

SNIPER POINT CAPITAL



NVIDIA

Still Calling It a Bubble?

Executive Summary

NVIDIA is the toll layer of accelerated computing, priced as though it were a late-cycle semiconductor company. The distance between those two descriptions is the opportunity.

- **The moat is the ecosystem, not the chip.** Three decades of hardware, CUDA, roughly 3.8 million developers and more than 35,000 companies building on NVIDIA AI compound into a switching cost that no single-workload design replicates.
- **Custom silicon is real, and it is being commoditized.** Google is already broadening beyond Broadcom toward Marvell, MediaTek and potentially AMD, while NVIDIA has bought its way into the inference-silicon conversation through Groq's technology and team. The ASIC layer is converging toward competition among suppliers, not displacement of the platform.
- **The demand base is wider than the hyperscaler debate.** Hyperscalers were only about 55% of Data Center revenue last quarter. The market scales with the number of economically valuable problems that respond to more compute: credit and fraud, drug discovery, digital twins, autonomy, agents, not with the capex budgets of five companies.
- **The economics already look like software.** Roughly 75% gross margin, 62% net margin and ROIC near 150%, with the balance-sheet flexibility to allocate capital while customers spend theirs.
- **Valuation is the anomaly.** About 15x forward earnings, below Broadcom at 18.8x and AMD at 30.0x, a 6.7% earnings yield against roughly 5% for the S&P 500, some 1.6 standard deviations below its own five-year average multiple, and materially underowned by active managers at about 5% versus a 7.7-8% index weight.

The framework. A 30% earnings CAGR, a 0.5% dividend yield and a partial re-rating from 15x to 25x imply roughly 44% annualized over five years. That does not require NVIDIA to return to its historical multiple — only for today's discount to stop widening.

I. The Qualitative Thesis

A monopoly hidden in plain sight

NVIDIA began in 1993 with a focused vision: bring 3D graphics to gaming and multimedia. What looked like a graphics-card company eventually became one of the most important infrastructure companies in computing. In 1999, NVIDIA invented the GPU, creating a new computing architecture built for parallel processing rather than traditional serial CPU workloads. That shift proved foundational: the same architecture that made video games more realistic later became the engine for scientific computing, simulation, and artificial intelligence.

The company's real strategic inflection came in 2006 with CUDA, NVIDIA's software architecture that opened the GPU beyond graphics and allowed developers, researchers, and enterprises to use GPUs for general-purpose accelerated computing. CUDA transformed NVIDIA from a hardware vendor into a computing platform. **Today, more than 3.8 million developers build applications for accelerated computing, and over 35,000 companies use NVIDIA AI technologies.**

This history matters because NVIDIA's moat is not just semiconductor design. It is the accumulated ecosystem of **hardware, software, developers, and deployment infrastructure built over three decades.**

With generative computing, computing has evolved from a cyclical hardware market into a high-margin, mission-critical infrastructure layer for intelligence.

What was once a cyclical semiconductor business is evolving into something closer to a Buffett-style compounder: protected by **high barriers to entry, strengthened by pricing power, valued by customers, predictable cash flows and high ROIC.**

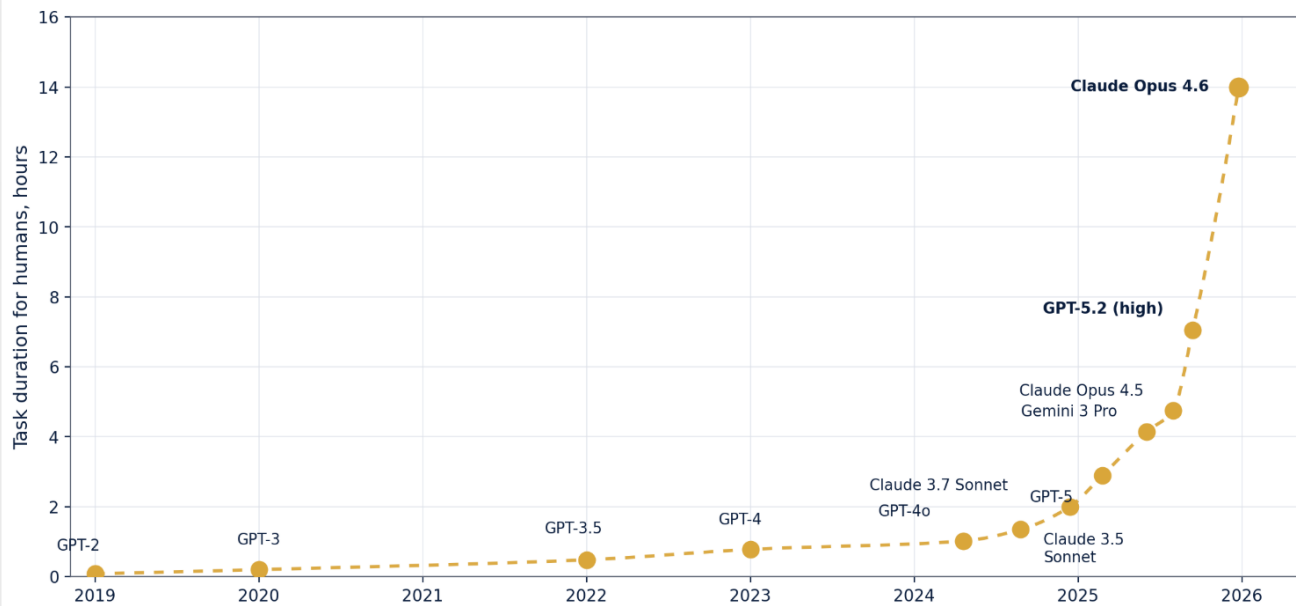
AI, software, agents, and robotics will scale only as fast as the hardware beneath them allows. The company that dominates this computing layer **becomes the tollbooth of intelligence ; the infrastructure provider through which future innovation must pass.**

NVIDIA's quality is amplified by **its customers'**. The hyperscalers are among **the most dominant businesses in the world,** and NVIDIA's GPUs are critical enough to their growth that NVIDIA increasingly looks **like a claim on their future free cash flow.**

NVIDIA sells hardware, but its economics look increasingly like software. As a fabless designer, the company generates roughly **75% gross margins and 62% net margins, above Visa at 51% and Mastercard at 45%.** These numbers alone hint at a sensational business model. The market's caveat is that semiconductors are cyclical; our argument is that NVIDIA's role in AI infrastructure is making that label increasingly outdated.

NVIDIA's edge is not only operational; it is allocative. **With ROIC around 150%, low internal capex needs, and unmatched visibility into AI bottlenecks,** NVIDIA increasingly behaves like **the asset manager of the AI economy** — placing capital across the layers that matter most: Corning for physical infrastructure, Intel for semiconductor supply, Synopsys for design and simulation, CoreWeave for cloud capacity, and OpenAI and Anthropic for frontier model demand.

Explosion of agentic capabilities

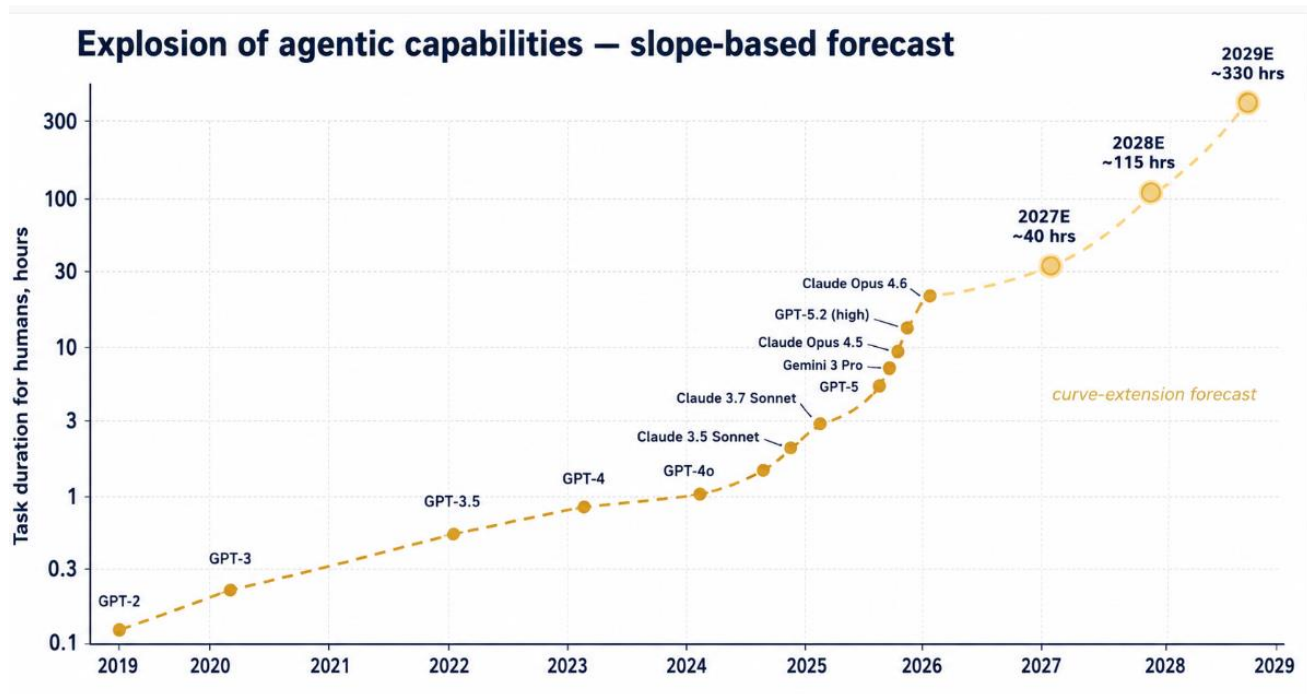


If AI agents become a new layer of digital labor, Nvidia is not merely selling hardware into a capex cycle. It is supplying the compute substrate **required for that labor layer to operate.**

The future is already starting to look agentic: agents will perform work, agents will coordinate sub-agents, and qualified humans will increasingly manage systems rather than execute every task directly. The marginal value of an AI agent relative to a human worker appears to be scaling non-linearly, and the recent jump in agentic capability suggests we may be near an inflection point.

Unlike human labor, there is no natural limit to how many agents a company can deploy. The real bottleneck becomes the compute required to run them. For AI application companies, **agents consume tokens**; for infrastructure providers, those tokens translate into **demand for accelerated computing**. If token costs continue falling at a rapid pace as hardware and software stacks improve, the economic case for deploying more agents becomes stronger, not weaker. Lower inference cost can unlock more usage, more workflows, and more total compute demand.

For the companies providing agentic services, agents cost tokens, and with the hardware of the likes of NVDA, based on their pace of iteration, token costs **will fall 90% annually**. To give context, Open AI and Anthropic are already profitable in a token basis and are likely to generate considerable operational cash flow in the coming years.



Extending the curve visually is not meant to be a precise forecast, but it highlights the key point: if agentic capability keeps compounding while inference costs keep falling, **compute estimates may be structurally too low**. The market may still be modeling AI demand as a tool-adoption cycle, when the better framework may be **the buildout of an intelligence grid**.

II. Competition

Competition kills profits and substitution eliminates your business. This concept is best expressed by Chris Hohn's (TCI) investment philosophy.

In my eyes, there are significant misconceptions in the semiconductor design field for that matter. The source of the misconception comes from the fact that Nvidia as a percent of Hyperscalers capex has fallen in the last year.

Let's reason from first principles. A capital investment is still an investment, not a guarantee. Even if the probability of success is high, concentrating everything in a single architecture is poor risk management. First, you expose yourself to obsolescence: a competitor can make a different technological bet that proves superior. Second, you become captive to one ecosystem, giving that supplier increasing control over your economics, roadmap, and strategic flexibility.

That said, hyperscalers operate at enough scale, and across enough overlapping workloads, that custom silicon can make a lot of sense. When a workload is large, repetitive, and well understood, designing hardware specifically around it can improve both performance and cost.

And that is exactly what we are seeing. Broadcom has had major success with Google's TPU program, OpenAI is moving into custom silicon with Jalapeño, and Amazon has scaled Trainium. Claude, arguably the surprise of the year, was also trained using significant amounts of TPUs.

However, that was only a point in time. Since then, NVIDIA has licensed Groq's technology and brought in much of the team behind it, including Jonathan Ross, **one of the creators of Google's original TPU**. NVIDIA has already turned that technology into its own LPU platform, **giving it a custom silicon offering alongside GPUs**.

Going back to the competition point, Google's strategy is telling. Despite Broadcom's success with TPUs, Google clearly **does not want to depend on a single custom silicon partner**. It has been broadening its supplier base across Marvell, MediaTek and potentially AMD, preserving competition and reducing dependence on Broadcom. In other words, **custom silicon is being commoditized**.

Even if custom silicon is competitive for a specific workload, it comes with an important tradeoff: **you lose optionality**.

With NVIDIA GPUs, that **capital has an alternative use**. There is a deep external market for compute, meaning excess capacity can potentially be leased rather than sitting underutilized. Meta is already considering exactly this. Mark Zuckerberg has said that external offers for its compute are so attractive that, **in some cases, it may make more sense to rent the capacity out than use it internally**.

Fungibility is an asset. NVIDIA GPUs are available across every major cloud and support one of the broadest ecosystems of open models. Their general purpose nature **allows NVIDIA to be far more ambitious**, serving not only hyperscalers but also thousands of smaller companies that need AI compute and will never have the scale, capital, or workload consistency to build custom silicon of their own.

You can already see that diversification in the numbers. In NVIDIA's latest quarter, **hyperscalers accounted for only about 55% of Data Center revenue**, while the remaining 45% came from AI clouds, industrial and enterprise customers. The opportunity is no longer limited to a handful of companies.

The more important way to think about the addressable market is **by asking how much of the economy is becoming compute intensive**. As AI moves from model training into inference, software, robotics, drug discovery, autonomous systems and everyday enterprise workflows, **compute becomes an input into a much larger share of economic activity**.

A bank can use more compute to underwrite credit and detect fraud. A pharmaceutical company can use it to screen molecules and design proteins. An industrial company can use it for digital twins, predictive maintenance and factory automation. An automaker can use it for autonomous driving and simulation. A software company can deploy agents against tasks that previously required human labor. Governments can use it for defense, weather modeling and scientific research.

The implication is that the **market grows with the number of economically valuable problems that can be attacked with more intelligence**. Ultimately, demand is driven by every industry that can convert additional compute into greater productivity, lower costs or entirely new products.

Lastly, NVIDIA has strong governance. Jensen Huang has instilled a constant fear of falling behind, keeping the company aggressive even from a position of strength.

III. Valuation

Price is what you pay and value is what you get. At its core, a stock is the present value of its future cash flows. The investment process therefore comes down to understanding what expectations are already embedded in the price and whether the return implied by those expectations is attractive relative to the alternatives.

Mechanically, a DCF can be used to estimate fair value based on current assumptions, while a reverse DCF works backward from the market price to show what assumptions are already embedded in it. (both analyses are available separately and can be provided upon request).

Some models are useful, but none are perfect, **and they often fail to capture how long a truly exceptional business can sustain superior economics.** NVIDIA is a good example: earnings expectations have been revised materially higher year after year as the duration of its growth repeatedly exceeded what was originally modeled.



Source: Bloomberg



This chart shows the other side of the equation: **while NVIDIA's earnings expectations have moved materially higher, the multiple investors are willing to pay for those earnings has collapsed**. NVIDIA now trades at roughly 17.7x blended forward earnings, versus a five year average of 35.1x, **putting the stock about 1.6 standard deviations below its historical mean** and close to the bottom of its five year range.

To avoid anchoring the analysis to a single price target, I prefer a more conceptual framework. Decomposing the expected rate of return is a useful way to think about the return a business can generate from today's valuation without relying on a precise estimate of fair value.

$$\text{Expected 5Y Annual Return} \approx (1 + g_{EPS}) \left(\frac{P/E_{Exit}}{P/E_{Entry}} \right)^{1/5} (1 + y_{div}) - 1$$

Over a five-year horizon, assuming a 30% earnings CAGR, a 0.5% dividend yield, and a conservative rerating from 15x to 25x earnings, the framework implies an expected annual return of roughly **44%**. Importantly, that does not require NVIDIA to return to its historical median multiple, only for the current discount to partially normalize.

Compared with other major semiconductor designers, NVIDIA trades at just **15x forward earnings**, below Broadcom at **18.8x** and AMD at **30.0x**, despite offering higher expected growth and, in my view, greater visibility into that growth over the coming years.

Benchmarking the valuation, even at a nominal forward P/E of 15x, NVIDIA offers an earnings yield of roughly **6.7%, versus about 5%** for the S&P 500 at 20x. Investors are therefore receiving a higher current earnings yield from a **business expected to grow materially faster than the broader market**.

Considering the index, NVIDIA is both its largest weight and one of the biggest contributors to expected earnings growth over the coming years. Yet, according to Morgan Stanley, active asset managers hold only about **5%** of their portfolios in NVIDIA versus roughly **7.7%–8%** in the S&P 500, leaving the stock materially underowned relative to its benchmark weight.

With that in mind, **underweighting NVIDIA becomes increasingly difficult to justify from a tracking error perspective**. At the portfolio level, adding the stock can simultaneously **lower the forward multiple relative to the index and increase expected earnings growth**, an unusual combination for the benchmark's largest constituent.

Valuation wise, the setup reminds me of Apple in 2017. Apple was trading at roughly 15x forward earnings, yet active managers were still materially underweighting the stock. Then, UBS found that global fund managers had a 1.3% investor weight in Apple but an active weight of negative 0.6% relative to their benchmarks.

From a capital allocation perspective, NVIDIA is uniquely positioned. While many of its customers are consuming most of their free cash flow to fund the AI buildout, NVIDIA retains substantial financial flexibility. That optionality gives the company the ability to move quickly when attractive opportunities emerge, whether through strategic investments, partnerships or acquisitions. At the same time, more than 50% of free cash flow is expected to be returned through share repurchases, which at current valuation levels should be meaningfully accretive to shareholders.

IV. Risks

The thesis is asymmetric, but it is not free. The risks that matter are the ones that attack the argument above: the durability of the moat, the breadth of the demand base, and the re-rating the return math leans on.

- **Custom silicon displacement (15-20% probability, -25-40% impact).** Hyperscaler ASIC programs — Google's TPU line with Broadcom, Amazon's Trainium, OpenAI's move into its own silicon — scale faster than expected and capture the highest-volume workloads inside their owners' data centers, capping NVIDIA's share of the incremental capex dollar. **Mitigant:** The same forces are commoditizing that layer: Google is already spreading work across Marvell, MediaTek and potentially AMD precisely so it does not trade dependence on NVIDIA for dependence on Broadcom. NVIDIA has also closed the inference gap by licensing Groq's technology and hiring much of the team behind it, giving it a purpose-built option alongside the GPU line. And a fixed-function part costs the buyer optionality — it cannot be re-pointed at a new workload or leased into the external compute market.
- **The re-rating never arrives (20-25% probability, -25-35% versus the base case).** The expected-return framework assumes the multiple normalizes from roughly 15x toward 25x over five years. If the market keeps underwriting NVIDIA as a late-cycle semiconductor name rather than an infrastructure compounder, the single largest contributor to the modeled return simply does not show up — and a multiple that compresses further would offset several years of earnings growth. **Mitigant:** The re-rating is the upside, not the thesis. With no multiple change at all, a 30% earnings CAGR from a 6.7% starting earnings yield still compounds faster than an index yielding roughly 5%.
- **Hyperscaler capex digestion (10-15% probability, -20-30%).** If inference monetization lags the buildout, the committees funding it slow down and orders push to the right regardless of NVIDIA's competitive position. A capex pause is the one risk that does not require anyone to out-engineer the company. **Mitigant:** The customer base is already broader than the debate assumes. Hyperscalers were about 55% of Data Center revenue in the latest quarter; the other 45% came from AI clouds, enterprise and industrial buyers whose spending is tied to their own productivity rather than to a handful of budgets.
- **Token deflation without volume elasticity (10-15% probability, -15-25%).** The agentic case assumes cost per token falls faster than demand for tokens is satisfied. If capability plateaus, or if price declines are passed through faster than usage scales, the installed base absorbs more of the workload and the replacement cycle stretches. Falling token costs are only bullish while they unlock new workloads. **Mitigant:** Every step down in inference cost so far has widened the set of tasks worth running rather than shrinking the bill, and the binding constraint on agent deployment is compute, not appetite.
- **Circularity between demand and the investment book (~10% probability, -15-25%).** NVIDIA's allocative advantage — positions across Corning, Intel, Synopsys, CoreWeave, OpenAI and Anthropic — points at counterparties that are also customers. Some portion of demand is, directly or indirectly, financed by the supplier. A repricing of the funding market for frontier labs and neoclouds would hit revenue and carrying values at the same time. **Mitigant:** These are minority positions funded out of free cash flow rather than leverage, alongside a buyback absorbing more than half of it. The exposure is to the return on that capital, not to the solvency of the company making the investment.
- **Supply chain concentration — foundry, HBM and advanced packaging (10-15% probability, -10-15%).** Fabless economics rest on one leading-edge foundry and a narrow memory and packaging base. Rising HBM pricing or

packaging constraints either compress the ~75% gross margin or cap shipments no matter what demand looks like. **Mitigant:** Long-dated supply commitments and an annual architecture cadence keep NVIDIA first in line for capacity, and the Intel position is in part an option on widening that base.

- **Policy, concentration and key-person risk (5-10% probability, two-sided).** Data Center is the overwhelming majority of revenue, so a single regulatory or competitive outcome at that layer moves the entire P&L. Export-control policy remains the clearest example: further restriction is largely headline risk against numbers that already assume very little, while a reopening is a call option nobody is paying for. Governance is a genuine strength — Jensen Huang has institutionalized a fear of falling behind that keeps the company aggressive from a position of strength — but a strength concentrated in one person is also a dependency. **Mitigant:** This is the structural cost of monopoly economics, not an argument against owning them. It is the reason the position is sized as a conviction holding rather than an unbounded one.

V. Conclusion

Investors continue to price NVIDIA as a story about the future, when much of that future is already the present. A “bubble” trading at a discount to the S&P 500. A “bubble” in earnings, with sold out demand and unusually strong visibility. A company supposedly losing share to competitors, while those same competitors are increasingly competing with one another for that share.

At the end of the day, the thesis is simple: a great business is being offered at an attractive valuation. NVIDIA combines a deep hardware and software moat, extraordinary economics, high returns on capital and a widening role in the infrastructure of AI. Custom silicon is real, but it does not eliminate NVIDIA’s advantage in flexibility, ecosystem breadth and the ability to serve demand far beyond the hyperscalers. At the same time, earnings expectations continue to move higher while the valuation has compressed, leaving the stock cheaper than both its own history and, on key measures, the broader market. The opportunity does not depend on heroic assumptions, only on continued execution, sustained earnings growth and a partial normalization in how the market values that growth. Narratives are permanent, opportunities are not, and fortunately this one will not last forever.

VI. Important Disclosures

Nature of this memo. Nothing in this memo constitutes investment, legal, tax, or accounting advice. Recipients should perform their own independent analysis and consult appropriate professional advisors before making any investment decision.

Position disclosure. Sniper Point Capital, its principals, employees, and/or affiliated entities may hold long, short, or derivative positions in NVIDIA Corporation (NASDAQ: NVDA) and in other securities discussed herein. Positions may change at any time without notice. The author has a personal long position in the subject security as of the date of this memo.

Sources and methodology. Information in this memo is derived from publicly available sources, including company SEC filings (10-K, 10-Q, 8-K), earnings call transcripts, third-party sell-side research, and public interviews and presentations given by NVIDIA management. Charts are sourced as labeled. The discounted cash flow and reverse discounted cash flow analyses referenced in Section III were produced by the Sniper Point Capital in-house valuation platform and are available separately upon request. While sources are believed to be reliable, no representation or warranty, express or implied, is made as to the accuracy, completeness, or timeliness of the information contained herein.

Forward-looking statements. This memo contains forward-looking statements, estimates, projections, and opinions based on assumptions that may prove materially incorrect. Actual results — including AI infrastructure demand, hyperscaler capex, regulatory outcomes (US/China export controls), competitive dynamics, software adoption, and capital allocation — may differ significantly from those expressed or implied. Past performance is not indicative of future results, future returns are not guaranteed, and a partial or total loss of capital may occur.

Risk warning. Concentrated single-name exposure carries materially higher risk than a diversified portfolio. The thesis depends on continued growth in AI infrastructure spending, on NVIDIA sustaining its pricing power and ecosystem advantage, and on at least a partial normalization of the earnings multiple. Structural impairment scenarios — meaningful competitive displacement by custom silicon, a sustained slowdown in hyperscaler and enterprise capex, supply-chain or export-control shocks, or a multiple that compresses further rather than normalizing — could result in equity-value impairment of 25-40% or more. These outcomes are not the base case but cannot be ruled out.

Valuation caveats. The valuation framework presented in Section III is sensitive to inputs including consensus EPS, applied P/E multiple, and revenue growth assumptions. Small changes in any of these inputs can produce materially different equity values. The Base Case relies on consensus growth estimates which may not be realized. The reverse-DCF gap analysis depends on current market price and is itself a moving target.

Confidentiality. This memo and its contents are proprietary. Reproduction, redistribution, or disclosure to third parties — in whole or in part, by any means — is prohibited without the prior written consent of Sniper Point Capital.

