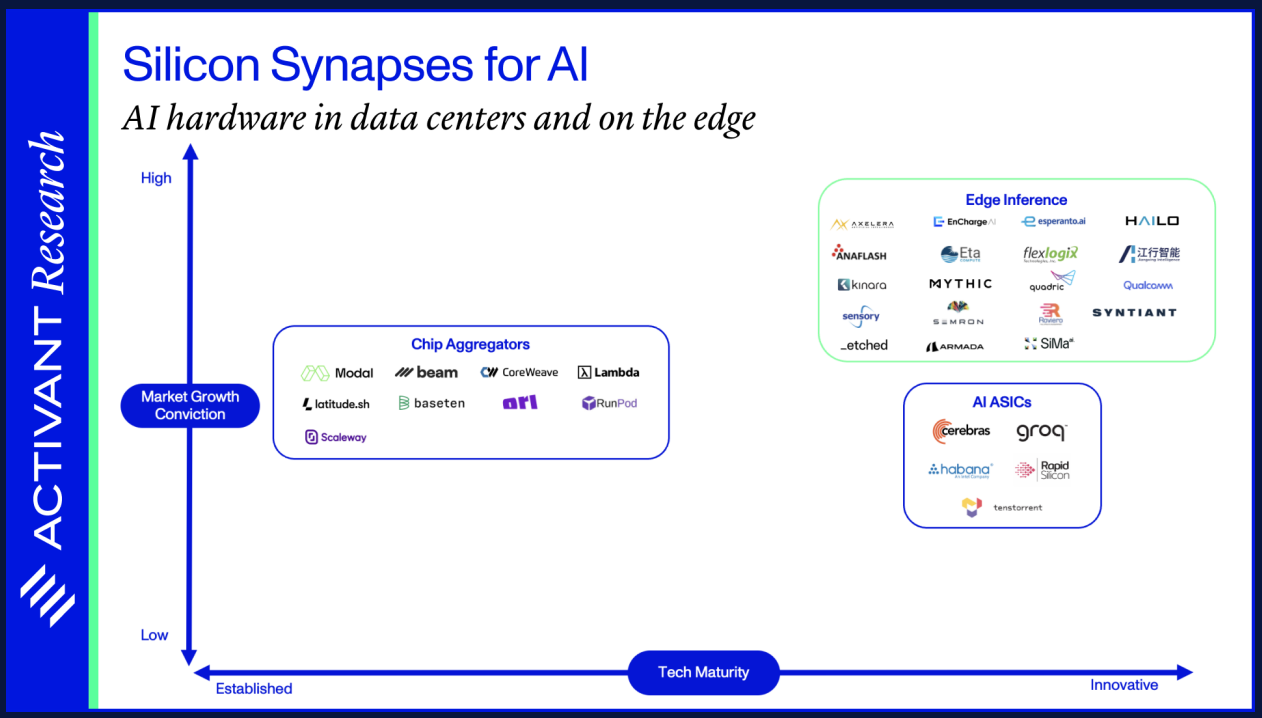
ACTIVANT RESEARCH

Silicon Synapses for AI

AI Hardware in data centers and on the edge

Simphiwe Msibi, Marc Wu



Q1 2024

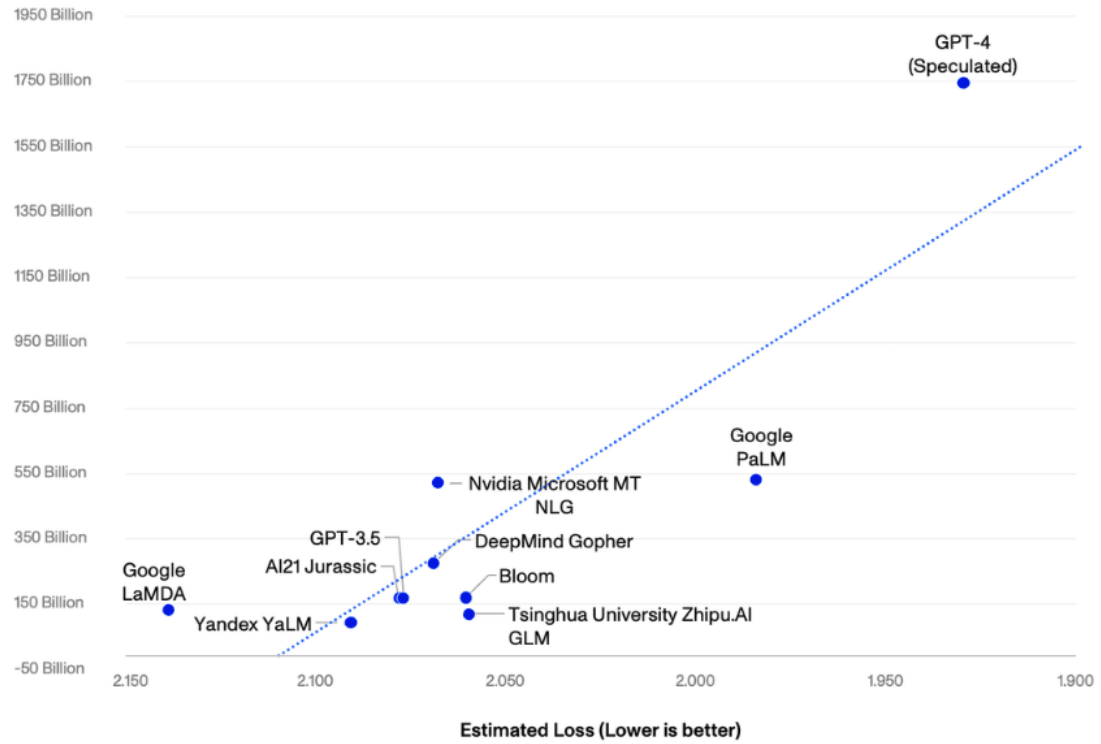
Bigger Models Need Bigger Compute Budgets

Large Language Models (LLMs) and ChatGPT have reaccelerated growth in demand for high-performance hardware. ChatGPT's rapid growth has showcased an immediately tangible use case for the new technology, sparking competition from other tech giants. Computational power, or compute, is crucial for the continued development of larger, more powerful models. Compute, for our purposes, is measured in floating point operations (FLOPs) and is used to refer to both computer processors and the computing performance thereof.

There are three phases in the lifecycle of a Large Language Model (LLM): Pre-training, fine-tuning and inference (the computation a model does to generate a response to user input). The most computationally intensive portion is pre-training wherein the model's parameter weights are optimized based on a large body of text data to learn the patterns, structure, and semantics of a language. Every AI model has some defining attributes: the model's size (measured in parameters), the size of the training data (measured in tokens), the cost to train the model, and the model's performance after training (measured in expected loss). Larger models are more performant than smaller models. However, larger models require more computing to be trained.

ACTIVANT *Research*

AI Model Performance Improves as Size Increases

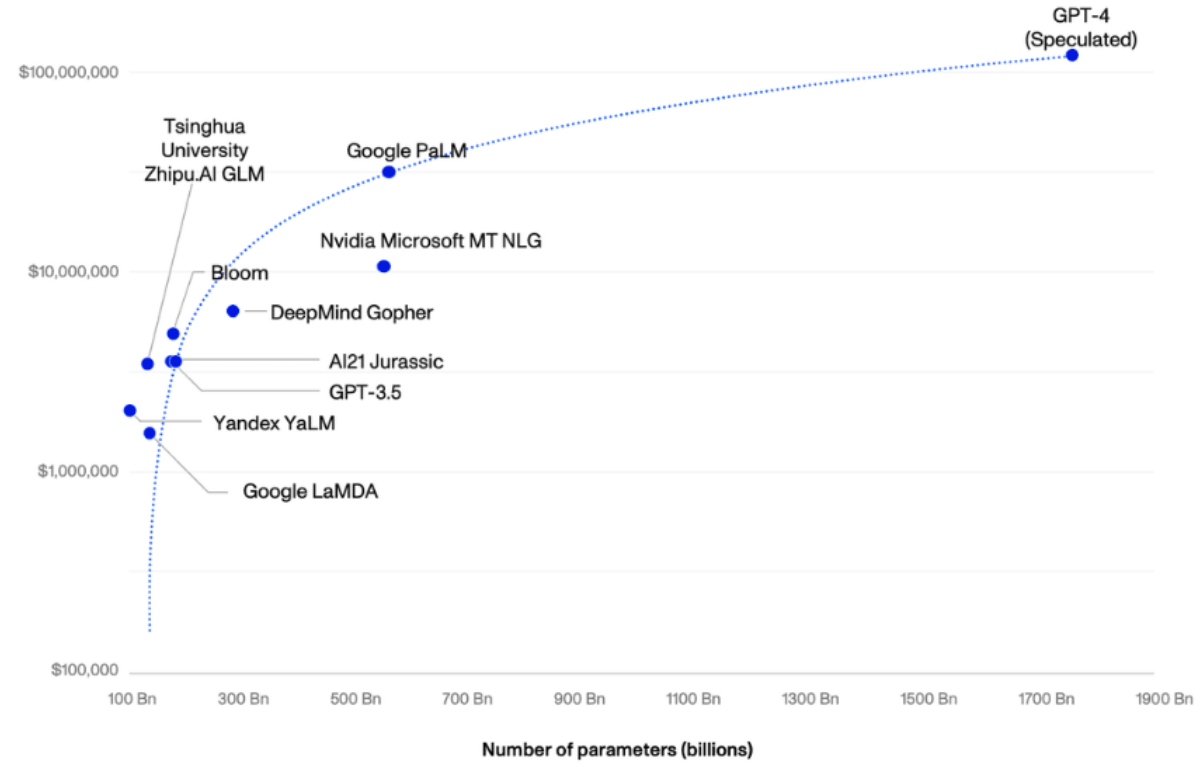


Source: Hoffmann et al, 2022; J. Sevilla et al, 2022; Activant Research

The underlying costs associated with training these models increase exponentially as models get bigger.[1] We estimate that OpenAI's 1.75tn parameter GPT-4 will cost ~$140mn to train, more than 20x greater than its 175bn parameter predecessor GPT-3.5 at $6mn. We further estimate that the subsequent 10x increase in parameters will require $7bn in training costs, $172bn in cumulative hardware, and the equivalent amount of electricity produced by three nuclear plants in a year (~26 000 GWh).[2]

ACTIVANT *Research*

Large Models are Exponentially more Expensive to Train



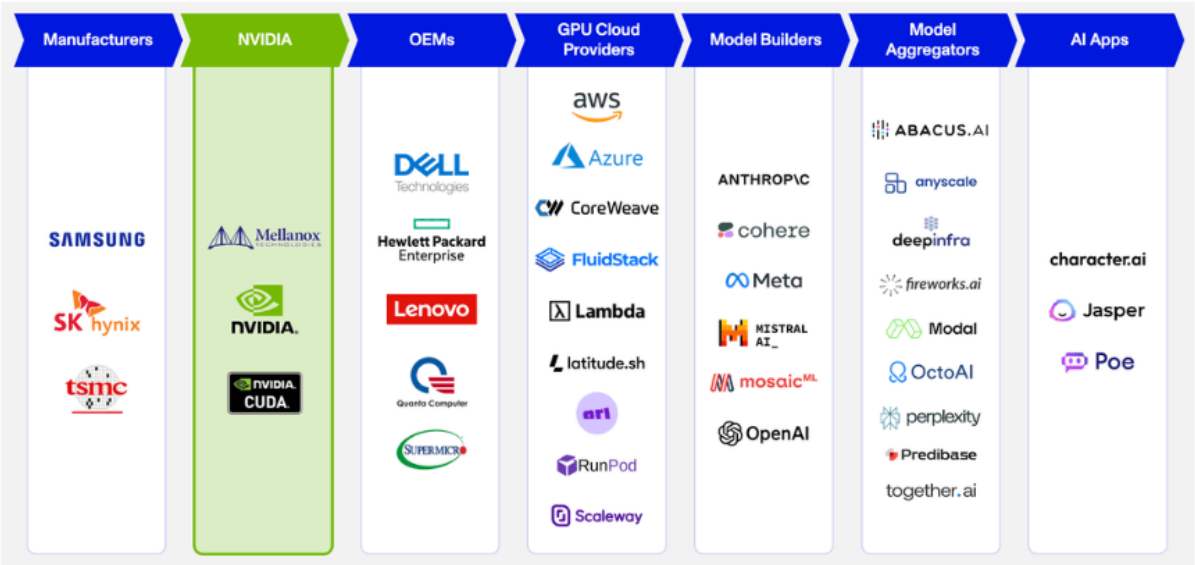
Source: Kaplan et. al, 2019; J. Sevilla et al, 2022; Activant Research

As models have gotten larger, **hardware has stagnated**. The doubling rate for hardware FLOPs has slowed materially since the advent of deep learning in c.2010 (from doubling every 1.5 years to doubling every 3 years).[3] This will naturally result in a stagnation in model performance or a bifurcation between compute-rich large model builders, who have the resources to acquire more compute, and compute-poor small model builders. The industry has previously resolved the mismatch between more performant models and stagnating compute by advancing our algorithmic techniques, data availability or **fundamentally changing our hardware.**[4]

NVIDIA's compute empire

ACTIVANT *Research*

NVIDIA is Deeply Entrenched in the AI Ecosystem



Source: Company records, Activant Research

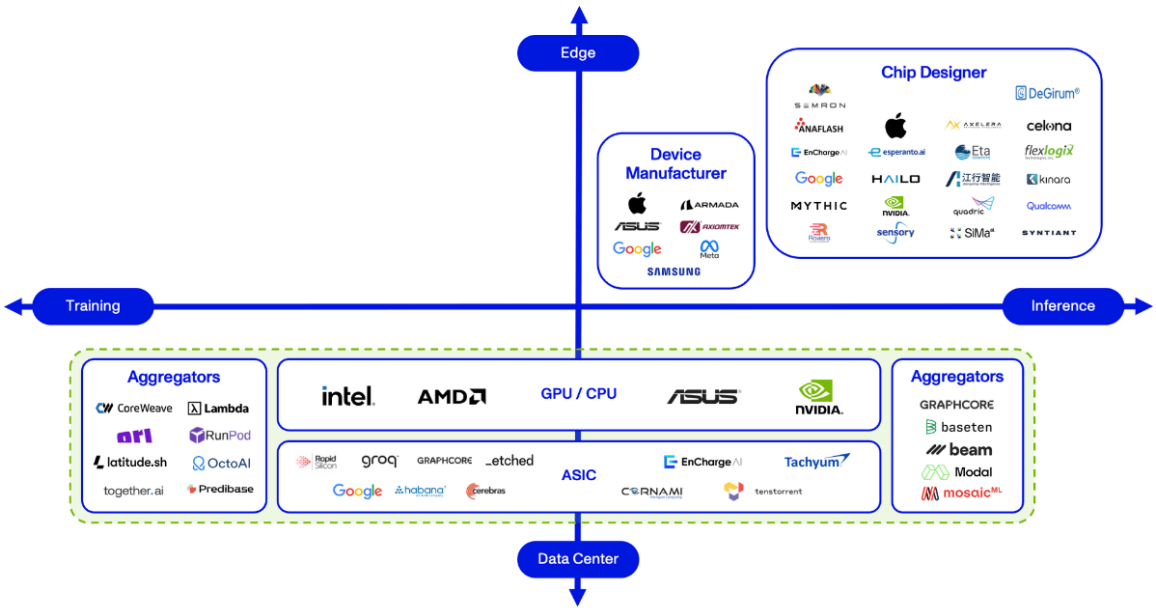
The industry's dependency on NVIDIA is worsened by the scarcity of compute and the dominance of NVIDIA at several points in the supply chain. Pictured above are the different players who are wholly dependent on the availability of NVIDIA's H100 GPU, the industry-leader in terms of performance and efficiency. Access to H100s is a meaningful advantage for AI firms and the supply of the chips is limited by the fact that TSMC is the only supplier capable of producing them. NVIDIA's H100 is in such high demand that NVIDIA has preferentially provided its chips to smaller players to avoid further concentrating its chips in the hands of a few large firms.

NVIDIA's market dominance extends beyond how powerful their GPUs are. NVIDIA has been laser-focused on deep-learning and AI since 2012 and has developed products to support AI model training on its platform. NVIDIA's CUDA is among their most notable innovations. The software allows developers to orchestrate the training of their models in parallel across many GPUs.

NVIDIA's deep ties into the AI hardware supply chain, first-mover advantage and highly specialized applications for niche high performance computing tasks have made NVIDIA the de facto standard for AI research and commercialization. **NVIDIA currently saturates the data center market,** and any challenger must contend with not only a deeply entrenched highly performant incumbent but also the non-trivial switching costs associated with moving to a new software and hardware stack.

ACTIVANT *Market Map*

AI Hardware



We describe the AI hardware market along the axes of "where the computation is taking place" (y–axis) and "what computation is taking place" (x-axis). We have included a green rectangle that circumscribes what we believe to be a market dominated by NVIDIA, with entry requiring head-on competition with the incumbent.

The only part of the market that can support new entrants is edge inference, i.e., hardware that allows for the inference of AI models at the location where the input data is being produced as opposed to at a remote server location. Often, data produced at industrial sites is either (1) too large to transport to remote server cost-effectively or (2) requires real-time processing to be of any use. A relevant example of the latter would be a self-driving car while the former could include the data from a drilling rig.

According to OpenAI, Inference is currently the primary use of their computing resource.[5] As the overall demand for generative AI applications grows year over year, we can expect the majority of the operating expense of gen AI models to become inference. Shifting the computation onto the end-user's device is cost-advantageous for an AI firm, and, provided that the device in question is capable of low-power edge inference, this helps to reduce AI's carbon footprint.

Ultimately, we believe edge inference players can overcome NVIDIA's dominance by selling directly to specific end customer profiles. Those customers need to leverage data in time, in place, and may have never previously leveraged AI solutions before.

We have identified several notable companies building in this space.

- Armada provides an edge computing platform that leverages ruggedized hardware and a portfolio of analytical tools. The company works with Starlink and operates a resilient, secure, and scalable infrastructure with a matching software component, providing users with a single control plane to manage and optimize data efficiently.
- Kinara develops low-power processors designed to deploy AI applications on video cameras or other devices. The company's processor offers a set of automated development tools to support the implementation of complex, streamlined AI applications, enabling users to get rich data insights to optimize real-time actions at the edge.
- Mythic develops integrated circuit technology designed to offer desktop-grade graphics processing units in an embedded chip. The company's technology utilizes in-memory architecture to store neural networks on-chip. This limits the need to shift data on and off the chip during inferencing.
- Etched.ai is creating the first-of-its-kind transformer architecture on the chip. If anything, Etched is among the very few companies that we would consider as exceptions to the rule listed above: they may disrupt NVIDIA.

After several months of work and dozens of interviews with individuals from ASML to AWS, we are confident that an opportunity exists in edge inference. Our original hypothesis that NVIDIA could be disrupted by a new competitor failed to account for NVIDIA's standing relationship with ASML itself. The hypothesis evolved and with it came the realization that hardware supply chains – not the chip designs themselves – are the greatest moat in this industry. If you're building the future of edge inference hardware and want to discuss supply chains, or costs and sales, reach out to Marc. We'd love to chat!

Endnotes

1 Estimated Loss herein refers to an approximation derived in: Hoffmann et. al, Training Compute-Optimal Large Language Models, 2022

2 Department of Energy, How much power does a nuclear reactor produce?, 2021

3 Epoch AI, Trends in GPU Price-Performance, 2022

4 Thompson et al, The computational limits of deep learning, 2022

5 Open AI, AI and Compute, 2018