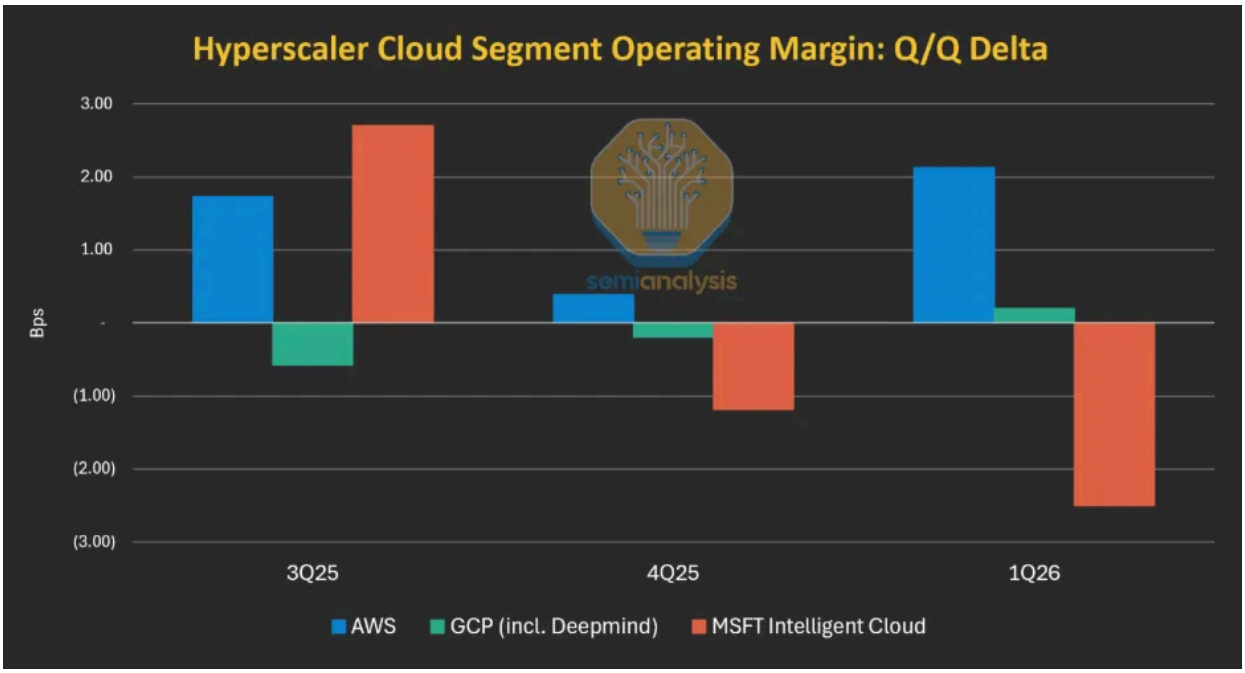
Anthropic Growth and Bedrock Mix Drive AWS Margins Higher While Peers Lag

Amazon's Bedrock Mix and Anthropic Deal Terms Combine to Show Greater Operating Leverage

JEREMIE ELIAHOU ONTIVEROS, JOEY BROOKHART, CRYSTAL HUANG, AND DYLAN PATEL

MAY 28, 2026 · PAID

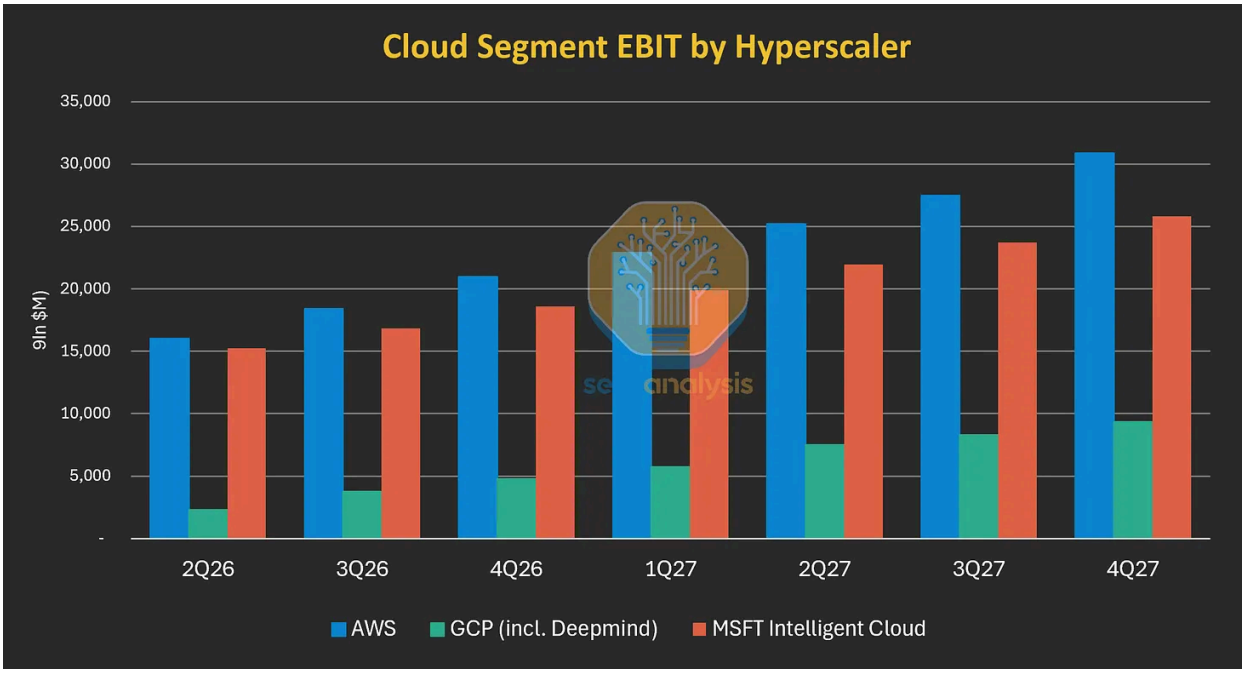
While other CSPs have seen declining-to-flat operating margins over the last several quarters, Amazon's AWS margins inflected this past quarter driven primarily by customer spending growth on Claude through Bedrock. AWS' higher share of 3P model API spend, Anthropic/Bedrock deal structure, and Anthropic's ARR outperformance in 1Q26 all contributed to EBIT margins increasing 213bp Q/Q while other CSPs lagged. SemiAnalysis' work in the new Tokenomics 2.0 model shows how AWS has pulled ahead of the pack and found a strong avenue to grow margins. Our model estimates quarterly revenue, profits, ROIC and compute requirements of every single business vertical of hyperscalers and AI Labs, e.g. Gemini API revenue & margins, Microsoft Copilot ARR, OpenAI ChatGPT subscriptions across plans, etc.



Source: SemiAnalysis Tokenomics Model

Although all CSPs are benefiting from increased AI and non-AI revenue, margins are a whole different story. Oracle and Coreweave both disappointed the market with lower-than-expected profits from their cloud arms. Azure is also seeing a downward trend in

margins. Google Cloud has had a great upwards climb recently, but margins are inflated since they do not include training costs from DeepMind in the GCP segment. The only CSP with a true rising trend is AWS – a remarkable achievement considering their server depreciation (5yrs) is the lowest of all CSPs.



Source: SemiAnalysis Tokenomics Model

The Amazon Story & Background

We believe that Amazon's margin success rests on a differentiated strategy that will be exploited further in coming quarters and years. The firm was late to wake up to the AI opportunity (we were the first to call out their leadership loss in 2023). Two years later, we were again the first to call out their change in trajectory, an upcoming revenue acceleration, when all the market was labelling them as an AI loser.

Amazon's AI Resurgence: AWS & Anthropic's Multi-Gigawatt Trainium Expansion

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL, AND 2 OTHERS • 2025年9月4日

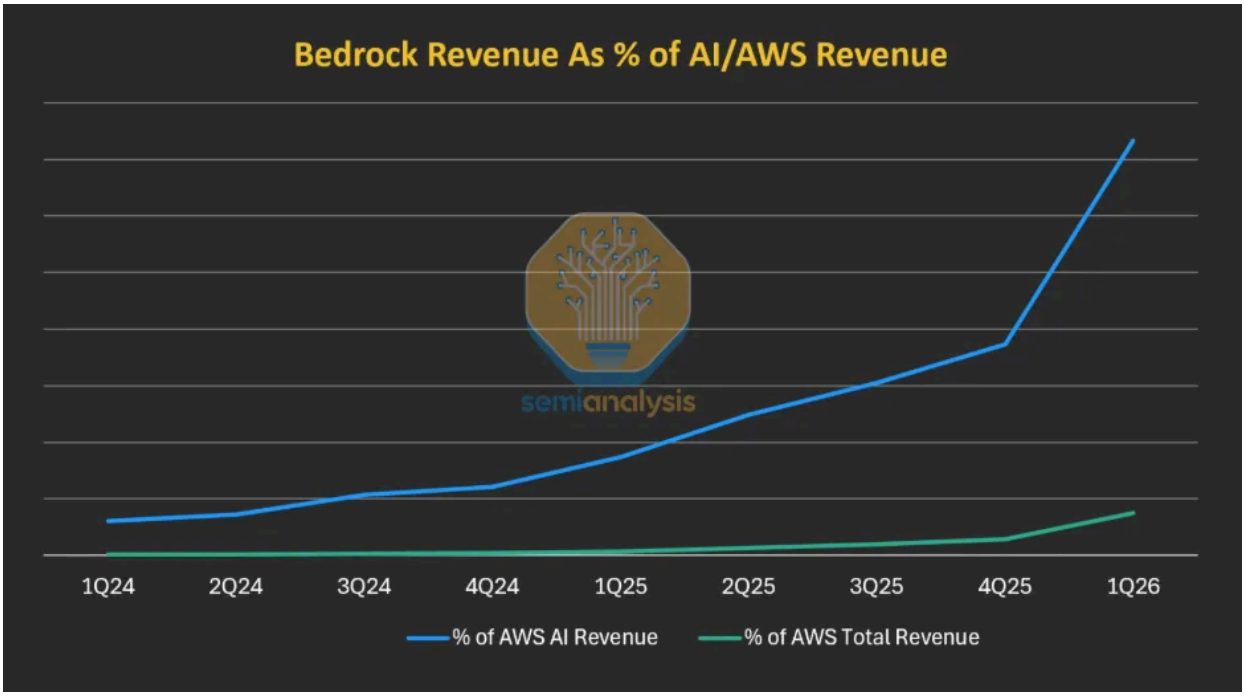
Read full story →

Now, we see a new era for AWS where the firm combines accelerating revenue growth AND outperforming margins. Amazon brings a unique combination of the following:

- Risk appetite: winners in the AI infrastructure landscape are not afraid of putting their balance sheet at work. As our Datacenter Industry Model demonstrates, Amazon has secured more power than any other cloud provider besides Google,

understanding before others that energy drives market share in this constrained environment, and that requires capital and multibillion dollar PPAs.

- Business Model: Amazon is the only CSP with token-as-a-service being the dominant share of its AI business, while all others are focused on multi-year IaaS deals. That demonstrates a higher risk appetite but also a better understanding of the unique opportunity provided by Bedrock, as we'll detail below.



Source: SemiAnalysis Tokenomics Model

- Scale & speed: No other provider will build more capacity than AWS in 2025, 2026 and 2027, as per our Datacenter Industry Model. AWS dwarfs rivals. Not only did the CSP procure a lot of power, it has also executed much more rapidly than peers and is rolling out a new datacenter design that will exacerbate its speed advantage.

- Vertical integration: we were first to cover the growing CPU constraints in December 2025 and explain the coming CPU surge driven by Reinforcement Learning and inference. Amazon is the best positioned CSP, with its custom chip Graviton providing better economics than merchant solutions. In the AI Accelerator market, Amazon hopes to replicate the same success, and is seeing good results with Trainium. As explained in our Deep Dives, Trainium is attractive for inference and RL workloads.

Let's dig in, starting by covering the economics and market drivers of Bedrock, Amazon's most differentiated product. We then dive into their datacenter footprint

relative to others, and then at the end of the report dive into the outlook for other CSPs.

Amazon Bedrock Deep Dive

Bedrock is an AWS service that enables customers to choose their favorite LLM among many options, benefit from AWS security and compliance and unified billing (among other things) and run AI workloads. This market, which we call "API endpoints", has many competitors, including Microsoft Foundry and Google Gemini Enterprise Agent Platform (previously Vertex), as well as many providers focused on open-source models such as TogetherAI, Fireworks, Baseten, etc.

Endpoints typically claim to be differentiated through the following items:

- Model library breadth: providers like to flex the number of LLMs their platform has available.
- Price: some providers have a differentiated inference stack, or cost structure, which enables them to offer more attractive prices while keeping margins viable.
- Interactivity: some vendors have better metrics, e.g. higher token throughput, lower TTFT, etc.

While some of these criteria do matter, they miss the single most important differentiator: access to frontier LLMs. As demonstrated by our Tokenomics Model, Frontier LLMs make up the vast majority of AI API industry revenue.

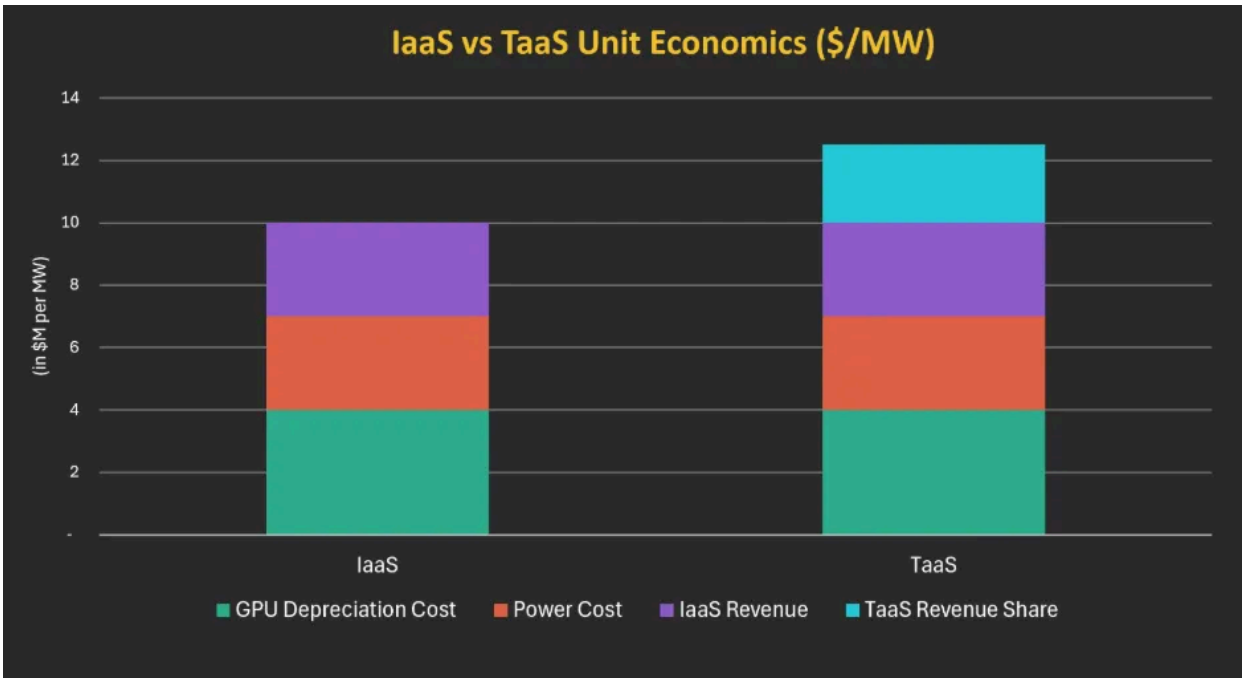
In the API endpoint market, this is the massive advantage that AWS, Microsoft, and Google boast over everyone else. For a long time, AWS had access to Claude, Google to both Claude and Gemini, and Microsoft to OpenAI models. Recently, AWS gained OpenAI access, while Microsoft gained Claude. No other CSP currently has the ability to sell OpenAI, Claude, and Gemini tokens.

Having access to these models is one thing, but building a substantial business around them is another. AI Inference notably has huge compute needs. To understand this, let's dig into the economics of Bedrock, Vertex, and Foundry.

Token-as-a-Service Platform Economics

The economics of TaaS platforms are very different if they own the IP (or can freely use it, eg open source), versus if they "distribute" the IP:

- IP ownership: the economics are the same as that of an AI Lab. The cloud/token vendor has a fixed cost which is the infrastructure. That cost is largely driven by GPU depreciation, margin of the CSP, datacenter costs, and electricity costs. Revenue is a function of tokens sold: to make good money, token pricing needs to be high enough, and hardware must be well utilized. Volumes and pricing must be large enough to absorb fixed costs and make some margin.
- Model distribution: in this example, Amazon sells Claude tokens to an Amazon customer that pays an AWS bill. However, the seller of record is Anthropic. Public terms on Bedrock are clear: the product is sold by Anthropic, and use of the model is governed by Anthropic's terms. However, customers are invoiced by AWS, and the terms state that the model is "Deployed on AWS". In practice, this means that:
- As Seller, Anthropic books full revenue of the sold tokens.
- As computer and marketplace provider, AWS gets both an infrastructure fee (akin to an EC2 IaaS fee) and a distribution or revenue share fee. The latter is what boosts margins and makes selling Claude on Bedrock a highly attractive business to AWS.



Source: SemiAnalysis Datacenter Model

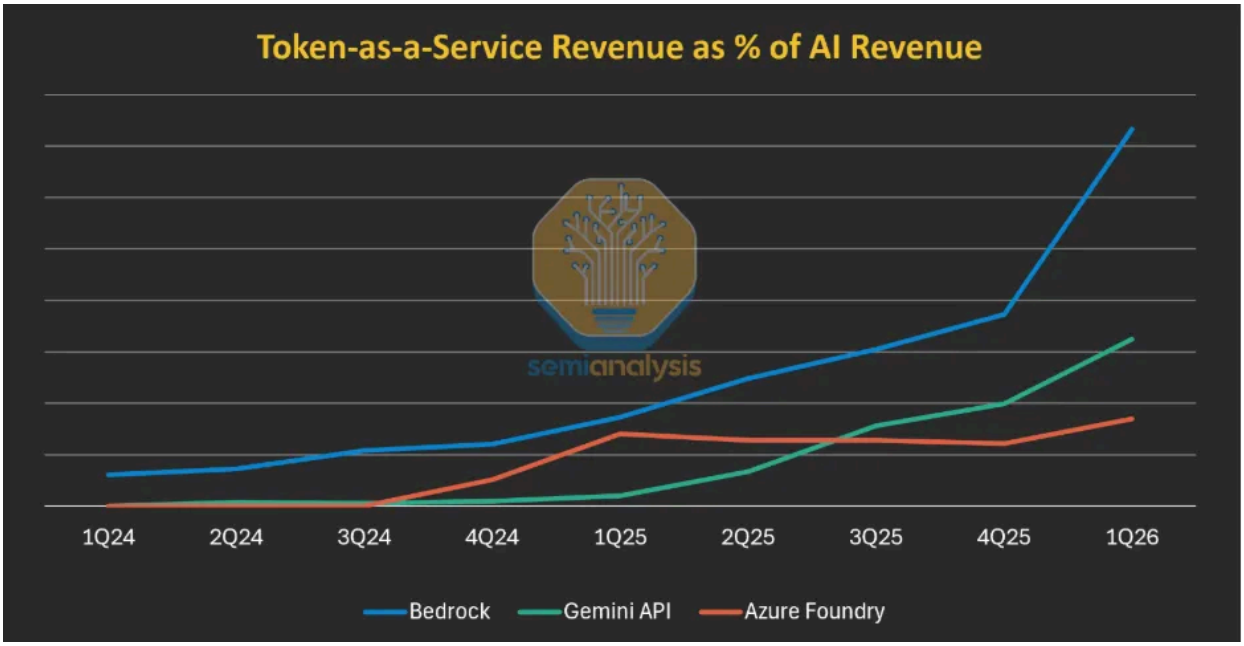
Anthropic was the first AI Lab to roll this out with AWS and Google, recently followed by OpenAI with AWS. Security is of maximum importance: ideally, the CSP doesn't

have access to model weights, or a very limited one, but can still use the model and run it on its infrastructure.

For the AI Lab, there are two main benefits:

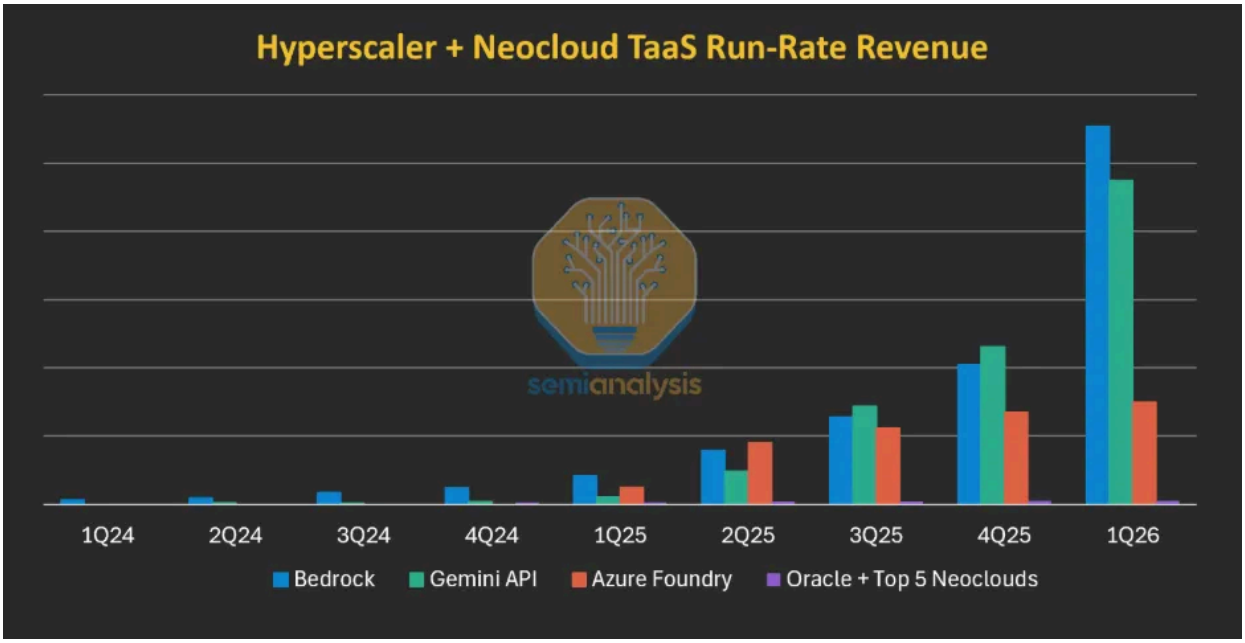
- Access to the CSP's customer base
- Access to compute capacity without the need of an expensive multi-year IaaS contract, but at a higher cost

For the CSP, this means more risk relative to a standard GPU IaaS deal, since revenue is not guaranteed via a 5yr IaaS take-or-pay deal. However, the margins are significantly more attractive. And Amazon with its deal structuring and execution has been the main beneficiary of its bet on Bedrock, dwarfing the size of their rival platforms.



Source: SemiAnalysis Tokenomics Model

And this is the main difference for the AMZN, MSFT, and GOOGL against ORCL and neoclouds. Their Token as a Service (TaaS) businesses are $10B+ in ARR today. Contrast this with neoclouds and Oracle at practically nothing. As we show in the Bedrock/Anthropic deal section, margins on these TaaS deals are significantly higher than being a wholesale seller of AI IaaS rental compute. The margin mix and distribution advantage these hyperscalers now possess is significantly widening the gap amongst the top hyperscalers and the rest of the field.



Source: SemiAnalysis Tokenomics Model

Vertical Integration Positions AWS for Industry-Leading Margins

Having a greater business mix towards Bedrock is a great combination with Amazon's bet on custom silicon. In prior reports, we've covered Trainium2 and 3 in great depth, arguing that the chips have leading perf/TCO in certain scenarios where memory bandwidth is most valuable, such as high-batch inference and RL.

In Bedrock, the underlying hardware is abstracted from customers; they only see tokens. This means that an inference-optimized chip is a natural fit with Bedrock, with the main caveat being that porting models can be longer than with Nvidia GPUs, but it's typically just a matter of days. This means that Trainium has a natural offtake in the form of Bedrock, and Amazon is taking full advantage of this, per the CEO:

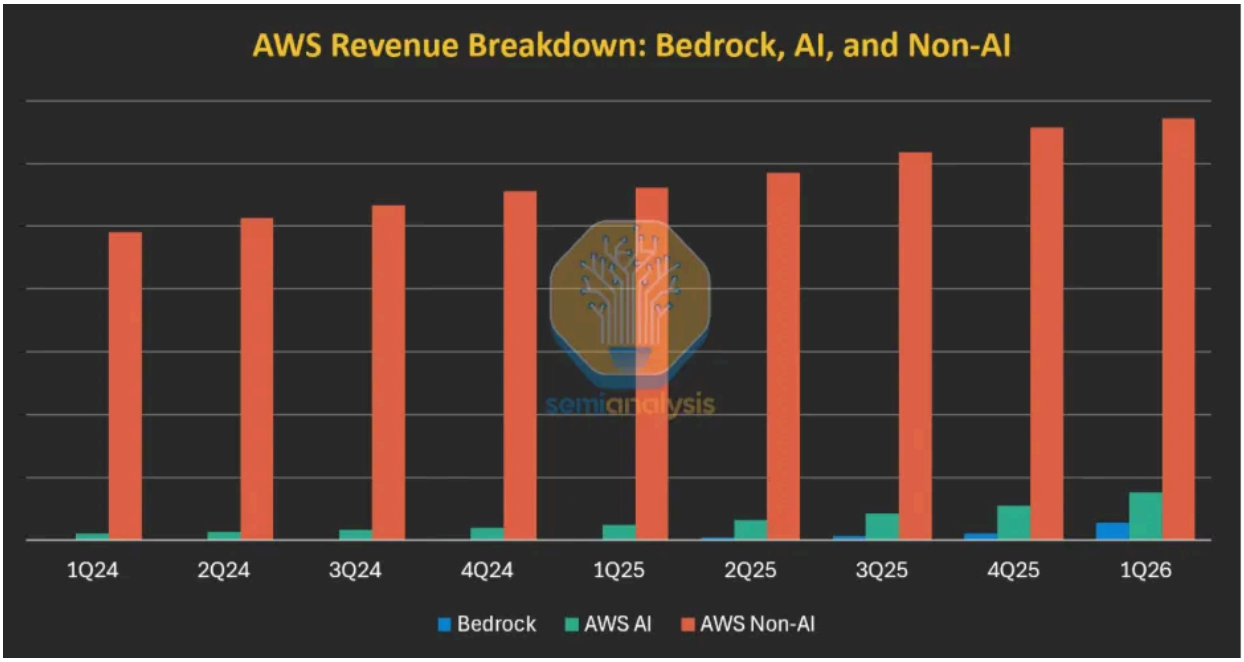
"Our AWS Trainium chips, designed in-house for AI workloads, now power more than 50% of Amazon Bedrock token usage." - Matt Garman, AWS CEO November 2025

Google benefits from the same advantage, but there are caveats, as we'll discuss below.

In addition to AI Accelerators, CPUs are also increasingly relevant to training and inference of frontier LLMs. In the world of CPUs, vertical integration is an even bigger differentiator, and Amazon's fifth generation custom chip, Graviton, is industry leading. We wrote a full deep dive on CPUs here, and we explained that Graviton4 and 5 provide a great perf/TCO advantage to Amazon. Graviton4 will be integrated to Trn3 as head node and will be used separately to handle RL workloads and agentic. Amazon

now has large CPU and Graviton deals with Anthropic, OpenAI, and Meta. Upselling Graviton to Bedrock customers will also be easier for Amazon. In our

As discussed above and as we dive deeper into the deal economics below, AWS' Bedrock mix vs Azure Foundry and GCP Gemini API is important. Bedrock has steadily grown as a % of total AWS AI revenue to 37% today, up from 9% in 1Q25 when IaaS dominated the AI business. In comparison, AI IaaS is still 80%+ of the AI business mix at Azure and GCP. Per Amazon's 1Q26 and 4Q25 earnings call, Bedrock revenue grew 170% and 60% Q/Q, respectively. SemiAnalysis believes Bedrock is a $5.5B run rate business today with the vast majority of customers (80-90%+) using Anthropic models.



Source: SemiAnalysis Tokenomics Model

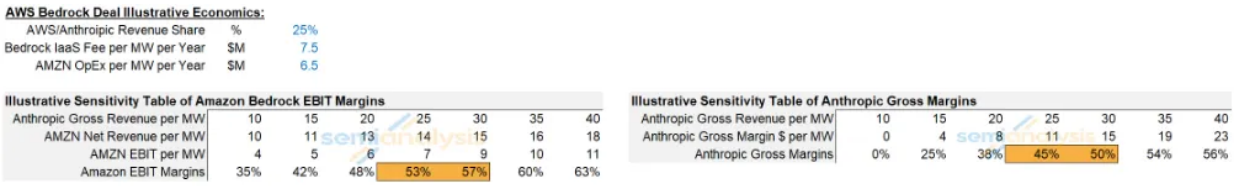
Clearly, Amazon has found a winner in terms of revenue growth and mix with Bedrock and the Anthropic deal. Microsoft, in contrast, is heavily IaaS driven, and other AI efforts like CoPilot in 365 and GitHub have failed to create traction. Google has done well with Gemini API within Gemini Enterprise Agent Platform, but has not benefited from the same coding market trends that have helped drive Anthropic's API revenue up ~13x y/y.

Bedrock/Anthropic Deal

Our understanding of Anthropic's Bedrock deal includes a flat IaaS fee, revenue share, and performance hurdles for outperformance above certain levels of token/spend throughput.

The Tokenomics team models Anthropic's and OpenAI's businesses extensively by segment (Consumer, Business/Team, Enterprise) and plan (Free, Pro, Max20, API, etc)

at an ARR, revenue, margin, token volume, and token pricing level to help investors and corporate strategy departments understand this flow of dollars and tokens between customers, labs, and CSPs. This work includes modeling the Anthropic and Bedrock deal, which we have previewed below to illustrate the Gross Margins of Anthropic and EBIT Margins of Bedrock at differing levels of Anthropic ARR per MW on Bedrock Compute. We illustrate that margins for Bedrock, or any Token as a Service product, have superior margins than the typical IaaS fee business. The takeaway for hyperscalers is that these TaaS arrangements are excellent margin accretive additions when structured beneficially for both the hyperscaler and lab to drive revenue higher.



AWS Bedrock Deal Illustrative Economics:

AWS/Anthroipic Revenue Share	%	25%
Bedrock IaaS Fee per MW per Year	$M	7.5
AMZN OpEx per MW per Year	$M	6.5

Illustrative Sensitivity Table of Amazon Bedrock EBIT Margins

Anthropic Gross Revenue per MW	10	15	20	25	30	35	40
AMZN Net Revenue per MW	10	11	13	14	15	16	18
AMZN EBIT per MW	4	5	6	7	9	10	11
Amazon EBIT Margins	35%	42%	48%	53%	57%	60%	63%

Illustrative Sensitivity Table of Anthropic Gross Margins

Anthropic Gross Revenue per MW	10	15	20	25	30	35	40
Anthropic Gross Margin $ per MW	0	4	8	11	15	19	23
Anthropic Gross Margins	0%	25%	38%	45%	50%	54%	56%

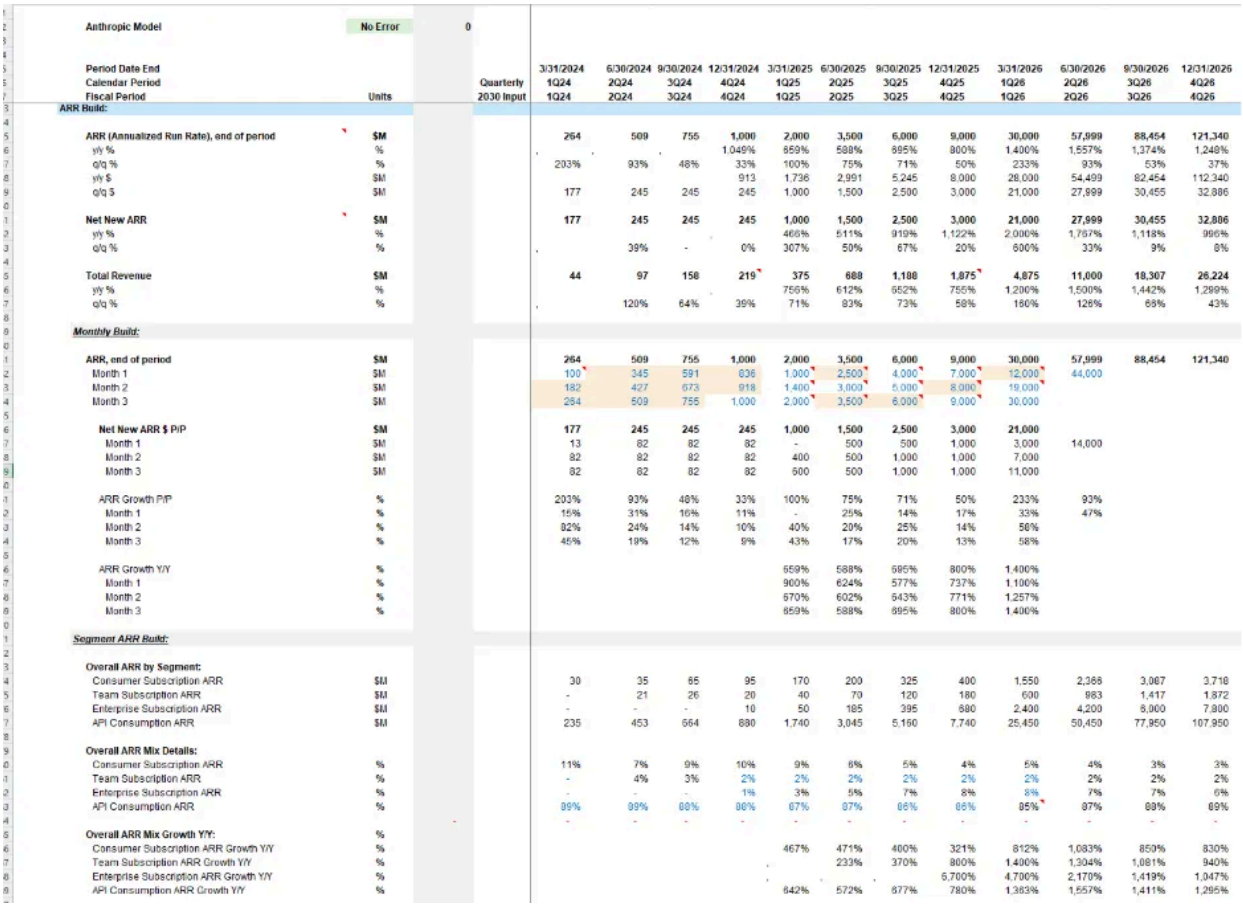
Source: SemiAnalysis Tokenomics Model

We believe that Anthropic's recent revenue outperformance (discussed below) puts Anthropic Bedrock revenue around $26M per MW in 1Q26, implying Bedrock EBIT margins of ~55% in the above analysis. For 1Q26 vs 1Q25 at AWS, we believe that reported incremental EBIT margins were primarily driven by Bedrock, with Bedrock accounting for 30% of the step-up in Gross Profit Dollars Y/Y despite accounting for only 4% of total AWS revenue. For 2Q we see Anthropic ARR per MW of Total Compute around $42M and Bedrock Revenue mix increasing to 53% of AWS AI Revenue contributing an incremental 9 points to total AWS revenue growth. We note that given Anthropic's ARR recognition accounting on a gross basis, that margins on Bedrock are slightly negative mix vs their current blended low 60% inference gross margins.

Anthropic's Blowout Q1

Anthropic's 1Q26 was characterized by rapid growth which helped propel Bedrock revenue and margins higher. The company added $21B in net new ARR in the quarter to reach $30B of ARR. Most of Anthropic's revenue is API and Enterprise based whereas OpenAI skews consumer and has a relatively heavy inference cost load from free users. This revenue explosion is driven by Claude Code, which has taken enterprises by storm. Consumers, too, have begun to flock to Claude. In a recent note to subscribers, we noted Anthropic's increasing (and now majority) market share of net

new customers vs OpenAI within card panels as well as higher average transaction values. We continue to see strong demand for Anthropic - primarily driven by enterprise API spending - through the end of the year. Our latest Tokenomics model estimates see potential for well over $100B in ARR by year.



Anthropic Model	No Error	0											
Period Date End			3/31/2024	6/30/2024	9/30/2024	12/31/2024	3/31/2025	6/30/2025	9/30/2025	12/31/2025	3/31/2026	6/30/2026	9/30/2026	12/31/2026
Calendar Period		Quarterly	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
Fiscal Period	Units	2030 Input	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
ARR Build:														
ARR (Annualized Run Rate), end of period	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000	57,999	88,454	121,340
y/y %	%					1,049%	659%	588%	695%	800%	1,400%	1,557%	1,374%	1,248%
q/q %	%		203%	93%	48%	33%	100%	75%	71%	50%	233%	93%	53%	37%
y/y $	$M					913	1,736	2,991	5,245	8,000	28,000	54,499	82,454	112,340
q/q $	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000	27,999	30,455	32,886
Net New ARR	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000	27,999	30,455	32,886
y/y %	%						466%	511%	919%	1,122%	2,000%	1,767%	1,118%	996%
q/q %	%			39%	-	0%	307%	50%	67%	20%	600%	33%	9%	8%
Total Revenue	$M		44	97	158	219	375	688	1,188	1,875	4,875	11,000	18,307	26,224
y/y %	%						756%	612%	652%	755%	1,200%	1,500%	1,442%	1,299%
q/q %	%			120%	64%	39%	71%	83%	73%	58%	160%	126%	68%	43%
Monthly Build:														
ARR, end of period	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000	57,999	88,454	121,340
Month 1	$M		100	345	591	836	1,000	2,500	4,000	7,000	12,000	44,000		
Month 2	$M		182	427	673	918	1,400	3,000	5,000	8,000	19,000			
Month 3	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000			
Net New ARR $ P/P	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000			
Month 1	$M		13	82	82	82	-	500	500	1,000	3,000	14,000		
Month 2	$M		82	82	82	82	400	500	1,000	1,000	7,000			
Month 3	$M		82	82	82	82	600	500	1,000	1,000	11,000			
ARR Growth P/P	%		203%	93%	48%	33%	100%	75%	71%	50%	233%	93%		
Month 1	%		15%	31%	16%	11%	-	25%	14%	17%	33%	47%		
Month 2	%		82%	24%	14%	10%	40%	20%	25%	14%	56%			
Month 3	%		45%	19%	12%	9%	43%	17%	20%	13%	58%			
ARR Growth Y/Y	%						659%	588%	695%	800%	1,400%			
Month 1	%						900%	624%	577%	737%	1,100%			
Month 2	%						570%	602%	643%	771%	1,257%			
Month 3	%						659%	588%	695%	800%	1,400%			
Segment ARR Build:														
Overall ARR by Segment:														
Consumer Subscription ARR	$M		30	35	65	95	170	200	325	400	1,550	2,388	3,067	3,718
Team Subscription ARR	$M		-	21	26	20	40	70	120	180	600	983	1,417	1,872
Enterprise Subscription ARR	$M		-	-	-	10	50	185	395	680	2,400	4,200	6,000	7,800
API Consumption ARR	$M		235	453	564	880	1,740	3,045	5,160	7,740	25,450	50,450	77,950	107,950
Overall ARR Mix Details:														
Consumer Subscription ARR	%		11%	7%	9%	10%	9%	6%	5%	4%	5%	4%	3%	3%
Team Subscription ARR	%		-	4%	3%	2%	2%	2%	2%	2%	2%	2%	2%	2%
Enterprise Subscription ARR	%		-	-	-	1%	3%	5%	7%	8%	8%	7%	7%	6%
API Consumption ARR	%		89%	89%	88%	88%	87%	87%	86%	86%	85%	87%	88%	89%
Overall ARR Mix Growth Y/Y:	%		-	-	-	-	-	-	-	-	-	-	-	-
Consumer Subscription ARR Growth Y/Y	%						467%	471%	400%	321%	812%	1,083%	850%	830%
Team Subscription ARR Growth Y/Y	%							233%	370%	800%	1,400%	1,304%	1,081%	940%
Enterprise Subscription ARR Growth Y/Y	%									5,700%	4,700%	2,170%	1,419%	1,047%
API Consumption ARR Growth Y/Y	%						642%	572%	677%	780%	1,383%	1,557%	1,411%	1,295%

Source: SemiAnalyis Tokenomics Model

In addition, Anthropic's margins on inference compute have shot up. We believe inference gross margins are now in the mid 60s, up from 38% in 2025 and -94% in 2024. Per a Wall Street Journal article on 5/20/26, Anthropic is expected to be Operating Income profitable in 2Q after adjusting for stock-based compensation. Our Tokenomics model for Anthropic allocates costs to Free and Paid User Inference, Training & Research Spend, and OpEx. Our latest estimates published to clients last week were extremely aligned across revenue, cost bucket mix and dollar amounts as well as overall GAAP (unadjusted for stock-based compensation) EBIT profitability.

Hyperscaler Takeaways: Amazon Met Demand With Capacity

Key to deploying a token-as-a-service platform at much larger scale than rivals was simply having more compute capacity, given the high compute-intensity of AI

inference. And as shown below, Amazon is adding significantly more capacity than rivals, with only Microsoft a close tie in 2024-26 but largely dwarfed in 2027. Our Datacenter Industry Model provides quarter-by-quarter capacity forecast by hyperscaler.

Source: SemiAnalysis Datacenter Model

However, total capacity is not the only relevant item. We need to break it down by end-user, which is what our Datacenter Industry Model can provide quarter by quarter. Our model quantifies the capacity going to Microsoft's internal AI efforts, which is higher than that of Amazon, reducing compute available to the broader customer base. In addition, the vast majority of Microsoft's AI compute goes to OpenAI via long-term compute contracts, as demonstrated by a remarkably large share of Microsoft's backlog, with OpenAI's backlog alone being 2.5x that of the total Azure annual revenue.

To deploy more capacity than peers, Amazon has been remarkably aggressive in growing its power pipeline, and signing multibillion PPAs with IPPs like Talen, Vistra, and NiSource.

On the other hand, Microsoft had a year-long datacenter pause (see our Datacenter Freeze article), which significantly lowered their 2027 capacity forecast. In addition, as we've covered in our Microsoft's AI Strategy piece, Microsoft has been remarkably slow at building large-scale AI clusters in Wisconsin, the opposite of AWS' lighting speed to build close to 2GW in Indiana and Mississippi. The only way for Microsoft to catch up is to contract significant amounts of capacity from Neoclouds, which is much more expensive and will reduce the margin advantage. This is happening all while Amazon keeps innovating to accelerate capacity growth. The company is rolling out a new datacenter design at a very large scale, with increased modularity and prefabrication as they continue to prosecute against the AI opportunity.

Amazon's AI Resurgence: AWS & Anthropic's Multi-Gigawatt Trainium Expansion

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL, AND 2 OTHERS • 2025年9月4日

Read full story →

The Google Pushback

The most natural pushback to our AWS outperformance thesis is that Google Cloud is also seeing increased margins, boasts the same vertical integration as AWS (if not more), and has been an even bigger outperformer, with revenue growth skyrocketing to >60% YoY in the latest quarter and margins a record high for GCP. We've been well ahead of the market in calling out Google Cloud's acceleration, with our first reports coming out in August 2025 in Accelerator, Datacenter and Core Research. Our TPU Deep Dive explained in great detail why Google Cloud is outperforming.

However, we think the margin rise is an illusion and is more akin to an "EBTIT" margin, i.e. Earnings Before Training, Interest and Taxes. Alphabet is seeing increased costs due to DeepMind/Gemini training expenses, but these costs are accounted for in "Alphabet-Level Activities," not GCP. Per Google's latest 10-Q, Alphabet-Level Activities, "primarily reflect expenses related to our shared AI research and development" and were $5.4B in 1Q26 up from $3.0B in 1Q25. Thus to say, all the Gemini API revenue is flowing to Google Cloud at higher-than-average margins, which boosts the margin profile of GCP while $10B+ of run rate costs are bucketed elsewhere.

In addition, Google Cloud might have benefited from one-off royalty payments due to the sale of TPUs to Anthropic, through Broadcom, with Google acting as an IP vendor.

Cloud vs AI Lab vs Hardware: How Can Google Satisfy All Demand?

Google is the ultimate supply-constrained business. This is a single company which attempts to compete simultaneously with AWS on Cloud, Nvidia on hardware, Anthropic & OpenAI on models, Meta on ads, Tesla on autonomous driving, and more.

When analyzing Google's capacity growth, we simply do not see a large enough buildout to serve all demand. Internal capacity is enough to build a successful AI Lab, but leaves little room for the Cloud business (excluding hardware) to grow to the same extent as AWS. In particular, Gemini Enterprise Agent Platform (previously Vertex) as a distribution platform for Claude (and maybe OpenAI in the future) **is seeing a significantly lower capacity additions than Bedrock**.

For Google, GCP appears to be a platform to upsell additional services. Met
best example of this: a large GPU deal that subsequently led to large-scale C
adoption, and then a massive TPU hardware deal.

Implications for Hyperscalers and Labs

We believe that margins at leading IaaS CSPs will only get better over the next 2-3 years. But, AWS' ability to show rising margins in a period of huge capacity ramp and CapEx inflation is remarkable and driven by strategic decisions their team made. Longer term, we expect AI Labs to increasingly verticalize their inference stack and pressure margins on pure IaaS vendors. We also expect CSPs to vertically integrate through custom silicon and strategic partnerships where they can add value through their large, diversified customer bases.

To have access to the full data, forward estimates, and talk with our Tokenomics team, reach out to sales@semianalysis.com.

Previous

A guest post by

Joey Brookhart

Discussion about this post