

The Compute Glut That Isn't

What Meta's move into compute rental actually signals for the AI trade

The market read Meta's pivot into renting compute as the first crack in the AI buildout. We read it as the opposite — the monetisation of a commodity that has become too scarce, and too valuable, to leave idle.

Background — The Announcement

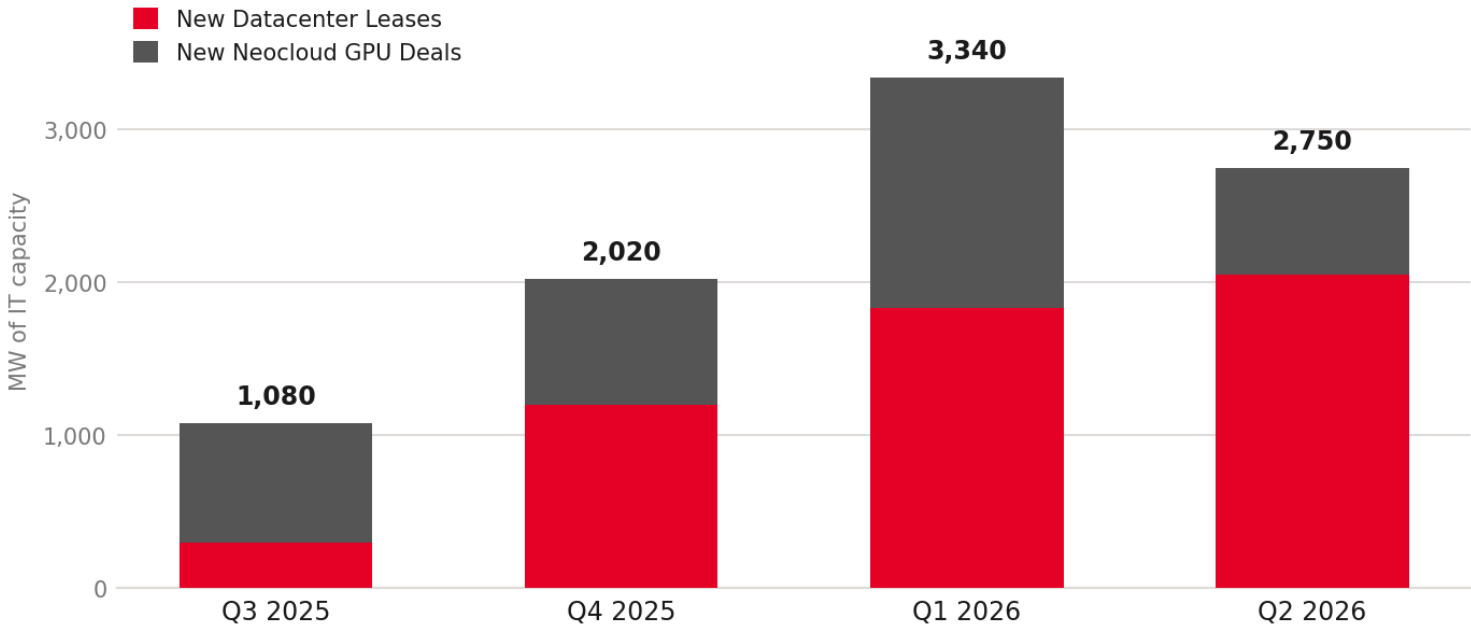
On 1 July 2026, Bloomberg reported that Meta is standing up a cloud infrastructure business — internally dubbed **Meta Compute** — to sell access to its AI computing capacity and models to outside customers. Two paths are reportedly under consideration: hosted access to Meta's models, including its closed-weight **Muse Spark**, in the mould of AWS Bedrock; or the sale of raw capacity in the mould of a neocloud such as CoreWeave. The effort is said to be led by infrastructure chief Santosh Janardhan, Superintelligence Labs' Daniel Gross, and president Dina Powell McCormick. Meta shares rose roughly 9-10% on the report — in a year where the stock had lagged on precisely the concern that its AI outlay, guided as high as ~USD 145bn for 2026, was a sink with no clear return.

Two points bear stating up front. First, **this is not a deal, a product, or even a confirmed business** — Meta has disclosed no pricing, timeline, or customer, and declined to comment. Zuckerberg himself flagged the possibility at May's shareholder meeting, noting that companies approached Meta almost weekly to buy access to its models or spare capacity. The substance here predates the headline by months; what changed on 1 July is that a wire ran a story. Second, **Meta is not first**. SpaceX, via xAI, has been renting Colossus capacity to Anthropic, Google, and Reflection AI — a playbook Bloomberg Intelligence estimates could generate north of USD 50bn by 2028, and USD 100bn by 2030.

The Street's Read — "The First Crack in the Bubble"

The bearish interpretation was immediate and, briefly, consensus. If Meta — the most aggressive spender in the group — has enough excess compute to rent out, then the buildout has outrun demand, capex is set to stall, and the AI infrastructure bubble is beginning to deflate. Commentators framed the news as the first hyperscaler to "admit" to excess capacity, with more expected to follow; semiconductors sold off on the print.

The accompanying logic leans on the familiar dot-com-fibre analogy: a debt-funded overbuild of rapidly depreciating hardware, destined to sit dark until demand catches up — if it ever does. On this view, a rental business is not opportunism but a distress signal — a company trying to salvage a return on infrastructure it can no longer justify on its own account. It is a tidy narrative. We think it is also the wrong one, and the chart below is the first reason why: Meta's contracted compute has been *accelerating*, not tailing off.



Meta's contracted compute per quarter, split between new datacenter leases and new neocloud GPU deals; over 5GW of cloud & colo contracted in 2026 year-to-date. Values indicative, redrawn by LC in brand style. Source: SemiAnalysis.

Our View — The Market Has This Exactly Backwards

We think the readout should be the opposite of consensus, and the available data does not support the glut thesis. Three observations frame our position — the first on why a rental business is a signal of scarcity, the next two on why demand is nowhere near rolling over.

1. This is the monetisation of scarcity, not a symptom of surplus.

The relevant precedent is SpaceX, and it is instructive. Compute has become scarce enough — and rentable at high enough rates — that leaving any capacity idle now carries a real opportunity cost. Renting externally is not what you do when compute is worthless; it is what you do when it is too valuable to sit unused. The economics of a rental business only work **because** the underlying commodity is scarce and pricing power is real — that is the tell. Far from admitting defeat, Zuckerberg is taking a page from Musk’s book and beginning to monetise a highly scarce, highly valuable asset he happens to be sitting on.

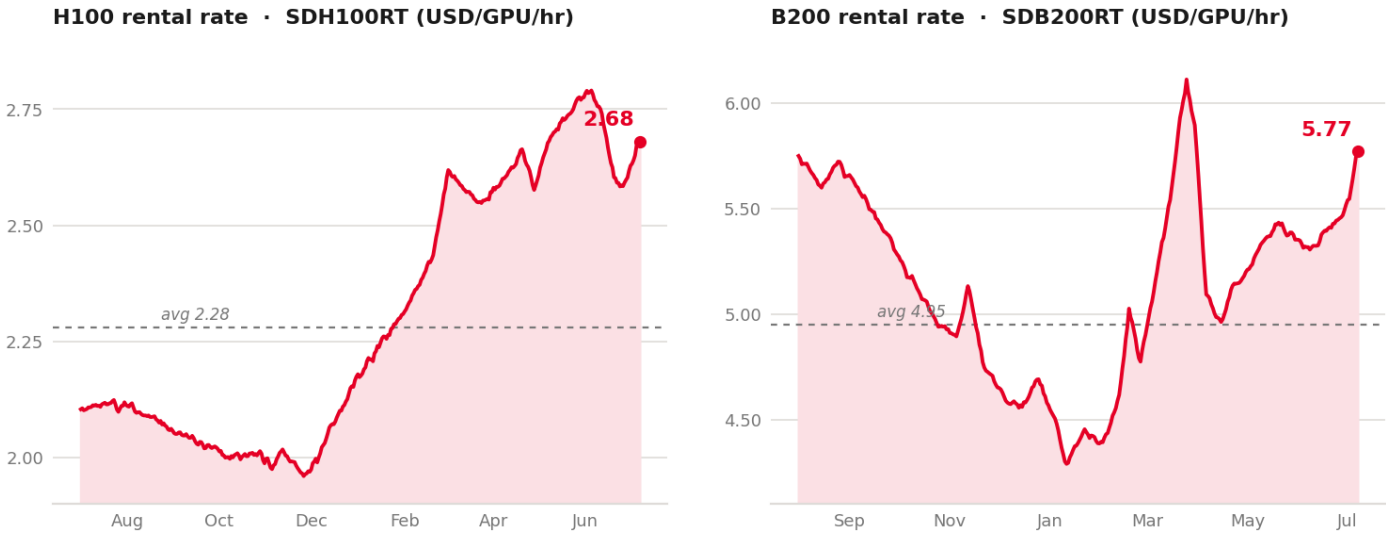
2. Demand is not rolling over — it is accelerating.

Every forward indicator points the same way. On our estimates token generation is compounding on the order of 70% month-on-month, with no sign of flattening — directionally consistent with Google’s disclosed 7x year-on-year jump to 3.2 quadrillion tokens per month, OpenRouter’s 5x in six months, and Goldman’s projection of a 24x rise in token consumption to 2030. On the supply side, Nebius — a leading neocloud — **raised** on-demand GPU rates in June, something almost impossible to reconcile with an over-capacity world. Prices simply do not behave this way in a glut.

Supply-Side Evidence — Rates Are Rising, Not Falling

NEBIUS ON-DEMAND GPU RATE (USD/HR)	PRIOR	NEW	CHANGE
NVIDIA H100	2.95	3.85	+30%
NVIDIA H200	3.50	4.50	+29%
NVIDIA B200	5.50	7.15	+30%
NVIDIA B300	6.10	7.85	+29%

Source: Nebius pricing notice, on-demand (PAYG) rates effective 1 June 2026. Preemptible (PVM) rates rose further (H100 +-51%).



SiliconData GPU rental-rate indices (USD/GPU/hr), trailing twelve months to 07 Jul 2026. H100 sits at 2.68 versus a 1.96 November trough; B200 at 5.77, near the top of its range. Dashed line marks the 12-month average. Both indices have stabilised materially above where they traded four-to-six months ago — the opposite of what an oversupplied market would show. Series redrawn by LC in brand style. Source: SiliconData / Bloomberg.

INDEX · LAST	1Y HIGH	1Y AVERAGE	1Y LOW	LAST VS LOW
2.68 H100 · SDH100RT	2.79 05 Jun 2026	2.28 Trailing 12M	1.96 26 Nov 2025	+37% off the trough
5.77 B200 · SDB200RT	6.11 25 Mar 2026	4.95 Trailing 12M	4.29 11 Jan 2026	+34% off the trough

3. The migration to cheaper open-source models is rationalisation, not value destruction.

The other leg of the bear case — that workloads are shifting to cheaper open-weight models, so frontier economics are broken — misreads a healthy development. What is unwinding is the brief, irrational phase of “token-maxing”: using AI for its own sake, regardless of return. That discipline is welcome, not ominous. And the conclusion that no value accrues to the frontier does not follow, for three reasons.

First, frontier labs continue to hold a roughly 6–9 month lead over open weights, which keeps them the default for the specialised, high-value tasks whose volume is itself exploding — the growth in token generation at Anthropic, Google, and OpenAI speaks to exactly that.

Second, even if most tokens ultimately run on open-source models, those tokens still need compute to be inferred — and the token count is growing exponentially. That is an unambiguous statement about demand for the full stack: compute, memory, networking, and power.

Third, Jevons’ paradox. As the cost per token falls — Goldman puts inference cost declines at 60–70% a year — usage rises by more than enough to offset it. Cheaper tokens mean more tokens, at both the open and frontier layers, and therefore **more** aggregate spend on infrastructure, not less. The shift to open-source models and the falling unit cost of a token are, on this logic, demand accelerants — not the trigger for a meltdown.

Bottom Line

Meta’s move is best understood not as a hyperscaler capitulating to overcapacity, but as one more owner of a scarce commodity choosing to monetise it — Zuckerberg following Musk into the compute-rental business. We would expect any rental contract Meta ultimately signs to carry **SpaceX-style termination and clawback provisions** — recall Musk framed the short Colossus lease as SpaceX’s own request, on the view that it might need the compute back — allowing Meta to reclaim capacity should its own frontier ambitions pay off. That optionality is the point: it demonstrates the inherent **fungibility of compute**, which can be monetised through multiple channels, and retains value even where the original objective is not met.

For the broader AI trade, the read-through is the reverse of the one that sold off semiconductors on 1 July. A hyperscaler electing to rent capacity into a market this tight is confirmation that compute clears at a premium — and that the structural beneficiaries remain the full stack that provisions it: accelerators, memory, networking, and power. The scarce, fungible asset is the story. The glut is not.

One qualification is worth making. While this does not signal an oversupply of compute infrastructure, it is in its own way a confession — from both SpaceX and Meta — that they are, at present, behind in developing a frontier model, and that at this point in time they see a better return on their investment by renting the capacity out rather than using it themselves. Nothing, however, stops them from eventually reaching a level where they might turn that same capacity inward to inference their own models as well.

What We Are Watching

Terms of any signed Meta contract — the presence and length of termination / clawback clauses as a read on how tight Meta itself judges capacity to be.

GPU rental indices — sustained SDH100RT / SDB200RT levels versus any rollover would be the cleanest real-time test of the glut thesis.

Token-volume disclosures — next prints from Google, and any figures from OpenAI / Anthropic, against our ~70% month-on-month working estimate.

Neocloud pricing actions — further on-demand rate moves from Nebius, CoreWeave and peers as a margin and scarcity signal.

Risks to Our View

Demand deceleration — the thesis rests on token growth continuing to compound; a sharp slowdown in agentic / inference workloads would validate the overbuild concern.

Financing and circularity — the debt-funded buildout and vendor-financing loops between chipmakers, neoclouds and labs could amplify any drawdown independent of physical utilisation.

Loss of open-source / frontier differentiation — if the frontier loses its 6–9 month lead, these trillion-dollar-valuation companies are in trouble, which would eventually reduce total demand for compute capacity.

Geopolitical risk — frontier model developers being completely banned from providing access to their latest models to the general public and broader enterprises.

Important Disclosures & Disclaimer

CONFIDENTIALITY

The contents of this document are confidential and are meant for the intended recipient only. If you are not the intended recipient, please delete all copies of this document and notify the sender immediately.

GENERAL COMMENTARY

This document, provided as a general commentary, is for informational purposes only and is not to be construed as an offer to sell or solicit an offer to buy any financial instruments in any jurisdiction. This does not constitute any form of regulated financial advice, and your independent financial advisor should be consulted prior to taking any investment decision(s).

This document is based on information from sources which are reliable but has not been independently verified by Lighthouse Canton Pte. Ltd. and its affiliates (“LC”). LC has taken reasonable steps to verify the contents of this document and accepts no liability for any loss arising from the use of any information contained herein. Please also note that past performances are not indicative of future performance. Information contained herein are those of the author(s) and does not represent the views held by other parties. LC is also under no obligation to update you on any changes made to this document.

REGULATION

This document is prepared by Lighthouse Canton Pte. Ltd. and its affiliate, Lighthouse Canton Capital (DIFC) Pte. Ltd., which are regulated by Monetary Authority of Singapore (“MAS”) and Dubai Financial Services Authority (“DFSA”) respectively. MAS and DFSA have no responsibility for reviewing, verifying and approving the contents of this document and/or other associated documents. The contents of this document may not be reproduced or referenced, either in part or in full, without prior written permission from LC.

Our Offices

SINGAPORE

16 Collyer Quay #11-02
Collyer Quay Centre
Singapore 049318
+65 6713 0570

UAE

The Exchange, Gate Village 11
Unit 802, DIFC
PO Box 507026, Dubai
+971 4 586 1500

UNITED KINGDOM

24 Hanover Square
London W1S 1JD
+44 1642 843 487

INDIA — MUMBAI

Unit 507/508, A Wing, INS Tower
G Block, BKC
Mumbai – 400051

INDIA — NEW DELHI

Unit 104A, Worldmark 2
Aerocity
New Delhi 110037

INDIA — BENGALURU

1st Floor, WeWork 37
Cunningham Cross Rd
Vasant Nagar
Bengaluru 560001

INDIA — CHENNAI

RK Swamy Centre, Hansa Building
Door No: 3, Thousand Lights
Chennai 600006

INDIA — HYDERABAD

Suite 502, Building 450
Central Plaza, Genome Valley
Shameerpet, Hyderabad 500078

INDIA — GIFT CITY

Unit FF-10, Pragya Accelerator
Block 15T, GIFT CITY
Gujarat 382355