Doubling I/O Performance with PAM4

Micron® Innovates GDDR6X to Accelerate Graphics Memory

Micron and Graphics Memory

Graphics double data rate type six (GDDR6) synchronous dynamic random-access memory (RAM) was introduced in 2018 as a natural evolution of GDDR5X and became a widely accepted standard not only for high-performance gaming and workstation-class graphics, but also for automotive, high-performance computing (HPC) and networking applications. GDDR6, synchronous graphics RAM (SGRAM) has been regarded as the fastest discrete memory offered by Micron to date, supporting per-pin data rates up to 16 Gigabits per second (Gb/s). With GDDR6X SGRAM Micron challenges the boundaries again by targeting per-pin data rates of 21 Gb/s and beyond.

GDDR6X (Figure 1) preserves the same data access granularity and form factor as GDDR6, so the new device easily integrates into the user's existing memory ecosystem for a fast and low-risk upgrade.

This technical brief highlights the **new features** introduced in GDDR6X, while the characteristics that remain unchanged are not covered. Please refer to the Highest Performance, Highest Bandwidth GDDR6 flyer for extended information on those topics.

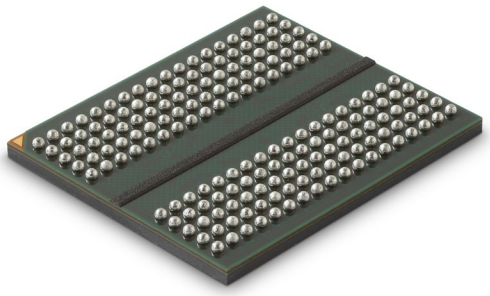

Figure 1: Micron GDDR6X SGRAM

Micron GDDR6X Fast Facts

Micron GDDR6X graphics memory doubles input/output (I/O) performance while minimizing the cost of memory. Working with AI-innovation leader NVIDIA®, Micron delivers higher bandwidth by enabling multi-level signaling in the form of four-level pulse amplitude modulation (PAM4) technology in this memory device.

The new GDDR6X SGRAM:

- Doubles the data rate of SGRAM at a lower power per transaction while achieving up to 1 Terabyte per second (TB/s) in system memory bandwidth;
- Is the first discrete graphics memory device that employs PAM4 encoded signaling between the processor and the DRAM, using ***four*** voltage levels to encode and transfer two bits of data per interface clock edge.
- Can be designed and operated stably at high speeds, and built in mass-production.

Micron and NVIDIA:

By bringing parallel processing to general-purpose computing, NVIDIA's GPUs evolved from gaming devices to accelerate HPC and AI. Micron and NVIDIA have a proven track record of successful collaboration. Starting with GDDR5X, Micron and NVIDIA broke performance boundaries for gaming with the successful GeForce® GTX 1080 Ti GPU, enabled by 11GB of GDDR5X. This technology was so successful that it drove the specification of the next JEDEC standard, GDDR6. Now Micron and NVIDIA have completely reimagined the memory/GPU interface to break more barriers.

From GDDR5 to GDDR6X

Micron has a long history of innovation in graphics memory. Table 1 compares the major features of GDDR5, GDDR5X, GDDR6 and GDDR6X. The benefits of GDDR6X's new features and improvements are discussed later in this technical note.

Table 1: Feature Comparison GDDR5 – GDDR5X – GDDR6 – GDDR6X

Feature	GDDR5	GDDR5X	GDDR6	GDDR6X
Density	From 512Mb to 8Gb	8Gb	8Gb, 16Gb	8Gb, 16Gb[1]
V_{DD} and V_{DDQ}	Either 1.5V or 1.35V	1.35V	Either 1.35V or 1.25V	Either 1.35V or 1.25V
V_{PP}	N/A	1.8V	1.8V	1.8V
Data rates (per pin)	Up to 8 Gb/s	Up to 12Gb/s	Up to 16 Gb/s	19 Gb/s, 21 Gb/s, >21 Gb/s[1]
Channel count	1	1	2	2
Access granularity	32 bytes	64 bytes 2x 32 bytes in pseudo 32B mode	2 ch x 32 bytes	2 ch x 32 bytes
Burst length	8	16 / 8	16	8 in PAM4 mode[2] 16 in RDQS mode
Signaling	POD15/POD135	POD135	POD135/POD125	PAM4 POD135/POD125
Package	BGA-170 14mm x 12mm 0.8mm ball pitch	BGA-190 14mm x 12mm 0.65mm ball pitch	BGA-180 14mm x 12mm 0.75mm ball pitch	BGA-180 14mm x 12mm 0.75mm ball pitch
I/O width	x32/x16	X32/x16	2 ch x16/x8	2 ch x16/x8[3]
Signal count	61 - 40 DQ, DBI, EDC - 15 CA - 6 CK, WCK	61 - 40 DQ, DBI, EDC - 15 CA - 6 CK, WCK	70 or 74 - 40 DQ, DBI, EDC - 24 CA - 6 or 10 CK, WCK	70 or 74 - 40 DQ, DBI, EDC - 24 CA - 6 or 10 CK, WCK
PLL, DCC	PLL	PLL	PLL, DCC	DCC
CRC	CRC-8	Modified CRC-8	2x CRC-8	2x CRC-8[4]
V_{REFD}	External or internal per 2 bytes	Internal per byte	Internal per pin	Internal per pin 3 sub-receivers per pin
Equalization		RX/TX	RX/TX	RX/TX
V_{REFC}	External	External or Internal	External or Internal	External or Internal
Self refresh (SRF)	Yes Temp. Controlled SRF	Yes Temp. Controlled SRF Hibernate SRF	Yes Temp. Controlled SRF Hibernate SRF VDDQ-off	Yes Temp. Controlled SRF Hibernate SRF VDDQ-off
Scan	SEN	IEEE 1149.1 (JTAG)	IEEE 1149.1 (JTAG)	IEEE 1149.1 (JTAG)

[1] Planned
[2] PAM4 encodes two data bits per UI
[3] Configured at power-up
[4] GDDR6X: half data rate cyclic redundancy check jedec.org scheme

Defying the GDDR6 Boundaries

The Micron GDDR6 architecture benefited from several enhancements first incorporated within GDDR5X. Among other things, challenges associated with DRAM array timing were mitigated through a doubling of the data prefetch. While this functionality relaxed the demands on the memory array, ever-increasing off-chip signaling speeds shifted the burden to the high-speed I/O and clocking schemes.

At GDDR6 per-pin data rates (e.g., 16Gb/s), the available timing window for reliably transmitting and capturing data reduces to 62.5ps (picoseconds) or less. As a result, operating at these frequencies requires additional levels of circuit precision and complexity, not to mention higher toggling rates, all leading to increased power usage. Given that complexity and power will soon outpace the achievable per-pin data rate, GDDR6 top speed today is at 16Gbps and exceeding that is very challenging and comes at the expense of reduced system margin.

To address the timing challenges associated with GDDR6, next-generation Micron GDDR6X memory replaces the existing binary signaling interface of GDDR6 (PAM2, often also referred to as non-return-to-zero or NRZ) with a PAM4-enabled scheme. Encoding 2 bits of data into every transmitted data symbol doubles the effective bandwidth for a given operating frequency. Stated another way, when supporting a common per-pin data rate, GDDR6 circuits must operate twice as fast as GDDR6X circuitry. As a result, the high-speed circuit techniques developed for GDDR6 are more than adequate to take GDDR6X far beyond the present GDDR6 target of 16 Gb/s, while simultaneously lowering I/O power consumption.

Figure 2 depicts how the same amount of data is transferred across the GDDR6X interface (below) at half the frequency when compared to the GDDR6 (above).

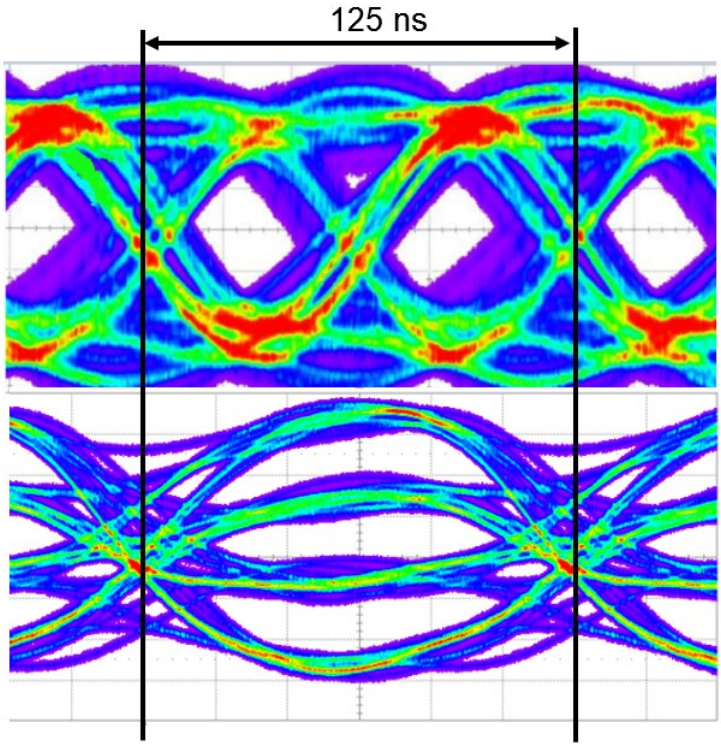

Figure 2: Data Eye Comparison Between GDDR6 (top) and GDDR6X (bottom) That Shows the Timing for a 2 Bits Data Transfer at 16Gb/s

Table 2: Each of the Four Physical Levels, or Symbols, of PAM4 Represents 2 Bits of Data

Logical	Symbol	Physical
10	+3	
11	+1	
01	-1	
00	-3	

With PAM4 signaling, the channel transmits 2 data bits per cycle using 4 distinct signal levels (Table 2). Each one of these levels is referred to as a symbol, and the data transfer rate is therefore expressed in symbols per second or bauds. The 2 bits per unit interval (UI) are gray-coded to ensure that any transmission error affects only one of the 2 bits within a symbol.

Figure 3 shows how the same amount of data transferred using PAM4 encoding needs half the interface cycles as compared to NRZ. As GDDR6 burst length is 16, to keep data granularity compatible, GDDR6X equivalent burst length is set to 8.

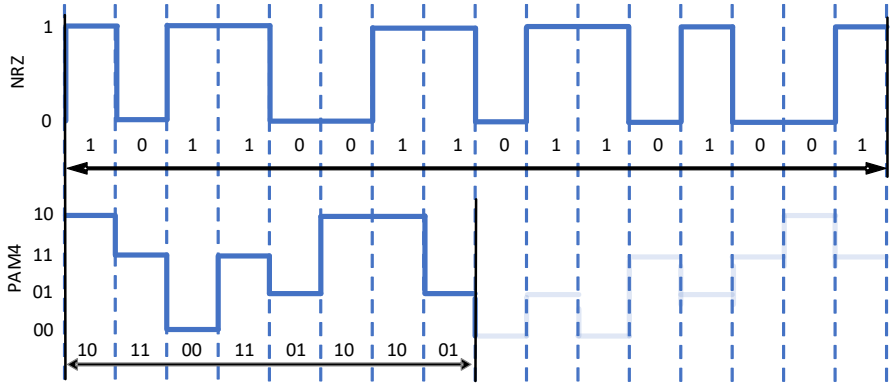

Figure 3: Example of Coding and Transfer of the Same Amount of Data (One Burst) for NRZ and PAM4 Interfaces

Write Data Capture

To enable PAM4, the receiver must be able to accurately sample and resolve the various levels of the multilevel input signal. For that purpose, the GDDR6X implements three input sub-receivers per I/O and data bus inversion (DQ/DBI) pins as illustrated in Figure 4 below.

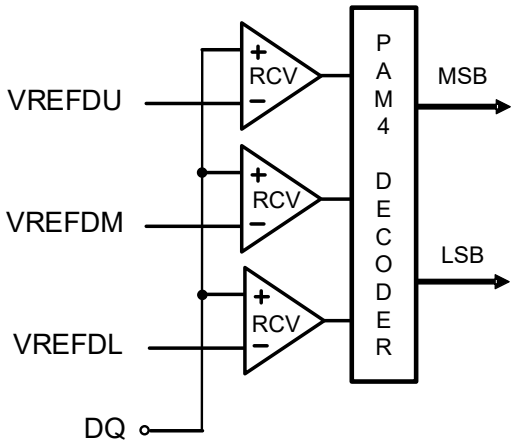

Figure 4: PAM4 Receiver

The reference voltages are generated internally and can be programmed separately for each sub-receiver of each input pin. They account for the data termination value considering that the GDDR6X on-die termination (ODT) allows for terminations of either 40Ω or 48Ω.

The VREFD levels are linear, with a total range of 64 steps. The midpoint for each receiver has been set to match its ideal vertical data eye center based on the PAM4 signaling. The reference voltages VREFDU, VREFDM and VREFDL (for the comparator at the receiver for the upper, middle and lower data eye of the PAM4 signal) are programmed via mode register bits, giving the host the ability to fine tune the VREFD levels, preferable but not exclusively, during the write training sequence.

Output Driver

For generations, DRAMs have used multi-leg architectures to support different driver strengths depending on how many legs are enabled in parallel. The same concept can be used to form several intermediate voltages levels through the pull-up (PU) and pull-down (PD) voltage divider (Figure 5 shows impedances).

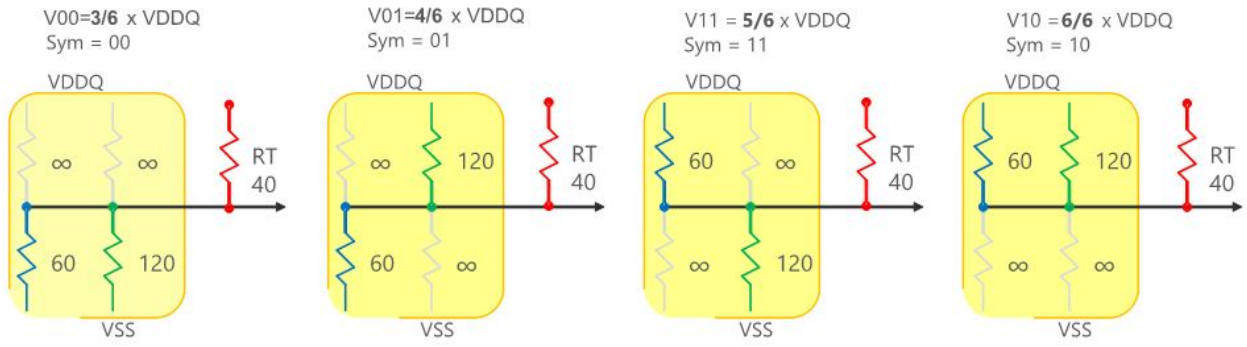

Figure 5: PAM4 Impedance Scheme

The PAM4 driver forms a constant 40Ω output impedance with four uniform levels which drive into a receiver-side termination resistor tied to V_{DDQ}. Figure 4 shows how to achieve that through four distinct combinations of 120Ω (PU or PD) and 60Ω (PU or PD) legs. A termination impedance might impact the overall swing, but it does not affect the relative signal levels. Note that all those pull-up and pull-down impedances may be derived from 120Ω legs.

Data Clocking

From the clocking perspective, GDDR6X supports two different operating modes that differ in the DQ/DBI/EDC pin to WCK ratio. Figure 6 illustrates the differences between both modes.

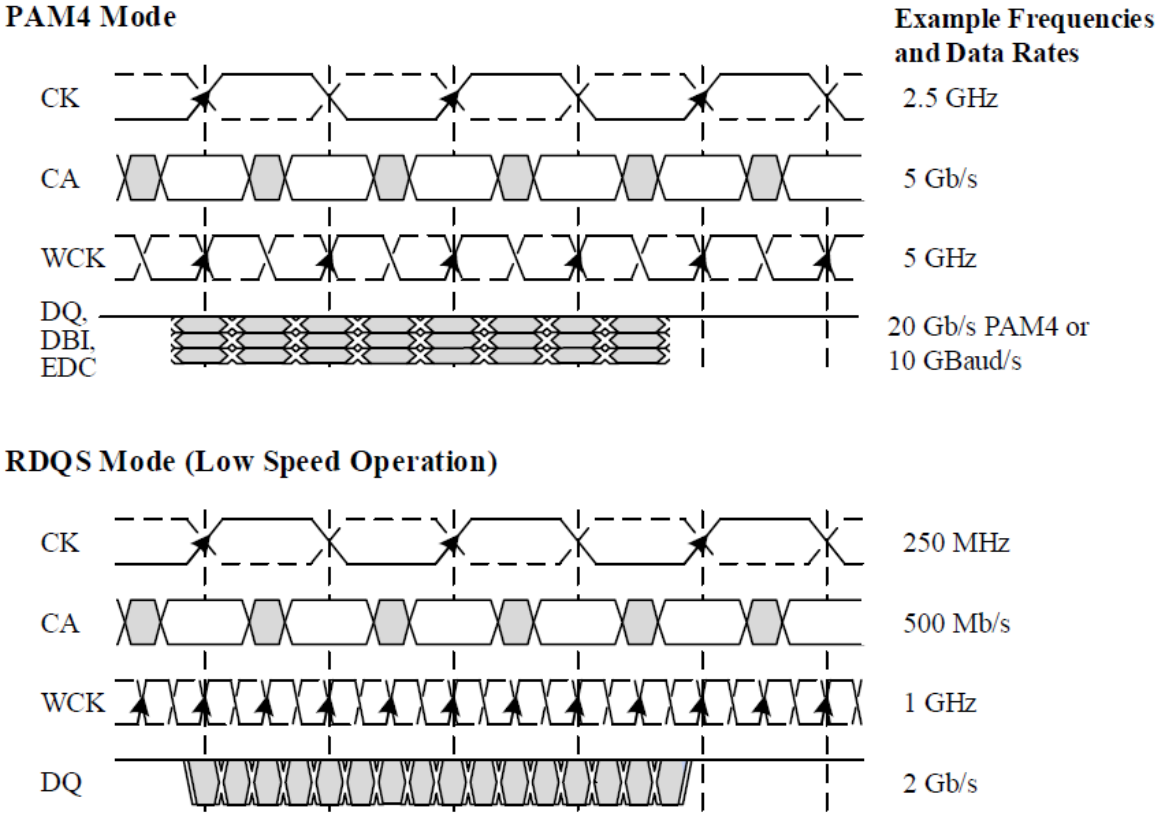

Figure 6: Clocking and Interface Relationship

While Figure 6 does not show a detailed timing diagram, it demonstrates the command/data bus clocking of a burst access, from which we can infer:

- In PAM4 mode the clock frequency is the same as the command clock:
  - The driven data is 2 UI wide, 8 symbols containing 2 bits of data in 2 command clock cycles, resulting in a burst length (BL) of 8.

The PAM4 interface is suited for high-speed operation; for data rates below 3Gbps, read strobe mode (RDQS) may be used to reduce the power consumption at the memory controller side.

- In RDQS mode, the clocking is equivalent to GDDR6:
  - The clock frequency is twice the frequency of the command clock.
  - The output is 1 UI wide, 16 bits in two command clock cycles (BL16).
  - The data signaling is NRZ encoded, aka PAM2.

(A detailed timing diagram that shows how the data burst of the two modes compare is presented later in the Memory Write and Read operations section.)

Memory Write and Read Operations

Figure 7 illustrates the memory array prefetch in the form of timing diagrams, where two seamless read accesses are shown for GDDR5, GDDR5X, GDDR6 and GDDR6X.

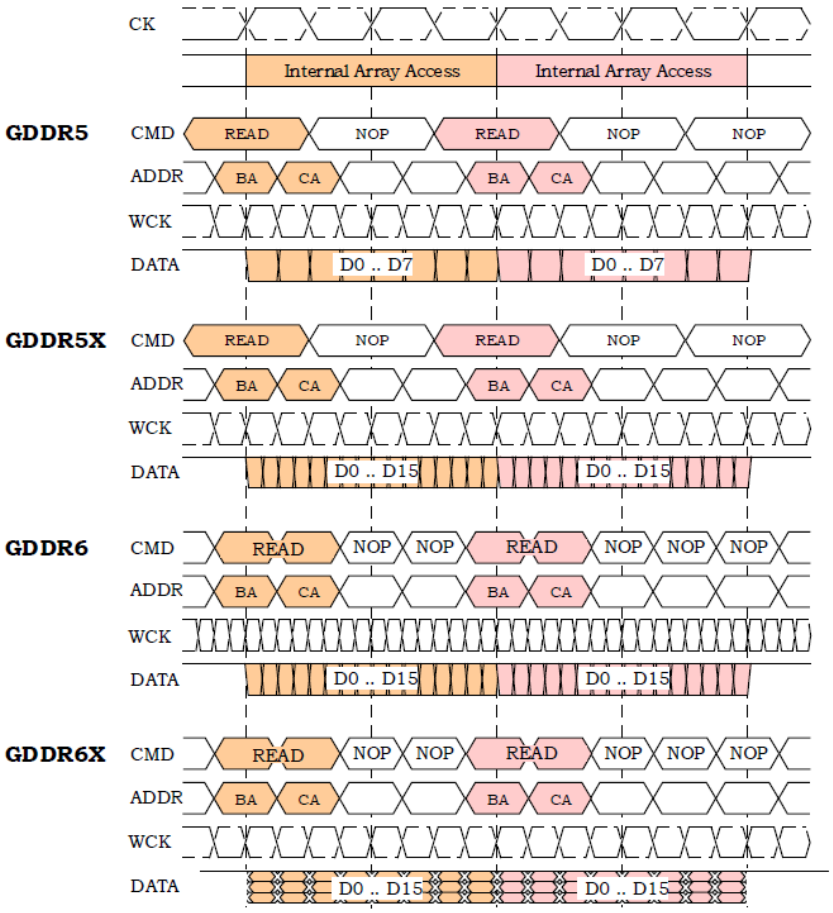

Figure 7: Seamless READs with GDDR5, GDDR5X, GDDR6 and GDDR6X

From the timing diagram in Figure 7 we can easily appreciate the main differences and similarities between the different standards with regard to memory access operation timings:

- With all four standards, internal write and read accesses are two clock cycles long (tCCD = 2 tCK). A 100% bus utilization is achieved when a WRITE or READ is issued every second cycle (e.g. READ - NOP - READ).
- GDDR6X achieves the same data throughput as GDDR6 while requiring just half of the WCK frequency.
- GDDR6X and GDDR6, receive commands and addresses as double data rate (DDR) referenced to both rising and falling CK clock edges, as opposed to GDDR5 and GDDR5X whose CA buses operate at single data rate (SDR) mode.

Power Efficiency

The boost in bandwidth is not the only advantage of GDD6X, there's also a significant improvement in power efficiency. Figure 8 demonstrates that a GDDR6X running at 21Gb/s requires 15% less power per bit transferred

than a GDDR6 running at 14Gb/s, even when the GDDR6X provides 50% more bandwidth. Note that the displayed data represents the efficiency gain calculated over the complete DRAM device power consumption. If the comparison were limited to the data interface alone, it would show the greater power/bit efficiency enabled by PAM4 signaling.

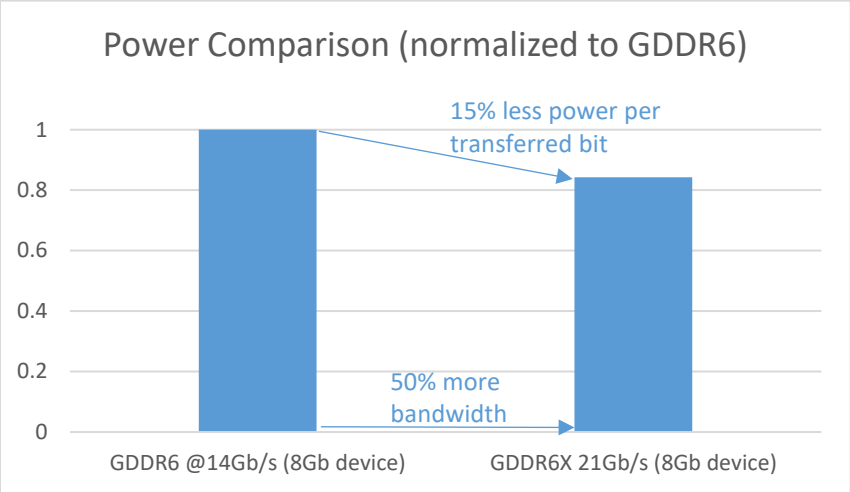

Figure 8: Power Comparison Between GDDR6 and GDDR6X, Normalized to GDDR6

Learn More

For the latest developments on Micron's graphic memory SDRAM innovation, follow us on @MicronTech on Twitter or on LinkedIn. For more details, visit micron.com/GDDR6X.

MICRON.COM