ENDOGENOUS DECENTRALIZATION

How Concentrated Capital Investment Finances the Learning Curves That Enable Distributed Alternatives

Connor Smirl

Department of Economics, Tufts University

February 2026

Working Paper

Abstract

This paper identifies and formalizes endogenous decentralization: a mechanism by which concentrated capital investment in centralized infrastructure finances the learning curves that enable distributed alternatives. The mechanism's distinctive property is $\partial T^*/\partial I < 0$: increased centralized investment accelerates the crossing time at which distributed architectures become cost-competitive. Unlike Arrow's (1962) learning-by-doing, where cost reduction benefits the same production paradigm, endogenous decentralization produces architectural substitution—the learning investments finance a different organizational form.

Five contributions are new. First, I formalize the mechanism as a continuous-time differential game in which the distance to the crossing point is a common-pool state variable depleted by cumulative production. Competing centralized firms, each maximizing individual rents in symmetric Markov Perfect Equilibrium, produce aggregate output that strictly exceeds the cooperative optimum at every interior state, accelerating T^* beyond what any firm would individually prefer (Proposition 1). Second, the pure cost-parity crossing condition generalizes to a self-sustaining adoption threshold: the distributed ecosystem's basic reproduction number R_0 must exceed unity. Third, the

model incorporates a structural distinction between training and inference workloads. Fourth, cross-domain empirical analysis identifies the operative learning curve as *3D memory stacking and advanced packaging*—not planar DRAM die fabrication. The packaging learning curve ($\alpha = 0.23$, measured from HBM product-level data, 2015–2024) is early-stage and consistent with cross-technology learning rates. The effective crossing threshold is simultaneously being reduced from above through algorithmic efficiency gains driven by open-weight model developers operating under binding compute constraints imposed by semiconductor export controls. Fifth, the model generates nine falsifiable predictions with specific timing and failure conditions for the current AI infrastructure buildout ($1.3 trillion cumulative hyperscaler capex, 2018–2025).

Keywords: endogenous decentralization, learning curves, Markov Perfect Equilibrium, architectural substitution, AI infrastructure, training-inference bifurcation, open-weight models, basic reproduction number

JEL: O33, L16, D43, C73

1. Introduction

Between 2018 and 2025, the five largest US technology companies—together with Oracle and the Stargate joint venture—committed an estimated $1.3 trillion in cumulative capital expenditure to construct centralized AI infrastructure. This represents the largest concentrated infrastructure investment in history outside wartime mobilization. The near-term revenue objective is to sell AI inference—running trained models to serve user requests—as a cloud service at premium margins. A second, longer-horizon objective is frontier model training at scales that may produce discontinuous capability advances. The mechanism identified in this paper applies to the inference objective; the training objective is addressed as an alternative specification of the firms' objective function (Section 3.1).

This paper argues that this investment is *endogenously self-disrupting*: the very act of building centralized AI datacenters finances the component learning curves—particularly in 3D memory stacking, advanced packaging, and model compression—that enable distributed alternatives to replicate datacenter-class inference on consumer hardware. The operative learning curve is not the mature planar DRAM die, whose cost trajectory is near-asymptotic after four decades of cumulative production, but the *packaging and stacking* technologies that hyperscaler HBM demand is financing through their early high-learning-rate phase. As of Q1 2026, the technology threshold for interactive 70B-class inference has been met at professional and enthusiast price points. Paradoxically, the same concentrated investment has also triggered the most severe DRAM supercycle in two decades, temporarily reversing consumer memory cost trends and inflating GPU prices far above MSRP—a boom-phase deviation that the model's capacity-constraint corollary predicts will resolve into overcapacity and below-trend pricing as new advanced packaging capacity ramps. The remaining constraint is price migration from professional to mass-market form factors—a market structure transition compounded by, but not permanently altered by, the current supply shock.

Two structural features of the current AI landscape sharpen the mechanism beyond what prior transitions exhibited.

First, AI workloads bifurcate into *training* (creating models via massive synchronized GPU clusters) and *inference* (running models to serve user requests on independent, atomizable tasks). The endogenous decentralization mechanism applies directly and powerfully to inference, which already constitutes 80–90% of AI compute cycles. Training may remain permanently centralized—not because learning curves fail to reduce its costs, but because the synchronization and bandwidth requirements are architectural constraints that cost reduction alone cannot address. The post-crossing equilibrium is partial decentralization: inference distributes while training persists centrally.

Second, the effective crossing threshold is being approached from two directions simultaneously. From below, the packaging learning curve reduces the cost of delivering memory bandwidth to inference workloads along the trajectory this paper models ($\alpha = 0.23$). From above, algorithmic efficiency gains—mixture-of-experts architectures, aggressive quantization, and distillation—reduce the effective hardware requirement for a given inference capability level. These software-side gains are driven primarily by open-weight model developers operating under binding compute constraints: US semiconductor export controls deny these firms access to frontier datacenter GPUs, creating a structural incentive to maximize inference capability per unit of available hardware. The result is a dual convergence in which cumulative packaging production $Q(t)$ rises toward the crossing threshold while the threshold itself $\bar{Q}_{\text{eff}}(t)$ falls.

The contribution is five-fold. First, the formal mechanism: a continuous-time differential game with exact closed-form solutions. Second, a generalized crossing condition: $R_0 > 1$. Third, the training-inference bifurcation. Fourth, dual-convergence empirical evidence. Fifth, nine falsifiable predictions with timing. The paper is organized as follows. Section 2 develops the mechanism. Section 3 presents the formal model. Section 4 establishes the training-inference structural distinction. Section 5 presents the empirical evidence. Section 6 validates parameter consistency across historical transitions. Section 7 offers predictions. Section 8 concludes.

2. The Endogenous Decentralization Mechanism

2.1 Three-Stage Structure

Stage 1: Centralized Investment. Firms with market power invest $I(t)$ in centralized infrastructure to capture scale economies, producing cumulative component production $Q(t)$.

Stage 2: Component Cost Decline. Cumulative production drives unit costs along Wright's (1936) learning curve:

$$c(Q) = c_0 \cdot Q^{-\alpha} \tag{1}$$

where α is the learning elasticity. The critical property is that α is a *technology* parameter, not a *firm* parameter: learning embodied in manufacturing process improvements transfers across applications. A crucial refinement: for mature technologies (such as planar DRAM die fabrication), cumulative production is sufficiently large that marginal cost reductions per doubling are negligible. The mechanism's force depends on *new* production processes—specifically, 3D memory stacking and advanced packaging—that are in their early high-α phase. The packaging techniques developed for datacenter HBM (TSV interconnects, hybrid

bonding, die thinning, thermal management of stacked dies) transfer directly to consumer memory form factors.

Stage 3: Architectural Recombination. When component costs cross a threshold c^*, the same components can be recombined into distributed architectures exhibiting network externalities. Beyond a crossing time T^*, the distributed paradigm dominates for workloads amenable to distributed execution.

2.2 The Self-Undermining Investment Property

The mechanism's distinctive feature is that each stage causally enables the next, and the final stage undermines the first. Define T^* as the first date at which distributed architecture cost-performance matches centralized provision for the marginal inference user. Then:

$$\frac{\partial T^*}{\partial I} < 0 \tag{2}$$

Increased centralized investment accelerates displacement of the centralized paradigm's inference revenue.

2.3 Dual Convergence

The current AI transition exhibits a feature absent from prior technological transitions: the effective crossing threshold is being approached from two directions simultaneously.

From below: the packaging learning curve. The cost of delivering memory bandwidth to inference workloads is driven by 3D stacking and advanced packaging, not by planar DRAM die fabrication. The die cost—historically the dominant component—is near-asymptotic: DRAM is among the highest-cumulative-volume semiconductor products ever manufactured. The packaging cost, by contrast, is in its early high-learning-rate phase: volume production of TSV-based stacked memory began circa 2015, and the learning curve ($\alpha = 0.23$ from HBM product-level data) is consistent with early-stage technologies across domains.

From above: algorithmic efficiency gains. Advances in model architecture and compression reduce the hardware *required* to achieve a given inference capability level. Mixture-of-experts (MoE) architectures activate only a fraction of total parameters per token, reducing effective memory bandwidth requirements by 3–6×. Quantization (INT4, INT2) reduces model memory footprint by 4–16×. Distillation transfers capability from large models to smaller ones.

Define $\bar{Q}_{\text{eff}}(t) = \bar{Q} \cdot f(\eta(t))$, where $\eta(t)$ indexes cumulative algorithmic efficiency gains and f is decreasing. The state variable becomes $x(t) = \bar{Q}_{\text{eff}}(\eta(t)) - Q(t)$, and the rate of

depletion exceeds what hardware learning curves alone would predict.

2.4 Distinction from Adjacent Theory

Table 1 summarizes the positioning. The distinctions are precise: Arrow's (1962) learning-by-doing benefits the same paradigm; Bresnahan and Trajtenberg's (1995) GPT spillovers enable applications across sectors rather than architectural self-replacement; Schumpeter's (1942) creative destruction comes from external entrants.

Table 1: Theoretical positioning of endogenous decentralization.

Framework	Learning Scope	Beneficiary	Disruption Source	Self-Undermining?
Arrow (1962)	Same paradigm	Same firms	N/A	No
Bresnahan-Trajtenberg (1995)	Cross-sector	Other sectors	External applications	No
Schumpeter (1942)	External	Entrant firms	External entrant	No
Christensen (1997)	Cross-market	Entrant firms	New value network	Partial
This paper	**Cross-paradigm**	**Different architecture**	**Self-financed**	**Yes**

3. Formal Model

3.1 Environment

Consider $N \geq 2$ symmetric centralized firms indexed by $i \in \{1, \ldots, N\}$. Time is continuous. The *state variable* is $x(t) = \bar{Q}_{\text{eff}} - Q(t) \in [0, x_0]$, measuring the remaining cumulative production until the effective crossing threshold at which distributed architecture becomes cost-competitive for inference workloads. When x reaches zero, inference crossing occurs. The state evolves as:

$$dx/dt = -\sum_i q_i(t) \tag{3}$$

where $q_i(t) \geq 0$ is firm i's output rate. Each unit of output serves the centralized market and simultaneously depletes the remaining distance to crossing—this dual role is the formal expression of the self-undermining investment property.

Flow profits for firm i are determined by linear inverse demand $P = a - bQ$, where $Q = \sum q_j$ is total output rate:

$$\pi_i(t) = (a - bQ)q_i \tag{4}$$

with $a > 0$, $b > 0$. Upon crossing ($x = 0$), each firm receives continuation value:

$$S = S_T + \frac{S_I}{N(r + \delta)} \tag{5}$$

where S_T represents the persistent training and model-licensing revenue that survives inference decentralization, $S_I = \bar{\pi}_I$ is the pre-crossing inference profit level, r is the discount rate, and $\delta > 0$ is the post-crossing inference displacement rate.

Remark on the objective function. The model assumes firms maximize discounted revenue from infrastructure services. An alternative specification treats centralized investment as purchasing an option on a discontinuous payoff: the first firm to achieve a capability threshold captures a prize V^* that dwarfs cumulative investment. Under this specification, S_T is the option value of retaining frontier training *capability*. The mechanism's core result ($\partial T^*/\partial I < 0$) is invariant to the firms' objective. The revenue-maximization model provides a *lower bound* on aggregate investment and a correspondingly conservative estimate of T^*. Section 3.7 calibrates both specifications.

The game has a *common-pool* structure: the state x is a shared resource (remaining time before inference disruption) that all firms deplete through production. This structure is analogous to the fishery or oil extraction commons (Levhari and Mirman 1980), with the critical distinction that the "resource" being depleted is the incumbent paradigm's remaining inference viability.

3.2 Markov Perfect Equilibrium

I restrict attention to symmetric stationary Markov strategies $q_i = q(x)$. Each firm's value function $V(x)$ satisfies the Hamilton-Jacobi-Bellman equation:

$$rV(x) = \max_{q_i}\{(a - b(q_i + (N-1)q(x)))q_i - V'(x) \cdot (q_i + (N-1)q(x))\} \tag{6}$$

The first-order condition under symmetry yields the equilibrium strategy:

$$q^N(x) = \frac{a - V^{N\prime}(x)}{b(N+1)} \tag{7}$$

Substituting back into the HJB yields the ODE:

$$rV^N(x) = \frac{(a - V^{N\prime}(x))(a - N^2 V^{N\prime}(x))}{b(N+1)^2} \tag{ODE-N}$$

with boundary condition $V^N(0) = S$.

3.3 Cooperative Benchmark

The cooperative planner maximizes total producer surplus $W(x) = NV^P(x)$, choosing total output rate Q:

$$rV^P(x) = \frac{(a - NV^{P\prime}(x))^2}{4bN} \tag{ODE-C}$$

with boundary condition $V^P(0) = S$.

3.4 Analytical Solutions

Both ODEs are autonomous and separable. The cooperative ODE yields the exact implicit solution:

$$x(V) = \frac{a \cdot \ln\left(\frac{a - 2\sqrt{bnrS}}{a - 2\sqrt{bnrV}}\right) + 2\left(\sqrt{bnrS} - \sqrt{bnrV}\right)}{2br} \tag{C-exact}$$

The Nash ODE is solved by the substitution $u = \sqrt{D + EV}$:

$$x(V) = \frac{4N^2}{E}\left[(u_0 - u) + A \cdot \ln\left(\frac{A - u_0}{A - u}\right)\right] \tag{N-exact}$$

Both solutions share the same functional form—$\sqrt{\cdot} + \log$—differing only in the constants governing shadow cost internalization. Both are verified to machine precision ($\max |x_{\text{exact}} - x_{\text{num}}| < 10^{-12}$).

3.5 The Overinvestment Result

Proposition 1 (Overinvestment in Markov Perfect Equilibrium)**.** *In the symmetric MPE, aggregate output $Q^N(x) = Nq^N(x)$ strictly exceeds cooperative output $Q^C(x)$ for all $x > 0$. Consequently, $T^{*,Nash} < T^{*,Coop}$: Nash equilibrium crossing occurs strictly earlier than the cooperative optimum.*

Proof. Step 1. At $x = 0$, $V^N(0) = V^P(0) = S$. Evaluating the boundary derivatives from (ODE-N) and (ODE-C), the planner's total shadow cost $N\mu$ strictly exceeds the Nash firm's private shadow cost λ for $N \geq 2$. This gap reflects the learning externality: each Nash firm internalizes only its own future profit loss from approaching crossing.

Step 2. By a standard comparison theorem for ODEs (Walter 1998, Theorem I.9.1), the ordering $N \cdot V^{P\prime}(x) > V^{N\prime}(x)$ propagates to all $x > 0$.

Step 3. From the output expressions, both the smaller numerator (higher shadow cost) and larger denominator of Q^C relative to Q^N ensure $Q^N(x) > Q^C(x)$ for all $x > 0$. □

Remark 1 (Irreversibility)**.** *At $Q = \bar{Q}$, a new basin of attraction—the distributed inference equilibrium—becomes accessible. Reversing the crossing would require cumulative production to decrease—which contradicts monotonicity. Once Q crosses $\bar{Q}$, the inference transition is topologically irreversible.*

Remark 2 (Niche Persistence)**.** *Irreversibility of inference crossing does not imply extinction of the centralized paradigm. IBM's mainframe business continues to generate approximately $3–4 billion annually as of 2025—decades after the PC revolution—serving high-reliability transaction processing.*

Economic interpretation. The overinvestment decomposes into a Cournot channel (price-depressing rival output) and a learning externality channel (private shadow cost $= 1/N$ of social shadow cost). The decomposition of S into $S_T + S_I/(N(r+\delta))$ reveals a moderating effect: when S_T is large, crossing is less catastrophic and the overinvestment gap narrows.

Welfare loss. At baseline calibration ($N = 5$, $S_T = 0$), the per-firm welfare loss under Nash competition is 34.1%. With S_T calibrated to estimated training revenue persistence, the loss moderates to approximately 22–28%.

3.6 Comparative Statics

Corollary 1 (Increasing N)**.** *Nash equilibrium aggregate output is strictly increasing in N for all $x > 0$.*

Corollary 2 (Asymmetric firms)**.** *If firm 1 has marginal cost $c_1 - \varepsilon$, aggregate equilibrium output is strictly increasing in ε.*

Corollary 3 (Asymmetric crossing valuation)**.** *If firm j has post-crossing value $S_j > S$, firm j produces strictly more than symmetric competitors, and aggregate output increases.*

Corollary 4 (Capacity constraint and boom-bust)**.** *Crossing time delay is bounded by the construction lag Δ for new capacity. The long-run packaging learning rate α is unaffected.*

The 2025–26 DRAM supercycle provides a real-time test. Consumer DDR5 prices have risen 300–400% above trend in under six months, driven by AI datacenter demand reallocating wafer capacity from consumer to HBM formats. The corollary predicts that (a) this

deviation is temporary, bounded by the construction lag for new advanced packaging capacity (Samsung P4, SK Hynix M15X, Micron Idaho, TSMC CoWoS expansion), and (b) the packaging learning rate $\alpha = 0.23$ is unaffected because the supercycle is a *demand allocation* shock, not a change in the stacking production function.

Remark 3 (Option-value amplification)**.** *Under the option-value objective function specification (Section 3.1), the overinvestment result is amplified. If firms invest to maximize the probability of achieving a discontinuous capability threshold, the marginal value of additional investment is governed by the prize V^* rather than by discounted market revenue. The model's quantitative predictions ($Q^N/Q^C \approx 3$–$4\times$, $T^* \approx 2028$) are then conservative.*

3.7 Calibration

The learning elasticity $\alpha = 0.23$ is estimated from the HBM packaging learning curve (Table 7), which measures the cost trajectory of 3D-stacked memory from first volume production (HBM1, 2015) through the current generation (HBM3E+, 2025). This estimate captures the relevant production process—through-silicon via (TSV) interconnects, die thinning, hybrid bonding, and thermal management—rather than the mature planar DRAM die (see Section 5.1 for the cost decomposition). Current HBM cost is approximately \$12/GB (HBM3E, 2025); the crossing threshold is \$5–7/GB. The calibration uses the conservative bound $\bar{Q} \approx 112$ EB (\$5/GB target).

Sensitivity of T^* to α. The model's timing predictions are sensitive to the learning elasticity. Table 2 reports T^* across the range of estimates in the literature, holding other parameters at baseline.

Table 2: Sensitivity of crossing time to learning elasticity.

α	Source / Label	T^* (yrs from 2024)	Calendar Year
0.12	Goldberg et al. (2024) w/ spillovers	93	2117
0.15	Conservative lower bound	74	2098
0.20	Irwin & Klenow (1994) canonical IV	56	2080
0.23	HBM packaging curve (baseline)	47	2071
0.25	Upper Irwin & Klenow range	45	2069
0.32	Irwin & Klenow OLS (likely biased up)	35	2059

Notes: T^* computed from hardware learning curve only, without algorithmic efficiency gains. Dual convergence (Section 5.2) shifts all dates earlier.

Post-crossing continuation value. The inference displacement rate $\delta \approx 0.30$ from the IBM trajectory (Section 6.1). Under revenue-maximization: S_T high (closed-model dominance), welfare loss ∼22%; S_T moderate (open-weight competition), ∼28%; $S_T \approx 0$

(commoditization), ∼34%. Under the option-value specification, S_T represents the option value of maintaining frontier training capability at scales no distributed architecture can replicate. The two specifications bracket the range of outcomes.

Quantitative predictions. Under Nash competition with $N = 5$, crossing at approximately 2028. The 2025–26 DRAM supercycle delays the cost threshold by an estimated 1–2 years during the boom phase, with potential acceleration during the subsequent bust. Under cooperation, ∼2042. Competition accelerates by 79%.

3.8 Note on Identification

The packaging learning curve is estimated by OLS regression of log cost on log cumulative output for HBM generations (Table 7). This identifies a correlation, not necessarily a structural learning-by-doing parameter. Endogeneity concerns (demand shocks driving both output and investment in cost reduction) are standard in the learning-curve literature (Irwin and Klenow 1994).

No published IV estimate exists for the packaging learning curve. The $\alpha = 0.23$ is identified from product-level HBM pricing that bundles die and packaging costs, with $n = 6$ generation-level observations—too few for formal structural estimation. This paper's empirical contribution is identifying *which* curve matters (early-stage packaging, not asymptotic die fabrication), not claiming precise estimation of its slope. The estimate's reliability rests on three indirect supports: cross-technology consistency of $\alpha \approx 0.21$–0.24 across independently estimated early-stage curves (Table 9); the physical cost decomposition showing packaging as the majority cost component at current HBM price points (Table 6); and the early-stage character of the process, where limited demand-side feedback reduces simultaneous-equations bias relative to the 41-year DRAM die series. The self-undermining property ($\partial T^*/\partial I < 0$) requires only that centralized investment contributes to cumulative Q and that $c(Q)$ is decreasing and stable. Refining the packaging α with firm-level production data as it accumulates is a natural next step.

Irwin and Klenow (1994) provide the most rigorous causal estimate for semiconductor learning: $\alpha = 0.32$ (SE = 0.05) using instrumental variables on a firm-level DRAM panel (1974–1992). Goldberg et al. (2024) estimate learning rates at the firm-technology-node level for microprocessor fabrication, finding $\alpha = 0.05$ at the firm-node level, rising to $\alpha = 0.12$ when cross-border spillovers are included. The model's $\alpha = 0.23$ is thus an *industry-level spillover-inclusive* estimate, consistent with the Goldberg et al. framework when cross-application spillovers are the dominant channel.

3.9 Generalized Crossing Condition

The model defines crossing at cost parity, but empirical evidence shows hardware crossing *precedes* architectural dominance by 3–5 years (Section 6.4). What is actually required is that the distributed ecosystem's basic reproduction number exceeds unity:

$$R_0 \equiv \frac{\beta(c, \lambda) \cdot \gamma}{\kappa + \mu} > 1 \tag{8}$$

where $\beta(c, \lambda)$ is the adoption rate, γ is the network effect multiplier, κ is the coordination friction, and μ is the churn rate. The latency advantage λ captures a structural property specific to inference: edge devices achieve <10ms response versus 50–200ms for cloud round-trip.

$R_0 = 1$ maps to a modified cumulative production threshold:

$$\bar{Q}^* = \bar{Q} \cdot \left(1 - \frac{\kappa + \mu}{\gamma \cdot c^*} + \frac{\lambda}{c^*}\right)^{-1/\alpha} \tag{9}$$

Replace $\bar{Q}$ with $\bar{Q}^*(\kappa)$ throughout the model. All propositions carry through identically.

Table 3: Coordination layer lag across transitions.

Transition	Hardware T^*	R_0 T^*	ΔT
Mainframe → PC	1987	1990–92	3–5 yr
ARPANET → Internet	∼1989	1993–94	4–5 yr
Cloud → Edge AI	2027–29[†]	?	2–3 yr (pred.)

[†] Hardware capability threshold met at professional price points Q1 2026; consumer cost threshold delayed by 2025–26 DRAM supercycle (Corollary 4).

4. The Training-Inference Structural Distinction

The $1.3 trillion in centralized AI infrastructure investment builds capacity for two structurally distinct workloads. Conflating them overstates the mechanism's scope; separating them sharpens it.

4.1 Two Workloads, Two Architectures

Training teaches models by processing massive datasets across tightly synchronized GPU clusters. It requires 10,000–100,000+ GPUs communicating at terabits per second via InfiniBand or NVLink, running continuously for weeks to months. Power density: 100–1,000

kW/rack. Latency-insensitive.

Inference runs trained models to serve real-time user requests. Tasks are independent and atomizable. Latency-sensitive: users benefit from <10ms local execution versus 50–200ms cloud round-trip. Frequency: continuous, scales with every user and query.

Table 4: Training vs. inference structural comparison.

Dimension	Training	Inference
Share of AI compute (2025)	∼50%	∼50%
Share of AI compute (2026, proj.)	∼33%	∼67%
Synchronization requirement	Massive (10K+ GPUs)	None (atomizable)
Latency sensitivity	Low	High (<10ms for UX)
Cost trajectory	Rising per frontier model	Declining ∼280× in 2 yr
Edge-viable?	No (architectural)	Yes (this paper's thesis)

Sources: Deloitte (2025), McKinsey (2025), MIT Technology Review (2025), Epoch AI (2025).

4.2 Training Does Not Decentralize

Frontier model training requires synchronized clusters of 10,000–100,000+ GPUs communicating at terabit-per-second speeds. A consumer device with excellent memory bandwidth cannot participate in a distributed training run because the inter-device communication latency (milliseconds over WiFi vs. nanoseconds over NVLink) creates a performance gap of 5–6 orders of magnitude. No plausible learning curve closes this gap because the constraint is topological (network diameter and synchronization protocol) rather than cost-based.

4.3 Inference Decentralizes

Inference tasks are *atomizable*: each user query is independent. Inference is *latency-advantaged*: local execution outperforms cloud round-trip on a quality dimension independent of cost. Inference is *bandwidth-bound*: token generation speed is determined almost entirely by the ratio of memory bandwidth to model size—exactly the constraint whose packaging learning curve the model tracks. Inference *scales with users*.

4.4 The Inference Revenue Pool

Inference dominates both compute cycles (80–90%) and ongoing revenue. The inference market is projected to grow from $106 billion (2025) to $255 billion by 2030 (MarketsandMarkets 2025). This is the revenue pool that the $1.3 trillion infrastructure buildout targets, and the revenue pool that edge devices will intercept.

4.5 Implications for the Model

5. Empirical Evidence: Dual Convergence

The inference crossing condition—≥70B-class output quality at ≥20 tok/s under $1,500—is being approached from two directions: hardware costs declining from below (Section 5.1) and algorithmic efficiency reducing the effective threshold from above (Section 5.2).

5.1 Convergence from Below: Hardware Cost Decline

5.1.1 Cost Decomposition: Die versus Packaging

The cost of delivering memory bandwidth to an inference workload decomposes into three components with distinct learning dynamics:

Die fabrication (mature, $\alpha \to 0$). Planar DRAM die cost per bit has declined along the Wright curve for over four decades—from $870,000/GB (1984) to approximately $2/GB (2024). At current cumulative production levels (∼3,200 EB through 2024), additional doublings yield marginal cost reductions. A 41-year OLS regression yields $\alpha = 0.66$ (SE = 0.04), but this estimate is inflated by simultaneous equations bias, product-generation transitions, and demand-side shocks (Irwin and Klenow 1994). Piecewise regression identifies structural breaks at 1995 and 2008, with the middle regime (1995–2007) yielding an implausible $\alpha = 1.15$. The bookend regimes yield $\alpha = 0.38$–0.39 (OLS), consistent with the Irwin and Klenow IV estimate of 0.32 after accounting for upward OLS bias. For the purposes of this paper, the critical observation is that the die cost is no longer the binding constraint or the operative learning curve.

3D stacking and advanced packaging (early-stage, $\alpha = 0.23$). This is the operative learning curve. Volume production of TSV-based stacked memory began with HBM1 in 2015. The techniques involved—through-silicon via drilling and filling, die thinning to $<50\mu$m, hybrid bonding for sub-2μm pitch interconnects, thermal management of multi-die stacks—are in their first decade of high-volume manufacturing. The critical property for the endogenous decentralization mechanism is that the packaging knowledge developed for datacenter HBM transfers directly to consumer memory form factors. Samsung and SK Hynix engineers solving yield problems on HBM4 stacking are generating process knowledge that flows to consumer product lines within the same companies. This is not abstract spillover—it is traceable intra-firm technology transfer through shared packaging R&D and manufacturing infrastructure.

System integration (declining with ecosystem maturity). PCB design, thermal management, power delivery, and firmware optimization. This component is declining but

Table 5: DRAM die cost trajectory (selected years).

Year	Generation	$/GB	Cum. Prod. (EB)	ln(Price)	ln(Cum.)
1984	64Kb	870,000	<0.001	13.68	−11.51
1990	4Mb	100,000	0.003	11.51	−5.81
1995	16Mb	30,000	0.10	10.31	−2.30
2000	256Mb	1,200	2.0	7.09	0.69
2005	1Gb	90	17	4.50	2.83
2010	2Gb	10	95	2.30	4.55
2015	8Gb	3.20	400	1.16	5.99
2020	16Gb	2.80	1,400	1.03	7.24
2024	32Gb	2.00	3,200	0.69	8.07
2025–26	32Gb†	10–16	∼4,200	2.30–2.77	8.34

OLS through 2024: $\alpha = 0.66$ (SE = 0.04), $R^2 = 0.96$. Piecewise: structural breaks at 1995 and 2008 (Bai-Perron). Regime 1 (1984–94): $\alpha = 0.39$. Regime 2 (1995–2007): $\alpha = 1.15$, implausible. Regime 3 (2008–24): $\alpha = 0.38$. Carlino et al. (2025) find structural breaks in 66% of technology learning curves; the DRAM die series is consistent with this pattern. †Supercycle pricing reflects demand allocation, not production cost.

not modeled explicitly.

Table 6: Approximate cost decomposition: memory bandwidth delivery ($/GB).

Component	HBM3E (2025)	Consumer DDR5 (2024, pre-cycle)	Consumer DDR5 (2026, supercycle)	Proj. consumer stacked (2029)
Die fabrication	∼3–4	∼1.50	∼1.50–2.00	∼1.00–1.50
Packaging & stacking	∼6–8	∼0.30 (planar)	∼0.30–0.50	∼1.50–2.50 (3D)
System integration	∼2	∼0.20	∼0.20–0.50	∼0.50–1.00
Total	**∼12**	**∼2.00**	**∼10–16**†	**∼3–5**

† Supercycle pricing reflects demand allocation, not production cost. Consumer stacked memory (2029) reflects post-boom pricing with packaging learning at $\alpha = 0.23$ and new capacity online.

5.1.2 The Packaging Learning Curve: HBM Cost Trajectory

HBM prices declined from $120/GB (2015) to $12/GB (2025). $\alpha = 0.23$ (SE = 0.06, $n = 6$). The packaging knowledge transfers to consumer form factors—the learning externality central to the mechanism.

The investment scaling behind this curve is concrete. TSMC's CoWoS advanced packaging capacity is growing at a >50% CAGR from 2022 to 2026 (Jun He, TSMC VP of Advanced Packaging, 2025), ramping from approximately 35,000 wafers/month (2024) to 75,000 (end 2025) to a target of 130,000 (end 2026). Total industry CoWoS demand is projected at 1 million wafers in 2026, up from 370,000 in 2024 (Morgan Stanley 2026). HBM yields cur-

Table 7: HBM packaging learning curve.

Year	Generation	$/GB	Cap./Stack (GB)	Stacking Technology
2015	HBM1	120	4	4-high TSV, 1024-bit
2016	HBM2	60	8	4-high TSV, improved yield
2018	HBM2E	35	8	8-high TSV
2020	HBM2E	25	16	8-high, die thinning
2022	HBM3	20	24	8-high, 2048-bit interface
2024	HBM3E	15	36	8-high, hybrid bonding
2025	HBM3E+	12	48	12-high, advanced thermal

$\alpha = 0.23$ (SE $= 0.06$). Estimated from log($/GB) regressed on log(cumulative HBM units shipped).

rently range from 50–60% (TrendForce 2025), indicating that the steep portion of the yield learning curve remains ahead. This is the packaging investment the model tracks—capacity tripling in two years on a process whose yields have not yet matured.

The learning rate $\alpha = 0.23$ is estimated from a short series ($n = 6$ generation-level data points, 2015–2025). The standard error (0.06) reflects this limited sample. However, three features support the estimate's reliability: (a) the cross-technology consistency documented in Table 9; (b) the estimate falls in the range expected for early-stage process technologies; and (c) the HBM series is less susceptible to simultaneous equations bias than the aggregate DRAM die series because HBM volumes are driven primarily by datacenter demand with limited consumer feedback.

5.1.3 Hyperscaler Capital Expenditure

Table 8: Hyperscaler capex ($B).

Company	2018	2020	2022	2024	2025E
Microsoft	11.6	15.4	23.9	44.5	80
Alphabet	25.1	22.3	31.5	52.5	75
Amazon	13.4	35.0	58.3	78.0	100
Meta	13.9	15.7	31.4	39.2	65
Stargate JV	—	—	—	—	100
Industry Total	**64**	**88**	**148**	**232**	**436**

Cumulative 2018–2025: $1,298B. Sources: company filings and guidance.

A significant fraction of this capex flows directly to the packaging learning curve: each NVIDIA H100/H200/B200 GPU contains multiple HBM stacks, each requiring TSV processing, die thinning, and advanced packaging. The Stargate project alone is estimated to demand approximately 40% of global HBM output.

5.1.4 Consumer Silicon and the Inference Crossing Condition

Token generation speed for inference is determined almost entirely by the ratio of memory bandwidth to model size in memory, making memory bandwidth the binding constraint. Four tiers of consumer and professional AI silicon now reveal both the trajectory and the constraint's shift from technology to market structure.

Edge tier. Rockchip's RK1828 (2025, 20 TOPS, 5GB 3D stacked DRAM co-processor) runs 7B-parameter models at 59 tok/s—a direct application of packaging techniques developed for HBM. Hailo's 10H (2025, 40 TOPS INT4, 2.5W) on the Raspberry Pi AI HAT+ at $130 runs 2B-parameter models at 10+ tok/s.

Consumer desktop tier. AMD's Ryzen AI Max+ 395 ("Strix Halo," ∼$2,000, 128GB unified LPDDR5X, ∼215 GB/s). MoE architectures with ∼20B active parameters achieve ∼31 tok/s at interactive speeds.

Discrete GPU tier. NVIDIA's RTX 5090 (Q1 2026, 32GB GDDR7, ∼1,792 GB/s, $1,999 MSRP) exceeds the speed threshold on models that fit in 32GB. However, street prices range $3,000–$5,000+ due to the DRAM supercycle. Memory now accounts for an estimated 80% of GPU BOM cost, up from ∼30–40% pre-shortage.

The gap is now three constraints, not one. (1) The segmentation premium on memory capacity, which is structural; (2) the supercycle premium on memory cost, which is cyclical; and (3) supply rationing, which is strategic. Constraints (2) and (3) are temporary. The packaging capacity expansion will, on historical precedent, produce overcapacity and below-trend pricing within 2–3 years of full ramp. The pivoting asset is primarily *advanced packaging capacity*—CoWoS and TSV lines designed for HBM, which will be available for consumer stacked memory when datacenter demand moderates.

5.2 Convergence from Above: Algorithmic Efficiency

5.2.1 The Incentive Structure

A binding compute constraint on a subset of model developers creates a structural incentive to maximize inference capability per unit of available hardware. US semiconductor export controls, beginning October 2022, denied a significant population of AI developers access to frontier datacenter GPUs. The theoretical prediction is that constrained firms should optimize for efficiency and pursue deployment strategies compatible with available hardware—including edge devices.

5.2.2 Scale and Adoption

Total downloads of the Qwen model family (Alibaba) exceeded 700 million on Hugging Face by January 2026. By August 2025, Qwen-derived models accounted for over 40% of all new Hugging Face language model derivatives (Lambert 2025). An empirical study of 100 trillion tokens processed through the OpenRouter aggregator found open-weight model share surging from 1.2% to peaks of ∼30% of weekly token volume within months (OpenRouter/Andreessen Horowitz 2025).

5.2.3 Mechanisms Reducing the Effective Crossing Threshold

Mixture-of-Experts (MoE). MoE architectures activate only a fraction of total parameters per token. DeepSeek V3 (671B total, ∼37B active) demonstrates that 70B-class output quality is achievable with 20–37B active parameters, reducing memory bandwidth required per token by 3–6×.

Quantization. INT4 quantization reduces model memory footprint by approximately 4× with modest quality loss.

Distillation. DeepSeek's distilled models (1.5B, 7B, 14B variants of R1) explicitly target edge deployment, maintaining reasoning capability at dramatically reduced hardware requirements.

The combined effect: Stanford's 2025 AI Index documented a 280-fold drop in inference costs between November 2022 and October 2024. The paper's $\alpha = 0.23$ captures the packaging learning curve alone; the effective cost decline including algorithmic optimization is significantly steeper.

5.3 The Demand Shock as Nash Overinvestment

The Stargate project alone demands approximately 40% of global DRAM output. Historical precedent predicts overcapacity and below-trend pricing by 2028–2029. The packaging lines built for datacenter HBM demand will pivot to consumer stacked DRAM and LPDDR6 when datacenter demand moderates—accelerating the very edge inference capability that drives the moderation. This is the Nash overinvestment dynamic operating through the supply side.

The 2025–26 DRAM supercycle is Corollary 4 operating through a novel channel: the boom phase temporarily *reverses* the consumer cost trajectory even as it finances the packaging capacity expansion that will eventually crash consumer prices below the pre-boom trend. The resolution is temporal: the boom phase adds 1–2 years to the hardware crossing

timeline, but the bust phase may compress the post-bust crossing timeline by a comparable amount.

6. Historical Validation and Parameter Consistency

6.1 Mainframe → Personal Computer (1975–2000)

IBM dominated mainframe computing with 75–80% market share through the 1970s. IBM's semiconductor investment drove the learning curves that reduced microprocessor and memory costs (Flamm 1993: $\alpha = 0.24$ for Intel microprocessors, 1974–1989). IBM's cumulative losses of $15.8B (1991–93) reflected a business model adaptation failure, not technology extinction. IBM's mainframe division persists today at $3–4B annual revenue. The $\delta \approx 0.30$ calibration: IBM lost ∼60% of its compute-service profit in three years.

6.2 ARPANET → Commercial Internet (1969–2000)

Government investment drove TCP/IP development and router cost reduction. Proprietary online service share collapsed from 60% (1989) to 2% (2000). Coordination layer lag: 3–5 years.

6.3 The Export-Control Natural Experiment

The October 2022 US semiconductor export controls provide an exogenous shock. Constrained developers converged on: (i) open-weight release strategies, (ii) efficiency-optimizing architectures (MoE), (iii) aggressive quantization and distillation, and (iv) model-hardware co-optimization for edge deployment. The resulting ecosystem functions as an exogenous accelerator of Stage 3: it reduces $\bar{Q}_{\text{eff}}$ from above, reduces κ through pre-crossing coordination layer construction, and potentially erodes S_T by commoditizing training output.

6.4 Cross-Domain Parameter Consistency

7. Falsifiable Predictions

The model generates nine predictions with timing. If these fail, the theory is wrong.

Prediction 1: Consumer Stacked Memory ≥16GB by 2027. HBM-derived 3D stacking in consumer products with ≥16GB on-chip stacked memory below $200. The Rockchip RK1828 (2025, 5GB 3D stacked) and Hailo-10H (2025, 8GB on-module at $130) confirm packaging technology migration is underway. Evidence against: ≤8GB through 2028.

Table 9: Cross-domain learning rates.

Industry	Product	α	SE	Period	Source
Semiconductor	HBM (3D stacking)	0.23	0.06	2015–2024	TrendForce
Semiconductor	NAND Flash	0.24	0.05	2003–2023	Micron/Samsung
Semiconductor	Intel microprocessors	0.24	0.04	1974–1989	Flamm (1993)
Semiconductor	DRAM (IV, causal)	0.32	0.05	1974–1992	Irwin & Klenow
Semiconductor	Microproc. (w/ spillovers)	0.12	—	2004–2015	Goldberg et al.
Energy	Solar PV cells	0.23	0.02	1976–2023	IRENA
Energy	Lithium-ion batteries	0.21	0.03	1995–2023	BloombergNEF
Internet	Cloud compute (AWS)	0.25	0.03	2006–2023	AWS pricing

Cross-technology central tendency: $\alpha \in [0.21, 0.25]$ for industry-level spillover-inclusive estimates. Goldberg et al.'s lower firm-node estimate measures learning within a single process at a single facility.

Prediction 2: 70B-Class Inference On-Device by 2028–2029 (Hardware Crossing). Consumer devices under $1,500 running inference at 70B-class output quality at $\geq$20 tok/s. As of Q1 2026, the technology capability threshold has been met at professional price points, but the DRAM supercycle has temporarily inflated consumer memory costs 300–400% above trend. Evidence against: not achieved by 2031.

Prediction 2* (Refined): $R_0 > 1$ for Distributed AI Inference by 2030–2032. Self-sustaining distributed inference adoption arrives 2–3 years after hardware crossing. Evidence against: distributed share stalling below 20% by 2033.

Prediction 3: Inference Capex Deceleration with Training Persistence by 2028–2029. At least one top-four US hyperscaler reduces inference-oriented capex by $\geq$20% YoY while maintaining or increasing training-oriented capex.

Prediction 4: Stablecoin-Treasury Holdings Exceed $300B by 2027. Tests coordination layer formation for distributed economic settlement. Evidence against: plateau below $200B.

Prediction 5: Packaging Learning Rate Stability. The 3D stacking / advanced packaging learning elasticity, measured by cost per GB of HBM and consumer stacked memory against cumulative stacked-memory shipments, remains in $[0.18, 0.28]$ through 2030. Evidence against: a rolling 3-year α estimate falling below 0.15 and not reverting within two years of the supercycle's resolution. If packaging $\alpha < 0.15$, all timing predictions shift outward. The prediction is framed around the packaging curve; structural breaks in the mature DRAM die series (Bai-Perron breakpoints at 1995 and 2008; Carlino et al. 2025) are irrelevant to this test.

Prediction 6: Distributed Inference Tipping Point at ~40% of Inference Workloads. Network effects reverse at ~40% distributed inference share. Evidence against: centralized inference remaining commercially stable with distributed share exceeding 50%

through 2032.

Prediction 7: Non-Monotonic Inference Adoption with Coordination-Layer Trough. Two-wave pattern: initial surge (2027–2030), coordination fragmentation trough (2031–2032), standardization-driven second wave (2033–2035).

Prediction 8: Open-Weight Models Exceed 50% of Global Inference Token Volume by 2028. Evidence against: proprietary closed models maintaining >60% of inference token volume through 2029.

Prediction 9: Training Remains Centralized Through 2035. Frontier model training (>10,000 synchronized GPUs, >7 days continuous operation) remains exclusively performed in centralized clusters. Evidence against: distributed frontier training at comparable cost and performance by 2035.

8. Conclusion

This paper has identified and formalized endogenous decentralization: a mechanism by which concentrated capital investment in centralized infrastructure finances the learning curves that enable distributed alternatives. The self-undermining investment property ($\partial T^*/\partial I < 0$) is distinct from learning-by-doing, GPT spillovers, and Schumpeterian creative destruction.

The mechanism operates through two convergence paths. Hardware cost decline from below follows the *packaging* learning curve—the early-stage trajectory of 3D memory stacking and advanced packaging ($\alpha = 0.23$), not the near-asymptotic planar DRAM die. The critical distinction is that planar DRAM die fabrication, after four decades of cumulative production, yields marginal cost reductions per doubling, while the packaging technologies being financed by hyperscaler HBM investment—TSV, hybrid bonding, die thinning, thermal management of stacked dies—are in their first decade of high-volume manufacturing, where Wright's law operates at its steepest. The technology transfer channel is concrete and traceable: packaging process knowledge developed for datacenter HBM migrates to consumer stacked memory within the same firms. Algorithmic efficiency gains from above reduce the effective crossing threshold through MoE architectures, quantization, and distillation, driven by developers operating under binding compute constraints.

The 2025–26 DRAM supercycle illustrates both the mechanism's predictions and its non-monotonic short-run dynamics: the same concentrated investment that finances long-run packaging learning curves has temporarily reversed consumer cost trends by reallocating memory and packaging capacity to datacenter formats—exactly the boom-phase deviation Corollary 4 predicts. The advanced packaging capacity expansion this demand shock has triggered will, on historical precedent, produce overcapacity and below-trend consumer pric-

ing within 2–3 years. The operative learning curve is the packaging process, not the die; and the capacity being expanded is packaging capacity that will pivot to consumer stacked memory formats.

The training-inference bifurcation sharpens the mechanism's empirical scope. The post-crossing equilibrium is partial decentralization: inference distributes to edge devices while training persists in centralized clusters. This coexistence is stable because the architectural constraints on training are topological, not cost-based. The generalized crossing condition ($R_0 > 1$) endogenizes the 3–5 year coordination layer lag observed in historical transitions and predicts compression to 2–3 years for the current AI transition.

What the mechanism predicts unambiguously is that concentrated investment endogenously produces inference decentralization, that this process accelerates with the number of competitors and is amplified by asymmetric players who benefit from crossing, and that training centralization and inference decentralization will coexist as stable features of the AI economic landscape.

A. Two-Period Pedagogical Model

Setup. Period 1: N symmetric firms choose investment I_i, earning Cournot profits. Period 2: if $\sum I_j$ exceeds $\bar{Q}$, distributed inference entry occurs. Incumbents earn $S = S_T + S_I$ per firm.

Nash equilibrium. $I^* = (a-c)/[b(N+1)]$. Total investment NI^* exceeds $\bar{Q}$ whenever N is sufficiently large.

Cooperative benchmark. $I^C = (a-c)/(2b) < NI^*$ for $N \geq 2$.

B. Overinvestment in Dollar Terms

Table 10: Overinvestment calibration.

	2024	2025 (prelim.)
Actual AI capex ($B)	∼230	∼436
Model Q^N/Q^C ratio	3–4×	3–4×
Implied cooperative ($B)	∼65–75	∼110–145
Excess investment ($B)	∼155–165	∼291–326

The excess is not deadweight loss—it transfers surplus to consumers through the learning curve.

C. Semi-Endogenous Coordination Dynamics

Under declining κ and declining $\bar{Q}_{\text{eff}}$, the state variable $x(t) = \bar{Q}_{\text{eff}}(\eta(t)) - Q(t)$ evolves as:

$$dx/dt = (d\bar{Q}_{\text{eff}}/d\eta) \cdot (d\eta/dt) - \sum q_i$$

Under the quasi-static approximation ($|d\bar{Q}_{\text{eff}}/dt| \ll |\sum q_i|$), the Nash equilibrium applies pointwise. Timescale separation: coordination dynamics evolve over years; production decisions are quarterly.

References

[1] Acemoglu, D., & Guerrieri, V. (2008). Capital deepening and nonbalanced economic growth. *Journal of Political Economy*, 116(3), 467–498.

[2] Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 29(3), 155–173.

[3] Bresnahan, T. F., & Greenstein, S. (1994). The competitive crash in large-scale commercial computing. NBER Working Paper No. 4901.

[4] Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: Engines of growth? *Journal of Econometrics*, 65(1), 83–108.

[5] Carlino, A., et al. (2025). Structural breaks in technology learning curves. *Joule*.

[6] Christensen, C. M. (1997). *The Innovator's Dilemma*. Harvard Business School Press.

[7] David, P. A. (1990). The dynamo and the computer. *American Economic Review*, 80(2), 355–361.

[8] Flamm, K. (1993). *Mismanaged Trade?* Brookings Institution.

[9] Goldberg, P. K., et al. (2024). Learning curves in semiconductor manufacturing. NBER Working Paper No. 32651.

[10] Greenstein, S. (1997). Lock-in and the costs of switching mainframe computer vendors. *Industrial and Corporate Change*, 6(2), 247–273.

[11] Irwin, D. A., & Klenow, P. J. (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of Political Economy*, 102(6), 1200–1227.

[12] Levhari, D., & Mirman, L. J. (1980). The great fish war. *Bell Journal of Economics*, 11(1), 322–334.

[13] Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy*. Harper & Brothers.

[14] Stokey, N. L. (1988). Learning by doing and the introduction of new goods. *Journal of Political Economy*, 96(4), 701–717.

[15] Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.

[16] Walter, W. (1998). *Ordinary Differential Equations*. Springer.

[17] Wright, T. P. (1936). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, 3(4), 122–128.