Generative AI Running On Edge And Hybrid Infrastructure

Karl Freund
Cambrian-AI Research, LLC
July 10, 2023

Introduction

When most people think of Artificial Intelligence (AI), they imagine a berserk Hollywood android, or more realistically, a massive data center filled with racks of GPUs, cranking out the answers to the meaning of life, the universe, and everything (which of course is 42). And the latter is undoubtedly true if incomplete. However, most are unaware we use AI when taking photos or playing games on our smartphones. The user isn't knowingly or directly interacting with artificial intelligence. Instead, edge AI is often hidden in an app, improving the performance or function.

But thanks to the explosive revolution caused by generative AI, Large Language Models (LLMs), and ChatGPT, the time has come when people want to directly interact with an AI application on a mobile device, in their vehicles, or in doctor's offices. Let's call this explicit AI. AI apps like Apple Siri or Google Assistant run primarily in the cloud today. But running them directly on an edge device could bring many benefits if those devices had the capacity and performance to do the job.

This paper will focus on the emergence of AI on the edge and in hybrid AI configurations, where local intelligence is augmented with cloud resources.

AI On The Edge

As cloud vendors begin to reckon with the "eye-watering"[1] costs of generative AI, the major players are looking for resources on edge to carry more of the load. While data center GPUs offer great performance, they can cost over $30K each. Inflection AI, a startup founded by the former head of Deep Mind, raised $1.3 billion from industry heavyweights to build a cloud supercomputer with 22,000 NVIDIA H100 GPUs, costing hundreds of millions of dollars.

To help lower the cost and increase access to the power of LLMs, Microsoft has introduced Office 365 Co-pilot, which uses AI hardware in both the cloud and locally, where possible, to help users across the Windows OS. In another example of seeking the benefits of on-device AI, Google has launched the Gecko version of the Palm 2

[1] Sam Altman, CEO of OpenAI https://www.reuters.com/technology/booming-traffic-openais-chatgpt-posts-first-ever-monthly-dip-june-similarweb-2023-07-05/

model. It is so lightweight that it can work on mobile devices and is fast enough for great interactive applications on-device, even offline. And Meta has released the LLAMA generative AI model, which has a version consisting of only 7B parameters intended for edge devices.

In addition to realizing significant cost reductions, these cloud providers are using artificial intelligence on devices close to the data source to help their customers realize other benefits as well, including:

- **Reduced latency**: Edge AI can provide reliable real-time insights and decision-making, essential for applications such as self-driving cars and medical diagnostics.

- **Improved privacy:** Edge AI can help to protect data privacy by keeping it local. This is important for applications that handle sensitive data, such as financial transactions, personal information, or medical records.

- **Increased accessibility:** Since Edge AI doesn't depend on the availability of networked cloud resources, it can be used when disconnected from the internet, or when internet bandwidth is inadequate. This is important for devices where connectivity varies, such as automobiles.

- **And of course, reduced costs:** Edge AI can help to reduce costs by minimizing the need to send data to the cloud. This is important for devices with limited bandwidth or data plans, such as smartphones and wearables. This also reduces the datacenter and server costs if you can move computing from the cloud.

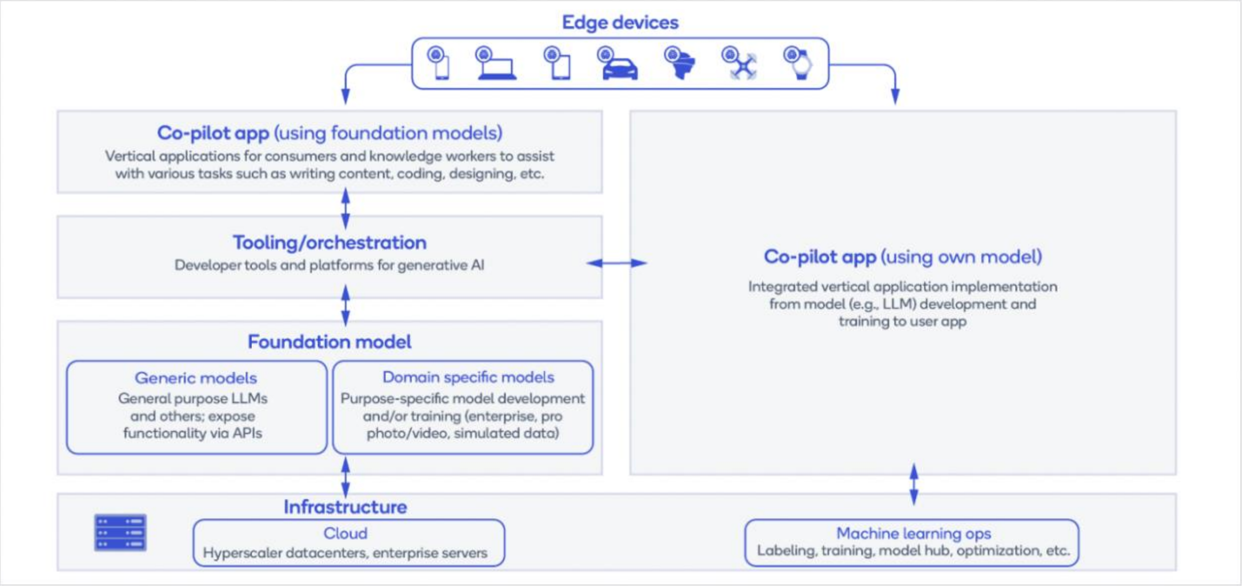

Edge devices
Co-pilot app (using foundation models)
Vertical applications for consumers and knowledge workers to assist with various tasks such as writing content, coding, designing, etc.
Tooling/orchestration
Developer tools and platforms for generative AI
Foundation model
Generic models
General purpose LLMs and others; expose functionality via APIs
Domain specific models
Purpose-specific model development and/or training (enterprise, pro photo/video, simulated data)
Co-pilot app (using own model)
Integrated vertical application implementation from model (e.g., LLM) development and training to user app
Infrastructure
Cloud
Hyperscaler datacenters, enterprise servers
Machine learning ops
Labeling, training, model hub, optimization, etc.

Figure 1: Using AI on the edge is far more complex than just adding AI features to a chip. Qualcomm.

So, if AI on the device is so compelling, why aren't we all using it? A few challenges must be addressed to provide performant AI solutions on edge devices.

- **Computational and memory constraints:** This is perhaps the biggest hurdle to running AI apps on the edge. Edge devices have significantly fewer computational and memory resources than cloud servers. This means that AI models need to be optimized for smaller devices. We will cover some of these techniques in the next section.

- **Heterogeneity:** Edge devices come in a variety of shapes and sizes, with different capabilities and limitations. This makes it difficult for application developers to deliver AI solutions that can run on a wide range of devices.

- **Security:** Edge devices are often connected to the internet, which makes them vulnerable to cyberattacks. Implementing security measures to protect data and devices from unauthorized access can minimize this risk.

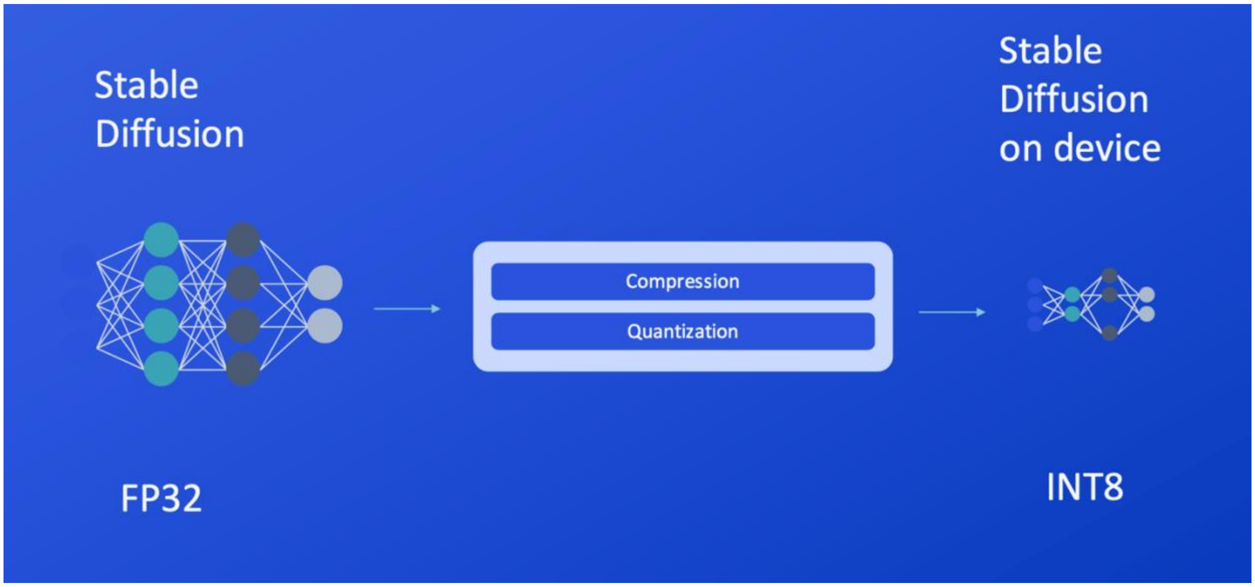
From everything we have seen, only one company seems to have realized the need and implemented adequate AI features on SoCs for edge AI: Qualcomm. Its competitors are almost certainly working hard to catch up, but Qualcomm Technologies has earned the pole position, delivering the benefits and addressing these challenges today.

The Road to Optimization

One of the most talked-about issues with the LLMs that power applications like ChatGPT is the huge cost of inference processing for trillion-parameter models. Submitting a query to ChatGPT using GPT3 or GPT4 can consume a cluster of 8 to 16 high-end GPUs. As a rule of thumb, most such queries cost at least an order of magnitude more than a traditional search would cost. We are not talking about model training here, which can consume thousands of GPUs for months. We are talking about using the LLMs, or inferencing processing.

Several approaches can combine to enable inferencing LLMs on an edge device. First and foremost, is to limit the domain of the model. For example, answering a query about a company's HR practices does not require knowledge about quantum computing or the current political environment that a model like GPT-3 covers. So, instead of running a model with hundreds of billions of parameters, you can do some very useful AI with, say, one billion parameters.

One example of smaller but useful models is the image creation application Stable Diffusion, which requires about a billion parameters. At the Microsoft Build conference, one could see a demonstration wherein Stable Diffusion ran on a laptop. They used model compression and quantization to 8-bit integers to reduce the model size to something that can fit and be processed by an edge SoC with reasonable performance.



Stable Diffusion

Compression

Quantization

Stable Diffusion on device

FP32

INT8

Figure 2: Stable Diffusion on device. Qualcomm

The impressive results show what is possible today with a state-of-the-art edge SoC and some creative software engineering. The demo below, running on a Qualcomm

Snapdragon-powered device without internet connectivity and shown at the recent Mobile World Conference, could generate high-quality images from simple text prompts in less than 15 seconds. A Qualcomm spokesperson said the company plans to handle models with over 10 billion parameters in the coming months.

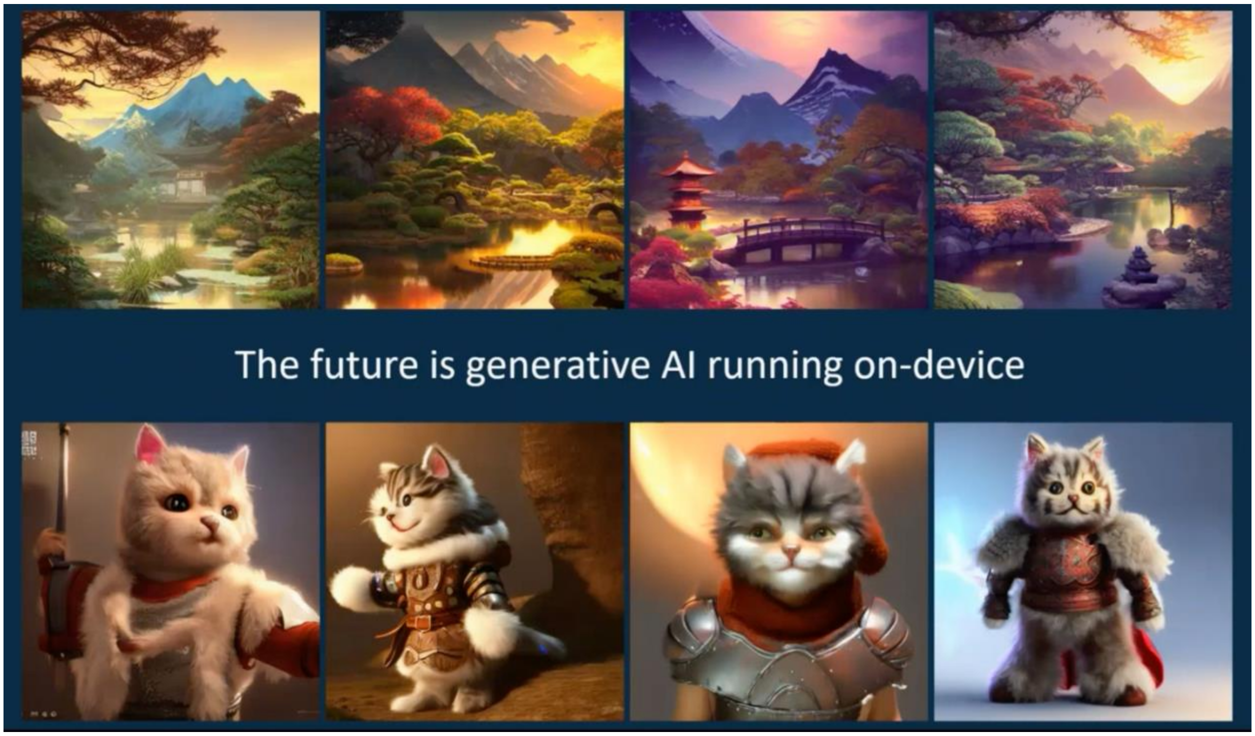


Figure 3: Watching a laptop or mobile phone generating high-res images from text is quite illuminating. Qualcomm

EDGE AI USE CASES

There are several use cases for edge AI, including:

- **Self-driving vehicles:** Edge AI is used in self-driving cars to detect objects, such as other cars, pedestrians, and cyclists. This information is used to make real-time decisions about how to navigate the road safely.

- **Fraud detection:** Edge AI can detect fraudulent transactions in real time. This information is used to prevent fraud and protect customers from financial losses.

- **Medical imaging:** Edge AI is used to analyze medical images, such as X-rays and MRI scans. This information is used to diagnose diseases and make treatment decisions.

- **Industrial automation:** Edge AI can automate industrial processes, such as quality control and predictive maintenance. This information is used to improve efficiency and reduce costs.

- **Real-time decision-making:** Edge AI can make real-time decisions, such as detecting objects in self-driving cars or identifying fraudulent transactions in real time.

- **Improved performance:** Edge AI can improve the performance of applications by reducing latency. This is important for applications that require fast response times, such as gaming and video streaming.

- **Smart productivity application:** Microsoft 365 Copilot is a new productivity feature that uses generative AI to help write and summarize documents, analyze data, or turn simple written ideas into presentations – all embedded in Microsoft apps, including Word, Excel, PowerPoint, Outlook, Teams, and more.

HOW DO WE GET THERE?

Optimizing and quantizing large language models is critical to making generative AI practical on edge devices. Large language models (LLMs) are a type of AI model that can be used for various tasks, such as natural language processing (NLP) and machine translation. However, LLMs can be computationally expensive to train and run.

Several techniques can be used to optimize and quantize LLMs for edge devices. One technique is to use a technique called "knowledge distillation," or “Domain reduction”, which involves training a smaller model to mimic the behavior of a larger model on a smaller data set. For example, a company building a customer service chatbot needs a model trained on its products. It does not need the hundreds of billions of parameters that are irrelevant to their business.

Another technique to reduce model size and improve performance is "quantization," which involves reducing the precision of the model's weights and activations without significantly impacting its accuracy. This can be tricky; you don’t want to accidentally opt for a highly efficient model using 8-bit integers that cannot give accurate answers.

Research has been underway for several years to enable better quantization, both in training and inference, to enable edge AI, including 8-bit floating point and integers for weights.

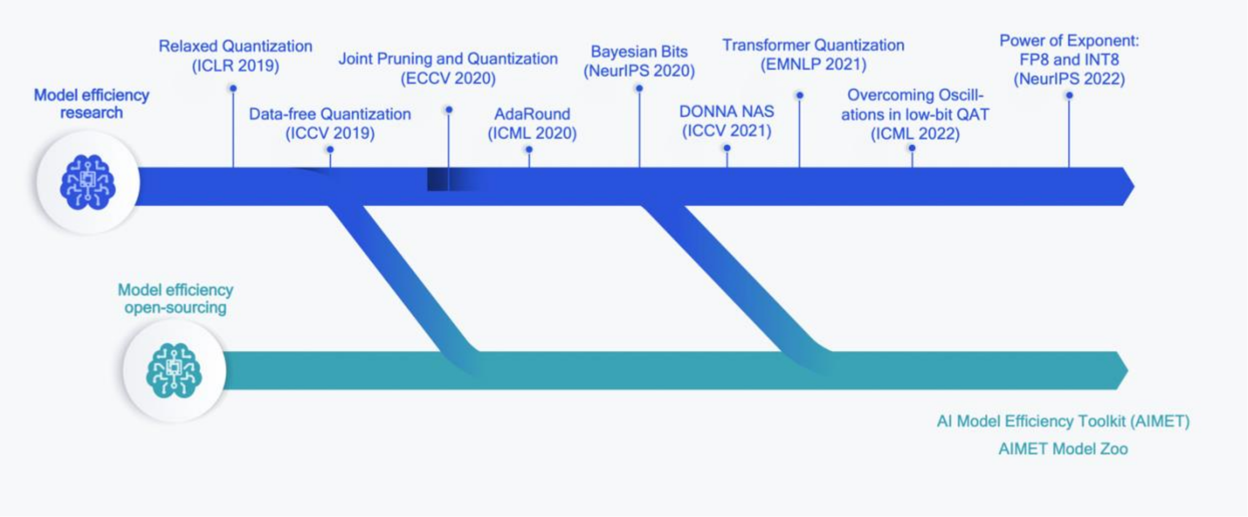
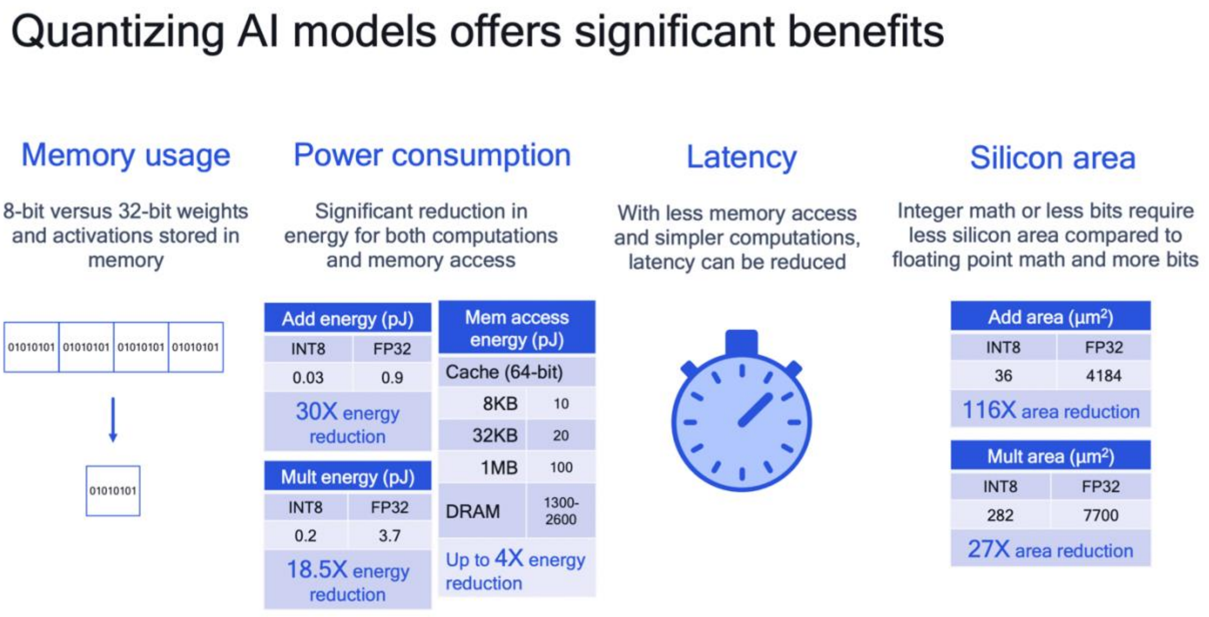

Model efficiency research
Relaxed Quantization (ICLR 2019)
Data-free Quantization (ICCV 2019)
Joint Pruning and Quantization (ECCV 2020)
AdaRound (ICML 2020)
Bayesian Bits (NeurIPS 2020)
DONNA NAS (ICCV 2021)
Transformer Quantization (EMNLP 2021)
Overcoming Oscill-ations in low-bit QAT (ICML 2022)
Power of Exponent: FP8 and INT8 (NeurIPS 2022)
Model efficiency open-sourcing
AI Model Efficiency Toolkit (AIMET)
AIMET Model Zoo

Figure 4: Techniques have been developed over the years to optimize complex models so they can execute efficiently on a mobile processor at the edge.

Model quantization can produce many benefits, including better memory optimization, reduced power consumption and latency, and smaller silicon die area. Today's state-of-the-art is mostly around going from the 32-bit floating point weights and activations produced by the training process to 8-bit integers. Researchers are seeing a significant reduction in model size and improved performance, while attaining accuracy within 1.0 percent of that achieved with 32-bit floating point. A more recent development is 4-bit processing which is equally promising. In fact, many LLMs on Hugging Face are already available with 4-bit quantization. Qualcomm has also demonstrated super resolution running in 4-bit hardware at https://www.youtube.com/watch?v=v27xs7boMtA.



Quantizing AI models offers significant benefits

Memory usage

8-bit versus 32-bit weights and activations stored in memory

01010101
01010101
01010101
01010101
01010101

Power consumption

Significant reduction in energy for both computations and memory access

Add energy (pJ)	
INT8	FP32
0.03	0.9
30X energy reduction	

Mult energy (pJ)	
INT8	FP32
0.2	3.7
18.5X energy reduction	

Mem access energy (pJ)	
Cache (64-bit)	
8KB	10
32KB	20
1MB	100
DRAM	1300-2600
Up to 4X energy reduction	

Latency

With less memory access and simpler computations, latency can be reduced

Silicon area

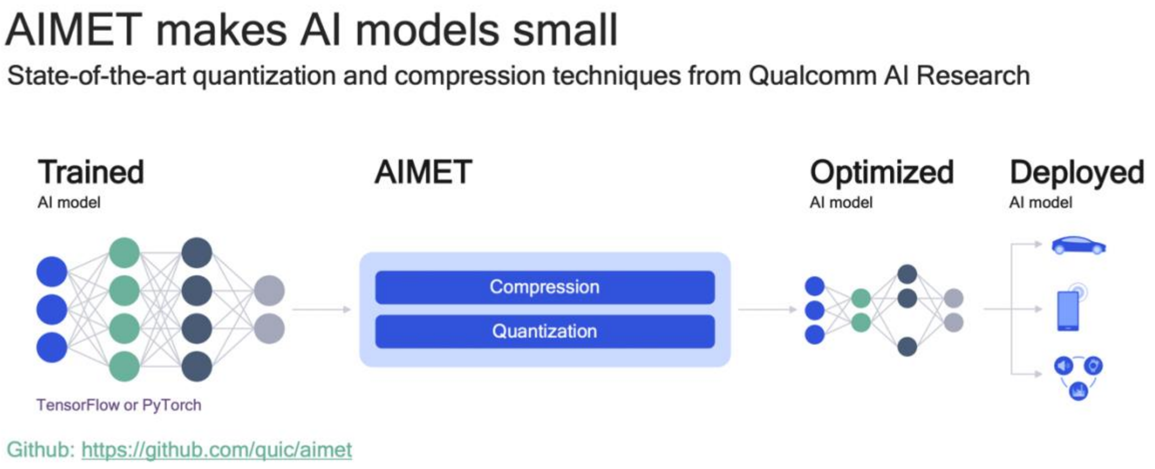
Integer math or less bits require less silicon area compared to floating point math and more bits

Add area (μm^2)	
INT8	FP32
36	4184
116X area reduction	

Mult area (μm^2)	
INT8	FP32
282	7700
27X area reduction	

Figure 5: The impact of quantization.

For its Snapdragon chips, Qualcomm offers an optimization platform called the AI Model Efficiency Toolkit, or AIMET, which includes both compression and quantization workflows to optimize AI models for edge deployment.



AIMET makes AI models small

State-of-the-art quantization and compression techniques from Qualcomm AI Research

Trained AI model

TensorFlow or PyTorch

AIMET

Compression

Quantization

Optimized AI model

Deployed AI model

Github: https://github.com/quic/aimet

In addition, quantization-aware training can significantly improve accuracy of 8-bit int models. When going to 4-bit int, however, the typical approach to rounding to the nearest value can induce noise into the predictions. A team in Qualcomm Netherlands has come up with Adaptive Rounding (AdaRound) that introduces a new way of rounding that can enable 4-bit models, increasing performance by 2X and reducing energy consumption by up to 4X as well (compared to 8-bit).

HYBRID SOLUTIONS:

In some cases, it may be necessary to use a hybrid solution, where some of the processing is done locally and some in the cloud. This can be a good option for applications that require high accuracy or that need to process larger amounts of data than the edge device can contain. The local processing needs to be augmented with cloud compute services and the application needs to know when to run what where to provide a seamless experience for the user.

The hybrid AI approach is applicable to virtually all generative AI applications and device segments – including phones, laptops, XR headsets, vehicles, and IoT. For example, if the model size, prompt, and generation length is smaller than a certain threshold and provides acceptable accuracy, inference can run completely on the device. If the task is more complex, the model can run across cloud and devices. Hybrid AI also allows for devices and cloud to run models concurrently – with devices running 'light' versions of the model for low latency while the cloud processes multiple tokens of the 'full' model in parallel and corrects the device answers if needed.

SOFTWARE FOR EDGE AI

Finally, we note that the software for edge AI needs to enable efficient model development and be able to deploy across a large range of devices. From the AI models down through optimization layers, and finally to a portfolio of supported devices, the programmer and data scientist need to see a rationalized stack of AI software that can help them get the model running with the desired level of accuracy, latency, power consumption, and model size. Qualcomm offers these capabilities in the Qualcomm AI Stack.

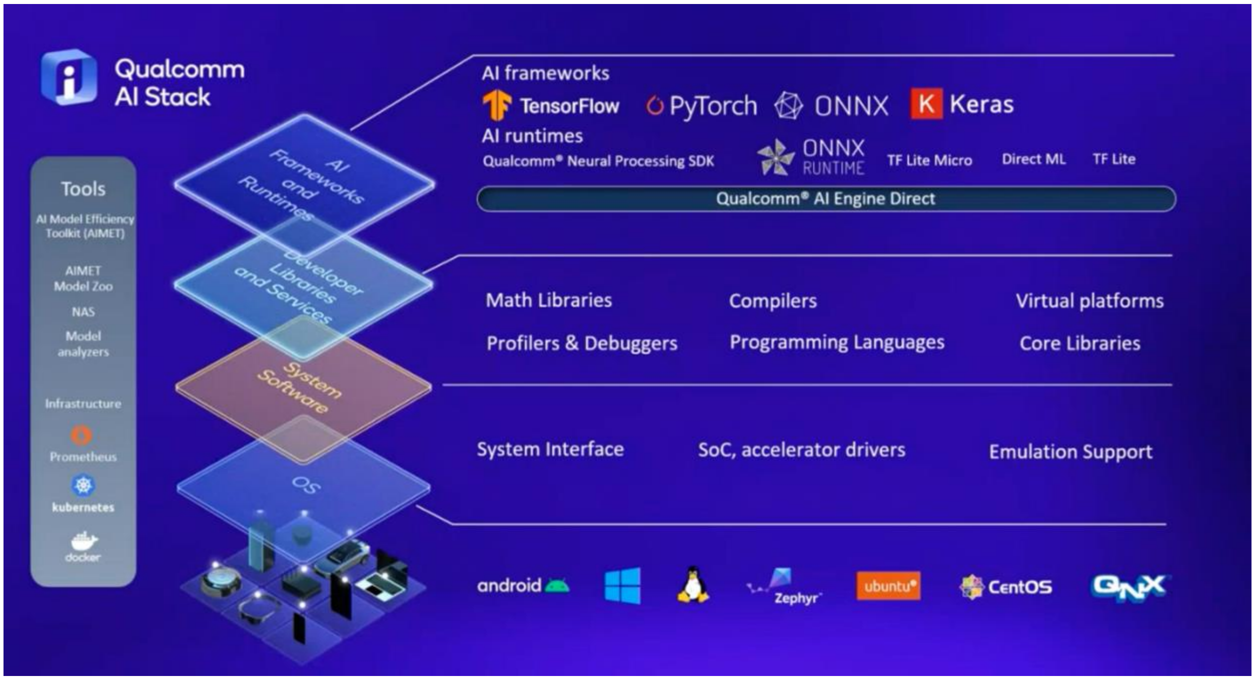


Figure 6: The Qualcomm AI Stack.

CONCLUSIONS

The world is rushing to embrace generative AI and LLMs, but most get a shock when they realize the costs involved, which can be 10X[2] higher than traditional search algorithms. Some say that the LLM explosion will fizzle if these costs cannot be constrained, reduced, and redirected. Edge AI shows the promise of utilizing AI-enabled edge devices to significantly offload the processing required while improving user experiences with high quality, low latency, and enhanced privacy. Mobile SoC vendors such as Qualcomm are developing silicon and optimization to deliver on the promise of AI on the Edge, and the leaders in AI development are building the smaller models that will power the shift.

Consequently, Edge AI is a promising technology that has the potential to revolutionize a wide range of applications. Advances in SoC design, model optimization, quantization, and hybrid solutions are addressing the challenges of using edge AI. As AI technology continues to develop, we can expect to see even more innovative and groundbreaking applications for edge AI in the coming years.

[2] https://arstechnica.com/gadgets/2023/02/chatgpt-style-search-represents-a-10x-cost-increase-for-google-microsoft/

Important Information About This Paper

Contributors and Publisher

Karl Freund, Founder and Principal Analyst, Cambrian-AI Research LLC

Inquiries

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission by Cambrian-AI Research for any citations.

Licensing

Disclosures

This document was developed with Qualcomm Inc. funding and support. Although the document may utilize publicly available material from various vendors, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.