Linaro Connect
MADRID 2024 | MAY 14-17 2024

Accelerate your AI workloads on Windows on Snapdragon

Devang Aggarwal
Senior Product Manager, Qualcomm

Generative AI is Here

180M
Monthly active ChatGPT users

55%
of adults interact with AI about once a day

61%
of professionals using or plan to use AI

Inference

Users now expect generative AI capabilities for everyday computing, predicating their patterns and providing custom responses & solutions.

Source: Number of ChatGPT Users (Feb 2024) (explodingtopics.com)
Source: Generative AI Statistics for 2024 – Salesforce
Source: What Americans Know About Everyday Uses of AI | Pew Research Center

Leader in AI Performance

On-device AI

AI Leadership with Snapdragon X Elite

Most Capable

AIPC with 45 TOPS NPU (+ CPU, GPU) available for AI

- Leading platform for “edge AI”
- Designed for AI performance
- Delivers unparalleled Efficiency
- Supports AIPC workload demands
- Supports Gen-AI large models
- Heterogenous AI - NPU, CPU, GPU

Up to 15X

Real world AI performance versus competition

- Benchmarked against competition for popular video and photo editing applications

60+

ISVs currently engaged for Snapdragon AI

- Includes solutions and apps
- Growing number

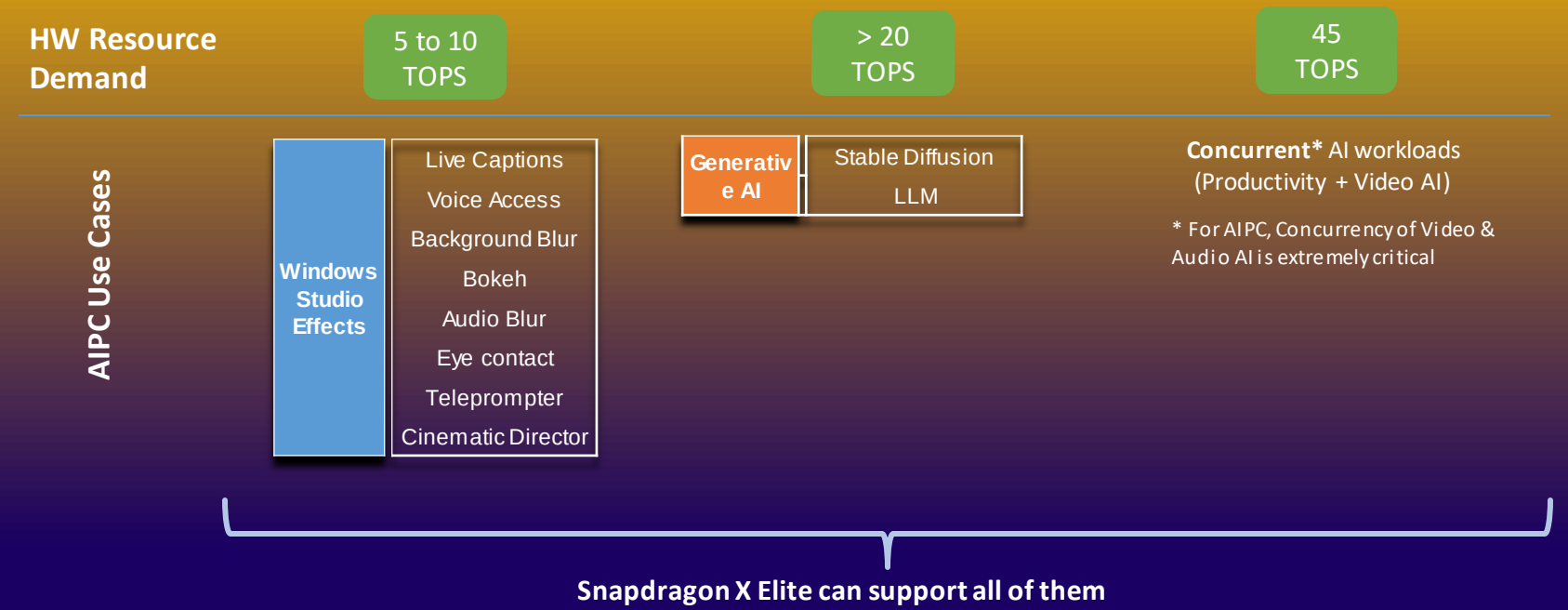
Most Resourceful

Toolset available for AI developers

- Comprehensive AI frameworks, SDK, samples & model support
- Tools for complete workflow

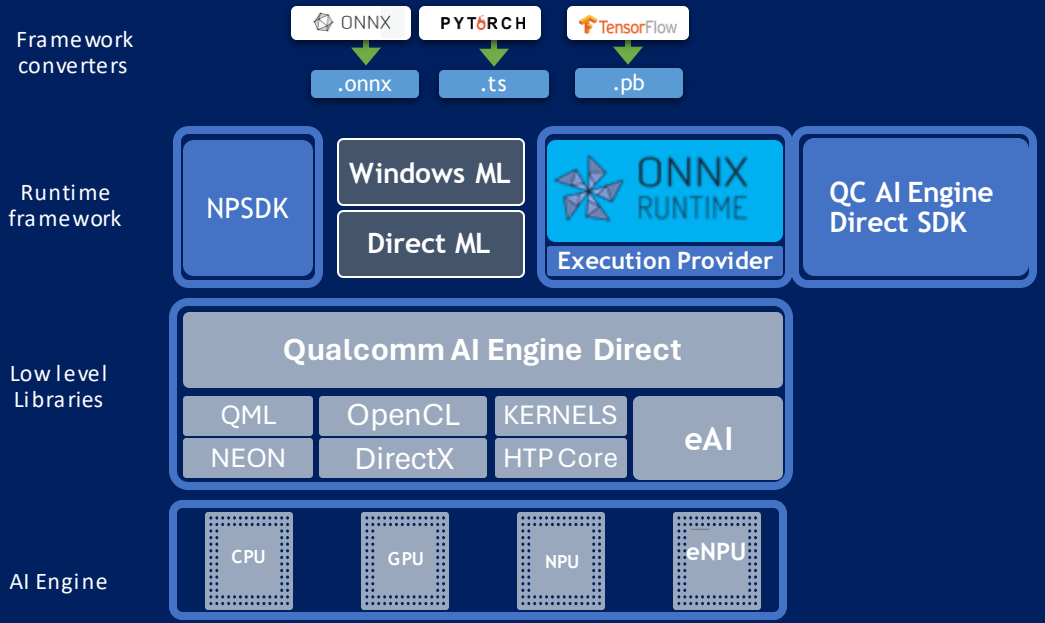
Snapdragon X Elite supports all AIPC workload demands

45 TOPS NPU is critical to support all Windows AI use cases with concurrency.



Resourceful Tools for AI Developers

Flexible Programmability



Powerful Optimization Workflow (AiMET)

Model Compression
- Spatial SVD
- Chanel Pruning
- Compression ratios

Visualization
- Weight Ranges
- Compression Sensitivity

Model Export

Model Quantization
- Cross Layer Equalization
- Bias Correction
- Quantization Simulation
- Fine Tuning
- Productization Quality Tests
- Quantization Aware Training
- Auto Mixed Precision

OSS Models

LLMs: Llama2, QWEN
LVMs: Stable Diffusion, Controlnet

Samples, Notebooks

AI Hub (QC, Github, Hugging Face)

Available Solutions/Use Cases

Image Classification
Object Detection
Pose Estimation
Super Resolution
Semantic Segmentation
Video Understanding
Speech Recognition
Natural Language Processing
Large Language Models
Text to Image

Live Llama2 Jupyter Notebook Walkthrough

3x Faster

Image Generation – Stable Diffusion
(powered by NPU)

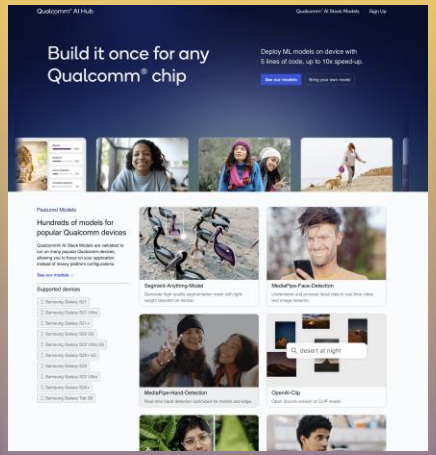
GIMP

Snapdragon X Elite — 7 seconds

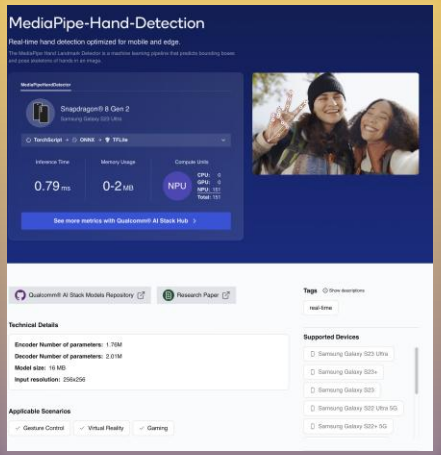
Intel Core Ultra 7 — 22 seconds

QUALCOMM AI HUB

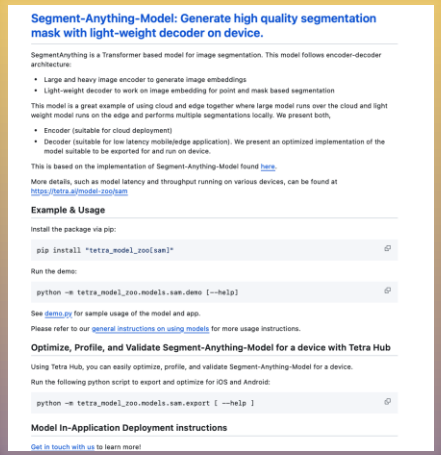
Don’t know where to start? We have the largest Collection of Reproducible Models for Qualcomm Edge Deployment



Search Models



See real results on Device



Reproduce on real devices in cloud


LinaroConnect
MADRID 2024 | MAY 12-17 2024
Thank you