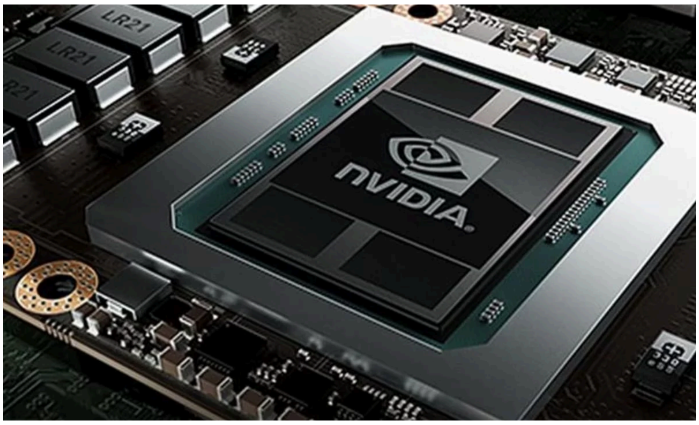
Edit article
View stats
View post



The Technical Risks to NVIDIA's MarCap are Fundamental

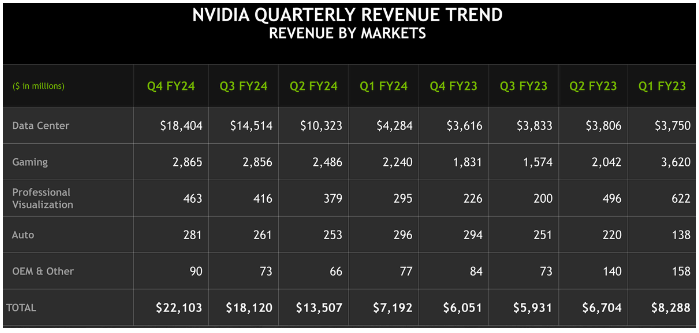
Jen Zhu Scott
Multifamily | Founding Partner | TED Speaker | Board Chair | Perpetual Student of Neuroscience, AI, Math, and History
2 articles

February 29, 2024

Open Immersive Reader

NVIDIA's astonishing ascend to a recent $2 trillion market cap makes it one of the most valuable companies in history. However, certain upcoming technical risks to the company might not be obvious to everyone but fundamental.

NVIDIA's market cap (and revenue) rise is mainly attributed to its success in the AI industry, which now impacts every other industry. The company's parallel processing capabilities, supported by thousands of computing cores, have contributed to its success in the GPU market. NVIDIA's focus on high-performance computing, gaming, and virtual reality platforms has also helped drive its market value.



NVIDIA QUARTERLY REVENUE TREND
REVENUE BY MARKETS

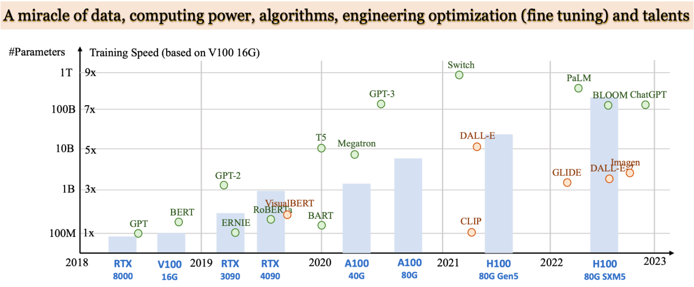
($ in millions)	Q4 FY24	Q3 FY24	Q2 FY24	Q1 FY24	Q4 FY23	Q3 FY23	Q2 FY23	Q1 FY23
Data Center	$18,404	$14,514	$10,323	$4,284	$3,616	$3,833	$3,806	$3,750
Gaming	2,865	2,856	2,486	2,240	1,831	1,574	2,042	3,620
Professional Visualization	463	416	379	295	226	200	496	622
Auto	281	261	253	296	294	251	220	138
OEM & Other	90	73	66	77	84	73	140	158
TOTAL	$22,103	$18,120	$13,507	$7,192	$6,051	$5,931	$6,704	$8,288

Image credit: NVIDIA

Suppose the current trajectory of AI development (driven by Generative AI based on Large Models) continues. In that case, NVIDIA will remain one of the most valuable companies, if not the most valuable, in the world. In other words, NVIDIA wins as long as absolute capacity to computing power is what is running the world. The assumption ignores the other side of the equation: efficiency. I share my arguments below. This article aims to spark healthy debates, inspire more good questions, and not serve as investment advice.

I. LARGE MODELS ARE NOT NECESSARILY THE FUTURE OF AI

GenAI based on LLMs is an incredible breakthrough in AI. The 'largeness', since the monumental 2017 paper **"Attention is All You Need"** that conceived Generative AI, has resulted from the enormous amount of data, computing power, algorithms, capital, and talents. It's a human engineering marvel that reduced human-machine interaction to natural language for the first time, hence OpenAI's ChapGPT's historical pace of user generation. But I don't believe the 'largeness' is the future. In fact, I prefer it is not.



A miracle of data, computing power, algorithms, engineering optimization (fine tuning) and talents

Image Credit: Linear Ventures

It is simple: we don't get on Mars by building taller and taller buildings on Earth - Mars being general purpose AI, or AGI. Currently, an LLM trained with several hundred billion parameters would qualify it in a tiny minority club globally in absolute capacity. **SambaNova System**, an enterprise GenAI company based in Palo Alto, just announced their **1 trillion parameter LLM**. Impressive - yet another taller building to reach to Mars. This planet's most complex and efficient general intelligence still operates inside your skull. A human brain, on average, processes 100 trillion parameters using 30-watt, only enough to power an average incandescent light bulb. Training a large language model like GPT-3, for example, is **estimated** to use just under 1,300 megawatt hours (MWh) of electricity, about as much power as is **consumed annually** by 130 US homes. In context, streaming an hour of Netflix requires around **0.8 kWh** (0.0008 MWh) of electricity. That means you’d have to watch 1,625,000 hours to consume the same power to train GPT-3.

The inefficiency is not only measured by the absolute consumption. The LLMs are still black boxes to us because we can't tell specifically which set of data really made a difference to allow the algorithm to 'understand', reason, and generate. So, the current approach is to include more and more data and train on more and more parameters, which means more and more AI chips from NVIDIA.

II. THE NEXT AI FRONTIER IS ALL ABOUT EFFICIENCY

Pedro Domingo, a Professor Emeritus of Computer Science & Engineering at the University of Washington and the author of **The Master Algorithm**, recently published an important paper, **"Every Model Learned by Gradient Descent Is Approximately a Kernel Machine"** that everyone who is half fluent in the language of mathematics should read carefully.

In a nutshell, solving the efficiency problem isn't as farfetched as it seems. Despite its many successes, deep learning remains poorly understood. In contrast, kernel machines are based on a well-developed mathematical theory, but their empirical performance generally lags behind deep networks. Gradient descent is the standard algorithm for learning deep networks and many other models. The paper shows that every model learned by this method, regardless of architecture, is approximately equivalent to a kernel machine. This kernel measures the similarity of the model at two data points in the neighborhood of the path taken by the model parameters during learning. Kernel machines store a subset of the training data points and match them to the query using the kernel. Deep network weights can thus be seen as a superposition of the training data points in the kernel's feature space, enabling their efficient storage and matching. Such results have significant implications for boosting algorithms, probabilistic graphical models, and convex optimization.

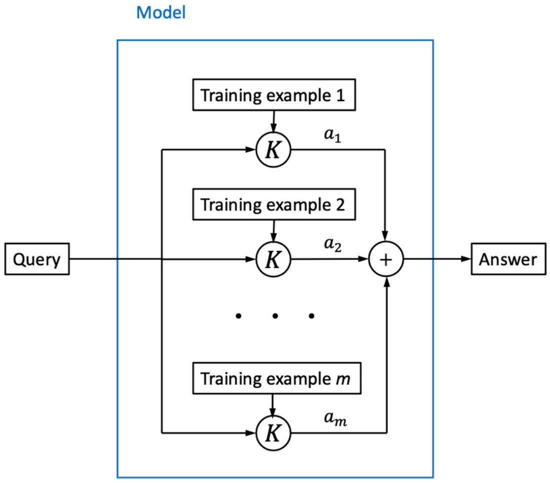


Figure 2: Deep network weights as superpositions of training examples. Applying the learned model to a query example is equivalent to simultaneously matching the query with each stored example using the path kernel and outputting a weighted sum of the results.

Image Credit: Pedro Domingos

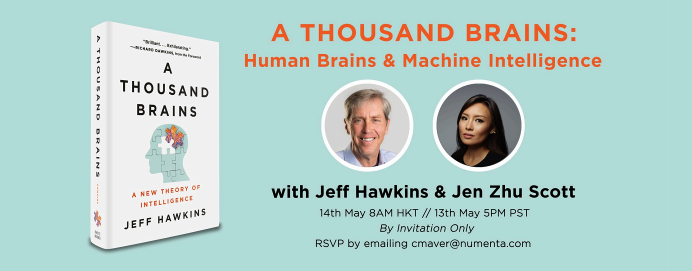
If the math in the paper is too thick to go through, you would then only need to understand this paragraph:

"Most significantly, however, learning path kernel machines via gradient descent largely overcomes the scalability bottlenecks that have long limited the applicability of kernel methods to large data sets. Computing and storing the Gram matrix at learning time, with its quadratic cost in the number of examples, is no longer required. (The Gram matrix is the matrix of applications of the kernel to all pairs of training examples.) ***Separately storing and matching (a subset of) the training examples at query time is also no longer necessary****, since they are effectively all stored and matched simultaneously via their superposition in the model parameters. The storage space and matching time are independent of the number of examples. (Interestingly, superposition has been hypothesized to play a key role in combatting the combinatorial explosion in visual cognition (Arathorn, 2002), and is also 8 Deep Networks Are Kernel Machines essential to the efficiency of quantum computing (Nielsen and Chuang, 2000) and radio communication (Carlson and Grilly, 2009).) Further,* ***the same specialized hardware that has given deep learning a decisive edge in***

scaling up to large data (Raina et al., 2009) can now be used for kernel machines as well."

Pedro is not the only one marching ahead to solve the efficiency problem. Jeff Hawkins, the inventor of Palmtop and author of On Intelligence and A Thousand Brains (one of my favorite books of all time, and I had the honor to help spread the word), has studied human brains his entire career. Jeff's AI application/research institute, Numenta, has been publishing its research on sparsity since 2021.

Less data, less computing power, better results.


"Brilliant.... Exhilarating."
—RICHARD DAWKINS, from the Foreword
A THOUSAND BRAINS
A NEW THEORY OF INTELLIGENCE
JEFF HAWKINS
A THOUSAND BRAINS:
Human Brains & Machine Intelligence
with Jeff Hawkins & Jen Zhu Scott
14th May 8AM HKT // 13th May 5PM PST
By Invitation Only
RSVP by emailing cmaver@numenta.com

Dialogues like this helped me stay sane during COVID.

There are external factors that will drive and accelerate the efficiency pursuit in addition to the obvious financial potential:

- **environmental concerns**: the energy consumption of the current LLM approach is not sustainable, especially how many nonsense queries go into GenAI models and how much plausible bullshit ChatGPT and Gemini generate.
- **geopolitical factors:** China can't currently access the most advanced AI chips due to the export ban by the US. The semiconductor sector in China is catching up at an astonishing speed, with heavy capital and policy support from the Central Government. But there are two ways to catch up with the US: build and scale your own advanced AI chips or work on the efficiency problem so you don't require the same level and quantity of advanced AI chips to get where you want to go. China is running fast on both.

Higher efficiency, less chips.

III. DECENTRALISED GPUS

Billions of GPUs are sitting idle in this world. The GPUs in your smartphones and laptops are not fully utilized. OTOY is the special effect tech company that enabled movies like Curious Case of Benjamin Button, Bladerunner 2049, and Star Wars. Humans perceive high-quality special effects, VR/AR, as realistic because every shadow, reflection, and movement is individually calculated. The team at OTOY constantly ran into capacity constraints. In 2017, the brains behind OTOY started **The Render Network** to use blockchain to enable idle GPU marketplace to utilize the untapped computing power. I was on the initial Advisory Board with JJ Abrams, Ari Emanuel, and Beeple. The Render project has come a long way, though it's still not mainstream. The idle GPUs are a giant goldmine representing enough incentives to attract mainstream solutions sooner or later.

More decentralized GPUs, less reliance on centralized GPU providers.

IV. CONCLUSION

I don't, for a second, pretend that NVIDIA's momentum will slow down soon. But if you are not paying attention to the above facts and trends, you shouldn't allocate any capital in this space. I admit some of the above trends are still early, and NVIDIA has the talents and war chest to be more future-proof. But assuming it remains on the current absolute capacity path, the risks are indeed fundamental, and the incentives and restraints are so real that one morning, we just might wake up and realize that one of them caught fire.

Published by

Jen Zhu Scott
Multifamily | Founding Partner | TED Speaker | Board Chair | Perpetual Stu...
Published • 2mo

2 articles

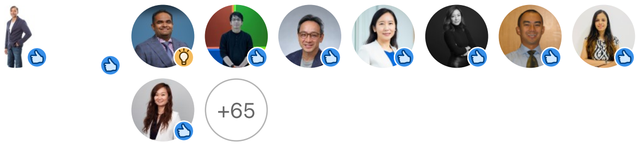
Largely driven by #GenAI built on #LargeModels, NVIDIA's recent brief touch on the $2 trillion market cap made it one of the most valuable companies on this planet. However, some of the technical risks in the path ahead are fundamental. I break them down in my following article.

Stephen McAlinden Raleigh Addington you both recently told me off for not writing more. So here is some red meat. CC'ing Marc Spenlé Huai (Harry) Wang bobby vedral V. Bunty Bohra Dr. Alex Yang who might be interested.

Like Comment Share

Bruno Sanchez-Andrade Nuño and 76 others • 33 comments

Reactions



33 Comments

Most relevant ▾

Add a comment...

Ken McMahon • 1st 2mo •••
Adding value with innovative next-generation products

Excellent summary which systematically lays out the instinctive thoughts I had about this opportunity! How will demand change and how will competition develop? And how fast will those changes happen...

Like · 👍 3 | Reply · 1 Reply

Ken McMahon • 1st 2mo •••
Adding value with innovative next-generation products

Just look at Mistral, for example! 😺

Like · 👍 1 | Reply

Bruno Sanchez-Andrade Nuño • 1st 2mo •••
AI for Earth

Loved this nuisance and deep article Jen Zhu Scott. I've learnt a lot! And got both book and an article to keep learning ❤️
On the topic, this also assumes that NVIDIA will retain its performance lead gap. The more they grow, the more incentive for others to invest. Just like Google made TPUs and Grok LPU, I expect more sp ...see more

Like · 👍 5 | Reply · 1 Reply

Jen Zhu Scott • You 2mo (edited) •••
Multifamily | Founding Partner | TED Speaker | Board Chair | Perpetual Student of Neuroscience, AI, Math, and History

Bruno Sanchez-Andrade Nuño that's a brilliant point Bruno! Thank you so much. The GPU poor crowd will find the next trillion dollar opportunities 🚀🚀🚀🚀

Like · 👍 1 | Reply

Show 10 more comments

Jen Zhu Scott

Multifamily | Founding Partner | TED Speaker | Board Chair | Perpetual Student of Neuroscience, AI, Math, and History

More from Jen Zhu Scott

Meta Llama 3: A Deep Dive & Zuck's Long Game

Jen Zhu Scott on LinkedIn