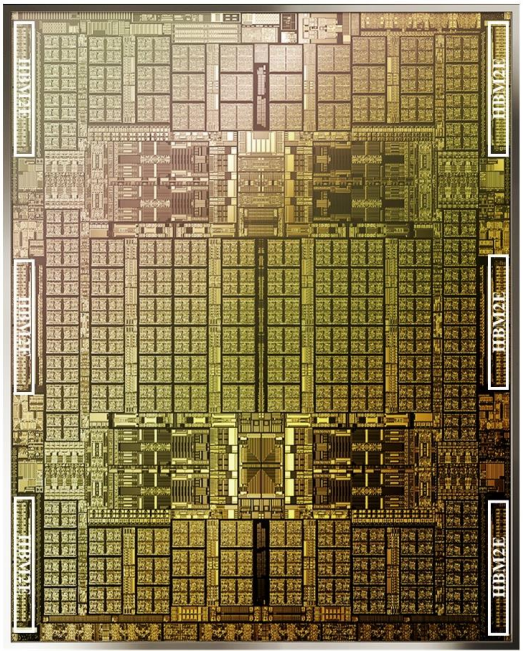
That takes care of DDR. What about chips that communicate with their DRAM via the very wide HBM channels? As an example of that we present the NVIDIA Ampere processor, which sports five HBM channels. It's shown in Figure 3.

This chip's five active HBM2E buses each achieve a read bandwidth of 200GB/s, and a similar write bandwidth, since HBM has separate read and write buses.

As the annotations on the figure illustrate, 11.4mm² of the die is consumed by each HBM2E interface, giving a read bandwidth of 17.5GB/s/mm² and a write bandwidth of the same magnitude.

Note that there are actually six HBM2E interfaces on this chip, but NVIDIA only supports five of them, so we determined that it was fair to disregard the area wasted on the unused port.

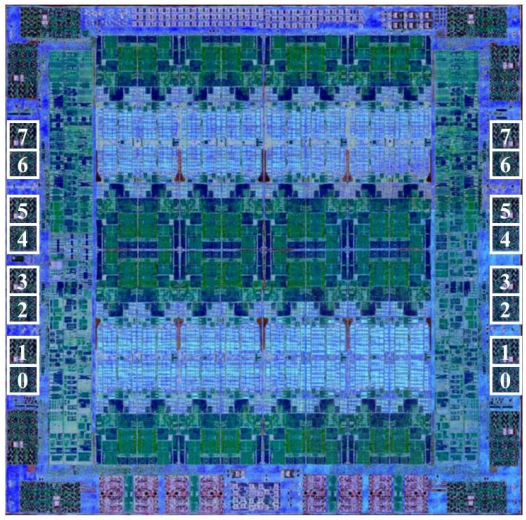
Figure 3. Ampere with HBM2E



HBM is a newer technology than DDR, and has been shipping in volume only since 2015 from SK hynix and since 2016 by Samsung. Micron Technology introduced its first HBM in 2021, after having shipped a competing technology, the Hybrid Memory Cube (HMC) that is based on the same manufacturing process.

There's another memory interface option in current production that is neither DDR nor HBM. IBM took a very different approach with its POWER10 processors by having neither a DDR nor an HBM interface on the chip, choosing instead to use OMI to reduce the area on the processor consumed by I/Os while still achieving a high data bandwidth. POWER10's OMI memory channel leverages industry standard 32Gbps PHY signaling, but, unlike PCIe, the protocol has been optimized for ultra-low latency memory transactions.

Figure 4. POWER10 with OMI



With OMI, the processor communicates with the memory through a separate transceiver chip that sits on the DIMMs. Microchip, who manufactures the transceiver chip, calls it a Smart Memory Controller. The processor communicates with this chip through a differential serial channel that moves data at a very high bandwidth. OMI DDIMMs (Differential DIMMs) are a mature technology produced by Samsung, SMART, and Micron.

The Smart Memory Controller is said to add only 4ns of latency to the DRAM's access time over a standard registered DIMM. IBM considers this latency increase to be a viable trade-off when compared to OMI's much higher bandwidth than that of DDR4, and the significantly larger DRAM capacities that this approach supports. The DRAM pin capacitance no longer loads down the I/O pins on the processor, but is broken up among the OMI transceivers on the DIMMs.

IBM's POWER10 processor chip, shown in Figure 4, uses sixteen OMI channels, eight on either side of the die, to get a total bandwidth of 1TB/s, broken down to 500GB/s of read bandwidth and 500GB/s of write bandwidth.

Table 1. Comparing the Options for Near Memory

Specification	DDR4 LRDIMM	HBM2E (8-High)	OMI
Protocol	Parallel	Parallel	Serial
Signaling	Single-Ended	Single-Ended	Differential
I/O Type	Duplex	Simplex	Simplex
Paths/Channel (Read/Write)	64	512R/512W	8R/8W
Data Transfer Rate	3,200MT/s	3,200MT/s	32,000MT/s
Channel Bandwidth (R+W)	25.6GB/s	400GB/s	64GB/s
Latency	41.5ns	60.4ns	45.5ns
Driver Area/Channel	7.8mm²	11.4mm²	2.2mm²
Bandwidth/mm²	3.3GB/s/mm²	35GB/s/mm²	29.6GB/s/mm²
Max Capacity/Channel	64GB	16GB	256GB
Connection	Multi-Drop	Point-to-Point	Point-to-Point
Data Resilience	Parity	Parity	CRC

The area consumed by these sixteen OMI channels is close to that required by three of the HBM2E channels on the Ampere chip or four of the DDR4 channels on the EPYC Rome I/O chip at 34.6mm², giving the POWER10 chip an efficiency of 29.6GB/s/mm². This is almost nine times the bandwidth per mm² of the DDR4 interface on the AMD EPYC Rome I/O chip of Figure 2 and close to that of the HBM2E channels on the NVIDIA Ampere chip.

The numbers behind the above calculations, along with some other information about the buses, is compiled in Table 1.

Capacity Limitations

HBM2E has other limitations that do not apply to DDR and OMI. Each HBM2E channel can support only a single stack of up to twelve chips, or 24GB of capacity using today's highest-density DRAMs, while each DDR4 bus can go up to 64GB and an OMI channel can support 256GB, over 10 times as much as HBM. Furthermore, both DDR and OMI, being modular, allow for flexible configurations, field upgrades, and repair, options that are impossible in an HBM-based system.

Additionally, the size of the HBM stack and its required proximity to the processor chip limits the number of HBMs that can be attached to the processor. Figure 5, dramatizes that point. The photograph shows an Ampere chip surrounded by its supporting HBM stacks on a GPU card. Each of the silver rectangles above and below the processor is an HBM stack. The silver color is the back of the stack's top DRAM die, illustrating that it would be impossible to use a smaller package. Although it may be possible to add a few more HBMs on the left and right sides of the Ampere die, this would reduce the amount of space available for the chip's other I/O.

As opposed to HBM, the OMI and DDR approaches use standard commodity DDR DRAM chips. These chips sell at a much lower price than HBM, first, because they aren't burdened with the extra processing that HBM stacks require with their through-silicon vias (TSVs), which add significantly to the manufacturing cost; and second, because the market for DDR DRAMs, being significantly larger

than the HBM market, doesn't come with the high price premiums that HBM chips can command.

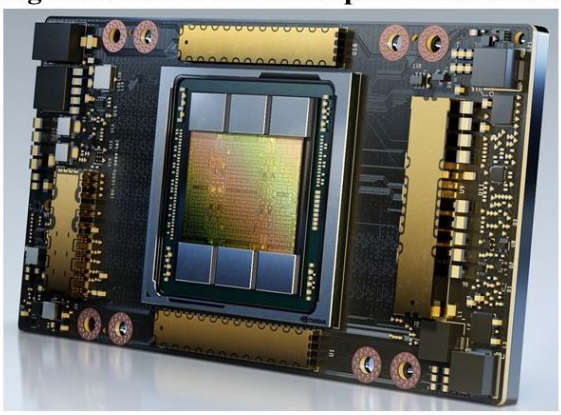
DDR supports upgrades and multiple system configurations, but it also has limitations, since the JEDEC DDR DIMM format was designed to fit within a 1U chassis. JEDEC's OMI DDIMM (Differential DIMM) mechanical specification defines not only a height to fit into a 1U chassis (Figure 6.) but it also defines larger card areas with taller form factors to be used in 2U and 4U chassis. The DDR DIMM's 1U format limits its capacity by the number of DRAM chips that can be fit onto the DIMM. Since larger DDIMM formats can hold larger numbers of chips, they can support higher-density modules than can be built within the DDR DIMM standard.

Also, the number of DIMMs that can be attached to a processor chip is limited by the number of DDR channels that the processor chip supports, and, as explained earlier, these ports consume valuable space on the processor die.

These two issues set a lower maximum capacity per DIMM and a maximum number of DIMMs that can be attached to the processor chip. Certain applications require more memory than can be supported in a DDR-based system.

OMI's ability to connect a very broad range of memory capacities via multiple DDIMM formats, coupled with the capability of fitting a very large number of OMI ports to a processor chip at a small expense in processor die area, provides a very simple memory upgrade path and supports a broader range of system configurations than does the standard DDR channel.

Figure 5. The NVIDIA Ampere with HBMs



Although the DDIMM does require a transceiver that is not found on a DDR DIMM, Microchip's Smart Memory Controller chip is relatively small, at 30mm², so it shouldn't add significantly to the overall cost of the DDIMM.

Technology Outlook

How is computer architecture likely to develop, then, with all of these options?

HBM prices are dropping, and we see that as a positive for the technology's more widespread adoption, but it will remain significantly more costly than standard DDR, since the addition of TSVs adds nearly 50% to the cost of processing a DRAM wafer. This cost is unlikely to be reduced significantly over time, even with higher production volume.

HBM also suffers from constraints of the maximum amount of memory that can be attached to a channel, and the number of channels that can be attached to a processor, since HBMs must be located extremely close to the processor chip. As a consequence, HBM will be relegated to applications similar to those where it is presently used – smaller systems that require very high bandwidth but only limited memory. Today nearly all HBM is used in high-end GPUs and specific classes of supercomputer processors. Over time it should migrate to high-end servers and midrange GPUs.

OMI stands a good chance of finding adoption in those systems that require high bandwidth, low latencies and large capacity memories. Today the technolo-

gy is used in high-end, midrange, and entry-level servers based upon POWER processors, and is supported within FPGAs, but it would be a good candidate for use in other servers, as well as broader adoption within FPGA accelerators, especially with the Open Source host and device RTL available today. The DDR interface has already been abandoned by IBM's POWER processors. OMI also supports the use of varying memory speeds, so that persistent memory, like Intel's 3D XPoint, which has significantly different read and write latencies, can be connected to the processor chip as Near Memory through the same sort of transceiver chip as is used to manage the DDR DRAM chips on a DDIMM. This removes the need for a specialized bus, like Intel's DDR-T, to accommodate 3D XPoint's different read and write speeds. DDR-T adds proprietary signals to a standard DDR4 memory channel, and is necessary to enable Intel's Optane Persistent Memory Modules to communicate with the processor.

Figure 6. An OMI DDIMM

DDR should remain around for a long time as it continues to dominate systems with modest bandwidth and memory capacity needs that are in the cost-driven lower-end of computing. This would include all PCs and many low-end servers. Smart phones also fit into this category, but they will continue to use variants of DDR like LPDDR for their special power-saving features.

Conclusions

Memory configurations for computers are undergoing changes to meet the demands of future Big Data workloads, and these changes are going to accelerate. Memory sizes must increase, while the bandwidth between memory and the CPU must grow.

This will be the driving force for an increasing number of systems to attach Near Memory to the processor either through HBM or OMI. Cost considerations make OMI an interesting argument, at least for systems with large memory capacity and bandwidth requirements, but DDR should remain in good favor for widespread use at computing's low end where only one or two DDR memory channels are required. HBM will see growing adoption, but will continue to remain an expensive niche technology for high-end computing thanks to its additional production cost and severe capacity limitations.

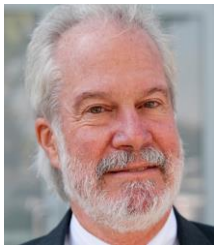
Jim Handy, Tom Coughlin, April 2021

Produced under the sponsorship of the OpenCAPI Consortium

All figures used with permission.
Figure 1. Nallasway, Inc.
Figure 2, Fritz Fritzchens
Figures 3 & 5, NVIDIA Corp.
Figure 4. IBM Corp.
Figure 6. SMART Modular Technologies, Inc.

Author Biographies

Jim Handy



Jim Handy, a widely recognized semiconductor analyst, comes to Objective Analysis with over 30 years in the electronics industry including over 20 years as an industry analyst for Gartner Dataquest, Semico Research, and Objective Analysis. His background includes marketing and design positions at market-leading suppliers including Intel, National Semiconductor, and Infineon.

Mr. Handy is a member of the Storage Networking Industry Association (SNIA), a Leader in the GLG Councils of Advisors, and serves on the Advisory Board of the Flash Memory Summit. He is the author of three blogs covering memory chips (www.TheMemoryGuy.com), SSDs (www.TheSSDguy.com), and semiconductors for the investor (www.Smartkarma.com). He contributes to a number of other blogs.

A frequent presenter at trade shows, Mr. Handy is known for his widespread industry presence and volume of publication. He has written hundreds of articles for trade journals, Dataquest, Semico, and others, and is frequently interviewed and quoted in the electronics trade press and other media.

Mr. Handy has a strong technical leaning, with a Bachelor's degree in Electrical Engineering from Georgia Tech, and is a patent holder in the field of cache memory design. He is the author of ***The Cache Memory Book*** (Harcourt Brace, 1993), the leading reference in the field. Handy also holds an MBA degree from the University of Phoenix. He has performed rigorous technical analysis on the economics of memory manufacturing and sales, discrediting some widely held theories while unveiling other true motivators of market behavior.

Tom Coughlin

Tom Coughlin has worked for over 40 years in the data storage industry and is President of Coughlin Associates, Inc.. He has over 500 publications and six patents and is a frequent public speaker. Tom is active with the IEEE, SMPTE, SNIA, and other professional organizations. Dr. Coughlin is an IEEE Fellow and HKN member. He is co-chair of the iNEMI Mass Storage Technical Working Group, Education Chair for SNIA CSMI, Past-President of IEEE-USA and a board member of the IEEE Consultants Network of Silicon Valley. His publications include the ***Digital Storage Technology Newsletter***, ***Media and Entertainment Storage Report, and The Emerging Memory Storage Report***. Tom is the author of ***Digital Storage in Consumer Electronics: The Essential Guide***, now in its second edition with Springer. He has a regular Forbes.com blog called ***Storage Bytes*** and does a regular digital storage column for the ***IEEE Consumer Electronics Magazine***.

He was the founder and organizer of the ***Storage Visions Conferences*** as well as the ***Creative Storage Conferences***. He was general Chairman of the annual Flash Memory Summit for 10 years. Coughlin Associates provides market and technology analysis as well as data storage technical and market consulting. For more information go to www.TomCoughlin.com