The AI Stack: Who Will Capture the Profits?

Introduction

The AI trade is no longer behaving as one single boom, as the debate over which parts of the industry will actually capture the profits has intensified among investors. After two years in which artificial intelligence pushed valuations higher across semiconductors, cloud platforms, model labs, and enterprise software, the market has become more selective. The recent "SaaSpocalypse" narrative captured this shift. Investors are no longer questioning AI adoption, but rather whether companies will capture durable value from AI development instead of simply absorbing higher costs while the profit pool migrates elsewhere or, even worse, be substituted by cheaper AI models in their offerings.

Within this paradigm shift, ServiceNow [NYSE: NOW] serves as a clear example. The company reported rising demand for its AI-based Now Assist suite and significant growth in AI-related annual contract value, yet its share price has sharply decreased over the past year. This reflects a broader underlying uncertainty around the AI stack: if models become cheaper, compute normalizes, and general-purpose agents become more capable, which layer retains pricing power?

In this article we argue that the AI economy should not be understood as one undifferentiated race. Compute, models, distribution, and applications each have different cost structures, switching costs, and competitive dynamics. The central question is therefore not whether AI creates value, but where that value might settle once the initial phase of hype, capital expenditure, and model competition normalizes.

The AI Stack

When people think about artificial intelligence, they immediately jump to companies like OpenAI, Anthropic, or maybe some major tech players that are trying to capture some of the value and integrate AI into their incumbent tech platforms. OpenAI's ChatGPT has become a household name, now being used as a verb for searching using AI, similar to how Google has become synonymous with search engines. These companies, however, represent just a portion of the AI stack, or value chain, that goes into bringing artificial intelligence solutions to everyday consumers. This stack consists of infrastructure, models and R&D, and applications at its most basic level. Each of these levels can be divided into ever more complex subcategories such as splitting infrastructure into hardware, software, and more.

The infrastructure of the AI stack consists of data centres, data storage, connectivity, chips, servers, energy generation and storage, and more. This layer begins with silicon, the physical chips that offer the immense amount of compute required to complete the complex tasks and queries that AI offers. The most significant player, Nvidia [NASDAQ: NVDA], has seen its market cap grow from $502bn in 2023 to $4.56tn in 2026, cementing itself among Google [NASDAQ: GOOGL], Apple [NASDAQ: AAPL], and Microsoft [NASDAQ: MSFT] as one of the most valuable companies in the world. Nvidia builds GPUs which train and run AI models, systems to build AI clusters, software stacks that run enterprise AI, as well as gaming and automotive chips. However, while Nvidia designs and builds these crucial chips, the production is done in Taiwan at TSMC [NYSE: TSM] fabrication facilities. TSMC holds a dominant position with about 90% market share of the most cutting-edge, 2nm chip manufacturing which allows Nvidia's chips to offer the power they do. TSMC is currently valued at $2.14tn.

A few steps up from this crucial process, but within the infrastructure portion of the AI stack, comes data centres and cloud computing. Data centres are a critical part of the AI stack as they are responsible for storing, processing, and distributing the large amounts of data that are required for modern queries and AI output. Power availability and data centre buildout have become the largest bottleneck within the AI industry, significantly hindered by regulatory and public pushback. Companies struggle to generate enough sustainable power to meet consumer demands and government mandates, while data centres are facing pushback from the public due to fears that they will drive the price of power up even higher. In March 2026, President Trump brought forward a directive aimed at having major tech firms shoulder the higher cost of electricity used in AI expansion. OpenAI, xAI, Oracle [NASDAQ: ORCL], Google, Meta [NASDAQ: META], and more have signed the voluntary directive, agreeing to build, buy, or bring their own electricity generation to satisfy data centre usage rather than pulling from the public grid. Data centres and power generation facilities also require storage and cooling systems, creating entire industries that capture value along the AI stack. In addition to this, AWS, CoreWeave, and others provide cloud computing to AI consumer-facing companies such as Anthropic and OpenAI.

In the next level of the AI stack comes the more popular names in the industry: OpenAI, Anthropic, xAI, Google's Gemini, etc. These companies have primarily been focused on the models for public and commercial usage, however in late 2025 and early 2026, this changed. Big tech companies such as Google and Amazon [NASDAQ: AMZN] are spending hundreds of billions in capex to expand and capture larger portions of the AI stack. In 2026, the top five hyperscalers are projected to spend a combined $602bn with the majority of that directed towards infrastructure. For these mega companies, it is about owning the structure behind AI and outspending market entrants to prevent any more dilution than that already caused by OpenAI and Anthropic. In the private markets, frontier labs face an interesting situation: they are both customers and competitors with these incumbent tech firms.

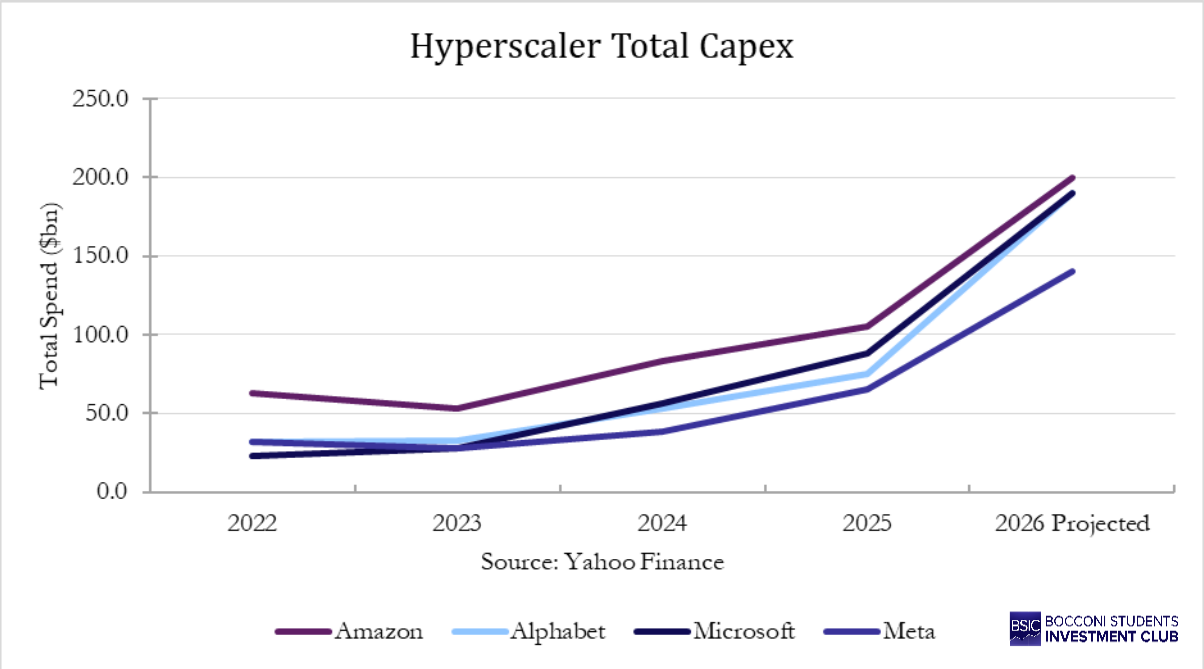
Above the model layer sits the fastest-growing and most diverse segment of the AI stack: use-case-specific applications and, ultimately, the consumer. This layer is where much of the visible commercial adoption is currently taking place. Vertical AI companies are building domain-specific models and interfaces that sit on top of foundation models, tailoring general-purpose intelligence to particular industries. In legal tech, companies like Harvey are using LLMs to automate contract review and due diligence. In healthcare, firms such as Abridge are deploying AI for clinical documentation, reducing the administrative burden on physicians. Software development is seeing perhaps the most dramatic adoption, with coding assistants Cursor (SpaceX has reportedly entered into a strategic agreement involving a potential acquisition of Cursor at a $60bn valuation), GitHub Copilot, and Anthropic's own Claude Code and Claude Cowork, now embedded in the daily workflows of millions of white-collar workers. These vertical players occupy a structurally attractive position: they build proprietary workflows and data flywheels on top of commoditising foundation models, meaning their moat compounds even as the underlying model cost falls. At the consumer end, the scale of adoption is remarkable. ChatGPT reached 100m users faster than any product in history, and AI assistants are increasingly becoming the default interface for search, writing, and decision-making for everyday users. The monetisation models at this layer range from subscription-based consumer products to enterprise SaaS contracts, with the latter proving significantly more defensible. Currently, enterprises are willing to pay much higher prices for AI workflow and the majority of the value capture sits at the commercial level.

Economics by Layer

At the infrastructure layer, the economics are clearest and, for now, the most favourable. Nvidia is the most visible beneficiary: data centre revenue now comprises approximately 90% of Nvidia's total sales, with the company generating operating margins of around 52% and gross margins that have sustained above 70%, figures that place

it comfortably ahead of every other large-cap technology company and supporting its $4.56tn valuation. The dynamic is simple: Nvidia designs the chips, TSMC manufactures them, and the entire industry is structurally dependent on both. Investment here is enormous and front-loaded, however both Nvidia and TSMC are now reaping the rewards, showcased by TSMC's free cash flow exploding from $292bn in 2023 to $1tn in 2025. Nvidia's FCF has grown from $27bn to $96bn, more than 3x, over the same period. Currently, demand could not be any higher and the bottleneck sits at the supply of compute; companies at this stage of the stack are capturing as much value as possible.

The cloud and data centre layer tells a more nuanced story. The AI hyperscalers have combined commitment to more than $700bn primarily in chips and infrastructure, seriously reducing FCFs to some of the lowest levels since 2014 for the major tech companies. Amazon's FCF declined 95% as the company pushes for gains in AWS. Yet, the margins on the cloud businesses themselves remain healthy. AWS generated $10.2bn in operating income on $30.9bn in net sales in Q2 2025, while Microsoft's Intelligent Cloud division produced $12.1bn in operating income, the highest of the three cloud giants. AI inference services are priced at a significant premium to commodity compute, making AI workloads margin-enhancing rather than margin-dilutive for the cloud platforms that host them. The capex spend demonstrates the big tech names buying themselves future margin rather than compressing today's, each dollar of infrastructure build-out locks in a customer relationship and a recurring revenue stream that compounds over time. These revenue streams also help block out new entrants that may dilute the future pot. With consumers always demanding more compute, there is more value for the hyperscalers to capture earlier in the stack than the level of the models.



The frontier model layer showcases the largest divergence between revenue trajectory and economic reality. Anthropic's annualised run-rate revenue reached $30bn in April 2026, up from $9bn at year-end 2025 and just $1bn fifteen months earlier. OpenAI sits at roughly $24bn on the same measure. These are extraordinary top-line numbers. The costs, however, are massive. Both companies burnt more than $5bn each in 2024, and the burn has continued to scale alongside revenue. Anthropic's gross margin improved significantly from deeply negative territory in 2024 to an estimated 40% in 2025, but still well below the 70–80% that characterises mature software businesses. The culprit is compute: training frontier models and running inference at scale against tens of thousands

of enterprise customers generates GPU bills that eat into every dollar of revenue before the business has a chance to leverage its operating model.

The strategic paths diverge sharply here, as they reflect fundamentally different theories of value creation. OpenAI projects operating losses of roughly $74bn in 2028, representing approximately 75% of that year's revenue, before pivoting to meaningful profitability around 2030. Anthropic, by contrast, forecasts reducing its cash burn to around one-third of revenue in 2026 and to 9% by 2027, with a break-even expected significantly sooner. The difference is strategic, not accidental. OpenAI is focused on consumer, aiming to scale and win in terms of model capability aggressively enough that the eventual market share and moat justify the losses in the meantime. Anthropic is making a revenue density bet, concentrating on enterprise contracts, where eight of the Fortune 10 are now paying customers and over 500 companies spend more than $1m annually, on the theory that durable, high value relationships compound faster than consumer acquisition costs. Both could prove correct, but for the time being, it seems that there is more willingness to spend on the enterprise side of the market.

At the application and vertical AI layer, the economics are the most contested. The revenue growth is real and often enormous; early vertical AI companies are growing at approximately 400% year-over-year while maintaining gross margins of around 65%, but the margin picture depends significantly on how deeply a company owns its workflow relative to the foundation models it sits on top of. The structural challenge is that AI-first application companies carry a cost that traditional SaaS never did: every query runs through an API that bills by the token. Traditional SaaS companies with 85% gross margins are adjusting to 60–70% as AI features are incorporated, while early-stage AI-native companies average around 25% gross margins, sometimes negative during usage surges. The path to SaaS-like economics requires either building proprietary model infrastructure, which is expensive and capital-intensive, or developing deep enough workflow integration that the company controls which model handles which task, routing cheaper models where quality tolerance allows. There is, however, a threat that the development and costs put into these practical use-case models will be for naught and they will be pushed out by the major players and powerful AGI. Anthropic has witnessed its ARR growth at astonishing rates over the last few years from $381m in 2024 to $9bn in 2025 up to $30bn in April of 2026. This demonstrates the fastest growth in ARR of a single company of all time. There is no guarantee this level of growth will remain sustainable; however, it demonstrates a serious risk for smaller models which will not benefit from the efficiency of scale that Anthropic will.

Strategy and control in the AI stack

Having mapped the economics of each layer, the next question is which companies control the bottlenecks that determine where AI profits ultimately settle. It is therefore more useful to view the AI race as a set of parallel contests over distinct control points: compute, models, distribution, and customer workflow. Each sits at a different layer of the stack, creates a different kind of bargaining power, and locks customers in through different mechanisms. Compute earns infrastructure rents without needing to own the end user, while models can produce genuine technical value and still fail to retain it if they are accessed through a rival's cloud or surfaced through someone else's interface.

Distribution and customer workflow are especially important because they determine how AI reaches the user. Search, browsers, operating systems, developer environments, productivity suites, collaboration tools, and device ecosystems all belong here. This layer often carries stronger switching costs than the model layer because its lock-in is behavioural and organizational rather than purely technical. In AI, the owner of distribution decides which model is surfaced, how it is bundled, and which monetisation logic remains attached to the user interaction. That

is why Search, Microsoft 365, Windows, iOS, Android, Chrome, and Slack matter strategically even when the headlines focus on models.

Customer control is narrower than distribution, and in enterprise software it is often more defensible. In AI it means owning the workflow, permissions, records, process logic, and interface through which work is authorized and executed. An agent that drafts an answer is one thing. An agent that can approve a procurement request, reroute a service issue, or update a regulated customer record is something else. That second function requires context, authority, and auditability. It sits much closer to enterprise lock-in than raw model capability does. This is where firms like ServiceNow and Salesforce matter: their control point is not the frontier model, but the governed environment in which model output becomes enterprise action.

These positions can reinforce one another, but they should not be collapsed into one contest. Upstream power over compute does not automatically translate into downstream monetisation. Model quality does not guarantee control over the customer interface. Distribution does not mean deep workflow ownership. Customer ownership in the enterprise does not mean possession of scarce infrastructure.

Big Tech is not competing for one single control point

Microsoft is pursuing the broadest form of control in this group because it combines compute with enterprise distribution. Azure gives it the infrastructure layer, with global scale that the company itself presents as larger than any other cloud provider. Azure AI Foundry extends that position into model access, offering a large catalogue of third-party and first-party models through a common enterprise interface. Then the company routes demand into Microsoft 365, GitHub, Teams, and Dynamics, which means the intelligence layer can be monetised both as cloud consumption and as software upsell. Management said in January 2026 that Microsoft 365 Copilot had reached 15m paid seats. The strategic logic is not that Microsoft must own the single best model at every moment. It is that it can make AI spend flow through a stack it already controls.

Google controls a different pair of assets. It has model strength through Gemini and infrastructure depth through its custom AI chips, but its central strategic problem is distribution. Search and Other advertising revenue reached $63.1bn in Q4, and Google Services remains the economic core that AI has to protect rather than disrupt. The same earnings call that highlighted Gemini's scale also highlighted 325m paid subscriptions across consumer services, more than 8m paid Gemini Enterprise seats, 750m monthly active users for the Gemini app, and a sharp rise in cloud backlog. Google is therefore not just trying to win the model race. It is trying to make AI deepen engagement across Search, Android, Chrome, Workspace, and YouTube without undermining the economics of the access points that already exist. That is different from Microsoft's logic. Microsoft is using AI to extend an enterprise suite and a cloud. Google is using AI to defend a distribution system whose largest profit pool predates generative AI.

Amazon's position is cleaner. Its strongest control point is compute, and its AI strategy is structured around monetising demand for infrastructure rather than trying to define the default interface through which most users experience AI. AWS generated $128.7bn of revenue and $45.6bn of operating income in 2025. Bedrock is used by more than 100,000 companies, while Trainium and Graviton together have crossed a $10bn annual revenue run rate. The important detail is not only scale. It is the design of the offering. Amazon explicitly emphasizes that Bedrock lets customers test and switch models without rewriting code. That is a strategic choice. It positions AWS as the infrastructure broker and access layer in a market where customers may not want to commit to one model family. Amazon is therefore competing for AI economics without trying to own the primary user interface in the way that Microsoft, Google, or Apple do.

Apple, by contrast, is almost entirely about distribution and interface ownership. Its installed base now exceeds 2.5bn active devices, and Apple Intelligence is framed around preserving trust, privacy, and seamlessness inside that ecosystem. The architecture itself makes the strategy visible. Apple pushes some intelligence on device, uses Private Cloud Compute for more demanding tasks, and integrates ChatGPT when external model breadth improves the user experience. That is not a sign that Apple lacks a strategy. It is the strategy. Apple is defending the operating system, the device, and the customer touchpoint. It does not need to become the neutral infrastructure provider for third parties, and it does not need to turn model access into a large standalone business. It needs to keep AI inside the experience users already associate with Apple hardware and services. With John Ternus set to become CEO in September 2026, it will be interesting to see whether Apple pivots its strategy regarding AI software and hardware within their existing ecosystem.

Taken together, these companies are targeting different bottlenecks. Microsoft wants full-stack capture, with particular strength in compute and enterprise distribution. Google wants to preserve distribution while reinforcing it with model leadership. Amazon wants to collect infrastructure rents and make model choice run through AWS. Apple wants to retain interface ownership and keep the device ecosystem central to AI use. None of those positions is identical, and that is why comparing firms only on model quality produces a distorted picture of strategic control.

Why enterprise software still matters

The enterprise software layer matters because the core of AI monetisation is not a raw answer, but the authorized actions inside a company's workflows. That is why companies like ServiceNow and Salesforce matter. Their control lies in workflow, permissions, records, metadata, governance, and the institutional memory of how work is actually done. That is why they should not be analysed as if they were simply weaker versions of the hyperscalers or failed attempts at foundation-model ownership. Their strategic asset is different. They are not trying to own the scarcest compute or trying to define the broad consumer interface. They are trying to own the governed enterprise context in which AI becomes operational.

ServiceNow has become explicit on this point: It wants to be the "AI control tower" for companies, connecting across different clouds, AI models, and data sources, reporting more than 95 billion workflows running on the platform each year. In the first quarter of 2026 it ended with 630 customers spending more than $5m per year on its platform, while the number of customers spending more than $1m per year on Now Assist grew by over 130%. Those metrics matter less as proof of short-term revenue explosion than as evidence of where the company wants to sit. ServiceNow is trying to own the orchestration layer, the audit layer, and the governance layer across model choices. Its recent partnerships reinforce that logic. OpenAI and Anthropic provide intelligence; ServiceNow keeps the workflow context. Microsoft provides another major enterprise surface; ServiceNow wants AI Control Tower to govern that sprawl as well.

Salesforce's position is similar in structure but different in enterprise emphasis. Its control point is the customer side of the workflow: customer records, service history, pipeline data, business metadata, collaboration flows, and the logic that ties those elements together across Customer 360. The company said in February 2026 that Agentforce ARR had reached $800m, that it had closed 29,000 Agentforce deals, and that more than 60% of Agentforce and Data 360 bookings in the quarter came from expansion within the existing base. The product language around Agentforce 360 makes the broader aim clearer. Salesforce is trying to connect agents, apps, data, metadata, and Slack in one trusted system, then strengthen that system with Informatica's cataloguing, governance, and master-data capabilities. That is a control strategy built around context and actionability. It assumes that customer ownership in AI will depend on who holds the cleanest data, the richest metadata, and the most direct tie between agent output and commercial workflow.

The point is not that enterprise software is automatically safe. The market has made that clear enough. ServiceNow's shares fell after both the January and April 2026 results despite strong operating numbers and visible AI traction, in part because investors remain unsure whether application vendors will keep the economics or merely carry the cost of integrating someone else's intelligence. That scepticism is rational. Still, it does not erase the strategic distinction. If AI agents move deeper into enterprise execution, the firms that control workflow, permissions, and context are not peripheral to monetisation. They are where monetisation becomes governable.

The shift in AI value creation

The first phase of the AI boom was a scarcity trade. GPUs were scarce, frontier models were scarce, and the market rewarded companies closest to the bottleneck. NVIDIA remains the clearest example, with FY2026 revenue reaching $215.9bn, up 65% YoY, GAAP net income of $120.1bn, and gross margin of 71.1%. In the near term, compute scarcity remains a real profit pool.

The strategic direction, however, is changing. Model performance is converging, API prices are compressing, open-weight models are improving, and hyperscalers are building their own chips. As the bottom of the AI stack becomes more competitive, value migrates upward to the layers that own workflows, data, customer relationships, and switching costs.

Model commoditisation: the model is becoming a component

The bear case for large language model economics lies in too many models improving at the same time. Stanford's 2025 AI Index shows how quickly performance gaps are narrowing. The gap between the leading closed-weight model and the leading open-weight model on Chatbot Arena fell from 8.04 percentage points in January 2024 to 1.70 percentage points by February 2025. The gap between the top ranked and 10th ranked models also narrowed from 11.9% to 5.4%, while the gap between the top two models shrank from 4.9% to 0.7%. On SWE-bench, a benchmark for software engineering tasks, AI systems improved from solving 4.4% of problems in 2023 to 71.7% in 2024.

Frontier intelligence still matters, especially for complex reasoning and agentic tasks. But the investment point is different: model leadership is becoming less durable as a standalone moat. If multiple models can satisfy a given enterprise use case, customers can route workloads across providers based on price, latency, accuracy, security, and contractual terms. The model becomes less like a proprietary product and more like a high-performance input.

Model	Input price, $/1m tokens	Cached input, $/1m tokens	Output price, $/1m tokens
GPT-4	30.0	n.a.	60.0
GPT-4 Turbo	10.0	n.a.	30.0
GPT-4.1	2.0	0.5	8.0
GPT-4.5	75.0	37.5	150.0
GPT-5	1.3	0.1	10.0
GPT-5.1	1.3	0.1	10.0
GPT-5.4	2.5	0.3	15.0
GPT-5.5	5.0	0.5	30.0

Source: BSIC, OpenAI

Pricing reinforces the same conclusion. Legacy GPT-4 API pricing was $30/$60 per million input/output tokens. GPT-4 Turbo reset that to $10/$30, GPT-4.1 is priced at $2/$8, while GPT-5 and GPT-5.1 are listed at $1.25/$10.

The latest official OpenAI pricing table also lists GPT-5.2 at $1.75/$14, GPT-5 mini at $0.25/$2, and GPT-5 nano at $0.05/$0.40 per million input/output tokens.

The important signal is not that prices are on the decline. The signal is that the market is becoming a segmented price performance curve. Enterprises can choose between frontier, mini, nano, open weight, proprietary, and task specific models. That increases customer choice and weakens the ability of any single model provider to sustain monopoly-like pricing.

For enterprise software companies, this is strategically useful. Focusing on ServiceNow's case, the latest message is already explicit: its AI Platform is designed to integrate with any model, any cloud, any data source. In other words, ServiceNow wants the model layer to be modular while its own platform controls orchestration, governance, and workflow execution.

NVIDIA wins now, hyperscalers attack the cost curve

Compute remains the strongest near-term profit pool because it is physically constrained. Data centres, power, networking, advanced packaging, and GPUs cannot be scaled overnight. NVIDIA's FY2026 numbers confirm that infrastructure scarcity is still monetising at exceptional margins.

However, the largest buyers of GPUs are also the firms most incentivized to reduce dependence on them. Microsoft, Amazon, and Google are scaling AI infrastructure while also internalizing more of the stack, with the objective of controlling their own cost curve. Google's TPU program and AWS Trainium reduce long-term reliance on external GPUs. This does not eliminate NVIDIA's advantage, but it does create a ceiling on how much value one supplier can capture indefinitely.

As noted earlier, AI infrastructure creates enormous capex requirements. The financing winners are companies with balance sheets large enough to fund data centres, chips, power procurement, and long-dated capacity commitments. Smaller infrastructure players may need project finance, strategic capital, or partnerships with hyperscalers.

For tech M&A, the infrastructure angle is therefore less about acquiring generic exposure to AI and more about buying control of constrained parts of the AI supply chain. Strategic buyers and sponsors will likely focus on assets with capacity that is hard to replicate, such as colocation and hyperscale data centres, power interconnection rights, GPU clusters, high-bandwidth networking, liquid cooling, memory and advanced packaging, and cloud security infrastructure.

How the application layer can retain value

Traditionally SaaS firms sold seats. Agentic AI can sell completed work. That creates a much larger monetisation opportunity if the software vendor owns the workflow. A model API sells inference whereas the application platform sells a resolved ticket, completed renewal, qualified lead, automated compliance check, or closed service request.

This is why switching costs matter. Enterprises do not want an AI model acting randomly across systems. They need identity, permissions, audit trails, governance, data context, and integration with existing workflows. Those assets sit in application platforms, not in the model itself. Hence software companies that successfully move from seat based pricing to usage based or outcome based pricing can expand TAM and defend premium multiples.

ServiceNow

ServiceNow is a clear example of value migration upward. It does not need to own the frontier model. Its advantage is that it already controls enterprise workflows across IT, HR, security, risk, and customer service.

In Q1 2026, ServiceNow reported $3.77bn of total revenue, up 22% YoY, subscription revenue of $3.67bn, current RPO of $12.64bn, and total RPO of $27.7bn. Importantly, customers spending more than $1m in *Now Assist* ACV grew over 130% YoY. The company also ended the quarter with 630 customers above $5m in ACV, up approximately 22% YoY.

ServiceNow now sells AI inside workflows customers already depend on. The company describes itself as an "AI control tower," integrating with any model, cloud, interface, data, and system.

Salesforce

Salesforce demonstrates the same logic in customer-facing workflows. Its advantage is ownership of CRM data, sales processes, service workflows, Slack, Tableau, MuleSoft, Data 360, and industry clouds.

In FY2026, Salesforce reported $41.5bn of revenue and $72.4bn of remaining performance obligation. More importantly, Agentforce and Data 360 ARR exceeded $2.9bn, up over 200% YoY, including $800m of Agentforce ARR, up 169% YoY. Salesforce also reported more than 29,000 Agentforce deals since launch, 2.4bn Agentic Work Units delivered, and more than 19tn tokens processed.

The key innovation is the development of Agentic Work Units. Salesforce is trying to move pricing away from seats and toward tasks completed by AI agents. If Salesforce can charge for digital labour inside CRM workflows, it can capture part of the productivity gains that AI creates.

Conclusion

AI does not eliminate corporate value creation; it redistributes it. At the bottom of the stack, compute remains extremely profitable but increasingly capital-intensive. This favours mega-cap balance sheets, hyperscalers, and infrastructure platforms with access to power, chips, and financing. In the middle of the stack, models face the hardest strategic position: high R&D intensity, rapid performance convergence, and declining unit prices. Unless a model company owns distribution, proprietary data, or a direct customer relationship, its bargaining power may weaken over time.

At the top of the stack, application platforms could capture durable value because they own the customer relationship and the workflow where AI becomes economically useful. At the same time, their business models may shift materially as pricing moves from seats to outcomes, while competition increases from model developers moving into final applications.

Overall, the question for firms is shifting from whether they use AI to which part of the workflow they control, and whether they can price the outcome.

TAGS: AI, TechM&A, SaaS, Cloud Computing, NVIDIA, ServiceNow