Qualcomm Linux AI/ML Guide

80-70018-15 AB

March 27, 2025

Contents

1 Overview

Qualcomm® Linux AI stack allows developers to optimally deploy pre-trained, deep learning models on Qualcomm hardware accelerators, such as Neural Processing Unit (NPU), Graphic Processing Unit (GPU), and Central Processing Unit (CPU). Qualcomm AI software offering contains software development kits (SDKs), APIs, sample applications, development tools, and third-party frameworks support such as GStreamer and TFLite, to ease application development.

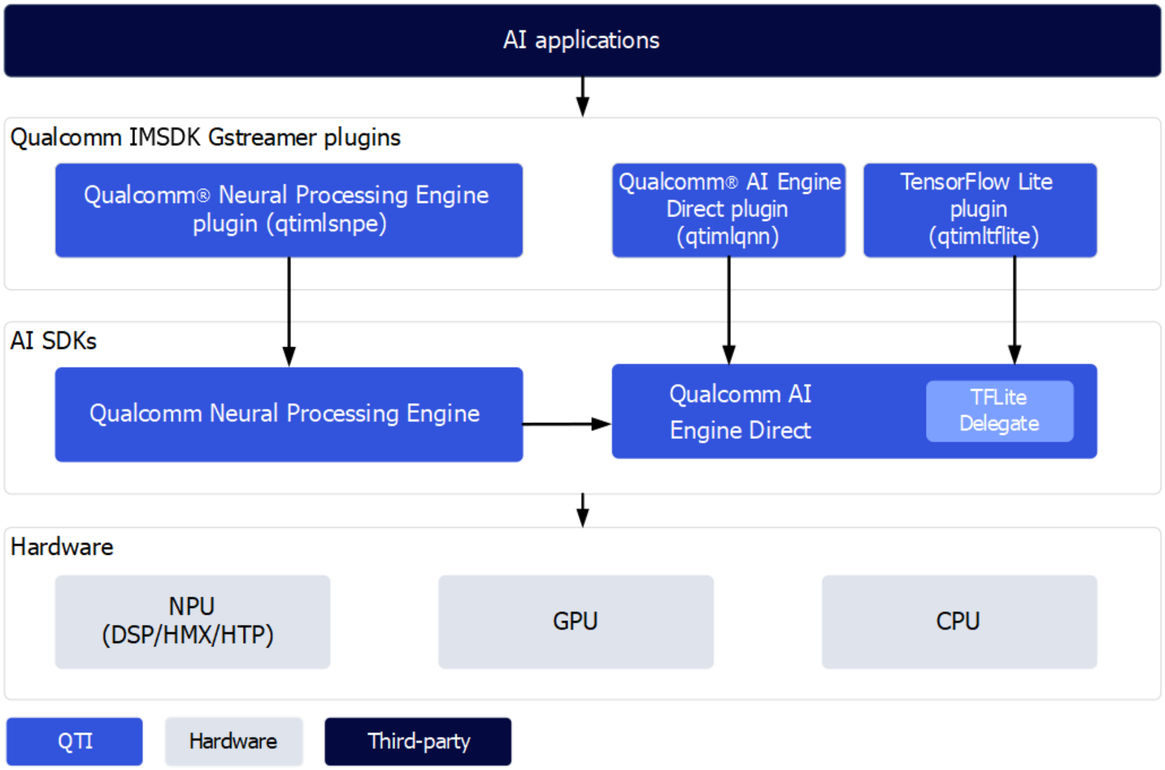

AI applications
Qualcomm IMSDK Gstreamer plugins
Qualcomm® Neural Processing Engine plugin (qtimlsnpe)
Qualcomm® AI Engine Direct plugin (qtimlqnn)
TensorFlow Lite plugin (qtimltflite)
AI SDKs
Qualcomm Neural Processing Engine
Qualcomm AI Engine Direct
TFLite Delegate
Hardware
NPU (DSP/HMX/HTP)
GPU
CPU
QTI
Hardware
Third-party

Figure1 Top-level AI hardware and software blocks of the Qualcomm Linux AI stack

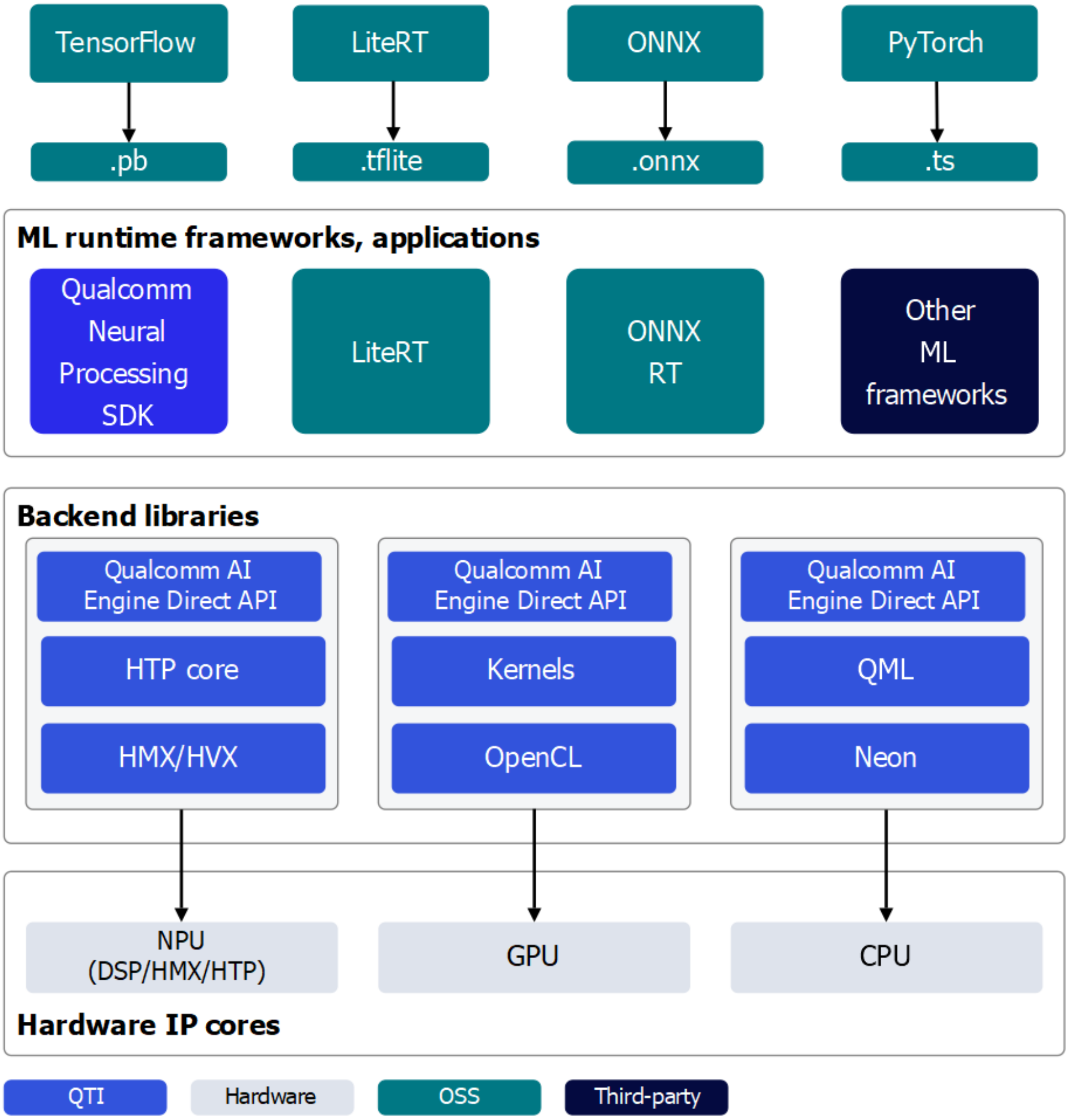
The key components of the Qualcomm Linux AI stack are:

- **AI applications** - Sample applications based on GStreamer that can be used or customized as needed.

- **GStreamer plugins** - Qualcomm Linux software offers GStreamer-based, machine learning plugins for accelerating AI inference using TFLite, Qualcomm® Neural Processing Engine SDK, etc., along with GStreamer plugins for pre- and postprocessing.
- **Qualcomm AI Stack** consists of two SDKs to accelerate AI workloads. The **Qualcomm Neural Processing Engine SDK** and **Qualcomm AI Engine Direct** provide tools, libraries, etc., to optimally accelerate AI models on multiple hardware accelerators.
- Qualcomm SoCs offer three **hardware cores** for AI loads.
  - **Neural Processing Unit (NPU)** - Also referred to as Qualcomm® Hexagon™ Tensor Processor (HTP) or DSP/HMX, is suitable for executing AI workloads with low-power and high-performance. To get optimized performance, pre-trained models need be quantized to one of the supported precisions.
  - **Graphics Processing Unit (GPU)** - Qualcomm® Adreno™ GPU is suitable for executing AI workloads with medium-power, and medium-performance. AI workloads are accelerated with OpenCL kernels. The GPU can also be used to accelerate model pre/post processing.
  - **Central Processing Unit (CPU)** - AI inferencing on the CPU can be used to benchmark model accuracy/performance against other hardware accelerators. The CPU can also be used to run model pre/post processing.

2 Architecture

The Qualcomm AI offering consists of hardware accelerators and AI SDKs to harness the power of hardware.



2.1 AI hardware accelerators

AI workloads can be accelerated on multiple hardware cores:

- Qualcomm® Hexagon™ Tensor Processor (HTP) - Also known as NPU/DSP/HMX, suitable to execute AI workloads with low-power and high-performance. For optimized performance, pre-trained models need be quantized to one of the supported precisions.
- Qualcomm® Adreno™ GPU - Suitable to execute AI workloads with medium-power, and

medium-performance. AI workloads are accelerated with OpenCL kernels. The GPU can also be used to accelerate model pre/post-processing.

- Qualcomm® Kryo™ CPU - AI inferencing on CPU can be used to benchmark model accuracy/performance against other hardware accelerators. The CPU can also be used to run model pre/post processing.

2.2 AI software stack

AI stack contains SDKs to harness the power of AI hardware accelerators. Developers can use the stack of their choice to deploy AI workloads. Pre-trained models (with the exception of TFLite models) need to be converted to an executable format with the selected AI Stack SDK before running them. Note that TFLite Delegate allows developers to directly run TFLite models.

- Qualcomm Neural Processing Engine (SNPE)

  An all-in-one SDK that provides C, C++, and Java APIs to support heterogenous computing, system-level configurations, and direct AI workloads to all accelerator cores. Provides developers with flexibility, including inter-core collaboration support and other advanced features.

- Qualcomm AI Engine Direct (QNN)

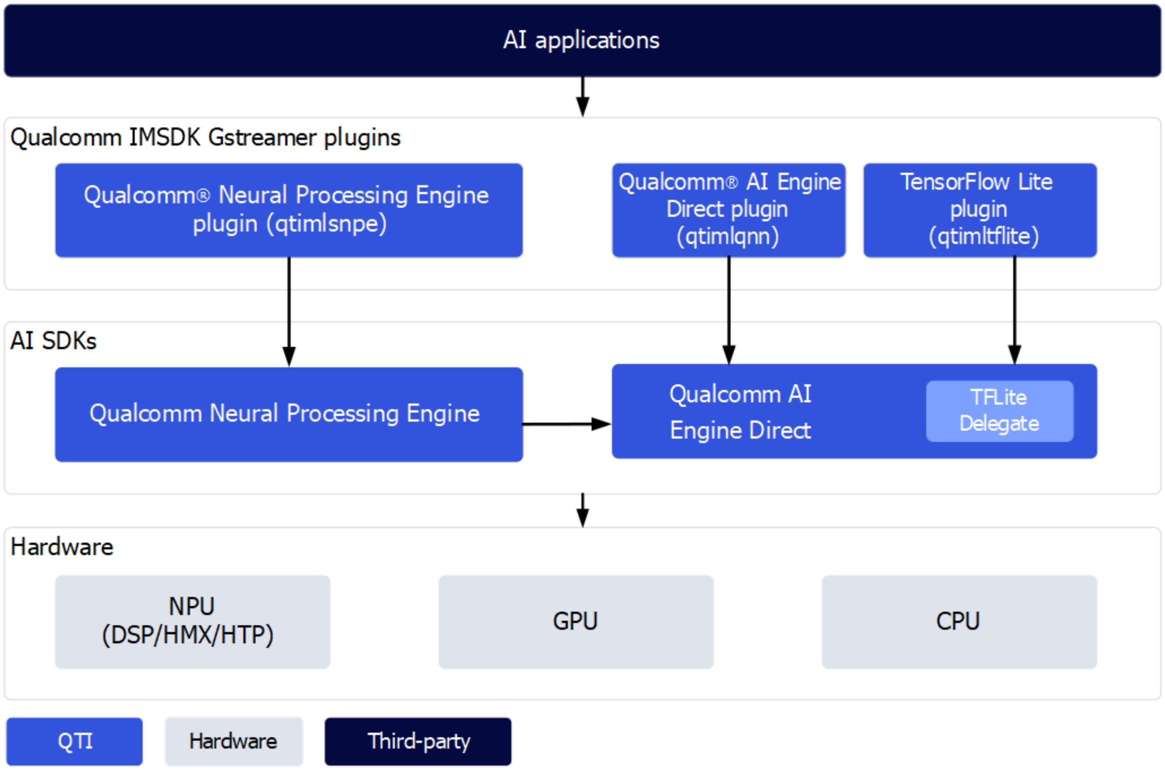
  Lower-level, highly customizable unified APIs that speed up AI models on all AI accelerator cores with individual libraries. Can be used directly to target a specific accelerator core or delegate workloads from popular runtimes including Qualcomm Neural Processing Engine SDK, TensorFlow Lite, and ONNX runtime. Low-level SDK provides more functionality and debugging abilities.

- AI Model Efficiency Toolkit (AIMET)

  Open-source library to optimize (compressing and quantizing) trained neural network models. This is a complex SDK designed to generate optimized quantized models. It's intended only for advanced developers.

3 APIs

Qualcomm Linux provides Qualcomm® Intelligent Multimedia SDK (IM SDK) GStreamer plugin APIs to interface with Qualcomm AI stack SDK APIs, to optimally run deep learning models on hardware modules such as NPU, GPU, CPU.


AI applications
Qualcomm IMSDK Gstreamer plugins
Qualcomm® Neural Processing Engine plugin (qtimlsnpe)
Qualcomm® AI Engine Direct plugin (qtimlqnn)
TensorFlow Lite plugin (qtimltflite)
AI SDKs
Qualcomm Neural Processing Engine
Qualcomm AI Engine Direct
TFLite Delegate
Hardware
NPU (DSP/HMX/HTP)
GPU
CPU
QTI
Hardware
Third-party

Three GStreamer ML plugin APIs are provided to support two AI stack SDKs, and external TFLite, which provides flexibility for developers to choose the right combination for their AI needs.

Qualcomm IM SDK plugin for Qualcomm Neural Processing Engine (qtimlsnpe)	Uses Qualcomm Neural Processing Engine APIs to load and execute models. Choose this plugin for quick prototyping and high-level API support.
Qualcomm IM SDK plugin for Qualcomm AI Engine Direct plugin (qtimlqnn)	Uses Qualcomm AI Engine Direct APIs, which provide low-level, unified API and improved performance to optimize and execute network models on the desired hardware accelerator. Choose this plugin for advanced graph execution options and optimizations.
Qualcomm IM SDK plugin for TensorFlow Lite (qtimltflite)	Accelerates TFLite models directly using Qualcomm AI Engine Direct APIs to load and execute models. Choose this plugin to directly run TFLite models, without the need of conversion.

SDK APIs - These SDKs provide AI APIs for application developers.

Qualcomm Neural Processing Engine SDK	C, C++, and Java APIs to support heterogenous computing, system-level configurations, and direct AI workloads to all accelerator cores. Provides developers with flexibility, including inter-core collaboration support and other advanced features.
Qualcomm AI Engine Direct	Low-level, highly customizable APIs that speed up AI models on all AI accelerator cores with individual libraries. This SDK can be used to target a specific accelerator core or delegate workloads from popular runtimes including Qualcomm Neural Processing Engine SDK, TensorFlow Lite, and ONNX runtime.

4 Sample Apps

Qualcomm Linux distribution provides sample AI/ML applications that demonstrate AI capabilities of the Qualcomm Linux platform.

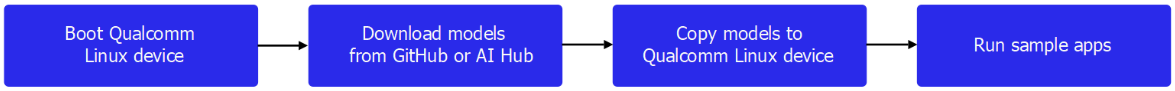

Boot Qualcomm Linux device
Download models from GitHub or AI Hub
Copy models to Qualcomm Linux device
Run sample apps

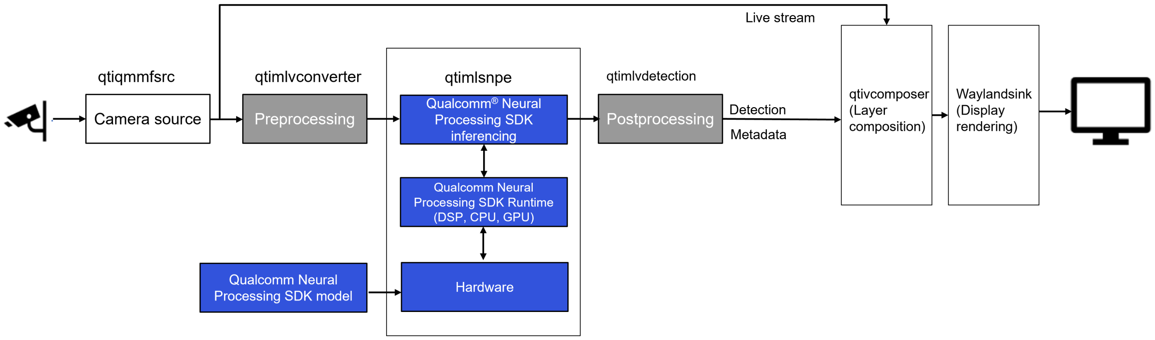
Figure1 Workflow for using the sample applications

The following AI sample applications are part of Qualcomm IM SDK.

Tip: For the complete list of supported sample applications, see Sample applications in the Qualcomm IM SDK documentation.

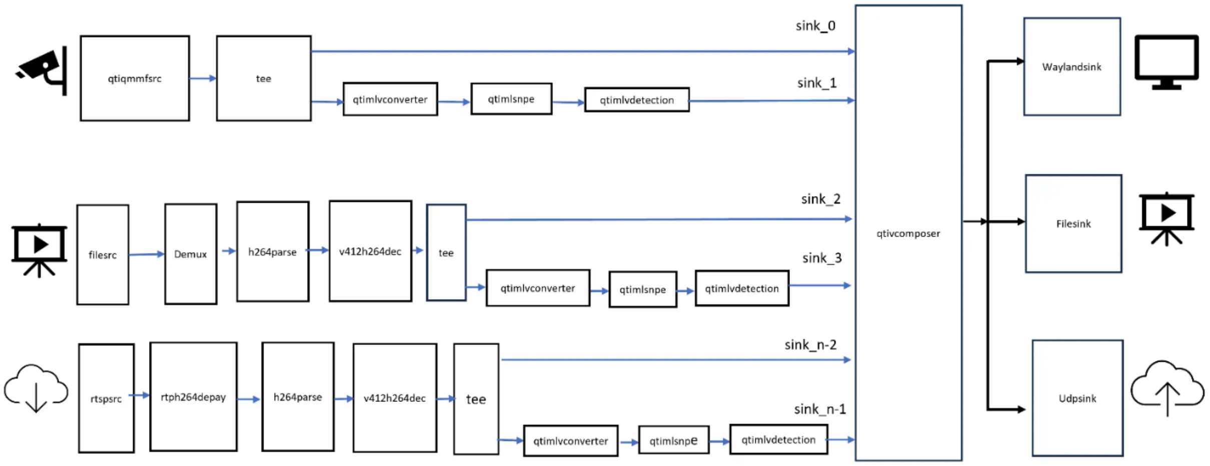
4.1 AI processing for object detection

The gst-ai-object-detection application demonstrates the capability of the hardware to perform object detection on input coming from the camera, file, or real-time streaming protocol (RTSP) source.


Live stream
qtiqmmfsrc
Camera source
qtimlvconverter
Preprocessing
qtimlsnpe
Qualcomm® Neural Processing SDK inferencing
Qualcomm Neural Processing SDK Runtime (DSP, CPU, GPU)
Hardware
Qualcomm Neural Processing SDK model
qtimlvdetection
Postprocessing
Detection Metadata
qtivcomposer (Layer composition)
Waylandsink (Display rendering)

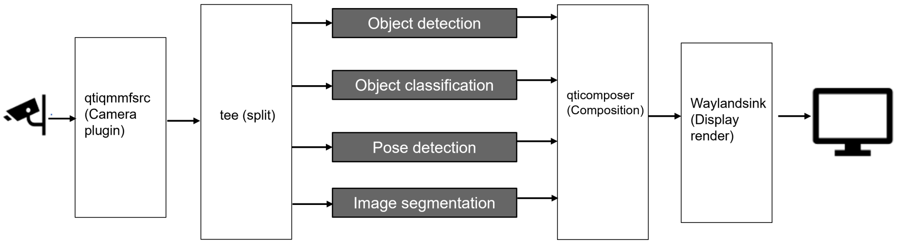
4.2 AI processing for multistream-inference

The gst-ai-multistream-inference application demonstrates the capability of the hardware to perform object detection or classification on multiple video streams from the camera, file, or RTSP source.



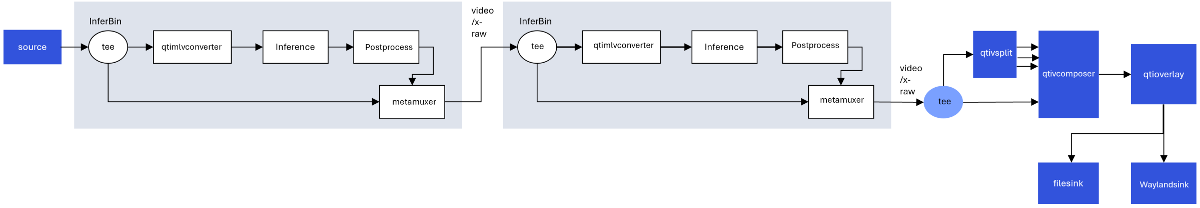
4.3 Parallel AI inference

The gst-ai-parallel-inference application demonstrates the capability of the hardware to perform four parallel AI inferencing functions on input coming from the camera, file, or RTSP stream source.



4.4 Daisychain of object detection and pose estimation

The gst-ai-daisychain-detection-pose application demonstrates the capability of the hardware to perform daisychained object detection and pose estimation on input coming from the camera, file, or RTSP source.


source
InferBin
tee
qtimlvconverter
Inference
Postprocess
metamuxer
video
/x-
raw
InferBin
tee
qtimlvconverter
Inference
Postprocess
metamuxer
video
/x-
raw
tee
qtivsplit
qtivcomposer
qtioverlay
filesink
Waylandsink

5 Developer workflow

The AI/ML developer workflow for Qualcomm Linux involves compiling and optimizing models and AI frameworks to run efficiently on Qualcomm hardware.

This process involves the following steps:

1. Selecting an AI model and framework.
2. Compiling and optimizing the model to run on Qualcomm hardware.
3. Building an application that integrates the optimized model.

See the complete AI Developer Workflow documentation

6 References

Title	Number
AI Hub	—
Qualcomm AI Model Efficiency Toolkit	—
Qualcomm Neural Processing Engine	80-63442-2
Qualcomm AI Engine Direct	80-63442-50
AI Engine Direct: TFLite Delegate	80-63442-50
Qualcomm Intelligent Multimedia SDK	80-70018-50

LEGAL INFORMATION

Your access to and use of this material, along with any documents, software, specifications, reference board files, drawings, diagnostics and other information contained herein (collectively this "Material"), is subject to your (including the corporation or other legal entity you represent, collectively "You" or "Your") acceptance of the terms and conditions ("Terms of Use") set forth below. If You do not agree to these Terms of Use, you may not use this Material and shall immediately destroy any copy thereof.

1) Legal Notice.

This Material is being made available to You solely for Your internal use with those products and service offerings of Qualcomm Technologies, Inc. ("**Qualcomm Technologies**"), its affiliates and/or licensors described in this Material, and shall not be used for any other purposes. If this Material is marked as "**Qualcomm Internal Use Only**", no license is granted to You herein, and You must immediately (a) destroy or return this Material to Qualcomm Technologies, and (b) report Your receipt of this Material to qualcomm.support@qti.qualcomm.com. This Material may not be altered, edited, or modified in any way without Qualcomm Technologies' prior written approval, nor may it be used for any machine learning or artificial intelligence development purpose which results, whether directly or indirectly, in the creation or development of an automated device, program, tool, algorithm, process, methodology, product and/or other output. Unauthorized use or disclosure of this Material or the information contained herein is strictly prohibited, and You agree to indemnify Qualcomm Technologies, its affiliates and licensors for any damages or losses suffered by Qualcomm Technologies, its affiliates and/or licensors for any such unauthorized uses or disclosures of this Material, in whole or part.

Qualcomm Technologies, its affiliates and/or licensors retain all rights and ownership in and to this Material. No license to any trademark, patent, copyright, mask work protection right or any other intellectual property right is either granted or implied by this Material or any information disclosed herein, including, but not limited to, any license to make, use, import or sell any product, service or technology offering embodying any of the information in this Material.

THIS MATERIAL IS BEING PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, WHETHER EXPRESSED, IMPLIED, STATUTORY OR OTHERWISE. TO THE MAXIMUM EXTENT PERMITTED BY LAW, QUALCOMM TECHNOLOGIES, ITS AFFILIATES AND/OR LICENSORS SPECIFICALLY DISCLAIM ALL WARRANTIES OF TITLE, MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, SATISFACTORY QUALITY, COMPLETENESS OR ACCURACY, AND ALL WARRANTIES ARISING OUT OF TRADE USAGE OR OUT OF A COURSE OF DEALING OR COURSE OF PERFORMANCE. MOREOVER, NEITHER QUALCOMM TECHNOLOGIES, NOR ANY OF ITS AFFILIATES AND/OR LICENSORS, SHALL BE LIABLE TO YOU OR ANY THIRD PARTY FOR ANY EXPENSES, LOSSES, USE, OR ACTIONS HOWSOEVER INCURRED OR UNDERTAKEN BY YOU IN RELIANCE ON THIS MATERIAL.

Certain product kits, tools and other items referenced in this Material may require You to accept additional terms and conditions before accessing or using those items.

Technical data specified in this Material may be subject to U.S. and other applicable export control laws. Transmission contrary to U.S. and any other applicable law is strictly prohibited.

Nothing in this Material is an offer to sell any of the components or devices referenced herein.

This Material is subject to change without further notification.

In the event of a conflict between these Terms of Use and the *Website Terms of Use* on www.qualcomm.com, the *Qualcomm Privacy Policy* referenced on www.qualcomm.com, or other legal statements or notices found on prior pages of the Material, these Terms of Use will control. In the event of a conflict between these Terms of Use and any other agreement (written or click-through, including, without limitation any non-disclosure agreement) executed by You and Qualcomm Technologies or a Qualcomm Technologies affiliate and/or licensor with respect to Your access to and use of this Material, the other agreement will control.

These Terms of Use shall be governed by and construed and enforced in accordance with the laws of the State of California, excluding the U.N. Convention on International Sale of Goods, without regard to conflict of laws principles. Any dispute, claim or controversy arising out of or relating to these Terms of Use, or the breach or validity hereof, shall be adjudicated only by a court of competent jurisdiction in the county of San Diego, State of California, and You hereby consent to the personal jurisdiction of such courts for that purpose.

2) Trademark and Product Attribution Statements.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Arm is a registered trademark of Arm Limited (or its subsidiaries) in the U.S. and/or elsewhere. The Bluetooth® word mark is a registered trademark owned by Bluetooth SIG, Inc. Other product and brand names referenced in this Material may be trademarks or registered trademarks of their respective owners.

Snapdragon and Qualcomm branded products referenced in this Material are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.