

Software Ate the World.

Hardware Must Not
Starve the *Dream*.

The choices leaders make about AI infrastructure will determine who can afford intelligence.

EXECUTIVE PERSPECTIVE

By June Manley

SEPTEMBER 2026

The next AI divide will be defined less by who can access intelligence than by who can keep its cost curve falling.

CONTENTS

- I The Physics Underlying the AI Boom
- II The Compute Divide
- III Efficiency Will Not Solve the Energy Problem
- IV The Strategic Asset: Compute Optionality
- V Three Decisions Leaders Need to Make
- VI Hardware Must Not Starve the Dream
- Sources and notes

Artificial intelligence is often treated as a model-selection decision. But as intelligence becomes a metered, infrastructure-dependent resource, the more consequential question is who controls the cost curve beneath it. Data-center power, custom silicon, inference efficiency, and the difficulty of moving workloads increasingly determine whether the cost of intelligence falls quickly or remains tied to another provider's economics.

This paper argues that the emerging divide will not simply be between organizations with and without AI. It will be between those able to follow a falling cost curve and those locked into one they cannot control. Recent moves reinforce the pattern: OpenAI is moving down the stack with its Jalapeño inference accelerator, while NVIDIA is moving up toward model distribution and developer workflow with its agreement to acquire Hugging Face.^{6,8,9}

For leaders, the response is not to avoid dependency. It is to make dependency deliberate. That means right-sizing the intelligence, placing the workload deliberately, and preserving leverage, while shifting the executive metric from cost per token to cost per useful outcome.

THREE EXECUTIVE DECISIONS

1

Right-size

Use the least expensive capability that can reliably deliver the required outcome.

2

Place

Match the environment to the workload rather than defaulting every task to the same provider or model.

3

Preserve leverage

Know where dependency is intentional, what value it creates, and what changing it would cost.

First published September 2, 2026. Updated September 20, 2026 to reflect NVIDIA's Hugging Face agreement, infrastructure commitments, and related developments.

Introduction

The printing press democratized the written word. The personal computer democratized productivity. The internet democratized access to information. Artificial intelligence (AI) could be the most consequential leap yet: the democratization of intelligence itself.

Its promise extends far beyond automated workflows or optimized balance sheets. Artificial intelligence could be a personalized, infinitely patient tutor for a child regardless of ZIP code; provide diagnostic support for a health worker far from a major medical center; or make research, analytical, technical, and strategic capabilities available to an entrepreneur who could never afford the team that once provided them. This is the promise of cognitive abundance.

But AI differs from the technological revolutions that came before it in an important way.

Over time, the marginal cost of distributing books, software, and information fell dramatically. Once software was built, providing it to another user was comparatively cheap. Capability spread because distribution kept getting easier.

In his landmark 2011 essay, Marc Andreessen captured the economic logic of that era in five words: “Software is eating the world.” He was right. Software moved into nearly every industry, in part because code could be replicated and distributed at extraordinarily low marginal cost.¹

AI breaks that pattern: intelligence has to be computed every time it is used.

Creating advanced intelligence is expensive. We expected that. What organizations are now discovering is that using it is expensive, too. Every inference consumes compute. That compute requires silicon and electricity. Electricity produces heat. Heat has to be removed. As AI moves from occasional prompts to persistent agents that reason, call tools, evaluate results, and try again, a single business outcome may require many model interactions rather than one.

When software is asked to reason, it collides with physical reality. Intelligence has become infrastructure.

That matters to business leaders because most companies are still treating AI architecture primarily as a model-selection decision: Which model is best? Which provider should we use? Which applications should we deploy?

Increasingly, however, their long-term economics will be determined one layer lower. What level of compute does the workload actually require? Where should that compute run? And how difficult will it be to change course as the economics evolve?

These may sound like infrastructure questions. They are becoming strategy questions.

The Physics Underlying the AI Boom

27 kW Average rack density in 2026, up from 16 kW a year earlier.²

945 TWh Projected data-center demand by 2030, from 415 TWh in 2024.⁴

4+ yrs Average wait for a grid connection in primary markets.⁵

The physical requirements of AI infrastructure are changing remarkably quickly. The Association for Computer Operations Management's 2026 State of the Data Center report put average rack density at 27 kilowatts, up from 16 kilowatts the year before. At the frontier, the figures are far higher: NVIDIA says its GB200 NVL72 rack consumes approximately 120 kilowatts.^{2,3}

That is not simply a matter of installing a larger server. It changes power delivery, cooling, facility design, location, and, ultimately, cost.

Electricity demand is rising with it. The International Energy Agency estimates that data centers consumed roughly 415 terawatt-hours of electricity globally in 2024. In its base case, demand rises to approximately 945 terawatt-hours by 2030, just under 3% of global electricity demand. AI is an important driver of that increase.⁴

Then there is a constraint that capital cannot solve on its own: power availability. Jones Lang LaSalle reports that the average wait for a grid connection in primary data center markets now exceeds four years. In some places, the facility can be built faster than the power needed to operate it can be secured.⁵

Time to power is becoming as important as time to build.

The shift toward agentic systems adds another dimension. A conventional application may respond to a request through a relatively predictable set of computations. An AI agent, by contrast, can reason through a problem, invoke external tools, inspect the results, revise its approach, and try again. The employee sees a single completed task; the underlying infrastructure may have performed dozens of model interactions to produce that result.

Multiply that across sales, engineering, finance, customer support, security, and operations, and what looked like a software licensing decision starts to look more like a decision about industrial capacity.

Brute force helped build modern AI. Now, AI is meeting the limits of physics.

But constraints on power and compute are only half the story. The more important strategic question is what happens to the economics of intelligence as the companies with the greatest scale begin optimizing around these constraints.

The Compute Divide

The companies with the greatest scale are no longer optimizing one layer of the stack. They are integrating across it.

OpenAI is moving down. Its Jalapeño inference accelerator is designed around the company's own models, and in OpenAI's InferenceX testing, the company reported 1.5 to 1.9 times more AI work per watt at peak throughput than the comparison systems across the public models tested.⁶ Its chief financial officer describes the strategy as preserving choice across providers, hardware, and deployment models while building first-party silicon for leverage.⁷

NVIDIA is moving up. On September 3, 2026, it announced a definitive agreement to acquire Hugging Face for \$12.93 billion. If completed, the transaction would bring the largest open model hub, used by more than 18 million developers, under the ownership of the largest supplier of AI accelerators. The transaction is expected to close in the first half of 2027.^{8,9}

NVIDIA has committed that Hugging Face will remain open: multi-cloud, multi-accelerator, with no requirement to use NVIDIA compute.⁸ That commitment matters, and leaders should take it at face value. They should also notice what it illustrates. Openness at that layer is now a policy set by an infrastructure owner, not a structural feature of the market.

It is also moving down. NVIDIA now tracks available power worldwide, invests in power producers, and has signed \$20 billion in data center

leases it intends to pass to customers, so that scarce electricity flows to facilities filled with its chips.¹⁵

The same week showed the pattern from other angles. NVIDIA invested \$3.5 billion in MediaTek, whose custom accelerators will connect through NVIDIA's NVLink Fusion platform, keeping NVIDIA's interconnect central even where it does not supply the chip.¹³ And Anthropic's reported \$35 billion capacity agreement with Lambda runs through a Texas data center on which NVIDIA reportedly holds the lease, making one company chip supplier, investor, and landlord in the same transaction.¹⁴

At this scale, a company can design hardware around recurring workloads, optimize software for that hardware, secure power at volume, and spread its investment across enormous inference volumes. Most enterprises cannot.

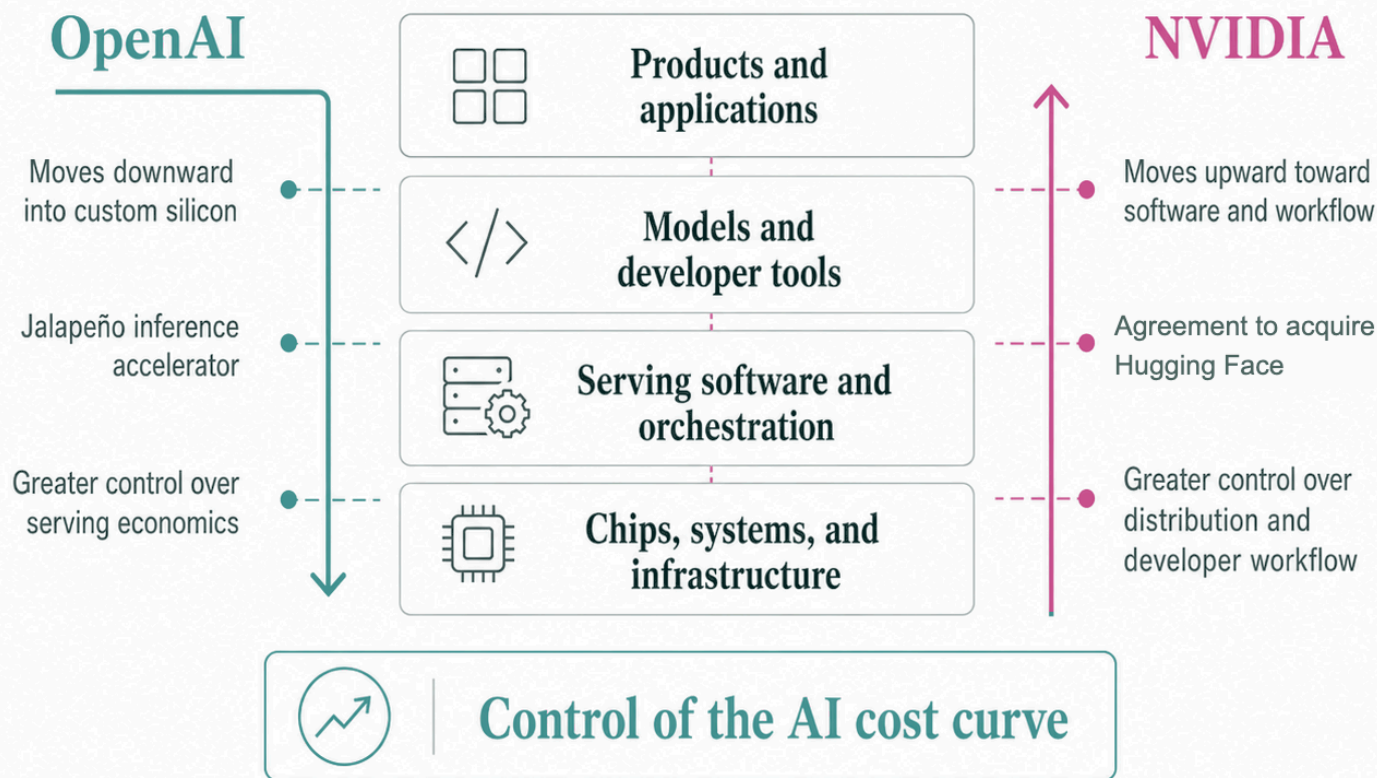
TEN DAYS OF CONVERGENCE

Aug 25	OpenAI publishes first Jalapeño inference results.
Aug 31	Anthropic's reported \$35B capacity deal with Lambda.
Aug 31	NVIDIA announces \$3.5B MediaTek investment and expanded NVLink Fusion partnership.
Sep 3	NVIDIA agrees to acquire Hugging Face for \$12.93B.

FIGURE 1

The AI Stack Is Converging

Vertical integration is moving from both directions



Frontier AI companies are becoming full-stack infrastructure companies.

Figure 1. Vertical integration is reshaping who captures the economics of intelligence.

Even the frontier labs are buyers in this market. The difference is that they negotiate at a scale that lets them shape their own cost curve.

NVIDIA has reported \$279 billion in supply and capacity commitments, primarily related to memory. Almost no one else can operate at that scale.¹⁵

Amazon offers Trainium through AWS, and Google offers TPU capacity through Google Cloud. So the risk is not that enterprises will be denied efficient compute. They can often rent it.^{10,11}

The issue is that the provider controls the cost curve.

The provider owns the infrastructure, determines the commercial terms, allocates capacity, and decides how much of an efficiency gain is passed on to the customer. An enterprise may benefit as the provider's economics improve, but it has no control over how much of that benefit it receives or how quickly.

This creates a different kind of AI divide. It is unlikely to be a simple divide between organizations that have AI and those that do not. AI capabilities will become broadly available.

The more consequential divide may be between organizations whose cost per useful outcome continues to fall and those whose costs decrease more slowly because they depend on infrastructure and commercial models that are difficult to change.

Lock-in is not new, nor is it inherently harmful. Enterprises have always accepted proprietary dependencies when the capability, speed, or economics justified them. The problem arises when platform

dependency becomes cost-curve lock-in, that is, when the technical and economic costs of moving prevent an organization from following a materially better cost curve elsewhere.

That alternative could be a different model, accelerator, or cloud provider; private infrastructure; or inference moved closer to the edge. The problem is not that an enterprise chooses a platform. The problem is discovering, too late, that the cost of changing that choice is greater than the economic advantage of the alternative.

This becomes more consequential as AI becomes more deeply embedded in the operating model. An inefficient choice may barely register when employees make a few thousand AI requests a day. The economics look very different when autonomous systems generate millions of model interactions across the business.

A small architectural disadvantage, repeated millions of times, becomes a business-model disadvantage.

The strategic question is therefore no longer simply: What does a token cost today? It is: How much control do we have over what intelligence will cost us tomorrow?

COST-CURVE LOCK-IN

When the technical and economic costs of moving prevent an organization from following a materially better cost curve elsewhere.

\$279B

NVIDIA supply and capacity commitments, primarily related to memory.

\$20B

Data center leases NVIDIA intends to pass to customers.

FIGURE 2

The Compute Divide

Who captures the efficiency gains from AI infrastructure?

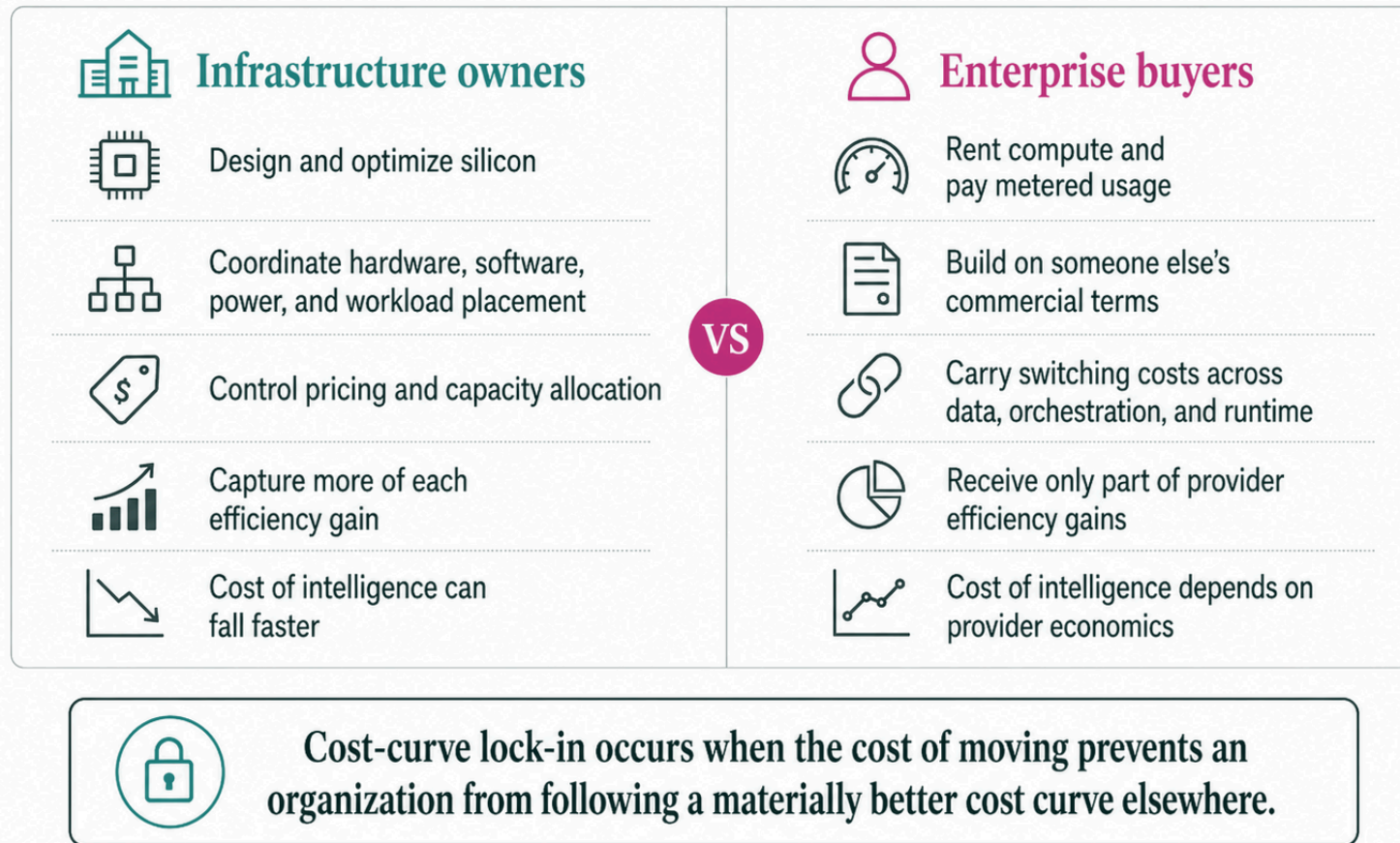


Figure 2. The emerging divide may be less about access to AI and more about the economics of intelligence.

Efficiency Will Not Solve the Energy Problem

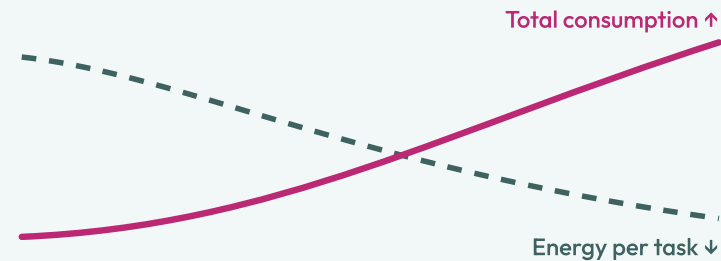
There is an obvious objection to the efficiency argument, and it deserves to be taken seriously.

Efficiency gains in computing rarely reduce total consumption. Cheaper computing drove more applications. Cheaper storage drove more data. Lower networking costs drove more traffic. There is little reason to assume cheaper inference will produce a different result.

As AI becomes more efficient, companies will find more applications for it. Agents that are uneconomic today will become viable, while intelligence will be embedded in products and processes whose costs cannot currently be justified. Total AI energy consumption may therefore continue to rise even as individual workloads become dramatically more efficient.

Pat Gelsinger, former CEO of Intel, has argued that inference needs to become at least 10,000 times more efficient, requiring advances across memory, interconnects, and system architecture. The figure suggests how far AI infrastructure may still need to evolve.¹²

JEVONS PARADOX



Efficiency can increase overall use. Falling cost per task unlocks demand faster than per-task resource use declines.

Even dramatic efficiency gains do not guarantee lower aggregate consumption if falling costs unlock demand faster than per-task resource use declines. This is Jevons paradox: efficiency can increase overall use.

So efficiency should not be judged by whether total AI consumption falls.

The goal is not less compute. The goal is more participants.

For business leaders, this changes the optimization question. The question is not whether a frontier model can perform a task, but which type of system can perform it reliably at the lowest cost.

10,000×

The efficiency gain Gelsinger argues inference still needs, across memory, interconnects, and system architecture.¹²

Sometimes the answer will be a frontier model. Sometimes it will be a smaller language model. In other cases, conventional machine learning will be sufficient. And sometimes the right answer will be deterministic software rather than AI at all.

That distinction matters in a clinic, factory, farm, vehicle, or retail environment, where a narrowly defined task may not require the most capable model available. If a smaller or specialized model can perform the task reliably, the economics may change considerably when it operates locally rather than requiring continuous access to a large remote model.

Compute is not the only obstacle to deploying AI in settings such as health care. Validation, regulatory compliance, integration, data quality, cybersecurity, maintenance, and workflow design all matter. But compute should not become an unnecessary obstacle simply because the default architecture routes every problem to the largest available model hosted in a third-party cloud.

Greater efficiency can increase overall AI consumption while enabling more organizations to participate. The two are not contradictory.

THE CAPABILITY LADDER

Use the least expensive capability that can reliably deliver the outcome.

Frontier model

HIGHEST COST

Smaller or specialized model

OFTEN ENOUGH

Conventional machine learning

BOUNDED TASKS

Deterministic software

LOWEST COST

Figure 3. The winning architecture is often the simplest tool capable of doing the job well.

IV The Strategic Asset: Compute Optionality

If the provider controls the cost curve, the enterprise needs something else to bargain with. That something is optionality.

Compute optionality is the practical ability to change models, accelerators, providers, or locations when the economics justify it. It is not a commitment to change. Most organizations will, and should, keep most workloads where they run today. The asset is the credible ability to move, because a credible alternative changes the terms on which you stay.

The largest players already behave this way. OpenAI describes its strategy as preserving choice across providers, hardware, and deployment models, even as it builds first-party silicon.⁷ Anthropic has reportedly spread its capacity commitments across several providers and more than one accelerator family.¹⁴ NVIDIA, in announcing its agreement to acquire Hugging Face, committed that the platform would remain multi-cloud and multi-accelerator, an acknowledgment that developers value the choice.⁸

If the organizations with the most leverage in the market treat optionality as strategy, enterprises with far less leverage have more reason to do so, not less.

Optionality has a cost. Abstraction layers add engineering effort. Portable architectures may give up some of the performance available from a tightly integrated stack. Spreading volume across providers dilutes discounts. The goal is not maximum portability. It is economically meaningful choice: enough freedom to follow a materially better cost curve when one appears, at a switching cost the business has measured in advance.

THE PRINCIPLE

The architecture should follow the work, rather than the work becoming permanently attached to the architecture.

Figure 4 overleaf: three decisions leaders need to make.

FIGURE 4

Three Decisions Leaders Need to Make

A framework for preserving compute optionality

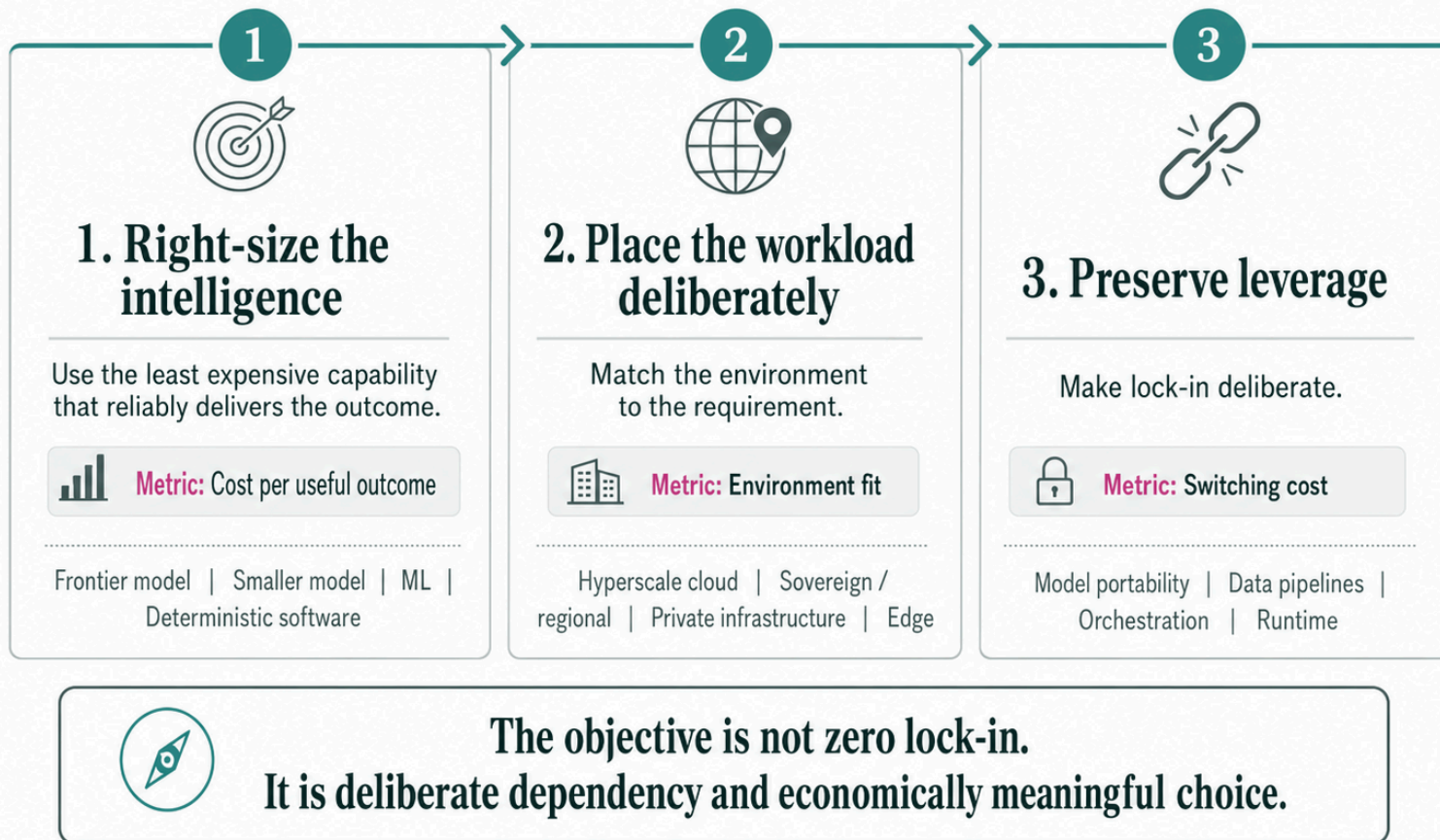


Figure 4. Compute optionality becomes a strategic asset when leaders preserve economically meaningful choice.



Three Decisions Leaders Need to Make

1. Right-Size the Intelligence: Measure Cost per Useful Outcome

The first question is simple: What level of capability does this task actually require?

The answer should not automatically be “the best model available.”

Many production workloads are bounded and repetitive.

Classification, extraction, summarization, routing, search, document processing, and other domain-specific tasks may not require frontier-level reasoning every time they run.

The principle is straightforward: use the least expensive capability that can reliably deliver the required outcome. Using the most capable model for every task is the AI equivalent of running every application on the largest cloud instance in the catalog. It works, but it is an expensive default.

Model routing, caching, retrieval, smaller specialized models, conventional machine learning, and deterministic software all create alternatives. Choosing among them requires a better economic measure than model benchmark performance alone.

The same logic that led cloud FinOps toward unit economics applies here. Cost per token is useful operationally, but it measures an input. The more meaningful business metric is cost per useful outcome.

What did the system actually accomplish for the money and compute consumed?

A customer issue resolved successfully, a software defect fixed, a qualified lead produced, or a document reviewed accurately matters more to a business than the number of tokens required to get there. CIOs, CFOs, and AI leaders should be able to connect model capability, inference volume, infrastructure cost, and business value at that level.

The winning architecture will not always use the smartest model available. Often, it will use the simplest tool capable of doing the job well.

2. Place the Workload Deliberately: Match Environment to Requirement

The second decision is where the intelligence should run.

During the first phase of generative AI, the default was often the provider’s cloud. For many workloads, this will remain the right answer. However, it is no longer the only answer.

Inference can increasingly run across a continuum: frontier models on hyperscale infrastructure, regulated workloads in sovereign or regional environments, sensitive workloads on enterprise-controlled infrastructure, and high-volume or latency-sensitive workloads closer to the edge.

Every placement choice involves tradeoffs among capability, cost, latency, privacy, sovereignty, reliability, and operational complexity. These tradeoffs should be explicit. A workload should not run somewhere simply because that is where the company's first AI vendor made an application programming interface (API) available.

Placement should follow the requirements of the work.

3. Preserve Leverage: Make Lock-In Deliberate

The third decision is where the organization is willing to accept dependency and what it receives in return.

No meaningful enterprise architecture is completely free of lock-in, nor should that be the objective. Proprietary cloud services, models, development tools, and infrastructure can create enormous value. The management question is whether leaders know where these dependencies are located and what changing them would cost.

AI makes this harder than it first appears because switching a model endpoint can create an illusion of portability. A company may be able to swap one model for another while its data pipelines, agent orchestration, identity controls, observability, application services, and runtime remain deeply embedded in one platform.

Portability therefore has to be evaluated at the level of the business capability, not simply the model API.

FOUR QUESTIONS TO ASK

Can the organization change models without rebuilding the application?

Could it move part of the workload to another accelerator or environment if the economics justified it?

Which data, orchestration, security, and governance dependencies would have to move as well?

How long would that take, and at what cost?

WHAT MOVES WITH THE MODEL

Data pipelines

Agent orchestration

Identity controls

Observability

Application services

Runtime

Swapping a model endpoint leaves these in place. Portability has to be evaluated at the level of the business capability.

THE DECISION BENEATH THE DECISION

When an enterprise chooses an AI architecture, it is not just deciding how a workload runs. It is choosing which cost curve it will participate in.

No meaningful architecture is free of lock-in. The question is whether you know where yours is.

These are not arguments against choosing a platform. They are arguments for choosing deliberately.

There is a significant difference between accepting a dependency because its value exceeds the switching cost and discovering the dependency only after the economics turn against you. Leaders should know where they are locking in, why they are doing it, and whether that decision could prevent them from following a better cost curve later.

These choices also send a market signal. Buyers that ask for portability and measure AI by useful outcomes create demand for more efficient, interchangeable architectures. Buyers that accept tightly coupled systems without understanding the switching economics encourage the opposite.

**Buyers that ask for portability**

Demand grows for more efficient, interchangeable architectures.

**Buyers that accept tight coupling**

The switching economics stay unexamined, and the opposite is encouraged.

**Either way, the choice compounds**

A small architectural disadvantage, repeated millions of times, becomes a business-model disadvantage.

Hardware Must Not Starve the Dream

The case for AI was never merely that companies could automate more workflows or improve margins. Its larger promise was that expertise, historically one of the scarcest and most unevenly distributed resources in the economy, could become abundant.

That promise is colliding with an uncomfortable reality. Software taught us to expect that once a capability was created, its cost would fall and access would spread. AI does not automatically inherit that economic advantage. Intelligence has to be computed every time it is used, and the infrastructure required to deliver it is becoming more specialized, capital-intensive, and concentrated.

One path forward treats AI primarily as an arms race: more parameters, more accelerators, larger clusters, bigger facilities, and more electricity. That path will continue to produce extraordinary systems. But if it becomes the only practical way to deliver meaningful intelligence, the organizations that control the infrastructure will also have the greatest ability to drive down its cost. Everyone else risks building on economics they do not control.

That is not democratized intelligence. It is democratized access to increasingly concentrated infrastructure.

The alternative does not require abandoning frontier computing. It requires an ecosystem that uses frontier models where frontier capability is genuinely needed, smaller models where they are sufficient, specialized accelerators where efficiency changes the

economics, distributed intelligence where privacy, sovereignty, connectivity, or latency demand it, and architectures that leave room to change course as technology and pricing evolve.

The printing press made knowledge reproducible. The personal computer made computation personal. The internet made information global. AI can make expertise abundant. But abundance is not guaranteed by increasingly capable models. It depends on whether the economic and physical foundation underlying them allows intelligence to become broadly affordable.

The tutor exists. The diagnostic capability exists. The entrepreneur can already access expertise once available only to institutions with enormous resources. The question is whether these capabilities become progressively cheaper and more accessible, or whether the infrastructure required to deliver them creates a new divide between those who can afford intelligence at scale and those who cannot.

Software ate the world because its economics allowed capability to spread. AI can go further and spread capability itself, but only if the infrastructure underlying it does not become the constraint that prevents that promise from reaching everyone.

Sources and notes

Selected evidence underlying the quantitative and time-sensitive claims in this paper.

Publication note. First published September 2, 2026. This edition reflects public information available through September 20, 2026. The Anthropic and Lambda agreement is based on press reports and had not been confirmed by the companies at the time of publication.

- 1 Marc Andreessen, “Why Software Is Eating the World,” Andreessen Horowitz, originally published in The Wall Street Journal, August 20, 2011.
- 2 AFCOM, 2026 State of the Data Center data as cited in “Power Is Not a Line Item,” June 17, 2026: average rack density of 27 kW, up from 16 kW the prior year.
- 3 NVIDIA Mission Control documentation: GB200 NVL72 designed rack power of approximately 120 kW.
- 4 International Energy Agency, Energy and AI: data centers consumed about 415 TWh in 2024; base-case consumption reaches about 945 TWh by 2030.
- 5 JLL, 2026 Global Data Center Outlook: average grid-connection wait in primary data-center markets exceeds four years.
- 6 OpenAI, “Jalapeño’s first results show industry-leading speed and efficiency in AI inference,” August 25, 2026. OpenAI reports 1.5 to 1.9x more AI work per watt at peak throughput across the public models tested.
- 7 Sarah Friar, OpenAI, “The full stack behind abundant intelligence,” August 25, 2026. OpenAI describes a strategy of preserving choice across providers, hardware, and deployment models while building first-party silicon for leverage.
- 8 Jensen Huang, “NVIDIA to Acquire Hugging Face,” NVIDIA Blog, September 3, 2026: \$12.93 billion; more than 18 million developers; platform to remain multi-cloud and multi-accelerator, with NVIDIA compute not required.
- 9 TechCrunch, “NVIDIA confirms it will buy Hugging Face for \$12.9 billion,” September 3, 2026. CNN Business, September 3, 2026, citing a securities filing: closing expected in the first half of 2027.
- 10 AWS, Trainium: purpose-built AI chips available through AWS services including EC2 and SageMaker.

- 11 Google Cloud, Tensor Processing Units: Google custom accelerators available through Cloud TPU.
- 12 Pat Gelsinger, “Inference Efficiency Accelerates at GTC,” LinkedIn, 2026: “we must make inference at least 10,000X more efficient.”
- 13 CNBC, “MediaTek jumps 10% after \$3.5 billion Nvidia AI chip deal,” September 1, 2026: NVIDIA purchases \$3.5 billion in MediaTek convertible bonds; MediaTek adopts NVLink Fusion for custom AI accelerators.
- 14 Bloomberg, “Anthropic Seals \$35 Billion Cloud Deal With NVIDIA-Backed Lambda,” August 31, 2026, following The Wall Street Journal: NVIDIA reportedly holds the lease on the Texas data center; Anthropic has reportedly signed capacity deals with several providers, including sites running Google TPUs.
- 15 Phoebe Liu, “How Nvidia Is Trying to Solve the Data Center Power Bottleneck,” The Information, September 20, 2026; NVIDIA Form 10-Q for the quarter ended July 26, 2026: supply and capacity commitments of approximately \$279 billion.

ABOUT THIS PAPER

This paper grew out of a question I kept returning to as I looked at the economics behind enterprise AI adoption: If intelligence becomes abundant, who controls the cost of producing it? The more I examined the infrastructure layer, the more convinced I became that the next AI divide will not simply be between those with and without access to AI. It will be between those able to follow a falling cost curve and those locked into one they cannot control.

A companion perspective, *The AI Diffusion Ecosystem*, places compute economics within the four conditions that allow intelligence to reach work.



June Manley

Technology and go-to-market executive and advisor

ABOUT THE AUTHOR

About June Manley

June Manley is a technology and go-to-market executive and advisor with more than two decades of experience across enterprise technology, strategic partnerships, ecosystems, and business transformation. Her work focuses on the commercial and operating implications of major technology shifts, including AI economics, infrastructure, sovereignty, and enterprise adoption.

More perspectives: junemanley.com

SEPTEMBER 2026

© 2026 June Manley. Executive Perspective.



Software ate the world.
Hardware must not starve
the dream.

The next AI divide will be defined less by who can
access intelligence than by who can keep its cost
curve falling.