AI at the Device Edge

Definitions, Insights, and Tools to bring Artificial Intelligence to Edge Devices

Qualcomm
developer network

Towards AI at the Device Edge

Data is at the heart of modern business, providing real-time insight and control over day-to-day operations. This is facilitated by the explosion of the Internet of Things (IoT), with billions of devices collecting zettabytes of data.* As IoT grows, so does the amount of data to be processed. This means sending data to centralized cloud servers for processing, along with concerns over latency, bandwidth, and security, is driving the demand for processing and intelligence close to or even at the very point where the data is collected—at the edge.

From smart cities, smart factories, wearables, and in-home automation, combined with the convergence of machine learning (ML) algorithms, edge devices with onboard AI processors, advances in sensor design, and next-generation communication technologies like 5G, are spawning a shift towards *AI at the device edge*.

In this eBook, we discuss what it means to process AI at the device edge, and highlight tools to get you started.

* https://seedscientific.com/how-much-data-is-created-every-day/

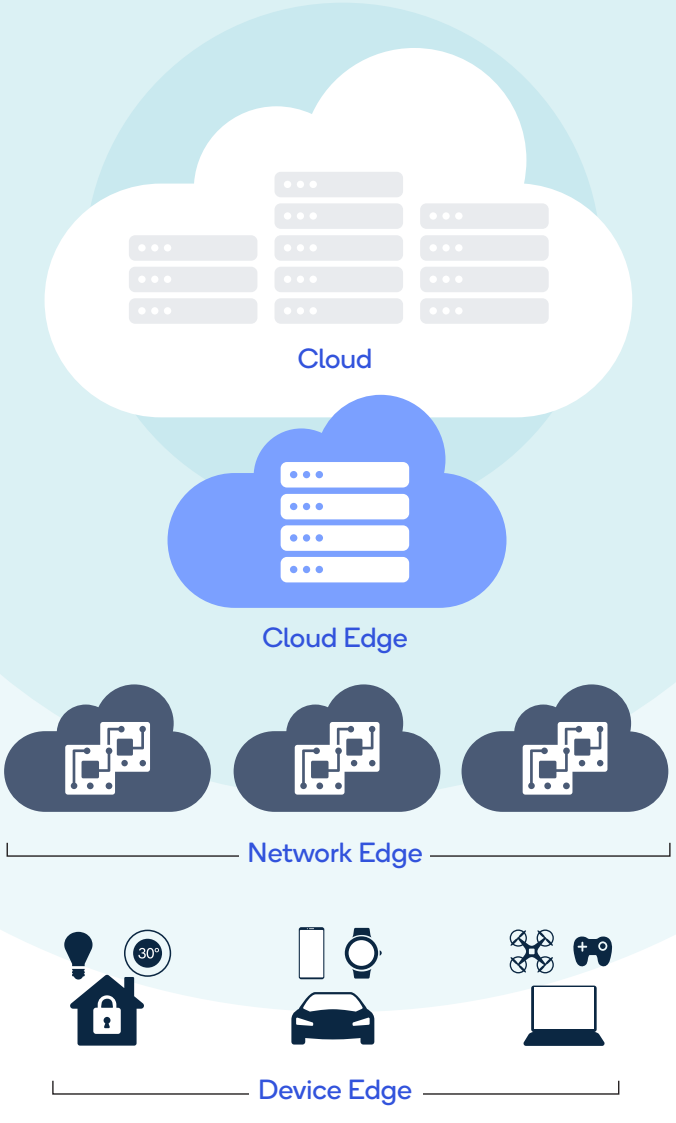
The Ever-Evolving Edge

In **edge computing**, data processing and storage are performed closer to the source of the data. Data is then transmitted back to the cloud, and the cloud may, in turn, return information or results back to the edge. Various *edges* have been devised to support this, such as:

The cloud edge architecture where network providers distribute cloud server resources closer to customers. This is leading to new business models, partnerships, and acquisitions between hyperscalers, telcos, start-ups, and other industry players.

Through network edge architectures, communication technologies like 5G private networks are now pushing server compute even closer to the customer. Devices, like on-premise servers that connect to the cloud, process data collected by local devices (sometimes called *first-level analytics*). This is prevalent in Industrial IoT and Industry 4.0 (e.g., smart factories), where devices monitor manufacturing processes.

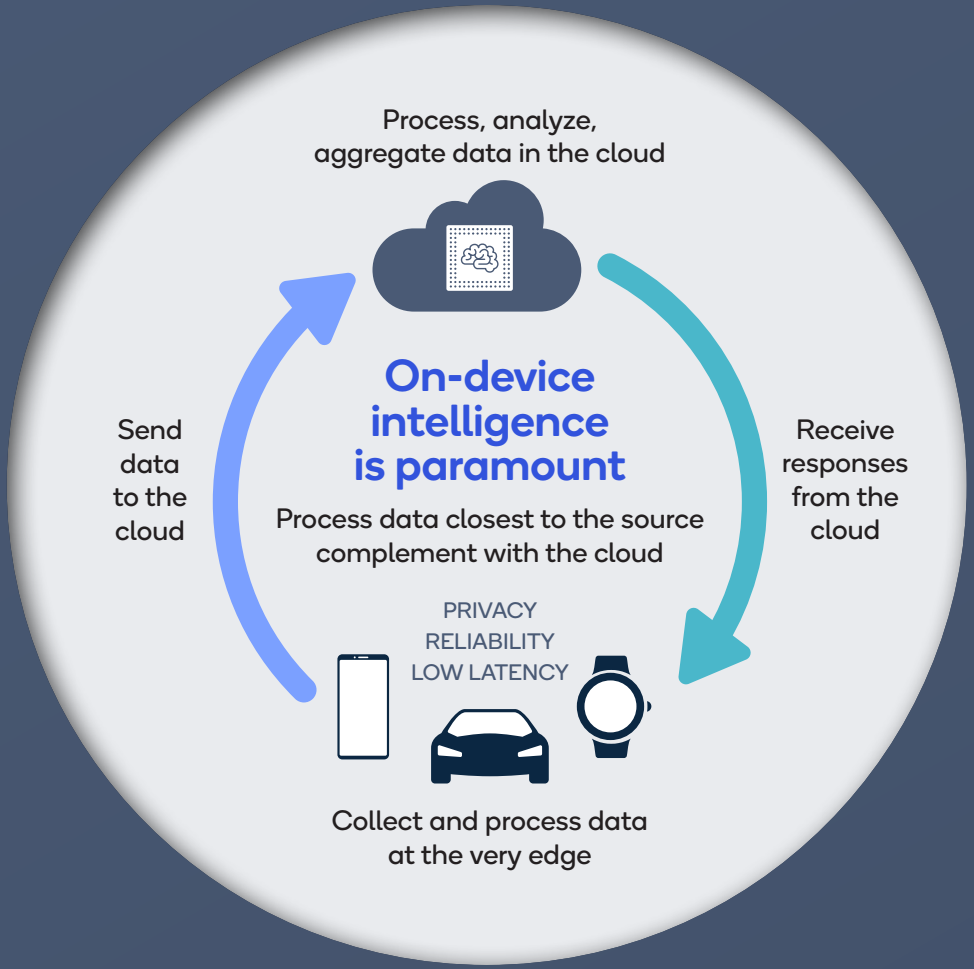
In the device edge architecture, data processing, storage, and intelligence sit right at the point where sensors collect the data. This can be achieved through IoT devices that physically or wirelessly connect to a sensor (or *fuse* multiple sensors) or via *smart sensors*, which incorporate *base sensors* and SoCs into the same package. Here, various ML models can make decisions about filtering or combining the data before sending it to the cloud for higher-level processing and analytics.



Edge and Cloud AI Complement Each Other

Edge AI is often used for *instant* data processing to make real-time decisions. For example, a self-driving car could use its telemetry and camera data to navigate by itself. It may then send processed data to the cloud, such as a snapshot of its current GPS position, vehicle operating state, etc.

In the cloud, powerful servers often perform a much broader analysis of the data. These servers can send responses back to the edge in near real-time over low-latency links like 5G (e.g., to notify the vehicle of an engine problem). The data on the server can also be aggregated to detect broader trends (e.g., traffic patterns).



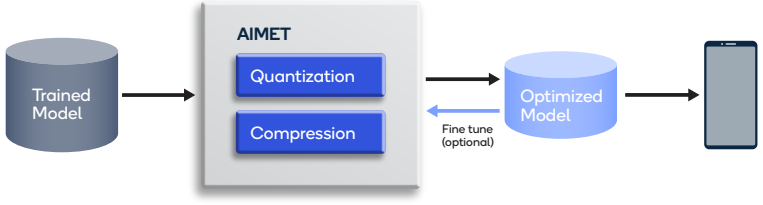
Optimizing Trained Models

ML models like **Convolution Neural Networks** (CNNs), **Recurrent Neural Networks** (RNNs), and **Generative Adversarial Networks** (GANs), are typically built and trained using ML frameworks. For example, TensorFlow and TensorFlow Lite (for mobile, embedded and IoT devices) provide a rich API for defining operations on tensors. These trained models are then exported for inference.

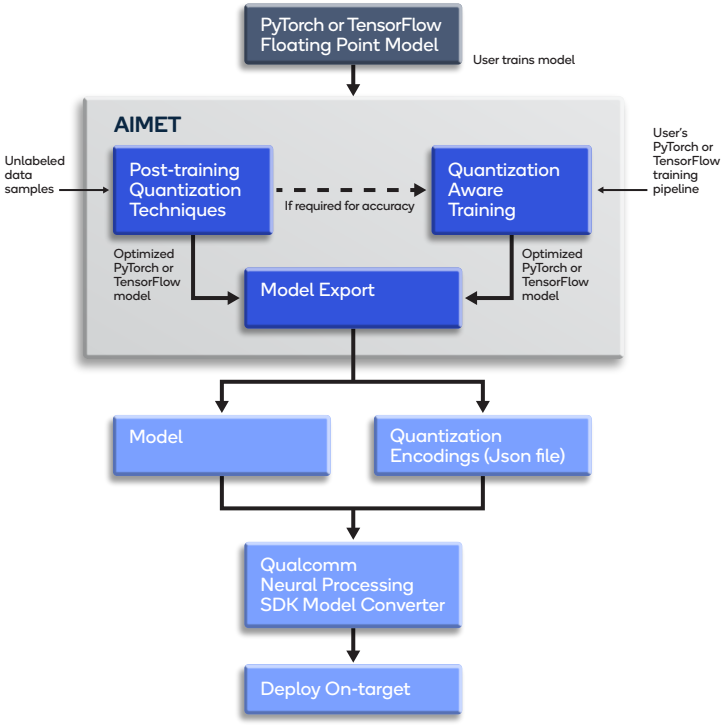
Edge devices typically have fixed-sized storage, limited memory, and power limitations due to small size of their battery. A key challenge for developers is to optimize the model for the edge device, while minimizing the loss of the model's predictive performance (e.g., accuracy).

Two techniques employed by developers to prepare and optimize their trained models for the edge are:

1. **Quantization**: reduces the bit size of the model's parameters (e.g., 32-bit floating point to 8-bit integers). This can reduce the model's file size and the required processing power and power consumption for inference.
2. **Compression**: removes redundant parameters or computations with little or no influence on predictions.



Through our Qualcomm Innovation Center (QuIC), we have done extensive research on model optimization and recently released our AI Model Efficiency Toolkit (AIMET). Using the AIMET library, developers can incorporate its advanced model compression and quantization algorithms into their PyTorch and TensorFlow model-building pipelines for automated post-training optimization, as well as for model fine-tuning if required.



AIMET is a product of Qualcomm Innovation Center, Inc.

Key Advantages and Considerations of AI at the Device Edge

Edge deployments have traditionally focused on the logistics of when to move data. However, the next wave of edge deployment and data processing planning will likely focus on how to derive and process the business value from all the IoT data. This will be based on the need to aggregate data, business dependencies, and the required value. AI at the Device Edge opens up possibilities and helps increase flexibility by providing several advantages.

Real-time Responsiveness and Performance

An increasing number of use cases now require low-latency processing, big data, and the adoption of AI, IoT, and 5G. Moving compute to the very edge eases the stress on bandwidth and speeds up processing and responsiveness, allowing more bandwidth-heavy technologies. This could potentially include additional use cases in AR and VR.

Form Factors and Power Constraints

Smart devices may be deployed in remote and even dangerous locations. This drives the need for devices to be compact, lightweight, and able to withstand harsh environments. Such devices may also need to be self-sufficient and able to recharge their battery or harvest energy and communicate wirelessly in the absence of a wired infrastructure. Self-healing functionality controlled by intelligence at the edge is also important, so that a device can switch to backup sensors or adjust for drift.

Reduced Cost

Larger deployments drive the need for reduced costs, both in terms of bandwidth usage and the unit cost per device. Processing at the edge can potentially reduce the amount of data sent out and received back from the cloud, reducing network data costs. At the same time, smart sensors capable of running inference, can be more cost effective than purchasing and integrating separate components.

Security

AI at the device edge can reduce the amount of sensitive data that is sent to the cloud. Using its onboard intelligence, the device can use the sensitive data collected to build and transmit information that is less sensitive and more anonymous.

Use Cases

The following examples illustrate how AI at the device edge could be used across different verticals:

Predictive maintenance in factories utilizes sensors on machines to measure properties like vibration, noise levels, etc. Onboard ML models then use this data to infer potential failures, maintenance requirements, etc.

Expression analysis via facial recognition can infer a person's mood, state, etc. There is a range of uses, from providing alerts in vehicles when the driver is not paying attention to analyzing responses to shopping ads or floor layouts.

Body monitoring through wearables can measure a variety of vital signs. This data can then be transmitted to the cloud to detect stress, healthy eating, etc., and responses returned to alert the wearer about their state or potential actions to take. The information can also be used by medical practitioners to monitor ailments or by insurance companies to encourage and reward healthy living.

Autonomous vehicles equipped with a variety of sensors and cameras can collect, process, and share data with other vehicles or infrastructure over broader networks (e.g., network edge). Data from the network edge can then be used to profile traffic patterns and other aspects to help workers plan for smarter cities.

Precision agriculture aims to grow crops more efficiently. By using various devices (i.e., cameras and sensors mounted out in the field and on farm equipment or other devices like drones) to gather data, insights are provided to farmers to help them make informed decisions about their crops.

Snapdragon at the Device Edge

Advances in mobile processors like **Snapdragon® mobile platforms** are allowing for on-device AI and compute at the Device Edge. Powering Snapdragon are the Qualcomm® Adreno™ GPU, Qualcomm® Kryo™ CPU, and Qualcomm® Hexagon™ DSP. Together with accompanying SDKs and tools for Snapdragon, these components are collectively known as the **Qualcomm® Artificial Intelligence (AI) Engine**. The following lists some key platforms which support this functionality:

Snapdragon mobile platform

Powers a range of mobile devices, most notably many of today's smartphones.

Qualcomm® Vision Intelligence Platform

Incorporates image processing and AI into smart-camera products for IoT.

Qualcomm® Robotics RB5

Supports the development of robots which incorporate on-device AI.

Qualcomm Flight™ Pro

Supports the development of drone and aerial robotic technology with on-device AI.

Snapdragon processors for XR and VR

Platforms like the Snapdragon XR2 5G Platform allow developers to incorporate AI into XR and VR experiences.

Snapdragon Wear™ 4100 platform

Brings the power of on-device AI to wearable technology like smart watches.

Develop AI at the Device Edge

Qualcomm Technologies, Inc. provides a comprehensive software stack for the Qualcomm AI Engine on Snapdragon:

Neural Processing Engine SDK for Artificial Intelligence (AI)

Tools to optimize ML models for Snapdragon mobile platforms.

AI Model Efficiency Toolkit (AIMET)

Open-sourced by the Qualcomm Innovation Center (QuiC), the AIMET library on GitHub provides a collection of advanced model compression and quantization techniques for trained neural network models.

Qualcomm® Computer Vision SDK

Library of mobile-optimized computer vision (CV) algorithms.

Qualcomm® Machine Vision SDK

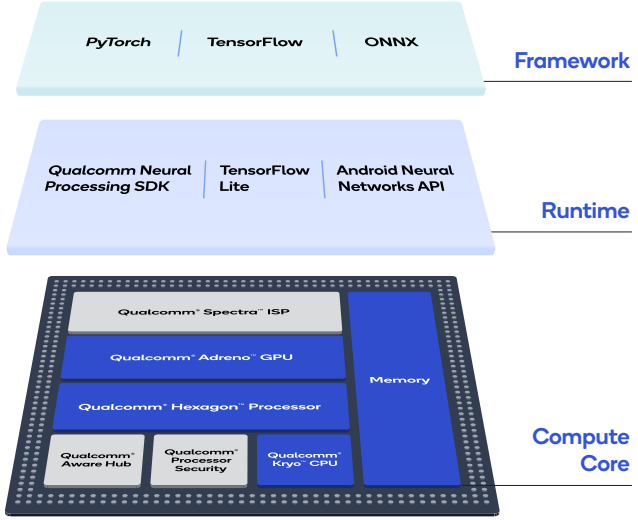
Optimized library of machine vision processing algorithms for robotics.

Snapdragon Profiler

Software to profile hardware resources on Snapdragon-based devices.

Qualcomm AI Software Stack

Covering every layer from applications to core



Innovate together

Qualcomm Developer Network is a collection of software and hardware tools, inspiring our community of developers to push the boundaries of mobile. We're continuously creating some of the most innovative, powerful and disruptive technologies in the world, and Qualcomm Developer Network is the gateway through which you can discover the tools you need, whether you're building high-performance apps, smart Internet of Things (IoT) devices, immersive virtual reality experiences or for other emerging technologies.

developer.qualcomm.com

Qualcomm
developer network