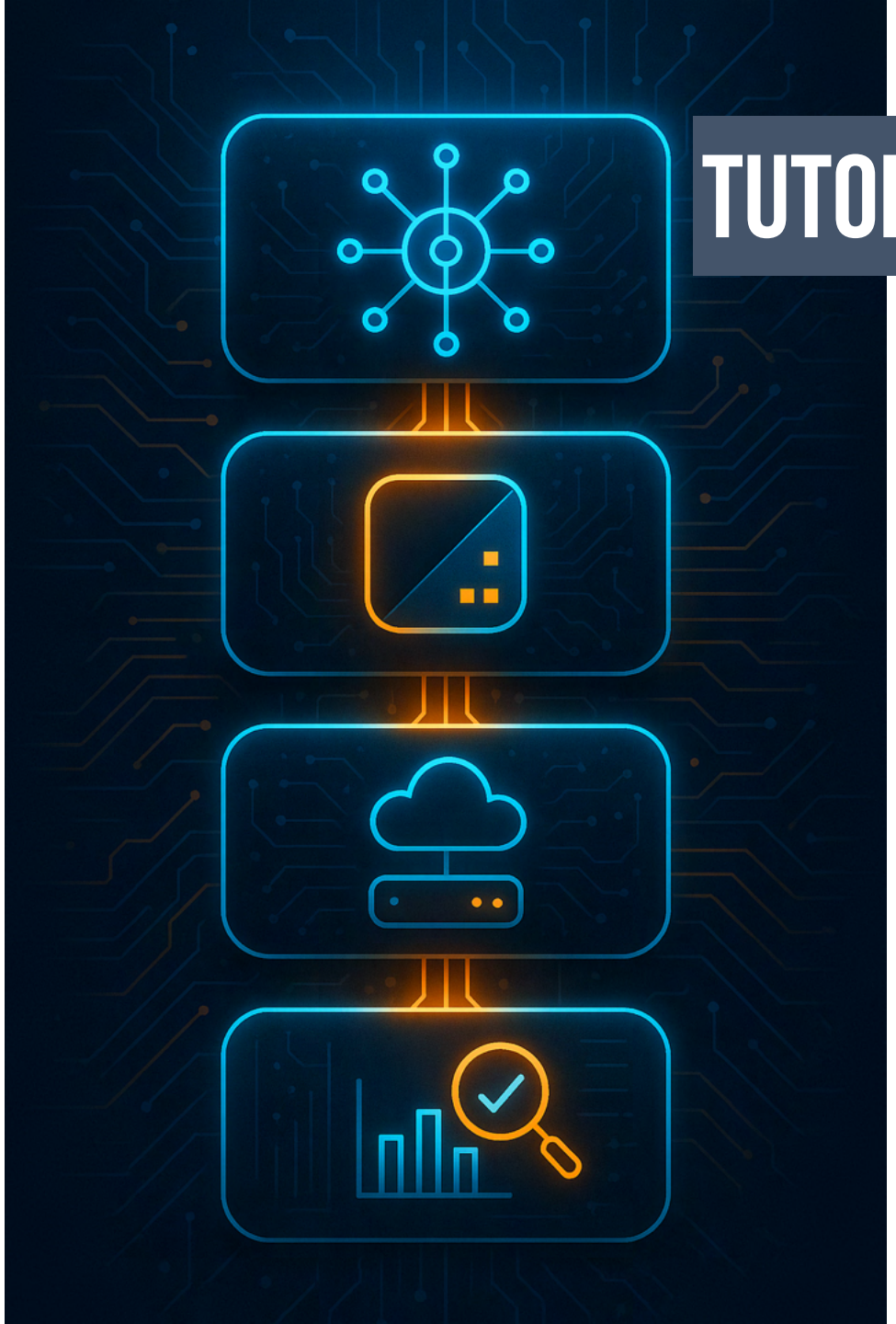
EDGE AI IN ACTION: TECHNOLOGIES AND APPLICATIONS

CVPR 2025 Tutorial

The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025

Nashville, TN, USA



TUTORIAL AGENDA

1 Introduction to Edge AI

2 Model Deployment for Edge AI

3 Edge AI with Qualcomm AI Hub

4 VLM Deployment on Edge AI

5 Closing Remarks and Joint Q&A

WHO ARE WE?

Organizers and Speakers

FABRICIO BATISTA NARCIZO
Senior AI Research Scientist

fbnarcizo@gn.com
Jabra / ITU

ELIZABETE MUNZLINGER
Industrial Ph.D. Candidate

munzlinger@itu.dk
Jabra / ITU

SAI NARSI REDDY DONTHI REDDY
Senior AI/ML Researcher

sdreddy@gn.com
Jabra

SHAN AHMED SHAFFI
AI and ML Researcher

shaahmed@gn.com
Jabra

HTTP://FABRICIONARCIZO.GITHUB.IO/CVPR2025-EDGE-AI-IN-ACTION

LEARNING OUTCOMES FOR THE TUTORIAL

Step 01

UNDERSTAND THE FUNDAMENTALS OF EDGE AI

Gain a solid understanding of Edge AI, including its motivations, benefits, and key challenges.

Step 02

LEARN MODEL DEPLOYMENT STRATEGIES

Learn the processes involved in converting and optimizing AI models for deployment.

Step 03

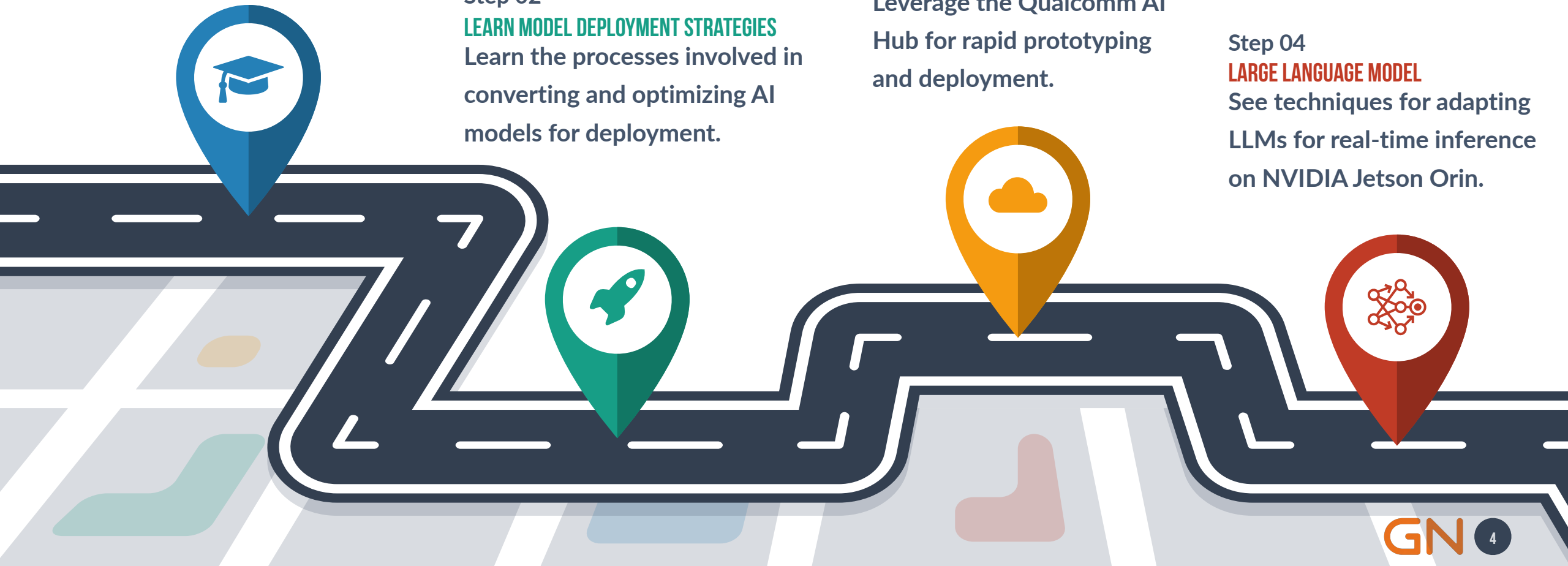
CLOUD ALTERNATIVES FOR OPTIMIZATION

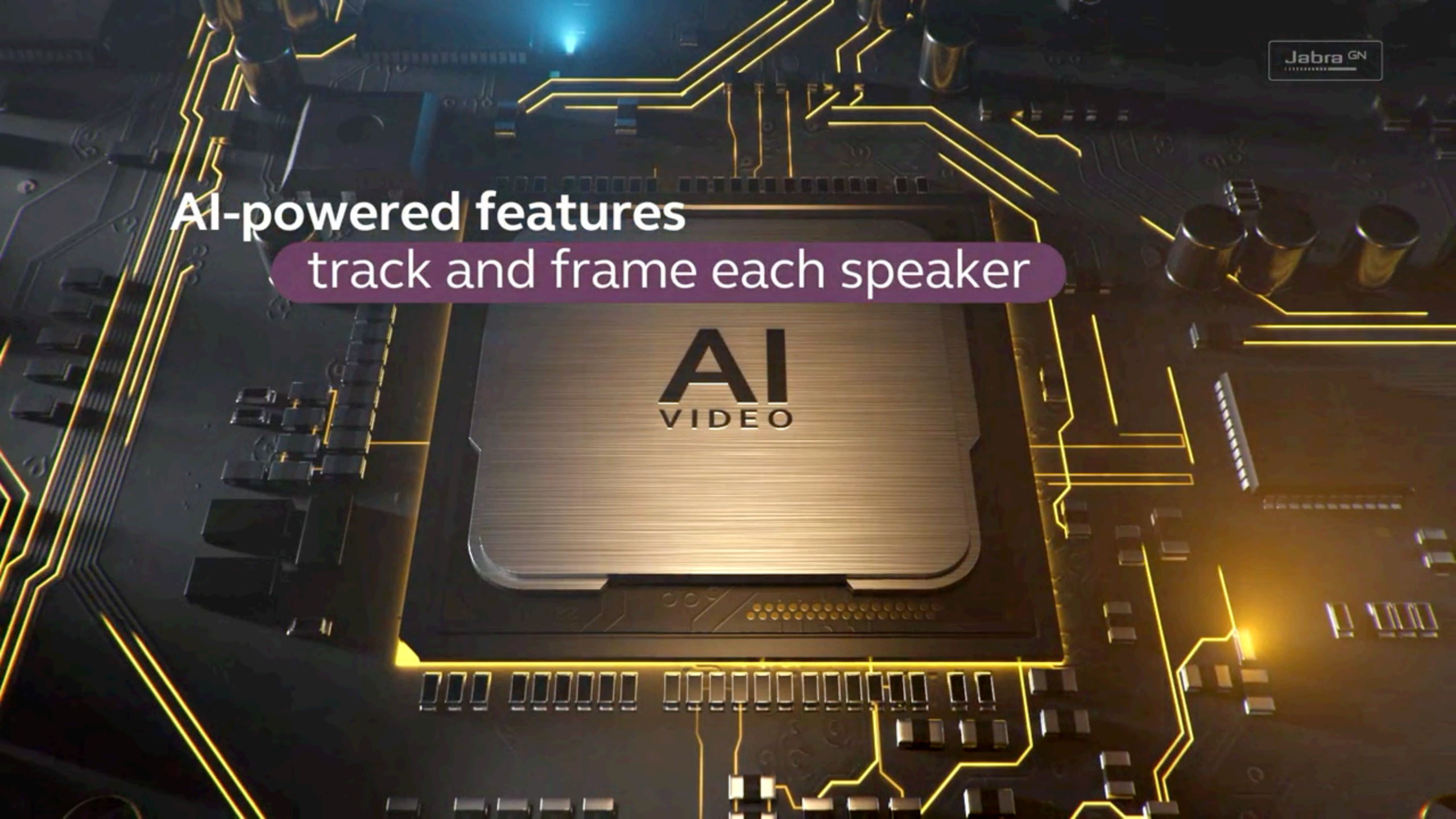
Leverage the Qualcomm AI Hub for rapid prototyping and deployment.

Step 04

LARGE LANGUAGE MODEL

See techniques for adapting LLMs for real-time inference on NVIDIA Jetson Orin.




Jabra GN
AI-powered features
track and frame each speaker
AI
VIDEO

DEMO SESSION

Edge AI in Action: Deploying Multimodal Models in Edge AI Devices

GN Advanced Science

EDGE AI
IN ACTION
DEPLOYING MULTIMODAL MODELS IN EDGE AI DEVICES

Sat, 14 June 2025
CVPR '25, NASHVILLE, TN, USA

OVERVIEW

Edge AI refers to the deployment of artificial intelligence models directly on edge devices—such as smartphones, cameras, sensors, drones, and wearables—enabling them to perform inference locally without continuous reliance on the cloud. This paradigm offers several significant benefits, including reduced latency, enhanced privacy, increased responsiveness, and improved energy efficiency.

However, building performant AI systems for edge environments introduces unique challenges. These include model compression, quantization, distillation, pruning, and runtime optimization, as well as effective orchestration across heterogeneous hardware in hybrid edge–cloud architectures.

MULTIMODAL

In this CVPR 2025 demo, we show a hands-on and practice-oriented guide to designing, optimizing, and deploying deep learning models for edge AI. Focusing on computer vision tasks, we will explore real-world use cases, including hand gesture recognition, object detection, and large language models. We emphasize multimodal AI, integrating inputs such as video, images, and text to enable intelligent, interactive systems.

We will demonstrate the use of leading tools and frameworks, including ONNX, Qualcomm SNPE, TensorFlow Lite for Android, vLLM, and Ollama, across a range of hardware platforms, such as Jabra intelligent cameras, Qualcomm dev boards, Android mobile phones, and NVIDIA Jetson AGX Orin.

CVPR Nashville

Head Body Detector

This model, based on YOLO5, detects the heads and bodies of participants in corporate meetings to support intelligent framing and analytics. It is optimized for real-time inference directly on Jabra PanaCast P40 cameras.

We use a customized YOLO5 model to detect participant heads and bodies in meeting rooms. The model was quantized and optimized to run efficiently on-device with Qualcomm QCS6490, enabling innovative video processing without the need for external computing.

Background Segmentation

Designed for privacy-aware video conferencing, this custom model separates the user from the environment and applies background blur without external processing. It runs directly on the Jabra PanaCast P20 using the Intel Movidius Myriad X VPU for on-device inference.

We blur the meeting background by segmenting the user's body in real-time. The model is deployed directly on the Jabra PanaCast P20, ensuring privacy and low-latency performance at the edge.

SNPE Optimizer Platform

Our SNPE Optimizer Platform streamlines model conversion, quantization, and deployment specifically for Qualcomm's QCS6490 edge chipset. It enables developers to generate performance-tuned DLC binaries for efficient on-device inference.

We have built an open-source toolchain that automates the conversion of models into Qualcomm SNPE containers. This pipeline handles quantization and runtime tuning to fully leverage Hexagon DSP, CPU, and GPU resources for edge devices.

InferSNPE Android Application

An Android app that performs on-device inference using quantized neural networks via Qualcomm SNPE, targeted at Snapdragon-based phones. It leverages the SNPE SDK to execute DLC models on DSP, GPU, or CPU, depending on hardware availability, ensuring accelerated and efficient edge performance.

This app captures camera frames, pre-processes them (e.g., resize, normalize), and runs inference on quantized models converted to DLC format in real-time.

YOLO-hagRID Model

Our model extends YOLO11 to detect and classify 34 static hand gestures using the HaGRIDv2 dataset, enabling rich gesture interaction. It operates in real-time and is robust to diverse environments, thanks to training on millions of hand gesture image samples.

The resulting model runs at real-time speeds on Snapdragon-based Android phones and performs reliably across varied lighting and backgrounds, making it ideal for interactive edge applications.

SATURDAY, JUNE 14 (10:30-12:30)

THANK YOU!