Where the Chip Really Begins

Why Materials Engineering and Improvements in Manufacturing Equipment Now Set the Timeline for AI Silicon

Summary

AI has rewritten the requirements for semiconductors. The buyers of AI infrastructure now measure performance in metrics like time to first token, time to outcome, and cost per token, and every one of those eventually maps to silicon. The chip sits at the center of that story, but it isn't where the story starts. Behind every production chip are years, and often more than a decade, of materials and equipment work.

That work has traditionally moved through a serial development model, with research in one organization, equipment and process development in another, and integration and pilot production in a third. Every boundary is a handoff, and every handoff costs time. The cost of those handoffs shows up twice: first in how long it takes a new technology to reach production, and again in how long it takes that technology to reach high yield once in production. With new AI compute generations now arriving roughly every year, a development model measured in decades is misaligned with the market it serves.

This research brief examines where that time goes and why the equipment and materials layer has become central to shortening the development timeline. It then looks at how Applied Materials designed its EPIC Center to change the development model itself, bringing together chipmakers, chip designers, equipment partners, and university researchers with Applied's materials scientists and process developers in one facility to work three to four generations ahead of production.

AI Has Changed What the Market Asks of Silicon

Compute has traditionally been bought primarily on peak throughput. However, enterprises and cloud providers deploying AI at scale now evaluate infrastructure based on time to first token, time to outcome, and cost per token at a given service level. These business metrics tie the return on an AI investment directly to how quickly and economically a model can respond.

These metrics are simultaneously user-facing and physics-bound. A latency target is an experience decision at the top of the stack, but meeting that target depends on memory bandwidth, packaging, and power delivery decisions at the bottom. The pressure travels

Copyright © 2026 Moor Insights & Strategy

the whole way down, from model to serving system to node to package to die to process to materials.

That's why inference economics increasingly surfaces as a materials question. The bandwidth, thermal headroom, and power delivery an AI accelerator can offer are now set by physical structures such as stacked high-bandwidth memory, advanced packaging, and backside power delivery. And each of those structures depends on depositing, shaping, bonding, and measuring materials at dimensions counted in atoms. These are integration problems before they're design problems.

As AI has changed what the market demands from silicon, the rhythm of the industry has shifted as well. Leading accelerator vendors have moved from multi-year product cycles to a roughly annual cadence, and memory and packaging roadmaps have been pulled along with them. The result is a structural mismatch. New compute generations arrive on an annual clock, while the process and materials innovations that make them possible are still developed on timetables measured in years, and often decades.

The consequences of that mismatch aren't abstract. When the development model can't keep up with demand, the costs show up as delayed products, constrained supply, capital tied up in long ramps, and gains in performance per watt that reach the datacenter later than they could.

Something has to give. The question is where.

THE CHIP IS STILL THE CENTERPIECE, BUT NOT THE STARTING POINT

Chip designs get the spotlight, but materials and manufacturing set the release date. A chip design is only as real as the process that can build it reliably and at yield in high-volume manufacturing. Two of the industry's defining transitions show how that plays out, and in both cases, the physics was the easier part.

FinFET: An Idea That Waited More Than a Decade for Its Tools

For decades, a transistor was essentially flat, a switch built into the surface of the silicon wafer with a gate that controlled current from the top alone. By the late 1990s, researchers could see where scaling was headed. As the gate shrank, it would increasingly lose its grip on the channel, and current would leak through even when the switch was off, wasting power and generating heat. The answer, which came out of UC Berkeley in 1999, was to stand the transistor up, turning the channel into a thin vertical

Copyright © 2026 Moor Insights & Strategy

fin that the gate could control from both sides. Its developers called it the FinFET. On paper, it was an elegant fix, and the device physics was well understood.

Manufacturing it was another matter. Making the transistor three-dimensional meant that nearly every step used to build it had to work on a tall, narrow structure rather than a flat surface. Films a few atoms thick had to coat the vertical sidewalls of the fin just as evenly as they coated the top. Etching had to cut straight down, stop precisely, and leave the delicate structure undamaged, all at the same time.

Polishing steps that once simply flattened a wafer now helped set critical device dimensions, and implant steps, which fire atoms into the silicon along a straight line to tune its electrical behavior, struggled to reach the fin's sidewalls evenly. All of these processes had to perform consistently at every point on a 300mm wafer and from one wafer to the next, across billions of transistors on a single die. There was no single hero step to solve. The burden landed on essentially the entire transistor-building flow at once, which made FinFET as much a toolmaker's problem as a chipmaker's. Intel began volume production in 2011, more than a decade after the FinFET was named, and the first FinFET-based processors reached the market in 2012.

Copper: When the Whole Wiring Flow Had to Be Rebuilt

Copper interconnect is an even stronger example. The wiring inside chips was aluminum for decades. As circuits grew denser, those wires grew so thin that aluminum's resistance began to slow signals, and the wires themselves wore out under the repeated application of current.

Copper's advantages had been recognized for years, and IBM researchers were pursuing copper wiring by the mid-1980s. The problem was building with it. Aluminum wiring was made by laying down a film and etching away everything that wasn't a wire. Copper doesn't lend itself to the plasma etch processes used for aluminum. It also contaminates the silicon devices beneath it if it escapes into areas where it doesn't belong.

Switching to copper meant reinventing how wiring is made. In the approach that became standard, known as dual damascene, the trenches and connecting holes are carved into the insulating layer first, then lined and filled with copper, and the excess is polished away.

Copyright © 2026 Moor Insights & Strategy

That created four new problems at once:

1. A barrier lining had to be thin enough to leave room for the copper yet continuous enough to seal it inside narrow, deep openings.
2. The copper itself had to fill those openings without voids or seams, a problem the industry solved with electroplating, a wet chemical process that had never been used for on-chip wiring.
3. Polishing had to remove copper and barrier materials that wear away at very different rates, or the resulting dishing and erosion would thin the wires and eat into the resistance advantage that justified copper in the first place.
4. Fabs had to prevent copper contamination of everything else, which required separate handling of copper-bearing wafers.

Solving each problem individually was never going to be enough. At IBM, an early approach that proved copper wiring could be built was judged non-manufacturable on yield, reliability, and cost. Getting to production took a flow in which the steps were developed as one system rather than qualified tool by tool. IBM announced the first copper chip technology in September 1997, more than a decade into its copper research, and shipped its first copper products the following year.

These two real-world examples bring the challenge to life. Across the architectural transitions the industry has actually completed, the path from a viable research result to a commercial chip has run 10 to 15 years. The equipment and the integrated process were the long pole.

And that's the layer where the calendar gets set.

Where the Years Actually Go in Chipmaking Development

The development timeline for a new chipmaking technology runs through five broad stages:

- **Materials research and pathfinding**: Identifying candidate materials, working from first principles, and demonstrating results at small scale
- **Equipment and process development**: Turning a promising result into an industrial process and the hardware that runs it reliably
- **Integration**: Making a new step work alongside every step around it
- **Pilot production and yield learning**: Finding out what breaks at scale, and why
- **High-volume ramp**: Maintaining yield and output at commercial volume

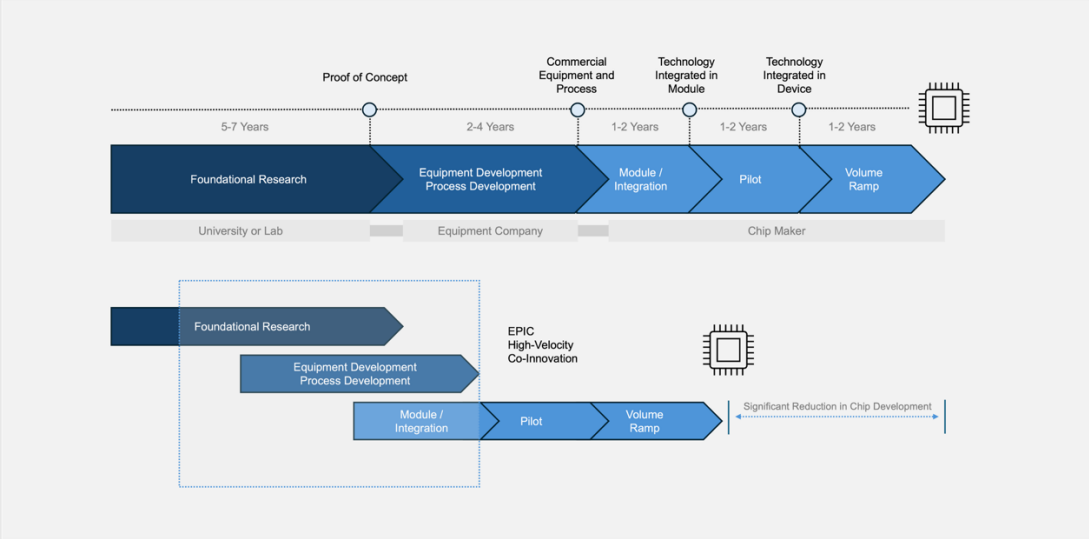
Copyright © 2026 Moor Insights & Strategy

Historically, these stages have been split across different organizations and often across different companies. The work doesn't necessarily flow from one stage to the next so much as it transfers. And every transfer carries a cost that rarely appears on a roadmap chart. Results have to be recharacterized on different tools, processes have to be requalified against a different baseline, and much of what was learned stays with the team that learned it.

A materials result proven in a university lab has to be reproduced on production-class equipment. A process developed on an equipment supplier's tools has to be rebuilt inside a chipmaker's flow, where it meets hundreds of other steps for the first time.

Physics is a real constraint, but it isn't the only one. A significant amount of time between the breakthrough and the chip arises from how the development process is organized. Physics can't be redesigned. A development model can.

FIGURE 1: THE SERIAL DEVELOPMENT MODEL VERSUS PARALLEL DEVELOPMENT



The high velocity co-innovation enabled by EPIC greatly reduces time to first chip
Source: Applied Materials

THE SECOND FINISH LINE: TIME TO YIELD

Chipmakers compete to deliver the best chip and packaging architectures, and that competition understandably gets most of the attention. But arriving first with a better architecture only matters if that architecture can quickly ramp to high yield. From an economics perspective, this isn't about the first wafer coming off the line. It's about how

Copyright © 2026 Moor Insights & Strategy

quickly production reaches the point where nearly every wafer coming off the line is working.

AI raises the stakes on that second finish line. As chipmakers move to an annual cadence of new products, each chip's commercial window shrinks, so months spent climbing the yield curve consume a larger share of the product's life. The packaging structures that AI depends on compound the challenge. A high-bandwidth memory stack or a multi-die accelerator package combines many components, and yield multiplies across all of them. A single defective bond can scrap an assembly full of otherwise good silicon. Supply that looks ample at the wafer level can shrink considerably by the time finished packages leave the line.

Yield learning is also where the handoff problem bites hardest. It's the stage where a process developed elsewhere first meets the full complexity of a production flow, and where unexplored interactions between steps surface as defects. The copper transition previously detailed makes the point: Individual steps worked well before the integrated flow was manufacturable.

Moor Insights & Strategy (MI&S) believes a development model that brings the manufacturer's engineers into the work earlier, on production-class equipment, attacks time to yield at its source rather than after the fact. And this, in turn, helps shorten development timelines.

Why Manufacturing and Its Toolchain Have Never Mattered More

Capacity is the constraint everyone already talks about. Leading-edge logic, memory, and advanced packaging capacity are all tight, and access to that capacity decides which AI products ship and when. That constraint is real, but it isn't the most consequential shift underway.

In MI&S's view, the more important change is that the boundary between architecture, process, and equipment is collapsing. Gate-all-around transistors, backside power delivery, 3-D memory architectures, hybrid bonding, and heterogeneous integration are all arriving within a few years of one another. FinFET and copper each took more than a decade as standalone transitions. The industry is now working through all five of these transitions at once, with materials and interfaces that have to behave as one system.

Copyright © 2026 Moor Insights & Strategy

Chip architectures have outgrown single-process development, and that changes the order of operations. A design can no longer be finished and handed down to be manufactured, because whether an architecture is possible at all now depends on materials and process capabilities that may not yet exist. The people building the tools must be in the conversation while the architecture is still being decided. As a result, the equipment and materials layer moves from supplying tools against a customer's specification to helping determine what that specification can be.

When the hardest problems sit between process steps rather than inside them, the advantage belongs to whoever can work across the whole flow at once.

The EPIC Center: Changing the Model, Not Just the Schedule

Applied Materials' Equipment and Process Innovation and Commercialization (EPIC) Center is designed around the need to work across the whole development flow at once.

What the EPIC Center Does

The EPIC Center is a collaborative R&D facility in Silicon Valley that opened in October 2026. It represents what Applied describes as the largest-ever U.S. investment in advanced semiconductor equipment R&D. Its premise is R&D without handoffs. The stages that have traditionally run in sequence across separate organizations run in parallel instead, inside a single secure environment built to protect each partner's roadmap and intellectual property. This model goes right to the heart of the operational timeline challenges detailed earlier.

How the Model Works

Each EPIC program pairs a partner's engineers with Applied's materials scientists and process developers to work on the partner's own technology challenges in the same building and on the same equipment. Programs are structured one-to-one between each partner and Applied, which allows companies that compete hard in the market to commit to the same facility with confidence. As the partner base grows, the concentration of expertise under one roof creates room for broader collaboration.

Partners gain access to Applied's most advanced chipmaking equipment years before it would typically reach a production fab. This lets them iterate on hardware and process simultaneously rather than sequentially, so a materials problem discovered during

Copyright © 2026 Moor Insights & Strategy

integration can be addressed at the tool level while the program is still underway rather than in the next development cycle.

Programs target technology three to four generations ahead of production, which is the window in which architectural influence is still possible. Work done there shapes what the eventual process can be rather than adapting to a design that has already been locked.

The payoff arrives at transfer. Because the technology has already been hardened by the customer's own engineers on production-class equipment, there is far less to relearn when it reaches the customer's fab. That means less rework at the handoff, fewer surprises during yield learning, and a shorter path to high yield.

It's the point where the development model and the second finish line meet.

The EPIC Roster Is the Evidence of Its Potential

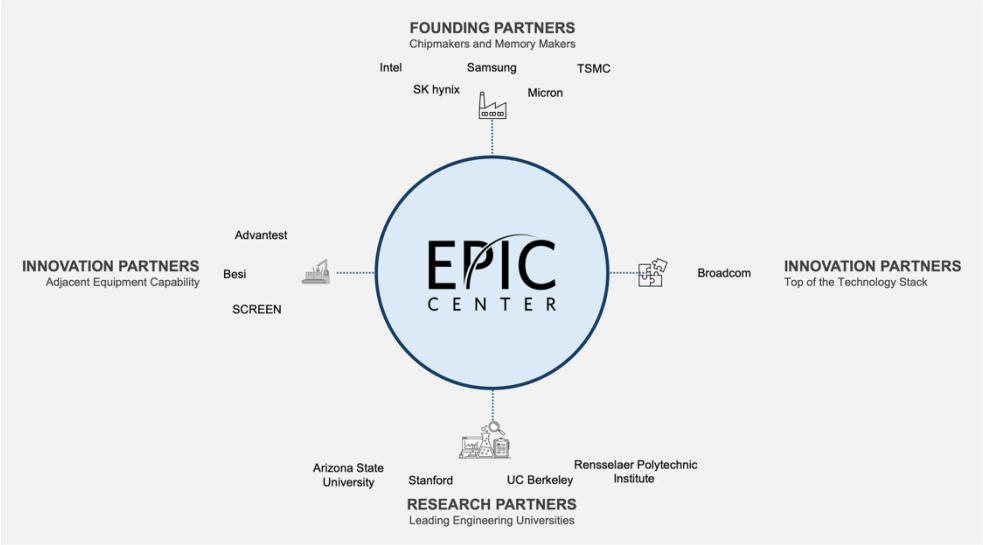
For anyone wondering whether this model has merit, the strongest early signal is who has signed on. Founding partners TSMC, Samsung, Intel, Micron, and SK hynix span leading-edge logic and memory. Innovation partners include Broadcom at the top of the technology stack, alongside Advantest, SCREEN, and Besi in adjacent equipment categories. Research partners include Stanford, UC Berkeley, Rensselaer Polytechnic Institute, and Arizona State University.

Berkeley's presence is a fitting link between the industry's past transitions and its next ones, given that the FinFET was first described there.

These are companies that guard their roadmaps more closely than almost anything else they own, and they don't commit engineers and roadmap visibility to a facility on a hypothesis. Their participation says more about the model's credibility than any claim Applied could make on its own.

Copyright © 2026 Moor Insights & Strategy

FIGURE 2: EPIC CENTER ROSTER



Applied Materials brings leaders from across the ecosystem into EPIC
Source: Moor Insights & Strategy

WHAT APPLIED MATERIALS BRINGS TOGETHER AT EPIC

Building a research facility is one thing. What makes the EPIC model work is the combination of capabilities behind it.

The first is breadth across the process flow. Applied's portfolio spans deposition, etch, planarization, materials modification, metrology and inspection, and advanced packaging. That breadth matters more now than it did a decade ago, precisely because each step increasingly depends on the steps around it.

The second is integrated systems. Applied combines multiple adjacent process steps inside a single controlled environment, so sensitive surfaces and interfaces aren't exposed to air between steps. That capability is what makes co-development of multi-step solutions practical rather than theoretical, and it reflects the same integration-first lesson the copper transition taught.

The third is commercial accountability. A shared research institute is typically focused on achieving a technological breakthrough. EPIC is operated by a company whose business depends on what happens after the breakthrough: whether the technology reaches high-volume manufacturing and ramps to yield in the customer's fab. That aligns the facility's incentives with the chipmaker's.

Copyright © 2026 Moor Insights & Strategy

The fourth is ecosystem density. Silicon Valley offers the world's greatest concentration of chipmakers, chip designers, systems companies, and research universities within reach of a single facility. Proximity lowers the cost of placing engineers, sharing context, and sustaining programs that run for years.

Call to Action

The cadence of the AI era is ultimately set by how quickly a materials breakthrough becomes a manufacturable process, and then how quickly that process reaches high yield. Those have historically been the industry's longest and least-examined lead times.

The time lost at the boundaries between organizations, rather than to the physics, is the part that can be addressed. The EPIC Center is Applied's attempt to remove those boundaries rather than run the same relay faster. Its model enables co-development that starts earlier, technology that arrives at the fab more mature, a shorter path to yield, and multi-generation roadmap visibility that tells Applied where to invest next.

For investors, EPIC is a structural bet on where value in the semiconductor supply chain is moving: toward the companies that help define what's manufacturable rather than only supplying the tools to manufacture it. For chipmakers and designers, it's a way to reduce the duration and risk of transitions they can't avoid. For the broader market, it's a test of whether the industry's longest lead time can be shortened by design.

The roster already in place shows that the industry sees value in changing the development model. The center opened in October 2026, and its first program outcomes will be what prove the model.

Copyright © 2026 Moor Insights & Strategy

Important Information About This Paper

Contributor(s)

Matt Kimball, Vice President and Principal Analyst, Datacenter Technologies

Publisher

Patrick Moorhead, CEO, Founder, and Chief Analyst at Moor Insights & Strategy

Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying the author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

This document, including any supporting materials, is owned by Moor Insights & Strategy. This publication may not be reproduced, distributed, or shared in any form without Moor Insights & Strategy's prior written permission.

Disclosures

Applied Materials commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

© 2026 Moor Insights & Strategy. Company and product names are used for informational purposes only and may be trademarks of their respective owners.