SpaceX 10GW in 2027 – Why It’s Real, Will Drive $500B ARR for SpaceX, and Why Microsoft Will Be the Largest Offtak…

SpaceX 10GW in 2027 – Why It’s Real, Will Drive $500B ARR for SpaceX, and Why Microsoft Will Be the Largest Offtaker Inference at 100B/GW/year, SpaceX's stellar pace, Microsoft's 10GW 2026 Awakening, Azure Can Grow Tiple-Digits

AUG 7 https://newsletter.semianalysis.com/p/spacex-10gw-in-2027-why-its-real

View in browser

QQ:1413221435

SpaceX 10GW in 2027 – Why It’s Real, Will Drive $500B ARR for SpaceX, and Why Microsoft Will Be the Largest Offtaker

Inference at 100B/GW/year, SpaceX's stellar pace, Microsoft's 10GW 2026 Awakening, Azure Can Grow Tiple-Digits

JEREMIE ELIAHOU ONTIVEROS, REYK KNUHTSEN, JORDAN NANOS, MAX KAN, DYLAN PATEL, AND MUHAMMAD ZUHAIR

AUG 7 • PAID

READ IN APP

Elon Musk shocked the world, once again, when he announced on SpaceX’s first earnings his Gigawatt ambitions for next year. He “conservatively” aims to build & deliver an incremental 6-8GW in 2027 alone, with potential for that number to be well above +10GW. At 50B per GW, that’s $300-500B in capex in 2027, on par with what we expect from AWS and Google – an unbelievable number for a company significantly less profitable than rival hyperscalers.

QQ:1413221435

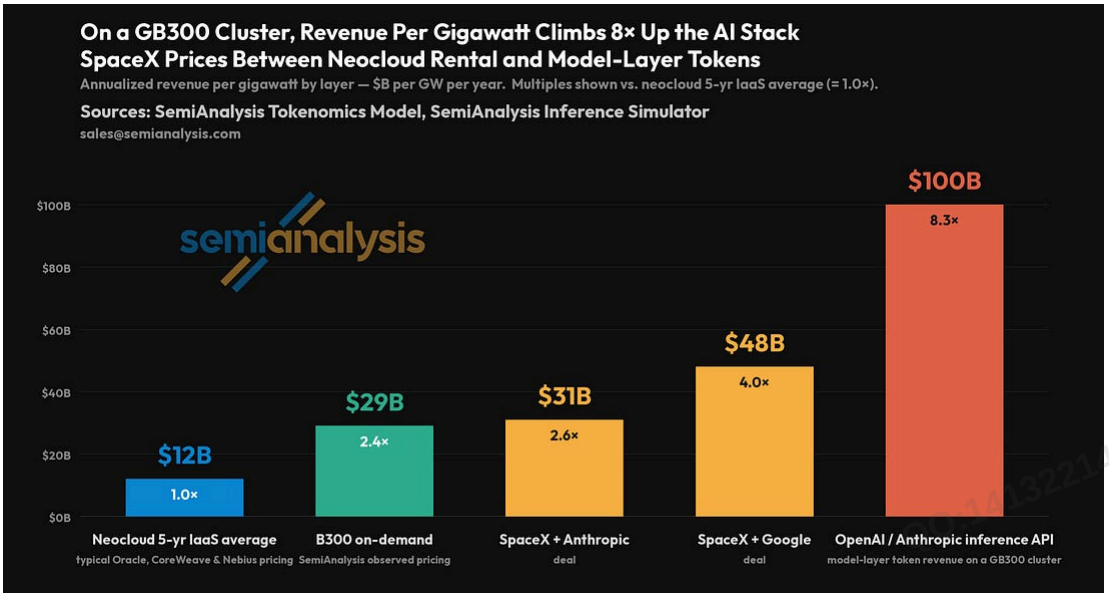
Yet, we believe that the number is real. We see SpaceX on track to build about 10GW by year-end 2027. We've evaluated all sites suitable for SpaceX and provided the list to our Datacenter Model subscribers. Our Energy Model subscribers also have the precise list of gas generation equipment available, quarter by quarter, by 30+ turbine, engine, fuel cell suppliers. We provided much of this data, before the market woke up to it.

SpaceX will develop anything they can and bring it online as fast as possible. As explained in our Meta Compute deep dive, large-scale + near-term compute is a remarkably scarce combination, and it's priced at a huge premium – up to $50B/GW/year. However, AI labs can handle it and make a good living off it.

Our Tokenomics Model and our Inference Simulator demonstrate that at realistic performance levels (e.g. tokens/sec per GPU), both OpenAI and Anthropic can generate over $100B/GW/year of revenue when selling API inference on a GB300 cluster. This is significantly more than the costs of renting a GB300 cluster for a year at current neocloud prices.

We assume around $12B/GW/year of cost per year, using a conservative rental pricing rate of $3/GPU-hr, and make a token production estimate using our Inference Simulator with a frontier-class model architecture and our agentic coding benchmark, AgentX (part of InferenceX), which is built by collecting real production coding traces. We blend that token production rate between input, cache-read, cache-write, and output token costs at our real workload ratios, and produce the final estimate, exceeding $100B/GW/year.

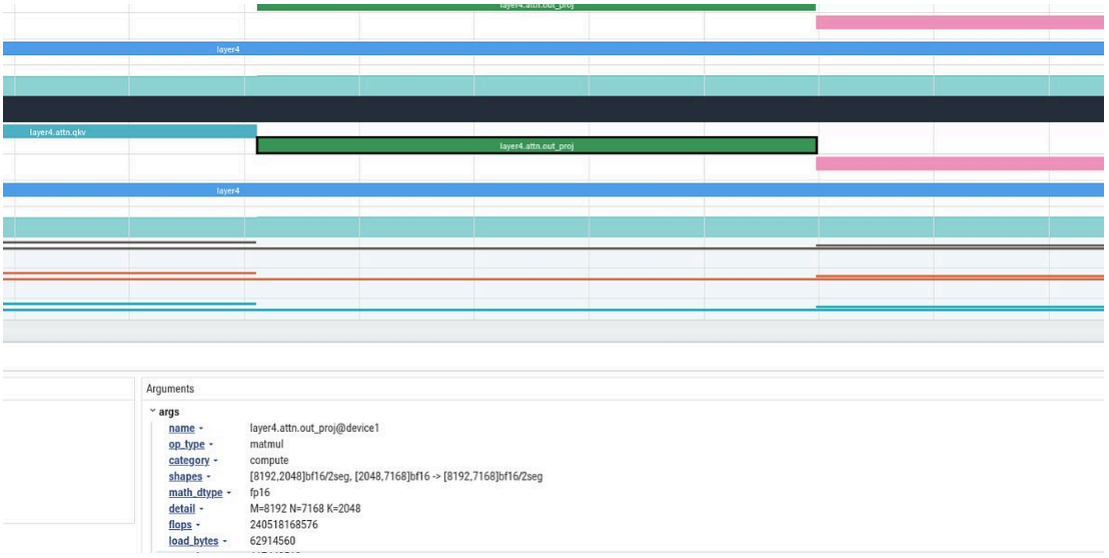
Serving inference tokens is unbelievably profitable for the frontier model companies.



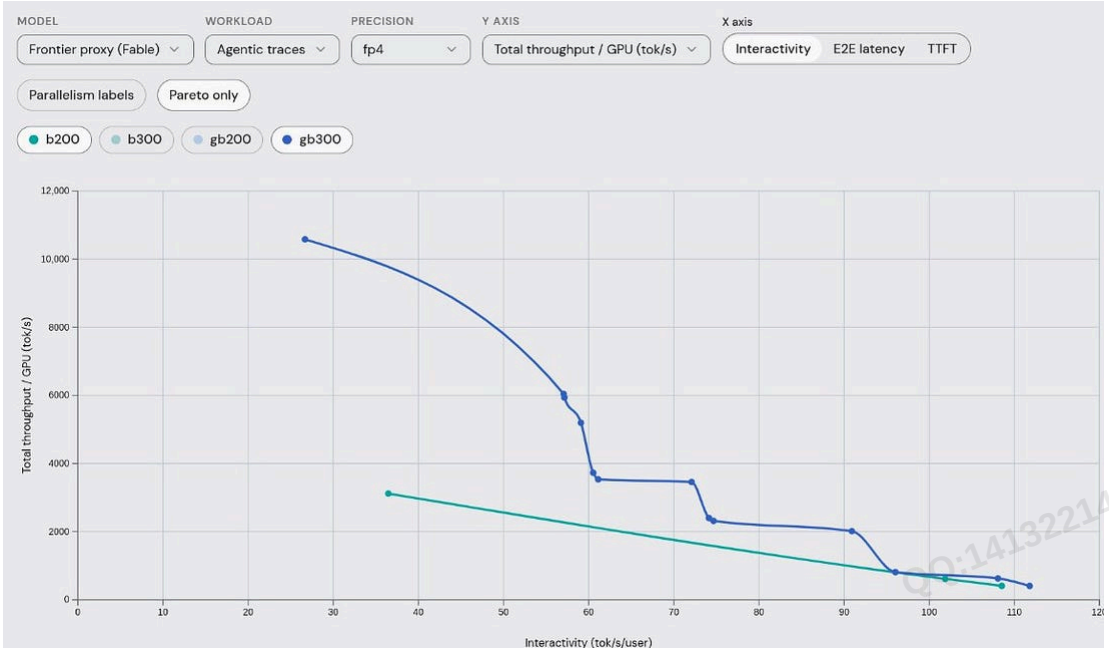
QQ:1413221435

Source: SemiAnalysis Tokenomics Model, SemiAnalysis Inference Simulator

For background, our Inference Simulator is built from the ground up with a fundamental understanding of how modern AI accelerators work. We build a roofline and realistic performance model for how frontier models work during inference, with timings for every operation and a real trace output. It is an end-to-end simulation of the actual workload executing on the actual silicon. We have validated the simulators fidelity on a wide range of accelerators and workloads and continue to improve its ability to accurately forecast performance of future accelerators based on design specifications.



Fine-grained data covering end-to-end simulated workload execution on silicon produces real profiler traces for analysis with standard tools such as Perfetto. Source: SemiAnalysis Inference Simulator



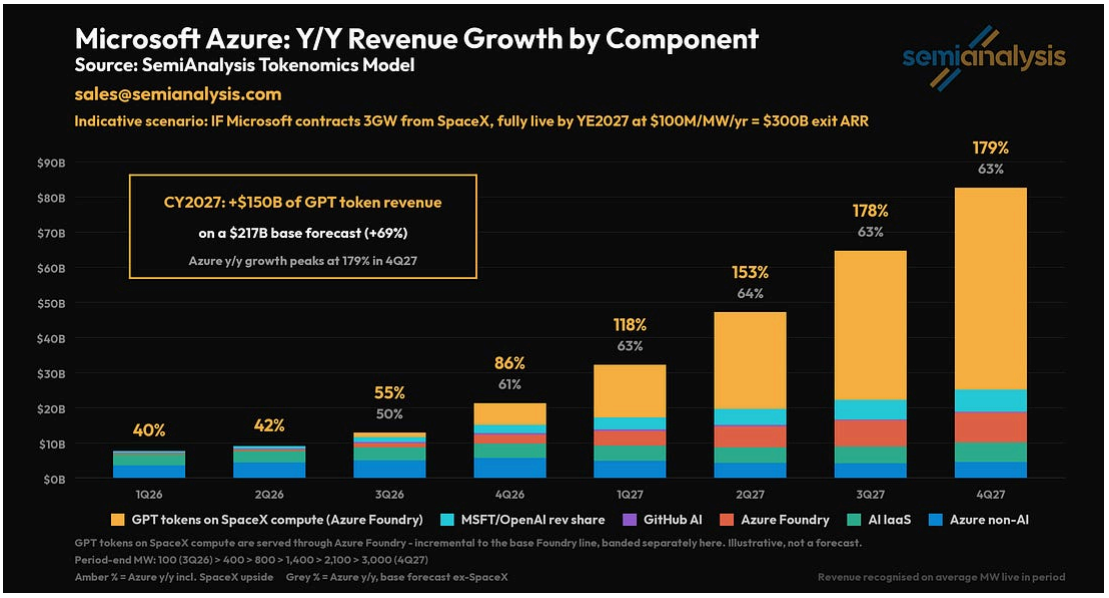
QQ:1413221435

High-level projections are produced across common inference workloads and hardware platforms across the pareto frontier. Source: SemiAnlaysis Inference Simulator

Please reach out to sales@semianalysis.com for more information on how we apply the Inference Simulator for custom research and analysis.

Beyond OpenAI and Anthropic, there is actually a third company in the world capable of printing such numbers: Microsoft. Having full access to OpenAI models, they can generate the exact same revenue and margin per MW, while paying none of the training costs. Satya nailed the negotiations with OpenAI: the deal reworked in April 2026 dropped the old 20% revenue share from the equation. Put simply, Microsoft has a giant incentive to procure as many MWs as possible, as fast as possible. While much of their datacenter capacity currently goes to OpenAI at ~14M/MW/year, they have the opportunity to improve that mix. The potential impact is Microsoft Azure accelerating revenue growth from ~42% to over 100% by next year. A once-in-a-generation opportunity, that SpaceX is incredibly well positioned to serve.

QQ:1413221435



Source: SemiAnalysis Tokenomics Model

For SpaceX, the next natural question is financing. How can Elon afford to pay so much CapEx without the balance sheet of the leading hyperscalers? We expect a combination of the two following items:

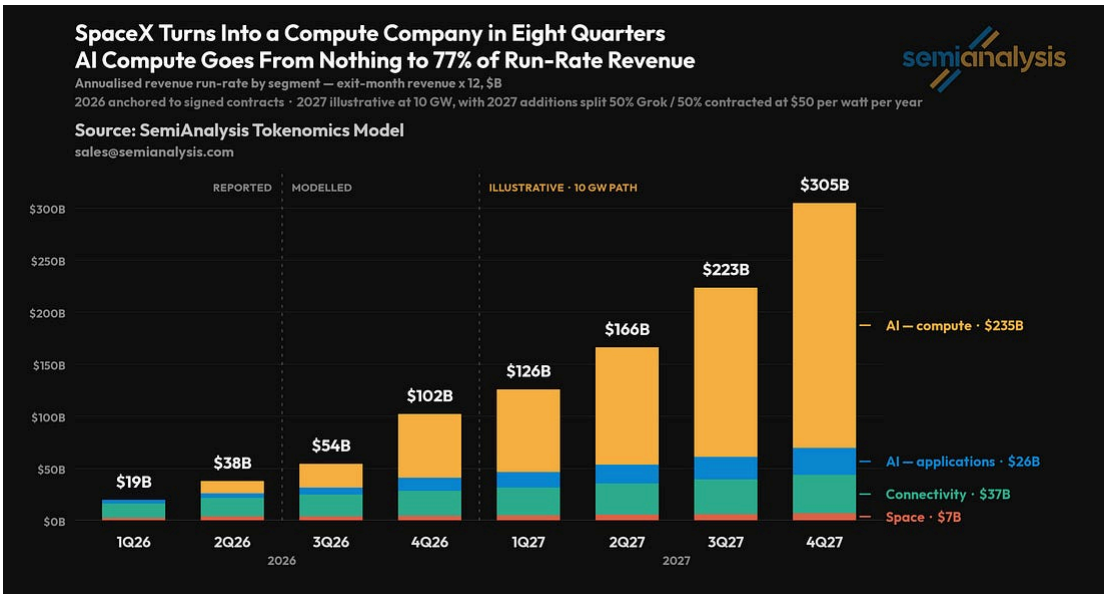
- 1/ Support from Nvidia, in the form of vendor financing to lower the upfront cash cost. This is likely why Elon declared to be Nvidia exclusive on the earnings call! As our Accelerator Model has repeatedly explained, xAI/SpaceX

QQ:1413221435

have actively evaluated alternatives like TPU and AMD – so the financial argument likely made them abandon these and focus on Nvidia.

- 2/ Industry-high pricing, enabled by fastest timelines: SpaceX will continue to sell large-scale compute with 3-5 months lead time, an unbeatable offering, and price it accordingly at 30-50M/MW/year. That pays back the capex in less than a year. We dived into this in our Meta Compute article.

Of course, the implications of this are a path to $300B of ARR by the end of 2027 for SpaceX. This assumes only 50% of their 2027 incremental compute is monetized, the reminder being for the Grok & Cursor teams for training (no inference revenue modelled).



Source: SemiAnalysis Tokenomics Model

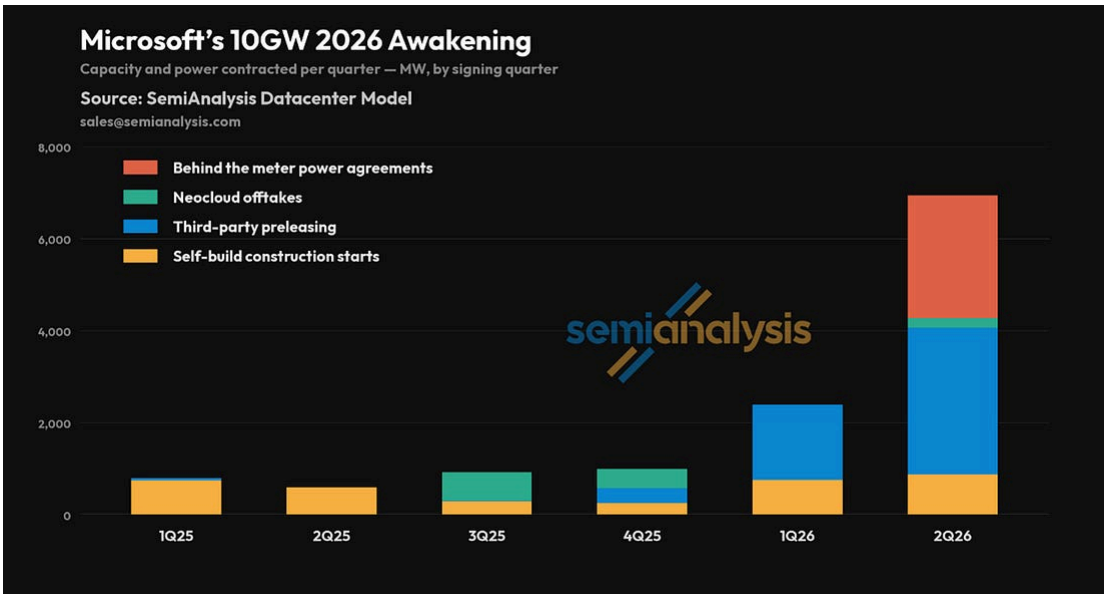
Let’s now dig in. We begin with Microsoft, who has spectacularly, finally, woken up: last year’s pause has reverted, with 10GW of signed binding contracts year-to-date. We’ll briefly discuss economics to get to $100M/MW/year of inference revenue. We then shift to SpaceX and analyze their datacenter ramp. Finally, we quantify the revenue and profit opportunity for SpaceX and Microsoft and discuss the meaning for OpenAI and Anthropic.

Microsoft’s 10GW awakening to capture the 100M/MW/year opportunity

In December 2024, we called out before anyone else in our Datacenter Model a dramatic pause in Microsoft’s leasing activity. Today, the giant awakened. Our

QQ:1413221435

models tracks quarter-by-quarter leasing activity, neocloud contracting, self-build construction starts, and large-scale binding PPAs and ESAs. We show below the outputs. Microsoft has contracted over 10GW across all these surfaces, which is the equivalent of ~$300B in new binding commitments.



Source: SemiAnalysis Datacenter Model

A key reason for this awakening is their desperate need for compute to capture a $100M/MW/Year revenue opportunity. Microsoft signed in October 2025 a $250B agreement with OpenAI, which we estimate at ~7GW in total in our Tokenomics Model – the world's best tool to understand the nuances of the dollar-to-watt math. This massive Infrastructure-as-a-Service deal has left Microsoft highly compute-constrained on their other use-cases. They've been unable to leverage their access to OpenAI models for their API business Foundry, or for their applications like Copilot.

Yet, these are the services that come at the highest margin and revenue per MW, by far. We've explained that in depth in our AI Value Capture piece.

AI Value Capture - The Shift To Model Labs

DANIEL NISHBALL, DYLAN PATEL, AND 7 OTHERS · MAY 1

QQ:1413221435
QQ:1413221435



A day in AI now feels like a year in any other industry. Model releases, software breakthroughs, and hardware improvements are compressing multi-year cycles for any other industry into weeks. Over just the past few months, agentic AI has crossed a real inflection point, driving a step-change in the value of tokens while software and hardware improvements have sharply reduced the cost of generating them.

Read full story

Over the past month, it's finally become consensus among sophisticated investors that serving frontier tokens at API prices is actually an extremely high margin business. We were the first to call this out to our Tokenomics Model subscribers back in Janurary, when we explained why inference gross margins are north of 60%. Then in June, we followed up with a deep dive that showed how Opus 4.8 in particular had 85%+ margins. This has since become the default number everyone cites when analyzing Anthropic.

To arrive at these margin estimates, we had to carefully synthesize leaked financials, InferenceX data, microbenchmarks on all the latest accelerators in the industry, papers, blogs and tweets from open source labs, and more. New datapoints such as the leaked DeepSeek investor call (which said they have a 10-month GPU payback period) confirm we're in the right ballpark, but we'll be the first to admit that the lack of granularity is extremely unsatisfying. **Rather than a single company wide inference gross margin number, what you really want to know is the gross margin for every (model, accelerator) combo along the entire throughput vs latency pareto frontier.** For example, what's the gross margin for serving Opus 5 Fast on Trainium3 vs Fable 5 on TPUv7?

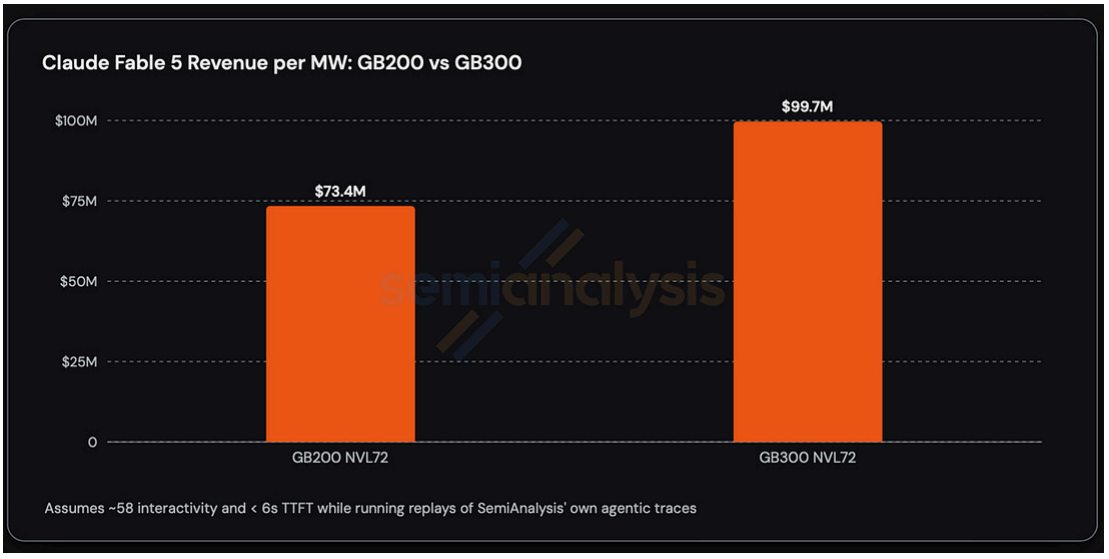
We answer that question with our Inference Simulator, available exclusively to SemiAnalysis consulting clients.

QQ:1413221435
QQ:1413221435

Our AI Cloud TCO model already answers the cost side of the equation, but the revenue side has historically been unknowable. To solve this, we wrote a simulation framework that simulates real model execution on virtual hardware, backed by tuned fine-grained performance models covering a variety of accelerators and operation types. We run each model on simulated XPUs across every possible serving configuration, with a mix of real-world and idealized serving conditions. This allows us to, given some well-informed assumptions about model architecture, accurately estimate the performance of any combination of software, hardware, and workload.

Thanks to this simulator, **our Tokenomics Model now includes high-level revenue per MW numbers for running the flagship OpenAI/Anthropic models on all the relevant chips.** Workload shape is obviously a huge factor, and we simulate running over $1M worth of agentic traces collected from our own usage while meeting the real interactivity and TTFT levels observed from hitting first party endpoints. As a teaser, here are our numbers for serving Fable 5 on GB200 vs GB300:

QQ:1413221435



Source: SemiAnalysis Tokenomics Model

This is Microsoft's $100M per MW opportunity. Given the recent surge in Codex demand and corresponding OpenAI ARR acceleration, we believe Microsoft would be able to monetize compute at similar rates by serving OAI models.

Now, to capture this once-in-a-lifetime opportunity, Microsoft and AI Labs datacenters, and they need them quick and big. SpaceX has already proven twice that they can build faster than others, but their compute capacity will "only" be 2GW by year-end 2026. Can they really build 10GW+ in just a single year, in 2027?

QQ:1413221435

SpaceX: building datacenters at a stellar pace

In our Meta Compute article, we explained in depth why Elon has proven, yet again, to be a commercial genius. He understands that AI lab margins have dramatically surged, and accordingly introduced a “value-based pricing” for his GPU clusters, as opposed to the more common “cost plus”.

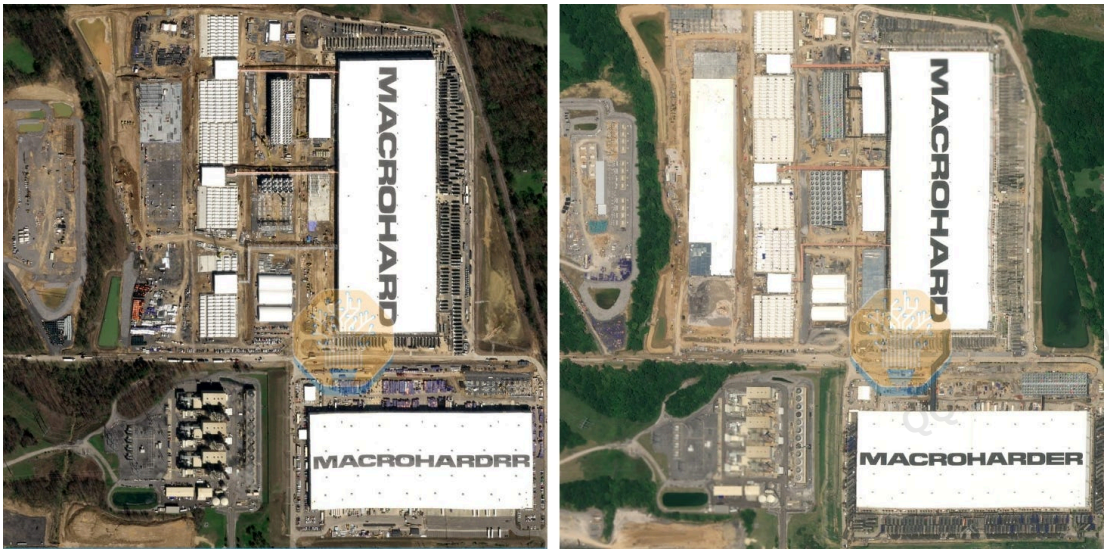
To keep the machine going, Elon needs to build datacenters faster than anyone else. We believe that he can. What gives us this confidence? We’ve written a few times about Elon’s speed, with 122 days to build Colossus 1’s 300MW, six months to build 200MW at Colossus 2, the decision to build an onsite generation plant 1km across the border to avoid permitting, and much more.

There’s been even more displays of speed since then. The power plant in Southaven has expanded from 27 turbines (~495MW) in February 2026, to 69 turbines (>1.2GW) in July 2026.



Source: SemiAnalysis Datacenter Industry Model; The Southaven Power Plant, February 2026 to July 2026

As well as the arrival of “MiniHard,” which upon vertical construction in March 2026, will likely reach 450-500MW in just ~5 months!



QQ:1413221435

Source: SemiAnalysis Datacenter Industry Model; MiniHard, March 2026 to July 2026

That leaves more than enough time to build many such shells by 2027. It also was his first true greenfield, so he can probably do better for the next. Another option is, of course, to retrofit. Colossus 1 and 2 have been built remarkably fast through retrofits, as explained in our xAI deep dive last year.

xAI's Colossus 2 - First Gigawatt Datacenter In The World, Unique RL Methodology, Capital Raise

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL, AND 3 OTHERS · SEPTEMBER 16, 2025

QQ:1413221435



Much has been written about xAI’s Colossus 1. The Memphis build belongs in the history books: the largest AI training cluster, erected from scratch in 122 days. With roughly 200,000 H100/H200s and ~30,000 GB200 NVL72, it remains, today, the largest fully operational, single-coherent cluster (setting apart Google,

Read full story

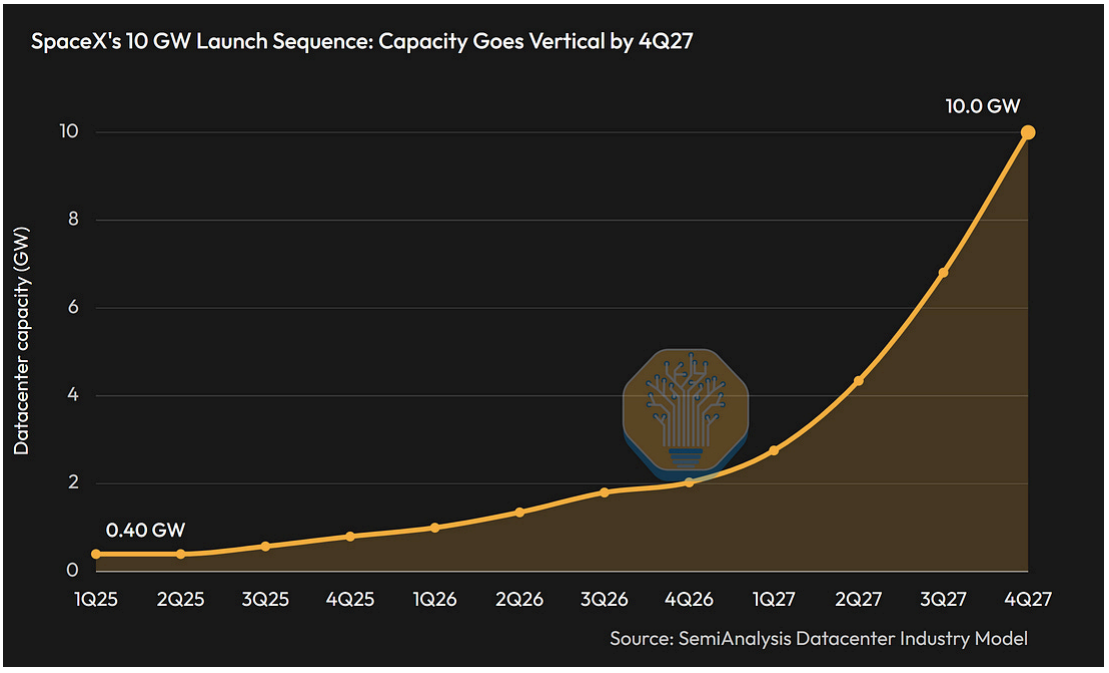
Building 10+GW in a year will be a different story. SpaceX will need to scout all over the country to find suitable land, with easy permitting and access to gas. We however believe that there are more than enough options to support a material ramp-up. This will, naturally, extensively rely on onsite gas generation – check our energy deep dives here to understand how it works and why it’s necessary.

QQ:1413221435

Beyond the paywall, we will discuss some of the sites that we suspect Elon might take.

How is this possible?

The question on everyone's mind: where and how can this happen? We believe SpaceX has likely found a handful of sites to deploy the remaining GW, and the 10GW 2027 target is within reach.



QQ:1413221435

But where will the new sites be? Our Datacenter Model subscribers got the full breakdown, but let's walk through a few fun ones.

Off the bat, we find two old warehouses, prime retrofit candidates, via liens against MZX (Elon).

Darana Hybrid, Inc. (hereinafter "Claimant"), a contractor, provided labor, services, equipment, and materials, claims, asserts, and holds a contractor's special lien pursuant to Miss. Code Ann. §85-7-401, et. seq. in the amount of $439,794.10 on the building, structure, house, factory, mill, machinery, or railroad, leasehold interest, and the premises or real estate on which it is erected or built, of ET Olive Branch V, LLC, pursuant to a contract with MZX Tech, LLC, located at 7700 Nail Rd, Olive Branch, Mississippi 38654 (the "Property"), and the Property is more particularly described as follows:

Source: SemiAnalysis Datacenter Industry Model

QQ:1413221435

These liens indicate that Darana Hybrid was performing electrical/mechanical work on the sites. One being an 863k sqft site in Olive Branch, Mississippi. This facility is owned by ElmTree Funds (Blackrock-owned since last year), which also sold the

Macrohard facility to MZX. Another is listed, but not owned by ElmTree, a 474k sqft building in Southaven, Mississippi. That size could provide ~1GW of capacity or more.



Source: Google Maps; Olive Branch, Mississippi

In wake of Elon's quiet APR acquisition, we also see the Pampa location – aiming to deploy 2GW of onsite generation – as a likely candidate. Kinder Morgan's NGPL interstate mainline passes about a mile from the Pampa site, and its just-announced, fully subscribed Panhandle expansion citing data center demand is the most likely firm-supply route for the 2GW permanent plant. No documents confirm this yet, but it seems quite possible.

QQ:1413221435
QQ:1413221435

QQ:1413221435



Source: Google Maps; Pampa, Texas



A guest post by

Muhammad Zuhair

DCmaxxing

Like Comment Restack

© 2026 Dylan Patel
548 Market Street PMB 72296, San Francisco, CA 94104
Unsubscribe

QQ:1413221435

Start writing

QQ:1413221435

QQ:1413221435