AI's $600B Question

The AI bubble is reaching a tipping point. Navigating what comes next will be essential.

BY DAVID CAHN
PUBLISHED JUNE 20, 2024

In September 2023, I published AI's $200B Question. The goal of the piece was to ask the question: "Where is all the revenue?"

At that time, I noticed a big gap between the revenue expectations implied by the AI infrastructure build-out, and actual revenue growth in the AI ecosystem, which is also a proxy for end-user value. I described this as a "$125B hole that needs to be filled *for each year of CapEx at today's levels*."

This week, Nvidia completed its ascent to become the most valuable company in the world. In the weeks leading up to this, I've received numerous requests for the updated math behind my analysis. Has AI's $200B question been solved, or exacerbated?

If you run this analysis again today, here are the results you get: AI's $200B question is now AI's $600B question.

	Q4 2023 ESTIMATE	Q4 2023 ACTUAL	Q1 2024 ACTUAL	Q4 2024 ESTIMATE
NVDA Data Center Run-Rate Revenue	$50	$74	$90	$150
Data Center Facility Build and Cost to Operate	*50%*	*50%*	*50%*	*50%*
Implied Data Center AI Spend	$100	$147	$181	$300
Software Margin	*50%*	*50%*	*50%*	*50%*
AI Revenue Required for Payback	**$200**	**$294**	**$363**	**$600**

Note: It's easy to calculate this metric directly. All you have to do is to take Nvidia's run-rate revenue forecast and multiply it by 2x to reflect the total cost of AI data centers (GPUs are half of the total cost of ownership—the other half includes energy, buildings, backup generators, etc)[1]. Then you multiply by 2x again, to reflect a 50% gross margin for the end-user of the GPU, (e.g., the startup or business buying AI compute from Azure or AWS or GCP, who needs to make money as well).

What has changed since September 2023?

1. **The supply shortage has subsided:** Late 2023 was the peak of the GPU supply shortage. Startups were calling VCs, calling anyone that would talk to them, asking for help getting access to GPUs. Today, that concern has been almost entirely eliminated. For most people I speak with, it's relatively easy to get GPUs now with reasonable lead times.

2. **GPU stockpiles are growing:** Nvidia reported in Q4 that about half of its data center revenue came from the large cloud providers. Microsoft alone likely represented approximately 22% of Nvidia's Q4 revenue. Hyperscale CapEx is reaching historic levels. These investments were a major theme of Big Tech Q1 '24 earnings, with CEOs effectively telling the market: "We're going to invest in GPUs whether you like it or not." Stockpiling hardware is not a new phenomenon, and the catalyst for a reset will be once the stockpiles are large enough that demand decreases.

3. **OpenAI still has the lion's share of AI revenue:** The Information recently reported that OpenAI's revenue is now $3.4B, up from $1.6B in late 2023. While we've seen a handful of startups scale revenues into the <$100M range, the gap between OpenAI and everyone else continues to loom large. Outside of ChatGPT, how many AI

products are consumers really using today? Consider how much value you get from Netflix for $15.49/month or Spotify for $11.99. Long term, AI companies will need to deliver significant value for consumers to continue opening their wallets.

4. **The $125B hole is now a $500B hole:** In the last analysis, I generously assumed that each of Google, Microsoft, Apple and Meta will be able to generate $10B annually from new AI-related revenue. I also assumed $5B in new AI revenue for each of Oracle, ByteDance, Alibaba, Tencent, X, and Tesla. Even if this remains true and we add a few more companies to the list, the $125B hole is now going to become a $500B hole.

5. **It's not over—the B100 is coming:** Earlier this year, Nvidia announced their B100 chip, which will have 2.5x better performance for only 25% more cost. I expect this will lead to a final surge in demand for NVDA chips. The B100 represents a dramatic cost vs. performance improvement over the H100, and there will likely be yet another supply shortage as everyone tries to get their hands on B100s later this year.

One of the major rebuttals to my last piece was that "GPU CapEx is like building railroads" and eventually the trains will come, as will the destinations—the new agriculture exports, amusement parks, malls, etc. I actually agree with this, but I think it misses a few points:

1. **Lack of pricing power:** In the case of physical infrastructure build outs, there is some intrinsic value associated with the infrastructure you are building. If you own the tracks between San Francisco and Los Angeles, you likely have some kind of monopolistic pricing power, because there can only be so many tracks laid between place A and place B. In the case of GPU data centers, there is much less pricing power. GPU computing is increasingly turning into a commodity, metered per hour. Unlike the CPU cloud, which became an oligopoly, new entrants building dedicated AI clouds continue to flood the market. Without a monopoly or oligopoly, high fixed cost + low marginal cost businesses almost always see prices competed down to marginal cost (e.g., airlines).

2. **Investment incineration:** Even in the case of railroads—and in the case of many new technologies—speculative investment frenzies often lead to high rates of capital incineration. The Engines that Moves Markets is one of the best textbooks on technology investing, and the major takeaway—indeed, focused on railroads—is that a lot of people lose a lot of money during speculative technology waves. It's hard to pick winners, but much easier to pick losers (canals, in the case of railroads).

3. **Depreciation:** We know from the history of technology that semiconductors tend to get better and better. Nvidia is going to keep producing better next-generation chips like the B100. This will lead to more rapid depreciation of the last-gen chips. Because the market under-appreciates the B100 and the rate at which next-gen chips will improve, it overestimates the extent to which H100s purchased today will hold their value in 3-4 years. Again, this parallel doesn't exist for physical infrastructure, which does not follow any "Moore's Law" type curve, such that cost vs. performance continuously improves.

4. **Winners vs. losers:** I think we need to look carefully at winners and losers—there are always winners during periods of excess infrastructure building. AI is likely to be the next transformative technology wave, and as I mentioned in the last piece, declining prices for GPU computing is actually good for long-term innovation and good for startups. If my forecast comes to bear, it will cause harm primarily to investors. Founders and company builders will continue to build in AI—and they will be more likely to succeed, because they will benefit both from lower costs and from learnings accrued during this period of experimentation.

A huge amount of economic value is going to be created by AI. Company builders focused on delivering value to end users will be rewarded handsomely. We are living through what has the potential to be a generation-defining technology wave. Companies like Nvidia deserve enormous credit for the role they've played in enabling this transition, and are likely to play a critical role in the ecosystem for a long time to come.

Speculative frenzies are part of technology, and so they are not something to be afraid of. Those who remain level-headed through this moment have the chance to build extremely important companies. But we need to make sure not to believe in the delusion that has now spread from Silicon Valley to the rest of the country, and indeed the world. That delusion says that we're all going to get rich quick, because AGI is coming tomorrow, and we all need to stockpile the only valuable resource, which is GPUs.

In reality, the road ahead is going to be a long one. It will have ups and downs. But almost certainly it will be worthwhile.

If you are building in this space, we'd love to hear from you. Please reach out at dcahn@sequoiacap.com

---

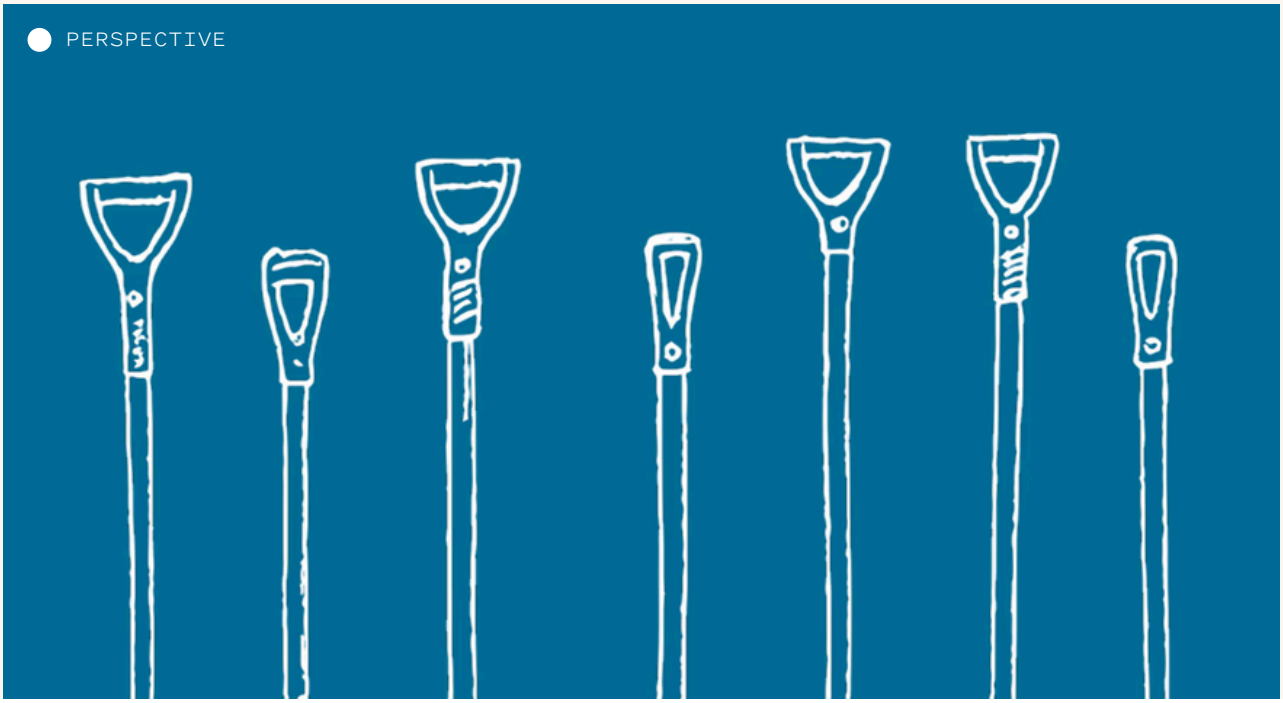
[1] Some commenters challenged my 50% assumption on non-GPU data center costs, which I summarized as energy costs. Nvidia actually came to the exact same metric, which you can see on Page 14 of their October 2023 analyst day presentation, published a few days after my last piece.

SHARE

RELATED TOPICS

#AI

PERSPECTIVE

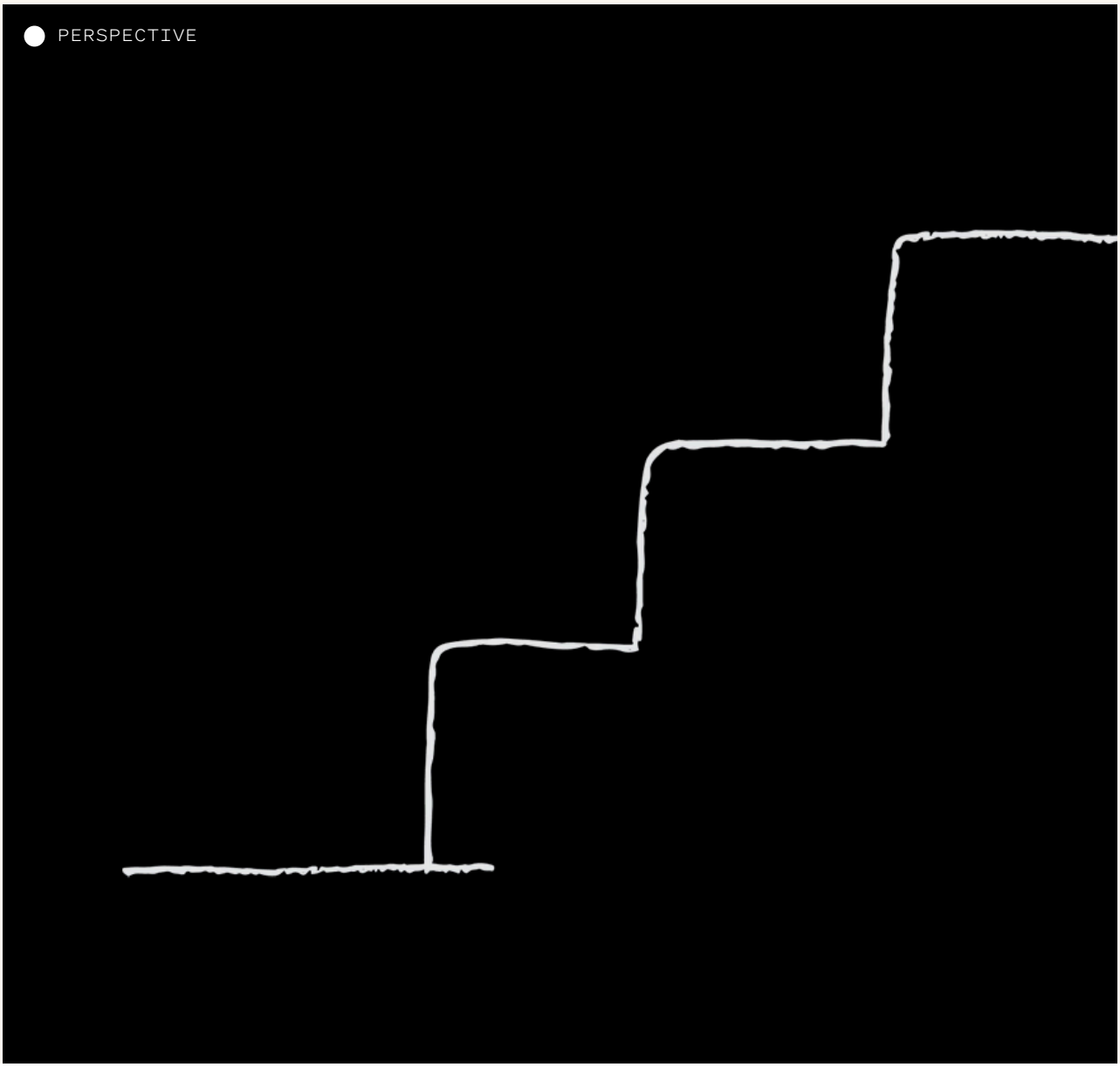




AI is Now Shovel Ready

By David Cahn

PERSPECTIVE



AI's $200B Question

By David Cahn

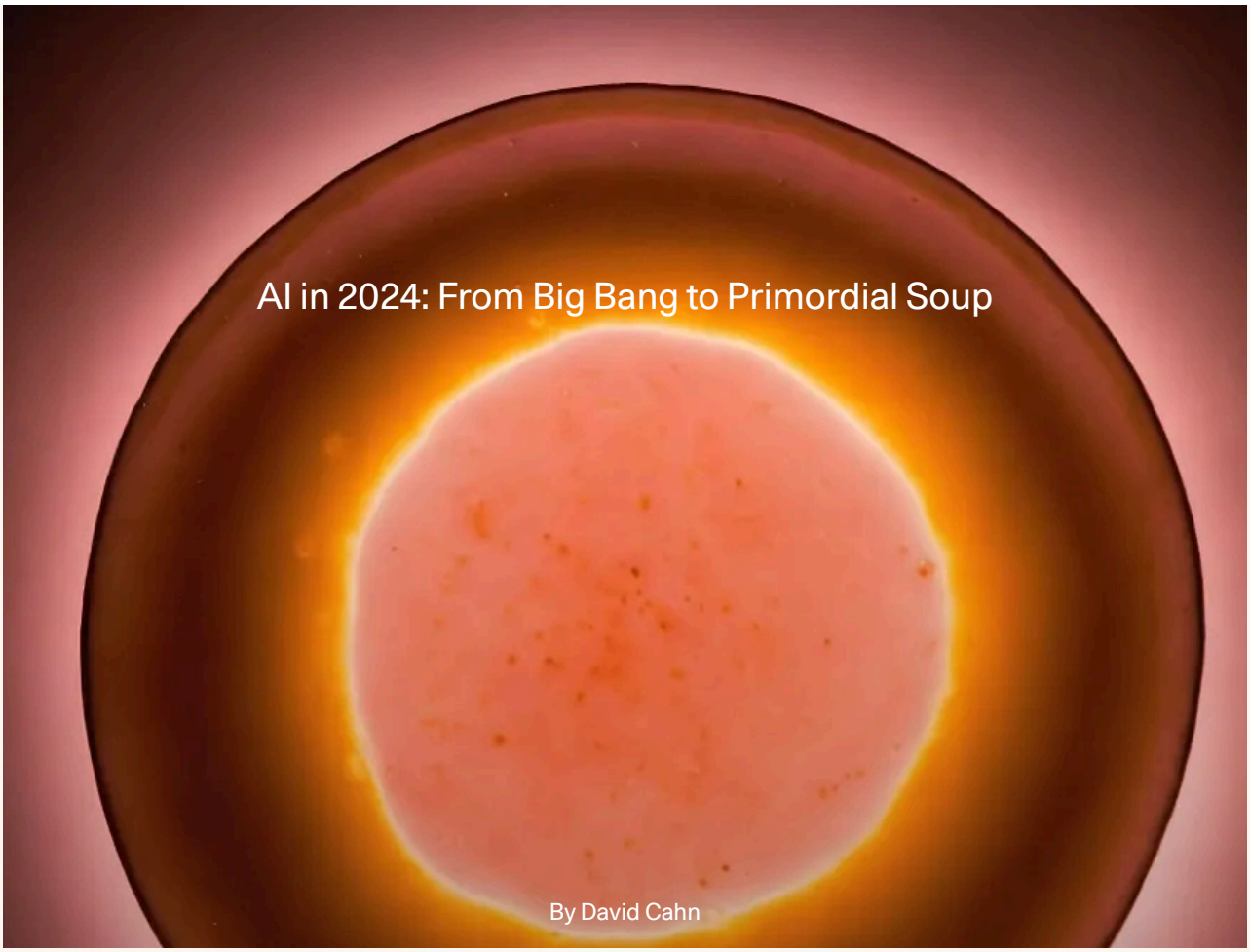
PERSPECTIVE



AI Ascent 2024

Video highlights from our AI conference.

PERSPECTIVE



© 2026 Sequoia Capital