Technical Note

GDDR5X: The Next-Generation Graphics DRAM

Introduction

Since its market introduction in 2008, GDDR5 SGRAM has been the choice for applications demanding the highest memory bandwidth.

Initially supporting a per-pin data rate of up to 4 Gb/s, Micron has continuously increased the data rate of their GDDR5 SGRAM devices, and today offers data rates of up to 8 Gb/s—a remarkable 2X speed improvement and more than 3X improvement compared to its predecessor, GDDR4. With this effort, Micron has kept pace with the industry's demand for continuously increasing memory bandwidth of high-end graphics cards and game consoles.

While this speed improvement has been widely embraced by the industry, some aspects of GDDR5 SGRAM have become bottlenecks, resulting in data rate saturation at 8 Gb/s. In other words, GDDR5 SGRAM has hit the memory wall.

Micron addressed these bottlenecks with what now is called GDDR5X SGRAM. As the name suggests, GDDR5X should be considered a speed-enhanced derivative of GDDR5 rather than a radical new DRAM standard. This approach was chosen to let users leverage their previous investment in the GDDR5 memory ecosystem and enable a fast and low-risk transition from GDDR5.

Micron offers GDDR5X SGRAM devices with data rates of 10 Gb/s to 12 Gb/s, and devices with 14 Gb/s are anticipated in the future.

This technical note compares GDDR5 and GDDR5X and discusses the evolution of GDDR5X.

Figure 1: Micron's GDDR5X SGRAM

GDDR5 vs. GDDR5X

With GDDR5X, Micron breaks the current GDDR5 memory wall—starting with data rates of 10 Gb/s to 12 Gb/s.

Micron is committed to opening up an unprecedented level of performance for discrete graphic DRAMs, supporting the bandwidth needs of future graphics cards, consoles and other applications. The table below summarizes the differences between GDDR5 and GDDR5X SGRAM.

Table 1: GDDR5 vs. GDDR5X

Feature	GDDR5	GDDR5X	Notes
Density	2Gb, 4Gb, 8Gb	8Gb	GDDR5X also supports 12Gb and 16Gb
V_{DD}, V_{DDQ}	1.5V, 1.35V	1.35V	GDDR5: 1.5V required for full bandwidth; 1.35V supported at reduced bandwidth
V_{PP}	N/A	1.8V	
Package	BGA 170 14 x 12 0.8mm ball pitch	BGA 190 14 x 10 0.65mm ball pitch	
I/O width	x32/x16	x32/x16	Configured at power-up
Signal count	61	61	Same signal count at memory controller
Data rates	6–8 Gb/s	10–12 Gb/s (first generation)	GDDR5X long-term targets 14–16 Gb/s
Access granularity	32 bytes	64 bytes 2x 32 bytes in pseudo 32B mode	
Burst length	8	16/8	GDDR5X BL16 in QDR mode, BL8 in DDR mode
V_{REFD}	External/internal per 2 bytes 10m V step size	Internal per byte 6m V step size	
V_{REFC}	External	External/Internal	GDDR5X: Half V_{REFC} mode for operation without ODT; CK ODT available
CRC	CRC-8	Modified CRC-8	
Scan	SEN	IEEE 1149.1 (JTAG)	

Breaking the GDDR5 Bottlenecks

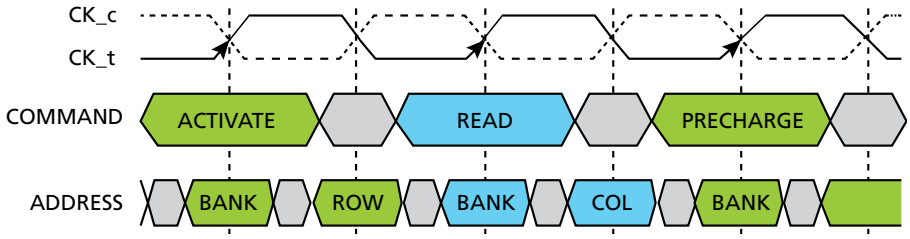
Bottleneck 1: Command and Address Protocol

Faster data rates are equivalent to shorter clock cycle times, a notion that has turned the command and address protocol into one of the bottlenecks in the GDDR5 memory subsystem.

A GDDR5 SGRAM must be able to receive and process a new command at every clock cycle, which makes it extremely difficult to implement when the device is operated with a data rate of 8 Gb/s and the clock cycle time is as short as 500ps.

One key aspect in the definition of GDDR5X SGRAM, therefore, was reducing the speed of the command and address bus while enabling data rates much faster than what was achievable with GDDR5.

Figure 2: GDDR5/GDDR5X Command and Address Timing

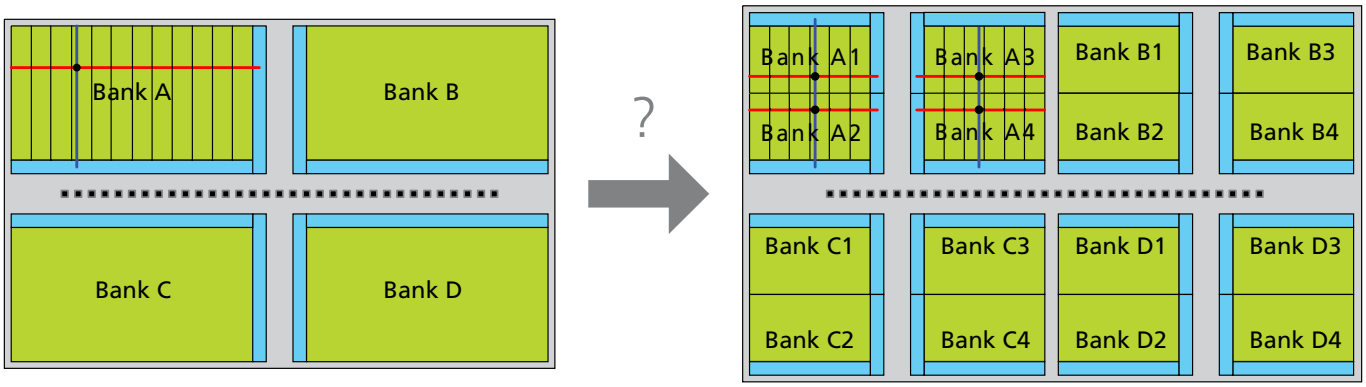


Bottleneck 2: Memory Array Speed

A second bottleneck of GDDR5 is the speed of the memory array. With GDDR5, a memory array access must be accomplished within two clock cycles. These shorter clock cycle times continuously lower the absolute time available to complete a memory array cycle. For example, at 4 Gb/s, two clock cycles equal 2ns, which is already much shorter than the array cycle time of commodity DDR3 and DDR4 DRAM devices. At 8 Gb/s, the same array access has to be accomplished within 1ns only.

In DRAM devices, there is a fundamental limit in scaling array speed, as shown below.

Figure 3: DRAM Array Architectures



The left architecture depicts a typical array with long word lines and bit lines routed across the memory array. The architecture on the right leads to shorter word lines and bit lines, which should enable the array to operate faster but at the expense of a much larger die.

The larger die not only results in higher cost of the device, but it also leads to longer data lines to connect the different array segments. These longer data lines eventually result in speed degradation, which contradicts with higher performance and also leads to higher operating power. The decision for GDDR5X, therefore, was that any further increase in bandwidth has to be achieved without an increase in array speed. GDDR5X accomplishes this by doubling the memory prefetch.

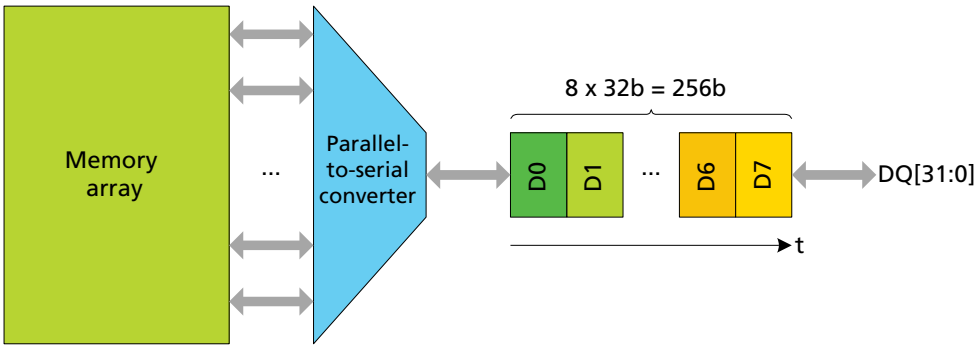
GDDR5 and GDDR5X Feature Comparisons

Array Prefetch and Access Granularity

The term *prefetch* describes a parallelism utilized in all state-of-the-art DRAM devices. The purpose of prefetch is to map the relatively low speed of an internal memory array access to the I/O data rate at the external interface.

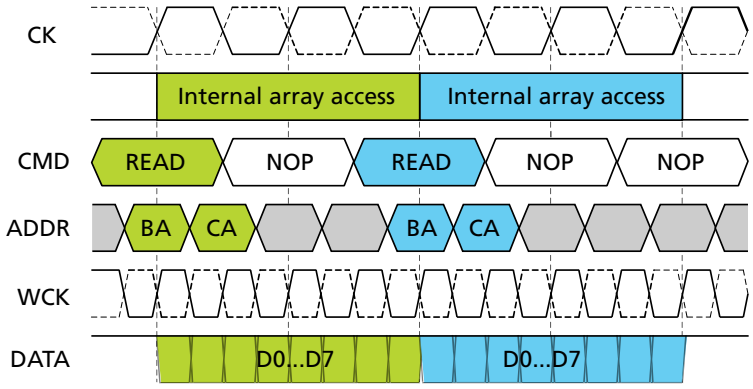
GDDR5 SGRAM uses an internal prefetch called 8n, as illustrated in the figure below. Each write or read memory access is 256 bits or 32 bytes wide. A parallel-to-serial converter translates each 256-bit data packet into eight 32-bit wide data words that are transmitted sequentially over the 32-bit-wide data bus.

Figure 4: GDDR5 8n Prefetch Memory Architecture



The timing diagram below illustrates this concept for two seamless read accesses with GDDR5. The 256 bits of data per memory access are transferred over the two CK clock cycles as 8 data words, D0 to D7. A zero-cycle read latency has been assumed for illustration purposes.

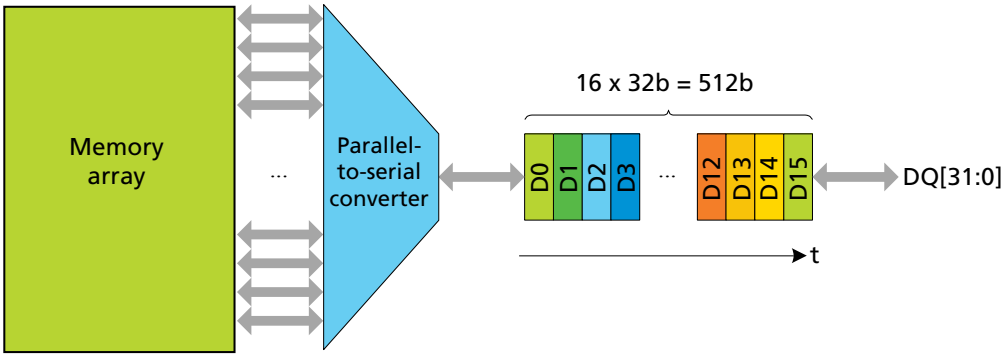
Figure 5: Seamless Read Accesses in GDDR5



GDDR5X SGRAM doubles this internal array prefetch. This approach is straightforward for DRAM devices and has been successfully implemented several times in DRAM evolution whenever the memory array speed ran into this limitation (for example, in DDR2 to DDR3 or LPDDR3 to LPDDR4 transitions).

GDDR5X SGRAM internal prefetch was doubled to 16, as illustrated in the figure below. Each write or read memory access is 512 bits or 64 bytes wide. A parallel-to-serial converter translates this 512-bit data packet into sixteen 32-bit wide data words that are transmitted sequentially over the same 32-bit-wide data bus.

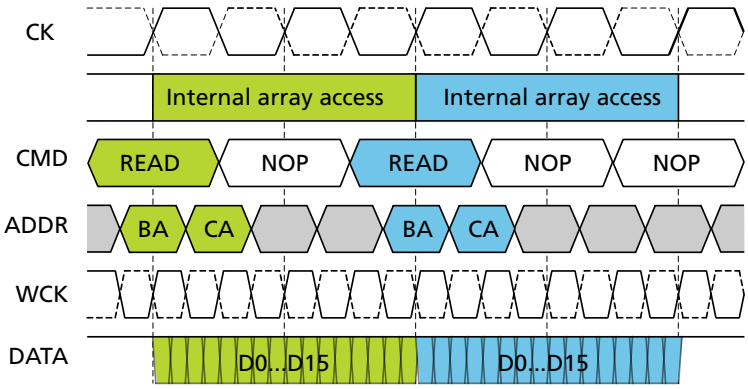
Figure 6: GDDR5X 16n Prefetch Memory Architecture



The timing diagram below shows the same two seamless read accesses as with GDDR5, but in this case for GDDR5X SGRAM. The 512 bits of data per memory access are transferred over the two CK clock cycles as sixteen data words, D0 to D15, each.

The diagram also illustrates that four data words are transmitted per WCK clock cycle, which is known as quad data rate (discussed later in this document).

Figure 7: GDDR5X 16n Prefetch Memory Architecture



When comparing the GDDR5 and GDDR5X timing diagrams, the commonalities and differences between GDDR5 and GDDR5X become obvious. GDDR5 and GDDR5X use the same command and address protocol, along with the same CK and WCK clock frequencies. The only apparent difference is the two-fold data rate on the data bus for GDDR5X. Even the EDC pins (not shown) used for transmitting CRC data between the SGRAM and the host, for example, are operated at the same data rate for both devices.

Table 2 (page 7) summarizes this outcome for a 12 Gb/s GDDR5X device and a 6 Gb/s GDDR5 device. Both devices are operated at the same CK and WCK clock frequencies and also at the same data rates on command, address and EDC.

The comparison also illustrates how GDDR5X SGRAM removes GDDR5 SGRAM bottlenecks: For the same 6 Gb/s data rate, GDDR5X SGRAM will be clocked at only half the frequency (in this case 750 MHz) of GDDR5 SGRAM (1.5 GHz). The divided-by-two command and address data rates enable GDDR5X SGRAM to theoretically achieve twice the bandwidth of GDDR5.

Table 2: Clock Frequencies and Data Rates

Pin	GDDR5 6 Gb/s	GDDR5X 12 Gb/s	GDDR5X 6 Gb/s
CK	1500 MHz	1500 MHz	750 MHz
WCK	3000 MHz	3000 MHz	1500 MHz
COMMAND	1.5 Gb/s	1.5 Gb/s	0.75 Gb/s
ADDRESS	3.0 Gb/s	3.0 Gb/s	1.5 Gb/s
DQ/DBI_n	6 Gb/s	**12 Gb/s**	6 Gb/s
EDC	6 Gb/s	6 Gb/s	3 Gb/s

Pseudo 32 Bytes Mode

The increase in access granularity from 32 bytes to 64 bytes is something that should be considered when transitioning to GDDR5X. Obviously, the maximum average performance is achieved when the 64 bytes transferred per memory access contain useable data, which is the case, for example, when a controller's cache line is 64 bytes wide.

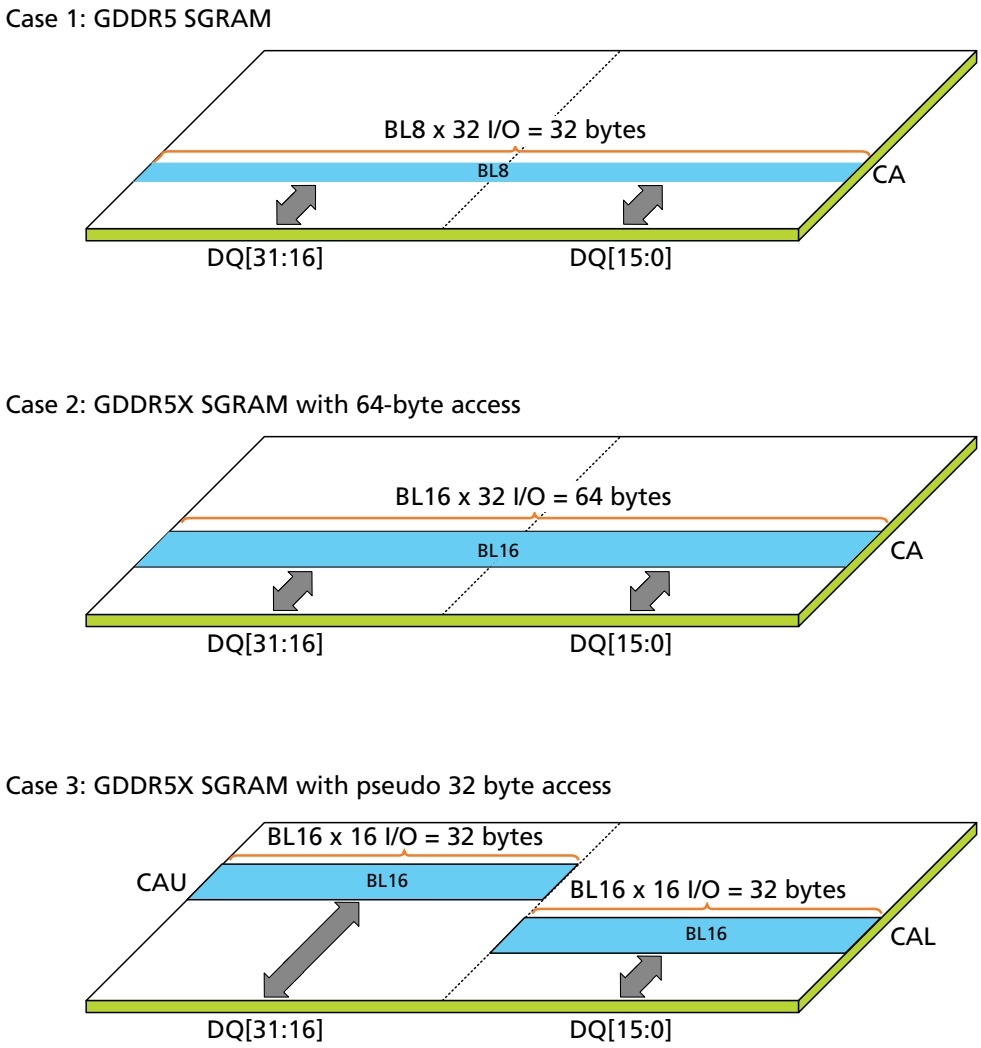
However, depending on the application, there may be cases where not all 64 bytes of a read burst contain usable data. The host in this case will ignore some of the data, which is known as *over-fetch*. Controller architecture optimizations try to minimize this over-fetch in an attempt to maximize the utilization of the available memory bandwidth.

To minimize the disadvantage of the GDDR5X SGRAM 64-byte access, some features have been added to the GDDR5X specification. When these features are efficiently used by the memory controller, GDDR5X is operated in what is called pseudo 32 bytes mode:

- Two column addresses (CAL and CAU): Instead of a single column address per read or write command, GDDR5X SGRAM supports two column addresses CAL (column address for lower DQ) and CAU (column address for upper DQ) per read or write command. This concept enables the memory controller to access two individual blocks of 32 bytes each within the same open bank and page.
- New write commands transmit data on the lower or upper 16 DQ: For 32-byte writes, new write commands transmit data on the lower or upper 16 DQ only while implicitly suppressing the write on the other 16 DQ. These commands are single-cycle commands (like a normal write) and therefore have no impact on the command bus utilization; these commands can be mixed with any regular 64-byte write command as desired.
- Masked writes are suppressed on lower or upper 16 DQs and then applied: For masked writes, additional write data mask bits efficiently suppress the write on the lower or upper 16 DQs while applying the actual write data mask bits on the other 16 DQs. This results in the same double-byte or single-byte mask granularity as with GDDR5 SGRAM, and again with no impact on the command bus utilization.

The following figure illustrates pseudo 32 bytes mode.

Figure 8: Pseudo 32 Bytes Mode



Case 1 represents GDDR5 SGRAM: Memory access comprises a read or write with a burst length of 8 over 32 DQs.

Case 2 represents GDDR5X SGRAM in regular 64-bytes mode: Memory access comprises a read or write with a burst length of 16 over the same 32 DQs.

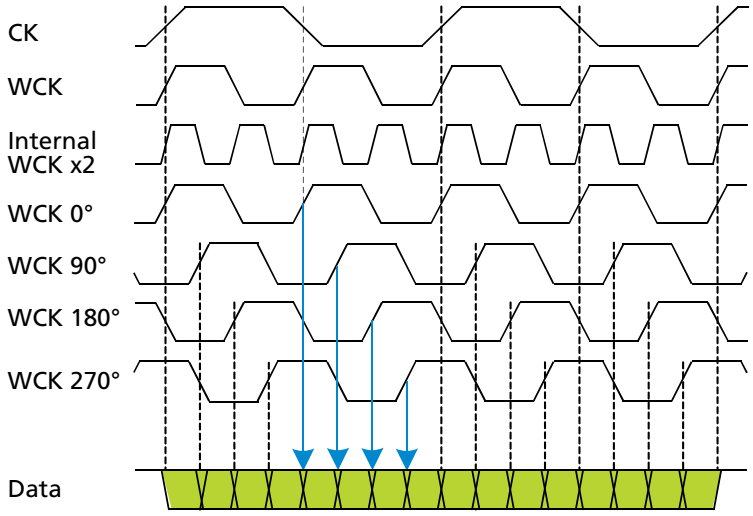
Case 3 represents GDDR5X SGRAM in pseudo 32 bytes mode: Memory access comprises two simultaneous reads or writes with a burst length of 16 over the lower and the upper 16 DQs. The two column addresses CAL and CAU are provided without any performance penalty over the existing address bus; they simply utilize "don't care" address bits of the GDDR5 read and write command protocol.

Quad Data Rate (WCK)

As already shown in the previous timing diagrams, GDDR5X SGRAM uses the exact same clocking scheme as GDDR5, with the WCK clock operated at exactly twice the CK clock frequency. Leveraging this proven clocking scheme enables users to migrate to GDDR5X with no or minimum changes to the clock generation circuits inside the memory controller, and to utilize the same WCK-to-CK training algorithm.

The quad data rate however requires the GDDR5X to internally generate additional clock edges for receiving and transmitting data. This is achieved by the use of a PLL, as shown in the figure below. The PLL doubles the external WCK clock frequency, and four internal WCK clock phases (0°, 90°, 180°, 270°) are derived from the PLL for receiving and transmitting data.

Figure 9: GDDR5X Data Clocking



GDDR5X Optimizations

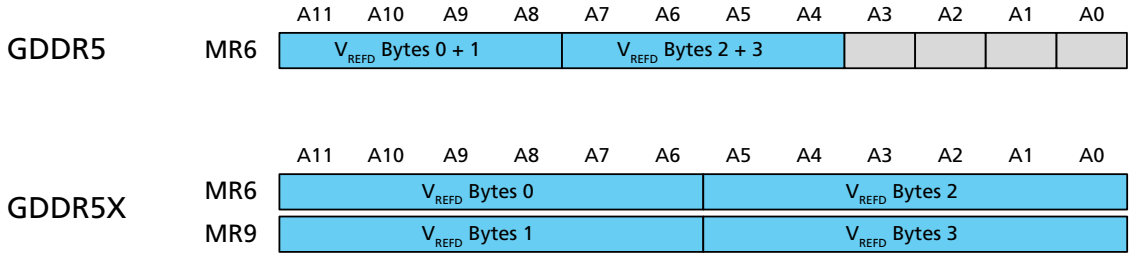
Data Receiver

The extremely fast data rates of GDDR5X require significantly more attention to be spent in the optimization of the actual data transmission at all aspects of the interface, the memory controller, the interconnect (PCB) and the DRAM design.

One critical circuit block is the GDDR5X SGRAM data receiver. While the actual receiver design is proprietary, some improvements have been incorporated into the GDDR5X specification to support the increased requirements for input receivers:

- GDDR5X provides registers that enable the data receiver reference voltage V_{REFD} to be programmed individually per each data byte, while GDDR5 had registers for two bytes grouped (see the figure below). The per-byte V_{REFD} capability reflects the fact that, for example, the board routing of two data bytes may be slightly different, leading to a slightly different vertical center of the data eye which suggests using different V_{REFD} levels for maximum performance.
- The additional bits enable the V_{REFD} step size to be reduced from 10mV to 6mV.
- In addition to the data receiver optimization, GDDR5X extends the same programmable on-die termination (ODT) to the CK_t and CK_c differential clock inputs.
- GDDR5X also offers improvements for the command/address receivers: The input receiver reference voltage V_{REFC} can now be set to be generated internally, which not only eliminates the need for an external resistor bridge to be placed close to each device, but also lets the command/address receivers benefit from the better supply voltage tracking of an internal voltage reference compared to external reference.

Figure 10: Mode Registers for Setting the V_{REFD} Level

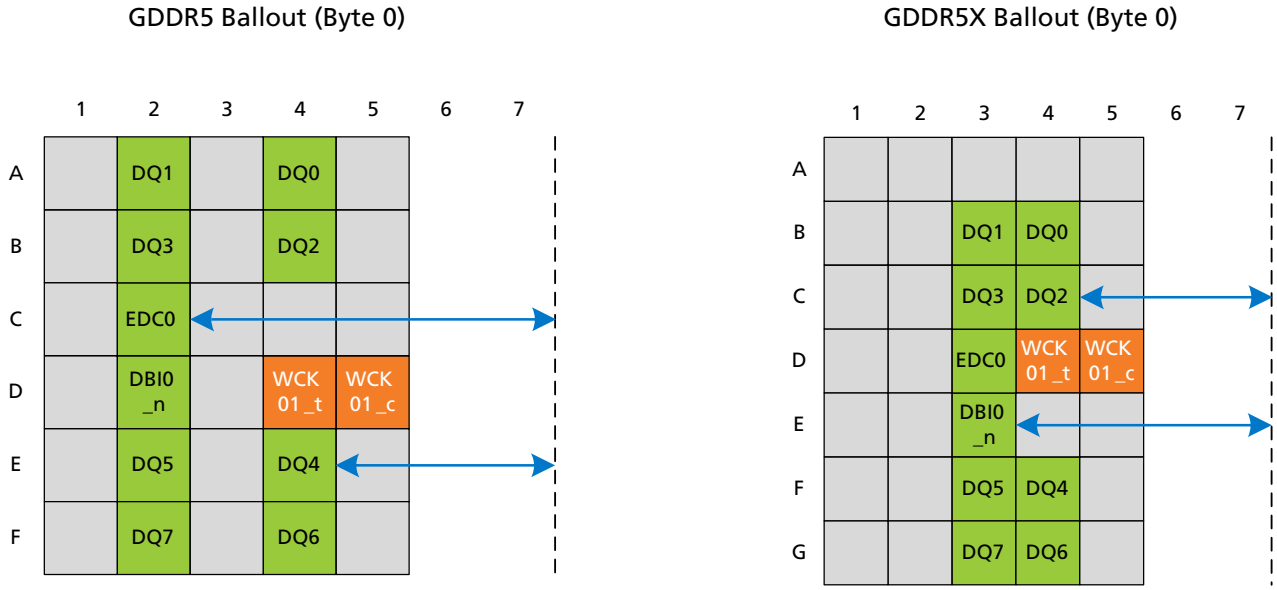


Package Ballout

The following figure shows excerpts from the GDDR5 and GDDR5X ballouts, covering the byte 0 data pins locations. With GDDR5X the outer data pins (DQ1, DQ3, DQ5, DQ7, EDC0, DBI0_n) were moved by one column towards the center of the package.

GDDR5X also uses a smaller ball pitch of 0.65mm compared to 0.8mm for GDDR5, which is reflected in the figure by proper scaling. Both measures result in shorter signal traces for all high speed data signals, which reduces crosstalk and also the pin-to-pin skew within a byte.

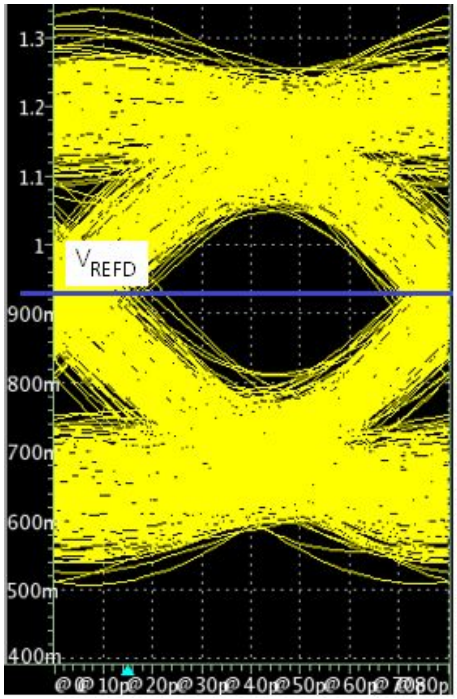
Figure 11: GDDR5 and GDDR5X Ballouts (Byte 0)



The smaller ball pitch of GDDR5X also enables a reduction of the package outline from 14mm x 12mm to 14mm x 10mm. Although not a big difference, it may lead to an improved DRAM device placement in dense PCB designs and enables power supply decoupling capacitors to be placed closer to the device for improved supply noise reduction.

All measures combined enable GDDR5X SGRAM to achieve unprecedented data rates. The figure below shows an input data eye simulated at 12 Gb/s using actual package parasitics and a typical channel model.

Figure 12: Data Eye

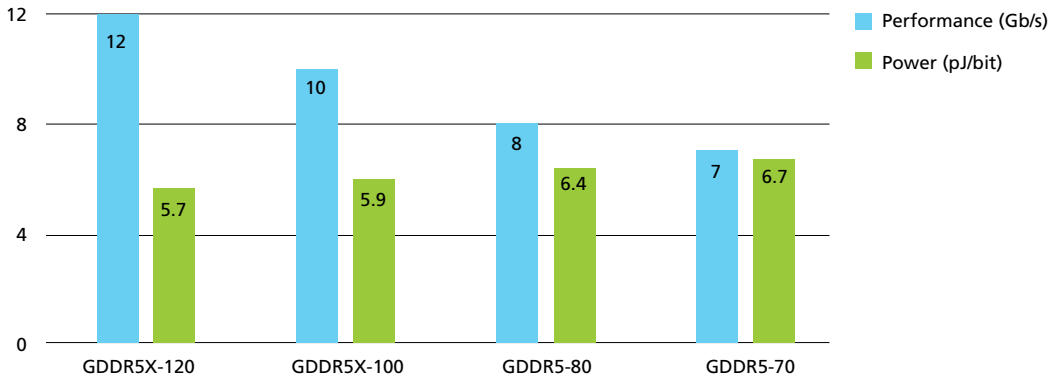


Supply Voltage and Energy Efficiency

GDDR5 SGRAM operates from a 1.5V supply voltage (V_{DD}, V_{DDQ}). With GDDR5X, the supply voltage is reduced to 1.35V, reflecting advances in both logic and DRAM transistor performance. GDDR5X also uses a separate 1.8V supply for driving the word lines (V_{PP}).

While the absolute power consumption of a single 12 Gb/s GDDR5X SGRAM device may be slightly higher compared to a conventional 8 Gb/s GDDR5 device, the energy per bit is significantly reduced.

Figure 13: Energy Efficiency Comparison for 20nm 8Gb GDDR5 and GDDR5X Devices



Low Power Features

GDDR5X SGRAM preserves all low power features of GDDR5 SGRAM and adds new features to address the increasing importance of energy saving:

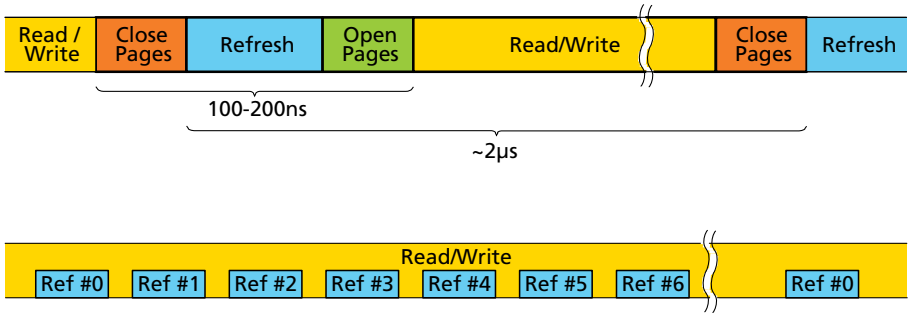
- The device can be operated over a contiguous frequency range starting at 50 MHz (equivalent to 200 Mb/s) up to the maximum specified data rate. This enables the controller to save a significant amount of power by adjusting the clock rate and bandwidth to the actual workload.
- For data rates below 2 Gb/s, the GDDR5X may be set into a strobe mode, which returns a data strobe like a conventional RDQS along with read data and thus helps the memory controller save power.
- On-die terminations for all high-speed inputs can be programmed to higher impedances, resulting in reduced static ODT currents, or they be fully turned off whenever appropriate.
- A DRAM memory cell's data retention is significantly better at lower temperatures. An integrated temperature sensor associated with the automatic temperature controlled self refresh (ATCSR) (a mandatory feature with GDDR5X) utilizes this fact and adjusts the time interval between internal refresh cycles, resulting in a lower average power consumption during idle states.
- Partial-array self refresh (PASR) addresses power consumption during idle states; the device can be programmed to exclude portions of its memory array from being refreshed.
- A so-called hibernate mode reduces the device's power consumption in self refresh mode even further, at the expense of a wake-up time similar to the time needed for device initialization.

Per-Bank Refresh

DRAMs store information in tiny capacitors that require regular refreshes in order to not lose data. The refreshes traditionally require the controller to stop and terminate any read or write operations, close all open pages, issue a REFRESH command, then reopen the pages and resume read and write operations. The REFRESH command refreshes one or more pages in all banks simultaneously (hence it is also referred to as *all-bank refresh*).

The time needed for each refresh depends on the DRAM vendor, DRAM density and the device's case temperature. Using 100-200ns as estimates, and comparing them to a 2µs interval for an all-bank refresh, these refreshes consume about 5% to 10% of the DRAM's overall bandwidth, as shown in the upper portion of the figure below.

Figure 14: All-Bank and Per-Bank Refreshes



GDDR5X SGRAM supports an alternative way to perform refreshes. The new per-bank refresh performs a refresh on a single bank only, while read and write transactions to the other banks may continue unchanged, as shown in the lower portion in the previous figure. To achieve the same amount of cell refresh, these per-bank refresh commands have to be issued at 16X the rate of the all-bank refresh due to the 16 banks in GDDR5X SGRAM.

Theoretically, per-bank refresh should be able to fully restore the performance loss of the all-bank refresh. Practically, the gain will be less because, on the one hand, each per-bank refresh command occupies command slots that otherwise could be used for page activations, and, on the other hand, one or more banks at a time are busy performing the refreshes and are thus not available for reads and writes.

Nevertheless, the per-bank refresh feature is expected to increase the average memory bandwidth compared to DRAMs, which do not support this feature.

Additional Features

GDDR5 (DDR) Mode

GDDR5X SGRAM may also be programmed to operate exactly as a GDDR5 SGRAM. This mode is called DDR mode, and it provides the same 32-byte access granularity, burst length (8), commands and write data mask as conventional GDDR5.

Users may decide to replace GDDR5 SGRAM devices with GDDR5X devices to leverage new GDDR5X features such as per-bank refresh. In this case, only a minor logic change would be required at the controller. While GDDR5 supports both a DESELECT command (CS_n = HIGH) as well as a NOP command (CS_n = LOW, RAS_n, CAS_n, WE_n = HIGH) for idle command cycles, GDDR5X abandoned DESELECT in favor of an additional address pin; thus GDDR5X supports the NOP command only.

JTAG Boundary Scan

GDDR5X SGRAM provides an IEEE1149.1-compliant boundary scan test port. This test port has been a state-of-the art test feature throughout the industry but supported by very few DRAM standards. The port enables signal connection testing between the host and DRAM using a standardized protocol.

References

- Micron 8Gb GDDR5X SGRAM Data Sheet
  https://www.micron.com/resource-details/a3cada97-b50f-46dc-bd2f-40a60a5f09ad
- Micron GDDR5 SGRAM Technical Note
  http://www.micron.com/~/media/documents/products/technical-note/dram/tned01_gddr5_sgram_introduction.pdf
- JEDEC Double Data Rate (GDDR5X) SGRAM Standard
  http://www.jedec.org/standards-documents/results/JESD232

Revision History

Rev. A – 05/16

- Initial release

8000 S. Federal Way, P.O. Box 6, Boise, ID 83707-0006, Tel: 208-368-4000
www.micron.com/products/support Sales inquiries: 800-932-4992