Deploying Large Models on the Edge: Success Stories & Challenges

Dr. Vinesh Sukumar

Sr Director, Product Management

Qualcomm Technologies Inc.

Qualcomm

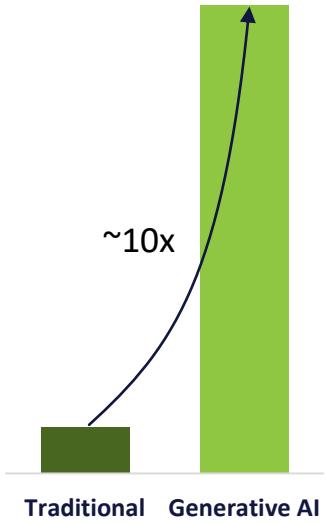
GEN AI is NOT scalable with cloud ONLY

Cloud economics will NOT allow Generative AI to scale

Edge GEN AI is becoming MORE than relevant NOW!

Cost Per Query[1]

(e.g. web search)



×

GEN AI Applications

Web search

Office copilot

Image & video creation

Coding assistant

Text summarization

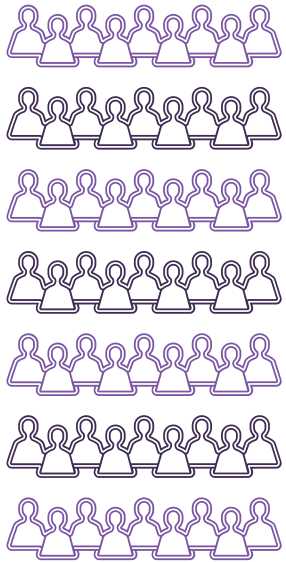
Conversational chatbots

Copy creation

...

×

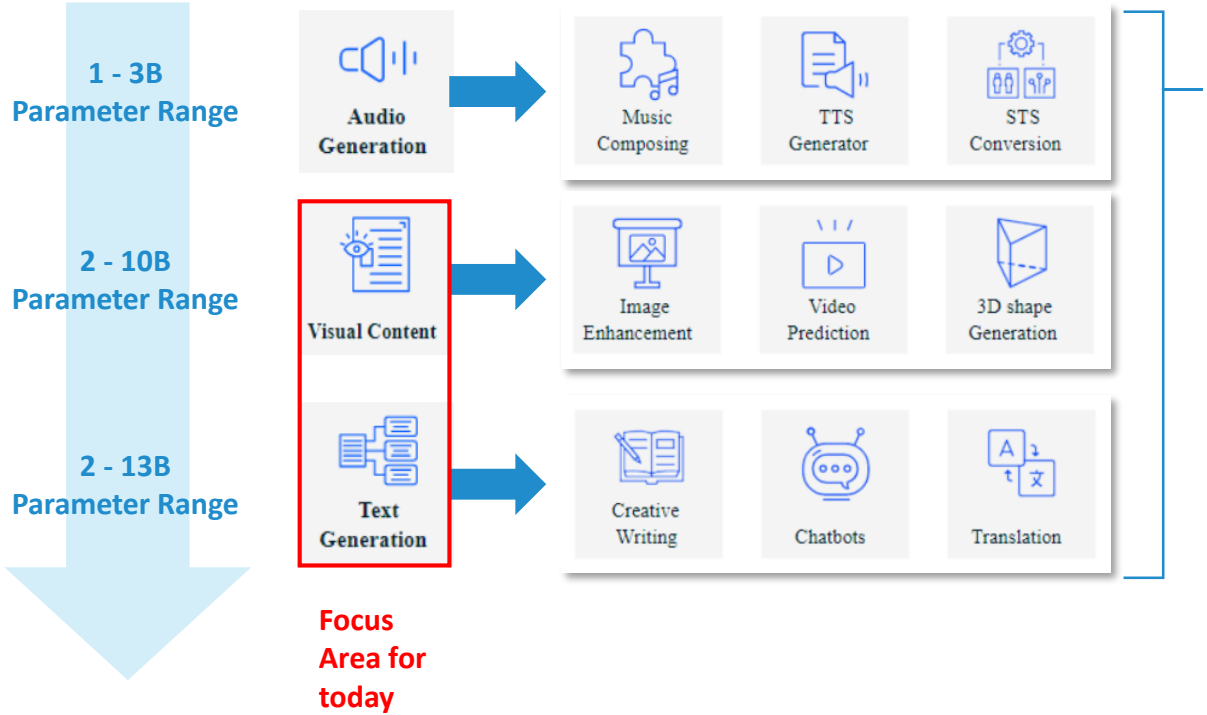
Billions of Users



1: Reuters 2023

GEN AI – Edge Application Deployment Trends

Anticipated to be deployed in 2023/24



+More around Avatar creation, Knowledge based QnA, Intelligent Search, Co-Pilot Assistance & more..

- **Trending towards MUST support larger models =>** Compute, BW & Memory with sustained performance
- **Multi modality fusion for better input prompt definition =>** Sensing + Vision + Text or various combinations
- **Concurrency of models for improved user experience =>** Texture + Stylization + Restructuring for Visual content

Focusing on LVMs

Language-Vision Models (LVMs): Models that combine vision and language

Trends/Attributes of recent generative LVMs:

Prompt-able

- Steerable, user guided, conditioned, grounded, …

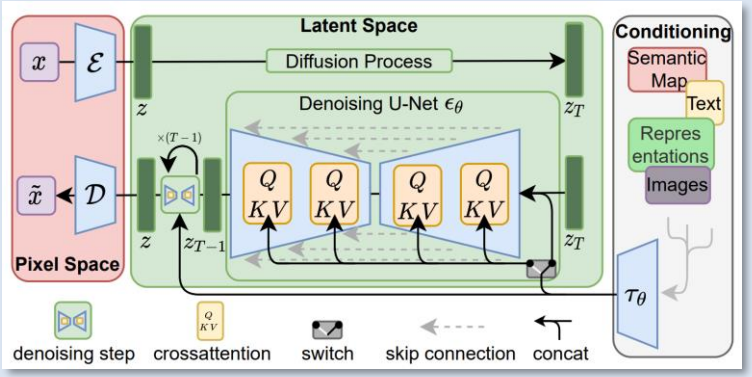
Multi-modal cross-attention

- text/audio/click/3D/image/video/…

Encoder-decoder

Relatively larger

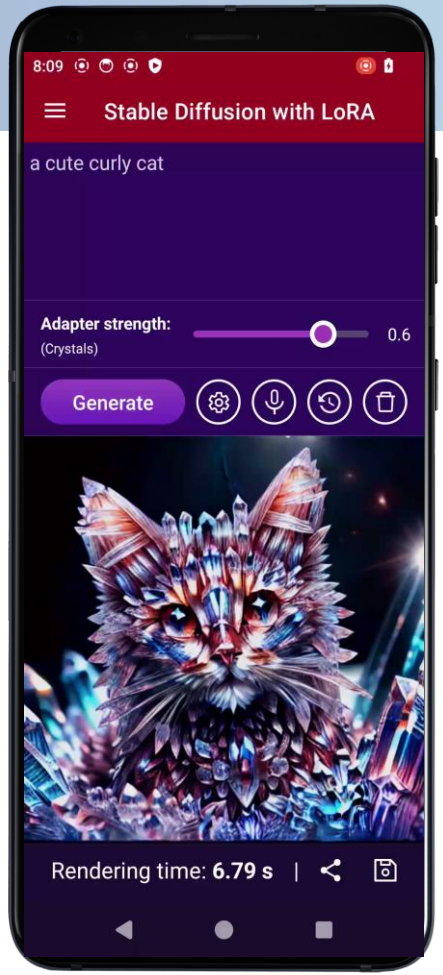
Example: Stable Diffusion (Stability.ai)



Low Rank Adaptation (LoRA)

Our first low rank adaptation (LoRA) on an Android phone done on LVMs

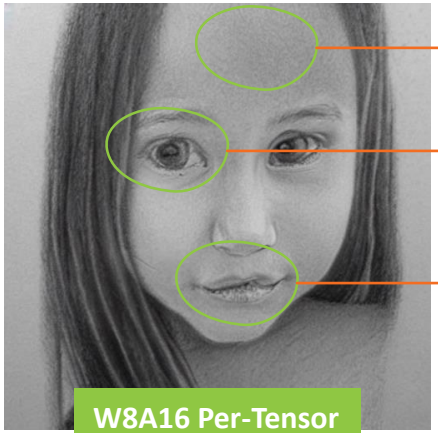
- 1+ billion parameter Stable Diffusion with **LoRA adapter for customized experiences**
- Create **high-quality custom images** based on personal or artistic preferences
- LoRA enables **scalability and customization** of on-device generative AI across use cases
- Full-stack AI optimization **to achieve high performance at low power and fast switching between adapters**
- **Enhanced privacy,** reliability, personalization, and cost with on-device processing



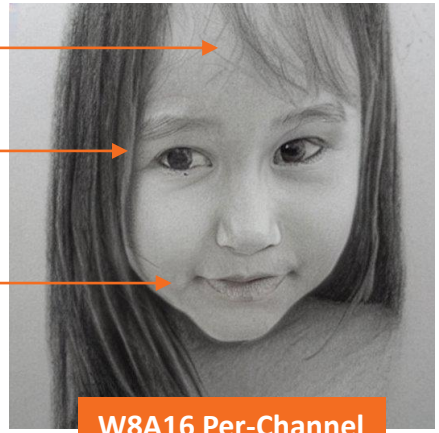
LVMs Challenges

Transitioning from floating point to fixed point

Prompt – I will draw realistic pencil portrait from a photo



X



√

√

Current Status:
Recommended path working with various partners

QAT – Quantization Aware Training

GEN AI LVM Applications

About 3 to 5B Parameters

Visual - LVM

Stable Diffusion + Control Net

Generate & manipulate → **Avatars/Sketch** animated, emotion adaptive

Human-like
AI companions or
Sketch to Image

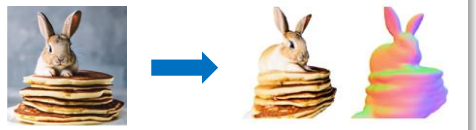
Applications: Generating custom wallpaper, replacing background, Content creation in gaming among others

Stable Diffusion + Control Net + NERF (3D)

Objects textured, stylized, 3D meshes → **Scenes** relighted, restructured

Entire 3D worlds from prompts

3D from a single image

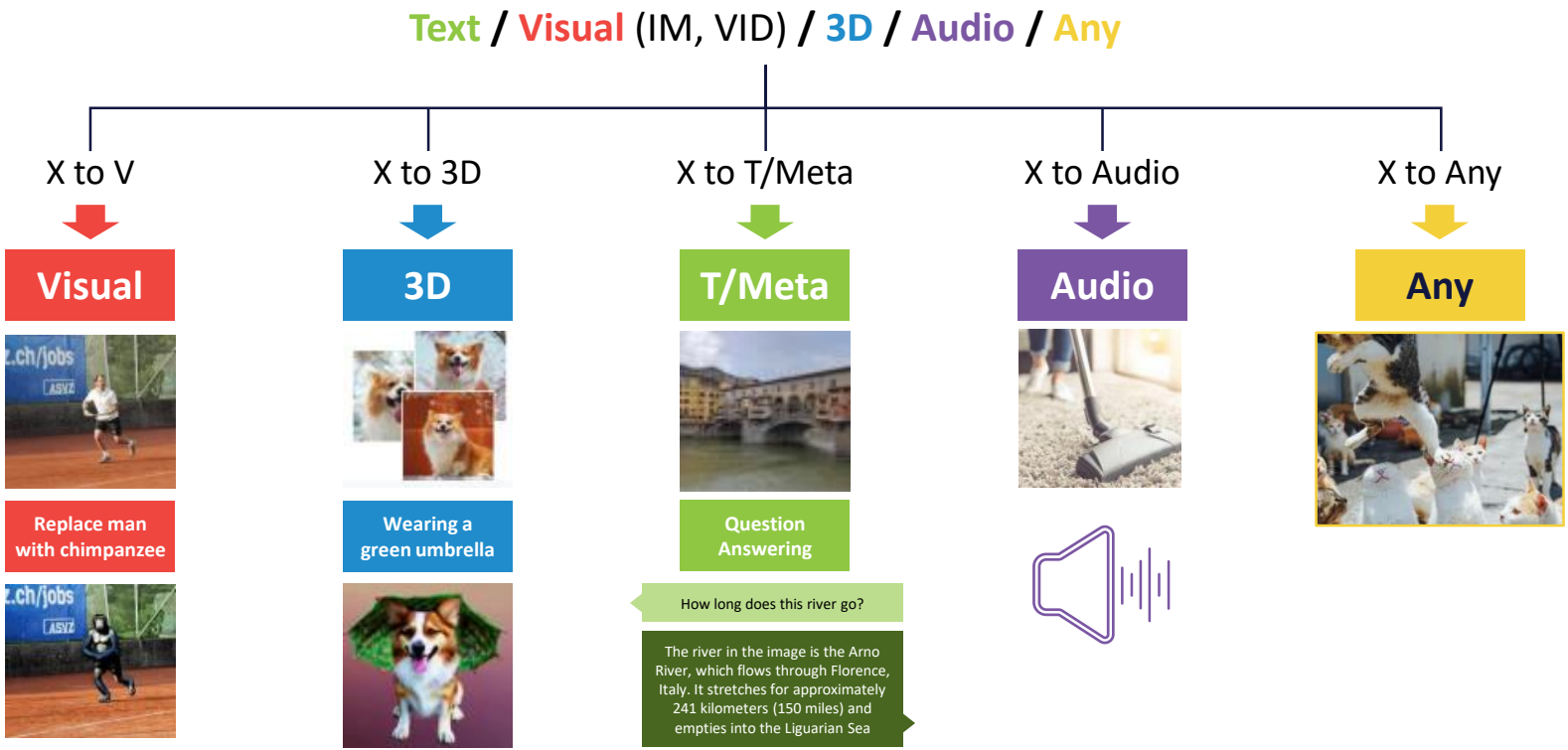


Modify with voice command

Applications: Generating 3D scenes for gallery applications, supporting content creators among others

What Next ?

Lay the foundation for new consumer applications based on creating synthetic content in any modality



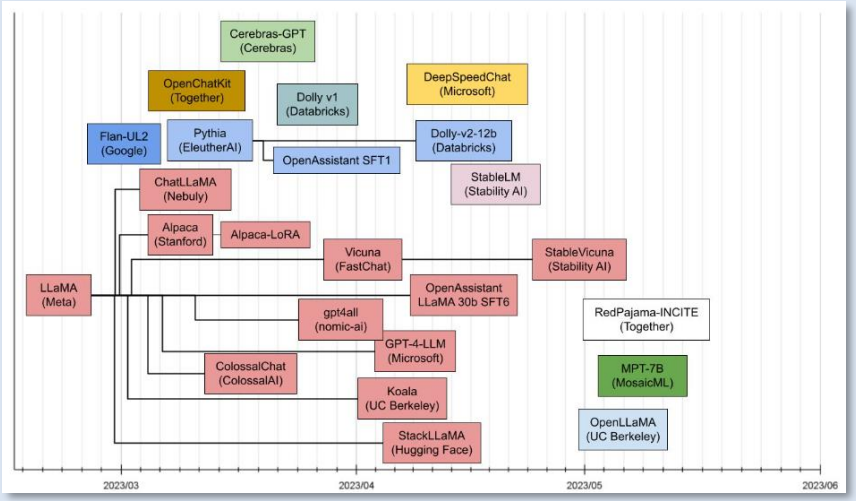
Focusing on LLMs

Large Language Models (LLMs): Models that focus on language

- **Move towards open-source models** (e.g. Llama, Phi)
- **Move towards multi modality** (e.g. GPT4)
- **Movement towards lower bit widths to reduce memory footprint**

Key Ecosystem Asks: move from generic foundational models to domain specific models

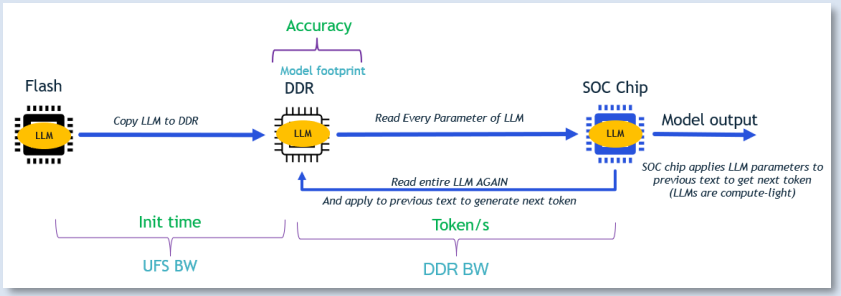
- **Data Ownership:** Depending on models trained on data of unknown origin = Safety concerns
- **Control:** Data is your IP. Own the model generated from that data → Control core IP
- **Model Ownership:** Own your weights – Better introspection, Explainability and Portability



LLM Deployment – KPIs & Memory

Key LLM KPIs

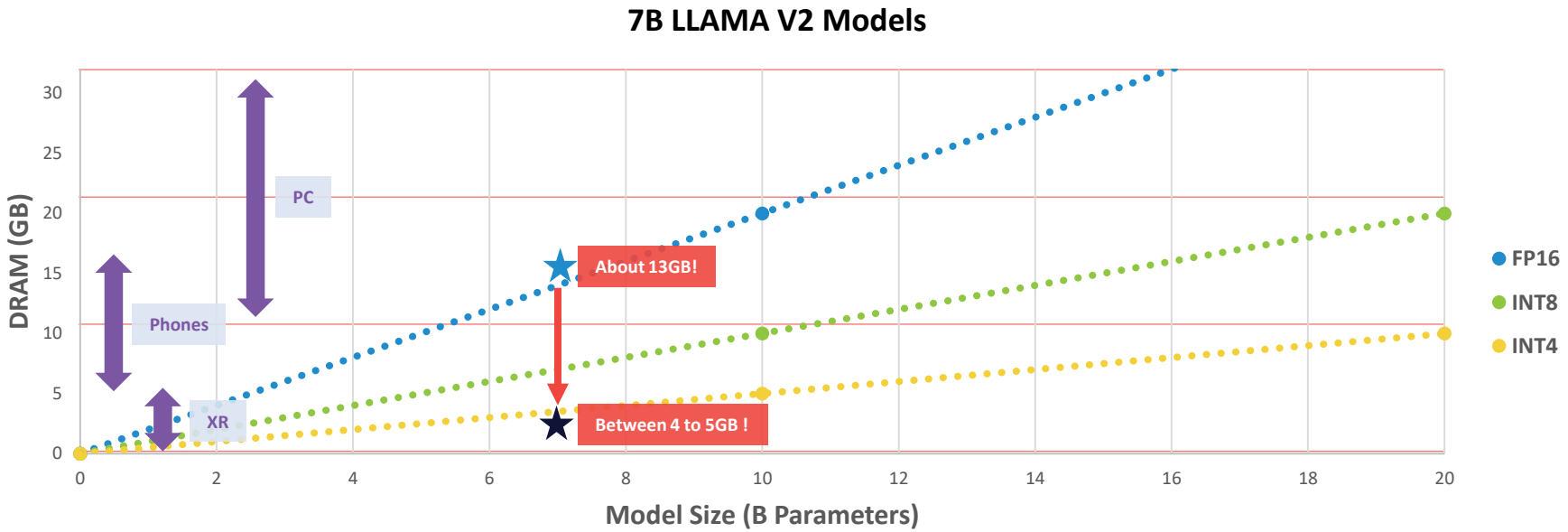
Accuracy →	Depends on model size and context length, which in turn drive DRAM GB needs
Time to First Token (TTFT) →	Mostly compute bound to produce first token very fast; does need high DDR BW
Token/s →	Typically, DDR BW bound as entire model and context needs to be moved from DRAM to AI Engine
Init Time →	Depends on UFS BW to transfer model from Flash to DRAM



Memory footprint: How does quantization Help ?

Mapping to various form factors

Observation : Reduction in memory needs is becoming important to really enable large models while maintaining accuracy



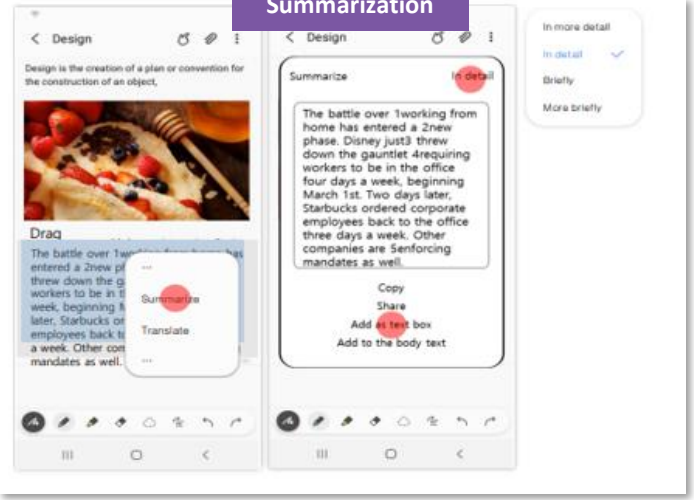
GEN AI LLM Applications – in Commercialization phase

About 1 to 10B Parameters

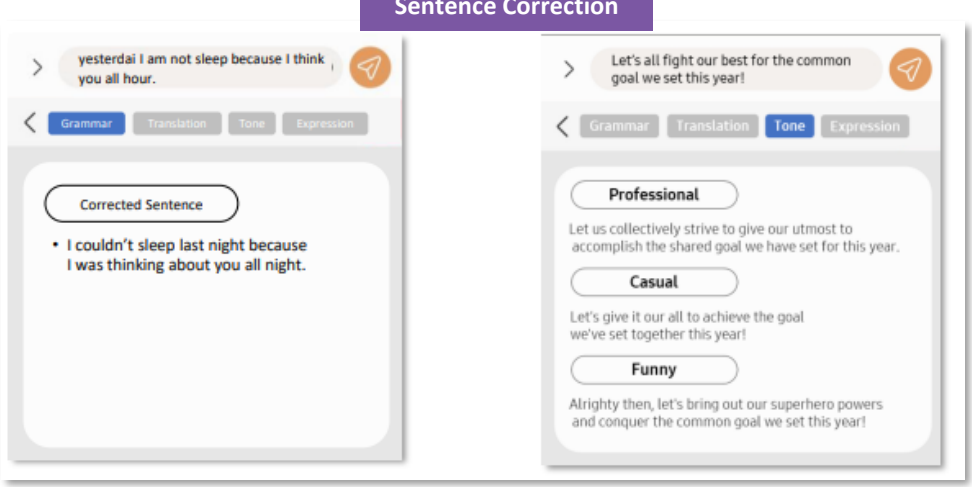
Standalone usage

In combination with other modalities

Summarization



Sentence Correction



Mobile Personal Assistant

Personalized experience integrated with other sensory information and using voice commands

Speech to Speech as interface

Edge AI

Sensing with ASR + LLM + TTS + Visual Avatar

Peripherals

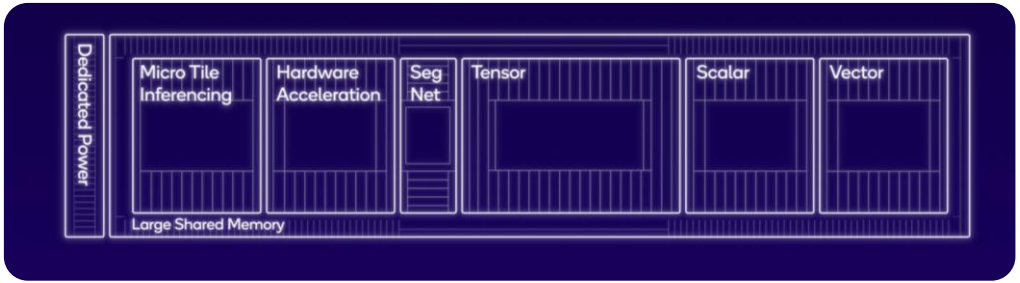
LLM (7-20B)
Fine-tuned data

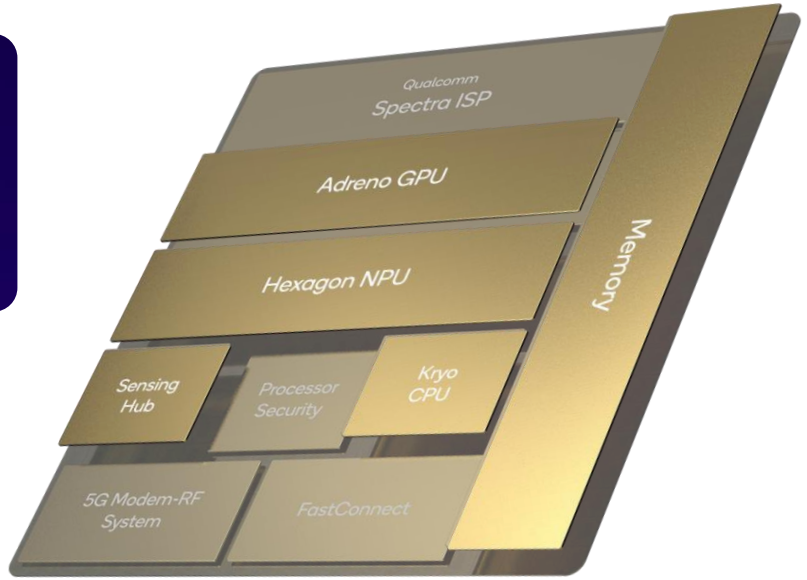
Productivity Assistant

- QnA for queries
- Extend to Plug Ins (Navigation, enterprise, entertainment..)
- Email Creation
- Document Summarization

Qualcomm® AI Engine and Qualcomm® Hexagon™ NPU

Hexagon NPU

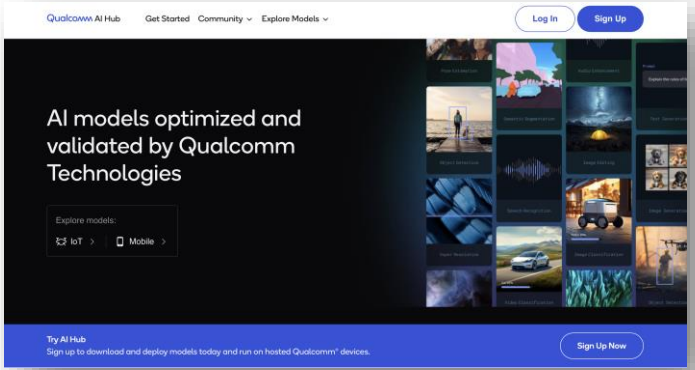




- Upgraded micro architecture
- Upgraded micro tile inferencing
- Peak performance cores
- Higher clock speeds
- 2X higher bandwidth on shared memory

Developer's Gateway to Superior On-Device AI

Qualcomm® AI Hub enables developers to easily quantize, optimize, and validate AI models in minutes

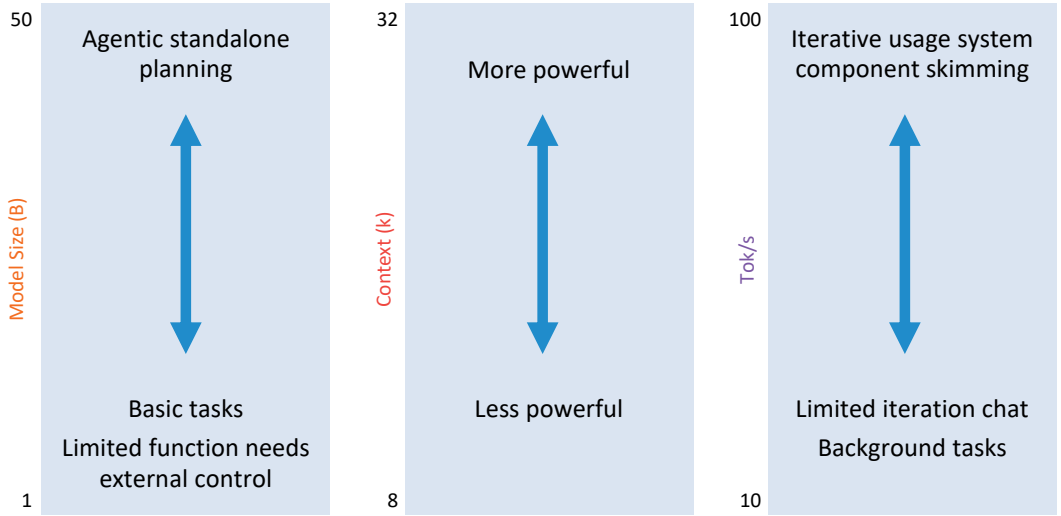


Pick a Model — Target Platform — Pick a runtime — Test and Validate — Deploy

What Next ?

Lay the foundation for new consumer applications based on need for personalization

- Larger models, larger context length enable more powerful use cases
  - E.g., LLMs in a system (e.g., with RAGs); LLMs as Agents - orchestrate sequence of complex tasks
- This also drives need for higher Tok/s due to need to iterate multiple times



RAG: Retrieval Augmented Generation

ICL: In-context learning

Conclusions

- Consumer and enterprise GEN AI applications **cannot scale ONLY with cloud**
- **Significant investments have been done on the edge/client side** that can enable many GEN AI experiences with support for user personalization
- Many ways to support personalization and **one among them is LORA (using Adaptors)**
- Deploying GEN AI applications at scale on the client side does come with **many challenges like accuracy, memory and performance** so focus on many SW optimizations is needed
- Plenty of innovation happening in the ecosystem side that is **expanding from traditional LVM, LLMs to LMMs** while **supporting the need for multi modalities**

Embedded Vision Summit 2024

Qualcomm AI Hub
https://aihub.qualcomm.com/

2024 Embedded Vision Summit

May 21st (1:00-4:00pm) **Workshop**
"Accelerating Model Deployment with Qualcomm® AI Hub" – Bhushan Sonawane

May 22nd (1:30-2:00pm) **Product Related Presentation**
"OpenCV for High-Performance, Low-Power Vision Applications on Snapdragon" – Xin Zhong

May 23rd (9:50-10:20am) **General Session Talk**
"What's Next in On-Device Generative AI" – Jilei Hou

May 23rd (10:20-11:10am) **Panel Session**
"Multimodal LLMs at the Edge: Are We There Yet?" – Jilei Hou (Panel session)

May 23rd (1:30-2:00pm) **Product Related Presentation**
"Deploying Large Models on the Edge : Success Stories & Challenges" – Vinesh Sukumar

Stop by our booth and live demos at exhibit hall booth 718

Thank You

Qualcomm