IDC ANALYZE THE FUTURE

IDC MARKET SPOTLIGHT

Sponsored by: Rambus

This IDC Market Spotlight highlights the history of DRAM in enabling major applications and describes how DRAM will be a key enabler for pervasive AI.

AI Requires Tailored DRAM Solutions

April 2020

Written by: Shane Rau, Research Vice President, Computing Semiconductors, and Soo-Kyoum Kim, Associate Vice President, Memory Semiconductors

Introduction

Artificial intelligence (AI) is the study and application of providing hardware and software that attempt to emulate human intelligence. Therefore, as human beings adapt to different tasks, AI must adapt. Indeed, if AI is to scale to the diversity of human tasks, it will not be via a single, monolithic solution. For every task a human can perform, AI must adapt through a unique combination of hardware and software components.

DRAM's History of Adaptation

The history of dynamic random-access memory (DRAM) is characterized by adaptation to the increasingly specialized needs of its applications. In the 1990s, PCs dominated the market and DRAM scaled in performance to meet the needs of those general-purpose systems. In the 2000s, PC graphics cards and gaming consoles demanded more specialized DRAM types to reduce delay (latency) between a user's actions and the rendering of high-resolution graphics and smooth video on screens. In the 2010s, mobile phones demanded low-power DRAM (LPDRAM), and servers demanded greatly increased capacity as cloud computing soared.

AT A GLANCE

KEY STATS

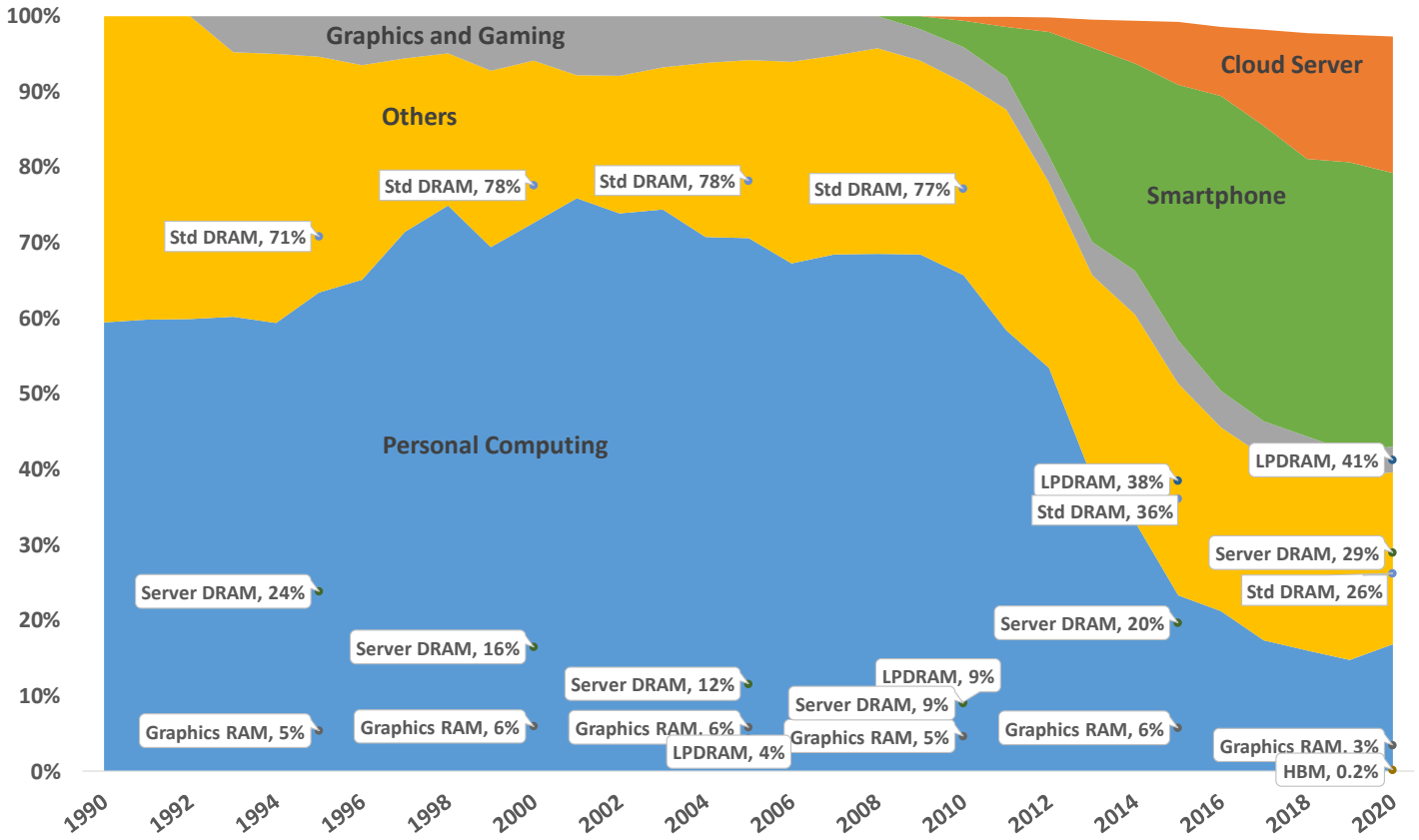
- In 2030, the DRAM required by electronic systems worldwide will amount to over 630,000 times that required in 1990.
- By 2030, non-PC applications will drive 90% of DRAM demand, and all of them will incorporate AI.

WHAT'S IMPORTANT

- DRAM's history follows the technology's adaptation to the needs of various system types.
- AI will become as ubiquitous in electronic systems as the human beings who use those systems.
- DRAM technologies such as GDDR6, HBM, and memory buffers that move, accelerate, and store data will be critical to enabling AI for the next 10 years and beyond.

Figure 1 illustrates how the DRAM market segmented with different types as new major system categories emerged.

FIGURE 1: ***DRAM Market Bit Share by Major System Category and Major DRAM Type, 1990–2020***



Note: New DRAM types emerge to serve the specialized needs of new major system categories. Standard DRAM, including SDRAM and DDR, has served much of the PC and server markets through DRAM's history. However, from about 2010 and beyond, LPDRAM emerged to serve the low-power needs of smartphones and server-specific DRAM emerged to serve the extensive capacity needs of off-premises cloud servers of cloud service providers.

Source: IDC, 2020

Adapting DRAM to AI: All of the Above

If DRAM has already adapted to a half dozen or so major system categories, why is AI so tough?

Why AI Is So Tough: The Technical Needs

AI is tough because each AI system application's unique combination of hardware and software components varies in its requirements for performance, latency, power, and capacity, in addition to the requirements for reliability, design complexity, and cost. Further, AI must continually adapt because the technical needs of its applications continually evolve.

Beyond the evolving requirements of each AI application are the unique requirements of the mode of AI itself. AI training is the mode of learning a new capability from existing data. AI inferencing is the mode of applying the capability to new data. For example, to expose an AI model to pictures of one million cats so the model can learn to recognize a cat is training. To have an AI model recognize a cat from a new picture of a cat is inferencing.

AI training and AI inferencing have different requirements. Training requires large amounts of data from which the AI model can learn and requires high performance to make the processing of the data practical. Moreover, the data models themselves are growing in complexity. In turn, large amounts of data and complex models require large amounts of power. For these reasons, AI training is typically done in servers within large datacenters.

Inferencing requires high performance to compare new data against the AI model and low latency to report the result in real time. AI inferencing can be done in servers for their performance, but the requirement for low latency means that AI inferencing is evolving toward edge infrastructure and Internet of Things (IoT) endpoint systems where they can be near or inside the system doing the inferencing. For this reason, systems doing AI inferencing tend to be adapted to the performance, power, capacity, reliability, design complexity, and cost constraints of the host system.

The difficulty of adapting to AI's technical needs can't be overstated. Across all the segments and modes of AI, the answer is often one of "all of the above," with bandwidth, latency, power, and capacity all having challenging requirements to be met. Solutions such as NVIDIA's Tesla V100 GPUs with up to 32GB of HBM2 memory have been introduced to support the needs of large training data sets and models. As training models continue to grow, a steady stream of increasingly powerful solutions will be required.

Why AI Is So Tough: Market Needs

AI is also tough because while its technical needs require specialization — each application has a unique combination of requirements — its market needs require technological generalization, meaning standardization and scalability, so that AI systems can be affordable.

IDC research tracks and forecasts more than 300 major electronic system categories for their market size and the technologies that they consume, including AI. No system category market is so large that it can justify the cost of a single, proprietary memory technology; a $500 memory chip for a $500 system is untenable. Instead, nearly all system categories across industries often share basic core technologies that are then adapted for the system. The core technology is standard; the adaptation is scalability.

DRAM is a case in point because, through its history, the memory core where data is stored has largely gone unchanged. However, its adaptability — the ability to scale to the need of its application — has had less to do with the technology of that core than with the doorway into the core (the DRAM interface), the path to that doorway (the memory bus), and the intelligence devoted to the memory (the memory controllers and memory buffers). For example:

"Since the fundamental DRAM cell and array has maintained the same basic structure for many years, the types of DRAM are mainly distinguished by the many different interfaces for communicating with DRAM chips."
— *Wikipedia*

- As shown in Figure 1, standard DRAM, such as synchronous DRAM (SDRAM) and double data rate SDRAM (DDR-SDRAM), has served much of the PC and server markets over time despite the different needs of each system type. When SDRAM replaced the preceding asynchronous DRAM, its innovation was the addition of a clock that synchronized the timing of access to the data core with the system clock. When DDR SDRAM replaced SDRAM, its innovation was the doubling of the clock, an innovation that has continued through generations of DDR1, DDR2, DDR3, DDR4, and so on.
- Memory controllers direct the reading and writing of data to DRAM. In 2003, CPU manufacturers began moving the memory controller from a satellite chip, often the northbridge, to their CPUs. Creating a direct link between the CPU and the memory, these integrated memory controllers accelerated DRAM performance.

- DRAM chips are usually contained on their own printed circuit board, a memory module. Memory buffers bring local intelligence to a DRAM memory module. By making the memory modules more intelligent, memory buffers expand the effective memory capacity of the module.
- Memory buses carry signals between memory controllers and DRAM chips in systems. Over many years, DRAM manufacturers have widened the memory bus to reduce data congestion and increase bandwidth back and forth between the controllers and the DRAM chips.

The Next 10 Years

The First Wave of AI

IDC's tracking of AI penetration reveals that initial AI systems were servers built by cloud service providers to train their AI algorithms. Notably, these cloud servers, hubs of data processing, often contain different kinds of processing chips. General-purpose microprocessors often connect to and share workloads with specialized processors, such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and custom application-specific integrated circuits (ASICs). CPUs, GPUs, FPGAs, and ASICs, however, share a need for external memories — such as DDR DRAM for a CPU's main memory and graphics DDR (GDDR) for a GPU's frame buffer— that hold data close to the processor. These needs thus mean that interfaces and memories, as well as intimacy between the two, are critical to making data signals flow efficiently. Thus, DRAM is adapting to different use cases in multiple forms within the same system, meeting requirements for capacity, performance, latency, and power management; this is the leading edge of what AI demands in a system.

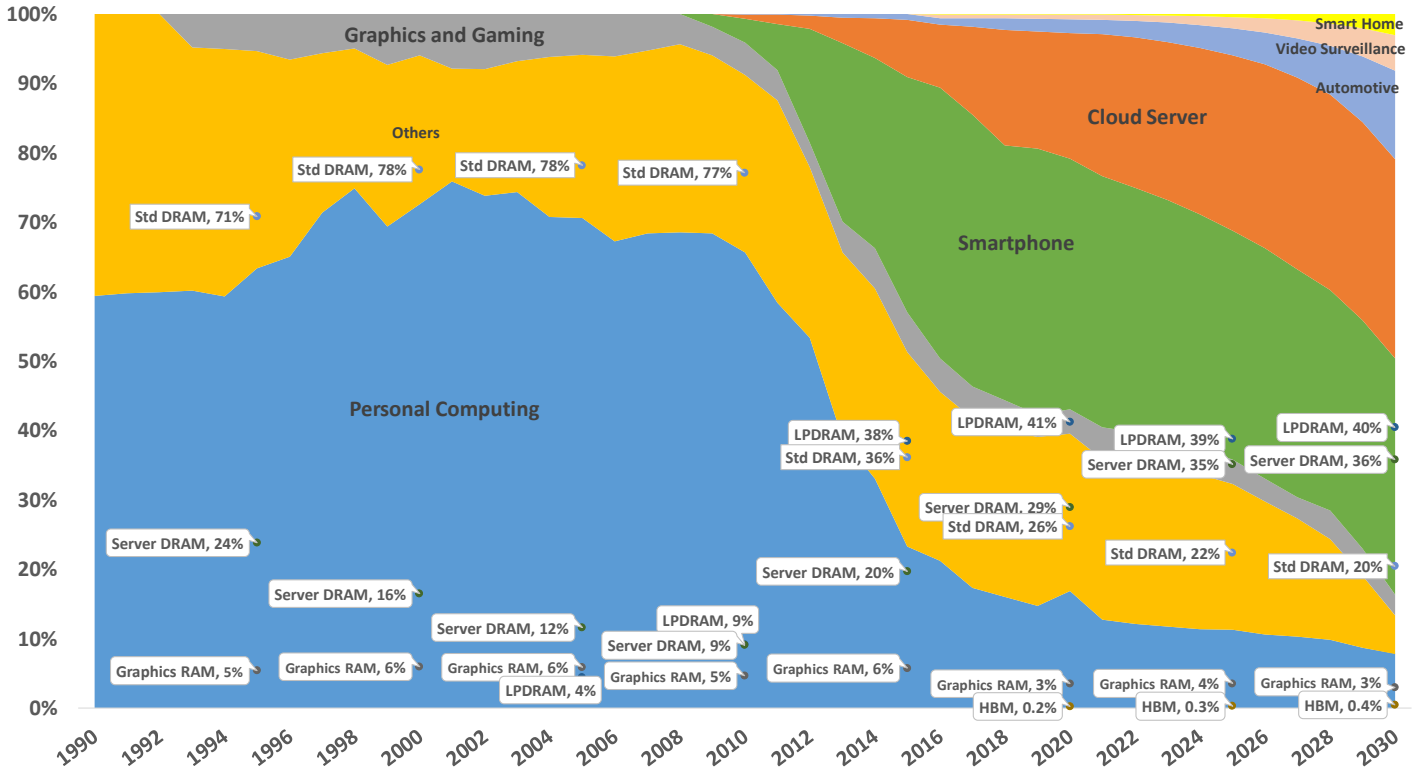
The Next Waves of AI

As Figure 2 illustrates, the second wave of AI penetration, which accelerates over the next five years, is in smartphones, which need to listen to and interpret user requests and respond in real time. In 2023, 100% of smartphones shipped worldwide will be AI enabled. IDC forecasts that in the next 5 to 10 years, several more markets will drive the DRAM market and, not coincidentally, drive AI penetration:

- Automotive systems, including advanced driver-assistance systems (ADAS) and automotive control devices, receive much attention now for their role in enabling fully autonomous vehicles. While the need for minimal latency is compelling, AI penetration that enables fully autonomous vehicles will take another 5 to 10 years to develop. In 2023, 100% of the ADAS (consisting of radar systems, control systems, and visual systems) that ship in automobiles will be AI enabled.
- Smart home speakers currently rely on the algorithms based in datacenters (today's generation of smart home speakers, such as Google Home, don't infer responses locally). However, reducing response times (latency) requires more local intelligence. IDC forecasts that from 2025 to 2030, 100% of smart home speakers will be locally AI enabled.
- Video surveillance systems require a large amount of data processing power to process incoming camera data. Interpreting data in real time will require local AI, either in the cameras or in local edge servers and storage systems.

Collectively, these waves of AI-enabled system categories mean that DRAM technologies such as standard DRAM, LPDRAM, GDDR6, HBM, and supporting technologies such as memory buffers and integrated memory controllers that move, accelerate, and store data will be critical to enabling AI for the foreseeable future. Notably, these applications are very different, and to adapt, AI will require different DRAM solutions for each application. As Figure 2 illustrates, a range of memories will be needed: DDR, GDDR, HBM, and LPDDR.

FIGURE 2: ***DRAM Market Bit Share by Major System Category and Major DRAM Type, 1990–2030***



Note: DRAM scales to serve the needs of new major system categories. The initial wave of AI penetration has been in servers deployed by cloud service providers, which have demanded more DRAM capacity and reduced latency. In the next five years, IDC expects smartphones to demand more DRAM capacity, higher performance, and reduced latency. IDC expects that from 2025 to 2030, automotive, video surveillance, and smart home categories, enabled by AI, will demand more DRAM capacity, higher performance, reduced latency, reduced power, and higher reliability.

Source: IDC, 2020

Definitions

- **Artificial intelligence (AI):** Artificial intelligence is the simulation of human intelligence by machines.
- **Cloud:** The cloud is not a physical entity; rather, it is a network of local and remote servers worldwide that are hooked together so that they operate as a single, virtual system.
- **Dynamic random-access memory (DRAM):** This is the generic term for memory that stores data, but the data needs to be refreshed consistently.
- **Single data rate SDRAM (SDR-SDRAM):** This is a legacy form of DRAM with higher bandwidth due to synchronizing its bus clock speed with that of the system's CPU.
- **Double data rate SDRAM (DDR-SDRAM):** This is a form of DRAM with higher bandwidth due to a doubling of its bus clock speed.

- **Graphics double data rate (GDDR):** GDDR is a form of DRAM but with higher bandwidth due to a wider memory bus than that of standard DRAM, such as DDR. For example, the latest version of GDDR, GDDR6, supports bandwidth up to 16 gigabits per second (Gbps). Notably, as GPUs have expanded from drawers of pixels on a screen to high-performance processors of highly parallelized data, so too has GDDR memory. For example, GPUs are seeing new applications in automotive and AI applications, and IDC expects GDDR memory to follow.
- **High-bandwidth memory (HBM):** HBM stacks multiple DRAM chips into a single chip package to increase the amount of memory that a system can address (talk to) per unit of time. High-bandwidth memory stacks multiple DRAM chips in a 2.5D package with a wider interface and lower clock speed than DDR4 to create a small form factor, higher bandwidth-per-watt solution for high-performance computing. The latest iteration, HBM2E, supports total bandwidth of 410 gigabytes per second (GBps) across eight 128-bit independent channels (1,024-bit wide data interface) and supports stacks of from 2 to 12 DRAMs.
- **Machine learning (ML):** ML is a form of AI; however, it focuses on the ability of machines to adapt to changing conditions on their own.
- **Memory buffer:** Memory buffers bring local intelligence to a DRAM memory module. By making the memory modules more intelligent, memory buffers expand the effective memory capacity of the module.

Conclusion

AI is one of the transformative trends of our time, and the overall journey will be long but characterized by rapid change by sector. AI will be ubiquitous but not monolithic. Notable among the system categories that will adopt AI are the disparate use cases they will serve. Servers, phones, smart home speakers, and cars vary significantly from each other in the kinds of AI solutions that they will need. Technology buyers will demand that technology providers deliver disparate solutions to enable these systems. No one solution will fit all. AI requires tailored solutions, including from DRAM.

About the Analysts

Soo-Kyoum Kim, *Associate Vice President, Memory Semiconductors*

Soo-Kyoum Kim is Associate Vice President for memory semiconductor research. Mr. Kim's research covers demand and supply analysis for DRAM and NAND, memory consumption for server workloads, next-generation memory, emerging memory market study for automotive, video surveillance, and other key IoT areas.

Shane Rau, *Research Vice President, Computing Semiconductors*

Shane Rau leads IDC's computing semiconductor research covering microprocessors and SoCs, discrete graphics processing units (GPUs), FPGAs, and artificial intelligence (AI) accelerators in systems across the internet, including in the datacenter, in PCs, at the edge, and at endpoints.

MESSAGE FROM THE SPONSOR

About Rambus

Rambus is a premier silicon IP and chip provider that makes data faster and safer. With 30 years of innovation, the company continues to develop foundational technology for all modern computing systems. Rambus is a leader in HBM memory interfaces and offers silicon-proven solutions for the latest generation HBM2E memory with an integrated interface consisting of a co-verified PHY and digital controller. For GDDR6 memory, Rambus has demonstrated the world's fastest GDDR6 interface operating at 18 Gbps. Here too, Rambus offers an integrated solution of co-verified PHY and controller which simplifies integration complexity and helps speed time to market. For more information, visit ***rambus.com***.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on **www.idc.com**.

IDC Research, Inc.
5 Speen Street
Framingham, MA 01701, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com