CMP United Business Media
NETWORK SYSTEMS DesignLine
techonline Community

QDR SRAM and RLDRAM: A Comparative Analysis

By Anuj Chakrapani, Cypress Semiconductor Corp.

Abstract

Today's high-speed networking applications require high-bandwidth and high-density memory solutions. For instance, typical networking line cards need memories for a variety of operations that include packet buffering, table lookup, and queue management among a host of other functions. Choosing the right memory solution is pivotal to ensuring that the memory bandwidth does not become a bottleneck on the throughput of the application. This white paper discusses memory solutions suitable for networking applications—specifically, Quad Data Rate Static RAM (QDR SRAMs) and Reduced Latency Dynamic RAM (RLDRAM)—and compares them in relation to the applications they are best suited for.

The Evolution of Networking SRAMs

Standard synchronous SRAMs, the earliest mainstream synchronous SRAMs, were ideal for cache applications. However, despite their extensive use, they weren't suited for networking applications that dictated a balanced READ/WRITE profile. A READ operation followed immediately by a WRITE operation would result in a contentious state on the data bus. The only workaround for the bus contention was to introduce "wait" or "no operation" (NOP) cycles to accommodate for bus turnaround. But these "wait" cycles affected the bus utilization and therefore resulted in underutilization of bandwidth. Because bandwidth utilization is a key factor, these synchronous SRAMs were not ideally suited for such networking applications.

To address the bus contention problem, No Bus Latency (NoBL), also known as Zero Bus Turnaround (ZBT), SRAMs were developed. These SRAMs contained data registers in the periphery to pipeline the READ and WRITE operations, thereby eliminating the "wait" cycles and achieving peak bus utilization. However, with line rates reaching the order of tens of gigabits per second, speed-, bandwidth- and interface-related bottlenecks had to be addressed. Several applications had emerged that not only demanded faster operation, but also needed simultaneous READ and WRITE operations to the memory. Although originally well suited for networking architectures, the NoBL SRAMs were unable to keep up with their performance requirements. Hence, the latest generation of networking memories—the QDR/DDR family of SRAMs—was developed to meet the speed, density, and bandwidth requirements of today's networking applications.

The QDR/DDR Family of SRAMs

QDR and QDR-II SRAMs, the latest generation of synchronous SRAMs, were developed by the companies that make up the QDR consortium (Cypress, Renesas, IDT, NEC, and Samsung). This family of network SRAMs, along with Double Data Rate (DDR) and DDR-II SRAMs, provides complete memory solutions for any networking system.

QDR and QDR-II SRAMs come in speeds up to 300 MHz and beyond and densities of 9 Mb to 72 Mb, with capability of future expansion up to 288 Mb and beyond. QDR and QDR-II SRAMs have separate ports for both READ and WRITE operations, which eliminates bus contention. The double data rate interface on these ports—data is written to or read from the SRAM on both edges of the clock— essentially doubles the bandwidth of each pin when compared to other SRAMs. The combination of having separate input and output ports and DDR interfaces on these ports provides a four-fold increase in overall bandwidth compared to earlier synchronous SRAMs.

DDR and DDR-II SRAMs belong to the same family as QDR SRAMs. They are similar to the QDR and QDR-II SRAMs, with the major difference being that DDR and DDR-II SRAMs do not have separate read and write ports. While QDR SRAMs can perform both READ and WRITE operations simultaneously, the DDR devices can perform only READ or WRITE, but not both at a given time.

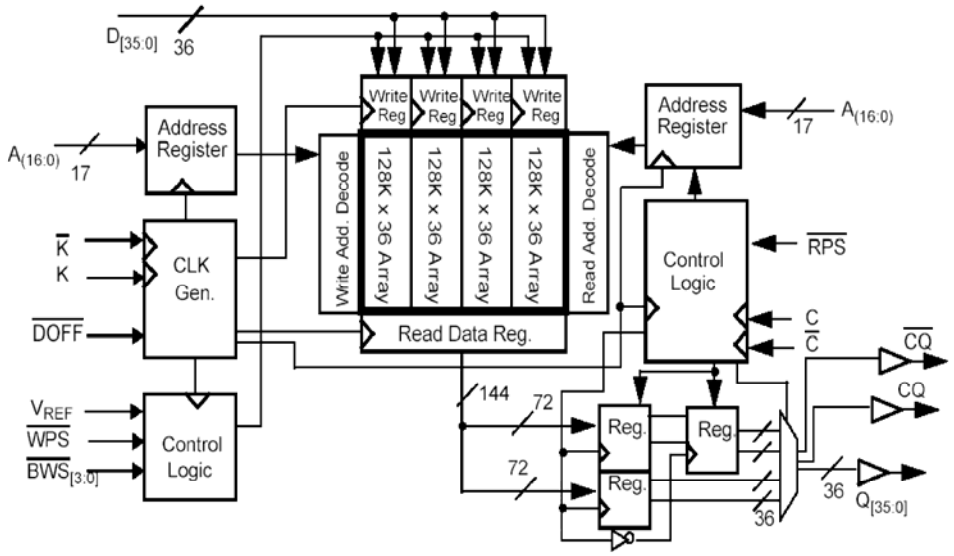


Fig. 1 Block diagram of a QDR-II burst-of-4 device

Several other features make the QDR family of SRAMs ideal for high-speed networking applications:

- **Output Clocks**: In addition to the input clocks K and K#, a pair of output data clocks, C and C#, can be used to synchronize data from the SRAM. The use of these output clocks is optional. In the single clock mode option, data is synchronized to the input clocks.
- **Programmable Output Impedance**: The QDR SRAMs are equipped with programmable impedance circuitry that can adjust their output driver strength to match the impedance of the transmission line. Matched impedance improves the signal integrity of the device.
- **Echo Clocks**: These SRAMs generate a pair of output clocks, CQ and CQ#, that closely match the data (edge-aligned with the data). Thus, these serve as output clocks from the SRAM that can be used for latching the output data into the controller. The echo clocks feature is available in QDR-II, DDR, and DDR-II products (not offered in QDR-I).

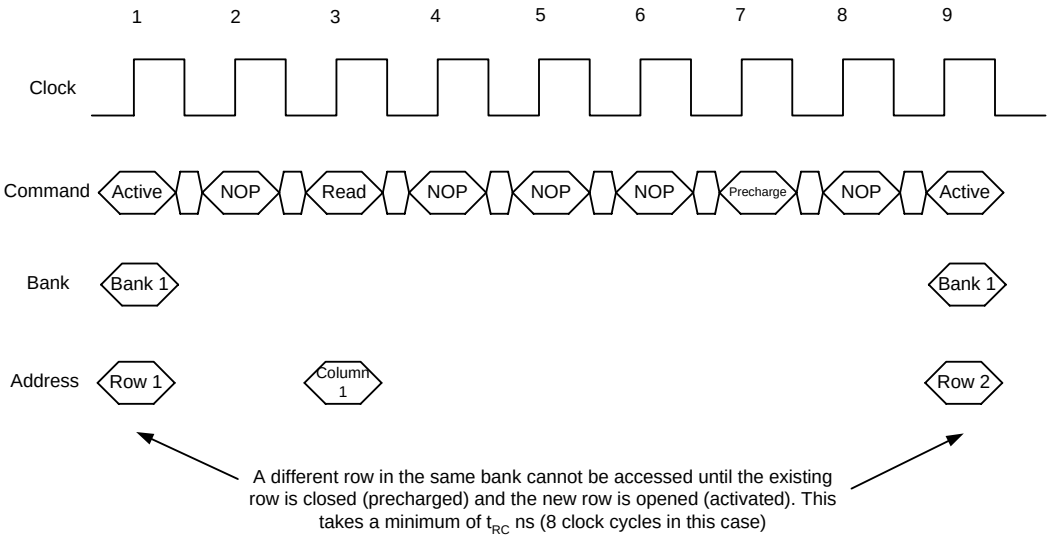


Figure 2. The t_{RC} limitation

The following section discusses the Reduced Latency DRAM (RLDRAM), a family of DRAM memories used in networking applications.

RLDRAM

Typically, the internal memory array in DRAMs is organized as “banks,” with a memory location specified in the form of bank, row, and column addresses. Before accessing a particular row in a bank, the bank (or specifically, the row) has to be opened or “activated” (cycle #1 in Figure 2). Following an access, the row has to be closed or “precharged” before opening another row in the same bank. Thus, between two accesses to different rows of a particular bank, the bank must have been precharged at the end of its previous operation (cycle #7) and the new row activated (cycle #9) before carrying out the next access. During this period, the bank is unavailable for access. The unavailability of a bank in turn limits the frequency at which accesses can be performed to the same bank. The minimum time delay between accesses (or bank activations, as shown in the figure) affects the bandwidth of the DRAM. This latency (8 clock cycles for the example shown in fig. 2), which affects on-the-fly accesses to banks, is called *random cycle time*, *active-to-active command period*, or *same bank latency* and referred to as t_{RC} in datasheet specs.

RLDRAM was designed to address this issue, thereby encroaching on the low-latency, high-bandwidth SRAM market.

The Reduced Latency DRAM (RLDRAM) is a DRAM architecture developed by Micron and Infineon that addresses the t_{RC} limitation with an improved architecture and interface design.

RLDRAM II devices use an eight-bank memory array architecture. DRAMs have traditionally been arranged in four banks, but the eight-bank arrangement in RLDRAMs helps achieve its peak bandwidth, albeit under specific conditions (discussed later). More banks translates to a higher probability that a bank will be available for access—that is, one of the banks may be in a precharged state already. This makes the probability of hitting an available bank higher in RLDRAM II.

Also, the RLDRAM II has an SRAM-like interface that makes it more suitable for networking applications than other DRAMs. The addressing of the device is similar to that of an SRAM—address supplied does not have to be in the form of row address and column address as is the case with standard DRAMs. In typical DRAMs, row activation needs to occur before a column address is provided, thereby making the array access a two-step process. In RLDRAM, with internal precharge and built-in activation, the entire addressing is done in a single cycle, which makes addressing much simpler.

In addition, RLDRAM II is equipped with double data rate interfaces that allow data to be transferred on both the rising and falling edges of the clock, thereby doubling the bandwidth compared to the standard single data rate interface.

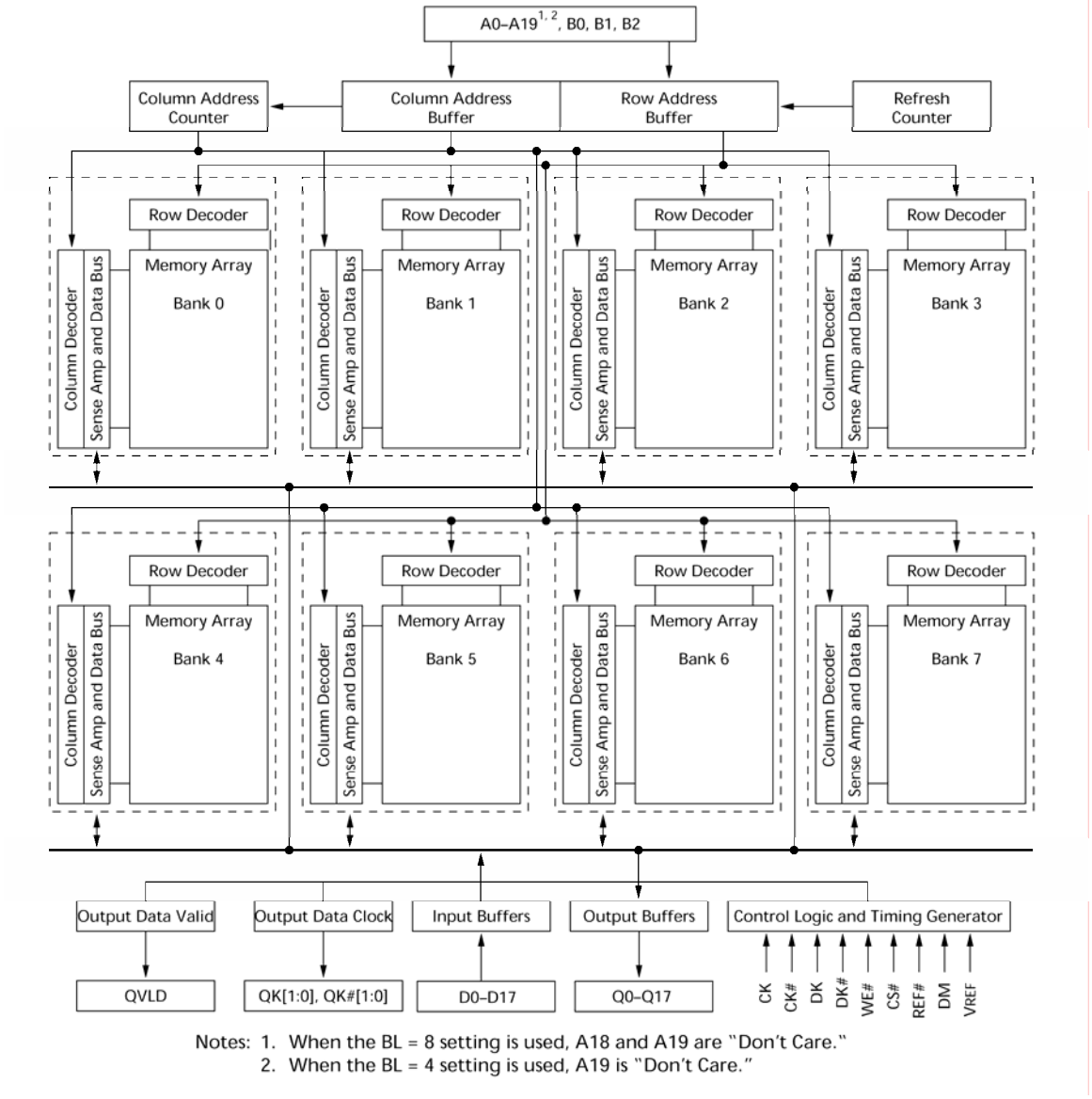


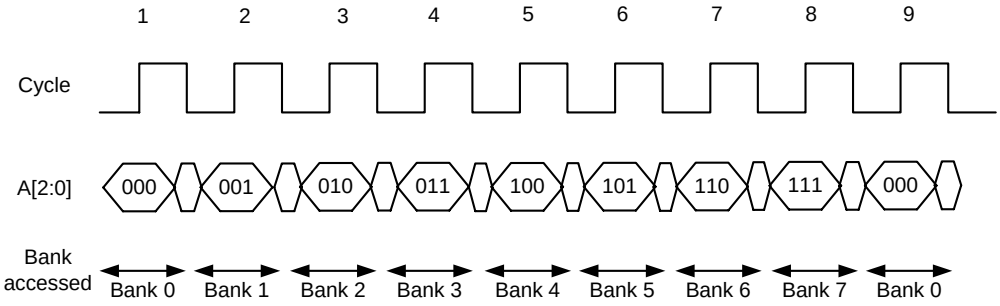
Fig. 3 Block diagram of RLDRAM II SIO architecture

Other features of RLDRAM II are as follows:

- Similar to the QDR/DDR family of SRAMs, the RLDRAM II architecture is available in separate I/O (SIO) and common I/O (CIO) versions. The SIO RLDRAM II architecture allows simultaneous READs and WRITEs like QDR, while the CIO architecture is similar to DDR SRAM.
- Although having an SRAM-type addressing, the RLDRAM can also employ the traditional DRAM multiplexed addressing scheme, enabled through a setting in the mode register. This feature allows the RLDRAM to be backward-compatible with older controller designs in terms of addressing and reduces the number of address pins used by the memory controller.
- An output signal, the *data valid signal,* indicates the data being read out on the I/O lines.
- The RLDRAM II design also employs data strobe clocks, a pair of free running clocks for latching output data (similar to the echo clocks CQ and CQ# of QDR-II).

The RLDRAM II architecture can achieve 100% bandwidth, although only under a specific access or address pattern. As mentioned above, the address lines in RLDRAM do not have to be multiplexed

(row/bank and column addresses do not have to be asserted at different points in time). So, using least significant bits (LSBs) of the address lines to reference the banks in a round-robin technique could ensure that the same bank is not accessed for a specific duration of time. This means that by using the LSBs of the address lines from the controller as bank inputs to the RLDRAM (pins B0, B1 and B2), incrementing them in order ensures that a different bank is accessed each cycle during a specific period. If this period is greater than or equal to the random cycle time (t_{RC}), as shown in figure 4, the t_{RC} will no longer limit the bandwidth of the device, resulting in full usage of the bandwidth. This round robin addressing technique may be beneficial in a situation where the accesses from the controller are sequential. However, in networking applications, the data accesses can be unpredictable and random in nature, as a result of which banks could be accessed in random order. This means that employing the round-robin address routing technique to access the RLDRAM may not be effective and could still result in the same bank being accessed before t_{RC} elapses, causing failed accesses. Thus, t_{RC} could limit the bandwidth of RLDRAM II in situations where the data pattern is unpredictable.



With a round robin addressing scheme (i.e., the least significant address lines connected to the bank inputs of RLDRAM), an increment of the address will result in a different bank being accessed every cycle, for 8 cycles. LSB 000 will hit Bank 0, 001 will hit Bank 1 and so on.

Figure 4. Sequential accesses with Round-robin addressing in RLDRAM

Another point to consider is the burst length. A higher burst length means more time for a new access. So, when a higher burst length device is used, fewer banks need to be accessed alternately to cover for the t_{RC} latency; however, with a smaller burst length device, more banks need to be used alternately to achieve maximum possible bandwidth.

Comparing QDR SRAM and RLDRAM

Having analyzed the architectural differences between the two high-speed memory solutions, we will now compare them based on suitability under various circumstances.

Randomness of application

Although RLDRAM II can achieve 100% bandwidth utilization using a round-robin addressing scheme and with a specific access sequence, it is not as effective when data accesses are random. While the architectural features of the RLDRAM II, like having more available banks and using internal precharge and built-in activation mechanisms ensure reduction in t_{RC}, they do not completely eliminate this latency and its effect on the bandwidth.

The waveform in Figure 5 shows how the t_{RC} latency affects it's bandwidth during short random bursts when the data pattern is unpredictable. A second access to bank A has to wait for several cycles before t_{RC} elapses, as a result of which the data bus goes unutilized. In such cases, the bandwidth of RLDRAM architectures will be affected due to the unpredictable nature of data accesses.

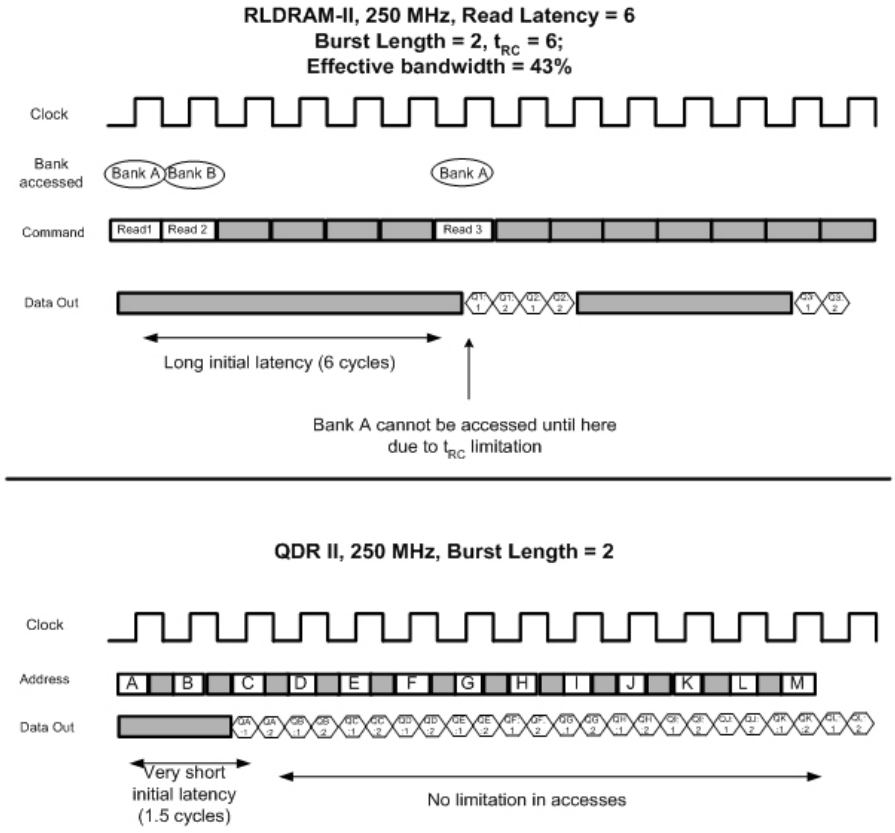


Figure 5. Comparison of RDLRAM II and QDR-II with a burst length of 2

In comparison, QDR SRAMs do not need any latency between accesses and are therefore not affected by randomness of application. They achieve 100% bandwidth utilization regardless of the access sequence or the randomness of data pattern.

Initial latency

RLDRAM II has a much higher initial latency compared to QDR SRAMs, as shown in figure 5. QDR and QDR-II SRAMs have an initial READ latency of just 1.0 and 1.5 clock cycles respectively; therefore, during a burst, the first piece of data comes out a lot sooner in QDR SRAM than RLDRAM II. This makes QDR SRAMs ideal for low latency applications. In RLDRAM II, the long initial latency is a problem when short data accesses occur back-to-back.

Figure 5 also shows how short burst lengths could limit the bandwidth utilization of RLDRAM II.

In comparison, the bandwidth of the QDR (or DDR) device is not affected by burst length at a given frequency.

Density and Cost

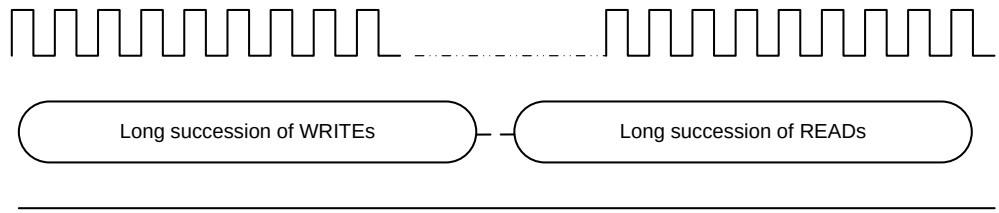
While deciding on a memory option, if density and cost per bit are more important considerations than the randomness of application and continuous peak bandwidth utilization, RLDRAM can provide a viable option due to the smaller 1T memory cell.

Bus utilization

Bus utilization is a key factor to consider in choosing the right memory solution. Figure 6 illustrates the following point.

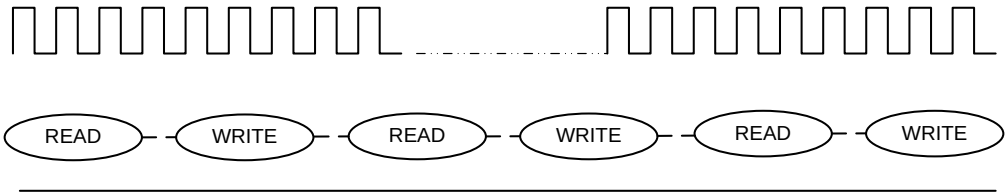
Long-term R/W ratio of 1:1

Long succession of multiple bursts of READs and WRITEs
Memories best suited: DDR CIO SRAM, RLDRAM II CIO



Short-term R/W ratio of 1:1

Short random bursts of alternate READs and WRITEs
Memories best suited: DDR SIO SRAM, QDR SRAM
RLDRAM II SIO also well suited if nature of access not random



Simultaneous/Overlapped READs and WRITEs

Memories best suited: QDR SRAM, RLDRAM II SIO

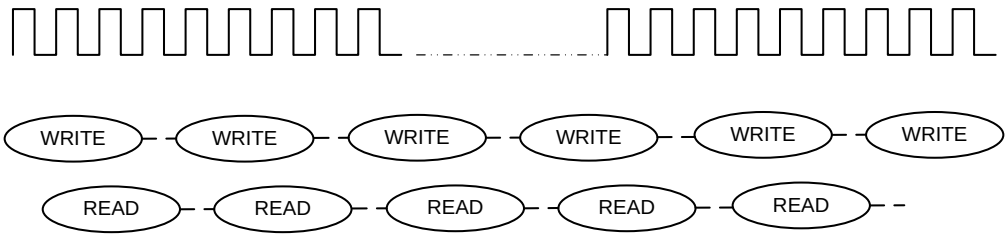


Figure 6. READ/WRITE patterns

Some systems may have a “near-term” READ/WRITE ratio of 1:1—small and equal number of READs and WRITEs, frequently interleaved. In such systems, inserting dummy cycles for bus turnaround would result in wasting significant number of cycles and adversely affecting the bandwidth. Therefore, DDR SIO SRAM, QDR SRAM, or RLDRAM II SIO would be the better option.

On the other hand, CIO devices like DDR SRAM, and RLDRAM II CIO would better suit an application that has a “long-term” READ/WRITE ratio of 1:1 with READs and WRITEs occurring in long bursts. If READs and WRITEs happen in long sequences without alternating frequently, the number of cycles lost to overcome bus contention would be very few compared to the number of cycles used for READs and WRITEs, thereby making a CIO device such as DDR SRAM or RLDRAM II CIO a suitable option. In such applications, choosing a SIO device would result in wasting I/Os for a significant portion of the cycles.

A third possible scenario is when READs and WRITEs occur simultaneously. In such systems, using SIO devices such as QDR SRAM and RLDRAM II SIO would be ideal.

In summary, a thorough understanding of the bus utilization needs of the application is essential in choosing the right memory solution in terms of the I/O architecture.

Figure 6 shows the different READ/WRITE patterns and which memory solutions are best suited.

The Big Decision: Choosing the Right Memory

With several high-speed synchronous memories available, the system designer is now presented with numerous viable memory solutions. Line cards typically need several memories for different functions (for example, table lookup, packet buffering, and queue management). Although each of these functions needs high-performance memory, not all the high-speed networking memories would be an ideal fit. This section describes the different memory requirements of a line card and proposes the memory solutions that best suit each application.

Figure 7 shows a high level view of a typical line card.

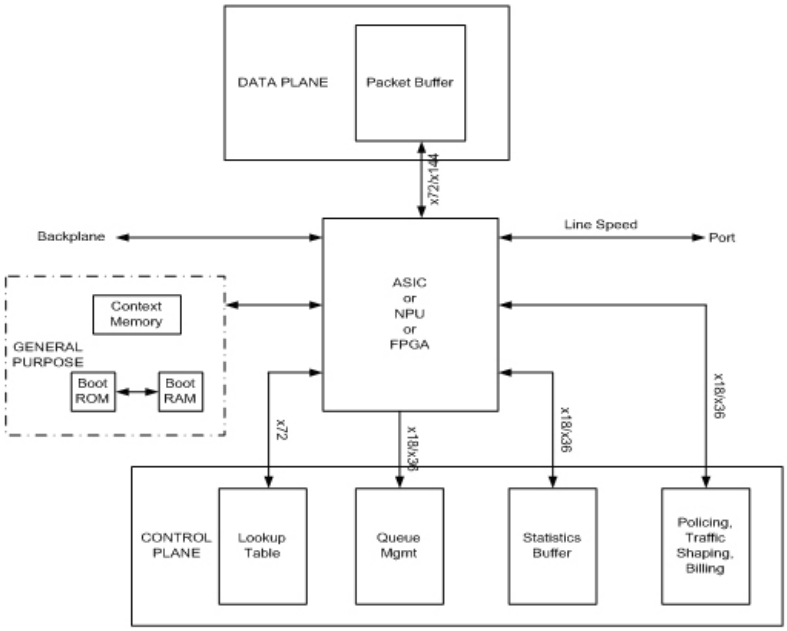


Figure 7. Components of a typical line card

Look-up table

The look-up table performs address translation while routing packets and resides in the control plane of the line card. Memory accesses to the look-up table are often random and characterized by short bursts of READ operations. Therefore, latency is the most critical factor in selecting the memory.

More recently, with core routers dealing with a large number of entries, density has also become a factor. While the ideal memory solution for look-up tables may actually vary from one architecture to another, the QDR/DDR family of SRAMs has numerous advantages. As explained earlier, QDR SRAM has a much shorter read latency compared to RLDRAM. This makes them more suitable for look-up tables, which are dominated by short READ bursts and need fast accesses.

In addition, randomness of the application and bus turnaround time during short bursts are critical factors that make QDR SRAMs the frontrunner choice for look-up tables.

On the other hand, if look-up tables are large and cost is a factor, RLDRAM II, with its low-latency, high-density, and low-cost features, is a good choice.

Queue/Packet Management

Queue management and flow control in a line card are characterized by random read and write operations. Hence, latency with unpredictable data patterns is a major factor to consider in choosing a memory.

A comparison of latency between QDR SRAM and RLDRAM II reveals that the QDR SRAM provides far superior performance over the networking DRAMs, especially when the data patterns are unpredictable. The shortcomings of RLDRAM II during random READ and WRITE operations have been explained before.

In applications such as queue management, where density is not a determining factor but latency is, QDR SRAMs are the better choice.

Statistics buffer

The Statistics buffer handles billing, diagnostics and a variety of other information. During packet processing, statistical data accesses need to be quick and hence, low latency is critical. However, statistical data is usually not large, and therefore, the operations are characterized by short bursts or no bursts. Both QDR SRAMs and NoBL SRAMs are well suited for this application.

Packet Cell Buffer

The packet buffer in the data plane is used to buffer packets in output ports and switch fabrics while the packet is being processed. Depending on the processing speed of the ASIC or the NPU, the packet buffer memory would need to be either very fast, very dense or both. In latency-critical designs, QDR SRAMs are preferred, while RLDRAM II would be a viable option where density is critical.

References

QDR Consortium Web site (2003). http://www.qdrsram.com/ Cypress Semiconductor Corporation, 18-Mb QDR™-II SRAM 4-Word Burst Architecture, data sheet CY7C1315AV18-200BZC (06/2004).

http://www.cypress.com/cfuploads/img/products/cy7c1315av18.pdf Micron Technology, Inc., 288Mb SIO REDUCED LATENCY (RLDRAM II), datasheet MT49H8M18C, Rev. 3 (05/2004).

http://download.micron.com/pdf/datasheets/rldram/MT49H16M18C.pdf Micron Technology, Inc., Exploring the RLDRAM II Feature Set, tech. note TN-49-02, Rev. A (01/2004).

http://download.micron.com/pdf/technotes/RLDRAMII/TN4902.pdf Micron Technology, Inc., RLDRAM II Design Guide, tech. note TN-49-01, Rev. A (03/2004).

About the Author

Anuj Chakrapani *is a senior applications engineer at the Memory and Imaging Division of Cypress Semiconductor. His responsibilities include creating behavioral simulation models of SRAMs, board-level failure-analysis debug, system-level testing, and applications support for customers. Anuj holds a master's degree in electrical engineering from Arizona State University (Tempe). He can be reached at:* *aju@cypress.com*