

Muse Scares Amazon, Doosan's Largest Bet, ASE's Test Buildout, TPU Ahead of GB200, Opus 5.5 Release

Timothy Chang, Nick Doyle | Friday, September 25, 2026

Executive Summary

- Positive: META, ASE & Test Equipment, Doosan, TPU Ecosystem, QCOM, NBIS
- Neutral: CRWV, Nscale
- Negative: AMZN, CUDA Moat

Muse Scales Personalized Agents and Scares Amazon

With nearly 400k GitHub stars, OpenClaw's success shows strong interest in personalized agents. Meta simply built on it. Nat Friedman acknowledged that Muse was "heavily inspired as a product by OpenClaw". As we discussed in our September 11 weekly, Meta put security and ease of use at the center of the pitch. That could help explain why Muse passed 2.5 million downloads within 13 days.

More broadly, Meta's competitive advantage is its ability to improve consumer experience. Rather than being first, it makes existing ideas easier to use and brings them to a wider audience. We believe much of Muse's upside will come from Meta's existing platforms, which reached 3.6 billion daily users in June.

Product / launch	Format precedent	Reported scale
Instagram Stories 2016	Snapchat	500M daily users Reported in January 2019
Reels 2020	TikTok	\$50B+ annual revenue run rate Instagram + Facebook, disclosed October 2025 More than half of Instagram ads ran in Reels in 2025
Threads 2023	Twitter / X	500M monthly active users June 2026 100M sign-ups in five days at launch in July 2023

Source: Meta; Sensor Tower via CNBC; Social Media Today; WSI; Reuters.

However, Amazon blocked Muse from purchases on September 20 and cited privacy and security concerns. We think this is a tacit acknowledgment of Muse's capabilities and its potential to take control of the online shopping experience. Amazon previously restricted ChatGPT's web access to its site in November 2025. Our September 17 Muse flash note already showed that Muse sent users straight to payment links. If the online shopping experience begins to run through agents, billions of dollars in advertising could move with it. Amazon would also have fewer opportunities to upsell shoppers or charge brands for sponsored placements.

Shopify chose to partner with Muse and enable Shop Pay checkout across its stores. We believe the key indicators now are whether other large online retailers follow Amazon or Shopify, and whether Meta and Amazon reach a deal.

ASE's Test Buildout Leads the OSAT Equipment Surge

ASE Technology Holding disclosed equipment purchases on September 18th and 22nd totaling \$260M, equivalent to nearly **half ASE's avg. quarterly CapEx** from 2020-24:

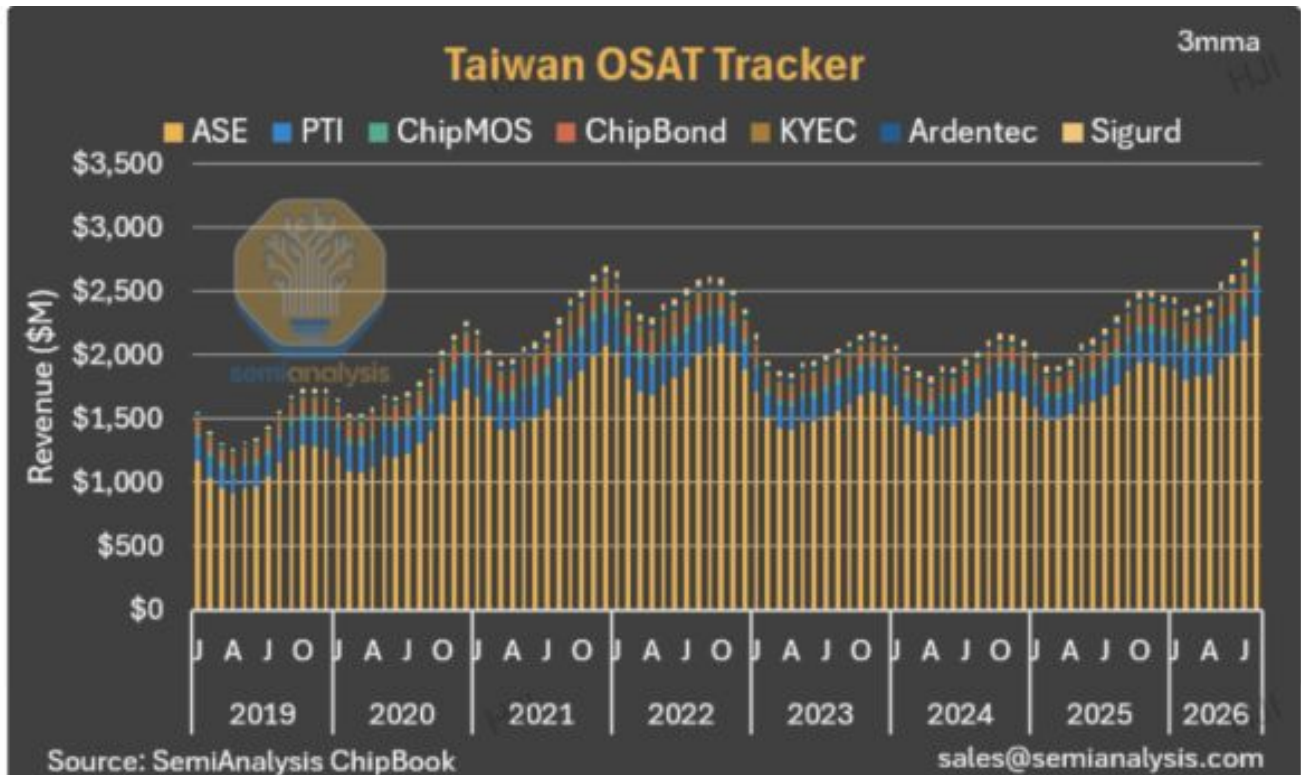
- \$133M of Advantest tools for SPIL
- \$48M of Teradyne tools for ASE (Singapore)
- \$37M of Murata Machinery equipment for SPIL
- \$42M of Mirle Automation for ASE

The \$181M committed to Advantest and Teradyne points to a substantial test-capacity build ahead of 2027. ASE said earlier this year that most incremental leading-edge advanced packaging (LEAP) machinery would go to wafer sort, with equipment deployed in 4Q26 supporting a '27 ramp. By July, ASE said both wafer sort and final test were near full. Our August note found that Taiwanese OSAT equipment orders had already exceeded 3x full-year 2025 and test-heavy OSATs were seeing overflow demand. **The second image shows Taiwan OSAT revenue following equipment orders;** the 3MMA rose to nearly \$3B in August, well above prior peaks. The Murata and Mirle orders are consistent with factory automation vs. test tools. Filings name suppliers and amounts, but do not identify specific tools.

China's aggregate equipment data obscures back-end strength. Total semiconductor equipment imports were – 5% YTD, while August OSAT equipment imports jumped 200% Y/Y. We see Chinese stockpiling as a possibility. Note these customs series include equipment categories outside pure WFE and record import timing, so we would not equate their growth rates with installed fab capacity.

We remain positive on ASE/SPIL and test-exposed equipment suppliers as more advanced packages require greater test capacity.





Doosan Makes Its Largest-Ever Bet on AI CCL

In High Density Interconnect, High Demand Incoming, we showed how rising layer counts and the shift to HDI increase material and manufacturing requirements. Doosan's existing CCL capacity is already stretched, with its Jeungpyeong and Gimcheon plants reporting utilization of 109.8% and 102.7%, respectively.

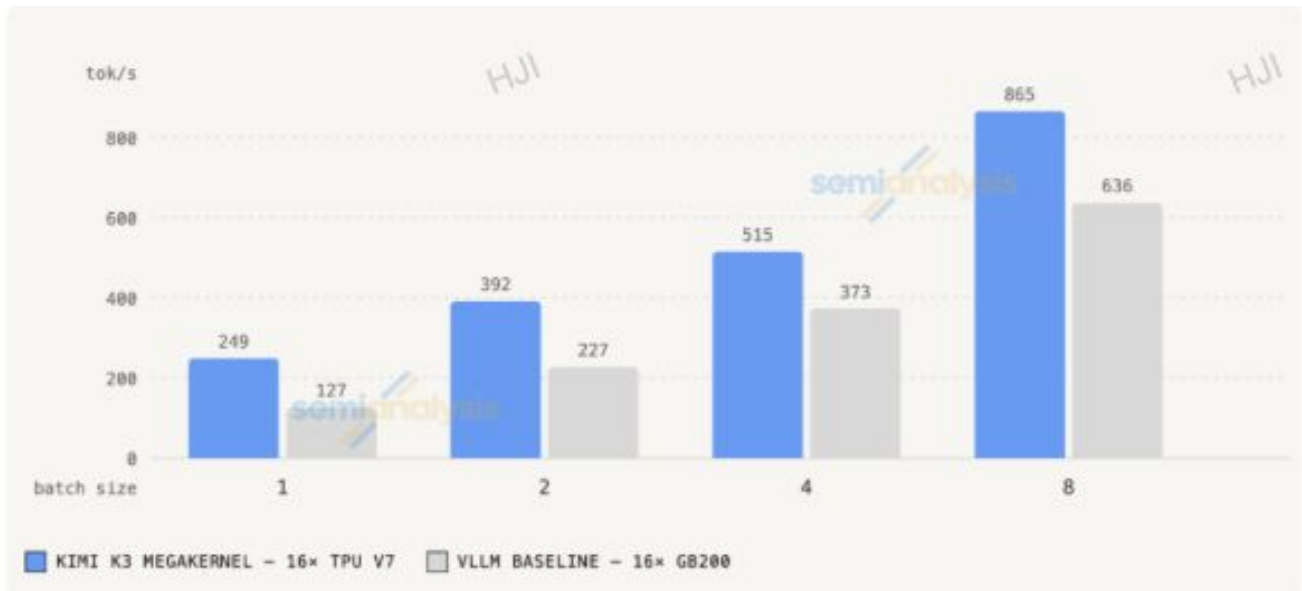
Doosan is responding with 970 billion won (\$700 million), the largest investment in its history. The larger piece, about 550 billion won, builds a new plant in China for AI accelerator and network switch CCL. Doosan chose China because that is where the global PCB makers are concentrated. Being close to them speeds up customer response and lets the plant go up faster. The other 420 billion won expand its Korean lines for optical-module CCL.

Our checks with Doosan show Nvidia makes up about 60% of its AI server CCL. Its DS-7409 DQN laminates are used by both Nvidia and AMD's MI450, and TPU and Trainium are testing them through PCB vendors. DS Investment expects VGT and Unimicron to be major customers, and industry chatter says VGT may co-invest in the China plant. EMC is still the main rival. In Nvidia's CCL Supply Chain Shake Up, we showed EMC overtaking Doosan at Nvidia with Rubin, though we expect Doosan to keep the Strata compute board.

Doosan's new lines don't start until 2H28, shortage in 2027 remains the worst. Even with aggressive expansion from Doosan, EMC, and others, **we still see shortages on high-end CCL in 2028.**

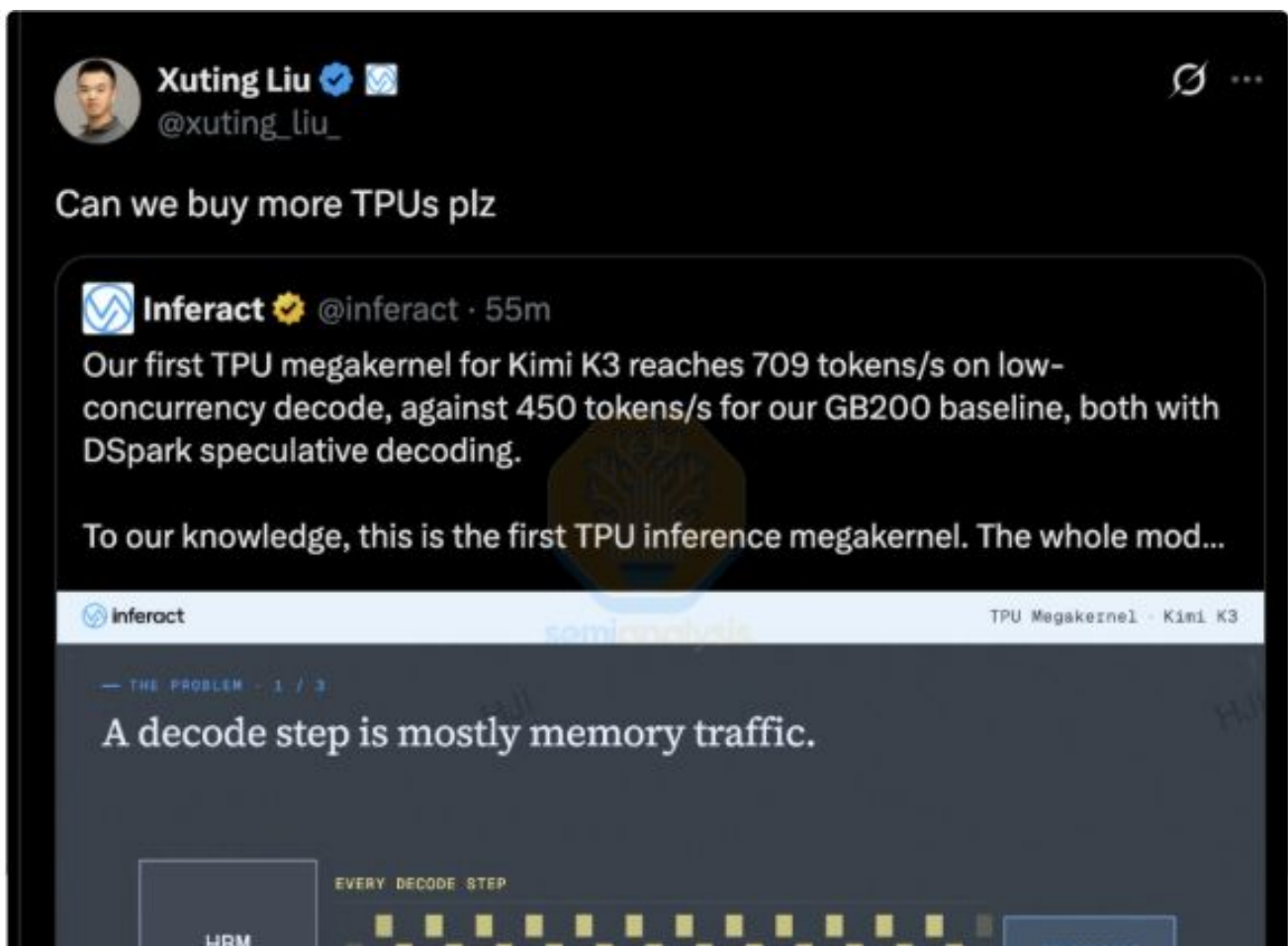
Popular Inference Engine Puts TPU Ahead of GB200

Following up on last week's note, the maintainers of a widely used open-source inference engine released open-source TPU v7 megakernels. On Kimi K3, the popular 2.8T-parameter model, 16 TPU v7 chips deliver 1.4x to 2x the decode throughput of 16 GB200s, with the biggest lead at small batch sizes. **On Google's internal cost basis, we estimate TPU is up to 3.5x better per dollar.**



Source: Inferact

The more important signal is who built it. The people who maintain open-source inference software have long written for CUDA first, and NVIDIA benefits from that work largely without paying for it. That free labor is a large part of why CUDA is hard to displace. These maintainers are now putting real engineering time into TPU and want to do more. One of the megakernel's authors answered the launch with "Can we buy more TPUs plz." As more of that work goes to TPU, NVIDIA's software lead gets harder to defend.



Source: X, Inferact, Xuting Liu

Opus 5.5: Cheaper Per Token, Not Per Task

This Tuesday, Anthropic launched Claude Opus 5.5 at \$4/\$20 per million input/output tokens, **20% below Opus 5 and Opus 4.8 (\$5/\$25), and 60% below Astra (\$10/\$50)**. Opus 5.5 also cut cached input from \$0.50 to \$0.20 (Astra at \$1.00). On Anthropic's launch table, Opus 5.5 beats Fable 5.1 (priced at \$10/\$50) on all nine benchmarks. On benchmarks where they overlap, **Opus 5.5 leads GPT-6 Astra on four of the six benchmarks**, trailing on AutomationBench (where Astra took the lead earlier this month), and on Terminal-Bench-Science. Artificial Analysis (AA) ranks it first on its Intelligence Index at 58, versus 53 for both Astra and Fable 5.1.



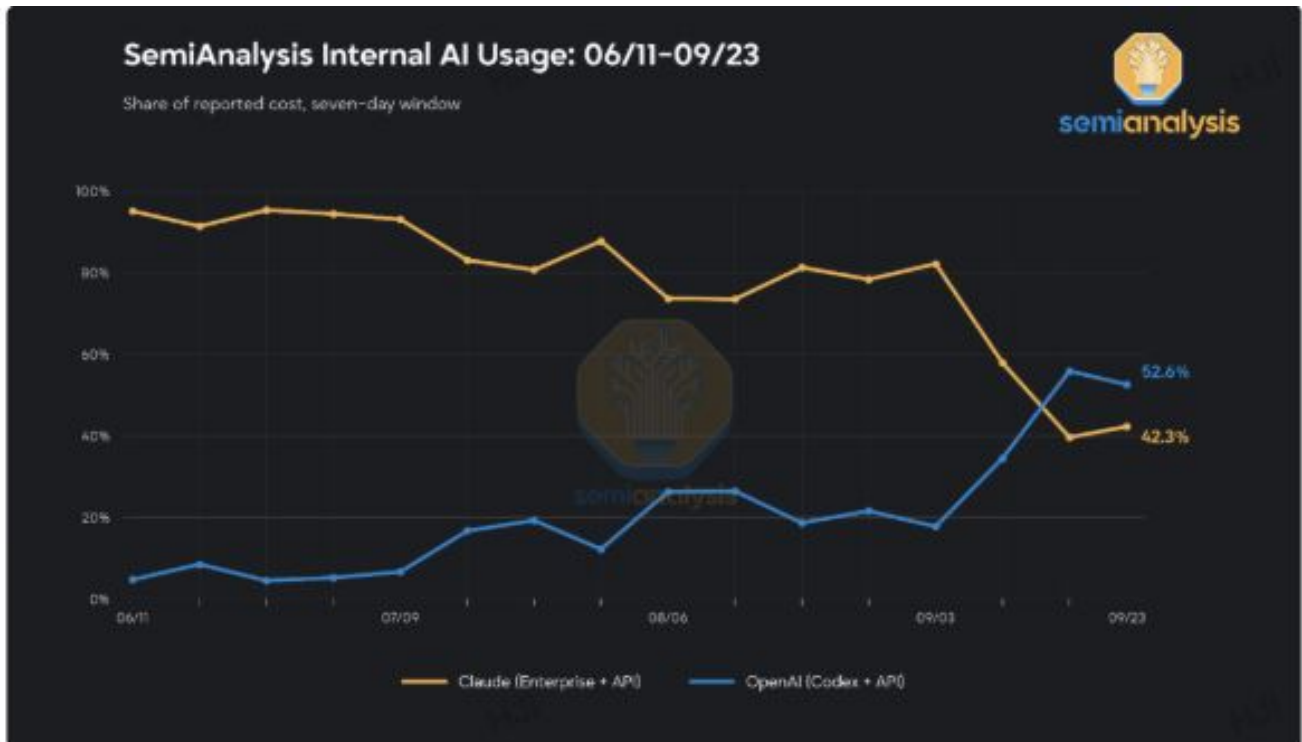
DATE	MODEL	ORGANIZATION	COUNTRY	CATEGORY	LICENSING
2025-09-22	GPT-6 Luna	OpenAI	US US	—	—
2025-09-22	Claude Opus 5.5	Anthropic	US US	—	—
2025-09-22	Baka EdgeQ	Baka AI	US US	—	—
2025-09-22	MMo-V2.6-Pro-UltraSpeed	Xiaomi AI (MMo)	CN CN	—	—
2025-09-22	MMo-V2.6-Fast	Xiaomi AI (MMo)	CN CN	—	—
2025-09-22	MMo-V2.6-Pro	Xiaomi AI (MMo)	CN CN	—	—
2025-09-22	Step 5 Preview	StepFun	CN CN	General	—
2025-09-21	H3 Max	Idolai	US US	—	—
2025-09-21	AlicAI-Foundation-80B-A3B-Base	Yandex	—	General	Open source
2025-09-21	Grok 4.7	xAI	US US	—	—
2025-09-21	Qwen-Image-2.1	Alibaba (Qwen)	CN CN	—	—

Source: Tokenomics Dashboard

Cheaper tokens do not guarantee cheaper tasks. At max effort, AA measures ~119K output tokens per Intelligence Index task for Opus 5.5 versus ~27K for Astra, **so output cost per task comes to ~\$2.38 versus ~\$1.35, roughly 1.8x more despite a 60% lower list price**. We continue to work on internal benchmarks.

We read the cut as a response to OpenAI's competitive pressure, while open models such as Kimi K3 (\$3/\$15) keep pressing from below. Pricing across providers is live on the Tokenomics Dashboard. Sonnet 5.5, expected in the coming weeks, will show whether the cut extends down the lineup.

Inside SemiAnalysis, Opus 5.5 is gaining traction, **reaching 30.7% of our daily reported tokens** on September 23, its first full UTC day after launch. However, the accompanying chart provides the broader context: **OpenAI's share of reported costs climbed sharply in September**, reaching 52.6% in the seven-day window ending September 23. Astra won the crowd, but Opus 5.5 has us looking back at ol' reliable.



Source: SemiAnalysis

Quick Hits:

- Apple-Qualcomm License Renewal:** Qualcomm renewed its global patent license agreement with Apple, effective April 1, 2027. The 2019 agreement was due to expire in March 2025 and included a two-year option, which Apple exercised in 2024 to extend the existing license through March 2027. The latest deal is a fresh renewal beyond that option. In Qualcomm Quits Apple, Keeps the Royalties note, we argued that Apple's shift to its own modems would phase out Qualcomm's chip sales while preserving its patent royalties. **We estimated roughly \$11 per Apple cellular device, or about \$3 billion annually across a 280 million-device base, regardless of whose modem Apple uses.** The renewal supports that call, though Qualcomm disclosed neither the new duration nor the royalty rate.
- ClusterMAX 3.0:** Our latest testing puts Nebius alongside CoreWeave in Platinum, while CRWV remains the technical benchmark. We favor NBIS for its combination of **managed-cluster quality, premium rental pricing, and substantial customer prepayments**, which should help monetize new capacity as GPU supply remains tight. Across 77 reviewed (323 tracked) providers, only 19 earned a medal: **reliable, secure, usable compute remains scarce**. Next, we plan to apply hands-on testing to bare metal, inference endpoints, post-training systems, sandboxes, models, and harnesses.
- Nscale's IPO:** Nscale has filed for a US listing with \$103.4bn of active and contracted TCV, highlighting the scale of its planned expansion. But we believe it is misleading investors by advertising managed clusters and inference endpoints it does not actually provide, as discussed in our ClusterMAX 3.0 note. We are also seeing permitting and construction risks from our datacenter team, that could delay turning contracts into revenue. However, we still believe Nscale's access to land and power, construction experience and strong demand for affordable bare metal give it substantial growth potential. **Our upcoming TCO and neocloud coverage will examine whether that growth can generate attractive returns after financing and infrastructure costs.**
- Murata MLCC Expansion:** Management is considering another MLCC expansion beyond March 2028, on top of ¥80 billion of FY26–FY27 equipment spending targeting more than 20% added capacity. We see **the proposed expansion as validation of AI-server MLCC demand, with little relief for the 2027 supply**

squeeze via our May thesis. The new plan is undecided, and we estimate new lines will take 18–24 months.

SemiAnalysis Weekly Published Research

Date	Vertical	Article	Key takes
9/24/2026	Memory	All About That Base: Custom HBM and the New Memory Economics	Custom HBM frees accelerator die area and shifts stack value toward Nvidia and TSMC, squeezing memory vendor take
9/23/2026	Core Research Expanded	ISIG EU Executive Summit 2026	Power suppliers showcase high-voltage DC hardware, but doubling lead times and slow qualification keep broad adoption distant
9/23/2026	TCO	ClusterMAX 3.0 Preview	Nebius joins CoreWeave at the top as a raised bar pushes most neoclouds down the ClusterMAX rankings
9/23/2026	WFE	Terafab Building at Full Speed; Phase 1 and Austin Fab will Purchase US\$9.8B and \$1.9B WFE over 2026-29	Terafab matching TSMC Arizona build pace, making SpaceX a major new WFE buyer benefiting ASML, KLA, AMAT and Lam
9/22/2026	Tokenomics	The Enterprise AI Token Budget Tracker: New Addition To The Tokenomics Model	Enterprises shift from AI access to token budgeting, with new tracker cataloging spend, caps and forecasts across companies
9/22/2026	Networking, Core Research Expanded	FIT (Foxconn Interconnect Technology) Tech Day 2026	FIT positions for XPO copper links while plastic waveguides, detachable engines and multidimensional multiplexing remain years from scale
9/22/2026	Tokenomics	A New Conformance Model	Jev turns small agent decisions into cheap calibrated probabilities, stealing high-volume calls without denting frontier lab ARR
9/22/2026	Tokenomics	September Tokenomics Update	New bottom-up capex build shows hyperscaler spending surging as model adds Oracle, SpaceX and refreshed lab ARR
9/22/2026	Datacenter	OpenAI's 2.7 GW Camellia Campus in Georgia	OpenAI's Georgia megacampus locks in curtailable utility power while navigating permitting, local opposition and a zoning lawsuit
9/21/2026	Foundry	Smartphone Model: Apple and Samsung Gain Share as Qualcomm and MediaTek Retreat	Memory costs gut entry Android volumes while Apple and Samsung gain share and AP value holds far better than units
9/18/2026	Accelerator	Broadcom Anthropic and OpenAI ASIC Upgrades; TPU Roadmap Update Across Whalefish, Bladerunner, AMD TPU v10, Humufish and Merope; Foxconn Assembly; 8-Hi HBM Implementation; SpaceX ODM	OpenAI and Anthropic ASICs lift Broadcom beyond TPU share loss at Nvidia's expense as roadmaps shift to shorter HBM stacks
9/18/2026	Accelerator, Memory	LPHBM Returns? Well, Almost: Qualcomm HBC Deep Dive	Qualcomm HBC trades external bandwidth for in-stack compute, costing more per GB with unproven software and capacity limits
9/18/2026	Core Research	Pacing the Frontier, Watch Forgent Power for Chase Potential, Optimized TPU, Micron's Labor Dispute	Positive: TPU Ecosystem, non-Micron DRAM suppliers, FPS Neutral: AI supply chain Negative: Frontier Labs, CUDA moat, Micron

Hard “Core” Hits of the Month

Date	Vertical	Article	Key Takes
9/14/2026	TCO	AI Cloud TCO Model Update: “Token Mode” for TaaS Revenue Projections, GPU Rental Prices Marching Up Again, Inclusion of OpenAI Jalapeno, Anthropic Cloud Deals	GPU rental prices rise as token serving economics strengthen; new token revenue modeling expands the cloud profitability toolkit
9/14/2026	Datacenter	SpaceX Update – New Sites Identified – 2.4GW Now Underway, TeraFab may bring +900MW	New SpaceX sites support aggressive compute expansion as rapid construction and turbine availability underpin capacity growth
9/14/2026	Tokenomics	What “Pacing the Frontier” Means for the Future of OpenAI and Anthropic	Slower frontier development erodes lab pricing power while compute demand stays strong as resources shift toward safety and inference
9/11/2026	Core Research	Meta's Muse, GPU Rental Pricing Up On OSS, High-NA for DRAM Benefits Lasertec, Rising Test Intensity Lifting Equipment and Consumables, TPUv7's TCO Advantage versus Blackwell	Positive: META, GPU Pricing, Open-Source Models, Lasertec, Test Ecosystem, TPU Ecosystem Neutral: TSMC Negative: KLAC, Nvidia's Inference Moat (TPU externalization)
9/9/2026	Core Research, Tokenomics	Nvidia's Backstop Universe – Heads I Win, Tails Who Loses?	Nvidia's off-balance-sheet backstops balloon, yet its cash generation supports far larger commitments to fund neoclouds and neolabs
9/9/2026	Core Research Expanded, Datacenter, Energy, Industrial	Sold-Out Turbines, Naked Leases, Backstop Rumors, and the Race to Power – Datacloud Austin 2026	Turbines sold out, naked leases burning engineering hours, and Nvidia backstops still propping neocloud financing despite wind-down talk
9/8/2026	Networking, Core Research Expanded, Foundry, WFE	SEMICON Recap: The Back End is Packaging More Value	Packaging captures most accelerator cost while FOPLP solves supply not cost, and CPU volume escapes CoWoS
9/4/2026	Datacenter	Mapping Every State and Local Datacenter Moratorium in the US, and What It Actually Delays	Datacenter moratoriums proliferating but delaying only a sliver of planned capacity, with behind-the-meter developers actually benefiting
9/1/2026	Networking	Innolight NPO Uses Low-Power Internal Lasers	Low-power internal lasers let NPO tap China's laser ecosystem, scaling past CPO's compounded yield and supply constraints
8/31/2026	Tokenomics	You can make over \$100M per MW per year selling open source tokens	Agentic InferenceX data shows open source serving margins fattening, pushing compute prices meaningfully higher
8/28/2026	Accelerator, TCO, Core Research	Nvidia Takes Half: Revenue Share, and 8-Hi Stays Our Call	We keep our Rubin HBM despec call despite Nvidia denials while revenue share emerges as an unmodeled pure-margin stream
8/26/2026	Core Research Expanded, Energy	SemiAnalysis Energy Webinar: Grid Constraints, Behind-the-Meter Power, & the IPP Tracker	Grid shortfall deepening with behind-the-meter generation the primary fix while grid reform crawls
8/25/2026	Accelerator, Core Research	Jalapeno: Spicing Things Up	Jalapeno proves OpenAI silicon competitive, impairing the CUDA moat while Broadcom emerges a clear winner
8/25/2026	TCO	Nvidia Captures More AI Value: VR NVL72 Server Price Hikes	Nvidia's planned Rubin rack price hike far exceeds component cost pass-through, lifting margins as neocloud returns run hot

Disclaimers

Analyst Certifications and Independence of Research.

Each of the analysts whose names appear in this report hereby certify that all the views expressed in this Report accurately reflect our personal views about any and all of the subject securities or issuers and that no part of our compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views in this Report. SemiAnalysis LLC (the “Company”) is an independent equity research provider. The Company is not a member of the Financial Industry Regulatory Authority or the Securities Investor Protection Corporation, and it is not a registered broker dealer or investment adviser.

Private Investment Disclosure.

Each analyst and/or author contributing to this Report hereby discloses that they may hold, directly or indirectly, private investments in companies, sectors, or asset classes discussed or referenced herein. Such holdings, if any, are disclosed externally with a third-party vendor. The Company requires that no author or analyst allow their personal investment positions to influence the views, opinions, analyses, or conclusions expressed in this Report. Readers should be aware that the existence of such holdings, even where disclosed, may present a potential conflict of interest, and are encouraged to weigh this accordingly when evaluating the contents of this Report.

Limitation of Research and Information.

This Report has been prepared for distribution to only qualified institutional or professional clients of the Company. The contents of this Report represent the views, opinions, and analyses of its authors. The information contained herein does not constitute financial, legal, tax or any other advice. All third-party data presented herein were obtained from publicly available sources which are believed to be reliable; however, the Company makes no warranty, express or implied, concerning the accuracy or completeness of such information. In no event shall the Company be responsible or liable for the correctness of, or update to, any such material, or for any direct, indirect, incidental or consequential loss, damage or lost opportunities (including loss of profits, revenue or goodwill) resulting from use of this data or from any investment decisions based on information or research obtained from the Company. Nothing contained in this Report or any distribution by the Company should be construed as any offer to sell, or any solicitation of an offer to buy, any security or investment. Any research or other material received should not be construed as individualized investment advice. Investment decisions should be made as part of an overall portfolio strategy and you should consult with a professional financial advisor, legal and tax advisor prior to making any investment decision.

Reproduction and Distribution Strictly Prohibited.

No user of this Report may reproduce, modify, copy, distribute, sell, resell, transmit, transfer, license, assign or publish the Report itself or any information contained therein. Notwithstanding the foregoing, clients with access to working models are permitted to alter or modify the information contained therein, provided that it is solely for such client's own use. This Report is not intended to be available or distributed for any purpose that would be deemed unlawful or otherwise prohibited by any local, state, national or international laws or regulations or would otherwise subject the Company to registration or regulation of any kind within such jurisdiction.

Copyrights, Trademarks, Intellectual Property.

The Company, and any logos or marks included in this Report are proprietary materials. The use of such terms and logos and marks without the express written consent of the Company is strictly prohibited. The copyright in the pages or in the screens of the Report, and in the information and material therein, is proprietary material owned by the Company unless otherwise indicated. The unauthorized use of any material on this Report may violate numerous statutes, regulations and laws, including, but not limited to, copyright, trademark, trade secret or patent laws.

ETF and Regulatory Status Disclaimer.

From time to time, The Company may collaborate with asset managers, issuers, sponsors, index providers or

other financial institutions in connection with the development, launch or support of exchange-traded funds (“ETFs”) and other investment products, including by providing research, market analysis, index-related research or other research related services. The Company does not sponsor, endorse, sell or promote any ETF, and makes no representation or warranty, express or implied, regarding the advisability of investing in any such ETF. In connection with the services described herein, the Company does not act as a fiduciary to any investor, fund, sponsor, issuer or other party. The Company is not registered with the U.S. Securities and Exchange Commission or any state securities authority as an investment adviser or exempt reporting adviser. The Company’s role is limited to providing research and related services, and it does not provide individualized investment advice, portfolio management, brokerage services or other services requiring registration as an investment adviser.