

This white paper describes some of the components involved in an external memory interface. Altera's portfolio of 28-nm FPGAs was developed to provide both the highest overall bandwidth of 921 Gbps and a highly efficient solution that allows a designer to get the most effective bandwidth possible.

## Introduction

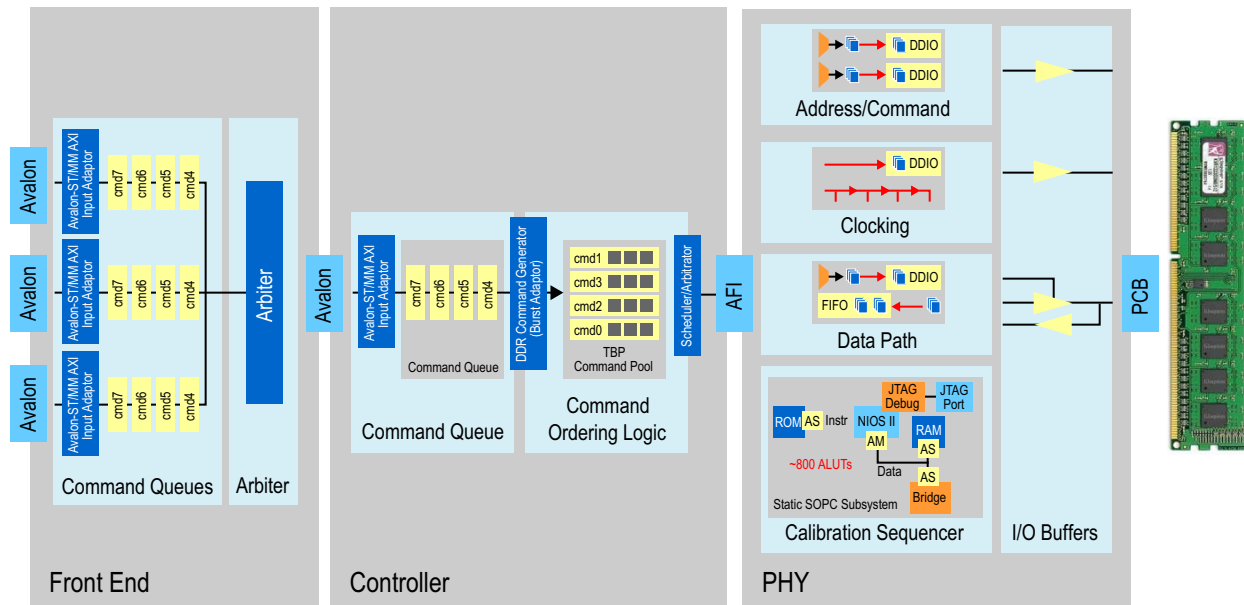
Many FPGA-based systems require an external memory interface. This memory interface often serves as a buffer between the external memory data path, which is often faster than the internal FPGA fabric, and the internal FPGA processing blocks. With the advent of transceiver-based FPGAs, the memory interface has become increasingly important. In order to ensure peak system performance, the memory must be able to store up to hundreds of gigabits of data as fast as the transceivers can provide that data to the FPGA. To keep pace, these data streams require a wide and fast memory interface .

Although I/O performance is important, it is not the complete story on bandwidth. The efficiency of the memory controller can be critically important in determining the actual system bandwidth that can be achieved. This "effective" bandwidth is a critical factor in determining the actual performance of a system.

## The Memory Interface

Altera's external memory interface controller consists of three blocks, as shown in [Figure 1](#). The multiported front end allows multiple processes inside the FPGA to share a common bank of memory, the memory controller implements all of the DDR3 command and addressing, and the physical layer interface (PHY) handles the timing on the data path itself. All three blocks are critical to the design and use of the memory interface block.

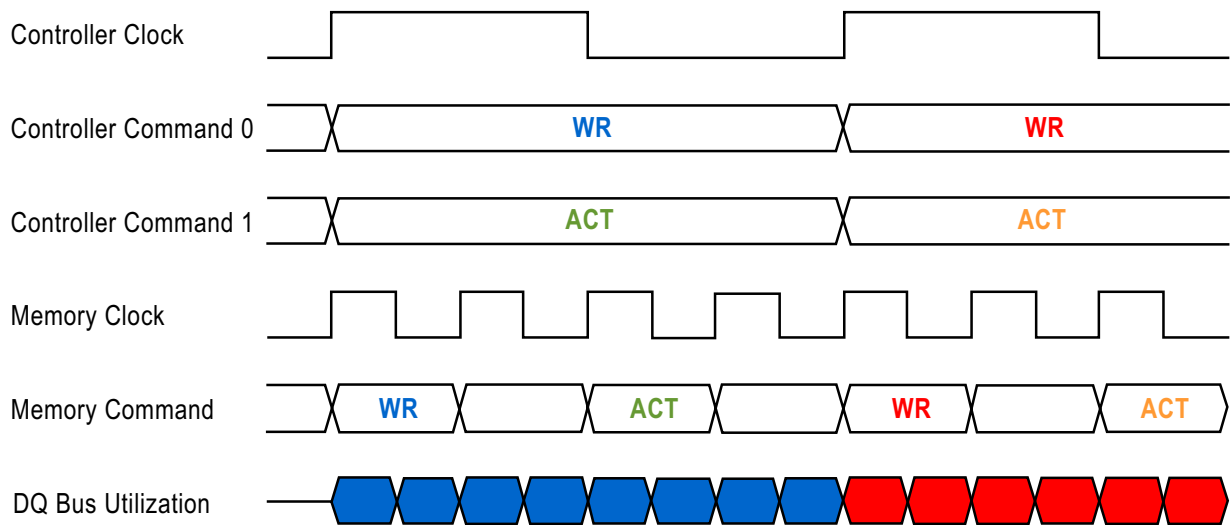
**Figure 1. Components of a Memory Interface**



## The PHY

In the Quartus® II development software version 11.0, Altera introduced a new type of controller to address the increasing speeds of DDR3 memories. The quarter rate controller allows the interface between the UniPHY version of PHY and the FPGA core fabric to run at a quarter of the rate of the DDR3 memory clock speeds. So a 2133-Mbps data interface with a 1066-MHz clock can be captured by the UniPHY and presented to the FPGA core fabric at a clock rate of only 266-MHz. This means that four bits of data would be presented to the core fabric on every 266-MHz clock edge. The quarter-rate controller supports 2T command timing as well. This means that a command is issued every two DDR3 clock cycles. [Figure 2](#) shows a timing diagram of the new quarter-rate controller with 2T command timing.

**Figure 2. Quarter Rate Controller with 2T Command Timing**



In addition, Altera supports full-rate and half-rate controllers, and which controller to use is dependent upon the speed of the interface. Table 1 shows the different DDR3 clock rates and the controller required.

**Table 1. DDR3 Clock Speeds vs. Controller Configuration**

DDR3 Clock	Quarter Rate	Half Rate	Full Rate
1066	✓		
933	✓		
800	✓		
667	✓	✓	
533	✓	✓	
400	✓	✓	
333	✓	✓	✓

Altera is a pioneer in the area of low-latency memory controllers. Altera's portfolio of 28-nm FPGAs implements a balanced clocked network in the periphery to reduce switching noise, while a hardened read-data FIFO buffer guarantees timing and makes it easier for the fitter to place the controller. These design changes have led to dramatic reduction in latency with the latest release of UniPHY. Table 2 shows the latency based on different command types for both the quarter-rate and half-rate controllers. These latencies are 60% better than the previous generations of UniPHY.

**Table 2. UniPHY DDR3 Memory Latency**

Controller Rate	Latency Type	Latency (Memory Clock Cycles)		
		Controller	PHY	Total
Half	Write command	12	3	15
	Read command	10	3	13
	Read data	0	7	7
Quarter	Write command	20	9	29
	Read command	20	9	29
	Read data	0	11	11

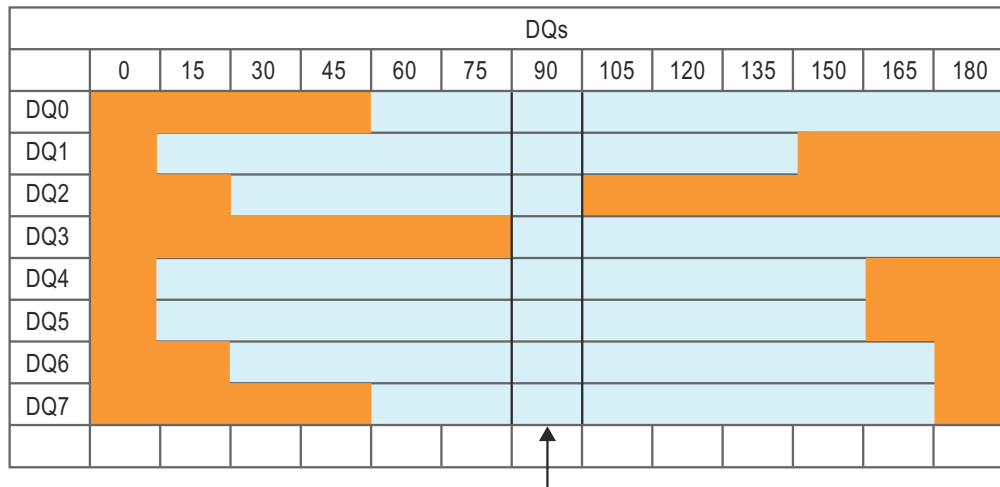
Table 3 shows that these latencies are also dramatically better than those of Altera's closest competition.

**Table 3. DDR3 Memory Latency—Altera vs. Competition**

Controller Rate	Latency Type	Latency (Memory Clock Cycles)		Advantage
		Competitor	Altera	
Quarter	Write command	46*	29	Altera
	Read command	46*	29	Altera
	Read data	31*	29	Altera

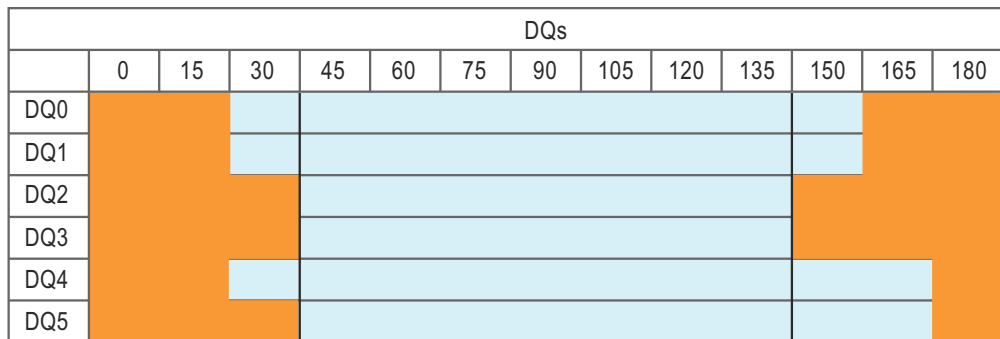
Board skew caused by variation in the traces can be significant when designing a 72-bit or larger DQ bus and can significantly impact the performance of the system. Figure 3 shows how board skew can significantly degrade the data valid window and cause errors on the memory bus. This is especially critical at higher data speeds.

**Figure 3. Reduced Data Valid Window Due to Board Skew**



Altera’s UniPHY has configurable delay chains that can adjust the delays of each DQ pin. This adjustment widens the data eye by lining up each of the individual DQ signals, then sending a PBRs pattern to memory and doing both a gross and fine calibration of the bus. This iterative pattern is performed by an embedded soft processor, part of the UniPHY PHY, which continues to skew the lines until it achieves the largest data eye without error. Figure 4 is the resulting eye after deskew.

**Figure 4. Increased Data Valid After Calibration**



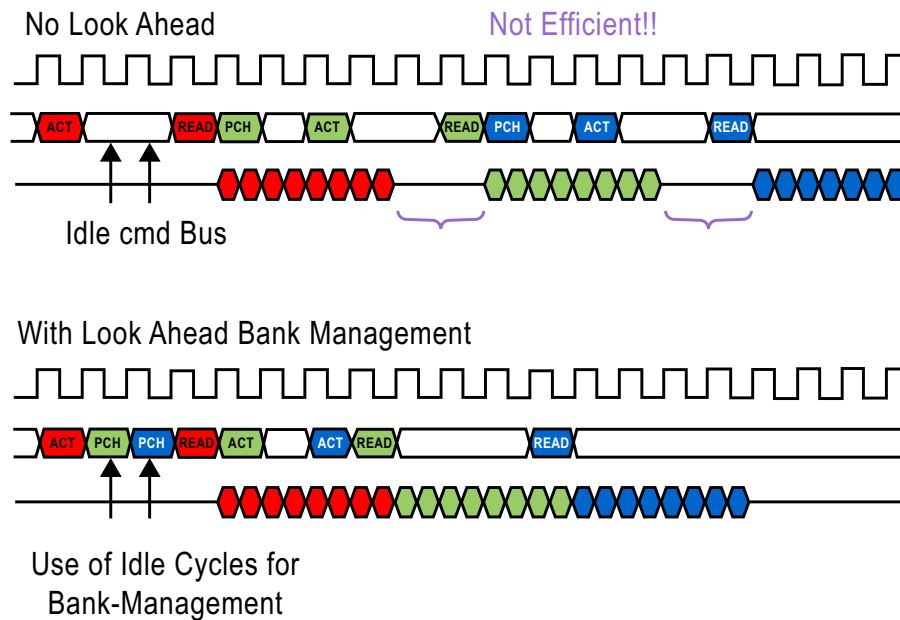
## The Controller

Investing in higher speed memory interfaces means an investment in memory bandwidth. Unfortunately, DDR3 memory, although lower in cost, is not designed to achieve maximum bandwidth, so the efficiency of the controller at managing the required SDRAM commands is critical in achieving maximum bandwidth through the interface. Efficiency is the amount of occupied DQ cycles (non-idle) that occur on the bus divided by the total number of cycles on the bus, or

$$\text{Efficiency} = \frac{\text{\# of clock cycles that DQ bus is occupied}}{\text{\# of clock cycles in the period}}$$

Increased efficiency on the bus can be achieved in two ways. The first is by reordering commands to take advantage of idle or dead cycles. Altera’s advanced bank management recognizes that precharges are required when the system switches banks and places those precharges to take effect during idle cycles. This minimizes the impact of bank switching on the DQ bus by eliminating the dead cycles that can occur between operations. Figure 5 shows the impact of reordering two precharges. In this example the number of idle cycles was reduced by four.

**Figure 5. Advance Bank Management**



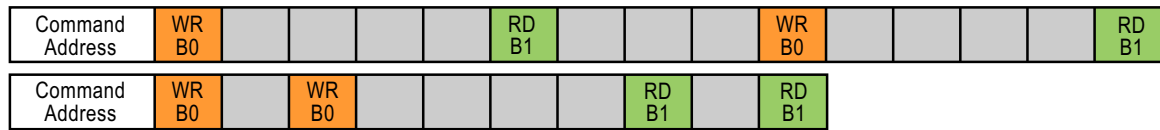
Command	Address	Condition
Read	Bank 0	Activate Required
Read	Bank 1	Precharge Required
Read	Bank 2	Precharge Required

In the case of a page hit, the Quartus II controller also has the ability to automatically cancel the precharge, so there is no need to change banks.

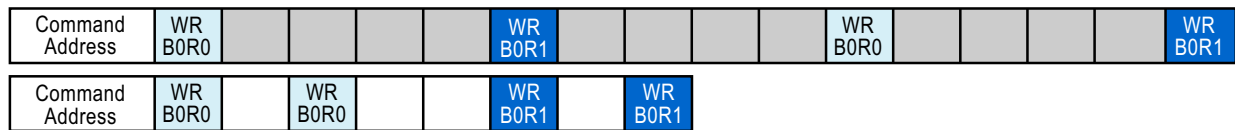
The second way to improve efficiency is to change the data order or commands. Bus turnaround time and the bank cycle time (tRC) can result in large idle cycles that reduce the overall efficiency of the bus. To minimize the bus turnaround time, it is important to group similar commands together. In other words, to make sure read and write commands are issued in groups to minimize the number of bus turnaround events that occur. Figure 6 shows that the 4-cycle time hit caused by bus turnaround time only needs to occur once in the transaction if the system can group the writes and reads together. The overall impact is to remove five dead cycles from the bus.

**Figure 6. Data Reordering to Minimize Dead Cycles**

Minimize bus turnaround time by grouping read and write transactions

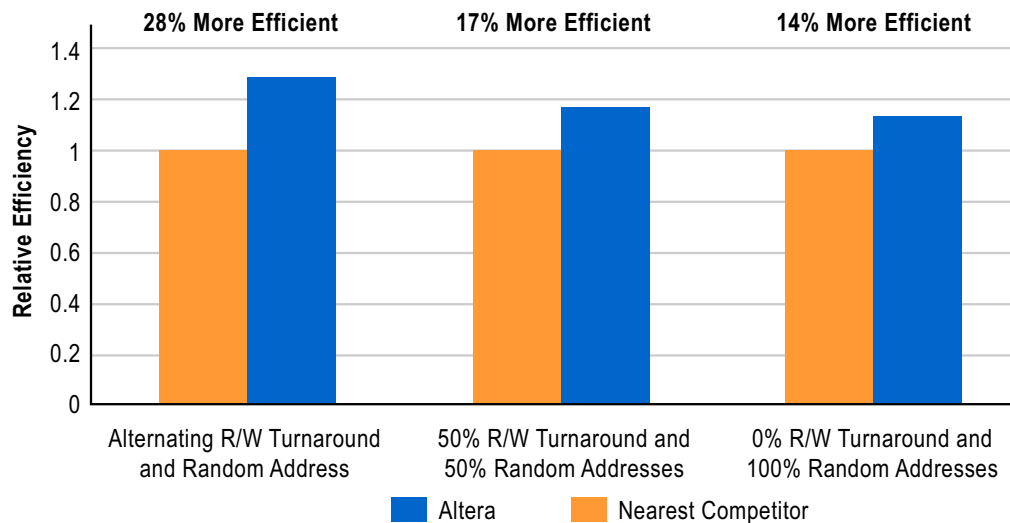


Minimize tRC impact by reordering transactions with bank conflicts



In order to minimize tRC, commands to the same bank are gathered together to remove the four idle cycles it takes to change banks on SDRAM memory. This type of operation results in the removal of eight idle cycles from these transactions. With these types of operation, Altera has been able to gain significant improvements in efficiency in the 28-nm portfolio of devices, as shown in Figure 7.

**Figure 7. Memory Efficiency of Altera’s New High-Performance Memory Controller**

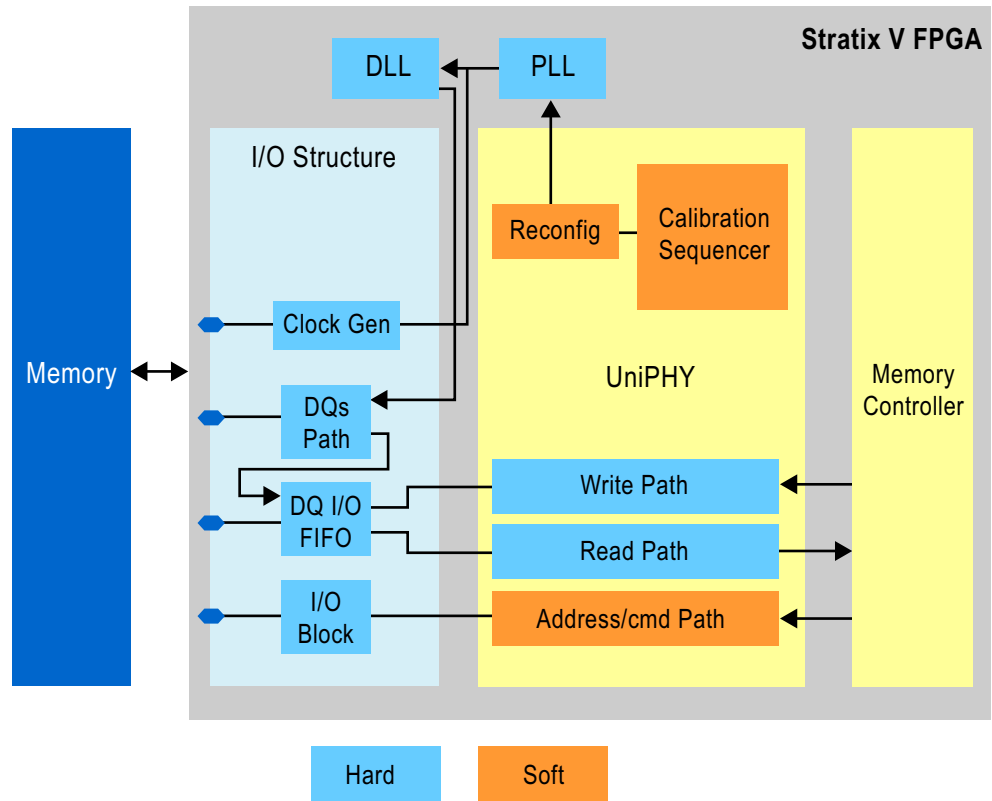


## Hard Memory IP vs. Soft Memory IP

Altera’s 28-nm portfolio of FPGAs provides two types of memory solutions: soft IP, which is provided in the Stratix® V, Arria® V, and Cyclone® V FPGA families, and a hard IP solution, which is provided in the Arria V and Cyclone V FPGA families.

Soft IP consists of the UniPHY and the high-performance memory controller. The hard read/write data paths ensure timing is met at the highest speeds. Figure 8 shows the hard paths, including the I/Os, the PLLs, the DLL, and the read/write FIFO buffers, in the soft IP.

Figure 8. Hard Paths in the Soft IP

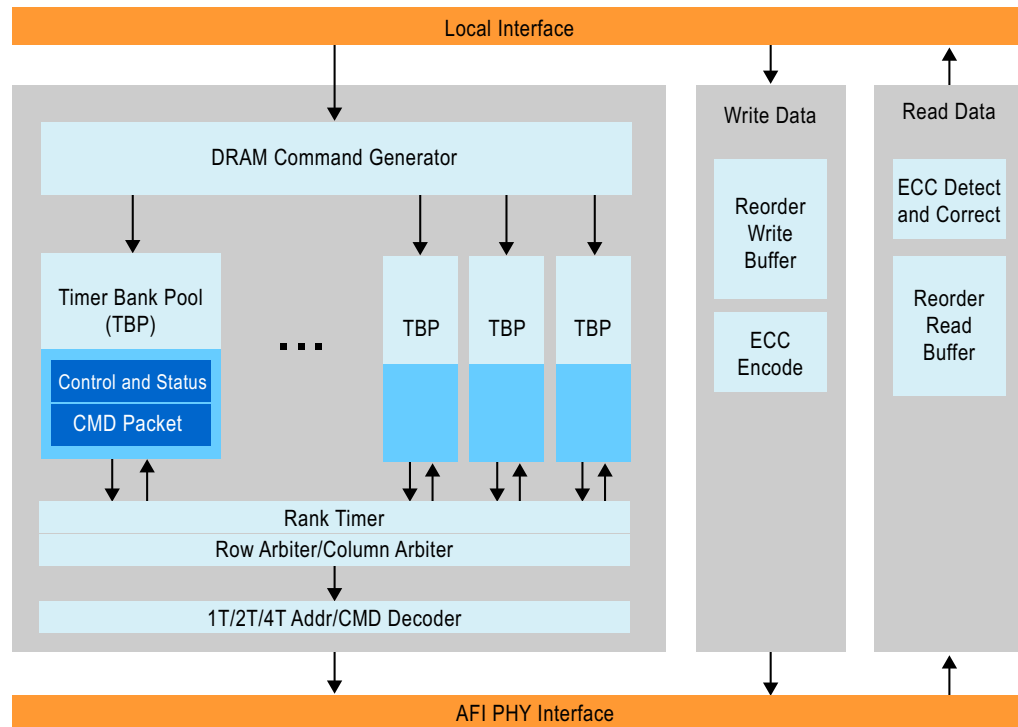


Altera offers soft IP instead of hard IP to allow the designer to choose where they can place the memory controller and the ability to size the memory controller based on the system requirements, especially in Stratix V FPGAs.

The hard memory controller IP consists of a hard UniPHY, a high-performance memory controller, and a multiported front end. The hard IP has a fixed location on the die and has a fixed maximum width: x16 in the case of Cyclone V FPGAs and x32 in the case of Arria V FPGAs. Furthermore, the hard IP runs at full rate, which allows for decreased latency and minimizes the required bus width of signals to the core of the device. This simplifies the overall memory design in Arria V and Cyclone V FPGAs and provides a truly out-of-the-box experience for the designer. Simply put, it just works.

## The MPFE

The hard multiported front end (MPFE) is a new feature in the Arria V and Cyclone V FPGAs. It allows multiple internal processes in the core to access the same hard memory interface. As shown in Figure 9, the MPFE has two types of ports: four bidirectional data ports and six address and command ports. The address and command port supports allow access for up to four bidirectional processes or up to six unidirectional processes. Each of these processes is considered either a read or a write transaction, and the interface to each port is the Avalon® Memory-Mapped (Avalon-MM) system interconnect.

**Figure 9. Block Diagram of MPFE with a Single Port, High-Performance Memory Controller**

The complete memory controller provides two levels of scheduling for the six address and command ports. Per-port scheduling takes place in the MPFE and DRAM burst scheduling takes place in the high-performance memory controller.

The per-port scheduler in the MPFE decides which of the active addresses and command ports to service next by first dividing the transactions on each port into a series of individually scheduled DRAM bursts. Then these bursts are serviced based on a deficit round robin (DRR) arbitration algorithm. The algorithm uses a per-port absolute priority and weight, which can be updated dynamically, and allows for temporary over and under service so the controller can smooth out memory traffic. A port's priority is automatically increased if its maximum latency is greater than the worst-case latency allowed in the controller. A port may also be configured for priority to allow multiple-burst transactions to keep a DRAM page open.

## Conclusion

Altera provides the fastest, most efficient, and lowest latency memory controllers, which allow designers to work with today's higher speed memories quickly and easily. Soft IP, provided in the Stratix V, Arria V, and Cyclone V FPGAs, gives the designer the flexibility to create interfaces that meet system requirements while benefiting from Altera's industry-leading performance. The hard IP, provided in Arria V and Cyclone V FPGAs, gives the designer a complete out-of-the-box experience when developing a memory controller, allowing them to get the design up and running quickly without having to worry about the memory interface. Altera understands that a fast and robust memory interface is crucial for many designers. Altera is committed to not only making those designs successful but ensuring that the implementation is both fast and easy.

## Further Information

- Arria V FPGAs: Balance of Cost, Performance, and Power:  
[www.altera.com/devices/fpga/arria-fpgas/arria-v/arrv-index.jsp](http://www.altera.com/devices/fpga/arria-fpgas/arria-v/arrv-index.jsp)
- Video: “Arria V FPGA Sneak Peek: Transceiver Operation at 6.375 Gbps and 10.3125 Gbps”:  
[www.altera.com/b/arria-v-fpga.html](http://www.altera.com/b/arria-v-fpga.html)
- Webcast: “Achieving 1066-MHz DDR3 Performance With Advanced Silicon and Memory IP”:  
[www.altera.com/education/webcasts/all/wc-2010-1066mhz-ddr3-silicon-memory-ip.html](http://www.altera.com/education/webcasts/all/wc-2010-1066mhz-ddr3-silicon-memory-ip.html)
- White Paper: *Achieving Lowest System Cost with Midrange 28-nm FPGAs*:  
[www.altera.com/literature/wp/wp-01175-lowest-system-cost.pdf](http://www.altera.com/literature/wp/wp-01175-lowest-system-cost.pdf)

## Acknowledgements

- Trung Tran, Staff Product Marketing Manager, High-Density Products, Altera Corporation

## Document Revision History

Table 4 shows the revision history for this document.

**Table 4. Document Revision History**

Date	Version	Changes
November 2011	1.0	Initial release.