12 Jul 2026 21:15:43 ET | 13 pages

US Semiconductors and Hardware

Weekly Investor Pulse: RAISE Summit Takeaways

CITI'S TAKE

We tuned into the RAISE 2026 AI Summit last week and recap key takeaways. We maintain AMD #1 (agentic CPU demand), CBRS #2 (fast inference pure play), and like NVDA #1/AVGO #2 in value mega cap compute semis.

Atif Malik[AC]

Asiya Merchant, CFA

Papa Sylla

AI Demand: The explosive growth of token-intensive agentic workflows and real-world code generation applications is driving unprecedented demand for AI inference infrastructure. **Machine learning compute requirements are currently expanding at roughly 10x year-over-year.** Unlike the dot-com era, today's AI infrastructure spending is being underwritten by companies generating hundreds of millions—and in many cases billions—of dollars in tangible revenue. While the demand outlook remains exceptionally strong, the increasing reliance on private credit and debt financing represents a risk factor for the investment cycle.

AI Interns Replacing Traditional Chatbots: Leading AI labs are increasingly moving beyond conventional chatbots and towards AI-native "interns" equipped with coding tools and autonomous research functions. Advanced models are now recursively accelerating their own development by assisting in engineering, software creation, and research workflows. **This self-reinforcing flywheel is compressing major model release cycles from the historical six-to-nine-month cadence toward monthly iterations, significantly increasing the pace of innovation.**

Hardware Disaggregation: Hyperscalers are rapidly adopting disaggregated inference architectures to extract greater efficiency from existing Nvidia GPU deployments. Rather than relying solely on GPU AI clusters, operators are building specialized infrastructure optimized for distinct workloads, including latency-sensitive inference, real-time tool use, and agentic task execution. **This architectural shift is becoming increasingly important as AI workloads diversify.**

The Resurgence of CPUs: While GPUs remain the dominant platform for model training, **CPUs are experiencing a significant resurgence as agentic AI workloads scale.** Complex inference pipelines, orchestration layers, reinforcement learning environments, and enterprise AI agents increasingly rely on CPU-intensive processing, expanding the addressable market for general-purpose compute.

Memory Squeeze: Data centers are on track to consume approximately 70% of global memory-chips, creating supply constraints that could ripple across the

See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations

broader electronics ecosystem. Major hyperscalers are already pursuing multi-year supply agreements to secure production capacity. **Industry supply-demand dynamics suggest that meaningful relief may not arrive until at least 2028, when new manufacturing capacity is expected to come online**.

Hyperscaler Custom Silicon: Traditional CPU and GPU architectures are increasingly constrained by networking, power, and efficiency bottlenecks. In response, hyperscalers are accelerating investment in custom silicon purpose-built for AI workloads. These chips are optimized around specific models, inference tasks, and agentic workflows running on internal infrastructure. **Google's latest roadmap illustrates this specialization, with separate silicon platforms dedicated to low-latency inference (8i) and large-scale model training (8t), highlighting the industry's broader move toward workload-specific compute architectures.**

The KV Cache Bottleneck: As the cost of multi-turn agentic workflows continues to rise, hyperscalers are increasingly focused on optimizing context memory management. A key strategy involves offloading KV cache data from expensive GPUs to lower-cost storage layers, reducing redundant computations and improving system utilization. **These techniques can deliver cache-hit rates exceeding 95%, reduce time-to-first-token latency by 20x, and materially lower inference costs.** Moreover, because flash is in shortage, hybrid storage architectures including hard disk drives are seeing a resurgence for massive AI datasets.

Latency as a Competitive Advantage: Within enterprise environments, inference speed is becoming a critical determinant of user adoption and productivity. Faster response times reduce workflow friction, minimize context switching, and encourage users to tackle increasingly sophisticated coding and analytical tasks. **LLM latest models can run at 750 tokens per second on Cerebras hardware or an order of magnitude faster than standard GPU fleets.**

Edge AI and Small Language Models: Growing pressure on centralized cloud infrastructure is accelerating the shift toward edge computing and specialized small language models (SLMs). Enterprises are increasingly deploying task-specific models that deliver domain expertise while operating on substantially lighter hardware footprints. **Chips like the Nvidia RTX Spark superchip are enabling Ai agents to run natively on consumer and commercial Windows devices without cloud reliance.**

Advanced Micro Devices

(AMD.O; US$557.89; 1; 10 Jul 26; 16:00)

Valuation

Our TP on AMD is $575. We value AMD using SOTP analysis with data center GPU/CPU biz at $281/$204 (35/30x 2028 PE), client $30 (15x), gaming $2 (8x), embedded $24 (20x), and net cash per share ~$35. Our target PE multiples are in-line with respective peer group averages.

Risks

Competition: AMD competes directly with Intel in the microprocessor market. Consequently, any fluctuations in market share between AMD and Intel could result in risk to our estimates. AMD competes directly with NVIDIA in the graphics and AI GPU market. Consequently, any fluctuations in market share between AMD and NVIDIA could result in risk to our estimates.

PC End Market: AMD derives roughly 30% of sales from the PC industry, which is highly dependent on IT spending. Therefore, any major uptick/downtick in IT spending could result in upside/downside to our estimates and rating for AMD.

Customer Risk: AMD derives roughly 15% of revenue from Sony and Microsoft. As a result, any major increases/decreases in orders from Sony or Microsoft could have a positive/negative effect on our estimates for AMD.

Valuation: While AMD is currently trading slightly above its historical range, revisions to consensus revenue estimates could result in high volatility to AMD stock.

Tender Offer: Any public or private tender offer for all or part of AMD's business could drive the stock higher and cause us to re-assess our rating.

If the impact on the company from any of these factors proves to be greater than we anticipate, the stock will likely have difficulty achieving our target price. However, should they be less than anticipated, the stock could trade above our target price.

Broadcom Inc

(AVGO.O; US$399.97; 1; 10 Jul 26; 16:00)

Valuation

Our PT of $500, or 20X F28 EPS, is at the low end of recent trading range of 20-40X to reflect intensifying competition.

Risks

We see the following key risks to our target price:

Customer: We estimate it derives ~35-40% of sales from its largest customer, Google. Therefore, any major uptick/downtick in demand for Google's products could result in upside/downside to our estimates and rating for Broadcom.

SG&A target: Broadcom's long-term SG&A target is <3% of revenue (excluding stock-based compensation), one of the lowest among semi peers. However, if Broadcom is unable to reach its SG&A target, it could result in downside to our estimates and rating for Broadcom.

Competition: Any fluctuations in market share between Broadcom and its competitors could result in upside/downside risk to our estimates.

M&A: If Broadcom pursues further acquisitions, it could lead to margin/EPS accretion/dilution and result in upside/downside to our estimates.

If the impact on the company from any of these factors proves to be greater than we anticipate, the stock will likely have difficulty achieving our target price. However, should they be less than anticipated, the stock could trade above our target price.

Cerebras Systems Inc

(CBRS.O; US$215.08; 1; 10 Jul 26; 16:00)

Valuation
We use EV/S valuation methodology to value CBRS given Cerebras is in its initial post-IPO hyper growth phase. We assume 2028 as our base line year as core revenues and gross margins normalize after the initial cloud build out.Our base case valuation applies NVDA's 3-year EV/S average of 15x to our 2028 projected core sales of $7.1B for a target price of $340.

Risks
This stock is High Risk based upon our quantitative model but assigning a High Risk rating is not supported by other qualitative factors such as Cerebras' first mover advantage in fast inference market with $130Bn TAM by 2030 with compelling token speed advantage thanks to its unique wafer technology and so a High Risk rating has not been applied.

Competition, customer concentration, and execution to ramp cloud service capacity are the biggest downside risks to our target price.

NVIDIA Corp
(NVDA.O; US$210.96; 1; 10 Jul 26; 16:00)

Valuation
Our price target for NVDA of $300 is based on ~28x P/E on C27E earnings power of ~$11.9 (incl. SBC) discounted back. Our 28x P/E multiple is in-line with 3 year average.

Risks
Downside risks to the attainment of our target price include: 1) competition on gaming could drive the stock lower if Nvidia loses market share; 2) slower-than-expected adoption of new platforms can drive lower data center and gaming sales; 3) lumpiness in auto and data center markets can add volatility to the stock/multiple; and 4) cryptomining impact on gaming sales.

