WP485 (v1.1) July 15, 2019

Virtex UltraScale+ HBM FPGA: A Revolutionary Increase in Memory Performance

By: Mike Wissolik, Darren Zacher, Anthony Torza, and Brandon Day

For system architects and FPGA designers of data center, wired, and other bandwidth intensive applications that need more performance than traditional DRAM technology, HBM is a memory that offers benefits in performance, power, and footprint —unlike available memory interfaces on the market.

ABSTRACT

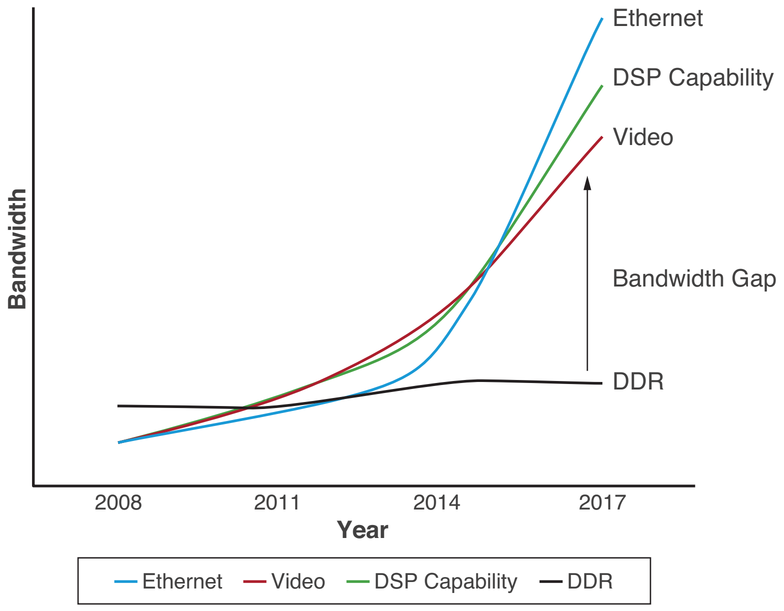
Electronic systems have experienced an exponential growth in compute bandwidth over the last decade. This massive increase in computational bandwidth brings with it a massive increase in memory bandwidth requirements to meet those computational needs. Designers of these systems often find that commodity parallel memories such as DDR4 no longer meet the bandwidth needs of their application.

Xilinx's high bandwidth memory (HBM) -enabled FPGAs are the clear solution to these challenges, providing massive memory bandwidth within the lowest power, footprint, and system cost envelopes. In designing these FPGAs, Xilinx has joined other leading semiconductor vendors in choosing the industry's only proven stacked silicon interconnect implementation—TSMC's CoWoS assembly process. This white paper explains how Xilinx® Virtex® UltraScale+™ HBM devices are addressing the demand for dramatically increased system memory bandwidth while keeping power, footprint, and cost in check.

Industry Trends: Bandwidth and Power

Over the last 10 years, the bandwidth capabilities of parallel memory interfaces have improved very slowly—the maximum supported DDR4 data rate in today's FPGAs is still less than 2X what DDR3 could provide in 2008. But during that same time period, demand for memory bandwidth has far outpaced what DDR4 can provide. Consider the trend in Ethernet: since the days of DDR3, Ethernet port speeds have increased from 10Gb/s to 40Gb/s, then to 100Gb/s, and now to 400Gb/s—a more than 10X increase in raw bandwidth.

Similar trends exist in high performance compute and video broadcast markets. FPGA machine-learning DSP capability has increased from 2,000 DSP slices in the largest Virtex-6 FPGA, to over 12,000 DSP slices in the largest Virtex UltraScale+ device today. For video broadcast, the industry has moved from standard definition to 2K, 4K, and now 8K. In each of these application areas, a significant gap exists between the bandwidth required by the application vs. that which can be provided by a DDR4 DIMM. See Figure 1.



WP485_01_051217

Figure 1: **Relative Memory Bandwidth Requirements**

To compensate for this bandwidth gap, system architects attempting to use DDR4 for these applications are forced to increase the number of DDR4 components in their system—not for capacity, but to provide the required transfer bandwidth between the FPGA and memory. The maximum achievable bandwidth from four DDR4 DIMMs running at a 2,667Mb/s data rate is 85.2GB/s. If the bandwidth required by the application exceeds this, then a DDR approach becomes prohibitive due to power consumption, PCB form factor, and cost challenges. In these bandwidth-critical applications, a new approach to DRAM storage is required.

Revisiting that same 10-year time period from a power efficiency perspective, it is clear that the era of “hyper-performance” at any cost is over. According to one MDPI-published paper, it was forecast that by 2030, data centers alone could be responsible for consuming anywhere from 3% up to potentially 13% of the global energy supply[Ref 1], depending upon the actual power efficiency of data center equipment in deployment at the time. Designers greatly value power efficient performance, especially in this era of multi-megawatt data centers. They also value efficient thermal solutions, because OPEX requirements for reliable airflow and cooling are significant—one-third of the total energy consumption[Ref 2]. Therefore, the vendor who can provide the best possible compute per dollar and compute per watt in a practical thermal envelope has the most attractive solution.

Alternatives to the DDR4 DIMM

To bridge the bandwidth gap, the semiconductor industry has introduced several clever alternatives to commodity DDR4. See Table 1. More recently, the industry has seen the rise of transceiver-based serial memory technologies such as hybrid memory cube (HMC). These technologies offer higher memory bandwidth, and as such, can provide the memory bandwidth of several DDR4 DIMMs in a single chip—although at the expense of allocating up to 64 ultra-high-speed serial transceivers to the memory subsystem.

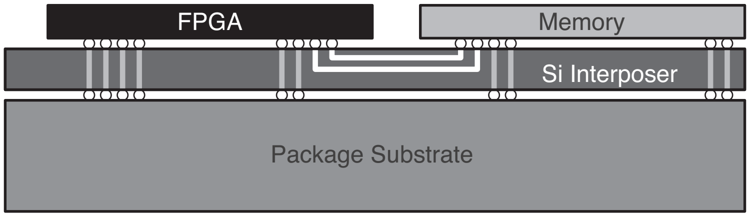
Table 1: **Comparison of Key Features for Different Memory Solutions**

	DDR4 DIMM	RLDRAM-3	HMC	HBM
Description	Standard commodity memory used in servers and PCs	Low latency DRAM for packet buffering applications	Hybrid memory cube serial DRAM	High bandwidth memory DRAM integrated into the FPGA package
Bandwidth	21.3GB/s	12.8GB/s	160GB/s	460GB/s
Typical Depth	16GB	2GB	4GB	16GB
Price / GB	$	$$	$$$	$$
PCB Req	High	High	Med	None
pJ / Bit	~27	~40	~30	~7
Latency	Medium	Low	High	Med

An Introduction to High Bandwidth Memory

HBM approaches the memory bandwidth problem differently, by removing the PCB from the equation. HBM takes advantage of silicon stacking technologies to place the FPGA and the DRAM beside each other in the same package. The result is a co-packaged DRAM structure capable of multi-terabit per second bandwidth. This provides system designers with a significant step function improvement in bandwidth, compared to other memory technologies.

HBM-capable devices are assembled using the de facto standard chip-on-wafer-on-substrate (CoWoS) stacked silicon assembly process from TSMC. This is the same proven assembly technology that Xilinx has used to build high-end Virtex devices for the past three generations. CoWoS was originally pioneered by Xilinx as silicon stacked interconnect technology for use with 28nm Virtex-7 FPGAs. CoWoS assembly places an active silicon die on top of a passive silicon interposer. Silicon-on-silicon stacking allows for very small, very densely populated micro-bumps to connect adjacent silicon devices—in this case, an FPGA connected to DRAM, with many thousands of signals in between. See Figure 2.



WP485_02_051217

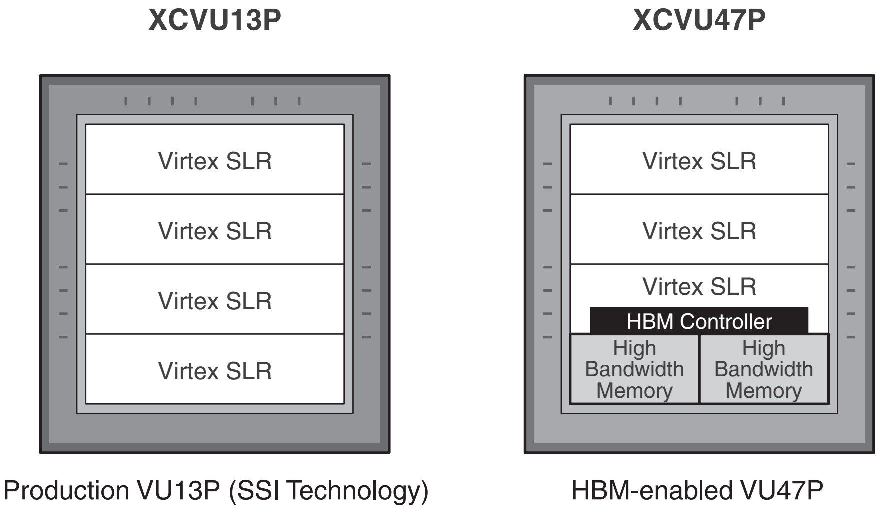
Figure 2: **TSMC CoWoS Assembly Allows for Thousands of Very Small Wires Connecting Adjacent Die**

With CoWoS assembly, the DQ traces connecting to HBM are less than 3mm in total length and have very low capacitance and inductance (LC) parasitic effects compared to typical DDR4 PCB traces. This allows for an HBM I/O structure that is 20X smaller in silicon area than a typical external DDR4 I/O structure. The HBM interface is so compact that a single HBM stack interface contains 1,024 DQ pins, yet uses half the I/O silicon area of a single DDR4 DIMM interface. Having 1,024 DQ pins with low parasitics results in tremendous bandwidth in and out of the HBM stack, with a latency similar to that of DDR4.

With an HBM-enabled FPGA, the amount of external DDR4 used is a function of capacity requirements, not bandwidth requirements. This results in using far fewer DDR4 components, saving designers both PCB space and power. In many cases, no external memory is needed.

Xilinx HBM Solution Overview

As illustrated in Figure 3, Virtex UltraScale+ HBM devices are built upon the same building blocks used to produce the Xilinx 16nm UltraScale+ FPGA family, already in production, by integrating a proven HBM controller and memory stacks from Xilinx supply partners. The HBM is integrated using the production-proven CoWoS assembly process, and the base FPGA components are simply stacked next to the HBM using the standard Virtex FPGA assembly flow. This approach removes production capacity risk since all of silicon, IP, and software used in the base FPGA family are already production qualified.



WP485_03_070219

Figure 3: **SSI Technology and HBM-enabled XCVU47P**

The only new blocks in the Virtex UltraScale+ HBM devices are the HBM, controller, and Cache Coherent Interconnect for Accelerators (CCIX) blocks. The transceivers, integrated block for PCIe®, Ethernet, Vivado® Design Suite, etc., are already production qualified, freeing designers to focus on getting the most from HBM features that will differentiate their products in the market.

Innovations for Timing Closure

Because the Virtex UltraScale+ HBM devices are built on an already-proven foundation, Xilinx engineers have been able to focus their innovation efforts on optimizing the HBM memory controller. The most obvious challenge in integrating HBM with an FPGA is in effectively harnessing all the memory bandwidth offered by HBM. Xilinx has innovated several key and unique features in these devices to help customers get the most usable bandwidth in and out of the HBM stacks.

Extended AXI Interface

The first innovation to highlight is the AXI Interface through which users connect to the memory controller. A typical integrated IP interfaces to the programmable logic immediately next to the embedded IP block. This is sufficient for most blocks, since the aggregate bandwidth of the local routing is plentiful to pass data in and out of that block. However, the bandwidth to and from the HBM is sufficiently high to motivate creating a new type of interface structure that physically extends into the programmable interconnect. This structure significantly increases the surface area of the interface, greatly increasing the available interconnect for the user's AXI interface, allowing for full 3.7Tb/s operation. See Figure 4.

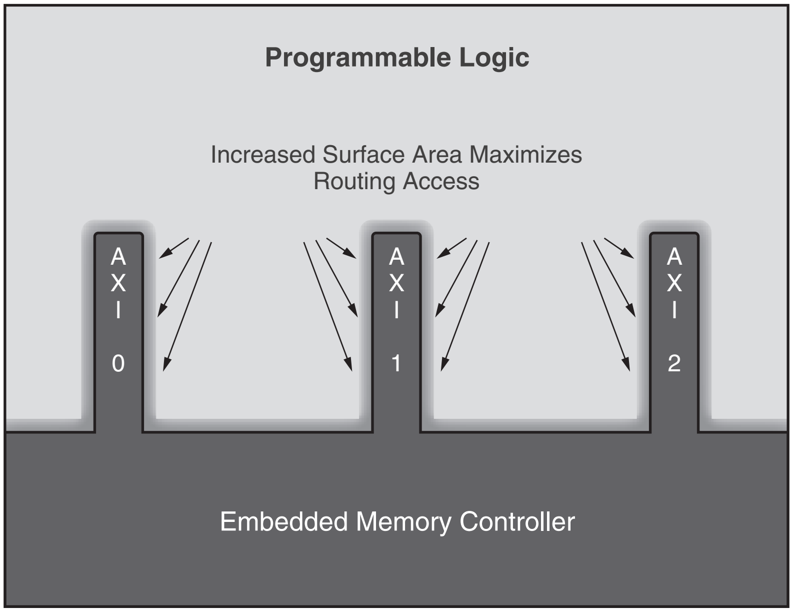

Programmable Logic
Increased Surface Area Maximizes Routing Access
AXI 0
AXI 1
AXI 2
Embedded Memory Controller
WP485_04_061317

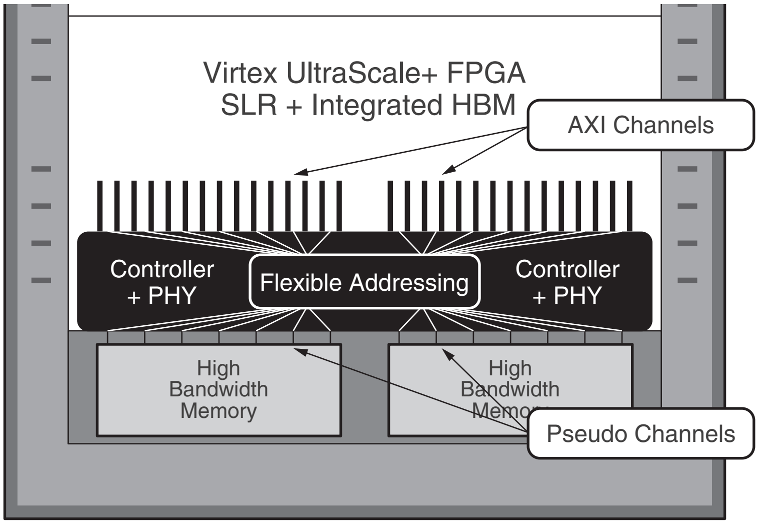
Figure 4: **Extended AXI Interface**

Flexible Addressing

The next innovation is flexible addressing, included in the HBM memory controller. The HBM stack architecture segments the memory address space into pseudo channels. This means that any given set of HBM DQ bits are allocated to a specific address region of the memory. So, if a designer wants to write data into a memory address, they can only write that data through the particular pseudo channel associated with that particular address.

This restriction is not ideal if a designer wants to treat the HBM stacks as a single contiguous memory, or partition them across pseudo-channel boundaries. To overcome this limitation, Xilinx has included an AXI switch network inside the embedded memory controller. This switch network routes memory reads and writes from any source AXI interface to any HBM pseudo channel based on address. This feature is called flexible addressing because it allows for any user AXI interface to access any HBM memory address.

Flexible addressing can be bypassed for those who want to optimize their own memory controller for very specific memory access patterns. See Figure 5.

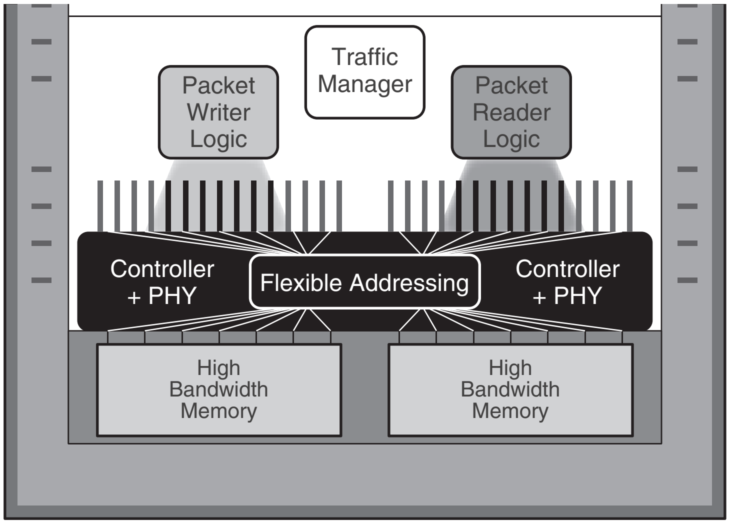


WP485_05_051717

Figure 5: **AXI Interfaces (to User Logic) and HBM Pseudo Channels (to HBM Stacks)**

Flexible addressing has several very important benefits:

1. **Gives users full control over how the HBM stacks are addressed.** Since the switch network can route the entire width of the device, the user does not need to adhere to the strict pseudo-channel requirements normally inherent to HBM. All 32 of the AXI interfaces can read or write from any HBM pseudo channel on either HBM stack, giving the user full control over address partitioning, regardless of pseudo-channel boundaries.
2. **Allows designers to connect using the most convenient AXI interface based on timing closure for the design.** For example, logic that writes to the memory no longer needs to be in the same location as logic that reads from the memory. In a basic traffic manager example, the AXI interfaces for packet writer and packet reader blocks can each be chosen to be in close proximity to the block. See Figure 6.



WP485_06_051917

Figure 6: **A Typical Ethernet Bridge Design**

With flexible addressing, the packet writer and packet reader logic can be physically separated so that they do not compete for routing resources.

3. **Provides a massive reduction in programmable routing resources.** There are many AXI buses in the memory controller, each of them 256 bits wide. So having 32 channels that route horizontally across the device in the integrated memory controller frees up considerable resources for the FPGA designer to use on higher value functionality. The AXI switch network, if fully implemented in FPGA logic, consumes 250,000 LUTs. With flexible addressing, the user gets a fully implemented switch network with zero LUT usage.
4. **Allows designers to use AXI interfaces more efficiently than the pseudo channels would otherwise allow.** A single HBM pseudo channel is subject to standard DRAM inefficiencies such as activate, pre-charge, and refresh delays. While the memory controller does re-order to improve efficiency, DRAMs are never 100% efficient. However, since a single AXI interface can pipeline multiple pseudo-channel accesses, it is possible to get higher AXI interface efficiency than HBM pseudo-channel efficiency. In many applications, this means fewer AXI interfaces are needed, which frees up more programmable resources.

Innovations for Power Efficiency and Thermal Management

Xilinx customers greatly value power efficient performance. The TSMC 16nm FinFET+ process enables dual-voltage operation, enabling customers to choose either the highest absolute performance or the highest performance-per-watt available. By leveraging that process, Xilinx offers the industry's lowest core voltage, resulting in 30% less total dynamic power, along with industry-leading transceiver technology, a beneficial mix of integrated blocks such as Ethernet, Interlaken, and PCIe in an FPGA.

HBM technology enables Xilinx to take power-efficient design to the next level, by eliminating scores of external memory interfaces and replacing them with on-interposer traces. Doing so reduces the capacitance of the memory interface and hence the power required to drive multiple Tb/s of memory bandwidth at 4X lower power per bit.

For thermal management, Xilinx offers several unique techniques to offset the inevitable increase in thermal density resulting from integrating HBM. Xilinx offers Virtex UltraScale+ HBM devices in heatsink-ready, lidless, bare die, flip-chip packages that significantly improve cooling to alleviate the expected higher thermal density. These lidless packages are already shipping in other Virtex UltraScale+ FPGAs, where they have been shown to improve the thermal design margin by about 10–20°C for most use cases. This results in higher compute ceilings and/or lower thermal design costs. Learn more by reading the Xilinx application note *Mechanical and Thermal Design Guidelines for the UltraScale+ FPGA D2104 Lidless Flip-Chip Packages*[Ref 3]. See Figure 7.

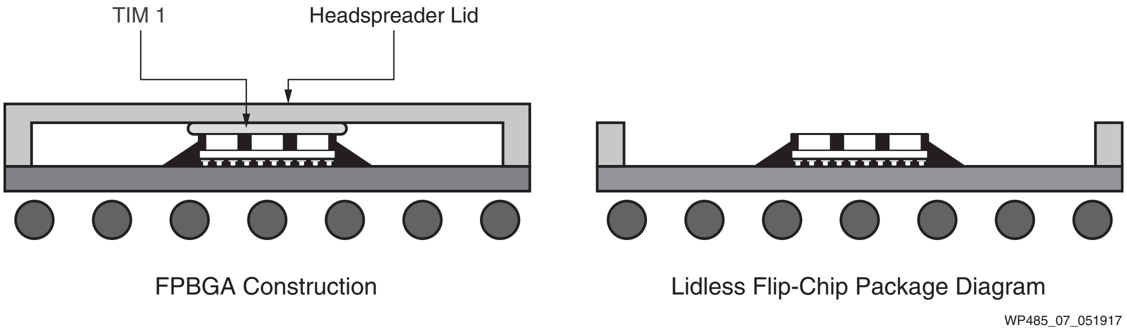


Figure 7: **Lidded vs. Lidless Flip-Chip Packages**

Application Example: Smart Network Interface Card

Marrying HBM with high-end programmable logic brings a great benefit to many applications across networking, data center, audio/video broadcast, radar, test and measurement, etc. One such application is the Smart Network Interface Card, or Smart NIC. A Smart NIC consists of one or more network ports, an interface to the CPU such as PCIe or CCIX, networking functions to be accelerated (e.g., OVS, GZIP, IPSec, SSL, etc.), and memory for packet storage and key lookup. A traditional Smart NIC implementation might require populating up to four 72-pin DIMMs on the PCB to provide sufficient memory bandwidth for servicing two 100G ports. Interfacing the four DIMMs requires driving 624 I/Os, contributing significantly to total power. Accommodating the four DIMMs requires a full-height, full-length (FHFL) form factor, which brings its own set of power and space efficiency challenges.

The same solution implemented in a VU45P with HBM can be reduced down to half-height, half-length (HHHL) because external DRAM components are replaced by the HBM stacks (see Figure 8). The VU45P solution (Figure 9) uses approximately 50% of the power since the I/O power drain of the DIMM interfaces is no longer required. An implementation using a VU45P device with two HBM stacks provides a 3X higher look-up rate thanks to the HBM bandwidth, and 4X more search entries than commercially available TCAMs. In addition to these inherent benefits for the end solution, an HBM implementation enjoys a simpler and lower risk design flow by simplifying the PCB and decreasing the memory subsystem complexity.

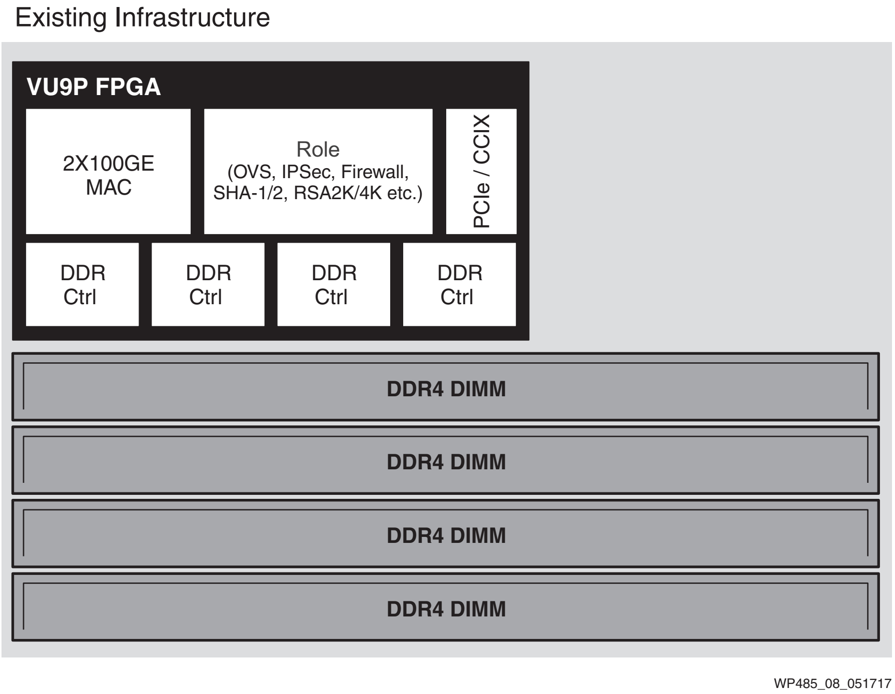


Figure 8: **Existing Infrastructure**

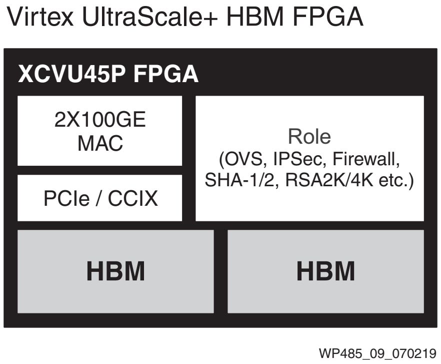

Virtex UltraScale+ HBM FPGA
XCVU45P FPGA
2X100GE MAC
PCIe / CCIX
Role (OVS, IPSec, Firewall, SHA-1/2, RSA2K/4K etc.)
HBM
HBM
WP485_09_070219

Figure 9: **Virtex UltraScale+ HBM Solution**

Conclusion

While many of tomorrow's systems exceed the bandwidth offered by DDR, HBM is the best alternative for dramatically boosting memory bandwidth with optimal performance per watt. Xilinx Virtex UltraScale+ HBM devices provide the right mix of memory bandwidth and programmable compute performance. With these devices, Xilinx focused its innovation efforts on helping designers make the most of HBM, while building upon a proven foundation of silicon process and architecture, assembly technology, and design tools. Designers and system architects alike will appreciate the benefits of adding HBM to their system using Virtex UltraScale+ HBM devices.

References

1. “On Global Electricity Usage of Communication Technology: Trends to 2030”, Anders S. G. Andrae and Tomas Edler, Apr 30, 2015.
   http://www.mdpi.com/2078-1547/6/1/117/htm
2. “Data Center Specific Thermal and Energy Saving Techniques”, Tausif Muzaffar and Xiao Qin, Jun 18, 2015.
   http://www.slideshare.net/xqin74/data-center-specific-thermal-and-energy-saving-techniques
3. Xilinx application note XAPP1301, *Mechanical and Thermal Design Guidelines for the UltraScale+ FPGA D2104 Lidless Flip-Chip Packages*.

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
07/15/2019	1.1	Updated Industry Trends: Bandwidth and Power, Table 1, Figure 3, Innovations for Power Efficiency and Thermal Management, Application Example: Smart Network Interface Card, and Figure 9.
06/14/2017	1.0	Initial Xilinx release.

Disclaimer

The information disclosed to you hereunder (the "Materials") is provided solely for the selection and use of Xilinx products. To the maximum extent permitted by applicable law: (1) Materials are made available "AS IS" and with all faults, Xilinx hereby DISCLAIMS ALL WARRANTIES AND CONDITIONS, EXPRESS, IMPLIED, OR STATUTORY, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE; and (2) Xilinx shall not be liable (whether in contract or tort, including negligence, or under any other theory of liability) for any loss or damage of any kind or nature related to, arising under, or in connection with, the Materials (including your use of the Materials), including for any direct, indirect, special, incidental, or consequential loss or damage (including loss of data, profits, goodwill, or any type of loss or damage suffered as a result of any action brought by a third party) even if such damage or loss was reasonably foreseeable or Xilinx had been advised of the possibility of the same. Xilinx assumes no obligation to correct any errors contained in the Materials or to notify you of updates to the Materials or to product specifications. You may not reproduce, modify, distribute, or publicly display the Materials without prior written consent. Certain products are subject to the terms and conditions of Xilinx's limited warranty, please refer to Xilinx's Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos; IP cores may be subject to warranty and support terms contained in a license issued to you by Xilinx. Xilinx products are not designed or intended to be fail-safe or for use in any application requiring fail-safe performance; you assume sole risk and liability for use of Xilinx products in such critical applications, please refer to Xilinx's Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos.

Automotive Applications Disclaimer

AUTOMOTIVE PRODUCTS (IDENTIFIED AS "XA" IN THE PART NUMBER) ARE NOT WARRANTED FOR USE IN THE DEPLOYMENT OF AIRBAGS OR FOR USE IN APPLICATIONS THAT AFFECT CONTROL OF A VEHICLE ("SAFETY APPLICATION") UNLESS THERE IS A SAFETY CONCEPT OR REDUNDANCY FEATURE CONSISTENT WITH THE ISO 26262 AUTOMOTIVE SAFETY STANDARD ("SAFETY DESIGN"). CUSTOMER SHALL, PRIOR TO USING OR DISTRIBUTING ANY SYSTEMS THAT INCORPORATE PRODUCTS, THOROUGHLY TEST SUCH SYSTEMS FOR SAFETY PURPOSES. USE OF PRODUCTS IN A SAFETY APPLICATION WITHOUT A SAFETY DESIGN IS FULLY AT THE RISK OF CUSTOMER, SUBJECT ONLY TO APPLICABLE LAWS AND REGULATIONS GOVERNING LIMITATIONS ON PRODUCT LIABILITY.