BLAZAR Family of Intelligent Data Accelerators

FAMILY BRIEF

BLAZAR SERIAL ACCELERATOR ENGINE INTRODUCTION

The **BLAZAR Family of Accelerator Engines** support high bandwidth, fast random memory access rates and *embedded **I**n **M**emory **F**unctions (**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.

Applications benefit from the fast random memory access rate, memory density, memory architecture and the "In memory functions" of the Blazar family:

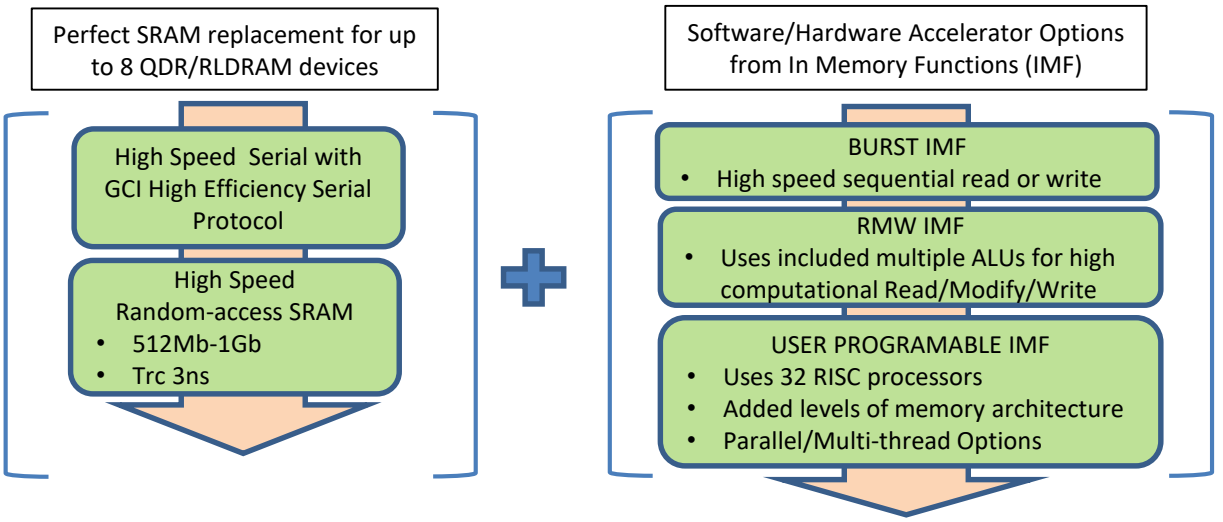
- Replaces up to 8 QDR/RLDRAM memory devices
- *Two types of In Memory Functions: BANDWIDTH Functions for high throughput applications and COMPUTE Functions for statistics, filtering, data manipulation and logical operations.*
- IMFs lowers latency up to 4x and increases available access rates 6x by avoiding memory bottlenecks
- Memory architecture allows up to 32 simultaneous accesses
- Accelerates applications by offloading functions from the FPGA into the Programable HyperSpeed Engine (PHE) freeing up FPGA resources to add features
- Total aggregate of *24 Billion memory operations per second*
- Support application acceleration for aggregate throughput rates ranging from 50Gb/s to over 400Gb/s per device.

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, no trace length matching required, operates over connectors
- 512Mb-1Gb of 1T-SRAM (8Mx72b or 16Mx72b)
- High access rate SRAM class memory 3.75B to 24B access/s with tRC, depending on device from 0.67ns to 3.2ns.
- Latencies as low as 8ns internal, 16ns external (pin to pin)
- **Bandwidth IMFs** - BURST sequential read and write functions for Data Movement nearly doubles bandwidth
- **Compute IMFs** - Atomic RMW functions for statistics, metering, filtering, mutex operations.
- Reduction of I/O up to 7X and avoids stale data problems
- 32 Programmable, 1.5Ghz, Multi-threaded RISC Processor Engines (PE)
  - **User programmable (IMF)** In-Memory Embedded *Bandwidth and Compute Functions* or Algorithms - with up to 256 threads

MoSys ACCELERATOR ENGINE Elements

MoSys Engines have Unique Memory Architectures that can replace SRAM/RLDRAM memories and embeds In Memory Functions (IMF) that execute many times faster. A single embedded function can replace several traditional memory accesses.



PROGRAMABLE HYPERSPEED ENGINE (PHE) ARCHITECTURE
(Differenced between other Devices are noted)

High speed serial I/O

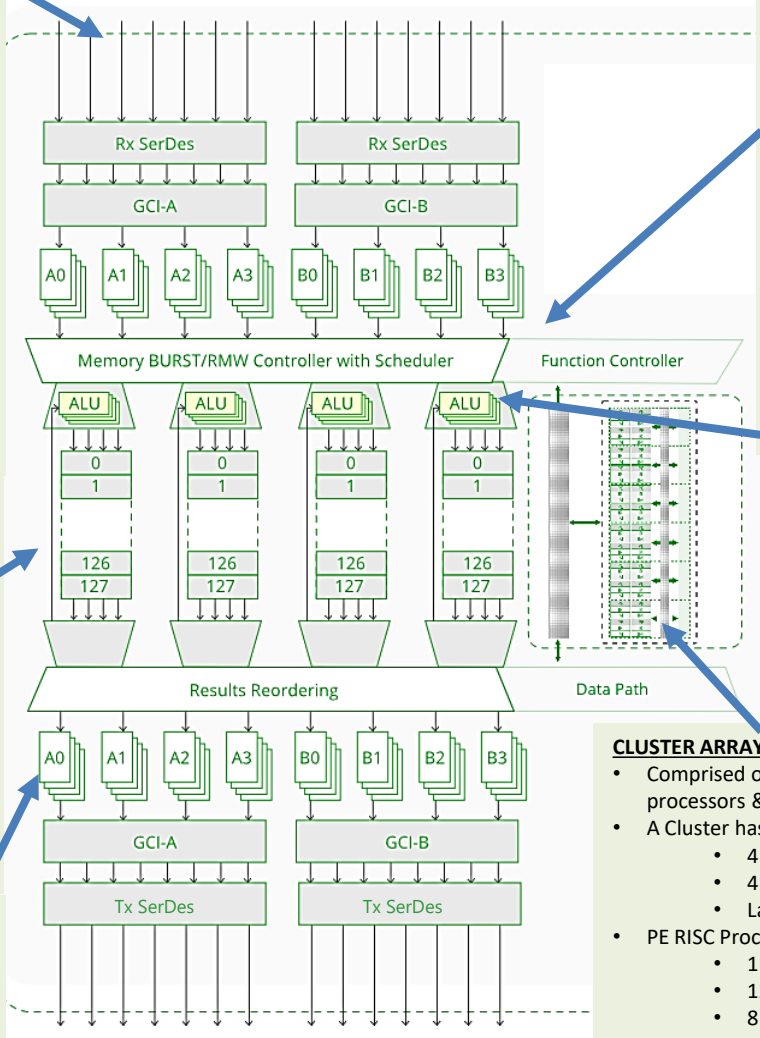
- GCI serial I/O versions of 10, 12.5, 15 and 28 Gbps for high bandwidth (up to 400Gbps)
- Device can operate with a minimum of 4 lanes.
- Has two, full duplex 8 lane ports that operate independently. Device can be used as a dual port memory
- Reduces number of signal pins over traditional memories, increases signal integrity allowing longer board traces to ease board signal routing

Level 1 Main Memory

- 512Mb (BE2 device ONLY)
  - 4 partitions/64 banks
  - 8 READ & 8 WRITE ports
- 1Gb (BE3/PHE Devices)
  - 4 partitions/128 banks
  - 16 READ & 16 WRITE ports
- From 0.67 to 3.2 ns tRC
- Allows parallel partition & Bank execution
- BE3/PHE: up to 6B rd/s + 6B wr/s simultaneously

Result Reordering

- Reorder buffers insure that results are returned in same order as commands are received in the correct scheduling domain



Memory/Function Controller

- Resolves localized bank conflicts
- Multiple scheduling domains minimize blocking short latency operation by long latency operations
- Directs read/write function execution to selected bank of memory
- Manages the sequence of operations to execute a RMW
  - 4-8x reduction in RMW accesses
- Directs User Defined Function execution
- Controls parallel function execution

ALU (BE2/3-RMW ONLY)

- Embedded RMW Functions utilize ALUs for in memory computational functions
  - Depending on the RMW device, there are 8 or 32 ALUs
  - Simultaneous operations

CLUSTER ARRAY (PHE ONLY)

- Comprised of 8 Clusters (total of 32 PE RISC processors & memory with 12Bt/s)
- A Cluster has
  - 4 PE RISC processors
  - 4k x 72b memory @ 1.5Bt/s
  - Latency 6ns
- PE RISC Processor, 1.5 GHz
  - 1k x 72b Instruction memory
  - 128 x72b Internal Registers
  - 8 threads (total of 256 threads managed by PHE device)
- User Defined In Memory Compute algorithms or Bandwidth Functions (IMF)

SOFTWARE DEFINED…HARDWARE ACCELERATED

SOFTWARE DEFINES PERFORMANCE OPTIONS

Each version of the Blazar family of accelerator engines offer many different speed options to the software architect to *Define where in memory various elements of the software should reside based on how best to accelerate system performance*.

- Data tables can use the 1Gb of high speed memory
- BURST data read/writes and computational R/M/W could best be executed in the engine by a single function call
- Using the 32 PE RISC processors, special algorithms can be embedded in the Cluster Array for faster execution, or several copies of the algorithm for parallel processing can be performed
- Simultaneous function execution in multiple partitions and/or banks of the main memory. Simultaneously, user defined functions can be executing in each PE within its cluster memory. And, if needed, function execution priorities can be defined with Domain Priority feature on the BE3 and PHE.

HARDWARE ACCELERATES THE EXECUTION

- Serial High speed I/O up to 28 Gbps provides high bandwidth and low pin count, simplifying the hardware design
- High speed (tRC=0,67 to 3.3ns), parallel capable memory
- 32 PE RISC processors running at 1.5 Ghz, 8 threads per PE, 256 total threads per device

WHAT ARE IN MEMORY FUNCTIONS (IMF)

MoSys makes available a new class of memory called **EFAM** (**E**mbedded In-Memory **F**unction **A**ccelerator **M**emory).

Acceleration is achieved by embedding two types of in-memory functions (IMF): Bandwidth and Compute Functions that execute much faster in-memory than could be executed outside of the memory.

IMF lowers latency up to 4X and increases available access rates 6X by avoiding memory bottlenecks.

Understanding the speed benefits of the embedded functions, combined with speed and parallel capability of the memory architecture, gives the Hardware/Software architect speed, latency and performance options to maximize performance.

MoSys implementation of IMFs

- Bandwidth - Fixed BURST Functions
- Compute - Fixed RMW Functions
- Programmable - User Defined Bandwidth and Compute (or algorithmic) Functions

Typical Applications

- Networking Search
- Statistics
- Buffering
- Security/Firewall
- Anomaly Detection
- Big Data Analysis
- 8K Video
- Genomics
- Graph/tree/list walking
- Monitoring
- High Speed Data Collection
- QDR/RLDRAM replacement

Fixed In Memory Bandwidth Functions (BE2, BE3, PHE)

The BURST Functions are focused on DATA MOVEMENT where increased bandwidth throughput is paramount. The separate Blazar I/O busses and architecture supports full duplex simultaneous read and writes.

The BURST Multi-Read/Multi-Write In-Memory Functions can combine up to 8 READS or 8 WRITES (x72b) into a single BURST function increasing data transfers per command. This reduces the command and address overhead, nearly *doubling the amount of data that can be moved with that same bandwidth*.

For example, buffering where transactions are larger than 72b

Fixed In Memory Compute Functions (BE2, BE3, PHE)

The RMW Functions are focused on accelerating atomic DATA MODIFICATION where coherent updates are needed by reducing the I/O overhead and insuring correct results with a short ECC protected pipeline utilizing data forwarding.

Provides an alternative to a long pipeline which sends one command to READ a memory location, a second operation to MODIFY the value, and a third command to WRITE the new value back to the memory location.

Saves time not having to move data in/out of memory and external modification time. RMWs are atomic, thus ensuring correct multi-threaded behavior.

I/O reduction up to 7X and avoids stale data problems.

Programable Function (Only PHE)

Each group of IMFs delivers different increments of performance acceleration. The PHE, with the User Defined Embedded in memory computing Functions, allows a design to achieve what we call *HyperSpeed.*

Using the 32 PE RISC processors, you can move functions into the Engine that are high use and time consuming, or specialize computation or matching algorithms that may not only be time consuming, but require considerable RTL resources. And, by *putting multiple copies of a function into the Cluster Array, significant performance is achieved* through the power of parallel processing.

Or, simply include functions/algorithms that consume a lot of FPGA RTL resources. *NOW, you can DO MORE in the FPGA!*

Accelerator Engine's Memory Architecture and In-Memory Functions have been used to significantly improve FPGA system performance and simplify hardware RTL and board signal routing while improving signal integrity.

Software Define - Hardware Accelerated

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

The different Accelerator Engine devices allow application tuning to achieve increasing levels of performance up to our most powerful engine… the Programable HyperSpeed Engine with 32 Processor Cores.

LEARN MORE:
www.mosys.com
https://mosys.com/blazar-family-of-accelerator-engines/

	Part Number	Description	Package	Interface					Memory		Access Rate	Commands /Functions		
			Pkg Size	Lanes	Rate per Lane			BW	tRC	Size	Billion			Custom
			mm	Tx/Rx	10-12.5G	15G	25-28G	Gb	ns	Gb	Transaction/s	R/W	RMW / ALU	32 RISC Cores
BURST	MSR620	Bandwidth Engine 2 Burst Serial 0.5Gb High Access Memory	FCBGA 19x19	16	✓			320	3.2	0.5	3.3	✓		
BURST	MSR630	Bandwidth Engine 3 Burst Serial 1Gb High Access Memory	FCBGA 27x27	16	✓	✓	✓	717	2.7	1	6.5	✓		
RMW	MSR820	Bandwidth Engine 2 RMW Serial 0.5Gb High Access Memory with ALU for RMW functiions	FCBGA 19x19	16	✓			320	3.2	0.5	3.3	✓	✓	
RMW	MSR830	Bandwidth Engine 3 RMW Serial 1Gb High Access Memory with ALU for RMW functiions	FCBGA 27x27	16	✓	✓	✓	717	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	✓	✓	✓

ACCELERATING FPGA APPLICATIONS with the BLAZAR Accelerator Engines

With more and more demand on performance and the increasing need for additional features, an FPGA's resources can easily be consumed.

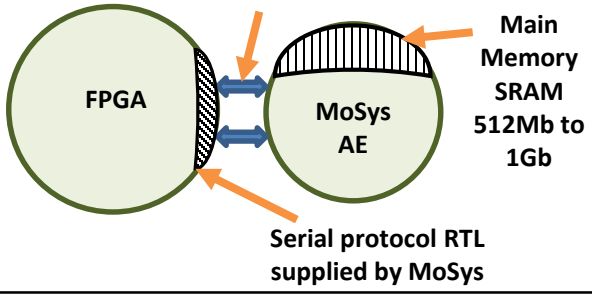
The Blazar accelerator gives the hardware/software architects many options to speed application performance or create unique system architectures that execute faster.

Options to increase the performance of an FPGA design with BLAZAR Engines:

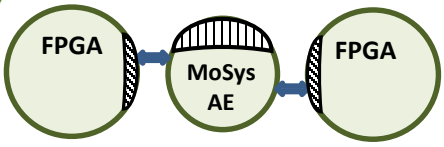
- Offload as many functions as possible
- Simplify software by moving complex functions or algorithms that consume execution time and RTL
- Take advantage of the High Speed Serial Interface to the high random access rate 1Gb memory (tRC 2.7-3.2ns device dependent)

REPLACE UP TO 8 QDR/RLDRAMs (BE2/BE3/PHE)

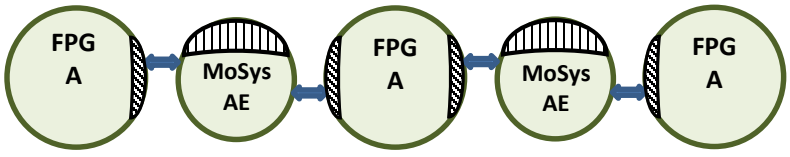
Two serial 8 lane independent I/O ports



DUAL PORT MEMORY (BE3/PHE)

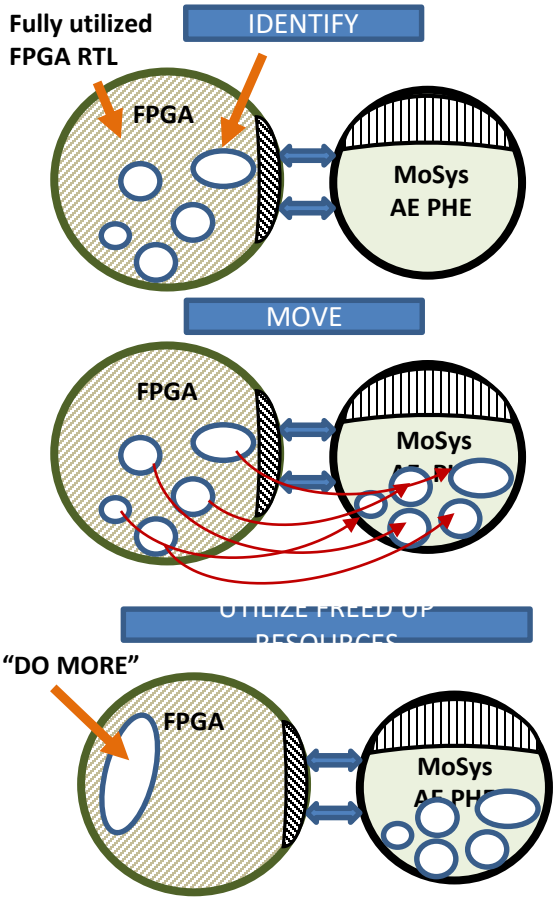


PIPELINE MEMORY (BE3/PHE)



MOVE FUNCTIONS/ALGORITHMS INTO THE PHE *AND* FREE UP RESOUCES TO "DO MORE IN THE FPGA"

"Software Defined…
Hardware Accelerated"



Many functions and algorithms can be moved into the PHE (Programmable HyperSpeed Engine) to *Achieve HyperSpeed by*

- Increasing performance
  - Will execute faster "In-Memory"
  - Take advantage of the parallel processing
- Freeing up space in the FPGA to allow More Features….*DO MORE*
  - Complex functions that are difficult to program in RTL
  - Save RTL space

Simplify software using the fixed BURST and RMW functions

- BURST READ or WRITE of multiple locations on single function call
- READ/MODIFY/WRITE (RMW) on a single function call
  - Statistical/counters
  - Atomic operations can be assured

FPGA tasks that execute faster In-Memory of the PHE using the 32 RISC processors

- Simplify
  - Provides flexibility for the System Architect to sort tasks between hardware RTL and software tasks
  - High usage RTL moved into the PHE
- Complex algorithms/functions that benefit from Software Defined approach
  - Associative look up functions like hash tables, flow caches, TCAM
  - Special Network functions like: L2 forwarding learning DB, L3 Longest Prefix match, ACL
  - Data analysis and monitoring function histograms
  - Data structure access function for graphs, radix trees, Mtries, Patricia trees or linked lists
  - Machine Learning functions like Random Forests of Trees

General tasks

- Need more memory than on-chip can support
- Memory access bottlenecked tasks
- Time consuming memory sequential tasks
- Time & sequence critical tasks

- Speed increase using
  - Parallel processing (32 processors) /optimized ISA & 256 threads
  - Higher access rates, lower latency, higher bandwidth
  - Utilize Engine scheduler to optimize execution
    - Install multiple copies of same algorithm/functions and scheduler will find available processors

Info: www.mosys.com Email: Sales@mosys.com