Technology Report 2024

Technology meets the moment as AI delivers results.

BAIN & COMPANY

Authors and acknowledgments

David Crawford, leader of Bain & Company's Global Technology practice, and a team led by Dana Aulanier, practice vice president of the Technology practice, prepared this report.

Bain Partners Saikat Banerjee, Bharat Bansal, Gregory Callahan, David Crawford, Matthew Crupi, Arjun Dutt, Matt Eldridge, Greg Fiore, Jonathan Frick, Erin Gillman, Adam Haller, Peter Hanbury, Karen Harris, Simon Heap, Anne Hoecker, Chris Johnson, Dan Levy, David Lipman, Sandeep Nayak, Christopher Perry, Bill Radzevych, Paul Renno, Michael Schallehn, Stuart Sim, Roy Singh, Velu Sinha, Colleen von Eckartsberg, and Jue Wang; Associate Partners Jay Bhatnagar and Kenzi Haygood; Expert Partners Purna Doddapaneni, Bala Oarameshwan, and Balaji Thirumalai; and Expert Senior Manager Martin Goette wrote its chapters.

The authors wish to thank Senior Managers Savi Joshi and Ishan Shrestha; Manager Mike Owen; Consultants John Petrie and Grey Pierce; Associate consultants Polly Moser, Nikhil Sriram, and Michael Yoo; Practice Vice President Jennifer Ferrigan; Practice Directors Lauren Brom and Glyn Truscott; Practice Senior Manager Tarun Gupta; Bain Capability Network Senior Manager Eva Gupta; Bain Capability Network Project Leaders Isha Kanna and Vaishali Sharma; and Bain Capability Network Associate Keshary Garg for their contributions; and John Campbell, David Diamond, Jeff Bauter Engel, Mike Oneal, and David Sims for their editorial support. The authors would also like to thank the teams at Aura and ClassifAI for their support.

Contents

Technology Meets the Moment as AI Delivers Results

By David Crawford

In 2024, the technology sector moved firmly into the AI phase of computing. Cloud service providers, enterprises, and technology vendors are spending more on AI than ever, and adoption rates are high. But skeptics are wary of AI's return on investment. What explains the dissonance? Our work with clients suggests that AI, more than other technology disruptions, generates little value from deployment alone. Creating value with AI requires changes in the working processes of hundreds or thousands of employees. Companies need to conduct business diagnostics, redesign processes, set targets, and manage change as they deploy this technology. But early proof points from our client work are encouraging, showing that generative AI initiatives could be worth up to 20% of EBITDA.

Bain's *Technology Report 2024* examines AI's sweeping impact on industry structure, enterprise value, data centers, geopolitical trading blocs, software, services, business opportunities, and resources and talent.

David Crawford
Leader of Bain's Global Technology Practice

Examples of generative AI benefits across functions

Customer service and contact centers

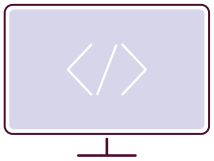
20%-35%

time reduction for manual responses

Sales and marketing

30%-50%

less time spent on content creation



Software product development

15%

time reduction in coding-related activities

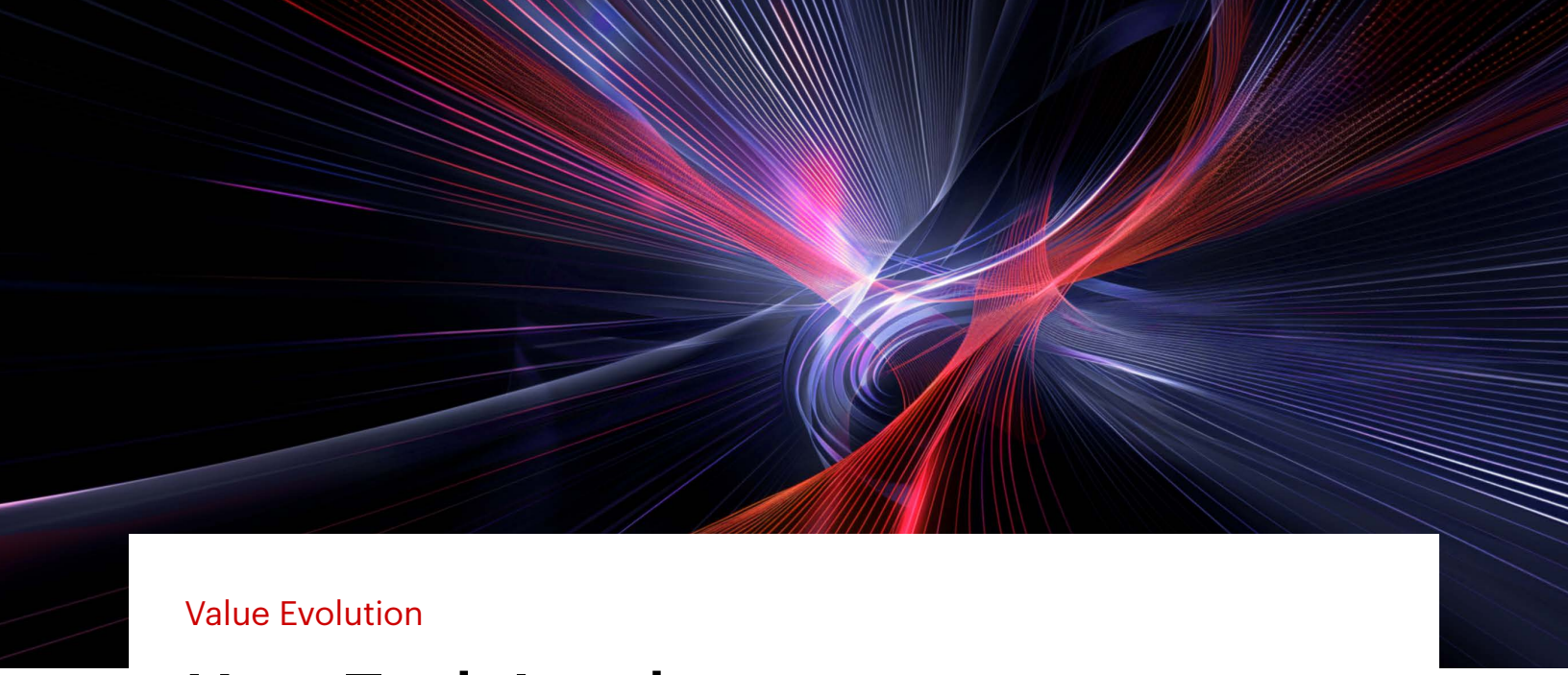
Back office and other productivity

20%-50%

task automation for document comparison

Source: Bain & Company

Value Evolution



Value Evolution

How Tech Leaders Commercialize Innovation

Today's incumbents know how to nurture new businesses even when it means disrupting existing ones.

By Matthew Crupi, Chris Johnson, and David Crawford

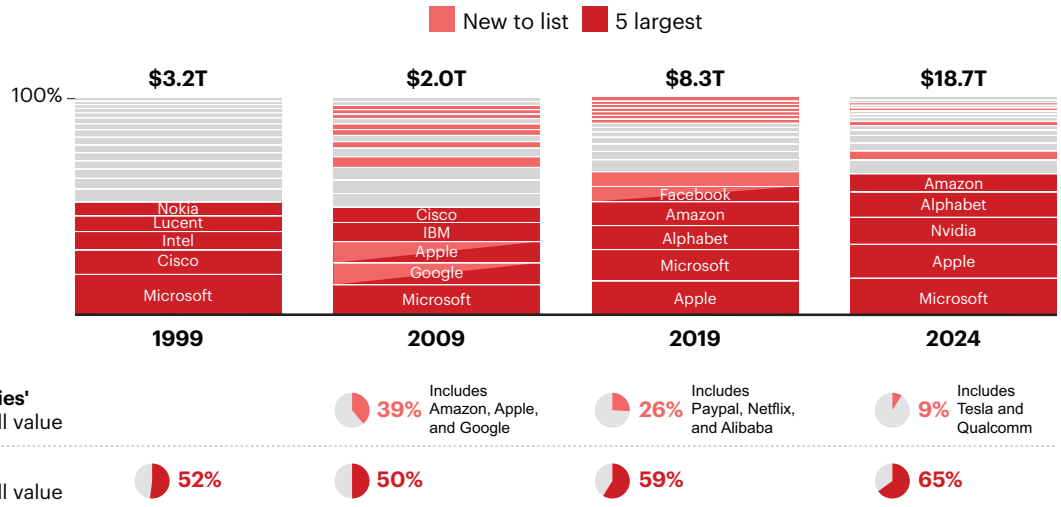
At a Glance

- Disruption rates are higher in technology than in other sectors. Typically, half of the 25 most valuable tech companies fall out of the top ranks every 10 years.
- Recently, the most valuable tech companies have shown remarkable resilience, holding spots at the top for many years and expanding their share of market value.
- Their success relies on their ability to identify disruptive trends and successfully scale and commercialize them, creating "winner takes most" dynamics.
- These companies may be seen as having a competitive advantage at a time when significant innovations require enormous resources in computational power, connectivity, and data.

Disruption in the technology sector hasn't slowed down, but turnover among the top companies has: Four of the top five most valuable technology companies in 2024 were among the top tech companies in 2019, and three of the five were on the list in 2009—Microsoft, Apple, and Alphabet, which was then Google *(see Figure 1)*.

Figure 1: The top five companies in the tech sector have increased their market shares over the past 15 years

Technology market caps by calendar year



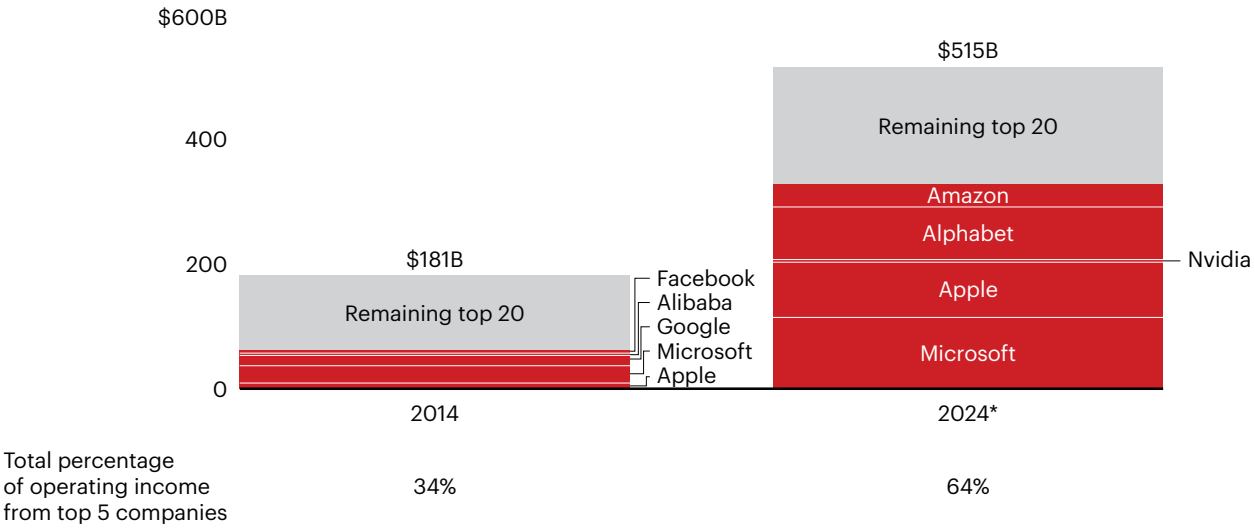
Notes: Market cap calculated on December 31 of year listed except 2024, where market cap is from May 20, 2024; top 20 technology and telecom equipment companies, excluding telecom services and consumer goods companies; Google was rebranded as Alphabet in 2015; Facebook was rebranded as Meta in 2021
Sources: S&P Capital IQ; Bain analysis

What's more, value is increasingly concentrated in these sector leaders. Together, the top five tech companies account for 63% of the market capitalization of the top 20 companies in the sector, up from only 53% a decade ago *(see Figure 2)*. The top five also represent 64% of the profit pool, nearly twice the amount it was 10 years ago (34%).

This durability among tech leadership is atypical because disruption often shifts the leaderboard as new entrants rise at the expense of incumbents that traditionally have been more invested in existing technology and less able or willing to pivot and embrace disruptive technology.

What changed? One reason these leaders have thrived through disruption is they have built businesses with scale effects—on both the supply and demand side. On the supply side, commercializing technology innovations requires large volumes of computational power. This plays to the strengths of companies that have large computing assets and relevant competence. On the demand side, these businesses exhibit network effects that arise from mining huge data sets. The combination can create a "winner takes most" market dynamic and raise barriers to entry. These leaders also benefited from a period of intense globalization, which allowed them to enter new markets, increase their customer base, and accumulate scale. Additionally, in post-pandemic capital markets, investors have favored earnings and safety over rapid growth.

Figure 2: Operating income for the top companies has grown disproportionately



*2024 rankings by market cap are from May 20, 2024, using 2023 operating income data
Notes: Google was rebranded as Alphabet in 2015; Facebook was rebranded as Meta in 2021
Sources: OPEXEngine; annual reports

Ripe for disruption?

What can executives anticipate next? The waves of disruption roiling the tech industry today are so significant that they will test the durability of the incumbents and are likely to shift value in the sector.

- **AI.** The rapid and massive adoption of AI by companies across sectors will force change in the tech sector and beyond. Incumbent leaders including Alphabet and Microsoft are disrupting their own core businesses to focus on the AI opportunity. GPU leader Nvidia has burst into a top-tier position, achieving market cap values of more than $3 trillion in the second quarter of 2024.

- **Geopolitics and national security.** National and regional technology ecosystems are developing as countries look to reward allies and guard against competitive threats. Many national governments are subsidizing or otherwise incentivizing investment in local and national tech endeavors. A major security or geopolitical event could accelerate the pace of this disruption.

- **Backlash against tech.** The sector is under scrutiny by lawmakers and regulators in several regions who want to check the power of the largest tech companies. Acquisitions are under heavy scrutiny. Attempts to break apart large companies could generate new sources of disruption.

The combination of these and other pressures has the potential to shift the leaderboard, as shown by Nvidia's rise to the top ranks over the past two years.

Takeaways for executives

The resilience of today's leaderboard may be partly due to the benefits of scale. However, today's technology leaders may also be skillful at identifying disruption and reinventing their businesses in ways that allow them to move from one strength to another. Running the core at full potential—ensuring that strategic, operational, financial, and organizational goals are met—is an essential foundation for achieving any growth ambitions. Sector leaders also demonstrate five important traits that have helped them stay at or rise to the top.

The resilience of today's leaderboard may be partly due to the benefits of scale. However, today's technology leaders may also be skillful at identifying disruption and reinventing their businesses in ways that allow them to move from one strength to another.

Be willing to self-disrupt. Incumbents must wrestle with the risk of cannibalizing existing businesses as they stand up new, competitive ones. Leaders find ways to fund new businesses, leveraging the momentum of existing ones. Today's incumbents are better at self-disruption, finding ways to challenge their core business and keep their insurgent mission alive.

Ten years ago, Microsoft was struggling to maintain morale in the face of dire predictions about the fate of PCs and middling results of its big push into mobile phones. New CEO Satya Nadella helped reposition Microsoft as a leader in a cloud-first world, relaunching its Azure platform and recommitting to selling Office 365 as a service, even though cloud competed with its traditional server business. More recently, by partnering with OpenAI and integrating advanced AI into products like Azure and Office, Microsoft continues to champion innovation and challenge its core business.

More recently, Nvidia saw the opportunity to leverage technology used in its gaming GPUs to support parallel processing, essential for AI workloads and mining crypto currency. This forward thinking allowed it to embrace new trends, funding its new ventures with the profits from its gaming GPU business. The move diversified Nvidia's revenue streams, solidified its role as a critical player in the future of technology, and in June 2024 made it the most valuable company in the world.

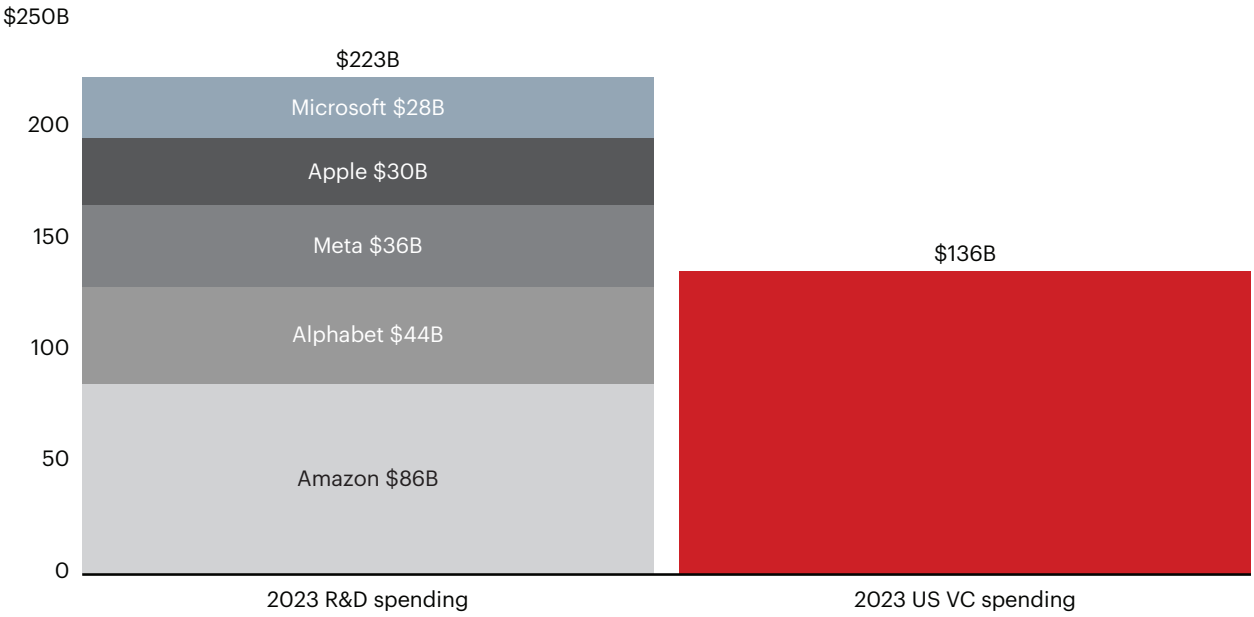
Identify new trends before they're mainstream. These leaders develop sophisticated future-sensing capabilities. They develop alternate scenarios of possible futures and monitor market signals that indicate the direction of travel.

For example, Netflix recognized the inevitable rise of streaming video earlier than most competitors and transitioned from a DVD rental service to a streaming platform. Its timely shift positioned it as a market leader in digital content delivery and funded its large investments in original content creation.

Invest in innovation. Companies strengthen their long-term resilience by continually exploring and evaluating new growth engines. Tech leaders have become extremely profitable and are able to reinvest in innovation. In 2023, the big five hyperscalers (Microsoft, Apple, Alphabet, Meta, and Amazon) spent $223 billion on R&D, about 1.6 times as much as all venture capital spending in the US *(see Figure 3)*. They have corporate venture capital capabilities and are investing in, working with, and sometimes acqui-hiring disruptive start-ups. Apple's substantial investments in R&D have delivered significant competitive advantage over the years. One example is the Apple silicon project, which led to the development of the M1 and M2 chips that boosted performance and efficiency in its Mac line.

Make skillful use of M&A. Large M&A deals in the technology sector can expect significant regulatory scrutiny, but most large tech deals ultimately create more competition and lower prices for customers.

Figure 3: In 2023, the top five tech companies spent $223 billion on R&D, about 1.6 times as much as all venture capital spending in the US



Notes: 2024 market cap rankings are from May 20, 2024; R&D spending is from 2023; venture capital investment value includes seed, series, and corporate VC rounds
Sources: S&P Capital IQ; annual reports; Startup Cruncher; Crunchbase

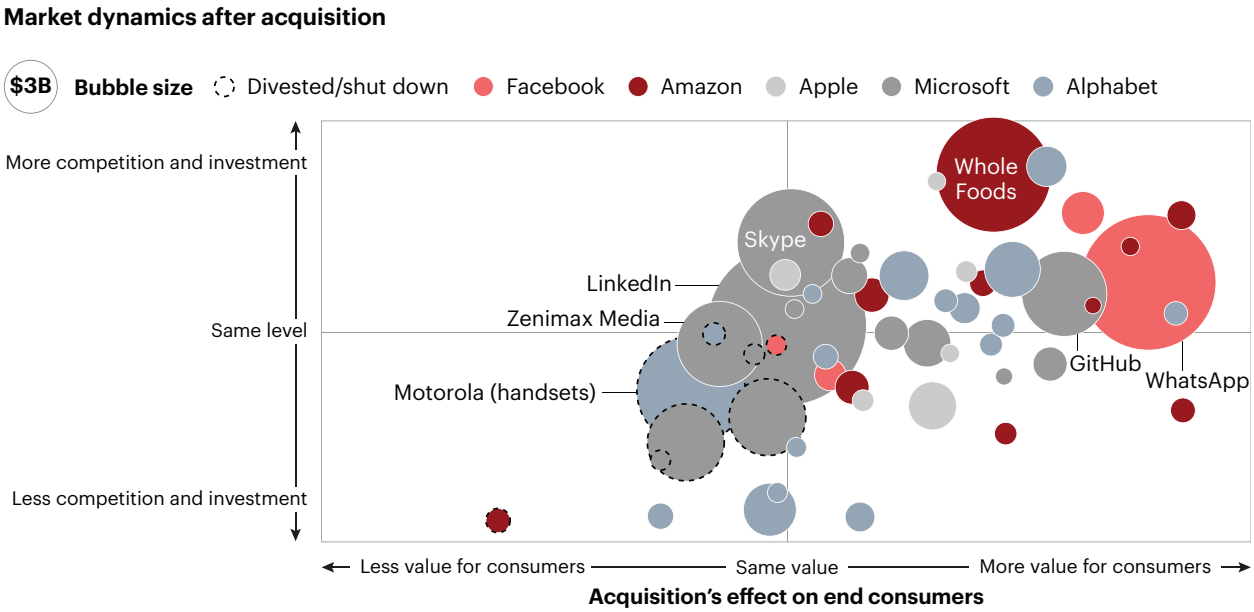
Bain's longitudinal analysis of M&A activity conducted by five large tech firms between 2005 and 2020 found that 72% of the deals increased competition and reduced prices *(see Figure 4)*.

Large tech companies are able to attract top talent while also funding and commercializing innovation in a way that national labs, universities, and other research institutions struggle to accomplish. The cost of training a leading-edge generative AI model is an excellent example: It requires data, computational resources, and funding on a scale that few can achieve.

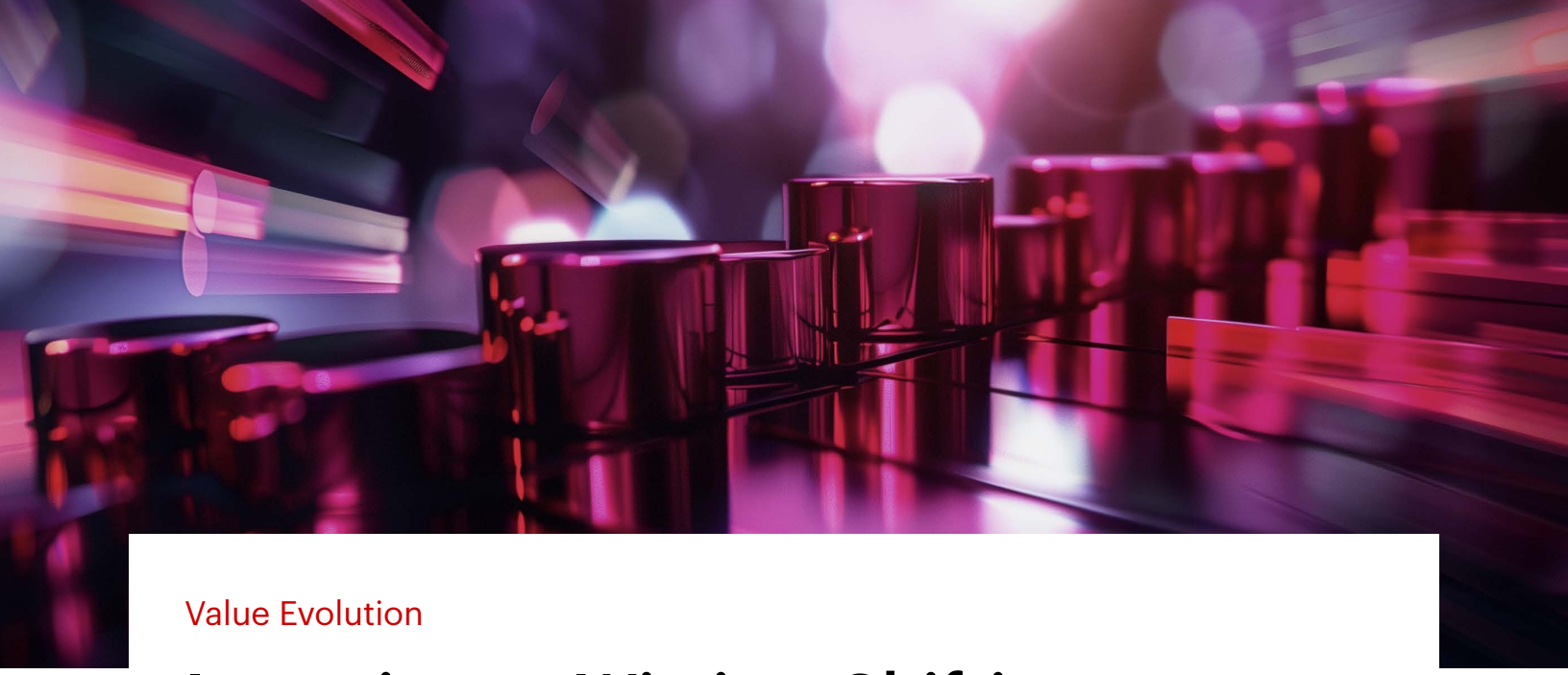
Build ecosystems and partnerships. Creating an ecosystem around disruptive innovations incentivizes other companies that can benefit financially from the incumbent's platform. Ecosystems distribute R&D costs and increase the likelihood that partners may develop innovations that directly benefit the incumbent's business. Amazon Web Services (AWS) has created a vast ecosystem that supports numerous start-ups and enterprises, fostering innovations that enhance its cloud services platform.

Even as the pace of technology continues to accelerate, the most valuable companies in technology are finding ways to embrace disruption and capture market value. At incumbents and new ventures alike, management teams should be looking ahead to identify the next innovation that could disrupt their core business or launch them into the top ranks of the sector.

Figure 4: Most M&A spending among US hyperscalers between 2005 and 2020 benefited consumers or enriched market dynamics



Notes: Includes all outright acquisitions over $300 million (2005–2020) by Amazon, Apple, Facebook, Alphabet, and Microsoft; acquisition labels are limited to the largest deals for legibility
Sources: Bain analysis; company websites; company financial reports; news articles; press releases; blog posts; equity analyst reports; industry reports



Value Evolution

Investing to Win in a Shifting Technology Market

With private equity investors targeting profits—not just growth—value creation is all about operational acumen.

By David Lipman, Christopher Perry, Jonny Holliday, Thibaud Chabrelié, and Jen Smith

At a Glance

- It's no secret that elevated interest rates and uncertainty about future growth continue to put a drag on private equity dealmaking in the tech sector.
- What's changed is that valuations on the deals getting done reflect an increased focus on both revenue and cash flow as investors pivot away from growth at any cost.
- Winning, in other words, will require different muscles: The winners coming out of this slump will be those investors adept at finding operational improvements that both boost margins *and* enhance growth.

It wasn't so long ago that private equity investors were convinced the slide in tech dealmaking would reverse itself by mid 2024. Stable, if not falling, interest rates would combine with aging portfolios and the industry's mountains of dry powder to prod the market forward again.

It hasn't turned out that way.

While deal markets bottomed out in the year's first half, private investors continue to wrestle with heavy uncertainty about when central bankers may finally ease rates. That, coupled with choppy

growth prospects for many software-as-a-service (SaaS) companies, has left buyers and sellers at odds over valuations, resulting in a waiting game. Until we see a meaningful reversal in rates, it is unlikely that tech dealmaking will regain anything like its former momentum.

What we do know already, however, is that investor expectations have shifted during the downturn in ways that have clear implications for how tech assets will have to be managed in the months and years ahead.

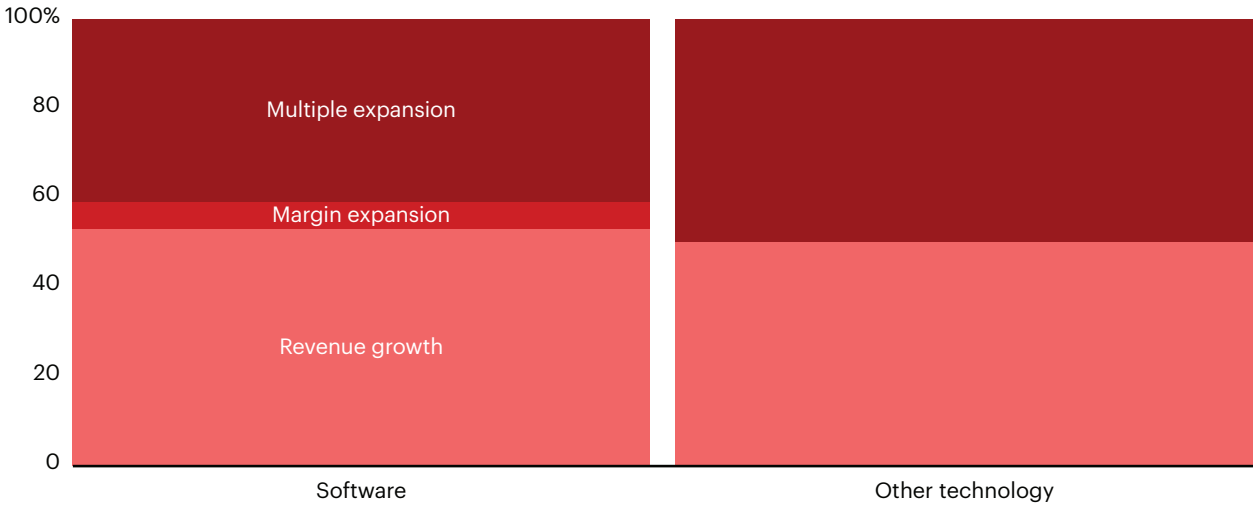
Tech investors have historically driven outsized returns in private equity through revenue growth and multiple expansion *(see Figure 1)*.

But in a higher-rate environment, multiple expansion is no longer a given, and investors are looking for a more balanced "Rule of 40"—the oft-used valuation formula suggesting that growth rate and profit margin should add up to 40% or more (*see Figure 2)*.

The importance of healthy growth isn't going away, but assets that exhibit strong growth prospects *and* robust cash flow are the ones rewarded with premium valuations in today's market. Consider EQT's recent acquisition of supply chain specialist Avetta, which sold for $3 billion (including debt), or 24 times its $125 million in projected 2024 earnings before interest, taxes, depreciation, and amortization (EBITDA).

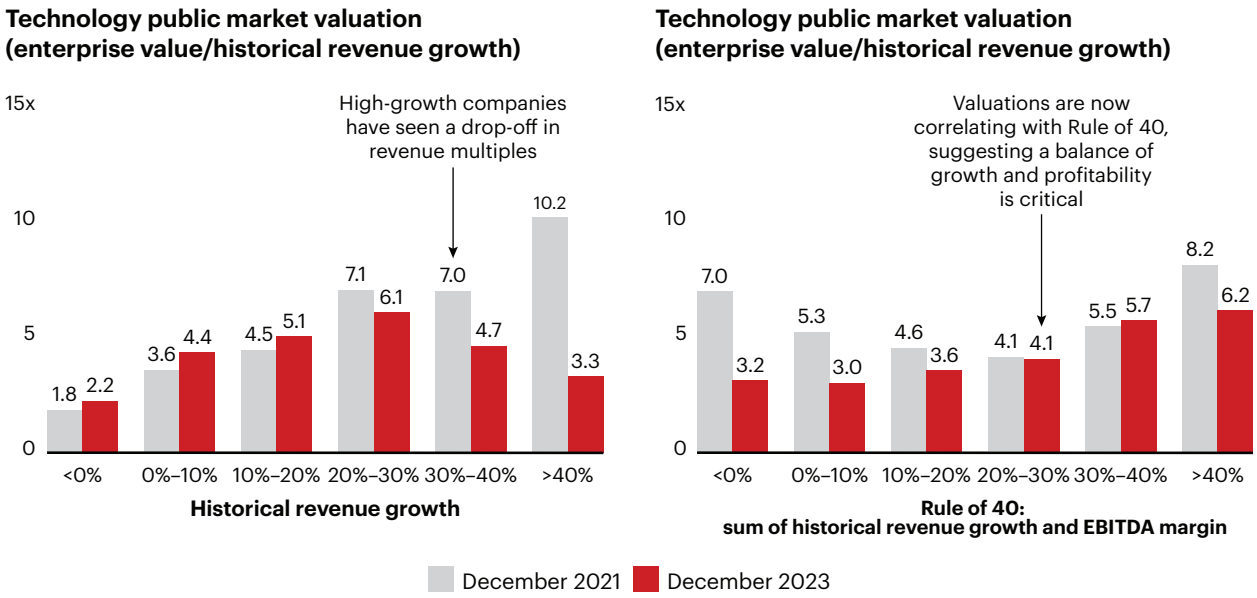
Figure 1: Historically, revenue growth and multiple expansion have largely driven private equity returns in technology

Proportion of deal value created (based on median value creation index)



Notes: Median value creation index by value creation lever and sector; includes buyout and growth deals, all sizes, fully realized; North America-based; year of investment 2010–2023; all figures calculated in USD; median value creation index attributes returns across multiple expansion, margin expansion, and revenue growth; negative value creation indexes set to 0 and net value gained removed proportionally from drivers with positive indexes
Source: DealEdge powered by CEPRES

Figure 2: Valuations today are rewarding a Rule of 40 that balances growth and cash flow



Notes: Includes US-based "Internet services and infrastructure" and "software" companies with greater than $50M in last 12 months (LTM) revenue; revenue growth based on revenue figures for current and prior years; n=15 or more for all growth segments for all years; multiples greater than 100x and less than 0x are excluded
Source: S&P Capital IQ

The renewed focus on profitability is hardly surprising. While dealmaking has slowed, the market is no less competitive. Many of the multi-sector funds that rushed into the software space during the post-Covid-19 boom remain in the hunt, ensuring that prices for any quality asset remain high. Avoiding the buyer's curse means underwriting the kind of performance necessary to justify those prices. And for the many maturing software segments where penetration curves are flattening, that requires identifying ways to displace existing products, not just selling into whitespace. That in turn dials up the importance of investments in strong R&D and go-to-market capabilities funded out of cash flow.

The investors with a clear advantage in this environment are those adept at boosting EBITDA through operational improvements. Doing so without compromising growth is easier said than done, but the firms getting it right follow a clear set of principles:

Invest (and cut) strategically. Portfolio companies need to match underlying demand and revenue growth targets at a segment or product level. They also need a clear understanding of where they have a unique right to win and what it will cost to get there. Leaders know where and how to rightsize without cutting into muscle. Lower revenue growth targets for a given product line, for instance, might mean rebalancing maintenance vs. new product expenses in R&D to optimize spend. Supporting margin targets by simply asking each function to cut 10% of costs is almost never the right answer.

Follow the money. Go-to-market and R&D are typically the areas of biggest spend for most software companies, making them obvious targets for rightsizing. "Hunting" vs. "farming" accounts, for instance, require different skills and compensation structures, so striking the right balance can point to savings. Fewer projected implementations for a given SaaS product also might signal fewer service requirements.

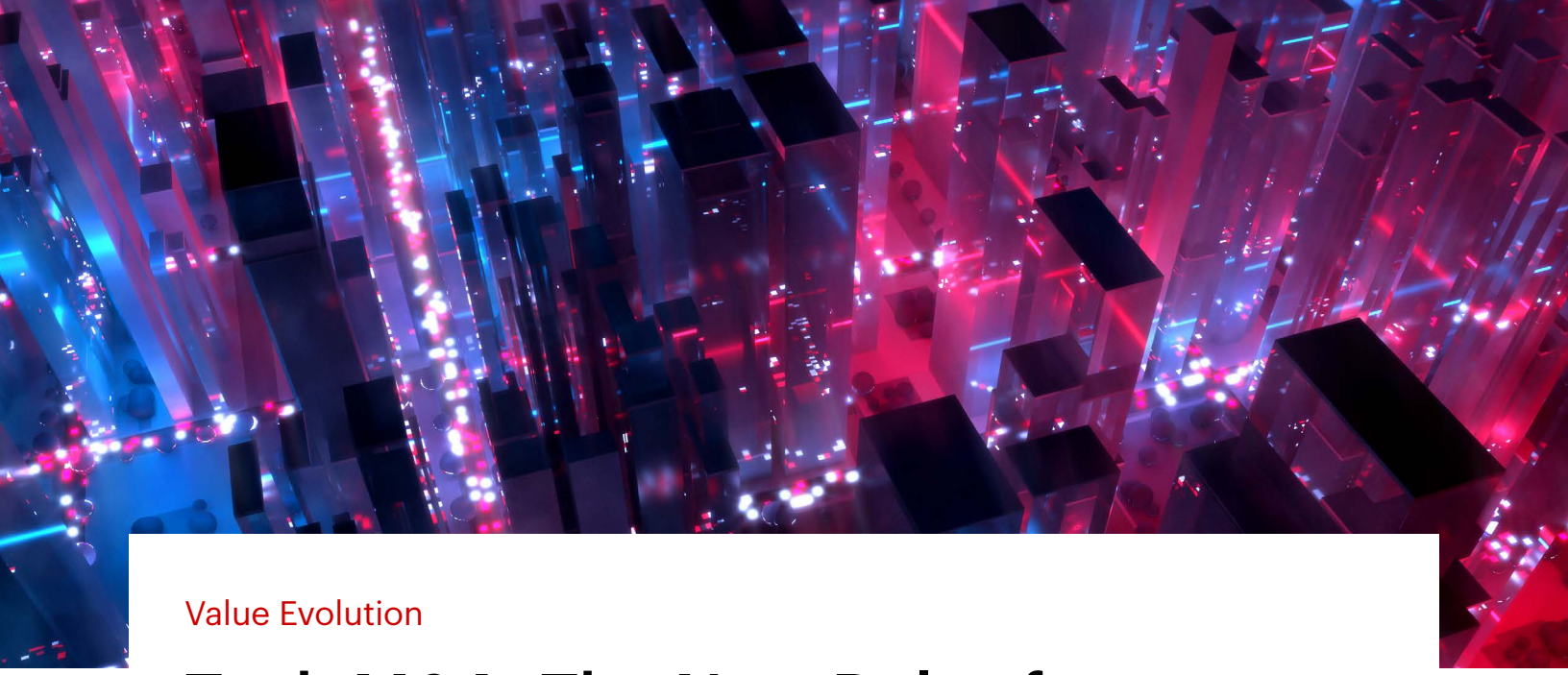
When one large private equity firm recently set out to combine two promising SaaS companies in the services space, it saw clear openings to cut costs by eliminating overlap in the commercial functions. Doing that without compromising growth, however, demanded the hard work of evaluating the combined pool of customers to understand each one's full potential and how best to go after it.

Working from a deeper understanding of customers and segments, the new management team designed the right product initiatives and go-to-market motions to focus sales and marketing on the deepest pools of revenue—effectively creating growth vs. just chasing it down. The strategic rightsizing not only captured $30 million in cost synergies, but it also helped the company identify $7 billion in untapped market whitespace while tagging 100 existing accounts primed for cross-selling. The new company emerged leaner, but it was also significantly more effective.

Get going on generative AI. It's easy to get lost in the hype surrounding these potentially transformative technologies. But the PE investors gaining the most traction recognize a couple of important things. First, they are accelerating plans to use generative AI tools to boost operational efficiency and effectiveness in areas where there is already evidence of measurable benefit—functions like software development and customer support. Second, they are exploring how to enhance or reimagine product offerings but are realistic about assumptions of near-term revenue uplift. AI needs to be part of long-term strategic planning for any software business, both in terms of offensive and defensive moves. Right now, though, it is critical to get moving on piloting and deploying these technologies in the areas that will pay off today.

Tackle change management head on. Shifting a company's focus from all-out growth to an emphasis on cash flow *and* growth inevitably involves the kind of cultural transformation that demands careful management and communication at all levels. When private equity-backed companies miss their objectives, it is often because a gap opens up between those in the boardroom making plans and those closer to the front line expected to execute them. Many organizations will need to change how people work and how they approach the business. They will also have to reevaluate talent based on the imperatives in new value creation plans. A clear strategy to mobilize the organization is critical to success.

Our crystal ball is no better than anyone else's when it comes to predicting when tech dealmaking will regain its momentum. But we can say this with confidence: The winners in the next upcycle won't just focus on revenue growth. Instead, they will help portfolio companies build the capabilities that produce *profitable* growth sustainably.



Value Evolution

Tech M&A: The New Rules for Scope Deals

Successful deals focus on revenue synergies, not just cost savings.

By Adam Haller, Erin Gillman, and Colleen von Eckartsberg

At a Glance

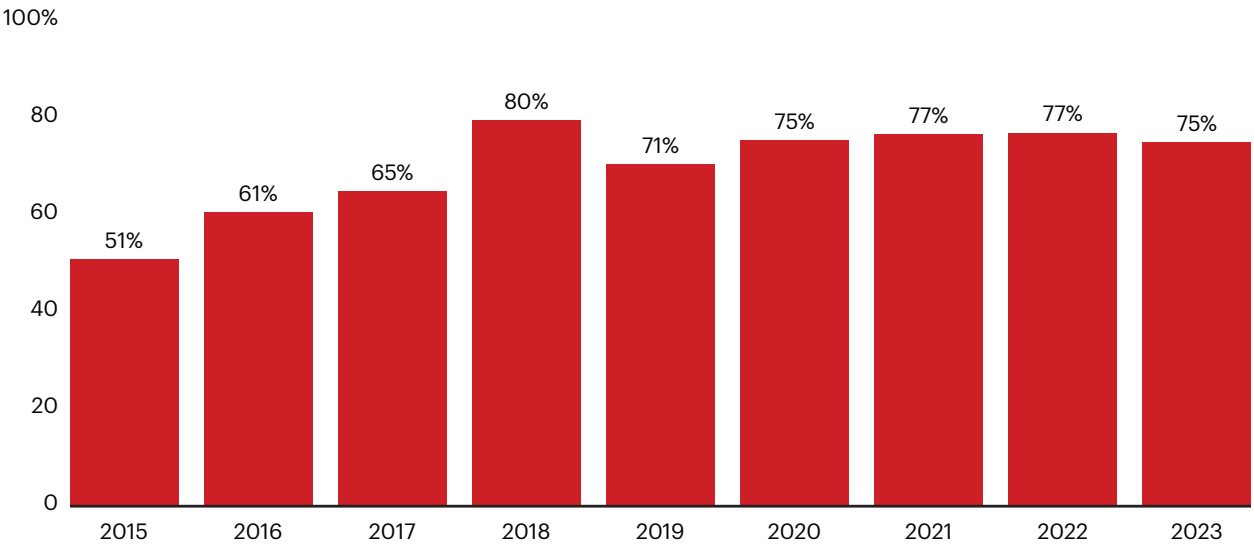
- There's no sign that the industry's reliance on scope deals to spur growth will end any time soon.
- Acquirers often fail to focus on the revenue synergies that will deliver a scope deal's intended value.
- An inability to integrate product portfolios is companies' most common challenge to capturing revenue synergies.
- As they pursue revenue synergies, companies must balance the need for cost synergies, too—without cutting critical capabilities.

When scale acquisitions in tech started encountering more regulatory obstacles, companies shifted their M&A activity to scope deals intended to give an acquirer access to new capabilities, products, or markets. The change has been so dramatic that over the past six years, scope deals have accounted for nearly 80% of all tech industry M&A *(see Figure 1)*. That's a bigger share than in most other industries.

And now there's no sign that the popularity of tech scope deals will give way to a return to massive scale deals any time soon. Big tech still is heavily scrutinized, and if anything, M&A in the industry has become more unpredictable. In addition to looking harder at scale acquisitions, regulators are

Figure 1: From 2015 to 2018, the percentage of tech industry scope deals increased from 50% to 80%, holding steady ever since

Percentage of tech scope M&A deals (by year)



Note: Chart displays all deals >$350M
Source: Bain Practice Operations, Mergers and Acquisition Database (2015–2023)

now challenging scope deals, too, requiring companies to endure a lengthy regulatory process that can delay closings for months.

The trouble is, while the types of deals have changed, too few tech companies have changed their M&A processes to accommodate. As a result, many have discovered that relying on a scale deal playbook almost guarantees that an acquirer won't deliver a scope deal's intended value.

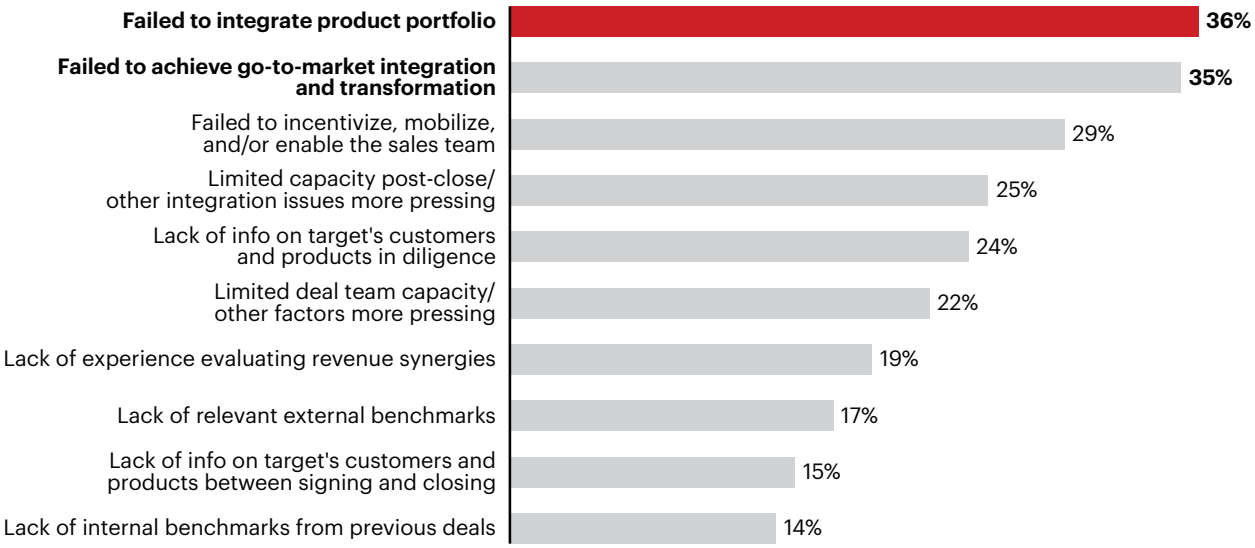
The big challenges

Unlike scale deals, which are primarily predicated on generating cost synergies, scope deals are heavily based on revenue synergies, with companies planning to grow revenues first by cross-selling and longer term by bringing products together. However, too frequently a deal's distractions cause decision makers to lose focus on that objective, and product integration never happens. It's to the point that a failure to integrate product portfolios is the most common challenge to capturing revenue synergies, according to our survey of tech M&A practitioners *(see Figure 2)*.

It gets even more difficult because, as they pursue revenue synergies, companies must balance the need for cost synergies, too. Especially in a high interest rate environment, cost synergies help a company pay off expensive debt as quickly as possible. Yet, by making the moves that enable an

Figure 2: M&A practitioners cite product integration as the top reason for falling short of revenue synergy capture

Top challenges to capturing revenue synergies



Source: Bain M&A Practitioners' 2022 Outlook Survey (n=281)

acquirer to meet targets for substantial cost synergies, companies risk cutting the very capabilities that would underpin revenue synergies.

There's also the talent issue. It's always a challenge to integrate teams when talent and culture vary from one type of business to another. But if there is no clarity on vision or growth in the short term, companies risk watching critical talent flee. Employees worry about the impending change and feel they aren't seeing the upside or the "better together" vision they expected.

Finally, while there's always a risk to the base business before a deal closes, the longer pre-close timeline caused by regulatory approval processes—even with scope deals—creates more uncertainty on all fronts. There's even more time for competitors to make inroads with customers, for example.

What can tech companies do?

The first move is to accept this new reality of tech M&A: In a world of scope deals and revenue synergies, you can't run the 20-year-old playbook that only focuses on taking out costs.

Instead, companies must look at the ways to change the playbook across the specific stages of a deal.

Update the diligence approach. Don't just pressure test the financials and market growth. Set the vision, pressure test scenarios, and align internal leadership on the deal thesis. And use diligence as an opportunity to develop a data-driven view of revenue synergy opportunities (both in go-to-market and product synergies) based on voice of the customer research. Include the non-financial elements of talent and culture in diligence as well.

Make the most of pre-close planning and alignment. Understand the expanding regulatory hurdles and adjust your transaction strategy and integration planning accordingly. The new mantra should be *plan for the worst, but be ready for the best*. That means pre-close integration planning that includes preparations for the earliest and latest possible closing outcomes, with a flexible, stage-gated approach to closing. Indeed, with heightened scrutiny even on scope deals, it's necessary to expect a prolonged timeline.

The new mantra should be *plan for the worst, but be ready for the best*. That means pre-close integration planning that includes preparations for the earliest and latest possible closing outcomes, with a flexible, stage-gated approach to closing.

Data shows that deals which don't face scrutiny usually close within about three months. If regulators request additional information, timelines are likely to be six or more months longer, with those deals turning into court cases reaching delays of up to two years. Challenges can sometimes come from multiple regulatory bodies across the globe, with large deals (those $10 billion or larger) facing more scrutiny. For example, the Microsoft-Activision Blizzard deal was challenged by multiple regulatory bodies, with concerns from the EC, and UK's CMA ultimately setting the restructuring terms and timeline.

Accelerate Day 1 revenue synergies. Another critical pre-close step requires data-driven sales planning aimed at enabling the sales team to hit the ground running. Focus on the highest-value opportunities—pinpointing Day 1 sales plays with a targeted list of reps and customers. Equip sales teams for success with a run book. Show salespeople how they can make money and how the *better together* story benefits them. Give them clear rules of engagement and the right comp structure.

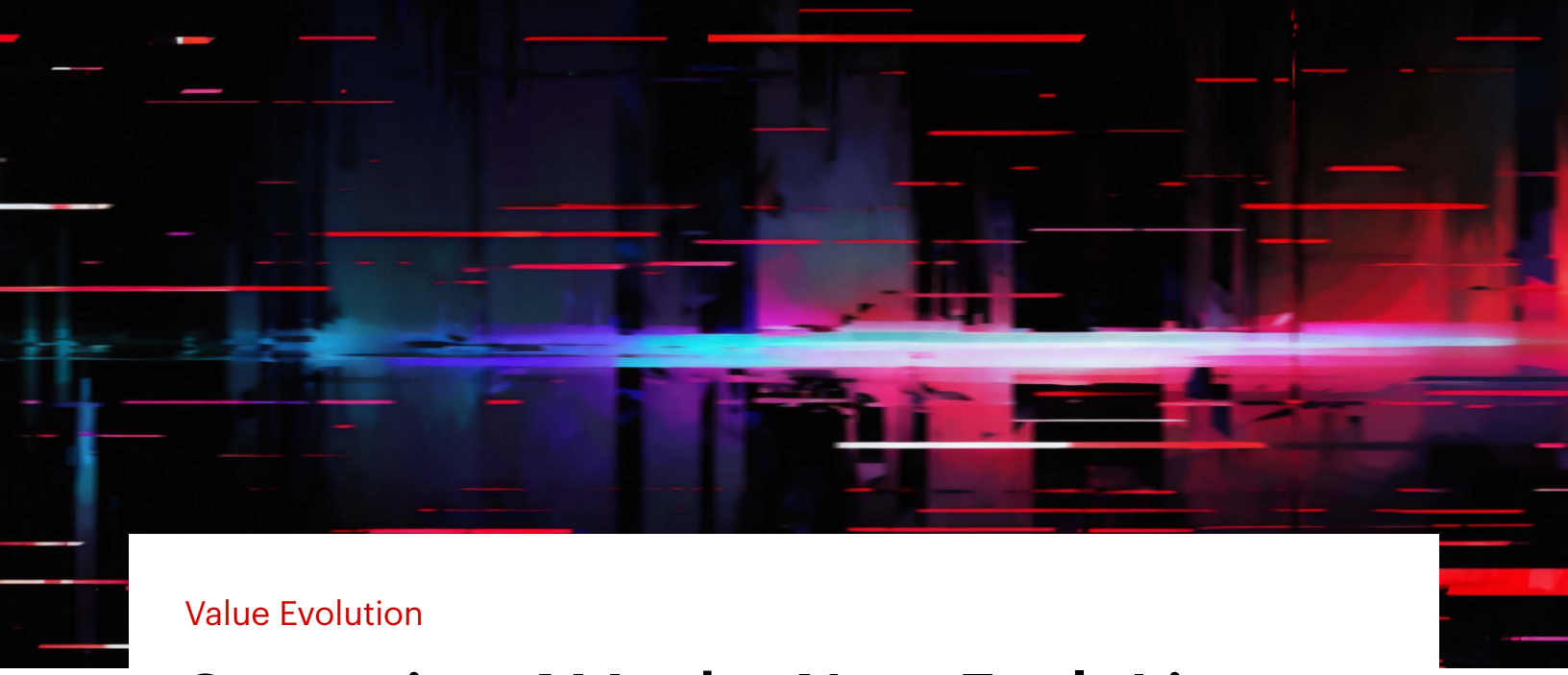
These elements of pre-close planning will buy you time to invest in bringing the products together. But it's also necessary to devise a thoughtful approach to talent and culture. Engage and inspire talent from the onset by tailoring your proposals for critical populations such as engineering and AI data scientists. And invest to understand and address the major cultural "fault lines"—potentially destabilizing differences—so integration planning teams can set the tone.

Hitachi's acquisition of Silicon Valley-based GlobalLogic serves as an example of what to do right. The Japanese company wanted to strengthen its digital engineering capabilities with GlobalLogic, but it knew that cultural differences could be an obstacle. One major risk was that the target would lose its unique strengths under Hitachi—or under any large corporation, for that matter. Hitachi invested heavily to pinpoint and resolve cultural fault line issues. That included several workshops, in-person visits, and a unique cross-cultural team staffed across geographies to smooth over potential misunderstandings. The investment not only preserved GlobalLogic's culture, but also helped Hitachi recognize how that culture encouraged innovation and then apply those learnings back to its own organization.

Know how to motivate immediately post close. Keep a laser focus on executing toward revenue synergy goals, setting the sales team up for success from Day 1. That means establishing a win room and providing sales training while also aligning sales incentives and rules of engagement on shared accounts.

Plan for longer-term integration, making product synergies key to revenue growth. The best companies strategically bring offers together, developing differentiated new customer value propositions and building integrations or unified platforms that, in turn, deliver the longer-term revenue synergies (see "M&A in Technology: Getting Serious about Product Synergies," a chapter in the *M&A Report 2024*). Proactively communicate with customers early on, showing the initial value of the new combined products and charting a clear joint roadmap. Accelerate strategic product planning with a dedicated cross-functional team as part of the integration management office (within regulatory guardrails). And continue to focus on talent, with emphasis on the moments in integration that matter.

In this new world, deals are more expensive than ever, and growth is riskier. But there's a key to capturing the value of any deal. A robust M&A capability, with a tailored and strategic approach to integration, allows any company to move earlier and with more confidence at every stage of the deal.



Value Evolution

Sovereign AI Is the Next Fault Line in the Global Tech Sector

The electronics supply chain was only the start of tech's global decoupling.

By Anne Hoecker, Jonathan Frick, Jue Wang, Balaji Thirumalai, and Karen Harris

At a Glance

- Sovereign AI blocs are emerging as governments worldwide spend billions of dollars subsidizing domestic computing infrastructure and AI models.
- Locally based data center providers account for nearly a quarter of new computing capacity coming online in the next few years.
- Despite sovereign AI momentum, tech incumbents' global scale and deep coffers provide significant advantages over domestic competitors.
- Data center operators will enjoy a short-term windfall, but there's a real risk of overcapacity.

As technology companies race to capitalize on breakthroughs in large language models (LLMs) and generative artificial intelligence (AI), executives must now grapple with an additional layer of complexity and opportunity: the emergence of "sovereign" AI blocs around the world.

De-globalization in technology began with the electronics supply chain, particularly semiconductors. Disruptions from Covid-19 and geopolitical tensions between the US and China (including export

controls and restrictive policies on trade and talent) pushed tech companies to rapidly invest in making their supply chains more resilient. They've expanded their manufacturing footprints beyond China and created more flexibility within their talent pools. With government support, companies are building new semiconductor hubs in places including the US, India, Germany, and Japan.

Now the post-globalization movement in technology is spreading to data, AI, security, and privacy. Governments worldwide—including India, Japan, France, Canada, and the United Arab Emirates—are spending billions of dollars to subsidize sovereign AI. In other words, they're investing in domestic computing infrastructure and AI models developed within their borders, trained on local data and languages.

Now the post-globalization movement in technology is spreading to data, AI, security, and privacy. Governments worldwide—including India, Japan, France, Canada, and the United Arab Emirates—are spending billions of dollars to subsidize sovereign AI.

While it's tempting to compare sovereign AI to the decoupling of semiconductor supply chains, the challenges are quite different. For example, compared to the semiconductor market, which has a complex supply chain with intellectual property fragmented throughout, the AI market is easier to enter. This is largely due to open-source LLMs, which make launching new AI products simpler.

As the sovereign AI push picks up steam, several factors will determine how it plays out.

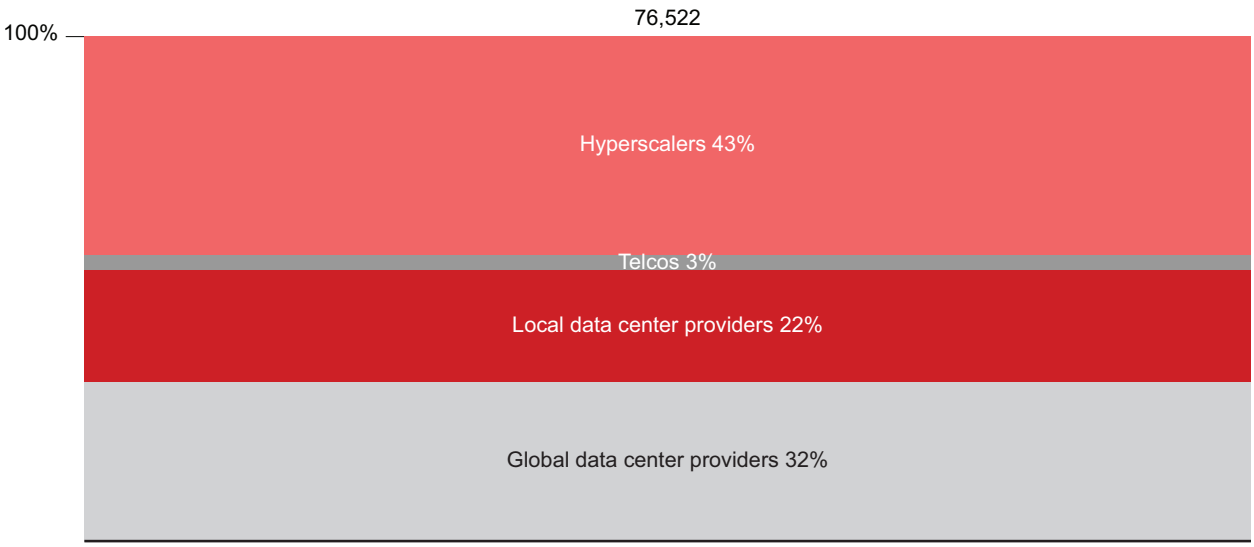
Factors favoring sovereign AI

1. **National interests:** Governments view localized AI as critical for protecting data privacy, ensuring national security, building or strengthening domestic high-tech ecosystems, and growing their economies. Countries can't afford to fully rely on others for AI and cloud computing capabilities due to the economic value at stake and the decoupling of the countries leading the AI race—the US and China.

2. **Infrastructure:** Like any utility, physical infrastructure for AI and cloud computing must be built somewhere and will require massive capital investments in data centers, computing capacity, and electrical grids. These investments intersect with other national infrastructure issues like the green transition in the electricity grid, which AI's significant power demand will complicate.

Figure 1: Nearly a quarter of new data center capacity will come from local providers, with hyperscalers planning the most capacity

Global forecasted new data center capacity through 2027, by provider (MW)



Note: Global providers operate on more than two continents
Sources: IDC 2023 Datacenter Deployment and Spend Forecast, 1H 2023; Bain analysis

Locally based data center providers account for nearly a quarter of new computing capacity coming online in the next few years, while technology hyperscalers are planning to add the most *(see Figure 1)*. It's also notable that national governments have ordered at least 40,000 graphics processing units (GPUs) themselves over the past year.

3. **Regulatory strategies:** AI regulatory strategies are diverging across borders. The leading AI markets—the US, EU, and China—are taking very different approaches so far.

4. **Localization:** Many AI models will need to be specific to local languages and context. Some applications will differ across countries to comply with security and privacy regulations and meet local market needs. AI use cases in healthcare, education, and agriculture, for example, will vary greatly between developed and emerging economies.

Factors working against sovereign AI

1. **Scope of subsidies:** Thus far, governments haven't subsidized national AI initiatives to the extent seen with semiconductor fabs or to the degree likely required to nurture local champions that could compete at scale with global incumbents. (The powerful open-source LLM series Falcon, backed by hundreds of millions of dollars from an arm of the United Arab Emirates government, is a notable exception.)

2. **Global scale:** This still provides critical advantages for developing a winning AI platform, including network effects (e.g., access to a large developer ecosystem), deep coffers, and the ability to spread R&D costs across worldwide operations. LLM training costs have grown exponentially over the past few years, with the most expensive models exceeding $100 million *(see Figure 2)*. Although smaller, more cost-efficient models are also being released, the cost dynamics continue to favor large global firms.

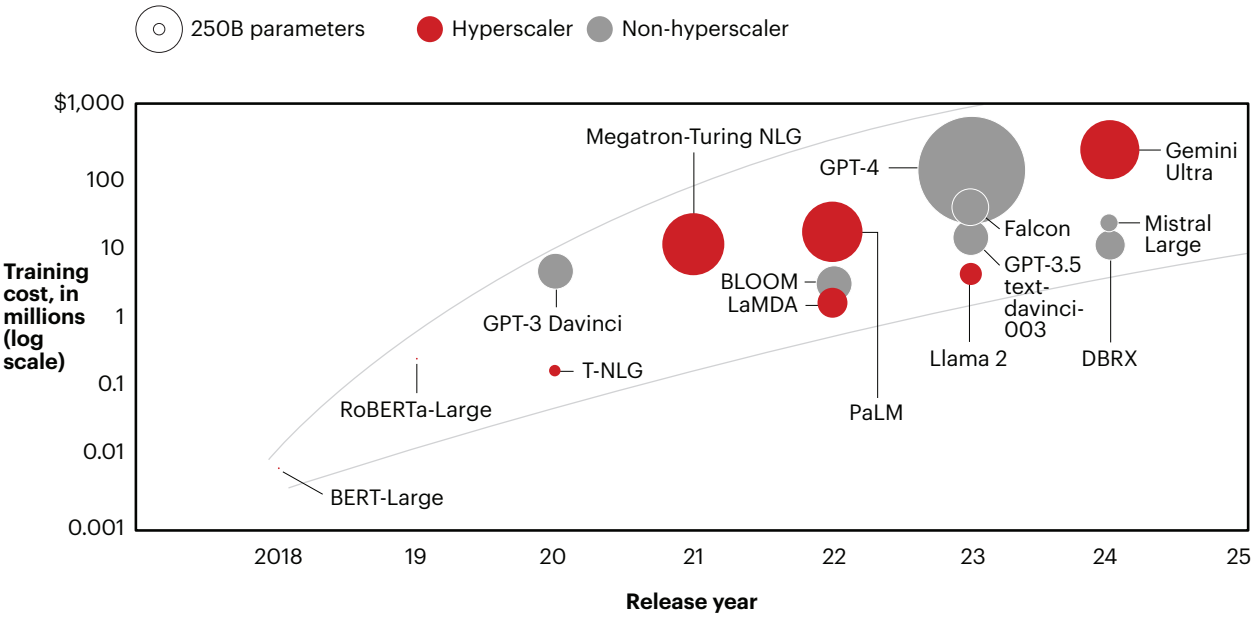
3. **Incumbents' adaptation:** Global tech companies are adapting to governments' push for sovereign AI by localizing operations, complying with local rules, and forming joint ventures with local firms.

4. **Practical realities:** Aspiring domestic competitors must navigate the same practical realities as multinational companies: significant investments in securing land, regulatory approvals, power, connectivity, and other key elements for AI initiatives.

Takeaways for executives

Establishing successful sovereign AI ecosystems will be time-consuming and incredibly expensive. While less complex in some important ways than building semiconductor fabs, these projects require more than securing local subsidies.

Figure 2: Large language model training costs have increased exponentially, but smaller, more cost-efficient models are also being released



Notes: Training cost estimates exclude staff costs; Mistral Large parameters are assumed to be equal to Mistral 8x7B due to similar capabilities; GPT-3.5 text-davinci-003 parameters are assumed to be same as GPT-3.5
Sources: Epochai.org; news articles; analyst reports; company websites; research papers; Bain analysis

Hyperscalers and other big tech firms may continue to invest in localized operations. This could fragment their ecosystems and R&D globally, though their scale will remain a significant advantage.

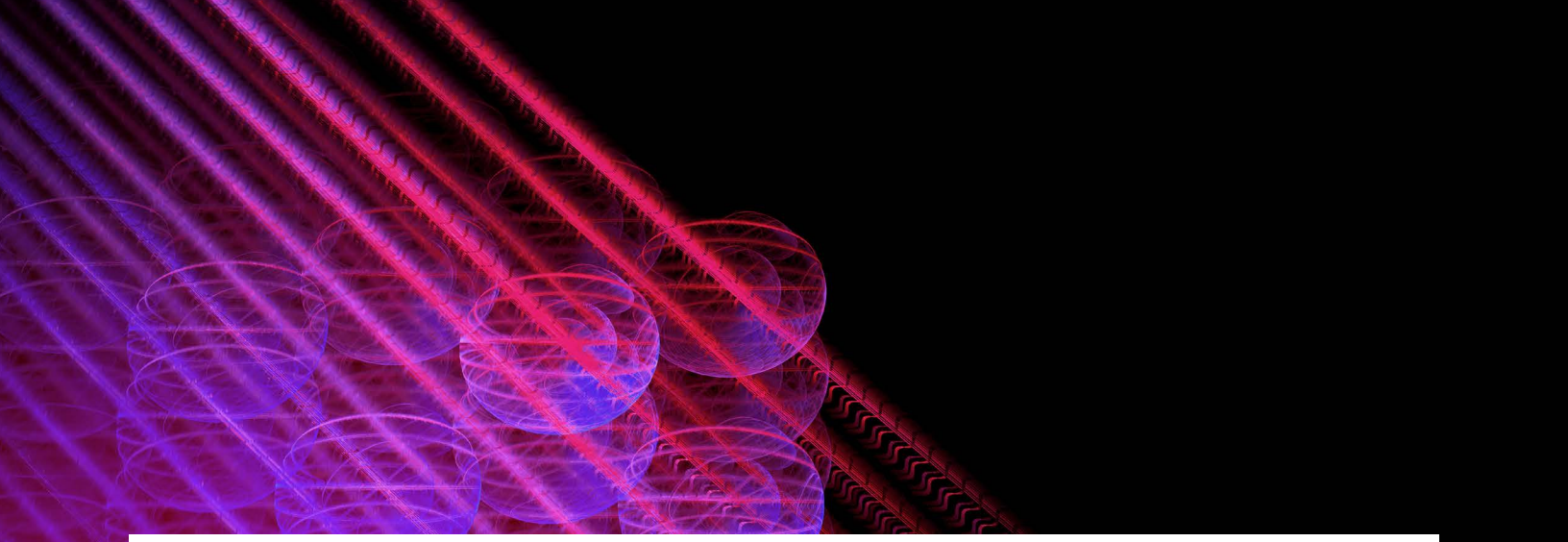
New AI workloads and fragmentation created by sovereignty could enable AI challengers to reach hyperscale. These challengers will need to recognize the power of the current hyperscaler ecosystem and prioritize business opportunities that capitalize on their competitive advantages, while partnering with big tech companies where possible.

Establishing successful sovereign AI ecosystems will be time-consuming and incredibly expensive. While less complex in some important ways than building semiconductor fabs, these projects require more than securing local subsidies.

Data center operators and hardware suppliers will enjoy a short-term windfall as companies and governments splurge on computing capacity. Nvidia, for example, projected $10 billion in revenue from governments' sovereign AI investments in 2024, up from zero last year. However, data center owners risk overcapacity, similar to telecom networks in the early 2000s. Suppliers of silicon and other hardware may see accelerated growth rates level off long-term.

Lastly, investors have a chance to stake high-value claims in a hot asset class, including new sub-asset classes. For example, secured financing tied to GPUs is becoming a more common form of corporate debt. Successful investors will base bets on a well-defined risk/return profile, deciding between lower-risk investments in "picks and shovels" like GPUs and data centers or higher-risk/higher-reward investments such as LLMs and cloud platforms.

Strategic Battlegrounds



Strategic Battlegrounds

Five Functions Where AI Is Already Delivering

Spurred on by early success, companies of all sizes are increasing their spending on generative AI.

David Crawford, Jue Wang, and John Kanan

At a Glance

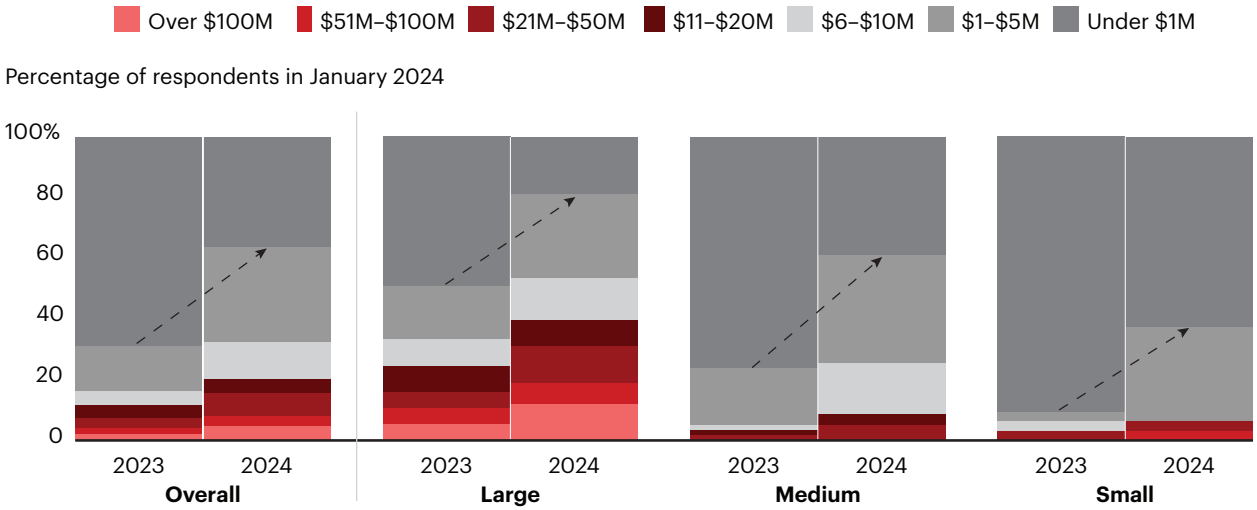
- Companies are ramping up spending on generative AI, especially in software development, customer support, and other areas.
- AI is delivering real efficiency gains across functions, reducing customer support response times by a third and cutting some code-generation times in half.
- More than most disruptions, AI requires some business redesign to capture value. Simply deploying the technology delivers little return on investment.

With some disruptions, fast followers gain a competitive edge by waiting to see what mistakes the first movers make. But that's not what we're seeing with AI: Early adopters are already starting to realize performance gains up to 20% of earnings in as little as 18 to 36 months. They're building capabilities and confidence that are likely to translate to a sustained, competitive advantage, empowering them to redefine operations and develop new business models. The last time we saw a new technology this powerful was when the Internet arrived in the 1990s. And this time, change is happening faster.

At the same time, some investors and analysts remain skeptical about returns on investments in AI. This may be because reaping value from AI requires more than just simply conducting trials or

Figure 1: Companies of all sizes are meaningfully increasing spending on generative AI

Spending on generative AI activities and supporting infrastructure



Notes: 2024 numbers are forecasts; large businesses=more than 10,000 full-time equivalents (FTEs); medium=1,000–10,000; small=less than 1,000; budget for generative AI includes spending on hardware infrastructure, large language models (LLMs), AI workbench and machine learning operations (MLOps) tools, off-the-shelf generative AI applications, and professional services
Source: Bain IT Decision Makers Survey, January 2024 (n=151)

deploying the technology. More so than previous disruptions such as the Internet or cloud, AI requires changes in business processes. Companies that conduct business diagnostics, set targets for business deliverables, redesign processes, and then develop and deploy AI tools are seeing extraordinary value.

These early successes are leading to greater investment: The number of large companies investing over $100 million to implement AI has more than doubled in the past year *(see Figure 1)*. These investments are spurring companies to experiment in hundreds of different use cases, but our research finds that most of the value today can be found in five core areas.

Software and product development

The top use cases for generative AI in software development include code generation, documentation, refactoring, debugging, testing, and run and maintenance. Some developer organizations are already saving 15% to 40% on code generation and documentation, and 30% to 50% or more on refactoring, select testing, and debugging use cases by utilizing the specific patterns and rich datasets that exist beyond the code base.

In some companies, AI deployment has served as the trigger to evaluate software development productivity, expanding their focus to more traditional improvement areas including product management, data-driven prioritization, process stage gate discipline, Agile, and QA shift-left efforts.

Intuit, a financial technology platform for consumers and small businesses, has been testing and scaling more than 30 different use cases to increase end-to-end development velocity throughout the company's software development life cycle with generative AI. By integrating generative AI technology and tools into its development platform, Intuit is improving productivity for product teams (software developers, designers, product managers, data engineers and analysts, technical program managers, etc.). For code generation, the company has seen greater-than-average efficiency gains by tuning its coding assistant tool on Intuit-specific code context patterns and repositories. It has also focused on a set of refactoring tasks to expedite its code base modernization efforts, further accelerating development velocity.

Customer support

Generative AI can do more than automate and optimize customer support; it can also reduce the amount of support needed in the first place. Generative AI's application in customer support includes analytics to anticipate, deflect, and address potential customer issues; chatbots to expand digital self-service offerings and automate interactions; algorithms to connect customers with the most appropriate representative; and knowledge assistant tools that help agents act more efficiently.

Generative AI can reduce adviser response time by up to 35%, support consultants during the resolution process by managing different sources of knowledge, and improve the quality of results by up to 40%.

For example, one technology and manufacturing company developed two cutting-edge generative AI prototype applications for field services. The company launched a maintenance assist copilot to boost productivity of field technicians performing maintenance and repair operations, and it developed new systems to analyze huge amounts of diverse and unstructured building sensor data and coordinate information and decision making for emergency responders.

Generative AI can reduce adviser response time by up to 35%, support consultants during the resolution process by managing different sources of knowledge, and improve the quality of results by up to 40%.

Sales and marketing

In sales and marketing, generative AI is deployed in generating dynamic, personalized content, personalized email marketing, social media engagement automation, automated account planning, and advanced training and support. By automating and optimizing these customer interactions,

generative AI is boosting the productivity of sales reps and other marketing staff, shortening cycle times, reducing churns, and delivering better click-through rates through hyper-personalization.

One technology hardware company, for instance, is transforming content management by simplifying content creation, automating systems and workflows that synthesize, assemble, and publish content, and adopting generative AI tools for some roles. The company aims to reduce time spent on content by 30%. Pilots have already delivered promising results in a variety of uses, meeting and exceeding this goal.

New products and features

Companies are deploying generative AI in product and feature development to create simpler and more user-friendly products and interfaces, and to deliver greater customization and personalization. For example, in healthcare, AI can quickly analyze patient data and offer personalized care plans. In other industries, generative AI enables voice or text chat interfaces for simpler interaction with products.

Carrefour's site, for example, offers a generative AI shopping assistant that can generate shopping lists and menu suggestions based on customer information and input. This simplifies the customer shopping experience while making it more engaging.

Back office

Back-office operations are particularly well suited for generative AI improvements, given the vast number of routine processes that are comparatively easy to automate. In the finance function, for example, generative AI can improve the efficiency of drafting internal audit reports, preparing documentation for tax audits, and running custom financial analyses.

Deutsche Telekom has developed a chatbot for its procurement department that is trained on the company's policies and historical procurement strategies. The chatbot can answer team requests about policy compliance and provide recommendations on vendors, contracts, or fair price for a specific request for proposal. Pilot results across the company suggest that the chatbot could save business users up to 2,000 hours per month and procurement users up to 5,000 hours per month.

Anticipating challenges

Deploying AI is a transformative journey that aims for significant productivity growth, but involves addressing challenges that span technological integration, human adaptation in ways of working, and reimagined business processes.

- **Preparing business processes.** In deploying AI, companies should avoid automating existing complexity into their operations. To do that, they should fix the processes before automating by streamlining, simplifying, and eliminating unnecessary steps. This frees up energy and capacity as they modernize operations.

- **Modernizing data and application environments.** Sprawling databases, multiple sources of truth, and complex application environments hinder the rapid deployment of reliable and productive AI. Investing in modernization and data governance before scaling AI applications releases an additional wave of productivity.

- **Finding technology and services support.** Companies implementing AI in the cloud and on premises need reference designs, large language model (LLM) recommendations, prompt engineering, and application development support. All of these resources are in short supply because so many technology providers are currently introducing foundation model AI into their own products. Graphics processing unit (GPU) infrastructure, in particular, is in high demand.

Leading an AI transformation

A strategic implementation of AI aligns initiatives with the organization's business goals. Whether the changes are incremental or transformational, several best practices are emerging.

- Prioritize AI as a way to generate value, from the CEO down. Set clear targets for return on investment (ROI) and hold teams accountable through the budgeting process for delivering savings and creating value.

- Conduct a business diagnostic. Don't automate bad processes. Invest in mapping out value opportunities and redesigning business processes before automating. Set targets and manage change to improve efficiency as the technology is deployed.

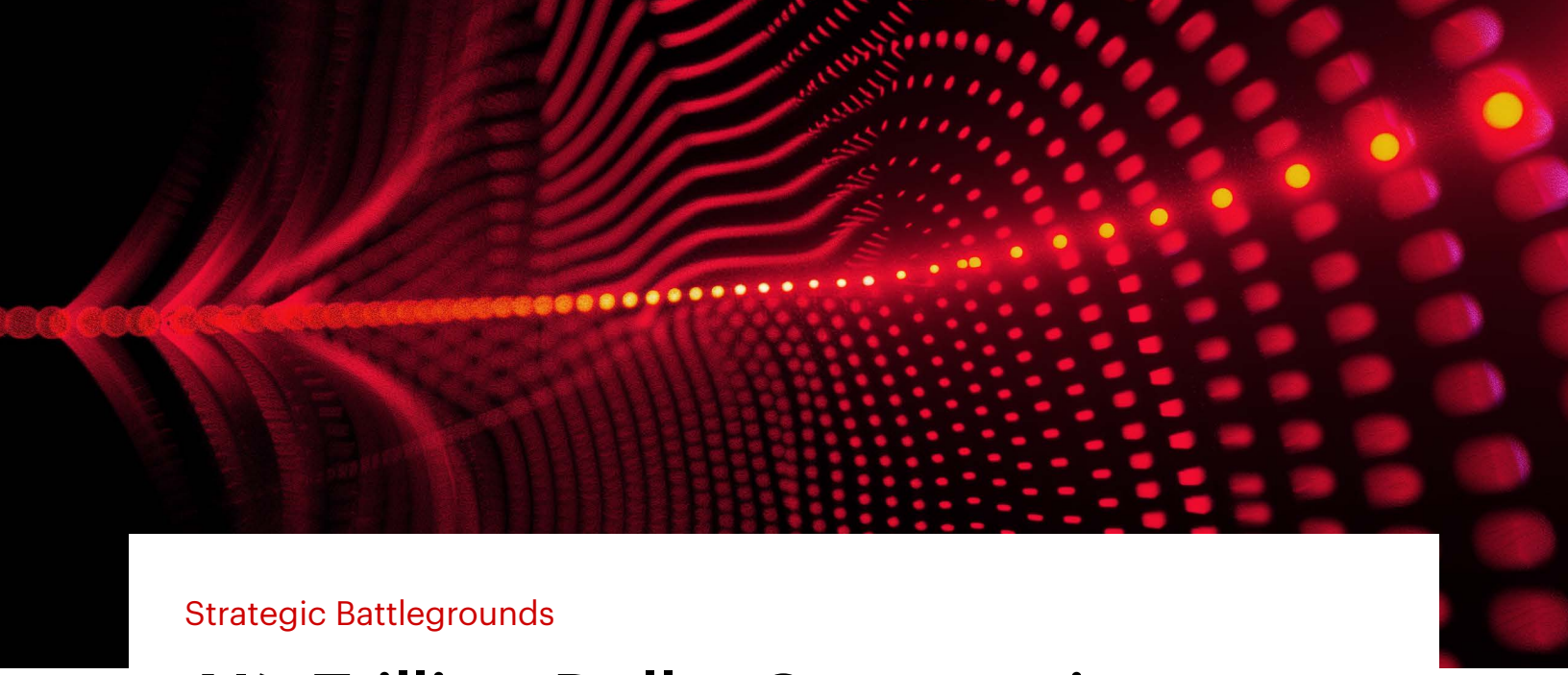
- Define a clear roadmap for use cases. Focus on functional areas with high value potential, such as sales and marketing, customer support, software development, and operations.

- Leverage multiple AI delivery models, including self-service knowledge worker tools (such as Microsoft 365 Copilot), prebuilt commercial AI systems from vendors, and custom AI models, when the need for differentiation and sensitivity of data is high.

- Build shared datasets, AI models, and technology components and platforms to ensure economies of scale across solutions. Improve product management, as well as Agile and DevOps processes, to support high-velocity AI development.

- Develop appropriate risk management, responsible AI, and governance roles, and ensure clear communication and talent strategies for the workforce.

For every enterprise, the AI journey will take a unique form. But across industries and markets, it's clear that the dramatic rise of AI is not a passing hype cycle. The strategic and innovative use of AI will play a key role in achieving competitive advantage over the next decade and beyond. Late adopters are out of time, and companies that fall too far behind the curve will find it difficult to maintain or regain their position.



Strategic Battlegrounds

AI's Trillion-Dollar Opportunity

The market for AI products and services could reach between $780 billion and $990 billion by 2027.

By David Crawford, Jue Wang, and Roy Singh

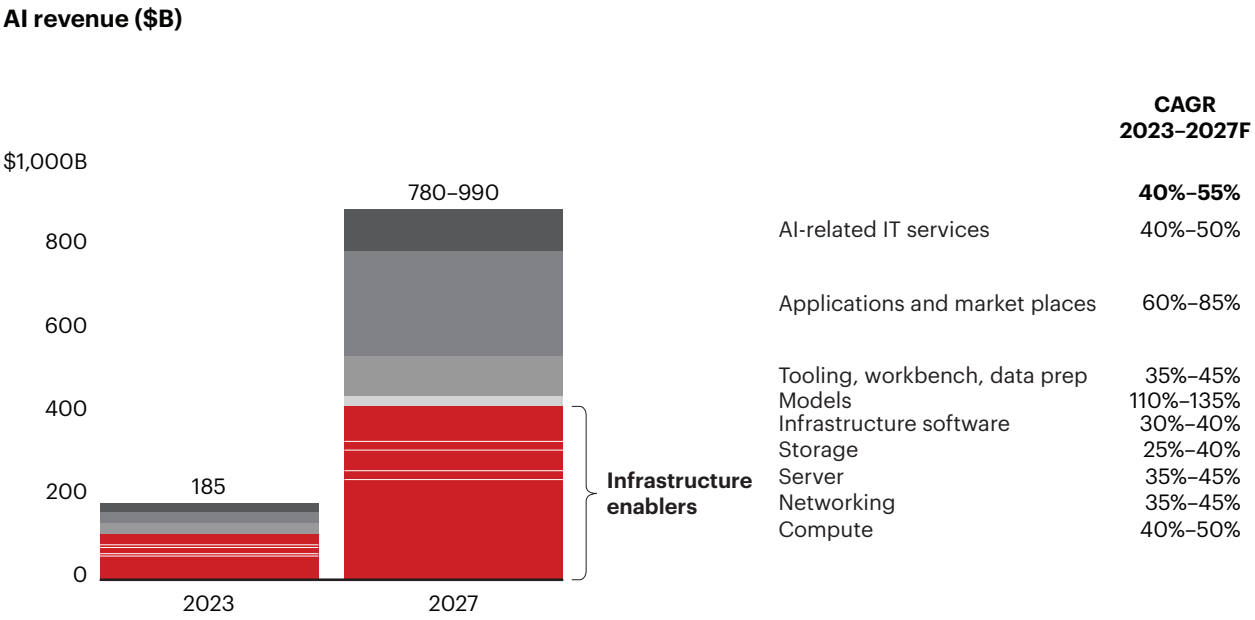
At a Glance

- The big cloud providers are the largest concentration of R&D, talent, and innovation today, pushing the boundaries of large models and advanced infrastructure.
- Innovation with smaller models (open-source and proprietary), edge infrastructure, and commercial software is reaching enterprises, sovereigns, and research institutions.
- Commercial software vendors are rapidly expanding their feature sets to provide the best use cases and leverage their data assets.

The pace of technological change has never been faster, and senior executives are looking to understand how these disruptions will reshape the sector. Generative AI is the prime mover of the current wave of change, but it is complicated by post-globalization shifts and the need to adapt business processes to deliver value.

Accelerated market growth. Nvidia's CEO, Jensen Huang, summed up the potential in the company's Q3 2024 earnings call: "Generative AI is the largest TAM [total addressable market] expansion of software and hardware that we've seen in several decades." Bain estimates that the total addressable market for AI-related hardware and software will grow between 40% and 55% annually for at least the next three years, reaching between $780 billion and $990 billion by 2027 *(see Figure 1)*. Fluctuations in

Figure 1: The AI market could reach $780 billion to $990 billion by 2027



Notes: AI defined as technology powered by neural networks/machine learning, excluding traditional business analytics and intelligence; compute category includes revenue from Nvidia and some server-like platforms using GPUs, leading to a category larger than servers; 2027 amounts are forecasts
Sources: IDC; Gartner; Bloomberg; Omdia; Morgan Stanley; BNP; market participant interviews; analyst reports; Bain & Company

supply and demand will create volatility along the way, but a long-term, durable trajectory seems like it is here to stay.

Three centers of innovation. So far, the largest cloud service providers (CSPs), or hyperscalers, have led the market in R&D spending, talent deployed, and innovation. They'll continue to lead but will look for more innovation from the next tier of CSPs, software-as-a-service providers, sovereigns, and enterprise as well as independent software vendors to fuel the next wave of growth.

- **High end: bigger models, better intelligence, more compute.** The big players will push ahead, developing larger and more powerful models and continuous gains in performance and intelligence. Their larger models will require more computational power, infrastructure, and energy, pushing the scale of data centers from today's high end (around 100 megawatts) to much larger data centers measured in gigawatts. This will strain the power grid and create readiness and resilience challenges in the supply chain for a wide spectrum of inputs, including graphics processing units (GPUs), substrates, silicon photonics, and power generation equipment and many others.

- **Enterprises and sovereigns: smaller models, RAG implementations, devices, tailored silicon.** Generative AI inference is set to become the killer app for edge computing as enterprises try to manage suppliers, protect data, and control total cost of ownership. Latency, security, and cost

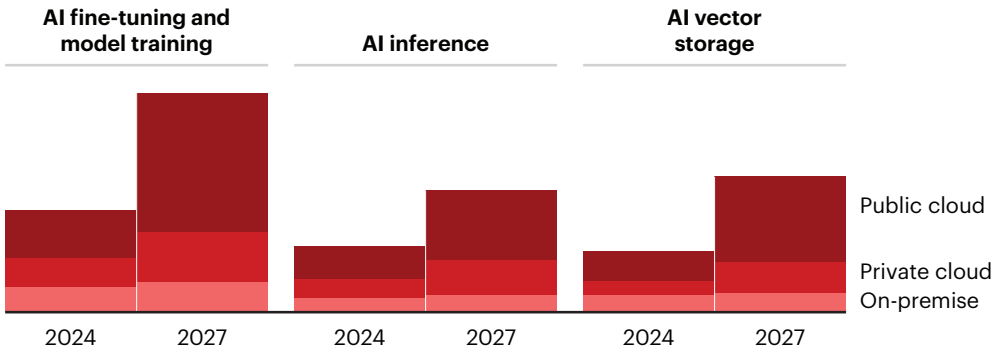
become increasingly relevant for inference workloads that need real-time processing and use owned data sets. Algorithms that use RAG (retrieval-augmented generation) and vector embeddings (numeric representations of data) handle a lot of the computing, networking, and storage tasks close to where the data is stored. This can reduce latency, lower costs, and keep data private and secure. Small language models that have been trained or tuned for a specific domain or task will become increasingly important in this context, as they can be less costly and more energy efficient to run than large general-purpose language models. The rapid growth of new models, both open-source (Meta's Llama, Mistral, TII's Falcon) and proprietary (Anthropic's Claude, Google AI's Gemini), is extending the range of cost- and energy-efficient options.

- **Independent software vendors (ISVs): racing to incorporate AI capabilities.** Large language model (LLM)-enabled software as a service is already providing AI-powered applications at Adobe, Microsoft, Salesforce, and many other companies. This will create a flood of new capabilities in the coming years, giving enterprises the option to deploy generative AI as part of their existing application suite rather than develop custom applications.

Disrupted industry structure with more verticalization. The AI workload is challenging and will continue to grow *(see Figure 2)*. The underlying matrix algebra and data-heavy computation strains parallelism, memory and system bandwidth, networking, infrastructure, and application software.

Figure 2: AI workloads could grow 25% to 35% per year through 2027

Infrastructure spending



Implied workload CAGR	25%–30%	20%–25%	30%–35%
Implied private cloud 2024–2027F CAGR	20%–25%	20%–25%	30%–35%
Implied on-premise 2024–2027F CAGR	5%–10%	5%–10%	0%–5%

Note: 2027 amounts are forecasts
Source: Bain & Company IT Workload Survey, May 2024 (n=283)

Technology vendors are responding by optimizing the technology stack vertically to deliver more efficiencies. For example, most hyperscalers have developed their own silicon for training and inference, like Amazon's Trainium and Graviton, Google's TPU, or Meta's MTIA. Nvidia has expanded its "unit of compute" beyond the GPU alone, now integrated with fabrics, hybrid memory, DGX, and cloud offerings. Nvidia is also enhancing its software stack and offering hosted services, providing tailored solutions that leverage its hardware and create a more efficient ecosystem for developers and users. Apple is developing its own on-device LLM and already has its own silicon.

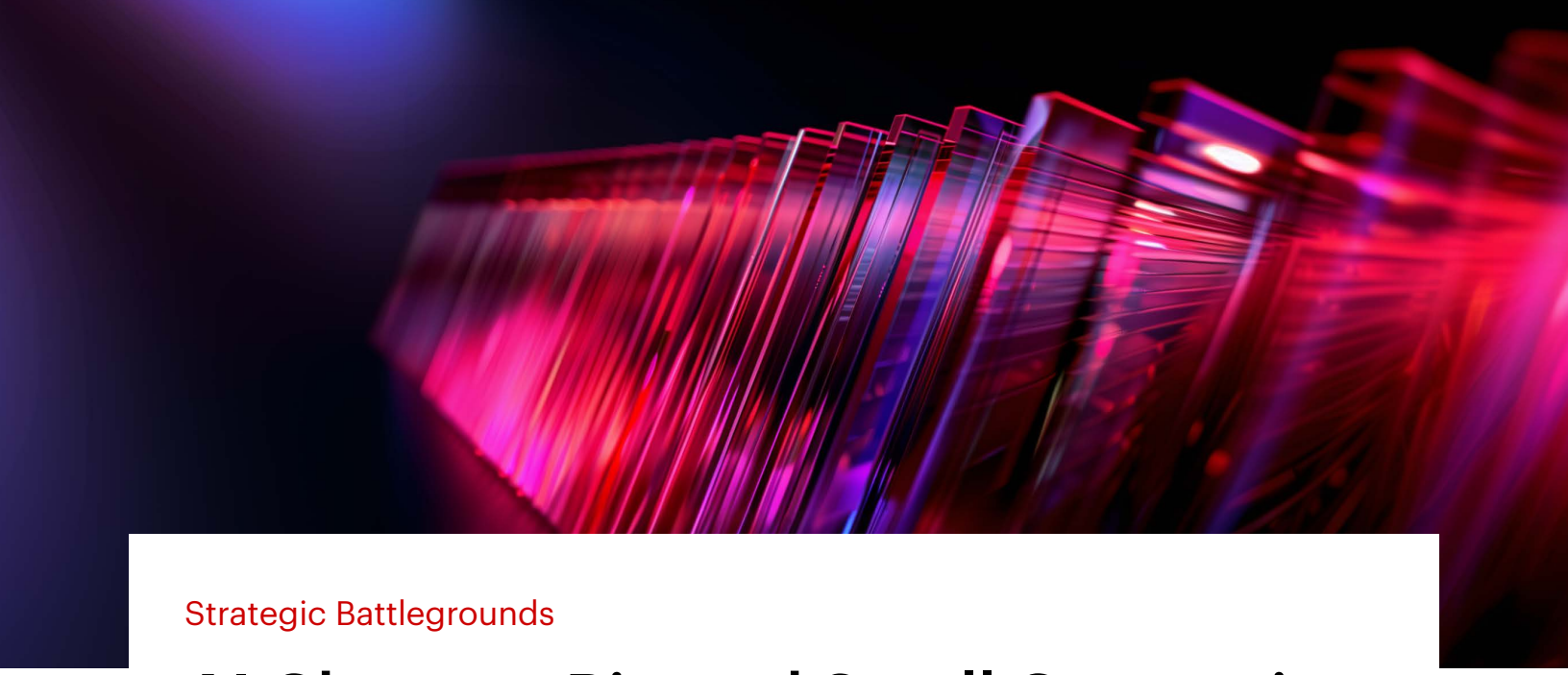
Generative AI is the prime mover of the current wave of change, but it is complicated by post-globalization shifts and the need to adapt business processes to deliver value.

Other segment-specific disruptions include:

- **Large language models (LLMs):** The underlying models are proliferating. OpenAI's ChatGPT held a near monopoly among production-grade generative AI solutions until 2023. Since then, the growth of open-source and proprietary models has improved to provide many more diverse options, including segmented versions of OpenAI's offerings.
- **Storage:** Storage technology will advance to accommodate the needs of generative AI, including accelerated consolidation of data siloes, increasing use of object vs. file and block storage, and selected upgrades to highly vectorized database capabilities.
- **Data management and virtualization:** The growing need for data preparation and mobility will spur growth in data management software. This will be particularly important as data-hungry AI apps mobilize data stored in public clouds with ingress and egress fees.
- **Tech services:** In the medium term, tech services will be in high demand while customers lack the skills and expertise needed for AI deployment and data modernization. Over time, significant portions of tech services themselves will be replaced by software. Clients in these domains are racing to design the new services to sustain their growth trajectories.

AI's disruptive growth will continue to reshape the tech sector, as innovation spreads beyond the hyperscalers (where it is centered today) to smaller CSPs, enterprises, sovereigns, software vendors,

and beyond. Bigger models will continue to push the boundaries, while smaller models will create new, more focused opportunities in specific verticals and domains. AI's workload demands will also spark innovation in storage, compute, memory, and data centers. As the market becomes more competitive and complex, companies will need to adapt rapidly to capture their share of this potential trillion-dollar market.



Strategic Battlegrounds

AI Changes Big and Small Computing

Data centers will get bigger, while more processing will move closer to the edge.

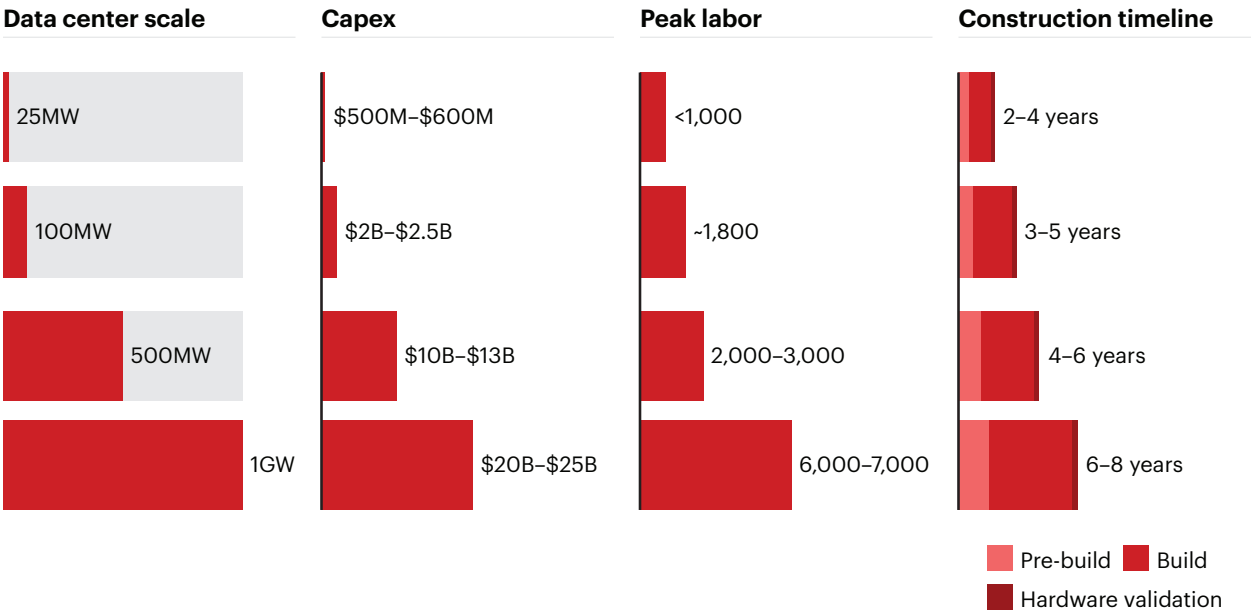
By Arjun Dutt, Paul Renno, and Velu Sinha

At a Glance

- AI's voracious appetite for computing power will spur growth in data centers, from today's 50–200 megawatts to more than a gigawatt.
- AI will also transform edge computing, as small, domain-specific language models will support tasks requiring lower latency.
- These changes will strain already-stressed supply chains as leaders vie for resources, especially labor and electricity.
- As data centers and edge computing evolve, enterprises may need to reassess market positions and revisit strategic ambitions.

AI's need for computing power will radically expand the scale of large data centers over the next five to 10 years. Today, big data centers run by hyperscale cloud service providers range from 50 megawatts to more than 200 megawatts. The massive loads demanded by AI will lead these companies to explore data centers in the 1 gigawatt and higher range. That will have huge implications on the ecosystems that support these centers (including infrastructure engineering, power production, and cooling), and affect market valuations. The architectural requirements for achieving the necessary computing, electrical power, and cooling density for gigawatt data centers will influence the design of many smaller data centers *(see Figure 1)*.

Figure 1: Data center requirements will rise significantly to meet AI's computing demands



Note: Peak labor is the maximum expected number of construction workers required at one time; capex includes servers and other data center equipment
Sources: Datacenter Dynamics; Top500; SemiAnalysis; company websites; industry interviews; Bain analysis

The ubiquity of AI will also change the nature of edge computing. Domain-specific language models—smaller, simpler, and optimized for specific purposes—will be necessary to handle computing loads that may require faster response, lower latency, or are able to use a simpler model due to a narrow focus. Innovation at the edge will extend to the form factor of user devices, which will also change to meet the needs of people engaging with AI.

The implications of these changes will be transformative across a number of critical dimensions, including speed of technology development, sector leadership, power generation and consumption, construction and industrial supply chains, environmental considerations, market economics, national security interests, and financing and investment. To remain in the top tier of the market, leaders will need to make unprecedented levels of investment in technology infrastructure. If large data centers currently cost between $1 billion and $4 billion, costs for data centers five years from now could be between $10 billion and $25 billion.

Strain on resources

The power demands and price tags of these large data centers will impose limits on how many can be built and how quickly. The scramble to acquire AI resources is already creating extreme competition

for resources at the high end of the market, and growing data center requirements will further strain capabilities.

Power consumption is one critical example. Utilities are already fielding requests from hyperscaler customers to significantly expand electrical capacity over the next five years. Their needs will compete with rising demand from electric vehicles and re-shoring of manufacturing, stressing the electric grid. Growth in electricity demand has been essentially flat for the last 15 to 20 years, but investments to expand and strengthen the grid and add new power sources (including on-site generation and renewables) will need to increase significantly.

Infrastructure providers and technology supply chains, including networking, memory, and storage, are also investing to meet the demands for high-performance compute from hyperscalers, digital service companies, and enterprises. Large data centers will push the limits and unleash innovation in physical design, advanced liquid cooling, silicon architecture, and highly efficient hardware and software co-design to support the rise of AI.

Large data centers are major construction efforts, requiring five years or more. Demand for construction and specialized laborers—as many as 6,000 to 7,000 workers at peak levels—will strain the labor pool. Labor shortages in electrical and cooling may be particularly acute. Many projects occurring at once will stress the entire supply chain, from laying cables to installing backup generators.

Innovation at the edge

As companies weigh the trade-offs between cloud and edge computing for AI, deciding where to handle inferencing is critical. One consideration is how closely to focus on specific domains and specific tasks, in order to use better curated and more focused data to build targeted models that reduce the compute infrastructure burden.

Another issue is how to move more computing power closer to the edge for AI in environments with low tolerance for latency, like autonomous driving. The rise of smaller models and specialized compute capable of running these models at the edge are important steps in this direction. Meanwhile, the industry is rapidly developing new form factors for the edge, including edge AI servers, AI PCs, robots, speakers, and wearables.

Preparations for expansion

The changing nature of data centers and edge computing increases the likelihood of AI reshuffling the technology sector and establishing a new order for the next era. Enterprises across the sector should be examining their market position and rethinking strategic ambitions to ensure they remain competitive in their chosen domains.

- **Cloud and data center service providers.** The overriding challenge for large players at this end of the market will be to find ways for their AI capabilities to meet the future demands of their

customers. Providers will need to decide what to deliver as a service and what to provide as enabling technologies at the industry level. Their efforts will also center on accelerating model development and working through the supply chain to construct large and distributed data centers. This will require the ability to refocus on compelling opportunities, build capabilities rapidly, and form partnerships that strengthen the platform.

Meta, for example, is competing with OpenAI, Alphabet, and others to secure a leadership role in large language models. To support these ambitions, Meta has massively increased the scale of its compute capacity over the past two years. Meta has also released Llama as an open-source language model, to serve as an enabler in the broader ecosystem.

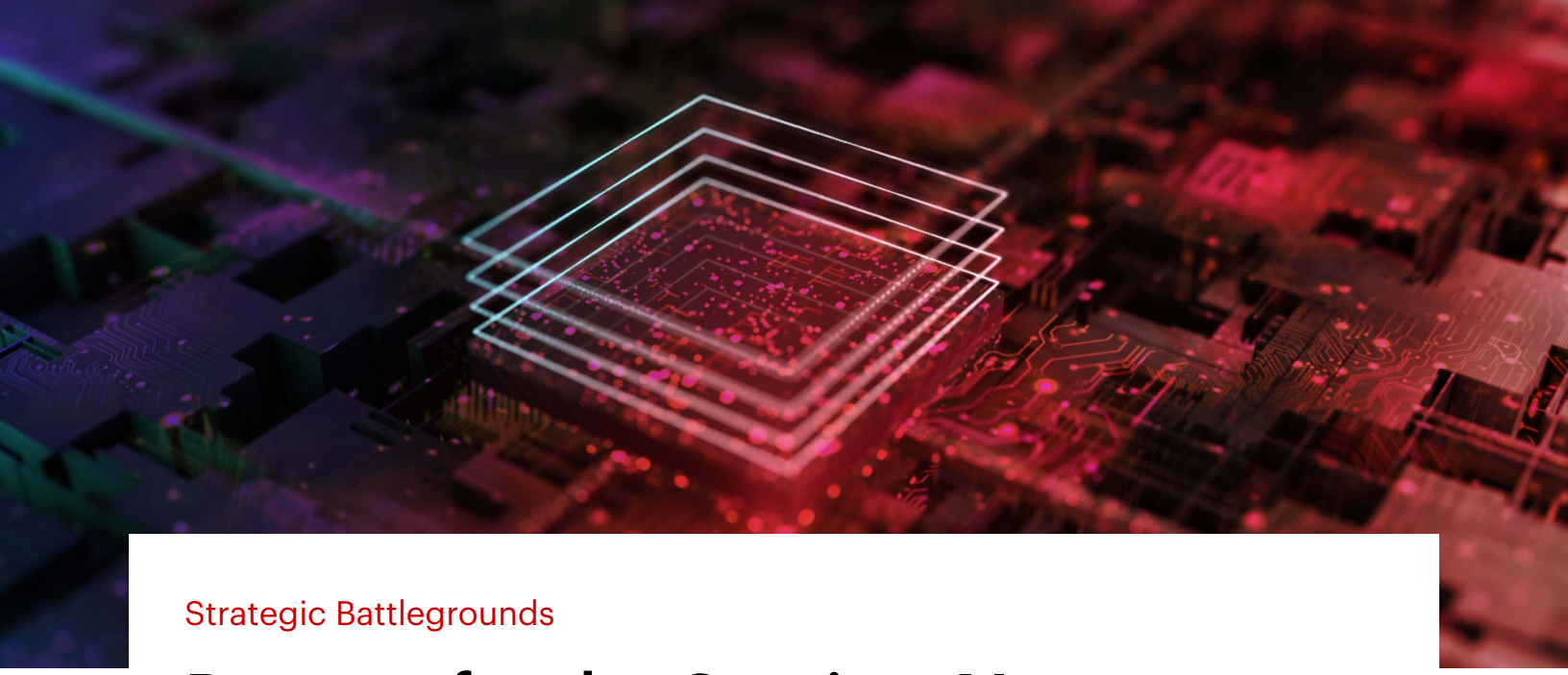
- **Infrastructure providers.** AI workloads require more specialization than prior generations of compute. Companies that design and manufacture servers, networks, storage, cooling, power, cabling, and all the other elements that go into building a data center will need to design their products to support AI. They will develop scale solutions to optimize compute and the performance of AI software. These companies also play significant roles in the delivery of infrastructure and services to customers. Accelerating the time to market of AI is an important opportunity for enterprises.

- **Software providers** will continue infusing AI into their core products to remain competitive. Increasingly, their business will need to focus on capturing and interpreting data insights while optimizing language models to deliver better (and faster) outcomes for customers. These aspects of their business will complement each other as software vendors build up their capabilities to augment the skills of their customers' workforce.

- **Edge device makers** will find ways to capitalize on innovation across the ecosystem, testing new form factors and interfaces, and using AI to increase personalization across devices. Sorting out users' privacy preferences will be critical to boosting adoption rates.

- **Data center supply chain providers** have a formative opportunity to reshape their roles in the market as mega centers proliferate and edge computing evolves. These players will focus on building capacity to scale and developing meaningful partnerships with engineering firms that can help meet the challenges of large data centers and more sophisticated edge computing.

As hyperscalers and other large companies plan for the large data centers necessary to accommodate AI's needs, additional factors will also require consideration. Paramount among these may be the investment requirements, as companies compete for funding of many massive projects at once. Stresses on the power grid are another area where companies have limited direct control. They may also have to manage the environmental implications of expanding data centers and electricity usage, including the effect on their carbon footprints and emission-reduction promises. The challenges are broad and complex, but as the global race to win in AI heats up, no company in this ecosystem can afford to stand by and wait; the time to act is now.



Strategic Battlegrounds

Prepare for the Coming AI Chip Shortage

While businesses couldn't predict the pandemic, they can guard against the next big threat to semiconductor supply chains.

By Peter Hanbury, Anne Hoecker, and Michael Schallehn

At a Glance

- The AI-driven surge in demand for graphics processing units alone could increase total demand for certain upstream components by 30% or more by 2026.
- Just as the pandemic created a surge in PC demand, a coming wave of AI-enabled devices will likely accelerate smartphone and PC upgrade purchases.
- These two trends, along with continued geopolitical tensions and other supply risks, could trigger the next semiconductor shortage.
- Proactive measures, including long-term purchase agreements and supply chain diversification, will be critical to mitigating looming risks.

The supply and demand of semiconductors is a delicate balance that can be quickly shaken, as the industry and its customers know all too well after the past few years. Although the pandemic-induced

chip shortage has passed, executives are starting to prepare for the next potential crunch caused by (you guessed it) artificial intelligence.

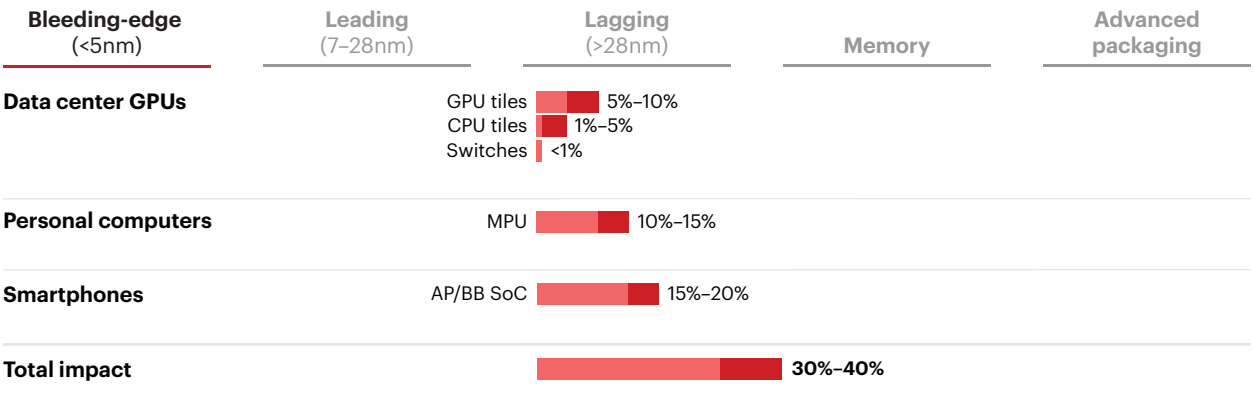
Accelerating adoption of AI across industries will pressure the supply of graphics processing units (GPUs) for data centers, as a seemingly insatiable demand for computing resources to train and operate large language models (LLMs) collides with supply chain constraints. In addition, the coming proliferation of AI-enabled devices appears poised to jumpstart a wave of purchases of new personal computers (PCs) and smartphones, which has major implications for the broader semiconductor supply chain.

The semiconductor supply chain is incredibly complex, and a demand increase of about 20% or more has a high likelihood of upsetting the equilibrium and causing a chip shortage. The AI explosion across the confluence of the large end markets could easily surpass that threshold, creating vulnerable chokepoints throughout the supply chain *(see Figure 1)*.

Balancing semiconductor supply and demand has always been difficult given the industry's fast-moving technologies, large capital requirements, and long lead times to add production capacity. But chip suppliers and buyers must act quickly to get ahead of this next, potentially massive crunch. Let's unpack how things could play out across potential demand and supply shocks.

Figure 1a: Surging demand for AI computing power will strain the supply chains for data center chips, personal computers, and smartphones

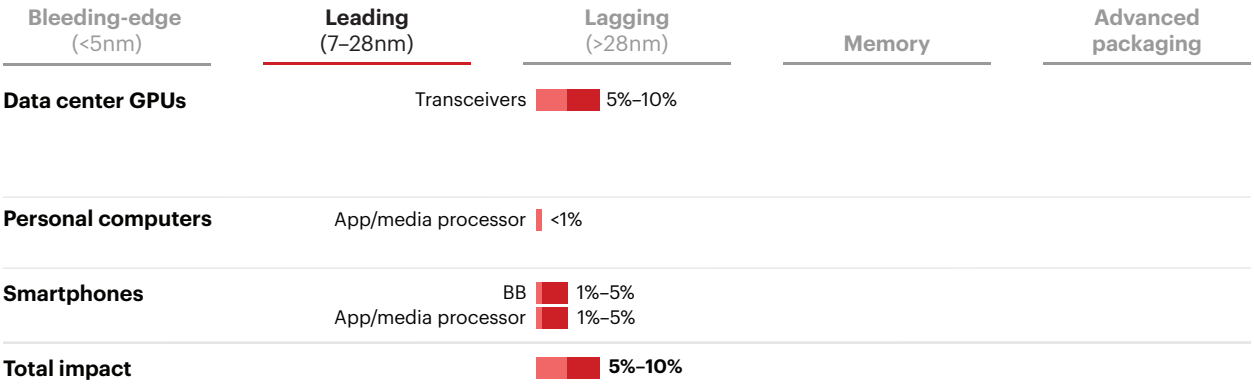
Projected percentage demand increase by 2026 in rapid AI adoption scenario



Notes: Data center projections based on GPU demand doubling from sales of 1.5 million H100 units in 2023 to 3 million GB200s in 2026; PC and smartphone projections based on 31% and 15% total unit sales growth, respectively, from 2023 to 2026; total impact values are the sum of each category's component growth projections divided by the sum of the component market sizes within the category; switches, transceivers, and interposer categories excluded from respective total impact values as the components were calculated using a different methodology; HDD (hard disk drive) is measured in gigabytes consumed as opposed to units; GPU is graphics processing unit; DRAM is dynamic random access memory; HBM is high-bandwidth memory; SSD is solid state drive; CoWoS is chip-on-wafer-on-substrate; BGA is ball grid array; MPU is microprocessing unit; RF is radio frequency; PMIC is power management integrated circuits; AP/BB SoC is application processor/baseband system-on-a-chip
Sources: IDC; Gartner; analyst reports; Bain semiconductor market forecasting model

Figure 1b: Surging demand for AI computing power will strain the supply chains for data center chips, personal computers, and smartphones

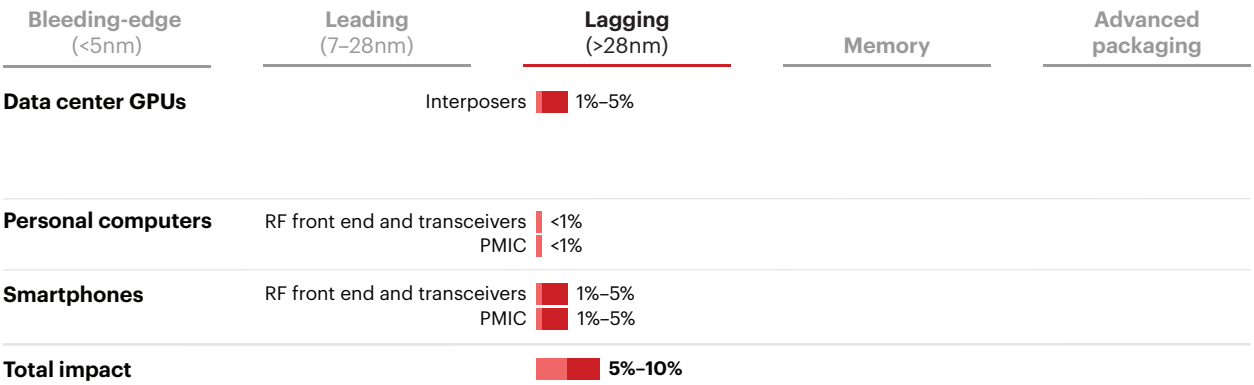
Projected percentage demand increase by 2026 in rapid AI adoption scenario



Notes: Data center projections based on GPU demand doubling from sales of 1.5 million H100 units in 2023 to 3 million GB200s in 2026; PC and smartphone projections based on 31% and 15% total unit sales growth, respectively, from 2023 to 2026; total impact values are the sum of each category's component growth projections divided by the sum of the component market sizes within the category; switches, transceivers, and interposer categories excluded from respective total impact values as the components were calculated using a different methodology; HDD (hard disk drive) is measured in gigabytes consumed as opposed to units; GPU is graphics processing unit; DRAM is dynamic random access memory; HBM is high-bandwidth memory; SSD is solid state drive; CoWoS is chip-on-wafer-on-substrate; BGA is ball grid array; MPU is microprocessing unit; RF is radio frequency; PMIC is power management integrated circuits; AP/BB SoC is application processor/baseband system-on-a-chip
Sources: IDC; Gartner; analyst reports; Bain semiconductor market forecasting model

Figure 1c: Surging demand for AI computing power will strain the supply chains for data center chips, personal computers, and smartphones

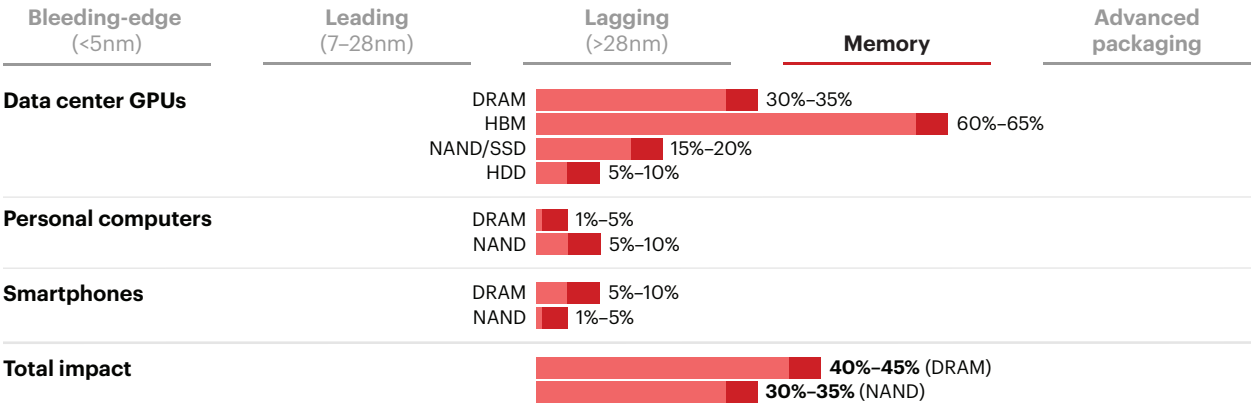
Projected percentage demand increase by 2026 in rapid AI adoption scenario



Notes: Data center projections based on GPU demand doubling from sales of 1.5 million H100 units in 2023 to 3 million GB200s in 2026; PC and smartphone projections based on 31% and 15% total unit sales growth, respectively, from 2023 to 2026; total impact values are the sum of each category's component growth projections divided by the sum of the component market sizes within the category; switches, transceivers, and interposer categories excluded from respective total impact values as the components were calculated using a different methodology; HDD (hard disk drive) is measured in gigabytes consumed as opposed to units; GPU is graphics processing unit; DRAM is dynamic random access memory; HBM is high-bandwidth memory; SSD is solid state drive; CoWoS is chip-on-wafer-on-substrate; BGA is ball grid array; MPU is microprocessing unit; RF is radio frequency; PMIC is power management integrated circuits; AP/BB SoC is application processor/baseband system-on-a-chip
Sources: IDC; Gartner; analyst reports; Bain semiconductor market forecasting model

Figure 1d: Surging demand for AI computing power will strain the supply chains for data center chips, personal computers, and smartphones

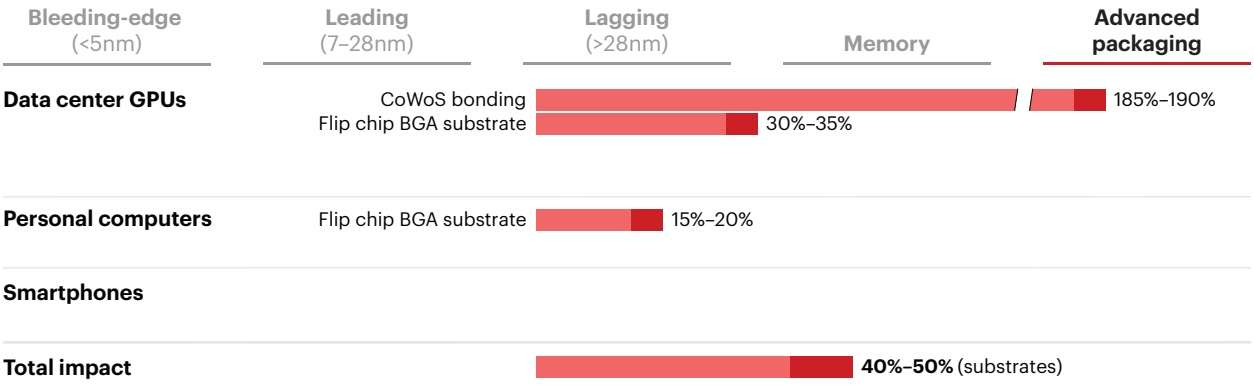
Projected percentage demand increase by 2026 in rapid AI adoption scenario



Notes: Data center projections based on GPU demand doubling from sales of 1.5 million H100 units in 2023 to 3 million GB200s in 2026; PC and smartphone projections based on 31% and 15% total unit sales growth, respectively, from 2023 to 2026; total impact values are the sum of each category's component growth projections divided by the sum of the component market sizes within the category; switches, transceivers, and interposer categories excluded from respective total impact values as the components were calculated using a different methodology; HDD (hard disk drive) is measured in gigabytes consumed as opposed to units; GPU is graphics processing unit; DRAM is dynamic random access memory; HBM is high-bandwidth memory; SSD is solid state drive; CoWoS is chip-on-wafer-on-substrate; BGA is ball grid array; MPU is microprocessing unit; RF is radio frequency; PMIC is power management integrated circuits; AP/BB SoC is application processor/baseband system-on-a-chip
Sources: IDC; Gartner; analyst reports; Bain semiconductor market forecasting model

Figure 1e: Surging demand for AI computing power will strain the supply chains for data center chips, personal computers, and smartphones

Projected percentage demand increase by 2026 in rapid AI adoption scenario



Notes: Data center projections based on GPU demand doubling from sales of 1.5 million H100 units in 2023 to 3 million GB200s in 2026; PC and smartphone projections based on 31% and 15% total unit sales growth, respectively, from 2023 to 2026; total impact values are the sum of each category's component growth projections divided by the sum of the component market sizes within the category; switches, transceivers, and interposer categories excluded from respective total impact values as the components were calculated using a different methodology; HDD (hard disk drive) is measured in gigabytes consumed as opposed to units; GPU is graphics processing unit; DRAM is dynamic random access memory; HBM is high-bandwidth memory; SSD is solid state drive; CoWoS is chip-on-wafer-on-substrate; BGA is ball grid array; MPU is microprocessing unit; RF is radio frequency; PMIC is power management integrated circuits; AP/BB SoC is application processor/baseband system-on-a-chip
Sources: IDC; Gartner; analyst reports; Bain semiconductor market forecasting model

Data center demand

Generative AI's breakthrough in late 2022 has so far been a boon for the semiconductor industry. The sales and valuation of chipmakers have grown enormously, from leading GPU sellers such as Nvidia to vendors who supply other chips into data centers, including Broadcom (switches) and SK Hynix (high-bandwidth memory). Spending on data centers and the specialized chips that power them shows no signs of slowing. Major cloud service providers are expected to increase their year-over-year capital spending by 36% in 2024, spurred in large part by investments in AI and accelerated computing. GPU demand will continue to grow as LLMs expand capabilities to processing multiple data types simultaneously (text, images, and audio) and as venture capitalists pour even more money into AI start-ups.

If data center demand for current-generation GPUs doubled by 2026—a reasonable assumption given current trajectory—suppliers of key components would need to increase their output by 30% or more in some cases, based on Bain's forecasting model that accounts for the intricacies of the multi-level semiconductor supply chain *(see Figure 1 previous page)*. This pull-through demand will be concentrated in advanced packaging and memory. In the scenario above, makers of chip-on-wafer-on-substrate (CoWoS) packaging components would need to almost triple production capacity by 2026.

To enable AI growth, a complex web of supply chain elements must come together, from constructing data centers and wafer fabs to securing access to advanced packaging and sufficient electricity. Obtaining many of these crucial elements involves long lead times that may make it impossible to keep up with demand *(see Figure 2)*.

Importantly, many of these supply chain elements are shared with other parts of the technology ecosystem, and they're all subject to capital, geopolitical, and timing risks. One missing chip could derail the entire system, like in the last shortage when new cars sat unsold in lots because they lacked a critical chip.

PC and smartphone demand

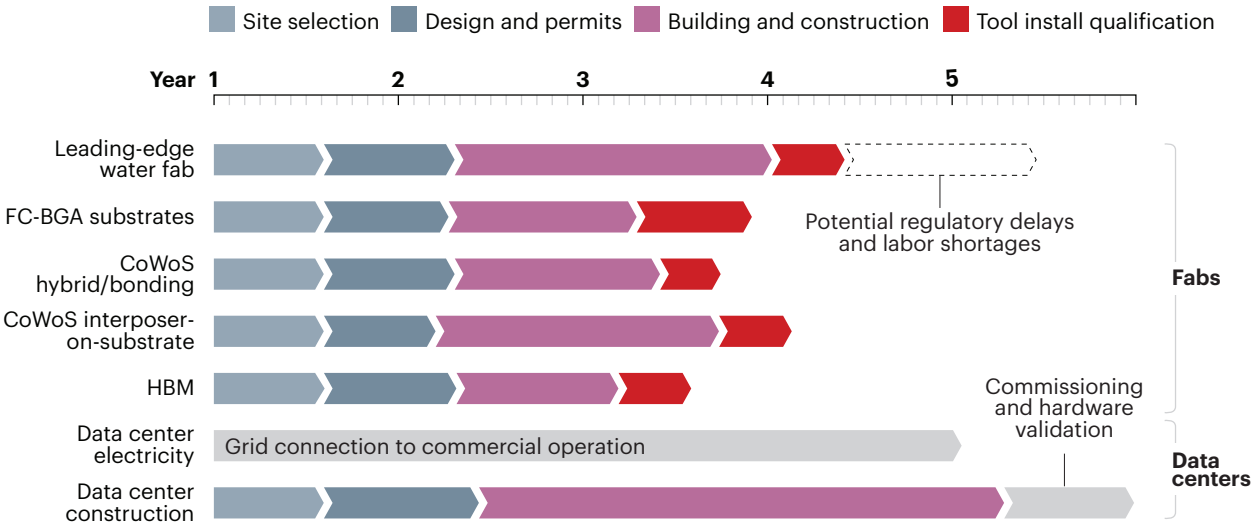
Personal device makers are already rapidly embedding AI capabilities directly into their products. To accommodate neural processing engines for on-device AI, the average notebook core processing unit (CPU) and smartphone processor have respectively added about 5% and 16% more silicon surface area, according to our benchmarking.

More importantly, as AI applications grow in usefulness, buyers looking to upgrade could accelerate their new device purchases, causing an uptick in demand similar to how the pandemic spurred a short-term surge in PC demand *(see Figure 3)*.

Compared with GPUs, AI demand will have a wider effect on semiconductor supply chains for smartphones and PCs given the long list of components associated with these devices. The most vulnerable link in these devices' supply chains will be bleeding-edge fabs that manufacture the most

Figure 2: A complex set of components, resources, and services must come together to meet demand for AI computing power

Development time from project launch to fully operational

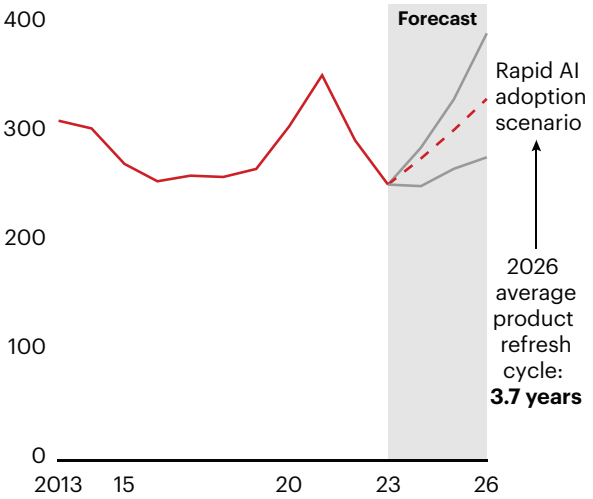


Notes: Estimated time for regulatory delays and labor shortages only applicable to expansions into new geographies; data center estimates based on 100-megawatt facility; FC-BGA is flip chip-ball grid array; CoWoS is chip-on-wafer-on-substrate; HBM is high-bandwidth memory
Sources: News reports; company websites; Bain analysis

Figure 3: AI could sharply accelerate PC and smartphone sales, similar to how the pandemic created a surge in PC demand

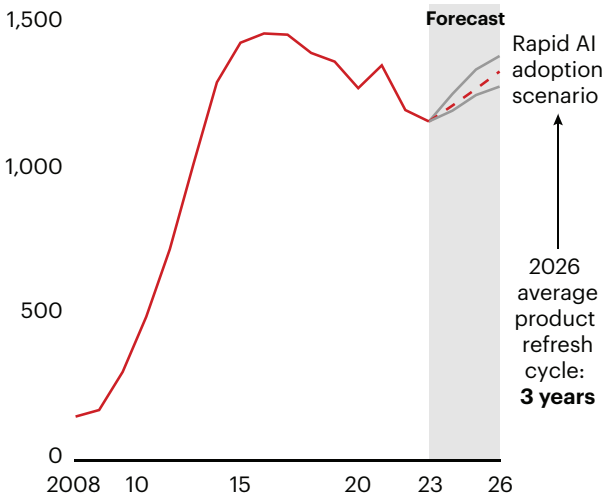
PC shipments expected to grow about 30% over three years

PC shipments worldwide, 2013–2026 (M units)



Smartphones expected to grow about 15% over three years

Smartphone shipments worldwide, 2008–2026 (M units)



Note: Average product refresh cycle is an estimate of how long the average buyer takes to upgrade or replace a device
Sources: Analyst reports; Bain analysis

advanced chips. In a rapid AI adoption scenario that spurs 31% PC sales growth and 15% smartphone growth between 2023 and 2026, bleeding-edge fabs would need to raise output by an estimated 25% to 35%. This would require building four or five additional bleeding-edge fabs costing an estimated $40 billion to $75 billion, which would help justify the many fabs that major foundries are already building.

Don’t forget the supply risks

Extreme weather, natural disasters, geopolitical strife, a pandemic, and other major disruptions over the past decade have made abundantly clear how supply shocks can severely limit the industry’s ability to meet demand. Much of the pressure on GPU supply over the last 18 months was caused by disruptions to less visible elements of the supply chain, such as CoWoS advanced packaging capabilities.

Extreme weather, natural disasters, geopolitical strife, a pandemic, and other major disruptions over the past decade have made abundantly clear how supply shocks can severely limit the industry’s ability to meet demand.

Geopolitical tensions, trade restrictions, and multinational tech companies’ decoupling of their supply chains from China continue to pose serious risks to semiconductor supply. Delays in factory construction, materials shortages, and other unpredictable factors could also create pinch points. Without accounting for these uncertainties, we expect the largest supply risks to come from larger demand for high-bandwidth memory components, advanced packaging fab and tool construction, and substrate fab construction.

Takeaways for executives

For semiconductor buyers across industries, navigating these supply chain intricacies starts with a deep understanding of the components being sourced. Effective leaders will pay extra attention to components that intersect with AI data centers, such as switches, transceivers, and power management integrated circuits. They’ll closely monitor PC and smartphone refresh cycles, as well as related peripherals like Wi-Fi routers and network equipment. A surge in these areas will have cascading effects across the supply chain that must each be closely tracked.

Leading companies will apply lessons from the most recent chip crunch to keep their inventories safely balanced between shortages and gluts. They’ll sign long-term purchase agreements to secure

access to chips and manufacturing capacity based on anticipated future needs (and they'll share this visibility with their suppliers). The "just-in-time" inventory strategy that dominated the past several decades will continue giving way to a "just-in-case" approach that's higher cost but more resilient. More companies will design products to use industry-standard semiconductors where possible instead of application-specific chips. They'll also continue to invest in supply chain resilience against geopolitical uncertainties such as tariffs or regulations. Lastly, they'll monitor silicon advanced packaging and substrate supply as closely as they do front-end semiconductor manufacturing capacity.

Executives may still feel weary from the semiconductor supply disruptions spurred by the pandemic, but there's no time to rest because the next big supply shock looms. This time, however, the signs are clear, and the industry has a chance to prepare. The path forward demands vigilance, strategic foresight, and swift action to reinforce supply chains. With proactive measures, business leaders can ensure their resilience and success in an increasingly AI-enabled world.

Strategic Battlegrounds

Thriving as the Software Cycle Slows

After a period of strong investment, software vendors can reset with a disciplined portfolio strategy.

By Simon Heap, Greg Fiore, Greg Callahan, Dan Levy, and Jay Bhatnagar

At a Glance

- Growth has slowed in the software market, so software companies must be more deliberate in their product portfolio strategy.
- Software companies have cut spending on sales and marketing, but spending on product and engineering has been more resilient.
- Efficiency in product and engineering is critical in order to free up capacity for investment in important products and features.

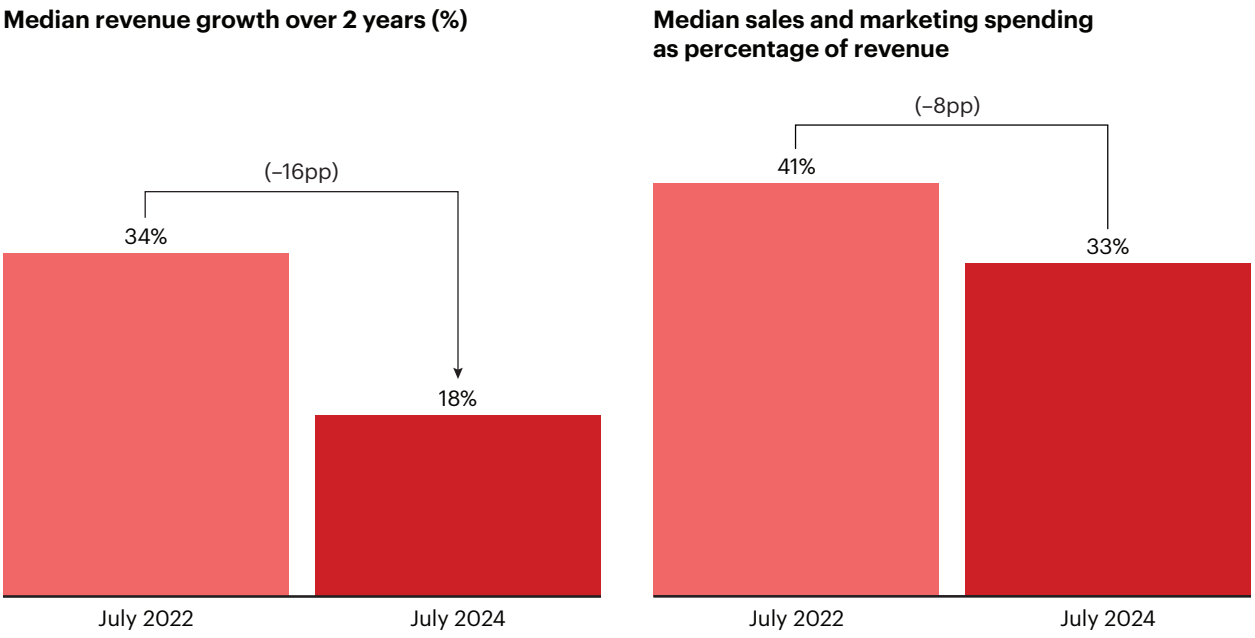
Since 2021, software companies have been on a spending spree. Flush with cash from investors in a low-interest-rate environment and motivated by the rich budgets of their customers, software companies made huge investments in research and development and sales and marketing. They believed their engineering teams could add endless features to their products and enter adjacent markets that their sales teams and product-led growth initiatives could easily sell into.

That is, until recently. Over the past year, while tech budgets remain healthy, CIOs are now much more disciplined in how they buy. Purchases are put under a microscope with competitive requests for proposals, and buying processes have lengthened. Companies are reducing the number of seats (or software licenses) based on their actual need and consolidating spending to strategic vendors. Employees find they need to make a business case to IT to justify buying a standalone product instead of one that comes bundled in another package. The one bright spot has been AI, where companies have been willing to spend aggressively. The result has been a deceleration of growth: We saw a 16-percentage-point decline in the median annual revenue growth for a group of about 90 publicly traded software-as-a-service (SaaS) companies over the past two years *(see Figure 1)*.

Consequently, software companies have tightened their own budgets. Sales and marketing budgets have shrunk from 41% of revenue to 33% of revenue. Despite budget pressure and the promise of generative AI co-pilots for developers, spending on engineering has been much more resilient: Spending on research and development has only declined 3 percentage points as a percentage of revenue *(see Figure 2)*.

CEOs and CFOs often lack visibility on spending. They may not know how much is being spent on innovation and new product development compared to the costs of maintenance for existing products. Often, they can't see when work is being duplicated, which can add up to a surprisingly large amount of spending.

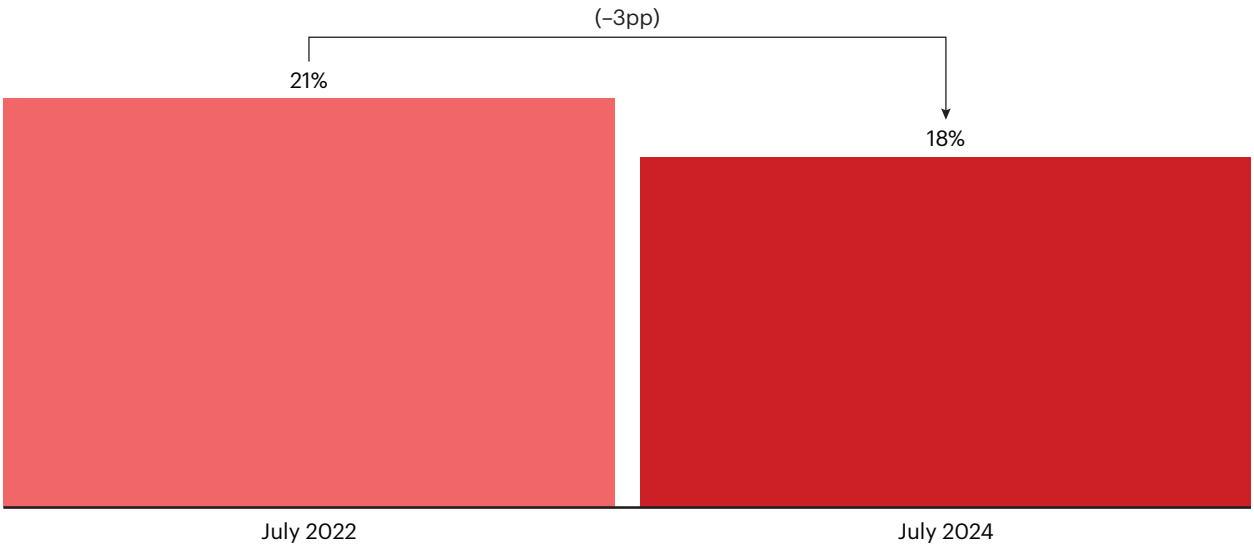
Figure 1: As growth slowed, SaaS companies significantly scaled back spending on sales and marketing



Note: Analysis based on about 90 public SaaS companies
Sources: Meritech Capital; Bain analysis

Figure 2: Spending on research and development has proved more robust

Median R&D spending as percentage of revenue



Note: Analysis based on about 90 public SaaS companies
Sources: Meritech Capital; Bain analysis

Disciplined portfolio strategy

Customers are not likely to return to an era of ambitious investment beyond AI anytime soon, so software companies will need to ensure they're producing what customers need, make the most of their research and development spend, and rein in operating expenses that may have inflated beyond optimal ranges.

Software vendors will need to become more disciplined in deciding what to build and sell, and be clearer about which product strategy they are pursuing. Our benchmarks indicate that companies typically spend about 25% of engineering resources on fixing defects in existing products, 25% on maintenance and technical debt remediation, and 50% on new features and new products. Of course, this ratio should vary based on product maturity.

To better allocate resources in pursuit of this optimal mix, leading companies follow a disciplined product portfolio strategy.

- Set the strategic vision for the business, defining which verticals and customer segments to focus on and with what products and sales motions. This requires a clear view of the size of the addressable market, the customers' needs, and the product's competitive edge. When targeting adjacent markets, it's critical to identify synergies with the core product and the extent to which buying cycles coincide.

- Develop a business case for each product initiative, clearly articulating the roadmap, resources required, and expected return on investment.
- Evaluate progress periodically and test product market fit.
- Coordinate between product, engineering, and go-to-market teams to align on the product strategy and the plan to create value.
- Take a hard line on end-of-service and end-of-life policies for products that are being deprecated and reinvest those resources in more productive efforts.

Efficient R&D

Dismantling the silos between product, research and development, sales and marketing, and customer success and support functions can help improve the efficiency of operations. An integrated operating model helps ensure that the right products are built (that is, products that fit well into the market) and are built right (supported by the most efficient architecture, development, and release programs). For example, selling directly to customers allows frontline staff to gather feedback, which can help continuous development of a product according to buyers' needs. In other cases, vendors will allow customers or customer-facing functions like sales or customer success to vote on the product roadmap.

Organizational structure also plays a critical role in improving efficiency. Getting the right ratios of managers to developers and other staff helps maintain the right levels of experience. Outsourcing and offshoring are also key factors in efficiency, although some vendors paid less attention to these during the last boom period. Software organizations can thrive by offshoring work on products in maintenance mode to low-cost geographies while investing more resources in development of new products closer to the center.

Valuation rewards

The software market is now in a slower growth part of the cycle, although the generative AI "gold rush" is distorting the maturation of underlying products. Productive growth and margin delivery are what matters now to drive valuation. These require a much more careful balance of growth investment and cost management than software companies have exercised in recent years. Product strategy, including AI, and product portfolio spending provide the foundation for achieving this balance. Investing in AI technologies doesn't preclude careful assessment of the rest of the portfolio.

These decisions then set the direction for how to efficiently deploy research and development spending and how to achieve productivity in sales, marketing, and services. Managing overhead costs efficiently helps free up budget dollars for investment in other areas. This is a difficult balancing act, but the leaders who are able to perform it may yet see the valuation rewards of previous highs.

Strategic Battlegrounds

How Generative AI Changes the Game in Tech Services

Tech service providers are using generative AI to operate and deliver better; leaders use this technology to help customers reinvent and innovate.

By Saikat Banerjee and Sandeep Nayak

At a Glance

- Companies are looking to their tech service providers to help them learn about and adopt generative artificial intelligence (AI).
- Providers are rising to the challenge, building generative AI capabilities so that they can deliver faster, more productively, and at higher-quality levels.
- Rather than looking at generative AI opportunities by individual use cases, leaders view them in families—that is, a set of use cases focused on a contiguous group of processes.
- Demonstrating expertise, hiring and upskilling talent to build AI capabilities, and developing and deploying solutions at the leading edge of innovation all signal the market about one's comfort with the technology.

In 2023, many technology service companies supported their clients' interest in generative artificial intelligence (AI) through proofs of concept aimed at reducing operational costs, completing tasks faster, or improving quality. These pilot programs focused on a variety of topics, including AI-enabled assistants for internal knowledge portal search, high-frequency marketing content generation, sales

collateral development, and knowledge interfaces supporting customer-facing agents and enhancing productivity and conversion. Some tech service companies executed dozens or hundreds of these pilots for clients, some with ticket costs up to $1 million.

In 2024, these clients are moving beyond the exploration stage, investing to scale up successful pilot programs. The focus this year appears to be on reaping the benefits of those pilots and demonstrating real business value from investments in AI. Bain's latest global survey on generative AI adoption found that generative AI is a top five priority for 85% of respondents. The percentage of companies planning to spend more than $5 million on generative AI is expected to rise from less than 20% in 2023 to 33% in 2024. Another one-third of companies said that they will spend between $1 million and $5 million on generative AI experiments, up from 15% in 2023.

Operate better: Signal expertise

Across industries, most companies expect their tech service company partners to play a key role in these efforts, particularly if these providers have already developed expertise using generative AI internally (to improve their own operations) or in how they deliver services—faster, more productively, and with higher quality.

Among the ways tech service companies are using generative AI:

- Some IT service companies use AI to customize sales collateral and their responses to requests for proposals by stitching together critical technology capabilities and success stories from existing repositories. This can speed up the process and showcase applicable strengths and successes most relevant to particular customers.
- Several tech service players have developed internal chatbots to support frontline human resources and IT queries through a conversational bot. These are designed to help employees at customer firms access information from data across the enterprise.
- Another company is using AI to improve its knowledge management and training by ingesting existing training materials and internal support chats to create something like digital twins of internal expert trainers.

Deliver better: Improve service

Examples of tech service companies using generative AI to improve delivery include:

- code generation, documentation, and testing, with productivity gains of up to 30%;
- process outsourcing, in which service providers working with their clients are reimagining processes, combining generative AI with automation to enhance productivity, customer satisfaction, and accuracy of solutions; and

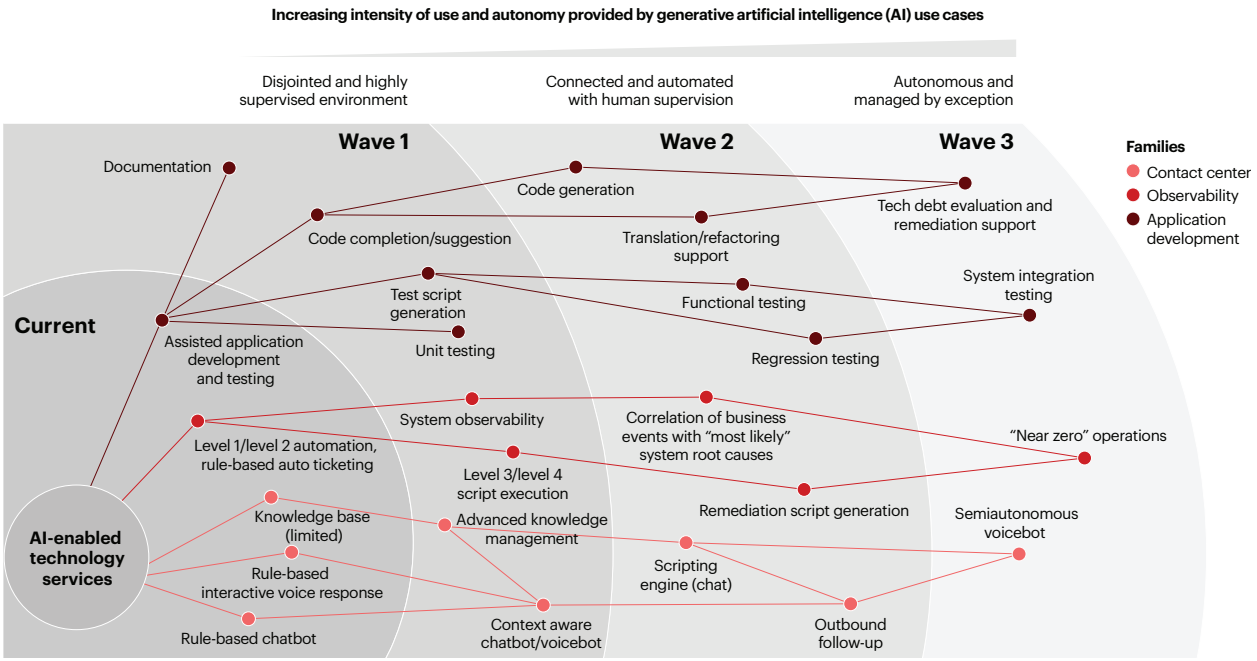
- invoice processing, in which one large, global outsourcing company has demonstrated how to cut processing time in half with redesigned processes, generative AI, and automation.

The contributions of generative AI are likely to roll out in waves, each building on the previous accomplishments *(see Figure 1)*. Consider contact centers: Over the next two to three years, half of all nonvoice interactions, 25% to 35% of simple voice interactions, and about 10% of complex voice interactions could be replaced with generative AI.

Contact centers had already developed rule-based response engines through chatbots. However, traditional chatbots have limited capabilities when responding to customer-specific questions or providing customer-specific recommendations, often simply offering a link to more information or giving them a callback option.

Generative AI's ability to assimilate and summarize large volumes of unstructured data creates a sharper knowledge management function. This should enable chatbots and chat assistants to provide more context-aware responses to customer queries faster, resulting in happier customers and more productive agents in the first wave of generative AI deployment. By the second wave, generative AI could help develop automated scripts for outbound calling, and in the third wave, we could see semiautonomous voicebots—both enhancements of the first wave rather than ground-up builds.

Figure 1: AI's contributions to delivery improvements could play out in waves



Source: Bain & Company

Change the game: Move beyond standalone use cases

Tech service companies that want to design, build, and deploy generative AI solutions customized to clients' needs can differentiate their approach by thinking in terms of collections of use cases—what we call use case families. A family is a set of use cases focused on a group of contiguous processes that need to work in tandem or in sequence and that require similar foundational technologies.

Use case families can be horizontal, with nuanced variations for different industries, or at the intersection of industries, with processes and geographies. Below are some examples.

- **Customer relationship management:** Summarizing customer meetings and generating customer outreach collateral, including recommendations and social media listening, are examples of a typical horizontal family.
- **Outbound marketing automation:** Micro-segmentation, persona development, best-offer prediction (based on usage and purchase knowledge base), outreach content generation, and automated outreach triggers are other horizontal examples of use case families. However, overlaying nuances related to business-to-business (persona-focused) or business-to-consumer (segment-focused) marketing adds a vertical flavor to this type of family.
- **Industry-specific families:** In financial services, for example, families focused on loan and mortgage origination might include automated preapproval, application completion, underwriting, processing, disbursal, and closure.

Tech service companies can differentiate themselves by choosing and delivering use case families that combine their domain expertise, process knowledge, and technical prowess to transform customer engagement modules.

Tech service companies can differentiate themselves by choosing and delivering use case families that combine their domain expertise, process knowledge, and technical prowess to transform customer engagement modules.

Families are a better approach because standalone use cases may not address all the sources of value that can be tapped using the same technology investment and enablement. Also, standalone use cases focus on internal productivity improvement while families of synchronized use cases deliver lasting customer value, which can lead to repeat buying, greater sales of the product or service, and new referrals.

Figure 2: Back-office automation, application development, RunOps, marketing automation, and contact centers rank highly in demand across industries

What are the top three use case families your company will focus on for generative artificial intelligence (AI)?

Generative AI use case family	Productivity increase	Wealth management	Commercial banking	Property & casualty insurance	Retail banking	Pharma	Discrete manufacturing	Retail	Telecom
Deliver better									
Application development	10%–20%	75%	63%	47%	68%	53%	60%	40%	47%
Testing automation	<10%	20%	37%	37%	63%	32%	60%	35%	53%
RunOps	10%–20%	75%	68%	37%	47%	63%	20%	15%	21%
Contact center automation	10%–20%	10%	37%	53%	37%	42%	40%	60%	63%
Change the game									
Back-office automation	10%–20%	75%	53%	68%	32%	63%	75%	70%	53%
Marketing automation	10%–20%	40%	42%	58%	53%	42%	45%	80%	63%

Note: RunOps includes the running and maintenance of IT systems
Source: Bain ITeS Survey (total N=155; n=between 19 and 21 for each vertical, November 2023)

Bain's research finds that customers are looking to tech service companies to address a few priority families, some that promise efficient delivery and others that are more likely to change the game *(see Figure 2)*.

Scaling generative AI

In our work, we find that clients are eager to engage with tech service players that have already begun to build their own AI skills and capabilities. Demonstrating expertise, hiring and upskilling talent to build AI capabilities, and developing and deploying solutions at the leading edge are great signals to the market about one's capabilities and comfort with the technology.

Five additional factors will also differentiate tech service players in the race to become preferred partners in generative AI solution deployment:

- a deep understanding of clients' domain and processes, as well as a proven track record of innovating and reengineering these through new technological advancements—this has been true for every major technology paradigm change, such as the advent of robotic process automation and public cloud adoption;
- an ability to articulate, prioritize, and sequence use cases to seamlessly realize near- and longer-term value;

- familiarity with technology options (including models, vector databases, and development frameworks) and deployment types (infrastructure and architecture choices) to help clients make correct cost-benefit trade-offs—clients often have existing cloud vendors with preintegrated solutions (for example, Azure AI Platform, including Azure OpenAI Service, and Amazon Bedrock + Anthropic Claude), and as a result, they want to work with partners that understand how to use these solutions, as well as their pros and cons;
- model-agnostic accelerators and building blocks ready to configure and deploy at scale and at speed; and
- a remuneration model based on outcomes, which ties the success of the program with realized returns on investment.

Using generative AI to operate and deliver better are table stakes now for tech service providers. Significant competitive advantage comes only from providing AI solutions that change the game for customers. But doing so requires not only new skills but also deliberation and focus, because no provider can be everything to everyone. As tech service providers build the capabilities to operate and deliver better with generative AI, they must also be choosing where to build best-of-breed abilities to deliver full potential to their customers.

Operational Transformations

Operational Transformations

To Deploy Generative AI Successfully, Look to Earlier Automations

The most experienced firms are widening their lead in cost savings and productivity.

By Michael Heric, Purna Doddapaneni, and Don Sweeney

At a Glance

- Technology companies investing most heavily in automation outperform others in savings and adoption of new disruptive technologies.
- The gap between leaders and laggards is widening as leaders increase investment as a share of IT budget.
- Leaders are planning to invest, on average, over three times more in generative AI than laggards.
- Successful automation programs include enterprise-wide rollout, combined technologies, value creation, and engaged staff.

Given Nvidia's long history of successfully scaling up automation and artificial intelligence (AI) in its engineering work, it came as little surprise last year when the company announced it was one of the first to test generative AI for boosting the productivity of its chip designers. ChipNeMo, as Nvidia calls it, takes publicly available large language models (LLMs), trains them on Nvidia's 30 years of data, and does some fine tuning. The resulting tools serve as a chatbot, an electronic-design-automation-tool script writer, and a summarizer of bug reports.

Like Nvidia, technology companies with a long track record of developing and scaling up programs in traditional forms of automation, such as robotic process automation (RPA) and analytical AI, are now applying the lessons learned to gain an early advantage in generative AI. As with traditional automation, true success comes only when pilots are converted into large-scale programs that deliver compelling returns on investment across the enterprise.

Bain's latest survey of 893 automation executives worldwide, including 124 in technology companies, finds that companies investing most heavily in automation outperform laggards in savings achieved and adoption of new, more disruptive technologies. (We define leaders as companies investing at least 20% of their IT budget in automation in the past two years, and this elite group achieved an average 22% in cost savings. Laggards are companies investing less than 5% of their IT budget in automation, and these firms achieved just under 8% in savings on average.)

Automation leaders at technology firms were able to reduce the cost of processes by 17% in 2023, whereas lagging companies managed only 8%. Respondents also cited the benefits of trimming the number of low-value tasks, speeding up process completion time, and improving service quality and accuracy.

Consider Microsoft's automation in finance over the years. From 2010 to 2020, Microsoft has grown revenue by 145% while growing finance headcount only 15%. AI has also made Microsoft's finance forecasts more accurate and faster—from 100 full-time-equivalent staff spending one month to 2 full-time employees spending two days.

Now the leaders are moving quickly into implementing generative AI, and plan to invest, on average, over three times more of their IT budget in generative AI than laggards *(see Figure 1)*.

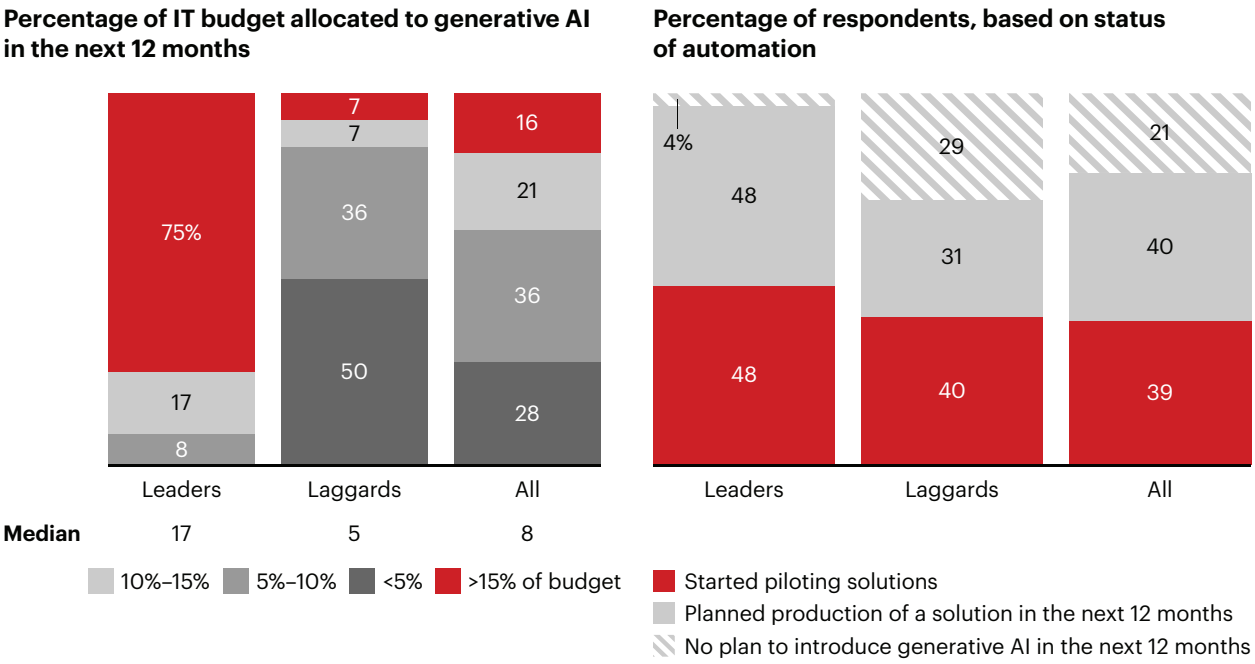
More than cost reductions

The continued wave of automation is generating significant value. AT&T, for instance, began working with RPA in 2015, making the company one of the earliest adopters of the technology, and has been applying AI across its operations for years. AI helps AT&T to optimize field technician routes, reducing fuel consumption while serving more customers; to translate and simplify documents; and to improve coder and developer productivity.

Companies that have successfully scaled up traditional forms of automation—workflow automation, RPA, scripting, and optical character recognition—have already embedded AI outside of LLMs, such as machine learning in document processing or natural language processing in job descriptions *(see Figure 2)*.

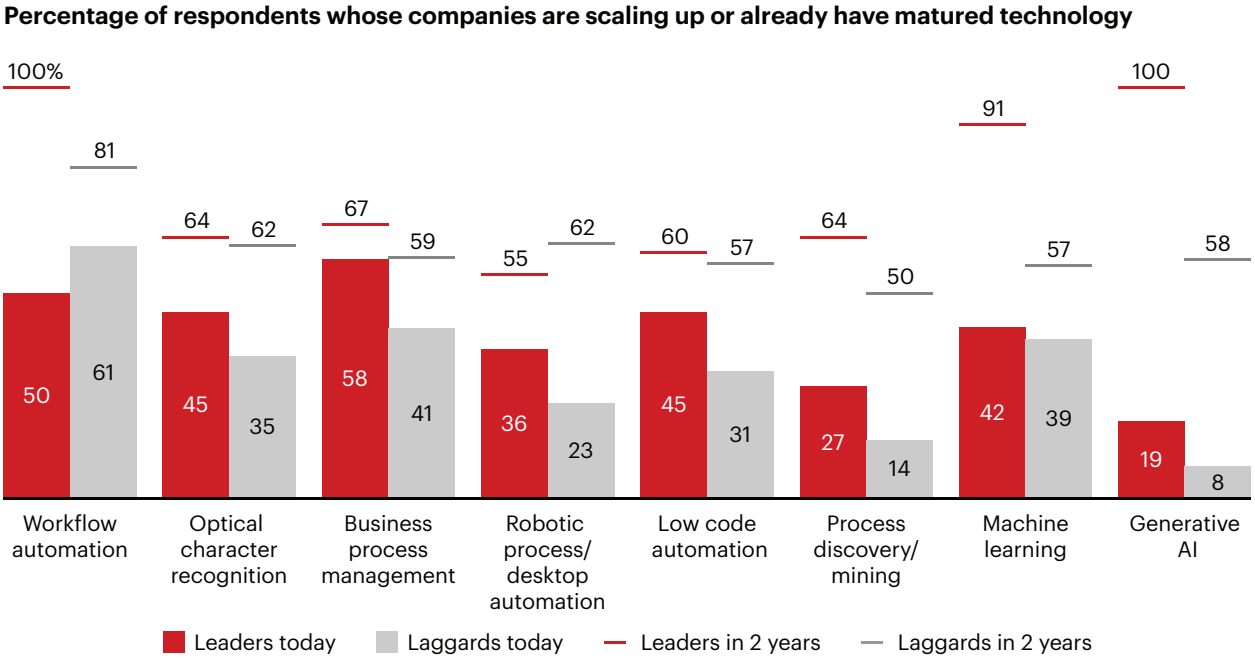
What's more, the gap between leaders and laggards at technology firms has widened and will likely continue to do so, as leaders plan to raise their investment as a share of IT budget while lagging companies plan to be more conservative. In our survey, 33% of leaders plan to invest significantly more in 2024, up from 21% in 2022, compared with only 13% of laggards, down from 19% in 2022 *(see Figure 3)*.

Figure 1: Automation leaders are out-investing other companies in generative AI and moving faster to implement the technology



Source: Bain Automation Pathfinder Survey, 2024, technology companies (n=124)

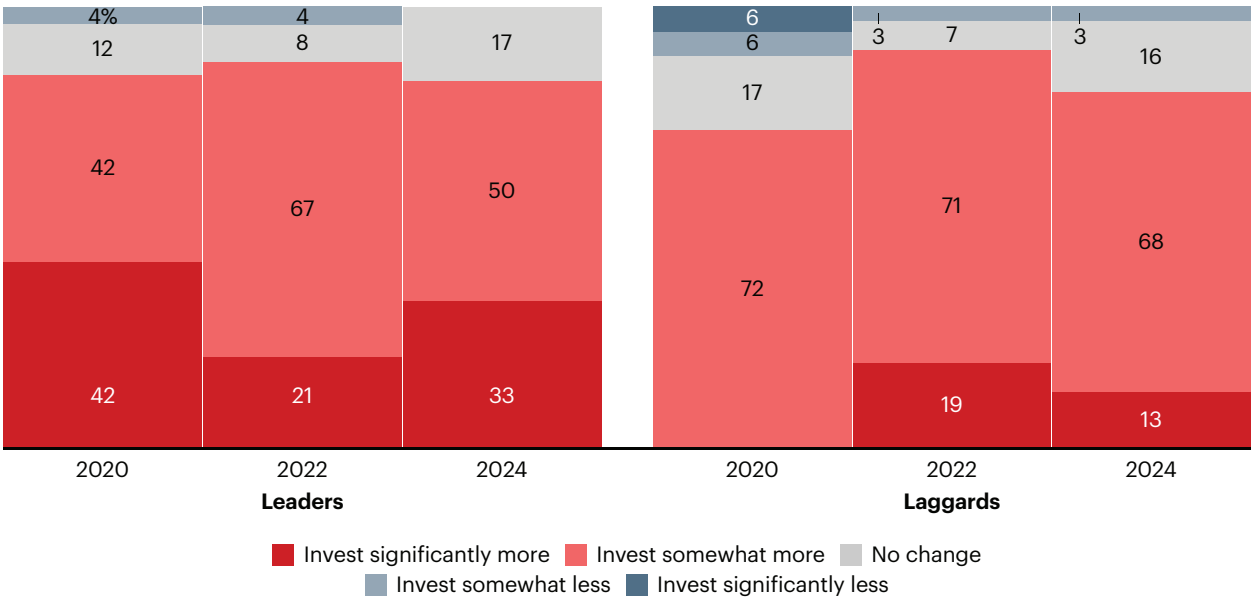
Figure 2: Leaders vs. laggards: technology type



Source: Bain Automation Pathfinder Survey, 2024, technology companies (n=124)

Figure 3: Leaders plan to invest more in automation than other firms in 2024

Percentage of respondents, based on their level of automation investment over the previous 12 months



Source: Bain Automation Pathfinder Survey, 2024, technology companies (n=124)

Generative AI will take automation to new levels of effectiveness and value. Most respondents are and will be using generative AI for three waves of use cases *(see Figure 4)*. In the first wave, they apply the technologies to use cases that were not possible in the past, such as creating new marketing content. For the second wave, they plan to replace technologies for current use cases, including order processing. A third wave will consist of enhancing current use cases, such as accounts payable and receivable. The logic here consists of companies wanting to apply generative AI to new areas, rather than start fresh with use cases where they have already invested resources, built integrations, and trained employees.

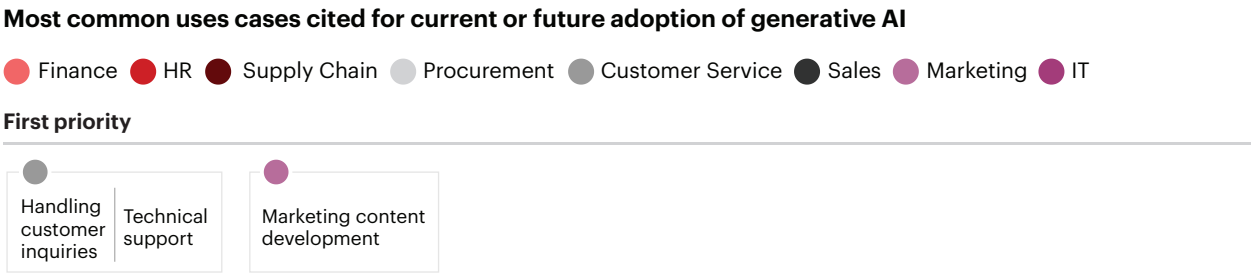
Automation principles that apply to generative AI

Companies that master the following principles will position themselves to rapidly take advantage of generative AI.

Elevate automation from narrow pilots to cross-company strategic initiatives. One common trap is crowdsourcing a long list of small automation projects, often within individual departments, then trying to execute them one by one. This makes it difficult to achieve major savings or other benefits.

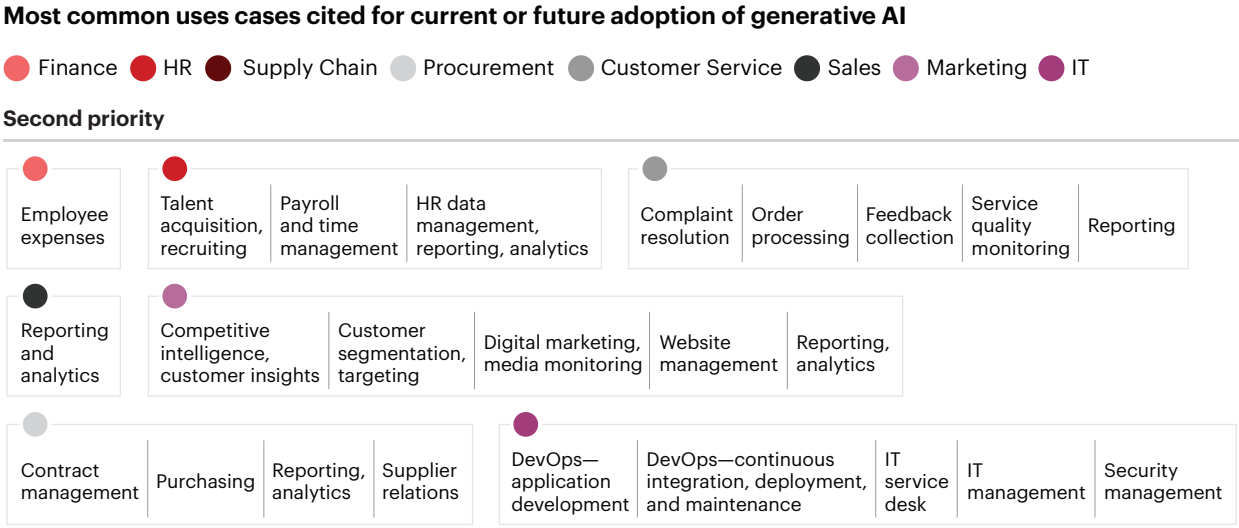
Automation leaders take a different approach. They set bold goals, framing the potential in the millions of dollars. They gain the sponsorship of senior executives and embed automation as a pillar of the overall strategic agenda.

Figure 4a: Companies are applying generative AI to completely new use cases first



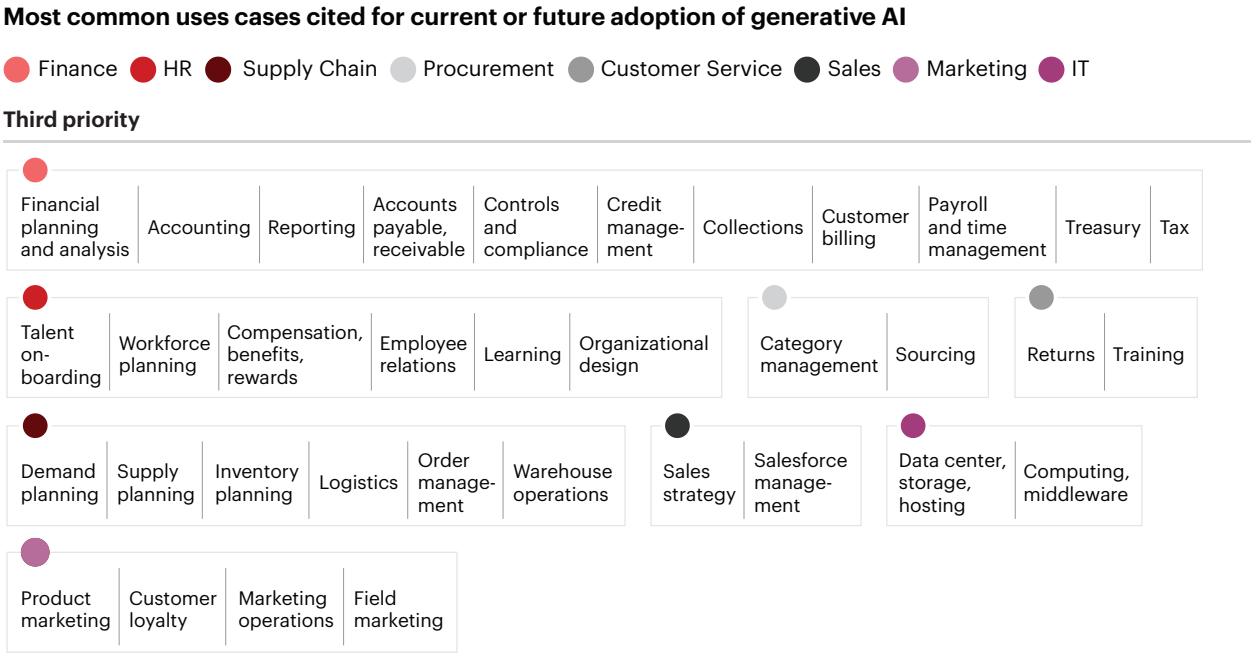
Source: Bain Automation Pathfinder Survey 2024, n=893

Figure 4b: Companies are applying generative AI to completely new use cases first



Source: Bain Automation Pathfinder Survey 2024, n=893

Figure 4c: Companies are applying generative AI to completely new use cases first



Source: Bain Automation Pathfinder Survey 2024, n=893

Combine automation technologies. When individual tasks are automated with different technologies, little value results. Worse, this can add more process complexity than the automation delivers in cost savings. Instead, automation leaders often combine technologies to deliver the best results. They start with the business needs and process, working back to determine the right combination.

Insist on realizing value from automation. Before investing in the software and implementation resources to build an automation, senior executives increasingly want a commitment from the people asking for the investment to achieve savings and other benefits, along with a plan to realize that value. Once automations are deployed, executives expect business processes to be redesigned, and they insist on seeing proof of how teams achieved the value claimed.

Coax and convince to reach full adoption. Managing how employees change their behaviors can make or break an automation program. Maximizing adoption of automation tools entails documenting and educating people on the new way to work, investing in training, tracking adoption rates, and taking steps to keep improving how people use the technologies.

• • •

The level of sophistication and maturity with automation varies widely. But companies that lag can catch up if they're willing to boost their investments and commit to a sustained effort that changes how people work.

The good news is that lessons learned from traditional automation technologies can inform fruitful deployment of new technologies, including generative AI. The techniques, governance issues, and process changes are all quite similar, so using generative AI offers a fresh approach to effectively manage costs and improve the customer experience.

Operational Transformations

Beyond Code Generation: More Efficient Software Development

Generative AI saves time, but meaningful improvements require a broader agenda.

By David Crawford, Bill Radzevych, Jue Wang, Purna Doddapaneni, and Martin Goette

At a Glance

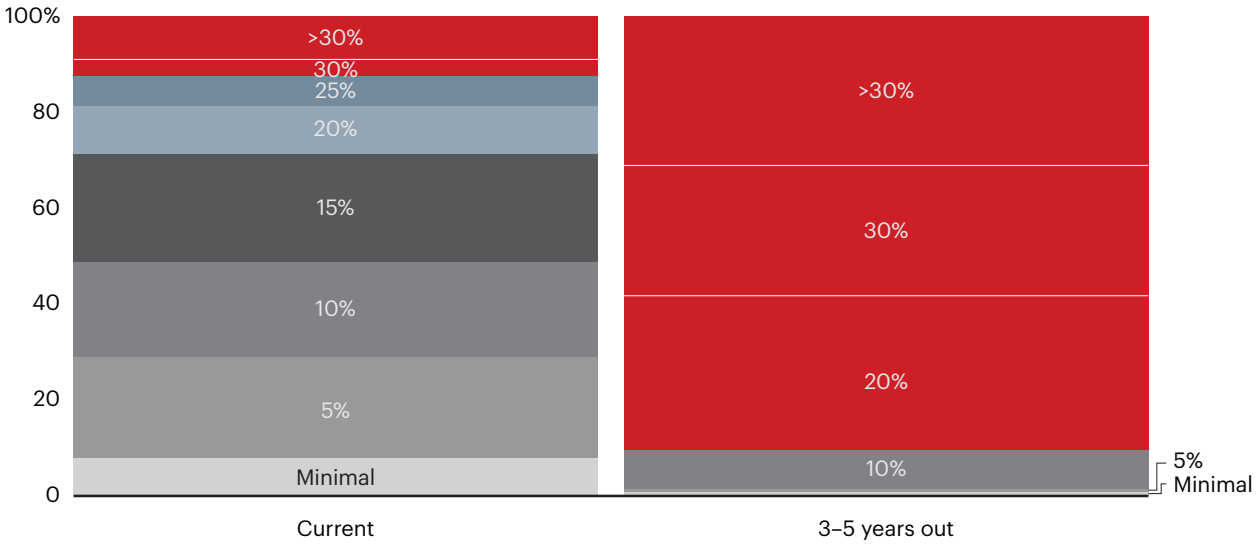
- The arrival of generative AI puts pressure on software development organizations to demonstrate greater efficiency.
- In practice, generative AI appears to save about 10% to 15% of total software engineering time, and a lot of companies aren't making profitable use of the savings.
- Improvements of 30% or more are possible, but they require using the full potential of generative AI and a broader agenda.

The introduction of generative AI coding assistants has raised expectations of improving the efficiency of software development. In practice, engineering organizations that are using such tools are seeing efficiency improvements of about 10% to 15% on average *(see Figure 1)*. In many cases, companies fail to monetize even these gains because they're unable to reposition the saved time and resources to productive uses.

But more is possible. Organizations that take a more comprehensive approach can see efficiency gains of 30% or more. The extra gains result from going beyond generative AI code generation, using

Figure 1: Companies already see some efficiency gains with generative AI, but they expect to see much more in the future

Percentage of respondents who indicated specified percentage of efficiency gain from using generative AI



Note: Excludes respondents who are not directly working with generative AI (n=199)
Source: Bain Generative AI Survey, 2024 (n=209)

generative AI for other tasks, and taking a more comprehensive approach to improving efficiency, including determining the right baselines and metrics.

Real efficiency gains

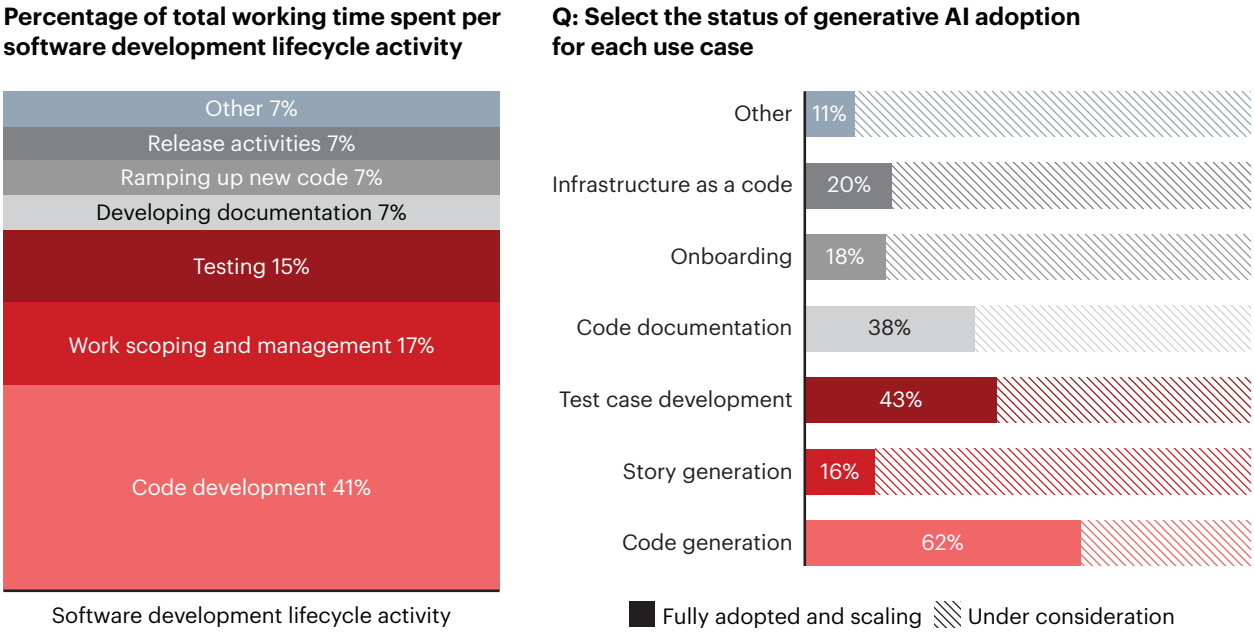
Developers spend about half their time writing and testing code, so although they report a 30% improvement from generative AI against those activities, this represents a net efficiency improvement of 15% across developers' total time *(see Figure 2)*. A more comprehensive approach to efficiency includes not only generative AI-assisted code generation and testing, but a comprehensive look at three dimensions: focusing on the right work, ensuring speedy, high-quality execution (including full potential use of generative AI), and optimizing resourcing costs.

Focus on the right work

By far, the fastest way to improve efficiency is to refocus efforts on the work that creates the most value, concentrating on several sets of actions:

- **Align investments with strategy across products and markets.** Does the allocation of engineering time match company strategy? Are strategy and roadmaps informed by customer and market insights?

Figure 2: Developers spend most of their time creating new products and improving existing ones; generative AI use cases focus on test development and code generation



Note: Excludes respondents who are not directly working with generative AI (n=199)
Source: Bain Software Developers Survey, 2024 (n= 209)

- **Weigh expense-to-revenue ratio across products.** Are older products taking resources from new developments? Should support be outsourced to reduce costs?
- **Balance resource allocation across new developments, product improvement, maintenance, technical debt, and quality.** Spending too little to address technical debt may eventually slow development.
- **Link product strategy to day-to-day developer priorities.** Unclear prioritization can lead to developers addressing the issues they believe are most critical, which are not always the same as those that support strategic goals.

Better visibility into how time is actually spent often reveals a mismatch between leadership's ambitions and the reality of how resources are allocated.

Ensure speedy, high-quality execution

There are many aspects to executing rapidly with high quality. Generative AI is top of mind today, but foundational elements such as continuous delivery and modern architecture can be more effective ways to drive efficiency *(see Figure 3)*.

Figure 3: Clear roadmaps, managing tech debt, and ensuring optimal resource allocation are the most effective ways to improve productivity

Percentage of respondents who selected this as a top three lever

Focus on the right work	**62%**	Clear link between product roadmap and company strategy
	62%	Roadmaps based on deep understanding of customer needs
	44%	Balanced investments across new development, sustainment, and technical debt
	41%	Work prioritized on standard set of criteria (cost, value, risk, etc.)
	40%	Clear decision rights and ways of working
Ensure speedy, high-quality execution	**56%**	Architectural soundness and technical debt management
	55%	Maturity of DevOps
	31%	Generative AI adoption across software development process
	31%	Robust quality management
	28%	Optimal developer infrastructure
Optimize resource costs	**64%**	Organizational structure suitable to needs
	53%	High satisfaction among developers
	38%	Competitive compensation
	37%	Servant leadership, investment in career progression
	36%	Transferability of skill set across teams and products

Source: Bain Generative AI for Development survey, 2024 (n=209)

Deploy full potential generative AI. Leaders in generative AI adoption can achieve up to 30% efficiency from optimal deployment. Intuit, a financial technology platform for consumers and small businesses, offers a good example with its initiative to move from "scrappy testing" to scale development.

Intuit set out to improve efficiency and productivity around two themes. First, it wanted to increase development velocity to deliver innovative products and solutions to its 100 million customers, with speed and at scale. Second, it wanted to take full advantage of the inherent benefits of generative AI on its modern development platform to streamline end-to-end development by "shifting left"—that is, bringing critical tasks forward in the software development life cycle. Among the key takeaways from testing and scaling more than 30 different use cases are:

- **Beyond code generation.** Intuit used its proprietary generative AI operating system (GenOS) to analyze developer support documentation, logs, and other records to understand how developers have solved common problems in the past, extracting this knowledge to accelerate development velocity. The company created tools that serve up solutions to common developer tasks, meeting development teams where they are in their day-to-day work (integrated development environments, development portal, Slack, etc.) to drive efficiencies.

- **Accelerating with context.** While Intuit's initial generative AI-driven code-generation tool sped up the process by 10% to 15%, by leveraging its generative AI tooling with Intuit-specific code

context patterns (repositories, component libraries, etc.), the company reduced integration task completion times by two or three times.

- **Improving the end-to-end development process.** Intuit used its generative AI tools to improve standardization of code and documentation for product development teams across personas (software developers, designers, data engineers and analysts, technical program managers, etc.).

Plan for continuous integration and delivery. Before developer teams deploy new code, they need to ensure that it won't break anything in the live product or create security risks. Manual testing is time-consuming, and deploying to a live environment would be risky. Automating the testing in a virtual product environment is a more efficient and safer way to confirm the viability of new code.

Generative AI is top of mind today, but foundational elements such as continuous delivery and modern architecture can be more effective ways to drive efficiency.

Continuous integration and delivery of new code also improves efficiency. It's a more efficient way to manage risk because developers can assess the effects of each new deployment, and it allows companies to address security threats as they are discovered, limiting potential harm that could occur if the patches had to wait. Customers also appreciate a quick response to identified issues and the consistency of ongoing product improvements.

Maintain a modern architecture. Modular architecture allows teams to adapt and improve products without reinventing the whole. A continuous investment in modular design avoids falling into technical debt—the cost incurred when companies fail to keep up with evolving technology and must invest heavily at some point to regain their competitive edge.

Optimize resource costs

Two software development organizations operating at similar speeds and quality can show very different cost profiles, depending on each organization's model and talent structure. Geographical footprint, outsourcing levels, ratio of senior engineers to other team members, and the roles that various functions play all help determine costs. For example, a staff overloaded with senior engineers can be costly and may be slower to adopt new practices, whereas a staff with too many junior engineers may lack technical depth and result in higher costs despite savings.

How to measure impact

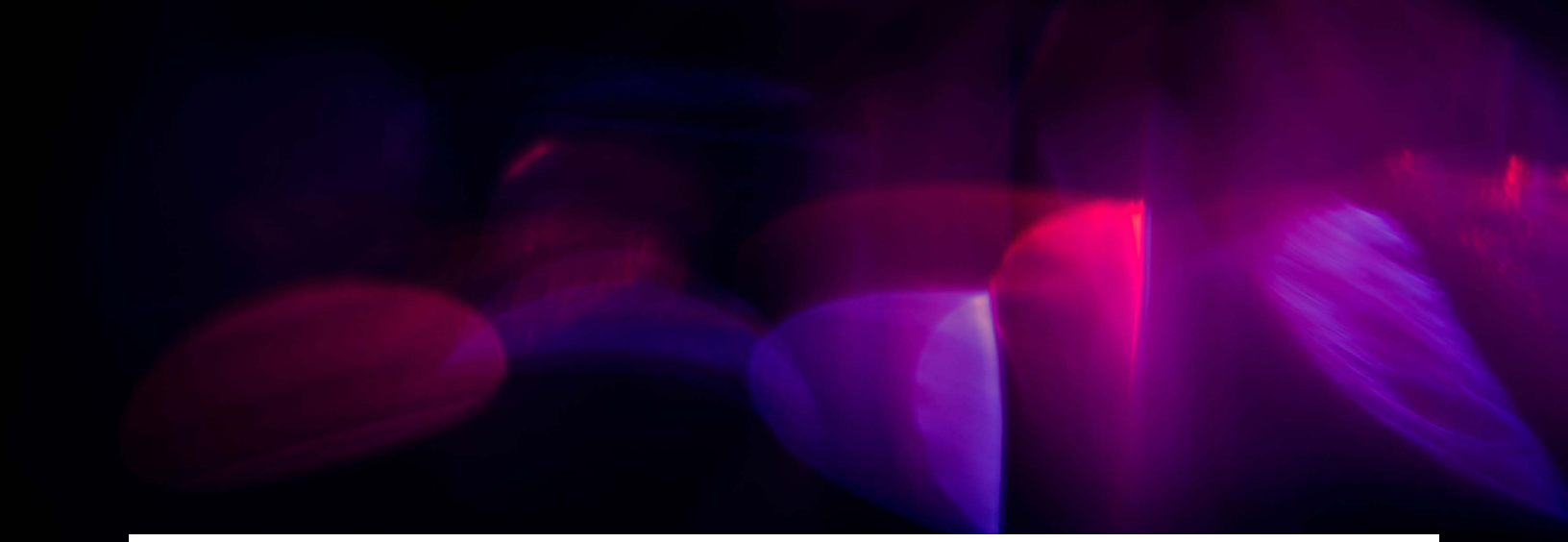
Many companies struggle to understand their baseline efficiency and measure the improvements they try to get from new initiatives like generative AI. About two-thirds of leaders surveyed aren't satisfied with the insights they're getting—or not getting. Many senior executives see software engineering as a black box: They don't know where the money's going.

Building an effective measurement system requires a bespoke approach and focused attention. To avoid overload, a good target is 3 to 5 KPIs for the senior executive level and up to 10 KPIs for engineering leadership. Tiered systems address different needs of different groups:

- Executives need to focus on product performance, cost, and resource allocation across priorities.
- Technical leadership needs a view of whether their efforts are achieving the right business outcomes, and needs to identify barriers and upcoming challenges.
- Teams need to know if their deliveries are in line with requirements.

A dedicated engineering productivity tool to measure efficiency is an essential enabler.

Meaningful improvement is possible in software development, but the effort required is more far reaching than introducing a generative AI coding assistant. Investments that increase efficiency, improve execution, and optimize costs consistently pay off, making the effort worthwhile for any R&D or other software development organization.



Operational Transformations

Why Software Companies' Customer Success Is Failing

Spending on customer success is up, but customer retention is down. Post-sales teams must evolve.

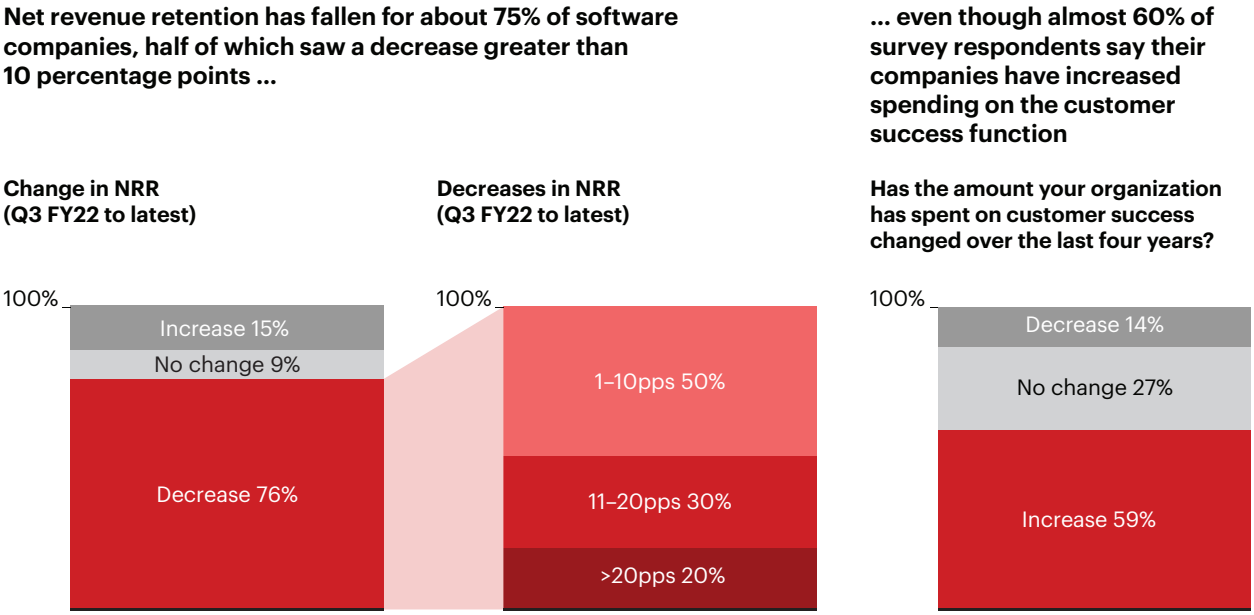
By Matt Eldridge, Greg Fiore, Simon Heap, and Kenzie Haygood

At a Glance

- Despite increased customer success investments, net revenue retention rates have declined for 75% of software firms in a recent Bain survey.
- In addition, nearly two-thirds of software customers feel their post-sales needs are only being moderately addressed or worse.
- There's a mismatch between how vendors provide support and what customers value, particularly in technical implementation assistance.
- Leading firms are developing a clear product and customer journey blueprint, better defining and coordinating post-sales roles, and investing in AI-enabled self-service tools.

Software companies are grappling with a surprising disconnect: Despite significant investments in post-sales personnel since the pandemic, customer retention has suffered. Net revenue retention (NRR) rates, a measure of how well companies retain and expand revenue from existing customers over a certain period, decreased for 75% of software companies in a recent Bain survey, even as nearly 60% increased customer success spending *(see Figure 1)*. Frustrated executives are

Figure 1: Net revenue retention rates have decreased for many software companies despite spending more on customer success



Notes: Net revenue retention (NRR) in Q3 FY2022 compared with latest reported figure as of June 2024; 74 software-as-a-service companies included in cohort
Sources: Bain Customer Success Practitioner Survey, June 2024 (n=150); Meritech Capital; company annual reports; Bain analysis

questioning why these investments haven't paid off and, worse yet, may have exacerbated the problem.

It turns out a huge disconnect exists between how customers want to be served and what vendors think they need. Additionally, many vendors silo the customer success function and miss the opportunity of a streamlined post-sales package with cohesion across tech support, customer success, and professional services.

Effective post-sales activities help buyers implement software, increase adoption, adapt their use as needed, and achieve ROI, all of which are more important than ever with the acceleration of software purchases during the pandemic and the increasing complexity of software-as-a-service (SaaS) products. Slowing post-pandemic sales and a shift from subscription- to customer usage-based pricing have further raised the stakes for retaining customers and enticing them to spend more.

Software vendors have largely relied on customer success teams to maximize product use. In addition to customer success practitioners' increased spending, more software companies are creating a dedicated customer success team. The share of US enterprise software companies with a customer success team reached 60% this year, up from about 40% four years ago, according to Bain analysis of LinkedIn and other data.

To deepen customer relationships, vendors have also emphasized more proactive, specialized customer success roles, including customer success managers, technical account managers, and success architects. Meanwhile, they've reduced spending on reactive tech support roles through automation, which generative AI could further accelerate. Consequently, customer success roles now constitute a larger portion of the post-sales workforce *(see Figure 2)*.

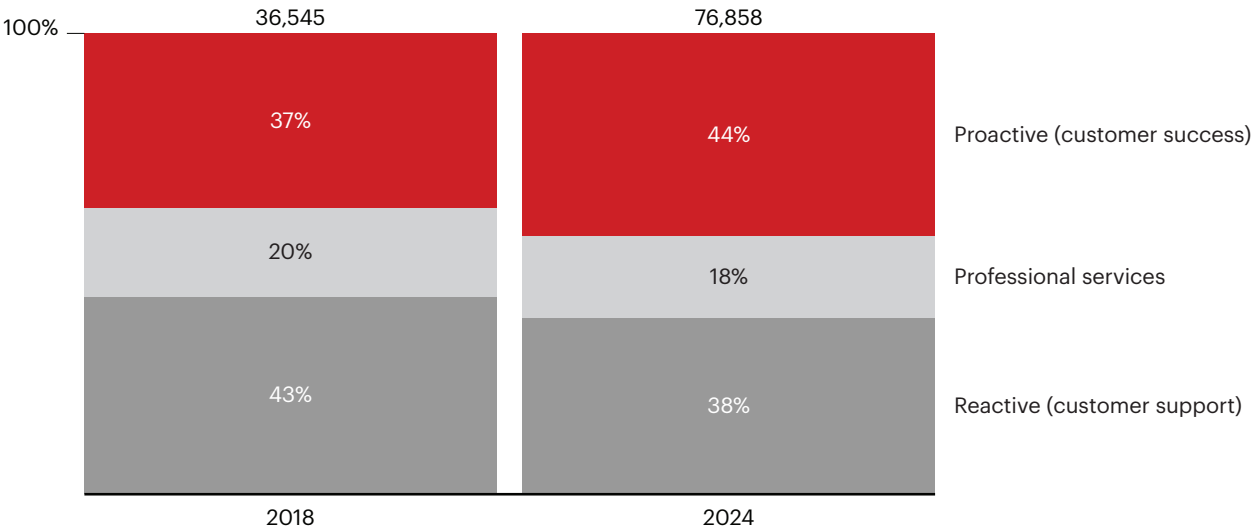
However, these investments haven't delivered the desired results. Software vendors' deteriorating NRR rates align with our customer success survey data: nearly two-thirds of software customers feel their post-sales needs are only being moderately addressed or worse *(see Figure 3)*.

Why? Our research found a mismatch between how vendors provide support and what customers value. In our recent survey, software buyers ranked assistance with technical implementation or deployment as their highest priority for customer success, while practitioners ranked it sixth *(see Figure 4)*. Vendors often provide general assistance but have abdicated too much technical implementation to systems integration partners. Clearly, customers see a role for vendors to provide architectural support and technical implementation best practices, even if systems integrators continue to do the heavy lifting.

Another disconnect is that customers prefer a technical role as their primary contact for customer success, while vendors often assign a non-technical customer success manager instead, according to our survey.

Figure 2: Software vendors are emphasizing more proactive, specialized customer success roles

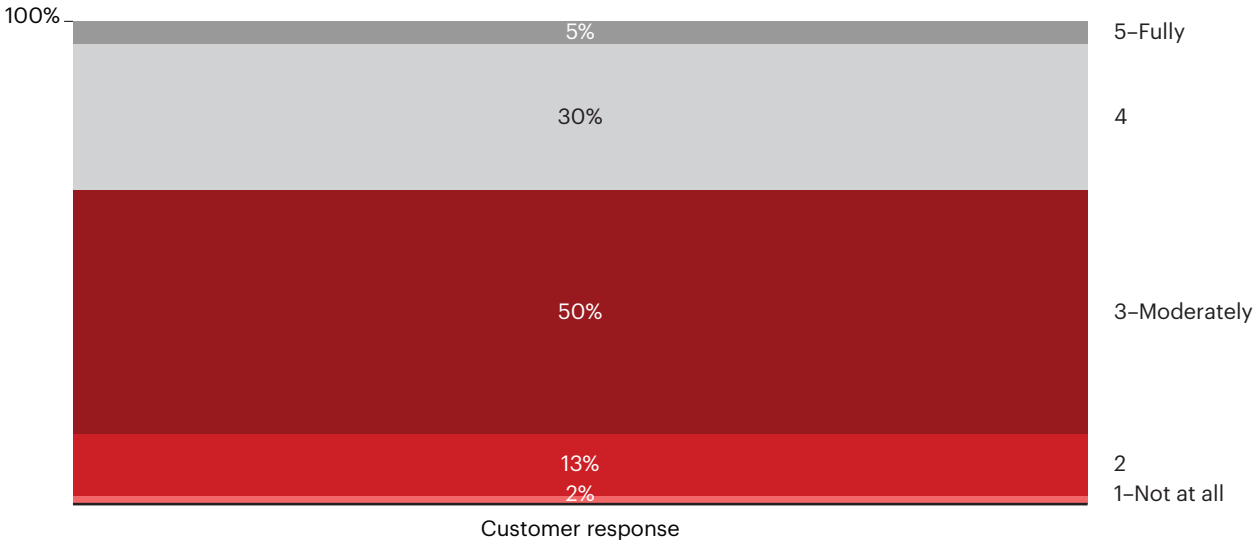
Share of post-sales headcount per function, by year



Notes: "Proactive" includes technology account managers, customer support managers, and other customer success roles; analysis includes 130 public software-as-a-service companies with annual revenue over $100 million; headcount data as of May 2024
Sources: Aura Intelligence; ClassifAI; Meritech Capital; Bain analysis

Figure 3: About 65% of customers feel their post-sales needs are only being moderately addressed or worse

Are your post-sales needs currently being met?

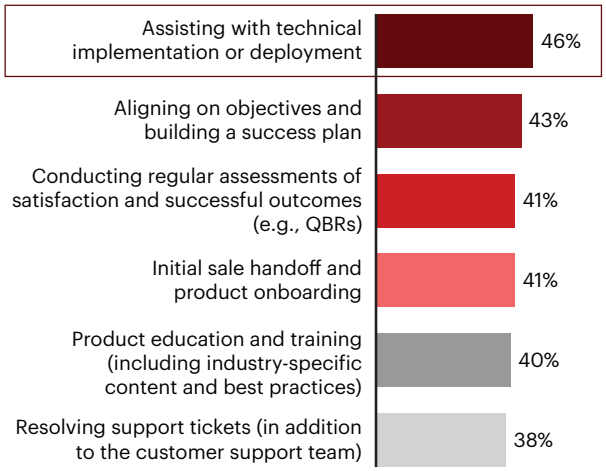


Source: Bain Customer Success Customer Survey, June 2024 (n=149)

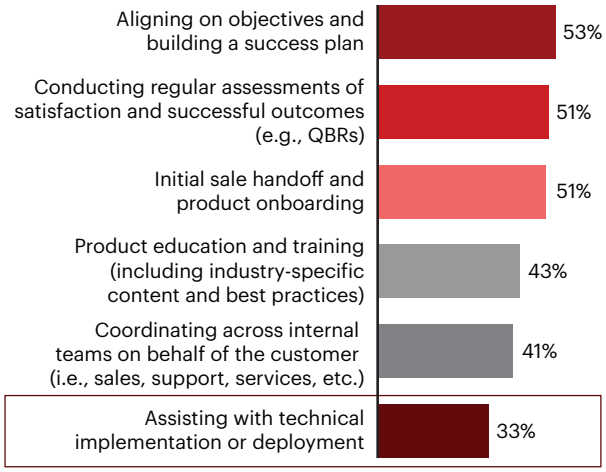
Figure 4: Software buyers highly value assistance with technical implementation, but vendors see it as a lower priority in their customer success activities

Please rank the top four most important/helpful activities for customer success to perform

Customer perspective



Practitioner perspective



Note: Ranked options were selected in the top four
Sources: Bain Customer Success Customer Survey, June 2024 (n=149); Bain Customer Success Practitioner Survey, June 2024 (n=150)

To be fair, customer success functions face intense pressure due to a broader push to reduce IT costs, increased scrutiny of internal budgets, and client procurement teams challenging add-on services such as paid, premium customer success offerings. These headwinds are compounding two long-standing challenges: It remains difficult to prove the return on customer success investments, and many of these teams are perceived as cost centers rather than revenue generators.

That said, many vendors have hired excessively in customer success without validating their post-sales model, missing opportunities to deploy all post-sales functions more efficiently.

What does good look like?

Based on our work with software companies worldwide and analysis of companies with high NRR, we've found that the emerging leaders are focusing on the following key steps.

1. **Step into your customer's shoes.** Emerging leaders first map how customers interact with and derive value from the product, followed by mapping the broader customer journey from initial consideration through purchase, post-sales support, and renewal. It sounds like a given, but many companies' product and engineering teams don't articulate what the product's value realization journey looks like or work to deeply understand customers' needs, desired outcomes, and key touchpoints. This understanding will help post-sales teams not only focus on the moments that matter, but also identify where the product falls short. Leading companies don't lose sight of the basic principle that product excellence is the true foundation of customer success.

2. **Redefine post-sales roles to better deliver on the customer success mandate.** Blurring lines between post-sales functions has created confusion for customers and inefficiency for vendors. Leading companies are taking a blank-sheet approach: They clearly define all activities required for the customer journey, rationalizing roles from the customer's perspective. Each role should have distinct responsibilities that collectively leave no gaps in customer success deliverables.

   Many leading companies are creating outcome-based success plans that ask customers to define what success looks like to them, which enables vendors to develop a catalog of post-sales service jobs that closely match customer needs. This improves efficiency while avoiding overly standardized activities that lose sight of the nuances of delivering value for individual customers with individual products.

3. **Coordinate the front lines.** Reorganization alone won't solve customer success issues; companies must better coordinate frontline teams. A chief customer officer can help oversee coherence across the customer success, tech support, and professional services teams, but simply creating the role isn't enough. The company must break down communication silos and design fluid handoffs between post-sales functions and partners that ensure a seamless customer experience.

   One key is to define the "swim lanes" within which each role supporting an account—including sales and post-sales—has a clear set of activities and interaction points with colleagues. This

helps ensure clear accountability and a blueprint for collaboration. Leading firms have also prioritized effectively training, empowering, and managing customer success managers, recognizing their integral role in a successful customer engagement model.

Because emerging leaders apply the right resource to the right activity at the right time, many have reduced spending while achieving better results. Unsurprisingly, our analysis found that companies with high NRR are much better than lower performers at cost-effectively deploying customer success resources.

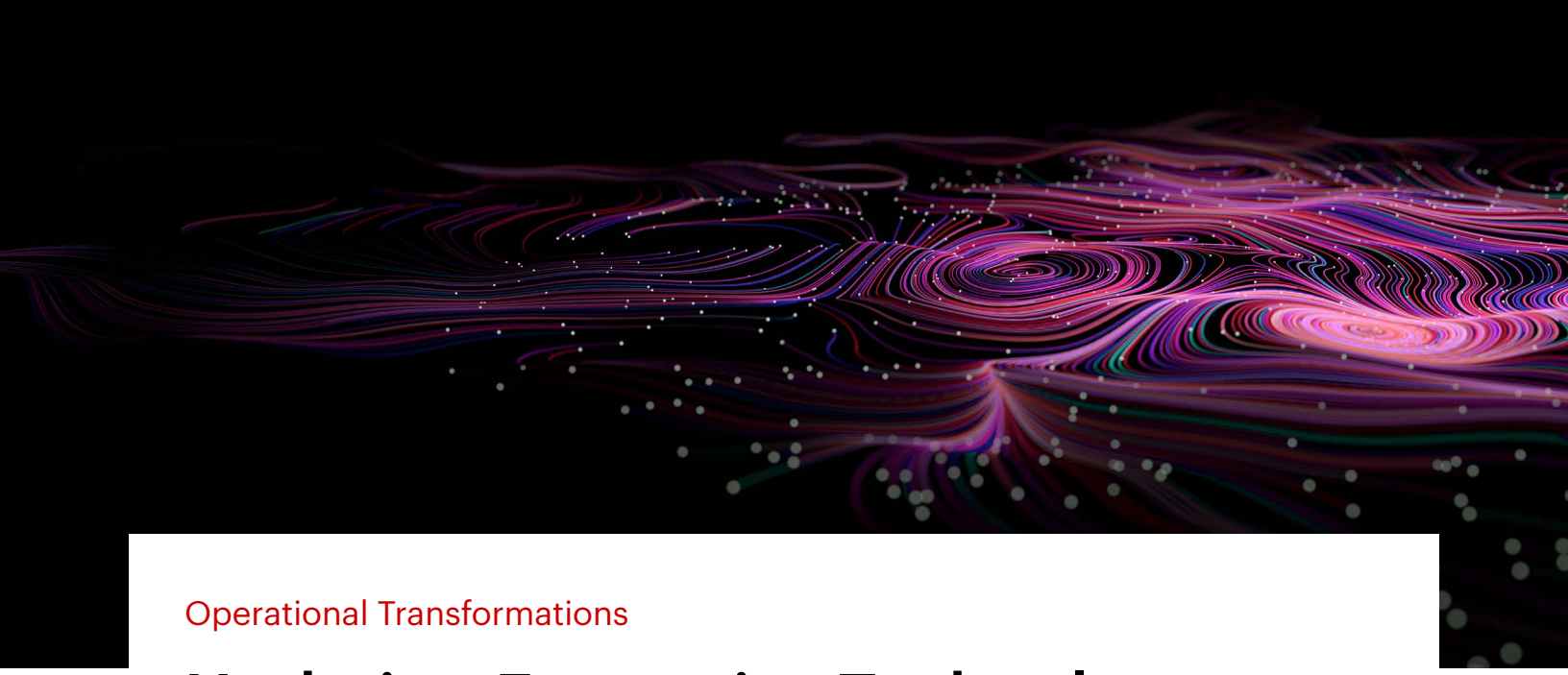
4. **Don't jump into generative AI applications before reassessing the underlying business process.** Customer success teams that are extracting the most benefits from generative AI start by thoroughly evaluating existing business processes to identify pain points and transformational opportunities. With this foundation in place, they redesign key processes to take full advantage of generative AI and other advanced tools, ensuring that inefficiencies and existing problems aren't automated (and thereby amplified). Leading companies are creating a comprehensive AI roadmap that strategically prioritizes use cases and quantifies the sources of productivity gains.

5. **Empower customers with digital self-service tools.** Many customers prefer self-service options for training, onboarding, and support. Robust digital self-service tools can improve customer satisfaction and allow post-sales employees to spend the bulk of their time helping customers solve their greatest challenges. Generative AI can be transformational in this arena, but only if the company has a well-defined product and customer journey blueprint.

The starting point

Going forward, the most effective post-sales organizations will nimbly adapt to customer needs and market trends. To identify the right strategy, customer success practitioners can start by asking themselves the following questions:

- Which activities most affect customers' return on software investments and subsequent purchase decisions, and how can we measure this?
- Are we packaging monetized customer success services in a way that doesn't lose sight of the activities that spur ROI for customers?
- Which activities are suitable for automation and digital self-service?
- How can post-sales functions be effectively coordinated to ensure a seamless customer experience?



Operational Transformations

Updating Enterprise Technology to Scale to "AI Everywhere"

The rapid adoption of generative AI has CIOs managing significant changes in the ways that work gets done.

By Bharat Bansal, Stuart Sim, and Bala Parameshwaran

At a Glance

- Companies can't scale their AI solutions without also reshaping their technology function to enable this massive shift.
- Taking an "AI everywhere" approach to re-architecting the tech stack is a critical, foundational step.
- Equally important will be upgrading current ways of working to make the best use of new AI solutions, which will require bringing the discipline of software development to the adoption of AI models.

Companies are moving beyond the experimentation phase of proofs of concept and minimum viable products, and beginning to scale up generative AI across the organization. As they do, CIOs will need to own, develop, and maintain production-grade AI solutions while efficiently delivering them at scale. At the same time, they will need to enhance their own function's productivity with the generative AI tools they are deploying to the rest of the organization.

This will fundamentally reshape the technology function across architecture, operating models, talent, and funding approaches, in several important ways:

- **re-architecting the entire tech stack** with an "AI everywhere" approach, integrating machine learning (ML) and generative AI;
- **upgrading ways of working** to incorporate AI solution development across product management, software development, operations, and support processes;
- **upskilling engineering teams** to integrate, test, and scale AI systems to production grade, while using AI tools to boost engineering productivity;
- **redefining the mix of tech spending** to support AI investments and infrastructure run costs, capturing the efficiencies from AI in areas like software development and service management; and
- **reviewing risk management and governance** to successfully deploy and upgrade AI models.

While all five of these processes will reshape the technology function, the first two—architecture with AI everywhere and upgrading ways of working—are the critical foundations to get right first.

Companies are moving beyond the experimentation phase of proofs of concept and minimum viable products, and beginning to scale up generative AI across the organization.

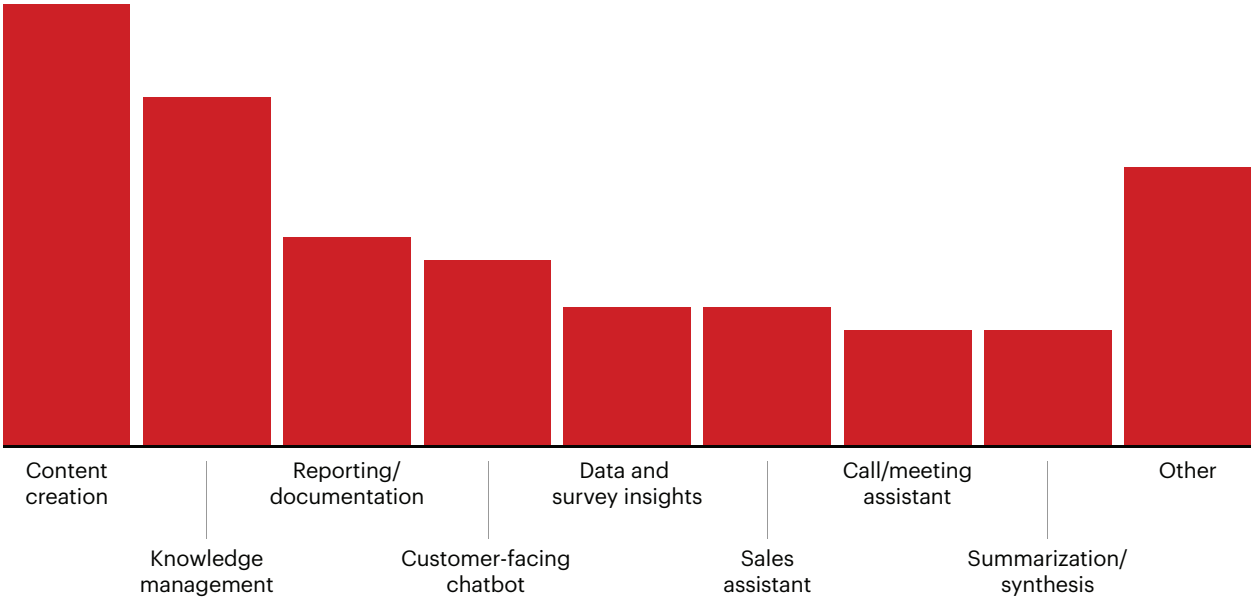
Architecture with AI everywhere

Generative AI will affect systems across the entire enterprise.

- **Operational systems** with significant unstructured data will face substantial re-architecting due to generative AI's ability to make use of previously underutilized data sources. In our experience, the most common solution patterns for generative AI use cases in operational systems fall within the areas of content generation, knowledge management, and reporting and documentation *(see Figure 1)*. CIOs and other tech buyers will need to decide between building or buying generative AI solutions for these uses, based on the potential competitive advantage and the cost and capability required. Currently, many companies are building or tailoring the solutions they need using foundation models because the necessary commercial solutions are not yet ready. Buying may become more practical and popular as existing software-as-a-service (SaaS) solutions incorporate generative AI.

Figure 1: Content creation, knowledge management, and reporting and documentation are among the most common applications of generative AI

Generative AI use cases, as a share of total related Bain casework, June 2023–July 2024



Source: Bain casework

- **Integration, workflow, and orchestration systems** will need to work seamlessly with AI models to enable more complex automation workflows. Additionally, generative AI accelerates the need for modernizing enterprise architecture, such as adopting API-driven integrations and cloud-first infrastructure, to deploy generative AI solutions more effectively. Over time, workflow and orchestration systems could be powered or replaced by agentic AI that can act semi-autonomously, as that capability matures.

- **Data analytics and ML systems** need to cover more unstructured data assets, as well as an AI as a service (AIaaS) platform and machine learning operations (MLOps) for reuse of common components and efficient deployment of new models. Data platform capabilities will need to be strengthened to incorporate more unstructured data sets (and treat them with the same discipline as structured ones), shared data catalogues, data versioning, and data lineage supported by data product teams. To enable use of approved models and common components (e.g., vector indexing or retrieval augmented generation) across use cases, an integrated AIaaS platform, rather than point solutions, needs to be created for each use case.

Upgraded ways of working

As generative AI model use cases get deployed across critical systems and complexity increases (for example, daisy-chained AI use cases), it will put further demands on collaboration, quality control,

reliability, and scalability. AI models will need to be treated with the same discipline as software code by adopting MLOps processes that use DevOps to manage models through their life cycle.

Companies should set up a federated AI development model in line with the AIaaS platform. This should define the roles of teams that produce and consume AI services, as well as the processes for federated contribution and how datasets and models are to be shared.

Given the pace of evolution of generative AI, it is also imperative to create AI-first software development processes that allow for rapid iteration of new solutions and architectures. Agile teams need to factor in dependencies between applications, AI models, and data teams.

Many of these choices will need to be made in a landscape of rapidly evolving generative AI technologies, necessitating some no-regret moves now while maintaining flexibility to adapt.

Software development and service management processes should also adopt generative AI tools, including coding assistants, knowledge management, and error detection. Clear guidelines are required on how to deploy these tools, regularly monitor their impact, and manage risks.

Many of these choices will need to be made in a landscape of rapidly evolving generative AI technologies, necessitating some no-regret moves now while maintaining flexibility to adapt. As a result, this topic will become a priority for CIOs, creating significant change in the function, far beyond what we have seen in recent years.

Bold ideas. Bold teams. Extraordinary results.

Bain & Company is a global consultancy that helps the world's most ambitious change makers define the future.

Across 65 cities in 40 countries, we work alongside our clients as one team with a shared ambition to achieve extraordinary results, outperform the competition, and redefine industries. We complement our tailored, integrated expertise with a vibrant ecosystem of digital innovators to deliver better, faster, and more enduring outcomes. Our 10-year commitment to invest more than $1 billion in pro bono services brings our talent, expertise, and insight to organizations tackling today's urgent challenges in education, racial equity, social justice, economic development, and the environment. We earned a platinum rating from EcoVadis, the leading platform for environmental, social, and ethical performance ratings for global supply chains, putting us in the top 1% of all companies. Since our founding in 1973, we have measured our success by the success of our clients, and we proudly maintain the highest level of client advocacy in the industry.

For more information, visit **www.bain.com**