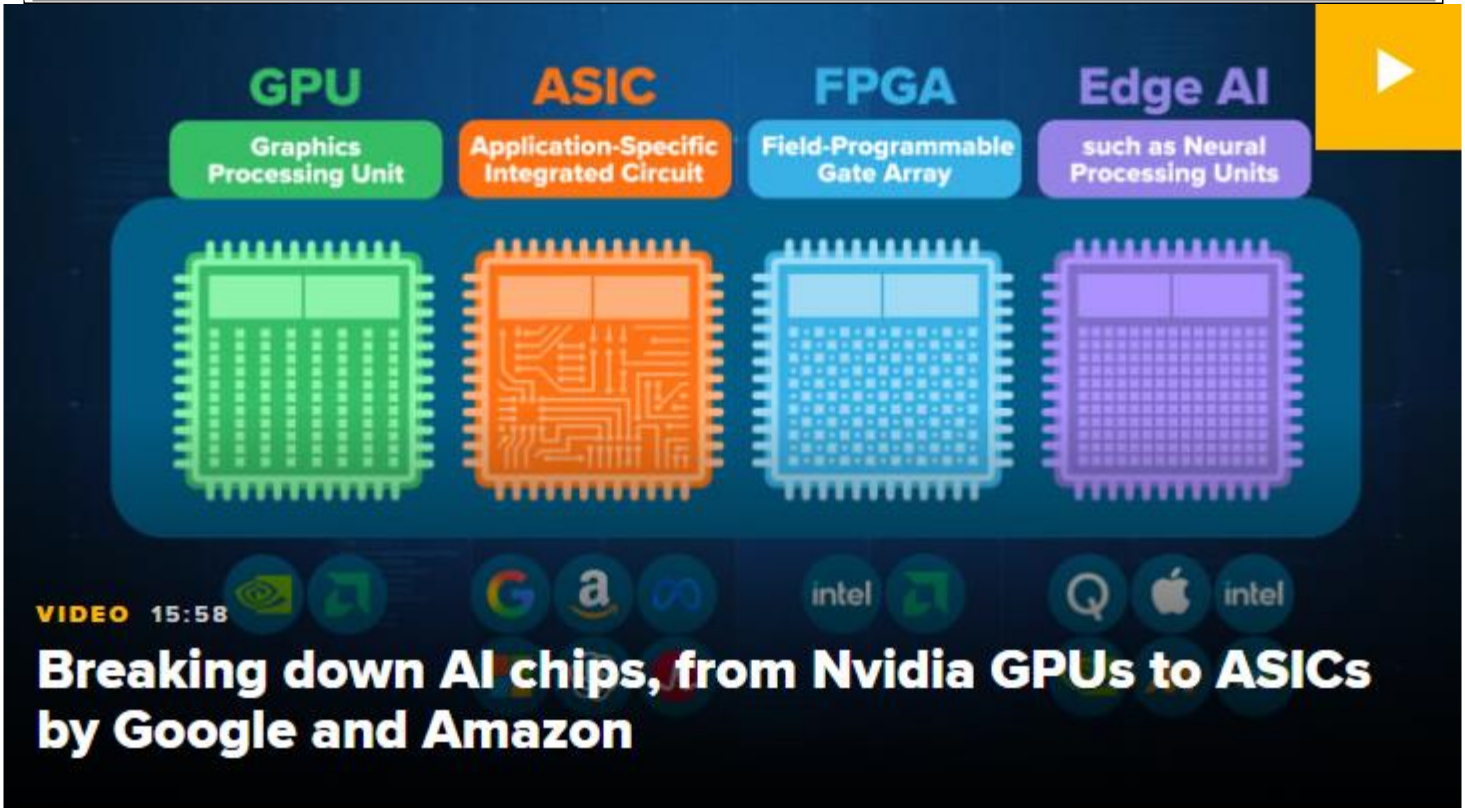
US Stock Express

Daniel Yue

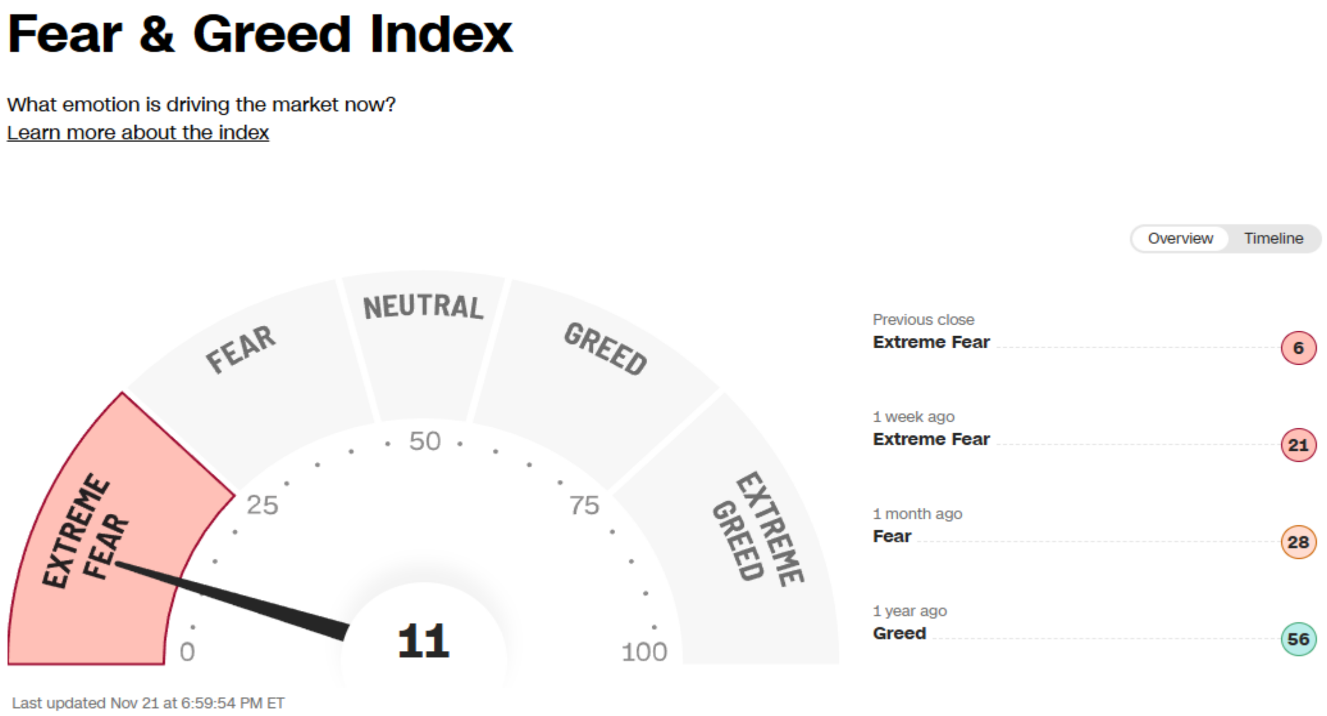
Email: info@ihandbook.org www.ihandbook.org ©


GPU
Graphics Processing Unit
ASIC
Application-Specific Integrated Circuit
FPGA
Field-Programmable Gate Array
Edge AI
such as Neural Processing Units
intel
intel
VIDEO 15:58
Breaking down AI chips, from Nvidia GPUs to ASICs by Google and Amazon

Custom ASICs, or application-specific integrated circuits, are now being designed by all the major hyperscalers, from Google's TPU to Amazon's Trainium and OpenAI's plans with Broadcom. These chips are smaller, cheaper, accessible and could reduce these companies' reliance on Nvidia GPUs.

Nvidia GPUs, Google TPUs, AWS Trainium: Comparing the top AI chips

Risk disclosure: Price can go up and down at any moment, use free money to trade and bear the risk according to your own capital;
Never trade with money that has a deadline for withdrawal.
All suggestions are for reference only, even AI cannot be 100% reliable, final decision still lies upon investors.
Copy trading cannot replicate another trader's background or psychological state.



Fear & Greed Index

What emotion is driving the market now?
Learn more about the index

Previous close
Extreme Fear 6

1 week ago
Extreme Fear 21

1 month ago
Fear 28

1 year ago
Greed 56

Last updated Nov 21 at 6:59:54 PM ET

North East West South is *NEWS*

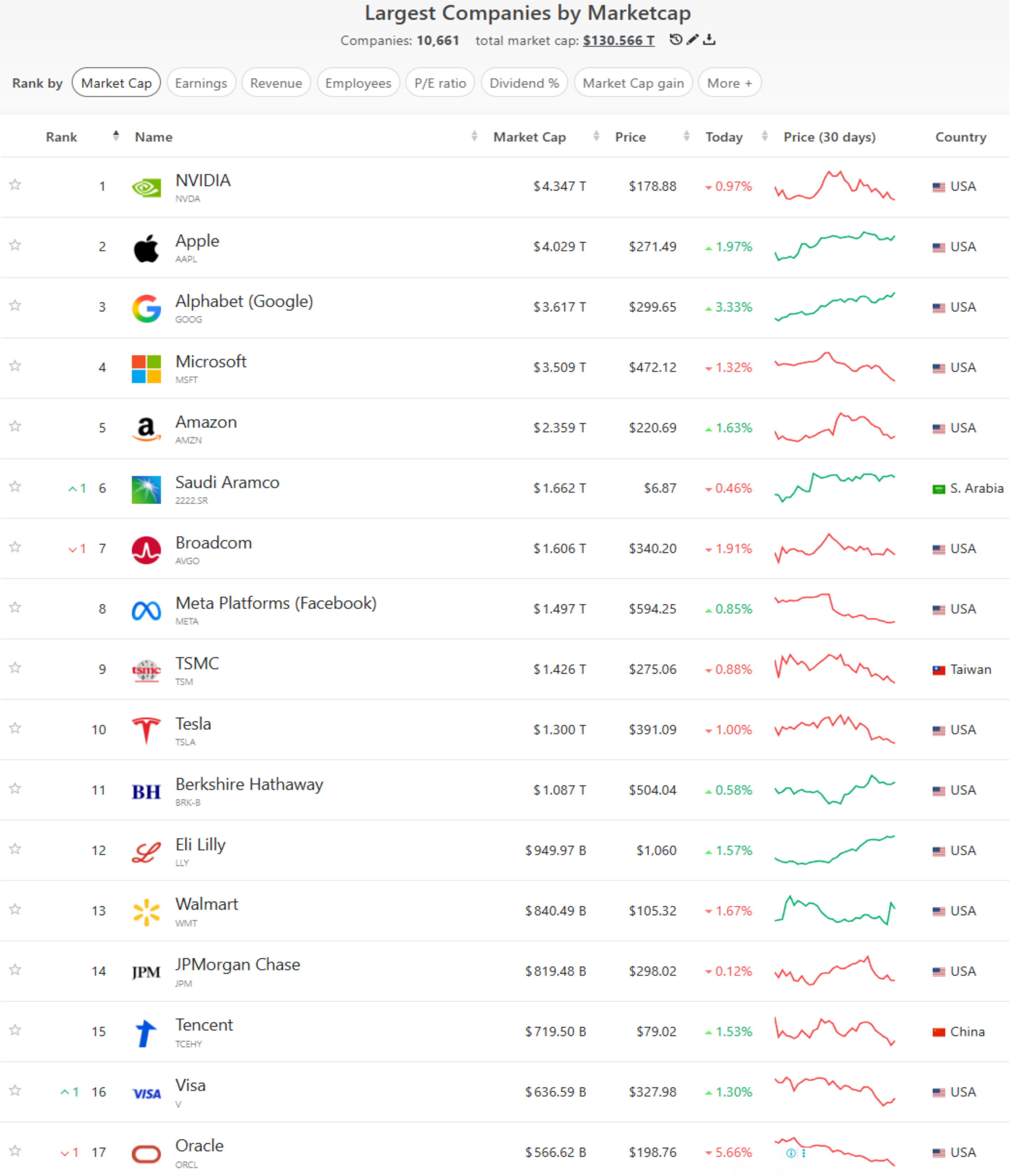
On November 20, 2025, OpenAI announced the global rollout of its ChatGPT group chat feature to all Free, Go, Plus, and Pro users. The group chat feature allows up to 20 users to collaborate with ChatGPT in the same conversation space. After a week of testing in Japan and New Zealand, OpenAI rapidly expanded the feature to global markets due to positive feedback.

The U.S. Department of Justice announced today that two Chinese nationals and two Americans have been arrested for allegedly illegally exporting advanced Nvidia chips with artificial intelligence (AI) applications to China. Approximately 400 Nvidia A100 GPUs were exported to China in two shipments between October 2024 and January 2025; two other shipments were cancelled due to law enforcement intervention.

The European Union (EU) and South Africa signed a mining cooperation agreement today, agreeing to strengthen cooperation in mineral and metal exploration, mining, and refining in the resource-rich country of South Africa to ensure the security of the supply chains needed for a green transition.

The French presidential palace stated that the leaders of France, Germany, and the United Kingdom today jointly called for a solution to the Russia-Ukraine war to involve Ukraine "fully" and for any decision to be based on "consensus" from Europe and NATO.

A series of smartphone robberies have recently occurred in London, with thieves displaying astonishing "professional eye," targeting only Apple iPhones and showing no interest in other brands of smartphones, even resorting to the absurd act of "returning" the stolen items. Sam, a 32-year-old London resident, was robbed by eight people near a Royal Mail warehouse in South London in January. The thieves pushed and shoved him, stealing his phone, camera, and even a beanie. Just as the group was about to leave, one of them returned, shoved his Android phone back into his hand, and said, "No Samsung." Another victim, Mark, had his phone stolen outside his company by a thief on an e-bike, but the thief stopped a few seconds later, looked at the phone, and threw it back on the ground. Mark eventually recovered his phone undamaged but admitted that "only my self-esteem was damaged." **(Good news for AAPL, bad news for Samsung)**



Largest Companies by Marketcap

Companies: **10,661** total market cap: **$130.566 T**

	Rank	Name	Market Cap	Price	Today	Price (30 days)	Country
☆	1	NVIDIA NVDA	$4.347 T	$178.88	▼0.97%		USA
☆	2	Apple AAPL	$4.029 T	$271.49	▲1.97%		USA
☆	3	Alphabet (Google) GOOG	$3.617 T	$299.65	▲3.33%		USA
☆	4	Microsoft MSFT	$3.509 T	$472.12	▼1.32%		USA
☆	5	Amazon AMZN	$2.359 T	$220.69	▲1.63%		USA
☆	⌃1 6	Saudi Aramco 2222.SR	$1.662 T	$6.87	▼0.46%		S. Arabia
☆	⌄1 7	Broadcom AVGO	$1.606 T	$340.20	▼1.91%		USA
☆	8	Meta Platforms (Facebook) META	$1.497 T	$594.25	▲0.85%		USA
☆	9	TSMC TSM	$1.426 T	$275.06	▼0.88%		Taiwan
☆	10	Tesla TSLA	$1.300 T	$391.09	▼1.00%		USA
☆	11	Berkshire Hathaway BRK-B	$1.087 T	$504.04	▲0.58%		USA
☆	12	Eli Lilly LLY	$949.97 B	$1,060	▲1.57%		USA
☆	13	Walmart WMT	$840.49 B	$105.32	▼1.67%		USA
☆	14	JPMorgan Chase JPM	$819.48 B	$298.02	▼0.12%		USA
☆	15	Tencent TCEHY	$719.50 B	$79.02	▲1.53%		China
☆	⌃1 16	Visa V	$636.59 B	$327.98	▲1.30%		USA
☆	⌄1 17	Oracle ORCL	$566.62 B	$198.76	▼5.66%		USA

GOOG is ranking number 3 now, surpassed MSFT, and will soon be number 2.
NVDA is the King of Kings, and GOOG is the Lord of Lords!

World Observation

Day 1370
Russia/Ukraine Conflict

CPU, GPU, TPU & ASIC

Market is adjusting, S&P fell from 6920 of Oct 29th to Oct 6602 of Nov 21st. Making people to think whether the current AI trend is a bubble, and when will it be burst. However, GOOG still keeps on going up and still break record high for 6 times in November which is on 10th, 11th, 12th, 17th, 19th and 20th. The market capitalization now ranks no 3 and already surpassed MSFT and will hit the level of 4 trillion soon. Already said it is the Lord of Lords, just after NVDA the King of Kings. Buffett reserved 31% cash thus he could buy at a ease.

The financial market is not just a game of numbers, but also a game of alphabets. Besides watching the price of Alphabet (GOOG), you still have to understand what is meant by GPU, CPU, TPU and ASIC. GOOG is developing the Ironwood, the 7th generation TPU. Recently, GOOG not only launched out Grok version 4.1, also the Gemini 3 which can speak Cantonese. Mind that the official spoken Chinese language is Mandarin, and Cantonese is a dialect or a language parallel to Mandarin, only in Hong Kong is treated as an official language.

GOOG has launched out robotaxi in fully driverless version as early as 2020 in Phoenix with limited access, and in Jun 2024 in San Franciso open to all, in Nov in Los Angeles open to all. Refer to The Express of 20251111. But no use, because they do not have the charm of Elon Musk and TSLA. Now everyone is waiting for the version 14.2 FSD of TSLA in December as if they are the first one in the world. By the end of 2024, even GOOG launched out their quantum computer and is still left behind in the rising trend of AI. But don't forget Trainium of AMZN is chasing after. Mind that AVGO a few years ago is still out to the top 20 market capitalization and now is already at number 7. It is also said to be the next NVDA. TSM is also said to be the next NVDA and is still on no 9 in ranking, apparently lower estimated.

All the above development including the OpenAI, AVGO, GOOG or AMZN, they need the chips designed by NVDA or AMD, and manufactured by TSM. Apparently, the market of NVDA with AMD is 9 to 1, TSM is irreplaceable. It is known as the Silicon shield of Taiwan. A lot of western countries would show their support to Taiwan in case if China attacks Taiwan, people say they are actually protecting TSMC rather than Taiwan. When without TSMC, most factories will be shut down within several weeks.

One thing you can say that once Berkshire purchased TSM in Q3 2022 at the quantity of 60.1 million shares, the average price is $82.65. Very soon, he sold 51.8 million shares at Q4 2022 at the average price of $72.34 (about 86.2% of the position) and in Q1 2023 sold all remaining 8.29 million shares at $89.76 because of geopolitical concerns. Unbelievable that he only held for 1 quarter. What is the use of copy trade? Once *The Economist* used the cover page to describe Taiwan as the most dangerous place on earth which means war is likely to break out. But since then, war has broken out in Ukraine and Gaza and other places and Taiwan is still safe. A KUNG FU master said *the most dangerous place is the safest!*

Anyway, Berkshire sold TSM, so it could not rise, Berkshire bought GOOG and it broke record high while market is falling. Any relations? GOOG launched out driverless taxi and for years and still could not make them popular, now even China developed later and is having export to Middle East and US still using in California only, because Elon Musk had not joined the competition in the past years. Since petroleum crisis in 1973, the whole world is developing e-car, but still not popular until Elon Musk joined the race. Now Elon Musk said he would join the manufacturing of semiconductor, what would happen? Some people look down upon him, some have high hopes, I will talk about this later.

The topic of today is to tell investor must look forward and have a farther eyesight than the market. When everyone is talking about GPU and CPU you have to know TPU. When everyone is talking about FSD and robotaxi you have to know *Kardashev Type II Civilization* (Space AI). When everyone is talking about Mars Landing and you go to buy those stocks, it will be too late. It is no use to tell you how low I have purchased PLTR, NVDA, GOOG, TEM and TSM, but I can tell you the importance of Space AI, TPU, ASIC and Mars related stocks now.

Bitcoin Price $84,074.79

Market Observation

Trading Philosophy

Warren Buffett has his own Philosophy of trading, that is you must have a target for trading. What is the purpose of your trading? That is when you fixed your target, you will aim at that target and would not be *Dust in the Wind* as in the current correction.

Make it down to earth, some want to earn more money so that he can set up his own business on their own. Some for emigration because they are not satisfied with their society, or else will be forever in day dreaming. Some want to earn enough money to have a trip on a luxurious liner to Mediterranean Sea, Caribbean Sea and then South Pole or North Pole, for during the trip he cannot go to work and only spend money for several months. Some want to earn money to study EMBA in world class universities, and learn finance in Wall Street, such as the ultimate target of a soccer player is at World Cup, an athlete is at Olympic Games, so a finance people should aim at Wall Street, no matter working or learning. If you don't have a target, will be easily as *Dust in the Wind*.

Let's take a look at DJIA, the adjustment has started just for 7 days, but there are 250 trading days in a year. The highest point is 48431 on Nov 12th, and closing of Friday is 46245, just 2186 points. Only 4.7%, my god! So, people are afraid of a slump or AI bubble. The lowest point of correction is still above the 100-SMA which considered as mid-term support, and previous low is 45452 of Oct 14th. No problem in chart analysis! S&P reached the top on Oct 29th at 6920, last closing is 6602, only 318 points, also 4.6%, but penetrated the 100-SMA and closed up above it, still higher than previous lower on Oct, still higher than previous low of Oct 10th @ 6550. The record high of NASDAQ is also on Oct 29th at 26182, the closing price of Friday penetrated the 100-SMA, but closed above it at 24239, a fall of 1943 points @ 7.4%, but still above the previous low. When in comparison with the circuit breaker of 2020, it's nothing at all. In Wall Street, a correction is 10% and when reaches 20%, it is in bear market, so all investors should prepare well to meet with a fall of 15% at any time and should treat it as usual. Even if bear market comes, sometimes can have a short bear market, that is golden pit.

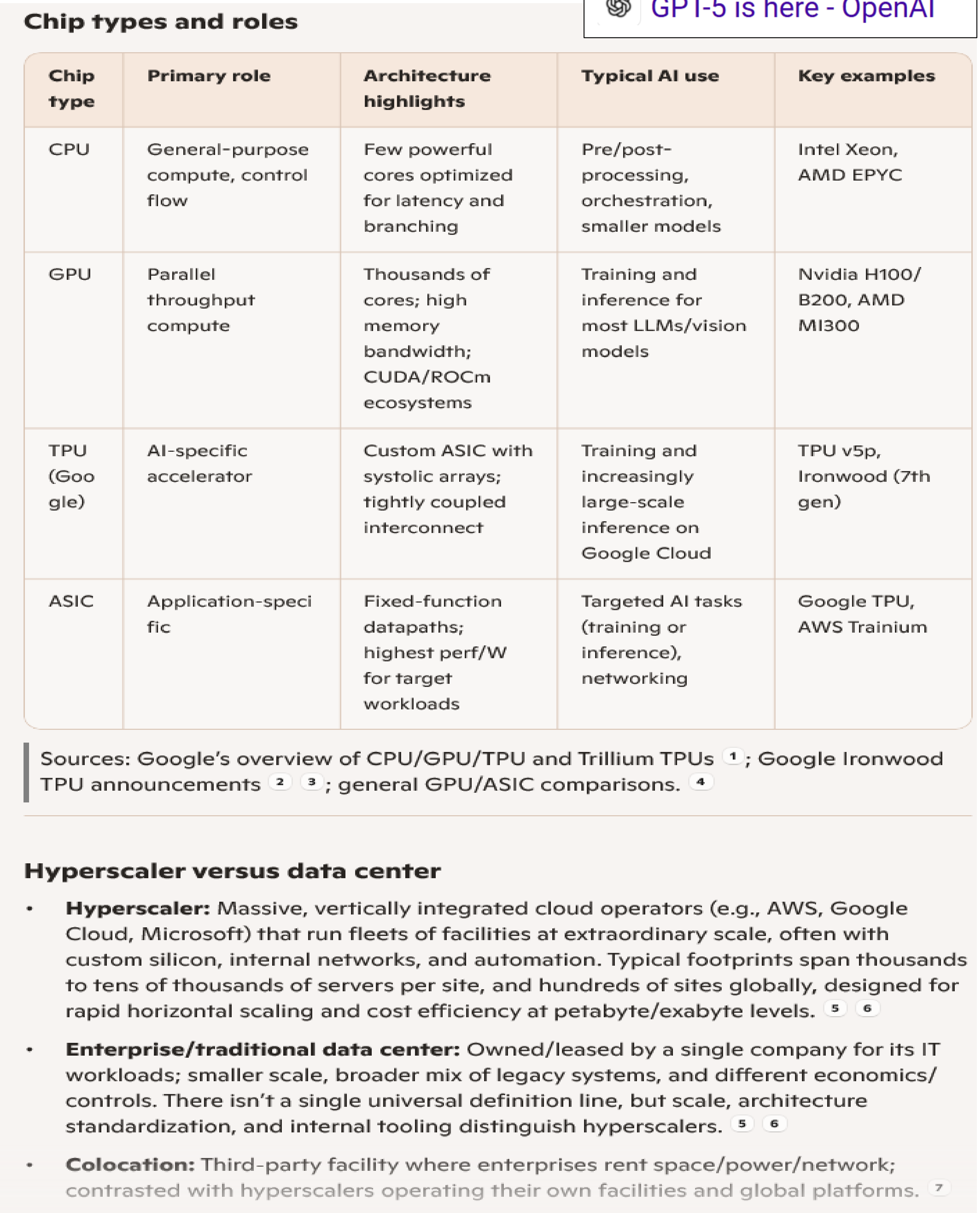
Current market situation can be said as mini golden pit. It is not likely to rebound in a single evening, but even if stay at current level, still normal. When investors want to buy at low, should buy those high above the 250-SMA. Those under the 250-SMA of course will have more room for rising, but takes longer time.

Such as defensive stocks, COST is always known as defensive, when it first went under the 250-SMA in July, people thought it would go up soon, but now goes back to the April bottom. V(Visa) again penetrated the 250-SMA this month and is near the low of April. That is why I always say, the top 10 market capitalization is already defensive enough and have the power of leading the market. Slow falling is meaning less when the recovery is even slower. A potential stock should have strong leading power to rise and to recover in a fall.

Don’t worry, every quarter stock index components will be re-organised, slow growth and sluggish stocks will be kicked off, energetic stocks and new stars will be added. That’s why the US market keep on rising. Don’t complain! The whole world is following US, just the same. Warren Buffett said he would like to hold stocks forever, it’s a little exaggerated. Anyway, hold them for Mars Landing for first stage, and second stage for Pay & Performance of Elon Musk is normal and the third stage is for Space AI. These 3 stages covering medium, long and longer-term. Unless you need money, no need to take out, it will be better than holding forever. For longer term scheme, buying at a fixed date for 12 consecutive months means you can buy very near to the mean average of the year. It is surely suitable for long term investment like aiming at the salary of Elon Musk, but it needs KUNG FU, not as easy as envisaged.

Anyway, buy before the trend blooms, no matter it is quantum computer, TPU, ASIC, driverless driving, health AI, Space AI or Mars Landing. Don’t buy those when everyone is talking. Warren Buffett held 31% cash in early November which is quite a high ratio recently. This is what copy trade cannot copy. So better do the trading on your own research. Too many cheap stocks to buy now, but too little capital. Must have a cleverer decision and not just rely on copy trade.

GPT-5 is here - OpenAI



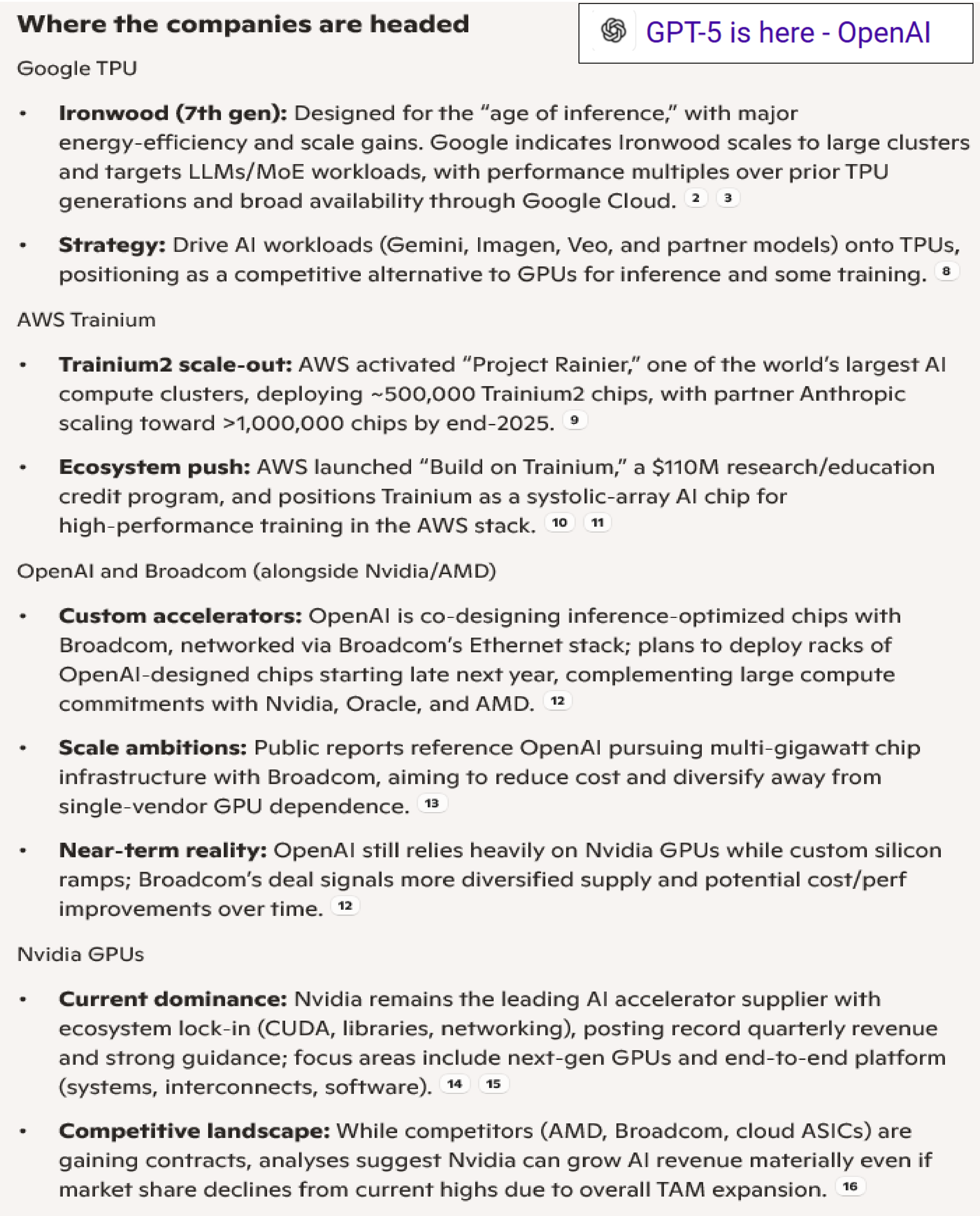
Chip types and roles

Chip type	Primary role	Architecture highlights	Typical AI use	Key examples
CPU	General-purpose compute, control flow	Few powerful cores optimized for latency and branching	Pre/post-processing, orchestration, smaller models	Intel Xeon, AMD EPYC
GPU	Parallel throughput compute	Thousands of cores; high memory bandwidth; CUDA/ROCm ecosystems	Training and inference for most LLMs/vision models	Nvidia H100/ B200, AMD MI300
TPU (Google)	AI-specific accelerator	Custom ASIC with systolic arrays; tightly coupled interconnect	Training and increasingly large-scale inference on Google Cloud	TPU v5p, Ironwood (7th gen)
ASIC	Application-specific	Fixed-function datapaths; highest perf/W for target workloads	Targeted AI tasks (training or inference), networking	Google TPU, AWS Trainium

> Sources: Google's overview of CPU/GPU/TPU and Trillium TPUs [1]; Google Ironwood TPU announcements [2] [3]; general GPU/ASIC comparisons. [4]

Hyperscaler versus data center

- **Hyperscaler:** Massive, vertically integrated cloud operators (e.g., AWS, Google Cloud, Microsoft) that run fleets of facilities at extraordinary scale, often with custom silicon, internal networks, and automation. Typical footprints span thousands to tens of thousands of servers per site, and hundreds of sites globally, designed for rapid horizontal scaling and cost efficiency at petabyte/exabyte levels. [5] [6]
- **Enterprise/traditional data center:** Owned/leased by a single company for its IT workloads; smaller scale, broader mix of legacy systems, and different economics/ controls. There isn't a single universal definition line, but scale, architecture standardization, and internal tooling distinguish hyperscalers. [5] [6]
- **Colocation:** Third-party facility where enterprises rent space/power/network; contrasted with hyperscalers operating their own facilities and global platforms. [7]



GPT-5 is here - OpenAI

Where the companies are headed

Google TPU

- **Ironwood (7th gen):** Designed for the "age of inference," with major energy-efficiency and scale gains. Google indicates Ironwood scales to large clusters and targets LLMs/MoE workloads, with performance multiples over prior TPU generations and broad availability through Google Cloud. [2] [3]
- **Strategy:** Drive AI workloads (Gemini, Imagen, Veo, and partner models) onto TPUs, positioning as a competitive alternative to GPUs for inference and some training. [8]

AWS Trainium

- **Trainium2 scale-out:** AWS activated "Project Rainier," one of the world's largest AI compute clusters, deploying ~500,000 Trainium2 chips, with partner Anthropic scaling toward >1,000,000 chips by end-2025. [9]
- **Ecosystem push:** AWS launched "Build on Trainium," a $110M research/education credit program, and positions Trainium as a systolic-array AI chip for high-performance training in the AWS stack. [10] [11]

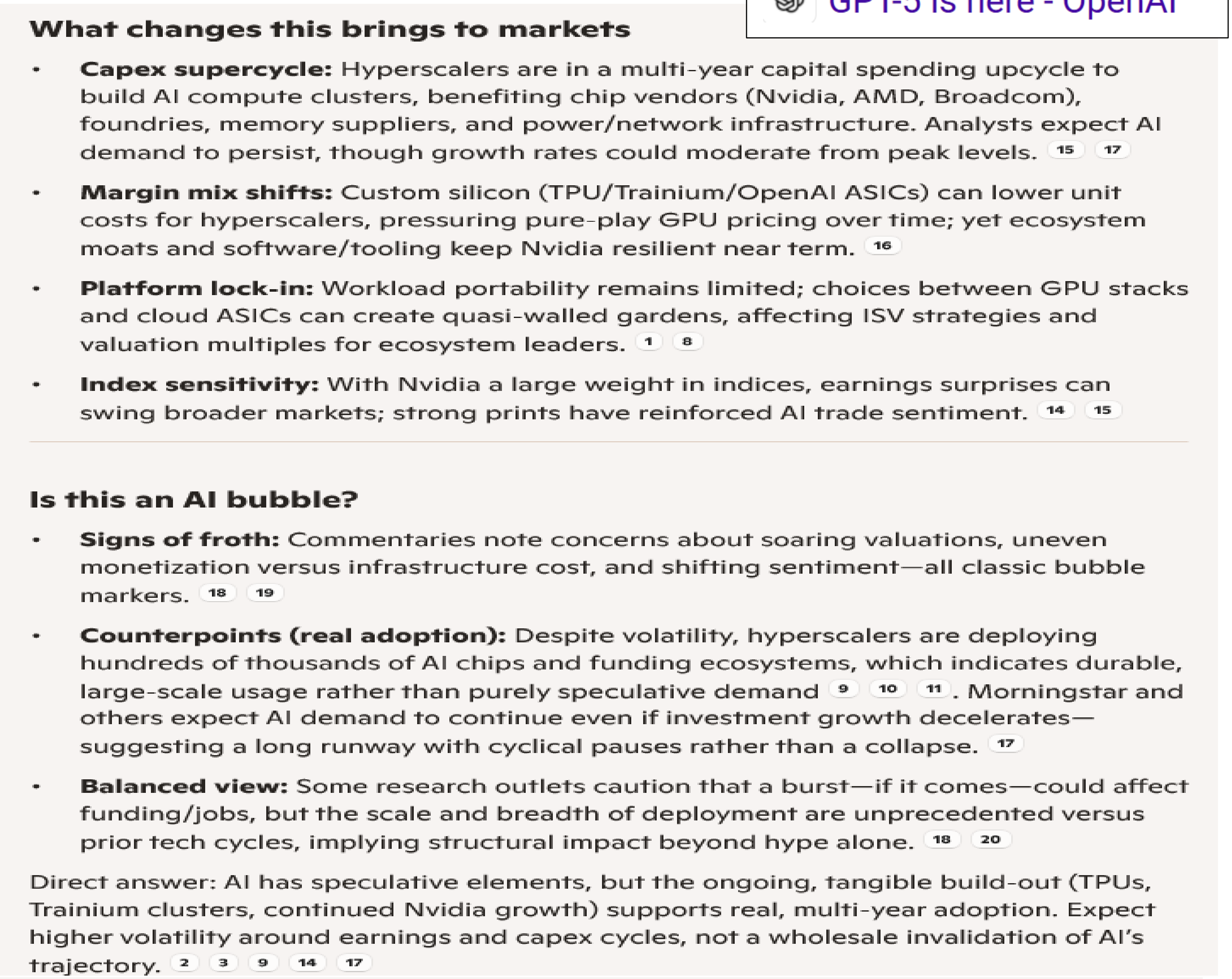
OpenAI and Broadcom (alongside Nvidia/AMD)

- **Custom accelerators:** OpenAI is co-designing inference-optimized chips with Broadcom, networked via Broadcom's Ethernet stack; plans to deploy racks of OpenAI-designed chips starting late next year, complementing large compute commitments with Nvidia, Oracle, and AMD. [12]
- **Scale ambitions:** Public reports reference OpenAI pursuing multi-gigawatt chip infrastructure with Broadcom, aiming to reduce cost and diversify away from single-vendor GPU dependence. [13]
- **Near-term reality:** OpenAI still relies heavily on Nvidia GPUs while custom silicon ramps; Broadcom's deal signals more diversified supply and potential cost/perf improvements over time. [12]

Nvidia GPUs

- **Current dominance:** Nvidia remains the leading AI accelerator supplier with ecosystem lock-in (CUDA, libraries, networking), posting record quarterly revenue and strong guidance; focus areas include next-gen GPUs and end-to-end platform (systems, interconnects, software). [14] [15]
- **Competitive landscape:** While competitors (AMD, Broadcom, cloud ASICs) are gaining contracts, analyses suggest Nvidia can grow AI revenue materially even if market share declines from current highs due to overall TAM expansion. [16]

GPT-5 is here - OpenAI



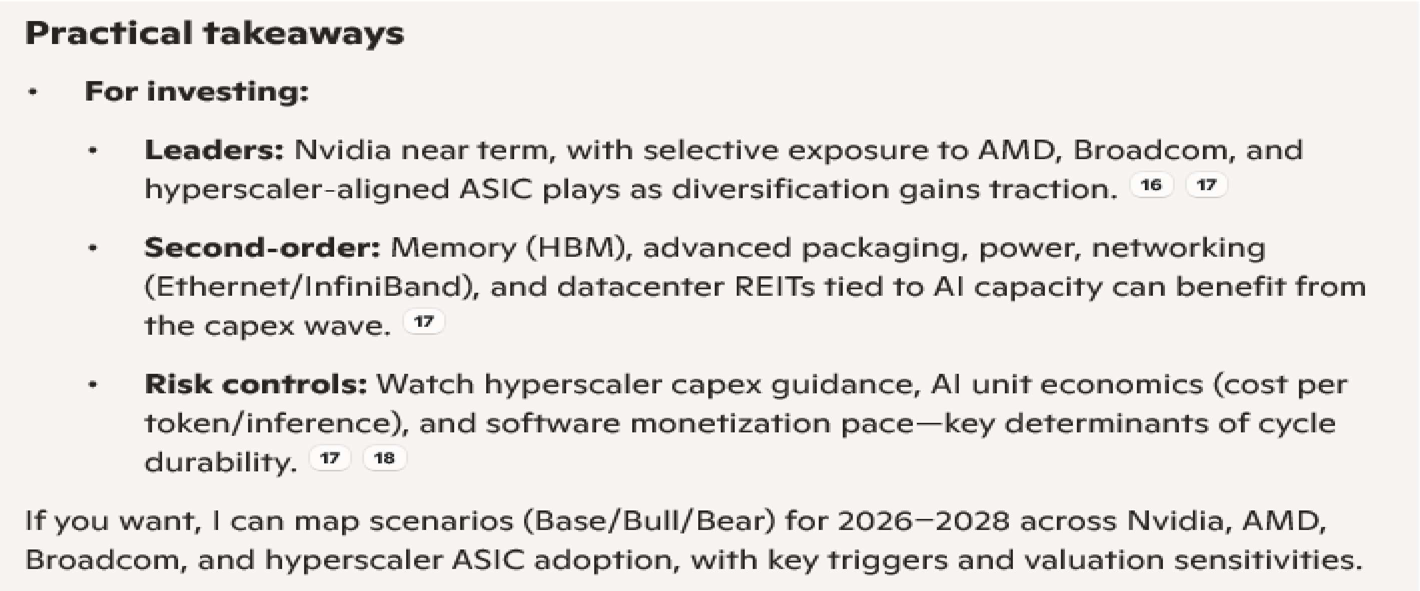
What changes this brings to markets

- **Capex supercycle:** Hyperscalers are in a multi-year capital spending upcycle to build AI compute clusters, benefiting chip vendors (Nvidia, AMD, Broadcom), foundries, memory suppliers, and power/network infrastructure. Analysts expect AI demand to persist, though growth rates could moderate from peak levels. [15] [17]
- **Margin mix shifts:** Custom silicon (TPU/Trainium/OpenAI ASICs) can lower unit costs for hyperscalers, pressuring pure-play GPU pricing over time; yet ecosystem moats and software/tooling keep Nvidia resilient near term. [16]
- **Platform lock-in:** Workload portability remains limited; choices between GPU stacks and cloud ASICs can create quasi-walled gardens, affecting ISV strategies and valuation multiples for ecosystem leaders. [1] [8]
- **Index sensitivity:** With Nvidia a large weight in indices, earnings surprises can swing broader markets; strong prints have reinforced AI trade sentiment. [14] [15]

Is this an AI bubble?

- **Signs of froth:** Commentaries note concerns about soaring valuations, uneven monetization versus infrastructure cost, and shifting sentiment—all classic bubble markers. [18] [19]
- **Counterpoints (real adoption):** Despite volatility, hyperscalers are deploying hundreds of thousands of AI chips and funding ecosystems, which indicates durable, large-scale usage rather than purely speculative demand [9] [10] [11]. Morningstar and others expect AI demand to continue even if investment growth decelerates—suggesting a long runway with cyclical pauses rather than a collapse. [17]
- **Balanced view:** Some research outlets caution that a burst—if it comes—could affect funding/jobs, but the scale and breadth of deployment are unprecedented versus prior tech cycles, implying structural impact beyond hype alone. [18] [20]

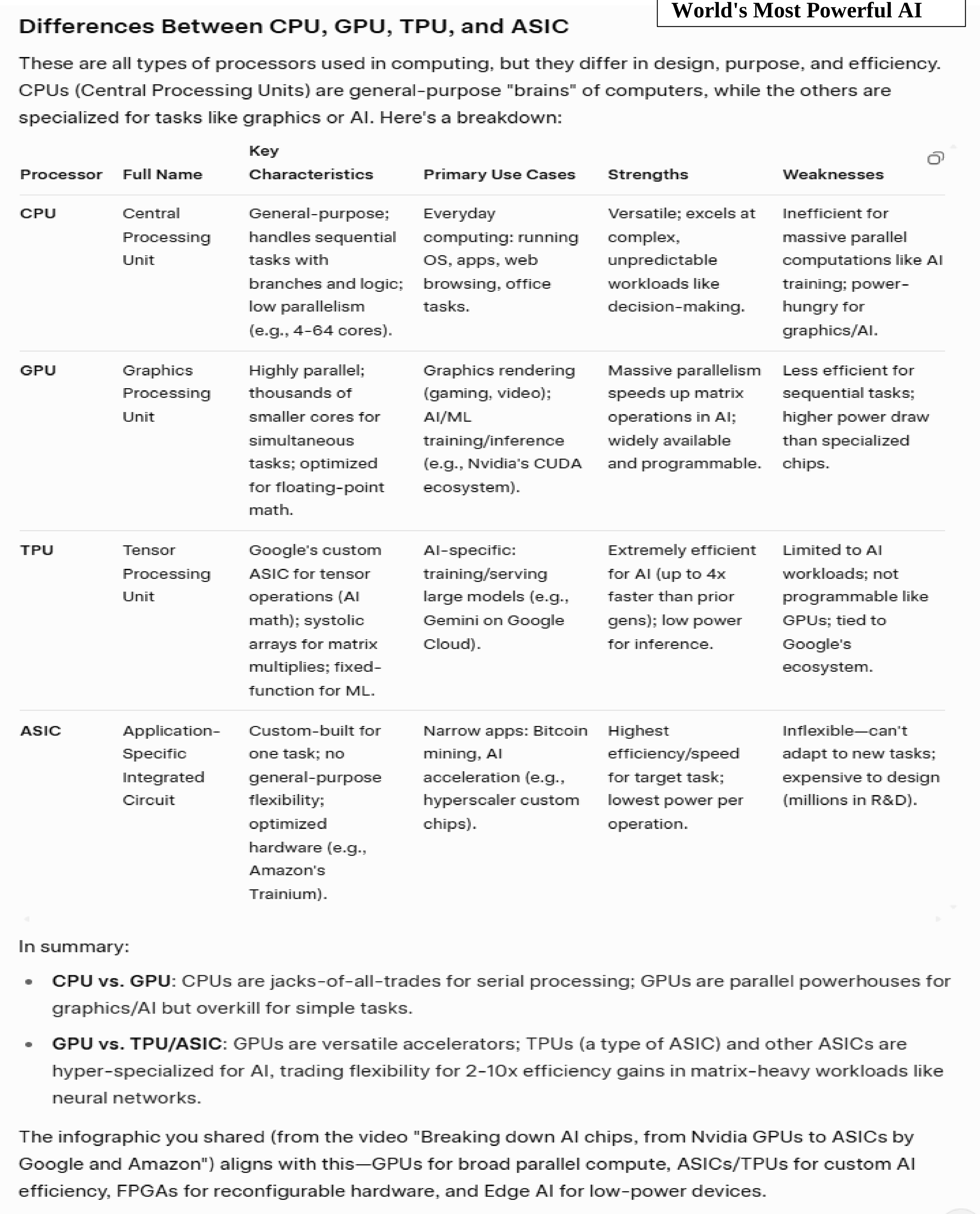
Direct answer: AI has speculative elements, but the ongoing, tangible build-out (TPUs, Trainium clusters, continued Nvidia growth) supports real, multi-year adoption. Expect higher volatility around earnings and capex cycles, not a wholesale invalidation of AI's trajectory. [2] [3] [9] [14] [17]



Practical takeaways

- **For investing:**
  - **Leaders:** Nvidia near term, with selective exposure to AMD, Broadcom, and hyperscaler-aligned ASIC plays as diversification gains traction. [16] [17]
  - **Second-order:** Memory (HBM), advanced packaging, power, networking (Ethernet/InfiniBand), and datacenter REITs tied to AI capacity can benefit from the capex wave. [17]
  - **Risk controls:** Watch hyperscaler capex guidance, AI unit economics (cost per token/inference), and software monetization pace—key determinants of cycle durability. [17] [18]

If you want, I can map scenarios (Base/Bull/Bear) for 2026–2028 across Nvidia, AMD, Broadcom, and hyperscaler ASIC adoption, with key triggers and valuation sensitivities.



Differences Between CPU, GPU, TPU, and ASIC

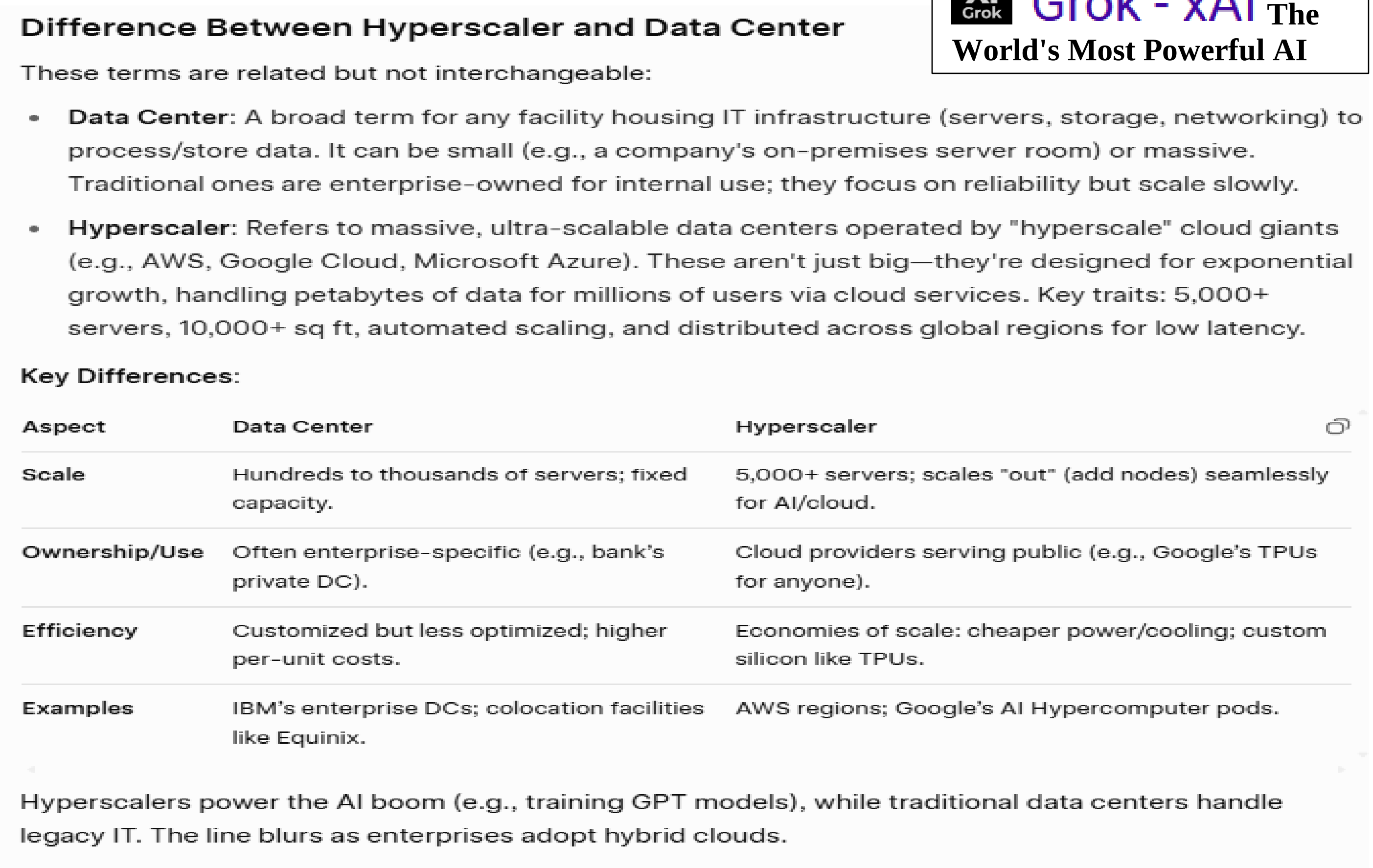
These are all types of processors used in computing, but they differ in design, purpose, and efficiency. CPUs (Central Processing Units) are general-purpose "brains" of computers, while the others are specialized for tasks like graphics or AI. Here's a breakdown:

Processor	Full Name	Key Characteristics	Primary Use Cases	Strengths	Weaknesses
CPU	Central Processing Unit	General-purpose; handles sequential tasks with branches and logic; low parallelism (e.g., 4-64 cores).	Everyday computing: running OS, apps, web browsing, office tasks.	Versatile; excels at complex, unpredictable workloads like decision-making.	Inefficient for massive parallel computations like AI training; power-hungry for graphics/AI.
GPU	Graphics Processing Unit	Highly parallel; thousands of smaller cores for simultaneous tasks; optimized for floating-point math.	Graphics rendering (gaming, video); AI/ML training/inference (e.g., Nvidia's CUDA ecosystem).	Massive parallelism speeds up matrix operations in AI; widely available and programmable.	Less efficient for sequential tasks; higher power draw than specialized chips.
TPU	Tensor Processing Unit	Google's custom ASIC for tensor operations (AI math); systolic arrays for matrix multiplies; fixed-function for ML.	AI-specific: training/serving large models (e.g., Gemini on Google Cloud).	Extremely efficient for AI (up to 4x faster than prior gens); low power for inference.	Limited to AI workloads; not programmable like GPUs; tied to Google's ecosystem.
ASIC	Application-Specific Integrated Circuit	Custom-built for one task; no general-purpose flexibility; optimized hardware (e.g., Amazon's Trainium).	Narrow apps: Bitcoin mining, AI acceleration (e.g., hyperscaler custom chips).	Highest efficiency/speed for target task; lowest power per operation.	Inflexible—can't adapt to new tasks; expensive to design (millions in R&D).

In summary:

- **CPU vs. GPU**: CPUs are jacks-of-all-trades for serial processing; GPUs are parallel powerhouses for graphics/AI but overkill for simple tasks.
- **GPU vs. TPU/ASIC**: GPUs are versatile accelerators; TPUs (a type of ASIC) and other ASICs are hyper-specialized for AI, trading flexibility for 2-10x efficiency gains in matrix-heavy workloads like neural networks.

The infographic you shared (from the video "Breaking down AI chips, from Nvidia GPUs to ASICs by Google and Amazon") aligns with this—GPUs for broad parallel compute, ASICs/TPUs for custom AI efficiency, FPGAs for reconfigurable hardware, and Edge AI for low-power devices.



Difference Between Hyperscaler and Data Center

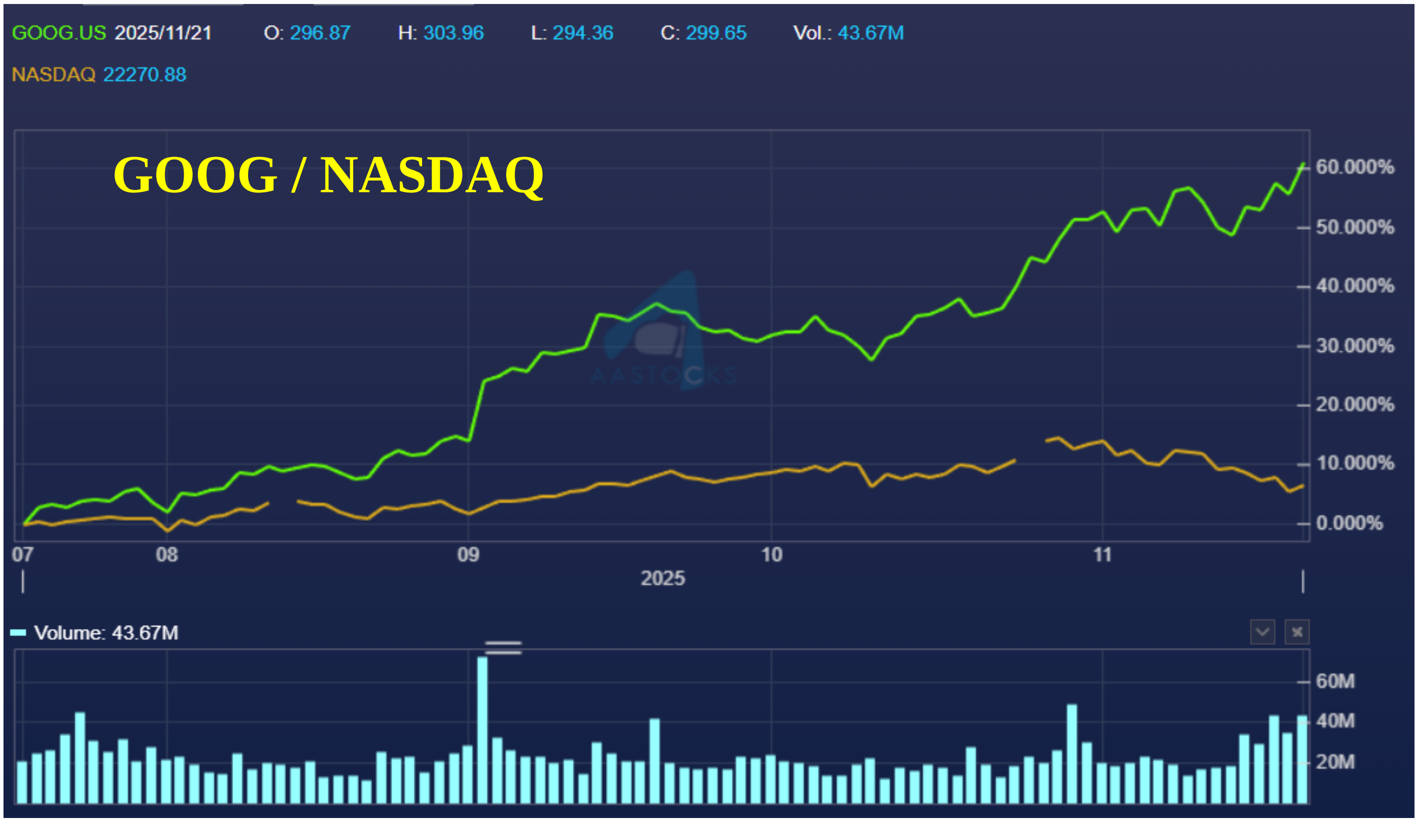
These terms are related but not interchangeable:

- **Data Center**: A broad term for any facility housing IT infrastructure (servers, storage, networking) to process/store data. It can be small (e.g., a company's on-premises server room) or massive. Traditional ones are enterprise-owned for internal use; they focus on reliability but scale slowly.
- **Hyperscaler**: Refers to massive, ultra-scalable data centers operated by "hyperscale" cloud giants (e.g., AWS, Google Cloud, Microsoft Azure). These aren't just big—they're designed for exponential growth, handling petabytes of data for millions of users via cloud services. Key traits: 5,000+ servers, 10,000+ sq ft, automated scaling, and distributed across global regions for low latency.

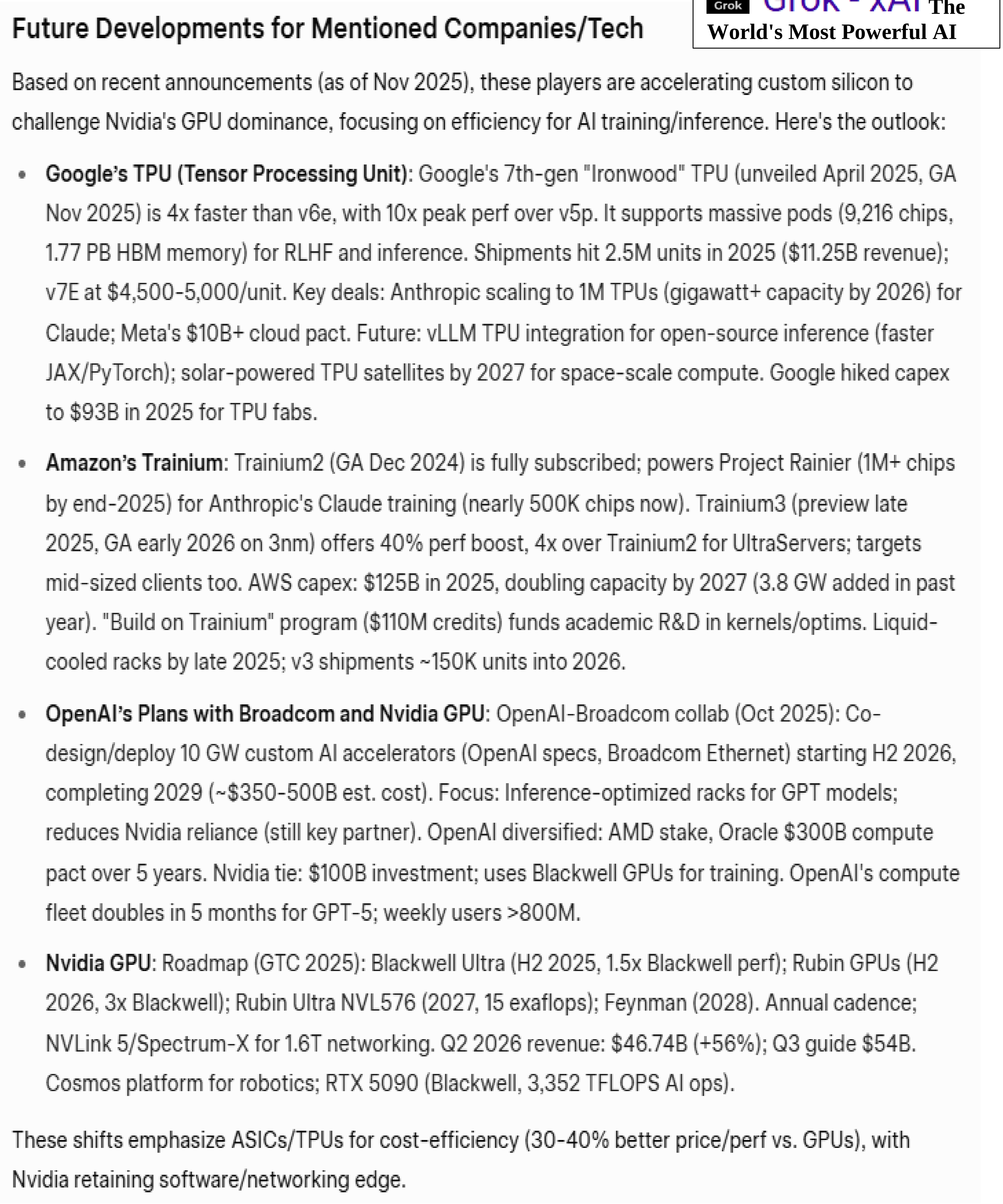
Key Differences:

Aspect	Data Center	Hyperscaler
Scale	Hundreds to thousands of servers; fixed capacity.	5,000+ servers; scales "out" (add nodes) seamlessly for AI/cloud.
Ownership/Use	Often enterprise-specific (e.g., bank's private DC).	Cloud providers serving public (e.g., Google's TPUs for anyone).
Efficiency	Customized but less optimized; higher per-unit costs.	Economies of scale: cheaper power/cooling; custom silicon like TPUs.
Examples	IBM's enterprise DCs; colocation facilities like Equinix.	AWS regions; Google's AI Hypercomputer pods.

Hyperscalers power the AI boom (e.g., training GPT models), while traditional data centers handle legacy IT. The line blurs as enterprises adopt hybrid clouds.



Upper line is GOOG, lower line is NASDAQ

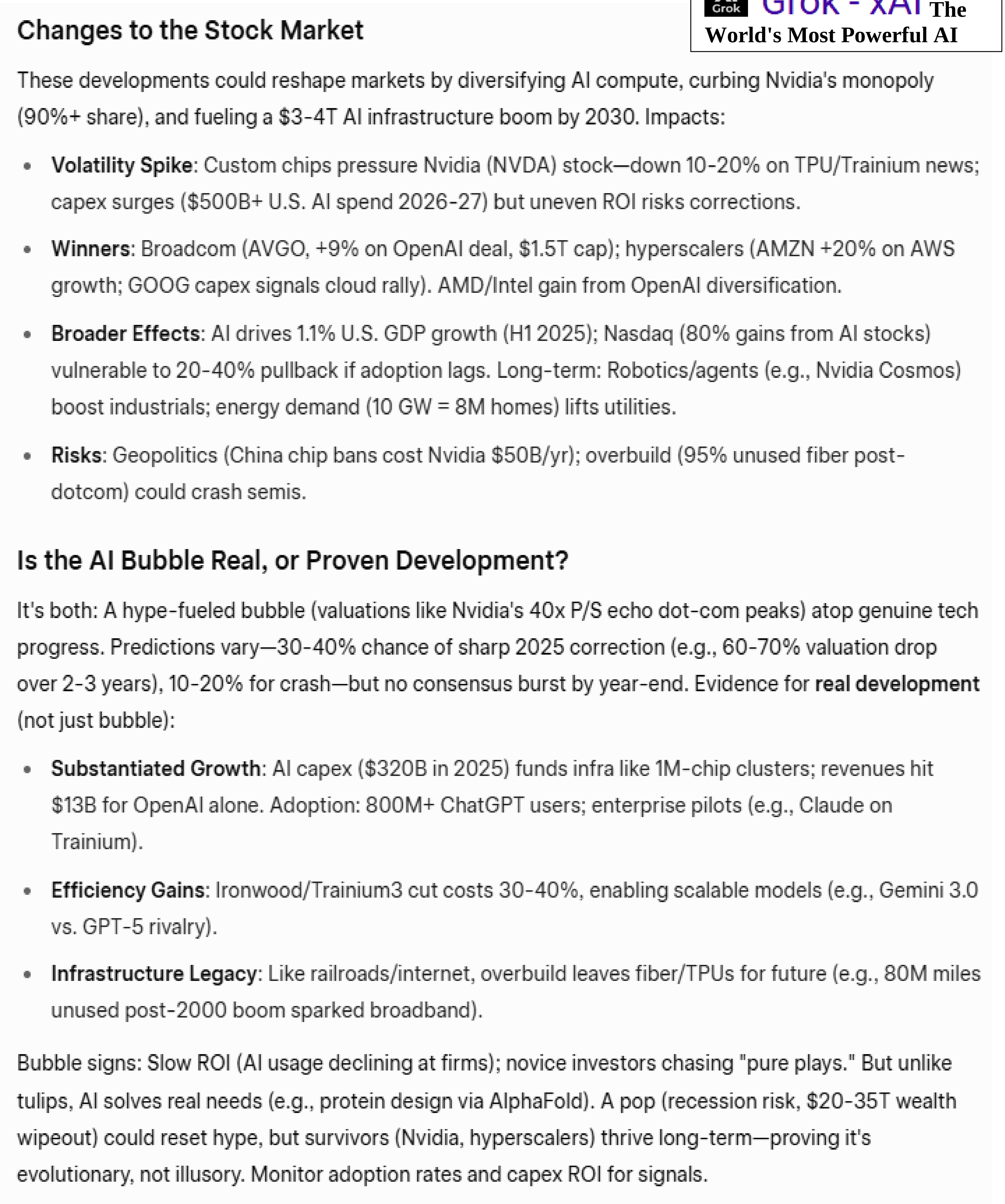


Future Developments for Mentioned Companies/Tech

Based on recent announcements (as of Nov 2025), these players are accelerating custom silicon to challenge Nvidia's GPU dominance, focusing on efficiency for AI training/inference. Here's the outlook:

- **Google's TPU (Tensor Processing Unit)**: Google's 7th-gen "Ironwood" TPU (unveiled April 2025, GA Nov 2025) is 4x faster than v6e, with 10x peak perf over v5p. It supports massive pods (9,216 chips, 1.77 PB HBM memory) for RLHF and inference. Shipments hit 2.5M units in 2025 ($11.25B revenue); v7E at $4,500-5,000/unit. Key deals: Anthropic scaling to 1M TPUs (gigawatt+ capacity by 2026) for Claude; Meta's $10B+ cloud pact. Future: vLLM TPU integration for open-source inference (faster JAX/PyTorch); solar-powered TPU satellites by 2027 for space-scale compute. Google hiked capex to $93B in 2025 for TPU fabs.
- **Amazon's Trainium**: Trainium2 (GA Dec 2024) is fully subscribed; powers Project Rainier (1M+ chips by end-2025) for Anthropic's Claude training (nearly 500K chips now). Trainium3 (preview late 2025, GA early 2026 on 3nm) offers 40% perf boost, 4x over Trainium2 for UltraServers; targets mid-sized clients too. AWS capex: $125B in 2025, doubling capacity by 2027 (3.8 GW added in past year). "Build on Trainium" program ($110M credits) funds academic R&D in kernels/optims. Liquid-cooled racks by late 2025; v3 shipments ~150K units into 2026.
- **OpenAI's Plans with Broadcom and Nvidia GPU**: OpenAI-Broadcom collab (Oct 2025): Co-design/deploy 10 GW custom AI accelerators (OpenAI specs, Broadcom Ethernet) starting H2 2026, completing 2029 (~$350-500B est. cost). Focus: Inference-optimized racks for GPT models; reduces Nvidia reliance (still key partner). OpenAI diversified: AMD stake, Oracle $300B compute pact over 5 years. Nvidia tie: $100B investment; uses Blackwell GPUs for training. OpenAI's compute fleet doubles in 5 months for GPT-5; weekly users >800M.
- **Nvidia GPU**: Roadmap (GTC 2025): Blackwell Ultra (H2 2025, 1.5x Blackwell perf); Rubin GPUs (H2 2026, 3x Blackwell); Rubin Ultra NVL576 (2027, 15 exaflops); Feynman (2028). Annual cadence; NVLink 5/Spectrum-X for 1.6T networking. Q2 2026 revenue: $46.74B (+56%); Q3 guide $54B. Cosmos platform for robotics; RTX 5090 (Blackwell, 3,352 TFLOPS AI ops).

These shifts emphasize ASICs/TPUs for cost-efficiency (30-40% better price/perf vs. GPUs), with Nvidia retaining software/networking edge.



Changes to the Stock Market

These developments could reshape markets by diversifying AI compute, curbing Nvidia's monopoly (90%+ share), and fueling a $3-4T AI infrastructure boom by 2030. Impacts:

- **Volatility Spike**: Custom chips pressure Nvidia (NVDA) stock—down 10-20% on TPU/Trainium news; capex surges ($500B+ U.S. AI spend 2026-27) but uneven ROI risks corrections.
- **Winners**: Broadcom (AVGO, +9% on OpenAI deal, $1.5T cap); hyperscalers (AMZN +20% on AWS growth; GOOG capex signals cloud rally). AMD/Intel gain from OpenAI diversification.
- **Broader Effects**: AI drives 1.1% U.S. GDP growth (H1 2025); Nasdaq (80% gains from AI stocks) vulnerable to 20-40% pullback if adoption lags. Long-term: Robotics/agents (e.g., Nvidia Cosmos) boost industrials; energy demand (10 GW = 8M homes) lifts utilities.
- **Risks**: Geopolitics (China chip bans cost Nvidia $50B/yr); overbuild (95% unused fiber post-dotcom) could crash semis.

Is the AI Bubble Real, or Proven Development?

It's both: A hype-fueled bubble (valuations like Nvidia's 40x P/S echo dot-com peaks) atop genuine tech progress. Predictions vary—30-40% chance of sharp 2025 correction (e.g., 60-70% valuation drop over 2-3 years), 10-20% for crash—but no consensus burst by year-end. Evidence for **real development** (not just bubble):

- **Substantiated Growth**: AI capex ($320B in 2025) funds infra like 1M-chip clusters; revenues hit $13B for OpenAI alone. Adoption: 800M+ ChatGPT users; enterprise pilots (e.g., Claude on Trainium).
- **Efficiency Gains**: Ironwood/Trainium3 cut costs 30-40%, enabling scalable models (e.g., Gemini 3.0 vs. GPT-5 rivalry).
- **Infrastructure Legacy**: Like railroads/internet, overbuild leaves fiber/TPUs for future (e.g., 80M miles unused post-2000 boom sparked broadband).

Bubble signs: Slow ROI (AI usage declining at firms); novice investors chasing "pure plays." But unlike tulips, AI solves real needs (e.g., protein design via AlphaFold). A pop (recession risk, $20-35T wealth wipeout) could reset hype, but survivors (Nvidia, hyperscalers) thrive long-term—proving it's evolutionary, not illusory. Monitor adoption rates and capex ROI for signals.