APOLLO

The Growing Compute Shortage

June 2026

Robert Bittencourt
Partner, Head of Apollo Thematic Investing

In Frank Herbert's *Dune*, spice is an indispensable resource—found only on the desert planet Arrakis—that powers interstellar civilization.

In 2026, compute is becoming the spice of our era.

Over the last few months, we have seen a marked shift in conversations around AI infrastructure, with debate now centered on whether the world can build enough compute fast enough to satisfy demand. Across the AI supply chain, constraints are emerging simultaneously: GPUs are scarce, memory prices are surging, TSMC capacity is effectively sold out,[1] and power infrastructure has become a critical bottleneck. The result is a new form of scarcity that is beginning to reshape corporate strategy, capital allocation, and market leadership.

Compute, in short, refers to the computational capacity required to train and run AI models. It depends on a deeply intertwined physical supply chain: accelerators, high-bandwidth memory, networking equipment, large-scale datacenters, and enormous amounts of electricity. AI may appear digital, but it is increasingly constrained by the physical world.

What changed over the past year is not only that more people are using AI. It is also that AI workloads themselves have become dramatically more compute-intensive. Traditional chatbot interactions consumed relatively small amounts of compute. Agentic systems, particularly coding agents, consume orders of magnitude more tokens because they reason iteratively, test outputs, execute workflows, use tools like web search, and repeatedly query models before producing a result. In some cases, agentic AI workloads can consume 100x–1,000x more tokens than traditional chatbot requests.

1. Source: JP Morgan Research. Data as of June 2026.

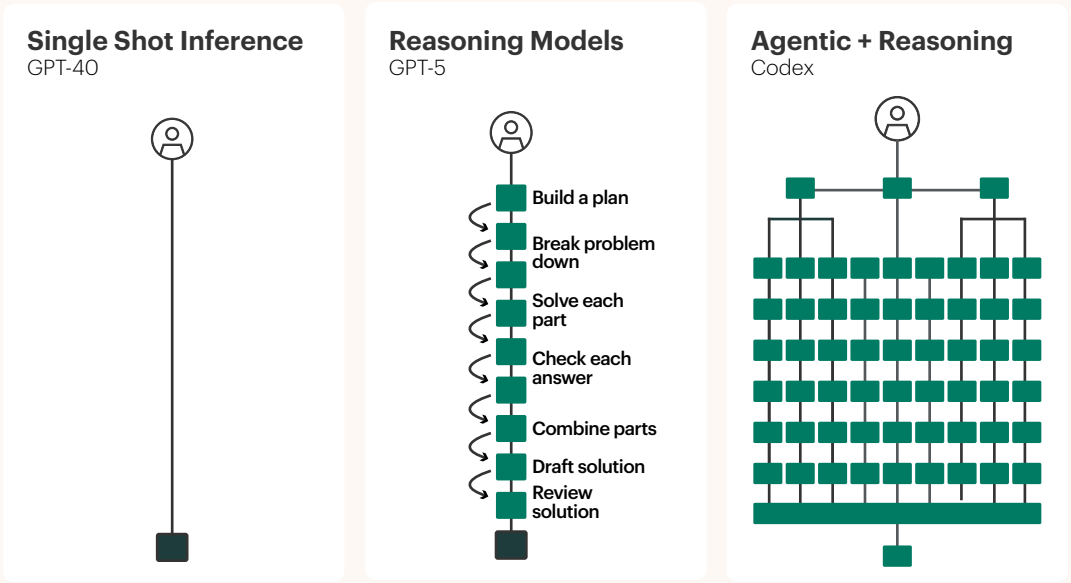
ATLWAA-20260615-5579472-17593507

Agents Use More Tokens Than Traditional Chatbot Requests

Tokens = Compute

Agentic models orchestrate multiple models, each performing search, reasoning, and synthesis.

Agentic models consume 100 to 1,000 times more tokens than base models.



Illustrative compute required for inference.	1x	~100x	>1,000x

Source: Cerebras S-1. Data as of April 2026.

At the same time, enterprise adoption remains in its early innings. Coding has emerged as the first major enterprise AI use case, but sectors ranging from legal services to financial analysis to healthcare are only beginning to deploy AI at scale. Demand is accelerating at a pace that is exceeding even a rapidly expanding supply base. That tension is beginning to reverberate across the compute ecosystem.

Emerging Bottlenecks

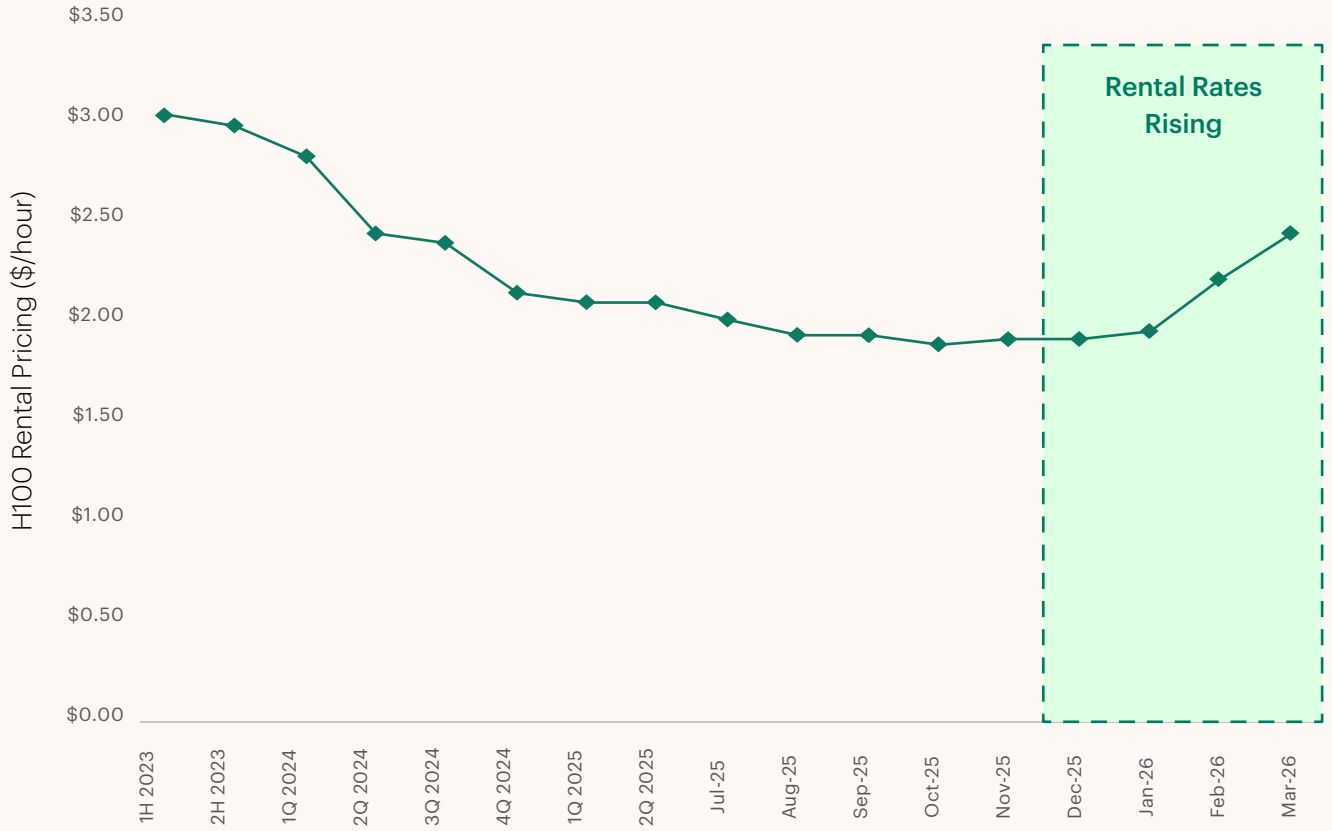
Historically, technology shortages have typically been traced to a single chokepoint. What makes the current environment unusual is that constraints are appearing across many layers of the supply chain at once.

The most obvious shortage remains deployed GPUs. On-demand GPU capacity is effectively sold out, and even older-generation chips are seeing rental rates increase to levels not seen since 2024.[2] Hyperscalers and model builders continue to spend aggressively to secure supply, in some cases signing multi-year infrastructure commitments well ahead of deployment. The race for compute has accelerated to such an extent that competing firms are partnering to secure capacity, with both Anthropic and Google signing large deals to rent compute capacity from xAI.[3]

2. Source: SemiAnalysis, April 2026.
3. Source: TechCrunch, June 2026.

ATLWAA-20260615-5579472-17593507

GPU Rental Rates are Rising



Source: SemiAnalysis, The Great GPU Shortage – Rental Capacity – Launching our H100 1 Year Rental Price Index. Data as of April 2026.

At the semiconductor level, TSMC's advanced-node capacity—particularly N3, which underpins much of the AI accelerator ecosystem—is approaching full utilization through at least 2027.[4] Advanced fabs take years to construct, and the EUV lithography machines required to equip them have similarly extended production timelines.

Memory has emerged as another major bottleneck. High-bandwidth memory (HBM), a specialized type of dynamic random-access memory (DRAM) essential for AI accelerators, remains structurally undersupplied as demand from hyperscalers exceeds existing production capacity. But AI systems also rely on large amounts of ordinary DRAM, and as manufacturers shift capacity toward HBM, the broader tightening of the memory market has driven spot DRAM prices up approximately 8x since early 2025.[5] Importantly, this dynamic not only impacts AI infrastructure providers. Industries competing for similar memory supply, including smartphones and consumer electronics, are already seeing upward pressure on component costs.

Even if silicon constraints ease, power infrastructure remains a critical limiting factor. Forecasts for global datacenter electricity demand now imply 200-300 gigawatts of data center power demand globally by 2030.[6] Hyperscalers are pursuing gigawatt-scale campuses while simultaneously racing to secure turbines, transformers, switchgear, and grid connectivity. GE Vernova and Siemens Energy have both said they are nearly sold out of gas turbines through 2029.[7] Transformer lead times have stretched to multiple years, and utilities and regional grids are increasingly struggling to respond to the rising demand from AI datacenters.

4. Source: JP Morgan Research. Data as of June 2026.
5. Source: Bloomberg data. Data as of June 2026.
6. Source: McKinsey. Data as of August 2025.
7. Source: GE Vernova, April 2026; Siemens, May 2026

ATLWAA-20260615-5579472-17593507

Scarcity Is Repricing Assets

As a result, the market is beginning to reprice anything perceived as a bottleneck to the AI buildout.

One of the clearest examples is bitcoin miners. For years, many crypto mining assets were viewed as structurally impaired following the collapse in crypto valuations and increasing competition within the sector. But in a compute-constrained world, those same assets suddenly look strategically valuable. Many miners already possess what hyperscalers need most: Access to power, land, and grid connectivity. Some companies like Crusoe and Core Scientific, that started as crypto-focused enterprises, have already made that transition, but more are likely to follow the same path now.

In other words, scarcity is creating "Lazarus assets"—assets once viewed as obsolete that are being revived by their relevance to AI infrastructure.

This rerating dynamic has spread across the broader economy. Power equipment manufacturers, cooling companies, electrical component suppliers, networking providers, and even industrial and brownfield infrastructure assets are increasingly valued through the lens of their ability to alleviate compute constraints.

Compute Is Becoming a Competitive Moat

In a compute-constrained world, access itself becomes a competitive moat. Some of the largest AI platforms and model builders have committed enormous amounts of capital to secure long-term compute supply. What looked like an overreach only months ago increasingly looks like strategic foresight. Companies with secured capacity can serve demand today; those without it could face rationing, slower deployment, or significantly higher costs.

This dynamic may also alter the economics of AI itself. For years, the dominant assumption surrounding AI was that models would become rapidly cheaper over time as hardware improved and efficiency gains accumulated. That may still prove true eventually. But in the near term, scarcity could temporarily reverse that trend. If demand continues to outpace supply, the immediate solutions are either materially higher pricing or some form of usage rationing for the leading-edge models. Early signs of both are already emerging across parts of the industry.

Higher inference costs would also likely slow the diffusion of AI across the economy, concentrating compute resources toward the highest-value applications first. Companies with the ability to pay for compute may gain an advantage, while smaller firms and experimental use cases could face increasing barriers to access.

Several developments could ultimately ease the compute shortage narrative: A sharper-than-expected improvement in model efficiency, a cyclical slowdown in AI demand, a faster ramp in semiconductor and power infrastructure capacity, or breakthroughs that reduce dependence on today's constrained hardware stack. But for now, the opposite dynamic appears to be dominating. Demand is accelerating faster than the physical world can respond.

The Physical World Still Matters

For the better part of two decades, the dominant assumption in technology was that software scaled infinitely while the physical world slowly faded into the background. AI is reversing that logic. The next phase of the technology cycle will not simply be determined by better models or smarter algorithms, but by access to scarce physical resources: Chips, memory, electricity, land, cooling, and industrial infrastructure. In other words, the AI race is no longer just a software race. It is an industrial-scale, infrastructure race.

And increasingly, the winners may not just be the companies with the best ideas, but the ones that secured the compute to turn those ideas into reality.

In *Dune*, spice scarcity reshaped the balance of power across the galaxy. In today's economy, compute may be beginning to do the same.

ATLWAA-20260615-5579472-17593507

Important Disclosure Information

All information contained in this material is as of May 8, 2026 unless otherwise indicated.

This material is for informational purposes only. This material should not be copied, distributed, published, or reproduced, in whole or in part, or disclosed by any recipient to any other person without the express written consent of Apollo Global Management, Inc. (together with its subsidiaries, "Apollo").

The views and opinions expressed in this material are the views and opinions of Apollo Analysts. They may not reflect the views and opinions of Apollo and are subject to change at any time without notice. This material does not constitute an offer of any service or product of Apollo. It is not an invitation by or on behalf of Apollo to any person to buy or sell any security or to adopt any investment strategy, and shall not form the basis of, nor may it accompany nor form part of, any right or contract to buy or sell any security or to adopt any investment strategy.

There can be no assurances that any of the trends described in this material will continue or will not reverse. Past events and trends do not imply, predict or guarantee, and are not necessarily indicative of future events or results. This material is not complete and the information contained in this material may change at any time without notice. Apollo has no responsibility to update any of the information provided in this material. Apollo has not made any representation or warranty, expressed or implied, with respect to fairness, correctness, accuracy, reasonableness, or completeness of any of the information contained in this material (including but not limited to information obtained from third parties unrelated to Apollo). The information contained in this material is not intended to provide, and should not be relied upon for, accounting, legal or tax advice or investment recommendations. Investors should make an independent investigation of the information contained in this material, including consulting their tax, legal, accounting or other advisors about such information. Apollo does not act for you and is not responsible for providing you with the protections afforded to its clients.

Certain information contained in this material may be "forward looking" in nature. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking information. As such, undue reliance should not be placed on such information. Forward-looking statements may be identified by the use of terminology including, but not limited to, "may", "will", "should", "expect", "anticipate", "target", "project", "estimate", "intend", "continue" or "believe" or the negatives thereof or other variations thereon or comparable terminology.

This material may reference trade names, trademarks, or service marks of companies that are not owned by Apollo or Apollo funds, or that may be held as investments by Apollo or one or more Apollo funds. The use or display of these companies' trade names, trademarks, or service marks is not intended to imply any relationship with, or endorsement or sponsorship of Apollo, by such companies. All company names and logos are trademarks of their respective holders.

Past performance is not necessarily indicative of future results.

Additional information may be available upon request.

ATLWAA-20260615-5579472-17593507