NVIDIA ORIN SYSTEM-ON-CHIP

MICHAEL DITTY | AUGUST 2022

INTRODUCING ORIN

Advanced CPU

12x ARM Cortex-A78AE Cores
ARM Arch V8.2

Ampere GPU

Up to 2 GPC / 8 TPC / 16 SMs

5.3 FP32 CUDA TFLOPs

10.6 FP16 CUDA TFLOPS

Higher DRAM BW

Up to 256-bit LPDDR5

205 GB/s

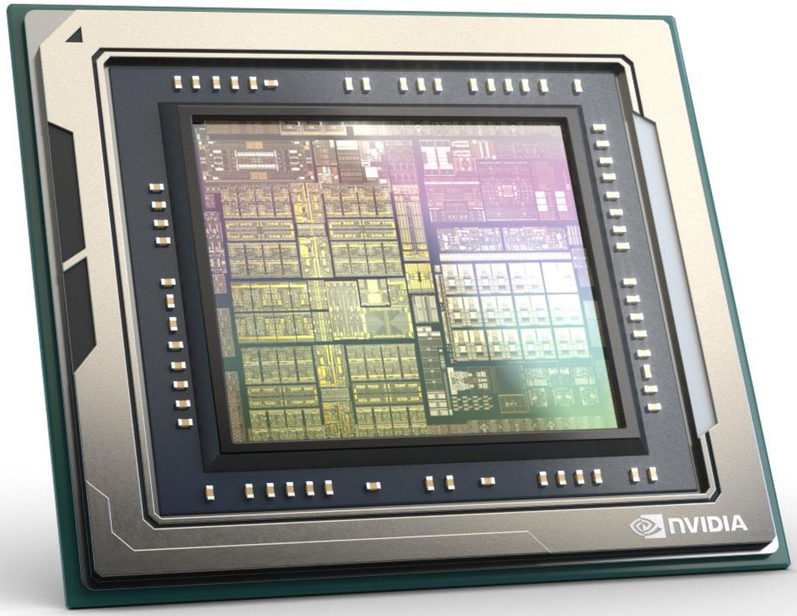
Process

Samsung 8nm

Safety Island

Up to 10K ASIL D DMIPS

4x Lockstep ARM Cortex-R52



Rich IO Connectivity

Up to x4 10 Gb Ethernet

x24 SERDES, x16 CSI

Strong DL Performance

Up to 275 INT8 DL TOPs
(170 GPU + 105 DLA)

85 FP16 DL TOPs (GPU)

Enhanced PVA

Up to 512 INT16 GMAC/s

2048 INT8 GMAC/s

SOC Safety

FUSA ASIL-B Chip

ASIL-D Systematic

ORIN CPU COMPLEX

- ARM Cortex-A78AE V8.2 high-performance CPU
- Lockstep Support
- 2.2 GHz frequency
- Cache hierarchy
  - L1 (per core): 64 KB I$, 64 KB D$
  - L2 (per core): 256 KB
  - L3 (per cluster): 2 MB L3 per cluster
  - System Cache (L4): 4 MB shared cache

CPU PERFORMANCE

Orin Silicon Based Measurements

Benchmark	Score
SPEC CPU2006 speed integer single core	31.8*
SPEC CPU2006 rate integer 12-core	269.5*
SPEC CPU2006 speed floating-point single core	41.6*
SPEC CPU2006 rate floating-point 12-core	332.0*
SPEC CPU2017 rate integer single core	4.04*
SPEC CPU2017 rate integer 12-core	39.36*
Geekbench 5 single core	754
Geekbench 5 12-core	7773

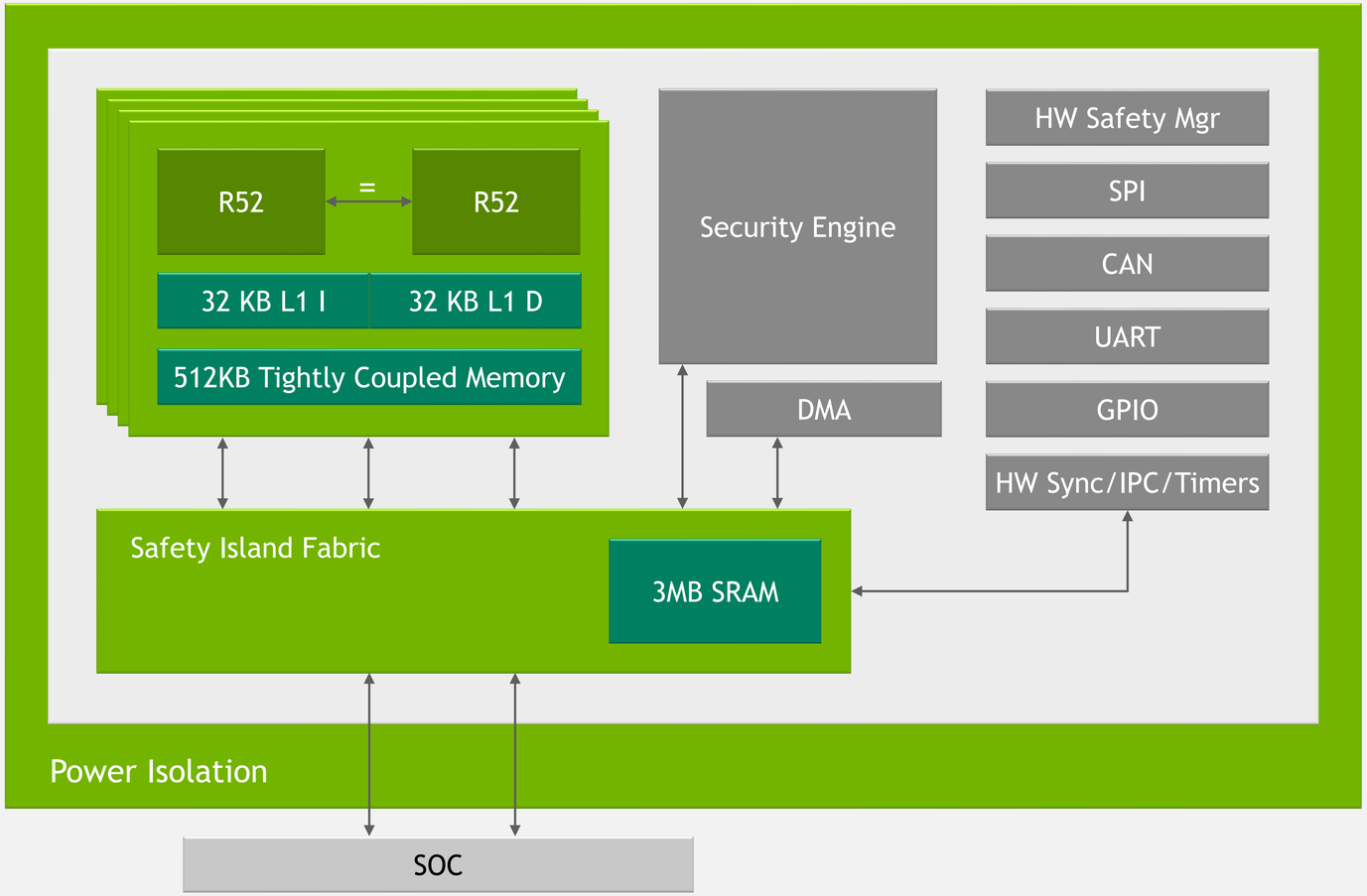
Notes:

- CPU clock speed used for testing is 2.2 GHz
- Memory running at 3200 MHz for all tests
- SPEC CPU2017 single-core uses rate, not speed

**SPEC scores are estimates*

ORIN SAFETY ISLAND

- Isolated ASIL-D Compute Subsystem
- Lockstep ARM Cortex-R52 CPUs
- Dedicated IO
- Dedicated Security Processor



ORIN SAFETY ISLAND

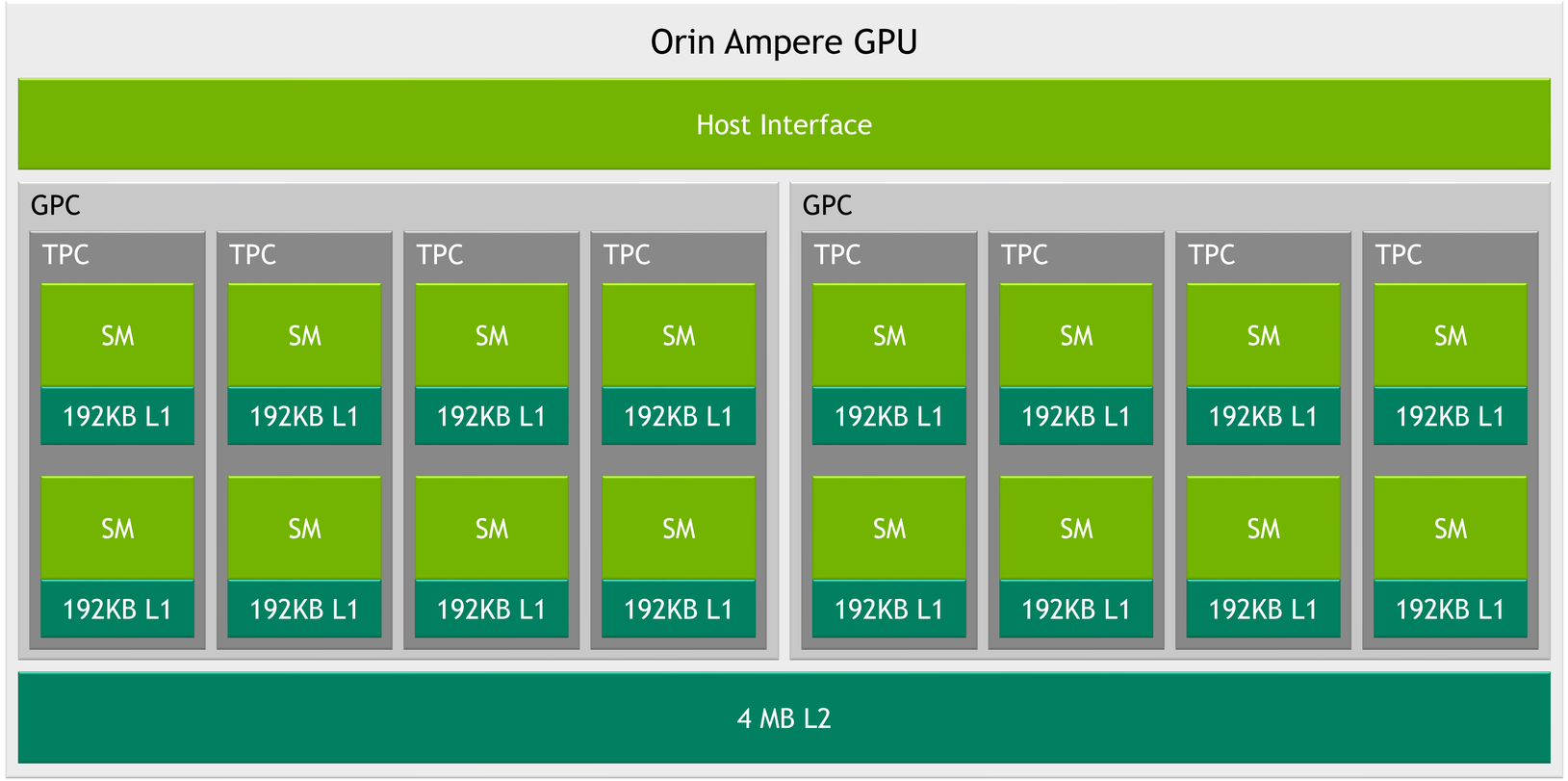
CPU & Memory Configuration

Safety Island Spec	Orin
CPU configuration	4x Cortex-R52 Lockstep pairs
Aggregate ASIL D DMIPs	10K
ICache / DCache per core pair	32KB / 32KB
Tightly Coupled Memory per core pair	512KB
Shared SRAM	3MB
Total Shared SRAM + TCM	5MB

AMPERE GPU

Orin features the Ampere GPU architecture with enhanced DL throughput, the latest graphics capabilities including ray-tracing, and improved power efficiency.

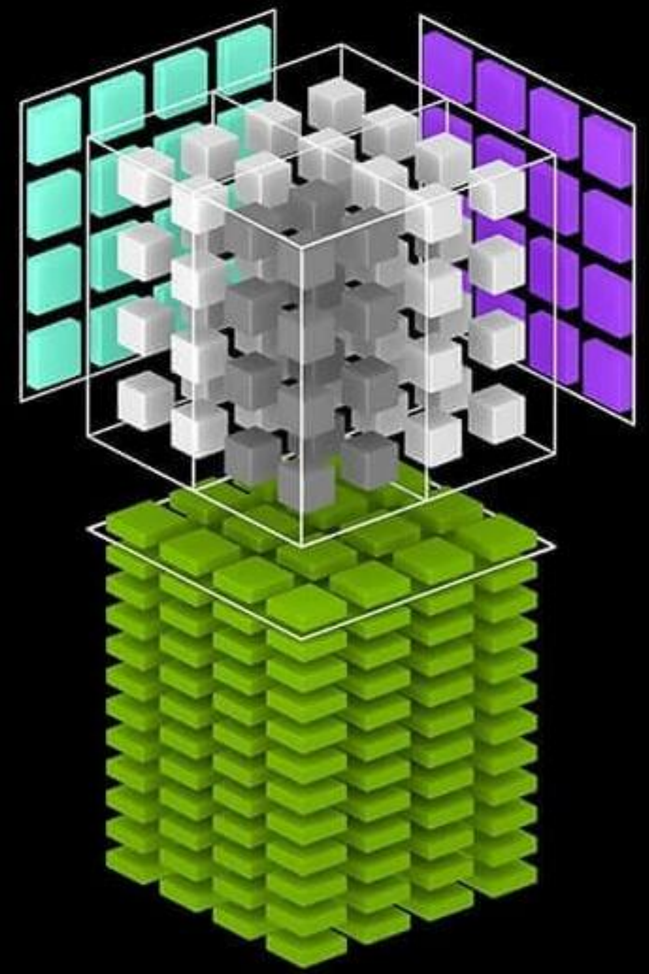
- 2 GPC / 8 TPC / 16 SM
  2x Xavier
- 192 KB L1-cache per SM
- 4 MB L2-cache
- Enhanced Tensor Cores



AMPERE GPU

Compute Enhancements

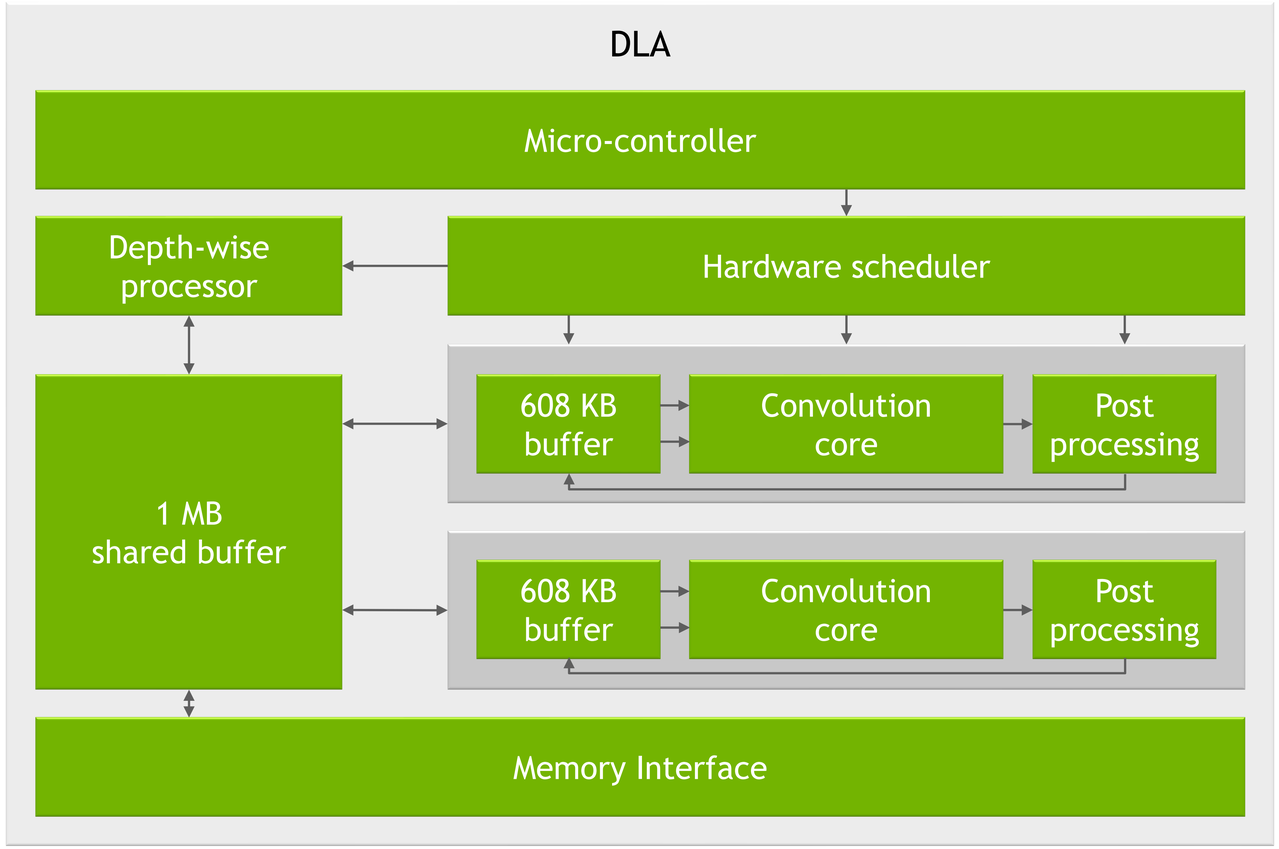
- MIG (Multi-Instance GPU): GPU can be split into two separate GPUs for compute
- Sparsity: fine grained structured sparsity doubles throughput and reduces memory usage
- 2x CUDA floating-point performance: higher compute math speed



NEXT GEN DLA

Deep Learning Accelerator Focused on INT8 Inference Performance

- Increased Performance to 52.5 TOPS (int8)
- Aligned with GPU
  - Compatible Math Pipeline
  - Structured Sparsity
  - TensorRT API
- Performance
  - Structured Sparsity
  - Larger SRAM
  - HW support for layer scheduling
  - Dedicated depth-wise convolution engine
- Additional native HW features supported



NEW FEATURES

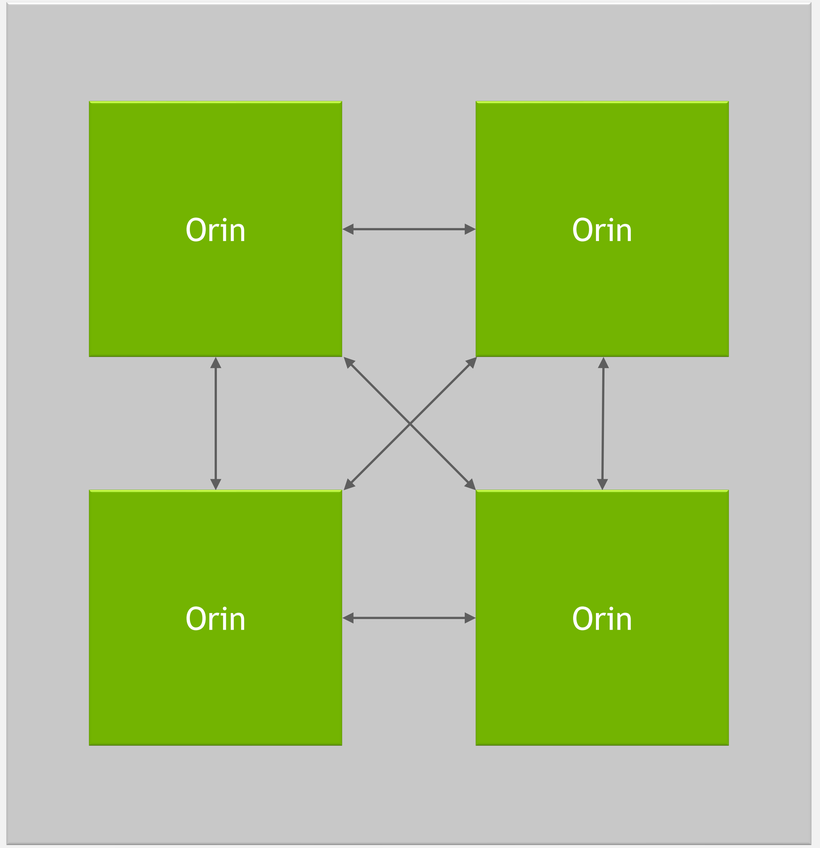
Compared to Xavier DLA

Feature	Comments
Softmax	New Function Support
Clamped RELU	New Function Support
Exclusive Average Pooling	New Function Support
Per channel scaling	New Function Support
Full-channel normalization	New Function Support
UINT8 support	New data format support
Support 3D Convolution	New Function Support
Hardware Scheduler	New Engine for optimization
Structured Sparsity	New Optimization Feature
Group function Optimization	Optimization for Group Function Performance
Depth-Wise Convolution Engine	Highly optimized dedicated engine for DW performance

MULTI-ORIN

High Speed Data Sharing

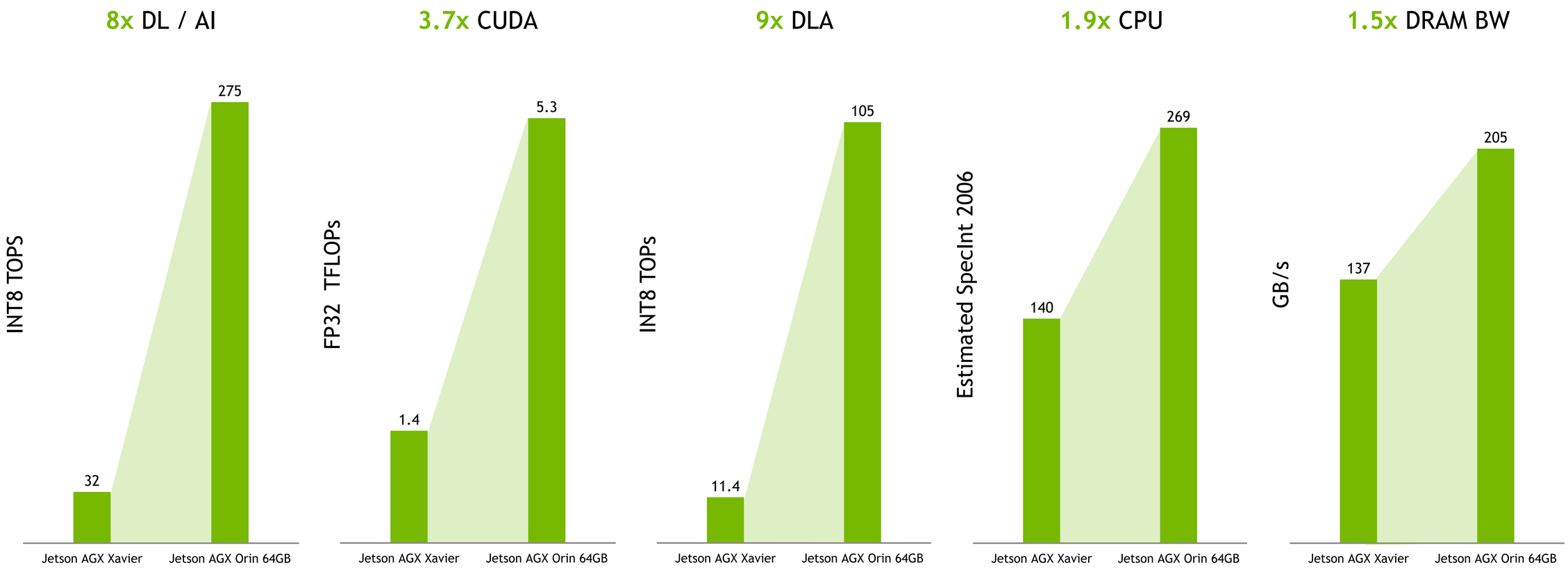
- Support for up to 4x Orin SOCs with direct high-speed connections
- Gen4 PCIe x4
  - Support Root Port and End Point Modes
- 10 Gb Ethernet



ADDITIONAL ENHANCEMENTS

- AV1 Video Encode & Decode support
- 8K60 Display Support
- 10Gb Ethernet
- Improved Optical Flow Accelerator
- Improved ISP
- Gen4 PCIe

JETSON AGX ORIN
UP TO 8X PERFORMANCE

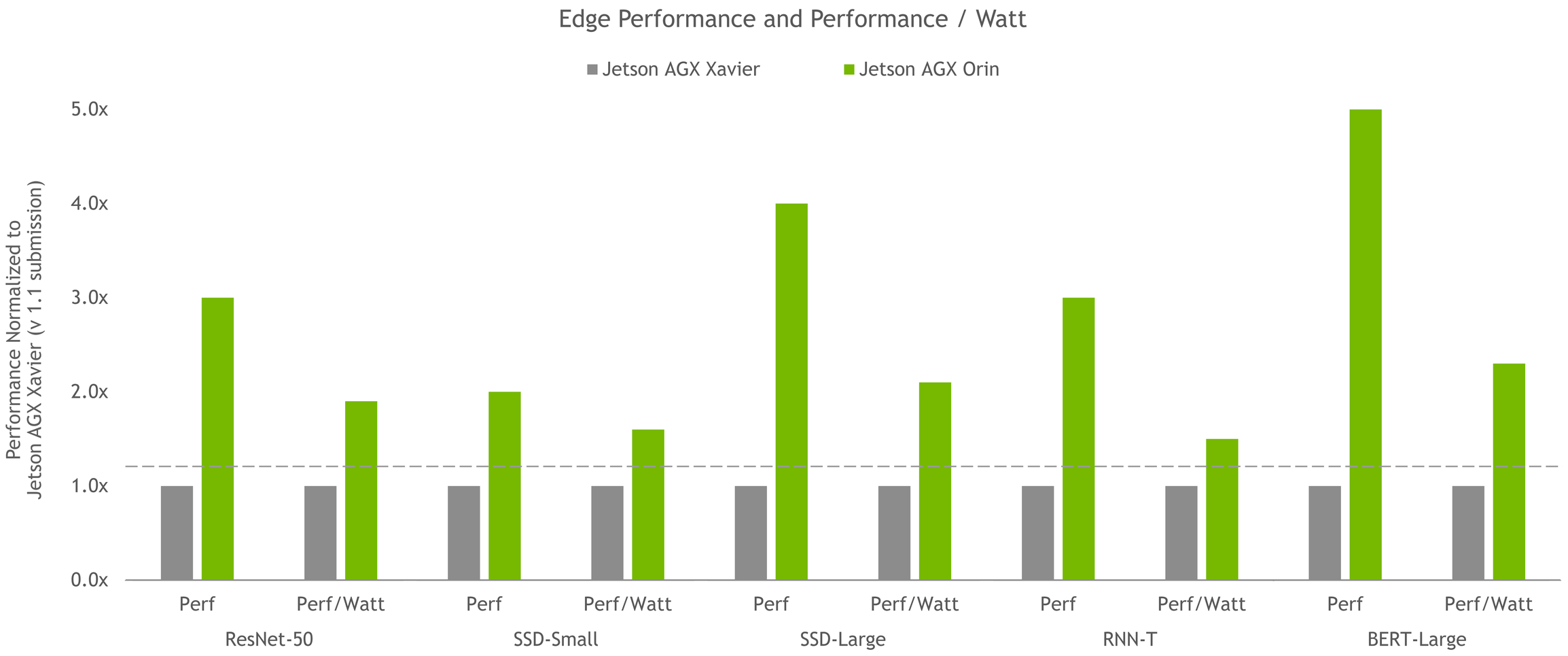


*MaxN performance

NVIDIA

JETSON BLASTS AHEAD

Delivers Up to 5x More Inference Performance and 2.3x Energy Efficiency

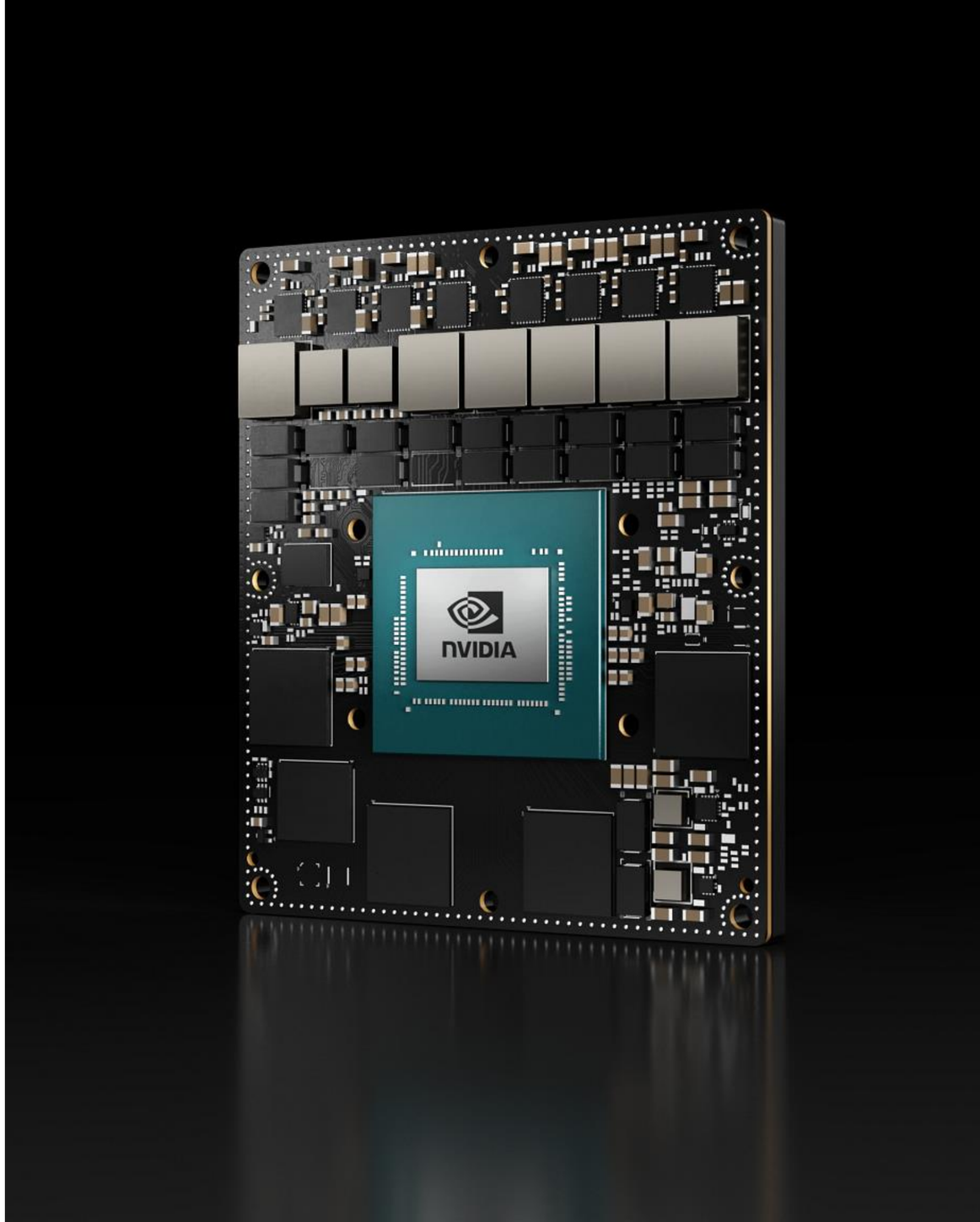


MLPerf v2.0 Inference Edge Closed and Edge Closed Power; Performance/Watt from MLPerf results for respective submissions for Data Center and Edge, Offline Throughput and Power.
NVIDIA Xavier AGX Xavier: 1.1-110 and 1.1-111 | Jetson AGX Orin: 2.0-140 and 2.0-141.

JETSON ORIN

Jetson AGX and Jetson NX Orin based products

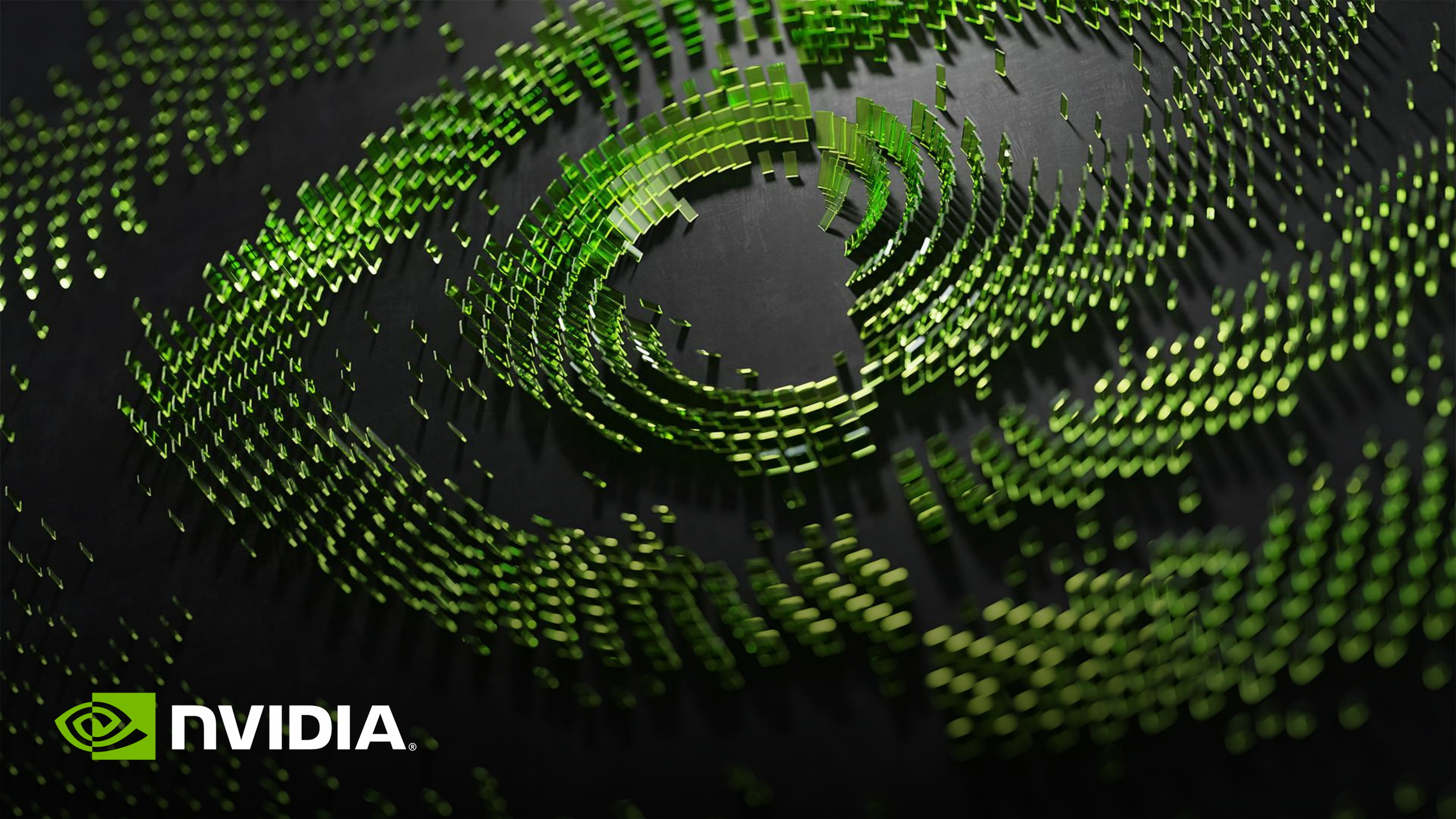
- Up to 275 INT8 TOPS powered by Ampere GPU +DLA
- Up to 12x A78AE ARM CPUs
- Up to 64 GB memory, 204 GB/s
- TDP from 10W – 60W



DRIVE ORIN AUTOMOTIVE

High Performance scaling to 4 Orin

- Up to 254 INT8 TOPS powered by Ampere GPU +DLA
- Up to 12x A78AE ARM CPUs
- Up to 204 GB/s
- 50-60W air cooled, 100W liquid cooled
- Scaling to 4 high BW connected Orin
  - Connectivity via Gen4 PCIe x4 or 10GbE


NVIDIA