



# BLAZAR BE2-BURST Accelerator Engine

## Intelligent In-Memory Computing

### 576Mb Memory



## PRODUCT BRIEF

**Acceleration Engines give Software and Hardware System Architects Acceleration Options not previously available**

### BANDWIDTH ENGINE (BE) INTRODUCTION

The **BLAZAR Family of Accelerator Engines** support high bandwidth, fast random memory access rates and embedded In-Memory Functions (IMF) that solve critical memory access challenges for memory bottlenecked applications like network search, statistics, buffering, security, firewall, 8k video, anomaly detect, genomics, ML random forest of trees, graph/tree/list walking, traffic monitoring.

The **Bandwidth Engine 2 (BE2-BURST)** combines the high speed serial memory with in memory **Bandwidth functions** called **BURST**.

#### System benefits of BE-2...

- **FPGA Acceleration for Xilinx and Intel**
- **QDR comparison**
  - *Replaces up to 4 QDR/RLDRAM memory devices*
  - *Equal to or outperforms QDR memories*
  - *Equivalent system latency*
- Memory architecture allows up to 16 simultaneous accesses
- **Bandwidth IMFs - BURST single commands for sequential read and write functions for Data Movement nearly doubles or triples bandwidth**
- The devices support application acceleration for aggregate throughput rates up to 320 Gb/s (160 Gb/s full duplex)

### KEY FEATURES / PRODUCT OPTIONS

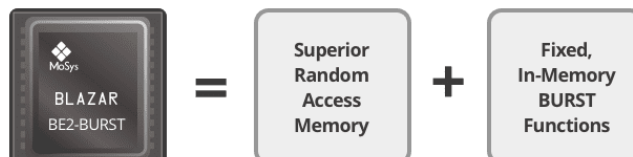
- High Bandwidth, low pin count serial interface
  - Highly efficient reliable transport command and data protocol optimized for 90% efficiency
  - Eases board layout and signal integrity, minimal trace length matching required, operates over connectors
- 576Mb SRAM (8M x 72b)
  - User defined WORD length
  - Typical 8x, 16x, 32x, 36x, ... 72x
- High access rate SRAM class memory
  - Up to 3.3 Billion transactions/sec
- High cycle rate memory
  - 3.2 ns tRC
- In-Memory BURST Bandwidth Functions
  - BURST sequential read and write functions for Data Movement
  - Burst length: 1, 2, 4, 8 words
  - Use of In Memory Burst Functions significantly outperform QDR
  - Reduction of I/O pins up to 7X
- Highest Single Chip Bandwidth – up to 320 Gb/s throughput

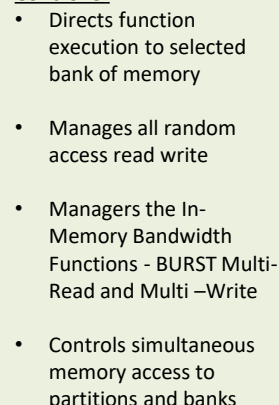
### APPLICATIONS FOCUS

- High bandwidth data access application where low latency and Movement of Data is a critical requirement.
- Applications needed large SRAMs.
- FPGA Acceleration for Xilinx and Intel

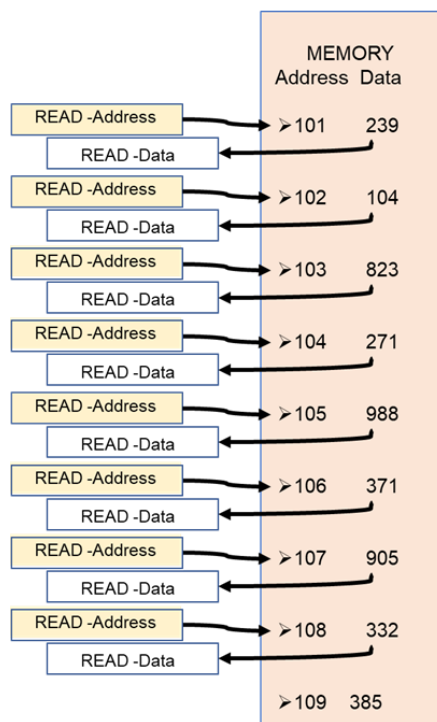
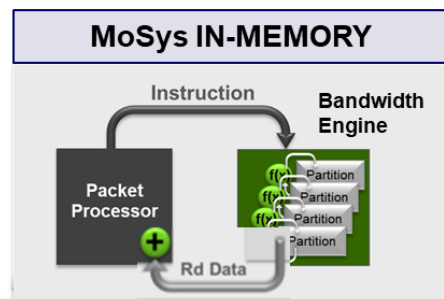
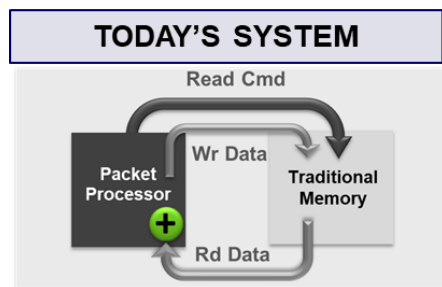
## MoSys ACCELERATOR ENGINE Elements of BE2-BURST

MoSys Engines have a Unique Memory Architecture that can replace SyncRAM/RLDRAM memories and Embeds In-Memory Functions (IMF) that execute many times faster. A single function replaces many traditional memory accesses.



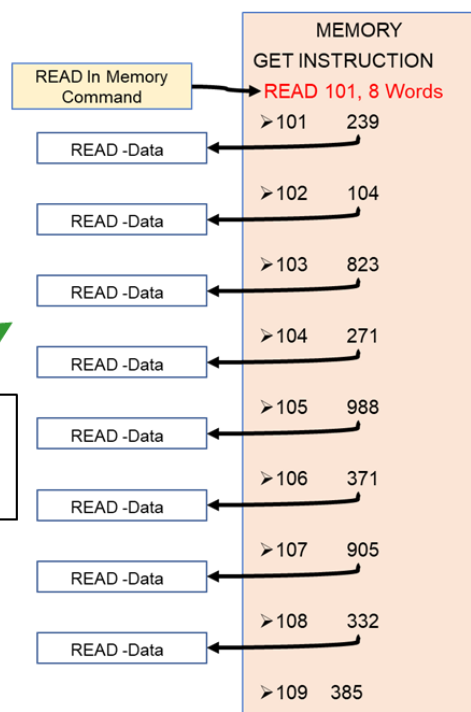


## PB-10 1218



- QDR requires 2 READ operations
- Total time estimate of 6ns

- BE2 requires 1 READ operation
- Total time estimate of 3ns



## Example In-Memory BURST time saving comparing 1 QDR to 1 BE2

### QDR...36b word width SINGLE READ

- Reads 144b/read
- **4 words of 36b**
- **Estimate of 3ns**

### BE2...36b word width SINGLE READ

- BE has to 8 lane ports (A & B)
- Port A can READS 288b/read
  - **8 words of 36b**
- Port B can READS 288b/read
  - **8 words of 36b**

### RESULT

- Total SINGLE READ "using" both A & B together
- **16 words of 36b**
- **Estimate of 3ns**
- **4 times a QDR Bandwidth**

## MoSys In-Memory Functions

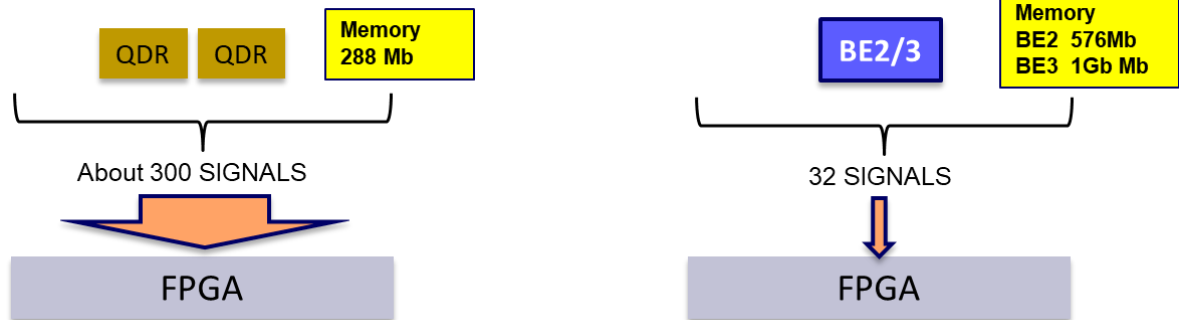
- BURST
  - Multiple Sequential READS
  - Multiple Sequential WRITES
  - Function types ~12

**Save significant system time with higher bandwidth**

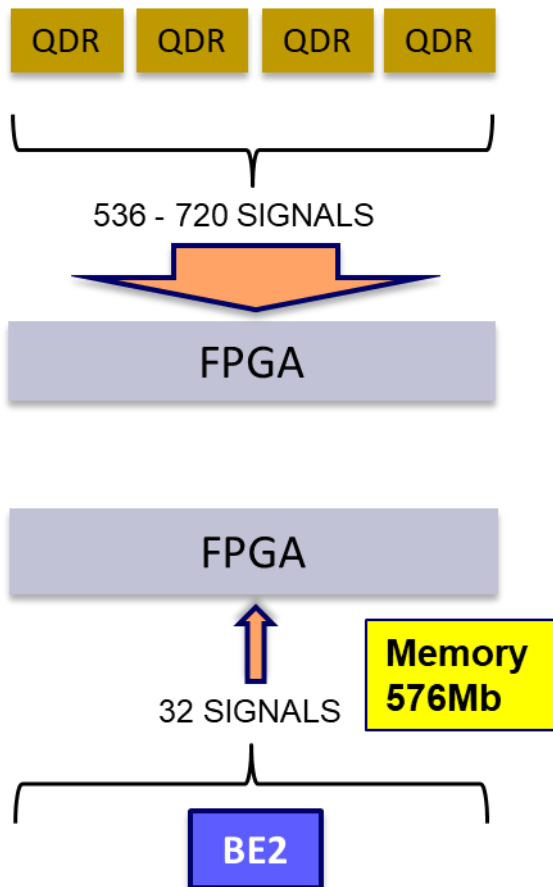


Serial memories bring many advantage over traditional parallel signal memory device like QDRs.

## Allows high bandwidth over a few pins



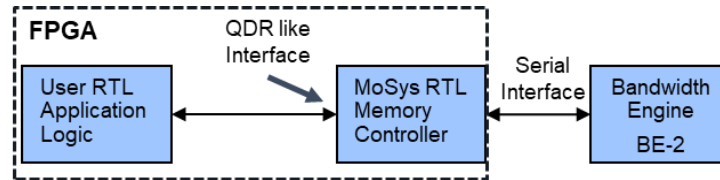
This is a typical 2 QDR design compared with a BE2 and BE3



## More High Speed memory generally allows acceleration options for software and hardware architects/Designers

### Overview comparison of BE2 to QDR

- Memory size
  - BE2 with 576Mb equivalent to 4 QDR at 144Mb per device
- Device PCB board Space Saving
  - 1 BE2 device vs 4 QDR device
- Signal Pins Reductions
  - 4 QDR...576Mb requires 500-700 pins/signals
  - 1 BE2...1Gb with as few as 32 signal pins (some systems require only 16 pins/signals)
  - All BE devices have *Auto-Adaptation* which handles on-board signal tuning, eliminating the need for any external components to insure a clean, reliable signals
- Costs
  - One BE-2 is approx. the price of 2 QDR memories with 4x the memory of single QDR
- Benefits
  - Larger Buffers, High Bandwidth
  - Allows Realtime operations and analysis at Line rate
  - Eliminates need for complex parallel operations using RLDRAM, HBM, or slow DRAM



## MoSys supplied RTL Controller simplifies the user interface with the BE.

MoSys can supply the FPGA RTL Memory Controller interfaces with the MoSys Bandwidth Engine. This controller is between the User Application logic and the BE device. It handles all the logic for the Serial GigaChip Interface (GCI) between the FPGA and BE as well as all memory addressing and commands and looks to the user like a QDR interface.

### MEMORY CONTROLLER

- Converts the Bandwidth Engine serial protocol to a FPGA parallel QDR like interface to the user and is provided at NO COST TO THE USER.
- Signal interface to the user from the MoSys RTL Controller is a simple SRAM memory read/write operation.
- *Supports all of the In-Memory BURST and RMW commands to achieve higher performance than a QDR*

### GRANULARITY OF MEMORY WORD WIDTH

Memory WORD width

- The RTL Memory controller allows the user to define a word width that best fits the application.
- Memory WORD width is user definable
  - Typical word sizes are 8, 16, 32, 36, 64 ...
- While the memory on the BE2 is organized as 8Mx72b and the BE3 is 16Mx72b, the address conversion mapping from the selected WORD width to the BE memory is handled by the RTL.
  - Address translation to BE2 memory organization is transparent to the application

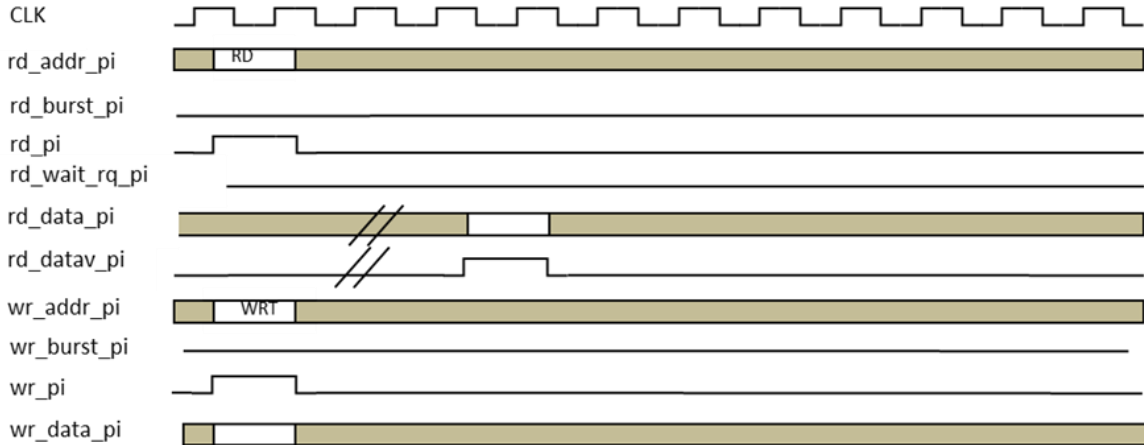
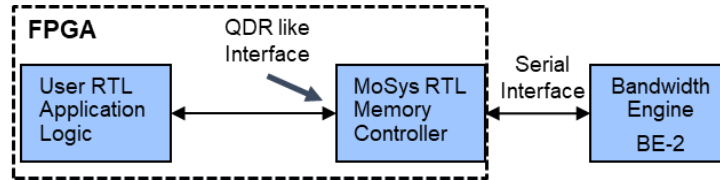
### SERIAL HIGH SPEED GCI DEVICE INTERFACE

- MoSys RTL handles all serial protocol conversion in the FPGA from the BE resulting in a parallel like QDR interface
- Systems use up to 16 SerDes lanes. (Can use as few as 4 lanes on one port)
- Controller supports 4, 8 or 16 lanes depending FPGA pins available and application bandwidth requirement
- If user would like to writes their own controller, the GCI protocol is available.

**The signal interface at the User Application is a simple SRAM memory Address, Data, Control structure with burst capability. This simple interface shields the users from the BE2 commands, serial interface and the scheduling logic for Bandwidth Engine memory partition timing.**



# High Speed GCI Serial Interface



SIGNAL NAME	WIDTH	DIR	DESCRIPTION
<b>Read Interface</b>			
rd_p	1	In	Assertion of this signal indicates that this is a read transaction.
rd_addr_p	32	In	Read address. Please refer to the Address section of this specification to see the detail of this address field.
rd_partsel_p	1	In	Indicates the BE-2 partition that this read command will be operated upon: 0 = Partition 0 for GCI port A, Partition 1 for GCI port B 1 = Partition 2 for GCI port A, Partition 3 for GCI port B
rd_data_p0	*	Out	Returned data from BE-2 memory. This data is qualified by the "rd_datav_p0" signal
rd_data_p1	*	Out	Returned data from BE-2 memory. This data is qualified by the "rd_datav_p1" signal. Note that rd_data_p1 will only have valid data if rd_data_p0 is valid as well. rd
rd_datav_p0	1	Out	The Memory Controller asserts this signal to indicate the current data in the "rd_data_p0" bus is valid
rd_datav_p1	1	Out	The Memory Controller asserts this signal to indicate the current data in the "rd_data_p1" bus is valid. Note that rd_data_p1 will only have valid data if rd_data_p0 is valid as well
rd_wait_rq_p	1	Out	The Memory controller asserts "rd_wait_rq_p" to indicate that it cannot accept the current read request from user. The User Application should hold all the request signals (rd_p, rd_addr_p ...) until the de-assertion of this signal.

SIGNAL NAME	WIDTH	DIR	DESCRIPTION
<b>Write Interface</b>			
wr_p	1	In	Assertion of this signal indicates that this is a write transaction.
wr_addr_p	32	In	Write address of the memory for this transaction. Please refer to the Address section of this specification to see the detail of this address field.
wr_partsel_p	1	In	Indicates the BE-2 partition that this write command will be operated upon: 0=Partition 0 for GCI port A, Partition 1 for GCI port B 1=Partition 2 for GCI port A, Partition 3 for GCI port B
wr_data_p	*	In	Write data from the User Application logic.
wr_wait_rq_p	1	Out	The Memory controller asserts "wr_wait_rq_p" to indicate that it cannot accept the current write request. The User Application should hold all the request signals (wr_p, wr_addr_p ...) until the de-assertion of this signal.

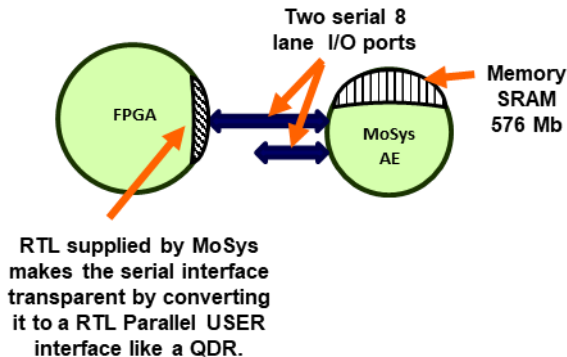






## STANDARD MEMORY INTERFACE USE

- REPLACE RLDram or UP TO 4 QDR SRAMS
- Most systems only require one Port as shown

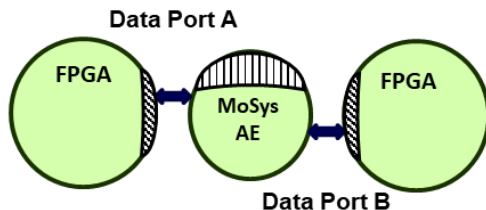


- Each Accelerator Engine Memory has two 8 lane serial Ports
- Typical QDR system needs only one port as shown
- Each port has 8 Data Lanes which is 32 signals
- Typical QDR application requires only 1 port on a BE-2 which is 32 signals (**Some applications can use only 4**)
- In addition the MoSys devices have *auto-adaptation which handles on-board signal tuning*
  - Eliminating the need for any external components to insure a clean, reliable signals

## DUAL PORT MEMORY USE

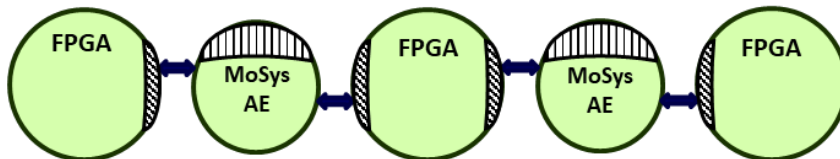
### Traditional DUAL PORT MEMORY

- True Dual Port operation
- Allows simultaneous memory access from each Port

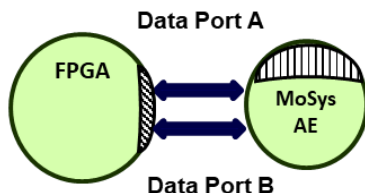


- Each Accelerator Engine Memory has two 8 lane serial Ports.
- Each port has 8 Data Lane which is 32 signals.
- A BE2 can operate as a true Dual Port with simultaneously memory access from each port
- In addition the MoSys devices have *auto-adaptation which handles on-board signal tuning*
  - Eliminating the need for any external components to insure a clean, reliable signals

### Dual Port used in a PIPELINE MEMORY Application



## SUPER HIGH BANDWIDTH MEMORY USE



- The devices support application acceleration for aggregate throughput rates up to 320 Gb/s (160 Gb/s full duplex)
- For extremely high bandwidth requirements, these two ports can be combined as one super high bandwidth port.
- Each Accelerator Engine Memory has two 8 lane ports. Using both ports, 64 signals
- In addition the MoSys devices have *auto-adaptation which handles on-board signal tuning*
  - Eliminating the need for any external components to insure a clean, reliable signals



# Accelerator Engine Family Overview

## Software Define - Hardware Accelerated

Software and System Architects can improve application performance by accelerating the memory access and utilizing the In-Memory BURST and In-Memory RMW Functions.

**BE2 with 576Mb** or the **BE3 with 1Gb** of memory comes in two version with different In-Memory acceleration functions.

- **BURST Functions ... High speed data movement and access functions**
- **RMW Functions ... Computing and Decision functions**

The different Accelerator Engine devices allow application tuning to achieve increasing levels of performance up to our most powerful engine... the Programmable HyperSpeed Engine (PHE).

**The Programmable HyperSpeed Accelerator Engine (PHE)** is essentially a BE3 with 1Gb of memory with BURST and RMW In-Memory Functions and has 32 RISC cores embedded in the device. *This is the ultimate in acceleration possibilities.*

- **User defined Functions**
- **Future- Standard functions from MoSys**

In-Memory	Part Number	Description	Package	Interface					Memory		Access Rate Billion Transaction/s	In-Memory Functions		
			Pkg Size	Lanes	Rate per Lane Gb/s				BW	tRC		BURST for Data Movement	RMW / ALU for Compute and Decision	Custom & User Functions with 32 RISC Cores
			mm	Tx/Rx	10.3	12.5	15.6	25	Gb	ns	Gb			
BURST	MSR622	Bandwidth Engine 2 Burst Serial 0.5Gb High Access Memory	FCBGA 19x19	16	✓	✓			320	3.2	0.5	3.3	✓	
	MSR630	Bandwidth Engine 3 Burst Serial 1Gb High Access Memory	FCBGA 27x27	16		✓	✓	✓	380	2.7	1	6.5	✓	
RMW	MSR820	Bandwidth Engine 2 RMW Serial 0.5Gb High Access Memory with ALU for RMW functions	FCBGA 19x19	16	✓	✓			320	3.2	0.5	3.3	✓	✓
	MSR830	Bandwidth Engine 3 RMW Serial 1Gb High Access Memory with ALU for RMW functions	FCBGA 27x27	16		✓	✓	✓	380	2.7	1	6.5	✓	✓
Program	MSPS30	Programmable Accelerator Engine Serial Interface, 1Gb Memory, 32 RISC Processor cores for custom algorithms, compute, functions	FCBGA 27x27	16		✓	✓	✓	717	2.7	1	24 Internal	✓	✓

[www.mosys.com](http://www.mosys.com)

LEARN MORE:

<https://mosys.com/blazar-family-of-accelerator-engines/>

