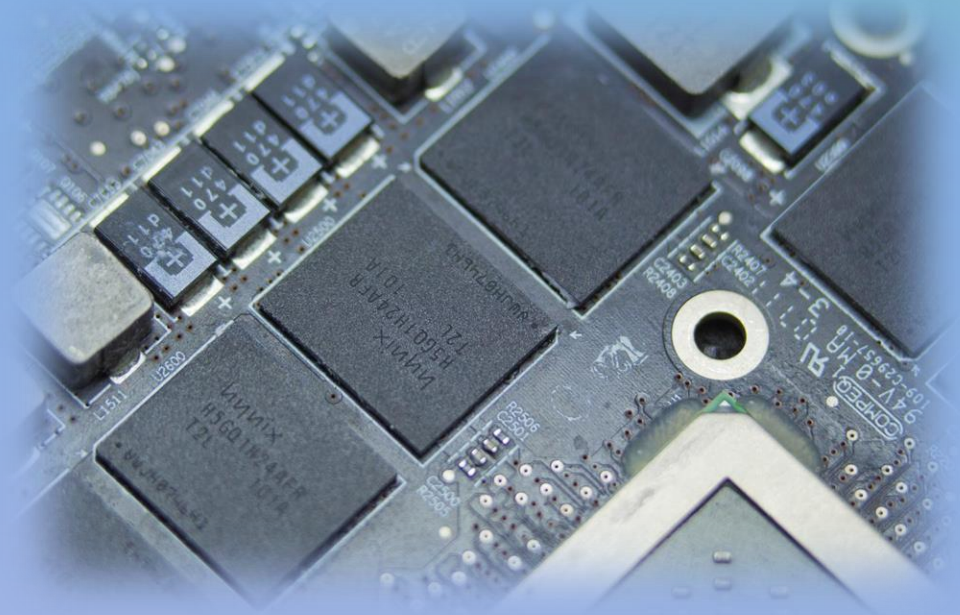
Discovering GDDR

Architectural Design Comparison with DDR



By Jason Silic

May 4, 2020

The History of GDDR

- Graphics DDR SDRAM, or GDDR, is a type of DRAM that is optimized for high bandwidth at the expense of other characteristics such as latency.

- Throughout the history of computing specialized chip have been designed with specific features to improve performance in specific workloads. In the 1980s the display driver circuitry was little more than a block of memory that would output characters or pixels to the screen.

- Typical resolutions were small, the original IBM Monochrome Display Adapter had a theoretical resolution of 720x350 and 4kB for the display buffer but could only output text characters contained in a ROM chip. [3]

- In the 1990s graphics adapters started becoming more complex, supporting hardware acceleration of common bitmap block transfer operations (“blitting”). As graphics processing became increasingly separated from the CPU the peculiar demands of storing texture data led to specialized memory requirements.

1998: First Chip Released

- Samsung released a DRAM chip in 1998 of what was known at the time as SGRAM (Synchronous Graphics RAM) that became the predecessor to today's GDDR. [1]
- GDDR smoothly developed from the existing DRAM technology and is similar enough that either type can be used for graphics or general purpose memory.
- For example, many systems today have integrated GPUs that use the system memory (e.g., DDR3) to store data for 3D graphics.
- Conversely, the Play Station 4 uses 8GB of GDDR5 memory as a pool for both CPU and GPU. [7]

2000s: Steady Development

- The GDDR standard developed in parallel with the standard DDR SDRAM. General trends visible over the last two decades include a decrease in supply voltage, a large increase in bandwidth per pin, and various power saving strategies such as DBI (Data Bus Inversion) and POD (Pseudo Open Drain) technologies.
- Process technology also played an important role, just as with CPUs. Current DRAM is manufactured on nodes below 20nm [10].
- This presentation will focus on DDR3 and GDDR5, which were both very successful contemporaneous technologies around 2015.

DDR standards

	Release Date	Notes
DDR	2000	
DDR2	2003 (Standardized)	First produced by Samsung
DDR3	2007	
DDR4	2014	
DDR5	2020?	

Data taken from [8], [9]

GDDR standards

	Release Date	Notes
GDDR	1998?	
GDDR2	2002	Developed by Samsung
GDDR3	2004	Based on DDR2
GDDR4	2006?	Short lived standard based on DDR3. Introduced Data Bus Inversion.
GDDR5	2008	
GDDR6	2018	

Data taken from [11], [12]

Major Differences

- On the next page we present a table with some of the major differences between GDDR and DDR.

- Perhaps the largest difference is the vast disparity in total bandwidth. Because graphics programs typically need to access large volumes of data (Textures) bandwidth has become a driving priority in GDDR development. As shown, GDDR memory architectures can yield almost an order of magnitude more bandwidth.

- A typical GPU workload is massively parallel. If a particular piece of data is not available it is easy to work on another area of the current problem. This characteristic make latency much less of a concern and DDR RAM typically has better latency.

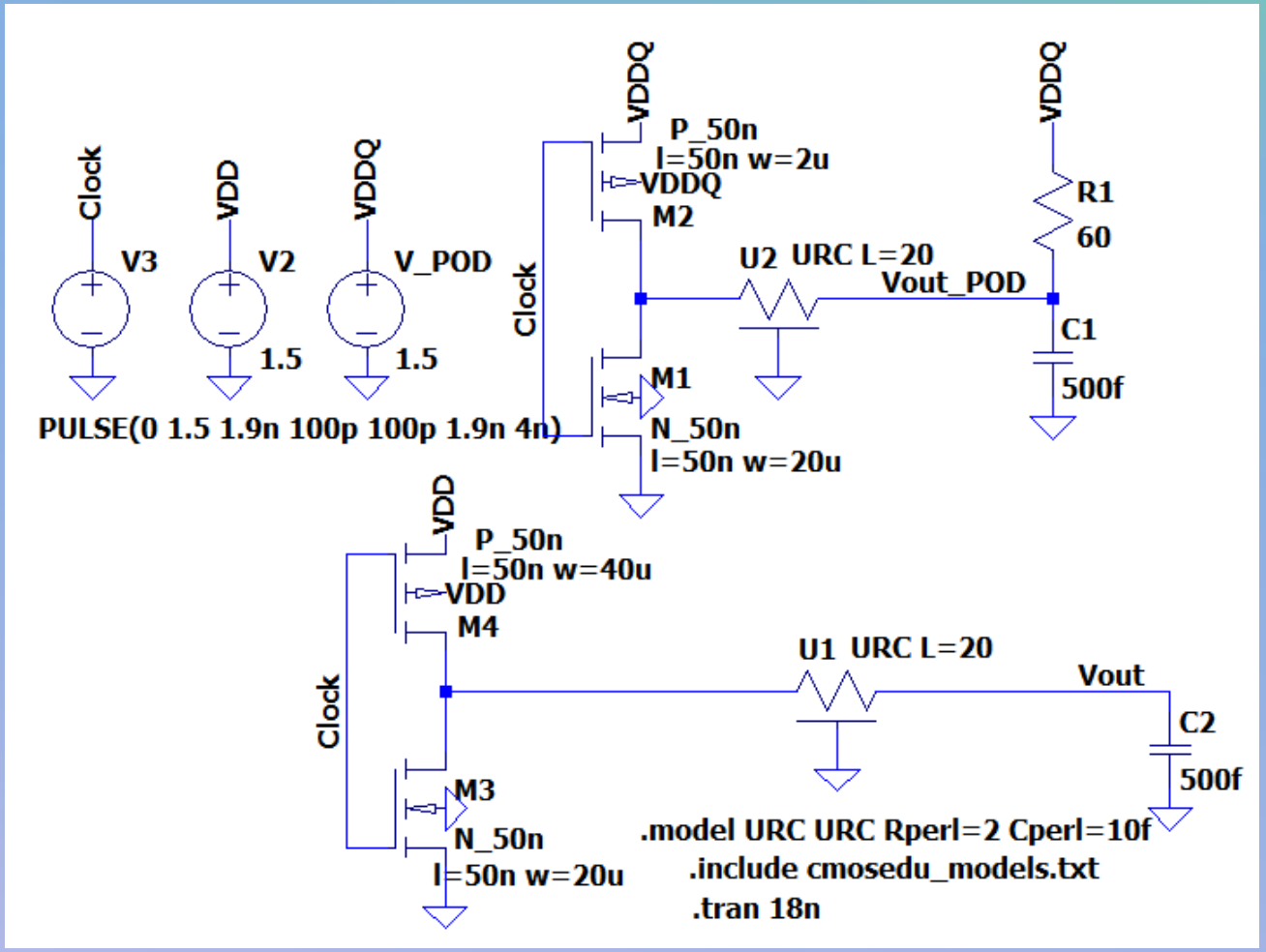
General comparison between GDDR5 and DDR3 standards.

	GDDR5	DDR3
Memory Bus Width (typical)	256-384 bits (8-12 chips)	128 bits (2 channels)
Bandwidth per pin per second	8Gbps (fast GDDR5) [6]	2.4Gbps (DDR3-2400)
Total Memory System Bandwidth	256GB/s (256 bit interface)	38.4GB/s
Typical Physical Packaging	Die soldered directly to PCB, within 3-4cm of GPU	Chips combined into DIMMs (Dual Inline Memory Module). Placed in socket on motherboard.
JEDEC standard document	Available with a free account	Available for a hefty fee
Operating Voltage (VDD)	GDDR5: 1.5V (optional 1.35V) [14]	DDR3: 1.5V (optional support for 1.35V)

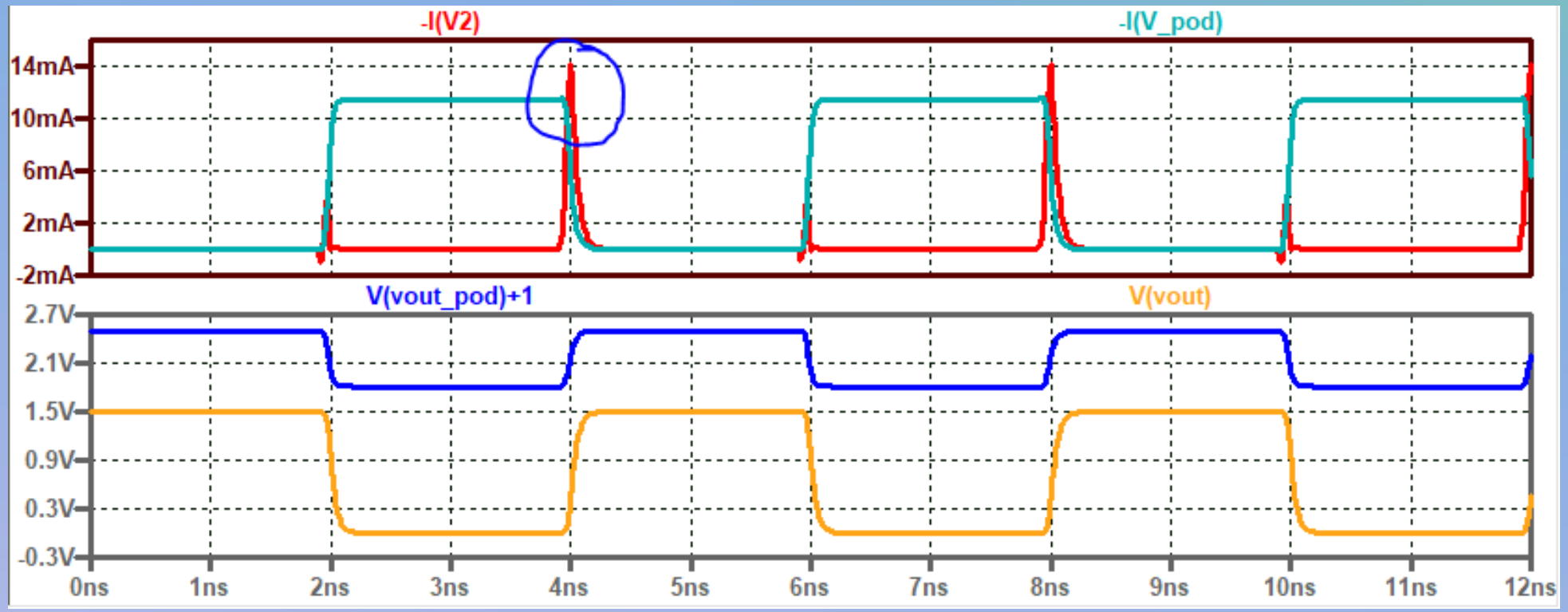
POD – GDDR's Secret Weapon

- The latest versions of GDDR have Pseudo Open Drain (POD) drivers for the data bus lines.
- This technology was apparently first used in GDDR5 and later adopted by the DDR4 standard, showing an instance where mainstream DDR4 has adopted technology pioneered by GDDR5.
- The main idea is to save power by having a much weaker transistor (PMOS) in the driver pulling the output signal high. Instead, a terminator on the other end of the line will pull the signal high. Therefore a very small amount of current will flow when the line is pulled high.
- This scheme is not as useful for low frequency data lines as the static drain from the pull up resistors/terminators will consume a lot of power when the line is being pulled low for extended periods of time.

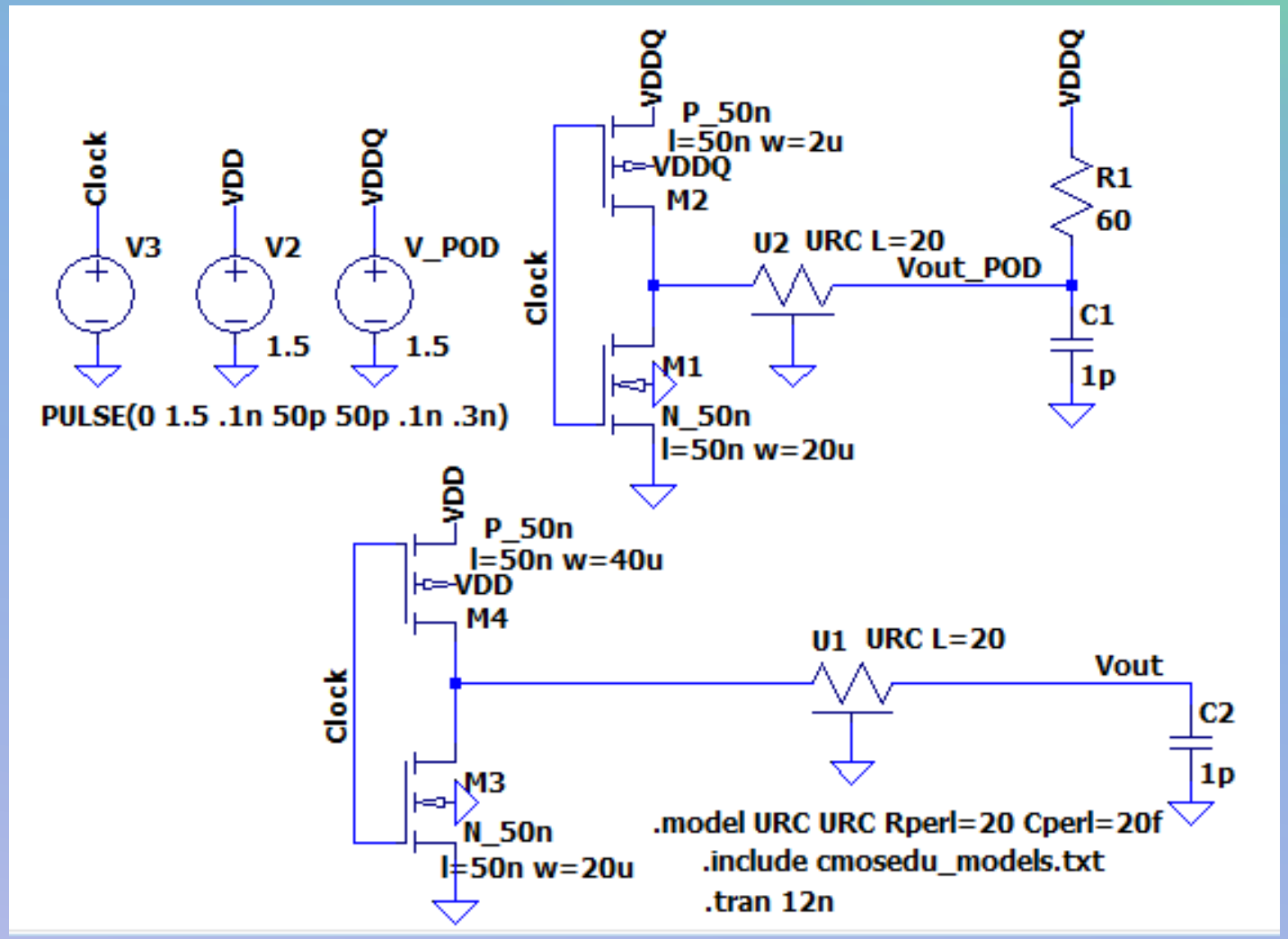
Consider a simple test circuit. We drive two data lines simulated as distributed RC elements with inverters. The lower circuit is a traditional push-pull inverter. The upper circuit has a weaker PMOS driver and a 60 ohm (See POD15 specification) resistor to pull the line toward VDDQ.



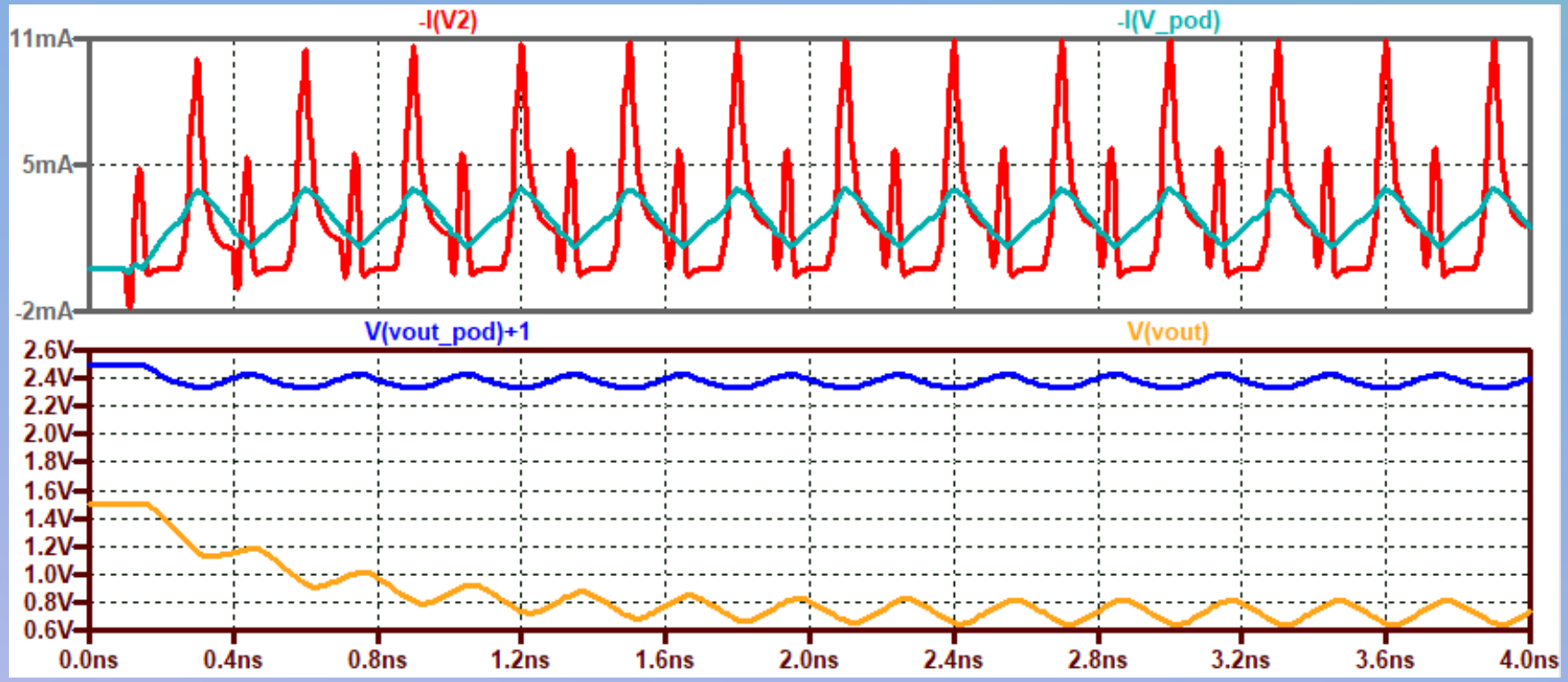
At the simulated low frequency (relative to delay of the transmission line) the static power draw of the POD driver dominates, but note the indicated current transients for V2 are significantly larger.



To simulate the high frequencies in a memory bus we increase the clock frequency and also increase the RC delay through the transmission line.



In this simulation we see that the output swing is slightly larger peak-to-peak for the push-pull driver, yet we finally have a smaller current draw (and thus power dissipation) for the POD driver.



Average -I(V2)	Average -I(V_POD)
2.405mA	2.3128mA

Shared Technology

- It is clear from reading [14], the voluminous standard document, that there are many small features of GDDR5 that are useful in practical circuits while being sparsely documented. Many of these features are also used by DDR memory standards.

- One example is the variable drive strength of the buffers driving the data lines. This has been a feature of memory since at least the DDR2 standard, yet is difficult to understand. Presumably this allows power savings when system design allows that margin.

- After being initialized the chips must apparently undergo a “training” sequence (at least for high frequency operation). From page 16 we learn that “Due to the high data rates of GDDR5, it is recommended that the interfaces be trained to operate with the optimal timings.” This is apparently an effort to synchronize signals times with associated clocks (e.g., WCK). Given the variable path length of traces in actual circuits, as well as variations from one circuit to the next, the necessity of this is quite understandable. What is less clear is how this is actually implemented. Presumably a programmable delay line is used internally to temporally offset the signals on different pins.

References

[1] https://www.samsung.com/semiconductor/newsroom/news-events/samsung-electronics-comes-out-with-super-fast-16m-ddr-sgrams/

[2] http://www.danielsevo.com/hocg/hocg_1990.htm

[3] http://www.seasip.info/VintagePC/mda.html Accessed 5/2/20.

[4] "Understanding DDR SDRAM memory choices“ https://www.techdesignforums.com/practice/technique/understanding-ddr-sdram-memory-choices/ Accessed 5/2/20.

[5] "AMD Radeon RX 5700 XT and Radeon RX 5700 Review: New Prices Keep Navi In The Game" https://www.tomshardware.com/reviews/amd-radeon-rx_5700-rx_5700_xt,6216.html

[6] Samsung 8Gb GDDR5 chip. https://www.samsung.com/semiconductor/dram/gddr5/K4G80325FC-HC25/

[7] PS4 specs. https://www.playstation.com/en-gb/explore/ps4/tech-specs/

[8] DDR standard history. https://en.wikipedia.org/wiki/DDR_SDRAM

[9] DDR5 still under development. https://www.jedec.org/category/technology-focus-area/main-memory-ddr3-ddr4-sdram Accessed 5/4/2020.

[10] https://news.samsung.com/global/samsung-now-mass-producing-industrys-first-2nd-generation-10-nanometer-class-dram

[11] https://en.wikipedia.org/wiki/DDR2_SDRAM

[12] https://en.wikipedia.org/wiki/GDDR4_SDRAM

[13] POD15 standard document. https://www.jedec.org/system/files/docs/JESD8-20A.pdf

[14] JEDEC Standard for GDDR5 (JESD212C) https://www.jedec.org/standards-documents/docs/jesd212c