How Big Tech Funds Its Own Replacement

Jon Smirl

2026-03-01

The Paradox

In 2024 and 2025, the major hyperscalers — Amazon, Google, Meta, Microsoft — collectively spent over $200 billion per year building centralized AI data centers. Thousands of GPUs per facility. Custom cooling systems. Dedicated power plants. The largest concentrated infrastructure buildout since the transcontinental railroads.

Here is the paradox: every dollar they spend brings forward the day when none of that infrastructure is necessary.

This is not a vague worry about "disruption." The CES framework makes it a theorem. Let T^* denote the date at which distributed AI hardware reaches cost parity with centralized data centers, and let I denote cumulative centralized investment. Then:

Theorem (Self-Undermining).

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

Increasing centralized investment strictly advances the crossing date at which distributed alternatives become cost-competitive.

The derivative is negative. More investment means an earlier crossing. The mechanism is simple and, once you see it, impossible to unsee.

The Four Channels

Why does centralized investment help distributed competitors? Through four distinct channels, each of which is individually sufficient and which compound when operating together.

Channel 1: Cost. The semiconductor learning curve, first documented by (Wright1936), says that unit costs decline as a power law of cumulative production: $C(Q) = C_0 \cdot Q^{-\alpha}$, where $\alpha \approx 0.23$ for planar semiconductor processes. When hyperscalers order millions of AI accelerators, they are pumping cumulative volume Q — and the resulting cost reductions benefit everyone, not just the firms placing the orders. A chip fabricated on the same process node costs the same whether it ends up in a data center or an edge device. Current estimates place the cost gap between centralized GPU inference and emerging crossbar memory architectures at 35–60×, but that gap is closing at a rate governed by α. The more volume the hyperscalers pull forward, the faster it closes.

Channel 2: Infrastructure. The software stack required for AI — PyTorch, ONNX, quantization toolkits, model compression libraries, inference runtimes — was developed overwhelmingly by

hyperscaler engineering teams. And it was released as open source. The rationale was sound at the time: open frameworks attract developers, developers build on your cloud platform, platform lock-in follows. But the infrastructure is now a public good. Anyone building distributed AI hardware can plug into the same software ecosystem without paying a licensing fee or reinventing the stack. The investment is sunk and the benefits are non-excludable.

Channel 3: Information. Key techniques — transformer architectures, attention mechanisms, distillation, mixture-of-experts — were published openly by the same firms that spent billions developing them. Training recipes and scaling laws diffuse within months. This is not altruism; it is the equilibrium of a talent market where researchers demand the right to publish. The consequence is that the knowledge required to build distributed systems is freely available.

Channel 4: Manufacturing. Programs like Intel's on-die memory integration and DARPA's investment in neuromorphic and crossbar architectures are explicitly designed to embed AI capability into every chip that rolls off the fabrication line. These programs are partially funded by, and certainly accelerated by, the demand signal from hyperscaler procurement. The end state is AI inference as a standard feature of commodity silicon — exactly the hardware base that makes centralized data centers redundant for most workloads.

Each channel converts private centralized investment into public distributed capability. The conversion is not incidental. It is structural.

The Prisoner's Dilemma

If concentrated investment undermines the investors, why do they keep investing? Because they are trapped in a classic N-firm prisoner's dilemma (Tirole1988).

Consider a hyperscaler deciding its annual AI capex. If it invests heavily and competitors do not, it captures a dominant market position — the payoff is enormous. If it holds back while competitors invest, it falls behind and loses both customers and talent — the payoff is catastrophic. The Nash equilibrium is for every firm to invest heavily, regardless of what others do.

The CES framework quantifies this. In the N-firm differential game where each firm chooses investment to maximize its discounted profit, the Nash equilibrium aggregate investment is 3–4$\times$ the cooperative optimum:

Theorem (Overinvestment Ratio).

$$\frac{I_{\text{Nash}}}{I_{\text{coop}}} \approx 3\text{–}4$$

Equilibrium aggregate investment in the N-firm AI capex game exceeds the jointly optimal level by a factor of three to four.

This is not waste in the usual sense. Each firm is acting rationally given its competitors' strategies. But the aggregate effect is to pour fuel on the learning curve far faster than any individual firm would choose if it could coordinate with the others. The overinvestment accelerates the cost decline. The cost decline accelerates the crossing. The crossing undermines the business model that justified the investment.

The firms know this. Their strategy is to build moats — proprietary data, customer lock-in, regulatory capture — that survive past the crossing. But the self-undermining theorem does not

depend on whether the moats hold. It depends only on whether cumulative semiconductor volume drives cost reductions, which it has done without interruption since 1965.

The One-Way Door

There is a critical asymmetry in this process: the crossing is irreversible.

Once distributed hardware costs drop below the centralized threshold, the economics permanently favor distribution. You cannot "un-learn" a learning curve. Cumulative production only goes up. The cost reduction is encoded in manufacturing process improvements, yield optimization, and design iteration that cannot be recalled.

This means the crossing date T^* is a one-way door. Before T^*, centralized infrastructure has a cost advantage and the data center model is economically rational. After T^*, the cost advantage flips and the pressure toward distribution is relentless. There is no scenario in which the economy passes through the crossing and then returns to centralization on cost grounds.

The *baumol_limit* provides the ceiling on the other side: once distributed AI capability exists, its growth rate converges to the frontier training rate but cannot exceed it. Distribution does not make AI magically better. It makes AI ubiquitously available at the cost floor set by semiconductor learning curves.

The IBM Parallel

This has happened before. In the 1960s and 1970s, IBM invested massively in mainframe R&D. That investment advanced semiconductor technology — integrated circuits, DRAM, microprocessor architectures — that directly enabled the personal computer. IBM itself launched the IBM PC in 1981, using commodity components that any manufacturer could source. Within a decade, the PC ecosystem had destroyed the mainframe monopoly that financed its creation.

(Perez2002) documented this pattern across five major technological revolutions: the firms that finance the buildout of a new technological paradigm are rarely the firms that dominate the deployment phase. The CES framework explains why this is not coincidence but mathematical necessity. The self-undermining derivative $\partial T^*/\partial I < 0$ operates in every case where (a) investment drives down costs via learning curves, (b) cost reductions are non-excludable, and (c) the new technology enables a structurally different mode of production.

All three conditions hold for AI. Condition (a) follows from the semiconductor learning curve. Condition (b) follows from the physics of fabrication — cost per transistor is set by process node, not by customer. Condition (c) follows from the fact that inference, unlike training, is embarrassingly parallel and does not require centralized data access.

What Comes Next

The self-undermining theorem does not say centralized AI is doomed tomorrow. For semiconductors ($d = 2$, $\alpha \approx 0.23$), the predicted transition duration is approximately 8 years from the crossing point. The crossing has not yet occurred — distributed hardware remains more expensive per unit of inference than data center GPUs. But the gap is closing at a rate proportional to hyperscaler investment, which is at an all-time high.

The prediction is that Big Tech is paying for its own disruption, and the receipt is the semiconductor learning curve — one of the most reliable empirical regularities in economics [@Arrow1962;

@Nordhaus2021].

Every GPU ordered is another step down the cost curve. Every step down the cost curve is a step closer to the crossing. And the crossing, once reached, does not reverse.

References