The AI Stack as the New OSI Model

Where the Trillions Are Flowing — And Where They Aren't

Views and opinions are my own, but all data is backed by publicly available references. Feel free to share, redistribute or use full or part of this report for your own use.

— Amit Gupta, a1009us@gmail.com

LinkedIn: https://www.linkedin.com/in/amit-gupta-85300774/

27th March 2026

If you studied computer networking in the late 90s, you likely remember this line: **"All People Seem To Need Data Processing."** It was the classic mnemonic used to memorize the seven layers of the OSI Reference Model top-down: **A**pplication, **P**resentation, **S**ession, **T**ransport, **N**etwork, **D**ata Link, **P**hysical.

Today, we are watching history repeat itself in real time. We are living through an era of **"All People Seem To Need AI Processing"** — where each letter maps to a layer of the modern AI technology stack: **A**pplications, **P**resentation/Agentic AI, **S**ession/LLMs, **T**ransport/Compute, **N**etwork/Chips, **D**ata Centers, **P**hysical/Energy.

Just as the OSI model defined how the internet was built — from physical wires at Layer 1 up to end-user applications at Layer 7 — the artificial intelligence industry is coalescing into an almost identical technology stack. Tracking where the capital is flowing across these layers reveals exactly what is happening today, and more importantly, who will capture the ultimate value tomorrow.

> *"The parallel between the OSI model and the AI technology stack lies in their shared, layered approach to managing complex systems — moving from raw physical hardware at the bottom to end-user applications at the top. While OSI structures network communication, the AI stack structures the creation, training, and deployment of intelligent models."*
>
> — Towards Data Science

The AI Stack: A Modern OSI Parallel

By mapping the AI stack to the OSI model, we can trace the flow of trillions of dollars and identify the critical bottlenecks. The table below presents the full seven-layer comparison, with investment intensity and primary players at each level.

OSI Layer	AI Stack Equivalent	Core Function	Investment Level
L1: Physical	Energy	Power generation, cooling, grid	Catching up — the ultimate bottleneck
L2: Data Link	Data Centers	Physical compute facilities	~$700B annually from hyperscalers
L3: Network	Chips / Silicon	Processing & routing computation (GPUs/TPUs)	Enormous — NVIDIA $5T+ market cap
L4: Transport	Compute / Cloud	Reliable delivery of compute power	Heavy — AWS, Azure, Google Cloud
L5: Session	Foundation Models (LLMs)	Managing intelligence sessions	High valuations, massive losses
L6: Presentation	Agentic AI / Middleware	Translating intelligence into actions	Growing — new enterprise frontier
L7: Application	End-User Applications	Delivering measurable ROI	Significantly underinvested

Layers 1 & 2: Energy and Data Centers — The Physical Foundation

Just as the early internet required laying thousands of miles of submarine cables, the AI era requires an unprecedented build-out of physical infrastructure. The hyperscalers — Amazon, Google, Meta, and Microsoft — are projected to spend nearly **$700 billion** collectively on capital expenditures in 2026 alone, with the vast majority directed toward data centers and the energy required to run them.

Elon Musk has become the poster child for Layer 1 and 2 investments. His company, xAI, built the Colossus data center in Memphis — the world's largest GPU cluster — and recently secured approval for 41 natural gas turbines in Mississippi to generate 1.2 gigawatts of power. Meta is building a $10 billion, 5-gigawatt site in Louisiana dubbed Hyperion, complete with a nuclear power arrangement. Energy has become the ultimate bottleneck; you cannot train a frontier model if you cannot plug it in.

Layer 3: Chips and Silicon — NVIDIA, the Cisco of the 90s

In the OSI model, Layer 3 handles routing. In the AI stack, this is the semiconductor layer, where massive datasets are routed through parallel processors. **NVIDIA is the Cisco of the 90s**, building heavily in the exact hardware required to make the entire ecosystem function. Just as Cisco provided the essential routers and switches that built the backbone of the internet, NVIDIA provides the essential GPUs that serve as the backbone of the AI era.

With a market capitalisation exceeding **$5 trillion** by late 2025, NVIDIA is not just selling chips; they are aggressively investing across the stack, including a $100 billion commitment to

OpenAI (paid in GPUs) and heavy investments in photonics to solve data transfer bottlenecks. Meanwhile, the hyperscalers are investing billions to develop their own custom silicon — Google's TPUs and Amazon's Trainium — to reduce their dependence on NVIDIA.

Layers 5 & 6: LLMs and Agentic AI — Intelligence and Action

This is where raw compute is transformed into intelligence. The Foundation Model layer (Layer 5) has captured the public's imagination, but the economics are brutal. OpenAI recently closed a **$110 billion fundraising round** and reached $20 billion in annual recurring revenue by the end of 2025, yet the company burned through an estimated $8 to $9 billion in cash to achieve it. Anthropic, valued at $380 billion, is similarly operating at a loss despite securing massive investments from Amazon.

Because the pure LLM layer is commoditising rapidly — a trend accelerated by open-source models like Meta's LLaMA and the emergence of DeepSeek — the smart money is moving to Layer 6: Agentic AI. Agentic AI acts as the presentation layer, translating raw model outputs into autonomous, multi-step workflows. Microsoft (Copilot), Salesforce (Agentforce), and emerging startups are pouring billions into this space. The agentic AI market is expected to reach **$9 billion in 2026**, as enterprises shift from chatbots to virtual coworkers that can execute tasks across software ecosystems.

Layer 7: The Application Layer — The Most Underinvested Opportunity

In the OSI model, Layer 7 is where HTTP and SMTP live — the protocols that actually deliver value to the user. In the AI stack, this is the application layer, and it is surprisingly the most underinvested layer of the entire stack.

> *"The AI application layer is significantly underinvested. Despite explosive progress in foundational models, the true value lies in building AI applications."*
>
> — Andrew Ng, January 2026

While venture capital poured **$258 billion** into AI in 2025 — accounting for 61% of all VC funding — the vast majority of those funds went to infrastructure and foundational models. AI-native application startups often operate at negative gross margins because they are effectively subsidising the compute costs of the lower layers. Yet this is precisely where the internet's value ultimately concentrated.

The Profitability Paradox

When analysing the AI stack through the OSI lens, a fascinating profitability paradox emerges. The layers receiving the most capital investment are not necessarily the ones generating the

highest margins. The table below illustrates the stark contrast.

Layer	Gross Margin (approx.)	Status
Chips (NVIDIA)	~74%	Highly profitable
Cloud / Compute	~35%	Margin compression
Data Centers	~18%	Capex-heavy
Applications (SaaS)	~72%	Mature, high-margin
LLMs (OpenAI)	**Negative (~-45%)**	Burning cash
Energy	~12%	Catching up

NVIDIA is the exception — guiding toward quarterly revenues of $65 billion with gross margins in the mid-70% range. However, the hyperscalers are seeing their margins compressed by massive capital expenditures. The model layer (OpenAI, Anthropic) is generating negative margins.

This mirrors the telecom boom perfectly. The companies that laid the **fiber optic cables of the 2000s** eventually went bankrupt or saw their services commoditised into "dumb pipes." The durable, high-margin value was ultimately captured by the application layer. If AI models become commoditised, the economic centre of gravity will inevitably shift upward.

A Bold Prediction for 2026–2036

The AI industry is currently trapped in the lower layers of the stack. We are constrained by power grids, data center real estate, and chip yields. Consequently, that is where the trillions are flowing today. But infrastructure is a means to an end.

> *The ultimate winner of the AI era has likely not even been founded yet — or is currently a small startup flying under the radar. Just as the massive infrastructure investments of the 90s paved the way for Google, Amazon, and Meta to dominate the next two decades, today's $700 billion infrastructure build-outs are paving the way for a new apex predator. This new giant will come from nowhere, capture the imagination and market share at Layer 7 (Applications), and become the most dominant force to reckon with from 2026 to 2036.*
>
> — Amit Gupta, March 2026

They will not win by building better chips or larger data centers; they will win by taking the commoditised intelligence provided by Layers 1 through 6 and turning it into an indispensable, everyday utility that changes how humanity lives and works.

The picks and shovels are being sold. The gold rush is about to begin.

References

1. TechCrunch. "The billion-dollar infrastructure deals powering the AI boom." February 2026.
2. Data Center Dynamics. "Musk's xAI gets go-ahead for 41 natural gas turbines in Mississippi." March 2026.
3. TechCrunch. "The billion-dollar infrastructure deals powering the AI boom." February 2026.
4. Reuters. "From OpenAI to Nvidia, firms channel billions into AI infrastructure." March 2026.
5. LinkedIn Pulse. "Where the AI Value Actually Accrues." February 2026.
6. Shanaka Anslem Perera. "The Growth Miracle and the Six Fractures: Anthropic at $380 Billion." February 2026.
7. Tech Insider. "Agentic AI in Enterprise 2026: $9B Market Analysis." March 2026.
8. Andrew Ng via LinkedIn. "Is there an AI bubble?" January 2026.
9. OECD. "Venture capital investments in artificial intelligence through 2025." February 2026.