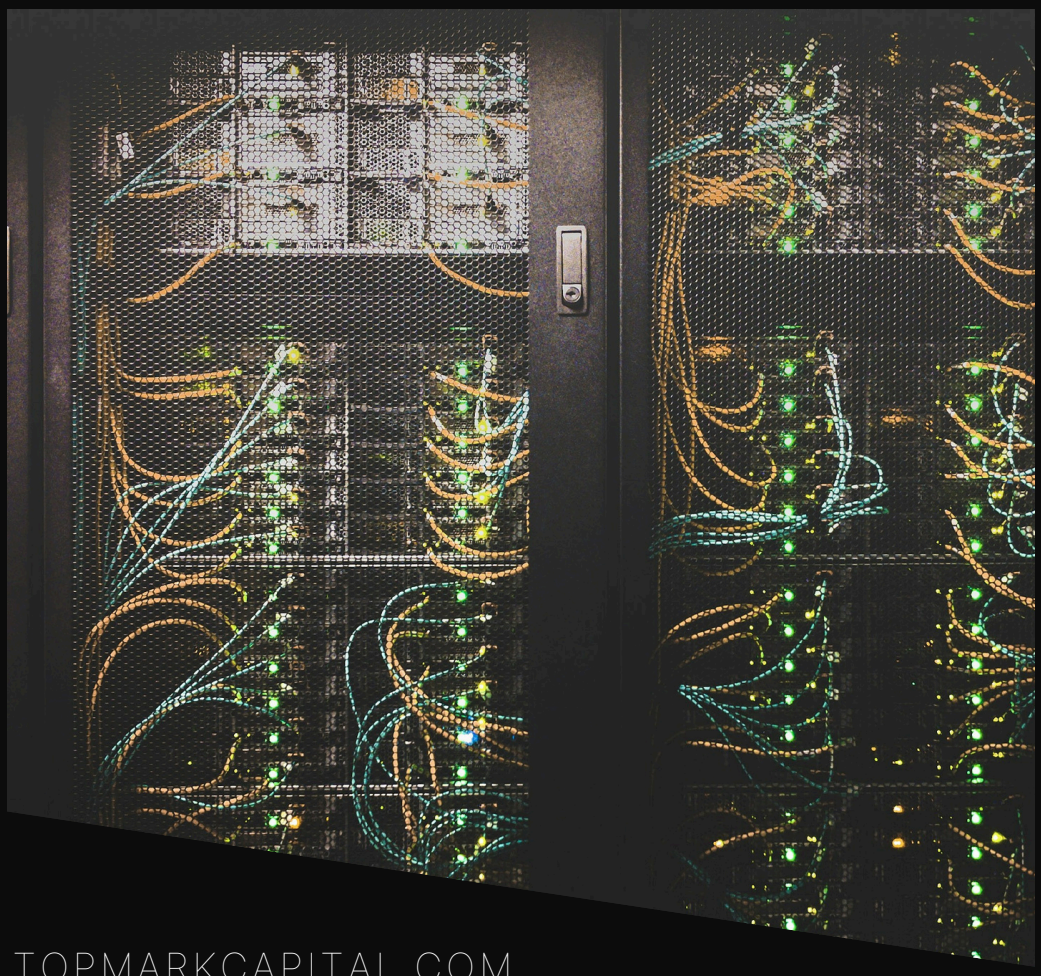
TOP MARK CAPITAL

HYPERSCALE CAPEX &

NVIDIA'S 5 RISKS



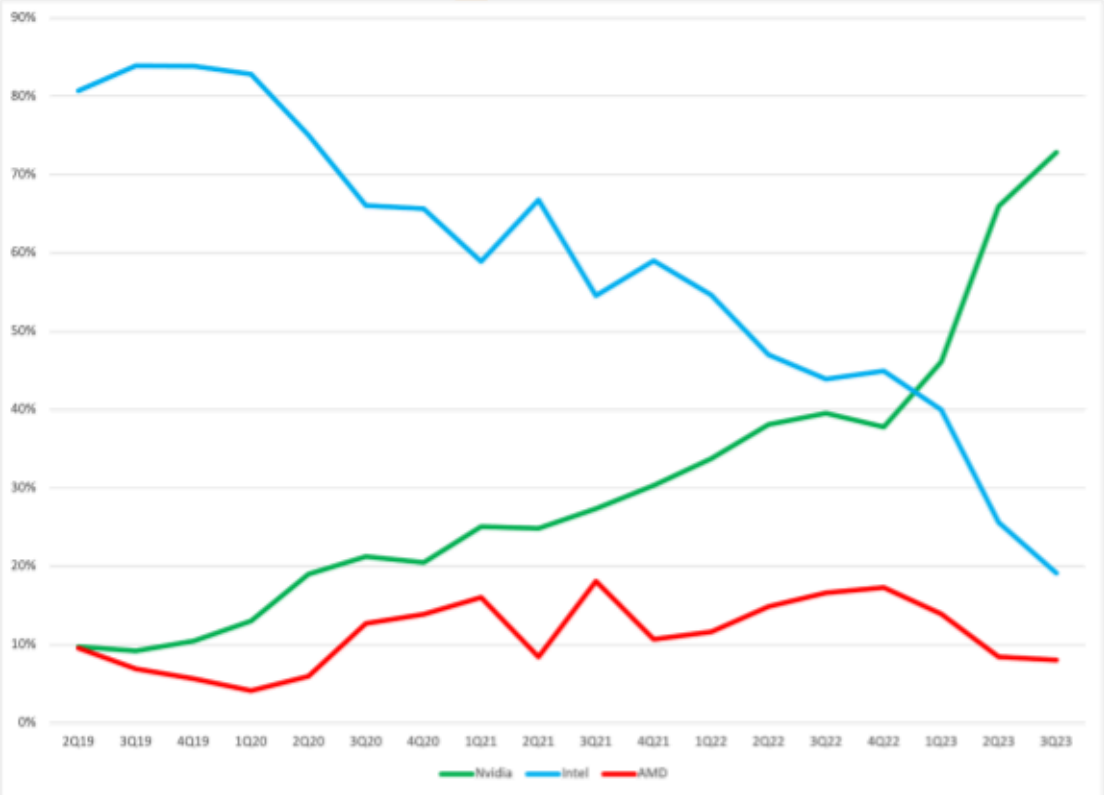
TOPMARKCAPITAL.COM
416 13th St.
2nd Floor
San Diego, CA 92101

HYPERSCALE CAPEX AND NVIDIA'S 5 RISKS

Theme: Software is Eating the World

Frequent readers of our letters will recall that we have been beating the drum on the prospects for artificial intelligence (AI) to reshape the software industry since 2015. Today, generative AI is at the cutting edge of this trend, and Nvidia's jaw dropping cash flow growth has stunned even us, its most ardent supporters. Nvidia represents at least ¾ of data center market[12], captures ~50% of total hyperscaler capital expenditures[13], and

Figure 1: Data Center Market Share



Source: Digits to Dollars

[12] D2D Advisory, "No Going Back."

[13] Note that Meta is the only one of these companies that is not a public cloud provider.

achieves margins in excess of 75%. Needless to say, many investors that missed the trend are scratching their heads.

"Trees don't grow to the sky."

- German Proverb

Trees don't grow to the sky. Or do they? Nvidia reports earnings next week and is expected to hit a run rate of $100 billion in data center revenue. **This article will assess the state of data center spending, through the lens of the hyperscalers** (Amazon, Google, Meta, Microsoft, and Oracle), **and suggest our top five risks to Nvidia's dominance in this market**.

"Software is eating the world" is the phrase we adopted to represent the long term trend of software encompassing all aspects of our professional and personal lives[14]. The shift from CPU based computing to GPU based computing fits squarely within this trend, and artificial intelligence is its largest accelerant. Nvidia CEO and co-founder Jensen Huang, has suggested that his vision for this shift in workloads dates back to the company's founding in 1993. But, back then, and for the next 25 years or so, Intel's CPUs would dominate the industry with market share figures similar to those of Nvidia today.

Traditional CPUs execute tasks sequentially using a single processing core. This is very effective for sequential programs with interdependencies. But for large tasks that can be divided into smaller sub-tasks, parallel computing increases overall processing speed and efficiency by spreading the work across multiple cores. Graphical rendering,

[14] Andreessen, "Why Software Is Eating the World."

which requires individual calculations for each pixel on a screen, was an early use case - hence the name Graphical Processing Unit or GPU. The computing shift from CPU to GPU wouldn't gain much steam until 2012 when AlexNet, a convolutional neural network written with CUDA[15], running on a Nvidia GPU, trounced the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[16]. ILSVRC pitted software programs against each other, to see which was most capable of correctly classifying objects and scenes in images. Many of the leading minds in the field of artificial intelligence can trace their experience, in one way or another, back to this AlexNet project.

Through the 2010's, Nvidia, and notably CUDA, chipped away at the CPU market, shifting workloads that were entirely impossible or exorbitantly expensive to run on CPUs, like weather forecasting, fluid dynamics simulations, and other scientific problems. AI and machine learning (AI/ML) were also major factors in this steady climb. Meta's recommender system, for example, predicts user behavior and ranks content in real-time, across the company's "Family of Apps". It is likely the most profitable AI workload in production today. But it wasn't until OpenAI released its large language model (LLM), ChatGPT, in November 2022, that the world woke up to the power of "AI" models. This triggered data center spending shifted quickly to Nvidia as the core of the data center shifted to GPUs rather than CPUs (see Figures 1 and 2).

Inference - the actual running of these LLMs - is the fastest growing parallelized workload today. These include ChatGPT, Claude, Gemini and other 'assistants', as well as APIs that OpenAI, Anthropic, Google, and others provide for developers to incorporate into their products. Walmart, for example, utilizes generative AI to

[15] CUDA is Nvidia's proprietary software language used to program the company's GPUs

[16] AlexNet was not the first AI model to run on GPUs, but it was the most notable given the *relative* fame of the ImageNet competition. All other competitors were running CPU based models.

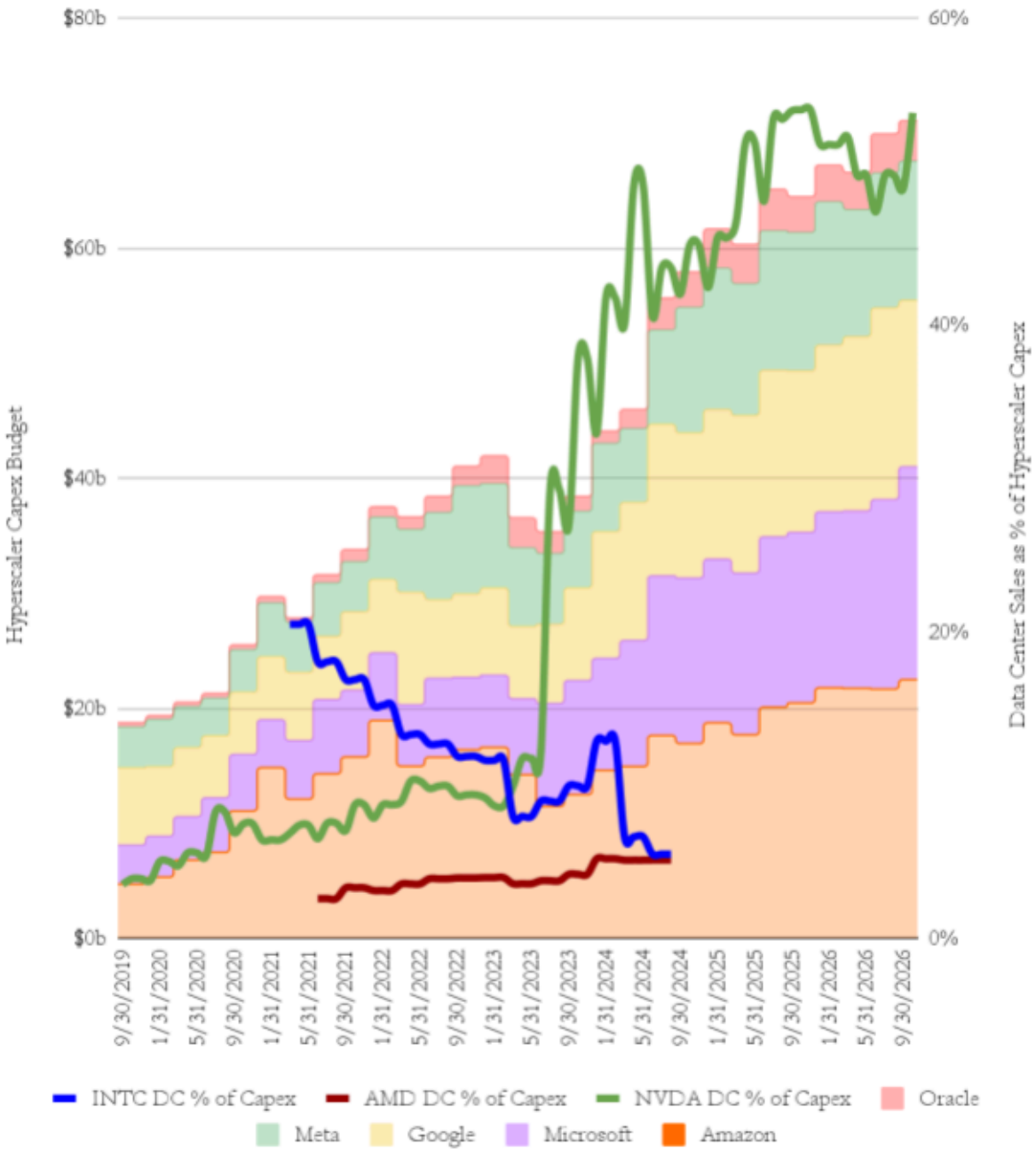
dynamically edit product descriptions. Some companies, like Goldman Sachs, are using a combination of open source models (like Meta's Llama) and Nvidia hardware to develop custom solutions using proprietary data[17]. Regardless of the implementation method, these LLMs act like a computer version of the human brain. They break down data into tiny pieces, analyze them in hidden layers, and make predictions based on what they've learned.

Training these models is the highest value and most compute-intensive task. Importantly, training heavily influences future hyperscale capital spending. Anthropic's CEO, Dario Amodei, says large language models in development today can cost up to $1 billion to train, up from *only* ~$100 million for ChatGPT-4o. He expects the cost to balloon to $10 or even $100 billion within three years for cutting edge models[18].

[17] "Goldman Sachs CIO on How the Bank Is Actually Using AI."

[18] Morales, "AI Models That Cost $1 Billion to Train Are Underway, $100 Billion Models Coming — Largest Current Models Take 'only' $100 Million to Train."

Figure 2: Hyperscale Capex & Chipmaker Share



The Hyperscaler Investment Decision

For hyperscalers to justify the costs associated with investing in the chips and infrastructure required to support each successive generation of models, they will need to continue to see an attractive return on investment (ROI) and a short payback period. Colette Kress, CFO of Nvidia made this abundantly clear on the company's most recent earnings call:

> *"Training and inferencing AI on NVIDIA CUDA is driving meaningful acceleration in cloud rental revenue growth, delivering an immediate and strong return on cloud provider's investment."*

So what is the ROI and payback period for a Nvidia H100 cluster for a hyperscaler? Jensen Hwang has suggested that hyperscalers can expect a 400% ROI over the course of five years. A more thorough analysis suggests that Microsoft, for example, expects an internal rate of return (IRR) between 38% and 110% over five years, and a payback period between 10 months and 2.1 years. The variation depends on if the usage is 'pay as you go' or on a five year fixed contract[19].

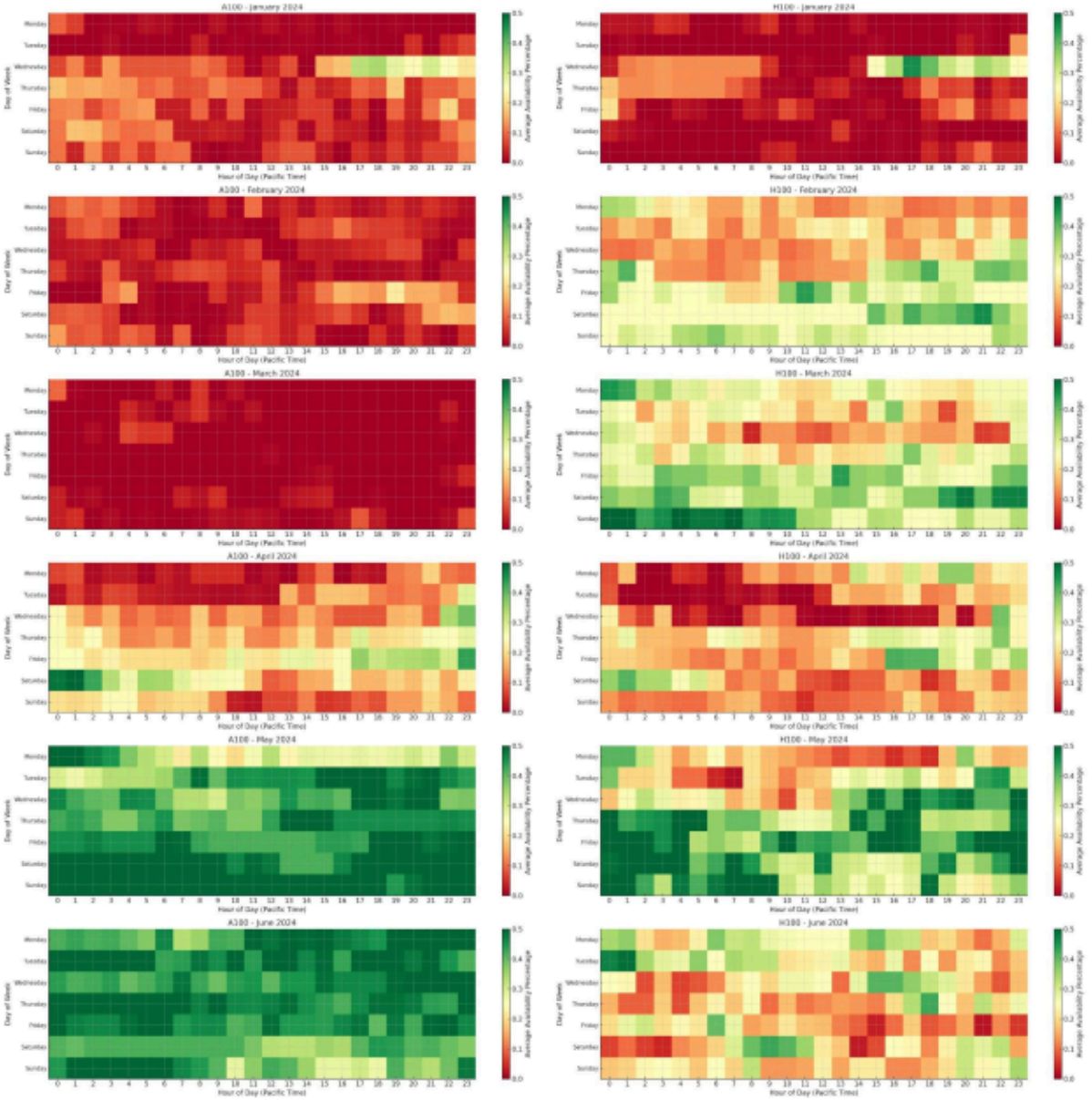
Demand ultimately drives hyperscaler purchasing decisions. GPU availability at public clouds can give us indications as to how much demand there is for GPUs. Figure 3, on the following page, shows A100 demand moderately easing as H100 capacity started to come online. What is quite remarkable is that the A100 first started to ship in May 2020! This means, four years into its lifecycle it was still practically impossible to get access to one. Every hyperscaler with A100s likely earned phenomenal ROIs.

The public cloud hyperscaler investment decision will be heavily weighted toward payback period. Having line of sight on the payback period creates a "heads I win, tails I

[19] "Return on Investment (ROI) on AI for Cloud Hyperscalers."

don't lose" scenario. If average payback periods creep much beyond two years, I would expect that the CFOs of the hyperscalers will tighten the reins on capital spending. This also means that so long as there is demand for GPUs, all public cloud hyperscalers will want as many Nvidia GPUs as they can get their hands on.

Figure 3: Mean GPU Availability Percentage by Hour/Day/Month



Nvidia's 5 Biggest Risks:

Today, Nvidia dominates not only the market for GPUs, but also nearly everything inside the data center, thanks to its acquisition of Mellanox. The company's enviable position today is largely dependent on Jensen Hwang's vision for the shift from CPU to GPU based computing that lead him to invest in CUDA, Nvidia's programming language that developers use to code parallelized applications that run on Nvidia GPUs[20]. CUDA gave Nvidia a 10+ year lead on competitors and created network effects in that every parallelized computing programmer has used CUDA. Defending this position into the future will depend on Nvidia maintaining its competitive advantage versus a litany of would be competitors, while simultaneously navigating the natural ebbs and flows that result from the underlying applications running on the company's GPUs. Below we list and detail our top five risks to Nvidia's dominance. Since the hyperscalers are ultimately the primary purchasers of these data centers, our analysis will be through the lens of a hyperscaler. Risks 1 and 2 cover the supply side decisions hyperscalers face, and risks 3-5 cover the demand side, which the hyperscalers will have to estimate.

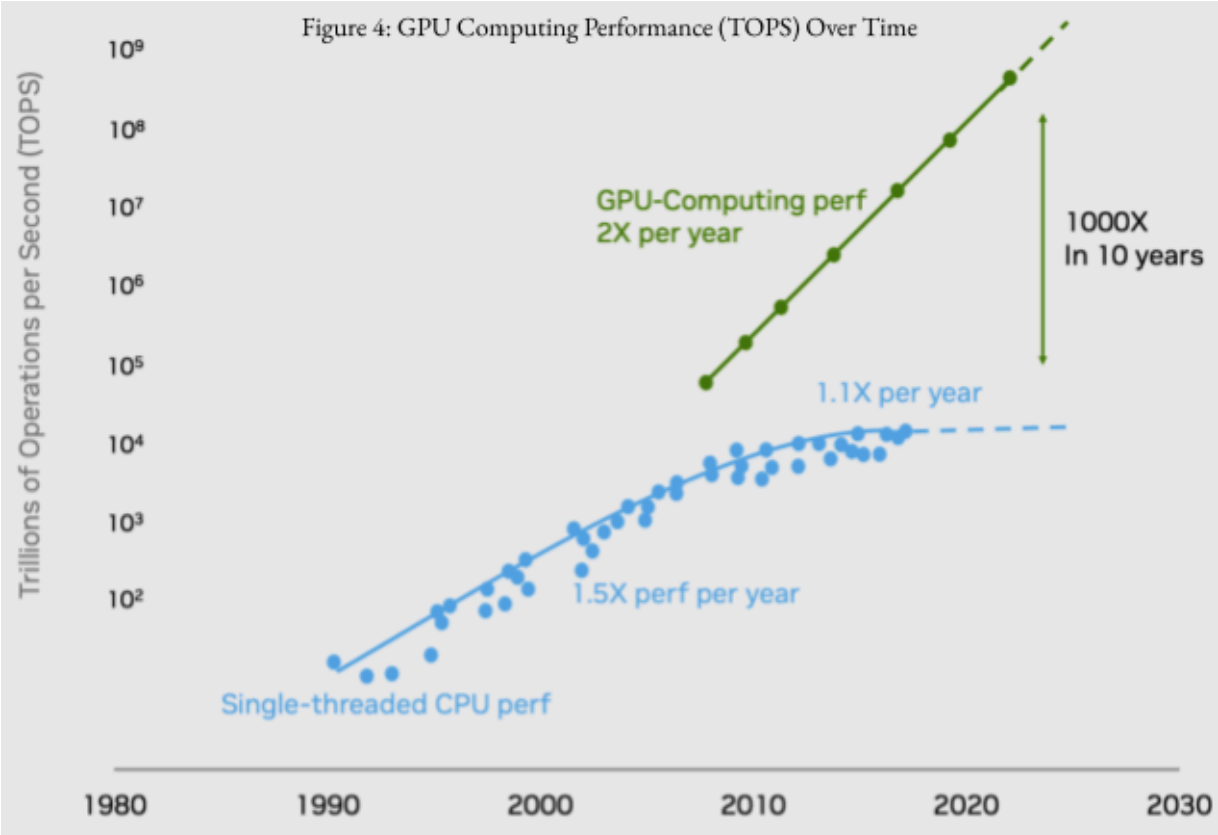
1. Hyperscalers Overestimate Demand

The case of investing in a Nvidia GPU cluster has been a very easy decision because the payback period is so short (<2 years, assuming the aforementioned assessment of a Nvidia H100 cluster at Microsoft is remotely accurate). When these servers are new, more of the resources are dedicated to training, therefore having line of sight on demand for training enables hyperscalers to confidently make this investment decision. This confidence breaks down if the hyperscaler doesn't have line of sight on the demand

[20] "NVIDIA's CORNERSTONE for UNMATCHED DOMINANCE in AI (CUDA)."

needed to cover its cost. Lower utilization and less confidence in the first two years of the forecasted demand will make this investment decision less obvious.

Training is the primary driver of adoption when new GPUs hit the market due to the level of compute intensity required to train frontier models. As those models are



Source: Nvidia Company Overview February 21, 2024

deployed into production, the inference process can run on the same GPUs[21]. Estimating this inference demand is much harder, given that we don't even know the capabilities of the next generation of models[22]. As the supply of leading edge Nvidia GPUs normalizes, hyperscalers may be tempted to overinvest in this equipment and accept lower rates of return. Finance departments at every hyperscale public cloud look

[21] There are of course exceptions to this rule, but in general this is what we are seeing

[22] One could argue that we don't truly understand the capabilities of the current frontier models.

at their investment decisions similarly. They estimate the cost of the project, then estimate the cash flows expected over the life of the investment. Typically, they have an internal hurdle rate that a project should meet. Since estimates in years 3-5 greatly influence the ending IRR, there is potential that hyperscalers could get too aggressive with their GPU buildouts. If these investments fail to meet prior target hurdle rates, hyperscaler finance departments may become more conservative in adopting the next generation of GPUs.

2. Hyperscalers are Overestimating the Useful Life of GPU Servers

Historically, servers are depreciated in a straight line basis over a three year period. However, starting in 2020, hyperscalers began stretching their useful life to as much as six years. This made sense. Intel had fallen off the leading edge and successive generations of CPUs were only marginally better than the prior generations. Moore's law was dead. However, as documented in this memo, hyperscale capex has shifted to Nvidia GPUs, which are increasing in performance at almost twice the rate of CPUs. Meaning, Nvidia's innovation cycle may render its older GPUs worthless faster. The implications of this would likely mean higher prices for cloud based computing, and cause

3. The Performance of Successive Generations of Frontier Models Do Not Scale

Successive generations of LLMs are generally improved today by increasing the amount of data (tokens) used to train the model and the parameters (complexity) of the model. As a result, new generations models are dependent on new generations of Nvidia hardware (A100 -> H100 -> B200). But, if performance increases of successive generations of models begin to taper, then we can expect that the appetite for investors to finance the costs associated with training new models will wane. At least until a new

approach arises that significantly moves the ball forward. This is a risk to Nvidia's data center sales, as AI could hit a proverbial "brick wall".

One should also remember that Moore's law was not a law at all. It was a voluntary challenge that Intel managed to meet year after year, until 2016 when it struggled to implement 10 nanometer technology[23]. Performance scaling of generative AI models will similarly be 'optional' and driven by the few companies that are developing foundational models. Improving capabilities are a necessity for Nvidia to maintain its share of data center, let alone expected growth.

What would cause the performance of successive generations of models to taper? The primary concern today is data availability. Successive generations of Large Language Models (LLMs) have shown that increasing the size of the training data set also increases the quality of the model output. Today's models are like giant vacuum cleaners, sucking up every bit of data they can. Web 2.0 companies are quickly realizing that they hold significant value for these large language models, hence the $60 million per year license fee Google is paying Reddit. Expect to see more deals like this, but don't underestimate the long term implications of this shift. In one sense, the internet as we know it today - the free and open internet - is dying. Interestingly, the sheer volume of data on the free and open internet is what made training these models feasible in the first place. Innovation in software, like synthetic data generation, transfer learning from data-rich domains, and data efficiency improvements are seen as potential solutions to overcome these data limitations. Fortunately, the total amount of data in the world doubles every two years, so we can expect that there will be lots of new information to train new models on. We hope that future AI models will have access to this data.

[23] Woods, "The Death of Moore's Law."

4. Hardware is No Longer the Limiting Factor for AI Scaling

Assuming that data availability and hardware capability (see 3 above) are not barriers to AI scaling, what other factors may hinder AI scaling? There are many, but we will focus on two here: training efficiency and power availability.

Training efficiency is the amount of compute and training time needed to achieve superior model performance versus the current state of the art. Increasingly powerful compute resources can mask inefficient software, as it did for Intel so long as Moore's law maintained its pace. Only when Moore's law started to break down (and total cost of ownership became increasingly important) did ARM based RISC CPU architectures become competitive. Are there opportunities for improving the software related to training that would render as good or better results than the current transformer architectures? It is fairly well documented that there are significant opportunities for improvement in the efficiency of AI software, but it is hard to ignore the history of Intel and Moore's law. We would argue that we are most likely to see efficiency innovations in AI software come from China. China is limited in its ability to adopt the latest semiconductor technology due to export restrictions. These export restrictions create a strong incentive for innovation within the Chinese domestic chip industry, but they also create an incentive to eke out improvements in other areas of the AI stack - notably software. These circumstances put China in a position where scarcity (of compute) may breed innovation[24].

A second factor that may hinder AI scaling is power availability. New data centers are facing significant delays of up to 5 years in getting power interconnections due to transmission infrastructure constraints, massive interconnection queue backlogs, and surging data center power demand. While some data centers are exploring alternative

[24] Notably, this creates the exact opposite effect intended by those who implemented the export restrictions

solutions like on-site generation[25] and co-locating with power plants[26], major transmission upgrades and interconnection process reforms will be critical to support the industry's long-term growth. The rapid increase in data center energy consumption increases risks associated with already problematic grid reliability. These issues are already becoming a major bottleneck for new data centers, and could ultimately hinder Nvidia's growth.

5. Competition

Nvidia currently dominates the data center market (see Figures 1 and 2). There are a litany of competitors that are attempting to gain an edge. AMD, Nvidia's long time competitor in the gaming GPU market, has released the MI300X GPU, Intel has launched the Gaudi 3 GPU, and Microsoft, Google, Amazon, and Meta are developing in house chips. Nonetheless, customers still heavily favor the H100 and only resort to alternatives if they cannot get access to H100s. These other products are currently not competitive with Nvidia. Nvidia has established process leadership and scale, enabling them to own practically all of the market. They also have strong network effects due to the fact that the entirety of AI software to date has been developed on CUDA. For Nvidia to maintain its dominance, it needs to maintain a rapid pace of innovation, releasing successive generations of chips that are not only better than previous generations, but also well ahead of their competition in the same way that Intel was able to dominate the x86 CPU market for decades (see figure 4).

A final, and frankly more likely, though longer term, competitive threat would come from an entirely new paradigm shift. In the same way that the CPU gave way to the

[25] Morales, "Elon Musk Powers New 'World's Fastest AI Data Center" with Gargantuan Portable Power Generators to Sidestep Electricity Supply Constraints."

[26] Fitch, "Why the AI Industry's Thirst for New Data Centers Can't Be Satisfied."

GPU due to the transition of workloads, quantum computing is widely seen as the next potential disruptive technology in the data center market. Quantum computers leverage principles of quantum mechanics like superposition and entanglement to perform certain computations exponentially faster than classical computers. Many computationally demanding problems don't parallelize efficiently due to sequential dependencies and communication bottlenecks between parallel processors. Quantum algorithms can perform certain complex computations with far fewer operations, reducing latency and energy consumption compared to massively parallel classical solutions. Problems with high combinatorial complexity, like optimization, molecular simulation, and machine learning, could see transformative speedups with quantum computing. However, quantum computing still faces many challenges in scaling qubits, improving error correction, and integrating with classical infrastructure. Nonetheless, the technology continues to advance, and it represents both an opportunity and a competitive threat that Nvidia will need to navigate strategically. With that said, Jensen's vision being forever forward looking, has pushed Nvidia to already begin work on CUDA-Q, their library for quantum workloads[27].

[27] "CUDA-Q."

Takeaways

Before we close, I'd like to draw your attention back to our initial premise: the data center market has shifted dramatically due to competition and changing workloads. Currently, there is only one player of importance in the data center market: Nvidia. As long as Nvidia stays substantially ahead of would be competitors, in the same way Intel stayed far ahead of AMD for decades, their position as the leader in AI hardware is secure. On the topic of changing workloads, we expect more AI/ML and training workloads in the future than we have today. In the same way that the typewriter became obsolete as a result of the PC and the iPod became obsolete because of the iPhone, we view generative AI as distinct from earlier versions of "AI" (machine learning, computer vision, etc) in that it has general purpose applications. As a result, generative AI will render many things obsolete (assuming performance continues to scale). Today's copilots and other AI based features (which are becoming table stakes in most software) will lead to completely different processes and business models across industries. This is a technological paradigm shift. The hyperscalers see this and are intent on investing heavily[28]. Generative AI has the ability to completely change the technology landscape, and ultimately render the cornerstones of cash flow generation in the technology

[28] Hyperscaler quotes from most recent earnings calls...

Mark Zuckerberg, Meta - "*Our expectation, obviously again, is that we are going to significantly increase our investments in AI infrastructure next year, and we'll give further guidance as appropriate.*"

Sundar Pichai, Google - "*...the risk of under-investing is dramatically greater than the risk of over-investing for us here, even in scenarios where if it turns out that we are over-investing, [inaudible] these are infrastructure which are widely useful for us*"

Sataya Nadella, Microsoft - "*To meet the growing demand signal for our AI and Cloud products, we will scale our infrastructure investments with FY '25 capital expenditures expected to be higher than FY '24. As a reminder, these expenditures are dependent on demand signals and adoption of our services that will be managed through the year.*"

Elon Musk, Tesla - "*I'm quite concerned about actually being able to get [inaudible] GPUs and when we want them.*"

industry (Google search, Microsoft Office, and Facebook/Instagram) obsolete. For each of these companies[29], their relentless pursuit of AI is purely selfish - they believe that the cost of being left behind far outweighs the cost of investing.

In conclusion, we see the market for hyperscale compute evolving rapidly. Nonetheless, our long term view remains intact: Nvidia[30] is the clear leader in the space, and there will be significantly more GPU based workloads in the future than there are today.

[29] Note that Apple, who generates more free cash flow than all of these companies is not exempt from the existential implications generative AI could have on its core iPhone business. Apple's current strategy is to provide the platform (iPhone) by which users access LLMs. This is a more rational approach from a cash flow generation perspective, but also leaves the door open for missing the (gen AI) boat entirely.

[30] Note that this discussion does not cover the all important topic of valuation

IMPORTANT DISCLAIMER

Past performance is no guarantee of future results. Investing in equities and fixed income involves risk, including the possible loss of principal. The investment performance presentation contains historical data and information relating to the performance of certain investments. These figures should not be considered as a guarantee or a reliable indicator of future performance. Investment returns and the value of an investment can go up or down, and there is no assurance that any investment strategy will achieve its objectives, generate profits, or avoid losses. Investing in financial markets inherently involves a certain degree of risk and speculation. The value of investments, and the income generated from them, can fluctuate due to various factors, including but not limited to market conditions, economic changes, interest rates, and political events. As such, there is always the potential for loss, and you should only invest funds that you can afford to lose. The S&P 500 Index is included to allow you to compare your returns against an unmanaged capitalization-weighted index of 500 stocks designed to measure performance of the broad domestic economy through changes in the aggregate market value of the 500 stocks representing all major industries. The NASDAQ 100 Index measures the performance of the largest 100 stocks on the NASDAQ.

IMPORTANT DISCLOSURES. The information herein is provided by Top Mark Capital Management LLC ("Top Mark Capital") and: (a) is for general, informational purposes only; (b) is not tailored to the specific investment needs of any specific person or entity; and (c) should not be construed as investment advice. Top Mark Capital makes no representation with respect to the accuracy, completeness or timeliness of the information herein. Top Mark Capital assumes no obligation to update or revise such information. In addition, certain information herein has been provided by and/or is based on third party sources, and, although Top Mark Capital believes this information to be reliable, Top Mark Capital has not independently verified such information and is not responsible for third-party errors. You should not assume that any investment discussed herein will be profitable or that any investment decisions in the future will be profitable. Investing in securities involves risk, including the possible loss of principal.

REFERENCES

AI Investing. "Return on Investment (ROI) on AI for Cloud Hyperscalers." Substack newsletter, August 8, 2024. https://aiinvesting.substack.com/p/return-on-investment-roi-on-ai-for.

Anderson, Timothy S., Wei-Hsuan Lo-Ciganic, Walid F. Gellad, Rouxin Zhang, Haiden A. Huskamp, Niteesh K. Choudhry, Chung-Chou H Chang, et al. "Patterns and Predictors of Physician Adoption of New Cardiovascular Drugs." *Healthcare (Amsterdam, Netherlands)* 6, no. 1 (March 2018): 33–40. https://doi.org/10.1016/j.hjdsi.2017.09.004.

Andreessen, Marc. "Why Software Is Eating the World." Andreessen Horowitz, August 20, 2011. https://a16z.com/2011/08/20/why-software-is-eating-the-world/.

D2D Advisory. "No Going Back: The New Data Center." Digits to Dollars, December 1, 2023. https://digitstodollars.com/2023/12/01/no-going-back-the-new-data-center/.

Fitch, Tom Dotan and Asa. "Why the AI Industry's Thirst for New Data Centers Can't Be Satisfied." WSJ, April 24, 2024. https://www.wsj.com/tech/ai/why-the-ai-industrys-thirst-for-new-data-centers-cant-be-satisfied-93c7eff5.

"Goldman Sachs CIO on How the Bank Is Actually Using AI." Odd Lots. Accessed August 13, 2024. https://omny.fm/shows/odd-lots/080624-odd-lots-marco-argenti-v1.

Morales, Jowi. "AI Models That Cost $1 Billion to Train Are Underway, $100 Billion Models Coming — Largest Current Models Take 'only' $100 Million to Train: Anthropic CEO." Tom's Hardware, July 7, 2024. https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-models-that-cost-dollar1-billion-to-train-are-in-development-dollar100-billion-models-coming-soon-largest-current-models-take-only-dollar100-million-to-train-anthropic-ceo.

Morales, Jowi. "Elon Musk Powers New 'World's Fastest AI Data Center" with Gargantuan Portable Power Generators to Sidestep Electricity Supply Constraints." Tom's Hardware, July 24, 2024. https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musks-new-worlds-fastest-ai-data-center-is-powered-by-massive-portable-power-generators-to-sidestep-electricity-supply-constraints.

NVIDIA Developer. "CUDA-Q." Accessed August 23, 2024. https://developer.nvidia.com/cuda-q.

Top Mark Capital | Linkedin. "NVIDIA's CORNERSTONE for UNMATCHED DOMINANCE in AI (CUDA)," August 23, 2023. https://www.linkedin.com/pulse/nvidias-cornerstone-unmatched-dominance-ai-cuda-top-mark-capital/.

Woods, Audrey. "The Death of Moore's Law: What It Means and What Might Fill the Gap Going Forward | CSAIL Alliances." Accessed August 22, 2024. https://cap.csail.mit.edu/death-moores-law-what-it-means-and-what-might-fill-gap-going-forward.