# NVIDIA Jetson AGX Orin Series

## A Giant Leap Forward for Robotics and Edge AI Applications

## Technical Brief

By Leela S. Karumbunathan

# Document History

TB_10749-001_v1.1

| Version | Date | Description of Change |
|---|---|---|
| 1.0 | November 2021 | Initial Release |
| 1.1 | March 2022 | Updated with the latest Jetson AGX Orin series |

# Table of Contents

# Introduction

Today's Autonomous Machines and Edge Computing systems are defined by the growing needs of AI software. Fixed function devices running simple convolutional neural networks for inferencing tasks like object detection and classification are not able to keep up with new networks that appear every day: transformers are important for natural language processing for service robots; reinforcement learning can be used for manufacturing robots that operate alongside humans; and autoencoders, long short-term memory (LSTM), and generative adversarial networks (GAN) are needed for various applications.

The NVIDIA® Jetson™ platform is the ideal solution to solve the needs of these complex AI systems at the edge. The platform includes Jetson modules, which are small form-factor, high-performance computers, the JetPack SDK for end-to-end AI pipeline acceleration, and an ecosystem with sensors, SDKs, services, and products to speed up development. Jetson is powered by the same AI software and cloud-native workflows used across other NVIDIA platforms and delivers the performance and power-efficiency customers need to build software-defined intelligent machines at the edge. For advanced robotics and other autonomous machines in the fields of manufacturing, logistics, retail, service, agriculture, smart city, and healthcare the Jetson platform is the ideal solution.

The newest members of the Jetson Family, the Jetson AGX Orin series, provide a giant leap forward for Robotics and Edge AI. With Jetson AGX Orin modules, customers can now deploy large and complex models to solve problems such as natural language understanding, 3D perception and multi-sensor fusion. In this technical brief we identify details on the new architecture of the Jetson AGX Orin series and steps customers can take to leverage the full capabilities of the Jetson platform.

# Jetson AGX Orin Series Hardware Architecture

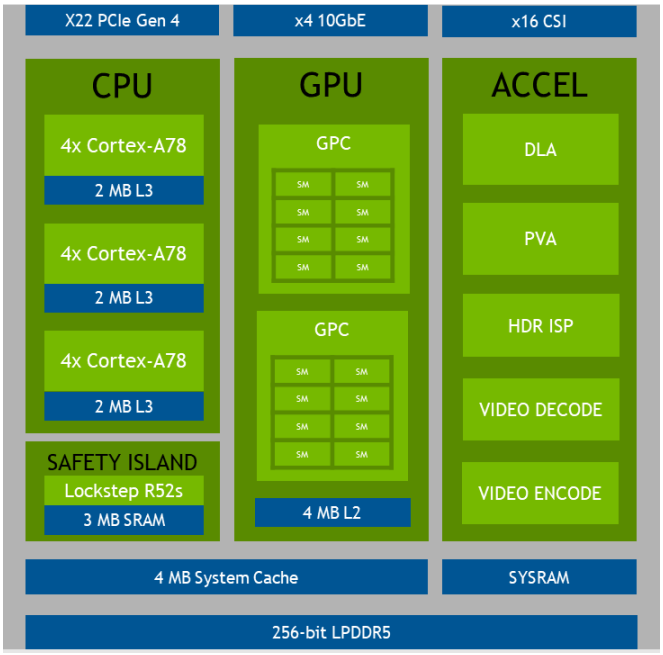
The NVIDIA® Jetson AGX Orin™ series provides server class performance, delivering up to 275 TOPS of AI performance for powering autonomous systems. The Jetson AGX Orin series includes the Jetson AGX Orin 64GB and the Jetson AGX Orin 32GB modules. These power-efficient system-on-modules (SOMs) are form-factor and pin-compatible with Jetson AGX Xavier™ and offer up to 8X AI performance. Jetson AGX Orin modules feature the NVIDIA Orin SoC with a NVIDIA Ampere architecture GPU, Arm® Cortex®-A78AE CPU, next-generation deep learning and vision accelerators, and a video encoder and a video decoder. High speed IO, 204 GB/s of memory bandwidth, and 32GB or 64GB of DRAM enable these modules to feed multiple concurrent AI application pipelines. With the SOM design, NVIDIA has done the heavy lifting of designing around the SoC to provide not only the compute and I/O but also the power and memory design. For more details, reference our Jetson AGX Orin Series Data Sheet[1].

Figure: 1 Jetson AGX Orin delivers 8X the AI performance of Jetson AGX Xavier


Note: Jetson AGX Orin 64GB Max Performance. Jetson AGX Orin 32GB Performance scales based on the number and frequencies of the CPU, GPU, and DLA.

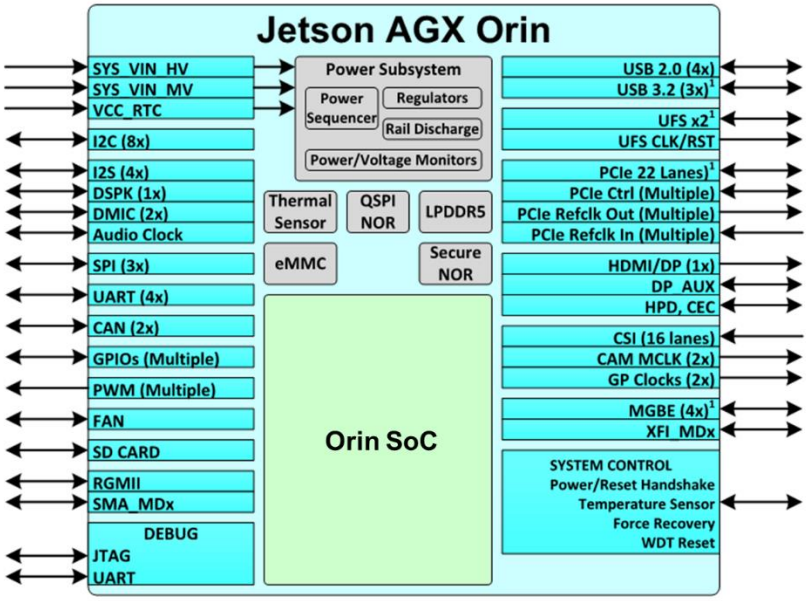
[1] Jetson AGX Orin Data Sheet

Figure 2: Orin System-on-Chip (SoC) Block Diagram




NOTE: Jetson AGX Orin 32GB will have 2x 4 Core Clusters, and 7 TPCs with 14 SMs

Figure 3: Jetson AGX Orin Series System-On-Module




NOTE: One USB 3.2 port, UFS, and MGBE shares UPHY lanes with PCIe

**Table 1: Jetson AGX Orin Series Technical Specifications**

| | Jetson AGX Orin 32GB | Jetson AGX Orin 64GB |
|---|---|---|
| **AI Performance** | 200 TOPS (INT8) | 275 TOPS (INT8) |
| **GPU** | NVIDIA Ampere architecture with 1792 NVIDIA® CUDA® cores and 56 Tensor Cores | NVIDIA Ampere architecture with 2048 NVIDIA® CUDA® cores and 64 Tensor Cores |
| **Max GPU Freq** | 939 MHz | 1.3 GHz |
| **CPU** | 8-core Arm® Cortex®-A78AE v8.2 64-bit CPU<br>2MB L2 + 4MB L3 | 12-core Arm® Cortex®-A78AE v8.2 64-bit CPU<br>3MB L2 + 6MB L3 |
| **CPU Max Freq** | 2.2 GHz | |
| **DL Accelerator** | 2x NVDLA v2.0 | |
| **DLA Max Frequency** | 1.4 GHz | 1.6 GHz |
| **Vision Accelerator** | PVA v2.0 | |
| **Memory** | 32GB 256-bit LPDDR5<br>204.8 GB/s | 64GB 256-bit LPDDR5<br>204.8 GB/s |
| **Storage** | 64GB eMMC 5.1 | |
| **CSI Camera** | Up to 6 cameras (16 via virtual channels*)<br>16 lanes MIPI CSI-2<br>D-PHY 2.1 (up to 40Gbps) \| C-PHY 2.0 (up to 164Gbps) | |
| **Video Encode** | 1x 4K60 \| 3x 4K30 \| 6x 1080p60 \| 12x 1080p30 (H.265)<br>H.264, AV1 | 2x 4K60 \| 4x 4K30 \| 8x 1080p60 \| 16x 1080p30 (H.265)<br>H.264, AV1 |
| **Video Decode** | 1x 8K30 \| 2x 4K60 \| 4x 4K30 \| 9x 1080p60\| 18x 1080p30 (H.265)<br>H.264, VP9, AV1 | 1x 8K30 \| 3x 4K60 \| 7x 4K30 \| 11x 1080p60\| 22x 1080p30 (H.265)<br>H.264, VP9, AV1 |
| **UPHY** | Up to 2 x8, 1 x4, 2 x1 (PCIe Gen4, Root Port & Endpoint)<br>3x USB 3.2<br>Single lane UFS | |
| **Networking** | 1x GbE<br>4x 10GbE | |
| **Display** | 1x 8K60 multi-mode DP 1.4a (+MST)/eDP 1.4a/HDMI 2.1 | |
| **Other I/O** | 4x USB 2.0<br>4x UART, 3x SPI, 4x I2S, 8x I2C, 2x CAN, DMIC & DSPK, GPIOs | |
| **Power** | 15W - 40W | 15W - 60W |
| **Mechanical** | 100mm x 87mm<br>699-pin Molex Mirror Mezz Connector<br>Integrated Thermal Transfer Plate | |

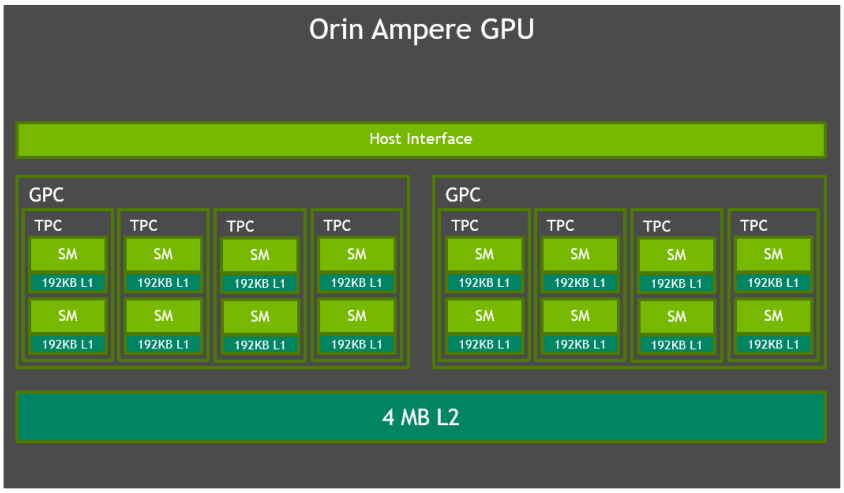
*Virtual Channel related camera information for Jetson AGX Orin is not final and subject to change.
NOTE: Refer to the Software Features section of the latest NVIDIA Jetson Linux Developer Guide for a list of supported features

# GPU

Jetson AGX Orin modules contain an integrated Ampere GPU composed of 2 Graphic Processing Clusters (GPCs), up to 8 Texture Processing Clusters (TPCs), up to 16 Streaming Multiprocessors (SM's), 192 KB of L1-cache per SM, and 4 MB of L2 Cache. There are 128 CUDA cores per SM for Ampere compared to the 64 CUDA cores for Volta, and four 3rd Generation Tensor cores per SM. Jetson AGX Orin 64GB has 2048 CUDA cores and 64 Tensor cores with up to 170 Sparse TOPs of INT8 Tensor compute, and up to 5.3 FP32 TFLOPs of CUDA compute. Jetson AGX Orin 32GB has 7 TPCs with 1792 CUDA cores and 56 Tensor cores with up to 108 Sparse TOPs of INT8 Tensor compute, and up to 3.37 FP32 TFLOPs of CUDA compute.

We have enhanced the Tensor cores with a big leap in performance compared to the previous generation. With the Ampere GPU, we bring support for sparsity. Sparsity is a fine-grained compute structure that doubles throughput and reduces memory usage.

Figure 4: Orin Ampere GPU Block Diagram




Note: The above diagram shows Jetson AGX Orin 64GB. Jetson AGX Orin 32GB will have 7 TPCs and 14 SMs.

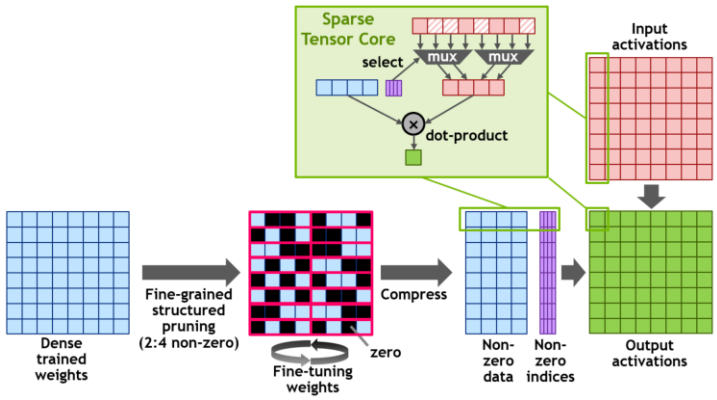
# 3rd Generation Tensor Cores and Sparsity

NVIDIA Tensor cores provide the performance necessary to accelerate next generation AI applications. Tensor cores are programmable fused matrix-multiply-and-accumulate units that execute concurrently alongside the CUDA cores. Tensor cores implement floating point HMMA (Half-Precision Matrix Multiply and Accumulate) and IMMA (Integer Matrix Multiple and Accumulate) instructions for accelerating dense linear algebra computations, signal processing, and deep learning inference.[2]

[2] NVIDIA Jetson AGX Xavier Developer Blog

Ampere brings support for the third-generation Tensor cores, which enable support for 16x HMMA, 32x IMMA, and a new sparsity feature.[3] With the sparsity feature, customers can take advantage of the fine-grained structured sparsity in deep learning networks to double the throughput for Tensor core operations. Sparsity is constrained to 2 out of every 4 weights being nonzero. It enables a Tensor core to skip zero values, doubling the throughput and reducing memory storage significantly. Networks can be trained first on dense weights, then pruned, and later fine-tuned on sparse weights.

Figure 5: Ampere GPU 3rd Generation Tensor Core Sparsity




## Get the most out of the Ampere GPU using NVIDIA Software Libraries

Customers can accelerate their inferencing on the GPU using NVIDIA TensorRT and cuDNN. NVIDIA TensorRT is a runtime library and optimizer for deep learning inference that delivers lower latency and higher throughput across NVIDIA GPU products. TensorRT enables customers to parse a trained model and maximize the throughput by quantizing models to INT8, optimizing use of the GPU memory and bandwidth by fusing nodes in a kernel, and selecting the best data layers and algorithms based on the target GPU.

cuDNN (CUDA Deep Neural Network Library) is a GPU-accelerated library of primitives for deep neural networks. It provides highly tuned implementations of routines commonly found in DNN applications like convolution forward and backward, cross-correlation, pooling forward and backward, softmax forward and backward, tensor transformation functions, and more. With the Ampere GPU and NVIDIA software stack, customers are able to handle new, complex neural networks that are being invented every day.

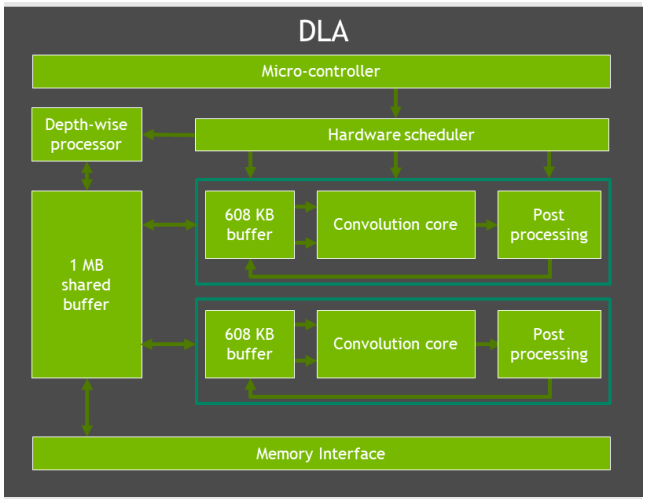
[3] NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf

# DLA

The NVIDIA Deep Learning Accelerator, or DLA, is a fixed-function accelerator optimized for deep learning operations. It is designed to do full hardware acceleration of convolutional neural network inferencing. The Orin SoC brings support for the next generation NVDLA 2.0 with 9X the performance of NVDLA 1.0.

DLA 2.0 provides a highly energy efficient architecture. With this new design, NVIDIA increased local buffering for even more efficiency and reduced DRAM bandwidth. DLA 2.0 additionally brings a set of new features including structured sparsity, depth wise convolution, and a hardware scheduler. This enables up to 105 INT8 Sparse TOPs total on Jetson AGX Orin DLAs compared with 11.4 INT8 Dense TOPS total on Jetson AGX Xavier DLAs.

Figure 6: Orin Deep Learning Accelerator (DLA) Block Diagram




# TensorRT supports DLA

Customers can use TensorRT to accelerate their models on the DLAs just as they do on the GPU. NVIDIA DLAs are designed for offloading deep learning inferencing from the GPU, enabling the GPU to run more complex networks and dynamic tasks. TensorRT supports running networks in either INT8 or FP16 on DLA, and supports various layers such as convolution, deconvolution, fully connected, activation, pooling, batch normalization, and more. More information on the DLA support in TensorRT can be found here: Working With DLA[4]. NVIDIA DLAs enables support for a diversity of models and algorithms to achieve 3D construction, path planning, semantic understanding, and more. Depending on what type of compute is needed, both the DLA and the GPU can be used to achieve full application acceleration.

[4] Working With DLA

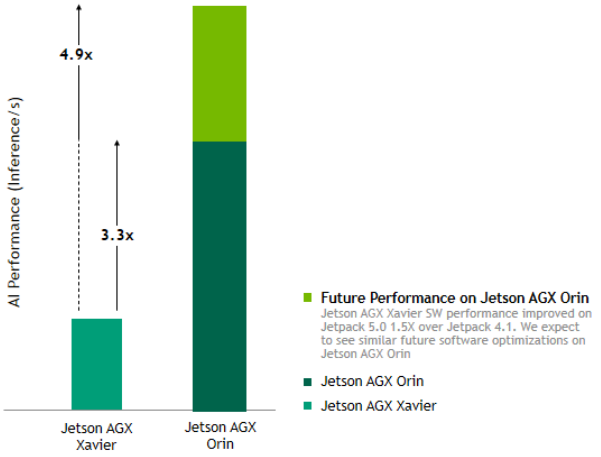
# Giant Leap Forward in Performance

With GPU and DLA enhancements, the Jetson AGX Orin series provides a giant leap forward in performance. A new age of robotics is emerging with computational requirements increasing by orders of magnitude for functionality such as multi-sensor perception, mapping and localization, path planning and control, situational awareness, and safety.

In particular, robotics and other Edge AI applications are requiring increased amounts of AI for computer vision and conversational AI. The Jetson AGX Orin module deliver up to 3.3 times the performance of Jetson AGX Xavier on real world AI applications, as can be seen with our pretrained models. We expect this will increase to an almost 5X performance improvement with future software updates. (Jetson AGX Xavier saw a 1.5X performance increase from when it was launched to now, with the most recent Jetpack software.)

Figure 7: Real World AI Performance on Jetson AGX Orin

| | Jetson AGX Xavier | Jetson AGX Orin |
|---|---|---|
| PeopleNet | 196 | 536 |
| Action Recognition 2D | 471 | 1577 |
| Action Recognition 3D | 32 | 105 |
| LPR | 1190 | 4118 |
| Dashcam Net | 671 | 1908 |
| Bodypose Net | 172 | 559 |
| ASR: Citrinet 1024 | 34 | 113 |
| NLP: BERT-base | 94 | 186 |
| TTS: Fastpitch-HifiGAN | 9 | 32 |

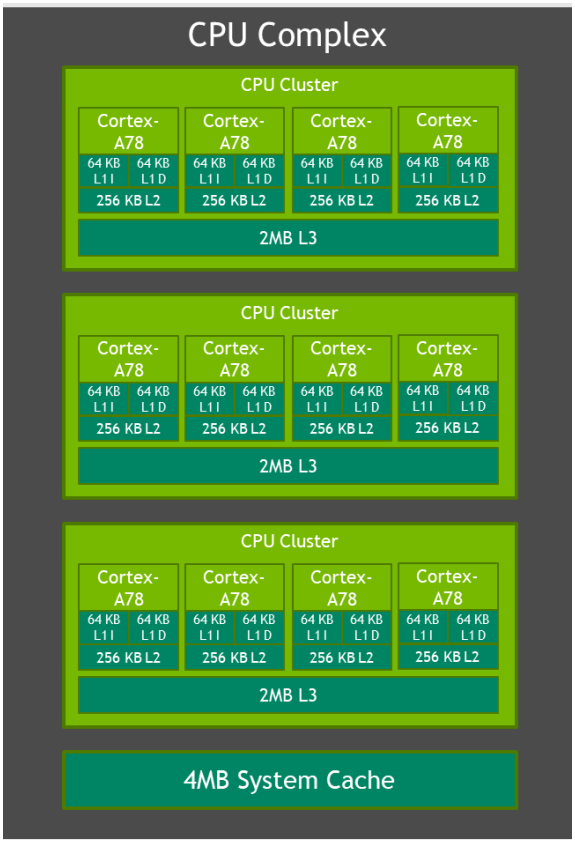



NOTE: These benchmarks were run on the Jetson AGX Orin Developer Kit.

# CPU

For Jetson AGX Orin series modules, we moved from the NVIDIA Carmel CPU to the Arm Cortex-A78AE. The Orin CPU complex has up to 12 CPU cores. Each core includes 64KB Instruction L1 Cache and 64KB Data Cache, and 256 KB of L2 Cache. Like Jetson AGX Xavier, each cluster also has 2MB L3 Cache. The maximum supported CPU frequency 2.2 GHz.

Figure 8: Orin CPU Block Diagram



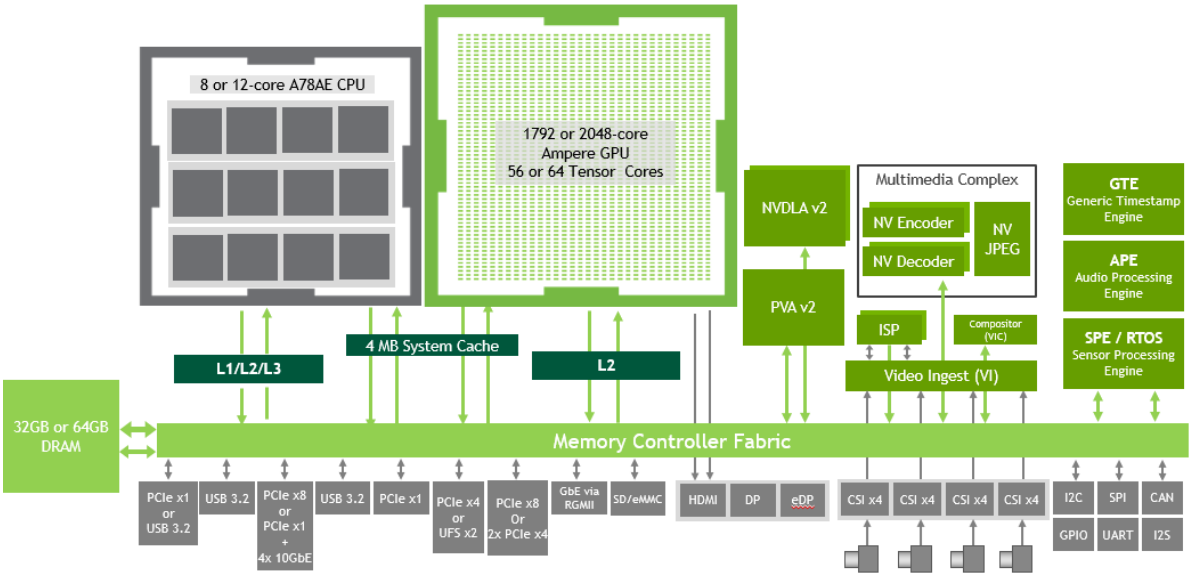

Note: The above diagram shows Jetson AGX Orin 64GB. Jetson AGX Orin 32GB will have 2x 4 Core Clusters.

The 12-core CPU on Jetson AGX Orin 64GB enables almost 1.9 times the performance compared to the 8-core NVIDIA Carmel CPU on Jetson AGX Xavier. Customers can use the enhanced capabilities of the Cortex-A78AE including the higher performance and enhanced cache to optimize their CPU implementations.

# Memory & Storage

Jetson AGX Orin modules bring support for 1.4X the memory bandwidth and 2X the storage of Jetson AGX Xavier, enabling 32GB or 64GB of 256-bit LPDDR5 and 64 GB of eMMC. The DRAM supports a max clock speed of 3200 MHz, with 6400 Gbps per pin, enabling 204.8 GB/s of memory bandwidth. Figure 8 highlights how each of the various components interact with the Memory Controller Fabric and the DRAM.

Figure 9: Jetson AGX Orin Series Functional Block Diagram




# Video Codecs

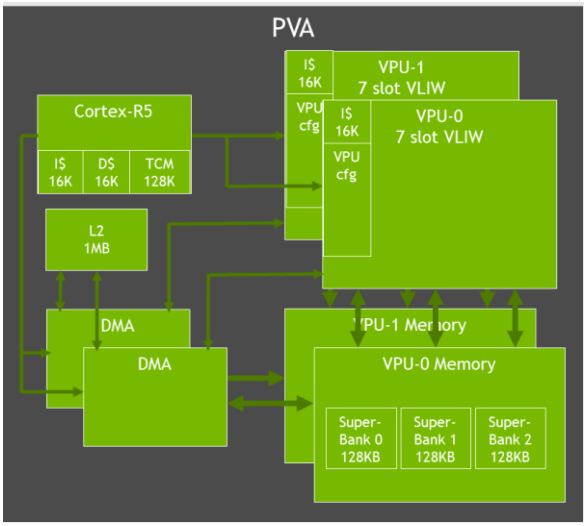
Jetson AGX Orin modules contain a Multi-Standard Video Encoder (NVENC), a Multi-Standard Video Decoder (NVDEC), and a JPEG processing block (NVJPEG). NVENC enables full hardware acceleration for various encoding standards including H.265, H.264, and AV1. NVDEC enables full hardware acceleration for various decoding standards including H.265, H.264, AV1, VP9. NVJPG is responsible for JPEG (de)compression calculations (based on the JPEG still image standard), image scaling, decoding (YUV420, YUV422H/V, YUV444, YUV400) and color space conversion (RGB to YUV). Please reference the Jetson AGX Orin Data Sheet[5]for a full list of standards. Customers can leverage NVIDIA Jetson's Multimedia API to power these engines. The Multimedia API[6] is a collection of low-level APIs that supports flexible application development across these engines.

[5] Jetson AGX Orin Data Sheet
[6] Multimedia API

# PVA & VIC

Jetson AGX Orin modules bring support for our next generation Programmable Vision Accelerator engine, PVA v2. The PVA engine includes dual 7-way VLIW (Very Long Instruction Word) vector processing units, dual DMA engines, and a Cortex-R5 subsystem. The PVA enables support for various computer vision kernels such as filtering, warping, image pyramid, feature detection, and FFT. Some common computer vision applications using the PVA include feature detector, feature tracker, object tracker, stereo disparity, and visual perception.

Figure 10: Orin PVA Block Diagram

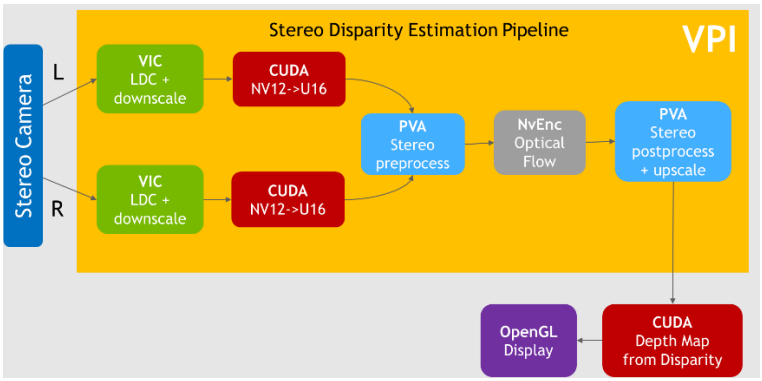



The Orin SoC also contains a Gen 4.2 Video Imaging Compositor (VIC) 2D Engine. The VIC enables support for various image processing features like lens distortion correction and enhanced temporal noise reductions, video features like sharpness enhancement, and general pixel processing features like color space conversion, scaling, blend, and composition.

Vision Programming Interface (VPI) is a software library which implements computer vision and image processing algorithms on several NVIDIA Jetson hardware components including PVA, VIC, CPU, and GPU. With VPI algorithm support on the PVA and the VIC, one can offload computer vision and image processing tasks to them and prioritize the CPU and the GPU for other tasks.

As an example, a complete Stereo Disparity Estimation pipeline using VPI can efficiently use several backends including the VIC, PVA, and NVENC. The pipeline receives the input from a stereo camera, which are left and right images of a stereo pair. The VIC works on this input to correct lens distortion and scale the image down, resulting in a rectified stereo pair. Then images get converted from color to grayscale using the GPU, with the results fed into a sequence of operations using PVA and NVENC as backends. The output is an estimate of the disparity between the input images, which is related to the scene depth.

Figure 11: Stereo Disparity Estimation Pipeline




VPI comes with several algorithms ranging from image processing building blocks like box Filtering, Convolution, Image Rescaling and Remap, to more complex computer vision algorithms like Harris Corners Detection, KLT Feature Tracker, Optical Flow, Background Subtraction, and more. Please check out the VPI Webinars here[7] to learn more about using VPI to accelerate computer vision applications.

[7] VPI Webinar

# I/O

Jetson AGX Orin series modules contain plenty of high speed I/O including 22 lanes of PCIe Gen4, Gigabit Ethernet, 4 XFI interfaces for 10 Gigabit Ethernet, a Display Port, 16 lanes of MIPI CSI-2, USB3.2 interfaces as well as various other I/O like $I^2C$, $I^2S$, SPI, CAN, GPIO, USB 2.0, DMIC and DSPK. Customers can leverage the UPHY lanes for USB 3.2, UFS, PCIe, and MGBE, and some of the UPHY lanes are shared between these interfaces. All 22 lanes of PCIe support root port mode, and some support endpoint mode as well. The Display Port can support 2 displays using the Multi-stream mode on DP1.4. Jetson is used across a variety of applications with various I/O requirements. For example, Autonomous Ground Vehicles could leverage the CSI cameras for a surround view around the robot, I2S for voice commands, HDMI for display, PCIe for Wi-Fi, GPIO & $I^2C$ and more. A video analytics application like traffic management at an intersection might require many GigE Cameras and Ethernet for networking purposes. As autonomous machines continue to perform more advanced tasks, more I/O is needed to interface more sensors.

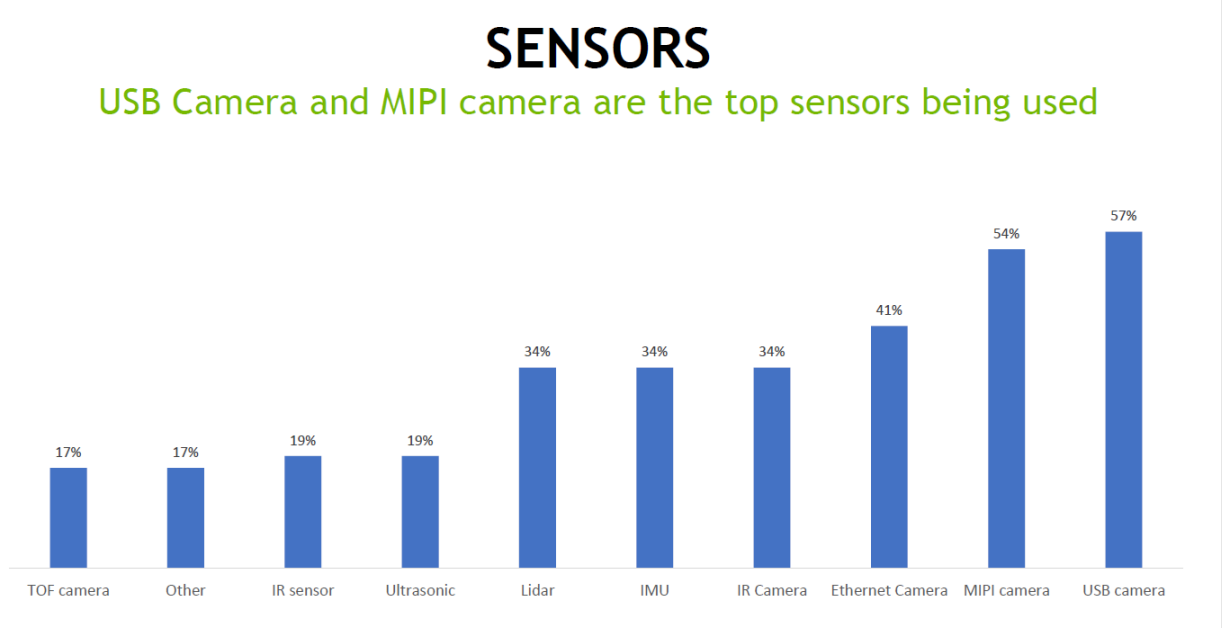
**Table 2: Jetson AGX Orin I/O**

| | |
|---|---|
| **CSI Camera** | Up to 6 cameras (16 via virtual channels*)<br>16 lanes MIPI CSI-2<br>D-PHY 1.2 (up to 40Gbps) \| C-PHY 1.1 (up to 164Gbps) |
| **UPHY** | Up to 2 x8 , 1 x4, 2 x1 (PCIe Gen4, Root Port & Endpoint)<br>3x USB 3.2<br>Single-lane UFS |
| **Networking** | 1x GbE<br>4x 10GbE |
| **Display** | 1x 8K60 multi-mode DP 1.4a (+MST)/eDP 1.4a/HDMI 2.1 |
| **Other I/O** | 4x USB 2.0<br>4x UART, 3x SPI, 4x I2S, 8x I2C, 2x CAN, DMIC & DSPK, GPIOs |

NOTE: Refer to the Software Features section of the latest NVIDIA Jetson Linux Developer Guide for a list of supported feature

In a survey NVIDIA conducted in March of 2021, when asked what sensors and sensor interfaces are being used in your projects, most customers responded that they used cameras via USB, MIPI, and Ethernet. Jetson AGX Orin not only enables all these interfaces, but also supports all the other sensors listed in the survey.

Figure 12: Jetson Customer and Developer Survey—Sensor Usage




# Power Profiles

Jetson AGX Orin series modules are designed with a high efficiency Power Management Integrated Circuit (PMIC), voltage regulators, and power tree to optimize power efficiency. Jetson AGX Orin 64GB supports three optimized power budgets: 15W, 30W, and 50W. Each power mode caps various component frequencies, and the number of online CPU, GPU TPC, DLA, and PVA cores. Jetson AGX Orin 64GB also supports a MAXN performance mode that can enable up to 60W of performance. Jetson AGX Orin 32GB supports power modes between 15W and 40W. Customers can leverage the nvpmodel tool in Jetson Linux to use one of these pre-optimized power modes or to customize their own power mode within the design constraints provided in our documentation.

# Jetson Software

The Jetson platform includes all the necessary software to accelerate your development and get your product quickly to market. At the base of the software stack is the JetPack SDK. It includes the Board Support Package (BSP), with boot loader, Linux kernel, drivers, tool chains and a reference file system based on Ubuntu. The BSP also enables various security features such as secureboot, trusted execution environment, disk and memory encryption and so on. Over the BSP we have several user level libraries for accelerating various parts of your application. These include libraries for accelerating deep learning like CUDA, CuDNN, and Tensor RT; accelerated computing libraries like cuBLAS and cuFTT; accelerated computer vision and image processing libraries like VPI; and multimedia and camera libraries like libArgus and v4l2.

On top of JetPack there are higher level, use case specific SDKs including DeepStream for Intelligent Video Analytics applications, Isaac for Robotics applications, and Riva for Natural Language Processing applications. Surrounding this NVIDIA has a growing partner ecosystem that can provide customers with specialty products and services to accelerate development.

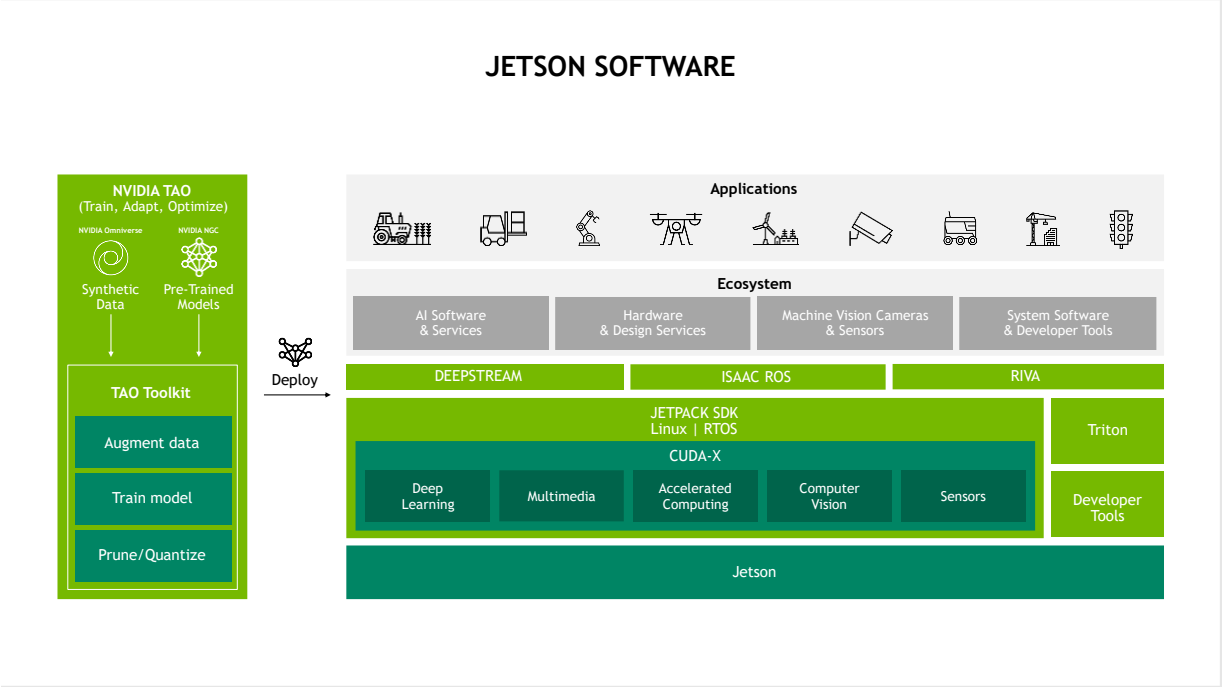
AI at the edge is growing, and the complexity of edge deployment is growing. We see various challenges arising from creating AI products. You need lots of data to train accurately, and you need your models to be optimized for inference. High Performance models from open-source options do not provide desirable results nor are optimized for highest inference throughput. There is also a need to support various frameworks, and a deep understanding of Deep Learning and Data Science. NVIDIA's TAO Toolkit and Pre-Trained Models (PTM) can help solve this challenge. NVIDIA pre-trained models provide customers with accurate models that have been pre-trained with millions of images to achieve state of the art accuracy out of the box. The pre-trained model library can be found here[8], and includes various PTMs like people detection, vehicle detection, natural language processing, pose estimation, license plate detection, and face detection. TAO toolkit enables customers with the ability to easily train, fine-tune, and optimize these pretrained models with their own data set. Customers can then easily deploy these models in production using our various inference SDKs.

Edge computing has historically been characterized by systems that rarely get software updates. With new cloud technologies, you need to be able to periodically update the software on your deployed products, and have the flexibility to adopt, and easily scale and deploy across various environments. Jetson brings Cloud-Native to the edge and enables technologies like containers and container orchestration. NVIDIA JetPack includes NVIDIA Container Runtime with Docker integration, enabling GPU accelerated containerized applications on Jetson platform. Jetpack also brings support for NVIDIA Triton™ Inference Server to simplify the deployment of AI models at scale. Triton Inference Server is open source and provides a single standardized inference platform that can support multi

[8] TAO Toolkit | NVIDIA Developer

framework model inferencing in different deployments such as datacenter, cloud and embedded devices. It supports different types of inference queries through advanced batching and scheduling algorithms to maximize performance of your AI application and supports live model updates with zero inferencing downtime.

Figure 13: Jetson Cloud Native Software Stack




JetPack 5.0 provides the software to power the Jetson AGX Orin and future Jetson modules, as well as existing Jetson modules based on the NVIDIA Xavier SoC. Jetpack 5.0 includes L4T with Linux kernel 5.10 and a reference file system based on Ubuntu 20.04. Jetpack 5.0 enables a full compute stack update with CUDA 11.x and new versions of cuDNN and Tensor RT. It will include UEFI as a CPU bootloader and will also bring support for OP-TEE as a trusted execution environment. Finally, there will be an update to DLA support for NVDLA 2.0, as well as a VPI update to support the next generation PVA v2.

# Jetson AGX Orin Developer Kit

The Jetson AGX Orin Developer Kit will contain everything needed for developers to get up and running quickly. The Jetson AGX Orin Developer Kit includes a Jetson AGX Orin module with heatsink, a reference carrier board, and a power supply. The Jetson AGX Orin Developer kit can be used to develop for all the Jetson Orin modules via emulation modes to emulate their performance. With up to 275 TOPS of AI performance and power configurable between 15 and 60 W, customers now have more than 8X the performance of Jetson AGX Xavier in the same compact form-factor for developing advanced robots and other autonomous machine products.

The Jetson AGX Orin Developer Kit is available today.

Figure 13: Jetson AGX Orin Developer Kit

**Table 3: Jetson AGX Orin Developer Kit Technical Specifications**

| MODULE: | |
|---|---|
| **GPU** | NVIDIA Ampere architecture with 2048 NVIDIA® CUDA® cores and 64 Tensor cores |
| **CPU** | 12-core Arm Cortex-A78AE v8.2 64-bit CPU<br>3MB L2 + 6MB L3 |
| **DL Accelerator** | 2x NVDLA v2.0 |
| **Vision Accelerator** | PVA v2.0 |
| **Memory** | 32GB 256-bit LPDDR5<br>204.8 GB/s |
| **Storage** | 64GB eMMC 5.1 |
| **Power** | 15W to 60W |

| REFERENCE CARRIER BOARD: | |
|---|---|
| **Camera** | 16 lane MIPI CSI-2 connector |
| **PCIe** | x16 PCIe slot supporting<br>x8 PCIe Gen4 |
| **M.2 Key M** | x4 PCIe Gen 4 |
| **M.2 Key E** | x1 PCIe Gen 4, USB 2.0, UART, I2S |
| **USB** | Type C: 2x USB 3.2 Gen2 with USB-PD support<br>Type A: 2x USB 3.2 Gen2, 2x USB 3.2 Gen1<br>Micro-B: USB 2.0 |
| **Networking** | RJ45 (up to 10 GbE) |
| **Display** | DisplayPort 1.4a (+MST) |
| microSD slot | UHS-1 cards up to SDR104 mode |
| **Others** | 40-pin header (I2C, GPIO, SPI, CAN, I2S, UART, DMIC)<br>12-pin automation header<br>10-pin audio panel header<br>10-pin JTAG header<br>4-pin fan header<br>2-pin RTC battery backup connector<br>DC power jack<br>Power, Force Recovery, and Reset buttons |
| **Dimensions** | 110mm x 110mm x 71.65mm<br>(Height includes feet, carrier board, module, and thermal solution) |

**Notice**

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this specification.

Unless specifically agreed to in writing by NVIDIA, NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

**VESA DisplayPort**

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

**HDMI**

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

**Bluetooth**

The Bluetooth® word mark and logos are registered trademarks owned b the Bluetooth SIG, Inc. and any use of such marks by NVIDIA is under license.

**ARM**

ARM, AMBA and ARM Powered are registered trademarks of ARM Limited. Cortex, MPCore and Mali are trademarks of ARM Limited. All other brands or product names are the property of their respective holders. "ARM" is used to represent ARM Holdings plc; its operating company ARM Limited; and the regional subsidiaries ARM Inc.; ARM KK; ARM Korea Limited.; ARM Taiwan Limited; ARM France SAS; ARM Consulting (Shanghai) Co. Ltd.; ARM Germany GmbH; ARM Embedded Technologies Pvt. Ltd.; ARM Norway, AS and ARM Sweden AB.

**Trademarks**

NVIDIA, the NVIDIA logo, CUDA, Jetson, Jetson Xavier, Jetson Orin, NVIDIA Maxwell, NVIDIA Volta, Xavier, and Tegra are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

**Copyright**