Technical Note

GDDR6: The Next-Generation Graphics DRAM

Introduction

Since its market introduction in 2015, GDDR5X has been the world's fastest discrete memory for high-performance gaming and workstation-class graphics cards, replacing GDDR5 SGRAM as the standard for these applications.

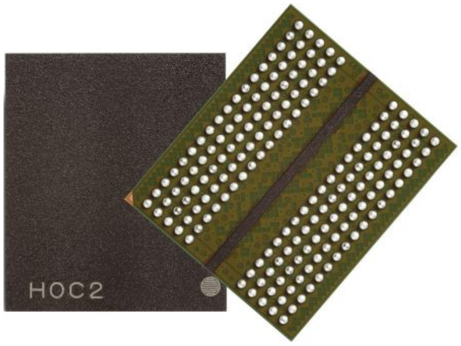
With GDDR6, Micron is taking the next step in accelerating this type of memory.

Today's GDDR5X SGRAM supports per-pin data rates up to 13 gigabits per second (Gb/s). Micron's GDDR6 raises per-pin data rates to 16 Gb/s — an unprecedented 2X improvement over GDDR5.

In addition to faster data rates, GDDR6 SGRAMs provide another major advantage over GDDR5X by fully supporting the same 32-byte access granularity as GDDR5. (The term access granularity refers to the amount of data written to and read from the memory core by a single WRITE or READ command.) Keeping the 32-byte granularity enables processor architectures that are optimized for GDDR5 to transition to GDDR6 with minimal effort.

This technical note describes how GDDR6 leverages features of GDDR5 and GDDR5X to be the best-suited memory for ultra-fast DRAM.

Figure 1: Micron's GDDR6 SGRAM



Products and specifications discussed herein are for evaluation and reference purposes only and are subject to change by Micron without notice. Products are only warranted by Micron to meet Micron's production data sheet specifications. All information discussed herein is provided on an "as is" basis, without warranties of any kind.

GDDR5 vs. GDDR5X vs. GDDR6

The table below compares the major features of GDDR5, GDDR5X and GDDR6. How GDDR6 benefits from feature improvements is discussed later in this technical note.

Table 1: Feature Comparison—GDDR5 vs. GDDR5X vs. GDDR6

Feature	GDDR5	GDDR5X	GDDR6	Notes
Density	(2Gb) 4Gb, 8Gb	8Gb	8Gb, 16Gb (32Gb)	
V_{DD}, V_{DDQ}	1.5V, 1.35V	1.35V	1.35V, 1.25V	GDDR6: initially 1.35V; 1.25V at reduced data rates
V_{PP}	N/A	1.8V	1.8V	
Data rates	Up to 8 Gb/s	Up to 12 Gb/s	Up to 16 Gb/s	
Channel count	1	1	2	
Access granularity	32 bytes	64 bytes 2x 32 bytes in pseudo 32B mode	2 ch x 32 bytes 2x 32 bytes in pseudo 32B mode	GDDR6: 32B per channel
Array prefetch	32 bytes	64 bytes	2x 32 bytes	
Burst length	8	16/8	16	
Package	BGA-170 14 mm x 12mm 0.8mm ball pitch	BGA-190 14mm x 10mm 0.65mm ball pitch	BGA-180 14mm x 12mm 0.75mm ball pitch	
I/O width	x32/x16	x32/x16	2 ch x 16/x8	Configured at power-up
Signal count	61 - 40 DQ,DBI_n,EDC - 15 CA - 6 CK, WCK	61 - 40 DQ,DBI_n,EDC - 15 CA - 6 CK, WCK	68 or 76 - 40 DQ,DBI_n,EDC - 16 or 24 CA - 12 CK, WCK	GDDR6: CA pin count depends on selected mode
ABI, DBI	Yes	Yes	Yes	
CRC	CRC-8	Modified CRC-8	2x CRC-8	
V_{REFD}	External/internal per 2 bytes ~10mV step size and 15 steps	Internal per byte ~7mV step size and 64 steps	Internal per pin ~7mV step size and 96 steps	
DFE	N/A	N/A	1-tap DFE	
V_{REFC}	External	External/Internal	External/Internal	GDDR6: with V_{REFC} offset
Self refresh (SRF)	Yes Temp controlled SRF	Yes Temp controlled SRF Hibernate SRF	Yes Temp controlled SRF Hibernate SRF	
Scan	SEN	IEEE1149.1 (JTAG)	IEEE1149.1 (JTAG)	

Memory Array Prefetch and Access Granularity

GDDR5, GDDR5X and GDDR6 devices provide a 32-bit wide data interface to the memory controller; however, there are major differences in the internal architecture.

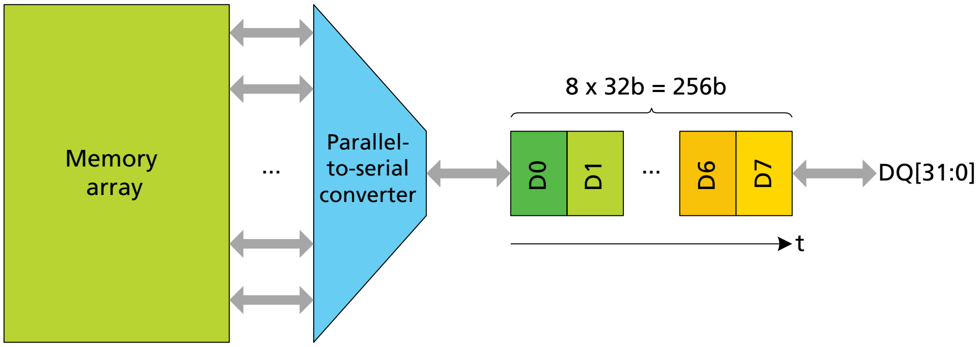
Memory Array Prefetch

The term *prefetch* describes a parallelism utilized in all state-of-the-art DRAM devices. The purpose of prefetch is to match the moderate speed of the internal memory array with the much faster I/O data rate of the external interface.

GDDR5 SGRAM uses an internal *8n* prefetch, as illustrated in the figure below. The term *8n* refers to the internal data bus being 8 times as wide as the device's I/O interface. 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 data words that are transmitted sequentially over the 32-bit data bus.

With this *8n* prefetch, an internal array cycle time of 1ns equals a data rate of 8 Gb/s at the I/O. The duration of a single data word at 8 Gb/s is 125ps, or 1/8 of the array cycle time.

Figure 2: GDDR5 8n Prefetch Memory Architecture

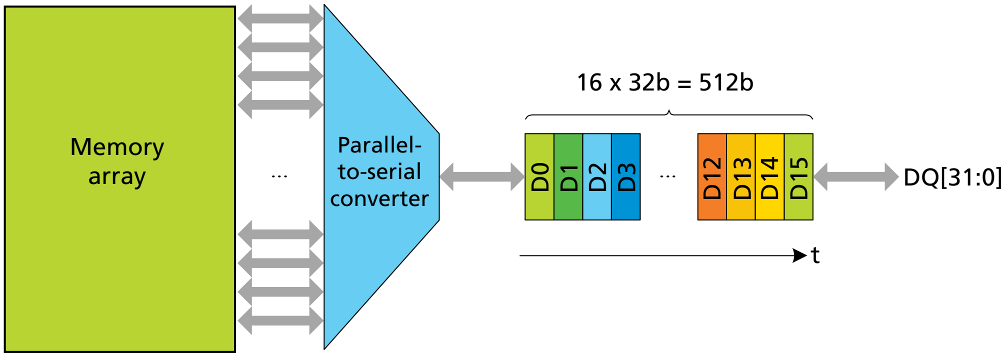


When we designed GDDR5X, we doubled this internal array prefetch. This approach is straightforward and has been successfully implemented in the development of mainstream DDR DRAM standards. This has allowed DRAM manufacturers to balance the design constraints placed on array cycle times with the ever-increasing demand for higher data rates.

GDDR5X uses an internal *16n* prefetch as illustrated in the figure below. The internal data bus is 16 times as wide as the device's I/O interface. Each write or read memory access is 512 bits or 64 bytes wide. A parallel-to-serial converter translates each 512-bit data packet into sixteen 32-bit data words that are transmitted sequentially over the 32-bit data bus.

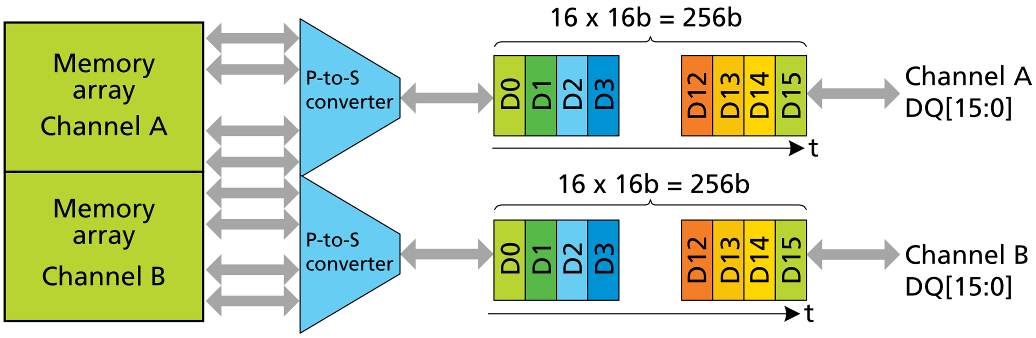
With this *16n* prefetch, the same internal array cycle time of 1ns equals a data rate of 16 Gb/s at the I/O. The duration of a single data word at 16 Gb/s is 62.5ps, or 1/16 of the array cycle time.

Figure 3: GDDR5X 16n Prefetch Memory Architecture



GDDR6 keeps the same *16n* prefetch of GDDR5X but logically splits the 32-bit data interface into two 16-bit channels, A and B, as shown in the figure below.

Figure 4: GDDR6 2-Channel 16n Prefetch Memory Architecture



The two channels are fully independent of each other. For each channel, a write or read memory access is 256 bits or 32 bytes wide. A parallel-to-serial converter translates each 256-bit data packet into sixteen 16-bit data words that are transmitted sequentially over the 16-bit data bus.

Due to this *16n* prefetch with GDDR6, the same internal array cycle time of 1ns equals a data rate of 16 Gb/s.

Access Granularity

The table below summarizes the access granularity of the different GDDR standards. Access granularity and array prefetch are synonymous and are the product of I/O width and burst length.

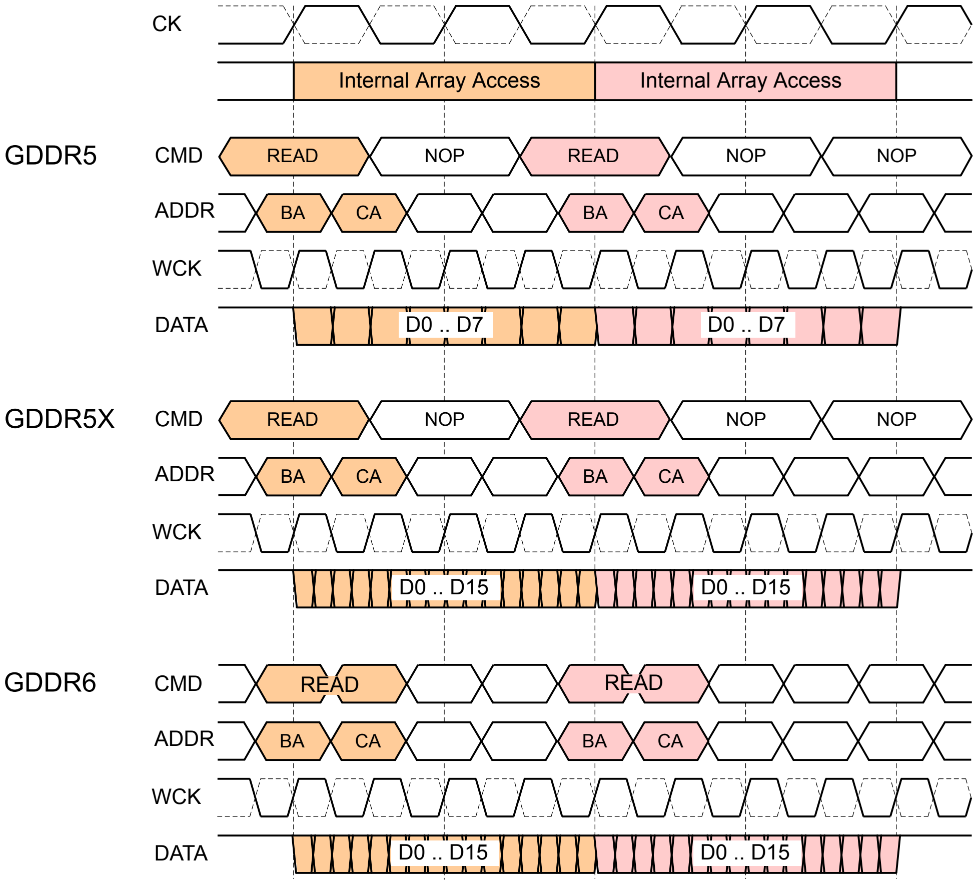
Table 2: Access Granularity GDDR5 – GDDR5X – GDDR6

Feature	GDDR5	GDDR5X	GDDR6 (Per Channel)
I/O width	32	32	16
Burst length	8	16	16
Access granularity	32 bytes	64 bytes	32 bytes

Memory WRITE and READ Operations

The figure below illustrates the memory array prefetch in the form of timing diagrams. Two seamless read accesses are shown for GDDR5, GDDR5X and GDDR6.

Figure 5: Seamless READs with GDDR5, GDDR5X and GDDR6



From the figure above it becomes apparent that the three GDDR standards have many similarities. In fact, taking GDDR5 as the parent GDDR standard, only select items have been modified from the migration of GDDR5 to GDDR5X and GDDR6 to allow as smooth a transition as possible to each next-generation standard.

- With all three GDDR standards, internal write and read accesses are two CK clock cycles long (t_{CCD} = 2 t_{CK}). A 100% bus utilization is achieved when a WRITE or READ is issued every second cycle (e.g. READ - NOP - READ).
- CK and WCK clock frequencies are the same, although GDDR5X and GDDR6 provide twice the I/O data rate. (See WCK clocking options with GDDR6 later in this tech note for more information.)
- GDDR5 and GDDR5X receive commands as single data rate (SDR), referenced to the rising CK clock edge, while addresses are received double data rate (DDR) referenced to both rising and falling CK clock edges. GDDR6 receives both commands and addresses as DDR, thus saving three CA pins.

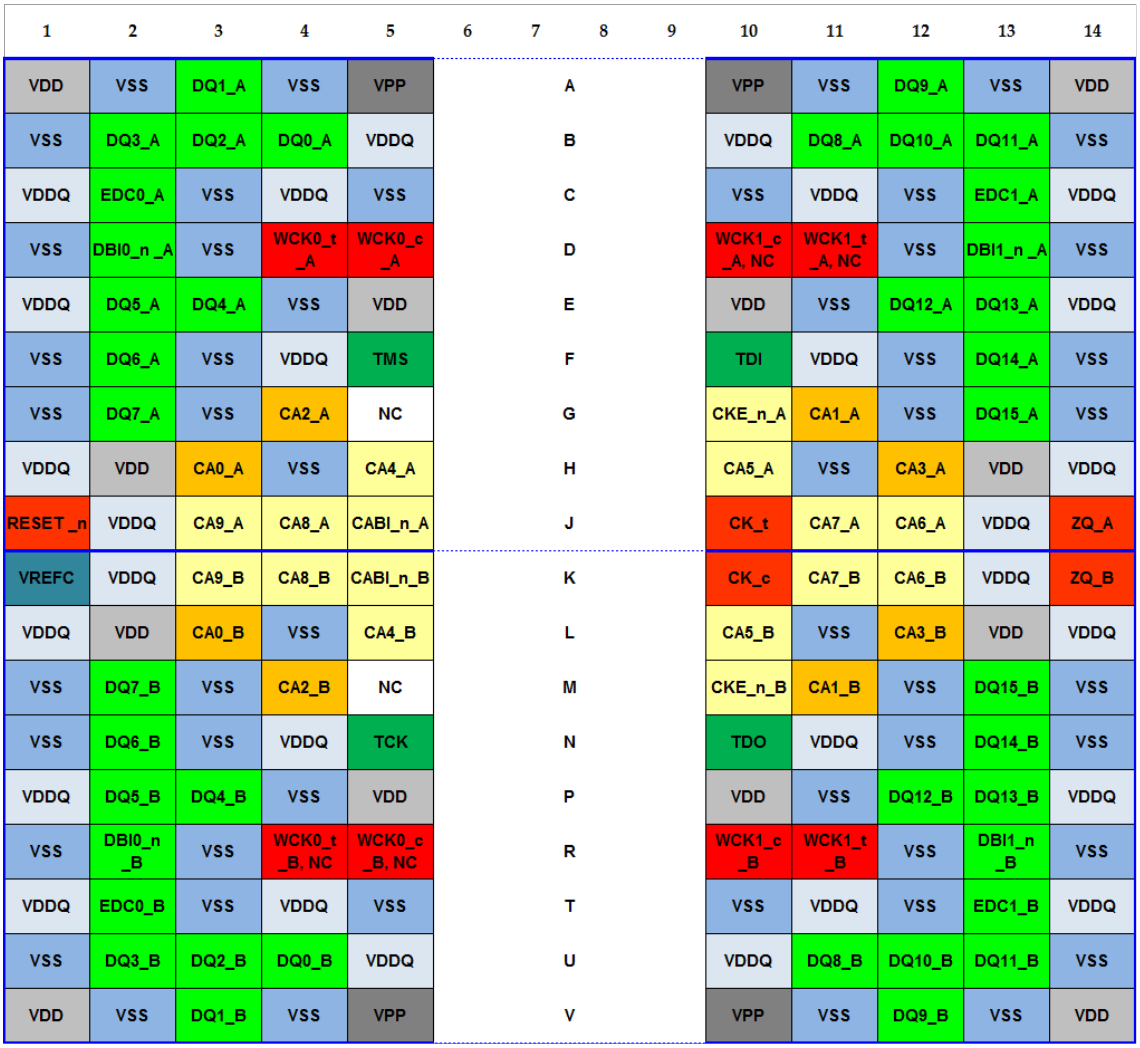
2-Channel and Pseudo Channel Modes

A GDDR6 device can be configured to operate in 2-channel mode or in pseudo-channel (PC) mode. Before discussing both modes, we will take a closer look at the ballout.

Ballout

The GDDR6 ballout reflects the 2-channel architecture of GDDR6. It has been developed with the GDDR5 and GDDR5X ballouts as the baseline.

Figure 6: GDDR6 Ballout



1	2	3	4	5	6	7		8	9	10	11	12	13	14
VDD	VSS	DQ1_A	VSS	VPP			A			VPP	VSS	DQ9_A	VSS	VDD
VSS	DQ3_A	DQ2_A	DQ0_A	VDDQ			B			VDDQ	DQ8_A	DQ10_A	DQ11_A	VSS
VDDQ	EDC0_A	VSS	VDDQ	VSS			C			VSS	VDDQ	VSS	EDC1_A	VDDQ
VSS	DBI0_n _A	VSS	WCK0_t _A	WCK0_c _A			D			WCK1_c _A, NC	WCK1_t _A, NC	VSS	DBI1_n _A	VSS
VDDQ	DQ5_A	DQ4_A	VSS	VDD			E			VDD	VSS	DQ12_A	DQ13_A	VDDQ
VSS	DQ6_A	VSS	VDDQ	TMS			F			TDI	VDDQ	VSS	DQ14_A	VSS
VSS	DQ7_A	VSS	CA2_A	NC			G			CKE_n_A	CA1_A	VSS	DQ15_A	VSS
VDDQ	VDD	CA0_A	VSS	CA4_A			H			CA5_A	VSS	CA3_A	VDD	VDDQ
RESET _n	VDDQ	CA9_A	CA8_A	CABI_n_A			J			CK_t	CA7_A	CA6_A	VDDQ	ZQ_A
VREFC	VDDQ	CA9_B	CA8_B	CABI_n_B			K			CK_c	CA7_B	CA6_B	VDDQ	ZQ_B
VDDQ	VDD	CA0_B	VSS	CA4_B			L			CA5_B	VSS	CA3_B	VDD	VDDQ
VSS	DQ7_B	VSS	CA2_B	NC			M			CKE_n_B	CA1_B	VSS	DQ15_B	VSS
VSS	DQ6_B	VSS	VDDQ	TCK			N			TDO	VDDQ	VSS	DQ14_B	VSS
VDDQ	DQ5_B	DQ4_B	VSS	VDD			P			VDD	VSS	DQ12_B	DQ13_B	VDDQ
VSS	DBI0_n _B	VSS	WCK0_t _B, NC	WCK0_c _B, NC			R			WCK1_c _B	WCK1_t _B	VSS	DBI1_n _B	VSS
VDDQ	EDC0_B	VSS	VDDQ	VSS			T			VSS	VDDQ	VSS	EDC1_B	VDDQ
VSS	DQ3_B	DQ2_B	DQ0_B	VDDQ			U			VDDQ	DQ8_B	DQ10_B	DQ11_B	VSS
VDD	VSS	DQ1_B	VSS	VPP			V			VPP	VSS	DQ9_B	VSS	VDD

- The ballout is symmetric with respect to the horizontal mirror axis between rows J and K. Channel A is located in rows A to J, and channel B is located in rows K to V.
- One data byte (8 DQ, DBI_n, EDC) is located in each quadrant of the ball array and layed out to be fully symmetric. A WCK pair is embedded within each byte, with two of the four WCK pairs being optional (denoted as NC).

- Separate command interfaces are provided for each channel. The different colors used for CA[3:0] pins and the other CA pins (CA[9:4], CABI_n and CKE_n) refer to the different system configurations (that is, either 2-channel or pseudo channel mode).
- The CK clock is common to both channels. The two channels, therefore, operate at the same frequency.
- Other pins common to both channels are RESET_n, VREFC and the JTAG port.

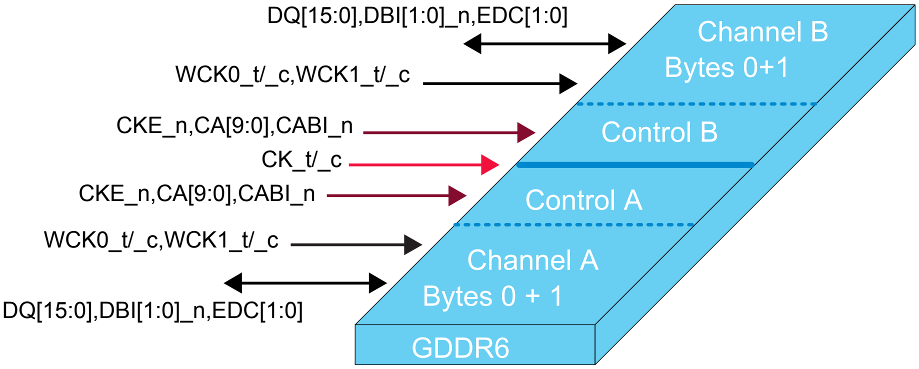
2-Channel Mode

The 2-channel operating mode of GDDR6 allows controller designers familiar with GDDR5 to view a single GDDR6 device simply as two GDDR5 devices. In this scenario, each 16-bit channel provides the same 32-byte access granularity as a single 32-bit GDDR5 device (as discussed previously).

In addition to the common features already discussed, the GDDR5 and GDDR6 command protocols also have significant commonalities in terms of command set, write data mask support with single and double mask granularity, as well as interface training commands. GDDR5's 15 CA pins are condensed into 12 CA pins with GDDR6, as GDDR6 operates all CA pins as DDR.

At the component level, as there are two separate command interfaces (one for each channel); the total number of data and control signals required to be connected to the host increases from 61 pins with GDDR5 (40 data, 15 CA, 2 CK, 4 WCK) to 74 pins with GDDR6 (2x 20 data, 2x 12 CA, 2 CK, 2x 4 WCK).

Figure 7: GDDR6 Pins in 2-Channel Mode



Pseudo Channel (PC) Mode

As an alternative to the 2-channel mode, a GDDR6 device can also be configured to operate in pseudo channel (PC) mode, providing an elegant transition option for users of GDDR5X.

The difference between PC mode and 2-channel mode is that 8 of the 12 CA pins (CKE_n, CA[9:4], CABI_n) are shared between both channels, while only the other 4 CA pins (CA[3:0]) are routed separately for each channel.

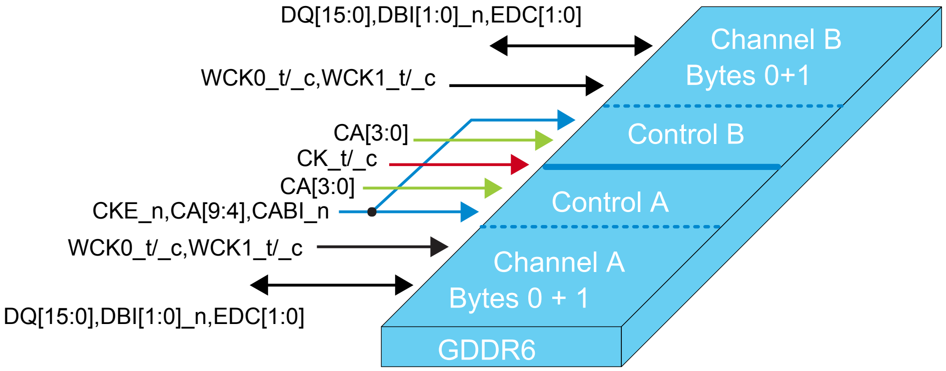
The shared pins encode the command and bank address; therefore, both channels always receive and execute the same command, similar to a GDDR5X device which has a single command interface.

The CA pins (CA[3:0]) routed independently provide the column address, allowing the host to simultaneously access different column locations in the same bank and same open row in each of the two channels. This functionality is consistent with GDDR5X operation for the lower and upper 16 DQ of the 32-bit data bus.

Note that this mode is also sometimes referred to as 32 byte pseudo mode.

The benefit of this GDDR6 PC mode is a substantial saving in the number of signals required to be connected to the host. This mode requires a total of only 66 pins (2x 20 data, 8 CA + 2x 4 CA, 2 CK, 2x 4 WCK), a single CA pin and 4 WCK pins more than what is required for GDDR5X.

Figure 8: GDDR6 Pins in Pseudo Channel Mode



Comparison

The table below summarizes the differences in signal pin count.

Table 3: Signal Pin Count GDDR5 – GDDR5X – GDDR6

Feature	GDDR5	GDDR5X	GDDR6
DQ, DBI_n, EDC	40	40	2x 20 = 40
Command/address (CA)	15	15	24 (2-channel mode) 16 (PC mode)
CK	2 CK	2 CK	2 CK
WCK	4 WCK	4 WCK	2x 4 WCK

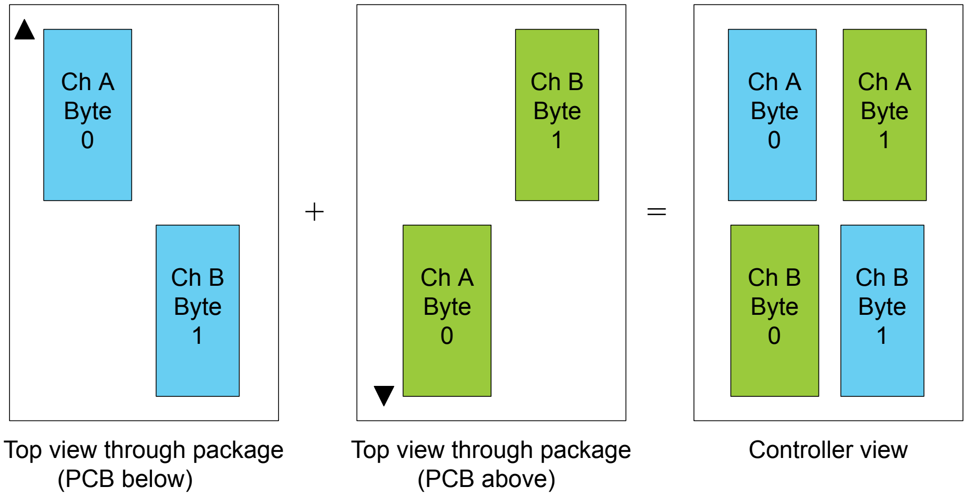
x16/x8 Configuration

GDDR5 and GDDR5X SGRAMs support a x16 mode in which two devices are assembled on the top and bottom on the PCB in a clamshell configuration. This x16 mode is an elegant way to double the density per memory channel as seen by the memory controller (host), as the two devices are connected to the same command/address bus and therefore always work in sync. The 32-bit data bus from the memory controller is split, with 16 bits routed point-to-point (P2P) to the top and bottom devices. This x16 mode is configured at boot time.

GDDR6 supports the exact same configuration option, except the name has changed from x16 mode to x8 mode, following the dual-channel architecture of GDDR6. In x8 mode, only one of the two data bytes per channel is enabled (byte 0 of channel A and byte 1 of channel B), while the other two data bytes are disabled. GDDR6 does not require a mirror function (MF) pin for this purpose, thanks to the dual-channel architecture.

The figure below illustrates how two GDDR6 devices in x8 mode are connected to give the memory controller the same view as a single GDDR6 device in normal (x16) mode:

Figure 9: GDDR6 Clamshell (x8) Configuration



- For memory channel A, byte 0 comes from the top device and byte 1 from the bottom device.
- For memory channel B, byte 0 comes from the bottom device and byte 1 from the top device.

The bottom device is flipped over and the two bytes of that device are logically swapped between the two channels. Channel A and B command/address pins are routed point-to-two-point (P22P) to both devices, preferably by simple vias.

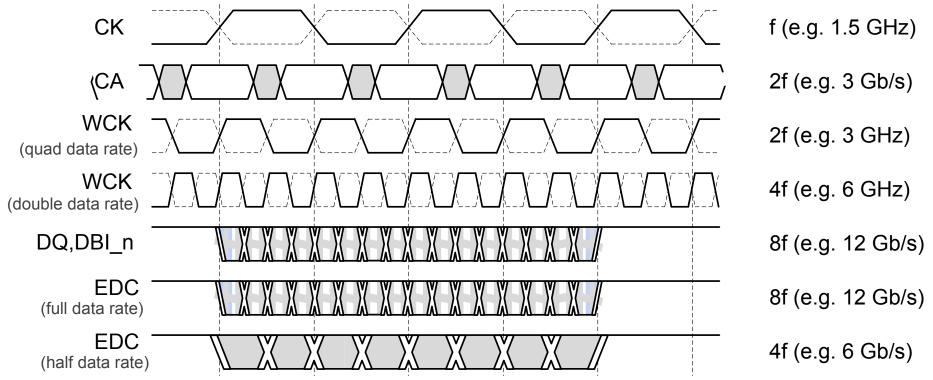
Data Clocking

GDDR6 provides two options related to data clocking (WCK).

QDR and DDR WCK Clocking

The first option determines the WCK clock frequency, as illustrated in the figure below.

Figure 10: WCK Clocking Frequency and EDC Pin Data Rate Options



- The quad data rate (QDR) operation option uses a PLL inside the GDDR6 SGRAM to double the internal clock rate. This mode is identical to the WCK clocking scheme of GDDR5X and, therefore, silicon-proven to work at the highest possible data rates. It does not require the designer, therefore, to route an 8 GHz WCK clock with lowest possible duty cycle distortion between the host and DRAM.
- The double data rate (DDR) operation option does not use a PLL and is primarily intended for mid and low data rates where a relatively small amount of WCK duty cycle distortion can be tolerated and where power consumption is key.

Micron's GDDR6 operates in QDR WCK mode at high data rates and DDR WCK mode at mid and low data rates. The selection of operating mode is made via a bit in a mode register. Details of the valid data rate ranges for each operating mode is available in the GDDR6 data sheets.

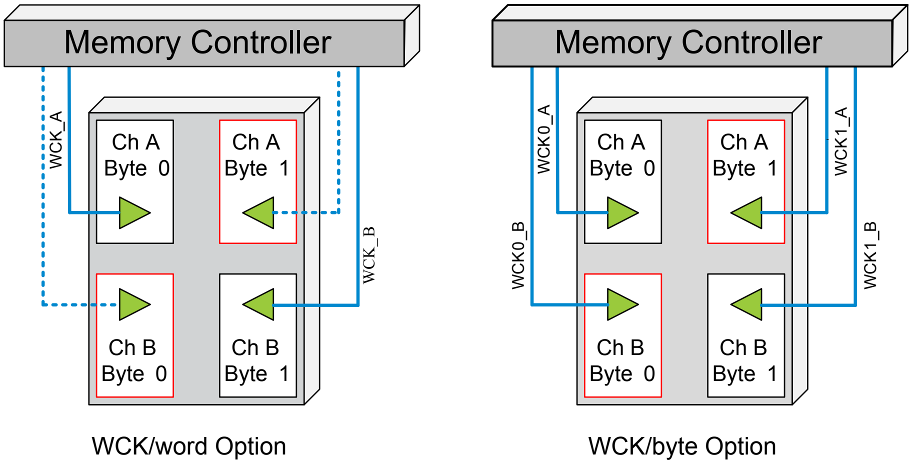
One other option related to the data rate on the EDC pin is discussed later in this technical note. This selection is again made via a mode register bit.

WCK Granularity

The second WCK option determines whether a WCK input pair is provided per word (2 bytes) or per byte as illustrated in the figure below. This allows the DRAM vendor to optimize each chip design to balance both performance and power consumption.

- The functionality of the WCK/word option is identical to GDDR5 and GDDR5X; that is, a WCK clock pair serves two bytes. Optimized circuit design and layout ensure that the clocking characteristics (for example, jitter, duty cycle distortion) are the same across both bytes, despite the asymmetric WCK ball locations of the GDDR6 package (pins D-4, D-5 for channel A, pins R-10, R-11 for channel B).
- The WCK/byte option provides a WCK pair for each byte.

Figure 11: WCK/word and WCK/byte Options



A PCB design can readily support both options as long as it is designed to support the WCK/byte option. If a GDDR6 device supports WCK/word, the host can simply deactivate the two unused WCK pairs.

The GDDR6 device indicates the support of WCK/word or WCK/byte during device initialization.

Write Data Capture

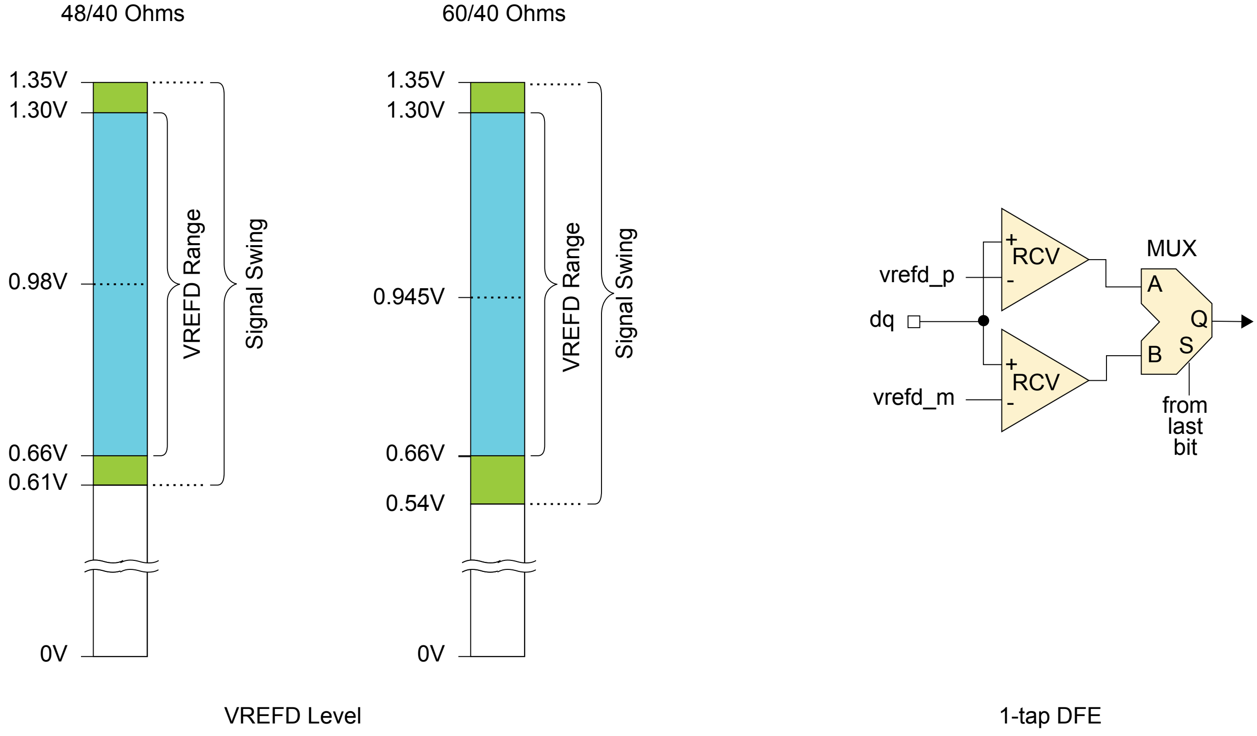
The ultra-high data rates of GDDR6 require an increased focus on the optimization of data transmission in all aspects of the interface, memory controller, interconnect (PCB) and DRAM design.

One critical circuit is the GDDR6 SGRAM's data receiver. The actual receiver design is Micron proprietary; however, some related functions have been incorporated into the GDDR6 standard to support the requirements for these high speed input receivers:

- GDDR6 allows the data receiver reference voltage (V_{REFD}) to be set individually for each data input, whereas GDDR5X provided this functionality only on a per-byte basis. The per-pin V_{REFD} capability reflects the fact that, for example, cross coupling of individual lanes may vary, resulting in a slightly different vertical center of the data eye per lane. The step size has been defined as 0.5% of V_{DDQ} (6.75mV) and the total range further enlarged as shown in the figure below. Two different termination strengths are also supported, 60/40 ohms (as per GDDR5X) and 48/40 ohms.
- Inter System Interference (ISI) on the channel may result in the data eye being closed at the receiver. GDDR6 adds a so-called 1-tap decision feedback equalization (DFE), which helps open the eye and reliably detect a symbol depending on whether the previous symbol was a 1 or a 0.

Both V_{REFD} and DFE are programmed by mode register bits. The DFE setting selects different (positive and negative) V_{REFD} levels for the two cases.

Figure 12: V_{REFD} and DFE

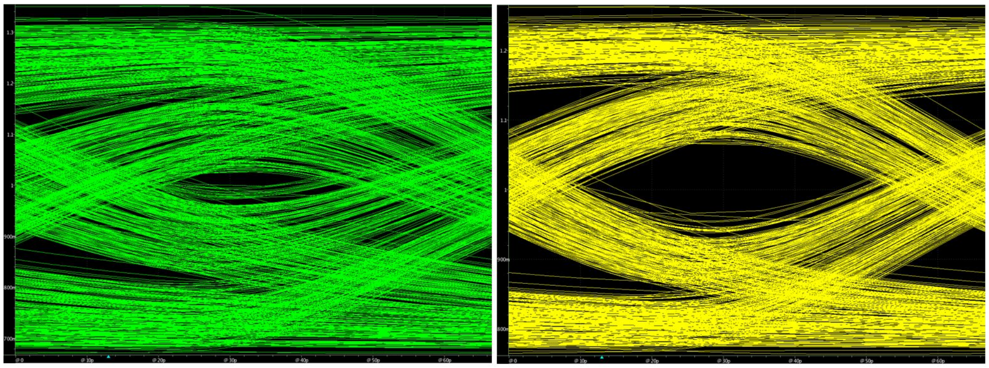


VREFD Level

1-tap DFE

The two images in the figure below illustrate how DFE opens a data eye that would otherwise be (nearly) closed:

Figure 13: Simulated Input Data Eye with and without DFE



- The green data eye shows the simulated data eye at the DQ pad over a typical PCB channel with no DFE.
- The yellow data eye shows the simulated data eye at the DQ pad over the same typical PCB channel with DFE enabled.

Other Input Signal-Related Features

GDDR6 also provides additional features to enhance high speed signaling:

- The CK clock can be configured to be terminated externally on the PCB (as per GDDR5) or internally terminated via ODT. This selection is made at power-up.
- The input reference voltage for the CA pins can be derived from an external resistor divider, or it can be generated internally. The selection is made at power-up, like with GDDR5X. The internal V_{REFC} can further be offset via mode register bits, which is new for GDDR6. The internal V_{REFC} is expected to provide better tracking with variations of V_{DDQ}.
- Data and WCK termination can also be programmed separately. GDDR6 adds a 48 ohm option for the data inputs to allow for better impedance matching to the channel.

Output Driver-Related Features

Output driver equalization function (TX EQ) has also been added to enable the programming of output drivers to better match system channel characteristics.

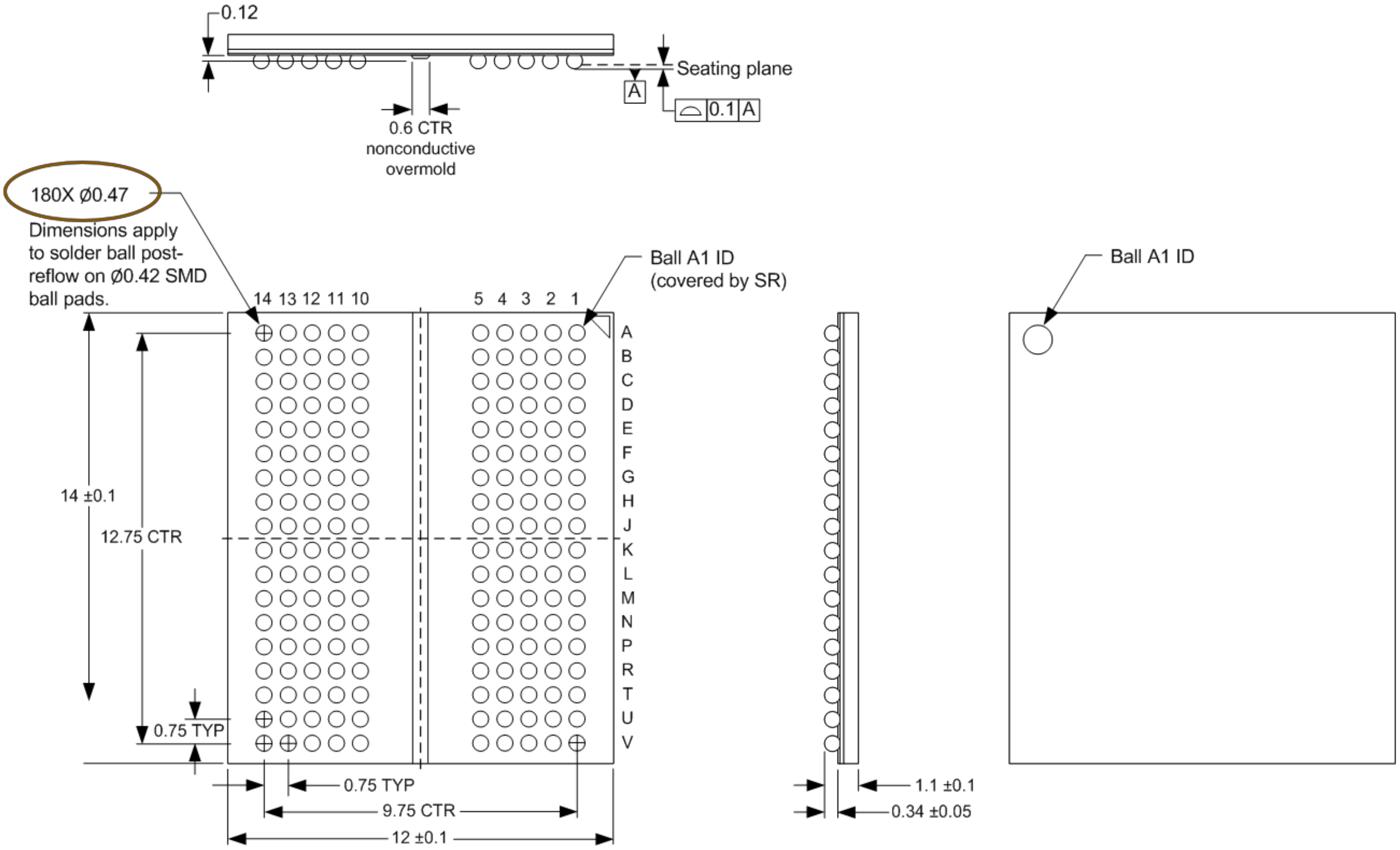
Package

GDDR6 SGRAMs support the same 14mm x 12mm fine-pitch ball grid array (FBGA) package as GDDR5, with the change that the ball count has been increased from 170 to 180, and ball pitch reduced from 0.8mm to 0.75mm.

The increased ball count allows 9 additional CA pins and 4 (optional) additional WCK pins without sacrificing the balanced supply pin budget. The reduced ball pitch balances the desire to keep the same package outline as GDDR5 while maintaining a pitch conducive to PCB manufacture cost effectiveness.

Micron will use the same solder ball and SMD pad opening sizes for GDDR6 devices, and so anticipates PCB-level reliability consistent with that seen with GDDR5.

Figure 14: Micron GDDR6 Package Outline



Miscellaneous Features

As mentioned previously, GDDR6 has been specified with the intent of carrying over proven features from GDDR5 and GDDR5X. The specification also adds new features to enhance functionality; for example, to provide additional flexibility for the host memory controller.

Command/Address Bus Inversion (CABI), Data Bus Inversion (DBI)

V_{DDQ} terminated signaling means that each signal line driven LOW draws a static current of 13.5mA (1.35V / (60+40) ohm); however, on the contrary, a signal line driven HIGH draws no static current.

The bus inversion functions, therefore, reduce the number of signal lines being driven LOW by either sending the data inverted or non-inverted. The result is not only lower average system power consumption, but also, crucially, much lower supply noise.

GDDR6 maintains the data bus inversion (DBI) function of GDDR5 and GDDR5X. It also maintains the address bus inversion function, which is now called command/address bus inversion (CABI) as it now covers all command and address (CA) pins.

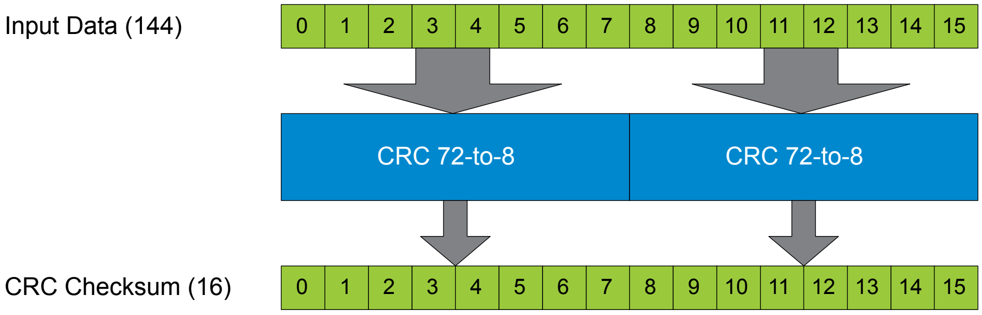
Data Link Protection (CRC)

GDDR6 also provides link protection for the data bus (DQ, DBI_n, EDC) as per GDDR5 and GDDR5X. The memory controller can select between two options:

- Full data rate: This option creates a 16-bit checksum per write or read burst, which is the concatenation of two 8-bit checksums calculated over the first 8 UI (0 to 7) and second 8 UI (8 to 15) of the burst (see the figure below). This option uses the same 0x83 8-bit CRC polynomial as GDDR5.
- Half data rate: This option creates an 8-bit checksum per write or read burst. The checksum uses a similar polynomial as the full data rate option to calculate two intermediate 8-bit checksums, but then compresses these two into a final 8-bit checksum.

Both options provide excellent error detection rates (for example, 100% fault detection for random single, double and triple bit errors, and >99% fault detection for other random errors).

Figure 15: GDDR6 Full Data Rate CRC



Refresh Options

All DRAM devices store information in tiny capacitor-based cells which require regular refreshes to maintain data. Refreshes traditionally require the controller to suspend all READ and WRITE operations, close any open pages, issue a REFRESH (REFab) command, and finally re-open the banks and resume READ and WRITE operations (shown in the figure below, Case 1). The REFab command refreshes one or more pages in all banks simultaneously (hence it is also referred to as *all-bank refresh)*.

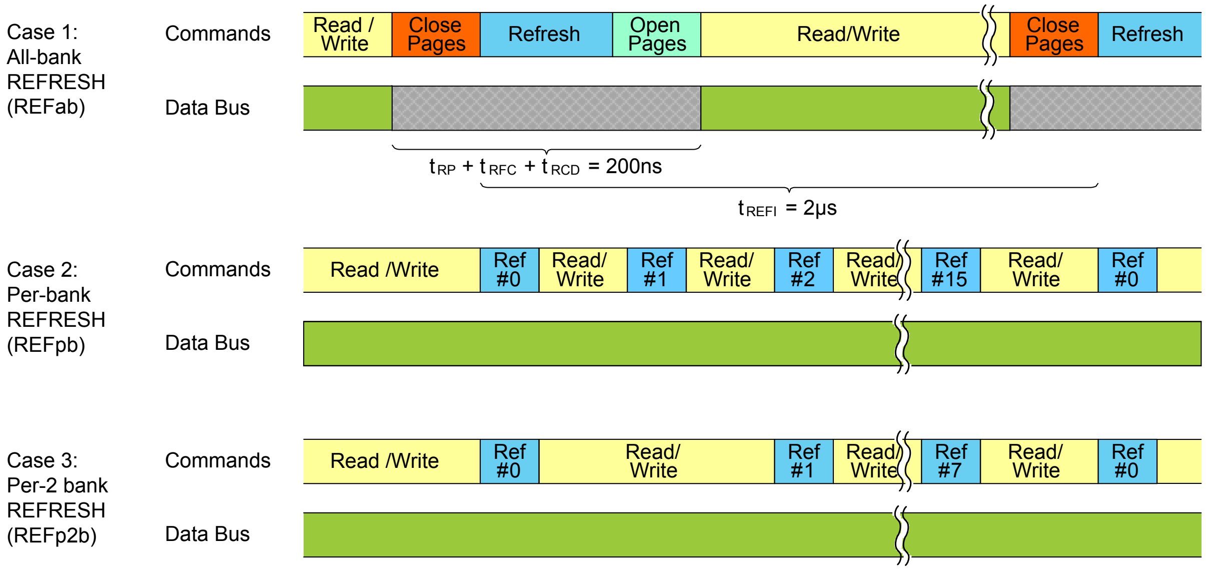
The interval required for each refresh cycle is device- (density-) specific. Taking 200ns as one example along with a 2µs interval between two all-bank REFRESH operations, it becomes apparent that refreshes consume around 10% of the DRAM's overall command bandwidth.

GDDR6 offers two additional mechanisms for issuing REFRESH commands:

- Per-bank REFRESH (REFpb; Case 2 in the figure below): This option refreshes a single bank only, while READ and WRITE operations to other unaffected banks may continue. 16 REFpb commands must be issued within the 2µs t_{REFI} interval.
- Per-2-bank REFRESH (REFp2b; Case 3 in the figure below): This option refreshes two of the 16 banks, while READ and WRITE operations to other unaffected banks may continue. Eight REFp2b commands must be issued within the 2µs window.

Theoretically, REFpb and REFp2b should be able to fully hide all refreshes. Practically, there will still be some performance impact due to command slots being used for the REFpb or REFp2b command, and the fact that a bank being refreshed is not available for reads and writes. Nevertheless, these command are an important tool to enable increased sustained memory bandwidth. The system designer is free to select the best option for their design.

Figure 16: Refresh Options



IEEE1149.1 Boundary Scan (JTAG)

GDDR6 supports an IEEE1149.1-compliant boundary scan. The feature was introduced with GDDR5X and can also be found in GDDR6.

Boundary scan allows testing of the interconnect on the PCB during manufacturing using state-of-the-art automatic test pattern generation (ATPG) tools.

In addition, Micron's GDDR6 devices have proprietary features associated with boundary scan that are beneficial to system qualification and debug.

Low Power Features and Energy Efficiency

Energy efficiency is of vital importance to all modern system designs and GDDR6 looks to support this goal by adding a number of energy saving features.

Supply Voltage

Initially, GDDR6 SGRAMs were specified to operate with a 1.35V V_{DD}, V_{DDQ} supply for graphic cards or game consoles.

Micron has extended these plans by enabling GDDR6 devices with a 1.25V V_{DD}, V_{DDQ} supply for applications such as networking and automotive.

Low Power Features

GDDR6 SGRAM supports all low power features of GDDR5 and GDDR5X SGRAMs, and adds new features to address the increasing importance of energy efficiency:

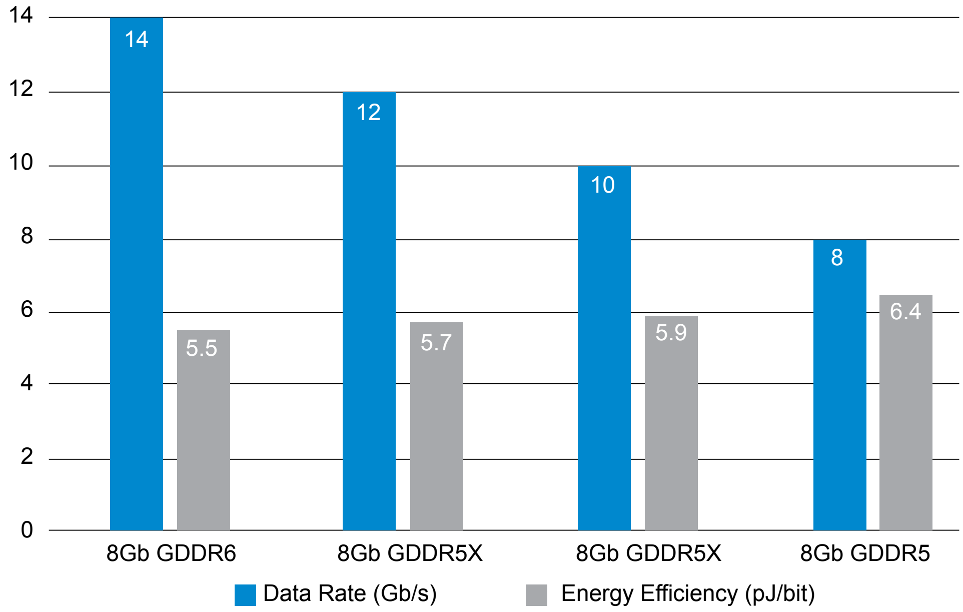
- Devices can operate over a contiguous frequency range starting at 50 MHz (equivalent to 400 Mb/s), up to the maximum specified data rate. This enables the memory controller to throttle the clock frequency whenever the current workload does not require the maximum bandwidth.
- For data rates below 2 Gb/s, the devices may be set into a strobe mode, which returns a data strobe like a conventional RDQS along with read data but keeps the EDC pin quiet during idle periods. This feature saves power on the device and also leads to lower power consumption at the memory controller.
- On-die termination for all high-speed inputs can be set to higher impedances to reduce the static ODT currents. These input terminations can also be turned off when appropriate.
- The device may be put into a power-down state during idle periods.
- For further power reduction during idle states, the device may be put into self refresh mode. During self refresh, the GDDR6 SGRAM retains the memory content without external clocking. This not only saves power inside the DRAM, but also at the memory controller.
- Power consumption during self refresh mode can further be reduced by enabling an internal temperature sensor. This sensor enables the device to adjust the refresh rate based on device temperature.
- Ultimate lowest power consumption is achieved by setting the device into so-called hibernate self refresh. Hibernate self refresh allows the DRAM to turn off additional circuits within the device, at the expense of a larger wake-up time which is similar in duration to the time required for device initialization.

Energy Efficiency

All features described above enable GDDR6 to be significantly more energy efficient than GDDR5.

The chart in the figure below compares the energy efficiency of GDDR5X- and GDDR5-based graphic cards. Micron internal measurements on early GDDR6 components confirm that GDDR6 will be even more energy efficient.

Figure 17: Energy Efficiency Comparison



References

- JESD250 Graphics Double Data Rate (GDDR6) SGRAM standard
- Micron 8Gb GDDR6 SGRAM data sheet
- Micron GDDR5X SGRAM Technical Note (TN-ED-02)

Revision History

Rev. A – 11/17

- 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