WP456 (v1.1) March 23, 2015

The Rise of Serial Memory and the Future of DDR

By: Tamara Schmitz

DDR memory is giving way to serial memory technologies like Hybrid Memory Cube (HMC) and other schemes still in the pipeline. The UltraScale™ Architecture Portfolio supports them all!

ABSTRACT

With no plans emerging to define a "DDR5" specification, it is certainly clear that the entire memory landscape is going to change profoundly over the coming years. The current DDR4 generation (including the mobile-focused LPDDR4 low-power model) continues to ramp up slowly, but it is essentially the last of its kind, the ultimate generation of memory technology based on the DDR model. LPDDR is being well-accepted in the wireless devices market, but with wireless technology evolving through a generation (or sometimes two) every year, the LPDDR model simply will not be able to keep up.

As they become competitively viable, multiple memory technologies are already vying for dominance of this chunk of the memory market space. The current DDR market is expected to be split, and eventually taken over, by new serial memory technologies, like Hybrid Memory Cube (HMC) and other schemes still in the development pipeline. Programmable logic devices like FPGAs and SOCs must maintain support for these new memory schemes as they emerge into the marketplace.

This white paper explains why the DDR model is now approaching the end of its useful lifetime, and how serial memory schemes can be expected to fill the memory needs of the future. Xilinx® UltraScale architecture-based devices and platforms have been engineered from the outset with the future in mind, providing a seamless transition to the newly emerging serial memory technologies.

Introduction

This white paper explains the limitations of parallel memory interfacing (e.g., DDR) and how serial memory schemes can be expected to fill some of the gaps between the current DDR4 generation and the multiple memory technologies on the horizon. Logic designers can rest assured that UltraScale™ architecture-based devices (i.e., FPGAs and MPSoCs) and platforms have been engineered from the outset with the future in mind, providing a seamless transition from commodity to emerging memory technologies (see Figure 1).

The universally popular DDR family of memories, a workhorse buffer used by 90% of Xilinx customers (see Figure 1) continues with the DDR4 generation. Currently, DDR3 is a comfortable choice for the majority of system boards, and DDR4, though still ramping up slowly, is expected to replace many of those sockets and serve them for years to come. Still, with the knowledge that DDR4 has no natural successor, there are some opportunities—accompanied by trade-offs, such as bandwidth, capacity, or power reductions—that designers should be aware of. One of the likely successors is the low-power LPDDR3/4 family of wireless-oriented memories; certain other application spaces are likely to prefer serial DRAM solutions such as Hybrid Memory Cube (HMC).

After a description of the market trends, this white paper focuses the limitations of parallel memory technology. The topic of DDR alternatives leads naturally from the wireless-focused LPDDR3/4 memories to the new serial memory models. Since serial memory is a new topic for many designers, it is described here more broadly, encouraging the reader to become a more active and informed customer as these new memory technologies are brought to market.

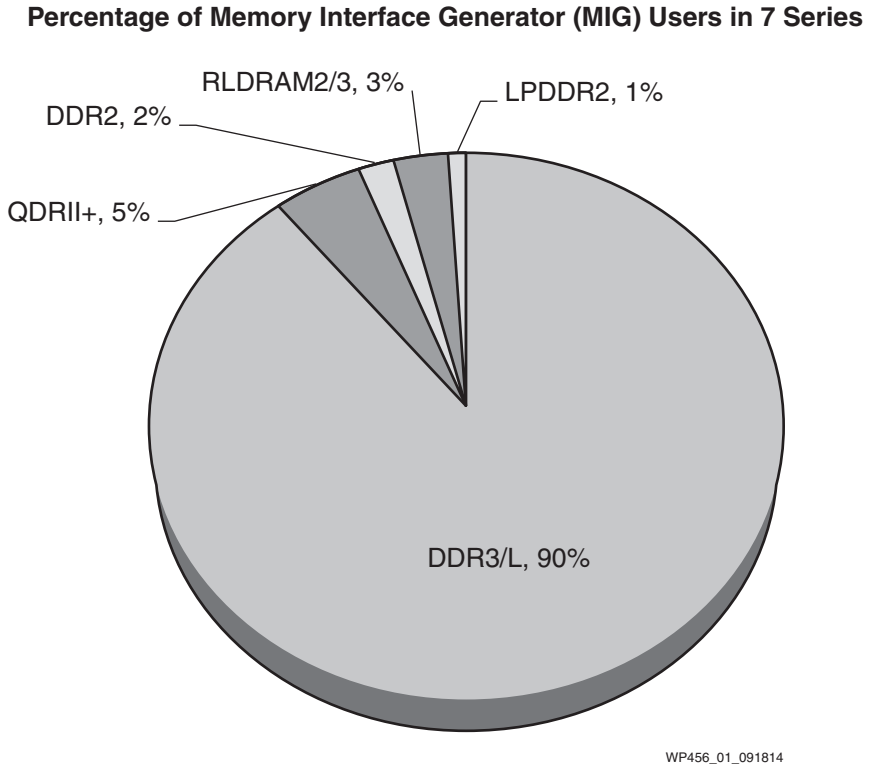

Figure 1: **Xilinx Customer Memory Usage, as Measured through the MIG GUI in 2013**

Market Trends

Usually, when customers are designing for their next generation, they look to the next generation of the same memory to give more capacity, speed, and throughput. The current and projected DRAM market share trends are depicted in Figure 2. DDR3 enjoys almost 70% of the total DRAM market. Its rise to domination was assured in the steep 40% rise in adoption between 2009 and 2010. DDR4 has been slower in adoption; part of the hindrance to DDR4 growth can be attributed to the slow but steady growth in Mobile DRAM, also known as LPDDR. DDR4 simply does not have as many sockets to claim if LPDDR is meeting the needs of the wireless market.

Looking at Figure 2, DDR4 growth is indeed increasing, because it does have advantages: lower supply voltage (which saves power) and higher speed. As a result, it will eventually supplant DDR3 in most markets, driven by the computing space. Despite the fact that the computing space no longer drives more than 70% of DRAM consumption, it has retained the position as the largest commodity device segment—until this year when the communications segment outsized it (Source: IC Insights). For now, according to memory vendors, DDR4 usage is localized more to the server market than to personal electronics. Still, DDR4 is an excellent choice. It is a well-known memory type and will be available for a very long time — in particular because there is no successor.

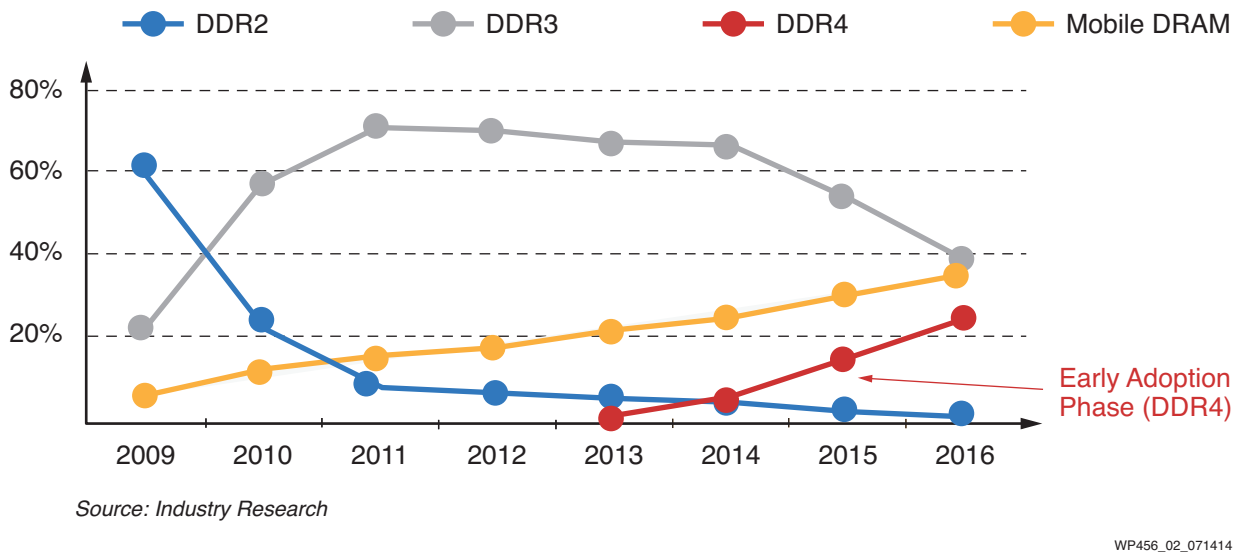

Figure 2: **Market Trends for DRAM Memory**

The Challenges with Parallel Memory Technology

When customers want a new product, they want it to have more memory. The consumer electronics industries have an insatiable demand for memory bandwidth. MP3 players need to hold 10,000 songs, rather than the couple dozen songs a cassette tape held. This applies just as well to the number of pictures or videos smartphones are expected to store. These expectations typically mean more components and more board space. Of course, consumers do not want the physical dimensions of their electronic devices to grow in proportion to their capacity or performance. There seems to be an expectation that as technology advances and devices can hold "more stuff," "more stuff" should fit in the same space as before — or maybe even in less!

Where FPGAs are concerned, specific guidelines exist about how to lay out the board—trace lengths, termination resistors, and routing layers—to ensure proper margins and overall system success. These rules limit how much the design can be compacted, or how close the parts can be placed together.

One way of getting to a smallest-possible board design is to demand some sort of ultra-compact, very bleeding-edge chip packaging scheme. Unfortunately, any really new packaging technology has to include such exotic implementations as die-stacking with through-silicon vias (TSVs). And leveraging any edge-of-the-art packaging on components that have historically been manufactured as commodities translates into *significantly* greater cost. Based simply on the economies of scale of the memory industry infrastructure, DDR memory is *not* seen as a justifiably high-cost system block. Adopting a radical design departure or absorbing an increased price point is simply not an option. Therefore, "improvements" of this sort are unlikely to be seen in DDR3/DDR4 systems any time soon.

Consumers also want more speed, and running a system at higher speeds has implications for board design. DDR is a memory with a single-ended signal that needs proper termination. The faster the system is operated, the shorter the traces between memory and FPGA must be to ensure proper functionality. This means that the devices themselves need to be placed in closer proximity to the FPGA. Of course, limiting memory distance from the FPGA limits the number of memory devices that can be used in a design. Many DDR4 designs are likely to reach the limits of crowding close to the FPGA as many memory devices as possible.

To sum up, then: If the designer wants more memory, more devices are needed. If the designer wants to go faster, components must be brought closer together. There is a limit to how many memory devices can be placed in a fixed amount of board space. Any speed improvements in a "DDR5" memory would reduce the area available for the memory devices themselves, thus reducing available storage capacity. In other words, DDR memory technology has "hit the wall": it has reached the intrinsic limits of its spacial and technical properties, and can go no further.

Successor to DDR3

DDR4 is not expected to completely replace DDR3. Trends show that the server market is adopting DDR4, while DDR3, with its lower cost, continues for now to be the predominate choice in the personal computing segment. There is no doubt that consumer appetite for more speed and capacity will continue to grow, and eventually PCs will migrate to DDR4 over time.

In the absence of DDR5, one likely choice to replace DDR3 and DDR4 is LPDDR4. The LP stands for *low power*. LPDDR4 is a type of DDR memory that has been optimized for the wireless market. LPDDR's advantages include its widespread adoption and availability and its well defined and stable specifications. The low-power optimization makes it only a little more expensive than DDR, and it still uses the I/O pins that DDR uses. Ease of migration is also a positive factor, because it runs in the same frequency range as DDR.

However, the biggest trade-off is its lifetime. Since the wireless market turns over its products approximately every 12 months, LPDDR memories change at a similarly rapid pace. If a big company sells products for 10–15 years, it is difficult to accommodate a memory device family that changes every 12 months. Possibly a manufacturer could guarantee to deliver one version of those devices for that company for 10–15 years under a special agreement. Such arrangements would

need to include preserving a process flow, an expensive endeavor that might be worthwhile only for the largest of opportunities.

If Not LPDDR, Then What?

Besides LPDDR, many other memory options are vying for the opportunity to be the next memory of choice. Serial memory is emerging as a viable alternative. In an FPGA, memory is the last functional block to go serial. The reason for that is *latency*. The time it takes to convert data from parallel to a serial stream, send it down the serial link, and then reverse the process by converting it back into parallel data bytes, has always taken much too long to be practicable. With increasing silicon speeds, however, the conversion trade-off is tolerable in applications where there are multiple writes and few reads, such as a test and measurement system for a CT scan or a set of telescopes scanning the sky. Alternatively, if the measurement of quality is to write data and immediately read that same data, then serial memory does not perform as well as parallel data in any form. However, if the measure of good memory is high bandwidth, storing lots of videos, or sending loads of information over the Internet, then considering a serial memory implementation can be tempting.

Latency aside, the same trade-offs deserve investigation. Lifespan is not a problem; these products will be made as long as there is an appetite for them, in comparison to the shorter availability of LPDDR. In fact, if the desire for serial memory grows, multiple vendors are likely to join in the business of making serial memory. A comparison of LPDDR vs. serial memory trade-offs is shown in Table 1.

Table 1: **Advantages/Drawbacks of LPDDR4 and HMC Serial as DDR Replacements**

Low-Power LPDDR4		HMC Serial	
Advantages:	**Drawbacks:**	**Advantages:**	**Drawbacks:**
Well-known, popular in wireless	Short lifetime (~12 mo.)	Longer lifetime	New and unfamiliar
Only slightly more expensive than DDR4	Consumes I/O pins	Leverages serial tech	Expensive
—	3,200 Mb/s max	15 Gb/s max	—

Instead of using I/O pins, serial memory leverages high-speed transceiver technology. In FPGAs, it is possible to use serial interfacing (transceivers) to run at high data rates. This well-developed serial technology can currently support a very high 15 Gb/s throughput. The next generation (in the case of HMC) is planned to reach 30 Gb/s. While high-speed transceivers are a well known technology, serial memory is very new, causing apprehension for some designers. In addition, with "newness" comes the inevitability of limited production runs at first, and of course, higher initial prices.

Hybrid Memory Cube (HMC)

The strongest serial memory candidate to replace DDR DRAM memory is HMC, being promoted by the HMC Consortium and spearheaded by Micron Technology, Inc. (see Figure 3). HMC has been well advertised. "HMC" for some designers has even become conversational shorthand for "serial memory." In fact, though, HMC is just *one type* of serial memory.

In addition to HMC:

- MoSys is developing its **Bandwidth Engine** technology, basically a kind of "serial SRAM"
- Broadcom is offering a range of **serial interface TCAMs**

At the other end of the future spectrum:

- Samsung is promoting **High Bandwidth Memory (HBM)**, a TSV-based DRAM stack with a massively wide parallel interface. It might seem lower risk since it uses a parallel interface. Read Other Memory Solutions for more details.

HMC, however, is still the leading contender to take market share from DDR3 and DDR4. It has four or eight stacks of DRAM connected together using through-silicon via (TSV) technology on top of a logic layer to create a 2G or 4G package. The logic layer creates a convenient interface. Up to eight devices can be daisy-chained if more capacity is needed. There is 256-bit access and enormous throughput considering the 1-to-4 link capability (in steps of half a link). Each link comprises sixteen transceivers (eight for a half-link)—all capable of handling 15 Gb/s, an extraordinary amount of bandwidth previously unavailable to memory designers. HMC claims to support chaining up to eight devices, but this has yet to be proven, and it will surely come with its own set of challenges.

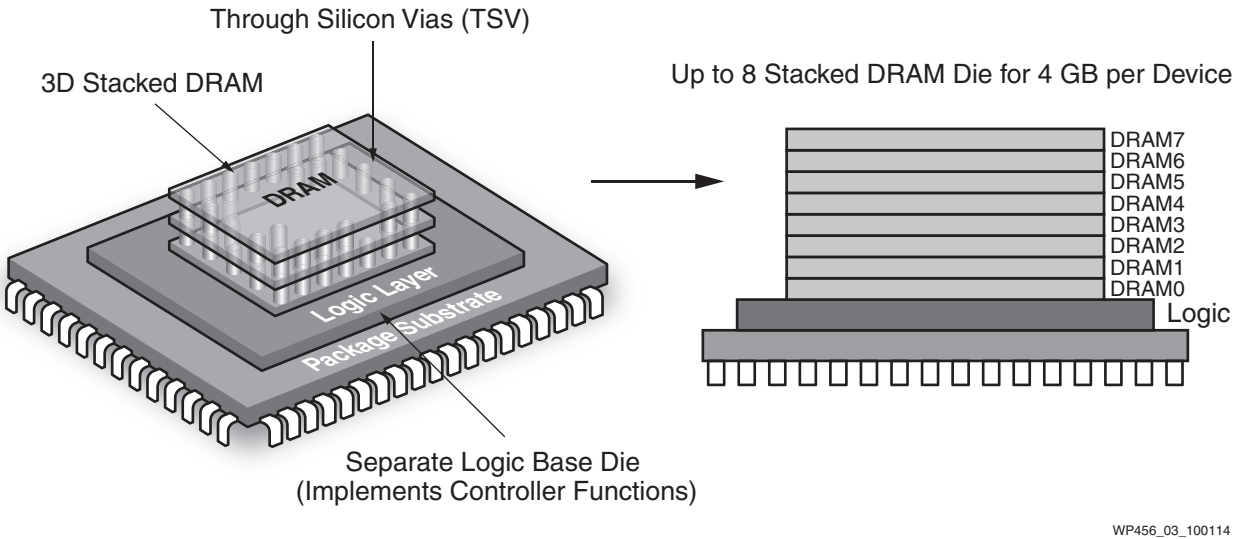

Figure 3: **Illustration of the Structure of HMC**

To see the improvement in bandwidth over the DDR solution, see Table 2, where three designs are presented. Each of the three designs (DDR3, DDR4, and HMC) is sized to support 60 Gb/s. Pin count is reduced in the HMC solution by at least a factor of 8, greatly reducing board complexity and routing (see Figure 4). The high bandwidth of the SerDes links allows fewer devices to be used—just one in the case noted. This single device and FPGA reduce board space needed by a

factor of nearly 20. Finally, the HMC solution consumes one-third the power per bit. These are eye-opening numbers that compel observers to envision HMC capturing a portion of the market previously earmarked for DDR4.

Table 2: **Comparison of Resources Needed to Support 60 Gb/s (Source: Micron)**

	DDR3	DDR4	HMC
Pin Counts (power and ground not included)	715	592	70
Board Area	8,250 mm^2	6,600 mm^2	378 mm^2
Power	64 pJ/bit	51 pJ/bit	16 pJ/bit
Bandwidth	18 MB/pin	29 MB/pin	857 MB/pin

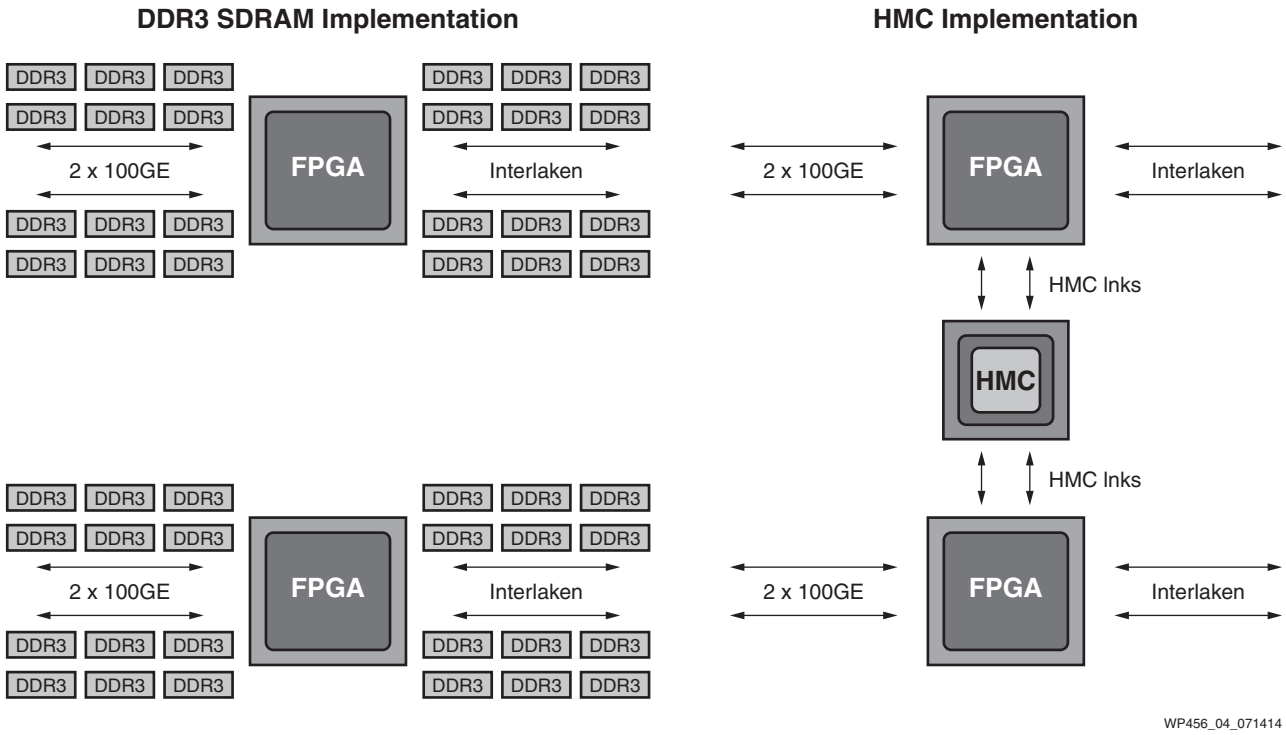

Figure 4: **Mock-up of a 2x100GE Design: Left, DDR3; Right, HMC**

Other Memory Solutions

Because "HMC" and "serial memory" are often used erroneously as interchangeable terms—and sometimes even to represent *any* new high-bandwidth memory—it is useful to explore some of these new non-HMC memories. The three top memories in this category are Bandwidth Engine, by MoSys; TCAM, by Broadcom; and HBM, promoted by Samsung, SK Hynix, and Intel.

- **Bandwidth Engine (BE2)** by MoSys works like a serial SRAM, using transceivers to achieve 16 Gb/s, but it is *not* a serial DRAM. It is not a likely replacement for DDR. Instead, with its 72-bit access and lower latency, BE2 targets QDR or RLDRAM. The most likely application is storage for packet headers, or a look-up table instead of a packet buffer, as in the case of DDR.

- **Ternary Content Addressable Memory (TCAM)** is a special high-speed memory that performs broad pattern-matching searches used in high-performance routers and switches. The high performance is paid for with expense, power, and heat. TCAM *is* high-speed, but its technology is *parallel* in nature; it does not use high-speed transceiver technology to reach high speeds. However, Broadcom is offering serial versions of this memory. In this way, the advantages found in serial memories (low pin-count and high speed) can still be associated with a TCAM solution.
- **HBM** is the third memory type. There is some "debate" about the relative features of HMC and HBM. However, there really is no such thing as an "HBM device." If a user wished to adopt HBM technology, a *die* would be obtained from a memory vendor; that die would then be mounted inside the user's package on an interposer. Connections between the user device and the memory would need to be included in the interposer design to enable this high-bandwidth parallel memory.

  On the business side of the debate, certain limitations exist at present that mitigate against HBM taking over the DDR market any time soon. Companies would need to decide what they would be willing to share in terms of trade secrets, as well as standards adoption agreements—for example, interposer design, heights, interfaces, tolerances, and so forth. Those details could, of course, all be worked out — but so far, they have not been. On the engineering design side, of course, HBM latency would be *extremely low* because the die resides *inside the package* with the user device—certainly a very attractive characteristic of this technology. Technically, therefore, HBM is a fantastic idea, but its unresolved implementation details put its widespread adoption well out into the future.

The sources of these solutions have been discussed. If successful, more suppliers are likely to join these pioneers to serve the industry. MoSys's Bandwidth Engine (BE2) is in production now. HMC is sampling now, and is planned to be in full production by Q3 of 2015. LPDDR4 should be sampling in 2015. HBM is not available as a stand-alone package, though there are talks about the possibility of serializing HBM into its own package. Prospective users can buy a die from Samsung, Hynix, or some smaller vendor and integrate it into their own package — which some customers are doing right now.

No matter how tempting LPDDR is for lower power applications, or serial memory is for enormous bandwidth, neither can truly replace DDR4 when density is considered. Technically, density refers to the capacity of a single device in the context of its size — but in more real-world terms, density refers primarily to the manufactured DIMM modules that provide ready-made groups of eight or sixteen DDR4 devices.

Conclusion

The take-away is that DDR3 is here and is very strong, while DDR4 is still in its growth and adoption phase. DDR4 is likely to experience its own staying power that might stretch even longer than that of the popular DDR3, simply because it has no successor. LPDDR4 is the most likely candidate to fill the gap, but will not replace DDR4 in all areas unless there are very rapid read/write iterations. Otherwise, serial memory is the newcomer to watch. HMC is poised to replace DDR, while Bandwidth Engine is the serial solution replacing QDR and RLDRAM.

Xilinx FPGAs support all of these memory types. The UltraScale architecture has the right mix of I/Os and high-speed transceivers to support any choice of parallel or serial memories that a customer might need. Now memory can travel at speeds previously unheard of, and make trade-offs of density, board space, latency, and availability. The high quality and design of the Xilinx UltraScale architecture-based devices allow the memory to set the limitations in the system when proper guidelines for PCB layout are followed.

For key documentation, hardware demonstrations, and other resources, go to: www.xilinx.com/memory.

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
03/23/2015	1.1	Updated with UltraScale architecture-based devices (FPGAs and MPSoCs) throughout document.
10/27/2014	1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer