Edge AI: The next step in AI infrastructure

REGIONS ASSET MANAGEMENT

July 2024

Introduction

Most of the development in generative AI, so far, has been in the Cloud with hyperscalers like Amazon Web Services (AWS), Microsoft (MSFT) and Alphabet (GOOG) leading the way. A key advantage of these hyperscalers is to provide the scale needed for adoption of AI in enterprise computing. However, in an effort to increase response time, there is now an emergence of Edge AI, for integrating AI in edge computing.

Edge AI

Edge computing is the process of storing and analyzing data close to the devices that produce the data and users who consume it. Edge computing provides some advantages over Cloud computing.

1. *Minimal Latency:* Latency is the time it takes for data to travel between two points in a network. If data analysis and AI is done in the Cloud, data has to travel from the point of generation, which is at the edge, to large data centers in the Cloud. This can take time. In Edge computing, data analysis is done close to where it is generated. Thus, latency is minimal.
2. *Data security:* While most hyperscalers provide data security, data breaches do occur occasionally. These breaches generally occur when data leaves the edge and travels to the Cloud. By computing at the edge, data can be analyzed and stored within local area networks (LAN) and behind company firewalls, thus reducing chances of data breach.
3. *Energy Efficiency:* Since Edge AI processing is done locally, there is no need to transport and process vast amounts of data to centralized data centers. This reduces energy consumption.
4. *Customization:* On device processing allows companies to build AI models specific to their needs on LANs.

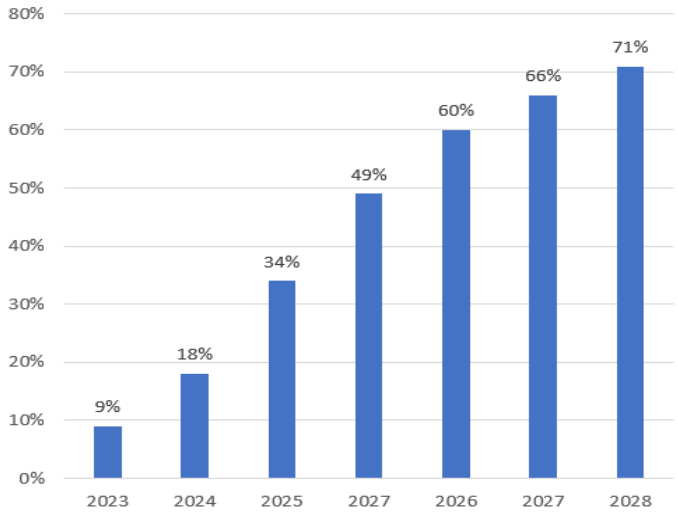
Global market for edge AI is expected to grow from about $19B in 2023 to $144B by 2033 for a CAGR of 26% over that time period [1]. The roll out of 5G networks along with the development of new chips is expected to stimulate this growth. Development of apps is expected to further fuel this growth. Edge AI allows for real time decision making, which is critical to many industries. For example, in manufacturing, predictive maintenance is crucial for minimizing down time. Edge AI allows for real time monitoring of systems to optimize production. In the automotive sector, Edge AI is important for driverless cars, where processing data from various inputs in real time is crucial for safety. In healthcare, it allows for monitoring of remote patients in real time for immediate medical interventions. In the energy and utility sectors, Edge AI can optimize energy distribution by analyzing usage data in real time. Smart phones, personal computers (PCs), sensors, and smart cameras are just some examples of devices that can be used for Edge AI[2].

Investment Case

An AI PC is a personal computer which can run AI software and applications locally, i.e., without connecting to the Cloud. AI PCs need significant processing power to run AI operations. Thus, in addition to Central Processing units (CPUs), these need Graphic Processing Units (GPUs) and Neural Processing Units (NPUs). The number of TOPs (trillion operations per second) of NPUs determine the performance of AI PCs. Microsoft recently suggested that a minimum of 40TOPs for Al PCs. The initial AI PCs are expected to have TOPs of about 40 while, those shipped after 2024 are expected to have TOPs greater than 40. Analysts expect AI PCs market share to increase from about 18% of total PC shipments in 2024 to more than 70% of total PC shipments in 2028, for a CAGR of almost 44% (Figure 1)[3].

Author: Anjali Tagare, PH.D., Microbiology
Regions Investment Management, Equity Analyst

Figure 1: Marketshare Estimate of AI PCs[3]



Also, due to productivity gains, early adoption, between 2024 through 2025 is expected to be in the commercial market followed by the consumer market. There are several companies that have developed or are developing chips for AI PCs. Qualcomm (QCOM), Apple (AAPL), Intel (INTC) and Advanced Micro Devices (AMD) are currently producing or developing chips that run near 40 TOPS (Table 1). To date, the most advanced AI PC app is MSFT's Copilot, which can help with various tasks such as creating documents, searching the web, and managing files. MSFT recently announced its **Microsoft Surface®** laptop and **Surface Pro®** tablet, its Copilot + PCs using QCOM's AI chipset, **Snapdragon X Elite™**. Other PC makers are also launching similar AI PCs using QCOM's **Snapdragon™** chipset. NVDA does not have a NPU chipset yet, however, its **GeForce RTX®** offers more than 1300 TOPS performance.

Source: State of the union: ARM MU well placed as AI expands on-device to Phones and PCs – Bank of America, June 17, 2024.

Table 1: AI PCs Chips [5]

Company	*Chipset*	*TOPS*
QCOM	**Snapdragon X Elite™**	45
AMD	**Ryzen AI 300 series™**	50
INTC	Lunar Lake	48
AAPL	M4	38
NVDA	**GeForce RTX®**	1300

Source: Company Sources

SourcesSmart phones are another device for Edge AI. AI smart phone market share of about 3% in 2023 is expected to increase to about 8% in 2024 and about 40% in 2027 for a CAGR of 83%4. Integrating AI features in a smart phone will allow users to edit photos as well as generate new images. It will also allow for editing text and be able to do live translation. There are several AI smartphones on the market including, **Google Pixle 8 Pro™**, and **Samsung Galaxy®**. APPL recently announced that its **iPhone 15 PRO®** and **iPhone 15PRO Max®**, which will be launched in October 2024, will have AI features.

QCOM is the leading producer of mobile chips. Most mobile chips are based of chip architecture designed by ARM PLC (ARM). ARM architecture allows for more efficient processing while lowering energy consumption. ARM licenses its architecture to semiconductor companies developing CPUs. For example, QCOM, AAPL and Samsung license ARM architecture to develop their AI chips. ARM licensing royalties for AI chips is higher than royalties on non-AI chips.

Summary

Edge computing reduces latency and energy consumption. It allows for AI customization while improving data security. As a result, Edge AI is expected to be the next frontier in the build out of AI infrastructure. Global market for edge AI is expected to grow from about $19B in 2023 to $144B by 2033 for a CAGR of 26% over that time period[1] . AI PCs and AI smartphones require specialized chips, which contain CPUs, GPUs and NPUs, to run AI algorithms. Companies developing these chips are expected to benefit from the growth in Edge AI.

References:

1. https://scoop.market.us/edge-ai-market-new/
2. https://stlpartners.com/articles/edge-computing/edge-computing-devices/
3. State of the union: ARM MU well placed as AI expands on-device to Phones and PCs – Bank of America, June 17, 2024
4. https://www.counterpointresearch.com/insights/over-1-billion-generative-ai-smartphones-to-be-shipped-cumulatively-during-cy-2024-2027/
5. Company Sources: https://www.qualcomm.com/products/mobile/snapdragon/laptops-and-tablets/snapdragon-x-elite, https://www.amd.com/en/partner/articles/ryzen-ai-300-series-processors.html, https://www.intel.com/content/www/us/en/newsroom/news/intels-lunar-lake-processors-arriving-q3-2024.html, https://www.apple.com/newsroom/2024/05/apple-introduces-m4-chip/, https://www.nvidia.com/en-us/geforce/rtx/