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

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC CPU2006 speed integer single core</td>
<td>31.8*</td>
</tr>
<tr>
<td>SPEC CPU2006 rate integer 12-core</td>
<td>269.5*</td>
</tr>
<tr>
<td>SPEC CPU2006 speed floating-point single core</td>
<td>41.6*</td>
</tr>
<tr>
<td>SPEC CPU2006 rate floating-point 12-core</td>
<td>332.0*</td>
</tr>
<tr>
<td>SPEC CPU2017 rate integer single core</td>
<td>4.04*</td>
</tr>
<tr>
<td>SPEC CPU2017 rate integer 12-core</td>
<td>39.36*</td>
</tr>
<tr>
<td>Geekbench 5 single core</td>
<td>754</td>
</tr>
<tr>
<td>Geekbench 5 12-core</td>
<td>7773</td>
</tr>
</tbody>
</table>

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

Diagram:
- R52 = R52
- 32 KB L1 I = 32 KB L1 D
- 512KB Tightly Coupled Memory
- 3MB SRAM
- Safety Island Fabric
- Power Isolation
- SOC
- HW Safety Mgr
- SPI
- CAN
- UART
- GPIO
- HW Sync/IPC/Timers
- DMA
### ORIN SAFETY ISLAND
#### CPU & Memory Configuration

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

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

8x DL / AI

3.7x CUDA

9x DLA

1.9x CPU

1.5x DRAM BW

INT8 TOPS

FP32 TFLOPs

INT8 TOPs

Estimated SpecInt 2006

GB/s

Jetson AGX Xavier

Jetson AGX Orin 64GB

Jetson AGX Xavier

Jetson AGX Orin 64GB

Jetson AGX Xavier

Jetson AGX Orin 64GB

Jetson AGX Xavier

Jetson AGX Orin 64GB

Jetson AGX Xavier

Jetson AGX Orin 64GB

Jetson AGX Xavier

Jetson AGX Orin 64GB

*MaxN performance
JETSON BLASTS AHEAD
Delivers Up to 5x More Inference Performance and 2.3x Energy Efficiency

Edge Performance and Performance / Watt

- Jetson AGX Xavier
- Jetson AGX Orin

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.
MLPerf name and logo are trademarks. See www.mlcommons.org for more information.
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