MoSys

Solution Notes and Technical Tips

Michael Miller
MoSys CTO

SOLUTION NOTE #101 Buffering up to 800 Gbps throughput with Bandwidth Engine Memory

Overcoming HBM and QDR performance and design limitations. Achieving Bandwidth at >100Gbps can be a challenge. Buffering up to 400G full duplex requires 800 Gbps of throughput (Read + Write).

Software and Hardware memory tradeoffs determine performance in applications such as packet capture, over subscription buffer, rate matching, stream merging, parking packet contents while doing header processing, managing data between two mismatched data rates or just simply buffering 8K video streams.

In the newer FPGA offerings, the solution is directly attached HBMs. Although this is a good solution for bulk storage, it has all the historical DRAM issues of long random access latency, little flexibility in designing the density, more complex thermal considerations and can be very costly.

Using external memories like SRAM, GDDR, DRAM or RLDRAM are fast, but the number of device and pin counts, layout of balanced traces and board space can become a real challenge.

At high rates the designer needs to be concerned with “bandwidth density”

- The number of devices
- Pin I/O efficiency
- The numbers of pins
- Board layout complexity
- Signal integrity
- Power

Where more than sequential access is needed, the designer will want to consider

- Random access cycle rates
- The random-access rate of a
    - DRAM based solution will be limited to tRC of ~45ns to 50ns.
    - SRAM speed may be overkill and may require many devices and still not be large enough

However, HBM and QDR are not the only solutions.

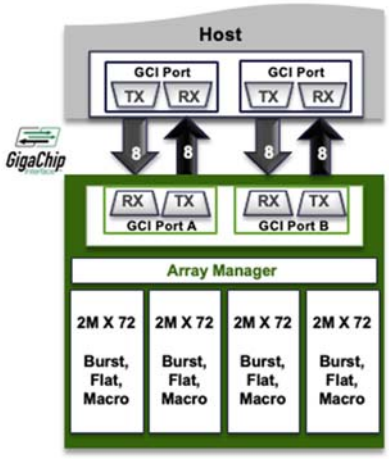
Alternative memory solution, MoSys’ single chip Accelerator Engines are serially attached memories with

- 576MB and 1Gb memory capacity,
    - tRC 3.2ns
- SerDes attached (up to 16 lanes a) at
    - 10-28Gbps
- Ease implementation
    - Less board space
    - Less signals to route
    - Solid signal integrity
    - Less heat sinks

    - Quicker board designs Full
- Embedded Intelligent memory acceleration functions
- Achieves 800 Gbps BANDWIDTH

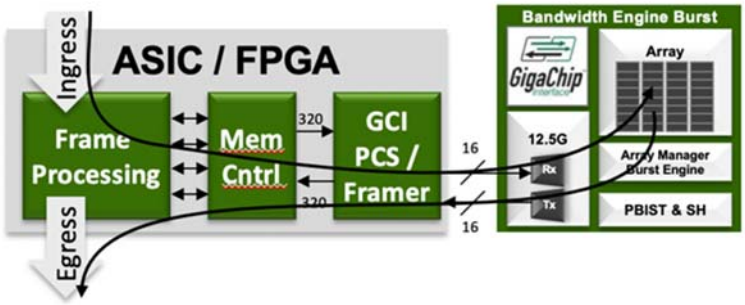
Basic Bandwidth Engine (BE) Architecture

The basic MoSys Bandwidth Engine architecture consists of a monolithic memory device with 16 SerDes lanes that can operate from 10Gbps to 28Gbps. The memory is arranged in 4 partitions 72b wide and the total density can be 576Mb or 1152Mb. The 72b width allows for optional ECC to be stored. All commands, address and data are transmitted between the host and the BE over the SerDes interface using a protocol called GigaChip Interface (GCI). GCI provides a reliable transport using CRC to detect errors and trigger a retransmit if required. Typical bit error rates are 10^{-18} for chip to chip.



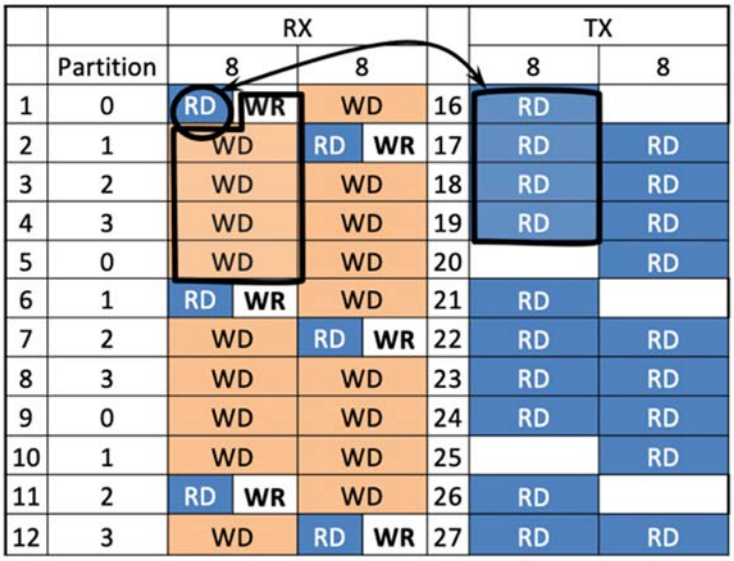
Bandwidth Density

When buffering data at high rates, the effective bandwidth of the interface tends to drive the number of devices that are needed to achieve the total throughput. Often designers find themselves in places where the resulting memory capacity is not required which results in unneeded resources. Therefore “protocol” or I/O efficiency to move data is key. With Bandwidth Engine and the GigaChip Interface there are 16 SerDes lanes which can be viewed as 16 Tx / Rx pairs that form two unidirectional buses to move write data and read data.



When using SerDes as a serial interface to a memory, the command (read or write) and the address must be sent with the write data, therefore the ratio of command overhead to write data becomes highly important. To this end, BE devices support a burst command for both read and write which maximize the amount of data transferred per transaction. In the diagram below, the “RX” and “TX” buses (SerDes lanes) are divided into two groups of 8 which are the two GCI links. Each link transmits a series of 80b bit frames with 72b payloads. The frames received on the RX “pins”, are either data or command frames, Command frames have 2 command slots which designate reads and writes. In the BURST format each command can request 2, 4 or 8 frames of data to be transferred, thus achieving up to 89% efficiency of the potential 72b frame payload bandwidth. Since the 72b payloads are in 80b frames, BE can achieve 80% efficiency of the raw bandwidth using burst of 8.

All the necessary FPGA RTL for the GigaChip protocol is supplied at no cost.



		RX			TX	
	Partition	8	8		8	8
1	0	RD WR	WD	16	RD	
2	1	WD	RD WR	17	RD	RD
3	2	WD	WD	18	RD	RD
4	3	WD	WD	19	RD	RD
5	0	WD	WD	20		RD
6	1	RD WR	WD	21	RD	
7	2	WD	RD WR	22	RD	RD
8	3	WD	WD	23	RD	RD
9	0	WD	WD	24	RD	RD
10	1	WD	WD	25		RD
11	2	RD WR	WD	26	RD	
12	3	WD	RD WR	27	RD	RD

On BE devices, the BURST command can be interspersed with single word transfers as needed. This gives the designer the ability to burst in large amounts of data and then do smaller random-access out of order transfers. This can be useful in storing frames of video or network packets and then accessing sub segments like header fields or sub images in the overall larger frame. In addition, the MoSys MSR820 and MSR830 devices implement RMW commands for in-memory-computation such that BURST commands can be mixed with RMW commands.

Comparing with Other Solutions:

Once the efficiency of the protocol (PCS/Framer) is understood and the memory transaction is known (command plus address) one can compute the actual data transfer rate of the BE device for a given lane count, burst size and SerDes bit rate.

For Full Duplex (read + write) buffer bandwidth, MSR620 and MSR630 Bandwidth Engines can be compared to alternative networking memories.

Throughput (Gbps)		Speed Grade				
		620-10	620-12	630-15	630-25	630-28
Width	Burst	10.3125G	12.5G	15G	25G	28.1G
16 lane	BL8	132.0	**160.0**	**192.0**	**320**	**360**
	BL4	118.8	144.0	172.8	288	324
	BL2	99.0	**120.0**	**144.0**	**240**	**270**
8 lane	BL8	66.0	**80.0**	**96.0**	**160**	**180**
	BL4	59.4	72.0	86.4	144	162
	BL2	49.5	60.0	72.0	120	135
4 lane	BL8	33.0	**40.0**	**48.0**	**80**	**90**
	BL4	29.7	36.0	43.2	72	81
	BL2	24.8	30.0	36.0	60	67.5

Sigma Quad IVe BL4 is:
93 Gbps Full Duplex
(192 Gbps I/O throughput)
8 x 16Mb single ported banks
~4.5W (device + I/O)

QDR IV XP is:
76.5 Gbps Full Duplex
(153 Gbps I/O throughput)
8 x 16Mb single ported banks
~7W (device + I/O)

RLDRAM 3:
33 Gbps Full Duplex
(76.8 Gbps I/O throughput)
8 x 16Mb single ported banks
~3.5W (device + I/O)

As can be seen from the table, a single BE device can replace multiple other memory devices for a given required buffer bandwidth resulting in a much lower I/O pin count: 16 SerDes lanes vs 100s of pins, less board area, easier board layout and signal integrity, all for about the same or slightly less power.

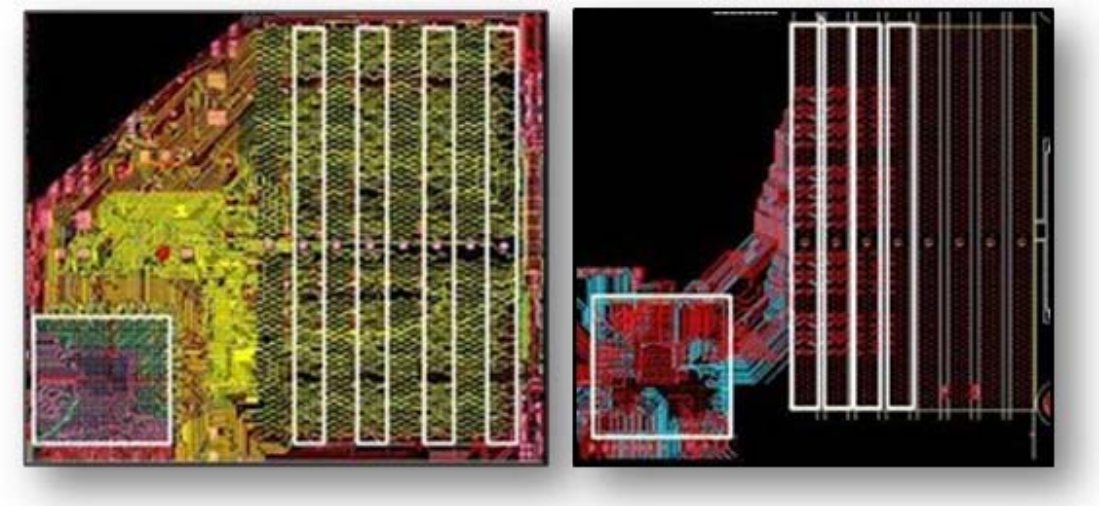
To go beyond a 360Gbps full duplex single chip buffer, a Programmable HyperSpeed Engine (PHE) can achieve 400Gbps+. The PHE can be programmed to reach higher efficiencies with longer bursts such as 32 words. In this case, the host device (FPGA or ASIC) will most likely keep head and tail buffers to assemble transfers into the large burst needed. This will be more fully described in a subsequent Solutions Notes detailing the use of PHE in high bandwidth applications.

Easing Layout, board area and layers

When laying out boards with classic devices like QDR SRAM, DRAM, GDDR or RLRAM at high rates, the board layout designer must be highly aware of the trace routing of data lines and control signals in order carefully match the length of traces which results in lane squiggling across the 64b or 72b of the data buses. Laying out 8 to 16 parallel I/O memory devices can be a challenge to minimize board area. Board area can be minimized with more layers, but layers at to the BOM cost.

The GCI interface and protocol includes features to further make layout easier compared to high frequency parallel buses. When devices are on both sides of the board or highly packed, board layout designers have to go to great lengths to route all the pins in the limited number of layers allowed, thus adding considerable time and expense to board design. When the BE device is reset, the SerDes and GCI train between the host device and the BE device to learn about channel characteristics, individual lane delays and lane ordering. The GCI interface of BE devices has automatic lane deskew which allows the designer to lay out Tx or Rx differential pairs with the shortest, most convenient routing and lets the BE match them up at link training time. In addition, the GCI protocol will also descramble and internally swizzle the lanes back into proper order. This allows a designer to easily layout topologies which have devices on either side of the board and/or route around obstacles which would normally require more layers to maintain strict lane ordering.

The following images show the layout of DRAM DIMM memory with matched trace lengths using squiggling vs the same effective throughput using SerDes traces and lane de-skewing.



Summary

The BE device can replace multiple QDR, GDDR, DRAM or RLDRAM devices with a much more relaxed board layout and reduced board area while reducing time to market and BOM costs. The BURST command format of BE can be mixed with other commands such as single word transfers or RMW commands, thus providing a powerful system solution which can mix the random-access rate of single transfers with the bandwidth efficiency of burst commands.

FUTURE SOLUTION NOTES...

-Is HBM Overkill?
-What is an FPGA Accelerator Engine?
-Are Dual Port Memories Dead?
-Managing Coherent or Stale data
-QDR SRAM Alternative...4-8x Capacity in a Single Chip
And more!

Visit the SOLUTION NOTE LIBRARY at https://www.mosys.com/solution-notes/

MoSys
2309 Bering Drive
San Jose, CA 95131
Tel: 408-418-7500
Fax: 408-418-7501