Meta Compute: Everyone Wants To Be A Cloud

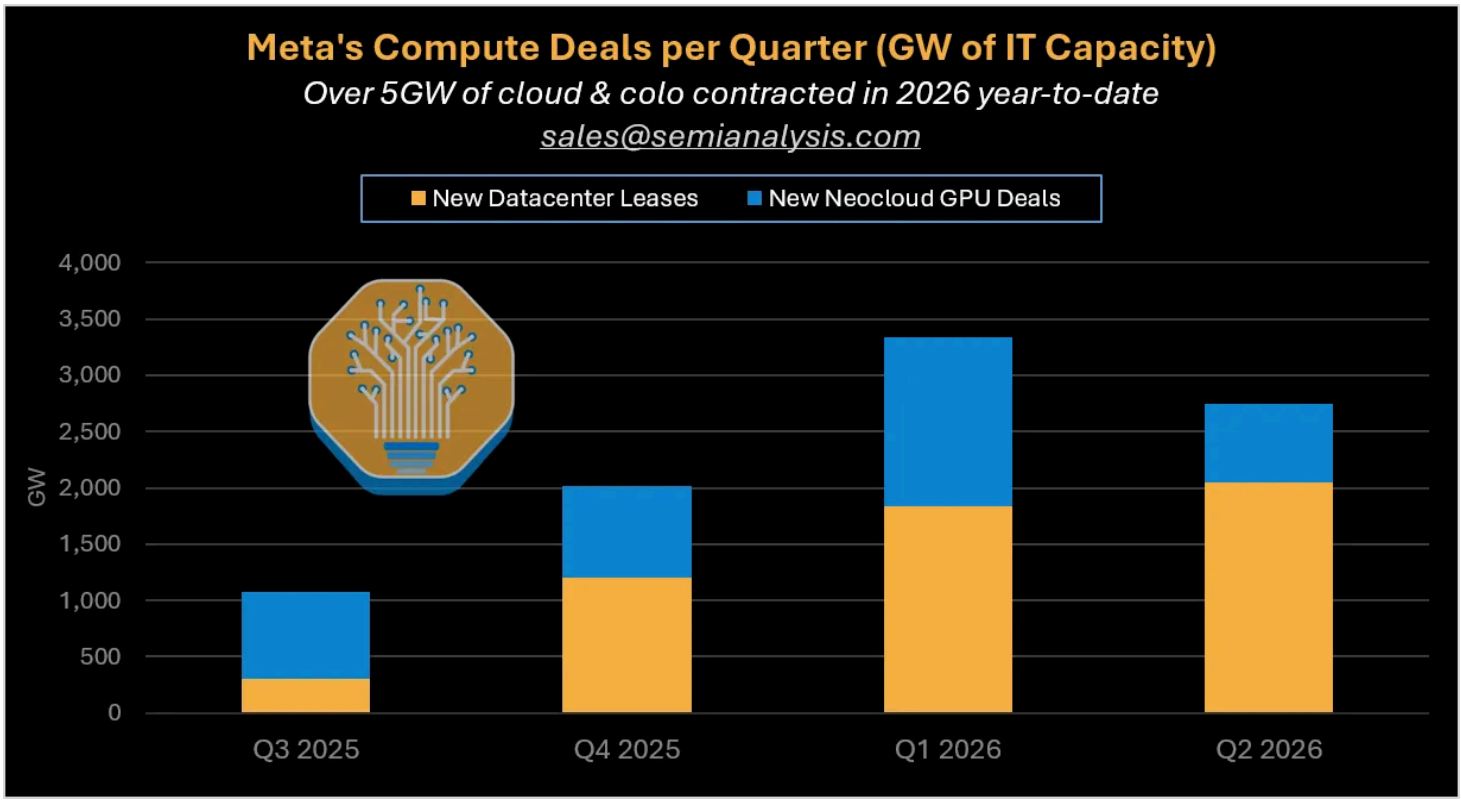
Zuck Takes Plan B? SpaceX 2.0, Bedrock 2.0, MSL Isn't Giving Up, Scaling RecSys by 10x... ClusterMAX ranking coming soon?

JEREMIE ELIAHOU ONTIVEROS, MAX KAN, JOEY BROOKHART, AND 2 OTHERS

JUL 03, 2026 · PAID

69 1 1 Share

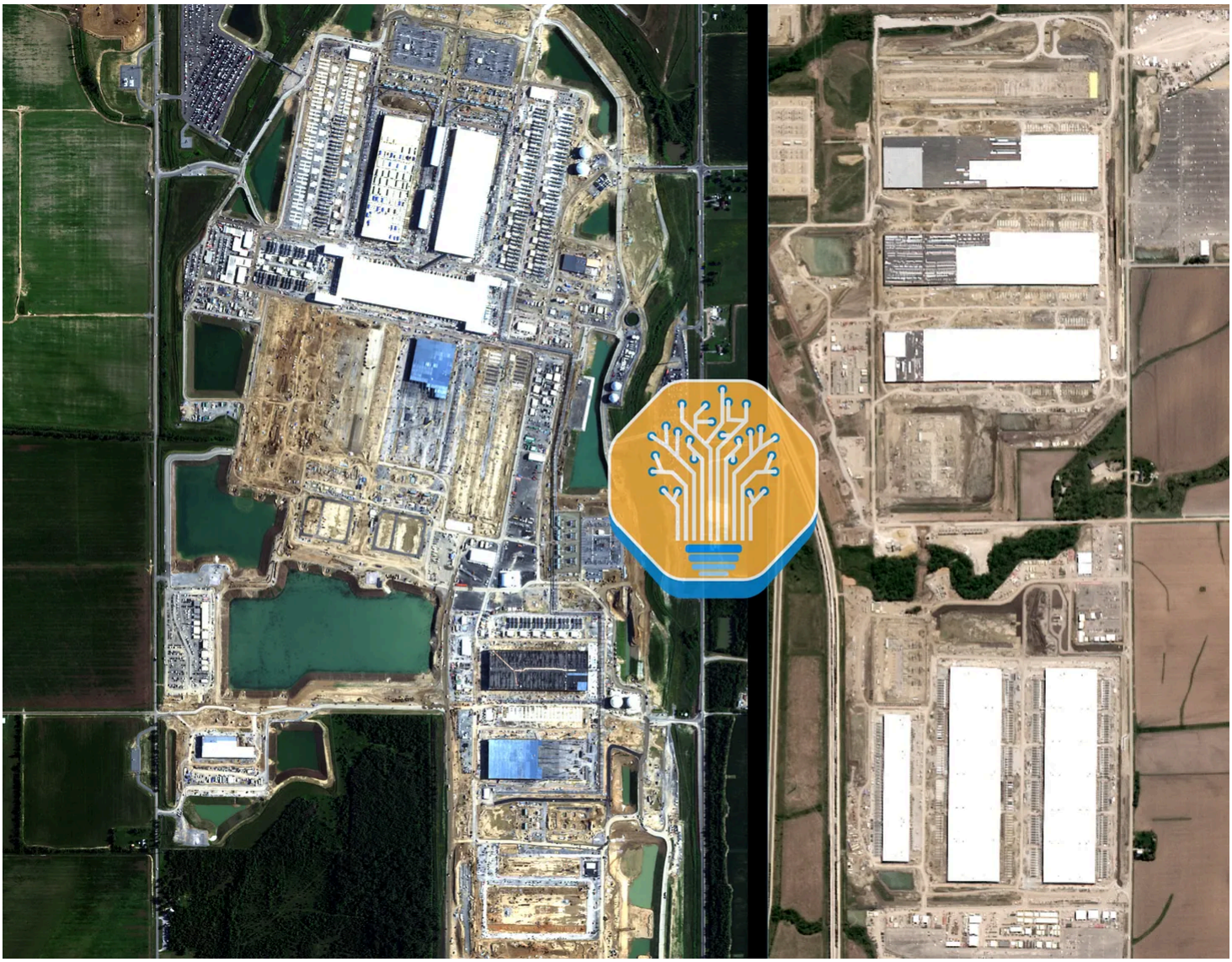
With Bloomberg headlines suggesting Meta could become a Neocloud, the market's reaction was immediate: aggressive sell-off of Neoclouds like Coreweave & Nebius, and debates of "overcapacity" coming back. Let's set the record straight – we believe that **both takes are erroneous** and that Meta's datacenter & compute procurement will ***accelerate***, not slow down. Capex in 2027 will be shockingly high. In just the first six months of the year, **Meta has contracted over 5GW of capacity across Cloud & Colo**, and that doesn't even include all their accelerating self-build activity. Everything is computer and everything is a neocloud.



Source: SemiAnalysis Datacenter Model

Meta's capacity under construction just keeps accelerating. We show below Meta's two largest campuses – these two pictures represent 2.5GW of capacity under construction!

By the way, if you believed the laughable headlines that “half of US datacenters and delayed and only 5GW are under construction”, we show you here that just two datacenters are half of this. Read our piece “Stop Saying Half of US Datacenters Are Delayed” for more on why these headlines are completely off.



Source: SemiAnalysis Datacenter Model

Of course, this naturally raises the questions of what Meta will do with this compute, and whether they’re going to flood the market with all of this supply if they turn into a Neocloud. Broadly speaking we see four major high-value use-cases, which are all differentied and very different relative to what traditional Neoclouds do:

1. Frontier AI Models: Meta has NOT given up on training frontier models. The bulk of incremental capacity still goes to Meta Superintelligence Labs, and we think the team is currently excited about their progress. A follow-up report will dive into MSL, their unique data strategy, and discuss their chances of catching up with Anthropic and OpenAI. Of course, our Tokenomics Model subscribers already know our takes and have access to all of this real-time.

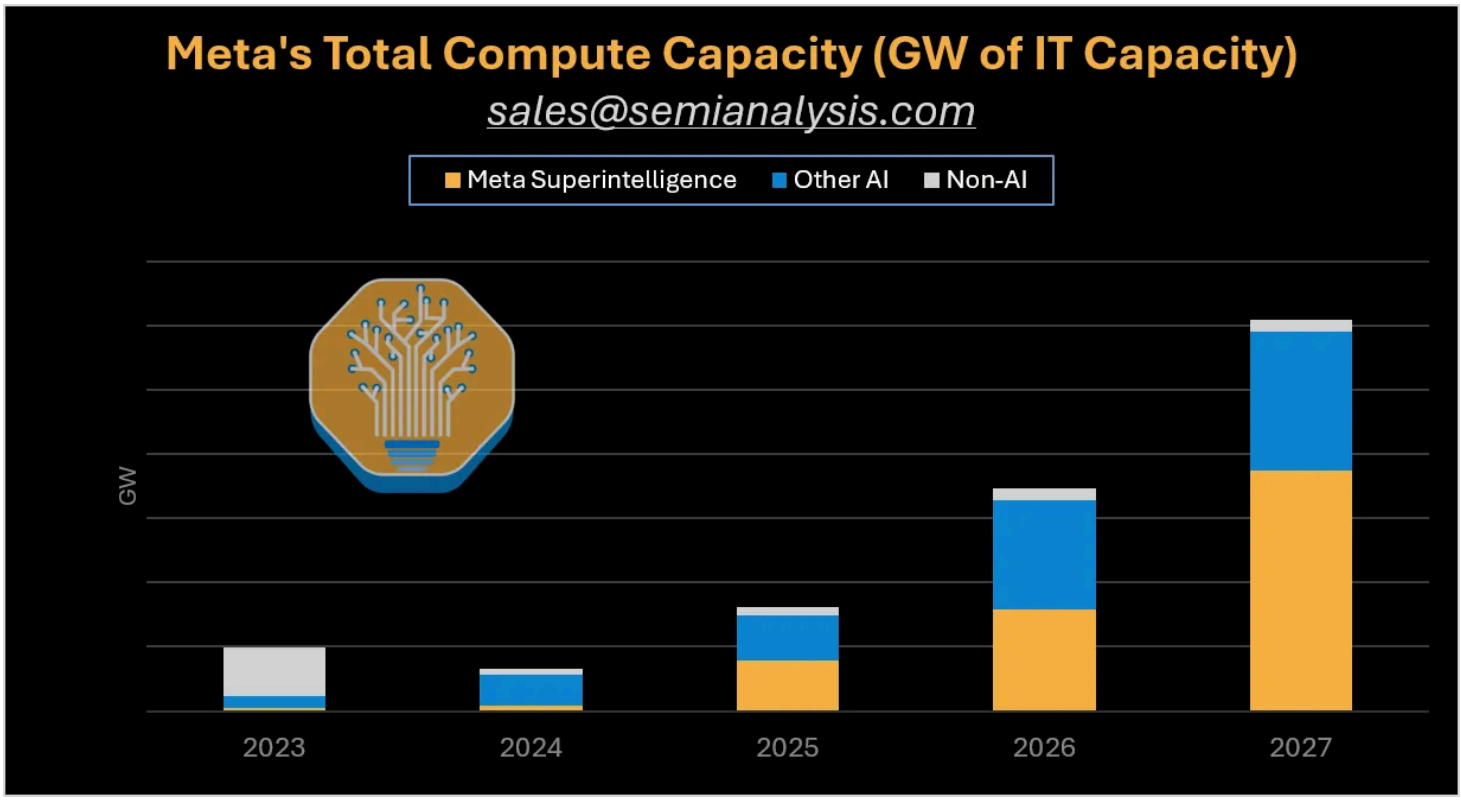
2. RecSys: We believe Meta thinks they can scale up Ads recommendation systems by >10x in complexity to accelerate revenue growth. That requires both inference & training compute for their RecSys models. They can also do more generative targeted ads.
3. (***SemiAnalysis Exclusive***) We believe that **Meta is in final talks with Anthropic to get access to private instances of Claude**. This would be akin to Bedrock, Foundry, Vertex from other hyperscalers (read our deep dive here). There are multiple use-cases for Meta, ranging from internal usage, to building the premier Sales & Marketing SaaS powered by Frontier AI Agents. We expect Meta to launch a token-as-a-service endpoint and increasingly move up the stack, leveraging its network and distribution. Initially it will be their own models externally and Anthropic internaly, but over time we believe they will serve Anthropic and OpenAI models externally
4. We expect Meta to strike a few "SpaceX-type" deals. Elon is a sales genius, and he created a **brand new market segment: large-scale on-demand compute at a huge pricing premium**. We think Meta wants in, but selectively. After all, just a couple hundred MWs can already drive >$10B of yearly revenue! We expect a ten billion dollar Anthropic deal to kick off the flywheel.

This high optionality, with four high-value-add options, makes it easy for Meta to keep aggressively contracting compute. Meta Superintelligence remains the core engine, but if it doesn't work out, there are many high-margin alternatives to sell compute. It is essentially a CFO's dream and makes it very easy to go all-in on compute – we bet Susan did a 180° flip when she saw the pricing of SpaceX compute deals! Meta won't be a normal bare-metal IaaS vendor with ~30% gross margins – all its options are high value, and enable to easily afford paying a margin to other Neoclouds in order to accelerate their fleet buildout - even if MSL doesn't work out.

Our Datacenter Model breaks down quarter by quarter their capacity additions across selfbuild, datacenter leasing, and cloud renting. After nearly 10GW of deals signed since early 2024, the bulk of their capacity additions are now through 3rd parties. We expect this to continue and believe that Meta will be a huge source of RPO growth for the likes of Coreweave, Nebius and others.

Our model also breaks down, quarter by quarter, Meta's capacity into MSL, other AI, and non-AI. While MSL's huge growth is understandable (they need to keep up with Anthropic & OpenAI!), the surge of Other AI in 2026 suggests plenty of optionality to

monetize compute across RecSys, “SpaceX equivalent”, and Bedrock-type token as a service. If RecSys scales a bit less than expected, they have other options. Of course, if MSL fails, Cloud capacity will skyrocket.

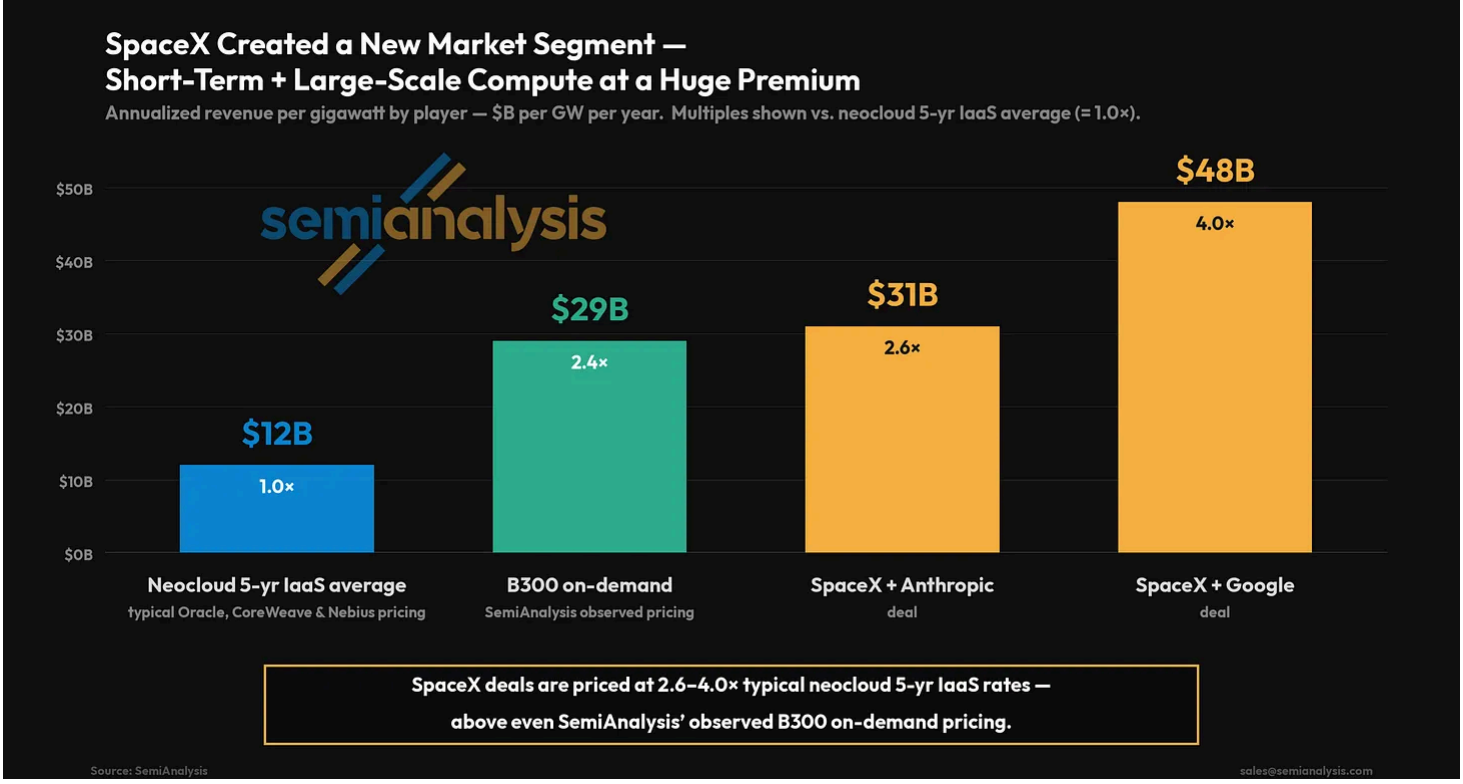


Source: SemiAnalysis Tokenomics Model, Datacenter Model

Let’s now dig into these four options. We start with the SpaceX-like deal, and the Bedrock-type ambition. Then, behind paywall, we discuss the outlook for MSL and for RecSys.

SpaceX dreams – Elon’s genius deals

Elon Musk shocked the AI infrastructure world when he announced the first Anthropic deal. He double-shocked the world when he came up with the Google deal. The reason is simple: the revenue per MW of these deals are respectively triple and quadruple what peers are charging, as shown below. The profit per MW gap is, of course, even larger given the cost structure is roughly the same.



Source: AI Cloud TCO Model

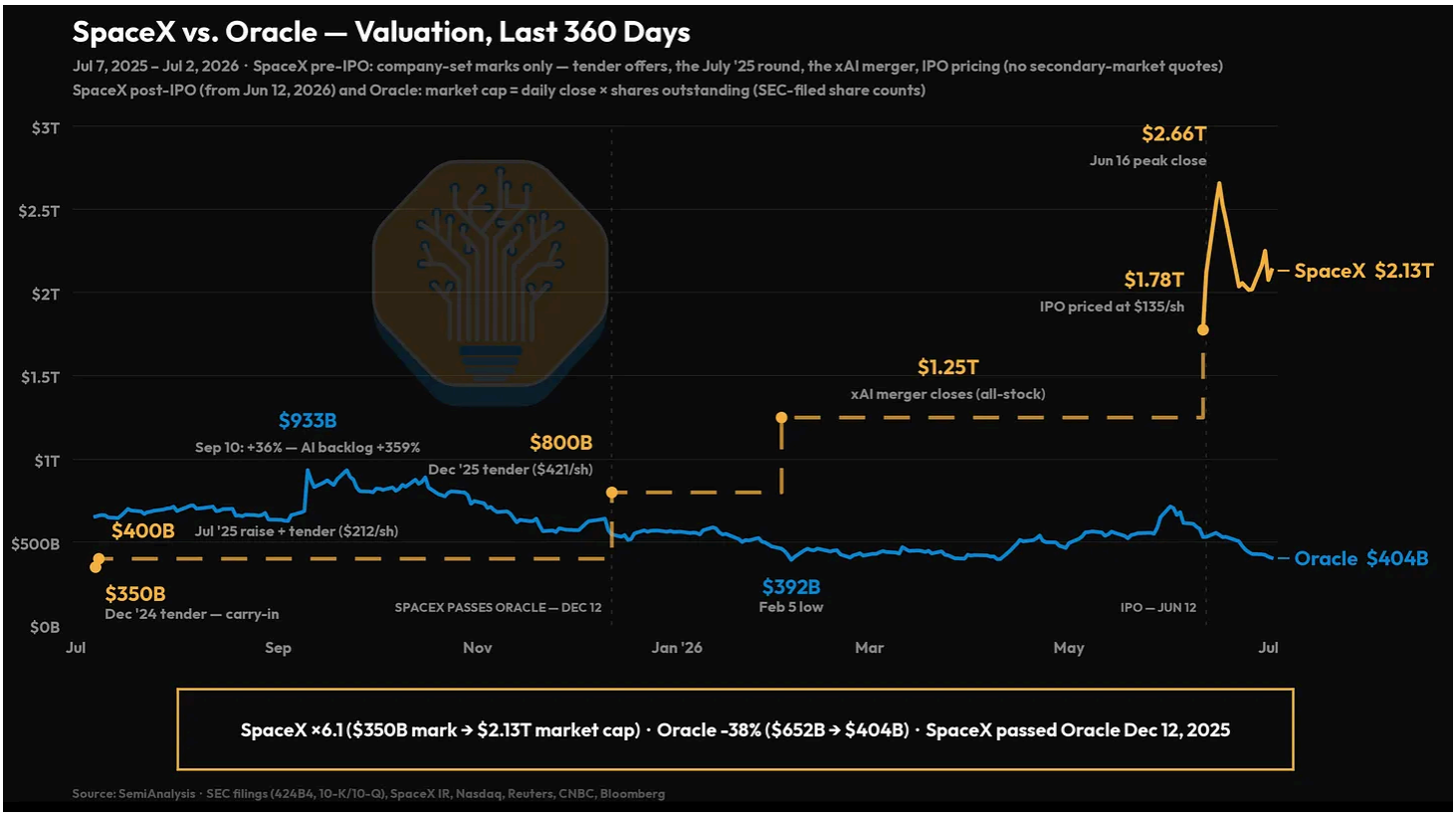
SpaceX's pricing on the Google deal is even higher than what we observe on the on-demand or short-term rental market. Elon essentially invented a new market segment. Our AI Cloud TCO model team tracks hundreds of GPU cloud deals every year, including all contract terms like SLA, pricing, contract length etc. That's why we have, by far, the world's best information on GPU pricing by contract length – see a preview in our free dashboard here.

We have never ever seen a deal that large, and that short: the deal is three years, but with option to cancel from both parties within 90 days – so it's effectively a 3-month deal with automatic renewal.

The reason this has never happened is because very few companies can do it. The financing burden excludes all the Neoclouds from participating in that market – for large clusters, they need to secure an offtaker over multiple years. The top 3 hyperscalers could've done it… but they all see higher-value long-term options, such as Microsoft getting OpenAI's IP in exchange for equity investments & compute, Amazon focused on increasing the adoption of Bedrock and Trainium, Google similarly with TPUs and Vertex (now Gemini Enterprise Platform).

This effectively leaves only two companies that can truly capitalize on this new market: Oracle and Meta. For the former, we see this as a massive blow. It's yet another piece of evidence that they could've done a much better job at monetizing all of their gigawatts of compute. A look at the valuation trajectory of Oracle vs SpaceX

illustrates the massive divergence. For both, Gigawatts are an increasing portion of their valuation, and a large driver of that evolution within the last year.



Source: SemiAnalysis, SEC, CNBC/Bloomberg/Reuters

Enter Meta. Given their laser-focus on building Superintelligence, they probably didn't put enough thought into the Cloud option. SpaceX and Elon paved the way for them. At $50B/Gigawatt of annual revenue, it becomes an easy call. Just allocating 200MW of compute to an external customer drives $10B/yr of revenue, at sky-high margin. And the ability to cancel that contract within 90 days makes it even easier – if they want to give MSL more compute, they can, in short notice.

That's also a perfect match for Meta's datacenter construction strategy. A year ago, we were the first to track their new "tent" ultra-fast datacenter design. Since then, Meta tents have been popping up all over the US! By bringing datacenters online fast, even if "lower quality", it'll be easier for Meta to monetize compute in a SpaceX fashion.



Source: SemiAnalysis Datacenter Model

As such, we expect Meta to announce such a deal soon. It's a great way to get the flywheel started. Anthropic is our prime suspect, but others could step in as well, such as OpenAI or Google.

Will Meta build the new Bedrock?

Another great option for Meta is to strike a deep partnership with frontier labs to sell their models with Meta's own compute. As said earlier, we believe they are in final talks with Anthropic to get private access to their LLMs, just like Amazon has through Bedrock (agreement explained in detail here). This means that for Meta, another way to monetize their compute will be to sell Claude. We see three main paths:

1. Part of it could be for internal usage only. Meta needs Claude tokens, and Anthropic can't keep up with demand. There are also security & privacy layers. In the future, we could see other very large enterprises striking such deals, e.g. a JPMorgan probably won't go all-in on Claude if they don't have the security and privacy guarantees that private instances enables (in their own datacenters).

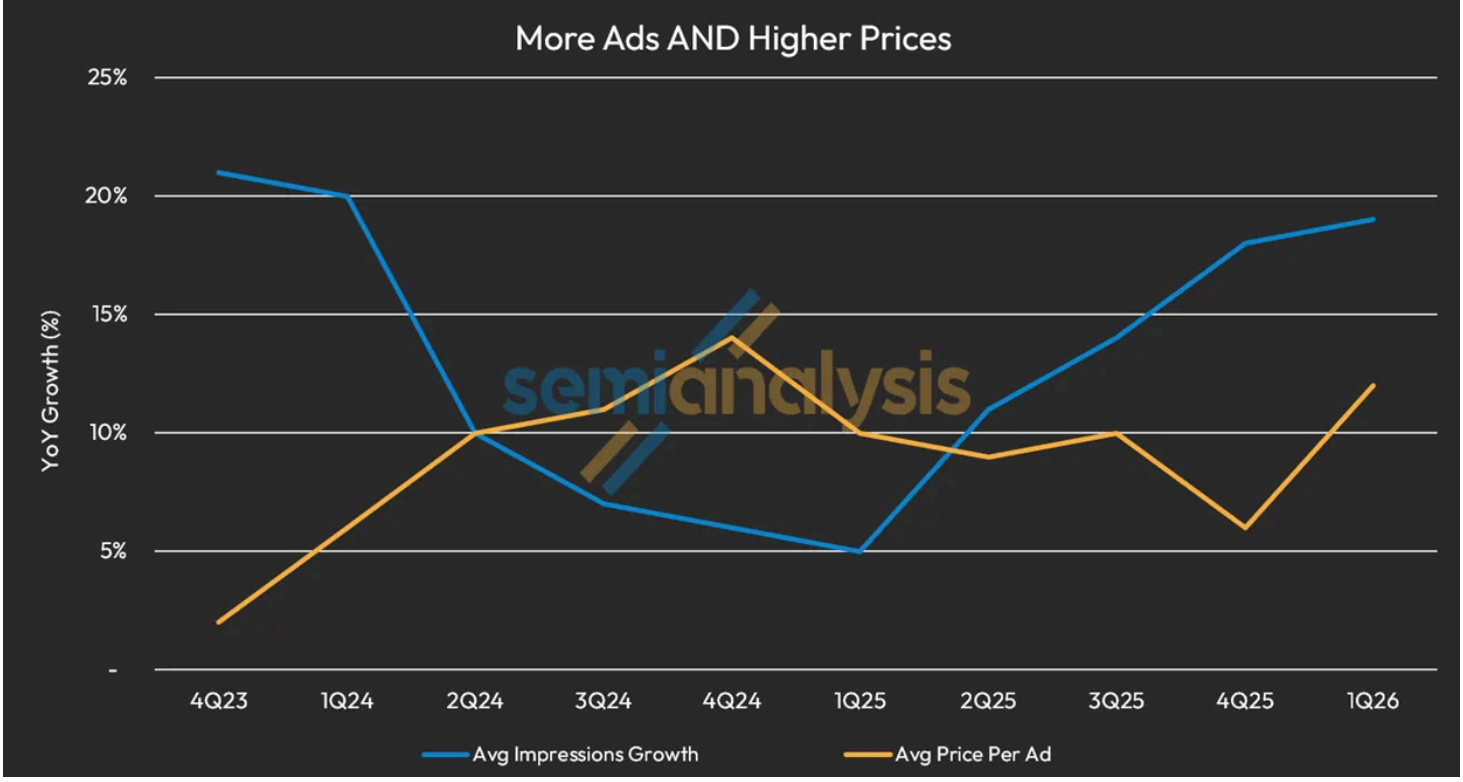
2. Meta could sell Claude as a service in the same way Bedrock does. They have capacity and own the full stack, from CPU to GPU to networking, with high security. However, as a new entrant, it'll be tough to create all the enterprise relationships that AWS has. But Meta has the option to leverage their customer base of advertisers and better integrate frontier agents and LLMs into their suite, creating a new large-scale distribution path.
3. Meta could go even more vertical and start building applications. As one of the world's largest advertising platforms, they have a path to building a Sales & Marketing powerhouse. Integrating frontier models and agents would increase their odds of building a world class solution.

There is, of course, also the opportunity to distribute models to free social media users, and to the broader Meta ecosystem including things like connected glasses. It appears more likely that they will prioritize use of their own models, but it's good to keep the optionality, and they likely have a great path to monetization. The huge distribution potential and network effect means that the likes of OpenAI and Anthropic are likely seeing this as highly strategic, and ready to make concessions to get a piece of the pie.

Scaling Ads Recommendation systems: how much can Meta accelerate?

For Meta, the key AI story is how RecSys is contributing to revenue acceleration at massive scale. At the end of 2022 and early 2023, the market widely saw the company as entering a stage of maturity and low growth amongst an investment cycle. Today, it's crystal clear that Meta has dramatically re-accelerated revenue growth, in large part due to GPU investments. They have been a key trigger, both on training and inference: RecSys models for ads have been getting bigger and more expensive to run, but much smarter, driving better yields for advertisers as seen in the ability for advertisers to pay higher prices while still seeing strong ROAS. At the s RecSys models for content have led to more time on platform across th Apps, improving the monetizable surface area and leading to strong ad growth.





Source: SemiAnalysis Tokenomics Model

The key question on everyone's mind is now – is this sustainable? How much can Meta accelerate core revenue growth? While Meta is playing catching up on the frontier lab side, the non-Meta SuperIntelligence AI chip fleet is producing outstanding ROI. We cover this split compute split between RecSys AI, Meta SuperIntelligence, and other Meta LLM products in our Tokenomics Model.

Behind paywall, we discuss the outlook for RecSys scaling, and then for Meta Superintelligence Labs.

Prior to GEM, ad ranking ran on DLRMs which don't follow the same exponential power laws seen in LLM and therefore failed to scale with compute. Meta's generative recommender work, Hierarchical Sequential Transduction Units (HSTU), changes that by reformulating ranking as a sequential prediction problem that scales in performance with compute. It beat prior baselines by ~66% on ranking metrics and is productionized as GEM, Meta's ads foundation model. Benefits of GEM includes:

· A scalable model that is 4x better at driving ad performance gains and compute efficient per unit of performance gain

· A new training stack that has 23x effective training FLOPs with ~1.4x increase in MFU using 16x more GPUs

We think Meta can profitably absorb a >10x increase in AdRec compute with this new model because each additional gigawatt correlates directly to better predictions, both on the training and inference side. This improvement is showing up through an increase in impressions delivered and average price per ad. In Q1 2026, ad impressions grew 19% YoY and average price per ad grew 12% YoY. This is a continuation CFO Susan Li has attributed their increased average price per ad to improved ad performance and targeting capabilities. After doubling GEM's training GPUs, ad conversion rate on Instagram increased on both Instagram and Facebook by 5% and 3%, respectively.

What makes the growth sustainable is that the higher prices are justified because ROAS is growing directly with CPMs. Meta Advantage+ Shopping delivers 32% higher ROAS with 17% lower cost/action relative to manual campaigns. Meta can grow ad revenue in one of two ways: (1) by showing more ads, or (2) charge more per ad. But there's a limit to how many ads it can show in a users' feed before it does more harm to the user's experience than good to ad revenue growth. As long as ROAS continues to grow with price, which has shown to be the case, Meta can keep lifting yield even as ad impressions growth eventually start coming up against ad-load ceiling.

Meta Superintelligence labs – another dropout, or a real threat to Anthropic and OpenAI ?

Compute is the lifeblood of every AI research org. Anthropic would never sell compute to any of its competitors, and it's therefore tempting to view Meta's decision as admitting defeat for MSL.

However, we believe this framing is too simplistic. There is a world where you can temporarily sell compute to your fiercest competitors AND still become a true frontier lab. Like so much else we've covered in this article, Zuck really needs to thank Elon for showing him this playbook.

We believe the following is the most likely explanation for the SpaceX/Cursor acquisition:

- Elon and Cursor are both still serious about building superintelligence. Cursor's messaging over the past 6 months has changed from "we want to build the best AI coding experience" to "we want to build RSI". This is the ethos of a frontier lab, not a wrapper startup.
- SpaceX still has up to 900MW left after their eye-watering compute deals. This is sufficient for taking a few true shots on goal at reaching the frontier. Both GPT 5.6 and Fable 5 were likely trained using less compute.
- The 90-day bilateral cancellation in all the SpaceX compute deals is crucial! Many were viewing this as a win for Anthropic/Google---because they can just cancel the deal if they find something cheaper---but it also means Elon can claw back all his compute if the Cursor team does well!
- In other words, SpaceX/Cursor still meet the minimum requirements for training a frontier model. Over the next 6 months, we will see if they can deliver on that potential. If they do, they will claw back all their compute and continue scaling to push the frontier. If they fail, then they'll give up for real and become a very large and profitable CSP instead. That's certainly not a bad backup plan, and the only true cost will have been the unmonetized GPU hours spent chasing the dream of RSI.

A future post will cover all of this in more detail, but we think Meta is taking the same approach with MSL. The only difference is that they have even more options to monetize the excess compute in the meantime! RecSys and TaaS compute can essentially be repurposed back to research at any time.

Compute, data, and talent are the three pillars of a frontier lab, and unlike the other hyperscalers, Zuck has put up the capital to try becoming world-class at all three. Success is far from guaranteed, but it'd be a mistake to count them out. The one thing that'd change our minds is if they sign a compute deal without any sort of SpaceX-style early cancellation clause. If that happens, then MSL is actually cooked.



Recommend SemiAnalysis to your readers

Bridging the gap between the world's most important industry, semiconductors, and business.

Recommend



69 Likes · 1 Restack

Previous

Discussion about this post

Comments **Restacks**

Write a comment...

Eddie Eddie 1h

100% agree. Thanks for such a quick and timely piece.

LIKE REPLY SHARE

© 2026 Dylan Patel · Privacy · Terms · Collection notice
Substack is the home for great culture