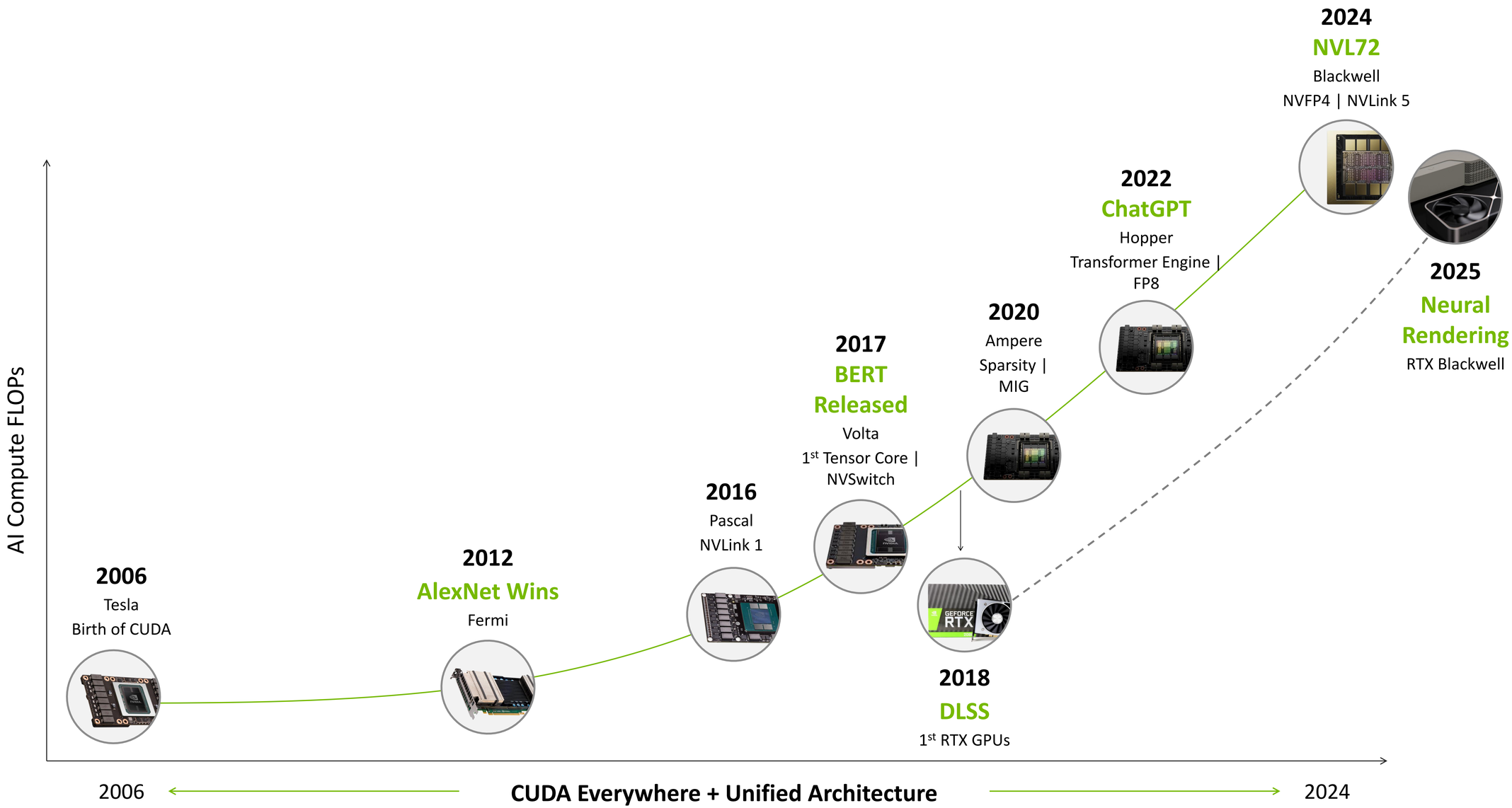
RTX 5090: Designed for the Age of Neural Rendering

Marc Blackstein | Hot Chips 2025

Building and Inspiring AI Across Generations



Scaling the Blackwell Architecture

GB300 NVL72 Rack
72 Blackwell Ultra GPUs
in a coherent domain

NVIDIA Spectrum-X
& ConnectX-8
800 Gbps Networking

Compute Density
4x more TFLOPs/mm
vs Hopper

FP4 Tensor cores
15 PetaFLOPS
Dense NVFP4

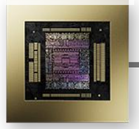
288GB HBM3e Memory
(8 Stacks, Up to 8 TB/s)

Chip Architecture for Perf at Scale

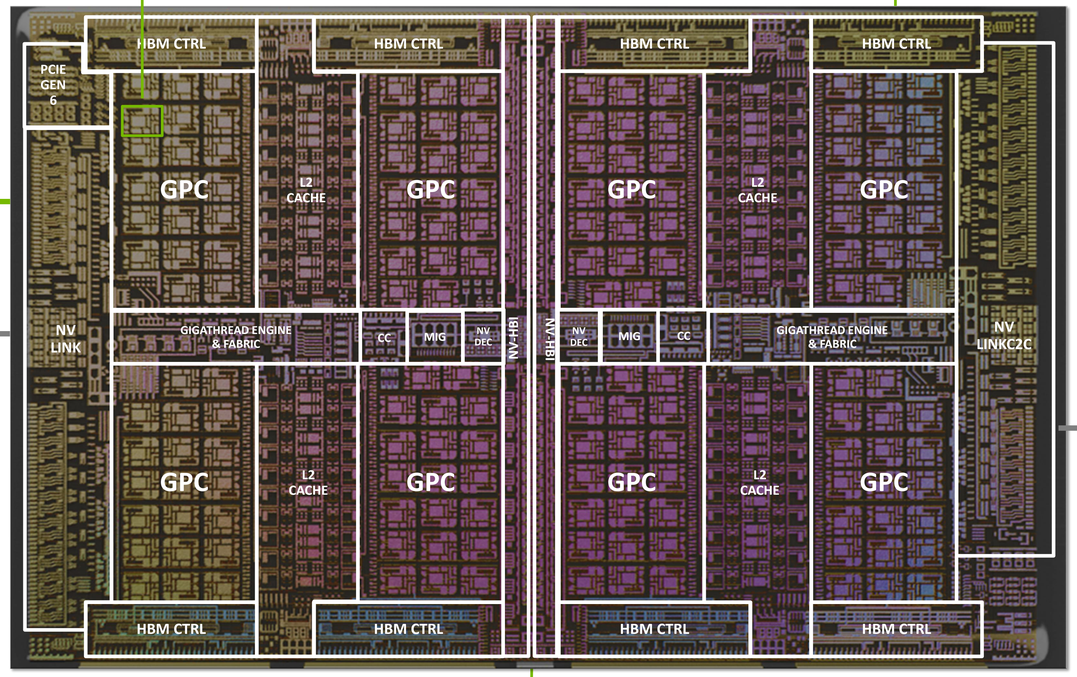
Same architecture,
optimized for neural rendering

RTX

NVSwitch



SHARP
In Network
Compute



Grace CPU

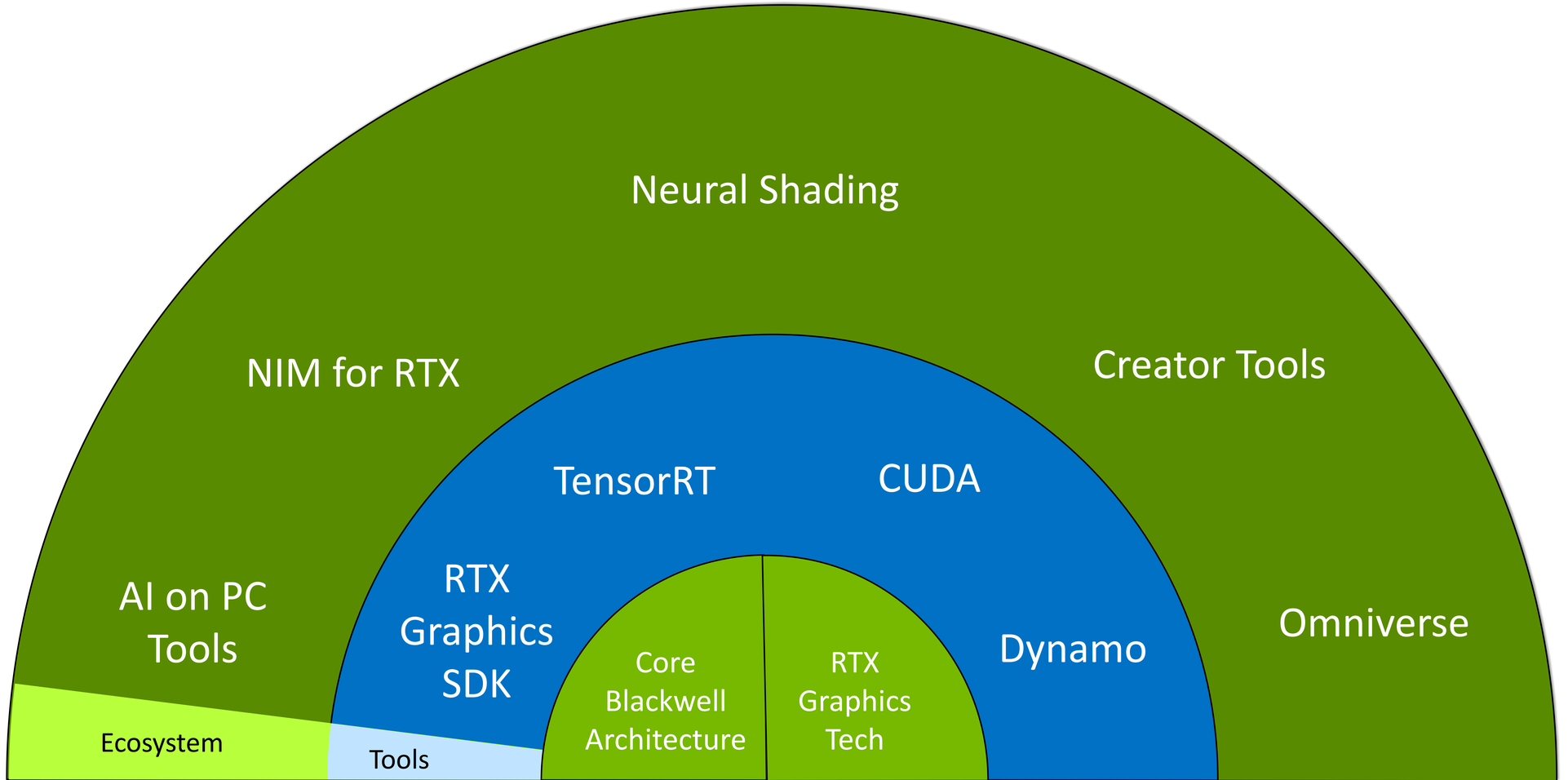
NVLink 5
1,800 GB/s

High Bandwidth Interface
10 TB/s Die-to-Die

NVLink-C2C
Up to 900GB/s Coherent CPU-GPU Interface

RTX Ecosystem Building on Blackwell

Optimizing RTX for AI Applications and Neural Rendering starts with the Core Blackwell Architecture and RTX



The Promise of Neural Rendering and Gaming

Exquisite visuals and worlds

- 10x amplification in performance, footprint, design cycle

Smooth, responsive rendering

- Up to 100% of pixels are AI generated. Scale performance to competitive gaming, battery life to the Gamer on the Go

Adaptive and personal

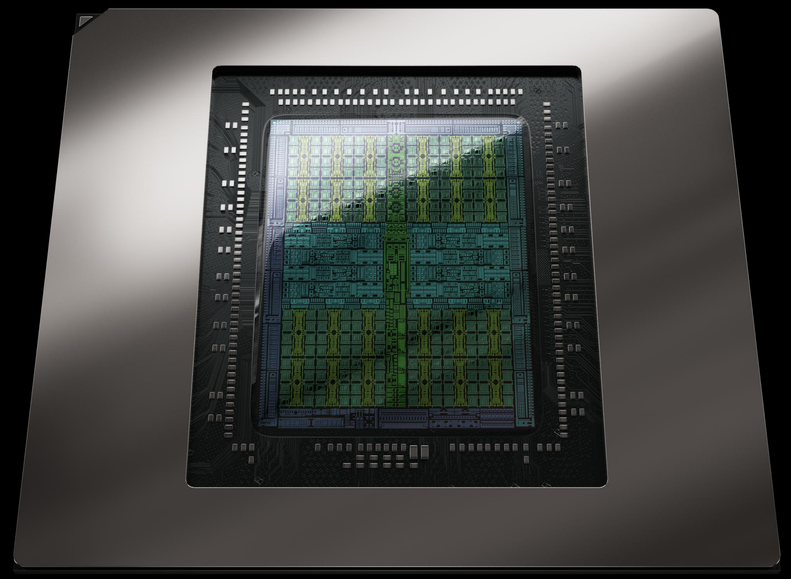
- Gameplay grows as you do

Surprise and imagination

- 1000s of thinking agents create emergent experiences

RTX Blackwell Design Principles

- Optimize for new neural workloads
- Reduce memory footprint
- Quality of service for neural + graphics
- Energy efficiency that scales



Putting It All Together

NVIDIA RTX Blackwell Neural Rendering Architecture

5th Gen Tensor Cores

- 4,000 AI TOPS | High speed FP4

4th Gen RT Cores

- 360 RT TFLOPS | Built for mega geometry

AI Management Processor

- Simultaneous AI models + graphics

BLACKWELL SM

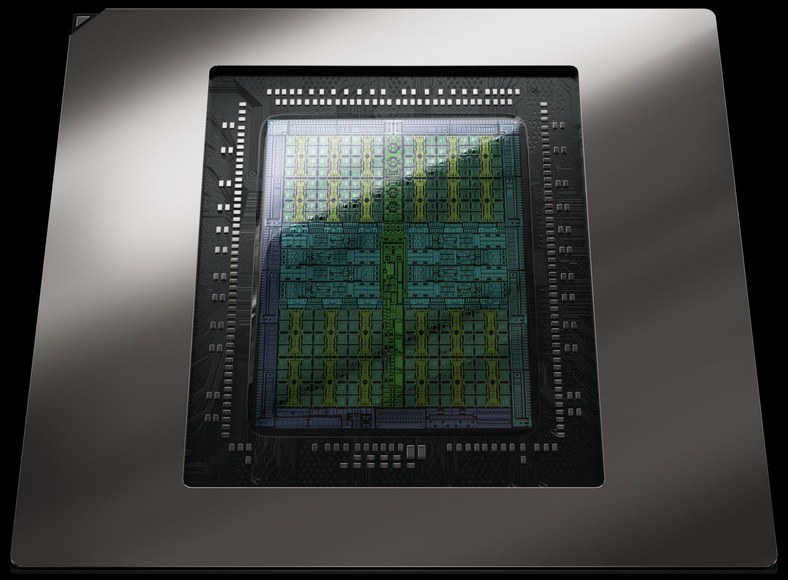
- 125 TFLOPS | Built for Neural Shaders

BLACKWELL MAXQ

- 2X Power efficiency

G7 Memory

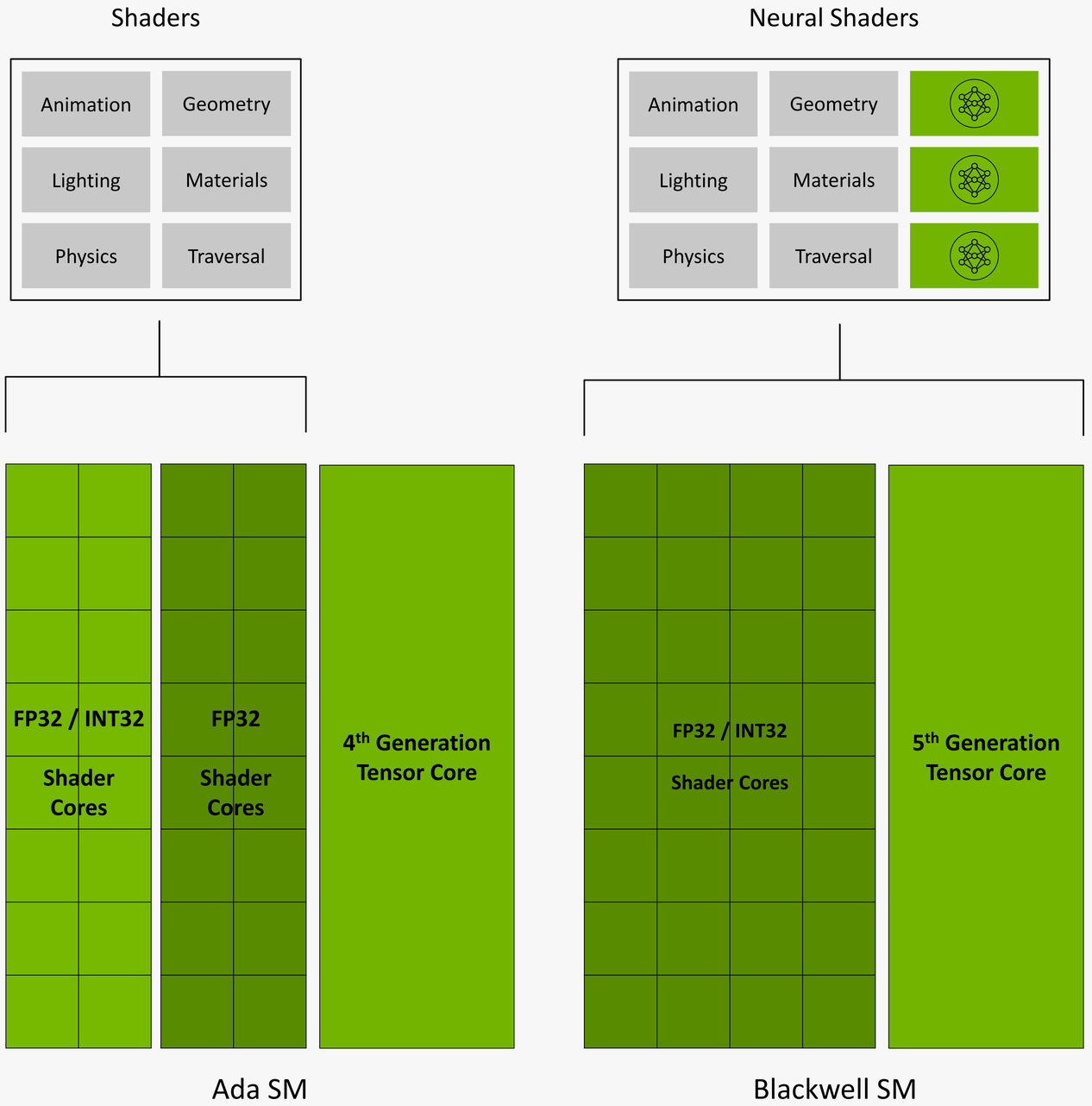
- 30 Gbps | World’s fastest



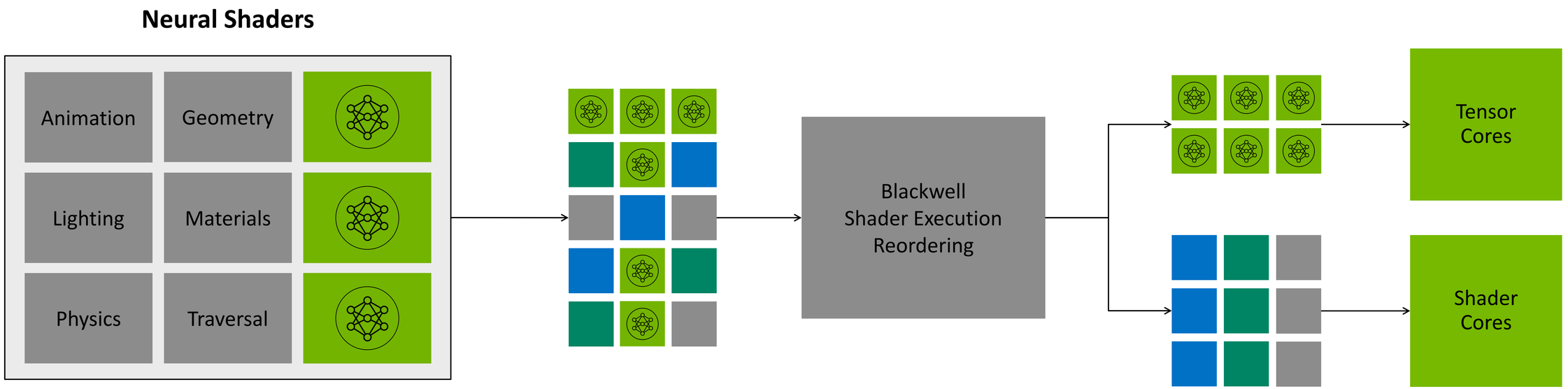
DisplayPort 2.1 UHBR20
PCIe Gen 5
4X NVDEC, 4X NVENC with 4:2:2

RTX Blackwell SM

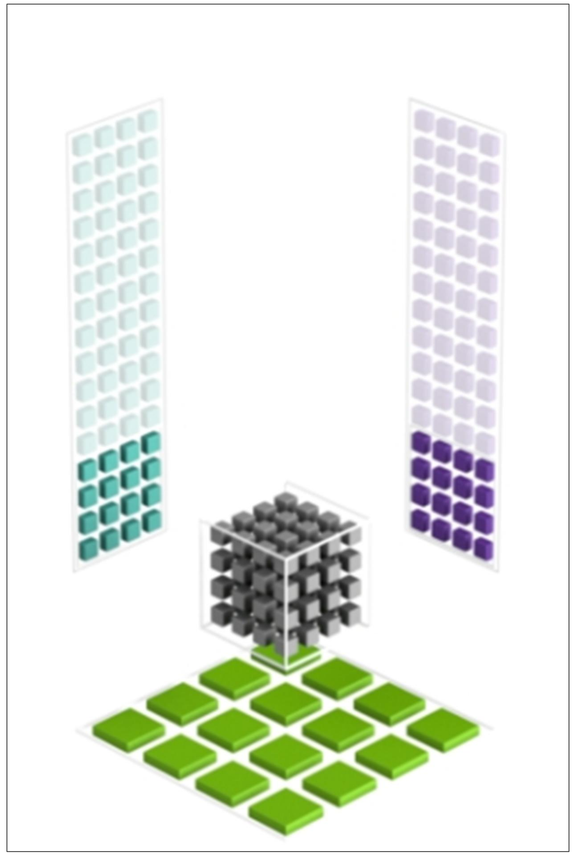
Built for Neural Shaders



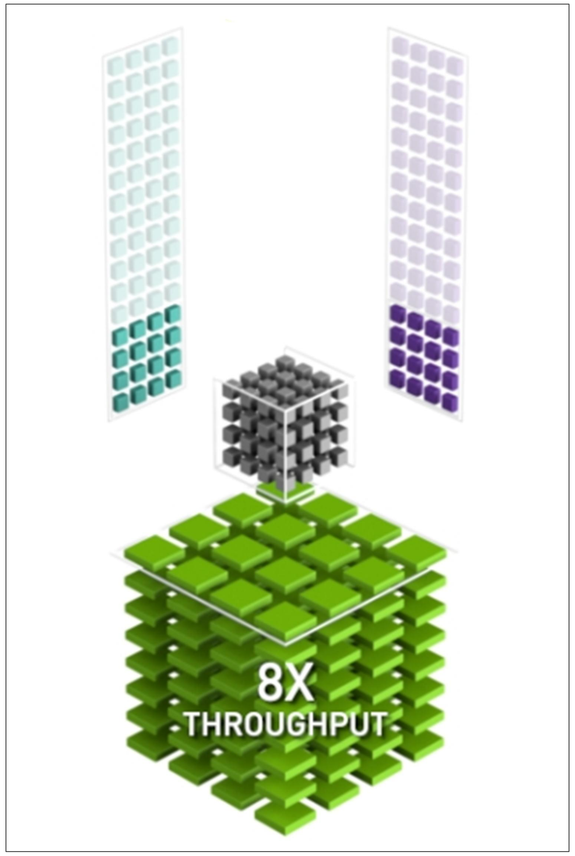
RTX Blackwell SM Improves SER by 2X



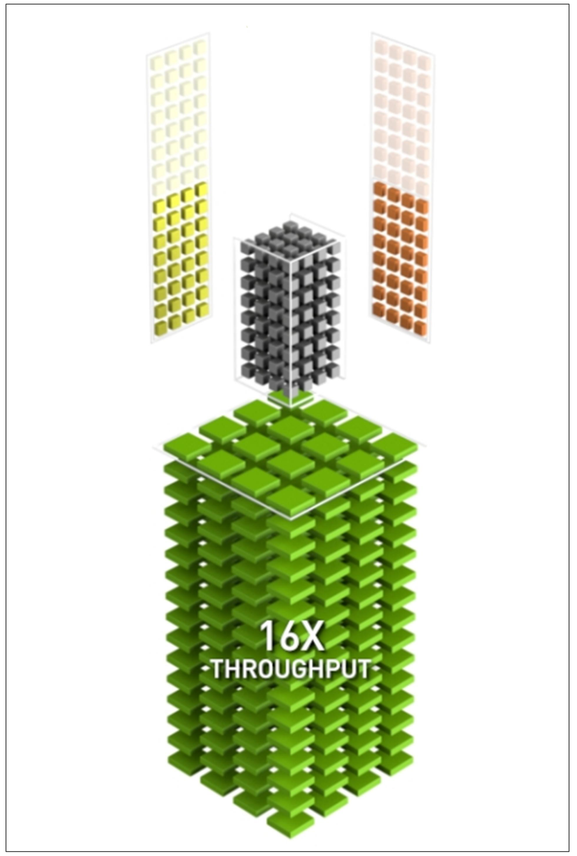
Blackwell 5th Generation Tensor Cores with FP4



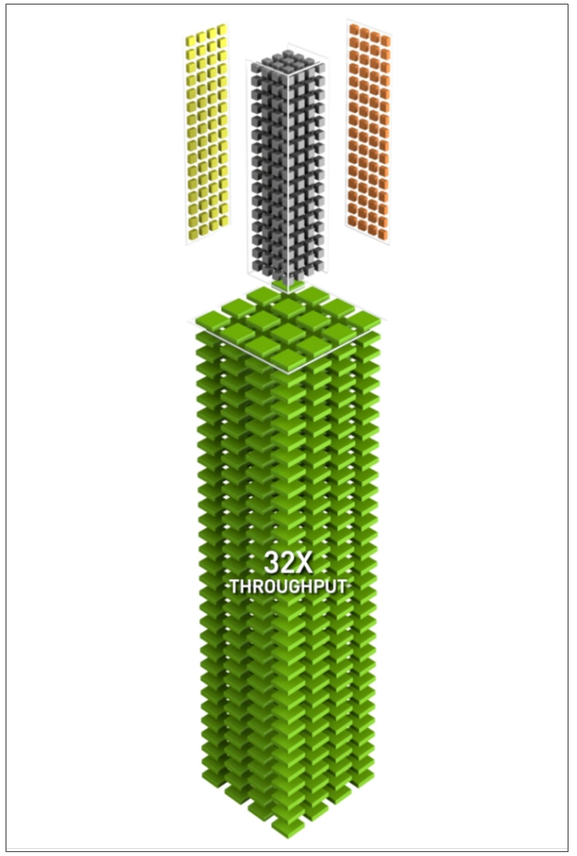
Pascal



Turing Tensor Core
FP16



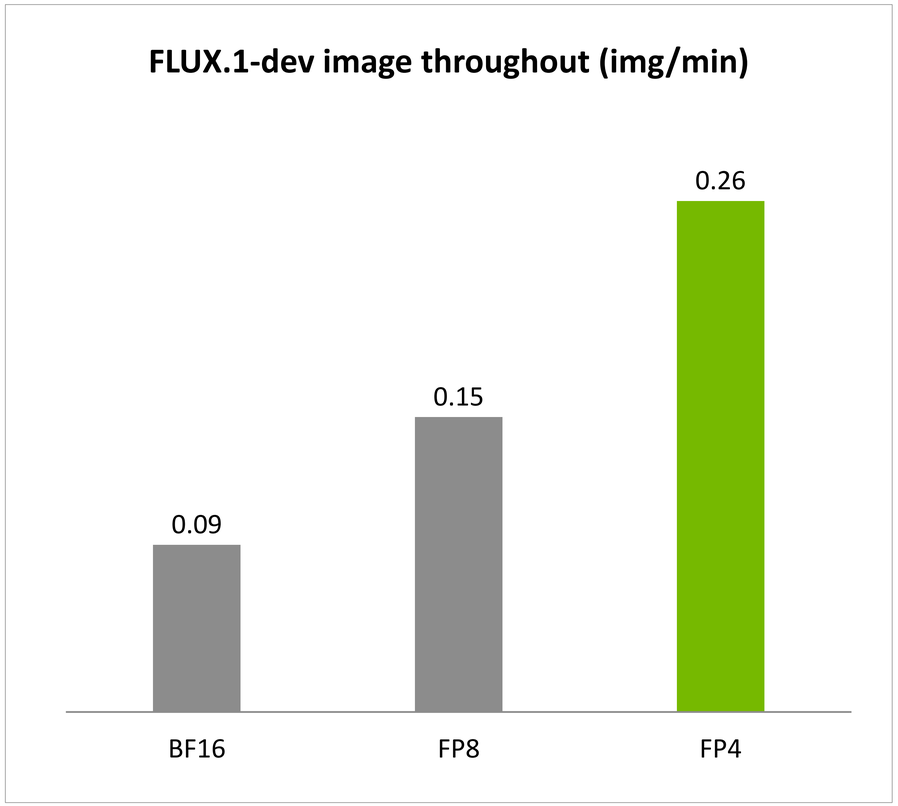
Ada Tensor Core
FP8

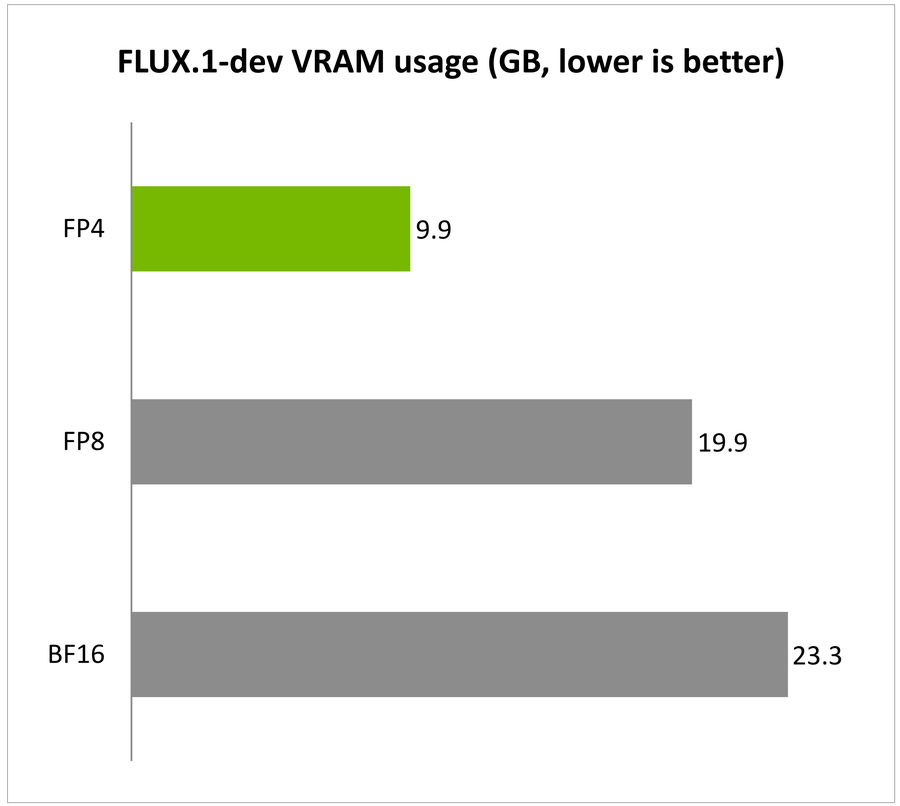


Blackwell Tensor Core
FP4

Blackwell FP4

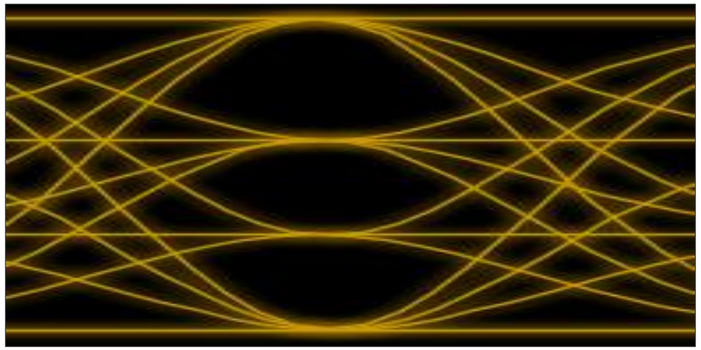
Over 2X performance and half the VRAM with comparable quality



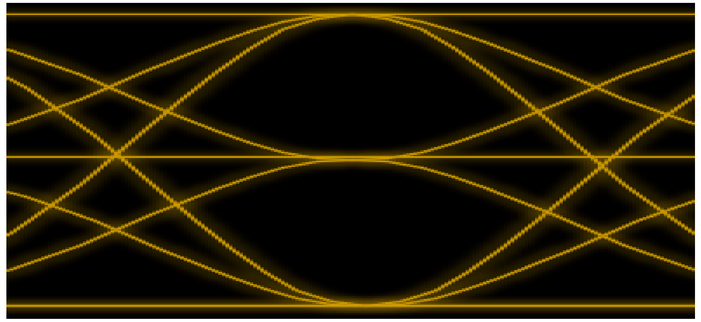


All measurements done using 30 steps with FLUX.1-dev on a GeForce RTX 5090 GPU.

GDDR7: The New Graphics DRAM Standard

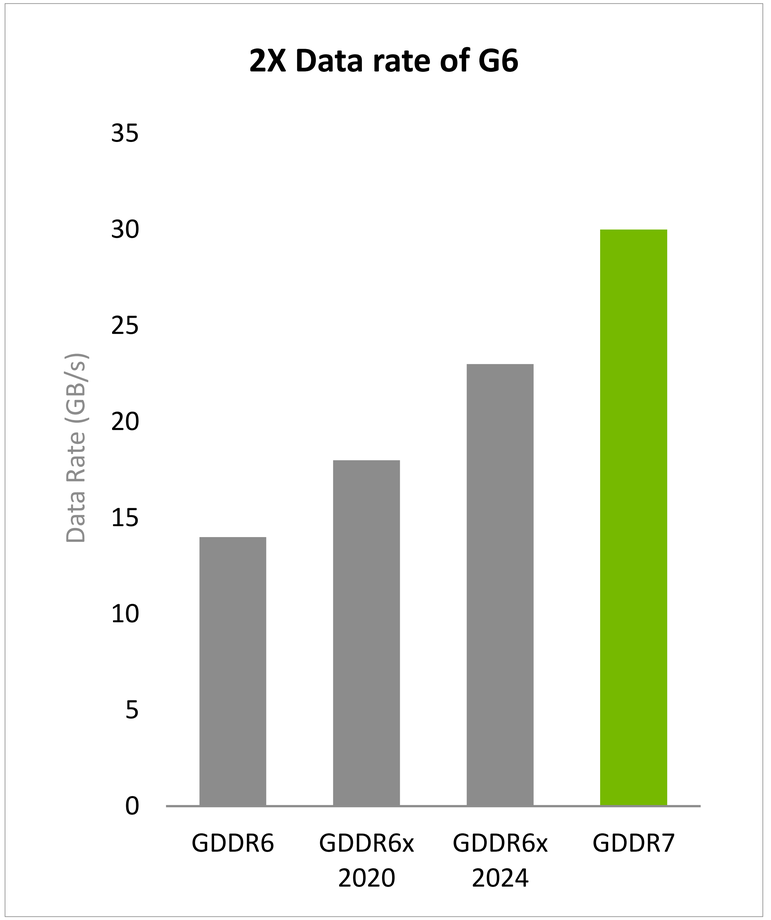


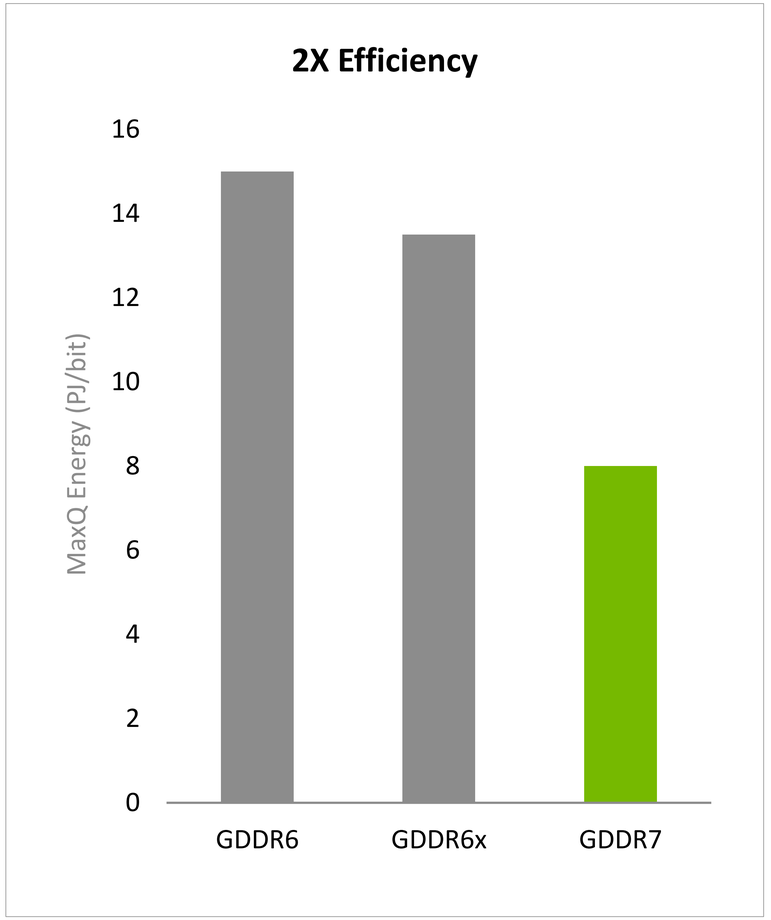
GDDR6x: PAM4



GDDR7: PAM3

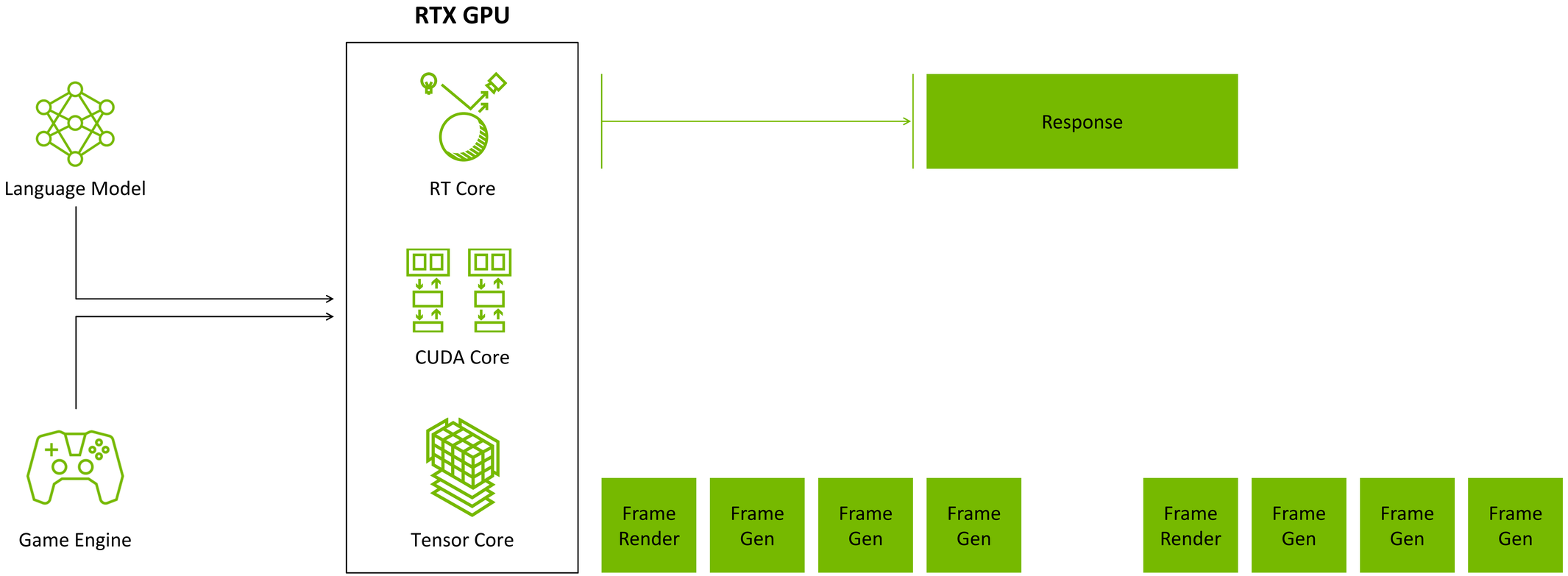
Higher frequency, lower voltage



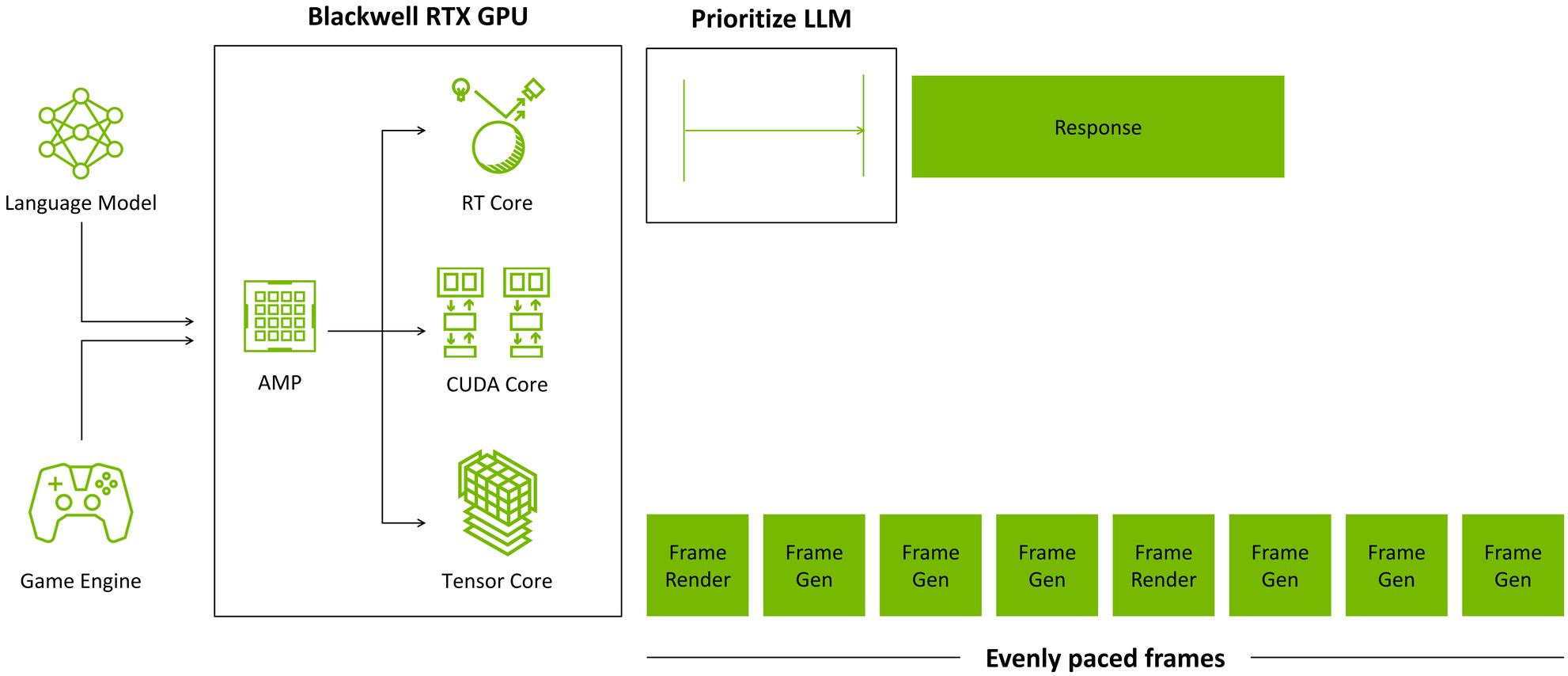


Energy efficiency reflects the average graphics application with 30% DRAM utilization

Simultaneous AI and Graphics Workloads



AI Management Processor



AI Workload Streaming in Action

VRAM

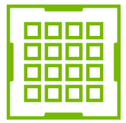
Real time workloads (>> game minspec)
Workloads always resident

Traditional render and path tracing | Neural rendering algorithms | Super resolution | Frame generation

Non-realtime carveout
Workloads streamed

World simulation | Thinking agents (100s)
AI co-players – speech + LLM pipelines

PCIE gen5 64GB/s streaming



AMP manages this all

Render to 16ms deadline

DLSS4 to match display

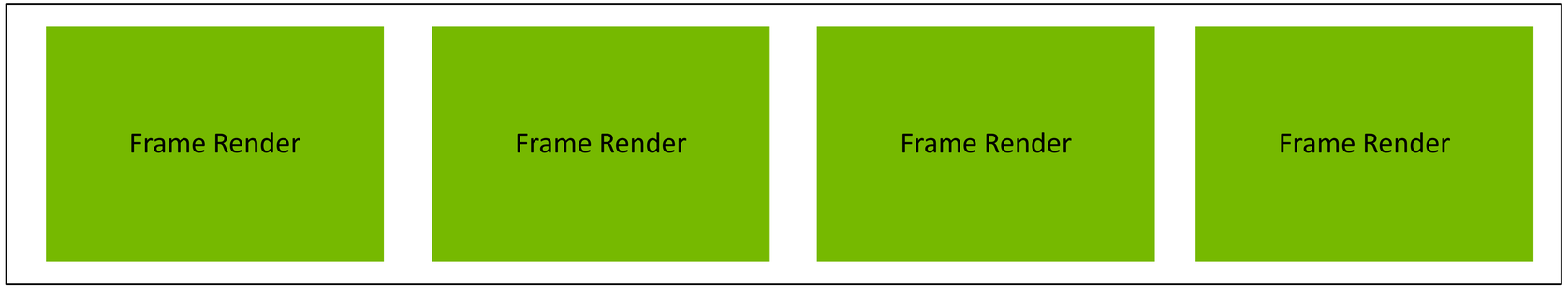
Double buffer

Run Model A | Stream in Model B

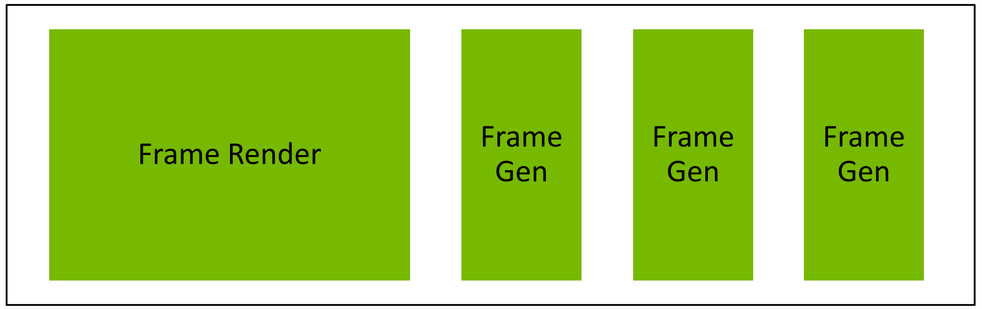
Gaming on the Go

Neural is both faster and more energy efficient

Traditional



DLSS4 + RTX Blackwell
Race to idle

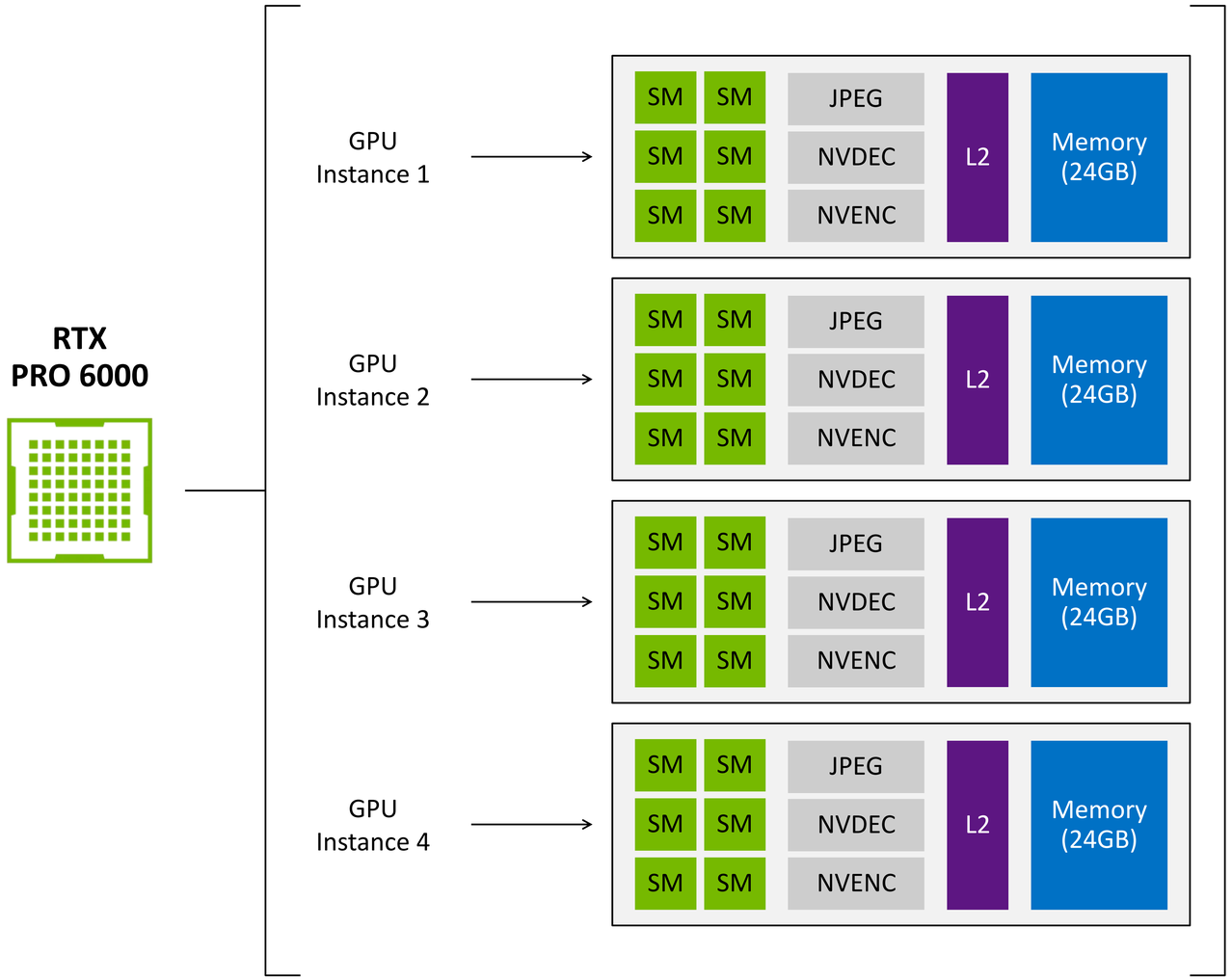


10x faster core rail gating
100x faster DRAM to self refresh

Up to 2x reduction in GPU power towards battery life

RTX PRO 6000 Introduces Universal MIG

Optimize GPU utilization with multi-instance GPU



Simultaneous AI and Graphics workloads

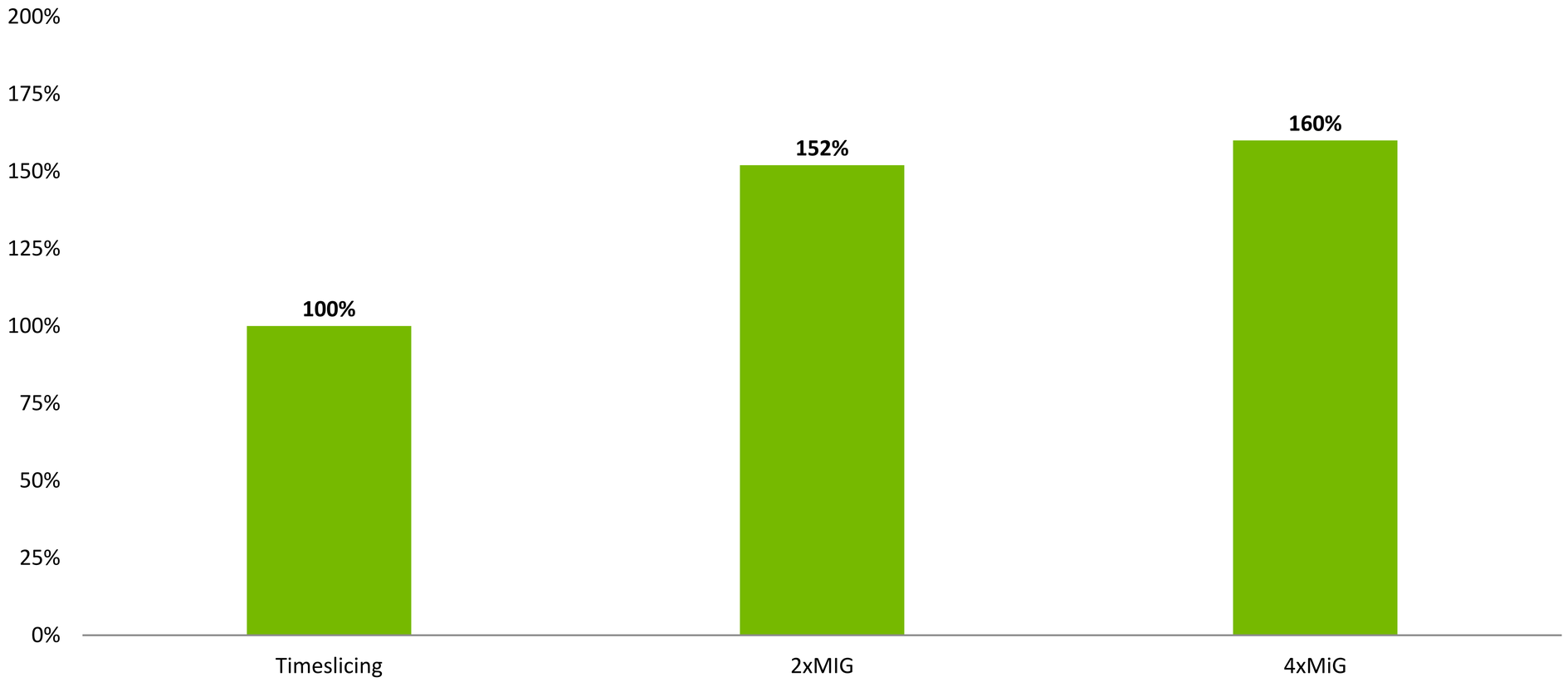
- Dedicated SM, Memory, L2 and BW for hardware QoS and isolation
- All MIG instances run in parallel with predictable latency and throughput
- Schedule used to decide which apps get run

Right Sized GPU Allocation

- Up to 4X instances on RTX PRO 6000 (24GB / instance)
- Different sized MIG instances based on target workloads
- All workloads share same OS

RTX PRO 6000 Scaling

Multi-tenant Scaling—Cyberpunk 2077 @ 1080p



Relative Frame Throughput of 4 instances on RTX PRO 6000

All measurements done on RTX PRO 6000, Cyberpunk 2077 at default/medium settings 1080p resolution

Conclusion

- Blackwell is one architecture that scales from data centers to desktops
- Neural rendering fuses traditional graphics with AI to reimagine real-time graphics, delivering new levels of visual fidelity and immersive worlds through a tight coupling of artists, models, and AI agents
- RTX Blackwell lays the foundation for the new era of neural rendering