cadence®

Five Emerging DRAM Interfaces You Should Know for Your Next Design

By Gopal Raghavan, Cadence Design Systems

Producing DRAM chips in commodity volumes and prices to meet the demands of the mobile market is no easy feat, and demands for increased bandwidth, low power consumption, and small footprint don't help. This paper reviews and compares five next-generation DRAM technologies—LPDDR3, LPDDR4, Wide I/O 2, HBM, and HMC—that address these challenges.

Contents

Introduction

Because dynamic random-access memory (DRAM) has become a commodity product, suppliers are challenged to continue producing these chips in increasingly high volumes while meeting extreme price sensitivities. It's no easy feat, considering the ongoing demands for increased bandwidth, low power consumption, and small footprint from a variety of applications. This paper takes a look at five next-generation DRAM technologies that address these challenges.

Mobile Ramping Up DRAM Demands

Notebook and desktop PCs continue to be significant consumers of DRAM; however, the sheer volume of smartphones and tablets is driving rapid DRAM innovation for mobile platforms. The combined pressures of the wired and wireless world have led to development of new memory standards optimized for the differing applications. For example, rendering the graphics in a typical smartphone calls for a desired bandwidth of 15GB/s—a rate that a two-die Low-Power Double Data Rate 4 (LPDDR4) X32 memory subsystem meets efficiently. At the other end of the spectrum, a next-generation networking router can require as much bandwidth as 300GB/s—a rate for which a two-die Hybrid Memory Cube (HMC) subsystem is best suited.

LPDDR4 and HMC are just two of the industry's emerging memory technologies. Also available (or scheduled for mass production in the next couple of years) are LPDDR3, Wide I/O 2, and High Bandwidth Memory (HBM). But why deal with all of these different technologies? Why not just increase the speed of the DRAM you are already using as your application requirements change?

Unfortunately, core DRAM access speed has remained pretty much unchanged over the last 20 years and is limited by the RC time constant of a row line. For many applications, core throughput (defined as row size * core frequency) is adequate and the problem is then reduced to a tradeoff between the number of output bits versus output frequency (LPDDR3, LPDDR4, Wide I/O 2, and

HBM are among the memory subsystems that address these concerns). However, if an application requires more bandwidth than the core can provide, then multiple cores must be used to increase throughput (HMC subsystems can be used in these scenarios).

Increasing DRAM bandwidth is not an effort without tradeoffs. While bandwidth is primarily limited by I/O speed, increasing I/O speed by more bits in parallel or higher speeds comes with a power, cost, and area penalty. Power, of course, remains an increasing concern, especially for mobile devices, where the user impact is great when battery life is short and/or the devices literally become too hot to handle. Additionally, increasing package ball count results in increased cost and board area.

The emerging DRAM technologies represent different approaches to address the bandwidth, power, and area challenges. In this paper, we'll take a closer look at the advantages and disadvantages of five memory technologies that are sure to play integral roles in next-generation designs.

LPDDR3: Addressing the Mobile Market

Published by the JEDEC standards organization in May 2012, the LPDDR3 standard (Figure 1) was designed to meet the performance and memory density requirements of mobile devices, including those running on 4G networks. Compared to its predecessor, LPDDR3 provides a higher data rate (1,600Mb/s), more bandwidth (12.8GB/s), higher memory densities, and lower power.[1]

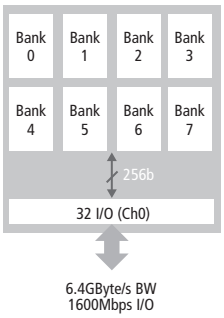
	LPDDR3 and LPDDR3E
Die Organization	1ch X 8 banks X 32 I/O 
Channel #	1
Bank #	8
Density	4Gb - 32Gb
Page Size	4KByte

Figure 1. LPDDR3 architecture

To achieve the goal of higher performance at lower power, three key changes were introduced in LPDDR3: lower I/O capacitance, on-die termination (ODT), and new interface training modes. Interface training modes include write-leveling and command/address training. These features help improve timing queues and timing closure, and also ensure reliable communication between the device and the system on chip (SoC). The mobile memory standard also features lower I/O capacitance, which helps meet the increased bandwidth requirement with increased operating frequency at lower power.[2]

LPDDR4: Optimized for Next-Generation Mobile Devices

LPDDR4 (Figure 2) is the latest standard from JEDEC, expected to be in mass production in 2014. The standard is optimized to meet increased DRAM bandwidth requirements for advanced mobile devices. LPDDR4 offers twice the bandwidth of LPDDR3 at similar power and cost points. To maintain power neutrality, a low-swing GND terminated interface (LVSTL) with data bus inversion has been proposed. Lower page size and multiple channels are other innovations used to limit power. For cost reduction, the standard LPDDRx core architecture and packaging technologies have been reused with selected changes such as a reduction of the command/address bus pin count.[3]

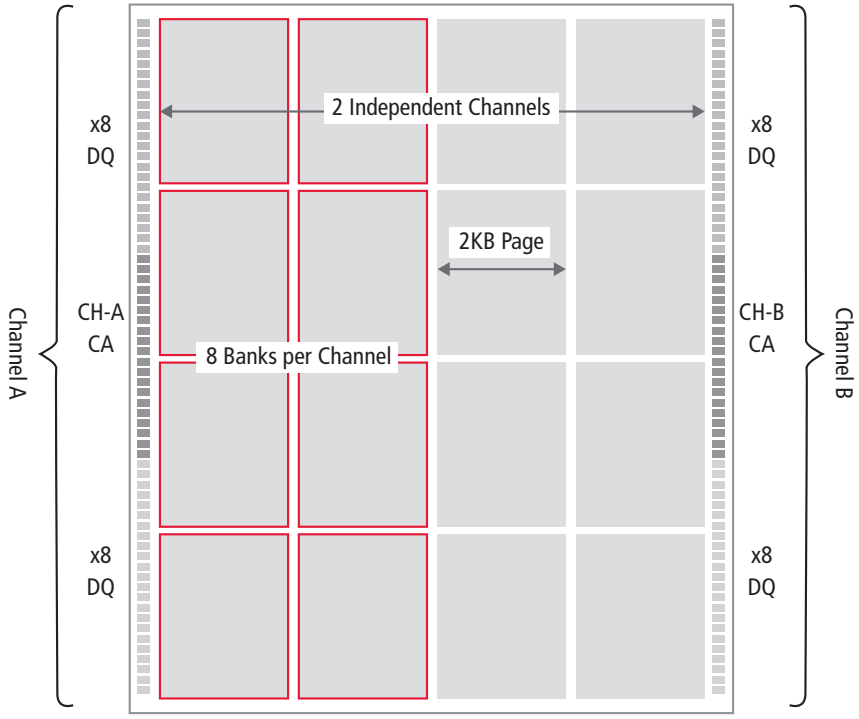


Figure 2. LPDDR4 architecture

Wide I/O 2: Supporting 3D-IC Packaging for PC and Server Applications

The Wide I/O 2 standard (Figure 3), also from JEDEC and expected to reach mass production in 2015, covers high-bandwidth 2.5D silicon interposer and 3D stacked die packaging for memory devices. Wide I/O 2 is designed for high-end mobile applications that require high bandwidth at the lowest possible power. This standard uses a significantly larger I/O pin count at a lower frequency. However, stacking reduces interconnect length and capacitance and eliminates the need for ODT. This results in the lowest I/O power for higher bandwidth.

	Wide I/O 2
Die Organization	4ch X 8 banks X 64 I/O Bank 0 Bank 4 Bank 1 Bank 5 Bank 2 Bank 6 Bank 3 Bank 7 — 256b — 64 I/O (ch0) — 6.4GByte/s BW 800Mbps I/O Bank 0 Bank 4 Bank 1 Bank 5 Bank 2 Bank 6 Bank 3 Bank 7 — 256b — 64 I/O (ch1) — 6.4GByte/s BW 800Mbps I/O Bank 0 Bank 4 Bank 1 Bank 5 Bank 2 Bank 6 Bank 3 Bank 7 — 256b — 64 I/O (ch2) — 6.4GByte/s BW 800Mbps I/O Bank 0 Bank 4 Bank 1 Bank 5 Bank 2 Bank 6 Bank 3 Bank 7 — 256b — 64 I/O (ch3) — 6.4GByte/s BW 800Mbps I/O
Channel #	4 and 8
Bank #	32 per die
Density	8Gb - 32Gb
Page Size	4KByte (4ch die), 2KByte (8ch die)

Figure 3. Wide I/O 2 architecture

In 2.5D stacking, two dies are flipped over and placed on top of an interposer. All of the wiring is on the interposer, making the approach less costly than 3D stacking but requiring more area. Heat dissipation is not much of a concern, since cooling mechanisms can be placed on top of the two dies. This approach is also lower cost and more flexible than 3D stacking because faulty connections can be reworked.

There are electronic design automation (EDA) tools on the market that help designers take advantage of redundancy at the logic level to minimize device failures. For example, Cadence® Encounter® Digital Implementation allows designers to route multiple redistribution (RDL) layers into a microbump, or to use combination bumps. In this scenario, if one bump falls, the remaining bumps can carry on normal operations.

With 3D stacking, heat dissipation can be an issue—there isn't yet an easy way to cool the die in the middle of the stack, and that die can heat up the top and bottom dies. Poor thermal designs can limit the data rate of the IC. In addition, a connection problem—especially one occurring at the middle die—renders the entire stack useless.

HMC: Breaking Barriers to Reach 400G

HMC (Figure 4) is being developed by the Hybrid Memory Cube Consortium and backed by several major technology companies, including Samsung, Micron, ARM, HP, Microsoft, Altera, and Xilinx. HMC is a 3D stack that places DRAMs on top of logic. This architecture, expected to be in mass production in 2014, essentially combines high-speed logic process technology with a stack of through-silicon-via (TSV) bonded memory die.[4] In an example configuration, each DRAM die is divided into 16 "cores" and then stacked. The logic base is at the bottom, with 16 different logic segments, each segment controlling the four or eight DRAMs that sit on top. This type of memory architecture supports more "DRAM I/O pins" and, therefore, more bandwidth (as high as 400G). According to the Hybrid Memory Cube Consortium, a single HMC can deliver more than 15X the performance of a DDR3 module and consume 70% less energy per bit than DDR3.

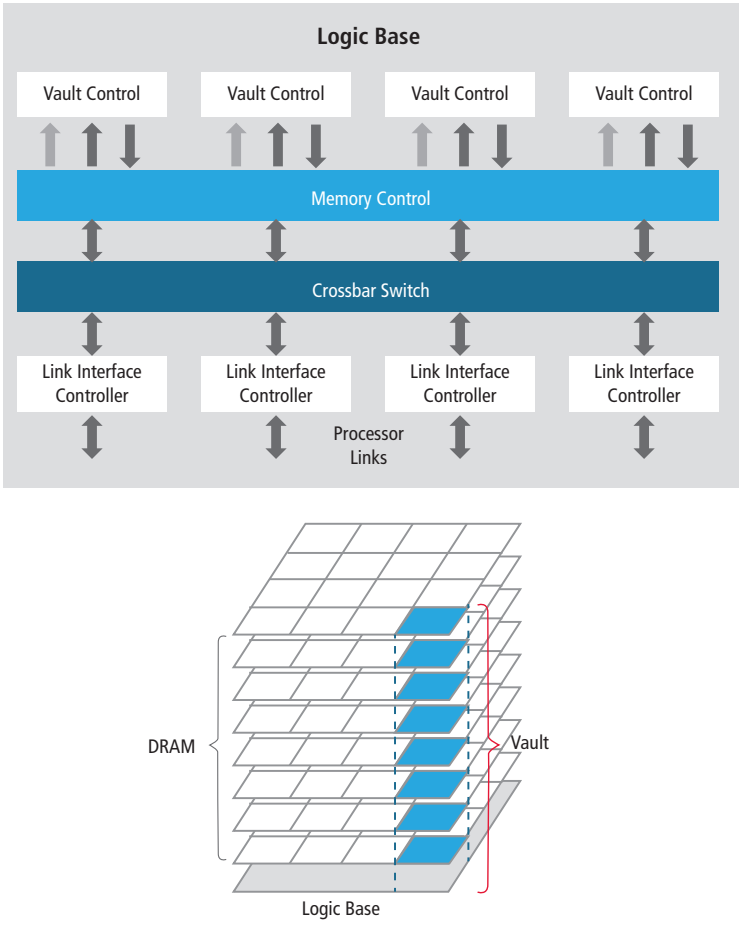


Figure 4. HMC architecture

HMC uses a packetized protocol on a low-power SerDes interconnect for I/O. Each cube can support up to four links with up to 16 lanes. With HMC, design engineers will encounter some challenges in serialized packet responses. When commands are issued, the memory cube may not process these commands in the order requested. Instead, the cube reorders commands to maximize DRAM performance. Host memory controllers thus need to account for command reordering. HMC provides the highest bandwidth of all the technologies considered in this paper, but this performance does come at a higher price than other memory technologies.

HBM: Emerging Standard for Graphics

HBM (Figure 5) is another emerging memory standard defined by the JEDEC organization. HBM was developed as a revolutionary upgrade for graphics applications. GDDR5 was defined to support 28GB/sec (7Gbps x32). Extending the GDDRx architecture to achieve a higher throughput while improving performance/watt was thought to be unlikely. Expected to be in mass production in 2015, the HBM standard applies to stacked DRAM die, and is built using TSV technologies to support bandwidth from 128GB/s to 256GB/s. JEDEC's HBM task force is now part of the JC-42.3 Subcommittee, which continues to work to define support for up to 8-high TSV stacks of memory on a 1,024-bit wide data interface.[5] According to JEDEC, the interface would be partitioned into eight independently addressable channels supporting a 32-byte-minimum access granularity per channel. There is no command reordering, which allows the graphics controller to optimize access to memory. The subcommittee expects to publish the standard in late 2013.

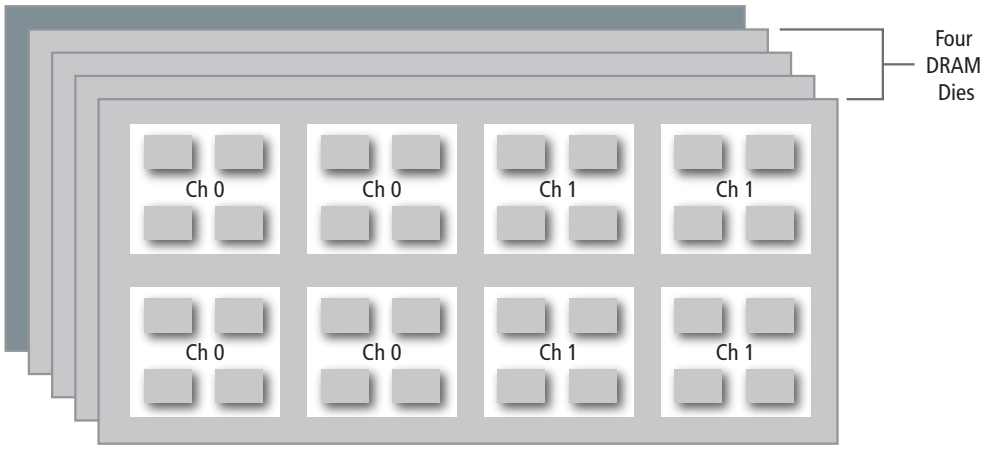


Figure 5. HBM architecture

Which Memory Standard Is Best for Your Next Design?

As this paper has discussed, each emerging memory standard tackles the power, performance, and area challenges in a different way. There are tradeoffs from one to another, with each optimized for a particular application or purpose. How do you select the right memory standard for your design?

Obviously, the basis for your decision will be your application's requirements. When considering a smartphone, for example, you may decide between Wide I/O 2 and LPDDR4. Because thermal characteristics are critical in smartphones, the industry consensus has turned to Wide I/O 2 as the best choice. Wide I/O 2 meets heat dissipation, power, bandwidth, and area requirements. However, it is more costly than LPDDR4. LPDDR4, on the other hand, also provides advantages in bandwidth, TSV readiness, and software support. Given its lower silicon cost, LPDDR4 may be more ideal for cost-sensitive mobile markets.

On the other end of the application spectrum, consider high-end computer graphics processing, where chip complexity is a given and high-resolution results are expected. Here, you might look to the higher bandwidth HBM technology. Computer graphics applications are less constrained by cost than, say mobile devices, so the higher expense of HBM memory may be less of an issue. Table 1 compares the features of the five standards discussed here.

Memory Standard	Mass Production Year	Bandwidth (GB/s)	Package Density (GB)	Power Efficiency (mW/GB/s)	Approximate Relative Cost Per Bit	Cadence Controller and PHY IP
LPDDR3	2012	17	2-4	67	1	Yes
LPDDR4	2014	25.6	4-8	~50	1.1	Yes
Wide I/O 2	2015	51.2	4-8	–	3	2014
HMC	2014	160	2-4	–	2	2014
HBM	2014	128	2-8	–	2	2015

Table 1: Where does each memory standard stand?

To help integrate your design at the register-transfer level, EDA companies like Cadence offer IP portfolios for memory subsystems. Cadence has controller and PHY IP for a broad array of standards, including many of those discussed in this paper. Cadence also provides memory model verification IP (VIP) to verify memory interfaces and ensure design correctness. Additionally, tools such as Cadence Interconnect Workbench allow the SoC designer to optimize the performance of memory subsytems through a choice of memory controller parameters such as interleaving, command queue depths, and number of ports. These tools can help speed up SoC development and ensure first-pass success.

Summary

Choosing the right DRAM technology for your design requires careful consideration. There are a variety of architectures that are either available now or will soon hit the market. Each has its strengths and weaknesses in terms of meeting bandwidth, power, cost, and other key specifications; your specific application and market requirements should guide you in making the right choice for your next design.

Footnotes

[1] Source: "Mobile DDR", Wikipedia: http://en.wikipedia.org/wiki/Mobile_DDR

[2] Source: Kristin Lewotsky, "Industry view: JEDEC on LPDDR3," EE Times; http://www.eetimes.com/electronics-blogs/other/4373861/Industry-view--JEDEC-on-LPDDR3

[3] Source: Daniel Skinner, "LPDDR4 Moves Mobile", JEDEC Mobile Forum 2013: http://www.jedec.org/sites/default/files/D_Skinner_Mobile_Forum_May_2013_0.pdf

[4] Source: About Hybrid Memory Cube, Hybrid Memory Cube Consortium: http://hybridmemorycube.org/technology.html

[5] Source: 3D-ICs, JEDEC: http://www.jedec.org/category/technology-focus-area/3d-ics-0