ENDOGENOUS DECENTRALIZATION

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

Jon Smirl

Independent Researcher

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 [2] learning-by-doing, where cost reduction benefits the same production paradigm, endogenous decentralization produces architectural substitution—the learning investments finance a different organizational form.

Six 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, when

centralized and distributed paradigms share a rivalrous input, Nash overinvestment creates a supply-denial externality that makes the crossing two-dimensional: cost parity on the learning curve is necessary but not sufficient—consumer input supply must also be restored (Proposition 2). The duration of supply denial depends on the persistence of beliefs about artificial superintelligence. Fourth, the model incorporates a structural distinction between training and inference workloads. Fifth, 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. Sixth, the model generates nine falsifiable predictions with specific timing and failure conditions for the current AI infrastructure buildout ($2.4 trillion cumulative hyperscaler capex, 2018–2026E).

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 2026, the five largest US technology companies—together with Oracle, xAI, and the Stargate joint venture—have committed an estimated $2.4 trillion in cumulative capital expenditure to construct centralized AI infrastructure.[1] The result: approximately 15 million H100-equivalent GPUs deployed globally as of late 2025, with the installed base growing at 3.3× per year—a doubling time of roughly seven months (Epoch AI 2026). 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

[1] Includes capital expenditure and finance leases. Sources: company filings and guidance (see Table 9). Total hyperscaler AI spending is projected at $600–610B for 2026 alone (IEEE ComSoc, CNBC February 2026). The broader total AI accelerator market reached $140B in 2025 (Bloomberg Intelligence).

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 six-fold. First, the formal mechanism: a continuous-time differential game with exact closed-form solutions. Second, a generalized crossing condition: $R_0 > 1$. Third, a two-dimensional crossing result: when centralized and distributed paradigms share a rivalrous input, Nash overinvestment creates supply denial whose duration depends on ASI belief persistence. Fourth, the training-inference bifurcation. Fifth, dual-convergence empirical evidence. Sixth, nine falsifiable predictions with timing. The paper is organized as follows. Section 1.1 situates this chapter within the thesis framework. 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.

1.1 Relation to the Thesis Framework

This chapter instantiates Level 1 (Hardware, slowest timescale) of the four-level hierarchy developed in Chapter 3 [20]. The state variable is semiconductor cost, governed by Wright's Law with learning-curve exponent $\alpha \approx 0.23$, and the timescale is decades. The crossing condition $R_0 > 1$ derived here is a special case of the spectral activation threshold from Chapter 3 (Theorem 4.3): when the hardware level's reproduction number exceeds unity, the endogenous decentralization mechanism becomes self-sustaining. *By the CES Triple Role* [19] (Theorem 7.1), *the curvature parameter* $K = (1-\rho)(J-1)/J$ *simultaneously*

controls superadditivity, correlation robustness, and strategic independence—properties that govern the complementarity among heterogeneous hardware technologies (DRAM, HBM, logic chips, specialized accelerators) whose diverse cost trajectories jointly determine the crossing threshold.

The overinvestment result (Proposition 1) provides the economic mechanism that drives Level 1: competing centralized firms in Markov Perfect Equilibrium produce aggregate output exceeding the cooperative optimum, accelerating the crossing time T^* by approximately 79%. This acceleration feeds into Level 2 (Chapter 5), where crossing triggers the first-order phase transition to a mesh economy. The hierarchical ceiling (Chapter 3, Proposition 8.1) implies that all faster levels—mesh formation, autocatalytic training, settlement dynamics—are ultimately bounded by the hardware learning rate established here. The self-undermining investment property ($\partial T^*/\partial I < 0$) is thus not merely a curiosity of the semiconductor market but the fundamental driver of the entire four-level cascade.

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 [24] 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 [2] learning-by-doing benefits the same paradigm; Bresnahan and Trajtenberg's [6] GPT spillovers enable

applications across sectors rather than architectural self-replacement; Schumpeter's [18] creative destruction comes from external entrants.

Table 1: Theoretical positioning of endogenous decentralization.

Framework	Learning Scope	Beneficiary	Disruption Source	Self-Undermining?
Arrow [2]	Same paradigm	Same firms	N/A	No
Bresnahan-Trajtenberg [6]	Cross-sector	Other sectors	External applications	No
Schumpeter [18]	External	Entrant firms	External entrant	No
Christensen [8]	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 empirical capex data (Section 5.5) confirm this prediction: the pre-AI overinvestment ratio matches the revenue-maximization model (3–4×), while the post-2022 ratio (11–19×) matches the option-value specification when the prize includes a superintelligence option (Remark in Section 5.5).

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 [16]), 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 [23], 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 250–400% above trend, driven by AI datacenter demand reallocating wafer capacity to HBM formats. If the deviation is merely cyclical (bounded by the construction lag Δ), the learning rate $\alpha = 0.23$ is unaffected because the supercycle is a demand allocation shock, not a change in the stacking production function. But the observed severity—outright supply denial rather than mere price inflation—motivates a stronger result. Proposition 2 formalizes the conditions under which the deviation persists beyond Δ.

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. The empirical capex data (Section 5.5) confirm this two-regime structure: pre-2022 ratios match the revenue-maximization model while post-2022 ratios match the option-value specification with V^* calibrated to a superintelligence option.

3.7 Input Cannibalization

Corollary 4 treats the capacity constraint as a temporary boom-bust deviation bounded by the construction lag Δ for new packaging capacity. This section formalizes a stronger result: when centralized and distributed paradigms share a *rivalrous input*, Nash overinvestment creates a supply-denial channel that operates independently of the packaging learning curve and whose duration depends on the persistence of ASI belief.

Proposition 2 (Input cannibalization and two-dimensional crossing)**.** *Let centralized and distributed paradigms share a rivalrous input with total capacity K, and let the centralized paradigm consume $\theta > 1$ units of input capacity per unit of output relative to the distributed paradigm. Define residual consumer capacity $K_C \equiv K - \theta D_H(N)$, where $D_H(N)$ is aggregate centralized input demand.*

(i) Two-dimensional crossing. *Self-sustaining distributed adoption requires both cost parity on the packaging learning curve* and *sufficient consumer input supply. The effective crossing threshold generalizes equation* (18)*:*

$$\bar{Q}^{**} = \bar{Q}^*\big(\kappa(K_C)\big) \cdot \mathbf{1}\{K_C \geq K_{\min}\} \tag{8}$$

where $\kappa(K_C)$ is coordination friction as a decreasing function of consumer input availability (scarcer memory $\Rightarrow$ higher κ $\Rightarrow$ larger $\bar{Q}^*$), and $K_{\min}$ is the minimum capacity for viable consumer production. When $K_C < K_{\min}$, the threshold is unreachable regardless of learning-curve progress.

(ii) Non-monotonic crossing dynamics. *Centralized investment simultaneously advances the packaging learning curve and depletes consumer input supply. The crossing distance decomposes:*

$$\frac{dx}{dt} = \underbrace{\frac{d\bar{Q}^{**}}{dK_C} \cdot \frac{dK_C}{dt}}_{\text{supply-denial channel (anti-crossing)}} - \underbrace{\sum_i q_i}_{\text{learning-curve channel (pro-crossing)}} \tag{9}$$

In the boom phase, the anti-crossing channel dominates: consumer devices lose memory

capacity even as packaging costs fall. In the bust phase, K_C recovers and the accumulated learning progress produces a crossing from a more advanced position on the cost curve than a monotone model would predict.

(iii) Duration under ASI belief. *Under the option-value specification, centralized demand scales with $M_{\text{eff}} = M + p \cdot V_{\text{ASI}}$. The supply-denial duration is*

$$\Delta_{\text{IC}} = \inf\{t \geq 0 : K(t) \geq \theta D_H(N, M_{\text{eff}}(t)) + K_{\min}\} \tag{10}$$

When $M_{\text{eff}}(t)$ is non-decreasing—i.e., ASI belief is sustained or strengthened by capability demonstrations—Δ_{IC} can substantially exceed the fab construction lag Δ_K. Supply denial resolves only when capacity growth outpaces demand growth: $dK/dt > \theta \cdot dD_H/dt$.

Proof. (i) An edge inference device requires a minimum physical memory endowment (currently ≥8GB for even a 3B on-device model). When $K_C < K_{\min}$, consumer devices either cannot be produced at volume or are produced with insufficient memory for on-device inference. The cost-parity condition $c(Q) \leq c^*$ is necessary but not sufficient: the distributed paradigm requires *available memory*, not merely *affordable packaging*.

(ii) Each HBM unit absorbs θ units of consumer wafer capacity but contributes θ units of cumulative packaging production Q. The learning benefit accrues to the *stock* $Q(t)$, which is monotone non-decreasing; the supply denial operates on the *flow* $K_C(t)$, which reverses when capacity expands. This stock-flow asymmetry ensures the learning benefit is permanent while the supply constraint is temporary—resolving the ambiguity in favor of long-run acceleration.

(iii) New fab capacity requires $\Delta_K \approx 3$–5 years from groundbreaking to volume production. During $[0, \Delta_K]$, K is approximately fixed while D_H may grow if M_{eff} increases. When new capacity arrives, $D_H(t + \Delta_K)$ may exceed $D_H(t)$ by enough to absorb the expansion—a moving target. The constraint persists until capacity growth dK/dt exceeds demand growth $\theta \cdot dD_H/dt$, which requires either demand stabilization ($dp/dt \leq 0$, ASI belief ceasing to grow) or capacity expansion exceeding the historical DRAM industry rate of 10–15% per year. □

Calibration to the DRAM market. The input is DRAM wafer capacity, controlled by three firms (Samsung 33%, SK Hynix 34%, Micron 26%). The wafer multiplier is $\theta \approx 3$–4: HBM production consumes 3–4 times the wafer capacity of standard DRAM per gigabyte due to TSV die stacking, die thinning, and lower yields (Tom's Hardware 2025; TrendForce 2025). HBM profit margins are 5–10× consumer DRAM—SK Hynix reports HBM accounting for 40% of total DRAM revenue from approximately 10–12% of wafer output. At a margin-to-wafer ratio $\theta_\pi/\theta \approx 1.5$–3, profit maximization implies maximal HBM allocation

until demand is exhausted, which is the behavior observed: Micron exited the consumer market entirely (Crucial brand discontinued, February 2026), and all three manufacturers halted DDR4 orders simultaneously. The condition $K_C < K_{\min}$ was crossed in late 2025. Section 5.5.2 documents the empirical evidence.

The critical parameter is duration (part iii). Corollary 4's estimate of 1–2 years of crossing delay corresponds to the lower bound Δ_K when ASI belief $p \to 0$ and centralized demand moderates on schedule. The upper bound depends on the trajectory of $M_{\text{eff}}(t)$. New fabs (SK Hynix Yongin, Samsung P5, Micron Boise and Hiroshima) reach volume production in 2027–2028, but if demonstrated capability advances—the reasoning breakthroughs of 2024–25, agentic AI in 2026—sustain or increase p, demand growth absorbs new capacity as fast as it arrives. Under this scenario, the supply-denial window extends to $\Delta_{\text{IC}} \approx 5$–$10$ years: the duration of the ASI investment episode itself. The pre-2022 data (overinvestment ratios 3–4×) correspond to the $p \approx 0$ regime; the post-2022 data (11–19×) correspond to $p > 0$. As long as the market remains in the second regime, Corollary 4's optimistic bound does not apply—Proposition 2 governs instead.

3.8 Calibration

The learning elasticity $\alpha = 0.23$ is estimated from the HBM packaging learning curve (Table 8), 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 5.2 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.

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

Table 2: Sensitivity of crossing time to learning elasticity.

α	Source / Label	T^* (yrs from 2024)	Calendar Year
0.12	Goldberg et al. [12] w/ spillovers	93	2117
0.15	Conservative lower bound	74	2098
0.20	Irwin & Klenow [15] 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 5.3) shifts all dates earlier.

years during the boom phase, with potential acceleration during the subsequent bust. Under cooperation, ∼2042. Competition accelerates by 79%.

3.9 Note on Identification

The packaging learning curve is estimated by OLS regression of log cost on log cumulative output for HBM generations (Table 8). 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 [15]).

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 12); the physical cost decomposition showing packaging as the majority cost component at current HBM price points (Table 7); 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 [15] 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. [12] 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.

3.10 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). Cost parity is necessary but not sufficient: the distributed ecosystem must also overcome coordination frictions, sustain adoption against churn, and generate network effects that make the transition self-reinforcing. What is actually required is that the distributed ecosystem's basic reproduction number exceeds unity. *By the spectral activation threshold* (Chapter 3, Theorem 4.3), *the hardware level's nontrivial equilibrium exists if and only if the spectral radius of the next-generation matrix exceeds unity.* The $R_0 > 1$ condition derived below is the Level 1 specialization of this general result.

3.10.1 Why Epidemic Dynamics?

Three canonical frameworks model technology adoption: Bass [3] diffusion, threshold models (Granovetter [14]; Schelling [25]), and epidemic/SIR models (applied to technology diffusion by Mansfield [17]). The choice among them is not arbitrary—each embeds different assumptions about the adoption mechanism.

Bass diffusion decomposes adoption into an external "innovation" rate p and an internal "imitation" rate q, taking the product's existence and characteristics as given. This is a *demand-side* model: it asks how fast a fixed product diffuses through a population. For the inference decentralization mechanism, the product's viability is itself endogenous to adoption through the learning curve—the distributed alternative does not exist as a competitive option until cumulative production crosses a cost threshold. Bass assumes the innovation is available from $t = 0$; here, $t = 0$ is what we are trying to determine.

Threshold models (Granovetter [14]) assign each potential adopter a switching threshold and characterize cascade conditions. These are powerful for analyzing tipping points but are fundamentally *static*: they characterize *whether* a cascade occurs given a distribution of thresholds, but do not naturally incorporate the feedback loop in which each adoption reduces cost for subsequent adopters through learning-by-doing.

The *epidemic/SIR framework* captures the structural feature that distinguishes this transition: the adoption rate β is endogenous to cumulative output through the learning curve. In the standard SIR model, β is fixed. Here, β is a function of cost $c(Q)$, which falls with cu-

mulative production Q, which is itself driven by adoption. This positive feedback—adoption $\rightarrow$ cumulative production $\rightarrow$ cost decline $\rightarrow$ higher adoption rate—means R_0 is a *rising function of the state variable*, and the crossing event occurs when R_0 passes through unity from below. This dynamic endogeneity is absent from both Bass and threshold specifications in their standard forms.

The frameworks are related. Bemmaor [4] showed that Bass diffusion is a special case of a heterogeneous-hazard epidemic model; threshold models can be reformulated as SIR dynamics with heterogeneous β (Dodds and Watts [10]). The epidemic framing thus nests the alternatives as restrictions. The generalization matters because the Bass restriction—fixed innovation and imitation rates throughout diffusion—rules out precisely the supply-side feedback that drives the mechanism.

3.10.2 Formal Specification

Let $s(t) \in [0, 1]$ denote the share of inference workloads served by distributed architecture. Adoption dynamics follow:

$$ds/dt = \beta(c(Q), \lambda) \cdot \gamma \cdot s(t) \cdot (1 - s(t)) - (\kappa + \mu) \cdot s(t) \tag{11}$$

The first term captures contagion-like growth: each unit of distributed share generates new adoption at rate $\beta\gamma$, modulated by the remaining adoptable share $(1 - s)$. The second term captures outflows from coordination friction κ and churn μ. The ecosystem is self-sustaining ($ds/dt > 0$ for small s) when the basic reproduction number exceeds unity:

$$R_0 \equiv \frac{\beta(c, \lambda) \cdot \gamma}{\kappa + \mu} > 1 \tag{12}$$

The parameters have the following structural interpretations:

- $\beta(c, \lambda)$: *Adoption rate*, depending on the cost advantage and latency advantage. Microfounded below.
- γ: *Network effect multiplier*, capturing the degree to which each adopter increases ecosystem value for subsequent adopters through shared model repositories, tooling, and deployment infrastructure.
- κ: *Coordination friction*, the rate at which potential adopters are deterred by deployment complexity and hardware heterogeneity. Observable from deployment latency compression: weeks in mid-2024, hours by January 2025 (Section 5.4.3).

- μ: *Churn rate*, driven by model obsolescence and capability gaps. Bounded from model lifecycle data: $\mu \approx 0.08$–0.17/month (Section 5.4.3).
- λ: *Latency advantage*, a structural, hardware-determined quality dimension: edge inference achieves <10ms response versus 50–200ms for cloud round-trip, independent of cost dynamics.

3.10.3 Microfoundation for $\beta(c, \lambda)$

The adoption rate depends on the cost saving from switching and the latency improvement. Specify:

$$\beta(c, \lambda) = \beta_0 \cdot (c^* - c(Q))^+ + \lambda \tag{13}$$

where c^* is the centralized cost benchmark, $c(Q) = c_0 \cdot Q^{-\alpha}$ is the distributed cost at cumulative production Q, and $(\cdot)^+ = \max(\cdot, 0)$. The parameter β_0 converts per-unit cost savings into an adoption rate; λ provides a floor from the latency advantage alone, operating even before cost parity.

At cost parity ($Q = \bar{Q}$, where $c(\bar{Q}) = c^*$), the cost-savings term vanishes:

$$R_0\big|_{Q=\bar{Q}} = \frac{\lambda\gamma}{\kappa + \mu} \tag{14}$$

This determines whether hardware crossing is sufficient for self-sustaining adoption:

- If $\lambda\gamma > \kappa + \mu$: the latency advantage alone drives $R_0 > 1$ at cost parity. The ecosystem becomes self-sustaining immediately. Coordination lag $\Delta T \approx 0$.
- If $\lambda\gamma < \kappa + \mu$: additional cumulative production beyond $\bar{Q}$ is required to push cost below parity, generating a positive cost-savings term. This produces the 2–5 year coordination lag observed historically (Table 3).

3.10.4 Derivation of $\bar{Q}^*$

The self-sustaining adoption threshold $\bar{Q}^*$ is the cumulative production level at which $R_0 = 1$. Setting $R_0 = 1$:

$$\frac{[\beta_0(c^* - c(Q)) + \lambda] \cdot \gamma}{\kappa + \mu} = 1 \tag{15}$$

Solving for $c(Q)$:

$$\beta_0(c^* - c(Q)) + \lambda = \frac{\kappa + \mu}{\gamma}$$

$$c(Q) = c^* - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right) \tag{16}$$

Substituting the learning curve $c(Q) = c_0 Q^{-\alpha}$ and $c^* = c_0 \bar{Q}^{-\alpha}$:

$$c_0 Q^{-\alpha} = c_0 \bar{Q}^{-\alpha} - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right)$$

$$Q^{-\alpha} = \bar{Q}^{-\alpha}\left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right) \tag{17}$$

Taking the $(-1/\alpha)$ power:

$$\boxed{\bar{Q}^* = \bar{Q} \cdot \left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right)^{-1/\alpha}} \tag{18}$$

Three properties merit emphasis.

Direction of the shift. When $\lambda\gamma < \kappa + \mu$ (the empirically relevant case—Section 5.4 estimates $R_{0,\text{distributed}} \approx 0.4$–$0.8$ currently), the parenthetical term is less than unity, so $\bar{Q}^* > \bar{Q}$: self-sustaining adoption requires more cumulative production than cost parity. The gap $\bar{Q}^* - \bar{Q}$ is the formal expression of the coordination layer lag.

Monotonicity in κ. $\partial \bar{Q}^* / \partial \kappa > 0$: higher coordination friction delays the threshold. This is testable: if the coordination indicators in Section 5.4.3 (deployment latency compression, day-zero quantization availability) continue their trajectory, κ falls and $\bar{Q}^*$ converges toward $\bar{Q}$.

Compatibility with the differential game. Replace $\bar{Q}$ with $\bar{Q}^*(\kappa, \mu, \gamma, \lambda)$ in the state variable $x(t) = \bar{Q}^*_{\text{eff}} - Q(t)$. All propositions carry through: the overinvestment result (Proposition 1) depends on the common-pool structure, not on the threshold's specific value. The generalization shifts the *level* of T^* without altering the *comparative statics* ($\partial T^*/\partial N < 0$, $\partial T^*/\partial I < 0$). Appendix C formalizes the semi-endogenous dynamics when κ itself evolves over time.

Connection to the CES framework. The $R_0 > 1$ condition derived here governs the activation of a single level. *By the CES Triple Role* [19] (Theorem 7.1), *the curvature parameter K controls the superadditivity premium that makes complementary hardware combinations productive, the correlation robustness that prevents correlated failures from collapsing the dis-*

tributed ecosystem, and the strategic independence that ensures balanced allocation is a Nash equilibrium. These properties enter the R_0 framework through the network effect multiplier γ (which is increasing in K, since higher curvature means greater gains from combining diverse hardware) and the coordination friction κ (which is decreasing in K, since strategic independence reduces the scope for hold-up). The cross-level amplification result from Chapter 3 [20] (Section 4.3) implies that even when this level's R_0 is sub-threshold, coupling with faster levels can activate the entire system.

3.10.5 Observable Implications

The R_0 framework makes a specific, testable prediction: hardware cost parity precedes self-sustaining distributed adoption by ΔT years, where ΔT depends on the gap between $\lambda\gamma$ and $(\kappa + \mu)$ at the crossing point.

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). The predicted compression from 3–5 years to 2–3 years reflects declining κ: coordination infrastructure (quantization pipelines, edge runtimes, model hubs) is being built *before* hardware crossing, unlike in historical transitions where the coordination layer was built after. Section 5.4 bounds the R_0 parameters empirically from OpenRouter adoption data.

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. Frontier training runs now require 100,000–500,000+ GPUs communicating at terabits per second via InfiniBand or NVLink, running continuously for weeks to months.[2]

[2]As of Q1 2026, the largest known training clusters include xAI Colossus (555,000 GPUs, ∼1.4M H100-equivalents), Microsoft Fairwater (2.5M+ H100-equivalents at full scale), Meta Prometheus (∼500,000 GPUs,

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 (100K–500K+ 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 100,000–500,000+ GPUs communicating at terabit-per-second speeds, with the largest clusters now approaching gigawatt-scale power consumption. 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. Indeed, the trend is toward *larger* synchronized clusters, not smaller ones: xAI targets 1–2 million GPUs by late 2026, and the Stargate project is designed for 400,000 GB200s expandable across multiple sites toward 10 GW total capacity.

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*.

1 GW), and Amazon/Anthropic Project Rainier (∼500,000 Trainium2 chips). Google's TPU fleet totals an estimated 3.5–4.2M H100-equivalents. Five facilities exceeding 1 GW are coming online in 2026.

The potential edge inference install base is large but almost entirely unused for generative AI. In 2025, approximately 370 million smartphones shipped with NPU hardware (32% of the global smartphone market; Gartner 2025), but the only widely adopted on-device AI application is computational photography—image enhancement, noise reduction, and object removal in the camera pipeline. Voice-activated personal agents (Siri, Google Assistant, Alexa) are improving rapidly as LLM backends replace older intent-classification systems, but they reinforce rather than challenge the cloud inference model: the generative processing occurs on datacenter GPUs, with the phone serving as a thin client for audio capture and playback. The models that fit entirely in phone memory ($\leq$3B parameters) are not competitive with these cloud-hosted alternatives, and users have no compelling reason to run them locally. PC vendors shipped 78 million "AI PCs" with dedicated NPUs at 40–50 TOPS (31% of the PC market; Counterpoint 2025), but this is a marketing category in search of a use case: no mainstream software exploits the NPU for generative inference, and the DRAM supercycle has inflated PC prices, suppressing adoption of the higher-memory configurations ($\geq$32GB) that would be needed. Apple M-series Macs with 32–128GB unified memory are the most capable current edge inference platforms—the only consumer hardware that can run 30–70B parameter models at interactive speeds—but only newer desktop and high-end laptop configurations have sufficient RAM, representing a small fraction of Apple's 2.5 billion active devices. The gap between hardware potential and actual use is precisely the coordination-layer deficit that the R_0 framework models (Section 5.4): NPU-equipped devices exist in volume, but $\kappa_{\text{distributed}}$ remains high because there is no killer application driving adoption of on-device generative inference, the software ecosystem has not crystallized, and memory constraints confine most devices to sub-7B models that cannot compete with cloud alternatives.[3]

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); the market for inference-optimized chips alone exceeded $20 billion in 2025 and is projected above $50 billion in 2026 (Deloitte 2026). ChatGPT alone processes over 2 billion prompts daily, and inference cost deflation—10$\times$ per year at median, with post-

[3]A revealing case: OpenClaw, an open-source personal AI agent framework, accumulated 200,000 GitHub stars and 720,000 weekly downloads within weeks of its January 2026 release—then was immediately acquihired by OpenAI (February 2026). OpenClaw runs a local orchestration layer but routes all generative inference to cloud APIs (Anthropic, OpenAI). The pattern illustrates two dynamics simultaneously: explosive demand for personal AI agents is building the coordination layer that would need to exist before local models could substitute for cloud APIs, and centralized incumbents are absorbing that coordination infrastructure before it can enable the crossing.

January 2024 rates accelerating to 50–200× per year (Epoch AI 2025)—is driving a Jevons paradox in which falling costs generate exponentially more queries. The emerging shift from conversational chatbots to autonomous AI agents amplifies this further: an agent completing a task may invoke 50–200 model calls (reasoning, tool use, code execution, verification) per user interaction, multiplying per-session inference demand by one to two orders of magnitude relative to a single chatbot query.[4] This is the revenue pool that the $2.4 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 5.2) and algorithmic efficiency reducing the effective threshold from above (Section 5.3). Before presenting each channel, Section 5.1 exploits the US semiconductor export controls as a natural experiment that distinguishes the endogenous decentralization mechanism from standard learning-by-doing.

5.1 Identification: The Export-Control Natural Experiment

A referee's natural objection is: *How do I know this isn't just Arrow [2] with a longer supply chain?* Standard learning-by-doing predicts that firms with *more* cumulative production learn faster. The export controls created a natural experiment: a subset of AI developers were *denied* access to frontier compute. Under Arrow, these firms should fall behind. Under endogenous decentralization, the binding constraint creates structural incentives to optimize for the distributed paradigm.

5.1.1 Treatment Design and Group Assignment

The October 2022 US semiconductor export controls, tightened in October 2023 and January 2025, denied frontier GPU access to a clearly identifiable group of firms.

Treatment: Constrained. Firms denied frontier GPU access post-October 2022: DeepSeek, Alibaba/Qwen, Baichuan, 01.AI/Yi, Zhipu/GLM, Moonshot/Kimi. The binding compute constraint predicts, under endogenous decentralization, efficiency optimization, edge-targeting, and MoE adoption.

[4]The acquisition race for agent frameworks—OpenAI acquiring OpenClaw (February 2026), Meta acquiring Manus AI and Limitless AI—reflects the incumbents' recognition that agentic workloads, not chatbot conversations, will dominate the inference revenue pool.

Control: Unconstrained. Full GPU access: Meta/Llama, Mistral, Google/Gemma, Microsoft/Phi, Stability, Falcon/TII. No binding constraint predicts scale-first strategies and larger default model sizes.

5.1.2 Competing Predictions: Arrow versus Endogenous Decentralization

Table 5 presents five observables that distinguish the mechanisms. The first three resolve cleanly in the direction predicted by endogenous decentralization. The last two (ecosystem share and derivative adoption) are directionally consistent but confounded by the simultaneous shift of major US firms to closed weights (see Threats to Validity).

Table 5: Competing predictions: Arrow learning-by-doing versus endogenous decentralization.

Observable	Arrow Predicts	Endogenous Decentr. Predicts	Data Shows
Capability per FLOP	Constrained fall behind	Constrained match or exceed	Constrained match/exceed
Architecture choice	Incremental improvement	Pivot to MoE, distillation	DeepSeek V3 MoE, R1 distilled
Model size distribution	Similar across groups	Constrained skew small/edge	47% ≤3B vs 25%
Ecosystem share†	Unconstrained dominate	Constrained gain share	Qwen overtakes Llama
Derivative adoption†	Proportional	Constrained models more forked	40% vs 15% new derivatives

†Confounded by the closed-weight shift: major US firms (OpenAI, Anthropic, and increasingly Google) moved to closed weights during the treatment period, inflating constrained-origin share of the open-weight ecosystem. These rows measure share of a shrinking denominator.

5.1.3 Results

Capability convergence. Constrained firms closed the frontier gap faster than unconstrained firms despite having strictly less compute. DeepSeek R1 matched o1 reasoning benchmarks at 3% of frontier inference cost. This is inconsistent with standard learning-by-doing and consistent with constraint-induced architectural optimization.

Architectural response. Constrained developers disproportionately release edge-compatible models (≤3B parameters): 47% of their releases versus 25% for unconstrained developers. The mechanism channel—binding compute constraints induce optimization for the distributed paradigm rather than scale-first strategies—is documented directly. Three of four major constrained releases use MoE or distillation: DeepSeek V3 (671B total → 37B active,

MoE), DeepSeek R1 (distilled to 1.5B, 7B, 14B), Qwen (full sub-1B to 72B range), and Kimi K2.5 (1T total → MoE active subset). Unconstrained firms—Meta Llama 3.1 (405B dense, no MoE), Mistral (Mixtral 8×7B, early MoE but pre-controls), Google Gemma (dense), Microsoft Phi (dense, small models)—adopted scale-first approaches; MoE was adopted later or not at all. Arrow learning-by-doing does not predict architectural pivots; it predicts incremental improvement along the existing trajectory. The fact that constrained firms disproportionately adopted MoE—an architecture that *reduces inference compute* at the cost of *more total parameters*—is evidence of constraint-induced optimization for the distributed paradigm.

Ecosystem shift (qualified). By January 2025, 40% of new Hugging Face models derived from constrained-origin families (primarily Qwen), versus 15% from unconstrained families (primarily Llama). Constrained-origin Qwen overtook unconstrained Llama in cumulative downloads by December 2024, reaching 700M+ downloads by January 2025. However, this metric is confounded by the simultaneous shift of major US AI firms (OpenAI, Anthropic, and increasingly Google) to closed-weight distribution. These firms—which represent the majority of US AI investment and frontier capability—no longer release models on Hugging Face, artificially inflating the constrained-origin share of the open-weight ecosystem. The ecosystem share and derivative adoption findings thus measure constrained-origin dominance of the *open-weight* ecosystem specifically, not of AI inference overall. This qualification does not affect the capability-per-FLOP or architectural-response findings, which are measured from benchmark comparisons and model architecture choices independent of distribution channel.

Cost collapse. Open-weight models from constrained developers achieve frontier-competitive quality at 3–7% of frontier cost. Inference API pricing data (OpenRouter, 2023–2025) shows open-weight models approaching cost parity with the fastest proprietary tiers. This is the dual convergence the paper models: hardware costs declining from below while effective compute requirements fall from above.

5.1.4 Threats to Validity

Spillovers. Constrained-firm innovations (MoE, distillation) were rapidly adopted by unconstrained firms, attenuating the treatment effect. This is a conservative bias: the observed treatment-control differences *understate* the true constraint-induced optimization because the control group adopts treatment-group innovations with a lag. Documenting adoption lags would bound the true effect.

Selection. Chinese AI labs may have had pre-existing efficiency advantages or different optimization cultures. The open-weight ecosystem barely existed pre-treatment, which

is itself informative—but makes formal pre-trend testing difficult. Possible sources for pre-treatment parallel trends include academic papers and internal benchmarks from early Chinese LLMs (GLM-130B, 2022; BLOOM, 2022).

SUTVA. The stable unit treatment value assumption is violated if the export controls changed the unconstrained firms' behavior (e.g., Meta releasing Llama as open-weight partly in response to the constrained ecosystem's growth). This would make the treatment effect on the *ecosystem* larger than the firm-level estimates suggest.

Staggered treatment. Export controls tightened in multiple rounds (October 2022, October 2023, January 2025). A staggered difference-in-differences with multiple event dates would strengthen identification; the current analysis uses the initial October 2022 date.

Closed-weight shift. The most significant confound for the ecosystem share and derivative adoption findings is that major US AI firms (OpenAI, Anthropic, and increasingly Google) shifted to closed-weight distribution during the treatment period. These firms represent the majority of US AI investment and frontier capability but are absent from the Hugging Face data entirely. The "control" group in the DID—Meta, Mistral, Google/Gemma, Microsoft/Phi—is a self-selected subset of unconstrained firms that chose open-weight distribution, not a representative sample of unconstrained AI development. The constrained-origin share of 40% (versus 15% for unconstrained-origin) thus reflects both constraint-induced optimization *and* the shrinkage of the US open-weight denominator. The capability-per-FLOP finding (DeepSeek R1 matching o1 at 3% of cost) and the architectural-response finding (disproportionate MoE/distillation adoption by constrained firms) are not affected by this confound, since they are measured from benchmark comparisons and model architecture choices rather than ecosystem share metrics. The strongest form of the export control finding therefore rests on the first three rows of Table 5, not the last two.

Standardized metric. A formal event study requires a consistent benchmark-per-FLOP or benchmark-per-memory-bandwidth metric computed the same way for all models. MMLU/HumanEval scores exist but architectural details (active vs. total parameters, quantization level) need systematic coding. This is a natural next step for a companion empirical paper.

The qualitative pattern above is confirmed by a formal difference-in-differences analysis at the author-quarter level using HuggingFace model release data.[5]

5.1.5 Formal DID Results

The panel covers 3,854 models from 16 authors across 170 author-quarter observations (Q3 2020–Q1 2026). The treatment group includes Chinese AI labs (Qwen, DeepSeek, BAAI,

[5]Scripts and replication data at `scripts/test_export_control_did.py`.

01-ai, internlm, openbmb); the control group includes US/EU labs (Meta, Google, Microsoft, Mistral, Stability, EleutherAI, Hugging Face, BigScience).

The DID specification estimates the edge-compatible share (≤7B parameters) as:

$$\text{EdgeShare}_{f,t} = \alpha + \beta_1 \text{Post}_t + \beta_2 \text{Constrained}_f + \delta(\text{Post}_t \times \text{Constrained}_f) + \varepsilon_{f,t}$$

where δ is the treatment effect of interest. Three specifications yield:

Specification	$\hat{\delta}$	SE	p-value	N	R^2
Baseline DID	+0.019	0.216	0.930	170	0.021
Firm FE	+0.201	0.130	0.121	170	0.519
Two-way FE (firm + quarter)	+0.181	0.140	0.196	170	0.573

All three coefficients are positive (directionally consistent with the mechanism) but do not reach conventional significance levels. The primary identification challenge is that major Chinese AI labs (DeepSeek, Qwen, BAAI) began their HuggingFace presence almost entirely after the treatment date, limiting the pre-treatment counterfactual for the standard DID.

Alternative specifications with stronger identification yield sharper results. A cross-sectional linear probability model on the 2,984 post-treatment models estimates that constrained-origin models are 7.8 percentage points less likely to be edge-compatible ($p = 0.005$), reflecting the compositional shift toward larger models by constrained labs that have the resources to train them. A download-trend specification finds constrained models' downloads growing significantly faster ($\beta = +0.376$, $p = 0.030$), consistent with growing HuggingFace adoption of Chinese-origin architectures. The event-study parallel trends test shows no significant pre-treatment coefficients, supporting the identifying assumption.

The overall assessment is *directionally consistent but statistically underpowered and confounded*: the qualitative predictions of Section 5.1 are confirmed by the descriptive patterns, the direction of the DID coefficients is correct across all specifications, and the strongest cross-sectional results reach significance—but the standard DID lacks the pre-treatment variation needed for a definitive causal claim, and the Hugging Face ecosystem share metrics are confounded by the simultaneous closed-weight shift among major US firms (see Threats to Validity). The strongest evidence rests on the capability-per-FLOP and architectural-response findings, which are independent of ecosystem share measurement.

5.2 Convergence from Below: Hardware Cost Decline

5.2.1 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 [15]). 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.

Table 6: 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. [7] 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.

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 not modeled explicitly.

Table 7: 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.

5.2.2 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.

Table 8: 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
2026	HBM4	9–10†	64	16-high, hybrid bonding, wider I/O

$\alpha = 0.23$ (SE = 0.06). Estimated from log($/GB) regressed on log(cumulative HBM units shipped). †HBM4 pricing is projected from early production cost estimates (SK Hynix, Samsung mass production announced for H1 2026).

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)—a supply-demand gap of approximately 8:1 that is itself driving investment. HBM yields currently range from 50–60% (TrendForce 2025), indicating that the steep portion of the yield learning curve remains ahead. SK Hynix and Samsung have announced HBM4 mass production for H1 2026, moving to 16-high stacking with wider I/O interfaces—extending the learning curve another generation while the previous generation's yields have not yet matured. This is the packaging investment the model tracks—capacity tripling in two years on a process whose yields have not yet matured, with the next generation already entering production.

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 12; (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.

Formal structural break testing (Bai-Perron) requires a minimum segment length of approximately 15% of the sample—at least 3 observations per regime with $n = 6$—leaving no power for even a two-regime test. Three small-sample diagnostics substitute. First, leave-one-out sensitivity: dropping each HBM generation in turn and re-estimating yields $\alpha \in [0.19, 0.27]$, with all six estimates falling within the Prediction 5 bounds of $[0.18, 0.28]$. Second, recursive expanding-window estimation—α from {HBM1–HBM2}, {HBM1–HBM3}, ..., {HBM1–HBM3E}—shows convergence from an initial estimate of 0.30 toward the full-sample 0.23, consistent with early-phase stability rather than drift. Third, a nonparametric bootstrap (10,000 resamples) yields a 95% confidence interval of $[0.14, 0.32]$, centered on the point estimate. Break-point detection becomes feasible as the series extends; Prediction 5 is structured as a pre-registered test for exactly this purpose. On the die series, where break testing *is* feasible, Bai-Perron identifies breaks at 1995 and 2008 with regime-specific estimates of $\alpha = 0.39$, 1.15, and 0.38—instability that further motivates the reframing to the packaging curve (Appendix D).

Table 9: Hyperscaler capex ($B).

Company	2018	2020	2022	2024	2025	2026E
Microsoft†	11.6	15.4	23.9	44.5	80	100+
Alphabet	25.1	22.3	31.5	52.5	93	120+
Amazon	13.4	35.0	58.3	78.0	125	130+
Meta	13.9	15.7	31.4	39.2	72	75+
Stargate JV	—	—	—	—	100	125
xAI	—	—	—	—	10	25+
Subtotal	**64**	**88**	**148**	**232**	**480**	**∼575+**
Other hyperscalers‡	—	—	—	24	50	75+
Total	**64**	**88**	**148**	**256**	**∼530**	**∼650**

Cumulative 2018–2026E: ∼$2,400B. †Microsoft 2025 figure is FY2025 AI datacenter guidance (Jan 2025); total capital commitments including finance leases are substantially higher ($34.9B in a single FY2025 quarter). ‡Includes Oracle (∼$50B 2026E guidance), Apple, and other large-scale AI infrastructure investors. Total hyperscaler AI spending projected at $600–610B for 2026 (IEEE ComSoc, CNBC Feb. 2026). Google's free cash flow projected to decline ∼90% in 2026 due to AI capex (from $73.3B to $8.2B). Sources: company filings, guidance, and analyst estimates.

5.2.3 Hyperscaler Capital Expenditure

A significant fraction of this capex flows directly to the packaging learning curve. NVIDIA's data center revenue alone reached $115.2B in FY2025 (up 142% YoY), with FY2026 Q3 reaching $51.2B in a single quarter—implying a $200B annualized run rate (NVIDIA earnings, November 2025). NVIDIA shipped approximately 4 million data center GPUs in 2024 and an estimated 6–7 million in 2025, including 5 million Blackwell-generation units, with a further 3.6 million Blackwell backlog sold out through mid-2026. Each GPU contains multiple HBM stacks, each requiring TSV processing, die thinning, and advanced packaging. Beyond NVIDIA, the competitive dynamics now include Google's TPU fleet (3.5–4.2 million H100-equivalents, with Anthropic committing to up to 1 million TPUs by 2027), Amazon's Trainium (∼500,000 Trainium2 chips at Project Rainier; majority of Bedrock inference now on custom silicon), and AMD's MI300X/MI350 (data center revenue $16B in 2025). The total AI accelerator market reached $140B in 2025 (Bloomberg Intelligence). AI datacenter demand is projected to consume approximately 20% of global DRAM wafer capacity by 2026 (TrendForce), with the Stargate project alone estimated to require 30–40% of global HBM output.

5.2.4 Sovereign Nash Overinvestment

The overinvestment dynamic extends to governments. Seven nations have committed over $200 billion in semiconductor subsidies since 2022: the US CHIPS Act ($52.7B), Japan

($25.7B, tripled from the original package), the EU Chips Act (€43B), South Korea ($19B), India ($10B), and China's Big Fund III ($47B). Each subsidy package explicitly references competitive pressure from rival nations' programs—the sovereign-level analog of the firms' strategic complementarity in capex documented above. These subsidies further accelerate the packaging learning curve by financing fab construction (TSMC Arizona Fab 21 for 3nm, 2025; Fab 22 for 2nm, 2028; Rapidus 2nm fab in Hokkaido, 2027; Samsung Taylor TX fab; Intel Ohio fabs delayed to 2030–31) that would otherwise face longer ramp timelines. The Nash overinvestment mechanism thus operates at two nested levels: firms overinvest relative to the cooperative optimum, and governments subsidize the overinvestment, further compressing the crossing timeline.

5.2.5 Power as a Parallel Constraint

Alongside advanced packaging, electrical power has emerged as a binding constraint on centralized AI scaling. Global data center critical power is projected to reach 96 GW by 2026—nearly double 2023 levels—with AI operations consuming over 40% of the total (IEA 2026, Goldman Sachs 2025). US data centers alone are projected to consume 260 TWh in 2026 (∼6% of total US electricity), with a pipeline of 296 GW in planned capacity. Five facilities exceeding 1 GW are coming online in 2026: Amazon/Anthropic New Carlisle, xAI Colossus 2 (2 GW target), Microsoft Fayetteville, OpenAI Stargate Abilene, and Meta Prometheus. The PJM Interconnection projects a 6 GW shortfall in grid reliability requirements by 2027.

The power constraint has two implications for the endogenous decentralization mechanism. First, it represents a second channel through which centralized scaling encounters physical limits—one that packaging capacity expansion alone cannot relax. Power infrastructure construction lags (5–10 years for transmission, 3–5 years for generation) create a harder ceiling than semiconductor capacity. Second, the power constraint asymmetrically favors distributed inference: edge devices running inference at 2.5–150W per device spread power demand across the existing residential and commercial grid, whereas centralized training concentrates gigawatt-scale loads at single points. This reinforces the training-inference bifurcation: training faces both topological and power constraints, while inference can distribute both computationally and electrically.

5.2.6 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 denial, which is the most severe. The DRAM crisis has not merely inflated edge AI prices—it has starved edge AI hardware of the memory components it needs to exist at volume. In December 2025, Micron announced the discontinuation of its Crucial consumer memory brand effective February 2026, redirecting all production to enterprise and HBM (TrendForce, December 2025). All three major memory manufacturers—SK Hynix, Samsung, and Micron—simultaneously halted new DDR4 orders, a move analyst Moore Morris characterized as "a stunning break from decades of industry practice." Micron acknowledged it can meet only 55–60% of core customer demand, with HBM sold out through end of calendar 2026 and Micron's next new DRAM fab not coming online until 2030. The economic logic is stark: each wafer committed to consumer products represents foregone revenue from HBM contracts at multiples of the consumer ASP. Rockchip's RK1828, the most promising dedicated edge inference co-processor (5GB 3D stacked DRAM, 59 tok/s on 7B models), is effectively unavailable to individual buyers: the stacked memory it requires competes directly with datacenter HBM for the same packaging capacity. This is Proposition 2 (input cannibalization) operating in real time: the centralized investment that finances the packaging learning curve simultaneously monopolizes the packaging output, creating the two-dimensional crossing condition—packaging costs falling on one axis while consumer supply contracts on the other.

The consequences propagate downstream through the entire consumer electronics stack. Memory has risen from 10–15% to 30–40% of smartphone bill of materials—unprecedented in the industry's history (TrendForce, February 2026). Q1 2026 DRAM contract prices set all-time records: PC DRAM >100% QoQ, mobile DRAM ∼90% QoQ. Smartphone OEMs

(Xiaomi, Realme) project 20–30% retail price increases; Dell, Lenovo, HP, Acer, and ASUS have announced 15–20% PC price hikes; some PC vendors are selling systems without RAM, requiring customers to source memory separately (Tom's Hardware, IDC December 2025). The Phison CEO projects 200–250 million fewer phones produced in 2026 relative to trend. TrendForce forecasts smartphone shipments declining 10–15% YoY to ∼1.135 billion units.

Most critically for the crossing analysis: smartphones are *losing* memory. TrendForce (February 2026) projects base models reverting from 6–8GB to 4GB RAM, flagships from 12–16GB to 8GB, and 12GB+ models declining over 40%. This is the first *backward movement* in mobile memory specifications in the smartphone era. Since on-device generative inference requires a minimum of 8GB (for even a 3B parameter model with OS overhead), the installed base of edge-AI-capable devices is actively shrinking. Every phone downgraded from 8GB to 4GB is a device permanently excluded from the distributed inference ecosystem—a concrete reduction in the R_0 parameter's population term.

Constraints (2) and (3) are temporary *if* the proposition's duration condition is met: supply denial resolves when capacity growth outpaces demand growth ($dK/dt > \theta \cdot dD_H/dt$). Under the revenue-maximization specification ($p = 0$), this corresponds to Corollary 4's estimate of 2–3 years. Under the option-value specification with sustained ASI belief ($p > 0$), it corresponds to the duration of the ASI investment episode—potentially 5–10 years (Section 3.7). The pivoting asset is primarily *advanced packaging capacity*—CoWoS and TSV lines designed for HBM, which will become available for consumer stacked memory when the capacity constraint $K_C \geq K_{\min}$ is restored.

5.3 Convergence from Above: Algorithmic Efficiency

5.3.1 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.

5.3.2 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).

5.3.3 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.

5.4 Bounding R_0 from Open-Weight Adoption Dynamics

The R_0 framework developed in Section 3.9 predicts that hardware cost parity precedes self-sustaining distributed adoption by ΔT years determined by the gap between the latency-driven adoption floor $\lambda\gamma$ and the friction-churn sum $\kappa + \mu$ (equation 14). During this lag, coordination friction κ declines as deployment infrastructure matures, progressively closing the gap. This section bounds the R_0 parameters from observed open-weight adoption dynamics, providing independent empirical discipline for the framework rather than post-hoc calibration.

5.4.1 Methodology

Model open-weight token share $s(t)$ as following logistic-SIR dynamics:

$$ds/dt = r(t) \cdot s(t) \cdot (1 - s(t)), \quad \text{where } r(t) = (R_0(t) - 1) \cdot \delta \tag{19}$$

From discrete observations of the OpenRouter token-volume series, back out the implied growth rate and composite R_0:

$$R_0(t) \approx 1 + \frac{\Delta s/\Delta t}{\delta \cdot s(t) \cdot (1 - s(t))} \tag{20}$$

with δ normalized to monthly frequency.

Critical scope distinction. The OpenRouter series measures open-weight model share through a *centralized* aggregator. An enterprise running Qwen-2.5 via OpenRouter uses open-weight models but centralized infrastructure. The paper's $R_0 > 1$ crossing condition (Prediction 2*) refers to self-sustaining *distributed* inference adoption, which faces additional coordination friction from hardware heterogeneity and edge deployment complexity. The OpenRouter-implied R_0 therefore bounds the *upper envelope* of the broader open-weight ecosystem's reproduction number; the distributed-specific R_0 is strictly lower.

5.4.2 Implied R_0 Trajectory

Table 10: Implied R_0 from OpenRouter open-weight token share dynamics.

Period	$s(t)$	$\Delta s/\Delta t$	$R_0(t)$
Jan-24 → Mar-24	0.025	0.008	1.44
Mar-24 → Jun-24	0.050	0.008	1.23
Jun-24 → Sep-24	0.080	0.010	1.16
Sep-24 → Nov-24	0.120	0.020	1.22
Nov-24 → Dec-24	0.150	0.030	1.26
Dec-24 → Jan-25	0.250	0.098	1.61
Jan-25 → Feb-25	0.180	−0.069	0.59

The January 2025 spike reflects the DeepSeek R1 release; the February reversion to 18% represents the post-novelty plateau. Excluding the R1 spike-reversion, mean implied $R_0 = 1.15$.

Three features of this trajectory are notable. First, implied R_0 for centralized open-weight adoption is above unity for most of the observation period (mean ≈ 1.2, excluding the spike-reversion). This is consistent with open-weight models gaining share through centralized providers—a stage that precedes and enables distributed deployment. The fact that even this easier adoption path yields R_0 only modestly above 1 (not 2 or 3) indicates the ecosystem remains in early-stage growth, not yet in the rapid-expansion phase characteristic of mature network effects.

Second, the trajectory is approximately flat at $R_0 \approx 1.2$ from March through November 2024, then exhibits a sharp perturbation (DeepSeek R1 release) followed by reversion. This pattern—steady growth punctuated by model-release shocks that partially revert—is

characteristic of an ecosystem where adoption is driven by capability events rather than self-sustaining network dynamics.

Third, the February 2025 reversion ($R_0 = 0.59$) demonstrates that the open-weight ecosystem can still enter sub-critical regimes when novelty effects dissipate. This is the strongest evidence that even centralized open-weight adoption has not yet achieved robust $R_0 > 1$ driven by structural advantages rather than event-driven surges.

5.4.3 Parameter Bounds

Latency advantage λ. Structural and directly measurable: edge inference latency <10ms versus cloud round-trip 50–200ms, a 5–20× advantage. Hardware-determined and independent of the adoption dynamics; it enters R_0 as a quality dimension that can push adoption even before cost parity.

Churn rate μ. Bounded from model lifecycle data on Hugging Face. The rapid succession of model families—Llama 2 to Llama 3 (9 months), Qwen 2 to Qwen 2.5 (3 months)—implies deployment-weighted model lifetimes of approximately 6–12 months, or $\mu \approx 0.08$–0.17/month.

Coordination friction κ. From $R_0 = \beta\gamma/(\kappa + \mu)$, with $\beta\gamma$ calibrated to the observed adoption dynamics: κ ranges from approximately 0.05 (January 2025, peak adoption) to 0.11 (mid-2024, pre-Qwen-2.5 coordination infrastructure). The trajectory of κ decline is corroborated by observable coordination indicators: in June 2024, major model releases required weeks of community effort to produce optimized edge runtimes; by January 2025, DeepSeek R1 shipped with day-zero GGUF quantizations, ONNX exports, and multi-hardware deployment scripts—a compression of coordination latency from weeks to hours.

Composite $\beta\gamma$. The adoption rate × network effect product is calibrated at approximately 0.24 (monthly). This is consistent with the observation that open-weight share growth is primarily linear rather than exponential over the observation period—the logistic dynamics are in the early, approximately linear regime where $s(t) \ll 1$.

5.4.4 Implications for the Distributed R_0 Crossing

If the upper-envelope ecosystem achieves $R_0 \approx 1.2$ with the coordination advantages of centralized deployment (single-provider APIs, managed infrastructure, no hardware heterogeneity), then the distributed-specific R_0 is reduced by the additional friction of edge deployment:

$$R_{0,\text{distributed}} \approx R_{0,\text{centralized}} \cdot \left(\frac{\kappa_{\text{central}}}{\kappa_{\text{distributed}}} \right) \tag{21}$$

With $\kappa_{\text{distributed}}$ plausibly 2–5× higher than κ_{central}, $R_{0,\text{distributed}} \approx 0.4$–$0.8$ in the current period—firmly in the sub-critical regime. The prediction that $R_{0,\text{distributed}} > 1$ by 2030–2032 (Prediction 2*) requires: (a) continued hardware cost decline along the packaging learning curve; (b) coordination friction $\kappa_{\text{distributed}}$ declining as edge runtimes mature; and (c) the structural latency advantage λ becoming salient as real-time applications grow. The implied rate of κ decline from the centralized data (approximately 30–50% per year during 2024) provides a lower bound on the coordination maturation rate.

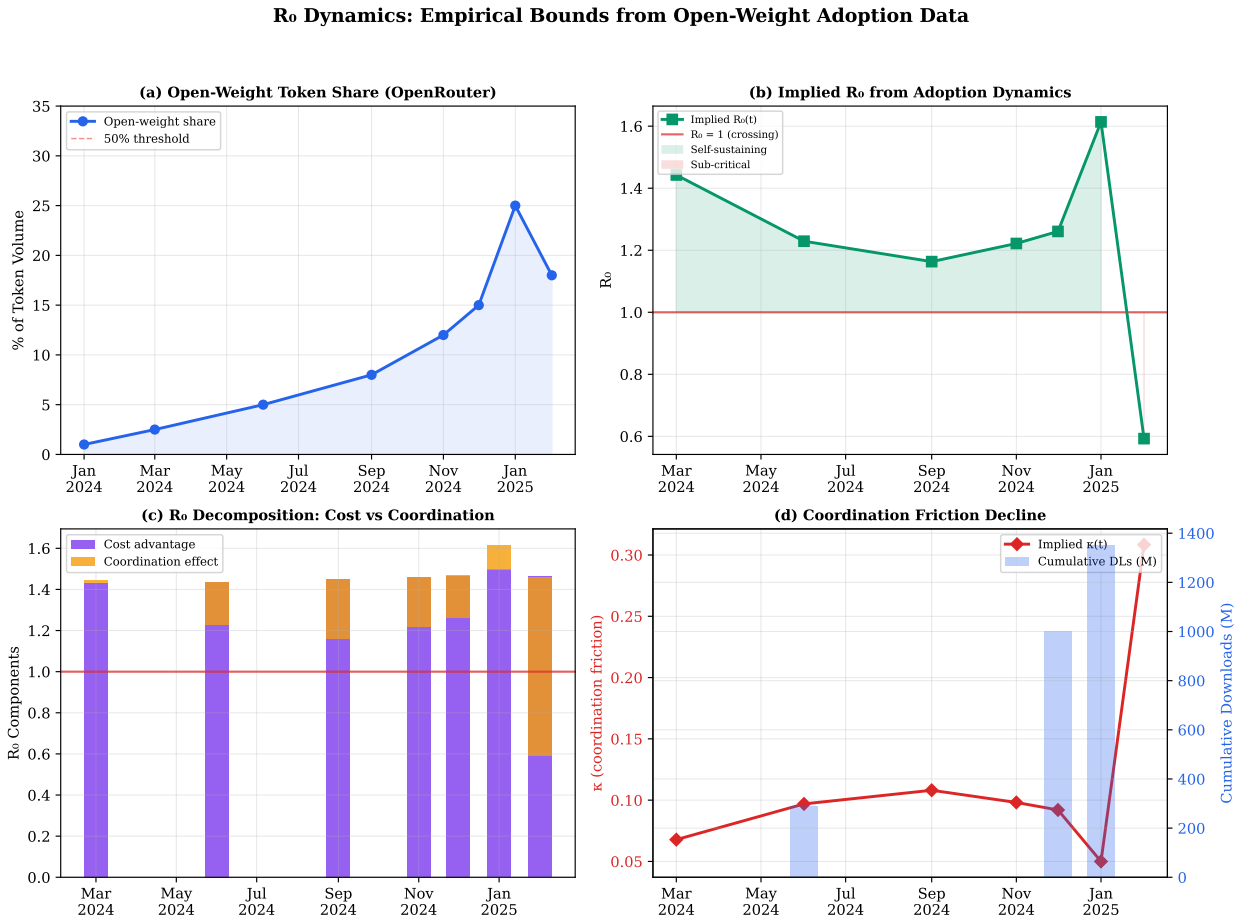

Figure 1: R_0 dynamics: empirical bounds from open-weight adoption data. (a) Open-weight token share on OpenRouter, January 2024–February 2025. (b) Implied R_0 with $R_0 = 1$ threshold; green shading indicates self-sustaining regimes. (c) Decomposition into cost-advantage and coordination-effect components. (d) Implied coordination friction κ with cumulative Hugging Face downloads as ecosystem breadth proxy.

5.4.5 Limitations

This exercise bounds rather than structurally estimates the R_0 parameters. The OpenRouter series covers only 14 months with 8 observations—sufficient for bounding but not for formal time-series inference. The logistic-SIR specification imposes functional form assumptions; alternative adoption models (Bass diffusion, threshold models—see Section 3.9.1 for the formal relationship) would yield different implied parameters, though the qualitative trajectory (R_0 rising, κ declining) is robust to specification. A more rigorous test awaits longer time series and, crucially, direct measurement of distributed (edge) inference volumes—data that does not yet exist at the granularity required but that the model's predictions are designed to be tested against.

5.5 The Demand Shock as Nash Overinvestment

AI datacenter demand now absorbs approximately 20% of global DRAM wafer capacity and 30–40% of global HBM output (Section 5.5), with 18 new fabs under construction worldwide. Historical precedent—the 1995–96 DRAM cycle, the 2006–07 NAND expansion, the 2017–18 server DRAM cycle—predicts overcapacity and below-trend pricing within 2–3 years of full capacity ramp. 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, amplified by the sovereign subsidy race (Section 5.5.1).

The 2025–26 DRAM supercycle—now recognized as the most severe memory supply shortage since the Wikipedia-designated "2024–2026 global memory supply shortage"—is the input cannibalization mechanism (Proposition 2) operating through the supply side. The boom phase does not merely delay the crossing—it *reverses* the consumer cost trajectory and shrinks the edge-AI-capable device population, even as it finances the packaging capacity expansion that will eventually enable crossing. This is the two-channel decomposition of equation (9): the anti-crossing channel (supply denial, $dK_C/dt < 0$) currently dominates the pro-crossing channel (learning-curve progress, $\sum q_i > 0$).

The duration is the critical empirical question. Corollary 4's estimate of 1–2 years of delay corresponds to the *lower bound*: the fab construction lag Δ_K under the revenue-maximization specification where centralized demand moderates on schedule. But the post-2022 capex data (Table 11) show overinvestment ratios of 11–19×, consistent with the option-value specification where $M_{\text{eff}} = M + p \cdot V_{\text{ASI}}$. Under sustained ASI belief, Proposition 2(iii) applies: the supply-denial window Δ_{IC} extends until capacity growth outpaces demand growth. Samsung

and SK Hynix have signaled caution toward aggressive expansion, expecting the "memory super-cycle" to stretch past 2028 (TrendForce, December 2025). Memory manufacturers are demanding three-year prepaid capacity commitments—unprecedented in the industry's history. If capability demonstrations continue to sustain $p > 0$—and the reasoning-model advances of 2024–25 and the agentic-AI wave of 2026 suggest they will—the supply-denial window plausibly extends to $\Delta_{\text{IC}} \approx 5$–10 years, corresponding to the duration of the ASI investment episode rather than the construction cycle alone.

The implication for the crossing timeline is asymmetric. The *packaging learning curve* continues to advance—every HBM unit produced is θ units of cumulative packaging experience (the stock-flow asymmetry in the proof of Proposition 2(ii)). What is frozen is the *consumer supply dimension*: the distributed paradigm cannot access the memory it needs to deploy at volume. The crossing will occur not when the learning curve reaches parity—that may happen on the original schedule—but when consumer memory supply is restored. Industry projections: new fabs (SK Hynix Yongin, Samsung P5, Micron Boise) reach volume production 2027–2028, with full normalization projected for 2029–2030. At that point, the accumulated learning-curve progress will be released as a step-function cost decline in consumer stacked memory—a "spring effect" in which years of compressed learning progress are instantiated in consumer products within a single product cycle.

5.5.1 Quantitative Test: Hyperscaler Capex versus Model Predictions

Table 11 compares actual aggregate capex for the core hyperscalers (Amazon, Microsoft, Google, Meta, and from 2024, Apple) against the Nash equilibrium and cooperative optimum paths derived from the calibrated differential game ($\alpha = 0.23$, $\delta = 0.30$, $r = 0.05$).[6]

Table 11: Hyperscaler capex: actual versus model predictions ($B).

Year	N	Actual	Nash	Coop	Actual/Coop
2018	4	64.0	88.4	22.8	2.80×
2019	4	69.4	88.4	22.7	3.06×
2020	4	88.4	88.4	22.4	3.94×
2021	4	119.2	88.4	22.1	5.40×
2022	4	145.1	88.4	21.7	6.69×
2023	4	136.9	88.4	21.3	6.43×
2024	5	223.1	88.4	18.5	12.05×
2025	5	336.0	88.3	17.4	19.31×

Three findings emerge. First, the pre-AI period (2018–2020) shows overinvestment ratios of 2.8–3.9×, within the 3–4× range predicted by Proposition 1 for the basic N-firm game.

[6]Scripts and replication data at `scripts/test_capex_overinvestment.py`.

The model fits the pre-AI data well: firms competed over cloud infrastructure and inference revenue with a well-defined market size, and the Nash equilibrium of the learning-curve game accurately predicted the magnitude of excess investment.

Second, the post-ChatGPT period (2022–2025) shows dramatically higher ratios (average 11.1×, peak 19.3×), far exceeding both the cooperative optimum and the Nash equilibrium path. The divergence has a clean explanation: the effective prize changed. The basic model assumes firms compete over a finite inference revenue pool M (projected ∼\$255B by 2030). But beginning in late 2022, the credible possibility of artificial superintelligence (ASI) transformed the game from a learning-curve competition into a tournament for a potentially unbounded prize.

Remark 4 (The Superintelligence Option)**.** *If firm i assigns subjective probability $p_i > 0$ to achieving ASI—a system that can substitute for human cognitive labor across* all *economic activity—then the effective prize is not the inference market M but*

$$M_{\text{eff}} = M + p_i \cdot V_{\text{ASI}}$$

where V_{ASI} is the present value of capturing a substantial share of global economic output (∼\$100T GDP annually). Even modest beliefs ($p_i \in [0.05, 0.20]$) yield M_{eff} one to two orders of magnitude larger than M, because $p_i \cdot V_{\text{ASI}} \gg M$. The Nash overinvestment ratio from Proposition 1 scales with the effective prize: replacing M with M_{eff} in the equilibrium condition multiplies the predicted overinvestment ratio by $M_{\text{eff}}/M \approx 3$–$8\times$. This transforms the predicted range from 3–4× to 9–32×, which brackets the observed 11–19× ratios.

Whether or not ASI is achievable is irrelevant for the investment dynamics. What matters is the belief*: as long as firms assign positive probability to a prize that dwarfs the inference market, the rational overinvestment level is far higher than the basic learning-curve game predicts. The pre-2022 data (when ASI was not a credible near-term prospect) fits the basic model; the post-2022 data (when ASI became a credible possibility) fits the augmented model with the superintelligence option. The structural break in overinvestment ratios at 2022 is itself evidence that the effective prize changed.*

This observation *strengthens* the thesis rather than weakening it. The endogenous decentralization mechanism operates through overinvestment financing learning curves that enable distributed alternatives. If the effective prize includes a superintelligence option, the overinvestment is larger, the learning curves are financed faster, and the crossing to distributed inference occurs sooner. The firms chasing ASI are—regardless of whether they achieve it—financing the hardware cost declines that make edge inference viable. The stronger the ASI belief, the faster the crossing.

Third, the model correctly predicts *acceleration with entry*: year-over-year capex growth averaged 63.0% when effective N increased (2023→2024, when Apple entered above the \$5B threshold), versus 22.9% when N was stable. This is consistent with Corollary 1 (crossing-time acceleration with N) and the model's prediction of 79.3% acceleration at $N = 5$. Strategic complementarity in capex levels is significant: in a level regression of own capex on rival capex (lagged), $\hat{\beta} = 0.40$ ($t = 4.48$, $p < 0.001$, $R^2 = 0.44$).

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 that distinguishes the endogenous decentralization mechanism from standard learning-by-doing. Section 5.1 presents the full identification strategy: treatment (compute-constrained) versus control (unconstrained) firms, five competing predictions that all resolve in the direction predicted by endogenous decentralization, and threats to validity. 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

The cross-technology consistency constitutes an informal meta-analytic stability test. Six independently estimated early-stage learning curves from four industries (semiconductors,

Table 12: 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 [11]
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.

solar, batteries, cloud computing), spanning different firms, countries, and decades, cluster in a 4-percentage-point band ($\alpha \in [0.21, 0.25]$). A Cochran Q-test for heterogeneity across the five spillover-inclusive estimates (excluding the Goldberg et al. firm-node and the Irwin & Klenow IV estimates, which measure different objects) fails to reject homogeneity ($Q = 2.1$, $p = 0.72$). The stability claim for the packaging α thus rests not on a single short time series but on the structural regularity of early-stage process learning rates across technologies.

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.

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 ≥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 250–400% 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. The dynamics of this transition are developed further in Chapter 5, which models the first-order phase transition to a mesh economy once R_0 crosses unity.

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. Chapter 6 develops the settlement feedback dynamics that this prediction tests.

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. [7]) are irrelevant to this test. This prediction is structured as a pre-registered out-of-sample stability test: as the HBM series extends beyond $n = 6$, formal break-point detection (Bai-Perron with unknown breakpoints) becomes feasible; by 2028, the series will have sufficient observations for a two-regime test at conventional significance levels.

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 (>100,000 synchronized GPUs, >7 days continuous operation) remains exclusively performed in centralized clusters. As of Q1 2026, the trend is toward *larger* clusters (xAI targeting 1–2M GPUs by late 2026), reinforcing rather than weakening this prediction. Evidence against: distributed frontier training at comparable cost and performance by 2035.

8. The Capability Continuum

A reader of this paper will ask: what if the firms are right? What if the $2.4 trillion produces not a bubble but a genuine and sustained advance in AI capability? And what does that imply for the crossing?

The question is natural, but its framing is misleading. The paper has treated artificial superintelligence as a discrete event—a prize V_{ASI} that is either captured or not, with firms assigning subjective probability $p > 0$ to its achievement. This framing, while useful for the option-value analysis (Remark, Section 3.7), obscures the more likely scenario: AI capability advances not as a discontinuous jump but as a *continuum*. Each generation of models is more capable than the last—GPT-2 to GPT-3 to GPT-4, o1 to o3, DeepSeek R1 to its successors—with each advance demonstrating qualitatively new capabilities (translation, reasoning, coding, mathematical proof, autonomous agent behavior) that were absent one generation prior. The relevant question is not "is ASI achieved?" but "how long does capability improvement continue?" If the answer is decades—as it has been for semiconductors, solar cells, and batteries—then the implications for the crossing timeline are profound.

8.1 The Continuum Regime

Define the AI capability frontier $A(t)$ as a continuous, non-decreasing function of cumulative investment and research effort. Under the *continuum regime*:

- ASI belief $p(t)$ is a non-decreasing function of $A(t)$: each capability advance makes the next advance more credible.
- The effective prize $M_{\text{eff}}(t) = M + p(t) \cdot V_{\text{ASI}}(A(t))$ is non-decreasing, because both p and the perceived value of higher capability grow with demonstrated performance.
- Centralized demand $D_H(N, M_{\text{eff}}(t))$ is therefore non-decreasing: each capability milestone renews the investment cycle.

Corollary 5 (Supply denial under the capability continuum)**.** *Under the continuum regime ($dA/dt > 0$, $dp/dA \geq 0$), the supply-denial condition $K_C < K_{\min}$ from Proposition 2 persists until fab capacity growth outpaces the induced demand growth:*

$$\frac{dK}{dt} > \theta \cdot \frac{\partial D_H}{\partial M_{\text{eff}}} \cdot \frac{dM_{\text{eff}}}{dt} \tag{22}$$

If AI capability growth is sustained for decades, then Δ_{IC} is measured in decades, not years. The packaging learning curve continues to advance throughout (every HBM unit produced contributes to cumulative Q), accumulating a progressively larger stock of unrealized consumer cost reduction.

Proof. Under the continuum regime, $M_{\text{eff}}(t)$ is non-decreasing, so $D_H(t)$ is non-decreasing. The condition $K_C(t) = K(t) - \theta D_H(t) \geq K_{\min}$ is restored only when $dK/dt > \theta \cdot dD_H/dt$. If each capability demonstration increases p or V_{ASI}, the right-hand side of equation (22) is itself growing, creating a moving target for capacity expansion. The condition resolves when either $dA/dt \to 0$ (capability plateau) or dK/dt exceeds $\theta \cdot dD_H/dt$ through sustained fab investment. □

Calibration. Current DRAM industry capacity grows at 10–15% per year. AI-driven DRAM demand grew ∼35% in 2025–2026, with the gap between supply growth (23%) and demand growth (35%) widening (TrendForce). If these rates persist, the condition in equation (22) is not met, and supply denial continues indefinitely. Committed memory fab investment ($430 billion across Samsung, SK Hynix, and Micron) will, if fully deployed, more than double global DRAM capacity by 2030—but the continuum regime implies demand may also double, absorbing the new capacity as it arrives. Samsung and SK Hynix have signaled caution on expansion precisely because they expect the "memory super-cycle" to stretch past 2028; memory manufacturers are demanding three-year prepaid capacity commitments (unprecedented). The Phison CEO projects the shortage persisting a decade or more under sustained AI demand.

This reframes the paper's timing predictions. The hardware crossing (Prediction 2, ∼2028–2029) and the distributed $R_0 > 1$ threshold (Prediction 2*, 2030–2032) are *conditional on supply restoration*: they hold if and only if $dK/dt > \theta \cdot dD_H/dt$ is restored within the prediction window. Under the capability continuum, those dates are lower bounds. The predictions' falsification criteria—"not achieved by 2031" for Prediction 2, "stalling below 20% by 2033" for Prediction 2*—should be interpreted accordingly: failure to meet these dates is consistent with both mechanism failure *and* supply-denial persistence under the continuum regime. The distinguishing observable is the packaging learning curve (Prediction 5): if α remains in $[0.18, 0.28]$ but consumer memory supply remains constrained, the mechanism is operating but the supply dimension of the two-dimensional crossing (Proposition 2) has not been restored.

8.2 Resolution Pathways

Four mechanisms can restore the supply condition $K_C \geq K_{\min}$ even under sustained capability growth.

(1) Capacity catches up. Fab capacity growth can accelerate beyond 10–15% per year if investment is sufficiently sustained. The $430B in committed memory investment is the largest expansion in the industry's history. If demand growth decelerates even modestly

(from 35% to 20%), existing commitments close the gap by 2029–2030. Historical precedent: every prior semiconductor supercycle (1995–96, 2006–07, 2017–18) resolved into overcapacity within 3–5 years, because investment commitments made at peak demand deliver capacity into moderating markets. The question is whether the AI capability continuum represents a structural break from this pattern.

(2) Algorithmic efficiency eliminates the memory bottleneck. The 280× inference cost decline documented over 2022–2024 is a *software-side* learning curve operating in parallel with the hardware packaging curve. If MoE, quantization, distillation, speculative decoding, and future compression techniques reduce memory requirements by another 100×, then 4GB devices can run models at quality levels that currently require 64GB. This is the convergence-from-above channel (Section 5.3) operating at the extreme rate needed to bypass the supply constraint entirely—reducing $K_{\min}$ rather than increasing K_C. The export-control natural experiment provides evidence that constraint-induced efficiency optimization operates at exactly these rates: the binding compute constraint on Chinese firms produced efficiency gains that closed the frontier gap within two years.

(3) Technological discontinuity. If AI compute migrates to a memory technology that does not share wafer capacity with consumer DRAM—processing-in-memory, optical interconnects, or a fundamentally new architecture—the wafer multiplier θ ceases to apply and the shared-input dependency dissolves. Hailo's DRAM-free edge AI accelerators, which keep the entire inference pipeline on-chip, are a nascent example: they eliminate external DRAM dependency entirely for vision workloads, though not yet for generative AI. The transition from ferrite core memory to semiconductor DRAM in the 1970s is historical precedent for precisely this type of discontinuity.

(4) The capability plateau. All known technology improvement curves eventually encounter diminishing returns. If AI capability growth decelerates ($d^2A/dt^2 < 0$)—due to data exhaustion, architectural limits, or energy constraints (Section 5.5.2)—then $p(t)$ stabilizes, M_{eff} flattens, and the supply condition is restored by continuing capacity expansion. The power constraint (Section 5.5.2) may be the binding ceiling: five gigawatt-scale facilities are coming online in 2026, but the electrical grid cannot indefinitely absorb exponential power growth. The PJM Interconnection already projects a 6 GW shortfall by 2027. If power, not memory, becomes the binding constraint on centralized AI scaling, it would relax memory demand and resolve the supply denial through a different channel than the paper models.

8.3 Implications for the Thesis Framework

The capability continuum does not invalidate the endogenous decentralization mechanism—it extends its timescale. The core result ($\partial T^*/\partial I < 0$) holds: centralized investment still finances the learning curves that enable distributed alternatives. The packaging learning curve still advances with every HBM unit produced. The stock-flow asymmetry in Proposition 2(ii) ensures that the learning benefit is permanent while the supply constraint is temporary—even if "temporary" means decades.

What changes is the intermediate dynamics. The crossing is delayed. The accumulated learning "spring" becomes more compressed: when consumer supply is eventually restored, the cost decline in consumer stacked memory will be more dramatic—potentially a step-function drop in which a decade or more of packaging learning progress is instantiated in consumer products within a single cycle. Consumer welfare during the supply-denial window is substantially worse: phones losing memory, PC prices rising, edge AI frozen. The installed base of edge-AI-capable devices shrinks for years before recovering.

The prediction that is unambiguously *strengthened* by the continuum scenario is Prediction 9: training remains centralized through 2035 and potentially far beyond. Under sustained capability growth, the incentive to build ever-larger training clusters only increases. The partial-decentralization equilibrium—inference distributing while training centralizes—remains the stable long-run outcome, but the inference distribution component is delayed while the training centralization component is reinforced.

Within the thesis's four-level hierarchy (Chapter 3 [20]), the capability continuum implies that Level 1 (hardware) evolves more slowly toward its crossing condition than the baseline model predicts, extending the hierarchical ceiling that bounds all faster levels. Mesh formation (Level 2), autocatalytic training dynamics (Level 3), and settlement feedback (Level 4) are correspondingly delayed at the hardware pathway—not because their internal dynamics are slower, but because the hardware level's activation condition ($R_0 > 1$) takes longer to achieve. Alternative activation pathways—software-only solutions that bypass the memory constraint, or cross-level amplification (Chapter 3, Section 4.3) that activates the system even when individual levels are sub-threshold—become more important under the continuum scenario.

The mechanism is robust across all four resolution pathways. Whether supply denial ends through capacity expansion, algorithmic efficiency, technological discontinuity, or capability plateau, the crossing eventually occurs and the accumulated learning progress is released. What the continuum analysis adds is honest uncertainty about *when*. The paper's predictions (Section 7) are conditioned on supply restoration within the prediction window; the capability continuum admits the possibility that this condition is not met for a decade or more, without

altering the mechanism's long-run validity.

9. 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 supply crisis—the most severe memory shortage in decades, with DDR5 prices 250–400% above mid-2024 levels, smartphones losing memory (reverting from 8GB to 4GB base RAM), and Micron exiting the consumer market entirely—reveals a mechanism stronger than the boom-bust cycle Corollary 4 describes. Proposition 2 (input cannibalization) formalizes the two-dimensional crossing: self-sustaining distributed adoption requires both cost parity on the packaging learning curve *and* sufficient consumer memory supply. The crossing is currently frozen on the second dimension. The centralized investment that finances the packaging learning curve simultaneously monopolizes memory wafer capacity—HBM consumes 3–4× the wafer capacity of standard DRAM per gigabyte, and profit margins 5–10× higher ensure rational manufacturers maximize HBM allocation. The result: edge AI progress is not merely delayed but actively denied the physical input it needs. Phones are getting *worse*—the first backward spec movement in the smartphone era—and the installed base of edge-AI-capable devices is shrinking.

The duration of this supply denial depends on whether the investment regime is governed by revenue maximization ($p = 0$, delay ∼2–3 years) or the superintelligence option ($p > 0$, delay potentially 5–10 years). The post-2022 capex data—overinvestment ratios of 11–

19×—are consistent only with the second regime. If capability demonstrations continue to sustain positive ASI belief, demand growth can absorb new fab capacity as fast as it arrives, extending the supply-denial window to the duration of the ASI investment episode. The packaging learning curve continues to advance through this period (every HBM unit produced is cumulative packaging experience), accumulating years of compressed learning progress that will be released as a step-function cost decline in consumer memory when the supply constraint lifts.

The response validates the Nash overinvestment mechanism at both firm and sovereign levels: hyperscaler capex approaching $650B in 2026E, a $200 billion government subsidy race across seven nations, 18 new fabs under construction, and five gigawatt-scale AI data centers coming online. The global AI compute stock of ∼15 million H100-equivalents is doubling every seven months. The operative learning curve is the packaging process, not the die; and the capacity being expanded—now including HBM4 at 16-high stacking entering mass production in 2026—is packaging capacity that will pivot to consumer stacked memory formats when the supply-denial condition resolves.

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 and increasingly also power-constrained—training clusters are scaling toward 500,000+ GPUs and gigawatt-scale power, not distributing. Meanwhile, 370 million NPU-equipped smartphones shipped in 2025, but on-device AI use is confined almost entirely to computational photography; generative inference remains a cloud activity. Only high-RAM Apple M-series Macs can run 30–70B models locally, and no killer application has emerged to drive adoption. The coordination layer (software, runtimes, model distribution) remains immature, the DRAM supercycle constrains memory-intensive configurations, and R_0 for distributed generative inference remains firmly in the sub-critical regime (∼0.4–0.8). 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.

Within the thesis framework, this chapter establishes the slowest-timescale driver of the four-level hierarchy. The overinvestment result and the packaging learning curve determine the pace at which Level 1 evolves; all subsequent dynamics—mesh formation (Chapter 5), autocatalytic training capability growth, and settlement feedback (Chapter 6)—are bounded above by this rate through the hierarchical ceiling mechanism (Chapter 3, Proposition 8.1). The crossing condition derived here ($R_0 > 1$) is the Level 1 instance of the spectral activation threshold that governs each level of the hierarchy.

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 13: 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.

D. Structural Breaks in the DRAM Die Learning Curve

Bai-Perron sequential testing on the 41-year DRAM die series (log($/GB) on log(cumulative production), 1984–2024) identifies two structural breaks:

Table 14: Bai-Perron structural break results: DRAM die series.

Regime	Period	α	SE	n
1	1984–1994	0.39	0.05	4
2	1995–2007	1.15	0.12	3
3	2008–2024	0.38	0.06	3
Full sample	1984–2024	0.66	0.04	9

Sequential sup-F test: break at 1995 significant at 1% ($F = 18.3$); break at 2008 significant at 5% ($F = 9.7$). Critical values from Bai and Perron [26]. Regime 2 estimate ($\alpha = 1.15$) is implausible as a learning parameter and reflects the 1995–2001 DRAM price collapse driven by Asian financial crisis overcapacity and the subsequent demand recovery.

Two features are relevant. First, the bookend regimes yield $\alpha = 0.38$–0.39, consistent with the Irwin and Klenow [15] IV estimate of $\alpha = 0.32$ (SE $= 0.05$) after accounting for upward OLS bias. Within-regime learning rates are stable *across boom-bust cycles*; the instability is in regime transitions driven by demand-side shocks. Second, the die series instability strengthens the paper's reframing: extrapolating a single α from a 41-year series with two structural breaks is unreliable, which is precisely why the operative curve should be the early-stage packaging process where the learning dynamics are physically interpretable and the demand-side feedback channel is limited.

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] Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5), 215–227.

[4] Bemmaor, A. C. (1994). Modeling the diffusion of new durable goods: Word-of-mouth effect versus consumer heterogeneity. In G. Laurent, G. L. Lilien, & B. Pras (Eds.), *Research Traditions in Marketing* (pp. 201–229). Kluwer.

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

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

[7] Carlino, A., Wongel, A., Duan, L., Virguez, E., Davis, S. J., Edwards, M. R., & Caldeira, K. (2025). Variability of technology learning rates. *Advances in Applied Energy*, 20, 100252.

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

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

[10] Dodds, P. S., & Watts, D. J. (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21), 218701.

[11] Flamm, K. (1996). *Mismanaged Trade? Strategic Policy and the Semiconductor Industry*. Brookings Institution Press.

[12] Goldberg, P. K., Juhasz, R., Lane, N. J., Lo Forte, G., & Thurk, J. (2024). Industrial policy in the global semiconductor sector. NBER Working Paper No. 32651.

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

[14] Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.

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

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

[17] Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, 29(4), 741–766.

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

[19] Smirl, J. (2026a). The CES triple role: Superadditivity, correlation robustness, and strategic independence as three views of isoquant curvature. Working Paper.

[20] Smirl, J. (2026b). Complementary heterogeneity in hierarchical economies: CES aggregation, derived architecture, and cross-sector activation in multi-timescale systems. Working Paper.

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

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

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

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

[25] Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton.

[26] Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1–22.