WHITE PAPER

Which DDR SDRAM Memory to Use and When

Author

Vadhiraj Sankaranarayanan
Sr. Technical Marketing Manager,
Synopsys

Overview

Memory performance is a critical component for achieving the desired system performance in a wide range of applications from cloud computing and artificial intelligence (AI) to automotive and mobile. Dual Data Rate Synchronous Dynamic Random-access Memory (DDR SDRAM) or simply DRAM has emerged as the de facto memory technology for the main memory due to its many advantages: high density with simplistic architecture using a capacitor as a storage element, low latency and high performance, almost infinite access endurance, and low power. The Joint Electron Device Engineering Council (JEDEC) has defined several DRAM categories of standards to meet the power, performance, and area requirements of each application.

Selecting the right memory solution is often the most critical decision for obtaining the optimal system performance. This whitepaper provides an overview of the JEDEC memory standards to help SoC designers select the right memory solution, including IP, that best fits their application requirements.

DDR DRAM Standards

The primary function of main memory in an SoC is to feed the host – CPUs and GPUs – with the necessary data or instructions as quickly and reliably as possible. While the demand for high performance is increasing, more cores and functionality are added to the System-on-Chip (SoC) that is growing the need to keep the overall silicon footprint small and system power down.

DDR DRAMs meet these memory requirements better than any other storage technologies overall, by offering a dense, high-performance, low-power memory solution. Furthermore, DDR DRAMs can be used in different form-factors depending on the system requirements — either on a dual in-line memory module (DIMM) or as a discrete DRAM solution. JEDEC has defined and developed the following three DRAM categories of standards to help designers meet their power, performance, and area requirements for their target applications:

- **Standard DDR** targets servers, cloud computing, networking, laptop, desktop, and consumer applications, allowing wider channel-widths, higher densities, and different form-factors. DDR4 has been the most popular standard in this category since 2013; DDR5 devices are expected to become available in the near future.
- **Mobile DDR** targets mobile and automotive applications, which are very sensitive to area and power, offering narrower channel-widths and several low-power operating states. The de facto standard today is LPDDR4 with LPDDR5 devices expected in the near future.
- **Graphics DDR** targets data-intensive applications requiring a very high throughput, such as graphics-related applications, data center acceleration, and AI. Graphics DDR (GDDR) and High Bandwidth Memory (HBM) are the standards in this category.

The three DRAM categories use the same DRAM array for storage with a capacitor as the basic storage element. However, each category offers unique architectural features to optimally meet the requirements of the target applications. These features include customizations around data rates/data widths, connectivity options between the host and DRAMs, electrical specifications, termination schemes for the I/Os (Input/Output), DRAM power-states, etc. Also, each category has multiple generations of standards, with the successor standard outperforming its predecessor by offering higher speed and density, as well as lower power. Figure 1 illustrates JEDEC's three categories of DRAM standards.

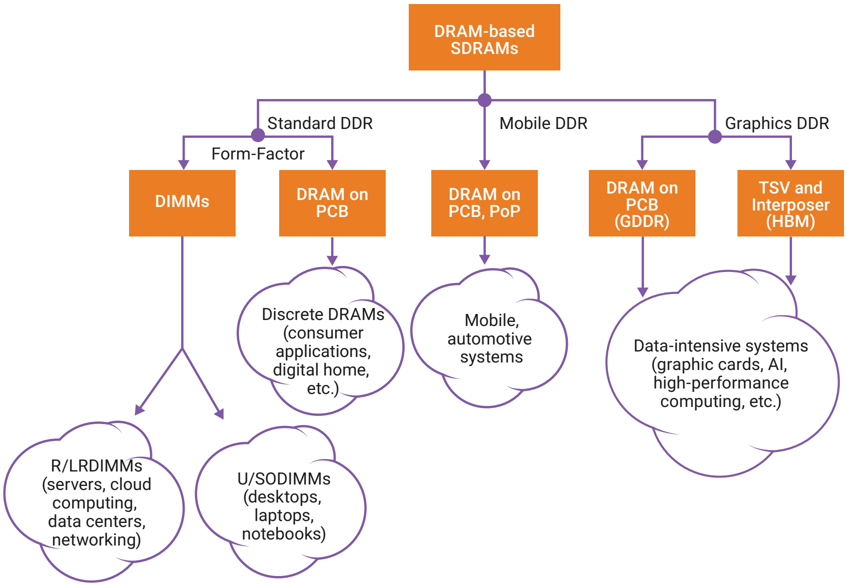
DRAM-based SDRAMs
Standard DDR
Mobile DDR
Graphics DDR
Form-Factor
DIMMs
DRAM on PCB
DRAM on PCB, PoP
DRAM on PCB (GDDR)
TSV and Interposer (HBM)
Discrete DRAMs (consumer applications, digital home, etc.)
Mobile, automotive systems
Data-intensive systems (graphic cards, AI, high-performance computing, etc.)
R/LRDIMMs (servers, cloud computing, data centers, networking)
U/SODIMMs (desktops, laptops, notebooks)

Figure 1: JEDEC has defined three widely used categories of DRAM standards to fit the needs of various applications

The following section provides a high-level overview of each DRAM category, and the common applications in which these devices are used.

Standard DDR

Standard DDR DRAMs, are ubiquitous in applications such as cloud servers, data centers, laptop, desktop, and consumer, providing high-density and high-performance in various types and form factors. DDR4 is the most popular standard in this category, offering several performance advantages over its predecessors – DDR3 and DDR3L (a low-power version of DDR3):

- Higher data rate, up to 3200 Mbps, as compared to DDR3 operating up to 2133 Mbps
- Lower operating voltage (1.2V, as compared to 1.5V in DDR3 and 1.35V in DDR3L)
- Higher performance (e.g., bank-groups), lower power (e.g., data-bus inversion), and higher Reliability, Availability, and Serviceability (RAS) features (e.g., post-package repair and data cyclic redundancy check)
- Higher densities due to an increase in the individual DRAM die sizes from 4Gb to 8Gb and 16Gb

Standard DRAMs support data widths of 4 (x4), or 8 (x8), or 16 (x16) bits. Servers, cloud, and data center applications use the x4 DRAM chips as they allow a higher density on DIMMs and support higher RAS features for minimizing the downtime of these applications during memory-related failures. The x8 and x16 DRAMs are less expensive and are commonly implemented in desktops and notebooks.

DDR5, actively under development at JEDEC, is expected to increase the operating data rates up to 4800Mbps at an operating voltage of 1.1V. DDR5 has several new architectural and RAS features to handle these high speeds effectively and minimize the system downtime due to memory errors. Integrated voltage regulators on modules, better refresh schemes, an architecture targeted at better channel utilization, internal error-correcting code (ECC) on DRAMs, increased bank group for a higher performance, and higher capacity are a few of the key features in DDR5.

DDP, QDP, 3DS DRAMs

Multiple DRAM dies are often packaged together to increase the density of standard DRAMs. Typically, individual DRAM dies are packaged as dual-die (DDP) or quad-die (QDP) packages to support 2 or 4 memory ranks respectively in the DRAM package. Expanding beyond 4 ranks limits the operating speed since the loading of the Command/Address (C/A) and data (DQ) wires increases significantly when connecting to all memory ranks in parallel. Instead, a Through-Silicon Via (TSV) stacking process supports higher memory ranks, where each memory rank (DRAM-die) is stacked vertically in a DRAM package. In TSV DRAM packages, the bottom die acts as an electrical buffer for the C/A and DQ wires, hence, a TSV DRAM acts as a single load to the host for both C/A and DQ wires. This allows TSV DRAM packages to have as many as 8 or even 16 stacked memory ranks. TSV DRAMs are also called 3D-stacked (3DS) DRAMs. A DIMM based on several TSV DRAMs can support hundreds of gigabytes of memory.

DIMM Types

DIMMs are printed circuit board (PCB) modules with several DRAM chips to support either a 64-bit data width or a 72-bit data width. 72-bit DIMMs are called ECC DIMMs since they support 8 bits of ECC in addition to the 64 bits of data. The ECC DIMMs are used in server and data center applications for protection against single-bit errors, while the non-ECC DIMMs (having a 64-bit data width) are used in desktops and laptops.

As servers and data centers need terabytes of memory, they typically support 2 or 3 DIMMs in every memory channel. Since DIMMs in the same memory channel share the same C/A and DQ wires, additional buffering is needed on each DIMM to reduce the electrical loading of these wires, before the outputs are distributed to the individual DRAMs. DIMMs that buffer only the C/A wires are called Registered DIMMs (RDIMMS), and DIMMs that buffer both C/A and DQ wires are called Load-Reduced DIMMs (LR-DIMMs). Both RDIMMs and LR-DIMMs are typically of the ECC type, supporting a 72-bit data width. Desktop applications are extremely power and cost constrained, usually requiring only a single DIMM per channel configuration and non-ECC unbuffered DIMMs (UDIMMs).

All the DIMMs discussed thus far (R/LR/UDIMMs) have the same form-factor. Applications with stringent area constraints (such as laptops, notebooks, office printers, etc.) use Small-Outline DIMMs (SODIMMs) as their memory solution since SODIMMs are roughly half the size of regular DIMMs (R/LR/UDIMMs). Table 1 shows a summary of different DIMM types with their typical densities.

DIMM Type	Description
RDIMM	• Only C/A is buffered on the DIMMs • Employed on Servers and data centers • Densities: 8, 16, 32GB, 64GB
LR-DIMM	• Both C/A and DQ are buffered on the DIMMs • Employed on Servers and data centers • Densities: 32GB, 64GB
TSV DIMM	• Provides the maximum density in standard DDR DRAM category. These DIMMs can be either of 'R' or 'LR' type • Employed on Servers and data centers • Densities: 128GB, 256GB
UDIMM	• Employed on desktops and home servers • Can be ECC or non-ECC based • Densities: 2, 4, 8, 16GB
SODIMM	• Employed on systems with space constraints, such as laptops, notebooks, high-end upgradable office printers, and networking hardware. • Can be ECC or non-ECC based • Densities: 2, 4, 8, 16GB

Table 1: Commonly used DIMMS provide different densities for the target application

Consumer applications that cannot accommodate a DIMM form-factor due to area constraints implement discrete DRAM solutions for the main memory. Such applications can use either standard DRAM devices or mobile DRAM devices supporting LPDDR5/4/4X/3 standards.

Mobile DDR

Mobile DDR, also called Low-Power DDR (LPDDR) DRAMs, offer identical memory storage array as in the standard DDR DRAM, however, LPDDR DRAMs have several additional features for achieving low power, which is a key requirement for mobile applications including tablets, mobile phones, automotive systems, and SSD cards. As such applications tend to have fewer memory devices on each channel and shorter interconnects, the LPDDR DRAMs can run faster than standard DRAMs, providing a higher performance. LPDDR DRAMs in low-power states help achieve the highest power efficiency and extend battery life. LPDDR DRAM channels are typically 16 or 32-bit wide, in contrast to the standard DDR DRAM channels which are 64 bits wide. Just as with standard DDR DRAM generations, each successive LPDDR generation targets a higher performance and lower power than its predecessor, and no two LPDDR generations are compatible with one another.

LPDDR5/4/4X DRAMs

LPDDR4 is the most popular standard in this category, capable of data rates up to 4267 Mbps at an operating voltage of 1.1V. LPDDR4X, a variant of LPDDR4, is identical to LPDDR4 architecturally, but provides additional power savings by reducing the I/O voltage (VDDQ) to 0.6V. LPDD4RX devices also run up to 4267 Mbps.

LPDDR4/4X DRAMs are typically dual-channel devices, supporting two x16 (16-bit wide) channels. Each x16 channel is independent and hence has its own dedicated C/A pins. The two-channel architecture provides a lot of flexibility to system architects while connecting the host or SoC to an LPDDR4/4X DRAM. LPDDR4/4X are point-to-point standards, and it is typical for these DRAMs to have no more than 2 ranks per channel. LPDDR4/4X DRAMs offer several options for the individual dies to obtain the various densities required for the two channels. The simplest die option has both channels in a x16 configuration. Two such dies can be packaged together to obtain a dual-ranked, dual-channel DRAM. Higher densities can be achieved by having dies that support both channels in a byte-mode (x8) configuration. Four such dies can be packaged together to give the same dual-ranked, dual-channel topology. Hence, the byte-mode-based LPDDR4/4X DRAMs allow higher densities such as 32Gb devices with four 8Gb DRAM dies, with each DRAM die supporting only a byte from both channels.

LPDDR4/4X DRAMs can be used as a discrete DRAM solution or as a Package-on-package (PoP) solution. For systems with a requirement for more channels, quad-channel-based x64 DRAMs are also available.

LPDDR5, the successor to LPDDR4/4X, is expected to run up to 6400Mbps, and is actively under development at JEDEC. LPDDR5 DRAMs are expected to provide many new low-power and reliability features, making them ideal for mobile and automotive applications. One such important low-power feature for extending battery life is the "deep sleep mode," which is expected to provide substantial power savings during idle conditions. In addition, there are several new architectural features that allow the LPDDR5 DRAMs to operate seamlessly at these high speeds at a lower operating voltage than LPDDR4/4X.

Connectivity Options with LPDDR4/4X DRAMs

The simplest option for connectivity to a dual-channel, dual-ranked LPDDR4 DRAM is shown in figure 2. The host operates the two DRAM channels as independent channels by sending separate C/A (CA_A and CA_B) wires to these two channels. CS_A[1:0] and CS_B[1:0] are the chip-select wires targeting the two ranks of each channel respectively. This configuration allows the host to implement each channel in different power states, targeting maximum performance and power efficiency.

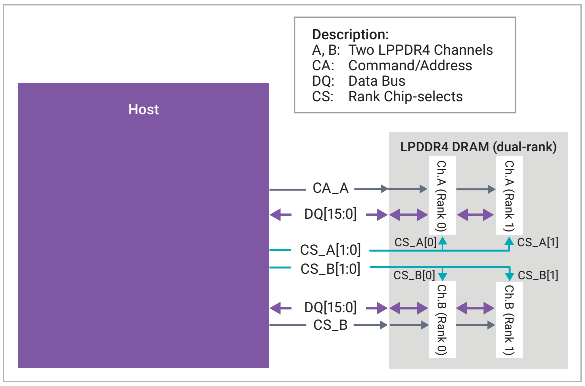

Figure 2: LPDDR4/4X Dual-channel connectivity options

Instead of operating the two x16 channels as independent channels, the host can alternatively view the two channels as a single-channel either with twice the data width (x32) or with the same data width (x16) but twice the channel density. In the first option, the host operates the DRAM as a x32 channel and sends only one pair of C/A (CA_A) signals to both x16 DRAM channels. This configuration can be used in applications requiring a larger data-width. In the second option, the host operates the two x16 DRAM channels as one x16 channel by sending one pair of C/A (CA_A) signals to both DRAM channels and connecting the DQ[15:0] from both DRAM channels together. The result is a single x16 quad-ranked channel. Although this connectivity option doubles the memory density of the channel itself, it significantly loads the C/A and DQ wires, which can limit the highest achievable data rate. Careful signal-integrity/power-integrity analysis of the memory channel is necessary to ensure channel robustness at desired speeds.

Graphics Standards

The next two sections explain GDDR and HBM, two disparate memory architectures targeting high throughput applications, such as graphic cards and AI. GDDR and HBM take different approaches to meet the high throughput needs of such applications.

GDDR Standard

GDDR DRAMs are specifically designed for GPUs and accelerators. Data-intensive systems such as graphic cards, game consoles, and high-performance computing including automotive, AI, and deep learning are a few of the applications where GDDR DRAM devices are commonly used. GDDR standards are architected as point-to-point (P2P) standards, capable of supporting 8 Gb/s and higher data rates. GDDR5 and GDDR6 are the most popular standards today supporting data-rates as high as 8 Gb/s and 16 Gb/s respectively.

GDDR5 DRAMs are 32-bits wide and always used as discrete DRAM solutions. They can be configured to operate in either ×32 mode or ×16 (clamshell) mode. GDDR5X, a variant of GDDR5, targets a transfer rate of 10 to 14 Gb/s per pin, almost twice that of GDDR5. The key difference between GDDR5X and GDDR5 DRAMs is that GDDR5X DRAMs have a prefetch of 16N, instead of 8N. GDDR5X also uses 190 pins per chip, compared to 170 pins per chip in GDDR5. Hence, GDDR5 and GDDR5X standards require different PCBs. GDDR6, the latest GDDR standard, supports a higher data-rate, up to 16 Gb/s, at a lower operating voltage of 1.35V, compared to 1.5V in GDDR5. GDDR6 DRAMs have two channels, each 16-bits wide, capable of servicing multiple graphic program threads simultaneously.

HBM/HBM2 Standards

HBM is an alternative solution to GDDR memories for GPUs and accelerators. GDDR memories target higher data rates with narrower channels to provide the needed throughput, while HBM memories solve the same problem through multiple independent channels and a wider data path per channel (128 bits per channel), operating at 2.4 Gb/s data rates. For this reason, HBM memories provide high throughput at a lower power and substantially smaller area than GDDR memories. HBM2 is the most popular standard today in the graphics standards category. HBM2 DRAMs stack up to eight DRAM dies, including an optional base die, offering a small silicon footprint. Dies are interconnected through TSV and micro-bumps. Commonly available densities include 4 or 8GB per HBM2 package. As HBM2 DRAMs support thousands of data wires, interposers are used to connect the memory and host. The interposer makes the routing shorter and direct, providing faster connectivity. A downside to the interposer requirement and the overall packaging process is the higher cost of the final product. Figure 3 shows a cross-section of an HBM memory.

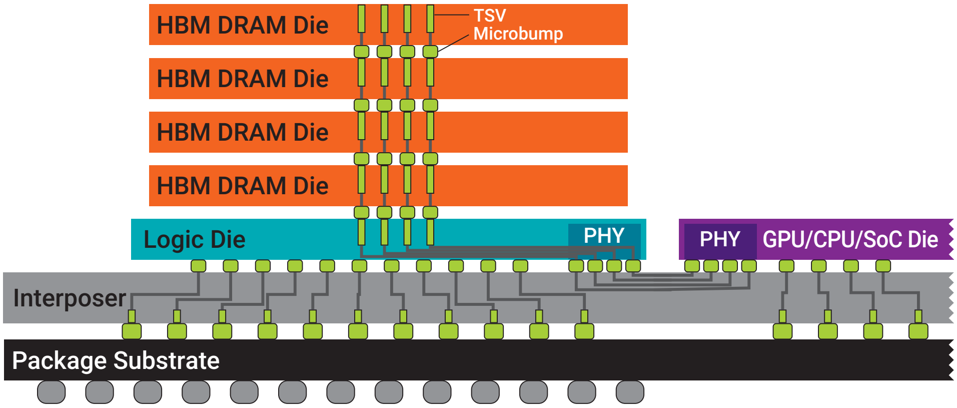

Figure 3: Cross-section of an HBM Memory stack (Source: AMD)

Besides supporting a higher number of channels, HBM2 also provides several architectural changes to boost the performance and reduce the bus congestion. For example, HBM2 has a 'pseudo channel' mode, which splits every 128-bit channel into two semi-independent sub-channels of 64 bits each, sharing the channel's row and column command buses while executing commands individually. Increasing the number of channels also increases overall effective bandwidth by avoiding restrictive timing parameters such as tFAW to activate more banks per unit time. Other RAS features include optional ECC support for enabling 16 error detection bits per 128 bits of data. HBM2E, a derivative of HBM2, is expected to provide a higher performance and higher density than HBM2. HBM2E devices are expected to be come available in the near future. Table 2 shows a high-level comparison of GDDR6 and HBM2 DRAMs:

Item	GDDR6	HBM2
DRAM density	16 Gb (per chip)	64 Gb (per stack)
# Channels / DRAM package	2 channels	8 channels
# Bits in a channel	16 bits	128 bits
Speed	16 Gb/s	2.4 Gb/s
Overall bandwidth	64 GB/s	307 GB/s
Power efficiency		Better than GDDR6
Cost	Lower cost than HBM2	
Packaging process	Traditional DRAM on PCB	Uses a 2.5D Interposer for connectivity between the host and the HBM2 DRAMs

Table 2: GDDR6 and HBM2 offer unique advantage for system architects

Summary

SoC designers can select from a variety of memory solutions or standards to meet the specific needs of their target applications. The selected memory solution impacts the performance, power, and area requirements of the SoC. To provide a wide selection of technologies with unique features and benefits, JEDEC has defined three main categories of standards for DDR: standard DDR, mobile DDR, and graphics DDR. Standard DDR targets servers, data centers, networking, laptop, desktop, and consumer applications, allowing wider channel-widths, higher densities, and different form-factors. Mobile DDR, or LPDDR, targets the mobile and automotive applications, which is very sensitive to area and power, offering narrower channel-widths and several low-power DRAM states. Graphics DDR targets data-intensive applications requiring very high throughput, such as graphics-related applications, data center acceleration, AI, etc. JEDEC has defined GDDR and HBM as the two standards for graphics DDR.

Synopsys offers complete solutions with silicon-proven PHYs and controllers, supporting the latest DDR, LPDDR, and HBM standards. Synopsys is an active member of the JEDEC work groups, driving the development and adoption of the standards. Synopsys' configurable memory interface IP solutions can be tailored to meet the exact requirements of SoCs for applications including AI, automotive, mobile, and cloud computing.

SYNOPSYS®
Silicon to Software™

02/25/19.CS313600928-ddr-white-paper.

Pub: Sept. 2019