2024 Planning Guide for Cloud, Data Center and Edge Infrastructure

Published 4 October 2023 - ID G00796448 - 48 min read

By Analyst(s): Simon Richard, Lydia Leong, Matthew Brisse, Tony Iams, Adrian Wong, Mohini Dukes, Stanton Cole, Jorge Aragon, Alexandru Giurovici, Ajay Chauhan, Douglas Toombs, Paul Delory, Ross Fomerand, John Fullbright, Paul DeBeasi, Ang Troy

Initiatives: Infrastructure for Technical Professionals

> Infrastructure and operations teams must pivot to support such trends as generative AI and the shift toward platforms, while navigating continued economic uncertainty. I&O technical professionals responsible for the cloud, data centers and/or the edge should focus on four technical planning trends.

Overview

Key Findings

- Cloud deployments built to prioritize business outcomes have accumulated technical debt through suboptimal technical choices. Paying down this debt should be prioritized to receive a directed effort.
- The demand for simplicity from end users and internal customers is pushing infrastructure and operations teams to build productized, easily consumable infrastructure platforms via a platform engineering approach.
- Rapid advances in artificial intelligence, especially generative AI, have ignited an urgent need to deploy and manage appropriate infrastructure in the cloud and on-premises, despite infrastructure and operations teams' struggles to keep pace with the velocity of AI change.
- Increasingly stringent latency and operational requirements for distributed applications and infrastructure, combined with geopolitical shifts, have increased the consideration of emerging technologies that address these requirements. These include edge, 5G private mobile networks and distributed cloud solutions.

Recommendations

In 2024, I&O technical professionals involved in cloud, data center and edge infrastructure should:

- Begin paying down cloud technical debt through an organized effort, rather than ad hoc fixes. Ensure that this includes plans and methods to adjust to a continual lack of cloud skills, to implement and optimize governance, and to apply cloud architecture best practices.
- Address GenAI requirements with off-the-shelf, cloud-based solutions first, if possible. If these off-the-shelf solutions fail to meet requirements, consider implementing GenAI solutions via public cloud infrastructure as a service or on-premises infrastructure, using a reduced-scale model.
- Focus on how they design and build infrastructure by prioritizing developer and user experience. Shift designs toward an infrastructure platform approach, rather than hobbling efforts with a traditional infrastructure model.
- Focus or reframe edge and emerging use cases around requirements for data management, security and sovereignty. Although network bandwidth and latency are important factors, they should not always take priority.

Strategic Planning Assumptions

- By 2027, more than 75% of the Fortune 1000 companies will have formal infrastructure platform organizations, which is a significant increase from fewer than 20% in 2023.
- By 2027, 80% of large organizations must embrace platform engineering to successfully scale DevOps initiatives in hybrid cloud environments, which is an increase from fewer than 30% in 2023.
- By 2028, large enterprises will triple their unstructured data capacity across their on-premises, edge and public cloud locations, compared with mid-2023.
- By 2026, storage-consumption-based platform, service-level agreement (SLA) guarantees will replace more than 50% of traditional on-premises IT capacity management, budgeting, assessment, sourcing and fulfillment activities, which is an increase from fewer than 15% in 2023.
- By 2027, 60% of new artificial intelligence (AI) clusters will use function accelerator cards to manage and control network bandwidth, jitter and latency, which is a major increase from 5% or fewer today.
- By 2025, 20% of 4G/5G private mobile networks will deploy edge computing platforms to optimize bandwidth, reduce latency and provide better security, which is an increase from 5% in 2022.
- By year-end 2026, 70% of large enterprises will have a documented strategy for edge computing, compared with fewer than 10% in 2023.

Cloud, Data Center and Edge Infrastructure Trends

2024 Planning Guide Web Graphics

In 2024, macroeconomic, geopolitical, and other uncertainties will continue to affect business' long-term planning, and short-term tactics. In this climate, infrastructure may not be the first target for organizations to re-evaluate, improve and optimize. However, ignoring infrastructure leads to long-term inefficiencies and incidents, stifling innovation. Organizations that leave infrastructure to languish in its current form will find it increasingly difficult to deliver fundamental IT services that meet business requirements. This will hinder the adoption of emerging technologies.

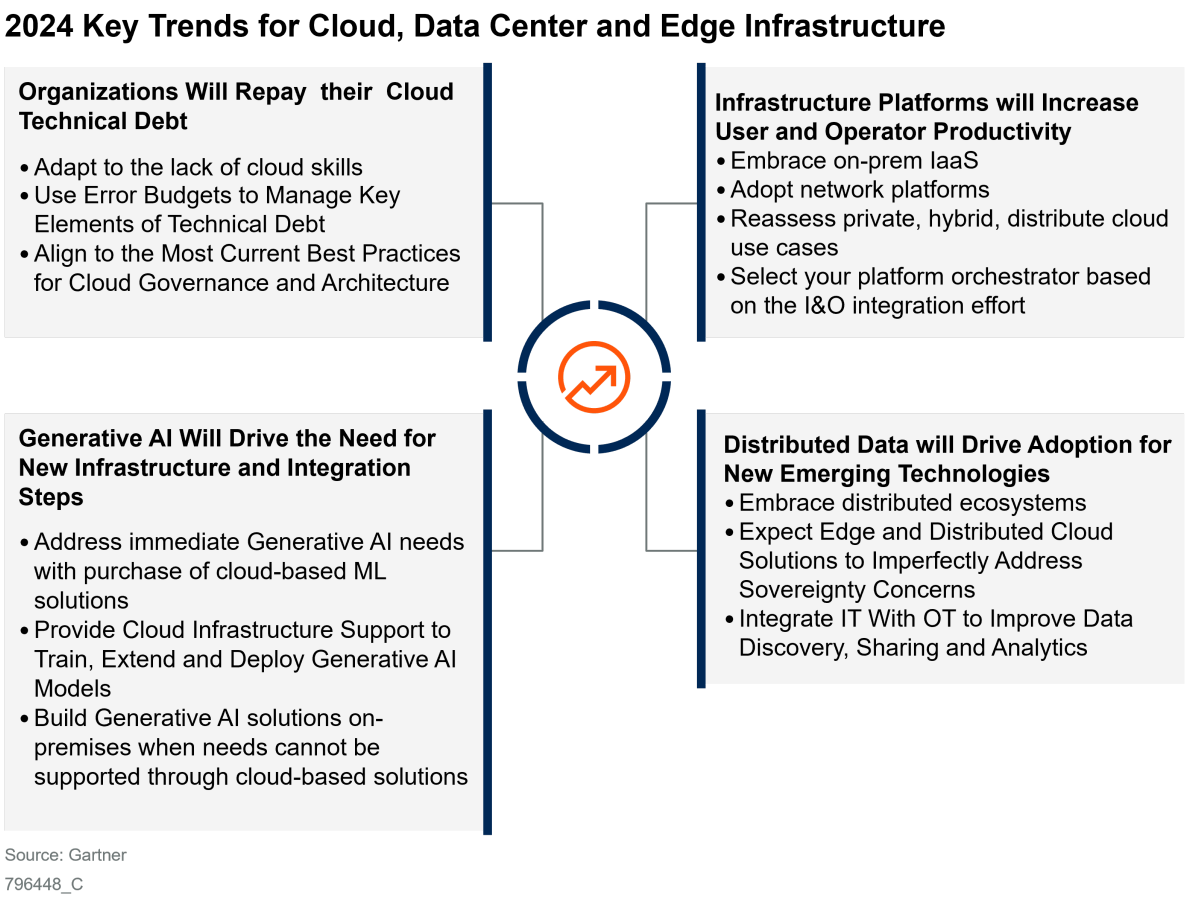
Challenges that infrastructure and operations (I&O) teams will face in 2024 include:

- **Cost management:** I&O teams have traditionally been asked to do more with less. In addition to managing their run rate, they need to harness the cost of their growing portfolio of services, and to govern diverse services and the increasing adoption of the cloud.
- **Increased geographical distribution of workload:** The global demand of customers for low-latency access and requirements by regulators and governments for control have led to competing requirements for privacy, sovereignty and distribution. I&O organizations need to increase the distribution of workloads, while managing the delicate balance among these priorities.
- **Skills shortage:** Most organizations will not be able to upskill their staff or hire to acquire all the skills they need. Cloud skills are expensive, and AI skills are scarce and becoming ever-more competitive. I&O organizations will need to accept the lack of skill and adjust to it.
- **Spiraling complexity:** Not only will I&O need a streamlined way to deliver and integrate services, but they also need to simplify how their customers consume services. There is an imperative to reduce the cognitive load of both the operators and the users of the system that I&O delivers.
- **Inflation:** Most countries expect widespread inflation in 2024, even as I&O teams' budgets may stagnate or shrink. Here too, I&O teams must focus on cost efficiency. Businesses will also need to revise salary expectations, if employee wages rise.

This year has been a time to refocus, retool and rethink your infrastructure. Although these activities will persist and encourage positive changes, new opportunities and challenges will arise. Businesses are rapidly advancing their digital strategies and eagerly anticipate IT's ability to deliver innovative technologies, such as generative AI (GenAI). However, these technologies will introduce challenges for more-traditional I&O teams. Developers also clamor for easily consumable infrastructure resources so they can take advantage of new capabilities as well. Transitioning to a new I&O approach, such as platform engineering, is no longer a "nice-to-have," but a necessity to meet business objectives and remain competitive.

In 2024, cloud, data center and edge infrastructure will be shaped by four long-term trends, which should inform the decisions of I&O technical professionals (see Figure 1).

Figure 1: 2024 Key Trends for Cloud, Data Center and Edge Infrastructure



Organizations Will Repay Their Cloud Technical Debt

The public cloud market is more than 15 years old. Regardless of when an organization first began to adopt cloud computing, it is likely to have accumulated some cloud-related technical debt. The probability of this technical debt increased during the COVID-19 pandemic, because many organizations radically accelerated cloud adoption to respond to an uncertain business environment that was changing rapidly. Three types of technical debt are likely to demand action (see Case Study: Stakeholder Incentives for Technical Debt Management [Intel]):

- **Deliberate:** Due to business necessity, a technical trade-off or shortcut was made during implementation, but an explicit plan was made to take corrective action in the future, and the governing metrics reflect this trade-off.

- **Accidental:** Due to a lack of necessary skills or knowledge, something was done incorrectly or suboptimally. The organization may or may not be aware of the specific issue, and the issue may not be reflected in metrics or key performance indicators (KPIs).
- **Rotten:** Without adequate business rationale, a deliberately poor choice was made that unnecessarily increased cost, complexity or risk. There was no plan for later corrective action, and metrics may have been selected to obscure the impact and cost of poor decision making.

Technical professionals should strive to identify, document and prioritize outstanding technical debt. High-priority accidental and rotten technical debt should be investigated, so that executable remediation plans can be created.

Cloud technical debt typically falls under the responsibility of the noted groups:

- **Cloud center of excellence (CCOE):** CCOEs are responsible for governance issues. Inadequate governance has led to unnecessary risk, excess cost or chaos.
- **Cloud operations:** This function is responsible for process, maintenance and life cycle issues. Suboptimal management has led to problems in the service management life cycle, unnecessarily high labor costs due to insufficient automation, and unnecessary risks and/or complexity.
- **Cross-functional:** Cross-functional cooperation is necessary to address implementation issues. Suboptimal architecture or implementation has led to stability, performance, cost or complexity problems. If necessary, the CCOE should sponsor a program to address these issues.

In 2024, I&O technical professionals should:

- Adapt to the lack of cloud skills
- Use error budgets to manage key elements of technical debt
- Align with the most current best practices for cloud governance and architecture

Planning Considerations

Adapt to the Lack of Cloud Skills

Many technical professionals review good practices for implementing cloud solutions, and realize that their organization lacks the necessary skills and knowledge, or there aren't enough skilled employees to perform the work that needs to be done. Gartner has observed that cloud adoption in such organizations may proceed nevertheless, driven by the business and possibly with third-party assistance. However, one of two things is likely to happen:

- Although the business proceeds, IT enters a state of paralysis, in which no useful cloud governance or long-term engineering work is completed, because the organization is unwilling to accept suboptimal completion of such work.
- A frantic and uncoordinated scramble ensues to deliver what the business needs, exploiting the heroics of individual employees who try their best to avert disaster.

Ideally, organizations would instead pursue some form of compromise solution, in which they:

- Methodically execute the "must haves" in the short term
- Build long-term plans
- Work toward long-term solutions

This requires appropriate staffing and budgeting, not merely hope. It also requires a careful consideration of immediate costs versus the available budget, the long-term total cost of ownership (TCO) and risk management. Quick, but thoughtful, prioritization of what can be practically accomplished is the key to these compromises.

> **It is reasonable to incur deliberate technical debt, as long as it is clearly identified and accounted for, and an executable plan to remediate that debt exists and will be funded at a known date.**

Organizations must be pragmatic and do what they can with the resources they have — but it is also critical for technical professionals to communicate when critical goals are not achievable with the resources available.

In addition to ruthless prioritization, extending timelines or reducing scope, consider the following:

- **Bring in third-party assistance:** Consultants and contractors can be useful for staff augmentation and may be used to obtain vital missing skills. However, only long-term employees should make strategic decisions for the organization.
- **Transition with a cloud managed service provider (MSP):** Many cloud MSPs have operations transition offerings. These typically include managed services for one to three years, a suite of cloud automation capabilities, and supporting a cloud-optimized or cloud-native style of management. Most importantly, they include assistance in reskilling staff through classroom and on-the-job training, with MSP personnel serving as senior mentors. The customer gradually takes over operations from the MSP during the term of the contract.
- **Use hybrid cloud operations to bridge the skills gap:** Workloads that have a low rate of change and were "lifted and shifted" (that is, rehosted) to the cloud are the best candidates for hybrid cloud operations. This is because their management in the cloud will still be effective if it remains largely unchanged from on-premises. This reduces cloud benefits and has an unattractive long-term TCO, but it also frees skilled personnel to focus on cloud-optimized or cloud-native workloads.
- **Push responsibilities to cloud consumers, such as application teams:** Cloud-native applications or other teams with significant cloud skills, along with DevOps experience and expertise, may be able to take on more responsibilities throughout the service life cycle. However, governance remains necessary, and there must be appropriate risk management. Such teams benefit from direct access to the cloud provider's technical support engineers.

Benefits

- Ability to proceed with cloud adoption, despite skill limitations, rather than falling into paralysis.
- Resets unrealistic expectations.
- Explicitly recognizes the trade-offs of short-term solutions and invests in executing on the long-term plan.

Cautions

- Transitional approaches can turn into suboptimal long-term operations, if they are not backed by strong efforts to continue to innovate or to transform the organization.
- Pursuing near-term methods to address cloud skill gaps should be paired with efforts to build a talent pipeline, but could be used as a reason to avoid skill development.

Recommended Reading

- Quick Answer: How Do I Overcome a Lack of Cloud Skills in My Organization?
- Bridge Cloud Skills Gaps With a Transitional Hybrid Cloud Operations Model
- Introducing the "You Build It, You Run It" Modern Operations Pattern

Use Error Budgets to Manage Key Elements of Technical Debt

In site reliability engineering (SRE), an "error budget" is the number of errors a service can accumulate over a specified period of time before service quality becomes unacceptable. When an application team has exceeded its error budget, it is not allowed to release any new features. Instead, the team must work on system stability, until the service again achieves its service-level objectives (SLOs).

Although error budgets are traditionally used to manage service availability, some Gartner clients have found that they can also be useful for managing cloud technical debt. Application teams often struggle to prioritize addressing technical debt in their backlog, because they are rewarded primarily for the release of new features, rather than stabilizing existing implementations. Application teams (and other technical teams) with significant responsibility for managing their cloud resources typically accumulate three key types of technical debt:

- **Availability:** Many organizations struggle with the balance between release velocity and system stability. This is the classic use case for error budgets.
- **Cost:** Continuous cloud financial management (CFM), sometimes implemented as "FinOps," frequently identifies cost inefficiencies that application teams can address through more efficient resource management, application modernization or performance optimization.

- **Security:** DevSecOps processes frequently identify security issues that application teams need to address. These may include cloud configuration issues, as well as vulnerabilities that exist in the infrastructure or application itself.

Each of these types of technical debt should be given its own error budget. Exceeding the error budget of any one of the types results in a halt to new feature releases and a requirement to address the accumulated debt.

In addition, there should be a *combined* error budget across the three individual budgets. Exceeding that combined error budget indicates that the overall level of technical debt is too high, halting feature development until enough debt is repaid.

Benefits

- Expanding the error budget concept for the quantification of technical debt enables the organization to manage toward a simplified single metric. This facilitates application team autonomy while aiding risk management.
- Error budgets can help the business owners of applications understand the importance and extent of technical debt. This makes it easier to get the business to agree to devote sprints to paying down technical debt.

Cautions

- Error budgets are only useful when the organization strictly enforces the requirement to pay down technical debt when the error budget is exceeded.
- Forms of technical debt that are not measured by the error budget metrics may be undesirably deprioritized.
- Error budgets should be tailored to each application. A “one size fits all” approach can result in a poor alignment with desired technical and business outcomes.

Recommended Reading

- Assessing Site Reliability Engineering (SRE) Principles for Building a Reliability-Focused Culture
- How Software Engineering Teams Should Work With Site Reliability Engineers

Align to the Most Current Best Practices for Cloud Governance and Architecture

Cloud governance has long been a challenging topic for organizations to address, frequently due to a mismatch of internal expectations and assumptions between decision rights and authority in the organization. When improperly governed, the conflicting goals of speed and agility versus stability and predictability create chaos in the organization.

As with any governance problem in an organization, the longer it is left unaddressed, the larger the problem tends to become. Organizations that did not start their cloud adoption with strong governance must optimize and automate their cloud governance in 2024. This is critical to ensuring that operations can be scaled effectively long-term.

If you do not have effective cloud governance practices — or you are still using the practices you established many years ago — adopt the most up-to-date cloud governance practices. There is no prerequisite to follow the evolution of cloud best practices. Skip to the most up-to-date practices and tools, as these have been shaped and refined by the lessons learned in their evolution. This will help you avoid the inadvertent accumulation of new technical debt. Some of the more impactful and proven practices that do not require evolutionary prerequisites are:

- **Implement CFM as a cultural practice, not a dedicated FinOps team.** Furthermore, ensure you have a business case to perform continuous CFM, rather than one-time "cloud cost hygiene" projects, to avoid wasting time and money on activities that will not have an adequate ROI.
- **Implement application ID tags and manage other metadata via a database, rather than trying to control a complex tagging scheme.** Mandate an application ID tag, but otherwise allow development teams to use tags as they wish in their environments. The application ID serves as a primary key that allows the application to be associated with other metadata in a database that can be more easily updated, as well as queried by management tools.
- **Use cloud security posture management (CSPM) to manage cloud policies at scale or within a multicloud strategy, rather than building manual workflows.** High-quality CSPM tools will leverage the cloud-native identity and access management (IAM) and policy capabilities. At the same time, it will handle workflow or integration with IT service management (ITSM) tools to manage granting exceptions, extensions to remediation deadlines and other complexities.

- **Tailor landing zones to specific needs, instead of taking a one-size-fits-all approach.** Development, testing and nonproduction data science environments do not need the same controls as production environments. “Spoke” landing zones (environment-specific landing zones connected to a central “hub” landing zone for the organization) should fit the needs of the team, the application and the application’s life cycle stage in that environment.
- **Separate self-service provisioning from self-service operations.** A team’s eager embrace of self-service access to cloud resources in development and testing does not imply that a technical team desires responsibility for operations, is capable of safely performing operations or benefits from doing so.

Benefits

- Building governance principles through a collaborative effort helps everyone understand that governance is the responsibility of everyone in the organization who handles applications or data.
- Strong automated controls and remediations can enable organizations to dramatically speed up agility and cloud adoption, while providing significant protection for the organization.
- More deliberate and controlled adoption of cloud capabilities lowers overall risk exposure for the organization.

Cautions

- Cloud providers tend to be overly permissive to users, as it serves their interest for customers to use as many services as possible. This makes it challenging to rein in permissions for user communities that have become accustomed to having “administrator level” rights.
- Governance controls that stand in the way of a user will often trigger a reflexive response to simply try and work around the problem to continue forward. This is why the cloud governance team must be imbued with appropriate authority in the organization to set and enforce hard rules on what is acceptable.

Recommended Reading

- Solution Path for Public Cloud Governance

- How to Empower Technical Teams Through Self-Service Public Cloud IaaS and PaaS
- Cloud Architecture Best Practices: Landing Zones
- How to Protect Your Clouds With CSPM, CWPP, CNAPP and CASB
- Quick Answer: Do You Actually Need FinOps to Manage Your Cloud Costs?

Infrastructure Platforms Will Increase User and Operator Productivity

Infrastructure platform engineering is the discipline of building internal software products that present IT infrastructure to users (or other platforms) in an easily consumable way. Infrastructure platforms are self-service tools that enable users to deploy and manage infrastructure themselves, whereas I&O builds such platforms and retains the responsibility and implementation of governance, security and compliance. Infrastructure platforms are often used as the foundation of higher-order self-service layers, such as internal developer portals.

An infrastructure platform is more than mere automation. It is an abstraction layer between the user and complex underlying infrastructure that presents infrastructure capabilities in whatever way is most beneficial to the user's unique needs and skill set. Because of this, the specific capabilities of an infrastructure platform vary widely. However, all platforms will share certain basic characteristics. Your infrastructure platform should be:

- **Productized:** The platform is an internal product with internal customers. Building an infrastructure platform is an exercise in product management and customer-centric software development.
- **User-centric:** The platform solves the user's problem, not IT's problem. Users want to focus on what matters most to them, such as deploying an application and not worrying about detailed infrastructure functions.
- **Self-service:** Users should be able to acquire resources themselves, without opening tickets or other delays. User productivity is the primary benefit of infrastructure platforms.
- **Consistent and compliant**: The platform must have guardrails built in. Make it so users can't do anything that would violate corporate policies or best practices.

Platform engineering represents a departure from traditional ways of working and a new operating model for I&O that focuses on the creation and management of infrastructure platforms. In 2024, I&O teams should lay the foundations of their infrastructure platforms by building the first-level abstractions atop the underlying complexity of devices and tools, which higher-order abstractions can then reference.

To establish or extend a platform engineering practice, I&O technical professionals must:

- Embrace on-premises infrastructure as a service (IaaS)
- Adopt network platforms
- Reassess composable infrastructure
- Select a platform architecture based on the level of I&O effort

Planning Considerations

Embrace On-Premises IaaS

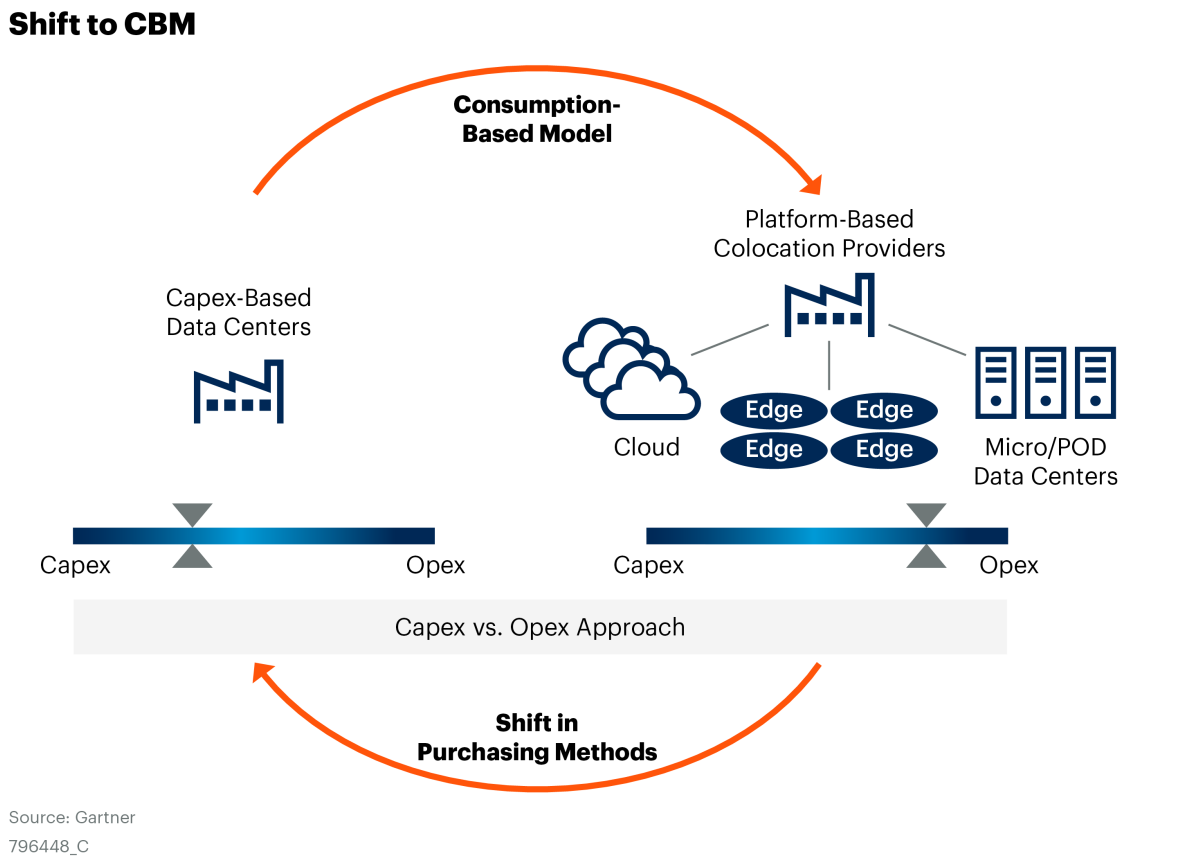
The public cloud has not displaced the data center entirely for many reasons. These include regulatory requirements, data gravity, the momentum of legacy infrastructure, limitations of staff skill sets, real-life deadlines and priorities and the reality that not all workloads belong in the cloud. However, IT organizations increasingly demand some of the benefits and operating models of the public cloud, even for those workloads that remain on-premises. Consequently, consumption-based models that mimic the infrastructure delivery and pricing models based on consumption have emerged for on-premises data centers. There are network, computing and storage-as-a-service (STaaS) solutions from a multitude of vendors. There are a variety of licensing and pricing plans balancing operating expenditures (opex) and capital expenditures (capex), based on the requirements of the customers. These models include:

- **Consumption-based models (CBM):** All the major data center hardware vendors now offer CBM programs. Many of these offerings include optional managed services.
- **Platform-based colocation:** Colocation platforms offer not just data center floorspace, but hardware as a service. They can orchestrate, instrument and provision single-tenant instances of networking, servers and storage in minutes or hours in specific physical locations spanning the globe.

- **Hyperconverged infrastructure (HCI):** The leading HCI products deliver the full stack of software-defined infrastructure (e.g., compute, storage and networking) and have become turnkey building blocks for private and hybrid clouds.
- **Distributed cloud:** Increasingly, the public cloud providers provide on-premises infrastructure as well. Distributed cloud — products such as AWS Outposts, Microsoft Azure Stack or the Google Distributed Cloud family — bring public cloud services to locations outside the provider's data centers.

Figure 2 illustrates these solutions in the context of the simultaneous shift from large-scale, owned data centers to colocation facilities and edge infrastructure.

Figure 2: Shift to CBM



Source: Gartner
796448_C

Gartner

Implications

- I&O prioritizes opex over capex for specific industries. On-premises IaaS is paid for on an ongoing basis, as opex, rather than via capex with large, and sometimes unpredictable, one-time bills.

- These delivery models mitigate supply chain risks. Because the infrastructure is delivered with excess capacity, customers have idle capacity on hand when they need to expand, rather than having to order it.
- On-premises IaaS requires a mindset shift. Evolving from sourcing products traditionally to an SLA-based operations model can be challenging not just for IT, but for many business units, such as procurement and finance.
- TCO for the same hardware is often higher, but may be offset by the operational benefits. Vendors now deal with the complexities of asset management and capital financing, which is a cost they, in combination with the costs of providing excess capacity, often pass on to the customer.

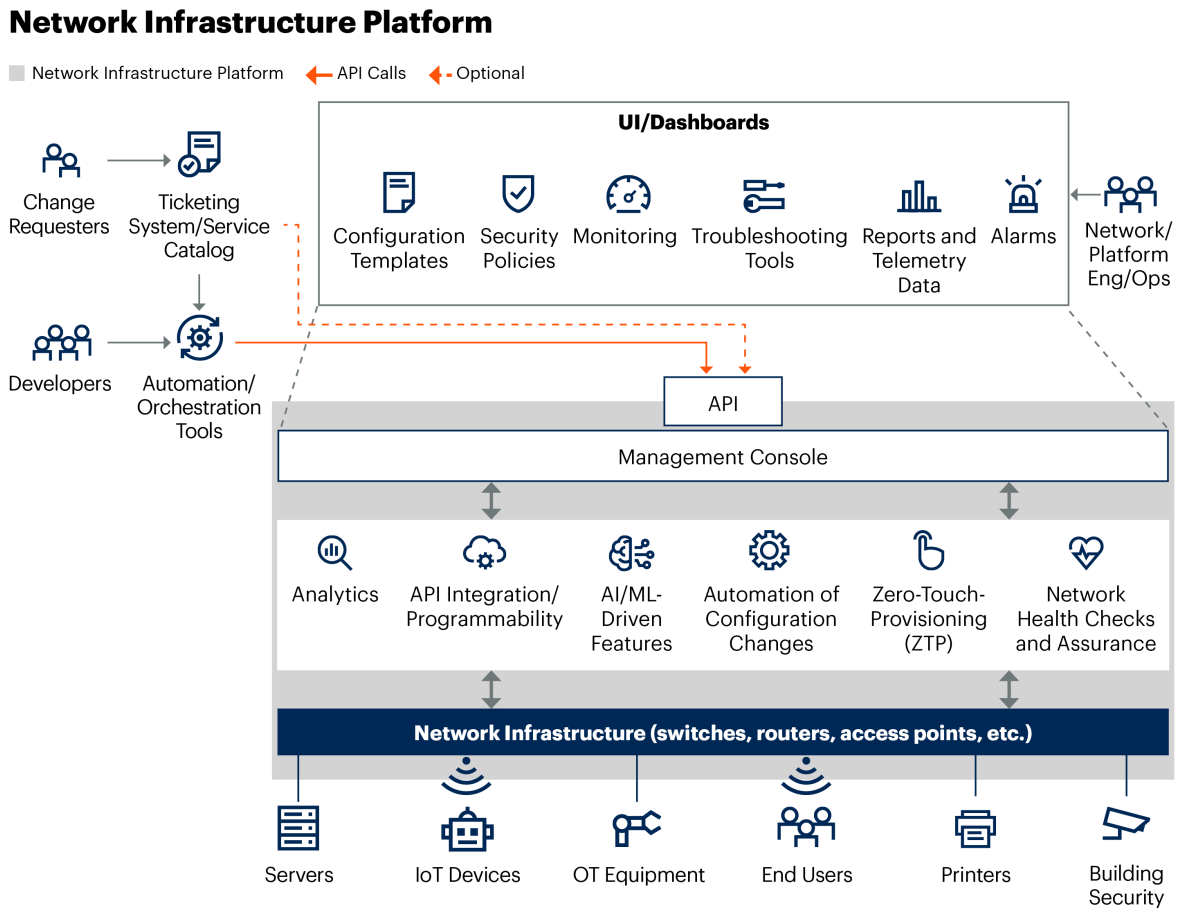
Related Research

- Toolkit for Estimating Data Center Build and Modernization Costs by Tier Level
- How to Evolve Your Physical Data Center to a Modern Operating Model

Adopt Network Platforms

Networking vendors have actively embraced platform engineering and strive to provide turnkey networking platforms, such as Cisco's Crosstalk. These platforms build multiple layers of abstraction on top of networking hardware and incorporate other software, such as tools for automation, observability and security. Users access network functions through a unified API with user-centric, intent-based semantics that the platform translates into network configuration automatically. This goes well beyond mere automation of network configurations. Figure 3 illustrates a typical network infrastructure platform.

Figure 3: Network Infrastructure Platform



On-premises networks have traditionally been rigid and have required manual or ad hoc changes when deploying new applications. Introducing automation to these environments is cumbersome or unreliable in some cases. Network platforms, however, offer a true cloud-like experience for on-premises networks. No longer will end users need to navigate the complexity of a collection of interconnected devices, open tickets for manual work by network engineers or worry that configurations are substandard.

Network platforms can be deployed for data center, campus/branch and software-defined wide-area networks (SD-WANs), as a unified platform or with separate platforms for each. Network platforms can be deployed alongside existing network gear for easier transitions. As older hardware is replaced, the new hardware can be onboarded into the network platform using zero-touch initial provisioning. Thus, network platforms can be implemented iteratively, even in brownfield environments.

Implications

- **Network platforms reduce the need for subject-matter expertise:** Users don't need to personally interact with the complexity of the underlying infrastructure, and platform operators/engineers can rely on the network API for routine changes and maintenance. If an issue arises that requires real domain expertise, then platform engineers/operators can simply open an incident ticket with the platform vendor to get support.
- **Network platforms are easier to integrate with other software:** Platforms make it easier to manage networks from other software, such as continuous integration/continuous delivery (CI/CD) pipelines, incident management tools or ITSM suites. The platform abstracts away the messy details of gathering an inventory of your network devices, IP addresses, the right configuration commands/syntax for each one and so forth. Rather, the entire network is exposed as a single API.
- **Traditional network management techniques are not suitable for network platforms, which are designed for use with configuration profiles and policies:** For example, use profiles, instead of per-device changes, when deploying network platforms. With these, you can create templates with baseline configurations that you can reuse on different parts of your network. This also enables you to build repeatable architectures. It also ensures that the underlay and overlay networks (tenants) are truly decoupled.

Related Research

- Best Practices for Modernizing Your Enterprise Network
- Solution Path for Evolving to Next-Generation Enterprise Networks

Reassess Composable Infrastructure

Composable infrastructure uses an API to create logical systems from shared pools of physical resources. Using a hardware fabric, it can connect disaggregated banks of processors, memory, storage devices and other resources. Composable infrastructure is an old idea, but it has been given new life for two reasons. First, high-speed interconnects, such as Compute Express Link (CXL), are finally available. Previously, it was impossible to sever the link between CPUs and memory, but new interconnects make it possible. Second, with the rapid adoption of GenAI and other flavors of AI, many IT organizations have an urgent need for graphics processing units (GPUs) and high-speed interconnects. Some may even deploy GPU fabrics on-premises in support of training models against large sets of local data. Composable infrastructure deserves consideration in light of these factors.

Composable infrastructure supports platform engineering by making compute and memory, including from GPUs, available programmatically. As with network platforms, compute fabrics can represent a first-level abstraction useful for infrastructure platform engineering efforts. Furthermore, GPUs are only part of the composable infrastructure story. Other non-x86 architectures continue to gain prominence. These include:

- Reduced instruction set computer (RISC) architectures, such as ARM and RISC-V, significantly enhance the value proposition of alternative CPUs.
- Function-accelerator cards (FAC) are a class of devices that have dedicated hardware accelerators with programmable processors to accelerate network, security and storage functions.

Including these new options, and others, as part of the composable infrastructure can enable superior efficiency for use cases that can take advantage of their unique properties. The ability to consume them from the same API as other infrastructure, and potentially have the composable infrastructure automatically select the right architecture, simplifies the user and operator experience. Table 1 offers some suggestions for where to incorporate non-x86 architectures into your environment and where to avoid them.

Table 1: Where to Incorporate Non-x86 Architectures Into Your Environment

Architecture	Consider For	Avoid For
ARM	Linux-based workloads; workloads based on microservices architecture; edge computing workloads	Windows servers; large, monolithic servers
GPU	AI/machine learning (ML) workloads; neuromorphic computing; analytics workloads; enhancing virtual desktop performance	General-purpose computing
FAC	Workloads that benefit from offloading functions from the CPU	Commodity workloads; non-performance-critical environments

Source: Gartner (October 2023)

The architectural requirements for these use cases are still emerging, and best practices are not yet well-understood, so I&O organizations may not actually deploy these technologies in 2024, but they should reassess composable infrastructure and emerging technologies. Furthermore, they should be prepared to deploy in 2025 and beyond, possibly on short notice.

Implications

- Composable infrastructure can form the bedrock of an infrastructure platform engineering initiative and provide substantial benefits to users. The availability of an API-addressable pool of resources provides a highly desirable layer of abstraction that can significantly improve user productivity.
- ARM-based silicon can offer advantages for certain workloads. It can be less power-hungry, which can support a green IT initiative, or it may offer higher price-to-performance ratios. Furthermore, specialized ARM chips can be especially appealing for AI/ML training and inference, even compared with GPUs.

- Choose the right architecture for your use case. Not all software is compatible with non-x86 architectures, and performance depends on the characteristics of the workload. Use-case selection is paramount, and software compatibility remains the primary barrier to non-x86 adoption. Factoring in use-case variance and alignment as part of composable infrastructure or platform can be difficult, but yield positive results.

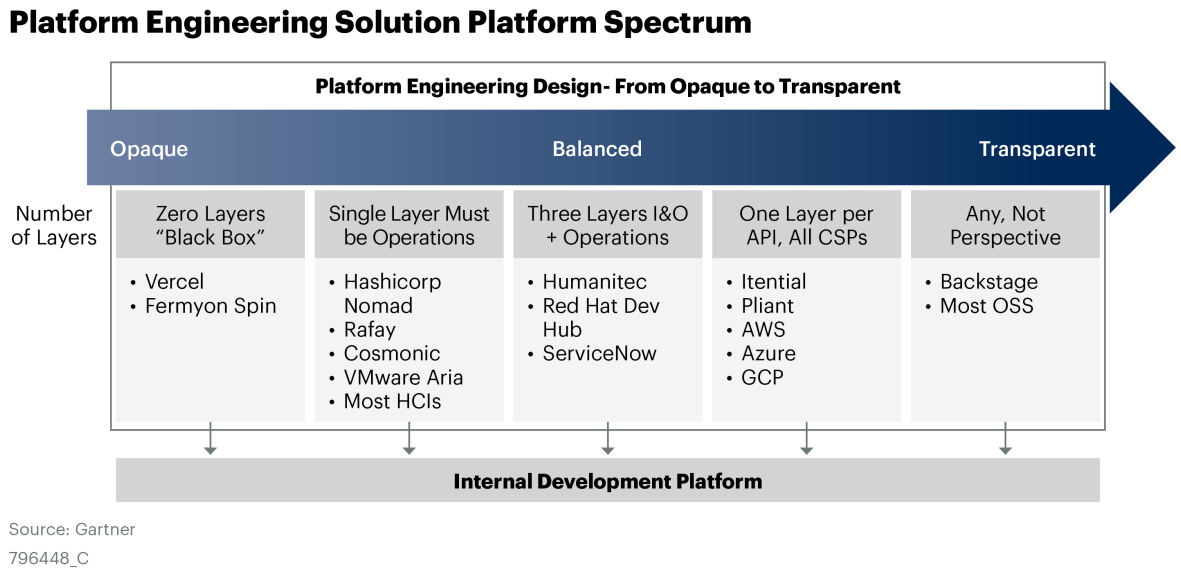
Related Research

- Best Practices for Modernizing Your Enterprise Network
- Emerging Tech: Top Semiconductor Technology Trends Impacting Data Centers for 2023

Select a Platform Architecture That Minimizes I&O Effort

Infrastructure platforms are complex software systems and tend to incorporate multiple, stacked layers of abstraction. Each time the platform team builds another layer of abstraction, it increases the power and flexibility of the platform, but it also increases the time and effort to build and maintain the platform. Furthermore, multilayered platforms require a platform orchestrator — a new level of automation tooling focused on integrating the various services. There is a spectrum of possible architectures and capabilities for the platform, and multiple architectural decisions to make, as described in the platform engineering design solution spectrum graphic shown below (see Figure 4).

Figure 4: Platform Engineering Solution Platform Spectrum



Source: Gartner
796448_C

Gartner

In general, platform teams should stay as far to the left of this spectrum as possible. That is to say; they should build the simplest possible platform, minimizing the number of API layers (as described in Choosing an API Format: REST Using OpenAPI Specification, GraphQL, gRPC or AsyncAPI) involved (one or, ideally, zero, if possible). Simpler platforms are more opaque, but easier to operate. Use a "transparent" or "balanced" design only where an "opaque" platform will not suffice.

Sometimes more transparency is required. A platform can't be opaque when infrastructure can't be fully abstracted, there are distinct use cases and/or architectural differences among the target infrastructure, when more than one vendor is involved, or formal platform operator roles have been defined (e.g., Cluster Admin in Kubernetes).

When building multiple layers of abstraction, start by choosing the right API format for the specifications of each layer as soon as possible. Every additional API layer requires substantial software development effort — usually more software engineering time and expertise than I&O has available. Furthermore, configuration drift and decision paralysis become more likely with each new layer added, especially in multicloud environments. To build a sustainable product, platform teams must match the level of effort that I&O is able and willing to put into its operation.

Implications

- A simpler platform that satisfies business requirements is preferable. Only when a real business case justifies new functionality should the platform team begin enhancing their products with more capabilities (thereby moving farther to the right of the spectrum).
- The platform must have a product owner. The product owner should have at least some subject-matter expertise in DevOps, plus the ability to manage technical workspace requirements.

Related Research

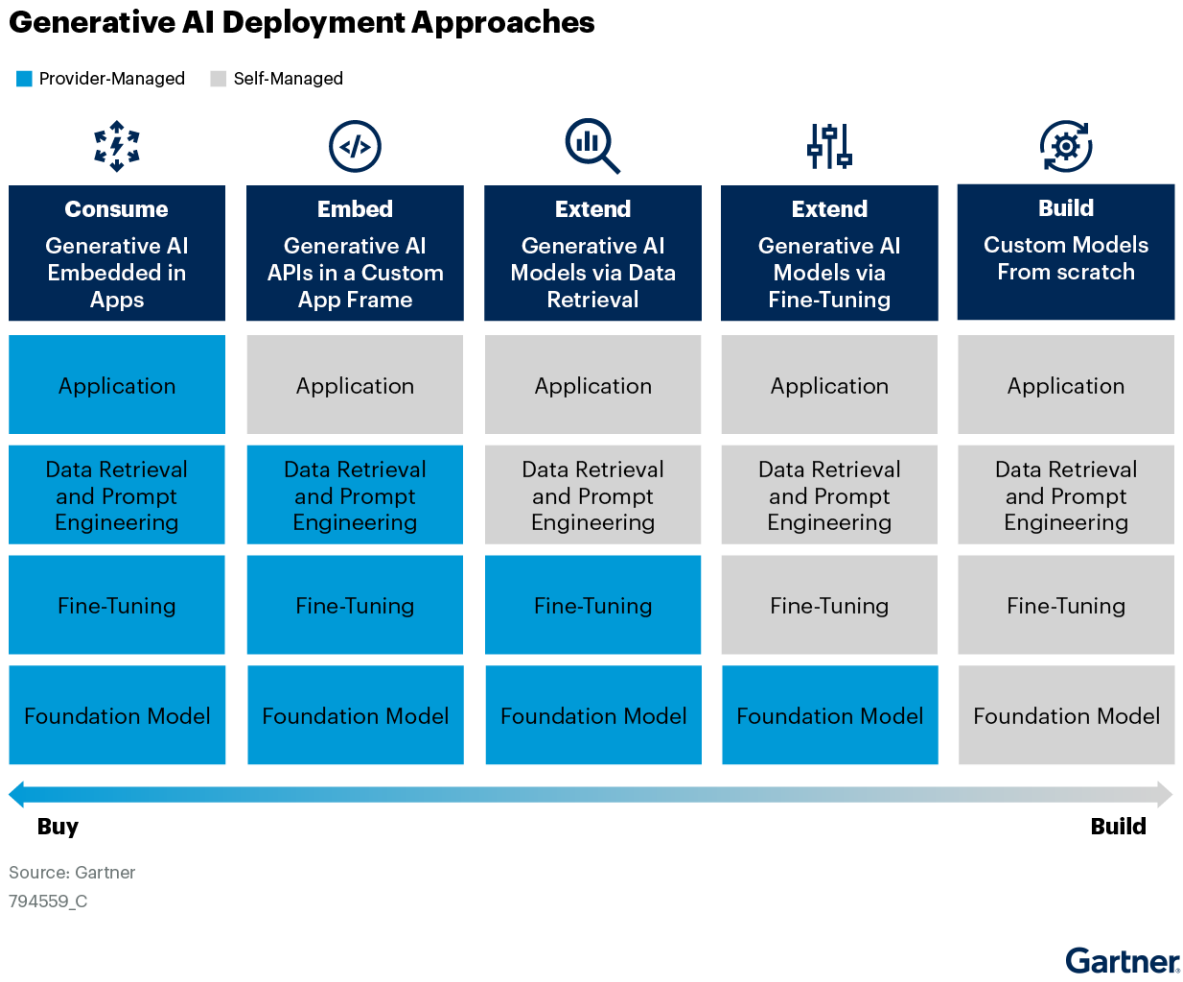
- Adopt Platform Engineering to Improve the Developer Experience
- Choosing an API Format: REST Using OpenAPI Specification, GraphQL, gRPC or AsyncAPI

Generative AI Will Drive the Need for New Infrastructure and Integration

Although GenAI had several niche applications in previous years, 2023 was a breakout moment, as OpenAI's ChatGPT garnered widespread adoption and mind share from consumers. Both enterprise buyers and technology vendors are now looking to use GenAI technology in their solutions. Promising initial use cases for generative AI include product development, creating new revenue channels, worker augmentation, long-term talent optimization and process improvement (see Innovation Guide for Generative AI Technologies).

The rise of GenAI introduces significant infrastructure requirements. Not all types of GenAI infrastructure will be directly implemented by customers themselves, but customers should consider the deployment approaches as part of their planning. Figure 5 shows five possible deployment approaches for GenAI, and provider-managed infrastructure plays a significant role in most of them.

Figure 5: Generative AI Deployment Approaches



Most enterprises will not train their own foundational models (FMs), due to the high cost and complexity of doing so. Instead, they will adopt existing commercial or open-source FMs, and more-sophisticated enterprises may fine-tune FMs for domain-specific applications. Enterprises will generally use APIs to interact with FMs delivered as a service. Some will deploy FMs on cloud IaaS, and some will use on-premises infrastructure.

For most GenAI initiatives, the main challenge for I&O technical professionals will be integrating services that provide access to FMs. In some cases, I&O teams must deploy and operate specialized infrastructure for inferencing, i.e., extending GenAI models through fine-tuning. To improve the accuracy and quality of model responses for domain-specific tasks, I&O teams must ensure that data can be provided from outside a foundation model, such as the organization's internal data.

Due to the significant resources required for GenAI processing, the use of hyperscaler cloud services will most likely be the easiest and most cost-effective approach. However, in some cases, organizations may require on-premises infrastructure for some of these functions, because of either compliance or privacy requirements, or because a particular solution can be achieved more cost-effectively with on-premises systems.

In 2024, I&O technical professionals should:

- Address immediate GenAI needs with cloud-based solutions
- Provide cloud infrastructure to train, extend and deploy GenAI large language models (LLMs)
- Build GenAI infrastructure on-premises when needs cannot be met through cloud-based solutions

Planning Considerations

Address Immediate Generative AI Needs With Cloud-Based Solutions

The majority of AI training and inference is happening in public cloud providers, due to the scalable, elastic and cost-efficient infrastructure options. However, hyperscale cloud providers — such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP) and Oracle Cloud Infrastructure (OCI) — have more to offer than just raw infrastructure. For example, they also provide:

- **Cloud AI developer services (CAIDS):** These services enable software developers who are not data science experts to use AI models via APIs, software development kits (SDKs) or applications. Hyperscale cloud providers are increasingly introducing developer-friendly FM services, although each provider has its own strategy. For example, Amazon Bedrock provides access to multiple FMs from Amazon and its partners, while Microsoft's Azure OpenAI Service offers access to multiple FMs from OpenAI.
- **Cloud data science and machine learning (DSML) platforms:** These suites of services offer comprehensive end-to-end modeling and engineering capabilities on managed infrastructure. They make it easier for DSML experts to carry out their work without needing to be infrastructure experts. Offerings include Amazon SageMaker, Microsoft Azure Machine Learning, Google Vertex AI and Oracle Data Science Service.

Broadly, the above services and the variety of other prepackaged cloud AI services, some of which are use-case-ready, provide organizations with easy and rapid on-ramps to address GenAI requirements and business initiatives. Although I&O technical professionals must ensure that these services are well-governed, they do not necessarily require a high-degree of I&O management. However, if the organization has not previously extended its data pipelines to the cloud, I&O technical professionals may need to ensure that a cloud solution is feasible, and to perform any necessary identity, security and network integration activities.

Benefits

- Adopting "model as a service" offerings significantly reduces I&O's burdens related to GenAI solutions. These services typically use token-based pricing and fully abstract the underlying infrastructure.
- Cloud-based platforms can significantly reduce I&O's responsibilities for the infrastructure used for training and inferencing. They also reduce help desk and I&O support requirements, since they make it easier for DSML experts to carry out their work without needing to be infrastructure experts.

Cautions

- Many GenAI startups offer their FMs as a service. In some cases, these services are hosted in hyperscale cloud providers, but not all are. Use of services from these startups requires I&O to methodically assess the suitability of these services for enterprise use, and related integration requirements (such as identity and networking).
- Many hosting and nonhyperscale cloud providers are now marketing infrastructure for GenAI training. Such providers are often pivoting from serving customers engaged in cryptocurrency mining. They may have little or no experience with traditional enterprise customers. Many such offerings are missing the cloud governance, management and security capabilities that enterprises have grown to expect.

Related Research

- Solution Comparison for Cloud Data Science and Machine Learning Platforms
- Magic Quadrant for Cloud AI Developer Services
- Tool: Vendor Identification for Generative AI Technologies

- Quick Answer: Which Cloud Provider Offers the Best Generative AI Business Outcomes?

Provide Cloud Infrastructure to Train, Extend and Deploy Generative AI models

When specific requirements of an organization's GenAI use case cannot be met with standard cloud-based services provided by hyperscaler cloud providers, I&O teams may be called on to deploy infrastructure themselves. The solution may have to support one of these scenarios:

- **Building a model:** This process is the same as "training" a data model, and it will be required if an organization wants to create a FM by using its own training data.
- **Extending a model:** This process starts with an existing FM, which is then enhanced with customizations reflecting the specific needs of the organization.

Building a new FM from scratch is extremely resource-intensive. For most organizations, attempting to implement the required infrastructure for building a FM in the public cloud will be cost-prohibitive. Using cloud resources for extending a model is more practical, and the two most common approaches are:

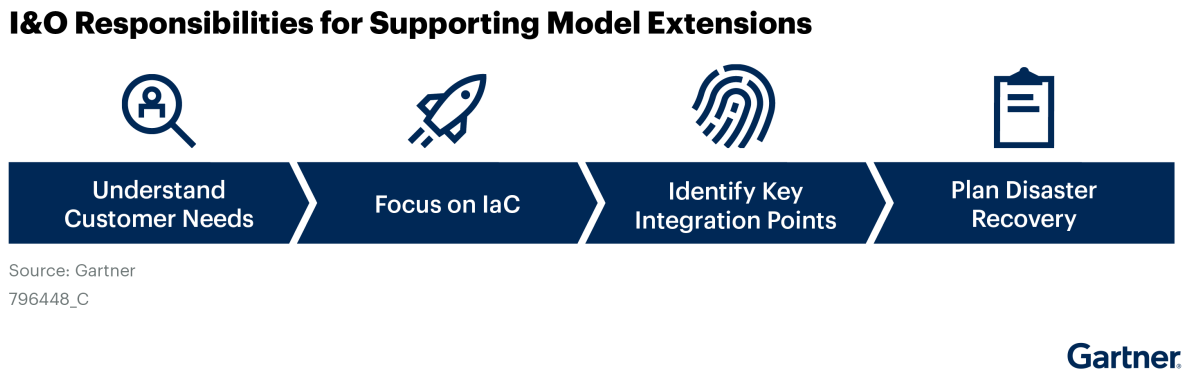
- **Fine-tuning a model:** The most popular fine-tuning approach is to retrain a portion of the FM with a specially prepared dataset, leading to improved performance or output.
- **Extending a model via data retrieval:** With this approach, the text prompt is augmented with additional context, such as including up-to-date data.

Each of these approaches introduces infrastructure needs that I&O professionals can meet using cloud-based resources, such as a search index, graph database or a similarly searchable solution. I&O and AI/data teams should work together to assess the business case for pursuing this methodology. I&O technical professionals will be expected to build GenAI infrastructure rapidly, which necessitates the use of infrastructure as code (IaC) as part of the solution. Furthermore, the many integration points a GenAI model will support must be factored into the design, and I&O must work with application and business teams to ensure that integration is successfully implemented.

To effectively meet the needs of users who want to extend models using cloud resources, I&O technical professionals in charge of cloud infrastructure should focus on these priorities (see Figure 6):

- **Identify the scale of required resources:** The infrastructure needed for building a new FM will require far more resources than infrastructure for extending a FM.
- **Provision GenAI resources using IaC:** Implement an IaC framework to orchestrate GenAI infrastructure with a special focus on computing and storage.
- **Identify the key integration points with GenAI models:** The upstream FM will change, so have a plan to adjust the integration points.

Figure 6: I&O Responsibilities for Supporting Model Extensions



Benefits

- Cloud infrastructure provides the necessary scalability for model training, especially for supporting pretraining needs. IaaS instances in the cloud offer strong support for GPUs with fast interconnects.
- Hyperscalers provide a variety of infrastructure abstractions for customers who want to deploy all GenAI-related software themselves, including virtual machines (VMs), containers, and Kubernetes-based orchestration services, as well as a wealth of platform as a service (PaaS)-based services.
- Hyerpscalers offer multiple storage solutions and access to other integration points for operating infrastructure supporting GenAI processes, such as logging and monitoring. The third-party ecosystem of AI vendors will also center around the ability to integrate with cloud solutions.

Cautions

- Although hyperscalers provide ML frameworks, security will be a concern for organizations entrusting intellectual property to build or extend models. Although cloud providers may provide appropriate protection options, the organization bears responsibility for implementing them, as well as their own data security.
- Deploying GenAI in the cloud is still a relatively new concept, which will require the development of new skills and capabilities. These will not be limited to AI and data disciplines, but will spill over into I&O roles that support those teams.
- Innovation of GenAI models is a continuous process, and I&O teams will have to provide tools and processes to support the management of model life cycles. This will be a joint process between I&O, application and AI, or data teams, which complicates matters in a relatively nascent market for these types of tools.

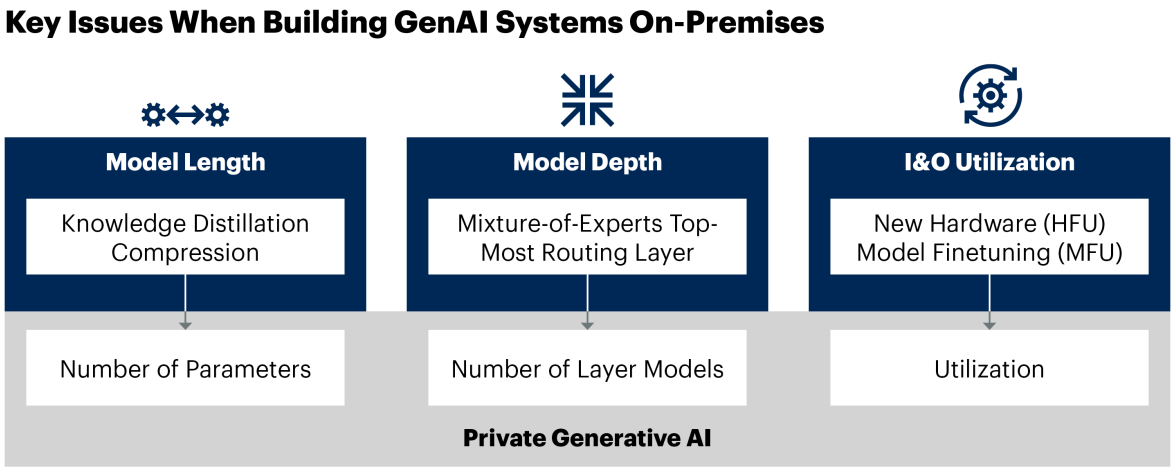
Related Research

- Solution Comparison for Strategic Cloud Integrated IaaS and PaaS Providers
- Solution Path for Assessing and Selecting Public Cloud IaaS and PaaS Providers
- Critical Capabilities for Cloud AI Developer Services

Build Generative AI Infrastructure On-Premises When Needs Cannot Be Met Through Cloud-Based Solutions

When I&O technical professionals cannot meet GenAI needs with cloud-based solutions at a reasonable cost, or if they fall short of business requirements, they will have to build or expand on-premises infrastructure to support GenAI. Successfully taking advantage of on-premises infrastructure for GenAI workloads is likely to require reducing the size and scope of models. It will require optimization of the use of on-premises hardware in support of fine tuning (see Figure 7). These optimizations will demand the development of new skills and tooling.

Figure 7: Key Issues When Building GenAI Systems On-Premises



Source: Gartner
796448_C

Gartner

A key prerequisite for deploying GenAI on-premises is to reduce the scale of models, so that they can accommodate on-premises hardware without considerable expense. The reduction can be accomplished by optimizing the length and depth of models using knowledge distillation and mixture-of-expert ML techniques. These techniques can reduce the model size by a factor of between 10 and 50, and enable fine-tuning output using fewer infrastructure resources by limiting the GenAI application to specific use cases.

Training failure, either when pretraining a model or fine-tuning an optimized model, will result in a significant waste of compute resources. The utilization of floating point operations per second (FLOPS), expressed in percent of success over time, describes the likelihood of the learning algorithm to yield results when going through an entire training dataset. Failure to update the internal model parameters to improve FLOPS utilization wastes both compute resources and the effort to create the model. Low FLOPS utilization may account for as much as 50% of operational costs, and even a small increase in utilization can save significant amounts of money in the long run. With an on-premises GenAI deployment, FLOPS utilization values should be in the range of 50% to 70% at a minimum.

I&O teams should expect to spend many weeks to months training one model. Hence, buying the infrastructure for training repeated cycles will be cheaper than renting over the long term, if the training stage is not a one-off endeavor. Platform teams should internally assign the appropriate roles and responsibilities to deal with automating tests with training cycles. The product owners for GenAI applications should have at least some subject matter expertise in test automation and the ability to manage the technical workspace requirements of I&O.

Benefits

- Owning and operating models on-premises is a potentially valuable capability for large organizations that plan to leverage GenAI for the foreseeable future and iteratively improve their models for the long term. Furthermore, it enables these organizations to get better value out of their private data, without the risk considerations of using cloud-based services.
- I&O technical professionals should plan on FLOPS utilization to increase through consolidation methods such as virtualization or containerization. Although this is an emerging field, the cost of model training is a focal point for cloud providers and customers alike. This should spur rapid innovation.

Cautions

- Skills in these disciplines are in very short supply. Operating GenAI is above typical DevOps professional's responsibilities and the job role will require cross-functional skills, such as SRE and coding. Moreover, the on-premises tooling is not as mature as in the cloud.
- There are supply chain constraints in GPU procurement. Cloud providers, and national governments, are buying the bulk of available GPUs.
- There is no formally accepted solution for rolling back models. Structural or logical bugs will require full retraining, because they cannot be undone, and training runs are not reproducible, which can greatly exacerbate training costs.

Related Research

- Innovation Insight: Transfer Learning

Distributed Data Will Drive Adoption for New Emerging Technologies

The genesis of emerging technologies such as edge, 5G private mobile networks (PMN), satellite connectivity and AI/ML correlates with the rise of data-intensive workloads that are widely distributed across various physical and virtual locations. Furthermore, geopolitical tensions, regulations and requirements for data sovereignty are driving organizations to look at distributed infrastructure, connectivity and security controls at the edge. Emerging technologies, most often edge solutions, are key to achieving these objectives.

I&O professionals should not restrict themselves to making traditional, silo-based infrastructure decisions related to these emerging technologies. They need to look beyond infrastructure to consider data management and analytics, because it is so deeply intertwined with infrastructure requirements. Emerging technology initiatives exist to fulfill specific business outcomes. Most often, these are tied to one or more of the following requirements:

- Rapid, predictable responses that demand local resources
- Reducing congestion is cost by leveraging network optimization
- Autonomy in the face of network disconnections or service outages
- Privacy and security to meet regulatory or sovereignty requirements

All of these requirements demand the efficient collection, manipulation and analysis of vast quantities of data. Solutions will involve multiple components and a broad scope and, therefore, inherent complexity. No single vendor can address all verticals or use cases, and a critical facet will be the partnerships among vendors and in organizations' IT, OT and business teams.

In 2024, I&O technical professionals should:

- Embrace distributed ecosystems
- Expect edge and distributed cloud solutions to imperfectly address sovereignty concerns
- Integrate IT with OT to improve data discovery, sharing and analytics

Planning Considerations

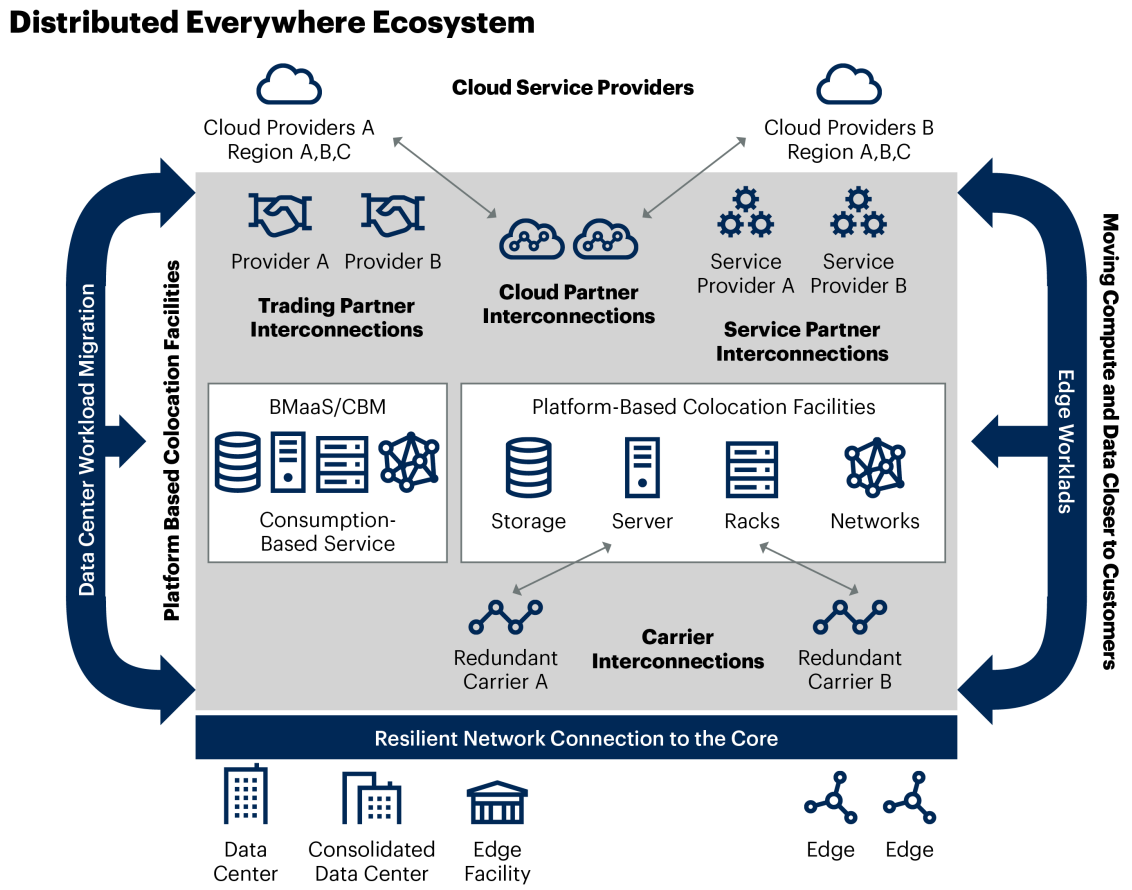
Embrace Distributed Ecosystems

Organizations are increasingly embracing a distributed ecosystem of infrastructure and solutions to address the growing need to serve a demanding, global user base. The spectrum of edge, on-premises data centers, colocation facilities and cloud services should be viewed as an ecosystem of distributed solutions. Each type of solution will have its role to play in this paradigm. For example:

- On-premises data centers may continue to support legacy, non-x86 workloads.
- Colocation facilities may be used in an opex, consumption-based model with prebuilt, as-a-service options delivered as a platform.
- Edge computing may be deployed via hyperscaler-integrated solutions, delivered on an opex basis and connected to resilient networks.

Colocation facilities already act as a glue between on-premises data centers and the public cloud, but can also become the glue between these resources and edge infrastructure, as shown in Figure 8. The partner ecosystem available through colocation providers and the network services available enables a strong set of capabilities. Furthermore, they are increasingly an attractive option for organizations as part of a cloud adoption plan by transferring noncloud solutions to those facilities.

Figure 8: Distributed Everywhere Ecosystem



In essence, I&O technical professionals cannot make independent choices in the variety of locations they must support. Instead, they must consider the entire distributed ecosystem as part of their decision-making and evaluate solutions in the context of wide-ranging requirements.

Benefits

- Data center and colocation facilities now offer bare-metal-as-a-service offerings transitioning from large, one-time capex to a fixed, more flexible on-demand pricing model. As organizations move toward opex-based models, these solutions offer a cloud-like experience outside the public cloud.

- Businesses are often no longer required to dedicate large amounts of capital to data center initiatives that don't contribute directly to the bottom line. This frees up I&O technical professionals to pursue other types of solutions that may match overall business demand better.

Cautions

- Organizations are unprepared. The distributed ecosystem requires a holistic view of operations. AI-assisted operations are increasingly needed to analyze operational telemetry, especially where the abundance of alerts generated across the entire ecosystem would overwhelm operations teams.
- Network connectivity is the backbone of the distributed ecosystem, but locality is a critical consideration for any transformation initiative. For example, if you add 30 milliseconds to a translation, it will reduce throughput approximately eightfold, and, with just a 2% packet loss, customers will notice an approximate 25x reduction in throughput. This increases the demand for local resources, which may be difficult to obtain and support.

Recommended Reading

- Toolkit for Estimating Data Center Build and Modernization Costs by Tier Level
- Market Guide for Consumption-Based Models for Data Center Infrastructure
- Create Successful Consumption-Based Pricing Models by Balancing Transparency, Predictability and Simplicity
- Predicts 2023: XaaS Is Transforming Data Center Infrastructure

Expect Edge and Distributed Cloud Solutions to Imperfectly Address Sovereignty Concerns

An increasing number of countries (and smaller jurisdictions, such as U.S. states) have imposed or intend to impose regulations for cloud sovereignty. The meaning and scope of "cloud sovereignty" varies by jurisdiction. However, technical professionals should assume that it encompasses data residency, both at rest and in motion, along with the location of the infrastructure used to process data and transactions. Questions of authority and ownership — including who is responsible for operations and support — depend on location, as well as potentially on the nationality and location of any involved personnel.

Cloud providers are doing their best to comply with emerging regulations, within the bounds of business constraints. For technical professionals, there is a material difference between a public cloud region that meets typical digital sovereignty requirements, and a sovereign cloud region designed to meet strict local ownership requirements. The latter will generally be more expensive, often be partner-operated, and will have a limited set of services.

More broadly, the promise of distributed cloud solutions (including cloud solutions at the edge) has not been fulfilled. Unfulfilled customer requirements include:

- The full range of public cloud IaaS and PaaS capabilities, available in their own data centers, including at the edge.
- Full API compatibility with the public cloud services, so that applications are portable between the core and the edge.
- Solutions that are fully functional and can be operated when disconnected, rather than depending on a "tether" to the vendor for control plane capabilities.

The distributed hybrid infrastructure (DHI) market has produced hyperconverged appliances with a limited set of IaaS capabilities or container orchestration capabilities. These solutions often are not API-compatible with a major public cloud provider's service APIs (even when delivered by the same provider). Furthermore, these solutions may depend on tethering.

Therefore, technical professionals must focus on selecting the solutions that have enough functionality to meet their immediate needs and specific edge use cases, without neglecting an architectural vision of their future edge platform. If no cloud solutions exist that can fulfill the organization's requirements, then noncloud solutions must be considered, or the scope of the project must be reduced to be achievable using the available solutions.

Benefits

- Edge solutions can reduce network latency and costs, improve resilience and help organizations comply with local regulatory requirements. Growing interest in edge solutions has encouraged DHI vendors to invest in more public cloud-like solutions. PaaS capabilities on the edge will improve, and interest in GenAI is encouraging the development of capabilities for DSML at the edge.

- Organizations must become compliant with emerging cloud sovereignty regulations. Digital sovereignty is important to many organizations, especially those headquartered outside the U.S., because it is related to potential business continuity risks. The implicit threat of looming regulations has pushed cloud providers to open regions in more countries, staff local sales and support, and offer payment in more currencies.

Cautions

- Some countries may require cloud providers to work with local partners, or otherwise conduct business in a way that results in high costs to deliver services. These costs will usually be passed on to customers. Therefore, you must be aware of the cost impact of cloud region choices — especially when using sovereign cloud regions.
- Sovereign cloud regions are unlikely to make customers safer. If you have a regulatory mandate, you will have to perform sovereignty theater that results in little, if any, risk reduction, while increasing cost (and likely increasing complexity).

Recommended Research

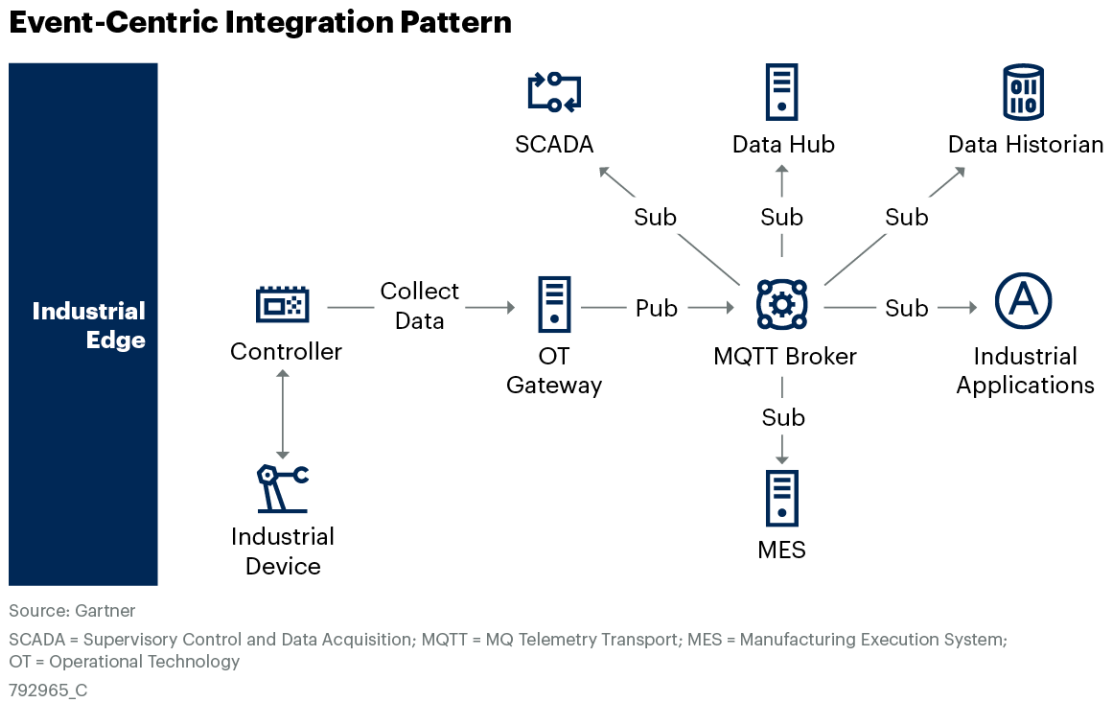
- Cloud Architecture Best Practices: How to Choose Cloud Regions
- Quick Answer: How Do I Obtain Isolated Private Cloud Services?
- Comparing On-Premises Public Cloud Appliances: AWS Outposts, Microsoft Azure Stack Hub and Google Distributed Cloud Edge

Integrate IT With OT to Improve Data Discovery, Sharing and Analytics

Data pipeline visibility is essential to support modern business processes. The data pipeline is the sequence of processing steps necessary to collect, transform, organize, analyze and store enterprise data. However, many organizations struggle to find, access and analyze their operational technology (OT) data. Even when OT systems share data, it is often unstructured, diverse and real-time. Pressure to deliver faster results, higher quality and greater resiliency is driving organizations to consider how to integrate IT with OT.

In response to this pressure, many organizations are developing their architectures to support an event-centric design pattern. A key component of the event-centric integration pattern is the event broker, which moves messages from event producers to event consumers. The event broker decouples producers and consumers in time, space and synchronization (see Figure 9).

Figure 9: Event-Centric Integration Pattern



The decoupling enables technical professionals to add producers and consumers without affecting the rest of the system, thus improving design flexibility and system scalability. Organizations will need to modernize edge architectures by transitioning toward an event-centric integration pattern using event brokers.

Threats to OT equipment are on the rise. As OT equipment became more connected, they exposed attack surfaces. Cyberphysical system protection platforms (CPS-PPs) have emerged to discover OT assets. Not only do CPS-PPs provide a front end to the security process, but they are also becoming a back end by capturing telemetry, utilization and operational data. Organizations will find they must use these platforms and integrate them with security information and event management (SIEM) tools and other processes.

Lastly, OT increasingly resembles IT, with such features as remote monitoring and software upgrades. However, OT assets largely haven't been managed in the same way as IT assets. As a result, successful IT-OT integration projects will require collaboration among teams with diverse skills. This often requires the IT team to become familiar with OT processes, people and technology, and vice versa. An architect responsible for IT-OT integration can drive transformation via collaborative work across stakeholders and, in particular, among IT and OT staff.

Benefits

- The event-centric pattern improves the ability to access and analyze OT data. The decoupling of producers and consumers improves data pipeline scalability, which will be necessary as the organization's needs grow.
- CPS-PPs not only discover OT assets, but also integrate with SIEM and security orchestration, automation, and response (SOAR) solutions to unify IT-OT security. This greatly improves the ability of organizations to secure and protect their OT assets and promotes integration between IT and OT.
- Integrating IT and OT systems can improve collaboration among IT-OT teams, business stakeholders and suppliers. Organizations will find this collaboration necessary, because business requirements demand that IT and OT systems communicate and interoperate well.

Cautions

- Event brokers are optimized for telemetry data publishing (one publisher to many subscribers). However, event brokers do not support command/response transaction exchanges and, thus, are not a panacea.
- The number of disclosed IT and OT vulnerabilities continues to grow, and remain difficult to manage. A major issue with vulnerabilities in production environments is the inability to patch at will.
- IT-OT collaboration will take time. Historically, IT and OT teams have been siloed, and collaboration requires trust. However, earning trust takes time, especially between teams that may not have interacted very much previously.

Recommended Reading

- Reference Architecture for Integrating OT and Modern IT
- Market Guide for CPS Protection Platforms
- Solution Path for Planning Edge Technology

Document Revision History

2023 Planning Guide for Cloud, Data Center and Edge Infrastructure - 13 October 2022

2022 Planning Guide for Cloud and Edge Computing - 11 October 2021

2021 Planning Guide for Cloud and Edge Computing - 9 October 2020

2020 Planning Guide for Cloud Computing - 7 October 2019

2019 Planning Guide for Cloud Computing - 5 October 2018

2018 Planning Guide for Cloud Computing - 29 September 2017

2017 Planning Guide for Cloud Computing - 13 October 2016

2016 Planning Guide for Cloud Computing and Virtualization - 2 October 2015

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

2024 Planning Guide for Security

2024 Planning Guide for Data Management

2024 Planning Guide for IT Operations and Cloud Management

2023 Planning Guide for Application Development

2024 Planning Guide for Identity and Access Management

2024 Planning Guide for Application Architecture, Integration and Platforms

2024 Planning Guide for Analytics and Artificial Intelligence

---

Table 1: Where to Incorporate Non-x86 Architectures Into Your Environment

Architecture	Consider For	Avoid For
ARM	Linux-based workloads; workloads based on microservices architecture; edge computing workloads	Windows servers; large, monolithic servers
GPU	AI/machine learning (ML) workloads; neuromorphic computing; analytics workloads; enhancing virtual desktop performance	General-purpose computing
FAC	Workloads that benefit from offloading functions from the CPU	Commodity workloads; non-performance-critical environments

Source: Gartner (October 2023)