



Top Cloud Trends for 2026

The year the cloud stopped being a destination
and became a decision.

Cloud & AI Infrastructure · Research Report



Contents



03	The old logic just broke
04	Key takeaways
06	Trend 1 · Agents stop suggesting and start doing, inside sandboxes
07	Trend 2 · Repatriation goes mainstream, and it is a strategy, not a surrender
08	Trend 3 · The training-versus-inference split becomes the core financial decision
09	Trend 4 · Geopatriation: sovereignty becomes a force of its own
10	Trend 5 · Infrastructure FinOps grows up
11	Trend 6 · Hybrid and multi-cloud become the default architecture
12	Trend 7 · The edge becomes tier-1 infrastructure
13	Trend 8 · AI-native platforms change who builds software, and how fast
14	Figures
16	How to decide: a placement framework for 2026
19	Resources and references
21	About Plaxonic

The old logic just broke



For over a decade, the cloud decision was simple, and it was made once. Move to the cloud, move everything, and treat any hesitation as legacy thinking. The economics backed it up: for bursty, unpredictable, variable workloads, renting elastic compute was almost always cheaper than owning it.



Two forces broke that logic at once, and 2026 is the year the break became undeniable. AI changed what runs on the cloud, and it changed who does the building. This report starts with the second, because it is the most visible sign that the ground has shifted, then works through the economics reshaping where all of it should live.



- **AI agents have started doing the work, not just suggesting it.**
Sandboxed execution environments now let autonomous agents write, run, and iterate on real code in isolation, turning agentic AI from a demo into production infrastructure.
- **AI broke the old cloud economics.** AI workloads are high-utilization, long-running, and GPU-dependent, the opposite of the bursty workloads public cloud was priced for. Renting everything is the wrong default for the first time in a decade.
- **Repatriation has gone mainstream, but it is not a retreat.**
Predictable, high-volume workloads move back to owned or colocated infrastructure, while bursty and global workloads stay in public cloud. The result is deliberate hybrid placement.
- **Sovereignty became a separate, jurisdictional force.** "Geopatiation" moves workloads to clouds legally incorporated in the customer's own regulatory zone, driven by data-residency law rather than the finance team.
- **FinOps is no longer optional.** With an estimated 29% of cloud spend now wasted, reversing a five-year improving trend, unit-cost economics have become a board-level discipline.

The winners in 2026 will not be the companies most "in the cloud" or most "out of it." They will be the ones who treat placement as a per-workload decision, and build the capability to move.



THE SHAPE OF 2026

The winners in 2026 will not be the companies most "in the cloud" or most "out of it." They will be the ones who treat placement as a per-workload decision, and build the capability to move.





Agents stop suggesting and start doing, inside sandboxes

The most striking shift of 2026 is the quiet arrival of infrastructure for AI agents that do not just write code but actually run it. This is the step past the code assistant: from an AI that hands a developer a suggestion, to an agent that is given a goal and an isolated environment and completes the work itself.

The enabling technology is the code-execution sandbox, and it is real, deployed infrastructure, not a concept. A sandbox is a remote runtime where an agent can actively write and execute application code, running deep inside an isolated microVM or secured container, walled off from the host operating system. The isolation is industrial-grade: leading providers build on Firecracker, the same microVM technology AWS uses to isolate Lambda functions, with each sandbox booting in under 200 milliseconds and receiving its own isolated kernel, filesystem, and network namespace.

And it is already in production at named companies, which separates this from hype. Ramp uses sandboxes for background coding agents that generate code changes and write them back as commits or pull requests, and Lovable uses them as preview environments for generated apps, on platforms already supporting more than 50,000 concurrent sessions.

The reason this matters is entirely about safety, and it is the honest heart of the trend. The whole category exists because AI agents generate and execute untrusted code autonomously, and running that code on production infrastructure creates serious security risks. If an agent makes a catastrophic mistake or is hijacked, the only thing destroyed is a disposable container, not your systems. There is a real enterprise caveat too: several sandbox providers do not yet carry enterprise compliance certifications like SOC 2, ISO 27001, or HIPAA, which matters enormously for regulated buyers.

PLAXONIC'S VIEW

This is the trend that turns agentic AI from a demo into infrastructure. An agent that can only suggest is a smarter autocomplete; an agent that can execute, test, and iterate in a safe, isolated environment is a genuinely new kind of worker. But the word that matters is "isolated." The organizations that adopt this well treat the sandbox as seriously as they treat production, with the same rigor on permissions, audit trails, and compliance. Giving an autonomous agent the ability to run code without a properly walled environment is not innovation, it is the fastest route to a breach.



Repatriation goes mainstream, and it is a strategy, not a surrender

Once agents and AI workloads run at scale, the next question is where all that compute should live, and the biggest answer in 2026 is that a lot of it is moving back out of the public cloud. This is not the anti-cloud backlash skeptics predicted. It is a maturing of judgment.

Gartner, IDC, and Deloitte highlight cloud repatriation accelerating in 2026 due to AI's compute-intensive nature, where public cloud costs for training and inference have become unpredictable, prompting CIOs to shift steady-state and sensitive AI workloads on-premises. The direction is consistent: a 2024 IDC survey found 86% of CIOs planned repatriation of some workloads, the highest rate recorded.

But the nuance is where most commentary gets it wrong. This is not companies leaving the cloud. The accurate framing is the "repatriation paradox": cloud spend grows despite repatriation, because new AI and analytics workloads flow in while traditional steady-state workloads exit. Enterprises do both at once. The emerging discipline: repatriate what is predictable and high-scale, keep buying what is bursty and global, and make every other workload a deliberate placement decision, not a default.

PLAXONIC'S VIEW

Repatriation is not a project you run once, it is a capability you build. The organizations that win are not the ones who move the most workloads home, but the ones who can move any workload in either direction as its economics change. That means owning your infrastructure as code, keeping your architecture portable, and refusing to lock so deeply into one provider's proprietary services that placement is no longer your decision to make.



The training-versus-inference split becomes the core financial decision

Underneath repatriation sits a more precise insight reshaping infrastructure strategy: training and inference are financially different activities and increasingly belong in different places.

The distinction is now explicit in how sophisticated teams plan. The fiscal strategy must differentiate between the capital-heavy phase of model training and the operational-heavy phase of inference. The pattern that has emerged is elegant: initial training happens in the cloud for its elasticity, but the final, high-volume inference is moved to on-premise or colocation grids to stabilize monthly spend.

This makes sense once you see it. Training is intermittent and spiky, you need vast compute for a while, then you do not, which is exactly what cloud elasticity is for. Inference is the opposite: it runs continuously, at high volume, for as long as the product lives. Paying cloud premiums for a workload that never stops is paying for flexibility you are not using.

PLAXONIC'S VIEW

For most enterprises deploying AI in 2026, the highest-leverage financial decision is not which model to use, it is where inference runs. Getting the training-in-cloud, inference-close-to-home split right is often the difference between an AI product that is economically sustainable and one that quietly bleeds margin with every request served.



Geopatriation: sovereignty becomes a force of its own

For years, "sovereign cloud" was a compliance checkbox. In 2026 it has become a distinct driver of architecture, operating on completely different logic from cost-driven repatriation.

The literature has given it its own name to separate it from cost. Geopatriation describes workloads moving back from globally distributed providers, typically US-parent hyperscalers, to sovereign regional clouds whose controlling legal entity is incorporated in the customer's own regulatory zone. The distinction is sharp: repatriation was a cost discussion; geopatriation is a jurisdictional one, where buyers run a legal exposure model against the EU AI Act, DORA, and Schrems II and decide sensitive workloads cannot live in US-controlled infrastructure.

This is driven by real, active regulation. The environment includes strict data laws like India's DPDP Act and Europe's GDPR driving geopatriation, and Gartner predicts that by 2030 a significant portion of enterprises will move workloads into regional or sovereign clouds. It carries a genuine cost: hyperscalers are charging up to 30% more for their sovereign cloud offerings, making sovereignty a real trade-off, not a free win.

PLAXONIC'S VIEW

Sovereignty is now an architecture requirement, not a legal footnote, and it must be designed in from the first diagram. For any organization handling regulated data across jurisdictions, particularly in India, Europe, and the Middle East, where we do significant work, which legal entity controls the infrastructure your data sits on is now as important as the uptime SLA. Teams that treat this as a design input win; teams that treat it as an audit-time checkbox rebuild.



Infrastructure FinOps grows up

The most quietly consequential trend of 2026 is that cost discipline has become a first-class engineering practice, because the numbers finally forced the issue.

After years of improvement, the trend reversed. Estimated wasted cloud spend ticked up to 29%, reversing a five-year downward trend, with complexity compounded by simultaneous migration and repatriation, SaaS proliferation, and the rapid adoption of AI. Organizations responded by formalizing the discipline: adoption of Cloud Centers of Excellence rose to 71%, and FinOps team prevalence climbed to 63%.

Its nature is changing too. The hidden tax of the cloud, specifically egress fees and the high cost of GPU idling, has led to the rise of Infrastructure FinOps, and a roadmap that does not account for unit-cost economics is destined for failure. The payoff is real: formal FinOps frameworks typically deliver a 20 to 30% reduction in unallocated waste within six months.

PLAXONIC'S VIEW

FinOps in 2026 is not a cost-cutting exercise you run when the bill gets scary. It is the discipline that makes every placement decision in this report possible, because you cannot decide whether a workload belongs in the cloud or on-prem until you know its unit economics. It is telling that this is one of the few areas where the honest answer to a client is sometimes "you are spending correctly, do not change anything."



Hybrid and multi-cloud become the default architecture

The debate about whether to be hybrid or multi-cloud is over. In 2026, nearly everyone is, and it is now a question of managing that reality well rather than deciding whether to enter it.

The figures are near-universal. Vendor lock-in is the leading concern for CTOs this year, and to mitigate risk, 90% of large organizations have now adopted a hybrid or multi-cloud approach, keeping sensitive data on-premises while leveraging specialized AI tools from different public providers. Multi-cloud is still inching upward.

The honest part is the cost: complexity. Being multi-cloud means more surfaces, more billing models, more security postures, and more integration. The naive version, spreading workloads across providers and hoping for the best, creates more problems than it solves.

PLAXONIC'S VIEW

Multi-cloud is now a given, but it is only an advantage if you have the operational maturity to run it. The organizations that benefit use multiple clouds deliberately, the right workload on the right provider for real reasons, with unified observability, governance, and cost management across all of them. Those who suffer ended up multi-cloud by accident and carry the complexity without the benefit. The differentiator is not whether you are multi-cloud. It is whether you meant to be.



The edge becomes tier-1 infrastructure

The edge has graduated. What used to be a fringe consideration for IoT sensors is now a Tier-1 infrastructure component, no longer just a collection of IoT sensors.

Two forces push compute to the edge simultaneously, and they reinforce each other. The first is AI inference that needs to happen close to where data is generated, for latency and cost. The second is sovereignty: enterprises are deploying "micro-grids" of compute, small high-performance clusters located within specific legal jurisdictions to ensure compliance with data sovereignty mandates. The edge is where the AI-cost story and the sovereignty story physically converge.

PLAXONIC'S VIEW

The edge in 2026 is where three of this report's trends meet: AI inference wanting to run cheaply and close to the data, sovereignty wanting data to stay in-jurisdiction, and FinOps wanting to escape cloud egress fees. For the industrial, utility, and public-sector systems we build, this is not abstract, it is the difference between a decision made in milliseconds on the floor and one that round-trips to a distant region and arrives too late to matter.



AI-native platforms change who builds software, and how fast

The report closes where it began, with how software gets built, because the same AI reshaping cloud economics is reshaping the tools of creation.

Gartner identifies AI-native development platforms as a top strategic trend for 2026, letting small, nimble teams build software using generative AI faster than ever. The forecast is striking: by 2028, 75% of enterprise software engineers will use dedicated AI code assistants. And Forrester notes 2026 marks the early fallout of the battle for agentic supremacy, where specialized cloud players challenge hyperscalers with more affordable, sovereignty-focused solutions.

There is a real constraint that deserves honesty. Over 90% of organizations are expected to face severe IT skills shortages this year, particularly in cloud security and AI/ML integration. The technology is racing ahead of the talent to deploy it responsibly.

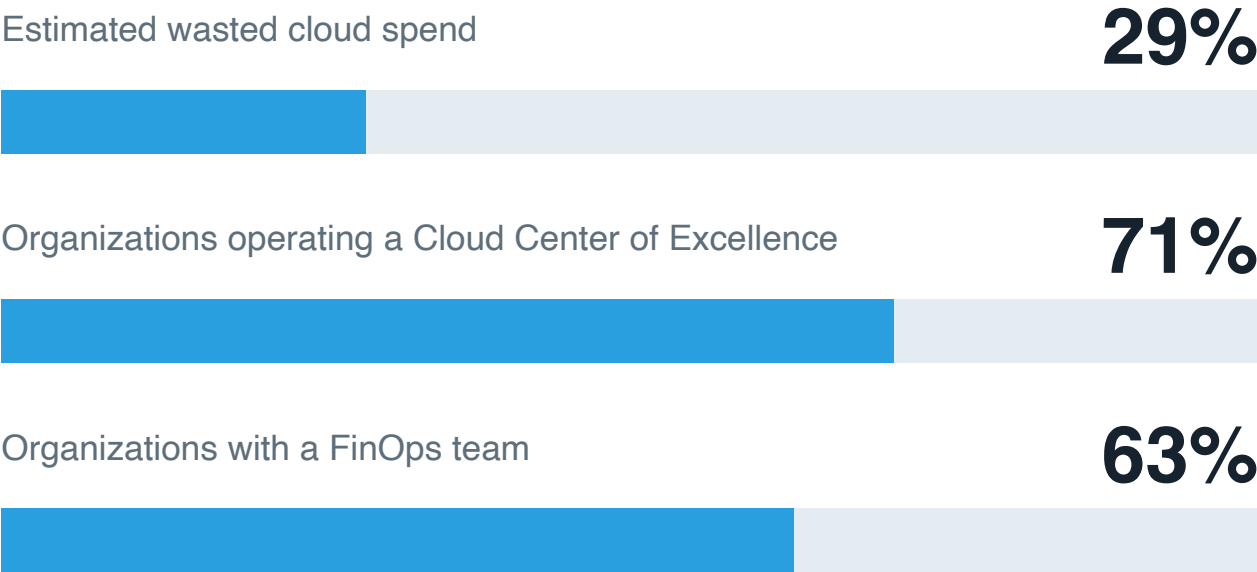
PLAXONIC'S VIEW

AI-native platforms are genuinely accelerating delivery, but the talent gap is the real story underneath. The bottleneck in 2026 is not the tooling, it is the shortage of people who can wield it well, especially at the intersection of cloud, security, and AI. This is precisely why build-versus-buy and cloud-versus-on-prem decisions have become harder, not easier: the tools are more powerful, but the judgment to use them well is scarcer.



Figure 1: Cost discipline becomes a formal function

Organizations responded by formalizing the discipline: adoption of Cloud Centers of Excellence rose to 71%, and FinOps team prevalence climbed to 63%.

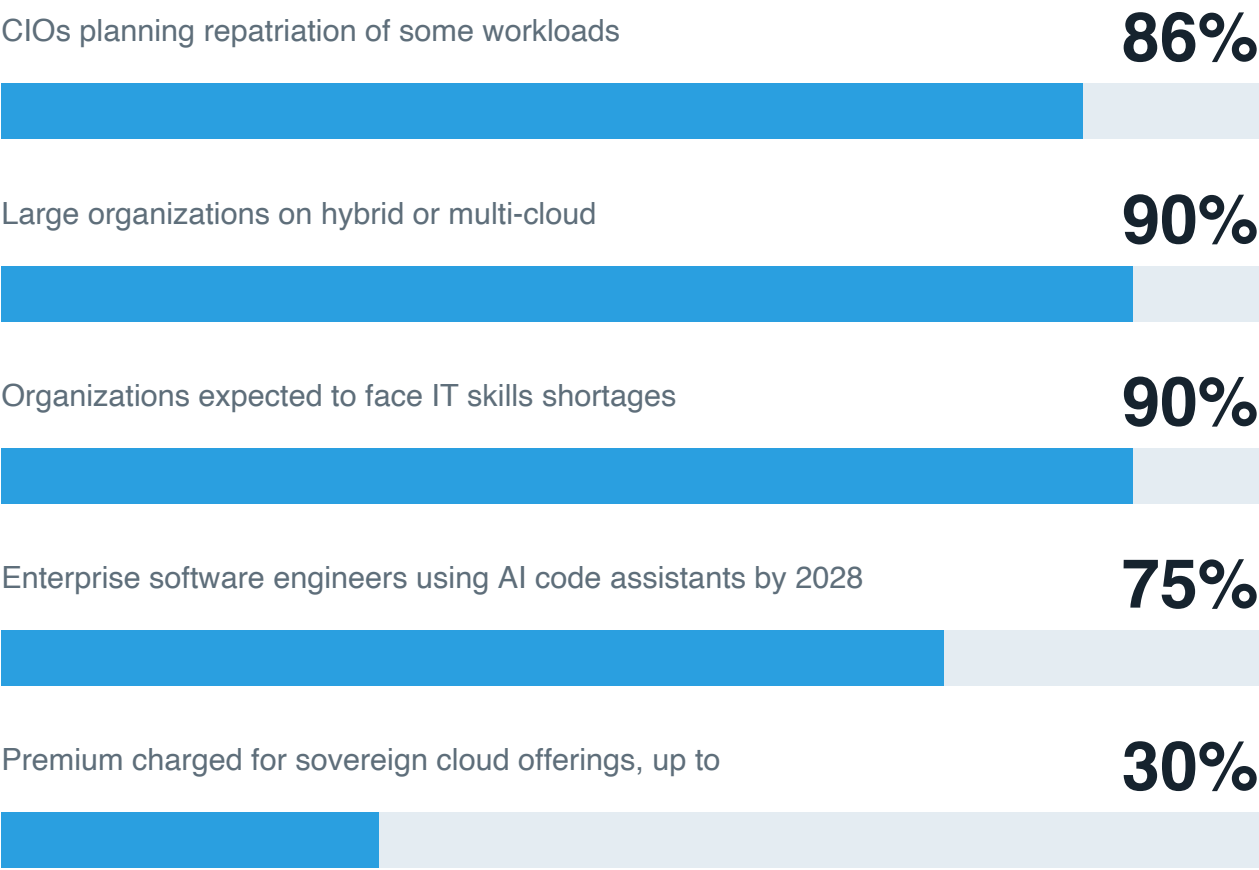


Source: Flexera 2026 State of the Cloud Report (N=753). See References 1 and 2.



Figure 2: Placement pressure, by the numbers

The winners in 2026 will not be the companies most "in the cloud" or most "out of it." They will be the ones who treat placement as a per-workload decision, and build the capability to move.



Sources as listed in the References section. These figures are drawn from separate studies covering different populations and time periods, and are not directly comparable with one another.

How to decide: a placement framework for 2026



If there is a single message in this report, it is that 2026 replaced a default with a decision. Here is the sequence of questions we work through with clients for any significant workload.

1. What is the workload's true profile?

Bursty and unpredictable, or steady and high-volume? Bursty favors cloud; steady and high-volume increasingly favors owned or colocated infrastructure. For AI, this usually splits training (cloud) from inference (closer to home).

2. What does the data's jurisdiction demand?

Before cost, ask whether regulation dictates where this data can legally live and which legal entity may control it. Sovereignty is a constraint, not a preference, and it can override the cost answer entirely.

3. What are the real unit economics?

Not the list price, the actual all-in cost including egress, idle GPU time, and operational overhead. You cannot decide this without FinOps visibility, which is why FinOps comes before the placement choice, not after.

4. What is the cost of being wrong, and how fast can you reverse it?

The best placement decision is one you can undo. Portability, infrastructure as code, and avoiding deep proprietary lock-in keep every decision reversible as the economics shift.

5. Do you have the people to run it?

Given the talent shortage, an architecture your team cannot operate securely is the wrong architecture, however good it looks on paper.



IN CLOSING

The organizations that will lead in 2026 are not the most cloud-committed or the most cloud-skeptical. They are the ones who made placement a deliberate, repeatable, reversible decision, and built the capability to change their mind as the ground shifts under them. In a year when the ground is shifting this fast, that capability is the real competitive advantage.



Resources and references



Every quantitative figure in this report is traced below to a named third-party source. Where a figure as stated in the report differs from the figure as published by its original source, the difference is noted. In those cases the original source should be treated as authoritative.

1. Flexera, 2026 State of the Cloud Report, published 18 March 2026. Survey of 753 cloud decision-makers worldwide. Source for the estimate that 29% of cloud spend is wasted, and for the reversal of a five-year downward trend. Cited in Key Takeaways and Trend 5.
2. Flexera, 2026 State of the Cloud Report. Source for Cloud Center of Excellence adoption at 71% and FinOps team prevalence at 63%. Cited in Trend 5.
3. Flexera, 2026 State of the Cloud Report. Source for continued upward movement in multi-cloud adoption. Cited in Trend 6.
4. IDC, Assessing the Scale of Workload Repatriation, June 2024. Approximately 80% of respondents expected some repatriation of compute and storage resources within twelve months.
5. IDC, Server and Storage Workloads Survey, October 2024. Approximately 8% of organizations were moving entire workloads off public cloud, indicating selective rather than wholesale repatriation.
6. E2B. Code-execution sandbox platform built on Firecracker microVMs, with sandbox start times under 200 milliseconds and per-sandbox kernel, filesystem and network isolation. Source for the isolation and boot-time description in Trend 1.
7. Gartner, Top Strategic Technology Trends for 2026, press release, 20 October 2025. Source for AI-native development platforms as a named strategic trend, cited in Trend 8, and for geopolatriation as a named strategic trend, cited in Trend 4. Gartner coined the term geopolatriation and placed it on this list.
8. Gartner, Top Strategic Technology Trends for 2026. Prediction that by 2030 more than 75% of European and Middle Eastern enterprises will geopolatriate their virtual workloads into solutions designed to reduce geopolitical risk, up from less than 5% in 2025. Cited in Trend 4.



9. Gartner, press release, 12 November 2025. Survey finding that 61% of Western European CIOs and IT leaders expect geopolitical factors to increase reliance on local or regional cloud providers, and prediction that by 2030 more than 75% of enterprises outside the United States will have a digital sovereignty strategy.
10. Gartner, Gartner Says 75% of Enterprise Software Engineers Will Use AI Code Assistants by 2028, press release, 11 April 2024. Based on a third-quarter 2023 survey of 598 global respondents. Cited in Trend 8.
11. Forrester, Predictions 2026: Cloud Computing, report RES185003. Source for the characterization of 2026 as the early fallout of the battle for agentic supremacy, with sovereignty concerns creating openings to challenge hyperscalers. Cited in Trend 8.
12. Forrester, 2026 Technology and Security Predictions, 28 October 2025. Related forecast that neoclouds will capture 20 billion dollars in revenue as specialized providers expand sovereign AI offerings.
13. IDC, Enterprise Resilience: IT Skilling Strategies, 2024, Doc #US52080524, May 2024. Prediction that by 2026 more than 90% of organizations worldwide will be affected by the IT skills crisis, at an estimated cost of 5.5 trillion dollars. Cited in Trend 8.
14. Deloitte, published guidance on FinOps tooling and cloud spend. Source for the finding that organizations implementing formal FinOps frameworks typically achieve a 20 to 30% reduction in unallocated waste within six months. Cited in Trend 5.
15. Boston Consulting Group, cost of cloud research, as reported by Civo, 2026. Source for the finding that hyperscalers charge up to 30% more for sovereign cloud offerings. Cited in Trend 4.

Note: Independent estimates of the sovereign cloud premium range between 10% and 30% depending on provider and configuration.

About Plaxonic



Plaxonic is a global AI-first engineering and product company. We help organizations turn ambition into working systems: intelligent, cloud-native, and automation-driven software built to perform in production, not just in a demo.

Our work spans the full journey of a modern digital system. We build production-ready AI that makes it into daily operations, engineer enterprise software that businesses depend on, and design scalable cloud platforms that hold up under real demand. Where legacy systems hold companies back, we modernize them. Where operations are slow or manual, we streamline and automate them.

With globally distributed teams, we bring together the engineering depth and range that complex problems require, working alongside our clients rather than at arm's length. The result is technology that is built for scale and measured by outcomes: systems that run reliably, operations that cost less to run, and digital products that move the business forward.

We build what actually ships, and we make sure it delivers once it does.



WEB www.plaxonic.com

LINKED IN linkedin.com/company/plaxonic

Disclaimer: The material in this document reflects information available at the time it was prepared, as indicated by the date on the front page. Conditions across cloud and AI infrastructure are changing rapidly and the position may change. This content is provided for general information purposes only, does not take account of any reader's specific circumstances, and is not intended to be used in place of consultation with a qualified professional adviser. Plaxonic disclaims, to the fullest extent permitted by applicable law, any and all liability for the accuracy and completeness of the information in this document and for any acts or omissions made on the basis of such information.

Figures cited in this document are drawn from named third-party sources listed in the References section and have not been independently verified by Plaxonic. A number of them are analyst forecasts rather than measurements of current conditions, and forecasts are frequently revised. Sections marked Plaxonic's View are opinion, offered as perspective rather than as findings.

Plaxonic does not provide legal, regulatory, audit, or tax advice. Readers are responsible for obtaining such advice from their own legal counsel or other licensed professionals.

This document refers to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied. Gartner does not endorse any vendor, product or service depicted in its research publications.

Copyright © 2026 Plaxonic. All rights reserved.