The Demand for High-Performance Memory

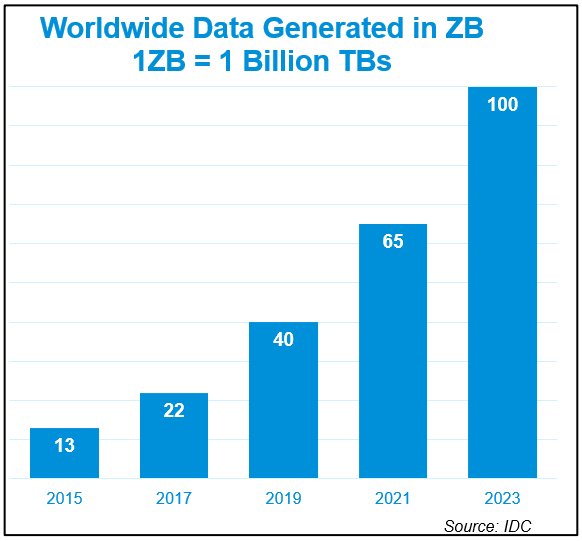
Micron believes memory is the heart of technology, the driving force that pumps life into every product, application and innovation it touches. Memory enables researchers to explore the microscopic detail of the human body and the infinite grandness of the universe. Memory helps make precision agriculture more productive, vehicles more autonomous, and smart homes more convenient. It is the heart of consumer experiences, making E-sports more active and virtual reality more interactive. And it is the heart of artificial intelligence (AI), enabling it to do everything from make us coffee to make us laugh.

In this data economy, vast amounts of information are created, stored and processed daily. Deep insights unlocked from that data can generate tremendous value and drive greater efficiencies. Innovation in memory technology is what makes these insights possible. This white paper highlights the need for and importance of high-performance memory in the market today and in the future. Read on for details on what challenges exist in the market and why high-performance memory is critical in a variety of data-intensive and bandwidth-intensive applications.

The High-Performance Memory Market

An Insatiable Demand for More Bandwidth

Artificial intelligence, machine learning (ML), deep learning (DL), autonomous driving, high-performance computing (HPC), virtual reality (VR), augmented reality (AR) and next-generation gaming are more than buzz words. Use of these applications is advancing significantly, and each requires an immense amount of data — data that is not just in high capacities but that also must be analyzed rapidly and repeatedly, a process that requires a tremendous amount of system bandwidth. Micron believes 2020 will be a defining year for next-generation developments in high-performance memory.



Building High-Performance Memory Systems

The evolution of GDDR (synchronous dynamic graphics double data rate) memory, specialized for fast rendering on graphics cards, began many years ago, but in this white paper we start our history with GDDR5 in 2008. The marketplace required higher memory data rates, in affordable packages, using known design methods and materials. Starting at 512Mb and reaching 8Gb in density, GDDR5 maxed out at 8Gb/s per pin in data rate performance. For a system bandwidth calculation, GDDR5 on a typical graphic card configuration (32-bit interface with 8 components) running at 8Gb/s per pin results in an 8GB frame buffer and a system bandwidth of 256GB/s. These results were acceptable for a time, but the market soon demanded more bandwidth.

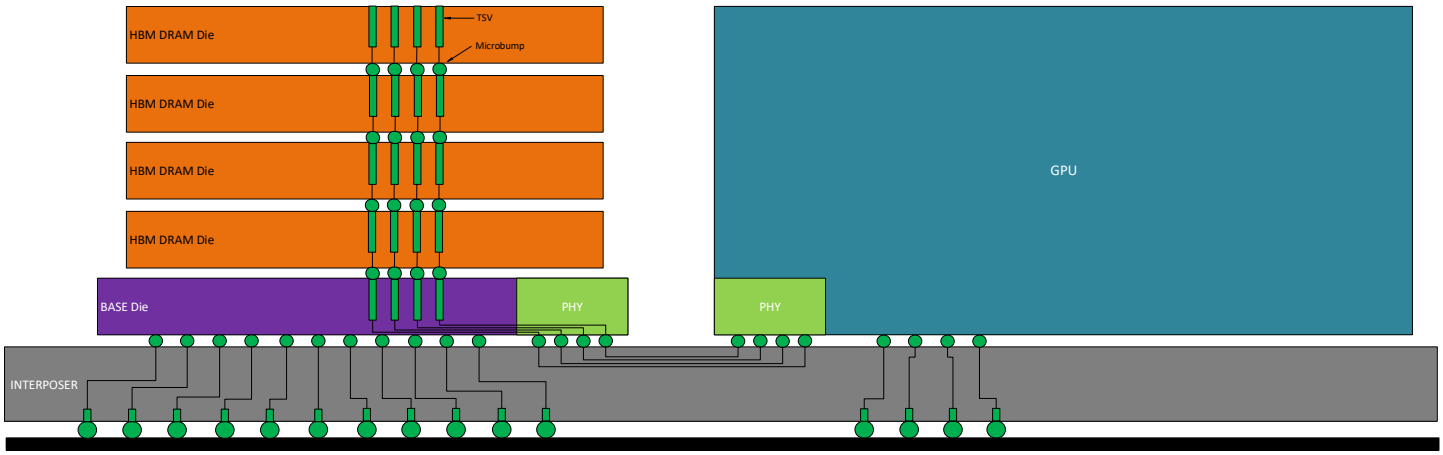
In 2015, in collaboration with NVIDIA®, Micron introduced a JEDEC innovation in GDDR5X, allowing users to reach a data rate of up to 12Gb/s per pin. After that, GDDR5X dominated the high-end graphics market for two years. For example, the NVIDIA Titan X® (32-bit interface, 12 components and 11.4Gb/s per pin data rate) reached a system bandwidth of 547GB/s.

Perhaps the most important result of GDDR5X was that it provided the framework for GDDR6. GDDR6 was introduced in the fall of 2018 and immediately became the market leader in performance. Micron was the launch partner of NVIDIA in 2018 and of AMD in 2019 with 8Gb GDDR6, delivering the high performance the market demanded. GDDR6 is still relatively early in its projected life span, but its current maximum data rate per pin is 16Gb/s. The maximum projected system bandwidth using GDDR6 is 768GB/s (32-bit interface, 12 components and 16Gb/s per pin data rate). GDDR6 is not only a high-performance solution but also a cost-optimized one, which can be used in a variety of applications.

High-Performance Memory Comparison	GDDR5 Graphics Card	GDDR5X Graphics Card	GDDR6 Graphics Card AI Inference Accelerator	HBM2 AI Training Accelerator	HBM2E AI Training Accelerator
Application Type (Example)	GTX 1070 RX 580	TitanX	Titan RTX RX5700 XT	Tesla V100 Radeon Instinct MI50	Expected
# of Placements	8	12	12	4	4
Gb/s per Pin	8	11.4	14-16	1.75-2	3.2
GB/s per Placement	32	45	56-64	225-256	410
Configuration (Example)	256 I/O per sec (8pcs x 32 I/O package)	384 I/O per sec (12pcs x 32 I/O package)	384 I/O per sec (12pcs x 32 I/O package)	4096 I/O per sec (4pcs x 1024 I/O cube)	4096 I/O per sec (4pcs x 1024 I/O cube)
Frame Buffer of Typical System	8GB	12GB	12GB	16-32GB	16-32GB
AVG Device Power (pJp)	9.00	8.0	7.5	7.0	6.0
Typical I/O Channel	PCB (P2P SM)	PCB (P2P SM)	PCB (P2P SM)	Si Interposer (2.5D integration)	Si Interposer (2.5D integration)

Source: Micron & analyst research

A discussion about high-performance memory would not be complete without mentioning HBM (high-bandwidth memory). HBM fills the gap for a memory solution by tightly integrating with compute and delivering lower power and higher bandwidth. Leveraging stacked memory components provides the density, and an extreme I/O count at lower clock rates delivers high bandwidth, all at a lower power profile. HBM is a powerful version of high-performance memory, but it is a relatively high-cost solution due to the complex nature of the product. HBM is targeted toward very high-bandwidth applications that are less cost-sensitive.



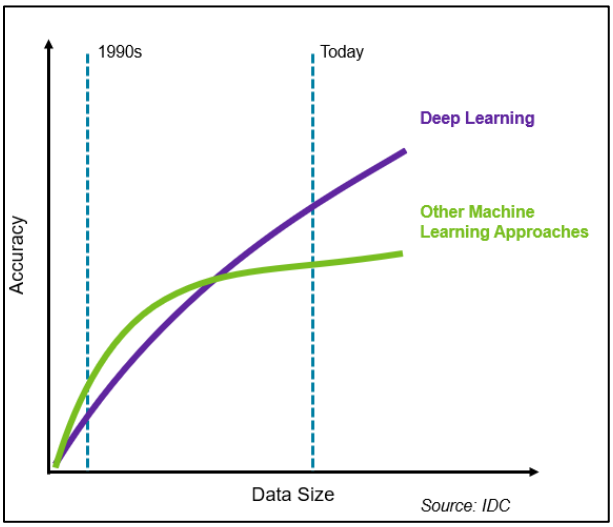
High-bandwidth memory leverages stacked memory components for density and high I/O counts

GDDR and HBM are the key products in the high-performance portfolio of memory. Let's take a look at the key trends in the markets.

New Market Trends & Applications

When listing applications that use high-performance memory, gaming usually comes to mind first. And while gaming is still very important, some exciting new market trends and applications are also driving demand in the graphics industry.

AI/ML and GPUs



Graphics processing units (GPUs) were used primarily for gaming. But now with the explosion of AI into almost all facets of industry, GPUs have become critical to delivering value and efficiency. GPUs predominantly demand high-performance memory. The algorithms used in training for machine learning and deep learning rely on sophisticated mathematical and statistical computations, and GPUs have proven more efficient at solving these complex computations than CPUs. When discussing AI, it is important to distinguish the different needs for inference and training. Training is very compute-intensive and requires the highest system bandwidth possible. Inference is more common and not quite as bandwidth hungry as training. Both are integral to the future and used together to achieve quality neural networks. GPUs and their enhanced memory are being leveraged as AI, ML and DL solve more real problems with better accuracy rates than humans can achieve.

High-Resolution Video

The next trend driving high-performance memory is 4K/8K content. High-end gaming pushes the envelope of graphics with higher resolution and faster response time (no lagging/buffering). Many of the top gaming rigs today are using 4K resolution, but 8K and higher is the future. Professional gamers often have multiple monitors, some driving 4K+ screens (a heavy workload that increases the need for graphics cards and large frame buffer).

Video rendering will continue to need large frame buffers and high bandwidth because video resolution grows. As the world moves to streaming options in both media content and gaming, data centers will need increasingly powerful data processing capabilities.

Gaming Innovation

Cloud gaming needs to be powered by the data center, and often these servers are supercharged with GPU power to optimize efficiencies. Google Stadia™, NVIDIA GeForce® Now, PlayStation Now™ and Microsoft® Project xCloud™ are some of the new platforms that have been introduced recently. Cloud gaming is projected to grow rapidly and continue to demand more innovation. Ray tracing is the holy grail of visual graphics. The ability to trace light from its origin to design realistic lighting environments has been a focus in the graphics industry for more than 20 years. This rendering technique is finally available in the latest iteration of graphics cards from NVIDIA and AMD along with the upcoming PlayStation 5 and Xbox™ Series X game consoles.

PC gaming is driving the highest specs in gaming. With the ability to update hardware yearly (or more frequently), PC gaming is the preferred choice of professional gamers. PC gaming continues to drive the need for increased graphics (4K/8K, ray tracing and variable rate shading) and the demand for maximum response time (minimal buffering/latency). As mentioned above about high-resolution video, professional gamers sometimes have multiple monitors running at the highest specs possible, a setup that requires continuous updates to their systems

to maintain a competitive edge. This promotes the use of gaming rigs with not just one graphics card but multiple graphics cards working in parallel to provide even better performance.

AR/VR

In both PC gaming and console gaming, virtual reality has become a popular option for a multitude of different games. Rudimentary at first, VR graphics and functionality are growing quickly and penetrating multiple new fields. Gaming will continue to grow at a healthy rate as the quality improves, but the more fascinating growth is in some exciting new areas.

Health care is a field where both VR and augmented reality (AR) are projected to become an important part of learning. Already, very interesting applications incorporate virtual objects onto real-life set pieces to teach in the medical field (using AR).

Micron's graphics solutions have moved beyond gaming into verticals like professional graphics, high-performance computing, automotive and networking.

Architecture, engineering and construction are obvious candidates for future use of VR and AR. The ability to virtually "see" or even tour an object or building before it has been built and understand how it may interact with its environment hold promise for multiple fields.

Education has enormous potential to drive increases in VR and AR business. As with medical applications, instructors and experts gain the ability to teach using virtual objects, giving virtual examples and interacting with virtual components.

High-end AR and VR headsets require a powerful PC and graphics card to maximize specs. The minimum graphics card for the HTC Vive™ Pro is a NVIDIA GeForce GTX 1060 or an AMD™ Radeon™ RX 480. As graphic requirements grow, the minimum graphics card will need to be updated, likely to one with the latest high-performance memory.

Workstations & Supercomputers (HPC)

For professional graphics and workstations, there is a large outstanding customer base in the workforce (video/photo graphic designers, CAD, engineering firms and others). Workloads that use a considerable amount of bandwidth need a strong graphics card and consistent upgrades to remain competitive. High-performance computing is building supercomputers that solve the world's most complex problems and are consistently requiring increases in memory performance. High-performance memory is a perfect match for HPC.

Automotive & Networking

High-performance memory has also entered the automotive and networking markets. Autonomous driving uses a massive amount of data to analyze the surroundings and processes that data extremely fast. GPUs, which use high-performance memory, are a perfect match for this work case. Autonomous driving continues to make significant progress and will drive substantial growth in the memory industry. On the networking side, high-end routers and switches need the performance and bandwidth capabilities of high-performance memory. For both automotive and networking, reliability and longevity are key focus areas.

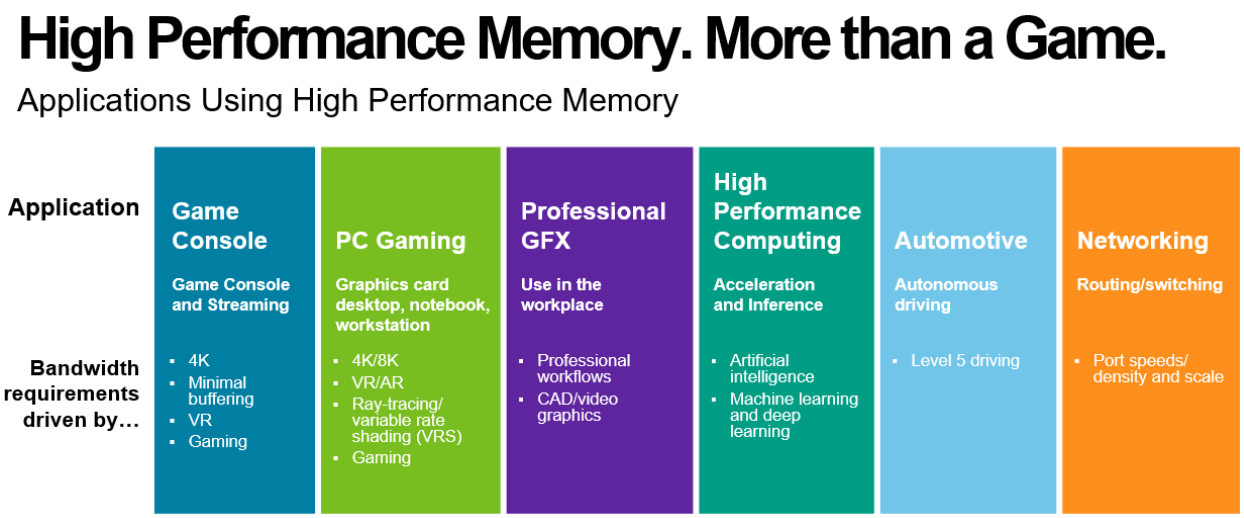
High-Performance Solutions in the Marketplace

High-performance solutions have moved beyond gaming into verticals like professional graphics, high-performance computing, automotive and networking. For many of the applications, memory and storage requirements differ. For instance, in game consoles, platform life span is long (5-7 years), so new consoles frequently come loaded with top-of-the-line specs to maximize bandwidth and performance, future-proofing the console and extending its use for multiple years. VR is a growing trend in consoles, and console VR is more cost-

friendly than PC gaming. VR headsets are an easy addition to the existing console customer base and individually more affordable since the VR compute needed comes from the game console already in place.

The **updateable graphics model** is used in the professional workplace, for CAD and video graphics, as well as in high-performance computing. PC gaming is also a potential area for continuous updates of graphics cards, but great options are also available for less enthusiastic gamers who don't need the maximum specs for each gaming experience. A new graphics card could be used multiple years and have a life span similar to a game console. HPC and data center applications should be able to upgrade platforms regularly to increase performance.

Networking and automotive will have much longer platform life spans, sometimes pushing 10 years. High-performance memory is well positioned to support this variety in requirements.



Characteristics That Drive Memory Systems

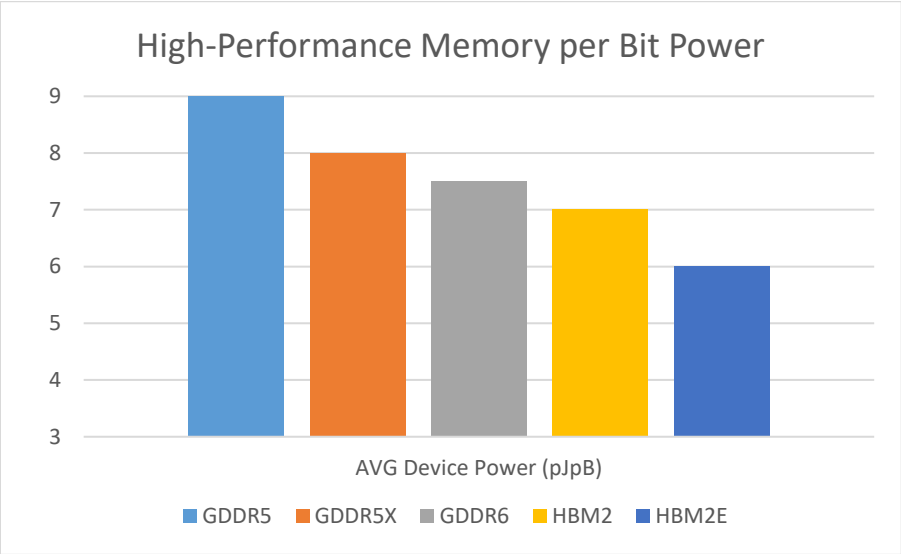
Analysis and trade-offs are necessary to decide what memory solution fits best for systems, considerations or parameters. Organizations must choose between mainstream memory or high-performance memory; they must even choose which form of high-performance memory best suits respective system requirements. In addition, the required frame buffer size has to be considered. GDDR type products offer a wide flexibility to configure system frame buffer size (from 4 pieces of memory to 24 pieces, any number of placements is possible) to match the system requirements.

Performance

As highlighted by market trends and history, continuously increasing performance needs to be supported. Graphics cards have upgrade cycles every year, which means higher bandwidth is in constant demand. High-performance memory is pushed to increase data rates each upgrade cycle. To maximize performance, users have to leverage the latest high-performance memory.

Power

While the performance requirement keeps increasing, power efficiency becomes more relevant. Users should look at power windows for their total solutions and match them to the bandwidth required.



Affordability

Many of the high-performance memory applications are consumer-based end applications that have cost-sensitive budgets. The memory needs to achieve bandwidth and frame buffer needs while maintaining a reasonable cost. Depending on the application and bandwidth needs, a discrete lower-cost option like GDDR may be the correct solution. If the bandwidth need outweighs the cost sensitivity, then HBM is likely the answer.

When considering affordability, users need to look beyond the device costs to system implementation costs as well. Discrete components that can be implemented with standard board material and mainstream assembly processes make an attractive total cost of ownership (TCO).

A Glimpse Into the Future

HBM2E/HBMnext

HBM2E is projected to enter the market in 2020. Micron plans to introduce HBM2E products in the second half of 2020. Our device will be fully JEDEC-complaint and available in 4H/8Gb and 8H/16GB densities, with data rates of 3.2Gb/s or potentially higher.

HBMnext is projected to enter the market toward the end of 2022. Micron is fully involved in the ongoing JEDEC standardization. As the market demands more data, HBM-like products will thrive and drive larger and larger volumes.

GDDR ... Next?

The need for an affordable high-bandwidth solution in a discrete package is not going away. GDDR6 is still young in its lifetime compared to GDDR5. But conversations about the future of GDDR are ongoing.

Conclusion

Pressure from the industry to increase data rates will continue and Micron will continue to flex our innovation muscle in the near future. Stay tuned for more from Micron on delivering high-performance memory solutions. Follow us on @MicronTech and LinkedIn.

Spencer Homan

Marketing Manager Graphics, CNBU
Micron Technology, Inc.

Spencer Homan drives marketing actions as a manager in Micron's global Graphics Memory Business department. Homan's marketing and execution of Micron's broad portfolio includes high-speed memory solutions that serve the game console and high-end graphics market. He has contributed his experience in the graphics segment for over five years since joining Micron in 2014. Homan's key roles include engaging with customers and market enablers on Micron's strategy around products, portfolio and market positioning, which have led to Micron's leadership in positioning graphics products.

About Micron Technology, Inc.

We are an industry leader in innovative memory and storage solutions. Through our global brands — Micron® and Crucial® — our broad portfolio of high-performance memory and storage technologies, including DRAM, NAND, 3D XPoint™ memory and NOR, is transforming how the world uses information to enrich life. Backed by more than 40 years of technology leadership, our memory and storage solutions enable disruptive trends, including artificial intelligence, 5G, machine learning and autonomous vehicles, in key market segments like mobile, data center, client, consumer, industrial, graphics, automotive, and networking. Our common stock is traded on the Nasdaq under the MU symbol. To learn more about Micron Technology, Inc., visit micron.com.

- Page 1 table, source: IDC, Global Datasphere reporting, such as https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/
- Page 6 table, source: Synopsys, https://www.techdesignforums.com/practice/technique/choosing-between-ddr4-and-hbm-in-memory-intensive-applications/

micron.com