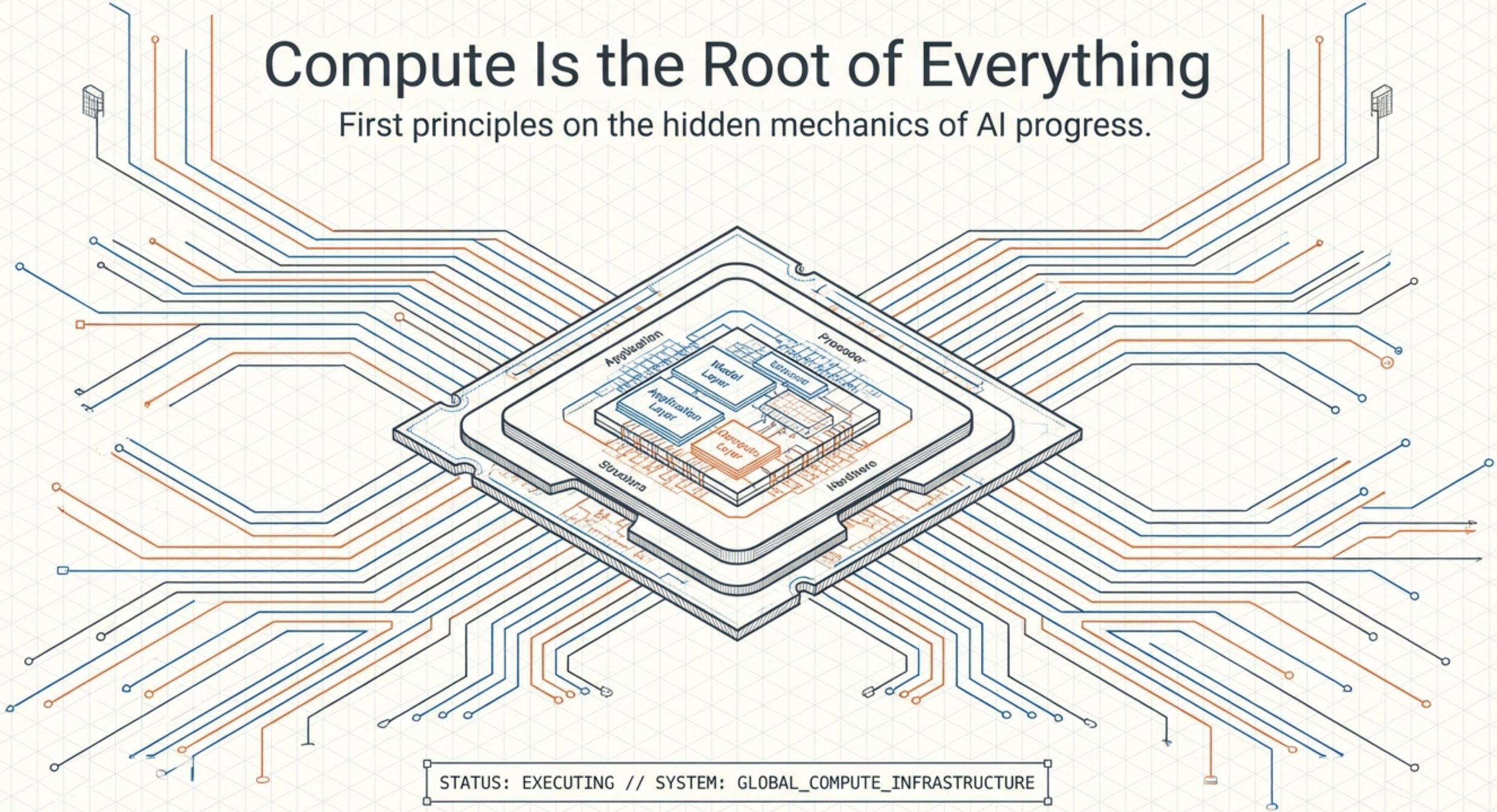
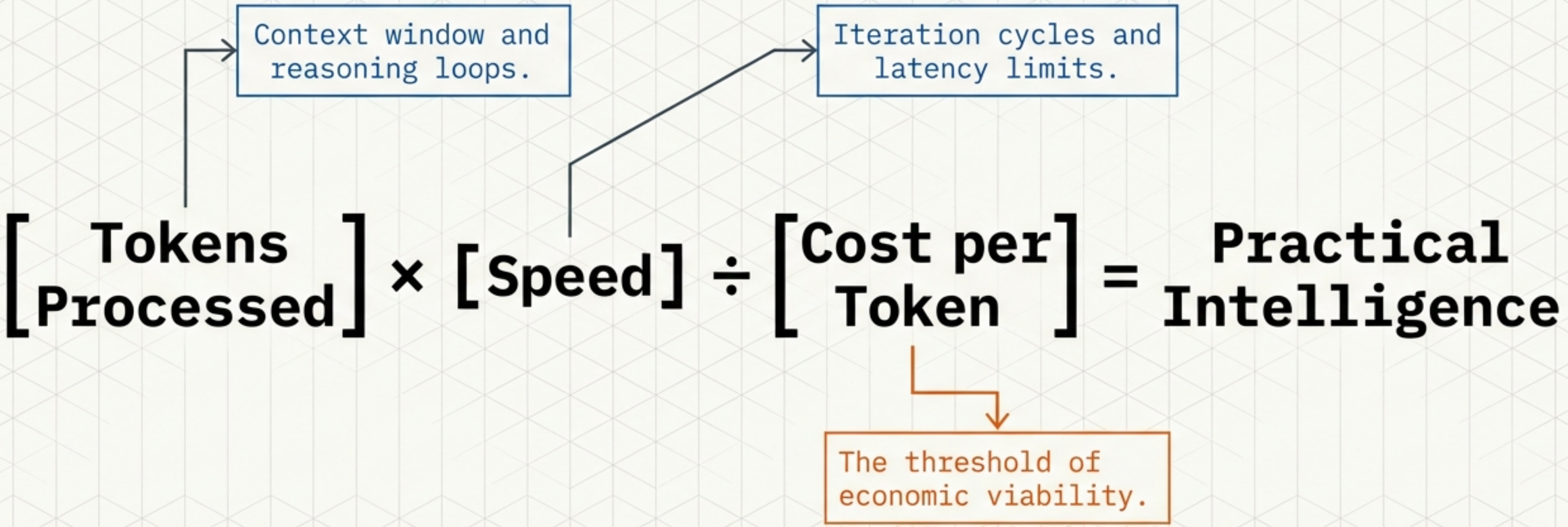
Compute Is the Root of Everything

First principles on the hidden mechanics of AI progress.



STATUS: EXECUTING // SYSTEM: GLOBAL_COMPUTE_INFRASTRUCTURE

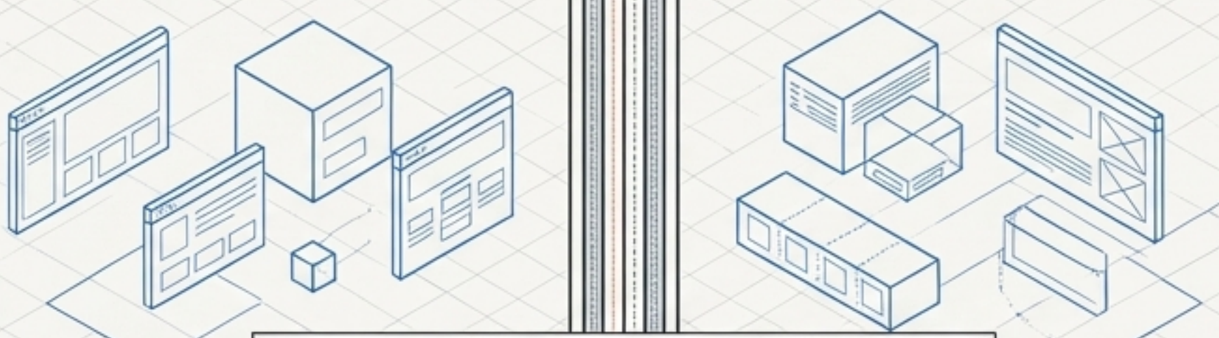
Intelligence is an engineering equation



Philosophers debated the soul. Engineers optimize the compute. For products, tools, and agents, frontier capability is useless if it cannot be afforded at scale.

Two altitudes observing the exact same variable

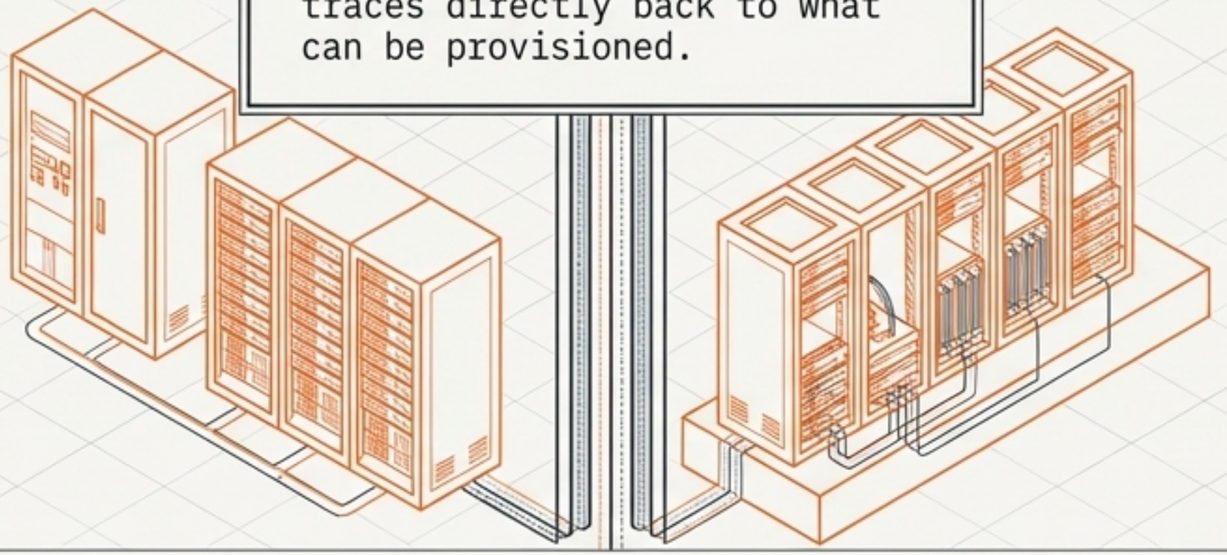
The Application Lens



Constraint:
The model isn't smart enough
OR
Inference is too expensive.

The root constraint for both is COMPUTE. What becomes possible traces directly back to what can be provisioned.

The Infrastructure Lens

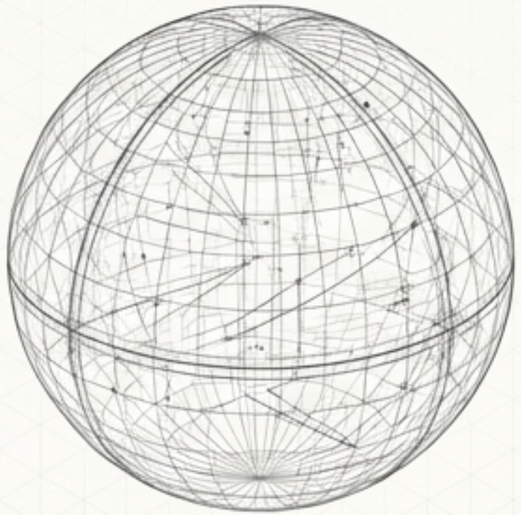


Constraint:
Financing hardware, building clusters, scaling chips.

The conventional narrative is incomplete

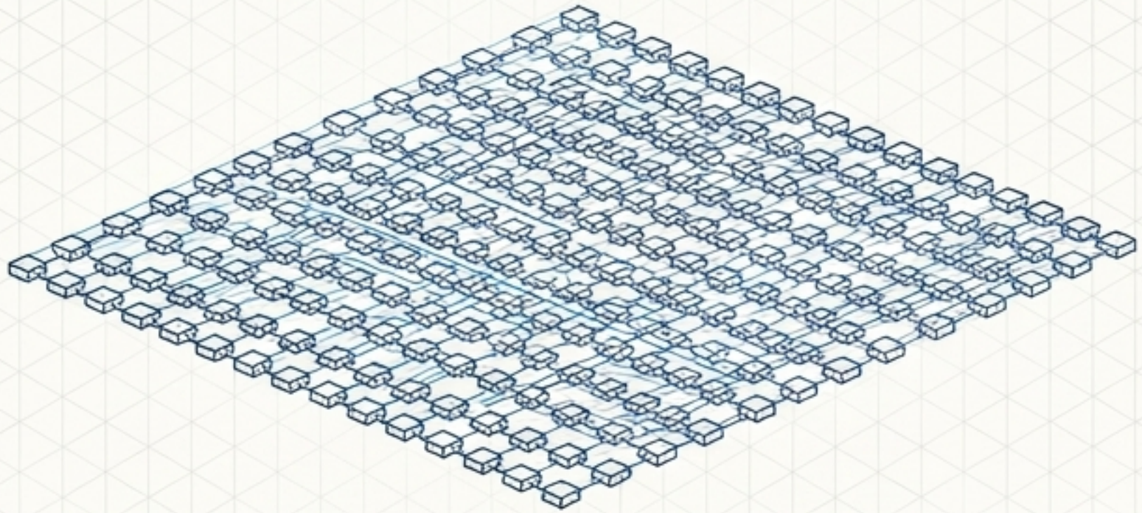
The Old Narrative

- **Core Belief:** Progress = Smarter Models
- **Focus:** Training compute breakthroughs
- **Result:** Frontier capability, but economically restricted



The Complete Reality

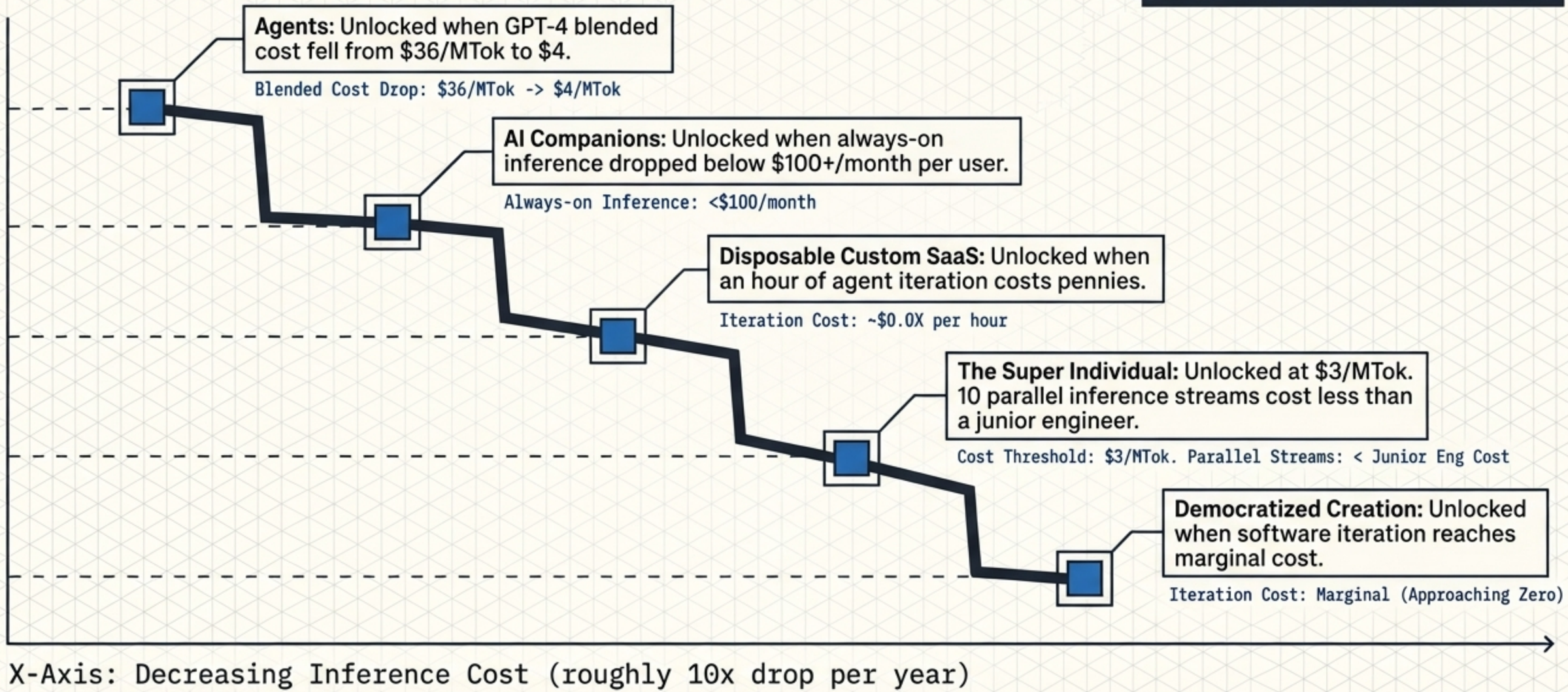
- **Core Belief:** Progress = Affordable Compute
- **Focus:** Inference compute efficiency (dropping 60%+ per generation)
- **Result:** Viable deployment, mass adoption, new product categories



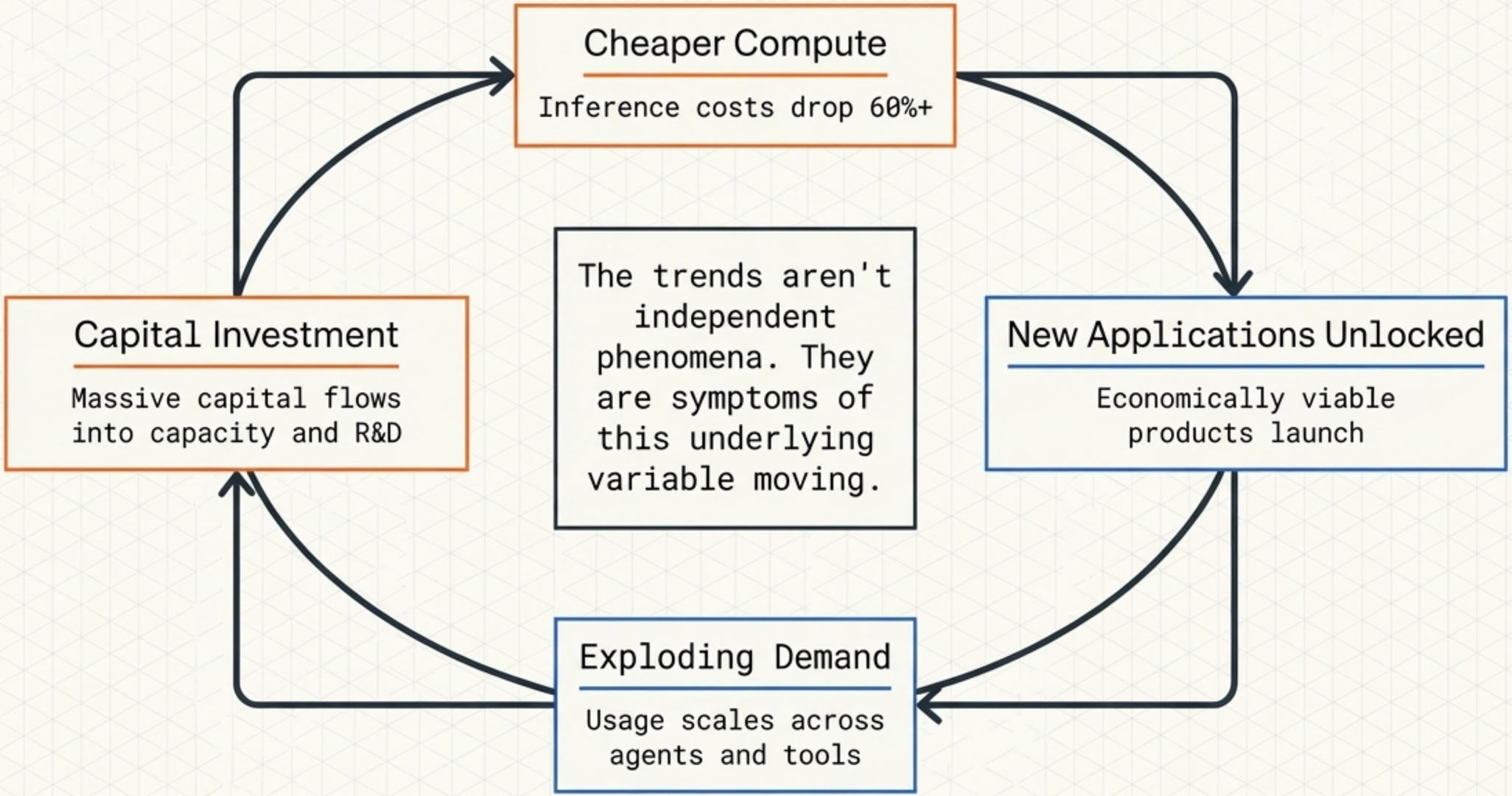
Intelligence isn't abundant. Affordable intelligence is even scarcer. Both rely on a single input.

The boundary of possibility expands as costs collapse

Every major trend in AI maps directly to a cost threshold being crossed.

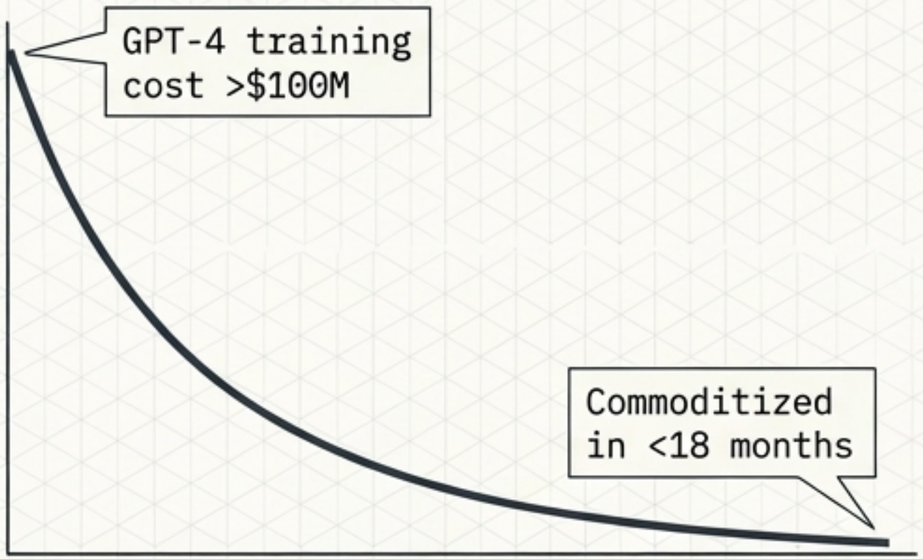


The causal flywheel of artificial intelligence



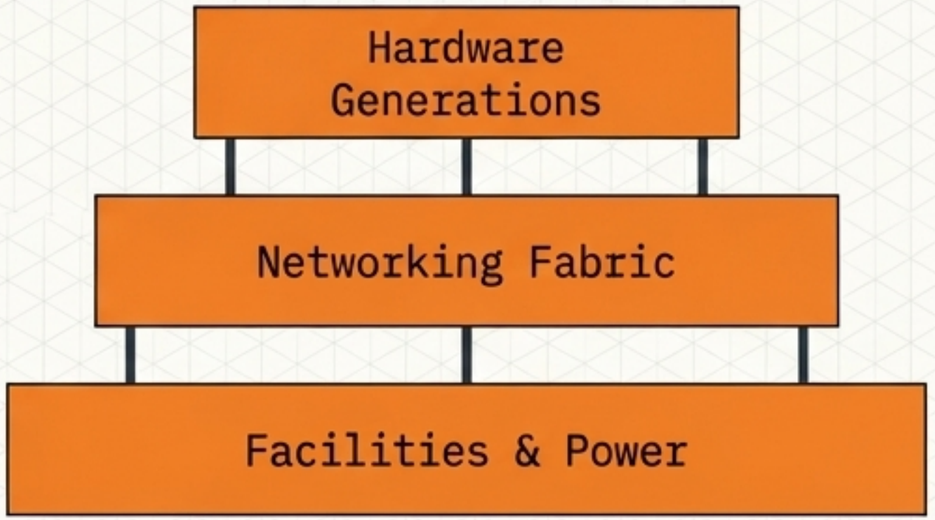
Models depreciate. Infrastructure appreciates.

The Commodity Layer



Models have a half-life of 12-18 months. The frontier model today is the legacy system of next year.

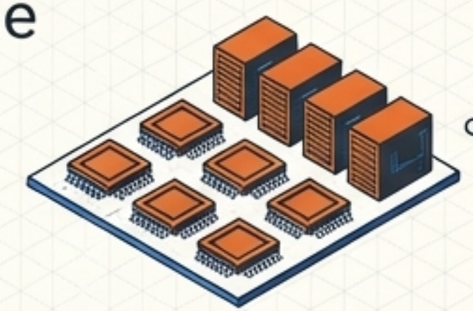
The Durable Layer



A data center built today runs tomorrow's models. Infrastructure and cognitive accumulation hold value across multiple architectural generations.

Betting on a specific model is like betting on a specific website in 1998. Betting on compute is betting on the internet backbone.

The true cost of intelligence is hidden below the surface



100 H100 GPUs
Purchase Cost: $3.0M
(35% of Total)

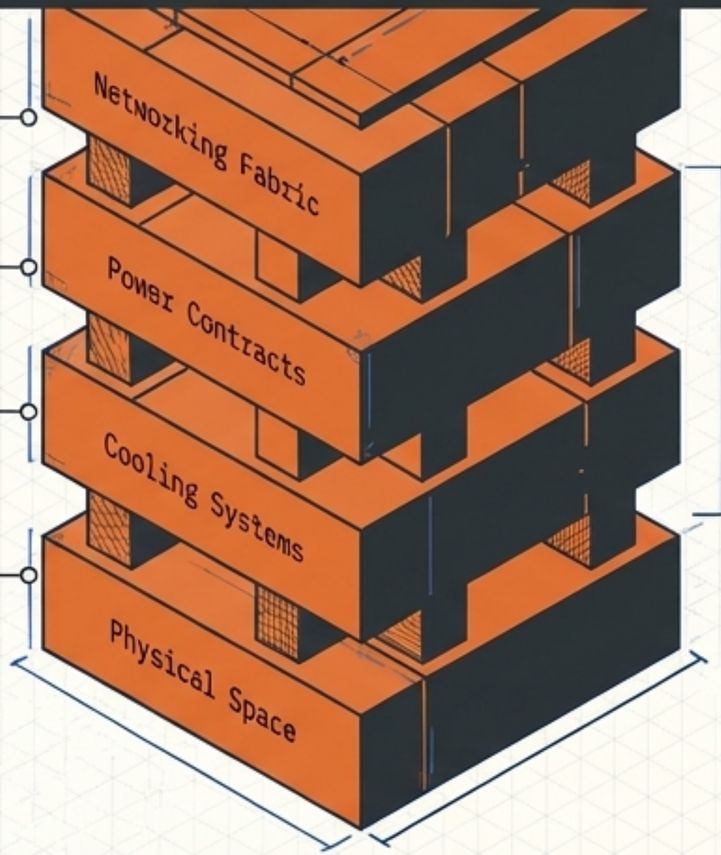
DATA CENTER FLOOR / DATAPLANE

Networking Fabric

Power Contracts

Cooling Systems

Physical Space



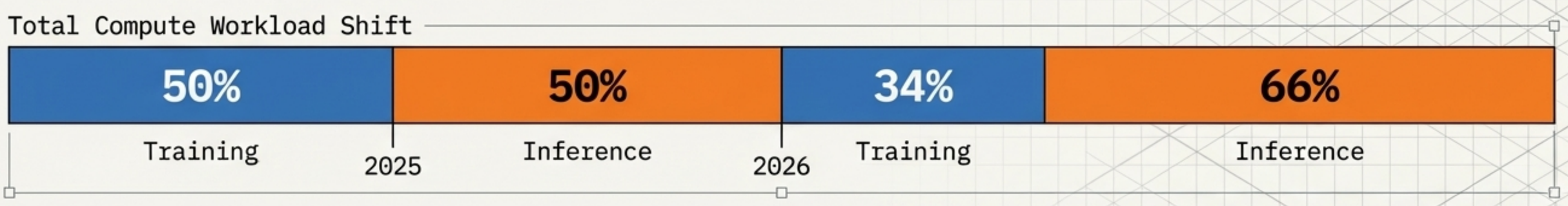
Hidden Infrastructure Costs:
$5.6M (65% of Total)

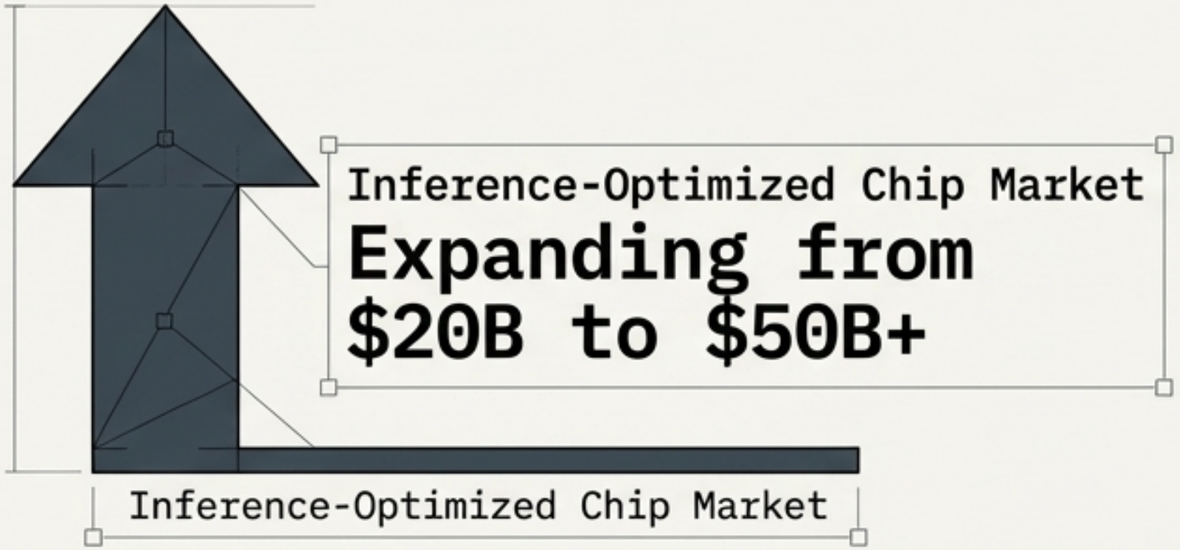
Takeaway

The GPU is only part of the picture. The durable infrastructure surrounding the silicon holds the real physical and financial weight.

5-Year Total Cost of Ownership = $8.6M

The permanent transition to **inference-dominated** workloads





Training is periodic. Inference is forever. Every user, every agent session, every always-on companion burns inference continuously.

The chip wars are a philosophical disagreement

	GPU (Nvidia)	TPU (Google)	ASIC (Groq)
Underlying Belief	Intelligence is divergent. The future is unpredictable.	Intelligence is convergent. Matrix math/Transformers are the core.	Intelligence is settled. Compute pattern is stable; execution is everything.
Primary Strategy	Maximize flexibility. Rely on CUDA cultural moat.	System-level co-optimization (chip, interconnect, software).	Pure engineering speed and power efficiency.
Risk Profile	High option value; safest in shifting landscape.	Path to dominance if Transformers persist 3-5 years.	Highest conviction, but vulnerable to paradigm shifts (expensive scrap).

`Regardless of who wins, the mechanism works: compute gets cheaper.`

The compute flywheel is choked by physical reality

The Demand

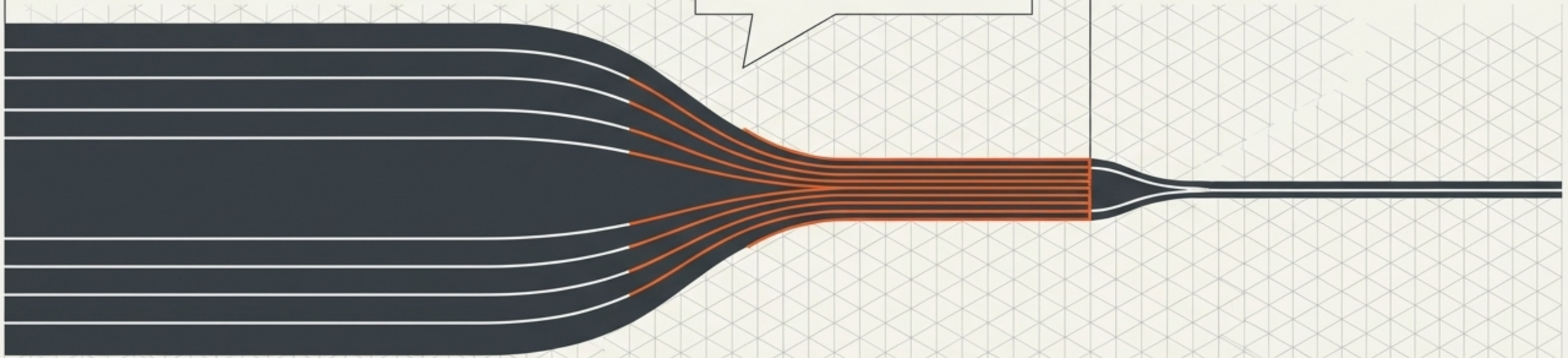
- $1 Trillion in orders (Double from a year ago)
- Chips are more of a limiter than even power.
- Chips are more of a limiter than even power.

Data center GPU lead times: 36-52 weeks.

Blackwell sold out through mid-2026 (3.6M backlog).

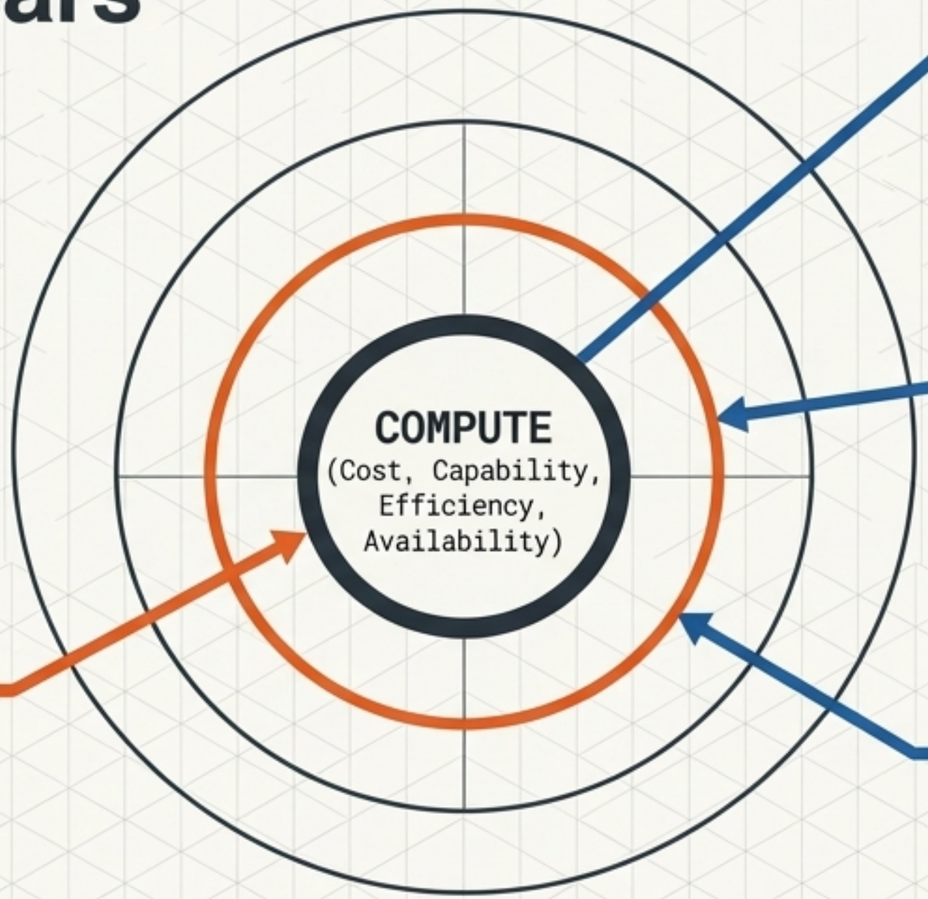
The Supply

- ~3M H100-class chips sold ’22-’24
- Meta stockpiling 1.3M by year-end
- Meta stockpiling 1.3M by year-end



Compute isn’t just a cost problem. It’s a supply problem. The industry can only move as fast as the physical chips come off the line.

The only reliable predictor for the next 5 years



The gap between “what AI can do” and “what is affordable to deploy” narrows to zero.

Agent workloads become universal; personalized AI becomes a utility.

Inference cost drops an order of magnitude every 18-24 months.

The gap between “what AI can do” and “what is affordable to deploy” narrows to zero.

Do not evaluate ideas based on the models of today. Evaluate them by asking:
“At what compute price point does this become viable?”
Everything traces back to the curve.