16 Sep 2026 04:58:36 ET | 14 pages

US Semiconductors

AI Infra Summit Day 1: AI Safety & Alignment Will Require MORE Compute, NOT Less

CITI'S TAKE

We attended Day 1 of AI Infra Summit in Santa Clara, CA and focused on compute specific takeaways. Fundamentally, frontier AI model developers see AI safety and alignment as a compute-intensive process, implying that safety concerns may drive more, rather than less, demand for compute. This runs counter to a key investor concern that tighter safety requirements could slow the pace of frontier model development - and, by extension, compute demand (see our preview).

Atif Malik[AC]

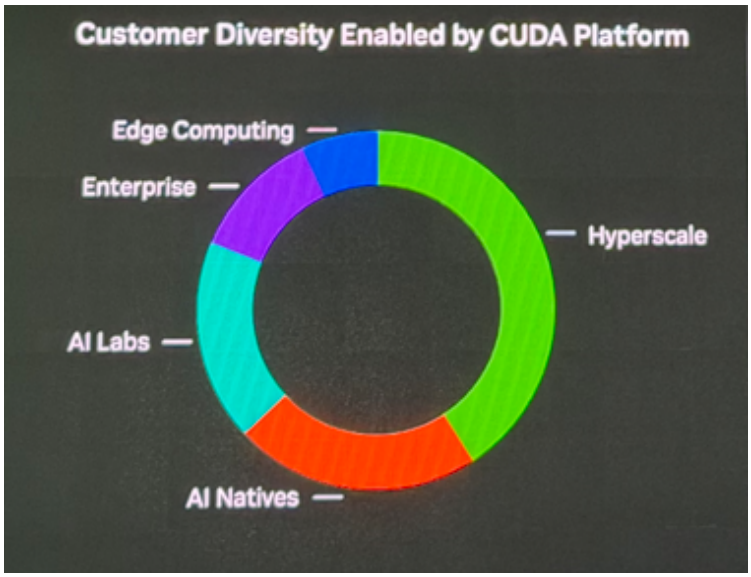
James Bowlin

Frontier AI Model Developers: Frontier AI model developers do not see compute scaling slowing down. Instead, the bottleneck is shifting from model architecture innovation alone towards securing massive amounts of compute, power, and infrastructure, while improving efficiency through platforms such as Astra and next-generation hardware like Vera Rubin. The implication is continued strong demand across the AI infrastructure stack: GPUs, networking, power, cooling, and data center capacity. Next-generation facilities could accommodate 30-40% more GPUs, improving infrastructure efficiency and helping support growing AI demand. Vera Rubin is a major upgrade with 3x token throughput improvement out of the box and 5x post optimization. **Importantly, safety and alignment require more compute not less. The expectation is that spending on both training and inference infrastructure will continue increasing as frontier models become more capable and require more extensive testing and alignment.** Astra as an example delivers returns through end-to-end system optimization, implying that performance gains come from both better hardware and better software orchestration.

Nvidia (Ian Buck, VP of Hyperscale and HPC): CUDA platform enables customer diversity (see Figure 1). Agentic AI is 100x more demanding than Chat in 2023 with ISL/KV Cache Size/Number of Turns/Sub-Agents moving to 142K/1M/65/4 from 1K/4K/3/0 respectively and new benchmarks are needed to represent today's workloads. **Nvidia Vera Rubin is extreme co-designed for agents and full stack co-design delivers up to 30x better AI factory throughput on agentic workloads.** Groq 3 LPX delivers long context with extreme token speed. Vera CPU is the fastest CPU and max single-threaded CPU at scale. NVLink Fusion connects XPUs to Nvidia AI Infrastructure e.g. NVLink fusion will connect next gen d-Matrix XPUs to Nvidia AI platform. D-Matrix is pursuing a memory-first AI inference architecture that complements GPUs rather than replacing them. Corsair provides an efficient SRAM-based inference platform today, while Raptor seeks to extend that advantage with high-capacity 3D stacked DRAM. D-Matrix strategy is centered on lowering inference costs and improving performance through more efficient data movement and heterogeneous system design.

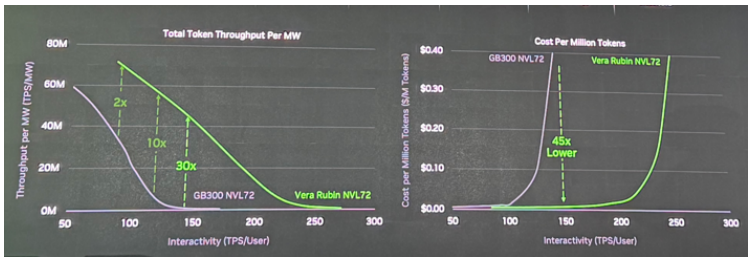
See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations

Figure 1. NVIDIA CUDA Platform Diversity



Source: NVIDIA

Figure 2. NVIDIA BW vs VR Token Performance Comparison



Source: NVIDIA

Intel (Lip-Bu Tan, CEO): The overarching message was that Intel is repositioning itself from a traditional CPU vendor toward a full-stack AI infrastructure company. The strategy relies on: a) organizational simplification, b) stronger software capabilities, c) advanced packaging leadership, and d) memory and system-level innovation. Intel sees the next phase of AI competition shifting from raw GPU performance toward systems optimization, memory architectures, software utilization, and infrastructure efficiency. Lip-Bu emphasized flattening Intel's organizational structure, reducing management layers from roughly 10–12 to 4–5 to improve decision-making speed. **Lip-Bu does not see current AI cycle as a reason for concern and believes AI's eventual impact will be larger than that of the internet.** Responsible adoption remains critical, with security and privacy highlighted as prerequisites for broader acceptance. CPUs remain highly relevant for reinforcement learning, agentic AI, orchestration layers, and general-purpose compute workloads. 18A is in volume production with 14A approaching launch, 0.9 PDK expected in October, and 1.0 PDK expected next quarter.

Meta (Santosh Janardhan, Head of Infrastructure): Meta shared its infrastructure build out scaling experience and lessons learnt on Titan AI training cluster Prometheus 1GW+ and towards Hyperion 5GW over the next few years. The biggest challenge according to the company is in building large cluster data centers built across multiple regions is the loss-less

network and reduce GPU cluster to GPU cluster interruptions. Meta deployed proprietary BAG (back-end aggregation) architecture to improve networking and meet its 2 interruptions per 8K GPUs per day target.



Intel Corp

(INTC.O; US$97.14; 1; 15 Sep 26; 16:00)

Valuation

We value INTC based on our SOTP valuation, our price target is $130. We apply a SOTP valuation framework to each of Intel's reportable segments and to highlight our growth expectations for Intel Foundry.

Risks

PC end-market: Intel derives roughly 90% of sales from the PC and server segments, which are highly dependent on IT spending. Therefore, any major uptick/downtick in IT spending could result in upside/downside to our estimates and cause us to reassess our rating on INTC.
M&A: If Intel engages in a substantial accretive acquisition, it could result in upside to our estimates.
Competition: Intel competes directly with AMD in the microprocessor market. Consequently, any fluctuations in market share between Intel and AMD could result in risk to our estimates.
Customer risk: Intel's largest PC OEM customers collectively represent roughly 45% Intel's revenue. As a result, any major declines in orders from PC OEMs could have a negative effect on our estimates for Intel.
Macroeconomic: Intel's geographic exposure spans multiple geographies including the U.S., Europe, and Asia. As a result, any prolonged macroeconomic downturn/upturn could result in downside/upside to our estimates and rating.
If the impact from the above risks turns out to be greater/less than we anticipate, the shares could fail to achieve or could exceed our target price.

Meta Platforms Inc

(META.O; US$670.24; 1; 15 Sep 26; 16:00)

Valuation

Our target price of $800 is based on ~22x our 2027E GAAP EPS of $36.92 and implies ~11x our 2027E EV/EBITDA. We justify our valuation methodology based on Meta's user scale, growing engagement, product super cycle and AI opportunity, ads innovation, and revenue growth, and we're watching investments and overall profitability.

Risks

Risks to our target price include: 1) Impacts from changes to ATT and IDFA could prove more difficult to overcome and structurally impact Meta's advertising business; 2) Online advertising spend could decelerate faster than we anticipate; 3) Competition could prove to be more impactful than we anticipate; 4) Meta may not be able to achieve meaningful contributions from its investments in non-advertising businesses (metaverse). 5) Adverse economic and/or consumer spending trends could hinder advertising spend; and 6) Regulatory pressures on anti-trust, data, and privacy practices could adversely impact on performance.

NVIDIA Corp

(NVDA.O; US$212.17; 1; 15 Sep 26; 16:00)

Valuation

Our price target for NVDA of $315 is based on ~21x P/E on C27E earnings power. Our 21x P/E multiple is in-line with 1 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.