Technical White Paper

Edge AI Smart Cameras Using Energy-Efficient AM62A Processor

Reese Grimsley

ABSTRACT

The AM62A Processor is designed for low to mid vision applications requiring one to two cameras. With its innovative AI accelerator, H264/H265 encode/decoder, and built-in image sensor processor (ISP) with RGB-Ir support, the AM62A is well suited for a wide variety of vision-based applications. This includes use cases across industrial and automotive such as security cameras, sport cameras, machine vision systems, retail scanners, in-cabin dash cameras, and front or side cameras for automobiles. Delivering up to two TOPS of AI performance at 2 Watts (typ.) under full load, this AI-accelerated processor enhances cost and power-constrained applications.

Table of Contents

List of Figures

List of Tables

Trademarks

1 Smarter Cameras at the Edge

Artificial Intelligence (AI) on embedded devices at the network edge is growing rapidly as complex data processing and analytics becomes essential in making cities, factories, automobiles, and homes smarter and more efficient. There is a wealth of information in imagery, which humans rely on heavily. Computer vision (CV) and machine learning (ML) extract meaning, for example, where a person is, from the information-dense image. CV and ML are invaluable in improving use cases like defect detection for machine vision, visual odometry and mapping for robots, lane detection for automobiles, and many more. Human-centric applications like identification, biometrics, fall detection, and behavior recognition further push the need for smarter cameras in building access and public security applications.

Cloud services have dominated vision analytics and machine learning inference in recent years as embedded devices lacked the processing power to operate on streams of camera data, such as a security camera producing 1080p at 30+ frames per second (FPS). In 2021, the number of home security cameras was estimated by Yole at 45.1 million, with an expectation to grow by 2.5x by 2026 [1], the recurring costs of processing in the cloud grows linearly with the number of devices, inhibiting scalability. By processing at the edge, this recurring cost is minimized or removed altogether. Edge AI also avoids additional network latency in time-sensitive applications and reduces privacy concerns in human-centric or consumer electronic applications.

SoCs capable of running deep learning and other complex analytics locally are gaining popularity. Market projections estimate 25% of home security cameras will employ Edge AI by 2026 [1] and that this subset of the market will grow with 88% CAGR. Embedded processors are now equipped to handle these applications, but must carefully manage the power vs. performance tradeoff. Battery powered and hand-held devices must stay within a few Watts of power to avoid too quickly draining the battery or becoming too hot to handle. However, they must also retain enough performance to a sufficient analytics processing rate, for example, 15 fps for a home security camera. The right choice for an embedded vision processor depends strongly on the AI performance, image preprocessing, video encode/decode, and power consumption requirements.

2 AM6xA Scalable Portfolio and the AM62A

This section introduces the AM62A, a quad-core Arm® Cortex® A53 microprocessor for embedded vision applications. This device occupies the lower range of the AM6xA ("A" standing for Analytics) portfolio, which features several other embedded vision capable devices, all of which includes deep learning acceleration, ISP, and video encode/decode capabilities. The scalable AM6xA portfolio ensures that there is a device to fit the application's desired point on the power vs. performance curve.

The AM62A is designed for power and cost-sensitive applications requiring one-to-two cameras, consuming less than 2 Watts under full load at room temperature and less than 5 Watts at 125°C.

Like other processors in TI's portfolio, variants of the AM62A are available for applications that do not require all possible device features and benefit from more cost optimization.

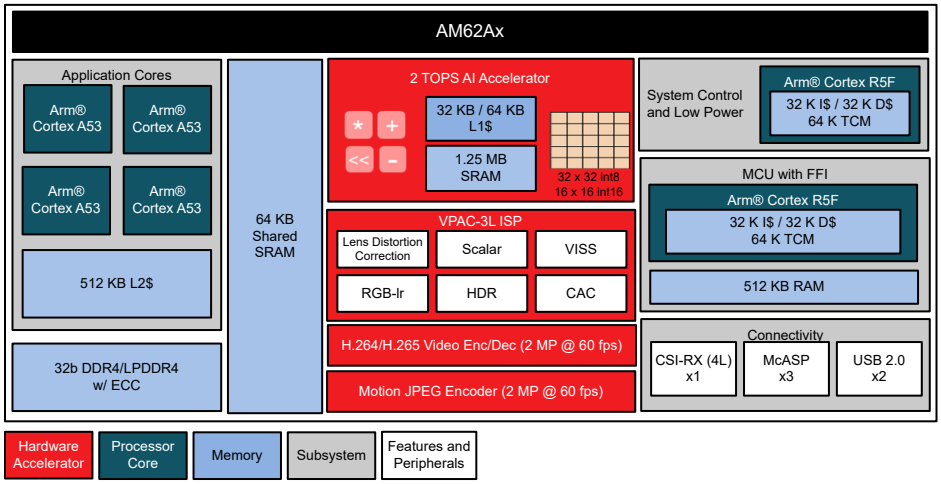

Figure 2-1. AM62A Block Diagram with Subsystems for Smart Camera Applications

3 Smart Camera Use Cases

Applications of smart cameras and computer vision span many domains including industrial, automotive, consumer, and public security. The application requirements and constraints depend on the use case, many of which include reporting an event, such as a security breach or intrusion, over the network to a cloud server. Computing at the edge reduces the impact on the network, yet these applications still often require video encoders, for example,H.265, to limit bandwidth usage when there is need to uplink video data.

Choosing the appropriate image sensor, such that, camera, is important to developing a robust smart system. Key parameters include resolution, frame rate, bit depth, pixel-size, etc. A home security camera may use a 5MP sensor giving 30 frames per second (fps) with rolling shutter whereas an infrastructure camera for tollways may require 2MP with 60fps and global shutter to capture plates on fast-moving vehicles, and a machine vision camera may need 2MP greyscale with 90 fps and global shutter to identiy defects on parts rapidly moving along an assembly line. Image sensors can include an image sensor processor (ISP) that internally preprocess images into a typical format like JPEG or YUV. However, many sensors do not include an ISP to reduce sensor cost and provide freedom in selecting an external ISP that can be more readily tuned than one within the sensor. Selecting a processor with its own integrated ISP, like those in the AM6xA family, gives the benefit of an external ISP, yet with simplified PCB design, lower BOM cost, lower end-to-end latency, and reduced DDR usage.

Applications also have varied computation requirements due to computer vision and machine learning complexity. Machine learning tasks like image classification requires fewer resources than object or keypoint detection, which identifies specific objects like people or vehicles, and their locations. More complex tasks like pixel-level segmentation require even more resources as each pixel is classified as part of an object or region ,such as the current lane for a lane detection algorithm in ADAS applications. Some applications may require multiple models. Increasing the resolution also dramatically increases the processing requirements. AM6xA processors include deep learning acceleration hardware to offload these compute-intensve tasks.

Table 3-1 lists end equipments for AM62A with ranges for required specs/features (min camera resolution, FPS range, low-med-high ML complexity, video enc/dec).

Table 3-1. End Equipments With Necessary Features and Typical Specifications

Use Case	Resolution (Megapixels)	Frame Rate (fps)	AI Complexity	Requires Video Encode	Camera Shutter Type
Surveillance	2-8 MP	10-30	Medium	Yes	Rolling
Machine Vision	5+ MP	60+	Low	No	Global
Infrastructure/Traffic Monitoring	1-5 MP	5-15	Low	Yes	Global
Automotive	2 MP+	30-60	High	No	Global
Driver Monitor / Dash Cam	2-5 MP	15-30	Medium	Yes	Rolling or Global
Sport Camera	2-8 MP	60+ fps	Medium	Yes	Rolling
Item / Code Scanner	1-5 MP	10-30 fps	Medium	No	Rolling

3.1 Security Camera Example

An illustrative example of the AM62A's capabilities is a security camera application. Video frames are captured through the MIPI CSI-2 port using a raw output camera. The processor sends configuration parameters for gain and white balancing back through the same interface. The raw output frame is denoised and de-mosaiced on AM62A's VPAC-3L image sensor processor (ISP), producing usable RGB images. If a wide-angle lens is used, then the lens distortion correction (LDC) accelerator within the ISP will transform the image to reduce warping. The VPAC-3L ISP can further preprocess the image to match the input specification of a machine learning model, such as an object detection network to recognize humans, vehicles, and animals. This ML model runs on the 2 TOPS deep learning acceleration hardware within the processor. The result of the model may be used to trigger an alarm or uplink an H.265-encoded clip of the event over the network. Figure 3-1 depicts this example at a high level, and Figure 3-2 shows the block diagram for example values for resolution, framerate, and pixel-format.

For smart camera applications like security or surveillance, object/keypoint detection deep learning models are most relevant as they may recognize and localize multiple things, such as humans and vehicles, within an image.

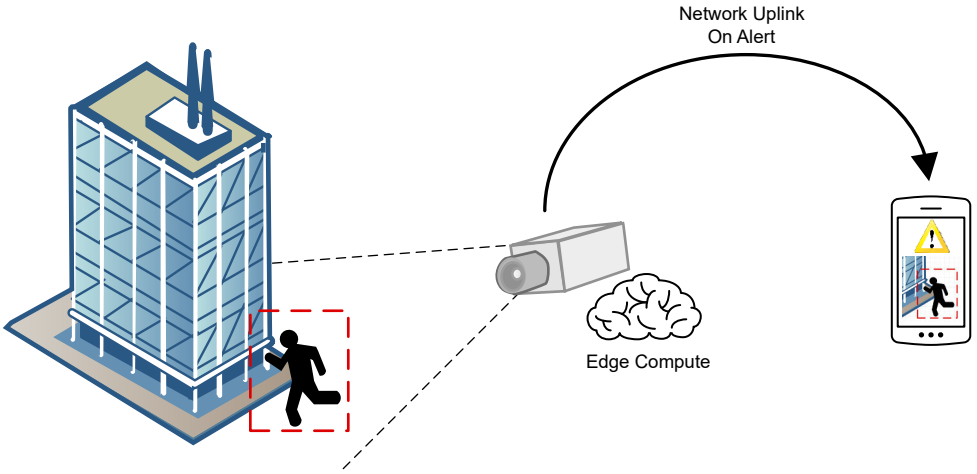

Figure 3-1. Smart Security Camera Recognizes Dangerous Events Locally With Edge AI Before Uplinking H.265-Encoded Video Over the Network

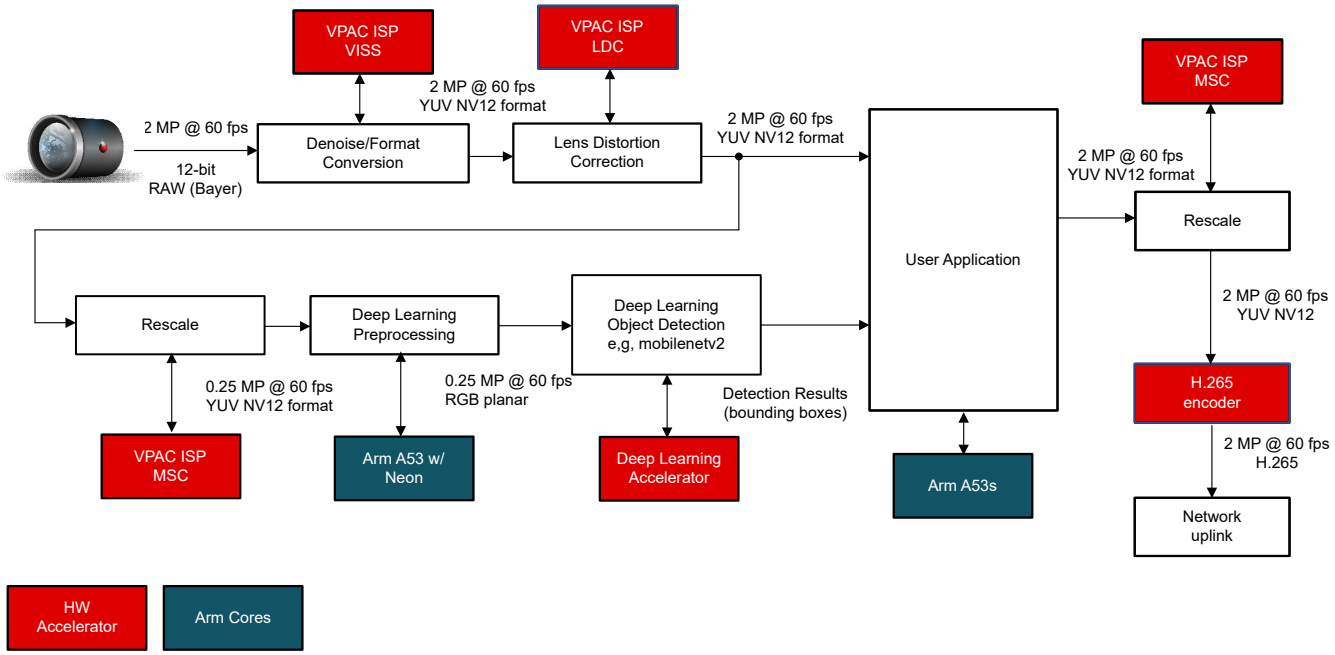

Figure 3-2. Block Diagram of Dataflows for Smart Security Camera Use Case Using Integrated Hardware Accelerators for Offloading ISP, AI, and Video Encoding Functions

4 Deep Learning on the AM62A

Edge AI, while not identical to deep learning (DL), is closely related to it. State-of-the-art neural networks continue to demonstrate superhuman accuracy on vision tasks like classification, object/keypoint detection, and semantic segmentation. Implementing these on devices at the Edge requires the right DL accelerator hardware and software to achieve suitable performance for reliable and efficient systems.

4.1 Deep Learning Accelerator

TI's processors use a state-of-the-art deep learning accelerator design. TI has a long history of digital signal processors (DSPs) that have become increasingly integrated other SoCs at TI; however, a DSP alone is insufficient for most vision deep learning models. Our deep learning accelerator is a tight-coupling of a C7x DSP and a custom matrix-multiply accelerator (MMA), which massively increases performance on neural networks (NN), especially convolutional NNs (CNN) that are common in vision AI.

The AM62A's deep learning accelerator uses a 256-bit C7x DSP and an MMA capable of performing 32x32 matrix multiplies on 8-bit integer values in a single clock cycle. When run at the maximum of 1 GHz, this provides a max compute capacity of 2 TOPS as the 32x32 matrix operation is 1024 multiply-accumulates (MACs, where each MAC is considered two operations). To ensure the MMA always has values to compute, the architecture includes multiple streaming engines that move 256-bits of data each clock cycle to the two input matrices from the single output matrix. Depending on the layers composing the neural network architecture, outputs from the MMA may be sent through the C7x for computing any non-linear functions within the layer. Developers need not program this themselves; API calls from the Arm cores reduce the complexity of programming the accelerator, as described in the Edge AI software section.

While TOPS is a common metric for quantifying machine learning performance on accelerators like TPUs, VPUs, NPUs, and GPUs, etc., one accelerator architecture may outperform another despite having a lower theoretical compute capacity. TI's architecture was designed to optimize power and performance by using a single large compute unit, the MMA, rather than many smaller ones in parallel. With many small units, more transfers to memory are required as there is less data reuse of the same data for subsequent execution cycles. More transfers equate to higher power expenditure. Specially designed data-streaming engines maintain the 256-bit buffers within the accelerator hold the necessary data. A well-optimized application will use a model whose dimensions at each layer completely fill the MMA.

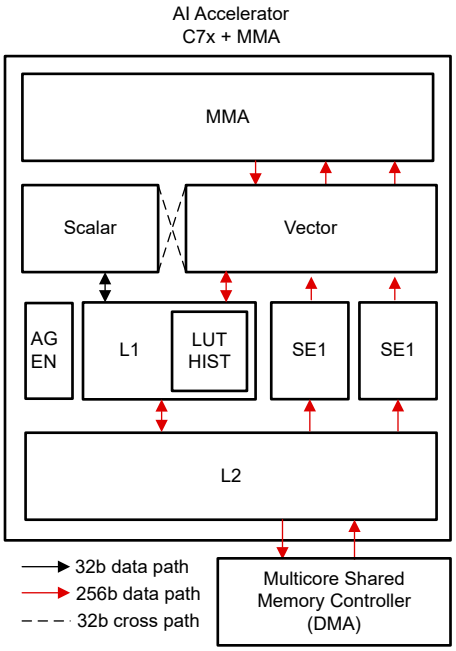

Figure 4-1. AI Accelerator Architecture

4.2 Edge AI Software

TI has put significant effort into simplifying Edge AI development and evaluation on processors, like the AM62A, that contain hardware accelerators for Edge AI [2].

As described in the referenced E2E blogpost, TI has tools to help select a model, train/refine, evaluate, and deploy to the processor with minimal increase in code complexity. Developers can invoke the deep learning accelerator with only a few extra lines to TensorFlow Lite (TFlite), ONNX, or TVM-DLR API calls.

TI's Edge AI Out-of-Box demos in Linux, the dominant operating system for Edge AI applications, further accelerates development of C++ and Python-based applications. These demos take a trained neural network model and an input-output description to run the model with full acceleration from both the C7xMMA and the ISP for a sample end-to-end application. For example, a developer can choose a MobileNetV2SSD trained on COCO dataset from the Texas Instruments Model Zoo as the model, a stored video file for the input, and the HDMI display as the output medium.

These demos are built using Gstreamer to efficiently pipeline image capture, preprocessing, deep learning inference, postprocessing, and further application specific software, including H.264/H.265 encode. TI's custom gstreamer plugins reduce overhead using zero-copy buffering, saving RAM/DDR bandwidth. In addition to Gstreamer and the open source runtimes (TFLite, ONNX, TVM), OpenCV is enabled in our default Linux builds to help developers perform computer vision operations not directly supported by hardware accelerators.

For users outside a Linux environment, the hardware accelerators are exposed via TI's implementation of the OpenVX standard, TIOVX.

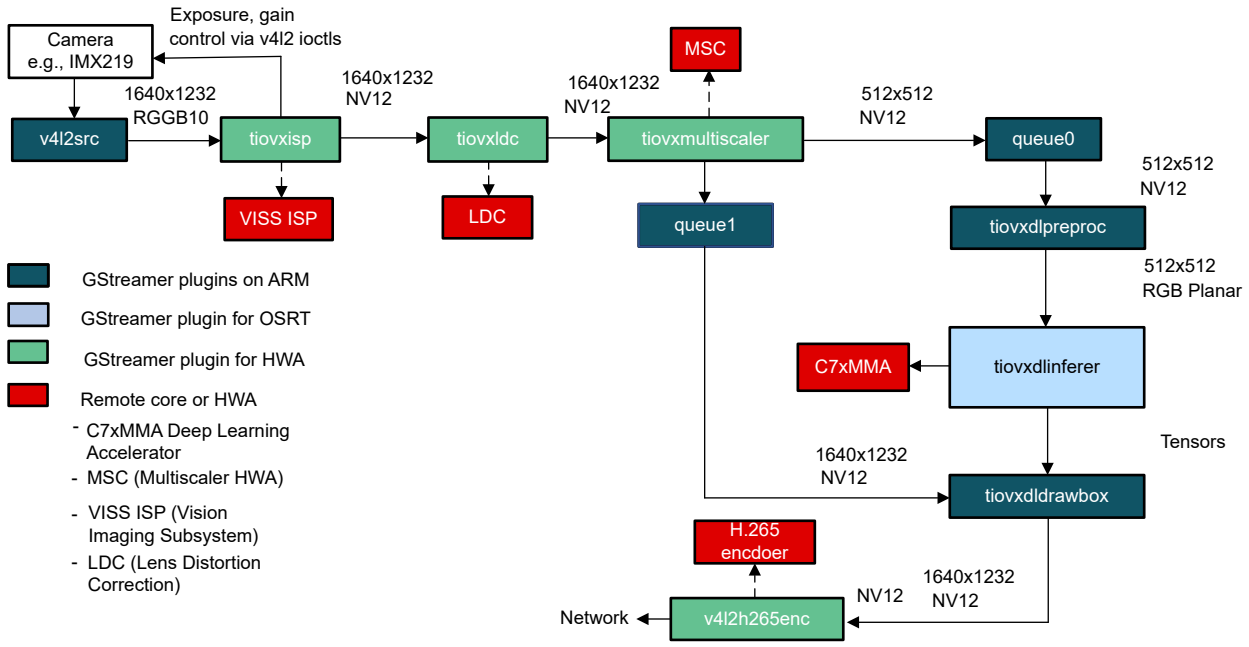

Figure 4-2. Example of Gstreamer Pipeline Using TI's Zero-Buffer Plugins Leveraging Hardware Accelerators and TI's OpenVX Implementation

5 VPAC Vision Accelerator and ISP

The VPAC-3L is a 7th generation image sensor processor (ISP) from TI. This hardware accelerator performs image processing functions to support raw output image sensors.

The Vision Imaging Sub-System (VISS) is a traditional ISP, and supports the following features: de-noising, de-mosaicing, defective pixel correction, gain and white balancing (manual and automatic), chromatic aberration correction (CAC), and wide dynamic range (HDR) encoding. The de-mosaicing feature supports RGB-Ir, allowing high resolution capture of infrared frames without a secondary camera and frame alignment.

VPAC-3L includes a Lens Distortion Correction subsystem for dewarping images that are captured through high field-of-view lens. Cameras with wide angle lenses like fish-eye require correction to bring the image back into a rectangular form and reduce the non-linear stretching effects closer to the edges of the camera frame.

Figure 5-1. Lens Distortion Correction: A) The unprocessed frame from a fisheye lens, B) Front perspective output image C) Intermediate mesh image - warp and interpolate to project image into full-resolution rectangluar format using equisolid fisheye projection model

VPAC-3L also includes a multiscalar (MSC) engine for downscaling images. The MSC allows two inputs and 10 outputs, such that inputs can be copied and scaled independently. One use-case of this is to downscale an image for machine learning and maintain the full-size version so that the output of machine learning can be overlaid on the full resolution frame. The MSC supports up to 4x reduction in each dimension and automatically interpolates where needed, including stretching/shrinking to accommodate changes in aspect ratio. Figure 4-2 makes use of this, in which the "tiovxmultiscalar" gstreamer plugin produces two output paths that are joined again at the"tiovxdldrawbox" plugin for drawing bounding boxes from object detection onto the original image.

TI supports a number of camera sensors and modules and partners with 3rd parties to help developers add new camera sensors with driver support and ISP tuning. The ISP tuning tools also help more experienced ISP tuners perform this work themselves.

6 Low-Power Performance

The AM62A processor is designed to meet the demanding computational requirements of deep learning algorithms while providing high energy efficiency. The device operates at a unique low-core voltage of 0.75 V, which enables the processor to achieve a high level of power efficiency, thus making it suitable for battery-powered and thermal-sensitive applications.

The AM62A can run deep learning processing with a 5 MP 60 fps camera within a power consumption range of 2 W under typical operating conditions, such as room temperature and nominal silicon, which is ideal for battery-powered edge devices that require real-time image recognition and object detection. The device also supports low-power modes, reducing its overall power consumption and extending the battery life in portable applications or relaxing thermal design requirements.

TI provides a power estimation tool to help engineers and designers estimate the power consumption of the AM62A processor in their specific application, taking into account factors such as operating voltage, frequency, junction temperature, and silicon characteristics. This information can be used to optimize power consumption, extend battery life, thermal design and ensure that the device meets the required power budget.

7 Call to Action

The AM62A is ready for developers to design their smart camera applications. Deep learning and image processing acceleration brings all the necessary performance for single and dual camera systems, and with TI's scalable portfolio, the same software can be scaled to processors with higher performance and more interfacing options. For users with less deep learning expertise, TI collaborates with a variety of 3rd parties that can provide their own end-equipment-specific expertise, datasets, and models for use-cases like people tracking, robotics, human-centric analytics, localization and mapping, sensor fusion, and more at Edge AI Demos on our developer Resource Explorer tool.

Designers can get started on their smart camera design with an AM62A starter kit EVM and the Processor SDK. Documentation, prebuilt images, and demos are available from the AM62A product page. Find more info about Edge AI across TI's processors at ti.com/edgeai.

8 References

1. Yole Report: Computing and AI Technologies for Mobile and Consumer Applications 2021
2. How to Simplify your Embedded Edge AI Application Development

IMPORTANT NOTICE AND DISCLAIMER

TI PROVIDES TECHNICAL AND RELIABILITY DATA (INCLUDING DATA SHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES “AS IS” AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT OF THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for skilled developers designing with TI products. You are solely responsible for (1) selecting the appropriate TI products for your application, (2) designing, validating and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, regulatory or other requirements.

These resources are subject to change without notice. TI grants you permission to use these resources only for development of an application that uses the TI products described in the resource. Other reproduction and display of these resources is prohibited. No license is granted to any other TI intellectual property right or to any third party intellectual property right. TI disclaims responsibility for, and you will fully indemnify TI and its representatives against, any claims, damages, costs, losses, and liabilities arising out of your use of these resources.

TI’s products are provided subject to TI’s Terms of Sale or other applicable terms available either on ti.com or provided in conjunction with such TI products. TI’s provision of these resources does not expand or otherwise alter TI’s applicable warranties or warranty disclaimers for TI products.

TI objects to and rejects any additional or different terms you may have proposed.

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265