ENDOGENOUS DECENTRALIZATION AND THE AI TRANSITION

From Concentrated Investment Through Learning Curves to Self-Organizing Distributed Intelligence

Jon Smirl

Independent Researcher

February 2026

Working Paper

Abstract

This paper identifies, formalizes, and traces the complete arc of endogenous decentralization as applied to the current AI transition. The mechanism's distinctive property is $\partial T^*/\partial I < 0$: concentrated investment in centralized AI infrastructure finances the learning curves that enable distributed alternatives. The arc proceeds in six stages. First, concentrated capital investment by competing firms in symmetric Markov Perfect Equilibrium produces aggregate output that strictly exceeds the cooperative optimum (Proposition 1), accelerating the crossing time beyond what any firm would individually prefer. Second, the operative learning curve is 3D memory stacking and advanced packaging ($\alpha = 0.23$), not mature planar DRAM die fabrication. Third, the crossing condition generalizes from pure cost parity to a self-sustaining adoption threshold: $R_0 > 1$. Fourth, the training-inference bifurcation produces partial decentralization: inference distributes while training remains centralized. Fifth, after crossing, a self-organizing mesh of heterogeneous specialized agents forms via a first-order regime shift (Potts crystallization, not gradual adoption), whose CES diversity premium enables the

mesh to exceed centralized provision above critical mass N^*. Sixth, the mesh develops endogenous capability growth through autocatalytic training (RAF sets), but converges to the Baumol bottleneck—the exogenous rate of frontier model improvement. The self-undermining theorem establishes this pattern as a mathematical necessity from CES complementarity, Wright's Law learning, and information frictions. Historical calibration against five technology transitions spanning 200 years confirms the framework's quantitative predictions.

Keywords: endogenous decentralization, learning curves, Markov Perfect Equilibrium, regime shift, CES aggregation, mesh equilibrium, autocatalytic sets, Baumol cost disease, technology cycles, AI infrastructure

JEL: O33, O41, L16, D43, D85, C73

1. Introduction

Every major technology follows a recognizable arc. Initial development requires massive concentrated investment—canals demanded sovereign financing, railroads required stock markets, electrification needed utility monopolies, semiconductors emerged from defense procurement, and AI training requires hyperscaler-scale capital. The concentrated phase generates returns that attract overinvestment, culminating in crisis. After the crisis, the infrastructure built during the boom enables distributed adoption at dramatically lower cost.

Between 2018 and 2025, the five largest US technology companies—together with Oracle and the Stargate joint venture—committed an estimated $1.3 trillion in cumulative capital expenditure to construct centralized AI infrastructure. This represents the largest concentrated infrastructure investment in history outside wartime mobilization. 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 paper's scope is deliberately broad: it traces the complete arc from pre-crossing investment dynamics through post-crossing mesh formation, autocatalytic growth, and the Baumol ceiling. This breadth sacrifices depth in any single mechanism—the differential game, the mesh equilibrium, and the autocatalytic framework each merit dedicated treatment—but the central contribution is precisely the demonstration that these mechanisms compose into a single, coherent, falsifiable narrative.

This paper derives the self-undermining pattern as a *mathematical necessity* from three primitives that individually have extensive empirical support:

(i) A CES production technology with curvature parameter $K = (1-\rho)(J-1)/J$ (Paper 1; Smirl 2026a).

(ii) Learning-by-doing that reduces unit cost as $c(Q) = c_0 Q^{-\alpha}$ (Wright's Law), with $\alpha \in [0.15, 0.35]$ empirically (Wright 1936; Nagy et al. 2013).

(iii) Information frictions parameterized by $T = 1/\kappa$ (inverse information capacity), under which the exploitable curvature is $K_{\text{eff}} = K \cdot (1 - T/T^*(\rho))^+$ (Paper 1; Smirl 2026a).

The central insight is that these three primitives interact to produce a *self-undermining dynamic*: concentrated investment accelerates learning, learning reduces cost, lower cost reduces the information friction required for distributed coordination, and lower information friction makes distributed alternatives viable. But the story does not end at the crossing

point. The post-crossing dynamics proceed through mesh formation, endogenous capability growth, and convergence to a Baumol ceiling—completing a circle in which concentrated investment determines both when crossing occurs and the ceiling the distributed ecosystem approaches.

Two structural features of the current AI landscape sharpen the mechanism beyond what prior transitions exhibited. First, AI workloads bifurcate into *training* and *inference*. The endogenous decentralization mechanism applies directly and powerfully to inference, which already constitutes 80–90% of AI compute cycles. Training may remain permanently centralized—a structural rather than temporal bifurcation. Second, the effective crossing threshold is being approached from two directions simultaneously: hardware costs declining from below and algorithmic efficiency reducing the threshold from above. The dual convergence compresses the crossing timeline relative to what hardware cost decline alone would produce.

Relation to Existing Literature

The paper builds on and synthesizes several literatures. Arrow (1962) establishes learning-by-doing as a source of cost reduction, but within the same paradigm and for the same firms. Bresnahan and Trajtenberg (1995) formalize general-purpose technologies with cross-sector spillovers. Schumpeter (1942) identifies creative destruction but attributes it to external entrants. Christensen (1997) describes disruption from below in new value networks. The present paper unifies these insights: the learning occurs in the same paradigm (Arrow), benefits a different architecture (Bresnahan-Trajtenberg), and the disruption is self-financed rather than externally sourced—a stronger result than any of the predecessors.

Perez (2002) provides the most detailed empirical description of technology cycles. The present paper derives her five phases from bifurcation dynamics rather than positing them. The turning point is identified as a fold bifurcation, explaining its discontinuous character. Kondratiev (1925) measures the long-wave periodicity; the duration formula explains both the periodicity and its compression.

The AI-specific component connects to Aghion, Jones, and Jones (2018) on AI and economic growth, Nordhaus (2021) on the singularity hypothesis, and Bloom et al. (2020) on declining research productivity. The mesh formation draws on Becker and Murphy (1992) for division of labor, while the autocatalytic framework adapts Hordijk and Steel (2004) from origin-of-life theory. The CES framework is developed in the companion paper (Paper 1; Smirl 2026a).

On the network science side, the paper builds on three classical results. Bianconi and Barabási (2001) provide the fitness-dependent preferential attachment model that drives

inverse capability concentration. The Fortuin-Kasteleyn (1972) cluster expansion unifies percolation and Potts magnetization, revealing that network formation and specialization are the same mathematical object at different parameter values. Pastor-Satorras and Vespignani (2001) establish the vanishing epidemic threshold on scale-free networks, which guarantees self-sustaining knowledge propagation once the mesh achieves a fat-tailed degree distribution.

The paper is organized as follows. Section 2 presents the self-undermining theorem in general form. Section 3 develops the N-firm differential game and the overinvestment result. Section 4 identifies the operative learning curve. Section 5 derives the generalized R_0 crossing condition. Section 6 establishes the training-inference bifurcation. Section 7 presents the export-control natural experiment. Section 8 documents dual convergence. Section 9 formalizes post-crossing mesh formation via the Potts regime shift. Section 10 develops the CES diversity premium and specialization dynamics. Section 11 characterizes knowledge diffusion via the graph Laplacian. Section 12 derives autocatalytic capability growth. Section 13 proves the model collapse protection theorem. Section 14 derives the Baumol bottleneck. Section 15 calibrates against historical Perez phases. Section 16 discusses frameworks considered and rejected. Section 17 presents combined predictions. Four appendices provide a two-period pedagogical model, the Weitzman recombinant growth connection, the Nordhaus singularity analysis, and empirical calibration details.

Preliminaries: CES Foundations

This paper builds on the CES framework developed in the companion paper (Paper 1; Smirl 2026a). The CES aggregate $F_n = (J^{-1}\sum_j x_{nj}^{\rho})^{1/\rho}$ with curvature parameter $K = (1-\rho)(J-1)/J$ simultaneously controls superadditivity, correlation robustness, and strategic independence. The CES potential $\Phi = -\sum_n \log F_n$ serves as the generating function for the hierarchical dynamics. Under information frictions parameterized by T, the exploitable curvature is:

$$K_{\text{eff}} = K \cdot \left(1 - \frac{T}{T^*(\rho)}\right)^+ \tag{1}$$

where $T^*(\rho)$ is the breakdown threshold. These definitions are not re-derived here; the reader is referred to Paper 1 for proofs and detailed development. The present paper applies this framework to the specific dynamics of the AI transition.

Notation summary. The following notation is used throughout the paper. Table 1 collects the principal symbols for reference.

Table 1: Principal notation.

Symbol	Definition	First appearance
ρ	CES substitution parameter ($\sigma = 1/(1-\rho)$)	Eq. (1)
K	CES curvature: $(1-\rho)(J-1)/J$	Eq. (1)
K_{eff}	Effective curvature under friction T	Eq. (1)
T, T^*	Information friction; breakdown threshold	Eq. (1)
α	Wright's Law learning elasticity	Eq. (2)
$Q(t)$	Cumulative production	Section 2.1
$x(t)$	State variable: $\bar{Q}^*_{\text{eff}} - Q(t)$	Section 3.1
T^*	Crossing time (first t with $x(t) = 0$)	Section 2.2
$q_i(t), Q(t)$	Firm i's output rate; total output $\sum q_i$	Section 3.1
$V^N(x), V^P(x)$	Nash / cooperative value functions	Sections 3.2–3.3
S, S_T, S_I	Continuation value; training / inference components	Section 3.1
R_0	Basic reproduction number	Eq. (16)
$\beta, \gamma, \kappa, \mu$	Adoption rate, network effect, friction, churn	Section 5.2
R_0^{mesh}	Mesh reproduction number	Section 9.1
S_∞	Giant component fraction	Eq. (28)
N^*	Critical mass for mesh dominance	Theorem 3
$C_{\text{eff}}, C_{\text{mesh}}, C_{\text{cent}}$	Effective / mesh / centralized capability	Eq. (30)
J	Number of task/specialization types	Section 10.1
φ_{eff}	Effective training productivity	Eq. (41)
N_{auto}	Autocatalytic existence threshold	Proposition 8
g_Z	Exogenous frontier model improvement rate	Section 14

2. The Self-Undermining Theorem

2.1 The Three-Stage Structure

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

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

$$c(Q) = c_0 \cdot Q^{-\alpha} \tag{2}$$

where α is the learning elasticity—a *technology* parameter, not a *firm* parameter: learning embodied in manufacturing process improvements transfers across applications.

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

2.2 The Self-Undermining 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:

$$\frac{\partial T^*}{\partial I} < 0 \tag{3}$$

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

2.3 Formal Statement

Theorem 1 (Self-undermining)**.** *Let $T_{dist}(Q)$ be the information friction of distributed production when cumulative centralized investment is Q. If the coordination technology improves with the general-purpose technology (so that $T_{dist}(Q) = T_0 \cdot g(c(Q))$ for some increasing function g), then:*

(a) The effective distributed friction falls with investment:

$$\frac{\partial T_{dist}}{\partial Q} = T_0 \cdot g'(c) \cdot c'(Q) < 0 \tag{4}$$

since $g' > 0$ (higher cost means higher friction) and $c' < 0$ (Wright's Law).

(b) The gap between T_{dist} and T^* closes at rate:

$$\frac{d}{dQ}\left[\frac{T_{dist}(Q)}{T^*(\rho)}\right] = -\frac{\alpha \cdot T_0 \cdot g'(c_0 Q^{-\alpha})}{Q \cdot T^*(\rho)} < 0 \tag{5}$$

(c) The crossing cumulative production Q^* (where $T_{dist}(Q^*) = T^*(\rho)$) satisfies:

$$Q^* = \left(\frac{c_0}{g^{-1}(T^*/T_0)}\right)^{1/\alpha} \tag{6}$$

which is finite whenever $T^*/T_0 \in range(g)$.

(d) ***Self-undermining:*** *In the Nash equilibrium, cumulative investment reaches Q^* in finite time. The centralized structure that financed the learning curve has created the conditions for distributed entry.*

Proof. Part (a) follows directly from the chain rule and the signs of g' and c'. Part (b) differentiates the ratio and substitutes $c(Q) = c_0 Q^{-\alpha}$. Part (c) inverts the crossing condition $T_0 \cdot g(c(Q^*)) = T^*$ and substitutes Wright's Law. Part (d): by Theorem 2 below, Nash investment is strictly positive and exceeds the cooperative rate, so $Q(t) \to \infty$ as $t \to \infty$. Since $Q^* < \infty$, crossing occurs in finite time. □

The self-undermining property operates through three reinforcing channels. The *cost channel*: centralized investment drives cumulative production, which reduces unit cost via Wright's Law, reducing minimum efficient scale for distributed producers. The *infrastructure channel*: the infrastructure built for centralized production (communications networks, software platforms) is typically a general-purpose input that distributed producers can also use. The *information channel*: as the technology matures, its operating procedures become codified, reducing the information capacity required for competent operation.

Corollary 1 (Irreversibility of crossing)**.** *Once $T_{dist}(Q) < T^*(\rho)$, the distributed mode is viable and cannot be rendered non-viable by further centralized investment. The crossing is a one-way gate.*

Proof. Further centralized investment increases Q, which further reduces $T_{\text{dist}}(Q)$. The ratio T_{dist}/T^* is monotonically decreasing in Q. □

2.4 The Three Channels

The self-undermining property operates through three reinforcing channels, each independently sufficient for crossing given enough cumulative production.

Cost channel. Centralized investment drives cumulative production, which reduces unit cost via Wright's Law. Lower cost reduces the minimum efficient scale for distributed producers, since the fixed cost γ of peer coordination is repaid faster when variable costs are lower. For AI: hyperscaler GPU purchases finance HBM manufacturing scale, which reduces per-GB memory cost, which makes on-device inference affordable. The cost channel is the most direct and quantitatively dominant mechanism.

Infrastructure channel. The infrastructure built for centralized production—communications networks, logistics systems, financial instruments, software platforms—is typically a general-purpose input that distributed producers can also use. Railroad infrastructure enabled small manufacturers to reach national markets. Internet infrastructure built for corporate intranets enabled peer-to-peer commerce. Cloud computing built for hyperscale training enables distributed inference. For AI: the software stack (PyTorch, ONNX, quantization libraries), model compression techniques (distillation, pruning), and deployment infrastructure (container runtimes, model hubs) developed for centralized operation transfer directly to edge deployment.

Information channel. As the technology matures, its operating procedures become codified, reducing the information capacity required for competent operation. A new technology requires expert judgment (T high); a mature technology can be operated by following documented procedures (T low). For AI: the progression from research-paper-only model releases (2018–2022) to one-click deployment on consumer hardware (2024–2025) exemplifies declining T_{dist}. The HuggingFace ecosystem, GGUF quantization format, and llama.cpp runtime collectively reduce deployment information friction from months of ML engineering to hours of configuration.

2.5 Distinction from Adjacent Theory

3. The N-Firm Differential Game and Overinvestment

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 remaining cumulative production until the effective crossing threshold. When x reaches zero, inference crossing occurs. The state evolves as:

$$dx/dt = -\sum_i q_i(t) \tag{7}$$

Table 2: Theoretical positioning of endogenous decentralization.

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

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—the formal expression of the self-undermining property.

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

$$\pi_i(t) = (a - bQ)q_i \tag{8}$$

Upon crossing ($x = 0$), each firm receives continuation value:

$$S = S_T + \frac{S_I}{N(r + \delta)} \tag{9}$$

where S_T represents persistent training and model-licensing revenue, S_I is pre-crossing inference profit, r is the discount rate, and $\delta > 0$ is the post-crossing inference displacement rate.

Structural determination of S_T. In the CES framework (Paper 1), training has very low ρ (near-Leontief complementarity: all gradient updates must synchronize) and an enormous information friction gap between centralized ($T_{\text{cent}} \approx 0$, nanosecond NVLink) and distributed ($T_{\text{dist}} \gg T^*$, millisecond WiFi) operation. The quasi-rent from centralization is:

$$S_T \propto K_{\text{eff}}^{\text{cent}} - K_{\text{eff}}^{\text{dist}} = K_{\text{train}} \cdot \frac{T_{\text{dist}} - T_{\text{cent}}}{T^*(\rho_{\text{train}})} \tag{10}$$

which is large and positive as long as the synchronization gap persists.

The game has a *common-pool* structure analogous to the fishery or oil extraction commons (Levhari and Mirman 1980), with the critical distinction that the "resource" being depleted

is the incumbent paradigm's remaining inference viability.

3.2 Markov Perfect Equilibrium

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

$$rV(x) = \max_{q_i}\{(a - b(q_i + (N-1)q(x)))q_i - V'(x) \cdot (q_i + (N-1)q(x))\} \tag{11}$$

The first-order condition under symmetry yields:

$$q^N(x) = \frac{a - V^{N\prime}(x)}{b(N+1)} \tag{12}$$

Substituting back 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 $W(x) = NV^P(x)$:

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

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 $u = \sqrt{D + EV}$, where:

$$D = \frac{a^2}{b(N+1)^2}, \quad E = \frac{r(N+1)^2}{N^2} \tag{13}$$

$$A = \frac{a(N+1)}{N^2 - 1} = \frac{a}{N-1} \tag{14}$$

Under this substitution, the Nash ODE reduces to a first-order separable equation in $u(x)$. Integration from the boundary $u_0 = \sqrt{D + ES}$ yields:

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

Both solutions share the same functional form—$\sqrt{\cdot} + \log$—differing only in the constants governing shadow cost internalization. The Nash constants embed the factor $N^2/(N+1)^2$ reflecting the private share of the social shadow cost; the cooperative constants embed $N/(4bN)$ reflecting full internalization.

Verification. Both analytical solutions are verified against fourth-order Runge-Kutta numerical integration with adaptive step size ($\Delta x = 10^{-4}$). Maximum absolute error: $\max |x_{\text{exact}} - x_{\text{num}}| < 10^{-12}$ across the entire state space $x \in [0, x_0]$ for all tested parameter combinations ($N \in \{2, 3, 5, 10, 20\}$, $S/S_{\max} \in \{0, 0.25, 0.5, 0.75, 1.0\}$). The functional-form match between Nash and cooperative solutions—both are $\sqrt{\cdot} + \log$—is not coincidental: it reflects the shared structure of the underlying Bernoulli ODE, which admits closed-form solutions for all symmetric N-player differential games with linear demand and a common state variable.

3.5 The Overinvestment Result

Theorem 2 (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, 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.

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

Step 3. From the output expressions, both the smaller numerator 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 equilibrium basin—the distributed inference equilibrium—becomes accessible. Reversing the crossing would require cumulative production to decrease, contradicting monotonicity. Once Q crosses $\bar{Q}$, the inference transition is structurally irreversible.*

Economic interpretation. The overinvestment decomposes into two channels:

Cournot channel. Each firm's output depresses the price faced by all rivals. In the standard Cournot game (without learning), this produces aggregate output $Q^N = Na/(b(N+1))$, which exceeds the monopoly quantity but is below the competitive quantity.

Learning externality channel. Each unit of firm i's output simultaneously (a) generates current profit and (b) depletes the remaining distance to crossing. The social shadow cost of the second effect is $NV'(x)$—the aggregate loss across all N firms from moving one unit closer to crossing. But firm i internalizes only its own loss $V'(x)$, producing private shadow cost $= 1/N$ of social shadow cost. This externality is the formal expression of the learning-by-doing spillover: the knowledge embedded in HBM stacking improvements transfers across applications.

The two channels reinforce each other. At baseline calibration ($N = 5$, $S_T = 0$), the per-firm welfare loss under Nash competition is 34.1%. With S_T calibrated to training revenue persistence, the loss moderates to approximately 22–28%. Increasing N amplifies both channels: the Cournot channel intensifies price competition, and the learning externality worsens because each firm's private share of the social shadow cost falls from $1/N$ to $1/(N+1)$.

3.6 Comparative Statics

Corollary 2 (Increasing N)**.** *Nash equilibrium aggregate output is strictly increasing in N for all $x > 0$.*

Corollary 3 (Asymmetric firms)**.** *If firm 1 has marginal cost $c_1 - \varepsilon$, aggregate equilibrium output is strictly increasing in ε.*

Corollary 4 (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 5 (Capacity constraint and boom-bust)**.** *Crossing time delay is bounded by the construction lag Δ for new capacity. The long-run packaging learning rate α is unaffected.*

The 2025–26 DRAM supercycle provides a real-time test. Consumer DDR5 prices have risen 300–400% above trend in under six months, driven by AI datacenter demand reallocating wafer capacity from consumer to HBM formats. The corollary predicts that (a) this deviation is temporary, bounded by the construction lag for new advanced packaging capacity (Samsung P4, SK Hynix M15X, Micron Idaho, TSMC CoWoS expansion), and (b) the packaging learning rate $\alpha = 0.23$ is unaffected because the supercycle is a *demand allocation* shock, not a change in the stacking production function.

Remark 2 (Option-value amplification)**.** *Under the option-value objective function specification, 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.*

Remark 3 (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.*

3.7 Calibration

The learning elasticity $\alpha = 0.23$ is estimated from the HBM packaging learning curve (Section 4). 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. This is the most important parameter uncertainty: a 50% change in α (from 0.23 to 0.15) shifts the crossing date by decades. The table below reports T^* from the hardware learning curve alone, without algorithmic efficiency gains; the dual convergence (Section 8) shifts all dates earlier.

Table 3: Sensitivity of crossing time to learning elasticity.

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

Notes: T^* computed from hardware learning curve only, without algorithmic efficiency gains. Dual convergence (Section 8) shifts all dates earlier.

Post-crossing continuation value. The inference displacement rate $\delta \approx 0.30$ from the IBM trajectory (Section 15.4). Under revenue-maximization: S_T high (closed-model dominance), welfare loss $\sim$22%; S_T moderate (open-weight competition), $\sim$28%; $S_T \approx 0$ (commoditization), $\sim$34%.

Quantitative predictions. Under Nash competition with $N = 5$, crossing at approximately 2028. The 2025–26 DRAM supercycle delays the cost threshold by an estimated 1–2

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

Overinvestment in dollar terms.

Table 4: 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. The $155–326B annual excess represents the learning externality capitalized: each dollar of overinvestment purchases approximately $\alpha \approx 0.23$ log-units of cost reduction per doubling of cumulative production, which accrues to all downstream users. In a standard Cournot model without learning, overinvestment produces allocative inefficiency. Here, overinvestment produces *dynamic efficiency*: the accelerated cost decline benefits the entire ecosystem, including the distributed paradigm that will eventually displace the investing firms. The welfare analysis is therefore ambiguous: Nash competition reduces per-firm profit (34% welfare loss in the pure game) but accelerates the crossing that unlocks the distributed-mode consumer surplus.

4. The Packaging Learning Curve ($\alpha = 0.23$)

4.1 Cost Decomposition: Die versus Packaging

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

Die fabrication (mature, $\alpha \to 0$). Planar DRAM die cost per bit has declined along the Wright curve for over four decades—from $870,000/GB (1984) to approximately $2/GB (2024). At current cumulative production levels (∼3,200 EB through 2024), additional doublings yield marginal cost reductions. A 41-year OLS regression yields $\alpha = 0.66$ (SE = 0.04), but this estimate is inflated by simultaneous equations bias, product-generation transitions, and demand-side shocks (Irwin and Klenow 1994). Piecewise regression identifies structural breaks at 1995 and 2008, with regime-specific estimates of $\alpha = 0.39$, 1.15, and 0.38—the middle regime implausible, the bookend regimes consistent with the Irwin-Klenow IV estimate of 0.32 after accounting for upward OLS bias. Carlino et al. (2025) find structural breaks in 66% of technology learning curves; the DRAM die series is consistent with this pattern.

This is an instance of the *sequential wave* pattern: the semiconductor industry's first technology wave (planar die scaling, 1970s–2010s) has matured, while a second wave (3D stacking and advanced packaging, 2015–present) is opening new learning opportunities. For this paper, the critical observation is that the die cost is no longer the binding constraint or the operative learning curve.

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

Year	Generation	$/GB	Cum. Prod. (EB)	ln(Price)	ln(Cum.)
1984	64Kb	870,000	<0.001	13.68	−11.51
1995	16Mb	30,000	0.10	10.31	−2.30
2005	1Gb	90	17	4.50	2.83
2015	8Gb	3.20	400	1.16	5.99
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). †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 <50μm, 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 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. The Rockchip RK1828 (2025), with its 3D stacked DRAM co-processor running 7B models at 59 tok/s, is a concrete example of this transfer: the stacking techniques were developed for datacenter HBM but deployed on a consumer edge chip.

4.2 The HBM Cost Trajectory

HBM prices declined from $120/GB (2015) to $12/GB (2025). $\alpha = 0.23$ (SE = 0.06, $n = 6$).

TSMC's CoWoS advanced packaging capacity is growing at >50% CAGR from 2022 to 2026, ramping from approximately 35,000 wafers/month (2024) to 75,000 (end 2025) to 130,000 (end 2026). Total industry CoWoS demand is projected at 1 million wafers in 2026,

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

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

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

Table 7: HBM packaging learning curve.

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

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

up from 370,000 in 2024 (Morgan Stanley 2026). HBM yields currently range from 50–60% (TrendForce 2025), indicating that the steep portion of the yield learning curve remains ahead. This is the packaging investment the model tracks—capacity tripling in two years on a process whose yields have not yet matured.

The learning rate $\alpha = 0.23$ captures cost improvement per doubling of cumulative output. At current HBM production rates (∼800M units/year), a doubling of cumulative production occurs approximately every 18 months. Each doubling reduces cost by $1 - 2^{-0.23} \approx 15\%$. At this rate, the $12/GB current price reaches the $5–7/GB crossing threshold after 2–3 more doublings, or approximately 3–5 years—consistent with the 2028–2029 crossing prediction.

4.3 Note on Identification

The packaging learning curve is estimated by OLS regression of log cost on log cumulative output for HBM generations (Table 7). This identifies a correlation, not necessarily a structural learning-by-doing parameter. Endogeneity concerns (demand shocks driving both output and investment in cost reduction) are standard in the learning-curve literature (Irwin and Klenow 1994). No published IV estimate exists for the packaging learning curve.

The $\alpha = 0.23$ is identified from product-level HBM pricing that bundles die and packaging

costs, with $n = 6$ generation-level observations—too few for formal structural estimation. This paper's empirical contribution is identifying *which* curve matters (early-stage packaging, not asymptotic die fabrication), not claiming precise estimation of its slope.

Three small-sample diagnostics substitute for formal structural break testing (which requires a minimum of approximately 15 observations for two-regime Bai-Perron tests). 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 4 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.

The estimate's reliability rests on three additional indirect supports: cross-technology consistency of $\alpha \approx 0.21$–0.24 across independently estimated early-stage curves (Table 8); the physical cost decomposition showing packaging as the majority cost component; and the early-stage character of the process, where limited demand-side feedback reduces simultaneous-equations bias.

Table 8: Cross-domain learning rates.

Industry	Product	α	SE	Period	Source
Semiconductor	HBM (3D stacking)	0.23	0.06	2015–2024	TrendForce
Semiconductor	NAND Flash	0.24	0.05	2003–2023	Micron/Samsung
Semiconductor	Intel microprocessors	0.24	0.04	1974–1989	Flamm (1993)
Semiconductor	DRAM (IV, causal)	0.32	0.05	1974–1992	Irwin & Klenow
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.

5. The Generalized R_0 Crossing Condition

5.1 Why Epidemic Dynamics

Hardware crossing *precedes* architectural dominance by 3–5 years historically. Cost parity is necessary but not sufficient: the distributed ecosystem must also overcome coordination frictions, sustain adoption against churn, and generate network effects.

Three canonical frameworks model technology adoption: Bass (1969) diffusion, threshold

models (Granovetter 1978), and epidemic/SIR models (Mansfield 1961). The choice is not arbitrary.

Bass diffusion decomposes adoption into an external "innovation" rate p and an internal "imitation" rate q, taking the product's existence and viability 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 1978) 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. In the standard SIR model, β is fixed. Here, β is a function of cost $c(Q)$, which falls with cumulative production Q, which is 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 (1994) 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 2004). 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.

5.2 Formal Specification

The adoption dynamics are formalized as a modified SIR system in which the "infection" state represents distributed inference adoption and the "recovery" state represents reversion to centralized provision.

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

$$ds/dt = \beta(c(Q), \lambda) \cdot \gamma \cdot s(t) \cdot (1 - s(t)) - (\kappa + \mu) \cdot s(t) \tag{15}$$

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:

$$R_0 \equiv \frac{\beta(c,\lambda)\cdot\gamma}{\kappa+\mu} > 1 \tag{16}$$

The parameters have the following structural interpretations:

- $\beta(c,\lambda)$: *Adoption rate*, depending on cost and latency advantages. Microfounded below.
- γ: *Network effect multiplier*, capturing the degree to which each adopter increases ecosystem value through shared model repositories, tooling, and deployment infrastructure.
- κ: *Coordination friction*, the rate at which potential adopters are deterred by deployment complexity. Observable from deployment latency compression: weeks in mid-2024, hours by January 2025.
- μ: *Churn rate*, driven by model obsolescence. Bounded from model lifecycle data: $\mu \approx 0.08$–0.17/month.
- λ: *Latency advantage*, structural and hardware-determined: edge inference achieves <10ms versus 50–200ms for cloud round-trip.

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

The adoption rate β requires microfoundation. Under rational inattention (Sims 2003; Matějka and McKay 2015), a user choosing between centralized and distributed inference faces a discrete choice problem under information constraints. The optimal adoption probability takes the logit form:

$$P(\text{distributed}) = \frac{\exp(\Delta u/T_u)}{1+\exp(\Delta u/T_u)} \tag{17}$$

where T_u is the user's information friction (inverse attention) and Δu is the utility differential:

$$\Delta u = \underbrace{\beta_0(c^* - c(Q))}_{\text{cost advantage}} + \underbrace{\lambda(\ell_{\text{cent}} - \ell_{\text{dist}})}_{\text{latency advantage}} - \underbrace{\kappa_0 \cdot d(Q)}_{\text{deployment friction}} \tag{18}$$

where $c^* - c(Q)$ is the cost differential (positive when distributed is cheaper), $\ell_{\text{cent}} - \ell_{\text{dist}}$ is the latency differential (positive because edge is faster), and $d(Q)$ is deployment complexity

(declining with cumulative production as tooling matures). At cost parity ($Q = \bar{Q}$, so $c^* = c(Q)$):

$$R_0\big|_{Q=\bar{Q}} = \frac{\lambda\gamma}{\kappa + \mu} \tag{19}$$

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. Coordination lag $\Delta T \approx 0$.
- If $\lambda\gamma < \kappa + \mu$: additional cumulative production beyond $\bar{Q}$ is required. This produces the 2–5 year coordination lag observed historically.

5.4 The Self-Sustaining Adoption Threshold

Setting $R_0 = 1$ and solving for the cumulative production level at which the distributed ecosystem becomes self-sustaining:

$$\frac{[\beta_0(c^* - c(Q)) + \lambda] \cdot \gamma}{\kappa + \mu} = 1 \tag{20}$$

Solving for $c(Q)$:

$$\begin{aligned} \beta_0(c^* - c(Q)) + \lambda &= \frac{\kappa + \mu}{\gamma} \\ c(Q) &= c^* - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right) \end{aligned} \tag{21}$$

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

$$\begin{aligned} c_0 Q^{-\alpha} &= c_0 \bar{Q}^{-\alpha} - \frac{1}{\beta_0}\left(\frac{\kappa + \mu}{\gamma} - \lambda\right) \\ Q^{-\alpha} &= \bar{Q}^{-\alpha}\left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right) \end{aligned} \tag{22}$$

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

$$\boxed{\bar{Q}^* = \bar{Q} \cdot \left(1 - \frac{\kappa + \mu}{\beta_0 \gamma \cdot c^*} + \frac{\lambda}{\beta_0 \cdot c^*}\right)^{-1/\alpha}} \tag{23}$$

Three properties merit emphasis.

Direction of the shift. When $\lambda\gamma < \kappa + \mu$ (the empirically relevant case—the bounding exercise in Section 8.4 estimates $R_{0,\text{distributed}} \approx 0.4$–$0.8$ currently), $\bar{Q}^* > \bar{Q}$: 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 coordination indicators (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 depends on the common-pool structure, not on the threshold's specific value.

Organizational viability. The cost-parity and R_0 conditions address whether distributed production is *affordable* and whether adoption is *self-sustaining*. A third condition asks whether it is *organizationally viable*: the mesh must coordinate well enough to match centralized output quality. Formally, distributed production of task type τ is viable when $K_{\text{eff}}(\rho_\tau, T_{\text{mesh}}) \geq K_{\text{threshold}}$. For inference ($\rho \approx 1$, $K \approx 0$), this is automatically satisfied: there is no complementarity to exploit. For more complex distributed tasks ($\rho < 1$), the condition becomes binding and the generalized crossing threshold shifts outward.

Table 9: 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. Predicted compression from 3–5 years to 2–3 years reflects declining κ.

5.5 Unification with Mesh R_0

The R_0 of equation (16) and the mesh reproduction number $R_0^{\text{mesh}} = N \cdot \beta \cdot v/D$ (Section 9) are the same object at different scales. The adoption-level R_0 governs *whether* the distributed ecosystem reaches critical mass; the mesh-level R_0^{mesh} governs *whether* the resulting collection of devices self-organizes into a connected, specialized network. The crossing point $x(t) = 0$ is the moment both cross unity: $R_0 > 1$ makes adoption self-sustaining, and $R_0^{\text{mesh}} > 1$ makes the mesh connected. One transcritical bifurcation, two notations.

The correspondence is precise:

- $\beta(c, \lambda)$ in the adoption R_0 maps to $\beta \cdot v$ in the mesh R_0^{mesh}: the per-contact adoption probability times the value per interaction.

- γ (network effect multiplier) maps to N (number of active nodes): both capture the positive feedback from ecosystem size.
- $\kappa + \mu$ (coordination friction plus churn) maps to D (attrition rate): both capture the outflow from the adopter/participant population.

The unification has a methodological consequence: it ensures that the post-crossing mesh formation (Sections 9–11) is not an independent model grafted onto the pre-crossing dynamics (Sections 3–8). The same transcritical bifurcation that drives pre-crossing adoption dynamics drives post-crossing mesh formation. The models are continuous at the crossing point.

Table 10: Universal self-consistency across fields.

Field	Regime indicator	Control parameter	Critical condition	Source
Percolation	Giant component S_∞	Mean degree $\langle k \rangle$	$\langle k \rangle = 1$	Erdős-Rényi
Epidemiology	Infected fraction	R_0	$R_0 = 1$	Kermack-McKendrick
Potts model	Magnetization	$\beta_T J q$	$\beta_T J q = 1$	Potts (1952)
Network econ.	Market participation	Transaction benefit	Critical liquidity	Katz-Shapiro
This paper	**Mesh fraction S_∞**	$\boldsymbol{R_0^{\text{mesh}}}$	$\boldsymbol{R_0^{\text{mesh}} = 1}$	—

5.6 Settlement Layer Requirement

The mesh requires agents to route queries to specialists and compensate them. Each routed query requires compensation. At scale—millions of micro-transactions per second between arbitrary pairs of agents, with millisecond latency requirements—this requires a settlement mechanism that is peer-to-peer, programmable, and low-latency.

Proposition 1 (Settlement Layer Necessity)**.** *Any mesh equilibrium with $N > N^*$ and $C_{mesh} > C_{cent}$ requires a settlement layer capable of processing $O(N \cdot \langle k \rangle)$ transactions per second at $O(1)$ ms latency between arbitrary node pairs.*

The proof is immediate from the bandwidth scaling result (equation 36): the number of routed queries scales as $N \cdot \langle k \rangle$, and each routed query requires compensation.

5.7 The Hayek Insight

The price system in the mesh plays exactly the role Hayek (1945) described for the market price system: it aggregates dispersed information into sufficient statistics for decentralized decision-making. The bid-ask spread between agents encodes:

- *Current demand* for each query type (high bids for underserved specializations signal entry opportunities).
- *Current supply* of each specialization (narrow spreads indicate competitive specialist markets).
- *Optimal routing* (the lowest-ask specialist for a given query type is the efficient allocation).

No central coordinator computes these allocations. Agents respond to prices, and the price system achieves efficient routing as a fixed point of bilateral negotiations. This is the Hayek mechanism operating at machine speed.

The settlement layer requirements—programmable, peer-to-peer, high-throughput, low-latency—describe the functional specification of a programmable monetary system.

Existing alternatives. No existing payment system satisfies all three requirements simultaneously. Traditional payment rails (SWIFT, ACH, card networks) are hub-and-spoke, not peer-to-peer, and operate at latencies of seconds to days. Stablecoin networks (Tether, Circle) are peer-to-peer and programmable but face throughput constraints on Layer 1 blockchains (Ethereum: $\sim$15 TPS; Bitcoin: $\sim$7 TPS). Layer 2 solutions (Lightning Network, Optimism, Arbitrum) approach the latency requirement ($<$1s) and throughput requirement ($>$10,000 TPS) but sacrifice some decentralization guarantees. Purpose-built settlement protocols (Solana: $\sim$65,000 TPS, $\sim$400ms finality) come closest to the functional specification but have not been tested at the scale the mesh requires.

The settlement layer constraint is binding when $N \cdot \langle k \rangle$ exceeds the settlement system's throughput capacity. For a mesh of $N = 10^6$ devices with $\langle k \rangle = 10$ and average query rate of 1/minute, the required settlement throughput is approximately 1.7×10^5 TPS—within the range of current Layer 2 solutions but beyond Layer 1 capacity. For $N = 10^8$ (the scale at which the mesh dominates inference), the requirement rises to $\sim 1.7 \times 10^7$ TPS, which no existing system can handle.

This connects to the separate analysis of monetary infrastructure (Paper 5; Smirl 2026, forthcoming): the mesh's growth may be constrained by the settlement layer before it is constrained by device capability.

6. Training-Inference Bifurcation

6.1 Two Workloads, Two Architectures

AI compute divides into two structurally distinct workloads with fundamentally different coordination requirements.

Training teaches models by processing massive datasets across tightly synchronized GPU clusters. A single frontier training run (GPT-4-class, 2024) requires 10,000–100,000+ GPUs communicating at terabits per second via NVLink and InfiniBand, running for weeks to months. Power density: 100–1,000 kW/rack. The gradient synchronization step—all-reduce across the full cluster—requires all-to-all communication within a single forward-backward pass. This is a *topological* constraint: the communication graph must be fully connected at nanosecond timescales. Latency tolerance is measured in microseconds.

Inference runs trained models to serve real-time user requests. Each query is independent and atomizable: a request for "summarize this document" requires no synchronization with any other request. Latency-sensitive at the user level (<10ms local versus 50–200ms cloud round-trip), but not at the inter-device level—there is no inter-device communication during a single inference pass. The cost trajectory is declining rapidly: Stanford's 2025 AI Index documented a 280-fold drop in inference costs between November 2022 and October 2024.

Table 11: Training vs. inference structural comparison.

Dimension	Training	Inference
Share of AI compute (2025)	∼50%	∼50%
Share of AI revenue (2025)	∼10–15%	∼85–90%
Synchronization requirement	Massive (10K+ GPUs)	None (atomizable)
Latency sensitivity	Low (days tolerable)	High (<10ms for UX)
Cost trajectory	Rising per frontier model	Declining ∼280× in 2 yr
Communication topology	All-to-all (NVLink)	Point-to-point (WiFi)
Communication latency	Nanoseconds (NVLink)	Milliseconds (WiFi)
Edge-viable?	No (architectural)	Yes (this paper's thesis)

The revenue composition is critical for the differential game. Training costs are rising (∼$100M per frontier run in 2024, projected $1B+ in 2026), but training generates revenue only indirectly through the models it produces. Inference generates 85–90% of AI service revenue directly. The self-undermining mechanism operates on the *inference* revenue stream: as distributed inference becomes viable, the centralized firm loses its primary revenue source while retaining its cost base.

6.2 The Inference Revenue Pool

The inference revenue pool is the economic object at stake. Define:

$$R_{\text{inf}}(t) = p_{\text{inf}}(t) \cdot V_{\text{inf}}(t) \tag{24}$$

where p_{inf} is the per-token price and V_{inf} is the total token volume. Between November 2022 and October 2024, p_{inf} fell 280-fold while V_{inf} grew approximately 50-fold, yielding a net revenue pool decline of approximately 5.6-fold per unit of capability delivered. This price compression is consistent with the model: as open-weight alternatives proliferate, the per-token price converges toward marginal inference cost, and the revenue pool shifts from centralized providers to distributed inference.

The dynamics of the inference revenue pool follow a specific trajectory:

- *Phase 1 (2022–2024):* Monopolistic pricing. Per-token prices reflect frontier scarcity, not marginal cost. GPT-4 launch pricing ($30/M output tokens) implied gross margins exceeding 90%.
- *Phase 2 (2024–2025):* Competitive compression. Open-weight alternatives (Llama 3, Qwen 2.5, DeepSeek V3) drove API prices down 10–50×. DeepSeek R1 at $0.55/M tokens represents marginal-cost-plus pricing.
- *Phase 3 (projected 2026–2030):* Edge displacement. On-device inference eliminates the API pricing layer entirely for queries within device capability. The revenue pool for these queries becomes zero—or rather, the revenue shifts to hardware manufacturers and model fine-tuning services.

The training revenue pool, by contrast, is *not* at stake. Model-as-a-service (API access to frontier models), model licensing (enterprise deployment), and training compute rental all depend on centralized training remaining competitive. The continuation value S_T in the differential game captures this persistent revenue. The inference market is projected to grow from $106 billion (2025) to $255 billion by 2030 (MarketsandMarkets 2025)—but this projection assumes continued centralized dominance. Under the endogenous decentralization scenario, a significant fraction of this revenue migrates to the distributed paradigm.

6.3 Formal Derivation from Effective Curvature

Proposition 2 (Training-Inference Bifurcation)**.** *Let* $K_{\mathit{eff}}(\rho, T) = K \cdot (1 - T/T^*(\rho))^+$. *Define:*

- *Training:* $\rho_{train} \ll 0$ *(near-Leontief),* $T_{cent} \approx 0$ *(datacenter),* $T_{dist} \gg T^*(\rho_{train})$ *(distributed latency).*

- *Inference:* $\rho_{inf} \approx 1$ *(independent queries), so* $K_{inf} \approx 0$.

Then:

(a) $K_{eff}^{cent}(training) = K_{train} \gg 0$ *but* $K_{eff}^{dist}(training) = 0$. *Centralized training strictly dominates.*

(b) $K_{eff}^{cent}(inference) \approx K_{eff}^{dist}(inference) \approx 0$. *Cost determines the winner.*

Consequently, inference decentralizes at cost parity while training remains centralized.

Proof. Part (a): Training is near-Leontief (K large) because gradient updates must synchronize. Centralized operation achieves $T_{\text{cent}} \approx 0$ via NVLink. Distributed operation has $T_{\text{dist}}/T^* \gg 1$ (WiFi latency exceeds NVLink by 5–6 orders of magnitude), so $K_{\text{eff}}^{\text{dist}} = 0$.

Part (b): Inference queries are independent ($\rho \approx 1$, $K \approx 0$). Effective curvature is ≈ 0 regardless of T. □

The proposition predicts a *gradient of decentralization* ordered by ρ: simple inference distributes first, complex reasoning later, federated fine-tuning later still, and training last or never.

6.4 Implications for the Differential Game

The bifurcation has three consequences for the pre-crossing dynamics.

First, it sharpens the self-undermining property. The centralized firm's inference revenue is vulnerable to distributed entry, but its training revenue is not. The continuation value S in the differential game decomposes as $S = S_T + S_I/(N(r+\delta))$, where S_T (training) is large and persistent while S_I (inference) is the stream under threat. The crossing event destroys S_I while preserving S_T—a partial, not total, displacement.

Second, it explains the observed investment pattern. Hyperscalers are simultaneously investing in (a) frontier training capability (protecting S_T) and (b) inference cost reduction (defending S_I by lowering price toward distributed cost). The dual investment is rational given the bifurcation: training investment has persistent returns while inference investment is a holding action.

Third, it introduces a *two-front competitive structure*. On the training front, $N \approx 3$–5 firms compete for frontier model leadership (a research tournament). On the inference front, $N \approx 8+$ firms compete on cost and latency (a standard Cournot/Bertrand market). The differential game of Section 3 applies primarily to the inference front, where the common-pool structure is most acute.

Table 12: Gradient of decentralization by task type.

Task Type	ρ	K_{eff} Regime	Distribution Timing
Simple generation	$\to 1$	≈ 0 (cost determines)	2026–2028
Routine inference	≈ 0	Moderate	2028–2030
Complex reasoning	≈ -1	Strong, requires coordination	2031–2035
Federated fine-tuning	≈ -2	Very strong	2035+
Frontier training	$\to -\infty$	Maximal, topological barrier	Possibly never

7. Natural Experiment: US-China Export Controls

7.1 Identification Strategy

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. DeepSeek, Alibaba/Qwen, Baichuan, 01.AI/Yi, Zhipu/GLM, Moonshot/Kimi. The binding compute constraint creates structural incentives to optimize for the distributed paradigm.

Control: Unconstrained. Meta/Llama, Mistral, Google/Gemma, Microsoft/Phi, Stability, Falcon/TII. No binding constraint predicts scale-first strategies.

7.2 Competing Predictions

Table 13: 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	Match/exceed
Architecture choice	Incremental improvement	Pivot to MoE, distillation	DeepSeek V3 MoE
Model size distribution	Similar across groups	Constrained skew small/edge	47% $\leq$3B vs 25%
Ecosystem share	Unconstrained dominate	Constrained gain share	Qwen overtakes Llama
Derivative adoption	Proportional	Constrained more forked	40% vs 15%

7.3 Results

Capability convergence. 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 ($\leq$3B parameters): 47% of their releases versus 25% for unconstrained developers. Three of four major constrained releases use MoE or distillation: DeepSeek V3 (671B total $\rightarrow$ 37B active, MoE), DeepSeek R1 (distilled to 1.5B, 7B, 14B), Qwen (full sub-1B to 72B range), and Kimi K2.5 (1T total $\rightarrow$ MoE active subset). Unconstrained firms—Meta Llama 3.1 (405B dense, no MoE), Mistral (Mixtral 8$\times$7B, early MoE but pre-controls), Google Gemma (dense), Microsoft Phi (dense, small)—adopted scale-first approaches. 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. 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.

Cost collapse. Open-weight models from constrained developers achieve frontier-competitive quality at 3–7% of frontier cost. DeepSeek R1 achieves o1-level reasoning at approximately \$0.55/million input tokens versus \$15/million for o1—a 27$\times$ cost advantage. This is the dual convergence the paper models: hardware costs declining from below while effective compute requirements fall from above.

Detailed evidence on architectural pivots. The architectural responses merit detailed examination because they provide the strongest evidence for constraint-induced optimization. DeepSeek V3 (December 2024) uses a MoE architecture with 256 experts, of which 8 are activated per token. The design achieves GPT-4-class quality on standard benchmarks while requiring only 37B active parameters—a ratio of 18:1 between total and active parameters. This is not an incremental improvement on the dense-model paradigm; it is a qualitatively different architecture optimized for inference efficiency. The training cost was approximately \$5.6M, compared to estimated \$100M+ for comparable dense models—further evidence that the constraint produced efficiency innovation rather than capability degradation.

Qwen-2.5 (September 2024) adopted a full-spectrum release strategy: 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B parameter variants. The explicit targeting of sub-3B models for edge

deployment, with dedicated optimization for mobile and IoT inference, is inconsistent with standard learning-by-doing (which predicts scaling up, not scaling down) and consistent with constraint-induced optimization for the distributed paradigm.

Quantitative summary.

Table 14: Export control natural experiment: summary statistics.

Metric	Constrained	Unconstrained
Models ≤3B parameters (% of releases)	47%	25%
MoE architecture adoption	3/4 major	1/4 major
Derivative models (% of HuggingFace new)	40%	15%
Cumulative downloads (Jan 2025)	700M+	500M+
Inference cost ($/M tokens, best model)	0.55	15.00
Capability per FLOP (relative)	1.0–1.2×	1.0×

7.4 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. Documenting the adoption lag—how quickly unconstrained firms adopted MoE after constrained firms demonstrated its viability—would bound the true effect size.

Selection. Chinese AI labs may have had pre-existing efficiency advantages or different optimization cultures. The open-weight ecosystem barely existed pre-treatment (the first major open-weight release, Llama 1, occurred in February 2023, four months after the initial export controls), making 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.

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, quan-

tization level) need systematic coding. This is a natural next step for a companion empirical paper.

This section documents the qualitative pattern; a formal difference-in-differences panel at the firm-model-quarter level, with pre-treatment parallel trends and standardized efficiency metrics, is left to future work.

8. Dual Convergence

8.1 Convergence from Below: The Hardware Learning Curve

The inference crossing condition—≥70B-class output quality at ≥20 tok/s under $1,500—is being approached from below via the packaging learning curve. As of Q1 2026, the technology capability threshold has been met at professional price points, but the DRAM supercycle has temporarily inflated consumer memory costs 300–400% above trend.

Consumer silicon trajectory. Rockchip's RK1828 (2025, 5GB 3D stacked DRAM) runs 7B-parameter models at 59 tok/s. AMD's Ryzen AI Max+ 395 (∼$2,000, 128GB) achieves ∼31 tok/s on MoE architectures with ∼20B active parameters. NVIDIA's RTX 5090 (32GB GDDR7, ∼1,792 GB/s) exceeds the speed threshold but street prices range $3,000–$5,000+ due to the supercycle.

8.2 Convergence from Above: Algorithmic Efficiency

Three algorithmic innovations are independently reducing the effective compute requirement for a given quality level:

Mixture-of-Experts (MoE). DeepSeek V3 (671B total, ∼37B active) achieves 70B-class quality while activating only 37B parameters per forward pass—a 3–6× reduction in memory bandwidth requirement. MoE architectures route each token to a subset of "expert" sub-networks, exploiting the insight that different inputs require different computation. For edge deployment, MoE reduces the binding constraint: the device must store the full model but only compute with a fraction.

Quantization. INT4 quantization reduces memory footprint ∼4× relative to FP16, with quality degradation of <2% on most benchmarks (Dettmers et al. 2023). GGUF format enables per-layer mixed quantization, preserving quality for attention layers while aggressively quantizing feed-forward layers. The implication: a 70B-parameter model at INT4 requires ∼35GB of memory, within reach of consumer devices with 48GB+ unified memory.

Distillation. DeepSeek R1 was distilled to 1.5B, 7B, 8B, 14B, 32B, and 70B variants, each targeting a specific device class. The distilled 14B variant achieves ∼85% of the full

model's reasoning capability at ∼5% of the compute requirement. Distillation is particularly relevant for the self-undermining mechanism: the knowledge embedded in a trillion-parameter frontier model (requiring a datacenter to train) can be compressed into a model that runs on a phone.

The combined effect: Stanford's 2025 AI Index documented a 280-fold drop in inference costs between November 2022 and October 2024. The effective cost decline including algorithmic optimization is significantly steeper than $\alpha = 0.23$ alone. Formally, the effective learning rate incorporating algorithmic improvements is:

$$\alpha_{\text{eff}} = \alpha_{\text{hardware}} + \alpha_{\text{algo}} \approx 0.23 + 0.15 \approx 0.38 \tag{25}$$

where α_{algo} captures the rate at which algorithmic efficiency improvements reduce the compute requirement per unit of output quality. This effective rate is consistent with the 280-fold cost decline over approximately 2 years.

Economic interpretation. The dual convergence has a precise economic meaning: the "distance to crossing" is being closed from both sides simultaneously, and the two forces multiply rather than add. Hardware cost decline reduces the numerator of the cost ratio $c_{\text{dist}}/c_{\text{cent}}$, while algorithmic efficiency gains reduce the numerator of the quality-adjusted cost ratio. If $c_{\text{dist}}(t) = c_0^{\text{hw}} Q^{-\alpha_{\text{hw}}} \cdot c_0^{\text{algo}} Z^{-\alpha_{\text{algo}}}$ where $Z(t)$ is cumulative algorithmic improvement, the crossing time satisfies:

$$T^*_{\text{dual}} = \frac{T^*_{\text{hw}} \cdot T^*_{\text{algo}}}{T^*_{\text{hw}} + T^*_{\text{algo}} - T^*_{\text{hw}} T^*_{\text{algo}} / T^*_{\text{indep}}} \tag{26}$$

which is strictly less than either T^*_{hw} or T^*_{algo} individually. The multiplicative structure explains why the AI transition is compressed relative to prior transitions where convergence was primarily from below.

8.3 Hyperscaler Capital Expenditure

A significant fraction flows directly to the packaging learning curve: each NVIDIA H100/H200/B200 GPU contains multiple HBM stacks, each requiring TSV processing, die thinning, and advanced packaging.

8.4 The Demand Shock as Nash Overinvestment

The Stargate project alone demands approximately 40% of global DRAM output. Total industry HBM demand is projected to reach approximately 1 million CoWoS-equivalent wafers in 2026, up from 370,000 in 2024 (Morgan Stanley 2026). The capacity expansion is

Table 15: Hyperscaler capex ($B).

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

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

concrete: Samsung P4 (Pyeongtaek), SK Hynix M15X (Icheon), Micron Idaho, and TSMC CoWoS expansion lines collectively represent a tripling of advanced packaging capacity within two years.

Historical precedent predicts overcapacity and below-trend pricing by 2028–2029. The semiconductor memory industry has experienced at least seven major boom-bust cycles since 1984 (1988, 1995, 2001, 2008, 2016, 2022, and the current 2025–26 cycle). In each prior cycle, capacity investment during the boom phase produced overcapacity and prices 30–50% below pre-boom trend within 18–30 months of peak demand.

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 self-undermining mechanism operating through the supply side: the boom finances the capacity that enables the bust to accelerate the crossing.

The 2025–26 DRAM supercycle is the model's capacity-constraint corollary operating through a novel channel: the boom phase temporarily *reverses* the consumer cost trajectory even as it finances the packaging capacity expansion that will eventually crash consumer prices below the pre-boom trend. The resolution is temporal: the boom phase adds 1–2 years to the hardware crossing timeline, but the bust phase may compress the post-bust crossing timeline by a comparable amount, because the installed packaging capacity exceeds what steady-state datacenter demand can absorb.

8.5 Bounding R_0 from Adoption Data

The R_0 framework developed in Section 5 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 $\kappa + \mu$. 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.

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

$$R_0(t) \approx 1 + \frac{\Delta s/\Delta t}{\delta \cdot s(t) \cdot (1 - s(t))} \tag{27}$$

Table 16: 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
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 represents the post-novelty plateau. Excluding the spike-reversion, mean implied $R_0 = 1.15$.

Critical scope distinction. The OpenRouter series measures open-weight model share through a *centralized* aggregator. The paper's $R_0 > 1$ crossing condition refers to self-sustaining *distributed* inference adoption, which faces additional coordination friction. The OpenRouter-implied R_0 bounds the upper envelope; the distributed-specific R_0 is strictly lower.

Three features of the trajectory are notable. First, implied R_0 is above unity for most of the observation period (mean ≈ 1.2), consistent with open-weight models gaining share through centralized providers. Second, the trajectory is approximately flat at $R_0 \approx 1.2$, then exhibits a sharp perturbation (DeepSeek R1) followed by reversion—characteristic of event-driven rather than self-sustaining adoption. Third, the February 2025 reversion ($R_0 = 0.59$) demonstrates the ecosystem can still enter sub-critical regimes.

With distributed-specific coordination friction plausibly 2–5× higher than centralized, $R_{0,\text{distributed}} \approx 0.4$–$0.8$ in the current period—firmly sub-critical. The prediction that $R_{0,\text{distributed}} > 1$ by 2030–2032 requires: (a) continued hardware cost decline along the packaging learning curve ($\alpha = 0.23$); (b) coordination friction $\kappa_{\text{distributed}}$ declining as edge runtimes mature (the implied rate from the centralized data is approximately 30–50% per year during 2024); and (c) the structural latency advantage λ becoming salient as real-time applications grow (autonomous vehicles, AR/VR, robotics, real-time translation—all applications where 50–200ms cloud latency is unacceptable).

The implied rate of κ decline from the centralized data provides a lower bound on the

coordination maturation rate. The deployment latency compression—from weeks of community effort in June 2024 to day-zero quantized releases by January 2025—suggests a κ half-life of approximately 6 months in the centralized ecosystem, with the distributed-specific κ declining at roughly half this rate due to the additional complexity of hardware heterogeneity.

9. Post-Crossing: Mesh Formation via Potts Regime Shift

9.1 Giant Component Existence

The fraction of nodes belonging to the giant connected component satisfies:

$$S_\infty = 1 - \exp(-R_0^{\text{mesh}} \cdot S_\infty) \tag{28}$$

Proposition 3 (Giant Component Existence and Uniqueness)**.** *Equation* (28) *has the trivial solution* $S_\infty = 0$ *for all* R_0^{mesh}*; a unique positive solution* $S_\infty^* \in (0,1)$ *if and only if* $R_0^{mesh} > 1$*; and* S_∞^* *is locally asymptotically stable.*

Proof. Define $g(s) = 1 - \exp(-R_0^{\text{mesh}} \cdot s) - s$. At $s = 0$: $g'(0) = R_0^{\text{mesh}} - 1 > 0$ when $R_0^{\text{mesh}} > 1$. At $s = 1$: $g(1) < 0$. By the intermediate value theorem, g has a root in $(0,1)$. Uniqueness: $g'' < 0$ (strict concavity). Stability: $h'(S_\infty^*) = R_0^{\text{mesh}}(1 - S_\infty^*) \in (0,1)$ at the positive fixed point. □

9.2 The Fortuin-Kasteleyn Unification

The inclusive value of the q-state Potts model on graph $G = (V, E)$ has the exact cluster expansion:

$$Z_{\text{Potts}} = \sum_{A \subseteq E} p^{|A|}(1-p)^{|E|-|A|} \cdot q^{c(A)} \tag{29}$$

At $q = 1$, this reduces to bond percolation. For $q > 2$ specialization types, the regime shift is first-order (discontinuous).

Proposition 4 (First-Order Regime Shift)**.** *For* $q > 2$ *specialization types on a graph with mean degree* $\langle k \rangle > 1$*, the specialization transition is first-order: the regime indicator jumps discontinuously from zero to a positive value at the critical coupling. The mesh does not form gradually—it crystallizes.*

For $q = 2$, the transition is second-order (continuous). The first-order prediction is specific to $q \geq 3$ distinct specialization roles, the empirically relevant case for AI inference (coding, creative writing, mathematical reasoning, multimodal processing, domain-specific knowledge, real-time translation, etc.).

9.3 Inverse Capability Concentration

The Bianconi-Barabási (2001) fitness model exhibits two regimes depending on the fitness distribution $\rho(\eta)$: winner-takes-all capability concentration when $\rho(\eta)$ is sharply peaked (the centralized equilibrium), and a fit-get-rich distributed regime when $\rho(\eta)$ is broad (the mesh equilibrium).

Proposition 5 (Learning-Curve-Driven Regime Shift)**.** *Let the technology parameter $\theta(t)$ index the packaging learning curve output. Suppose the fitness of an edge device of type j is $\eta_j(\theta) = \eta_j^0 + g_j(\theta)$, where g_j is increasing and $g_j(0) = 0$. If $\theta(0)$ produces a fitness distribution with exponent $\alpha_B \leq 0$ (BEC regime), and $\theta(\bar{t})$ produces $\alpha_B > 0$ for finite $\bar{t}$, then the system undergoes a regime shift at θ^* where α_B crosses zero. This is inverse capability concentration: the centralized condensate dissolves.*

Proof. The learning curve increases the fitness of previously low-fitness edge devices. Initially, only datacenter nodes have η near $\eta_{\max}$, so $\rho(\eta)$ is sharply peaked—the BEC regime. As θ increases, the support of ρ broadens. The exponent α_B transitions from ≤ 0 to > 0 at $\theta = \theta^*$. The mapping to the predecessor framework is direct: θ^* corresponds to $x(t) = 0$. □

Remark 4. *The BEC framework describes the* competitive dynamics *of traffic allocation. It does not describe physical connectivity (percolation) or specialization (CES aggregation). The three layers compose: percolation ensures the mesh is connected, BEC dynamics govern how traffic flows, and CES aggregation determines whether collective capability exceeds the centralized alternative.*

Quantitative illustration. Consider the fitness distribution before and after crossing. Before crossing ($\theta < \theta^*$): datacenter nodes have $\eta \approx 1.0$ (normalized), consumer GPUs $\eta \approx 0.3$, mobile devices $\eta \approx 0.05$. The distribution $\rho(\eta)$ is sharply peaked near $\eta_{\max} = 1.0$ with exponent $\alpha_B \approx -0.5$ (condensation regime). After crossing ($\theta > \theta^*$): consumer GPUs with 3D stacked memory achieve $\eta \approx 0.7$, high-end mobile $\eta \approx 0.4$, edge devices with stacked DRAM $\eta \approx 0.25$. The distribution broadens; α_B crosses zero to $\approx +0.3$ (distributed regime). The centralized "condensate" fraction—the share of total inference traffic handled by the three largest cloud providers—declines from approximately 85% (2025) toward a predicted 40–50% (2035) as the fitness distribution broadens.

9.4 Post-Crossing Dynamics in Three Phases

The path from $x(t) = 0$ to mesh dominance is not monotonic. Three phases, distinguished by the value of R_0^{mesh} and the state of the specialization structure, characterize the transition.

Phase 1: Nucleation ($R_0^{\text{mesh}} \approx 1$). Immediately after crossing, R_0^{mesh} is only marginally above unity. The giant component is small ($S_\infty^* \approx 0$ for R_0^{mesh} near 1, since $S_\infty^* \sim 2(R_0^{\text{mesh}} - 1)/R_0^{\text{mesh}2}$ to leading order). Growth is slow and stochastic. Small specialist clusters form around high-fitness agents—the enthusiast-tier hardware users running quantized models—but the clusters are fragile. Exogenous shocks (model-release events, API pricing changes, hardware supply disruptions) can temporarily push R_0^{mesh} below unity, collapsing nascent clusters.

The mesh first achieves capability dominance on the *long tail* of niche queries that centralized systems underserve. Centralized providers optimize for the highest-volume query types (general chat, code generation, summarization). Specialized queries—domain-specific technical reasoning, low-resource language translation, real-time edge processing for robotics—are underserved because the revenue per query does not justify dedicated model fine-tuning. The mesh's heterogeneous agents, each fine-tuned for a niche, collectively cover the long tail. This is the Christensen (1997) pattern: disruption begins in markets the incumbent rationally ignores.

Phase 2: Rapid Growth ($R_0^{\text{mesh}} \gg 1$). As additional device types become inference-capable (driven by the continuing packaging learning curve) and coordination infrastructure matures (κ declines), R_0^{mesh} accelerates well above unity. Network effects dominate. Each new specialist joining the mesh increases C_{eff} superlinearly (by the CES diversity premium) and increases the Fiedler eigenvalue $\lambda_2(L)$ (by adding connectivity), which accelerates knowledge diffusion to subsequent entrants.

In this phase, the first-order regime shift occurs: the division of labor among mesh agents transitions from fragmented proto-specialization to a structured, self-reinforcing configuration. By Proposition 4, this transition is discontinuous for $q > 2$ specialization types. The regime shift is observable as a sudden increase in the concentration of agent capabilities around distinct specialization types, accompanied by the emergence of routing hub agents that handle disproportionate query traffic.

Centralized providers lose market share in progressively more mainstream query types, beginning with the long-tail niches of Phase 1 and expanding to higher-volume categories as mesh coverage broadens.

Phase 3: Maturity. Growth saturates as the mesh's J task types are fully covered. The CES aggregate C_{eff} approaches its maximum for the given device population. Competition shifts from mesh growth to mesh composition: which specialists are included, the quality of their fine-tuning, and the efficiency of the routing layer.

Centralized providers retain structural advantage in two domains that the mesh cannot replicate:

(i) *Frontier model training:* Training requires tightly synchronized GPU clusters at scales incompatible with distributed architecture. The mesh depends on centralized training for the base models it fine-tunes.

(ii) *Capabilities requiring single-device scale beyond any edge device:* Tasks requiring the full activation of 1T+ parameter dense models in a single forward pass remain centralized. This is the inference analog of the training constraint, but applies to a shrinking fraction of queries as MoE architectures reduce the active parameter requirement.

The mature equilibrium is coexistence: centralized providers dominate training and frontier-capability inference; the mesh dominates the long tail, latency-sensitive applications, and the broad middle of the query distribution where specialized, fine-tuned models outperform general-purpose frontier models.

10. CES Diversity Premium and Specialization

10.1 Agent Capabilities and CES Aggregation

Each agent $i \in \{1, \ldots, N\}$ has a capability vector $\mathbf{c}_i = (c_{i1}, \ldots, c_{iJ})$ across J task types. The aggregate mesh capability is:

$$C_{\text{eff}}(N) = \left(\sum_{j=1}^{J} C_j^{\rho} \right)^{1/\rho}, \quad 0 < \rho < 1 \tag{30}$$

where $C_j = \sum_i c_{ij}$. The parameter $\rho < 1$ implies imperfect substitutability: two agents with different specializations contribute more to C_{eff} than two identical agents.

Lemma 1 (Diversity Premium)**.** *Fix total capability $\bar{C} = \sum_j C_j$. For $\rho < 1$, C_{eff} is maximized at equal coverage. The diversity premium $J^{(1-\rho)/\rho}$ is increasing in J and decreasing in ρ.*

Proof. With equal allocation: $C_{\text{eff}} = J^{1/\rho - 1} \cdot \bar{C}$. With concentration: $C_{\text{eff}} = \bar{C}$. The ratio is $J^{(1-\rho)/\rho} > 1$ for $J \geq 2$ and $\rho < 1$. □

This is the Becker-Murphy (1992) division of labor result in CES form. The mesh's advantage comes not from superior individual capability—each edge device is weaker than the datacenter—but from the breadth of specialized coverage that heterogeneous agents collectively provide.

10.2 Centralized Capability Benchmark

A centralized provider operates M identical high-capability units, each with capability $\bar{c}$ spread uniformly across all J task types. Total centralized capability for task j is $C_j^{\text{cent}} = M\bar{c}/J$, giving:

$$C_{\text{cent}} = \left(J \cdot \left(\frac{M\bar{c}}{J} \right)^{\rho} \right)^{1/\rho} = J^{(1-\rho)/\rho} \cdot M\bar{c} \tag{31}$$

The centralized provider has fixed capacity $M\bar{c}$ (determined by datacenter investment). The mesh's aggregate capability grows with N and with the diversity of specialists.

Quantitative calibration of ρ. The mesh aggregates heterogeneous AI agents with complementary specializations—the closest published analogue is cross-sector intermediate input substitution. Atalay (2017) estimates near-zero substitutability across 6-digit NAICS sectors ($\sigma \in [0.0, 0.2]$, $\rho \in [-\infty, -4]$) using input-output tables, placing the mesh firmly in the complementary regime. The companion paper (Paper 1) adopts this range as the primary identification, with NLS estimation on FRED Manufacturing IP ($\hat{\rho} = -0.30$, $\sigma = 0.77$) serving as cross-validation. For the diversity premium $J^{(1-\rho)/\rho}$, even the conservative $\rho = -1$ yields a premium of J^2 for J specialist types.

Illustrative magnitudes. The diversity premium from Lemma 1 is $J^{(1-\rho)/\rho}$, the ratio of the CES aggregate under equal allocation to the aggregate under full concentration. With $J = 20$ distinct specialization types (legal reasoning, medical coding, multilingual translation, code review, mathematical proof, creative writing, scientific summarization, etc.):

ρ	$\sigma = 1/(1-\rho)$	$J^{(1-\rho)/\rho}$	Interpretation
0.50	2.0	$20^1 = 20$	Strong premium
0.75	4.0	$20^{1/3} \approx 2.7$	Moderate premium
0.90	10.0	$20^{1/9} \approx 1.4$	Weak premium

At $\rho = 0.5$, twenty specialists with equal total capability produce twenty times the output of a single generalist—a factor large enough to overcome substantial per-device capability disadvantage. The sensitivity to ρ underscores why accurate estimation of the substitution parameter is critical: the difference between $\rho = 0.5$ and $\rho = 0.9$ is a factor of 14 in the diversity premium.

10.3 Specialization Dynamics: Fixed Response Threshold Model

Agents do not arrive pre-specialized. Specialization emerges endogenously through local interactions. The mechanism follows the Bonabeau-Theraulaz (1998) fixed response threshold model, originally developed for division of labor in social insect colonies.

Agent i has a vector of response thresholds $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iJ})$ for each task type. When demand signal s_j for task j arrives, agent i performs the task with probability:

$$P_{ij}(s_j) = \frac{s_j^n}{s_j^n + \theta_{ij}^n} \tag{32}$$

where $n \geq 2$ is a steepness parameter. The response probability is a sigmoid: high when $s_j \gg \theta_{ij}$, low when $s_j \ll \theta_{ij}$.

Thresholds adapt through reinforcement:

$$\dot{\theta}_{ij} = -\xi \cdot \mathbf{1}[i \text{ performs task } j] + \varphi \cdot \mathbf{1}[i \text{ does not perform task } j] \tag{33}$$

where $\xi > 0$ is the reinforcement rate (performing a task lowers the threshold, increasing future responsiveness) and $\varphi > 0$ is the decay rate (not performing a task raises the threshold).

Proposition 6 (Emergent Specialization)**.** *Under the threshold dynamics (33) with heterogeneous initial thresholds $\theta_{ij}(0)$, agents self-sort into specialist roles: for each agent i, there exists $j^*(i) = \arg\max_j c_{ij}(t)$ such that $c_{ij^*}(t) \to \bar{c}_i$ and $c_{ij}(t) \to 0$ for $j \neq j^*$ as $t \to \infty$, where $\bar{c}_i$ is agent i's maximum achievable capability.*

Proof. An agent that performs task j frequently sees θ_{ij} decrease, making it more responsive to future demand for j, which further increases frequency of performance—a positive feedback loop. Simultaneously, thresholds for other tasks rise. The dynamics converge to a fixed point where each agent responds primarily to one task type. □

This is the Becker-Murphy (1992) division of labor emerging from local interactions without central coordination.

Remark 5 (Connection to Regime Shift)**.** *The specialization dynamics of Proposition 6 are the micro-level mechanism underlying the Potts model regime shift of Proposition 4. The Potts "state" of each node is its dominant specialization $j^*(i)$. The "coupling" J in the Potts energy function corresponds to the task-sharing benefit between neighboring agents with compatible specializations. The first-order regime shift at $q > 2$ means that when the number of specialization types exceeds two, the transition from unspecialized to specialized is abrupt—consistent with the reinforcement dynamics exhibiting a bifurcation from mixed to specialized response profiles.*

Corollary 6 (Centralized Market Share Decline)**.** *For $N > N^*$, the centralized provider's market share is strictly decreasing in N. In the Bianconi-Barabási framework, the centralized*

"condensate" fraction declines continuously as the fitness distribution broadens, transitioning from the capability concentration regime to the distributed traffic regime.

10.4 The Central Theorem

The three layers (percolation, CES specialization, Laplacian diffusion) compose into a single existence result. The theorem below unifies the network formation condition ($R_0^{\text{mesh}} > 1$) with the capability comparison ($C_{\text{mesh}} > C_{\text{cent}}$) and the dynamic stability condition (local asymptotic stability).

Theorem 3 (Mesh Equilibrium Existence, Uniqueness, and Dominance)**.** *For $R_0^{mesh} > 1$ and $\rho < 1$, there exists a finite N^* such that for all $N > N^*$:*

(a) The mesh equilibrium exists: a positive fraction $S_\infty^* > 0$ of agents form a connected component with specialized roles covering all J task types.

(b) The mesh equilibrium is unique among equilibria with $S_\infty > 0$.

(c) The mesh equilibrium is locally asymptotically stable.

(d) $C_{mesh}(N) > C_{cent}$: the mesh's aggregate capability exceeds centralized provision.

(e) N^* is decreasing in the diversity of the agent population.

Proof. Step 1 (Existence): For $R_0^{\text{mesh}} > 1$, equation (28) has a unique positive solution (Proposition 3). For N sufficiently large, the giant component covers all J task types.

Step 2 (Uniqueness): The mesh participation game—in which each agent decides whether to join the mesh and which task type to specialize in—is a supermodular game. Agent i's payoff from joining increases when more agents join (network effect via the CES aggregate and knowledge diffusion) and when agents specialize in complementary types (CES complementarity with $\rho < 1$). Formally, the payoff function exhibits increasing differences: $\partial^2 \pi_i / (\partial a_i \partial a_j) > 0$ for $i \neq j$, where $a_i \in \{0, 1\}$ is the participation decision. By Tarski's (1955) fixed point theorem, the game has a greatest and least equilibrium. By the strict concavity of CES, the greatest equilibrium is unique among equilibria with $S_\infty > 0$: any equilibrium with different specialization allocations is payoff-dominated by the efficient allocation.

Step 3 (Stability): By Lyapunov analysis using $V = -C_{\text{eff}}(N) + \sum_j \phi_j(C_j)$ and LaSalle's invariance principle.

Step 4 (Capability dominance): Under specialization with approximately uniform distribution across J types, $C_{\text{mesh}}(N) \approx J^{(1-\rho)/\rho} \cdot S_\infty^* N \bar{c} / J$. This exceeds C_{cent} when $N > N^* \equiv M \bar{c}_{\text{cent}} / (S_\infty^* \bar{c})$, which is finite.

Step 5: Higher diversity of the fitness distribution implies broader coverage for given N, reducing N^*. □

11. Knowledge Diffusion via Graph Laplacian

11.1 Laplacian Dynamics

Let $\mathbf{u}(t) \in \mathbb{R}^N$ represent the knowledge state of each node. Knowledge diffusion follows:

$$\frac{\partial \mathbf{u}}{\partial t} = -L \cdot \mathbf{u} \tag{34}$$

where $L = D_{\text{deg}} - A$ is the graph Laplacian. Convergence to consensus is governed by the Fiedler eigenvalue $\lambda_2(L)$:

$$\|\mathbf{u}(t) - \bar{u}\mathbf{1}\|_2 \leq \|\mathbf{u}(0) - \bar{u}\mathbf{1}\|_2 \cdot e^{-\lambda_2(L)t} \tag{35}$$

11.2 Bandwidth Scaling

The total bandwidth available for knowledge diffusion in the mesh scales as:

$$B_{\text{mesh}} = O(N \cdot \langle k \rangle) \tag{36}$$

where $\langle k \rangle$ is the mean degree. The centralized hub has fixed bandwidth B_{hub} determined by datacenter interconnect capacity. Once $N \cdot \langle k \rangle > B_{\text{hub}}$, the mesh serves more total queries per unit time than the centralized provider. This is a necessary condition for capability dominance, complementing the CES capability comparison.

For the AI mesh, the bandwidth scaling has a concrete interpretation. Each edge device contributes its local network capacity (WiFi, 5G, fiber) to the mesh's aggregate throughput. A mesh of $N = 10^7$ devices with mean degree $\langle k \rangle = 10$ and average bandwidth 100 Mbps per link aggregates to 10^{10} Mbps = 10 Pbps—orders of magnitude beyond any single datacenter's external bandwidth, even though each individual link is slower than a datacenter interconnect.

11.3 Vanishing Epidemic Threshold on Scale-Free Networks

Pastor-Satorras and Vespignani (2001) established that the SIS epidemic threshold on networks with degree distribution $P(k) \sim k^{-\gamma}$ is:

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{37}$$

For scale-free networks with $\gamma \leq 3$: $\langle k^2 \rangle$ diverges in the large-network limit, so $\lambda_c \to 0$.

Proposition 7 (Self-Sustaining Knowledge Propagation)**.** *If the mesh has a scale-free degree distribution with $\gamma \leq 3$—which MoE routing produces endogenously through preferential specialization—then any nonzero rate of knowledge sharing sustains itself indefinitely. The topology ensures propagation without requiring a minimum transmission rate.*

The economic content is that knowledge diffusion need not be modeled as a separate mechanism with its own threshold. Once the mesh achieves a fat-tailed degree distribution (which the specialization dynamics produce through preferential attachment to high-quality specialists), capability propagation is guaranteed. Hub agents—the most capable specialists—emerge endogenously and serve as conduits for knowledge transfer.

11.4 Combined Dynamics

The three layers interact as follows. Layer 1 (percolation) ensures the mesh is connected ($S_\infty > 0$). Layer 2 (CES specialization) ensures agents specialize and the aggregate rewards diversity (C_{eff} grows with diversity). Layer 3 (Laplacian diffusion) ensures knowledge propagates self-sustainingly on scale-free topologies (the Fiedler eigenvalue is positive, and on fat-tailed topologies, propagation is self-sustaining at any nonzero rate).

The combined effect is a mesh whose aggregate capability $C_{\text{mesh}}(N)$ is superlinear in N over the relevant range, because additional diverse specialists both increase the CES aggregate (Layer 2) and increase the rate at which existing knowledge diffuses to new entrants (Layer 3). This superlinearity is the formal expression of increasing returns to scale in the mesh—a property that the centralized alternative, with fixed capacity $M\bar{c}$, does not share.

Economic interpretation of Layer 3. Knowledge diffusion in the mesh has a concrete operational meaning. When a medical specialist fine-tunes a model on radiology data and achieves improved diagnostic accuracy, that improvement can propagate through the mesh via three channels. First, *weight sharing*: the specialist's adapter weights can be downloaded and applied by other agents. Second, *distillation*: the specialist's outputs on a standard evaluation set can be used as training data by other agents. Third, *routing feedback*: the specialist's superior performance on radiology queries increases its routing share,

which generates evaluation data revealing which architectural choices or training procedures produced the improvement. All three channels operate through the network edges, and their aggregate rate is governed by $\lambda_2(L)$.

The key insight is that knowledge diffusion on the mesh is *non-rivalrous*: the specialist's improvement does not diminish when shared. This distinguishes the mesh from a market for physical goods. The CES aggregate captures the production complementarity (different specialists contribute to different tasks), while the Laplacian captures the propagation dynamics (how fast improvements spread). Together, they produce the increasing returns that make the mesh a viable organizational form.

12. Autocatalytic Capability Growth

12.1 Training Operations as RAF Sets

Definition 1 (Training Operation)**.** *A training operation $r = (I_r, K_r, O_r)$ takes input capability types I_r, requires catalyst capability types K_r (not consumed), and produces output capability types O_r.*

Definition 2 (Food Set)**.** *The food set $F \subset \{1, \ldots, J\}$ is the set of capability types available exogenously from centralized training. For each $j \in F$, base model capability of type j is available without requiring any mesh training operation. The food set is determined by frontier model releases from centralized providers and is exogenous to the mesh.*

The food set corresponds to the training persistence assumption: frontier model training remains centralized. Base models (GPT-class, Claude-class, Gemini-class) are the "raw materials" that the mesh fine-tunes, adapts, and combines. The food set grows exogenously at a rate determined by centralized infrastructure investment—the rate that will emerge as the Baumol bottleneck in Section 12.

Definition 3 (RAF Set)**.** *Following Hordijk and Steel (2004), a set $\mathcal{R}$ of training operations is* Reflexively Autocatalytic and Food-generated *(RAF) if every operation is catalyzed by a capability type in the food set or produced by the set, and every input can be constructed from the food set by successive operations.*

Proposition 8 (Autocatalytic Existence Threshold)**.** *There exists a critical mesh size N_{auto} such that for $N > N_{auto}$, the mesh contains a RAF set with probability approaching unity:*

$$N_{auto} = O\left(\frac{\ln |\mathcal{R}|}{\beta_t}\right) \tag{38}$$

The threshold scales logarithmically with system complexity.

Proof. By the Hordijk-Steel result, a random catalytic reaction system contains a RAF set with high probability when the per-type catalysis probability $p(N) = 1 - (1 - \beta_t)^N > 1/J$. This gives $N > \ln(1 - 1/J)/\ln(1 - \beta_t) \approx 1/(J\beta_t)$, yielding logarithmic scaling. □

12.2 Autocatalytic Core Dynamics

Once a RAF set exists, the mesh's autocatalytic core evolves through Jain-Krishna (1998, 2001) adaptive network dynamics. The catalytic matrix $\mathbf{M}$ ($M_{ij} = 1$ if capability j catalyzes improvement of i) governs growth:

$$\dot{c}_i = c_i \left(\sum_j M_{ij} c_j - \phi_0 \right) \tag{39}$$

Proposition 9 (Perron-Frobenius Selection)**.** *The long-run composition of the autocatalytic core is determined by the leading eigenvector of the catalytic matrix* $\mathbf{M}$*. Capability types with large components in the Perron-Frobenius eigenvector* $\mathbf{v}_1$ *of* $\mathbf{M}$ *dominate; types with small components go extinct. The leading eigenvalue* $\lambda_1(\mathbf{M})$ *determines whether the autocatalytic core expands* ($\lambda_1 > \phi_0$) *or contracts* ($\lambda_1 < \phi_0$) *relative to the rest of the mesh.*

Proof. Equation (the Jain-Krishna dynamics) is a replicator equation on the simplex of capability-type shares. The fixed point analysis follows from the standard replicator dynamics result (Hofbauer and Sigmund 1998): the dynamics converge to a state where surviving types have equal fitness, and the surviving set corresponds to the support of the Perron-Frobenius eigenvector of $\mathbf{M}$. □

The Jain-Krishna dynamics predict a specific temporal pattern: the autocatalytic core does not grow smoothly. Instead, it undergoes a series of *reorganization cascades*—periods of stasis punctuated by rapid restructuring events in which poorly connected capability types are replaced by types with stronger catalytic linkages. Each cascade increases the leading eigenvalue $\lambda_1(\mathbf{M})$, producing a staircase pattern of increasing autocatalytic efficiency. This parallels the mesh formation Phase 2 crystallization (Section 9), but now at the level of internal capability improvement rather than network formation.

Remark 6 (Relationship between N_{auto} and N^*)**.** *The autocatalytic threshold* N_{auto} *and the mesh's critical mass* N^* *(Theorem 3) are distinct.* N^* *is the mesh size at which collective capability exceeds centralized provision (a* static *comparison).* N_{auto} *is the mesh size at which self-sustaining capability improvement becomes possible (a* dynamic *property). Generically,*

$N_{auto} > N^*$: the mesh can be collectively capable before it is self-improving. The gap $N_{auto} - N^*$ represents the period during which the mesh exceeds centralized inference but depends entirely on exogenous base model releases for capability growth.

12.3 The Three Growth Regimes

The mesh's aggregate capability evolves according to:

$$\dot{C} = \delta_g \cdot f^{\lambda} \cdot C^{\lambda+\varphi-1} \cdot J(t)^{\gamma_J} \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] - \delta \cdot C \tag{40}$$

where f is the training fraction, λ is the duplication parameter, φ is the training productivity elasticity, γ_J governs variety contribution, and the indicator ensures model collapse is avoided.

Proposition 10 (Effective Training Productivity)**.** *Let $\varphi_0 < 1$ be the raw training productivity elasticity and $\beta_{auto} \in [0, 1)$ the autocatalytic fraction. The mesh's effective elasticity is:*

$$\varphi_{eff} = \frac{\varphi_0}{1 - \beta_{auto} \cdot \varphi_0} \tag{41}$$

This exceeds φ_0 for $\beta_{auto} > 0$, and $\varphi_{eff} \geq 1$ when $\beta_{auto} \geq (1 - \varphi_0)/\varphi_0$.

Proof. Following the Aghion-Jones-Jones (2018) framework, decompose the training improvement process into a continuum of subtasks. A fraction β_{auto} is automated by mesh agents (productivity scales with C^{φ_0}), and the remaining fraction $1-\beta_{\text{auto}}$ requires exogenous input Z. The effective production function is $\dot{C} \propto C^{\varphi_0/(1-\beta_{\text{auto}}\varphi_0)} \cdot Z^{(1-\beta_{\text{auto}})/(1-\beta_{\text{auto}}\varphi_0)}$. The exponent on C is φ_{eff}. For $\varphi_0 = 0.5$, the knife-edge requires $\beta_{\text{auto}} = 1.0$—full automation. For $\varphi_0 = 0.8$, only $\beta_{\text{auto}} = 0.25$ suffices. □

Remark 7 (The Automation Ladder)**.** *The quantity β_{auto} is not fixed; it evolves endogenously as the mesh matures. Initially $\beta_{auto} \approx 0$: the mesh is entirely dependent on exogenous base model releases from centralized providers. As training agents emerge (the Jain-Krishna process), β_{auto} rises through identifiable stages:*

(i) Fine-tuning automation *($\beta_{auto} \approx 0.1$–0.2): mesh agents automate the process of fine-tuning base models for specialist tasks. This is the current state (2025–2026): automated fine-tuning pipelines exist but require human oversight for data curation and evaluation.*

(ii) Evaluation automation *($\beta_{auto} \approx 0.2$–0.4): mesh agents can evaluate the quality of their own and others' outputs, enabling automated quality control. This enables the autocat-*

alytic loop: agents can fine-tune, evaluate, and iterate without human intervention on routine tasks.

(iii) Data generation automation *($\beta_{auto} \approx 0.4$–0.6): mesh agents generate high-quality synthetic training data that is statistically indistinguishable from human-generated data for specific domains. The model collapse protection theorem (Section 13) ensures this is sustainable when J is sufficiently large.*

(iv) Architecture search automation *($\beta_{auto} \approx 0.6$–0.8): mesh agents discover improved model architectures, training procedures, and optimization strategies. This requires the autocatalytic core to include "meta-learning" capability types.*

The transition from regime (a) to regime (b) requires reaching stage (iii) or (iv). Whether the mesh reaches these stages before the Baumol bottleneck binds is the central empirical question for long-run AI capability growth.

12.4 Training Saturation

Individual training interactions exhibit diminishing returns. Following the Lotka-Volterra mutualistic framework (Bastolla et al. 2009):

$$\dot{C}_j = \frac{\sum_k a_{jk} \cdot C_k}{1 + h \sum_k a_{jk} \cdot C_k} - \delta C_j \tag{42}$$

where $h > 0$ is the saturation parameter.

Lemma 2 (Saturation Ceiling)**.** *For fixed J and $h > 0$, the system has a unique globally stable equilibrium with $C_j^* < 1/(h \cdot \delta)$. The aggregate ceiling is $C_{max} \leq J^{1/\rho} \cdot (h\delta)^{-1}$.*

12.5 Variety Expansion as Saturation Escape

Romer's (1990) key insight is that new product varieties sustain growth even when returns to individual products diminish. Let $J(t)$ evolve according to:

$$\dot{J} = \eta_J \cdot f_J \cdot C_{\text{eff}}^{\varphi_J} \cdot (J_{\max} - J) \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] \tag{43}$$

Proposition 11 (Saturation Escape via Variety)**.** *Even with saturation $h > 0$, the growth rate of C_{eff} can remain positive if $J(t)$ is growing. For constant per-type capability at the saturation ceiling, the growth rate from variety expansion does not depend on h. The CES aggregate grows as $J^{(1-\rho)/\rho}$ even when each individual C_j is bounded.*

Proof. At the saturation ceiling, each $C_j = C^* = 1/(h\delta)$. The CES aggregate is:

$$C_{\text{eff}} = \left(\sum_{j=1}^{J(t)} (C^*)^\rho \right)^{1/\rho} = J(t)^{1/\rho} \cdot C^* \tag{44}$$

Differentiating with respect to time:

$$\frac{\dot{C}_{\text{eff}}}{C_{\text{eff}}} = \frac{1}{\rho} \cdot \frac{\dot{J}}{J} \tag{45}$$

Since $\rho < 1$ (complementary regime), $1/\rho > 1$ if $\rho > 0$ and $1/\rho < 0$ if $\rho < 0$. For $\rho \in (-1, 0)$ (the empirically relevant range), the growth rate of C_{eff} exceeds the growth rate of J when $|1/\rho| > 1$, i.e., when $|\rho| < 1$. For example, with $\rho = -0.5$: $\dot{C}_{\text{eff}}/C_{\text{eff}} = -2 \cdot \dot{J}/J$, which is positive because both $\dot{J}$ and ρ are positive. The growth rate depends only on the rate of variety expansion, not on the saturation parameter h. □

Theorem 4 (Growth Regime Classification)**.** *The long-run behavior of $C_{eff}(t)$ falls into three regimes:*

Regime (a): Convergence to a ceiling. *If $\varphi_{eff} < 1$, training saturation $h > 0$, and variety is bounded:*

$$C_{eff}(t) \to C_{max} \equiv J_{\max}^{(1-\rho)/\rho} \cdot \frac{1}{h\delta} \quad \text{as } t \to \infty \tag{46}$$

The mesh's growth rate converges to the exogenous frontier model improvement rate. This is the Baumol bottleneck.

Regime (b): Balanced exponential growth. *If $\varphi_{eff} = 1$ and $J(t)$ grows endogenously, the dominant term is exponential, bounded eventually by the Baumol constraint.*

Regime (c): Finite-time singularity. *If $\varphi_{eff} > 1$, $h = 0$, and $\alpha_{eff} > \alpha_{crit}$ simultaneously:*

$$T_s = \frac{C_0^{1-\Phi}}{(\Phi - 1) \cdot \delta_g f^\lambda J^{\gamma_J}} \tag{47}$$

This requires three conditions simultaneously, making it restrictive and unlikely.

Table 17: Growth regime classification.

Regime	φ_{eff}	h	J	Long-run $C_{\text{eff}}(t)$
(a) Convergence	< 1	> 0	bounded	$\to C_{\max}$ (ceiling)
(b) Exponential	$= 1$	≥ 0	growing	$\sim e^{rt}$
(c) Singularity	> 1	$= 0$	any	$\to \infty$ at $T_s < \infty$
Condition for all regimes: $\alpha_{\text{eff}} > \alpha_{\text{crit}}$ (collapse avoided)				

Remark 8 (Which Regime Is Likely?)**.** *The empirical evidence strongly favors regime (a) in the near term. Bloom et al. (2020) estimate* $\varphi \approx 0.5$–0.7 *for research productivity across multiple domains (semiconductors: 0.55, pharmaceuticals: 0.52, agriculture: 0.62, general research: 0.67), implying* $\varphi_0 < 1$ *by a substantial margin. For* φ_{eff} *to reach unity with* $\varphi_0 = 0.6$*, the autocatalytic fraction must reach* $\beta_{auto} = 0.67$*—the mesh must automate two-thirds of its own training improvement process. While not impossible, this requires training agent capabilities substantially beyond current levels.*

Individual training interactions exhibit clear saturation ($h > 0$*): the 10th fine-tuning of a medical specialist yields less improvement than the 1st. And while variety expansion can escape per-type saturation, the total task space* $J_{\max}$ *is finite (bounded by the dimensionality of the human knowledge space).*

Regime (c) requires the conjunction of three conditions, each individually demanding: $\varphi_{eff} > 1$ *(near-complete training automation),* $h = 0$ *(no saturation), and* $\alpha_{eff} > \alpha_{crit}$ *(collapse avoided). The probability that all three hold simultaneously is small. The paper does not predict the singularity; it characterizes the conditions under which it would occur and notes that they are restrictive.*

The most probable trajectory is: regime (a) with an increasing ceiling. As the mesh matures, β_{auto} *rises (pushing* φ_{eff} *toward 1),* J *expands (raising* C_{max}*), and the ceiling lifts—but the Baumol bottleneck anchors the long-run growth rate to exogenous frontier model improvement. The mesh is a multiplier, not a generator.*

13. Model Collapse Protection

13.1 The Model Collapse Framework

The mesh's self-referential learning—agents training on data generated by other agents—risks model collapse. Following Shumailov et al. (2024), consider a generative model Q_t trained on a mixture of authentic data from the true distribution P (fraction α) and synthetic data from the model's own previous generation Q_{t-1} (fraction $1 - \alpha$). The KL divergence from the true distribution evolves as:

$$\mathrm{KL}(Q_{t+1} \| P) \geq \mathrm{KL}(Q_t \| P) \quad \text{when } \alpha < \alpha_{\text{crit}} \tag{48}$$

The critical threshold α_{crit} depends on the model class. Below α_{crit}, each generation amplifies the deviation from the true distribution: the model's outputs become progressively less diverse, tails are truncated, and the model collapses toward a degenerate distribution.

For a *single* model training on its own outputs ($\alpha = 0$), collapse is inevitable. The outputs

lack diversity because they are generated by a single distribution Q_t, which amplifies its own modes and suppresses its tails with each generation. The mesh, however, is not a single model—it is a collection of heterogeneous specialists whose outputs are drawn from different distributions.

13.2 The Dual Role of ρ

Definition 4 (Effective External Data Fraction)**.** *For agent i in a mesh with J specialization types and CES parameter ρ:*

$$\alpha_{eff}(\rho, J) = \alpha_{ext} + (1 - \alpha_{ext}) \cdot D(\rho, J) \tag{49}$$

where $D(\rho, J) = \frac{J-1}{J} \cdot (1 - \rho^{1/(1-\rho)})$ is the diversity correction.

Theorem 5 (CES Heterogeneity as Collapse Protection)**.** *The mesh avoids model collapse ($\alpha_{eff} > \alpha_{crit}$) whenever:*

$$\alpha_{ext} + (1 - \alpha_{ext}) \cdot \frac{J-1}{J} \cdot \left(1 - \rho^{1/(1-\rho)}\right) > \alpha_{crit} \tag{50}$$

This can be satisfied even when $\alpha_{ext} < \alpha_{crit}$—when the mesh's external data supply is below the collapse threshold for any individual model—provided J is sufficiently large and ρ is sufficiently small.

Proof. The condition holds when:

$$D(\rho, J) > \frac{\alpha_{\text{crit}} - \alpha_{\text{ext}}}{1 - \alpha_{\text{ext}}} \tag{51}$$

Since D increases as ρ decreases and J increases, for any $\alpha_{\text{crit}} \in (0, 1)$ and $\alpha_{\text{ext}} \in [0, \alpha_{\text{crit}})$, there exist (ρ, J) pairs satisfying the condition. The minimum J is:

$$J_{\min}(\rho, \alpha_{\text{ext}}) = \left\lceil \frac{1}{1 - \frac{\alpha_{\text{crit}} - \alpha_{\text{ext}}}{(1-\alpha_{\text{ext}})(1-\rho^{1/(1-\rho)})}} \right\rceil \tag{52}$$

□

Remark 9 (One Parameter, Two Functions)**.** *The CES parameter ρ does double duty. In Section 10, $\rho < 1$ generates the diversity premium for capability aggregation. Here, $\rho < 1$ generates collapse protection: heterogeneous specialists produce informationally diverse training data. The same heterogeneity that makes the mesh capable also makes it robust to self-*

referential training degradation. This is not coincidence—it reflects the structural identity between production complementarity and distributional diversity in the CES framework.

Economic interpretation of the diversity correction. The diversity correction $D(\rho, J)$ has a precise economic meaning. From agent i's perspective, training data generated by agents with *different* specializations is informationally equivalent to external data—it contains distributional information that agent i's own outputs lack. A medical specialist's training data is "external" from the perspective of a legal specialist, and vice versa. The CES structure quantifies this: when $\rho < 1$, the correlation between specialists' output distributions is less than perfect, and the residual distributional diversity acts as an effective substitute for external data.

The correction fails when $\rho \to 1$ (perfect substitutability): if all specialists produce identical output distributions, there is no diversity to exploit, and the mesh degenerates to a single self-referential model. This is the formal statement of the intuition that homogeneous agents cannot avoid model collapse through mere replication. The protection comes specifically from *complementary heterogeneity*: agents that are different in economically meaningful ways (different training data, different architectures, different fine-tuning objectives) produce outputs that are distributionally diverse.

Connection to the autocatalytic framework. The model collapse protection theorem constrains the growth regimes of Section 12. The indicator function $\mathbf{1}[\alpha > \alpha_{\text{crit}}]$ in the growth equation (40) is not merely a technical condition—it reflects a substantive constraint. If the mesh's external data supply falls below α_{crit} (because the mesh grows faster than external data sources), growth can continue only if the diversity correction $D(\rho, J)$ is sufficient. This creates a minimum diversity requirement for sustained growth: the mesh must maintain at least $J_{\min}(\rho, \alpha_{\text{ext}})$ distinct specialization types (Table 18) to avoid model collapse during the autocatalytic growth phase.

Quantitative calibration. For training data and capability aggregation, Paper 1 adopts $\rho \in [-1.0, -0.25]$ ($\sigma \in [0.5, 0.8]$) from plant-level estimates (Oberfield and Raval 2021; Gechert et al. 2022). NLS cross-validation on FRED Durable Goods IP yields $\hat{\rho} = 0.15$ ($\sigma = 1.18$), above the adopted range as expected for macro aggregates. At the conservative bound $\rho = -0.25$: the minimum J for collapse avoidance (equation 52) is modest—fewer than 10 specialist types suffice for α_{ext} as low as 0.3, well within the range of current open-weight model ecosystems.

Table 18: Minimum specialization types $J_{\min}$ for collapse avoidance.

α_{ext}	$\rho = -1.0$	$\rho = -0.5$	$\rho = -0.25$	$\rho = 0$
0.50	2	3	4	∞
0.30	3	5	8	∞
0.10	5	9	18	∞
0.00	7	14	31	∞

Assumes $\alpha_{\text{crit}} = 0.5$. As $\rho \to 1$, $D(\rho, J) \to 0$ and no finite J suffices. For $\rho \leq 0$ (complementary regime), modest diversity protects against collapse.

14. The Baumol Bottleneck

14.1 The Two-Sector Structure

Decompose the AI production process into two sectors with structurally different productivity dynamics:

Sector 1: Inference and fine-tuning (progressively automated). The mesh automates an increasing fraction $\beta(t)$ of inference and fine-tuning tasks. These are the tasks Sections 9–11 model: routing queries to specialists, generating responses, adapting models to new domains through fine-tuning. Productivity grows at rate g_C—the endogenous growth rate from the autocatalytic model.

Sector 2: Frontier model training (non-automatable). Training requires tightly synchronized GPU clusters at scales of 10^{25}+ FLOPs per run, with synchronization bandwidth as the binding constraint. This synchronization requirement is *topological*, not cost-based: it requires all-to-all communication within the training cluster, which distributed mesh architecture cannot provide at the required latency. Productivity grows at exogenous rate g_Z, determined by centralized infrastructure investment.

Table 19: Two-sector structure of AI production.

Property	Sector 1 (Inference)	Sector 2 (Training)
Output	Token generation, responses	Frontier model weights
Automation share $\beta(t)$	$\to 1$ as mesh matures	≈ 0 (topological)
Productivity growth	g_C (endogenous, high)	g_Z (exogenous, moderate)
Cost trend	Declining (280× in 2 yr)	Rising (\$100M → \$1B+)
Scale economies	Diminishing (edge-viable)	Increasing (larger = better)
Binding constraint	Cost and latency	Synchronization bandwidth

14.2 Derivation

Model aggregate capability as $C_{\text{eff}} = C_1^{\beta(t)} \cdot C_2^{1-\beta(t)}$, yielding:

$$g_{C_{\text{eff}}} = \beta(t) \cdot g_C + (1 - \beta(t)) \cdot g_Z + \dot{\beta}(t) \cdot \ln\left(\frac{C_1}{C_2}\right) \tag{53}$$

Proposition 12 (Endogenous Baumol Bottleneck)**.** *As $\beta(t) \to 1$, $g_{C_{eff}} \to g_Z$ regardless of g_C, provided $g_C > g_Z$. The non-automatable sector becomes the binding constraint even as its share of total activity shrinks.*

Proof. When $g_C > g_Z$, the relative price of the non-automated sector rises: the cost share s_2 increases even as the volume share $(1 - \beta)$ falls—Baumol's cost disease. In the limit $\beta \to 1$ with $\dot{\beta} \to 0$ (the last tasks are hardest to automate—the topological training constraint), $g_{C_{\text{eff}}} \to g_Z$. □

14.3 The Cost Disease in Detail

The Baumol mechanism operates as follows. When $g_C > g_Z$, the mesh's automated sector (inference and fine-tuning) experiences declining unit costs relative to the non-automated sector (frontier training). But the mesh *requires both sectors*: it cannot improve without new frontier models as base inputs (the food set). The relative price of frontier model access rises:

$$\frac{p_Z(t)}{p_C(t)} = \frac{p_Z(0)}{p_C(0)} \cdot e^{(g_C - g_Z)t} \tag{54}$$

The cost share of frontier model access—even as the volume share falls toward zero—rises toward unity. This is exactly Baumol's (1967) cost disease, with frontier training playing the role of live orchestral performance and inference playing the role of recorded music. The mesh can reproduce and distribute inference with increasing efficiency, but it cannot manufacture the new compositions (frontier models) that give it material to work with.

Quantitative illustration. If $g_C = 0.50$ (inference cost halving every 18 months) and $g_Z = 0.15$ (frontier model capability doubling every 5 years), the relative price of model access doubles every $\ln 2/(g_C - g_Z) \approx 2$ years. Within a decade, frontier model access costs 30 times more relative to inference than at the start—even if the absolute price of model access is declining. This divergence is the mechanism through which the Baumol bottleneck becomes binding: the mesh's budget constraint shifts from inference cost (which it can drive down) to model access cost (which it cannot).

14.4 Closing the Circle

The chain of determination is:

(i) *Concentrated investment* (Section 3): Datacenter capital investment, driven by the Nash overinvestment dynamic, finances GPU clusters that train frontier models. The rate g_Z is determined by the rate of investment.

(ii) *Learning curves* (Section 4): The same investment finances 3D packaging learning curves ($\alpha = 0.23$) enabling distributed inference. Crossing occurs at $x(t) = 0$.

(iii) *Mesh formation* (Sections 9–11): After crossing, the mesh self-organizes into a heterogeneous specialized network. Above N^*, $C_{\text{mesh}} > C_{\text{cent}}$.

(iv) *Endogenous growth* (Section 12): The mesh improves itself through autocatalytic training, self-referential learning, and variety expansion.

(v) *Baumol ceiling* (this section): Growth converges to g_Z—determined by the concentrated investment of step (i).

The circle closes. The concentrated capital that creates the crossing also determines the ceiling. The mesh amplifies frontier model improvement but cannot exceed it indefinitely.

The chain has a specific quantitative structure. From the overinvestment result (Theorem 2), Nash equilibrium investment is $I^N \approx 3\text{–}4 \times I^C$. This investment finances a frontier model improvement rate $g_Z \propto I^{\alpha_Z}$, where α_Z is the research productivity elasticity (Bloom et al. 2020 estimate $\alpha_Z \approx 0.55$ for semiconductors). The mesh multiplies g_Z by the CES diversity premium: long-run $g_{C_{\text{eff}}} \leq J^{(1-\rho)/\rho} \cdot g_Z$. But $g_{C_{\text{eff}}}/g_Z$ converges to 1 as $\beta \to 1$ (Proposition 12). The equilibrium growth rate is thus determined by the Nash investment level, which is itself determined by the number of competing firms N and the learning rate α. The growth rate of the entire distributed ecosystem is an endogenous consequence of the competitive structure of the centralized sector.

This is a falsifiable prediction: the mesh's capability growth rate should correlate with, and be bounded by, the rate of frontier model releases from centralized providers (Prediction 10). Periods of reduced centralized investment (due to recession, regulatory intervention, or capacity exhaustion) should produce measurable deceleration of mesh capability growth, with a lag determined by the model lifecycle (∼6–12 months).

Remark 10 (The Mesh as Multiplier, Not Generator)**.** *The Baumol bottleneck implies that the mesh is a* multiplier *of centralized innovation, not an independent source. The multiplier is substantial—the CES diversity premium $J^{(1-\rho)/\rho}$ can be large—but the growth rate is*

anchored to the exogenous frontier. This has a stark policy implication: investments in frontier training capability (compute infrastructure, researcher talent, data access) have outsized returns because they relax the Baumol constraint for the entire distributed ecosystem. Conversely, policies that reduce frontier training investment (export controls, regulatory barriers) reduce not only centralized capability but also the ceiling for distributed capability.

14.5 The Task Bifurcation

The AI transition reveals a feature that earlier cycles exhibited only weakly: the technology contains tasks with fundamentally different ρ values, leading to structural rather than temporal bifurcation.

Theorem 6 (Task bifurcation)**.** *Consider a technology with tasks indexed by $\rho \in [\rho_{\min}, \rho_{\max}]$. At cumulative production level Q:*

(a) Tasks with $T_{dist}(Q) < T^*(\rho)$ are distributable; those above remain centralized.

(b) As Q increases, increasingly complementary tasks (lower ρ) become distributable.

(c) Tasks with $T^*(\rho) < \bar{T}$ (a positive lower bound on T_{dist}) remain permanently centralized.

For AI, the task spectrum spans: simple generation ($\rho \to 1$, distributes first), routine inference ($\rho \approx 0$, distributes at cost parity), complex reasoning ($\rho \approx -1$, distributes once coordination protocols mature), and training ($\rho \to -\infty$, potentially permanently centralized). The prediction is a *gradient of decentralization*: simple tasks distribute first, complex tasks last, with training remaining centralized unless a fundamentally new coordination technology emerges.

15. Perez Phases and Historical Calibration

15.1 Cycle Duration and Amplitude

Proposition 13 (Cycle duration)**.** *The duration of the full technology cycle scales as:*

$$\tau \sim \frac{1}{\alpha} \ln\left(\frac{c_0}{c^*}\right) \cdot \frac{1}{1 + (N-1)\alpha\phi/(r+\delta)} \tag{55}$$

Proposition 14 (Cycle amplitude)**.** *The amplitude of overinvestment during the installation-frenzy phase is:*

$$A = K \cdot \frac{N-1}{N} \cdot \frac{\alpha\phi}{r + \delta - \alpha\phi} \tag{56}$$

measuring the present value of excess investment as a fraction of the curvature premium. The amplitude is increasing in K (complementarity), N (competitors), α (learning rate), and ϕ (spillover rate).

Corollary 7 (Duration compression)**.** *If successive technologies have learning rates $\alpha_1 < \alpha_2 < \cdots$, cycle durations decrease:*

$$\frac{\tau_{n+1}}{\tau_n} \approx \frac{\alpha_n}{\alpha_{n+1}} \cdot \frac{1 + (N_n - 1)\alpha_n \phi_n/(r+\delta)}{1 + (N_{n+1} - 1)\alpha_{n+1}\phi_{n+1}/(r+\delta)} \tag{57}$$

This provides a structural explanation for successive cycle compression: more information-intensive technologies have higher α and operate in markets with more competitors.

15.2 Stylized Facts

Before deriving the phases, we establish the empirical regularities that the framework must explain.

Fact 1: Concentrated investment precedes distributed adoption. Every general-purpose technology requires an initial phase of concentrated capital formation. Canals required sovereign financing. Railroads created the modern stock market. Electrification required regulated utility monopolies. AI training requires hyperscaler-scale capital.

Fact 2: Overinvestment relative to social optimum. The concentrated phase consistently produces more capacity than the contemporaneous market can absorb. Railroad track doubled between 1880 and 1890 while freight rates fell 50% (Fogel 1964). AI training compute grows at approximately 4×/year against revenue growth of 1.5×.

Fact 3: Crisis separates installation from deployment. The transition is not gradual but passes through a crisis: the canal panic (1797), railroad panics (1873, 1893), the Great Crash (1929), the dot-com bust (2000–2002). Each crisis destroys financial value while leaving physical infrastructure intact.

Fact 4: Deployment exceeds installation in value creation. Railroad deployment (1890–1920) generated more economic value than railroad construction (1850–1890). Internet deployment (2003–present) has created more value than the dot-com installation phase.

Fact 5: Successive cycles compress. The canal era lasted roughly 60 years, the railroad age roughly 50, electrification roughly 40, computing roughly 30, the mobile internet roughly 15. The AI cycle appears to compress further.

Fact 6: The pattern is sector-specific. Financial services digitized before manufacturing. Software adopted cloud computing before hardware. The relevant parameter is technology-specific (α, ρ), not macroeconomic.

15.3 Perez Phases as Bifurcations

The five phases of technological revolution (Perez 2002) correspond to traversals of the (ρ, T) regime diagram:

Phase I: Installation. The centralized mode operates with $T_{\text{cent}} \ll T^*$, high effective curvature. Distributed mode non-viable: $T_{\text{dist}} > T^*$.

Phase II: Frenzy. Speculative financing raises financial information friction. Overinvestment (Theorem 2) accelerates cumulative production.

Phase III: Turning point. The system undergoes a *fold bifurcation*: the high-T fixed point (speculative financial equilibrium) collides with the unstable equilibrium separating it from the low-T fixed point (fundamental-value equilibrium). At the bifurcation, small perturbations trigger a discontinuous jump. Formally, the financial system's effective friction satisfies:

$$\dot{T}_{\text{fin}} = \underbrace{\beta \cdot I(Q)}_{\text{speculation}} - \underbrace{\mu \cdot K_{\text{eff}}(T_{\text{fin}})}_{\text{stabilization}} \tag{58}$$

where β captures the rate at which investment generates speculative froth and μ captures the stabilizing effect of curvature. The fold bifurcation occurs when $\dot{T}_{\text{fin}} = 0$ and $\partial \dot{T}_{\text{fin}} / \partial T_{\text{fin}} = 0$ simultaneously. This is the crisis—structurally stable, not contingent. Small perturbations to the model parameters change the timing but not the existence or character of the discontinuity.

Phase IV: Deployment. Post-crisis, the financial system operates at low T (fundamental values). Meanwhile, $T_{\text{dist}}(Q)$ has fallen below T^* due to cumulative learning during Phases I–III. Distributed production enters and expands. The trajectory moves leftward in (ρ, T) space as distributed producers address increasingly complementary tasks (lower ρ).

Phase V: Maturity. Both modes coexist, with centralized production dominant for high-ρ tasks (scale-economy, substitutable) and distributed production dominant for low-ρ tasks (complementary, diversity-premium). Learning exhaustion ($\alpha_{\text{eff}} \to 0$) stabilizes the cost curve. The bifurcation condition for the next technology begins to emerge.

15.4 Crisis Sequence

The three roles of curvature fail in a fixed order as information friction rises during the frenzy phase:

(a) **Correlation robustness fails first**: degrades as $(1-\theta)^2$ (quadratic). Portfolio diversification breaks down.

(b) **Superadditivity fails second**: degrades as $(1-\theta)$ (linear). Production complemen-

tarities degrade.

(c) **Strategic independence fails last**: persists until $K_{\text{eff}} = 0$.

The ordering $(1-\theta)^2 < (1-\theta)$ for $\theta \in (0,1)$ produces the universal sequence: *financial crisis → production disruption → governance failure*. This matches railroads (1873 panic → rate wars → Interstate Commerce Act), electrification (1929 crash → industrial collapse → SEC), and the internet (dot-com crash → WorldCom → Sarbanes-Oxley).

15.5 Five Transitions

Table 20: Technology cycle parameters and predictions.

Technology	α	ρ	K	N	Duration (years)	Predicted τ	Actual τ
Railroads	0.18	-2.0	0.71	8	1840–1890	52	50
Electrification	0.22	-1.0	0.50	12	1890–1930	42	40
Telephony	0.20	-0.5	0.38	5	1880–1920	44	40
Internet	0.30	0.0	0.25	20	1990–2010	18	20
AI (projected)	0.35	varies	varies	8	2020–?	12–15	—

Railroads (1840–1890). The canonical technology cycle. $\alpha \approx 0.18$ from the decline in construction cost per mile as cumulative mileage increased (Fogel 1964). Strong complementarity ($\rho \approx -2$): track, rolling stock, stations, signaling, and trained personnel must coordinate—failure of any one component renders the others worthless.

Installation (1840–1860): Railroad construction required concentrated capital on a scale unprecedented for private enterprise, driving the creation of modern capital markets (Baskin 1988). $N \approx 8$ major systems competed for trunk routes. The stock market emerged largely to finance railroad construction.

Frenzy (1860–1873): Track mileage doubled while freight rates fell 50%. Financial innovation (mortgage bonds, preferred stock) expanded the investor base, raising T_{fin}. Speculative railroad financing became the dominant activity of New York capital markets.

Turning point (1873): The failure of Jay Cooke and Company—the leading railroad financier—triggered a banking panic. Crisis sequence matches the theorem precisely: financial crisis (bank failures and credit contraction) preceded productive disruption (rate wars among railroads, 1877–1886) and governance restructuring (Interstate Commerce Act, 1887).

Deployment (1880–1910): With infrastructure in place and construction costs dramatically reduced, small shippers and manufacturers gained access to national markets. The

distributed adoption phase generated the Second Industrial Revolution—the value created during deployment vastly exceeded the installation-phase stock returns.

Predicted $\tau \approx 52$ years; actual 50.

Electrification (1890–1930). $\alpha \approx 0.22$ from electricity generation cost per kWh (Fouquet 2014). Moderate complementarity ($\rho \approx -1$): generation, transmission, distribution, and end-use equipment must coordinate, but are more modular than railroads. *Installation:* Required regulated utility monopolies with guaranteed returns to justify generation and transmission investment. $N \approx 12$ regional utility systems. *Frenzy:* Utility holding companies (Insull empire, Associated Gas and Electric) used leverage ratios exceeding 10:1. T_{fin} rose as leverage ratios exceeded prudent levels. *Turning point:* 1929 stock crash (financial) → industrial production collapse (productive) → SEC and PUHCA (governance). *Deployment:* Rural electrification (REA, 1935 onward) brought electricity to distributed users at dramatically lower cost, enabling the appliance revolution and suburban manufacturing. Predicted $\tau \approx 42$ years; actual 40.

Telephony (1880–1920). $\alpha \approx 0.20$ from per-line installation and switching cost (Mueller 1997). Moderate complementarity ($\rho \approx -0.5$): switching equipment, copper lines, operator training, and billing systems must coordinate, but are more modular than heavy infrastructure. *Installation:* AT&T's Bell System required monopoly protection (natural monopoly argument) to finance the long-distance network. $N \approx 5$ (Bell plus independents). *Frenzy:* Duplication of local networks by competing independents in the 1890s–1900s. *Turning point:* Kingsbury Commitment (1913) and subsequent regulation replaced market crisis with regulatory restructuring—the one historical case where the crisis was preempted by institutional intervention, reducing the amplitude but not eliminating the regime shift. *Deployment:* Universal service policy extended telephone access from businesses to households. Predicted $\tau \approx 44$ years; actual 40.

Internet (1990–2010). $\alpha \approx 0.30$ from bandwidth cost per Mbps and storage cost per GB (Nagy et al. 2013). Approximate Cobb-Douglas ($\rho \approx 0$): heterogeneous components (fiber, routers, servers, software, content) combine with significant modularity through layered protocols. $N \approx 20$ in the broader ecosystem. *Installation (1993–1998):* Building the internet backbone required concentrated investment by a small number of telecom carriers and ISPs. *Frenzy (1998–2000):* Venture capital and IPO markets financed speculative expansion. "Eyeball" valuations replaced revenue multiples. T_{fin} rose dramatically. *Turning point:* NASDAQ crash (March 2000, financial) → WorldCom/Global Crossing collapse (2002, productive) → Sarbanes-Oxley (2002, governance). *Deployment:* cloud computing, mobile broadband, SaaS enabled distributed adoption at near-zero marginal cost. The infrastructure built during the bubble (fiber, data centers) supported decades of growth. Predicted

$\tau \approx 18$ years; actual 20.

AI (2020–projected). $\alpha \approx 0.35$ from the effective learning rate combining hardware ($\alpha_{\text{hw}} = 0.23$) and algorithmic ($\alpha_{\text{algo}} \approx 0.15$) improvements. The complementarity parameter ρ *varies by task*: training is near-Leontief ($\rho \to -\infty$) while inference approaches linear substitution ($\rho \to 1$). This is the training-inference bifurcation (Section 6). $N \approx 8$ hyperscale firms (Microsoft, Google, Amazon, Meta, Oracle, xAI, Anthropic, and Stargate JV). Overinvestment ratio ≈ 3–$4\times$ from the differential game (observed ratio $\sim 11\times$, consistent with option-value amplification).

Installation (2017–2024): Construction of hyperscale GPU clusters. Cumulative capex: \$1.3T (2018–2025). The investment financed HBM production scaling, CoWoS packaging expansion, and model compression research. The installation phase created the shared component base (HBM, advanced packaging, transformer architectures) that the distributed paradigm will exploit.

Frenzy (2024–2026): The current period. Capex commitments accelerate beyond what current revenue can justify. Stargate (\$100B commitment), Meta AI infrastructure (\$65B 2025 guidance), and competitive announcements indicate the frenzy phase. The 2025–26 DRAM supercycle is the frenzy's supply-side manifestation.

Turning point (predicted 2027–2029): The model predicts a financial correction in AI-linked equities, following the crisis sequence (Section 15.3): correlation robustness fails first (AI stocks become increasingly correlated), then production disruption (overcapacity in inference infrastructure), then governance restructuring (regulatory response). The fold bifurcation (equation 58) predicts the crisis is structurally stable.

Deployment (predicted 2029–2035): Post-crisis, the infrastructure built during installation and frenzy enables distributed deployment at dramatically lower cost. Consumer 3D stacked memory, mature model compression, and standardized edge runtimes enable the mesh formation described in Sections 9–11.

The cycle should be the shortest yet: $\tau \approx 12$–15 years (2020–2032/35). This follows from $\alpha = 0.35$ (the highest effective learning rate of any major technology) and $N = 8$ (high competitive intensity). The duration formula (Proposition 13) yields $\tau = 12.8$ years at baseline parameters.

Remark 11 (Preliminary Empirical Evidence)**.** *Several predictions have been confronted with data. Key findings: (i) the overinvestment ratio averages* $11.12\times$ *for 2022–2025, exceeding the 3–4× Nash prediction—the arms race has intensified beyond standard Nash dynamics, consistent with option-value amplification (Remark 5); (ii) crossing-time acceleration is 79.3% at* $N = 5$*, matching the theoretical prediction; (iii) the duration formula achieves* $MAE = 2.5$ *years across 4 historical cycles* ($R^2 = 0.99$)*; (iv) successive cycles compress as*

α rises (Kendall $\tau(\alpha, \tau_{cycle}) = -0.91$, $p = 0.07$); (v) 3/3 testable historical cycles follow the predicted financial → production → governance crisis sequence; (vi) consumer silicon trajectory is on track for ∼2028 inference-cost crossing.

15.6 Historical Validation: Mainframe → PC

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). The self-undermining mechanism operated through a specific channel: IBM's own semiconductor division produced components (memory chips, logic circuits) whose cost trajectory enabled the PC architecture that displaced IBM's mainframe revenue.

Phase mapping. *Installation (1960–1975):* IBM invested billions in System/360 development (estimated $5B in 1964 dollars, the largest private industrial investment in history at that date). The investment created the shared component base—integrated circuits, DRAM, magnetic storage—that would enable microcomputers. *Frenzy (1975–1987):* The PC revolution produced massive entry. IBM's own PC division (1981) accelerated component learning curves while cannibalizing mainframe revenue. By 1986, the PC market exceeded the mainframe market in unit volume. *Turning point (1991–1993):* IBM recorded cumulative losses of $15.8B, the largest corporate loss in American history at that time. Revenue declined from $69B (1990) to $62.7B (1993). *Deployment (1993–2000):* Under Gerstner's restructuring, IBM pivoted to services. Mainframe revenue stabilized at $3–4B annually—a niche, not extinction.

The $\delta \approx 0.30$ calibration: IBM lost approximately 60% of its compute-service profit in three years (1990–1993). This displacement rate enters the continuation value $S = S_T + S_I/(N(r + \delta))$ and governs the pace at which inference revenue erodes post-crossing.

Parallel to the AI transition. The structural parallel is precise. IBM's semiconductor investment financed the component learning curves (microprocessors, DRAM) that enabled a different organizational form (PC ecosystem) to displace the centralized form (mainframe timesharing). IBM retained training-equivalent revenue (enterprise services, middleware) while losing inference-equivalent revenue (compute-as-a-service). The mainframe did not disappear; it occupies a niche defined by high-reliability transaction processing—exactly the centralized residual predicted by the training-inference bifurcation.

16. Frameworks Considered and Rejected

Several candidate frameworks were evaluated for the formal model and rejected for specific technical reasons. Documenting these decisions clarifies the modeling choices and distinguishes this paper's approach from adjacent literatures. The common theme is that each rejected framework captures one aspect of the mechanism but fails to accommodate the specific structural features—heterogeneity, openness, non-conservation—that the CES + RAF + Potts combination handles.

Mean Field Games (Lasry-Lions 2007). MFG assumes a continuum of exchangeable (identical) agents whose individual optimization depends on the population distribution. The mesh's agents are heterogeneous specialists—heterogeneity is the source of the CES diversity premium that drives Theorem 3. Replacing heterogeneous agents with a continuum of identical agents eliminates the mechanism. The supermodular game framework (Topkis 1998; Milgrom and Roberts 1990) handles heterogeneity naturally through lattice-theoretic monotone comparative statics.

Spin Glasses (Edwards-Anderson 1975; Sherrington-Kirkpatrick 1975). Spin glass models require frustrated interactions—a mix of positive and negative couplings. In the mesh, all interactions are positive: each agent benefits from others joining the network (network effect) and from others specializing in complementary tasks (CES complementarity). There is no frustration. The appropriate statistical mechanics model is the Potts model (positive couplings, heterogeneous external fields), not a spin glass.

Eigen's Hypercycle (Eigen and Schuster 1977). The hypercycle describes a cyclic network of autocatalytic replicators. The autocatalytic structure is conceptually correct for the mesh. However, the hypercycle imposes a conservation law: $\sum_i x_i = \text{const}$ (total concentration is fixed). This forces zero-sum dynamics among capability types. The mesh has no such conservation law—it is an open system where total capability can grow. The RAF framework (Hordijk and Steel 2004) provides the autocatalytic structure without the conservation constraint.

NK Fitness Landscapes (Kauffman 1993). The NK model of rugged fitness landscapes with epistatic interactions is conceptually appropriate for co-evolutionary dynamics. However, the NK framework lacks analytical results on convergence or divergence: whether the co-evolutionary dynamics converge, cycle, or exhibit chaotic behavior is an open problem in complexity science. Useful as motivation for variety expansion, but not as formal machinery for growth regime characterization.

Ecological Niche Models (Tilman 1982; Loreau-Hector 2001). The conceptual analogy is precise: diverse specialist communities outperform monocultures, exactly as in the

mesh. The formal ecological models, however, are calibrated to plant biomass dynamics with resource-competition mechanics (light, nutrients, water) that do not transfer to inference economics. CES aggregation captures the identical qualitative result—diversity premium from imperfect substitutability—while being native to the economics literature.

Chemical Reaction Network Theory (Feinberg 2019). CRNT provides powerful results on equilibrium existence via the deficiency zero theorem. However, CRNT assumes closed systems with stoichiometric conservation laws. The mesh is open (it receives exogenous base models and produces capability improvements that are not conserved stoichiometrically). Moreover, CRNT characterizes equilibrium existence, not growth trajectories—and the central question of Section 12 is the growth trajectory.

Remark 12 (Mean-Field Exactness)**.** *For systems on networks with spectral dimension $d_s > 4$, mean-field theory is exact—not an approximation, but a rigorous result (Dorogovtsev et al. 2008). Real-world networks, including the internet and social networks, generically have $d_s > 4$ due to the small-world property. This means the mean-field percolation result (Proposition 3), the mean-field Potts regime shift (Proposition 4), and the Katz-Shapiro network goods model that underlies the supermodular game are exact characterizations for the mesh, not approximations. The mean-field framework requires no apology.*

17. Predictions

The model generates predictions spanning the complete arc. If these fail, the theory is wrong.

17.1 Pre-Crossing Predictions

Prediction 1: Consumer Stacked Memory ≥16GB by 2027. HBM-derived 3D stacking in consumer products below $200. Evidence against: ≤8GB through 2028.

Prediction 2: 70B-Class Inference On-Device by 2028–2029 (Hardware Crossing). Consumer devices under $1,500 at 70B-class output quality at ≥20 tok/s. Evidence against: not achieved by 2031.

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

Prediction 4: Packaging Learning Rate Stability. The 3D stacking learning elasticity remains in $[0.18, 0.28]$ through 2030. Evidence against: rolling 3-year α below 0.15 not reverting within two years of supercycle resolution.

Prediction 5: Open-Weight Models Exceed 50% of Inference Token Volume by 2028. Evidence against: proprietary models maintaining >60% through 2029.

17.2 Crossing and Mesh Formation Predictions

Prediction 6: First-Order Regime Shift, Not Gradual Adoption. Distributed inference share transitions from <5% to >25% within 18 months. Timing: 2030–2033. Evidence against: smooth growth at <5 percentage points per year through 2035.

Prediction 7: Specialization Precedes Generalization. Early mesh agents are narrow specialists (legal reasoning, medical coding, specific-language code review). Evidence against: early participants predominantly running general-purpose base models.

Prediction 8: Long-Tail Niche Dominance First. Distributed inference exceeds 50% for long-tail categories while below 20% for mainstream. Evidence against: mesh competing first on mainstream query types.

17.3 Post-Crossing Growth Predictions

Prediction 9: Autocatalytic Threshold Timing. Mesh achieves self-sustaining capability improvement within 3 years of crystallization. Observable as: mesh benchmark scores improving $\geq 5\%$ during ≥ 6 months without major frontier model releases. Timing: 2033–2036. Evidence against: mesh capability plateauing during 12 months without new releases through 2038.

Prediction 10: Baumol Bottleneck Binding. Mesh capability growth rate tracks within 1.5× of frontier model improvement rate. Timing: 2034–2040. Evidence against: mesh growth exceeding 3× frontier rate sustained over >2 years.

Prediction 11: Diversity-Collapse Protection. Heterogeneous meshes ($J \geq 10$, $\rho \leq 0.5$) maintain capability when training on >50% synthetic data, while homogeneous networks ($J \leq 3$) exhibit model collapse. Evidence against: homogeneous networks showing no degradation, or heterogeneous meshes degrading despite high diversity.

17.4 Structural Predictions

Prediction 12: Training Remains Centralized Through 2035. Frontier training (>10,000 synchronized GPUs, >7 days) remains exclusively in centralized clusters. Evidence against: distributed frontier training at comparable cost by 2035. This prediction follows directly from the training-inference bifurcation (Proposition 2): the near-Leontief complementarity of training ($\rho \ll 0$) combined with the topological synchronization requirement places training at the extreme end of the decentralization gradient.

Prediction 13: AI Cycle Duration 12–15 Years. Full cycle (installation through deployment equilibrium) completes by 2032–2035, the shortest major technology cycle in

history. This follows from the duration formula (Proposition 13) with $\alpha = 0.35$—the highest learning rate of any major technology. Evidence against: cycle extending beyond 20 years.

Prediction 14: Financial Crisis Precedes Production Adjustment. The correlation robustness of AI-linked equity portfolios degrades before the productive value of AI complementarities. This follows from the crisis sequence (Section 15.3): correlation robustness degrades as $(1-\theta)^2$ while superadditivity degrades as $(1-\theta)$, ensuring the financial system fails first. Specifically, the diversification benefits of holding multiple AI firms' stock will diminish as their return correlations increase toward 1, preceding any reduction in the productive value of AI applications. Evidence against: production disruption preceding financial stress.

Prediction 15: Settlement Layer as Binding Constraint. The settlement layer (routing compensation and micro-transaction infrastructure) becomes the binding constraint on mesh growth before device capability, network connectivity, or model quality bind. Observable as: mesh growth stalling despite available device capacity, with growth resuming when settlement infrastructure improves. Timing: 2031–2034. Evidence against: mesh growth constrained by device capability or bandwidth through 2035.

Prediction 16: Endogenous Hub Emergence. The mesh's degree distribution becomes fat-tailed ($\gamma \leq 3$) within 3 years of the regime shift, with <1% of nodes handling >30% of routing traffic. These hub agents emerge from the Bianconi-Barabási preferential attachment dynamics, not from central design. Evidence against: the degree distribution remaining thin-tailed (exponential or Gaussian) through 2036.

17.5 Summary of Timing

Table 21: Prediction timeline.

Event	Prediction	Timing	Key Parameter
Consumer stacked memory	≥16GB	2027	Packaging $\alpha = 0.23$
Hardware crossing	70B on-device	2028–2029	Cost threshold \$1,500
Open-weight >50% tokens	Token share	2028	R_0 centralized
$R_0 > 1$ distributed	Self-sustaining	2030–2032	κ decline
Regime shift	5% → 25% in 18mo	2030–2033	Potts $q > 2$
Autocatalytic threshold	Self-improving	2033–2036	N_{auto}
Baumol bottleneck binds	Growth $\leq 1.5 \times g_Z$	2034–2040	$\varphi_{\text{eff}} < 1$
Cycle completion	Deployment equil.	2032–2035	$\alpha = 0.35$

Conclusion

This paper has traced the complete arc of endogenous decentralization as applied to the AI transition: from concentrated investment through learning curves, through cost crossing and self-sustaining adoption, through mesh formation and autocatalytic capability growth, to the Baumol ceiling where the circle closes.

The central mechanism is self-undermining: concentrated investment finances the learning curve that reduces the information friction required for distributed alternatives. This is not a claim that centralized structures are inefficient—they are *necessary* during the installation phase. The claim is that their success undermines the conditions for their dominance.

Six results deserve emphasis. First, the self-undermining theorem (Theorem 1) establishes the pattern as a mathematical necessity. Second, the overinvestment result (Theorem 2) shows that Nash competition accelerates crossing by 79%. Third, the training-inference bifurcation (Proposition 2) predicts partial, not total, decentralization. Fourth, the mesh equilibrium theorem (Theorem 3) proves that heterogeneous specialists exceed centralized provision above finite N^*. Fifth, the CES parameter ρ does triple duty: governing the diversity premium, collapse protection, and the gradient of decentralization. Sixth, the Baumol bottleneck (Proposition 12) emerges endogenously rather than being assumed, closing the circle.

The export-control natural experiment distinguishes this mechanism from standard learning-by-doing: constrained firms match or exceed unconstrained firms on capability per FLOP, disproportionately release edge-targeted models, and gain ecosystem share—all inconsistent with Arrow and consistent with constraint-induced optimization for the distributed paradigm.

Historical calibration against five technology transitions spanning 200 years confirms the quantitative predictions: the duration formula achieves MAE = 2.5 years, the crisis sequence matches all examined transitions, and successive cycles compress as learning rates increase. The AI transition, with $\alpha \approx 0.35$, should produce the shortest cycle yet.

The training-inference bifurcation sharpens the mechanism's scope. The post-crossing equilibrium is partial decentralization: inference distributes while training persists centrally. This coexistence is stable because training's near-Leontief complementarity ($\rho \ll 0$) is destroyed by the information friction of distributed coordination ($T_{\text{dist}} \gg T^*$), while inference's near-perfect substitutability ($\rho \approx 1$) makes it indifferent to coordination quality. 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.

The organizational form that emerges is not designed; it is the equilibrium. The mesh

is not merely a static division of labor but a dynamical system capable of self-improvement, whose growth rate is bounded by the concentrated investment that created it. This is the AI transition's ultimate irony: the infrastructure arms race between centralized firms is the engine that powers their own partial displacement.

What the mechanism predicts unambiguously is that concentrated investment endogenously produces inference decentralization, that this process accelerates with the number of competitors, and that training centralization and inference decentralization will coexist as stable features of the AI economic landscape. The framework further predicts a *gradient of decentralization* for intermediate tasks—federated fine-tuning, multi-agent evaluation, distributed RLHF—ordered by each task's CES complementarity parameter, with inference decentralizing first and training last.

Policy Implications

The analysis suggests three policy implications. First, investment in frontier training capability (compute infrastructure, researcher talent, data access) has outsized returns because it relaxes the Baumol constraint for the entire distributed ecosystem. Policies that reduce frontier training investment—export controls, regulatory barriers, taxation of AI compute—reduce not only centralized capability but also the ceiling for distributed capability. The Baumol bottleneck means that investments in the concentrated sector propagate to the distributed sector.

Second, the coordination layer lag ($\Delta T \approx$ 2–3 years) is amenable to policy intervention. Reducing κ (deployment complexity) through standardization, interoperability mandates, and open protocol development compresses the lag between hardware crossing and self-sustaining adoption. This is the lowest-cost lever for accelerating the benefits of decentralization.

Third, the crisis sequence prediction (financial $\rightarrow$ productive $\rightarrow$ governance) suggests that prudential regulation of AI-linked financial instruments should receive priority. Correlation robustness fails first; policies that reduce the financial system's exposure to concentrated AI equity positions reduce the amplitude of the crisis without slowing the underlying technology transition.

Limitations

The analysis has five principal limitations. First, the differential game assumes symmetric firms. In practice, the AI landscape exhibits significant asymmetry (NVIDIA's GPU dominance, OpenAI's first-mover advantage in chatbots). Corollary 2 addresses cost asymmetry;

a full treatment of heterogeneous firms with different continuation values, learning rates, and capability trajectories is left to future work.

Second, the model treats the learning rate $\alpha = 0.23$ as constant. In practice, learning rates exhibit regime transitions (the structural breaks in Table 5). A piecewise learning curve with endogenous regime transitions would improve the quantitative predictions.

Third, the mesh formation analysis uses mean-field approximations. While these are exact for networks with spectral dimension $d_s > 4$ (Appendix remark on mean-field exactness), the actual mesh network during the nucleation phase may have lower effective dimension, making mean-field predictions unreliable precisely when precision matters most.

Fourth, the settlement layer requirement (Proposition 1) is stated as a functional specification without analyzing which existing or prospective monetary systems satisfy it. This is the subject of the companion paper (Paper 5; Smirl 2026, forthcoming) and represents a significant gap in the present analysis: the mesh's viability depends on a settlement infrastructure whose existence is not proven.

Fifth, the Perez phase calibration uses ex-post parameter estimates. The duration formula achieves MAE = 2.5 years across historical transitions, but this is in-sample validation. Out-of-sample predictive power requires the AI transition to unfold as predicted. Moreover, the crisis sequence prediction (financial $\rightarrow$ productive $\rightarrow$ governance) is based on a curvature-degradation ordering that has been tested against only three fully resolved transitions—a sample too small for statistical confidence, despite the perfect match.

Despite these limitations, the framework generates 16 falsifiable predictions with specific timing and failure conditions spanning 2027–2040. If the AI transition follows the predicted pattern—overinvestment, hardware crossing by 2028–2029, $R_0 > 1$ by 2030–2032, first-order regime shift, Baumol bottleneck binding by the mid-2030s—it would provide real-time confirmation of a theory calibrated against 200 years of prior transitions.

Future Work

Several extensions are immediate. First, the differential game assumes symmetric firms; a heterogeneous-firm extension with firm-specific learning rates α_i and continuation values S_i would capture the asymmetric competitive landscape (NVIDIA's GPU dominance, OpenAI's first-mover advantage, Meta's open-weight strategy). Second, a formal difference-in-differences panel at the firm-model-quarter level, with pre-treatment parallel trends and standardized efficiency metrics (benchmark-per-FLOP, benchmark-per-memory-bandwidth), would strengthen the export-control natural experiment (Section 7). Third, the mesh formation model assumes homogeneous connection probability; a heterogeneous-degree model with realistic Internet topology (power-law degree distributions, community structure, geographic

clustering) would refine the nucleation dynamics of Section 9. Fourth, the Baumol bottleneck derivation assumes a clean separation between automatable and non-automatable sectors; a continuous-ρ extension where the automation frontier $\rho^*(t)$ advances endogenously would produce richer growth dynamics. Fifth, the settlement layer requirement (Proposition 1) is stated as a functional specification; the companion paper (Paper 5; Smirl 2026, forthcoming) develops the monetary implications in detail. Sixth, the interaction between this paper's framework and macroeconomic monetary policy—the "monetary schism" hypothesis—is deferred to Paper 5.

The most valuable empirical extension would be direct measurement of distributed (edge) inference volumes. No existing data source tracks this at the granularity required. As edge inference platforms mature, inference-volume telemetry should become available, enabling direct tests of the R_0 crossing prediction (Prediction 3), the first-order regime shift prediction (Prediction 6), and the Baumol bottleneck prediction (Prediction 10).

References

[1] Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, 60(2), 323–351.

[2] Aghion, P., Jones, B. F., & Jones, C. I. (2018). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence* (pp. 237–282). University of Chicago Press.

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

[4] Atalay, E. (2017). How important are sectoral shocks? *American Economic Journal: Macroeconomics*, 9(4), 254–280.

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

[6] Bianconi, G., & Barabási, A.-L. (2001). Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86(24), 5632–5635.

[7] Bastolla, U., Lässig, M., Manrubia, S. C., & Valleriani, A. (2009). Biodiversity in model ecosystems. *Journal of Theoretical Biology*, 235(4), 521–530.

[8] Baskin, J. B. (1988). The development of corporate financial markets in Britain and the United States, 1600–1914. *Business History Review*, 62(2), 199–237.

[9] Baumol, W. J. (1967). Macroeconomics of unbalanced growth. *American Economic Review*, 57(3), 415–426.

[10] 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.

[11] Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *Quarterly Journal of Economics*, 107(4), 1137–1160.

[12] Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4), 1104–1144.

[13] Bonabeau, E., Theraulaz, G., & Deneubourg, J.-L. (1998). Fixed response thresholds and the regulation of division of labor in insect societies. *Bulletin of Mathematical Biology*, 60(4), 753–807.

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

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

[16] Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4), 1275–1335.

[17] Eigen, M., & Schuster, P. (1977). The hypercycle: A principle of natural self-organization. *Naturwissenschaften*, 64(11), 541–565.

[18] Feinberg, M. (2019). *Foundations of Chemical Reaction Network Theory*. Springer.

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

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

[21] Fogel, R. W. (1964). *Railroads and American Economic Growth*. Johns Hopkins Press.

[22] Fouquet, R. (2014). Long-run demand for energy services. *Review of Environmental Economics and Policy*, 8(2), 186–207.

[23] Fortuin, C. M., & Kasteleyn, P. W. (1972). On the random-cluster model. *Physica*, 57(4), 536–564.

[24] Gechert, S., Havranek, T., Irsova, Z., & Kolcunova, D. (2022). Measuring capital-labor substitution: The importance of method choices and publication bias. *Review of Economic Dynamics*, 45, 55–82.

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

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

[27] Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, 35(4), 519–530.

[28] Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[29] Hordijk, W., & Steel, M. (2004). Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4), 451–461.

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

[31] Jain, S., & Krishna, S. (1998). Autocatalytic sets and the growth of complexity. *Physical Review Letters*, 81(25), 5684–5687.

[32] Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

[33] Lasry, J.-M., & Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1), 229–260.

[34] Jones, C. I. (1995). R&D-based models of economic growth. *Journal of Political Economy*, 103(4), 759–784.

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

[36] Milgrom, P., & Roberts, J. (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58(6), 1255–1277.

[37] Mueller, M. L. (1997). *Universal Service: Competition, Interconnection, and Monopoly in the Making of the American Telephone System*. MIT Press.

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

[39] Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices. *American Economic Review*, 105(1), 272–298.

[40] Nagy, B., Farmer, J. D., Bui, Q. M., & Trancik, J. E. (2013). Statistical basis for predicting technological progress. *PLoS ONE*, 8(2), e52669.

[41] Nordhaus, W. D. (2021). Are we approaching an economic singularity? *American Economic Journal: Macroeconomics*, 13(1), 299–332.

[42] Oberfield, E., & Raval, D. (2021). Micro data and macro technology. *Econometrica*, 89(2), 703–732.

[43] Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14), 3200–3203.

[44] Perez, C. (2002). *Technological Revolutions and Financial Capital*. Edward Elgar.

[45] Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71–S102.

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

[47] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

[48] Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3), 665–690.

[49] Smirl, J. (2026a). Emergent CES: The unified mathematical framework. Working Paper.

[50] Tilman, D. (1982). *Resource Competition and Community Structure*. Princeton University Press.

[51] Topkis, D. M. (1998). *Supermodularity and Complementarity*. Princeton University Press.

[52] Tarski, A. (1955). A lattice-theoretical fixpoint theorem. *Pacific Journal of Mathematics*, 5(2), 285–309.

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

[54] Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, 113(2), 331–360.

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

A. Two-Period Pedagogical Model

This appendix presents a simplified two-period version of the differential game that captures the overinvestment result and the self-undermining property in a framework accessible without continuous-time machinery.

A.1 Setup

Two periods, $t \in \{1, 2\}$. N symmetric firms. In period 1, each firm chooses output q_i in a Cournot market with inverse demand $P = a - b\sum_j q_j$. Cumulative output $Q_1 = \sum_j q_j$ drives the learning curve: period-2 cost is $c_2 = c_1 \cdot (1 - \alpha \cdot Q_1/\bar{Q})$ for $Q_1 \leq \bar{Q}$, where $\bar{Q}$ is the crossing threshold.

If $Q_1 \geq \bar{Q}$, crossing occurs: distributed inference becomes viable and the centralized firm's period-2 inference profit drops by a fraction δ. Each firm receives:

$$\Pi_i = \pi_i^1 + \beta \cdot \pi_i^2(Q_1) \tag{59}$$

where β is the discount factor and π_i^2 depends on whether crossing occurred.

A.2 Nash Equilibrium

Each firm maximizes Π_i taking rivals' first-period output as given. The first-order condition is:

$$\frac{\partial \pi_i^1}{\partial q_i} + \beta \cdot \frac{\partial \pi_i^2}{\partial Q_1} = 0 \tag{60}$$

The second term is the learning externality: firm i's output reduces *all* firms' period-2 costs, but firm i captures only $1/N$ of the benefit. In the symmetric Nash equilibrium:

$$q_i^N = \frac{a - c_1 + \beta\alpha c_1/N}{b(N+1)} \tag{61}$$

A.3 Cooperative Solution

A planner maximizing total surplus sets:

$$q_i^C = \frac{a - c_1 + \beta\alpha c_1}{b(N+1)} \tag{62}$$

The ratio $Q^N/Q^C = (a - c_1 + \beta\alpha c_1/N)/(a - c_1 + \beta\alpha c_1)$. For $\beta\alpha c_1$ small relative to $a - c_1$:

$$\frac{Q^N}{Q^C} \approx 1 + \frac{(N-1)\beta\alpha c_1}{N(a - c_1)} \tag{63}$$

confirming overinvestment: $Q^N > Q^C$ for all $N \geq 2$.

A.4 Self-Undermining

If $Q^N > \bar{Q}$ but $Q^C < \bar{Q}$, then Nash competition triggers crossing while cooperation does not. The firms' aggregate behavior creates the conditions for distributed entry—each firm individually prefers less total output (to delay crossing), but cannot commit to restraint. This is the commons structure of Section 3 in its simplest form.

The period-2 profit loss from crossing is $\delta \cdot \pi^2_{\text{monopoly}}/N$ per firm. The period-1 benefit from the extra output is $(Q^N - Q^C) \cdot (P - c_1)/N$ per firm. In the Nash equilibrium, the marginal firm equates these, producing exactly the crossing it would prefer to avoid.

A.5 Welfare Analysis

Despite the crossing being privately costly for incumbent firms, total welfare (producer surplus plus consumer surplus) may increase. Consumer surplus from lower inference costs is:

$$\Delta CS = \int_0^{Q^N} (P(Q) - P(Q^N))\, dQ - \int_0^{Q^C} (P(Q) - P(Q^C))\, dQ = \frac{b}{2}\left[(Q^N)^2 - (Q^C)^2\right] \quad (64)$$

The producer surplus loss is $N \cdot (\Pi_i^C - \Pi_i^N)$. The net welfare effect depends on the relative magnitudes:

$$\Delta W = \Delta CS - N \cdot \Delta\Pi = \frac{b}{2}\left[(Q^N)^2 - (Q^C)^2\right] - N(\Pi_i^C - \Pi_i^N) \quad (65)$$

For standard Cournot parameters, $\Delta W > 0$: the consumer surplus gain from overinvestment exceeds the producer surplus loss. The learning curve reinforces this effect: the "excess" production that drives cumulative output past $\bar{Q}$ creates permanent consumer benefits (lower inference costs) while the producer losses are bounded (firms retain training revenue S_T). The self-undermining mechanism thus has a positive welfare sign: Nash overinvestment accelerates a welfare-improving transition.

This welfare result qualifies the common interpretation of overinvestment as pure waste. The excess investment relative to the cooperative optimum transfers surplus from producers to consumers through the learning curve. In the AI context, this means that the "arms race" among hyperscalers—which analysts characterize as wasteful—is in fact accelerating a consumer-beneficial technology transition at the expense of incumbent margins.

B. Weitzman Recombinant Growth Connection

Weitzman (1998) models the growth of ideas as a combinatorial process: new ideas are produced by recombining existing ideas, and the number of potential recombinations grows faster than the number of existing ideas. This produces growth rates that accelerate over time.

The mesh's variety expansion mechanism (Section 12, Proposition 11) has a Weitzman interpretation. New specialization types are produced by combining existing specializations: a medical-legal specialist combines medical reasoning and legal analysis capabilities. The number of potential combinations grows as $\binom{J}{2} \sim J^2/2$, so the potential for variety expansion accelerates with existing variety.

The growth rate of variety under recombinant dynamics is:

$$\dot{J}_{\text{recomb}} = \eta_J \cdot \binom{J}{2} \cdot p_{\text{viable}} \cdot \mathbf{1}[\alpha > \alpha_{\text{crit}}] \tag{66}$$

where p_{viable} is the probability that a random combination produces a viable new specialization. Even with small p_{viable}, the J^2 scaling ensures that variety expansion accelerates.

However, two forces limit recombinant growth in the mesh. First, not all combinations produce viable specializations: as J grows, the fraction of viable combinations may decline (the curse of dimensionality). Second, the Baumol bottleneck constrains the rate at which new specialization types can be trained—each requires fine-tuning on base models from the food set, which grows at exogenous rate g_Z.

C. Nordhaus Singularity Analysis

Nordhaus (2021) asks whether we are approaching an economic singularity—a regime in which economic growth becomes superexponential. His analysis identifies conditions under which standard growth models produce accelerating growth, and concludes that current empirical trends do not support the singularity hypothesis.

This paper's Regime (c) in Theorem 4 is precisely Nordhaus's singularity condition applied to the mesh's internal dynamics. The contribution is identifying the three specific parameters (φ_{eff}, h, α) whose conjunction determines whether the singularity obtains for the mesh. The paper's conclusion aligns with Nordhaus: the conditions are restrictive and unlikely to hold simultaneously.

Specifically, Nordhaus identifies two requirements for singularity: (i) the share of capital (broadly defined to include AI) in income must approach unity, and (ii) the elasticity of

substitution between capital and labor must exceed unity. In the mesh framework, condition (i) corresponds to $\beta_{\text{auto}} \to 1$ (full automation of the training process), and condition (ii) corresponds to $\varphi_{\text{eff}} > 1$. The mesh adds a third condition absent from Nordhaus's analysis: $\alpha_{\text{eff}} > \alpha_{\text{crit}}$ (avoiding model collapse). The conjunction of all three is more restrictive than either (i) or (ii) alone.

D. Empirical Calibration Details

D.1 Overinvestment Ratio

The Nash overinvestment ratio from Theorem 2 predicts $Q^N/Q^C \approx 1+(N-1)\alpha\phi/(r+\delta) \approx 3$–$4\times$ for baseline parameters ($N = 5$, $\alpha = 0.23$, $\phi = 0.5$, $r = 0.05$, $\delta = 0.10$). The observed ratio is higher ($\sim 11\times$), which is consistent with option-value amplification: if firms invest to maximize the probability of achieving a discontinuous capability threshold (frontier model leadership), the marginal value of additional investment is governed by the prize V^* rather than by discounted market revenue.

D.2 Duration Formula Fit

Table 22: Duration formula validation.

Transition	Predicted τ	Actual τ	Error	Crisis Sequence
Railroads	52	50	+2	Financial → Productive → Governance ✓
Electrification	42	40	+2	Financial → Productive → Governance ✓
Telephony	44	40	+4	Regulatory preemption (unique)
Internet	18	20	−2	Financial → Productive → Governance ✓
MAE			2.5 yr	3/3 testable match

D.3 Consumer Silicon Trajectory

The consumer silicon trajectory toward the crossing threshold ($\geq$70B-class quality, $\geq$20 tok/s, $\leq$\$1,500) is tracked quarterly:

E. Semi-Endogenous R_0 Dynamics

In the main text, the coordination friction κ in the R_0 expression is treated as declining exogenously. This appendix formalizes the semi-endogenous dynamics when κ itself evolves with the state of the ecosystem.

Table 23: Consumer silicon trajectory toward crossing.

Date	Device	DRAM (GB)	tok/s	Price	Status
Q4 2024	Apple M4 Pro	48	∼15	$2,499	Below speed threshold
Q1 2025	AMD Ryzen AI Max+	128	∼31	∼$2,000	Above speed, above price
Q1 2025	Rockchip RK1828	5	59	∼$100	7B only (below quality)
Q1 2025	NVIDIA RTX 5090	32	∼45	$3,000+	Supercycle pricing
Q1 2026	Apple M5 Ultra (proj.)	192	∼50	∼$2,500	Meets quality, above price
2028–29	Post-supercycle	≥64	≥30	≤$1,500	Crossing predicted

Quality threshold: 70B-class output quality via MoE (∼20B active parameters). Speed threshold: ≥20 tok/s. Price threshold: ≤$1,500. The capability threshold has been met at professional price points; the cost threshold is delayed by the 2025–26 DRAM supercycle.

E.1 Coordination Friction as a Function of Ecosystem Maturity

Let $E(t)$ be an ecosystem maturity index—an aggregate of standardization, tooling availability, documentation quality, and deployment automation. The coordination friction declines with maturity:

$$\kappa(t) = \kappa_0 \cdot \left(\frac{E_0}{E(t)} \right)^{\gamma_E} \tag{67}$$

where $\gamma_E > 0$ is the friction elasticity with respect to ecosystem maturity. Ecosystem maturity evolves with cumulative adoption:

$$\dot{E}(t) = \eta_E \cdot s(t) \cdot (E_{\max} - E(t)) \tag{68}$$

where $s(t)$ is the distributed adoption share and $E_{\max}$ is the maximum attainable maturity. This produces a positive feedback loop: adoption improves the ecosystem, which reduces friction, which accelerates adoption.

E.2 The Coupled System

The full dynamics couple the adoption equation (15) with the ecosystem equation:

$$\dot{s} = \frac{\beta(c(Q), \lambda) \cdot \gamma}{\kappa(E) + \mu} \cdot s(1 - s) \cdot (\kappa(E) + \mu) - (\kappa(E) + \mu) \cdot s \tag{69}$$

$$\dot{E} = \eta_E \cdot s \cdot (E_{\max} - E) \tag{70}$$

The system has two equilibria. The *trivial equilibrium* $(s^*, E^*) = (0, E_0)$ is stable when $R_0(E_0) < 1$. The *nontrivial equilibrium* has $s^* > 0$ and $E^* > E_0$ and exists when $R_0(E_0)$ is sufficiently close to unity that the ecosystem feedback pushes the system past the threshold.

Proposition 15 (Ecosystem Bootstrap). *Even if $R_0(E_0) < 1$ (the ecosystem is sub-critical at initial maturity), the nontrivial equilibrium exists if:*

$$R_0(E_{\max}) = \frac{\beta(c(Q), \lambda) \cdot \gamma}{\kappa(E_{\max}) + \mu} > 1 \tag{71}$$

and there exists a "seed" adoption level s_{seed} such that the coupled dynamics (69)–(70) escape the basin of attraction of the trivial equilibrium.

The economic interpretation: even before hardware crossing makes the distributed ecosystem self-sustaining, ecosystem development (tooling, standards, model hubs) can pre-position the coordination infrastructure so that κ is already low when crossing occurs. This is precisely what the 2023–2025 period represents: the HuggingFace ecosystem, GGUF quantization format, llama.cpp runtime, and ONNX export pipelines are reducing κ *before* hardware costs reach the crossing threshold. The coordination lag ΔT is compressed because the ecosystem matures in advance of crossing.

This formalizes the observation in Section 5 that the AI transition's coordination lag is predicted to be 2–3 years rather than the historical 3–5 years: the information channel (Section 2.3) is operating faster than in previous transitions because the coordination infrastructure is digital and can be developed in parallel with the hardware learning curve, rather than sequentially.