



Solution Notes and Technical Tips



Michael Miller
MoSys CTO

SOLUTION NOTE #1005: QUAZAR QPR Features and Benefits
TOPIC SUMMARY:

TABLE OF CONTENTS

Overview

Architecture Functional Overview

- QPR Operational Advantages
- Memory Modes of Operation
- QPR vs QDR

Total Memory Solution

Architecture Technical Features

- Overview
- Driver side
- Power consumption
- Power vs Speed
- Die size
- Quad Partition Rate
- FPGA Interface
- Memory controller
- Cost

Summary

OVERVIEW

In reviewing current QDR devices that are presently on the market, the issues designers face, include:

- Density of 144Mb with one offering at 288Mb
- Use of wide parallel buses that run at very high frequency
- Strict rules to layout boards to accept these fast-wide buses
- Sourced by multiple vendors (Cypress (Infineon) and GSI) but each vendor uses a slightly different pinout
- No commitment to a future roadmap

The QUAZAR (QPR) devices are members of the second-generation Bandwidth Engine family of high-capacity, high-bandwidth, low-latency memory devices from MoSys, Inc. MoSys uses its robust and proven 1T-SRAM memory technology in QPR (Quad Partition Rate) devices to provide higher density and better reliability than SRAMs and lower latency than bulk DRAMs.

The QPR devices overcome most of the critical limitation of the QDR architecture, provides devices that replace 4 or 8 QDR on a single device and at significantly lower cost.

ARCHITECTURE FUNCTIONAL OVERVIEW

QPR Operational Advantages:

- Low Cost QDR alternative
- High capacity replaces... 4-8 QDR devices
- Higher bandwidth
- Low tRC
- Significantly Lower power than QDR
- Simplified design effort – only 32 FPGA pins
- System performance equal/better than a QDR
- *Costs significantly less than the equivalent cost of QDR components*
- *32 Pin FPGA I/F take less design/debug time*
- Xilinx & Intel FPGA compatible
- Compatible with other FPGA families
- Pin compatible with MoSys High Bandwidth Accelerator Engines (BE3 & BE3)
- Lower pin count and ease of PCB design
 - Highly efficient serial protocol

- Reduction of I/O pins, 20-40x of a QDR
- *Typical system uses 32 pins*
- Signal Auto-Adaption feature
 - Eases board layout & signal integrity concerns
 - Operates over connectors
- Pin compatible with MoSys higher performance Blazar family of Bandwidth Engine Memories

Memory Modes of Operation

Each Partition operates as an independent Random-Access Memory. All four of the partition are accessed each tRC clock cycle.

Bandwidth is determined by combining access across multiple partitions with word width up to 576b.

A selected MoSys RTL Memory Controller for the QPR4 provides for one of two modes of operation.

- 1 Port A

DEEP Partition Mode

Maximum word width is 288b each tRC

Allows Up to 8 accesses 4 READ & 4 WRITES in one tRC.

- 2 Port A & Port B

WIDE Partition Mode

Maximum word width is 576b

Allows up to 16 accesses 8 READ & 8 WRITES in one tRC.

Random-Access Bandwidth

- Bandwidth

QPR4

Maximum bandwidth is 320 Gb/s (160 Gb/s full duplex)

QPR8

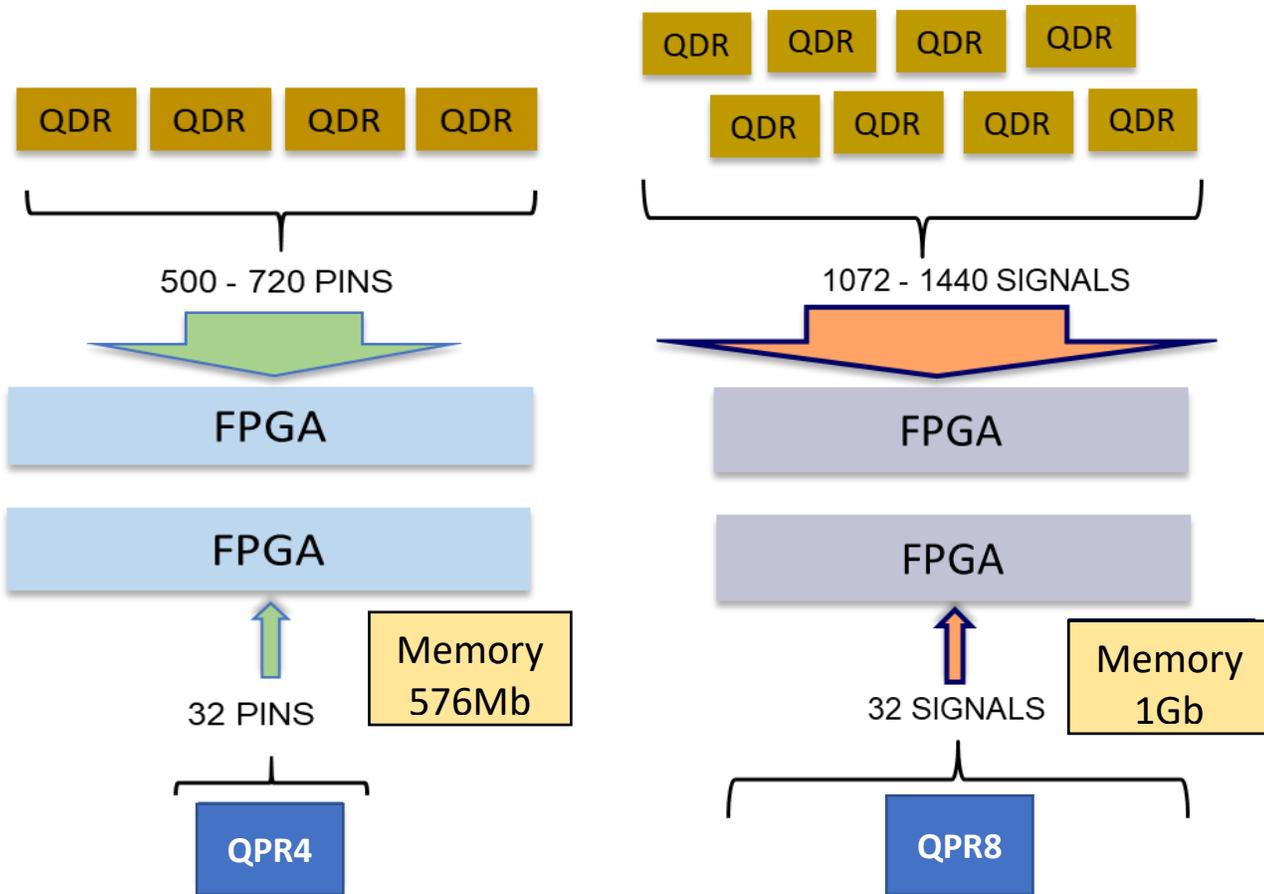
Maximum bandwidth is 640 Gb/s (320 Gb/s full duplex)

The discussion so far has been on bandwidth horizontally across the partitions for wider word width for higher bandwidth in one tRC.

t should be clearly stated that at this high bandwidth, each partition can still address any location within its partition independently of the other partitions accessing any location within its partition in the same tRC cycle.

In effect, true 4 or 8 independent random-access memories.

QPR vs QDR



- Memory size
 - MSQ220: 576Mb capacity equivalent to 4 QDR-144Mb capacity per device
 - MSQ230: 1Gb capacity equivalent to 8 QDR-144Mb capacity per device
- Device PCB board space saving
 - 1 MSQ220 (361 mm²) vs 4 QDR devices ~1000 mm²
 - 1 MSQ230 (729 mm²) vs 8 QDR devices ~2000 mm²
- Signal pin reductions
 - 4 QDRs: 500-720 pins
 - 8 QDRs: 1071-1440 pins
 - Typical MoSys system 32 pins
 - All MoSys devices have Auto-Adaptation which handles on-board signal tuning, eliminating the need for any external components to ensure clean, reliable signals
 - Cost
 - One MSQ220 with 4x the memory capacity is less than the price of 2 QDR memories

TOTAL MEMORY SOLUTION

MoSys memory products can be used with the MoSys FPGA RTL Memory Controller reference design to simplify the design of upgrading from QDR or with a user implemented design either of which will result in higher capacity, with less design effort and minimal impact on software.

Three System Elements

- MoSys supplied RTL Memory Controller

QDR type RTL interface

Provides RTL register set for each partition

Support user selectable word widths up to 576b

Controls the SerDes protocol (transparent to user)

- SerDes high-speed Memory FPGA I/F (GCI)

Handled by the MoSys RTL Memory Controller

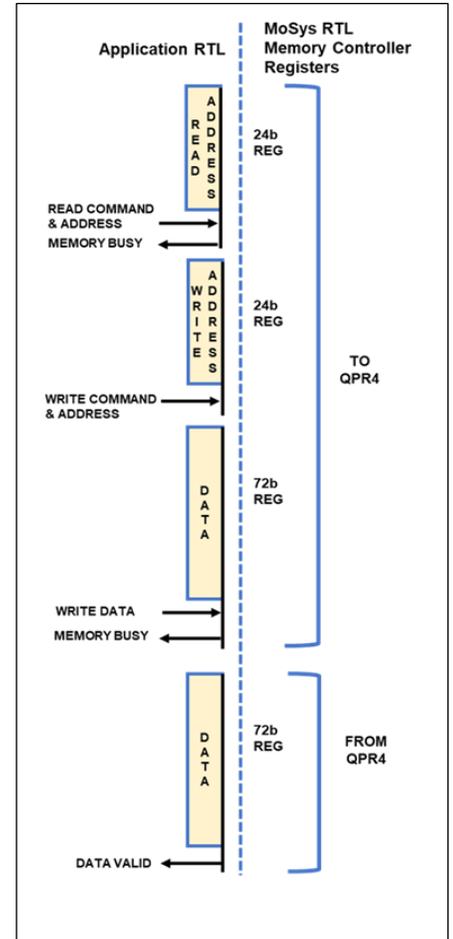
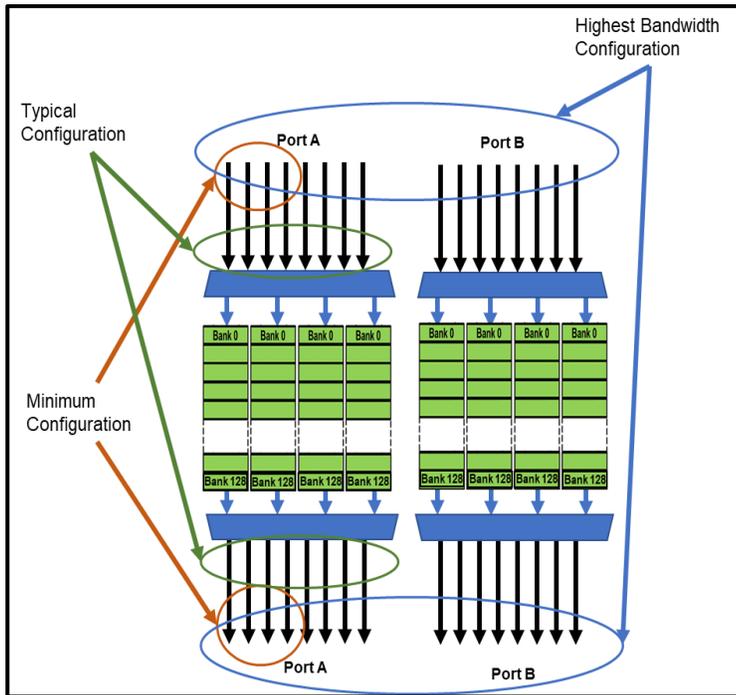
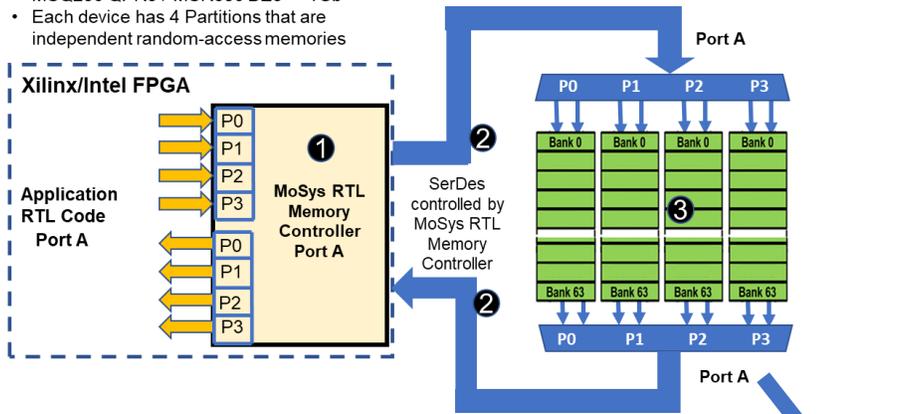
Is the preferred FPGA interface today with the Lowest pin count

Has become the future preferred I/F

Lowest pin count 4/ typical 16/ highest bandwidth 32.

MoSys QPR or Bandwidth Engine Memory

- MSQ220 QPR4 / MSR222 BE2 = 576mB
- MSQ230 QPR8 / MSR630 BE3 = 1Gb
- Each device has 4 Partitions that are independent random-access memories



ARCHITECT TECHNICAL FEATURES

The QPR architecture is a highly parallel, multi-ported memory array coupled with a high efficiency serial interface, delivering the highest access rate and data throughput of any single chip solution on the market. The array architecture enables up to 16 accesses per memory cycle.

The full-duplex SerDes interface, in combination with the Array Manager, allows simultaneous read and write accesses into each memory partition and eliminates the timing penalties associated with bus turnaround, by utilizing separate Rx and Tx busses.

The cell design that MoSys has chosen to use in the Quazar product line uses an embedded DRAM design. The benefits of this design vs either a straight DRAM or SRAM cell is two-fold:

1. By using an E-DRAM cell it is designed using a “logic” process at fabrication facilities like TSMC. By using a logic vs. a DRAM process the cell is slightly larger than a pure DRAM cell would be but the process, as it is named, also allows for integration of large amounts of Logic to surround the memory array (in the case where additional embedded functionality is desired).
2. By using this version of the process, it is also possible to design the array with the desired characteristics to enable very fast access.

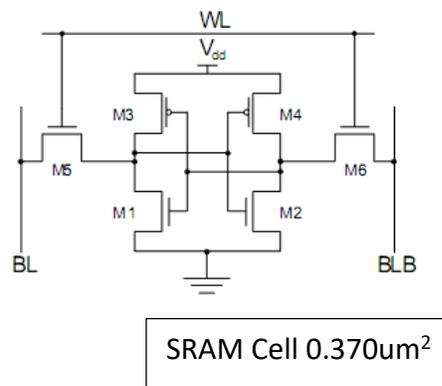
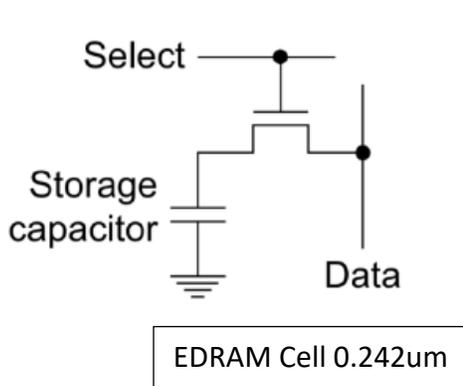
What is meant by this is that to achieve the desired speed (which in the MoSys case is close to SRAM speed) it is necessary to keep bit lines and word lines at a reasonable length. By doing this, the capacitance of these metal lines is kept to a minimum which in turn, defines the size of the line drivers and sense amps that will be needed to drive these lines. In the case of the MoSys devices, we designed the array with lines that had only 144 bits per line vs. the approx. 2000 bits per line that is standard in normal DRAM devices. This has the impact of reducing the load that both needs to be driven by line drivers and by the cells themselves when they are activated.

Reduced driver size. When the line drivers do not have to drive a large capacitive load the sizing of the drivers can be reduced. This has a positive effect in that the drivers can be smaller

Power is also impacted. When reviewing the power impact of utilizing E-DRAM' any one access will require less power than an access of a larger number of bits. The equation of $P=CV^2F$ is directly applicable. *Since the C factor is reduced on any one access, to drive shorter bit and word lines, the resultant power dissipation is reduced.*

Power vs speed. As mentioned in the previous paragraph, by reduction of the capacitive loads that need to be driven. the individual power of a single access is reduced., By reducing the capacitance, the resulting speed is increased by being able to drive and recover the lines faster. This allows the array to run faster, which results in a higher bandwidth device. Even though the impact of this is a slightly higher power dissipation in that the same $P=CV^2F$ equation now increases, *the result is still a reduction in power over an equivalent density of QDR devices.*

Cell Size. An E-DRAM cell in 40nm TSMC process each takes approx. $0.242\mu\text{m}^2$. Whereas a single SRAM cell in a comparable technology takes approximately $0.370\mu\text{m}^2$, which is approximately 53% larger than the E-DRAM cell. When one places 576 Mb of these cells on a die it has a very large impact on overall die size. (This is the major reason why even with newer technologies, the size of an SRAM cell will be the limiting factor in the density of the available arrays. The densest SRAM is 288Mb and DRAM are in the Gb and higher density).



The result of using this cell structure to be the basis of the MoSys Quazar MSQ220 (576Mb) and MSQ230 (1.1Gb) devices, is that one can achieve the density of 2X to 8X that of available SRAM devices at a comparable technology node. *The device can run at comparable system speeds and achieve a lower power consumption than an equivalent SRAM while still saving the user money.*

Quad Partition

The next innovation in the MoSys memory architecture is the division of the memory array into Quad Partitions. A partition is 2M x 72 (144Mb) in the case of the MSQ220 (the equivalent of 1 QDR-144Mb), or 4M x72 (288Mb) in the case of the MSQ230 (the equivalent of 2 QDR-144Mb). In total memory capacity, since the MoSys device has 4 partitions, the MSQ220 is equivalent to 4 QDR devices in one package and the MSQ230 is equivalent to 8 QDR devices in one package.

In addition to the density benefit, the bus structure allows each partition to be accessed as a fully independent memory structure or as part of a unified memory. This enables the user to access 4 independent 72-bit words, one in each of the 4 partitions, from each of the GCI ports. This equates to 288 bits of data from each of the ports or potentially a total of 576 bits of data during each of the associated FPGA clock cycles. As with all previous access patterns it remains important that each of the accesses into any one partition be to a separate bank within the partition. This is because even though a partition can support multiple reads and multiple writes (simultaneously) any individual bank is limited to a single access at a time (banks are single ported, limiting it to a read or write per cycle.)

FPGA Device Interface

QPR devices, like other MoSys Bandwidth Engine devices, employ the GigaChip Interface (GCI) instead of a parallel LVCMOS memory interface. The GigaChip Interface uses high-speed serial signaling and an acknowledgment/replay protocol to provide reliable point-to-point communication of fixed-size frames over short distances.

Use of one or two GCI ports as follows:

- Each GCI port uses 4 TX and 4 RX lanes.
- Each GCI port uses 8 TX and 8 RX lanes

- The lanes have a bit rate of between 10.3125Gbps up to 25 Gbps, depending on the speed grade of the device.

The most full-featured configuration used 2 GCI ports with 8 TX lanes and 8 RX lanes per port. A GCI port transfers 80-bit frames in both directions, striped over 8 RX lanes and 8 TX lanes. The electrical and timing specifications of the serial lanes are compatible with the XFI standard and CEI-11G-SR standard. The Array Manager steers commands and data between the GCI ports and the memory partitions, and detects errors, such as illegal commands and memory bank conflicts.

GigaChip™ Interface (GCI) standard

- 10.3125 or 25 Gbps SerDes Interface
- 90% efficient transport protocol
- 80 bit frames
- PRBS-48 scrambling
- CRC error detection
- Automatic Error Recovery
- One or two GCI ports
- Up to 8 transceivers (TX/RX) per port
- Lane configuration can be modified
- Mesochronous clocking for low latency
- XFI & CEI-11G+ compatible electrical IO

Device Performance:

- Deterministic read latency

Memory Controller

The Memory Controller in the host device (ASIC, ASSP, or FPGA) connects to the Quad Partition Rate (QPR) device through one or two GigaChip Interface ports. The memory controller presents a SRAM like RTL interface and manages the GCI signals to the QPR device so the GCI signals are transparent to the user.

Cost

These new chips are targeted at accelerating Intel and Xilinx FPGA designs and offer similar performance to traditional QDR SRAMs at a significant price advantage. The devices come

in two capacities, 576Mb or 1.1Gb, so you can replace 4 to 8X QDR devices in a single monolithic package. The parts are easy to design in and have many benefits over traditional QDR SRAMs (which are stated below). If you are considering QDR SRAMs, you should consider these Quazar parts. Think of getting 4 to 8 times as many QDR SRAMs in a single package and saving lots of money while doing so. The lowest cost version in the family sells for less than \$200 in volume quantities.

SUMMARY

The Quazar family of low cost QPR SRAMs are a great fit for an even broader range of applications than MoSys' traditional higher end Intelligent Memory ICs and our recently announced Packet Classification IP portfolio address. Whereas those MoSys solutions address traditional networking, security, video, search, TCAM, Aerospace & Defense, Test & Measurement, Cloud Infrastructure markets, the new Quazar QPR SRAMs expand MoSys into even more markets like industrial, appliances, IoT and others.

MoSys is always interested in how you found the ideas presented in this solution note, so any feedback would be greatly appreciated and will support us in what future topics and data will be addressed in future solution notes.

If you need to free up resources on your FPGA or would like more flexibility in your FPGA part selection options, please contact MoSys and we can do a memory architecture design tradeoff review with you. Contact: AppContact@Mosys.com for a memory architecture discussion! [Email us](#) and we will arrange to have one of our technical specialists speak with you. You can also sign up for our [newsletter](#). Already convinced? You can request a quote from [sales](#). Finally, please follow us on social media so we can keep in touch.

