WOLF Advanced Technology

GDDR7 VS GDDR6 — WHAT CHANGES AT THE MEMORY LEVEL

– WHITEPAPER

Written by: Marc Henney
Product Manager, Product Management

WOLF ADVANCED TECHNOLOGY

INTRODUCTION

Graphics Double Data Rate (GDDR) memory is the high-bandwidth VRAM that feeds a GPU's compute cores with pixels, tensors, and sensor data. Its two core dimensions are how much it can hold (**capacity**, in GB) and how fast it can move data (**bandwidth**, driven by bus width × per-pin data rate). The newest generation, GDDR7, raises per-pin speed and adds stronger reliability features (on-die ECC and richer error reporting), enabling equal or higher bandwidth than many wider-bus GDDR6 designs—often with fewer pins.

Why this matters: system performance and reliability hinges on choosing the right combination of capacity and bandwidth for your workload, not just on raw GB alone. This whitepaper explains the memory-level differences between GDDR7 and GDDR6, provides practical sizing rules, and uses WOLF modules (134S, 167S, 166S, 163S) as concrete examples to help you pick the most appropriate VRAM configuration for your program.

KEY TAKEAWAYS

- **Per-pin speed** jumps with GDDR7 (typical early bins ~28–32GB/s), so a **128-bit GDDR7** design can match or beat a **256-bit GDDR6** design on bandwidth—pin-for-pin, it's just faster.
- GDDR7 adds **RAS/ECC** enhancements (on-die ECC and richer error reporting), improving mission reliability at high speeds.
- The real decision pivots on **two knobs**:
    - The **bandwidth target** (bus width × pin rate)
    - The **resident working set** (total **capacity** you truly need).

MEMORY CONFIGURATIONS ACROSS MODULES:

- **134S:** 16GB **GDDR6**, **256-bit** bus (up to ~448GB/s at 14Gb/s per pin).
- **167S:** 8GB **GDDR7**, **128-bit** bus (ECC), bandwidth depends on bin (e.g., 28–32Gb/s per pin).
- **166S:** 16GB **GDDR7**, **256-bit** bus (ECC), for programs that need both capacity and next-gen bandwidth.
- **163S:** 24GB **GDDR7**, **256-bit** bus (ECC), for programs that need the highest capacity and next-gen bandwidth.

GDDR BASICS

When you compare VRAM across parts or generations, four levers matter:

1. **Capacity (GB)** — how much data can live in-place (resident working set) without streaming or tiling. If the working set is **>8GB** and **can't be partitioned**, you need **16 GB** (or more), regardless of generation.
2. **Bus width (bits)** — How many bits the GPU can move in parallel per transfer across the VRAM interface. (e.g., **128-bit** vs **256-bit**).

3. **Per-pin data rate (Gb/s)** — speed grade per IO
    - GDDR6 = **14–16Gb/s** (common for many GDDR6 deployments)
    - GDDR7 = **28–32Gb/s** (typical early GDDR7 bins).
4. **ECC/RAS** — error detection/correction, scrubbing, and reporting that influences reliability. It can prevent silent corruption and speed up root-cause analysis—often saving more time than a small bandwidth delta.

$$\textbf{Peak Bandwidth}(GB/s) = \left(\frac{\textbf{Bus Width (bits)}}{8}\right) * \textbf{Per Pin Rate } (Gb/s)$$

NOTE: (For both GDDR6 and GDDR7; sustained rates depend on controller scheduling, refresh, and thermals.)

WHAT CHANGES FROM GDDR6 → GDDR7

- **Signaling:** GDDR6 uses NRZ; **GDDR7** moves to **PAM3**, enabling more bits per unit interval at similar clocks and better energy/bit at high speeds.
- **Concurrency:** GDDR7 increases per-device channelization (more independent channels), which helps sustained throughput when many small bursts compete.
- **RAS/ECC:** GDDR7 standardizes **on-die ECC** and richer error reporting (e.g., poison, scrub), boosting resilience in harsh environments.

BANDWIDTH SIZING — QUICK, GROUNDED EXAMPLES

- **GDDR6, 256-bit @ 14Gb/s →** (256/8)×14 = **448GB/s** ← (WOLF-134S example bin)
- **GDDR7, 128-bit @ 28Gb/s →** (128/8)×28 = **448GB/s** ← **parity** with the above on half the bus (WOLF-167S example)
- **GDDR7, 128-bit @ 32Gb/s → 512GB/s ←** (WOLF-167S at max Pin Data Rate)
- **GDDR7, 256-bit @ 28Gb/s → 896GB/s ←** (WOLF-166S and WOLF-163S parity example)
- **GDDR7, 256-bit @ 32Gb/s → 1,024GB/s ←** (WOLF-166S and WOLF-163S at max Pin Data Rate)

Rule of thumb: if you previously sized for **256-bit GDDR6 @ 14–16Gb/s** (≈448–512GB/s), then **128-bit GDDR7 @ 28–32Gb/s** generally meets or beats it. The remaining decision is **capacity** (8GB vs 16GB+) and **RAS** needs.

HOW TO BUDGET CAPACITY (8GB VS 16GB) WITH REAL NUMBERS

When deciding whether 8GB or 16GB of memory will serve your needs, it helps to ground the discussion in something more concrete than specs on a datasheet. Thinking in terms of **"footprints"** makes it easier to see where the bytes really go. It helps to think of these footprints as weights, activations, frames, and maps. Below are **uncompressed** sizes to keep math honest (your pipeline may compress or reuse memory, which improves headroom).

NOTE:

MiB = mebibyte: a **binary** unit.
- **1 MiB = 1,048,576 bytes = 1024 × 1024.**

MB = megabyte: a **decimal** unit (marketing/standards SI).
- **1 MB = 1,000,000 bytes.**

Quick conversion tips
- **MiB → MB:** multiply by **1.048576** (or add ~4.86%).
- **MB → MiB:** divide by **1.048576** (or subtract ~4.64%).
- **GiB ↔ GB** use the same factor.

Real world example
Frames & buffers (per frame, uncompressed):
- 4K (3840×2160) **RGBA8: ~31.6 MiB.** Triple-buffered: **~95 MiB** per stream.
- 4K **RGBA16F: ~63.3 MiB.** Double-buffered: **~126.6 MiB** per stream.
- 8K (7680×4320) **RGBA8: ~126.6 MiB. RGBA16F: ~253.1 MiB** (double-buffered: **~506 MiB**).

Point clouds / LiDAR (XYZ float32 + intensity):
- **10 M points × 16 B/pt ≈ 153 MiB** (single copy). Two levels of detail: **~306 MiB**.

Model weights (rule of thumb):
- **1.0 B parameters → ~3.73 GiB (FP32), ~1.86 GiB (FP16), ~0.93 GiB (INT8).**
- **2.0 B parameters → ~7.45 GiB (FP32), ~3.73 GiB (FP16), ~1.86 GiB (INT8).**

Reserve overhead. *In practice, set aside* **~10–20%** *of VRAM for drivers, heaps, alignment, and fragmentation. On an* **8 GB** *card, expect roughly* **6.5–7.3 GB** *truly usable for your app.*

What "fits comfortably" in 8 GB vs 16 GB
8GB class (example: 167S, 128-bit GDDR7):
- Multiple **4K60** pipelines with **double/triple buffering** (hundreds of MiB), **+** mid-sized DL models (**≤~1–2 GB** weights at FP16/INT8), **+** working tensors, **+** point-cloud tiles (hundreds of MiB).
- **Streaming/tiled** algorithms shine here: keep only the **active tiles/frames/segments** in VRAM; everything else streams over PCIe/Ethernet.

16GB class (examples: 134S GDDR6, 166S GDDR7):
- **Very large resident sets:** big mosaic/SLAM maps, multi-sensor **8K** buffers, or **larger models** (e.g., **~2 B FP16 ≈ 3.7 GiB** weights plus activations) without aggressive tiling.
- More leeway for **multiple concurrent models** and **larger batch sizes** when latency constraints allow.

Decision shortcut (capacity-first, then bus width)

1. **Size the resident working set** (worst-case, uncompressed). If **>8GB** and not tileable/ streamable or cant be partitioned → **choose 16GB** (e.g.,163S, 166S or 134S).
2. **Hit the bandwidth target** using **bus width × pin rate:**
   - Need **≈448–512GB/s?** Either **256-bit GDDR6 @ 14–16Gb/s** or **128-bit GDDR7 @ 28–32Gb/s.**
   - Need **>512GB/s?** Favor **GDDR7** and/or **256-bit** buses.
3. **Apply RAS/ECC requirements** for mission profile (GDDR7 helps here).

WOLF SWITCH FAMILY MEMORY SPECS

WOLF MODULE	DRAM GEN	CAPACITY	BUS WIDTH	ECC/RAS NOTE	BANDWIDTH (ILLUSTRATIVE)
134S	**GDDR6**	**16 GB**	**256-bit**	No ECC	Up to **~448 GB/s** (14 Gb/s bin)
167S	**GDDR7**	**8 GB**	**128-bit**	ECC support (module level)	**~448–512 GB/s** @ 28–32 Gb/s per pin
166S	**GDDR7**	**16 GB**	**256-bit**	ECC support (module level)	**~896–1,024 GB/s** @ 28–32 Gb/s per pin
163S	**GDDR7**	**24 GB**	**256-bit**	ECC support (module level)	**~896–1,024 GB/s** @ 28–32 Gb/s per pin

(Exact bins are configuration-dependent; use the calculator below to size conservatively.)

FAQ — MEMORY CONCERNS WHEN MOVING FROM 134S TO 167S

Below are common concerns a customer may have regarding the shift to 167S:

1. **Why per-pin rate varies?** (GDDR6 = 14–16Gb vs GDDR7 = 28–32Gb/s):
   Vendors offer multiple **speed bins**, and the achievable rate depends on **temperature, board SI/PI, channel loading, controller training, and reliability margin.** Programs typically select a **conservative bin that meets bandwidth goals across worst-case conditions.**

2. **How does bus width impact the memory?**
   - **Bus width → bandwidth.** At a fixed per-pin rate, a **256-bit** bus moves **2×** the bytes of a **128-bit** bus.
   - **Bus width ≠ capacity. Capacity** is set by **DRAM density × number of packages × how the controller wires them.**

However, board reality ties them together. A **wider bus** usually implies **more packages** (routing/pin budget), which often makes **higher total capacity** easier to achieve at a given generation. Conversely, a **narrower bus** simplifies routing and power but may **cap max capacity** until higher-density devices are available.

3. **Will 8GB on 167S choke my workload that ran on 16GB (134S)?**
   Only if your **resident working set** truly exceeds **~6.5–7.3 GB usable** (after overhead). If you can **tile/stream** frames, maps, or model shards, 8GB is typically fine—and you gain GDDR7's higher per-pin bandwidth plus on-die E`CC. If it must be **> 8GB** and can't be partitioned, step to **16GB** (e.g., a GDDR7 256-bit option).

4. **We're worried about bandwidth dropping with a 128-bit bus. Is that real?**
   No. Bandwidth **= bus width × per-pin rate. 128-bit GDDR7 @ 28–32Gb/s** matches or beats **256-bit GDDR6 @ 14–16Gb/s.** Many pipelines are limited by **compute or I/O**, not VRAM bus width.

5. **How do we know if our model fits in 8GB?**
   Add **weights + activations + buffers** (uncompressed) and add **10–20% headroom.** Quick rule of thumb:
   - **~1 B parameters ≈ ~1.9GB (FP16) / ~0.95GB (INT8)** for weights (plus activations). If total > **~7GB** after headroom, pick **16 GB** or use **tiling/quantization/micro-batching.**

6. **We run multiple high-res video streams. Will 8GB be enough?**
   Usually, yes, if you **process in a streaming fashion. 4K RGBA8 ≈ 32 MiB/frame; RGBA16F ≈ 63 MiB/frame.** Even with double/triple buffering across several streams, you're typically in the **hundreds of MiB,** leaving room for models and working tensors—provided you don't cache many frames simultaneously.

7. **Is GDDR7 more heat-sensitive?**
   PAM3 shrinks noise margins and **speeds are higher,** so **thermals + SI/PI** discipline matter more. But vendors actually report **better efficiency and even lower thermal resistance vs GDDR6.** Bottom line: with **adequate cooling and PDN design,** GDDR7 hits its bins reliably; with marginal cooling, however it may throttle earlier because margins are tighter.

CONCLUSION

In conclusion, GDDR7 represents more than just an advantage over GDDR6—it changes how engineers think about balancing bus width, capacity, and reliability. With higher per-pin rates, integrated ECC, and stronger RAS features, GDDR7 can deliver equal, or even greater bandwidth on fewer pins while improving system resilience. In practice, many workloads that previously required 16GB of GDDR6 can now run comfortably on 8GB of GDDR7, while data-heavy or non-partitionable applications still benefit from 16-24GB configurations. Selecting the right module means matching memory footprints and reliability needs with the strengths or each generation to ensure both performance efficiency and long-term robustness.

WOLF
ADVANCED TECHNOLOGY