Gartner®

Hype Cycle for Infrastructure Strategy, 2023

Published 25 July 2023 - ID G00790892 - 112 min read

By Analyst(s): Philip Dawson, Nathan Hill

I&O leaders should use this Hype Cycle to drive innovation in their platform strategies to achieve infrastructure anywhere. Generative AI, new consumption models, intelligent platform automation and sustainability must be factored in. Hybrid cloud and edge delivery continue to help optimize costs.

Additional Perspectives

- Summary Translation: Hype Cycle for Infrastructure Strategy, 2023 (11 September2023)

More on This Topic

This is part of an in-depth collection of research. See the collection:

- 2023 Hype Cycles: Deglobalization, AI at the Cusp and Operational Sustainability

Strategic Planning Assumptions

By 2027, 35% of data center infrastructure will be managed through a cloud-based control plane, up from less than 10% in 2022.

By 2026, infrastructure and operations (I&O) will spend more than half its budget working with technology the organization does not directly own.

Analysis

What You Need to Know

This document was republished on 27 July 2023. The document you are viewing is the corrected version. For more information, see the *Corrections* page on gartner.com.

For 2023, this Hype Cycle reflects the need to adapt to AI to deal with automation and inflationary pressures, with a focus on infrastructure sustainability. Additional related innovations (e.g., off-grid power, net-zero data centers and consumption-based models) have risen in importance as platforms and infrastructure delivery options become more comprehensive and more complex.

While we often focus our comments on the emerging technologies new to this Hype Cycle, this year we highlight an innovation that has reemerged. Viewed as traditional “virtualization” through hyperconvergence, private cloud computing is being refreshed and reiterated (rather than off the Hype Cycle). This trend is made possible by the revirtualization of projects concerning storage and network virtualization. It is further driven by merger and acquisition events such as Broadcom’s acquisition of VMware which was split away from Dell (see Quick Answer: How Should VMware Customers Prepare for the Broadcom Acquisition?).

Revirtualization is also tied to the innovation called software-defined infrastructure (SDI), which continues to abstract software from hardware. SDI, now mature, is still considered obsolete before the Plateau of Productivity, because the innovation is tied to vendors’ individual offerings and domains, with limited common terminology, a lack of common integration, and proprietary lock-in.

Cloud-native applications with containers driven by microservices are mature as an alternative abstraction with or without virtual machines (VMs) for cloud-native applications that quantify the measurement and consumption of resources, instead of just provisioning them.

API-led programmable platforms and infrastructure, hyperconvergence, composable infrastructure and intelligent infrastructure are paths to more intelligent platforms. This continuous trend captures the progression of infrastructure through traditional, manually intensive administration, to highly automated intelligent infrastructure. It continues to focus on the expansion of on-premises distributed platforms, infrastructure integration, and learning from the location and proximity demands of edge computing.

For information about how I&O leaders view the technologies aligned with this Hype Cycle, see Infographic: 2023 Technology Adoption Roadmap for Infrastructure and Operations.

The Hype Cycle

With generative AI, variable inflationary pressures and renegotiated subscriptions driving up costs, hybrid infrastructure is now “stretched” with edge and elastic with cloud, blurring the boundaries between infrastructure, management, cloud, sourcing and virtualization.

Hybrid delivery models continue to transform I&O leaders’ project silos and team skills into a platform-and-product-driven strategy as they shift to repeatable everything-as-a-service (XaaS) models. This infrastructure simplification rationalizes core IT, while accommodating and planning IT expansion, keeping new things simple and repeatable through standardization and being more cloudlike on-premises.

This research describes the 33 most hyped infrastructure innovations. For each innovation, we define and analyze the value to enterprises, level of adoption and anticipated rate of future growth. I&O leaders should use this research to determine whether and/or when to invest in these innovations.

New entrants: Generative AI and related infrastructure innovations such as programmable platforms, augmented reality in data centers and operational AI systems represent innovations that highlight a bifurcation between AI programmability and infrastructure consolidation in cloud and platform deployment with expansion and management integration with edge computing delivery.

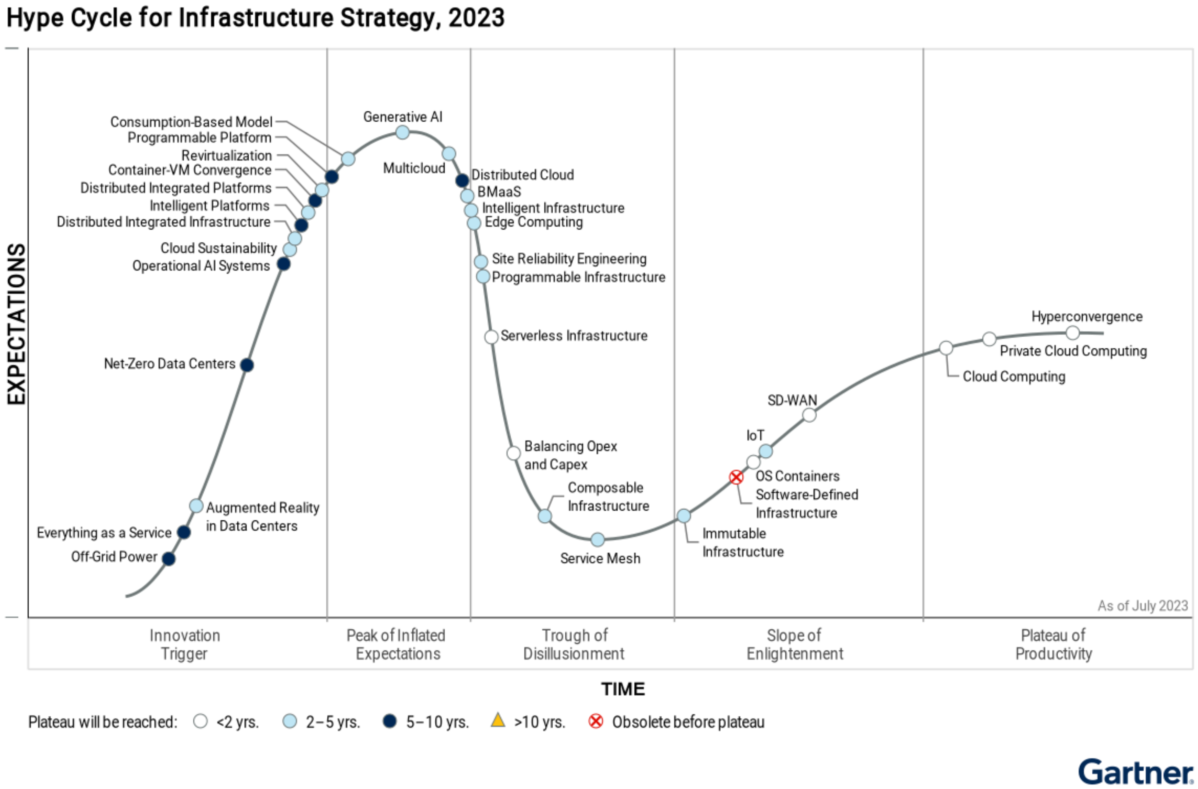
On the rise: Everything as a service (XaaS), revirtualization, intelligent platforms and cloud sustainability are moving up the slope and toward peak hype. Real-world use cases, projects and implementations are needed to take them over the peak.

Peak hype: From the peak, consumption-based models, generative AI, multicloud, distributed cloud remain hyped and bare metal as a service (BMaaS), gaining traction as they move toward the Trough of Disillusionment.

Fast movers: The three fastest movers — distributed cloud, BMaaS and intelligent infrastructure — quickly jumped the peak since 2022. These three innovations demonstrate the importance of packaging of cloud applications with services and security in hybrid delivery models.

In the trough: Composable infrastructure and service mesh are in the Trough of Disillusionment, and immutable infrastructure is pulling out of the trough as it matures as an offering that must shift from hype to delivery.

Figure 1: Hype Cycle for Infrastructure Strategy, 2023

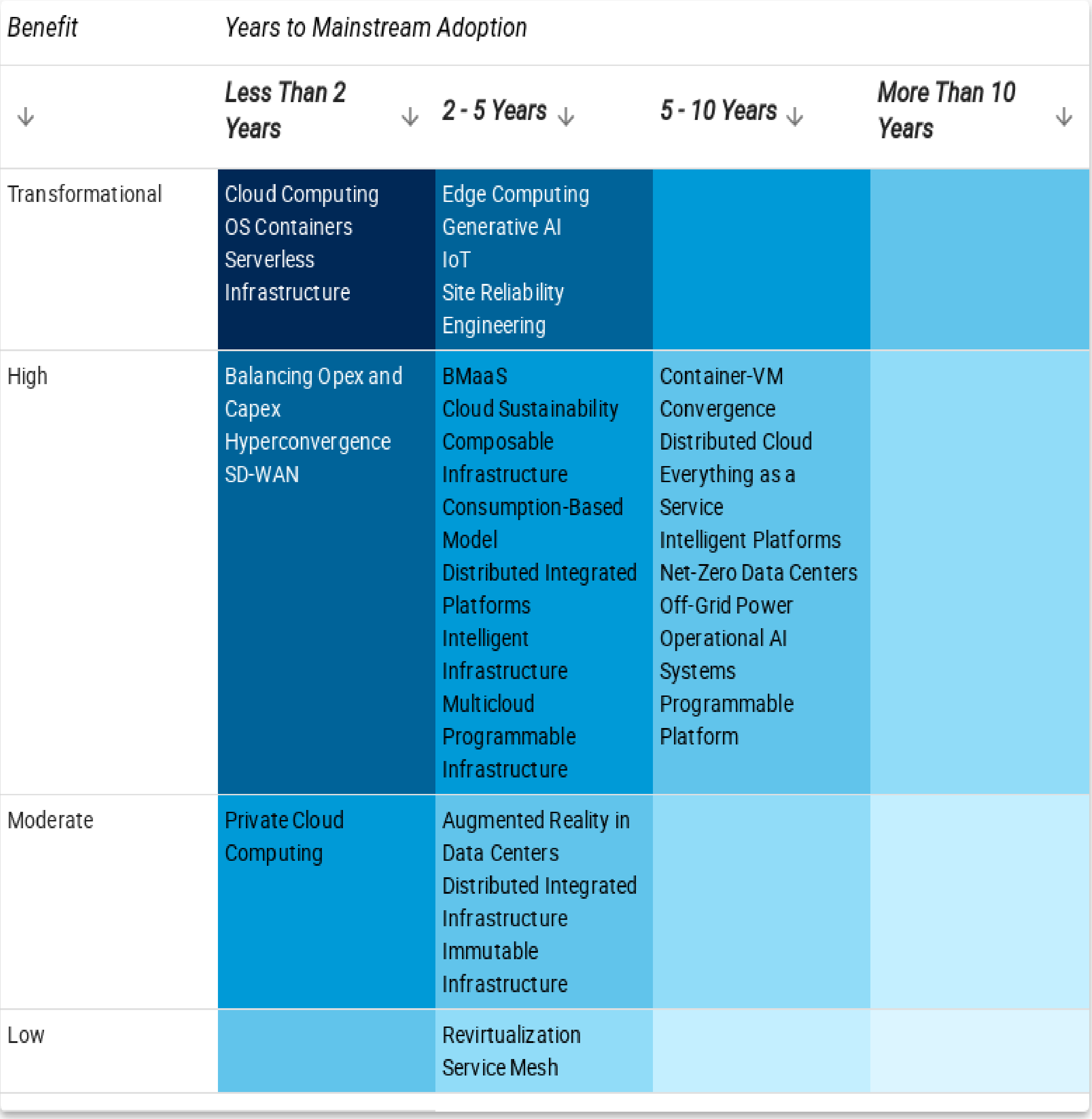


The Priority Matrix

Many of the innovations covered here are transformational, mainly centered around cloud, edge, the Internet of Things (IoT), containers and serverless infrastructure. High-benefit technologies include hyperconvergence, container-VM convergence, cloud sustainability, programmable platforms and other data center initiatives. Other infrastructure technologies provide moderate impact, such as augmented reality in data centers and distributed integrated platforms/infrastructure.

Table 1: Priority Matrix for Infrastructure Strategy, 2023

(Enlarged table in Appendix)



Benefit	Years to Mainstream Adoption			
↓	Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational	Cloud Computing OS Containers Serverless Infrastructure	Edge Computing Generative AI IoT Site Reliability Engineering		
High	Balancing Opex and Capex Hyperconvergence SD-WAN	BMaaS Cloud Sustainability Composable Infrastructure Consumption-Based Model Distributed Integrated Platforms Intelligent Infrastructure Multicloud Programmable Infrastructure	Container-VM Convergence Distributed Cloud Everything as a Service Intelligent Platforms Net-Zero Data Centers Off-Grid Power Operational AI Systems Programmable Platform	
Moderate	Private Cloud Computing	Augmented Reality in Data Centers Distributed Integrated Infrastructure Immutable Infrastructure		
Low		Revirtualization Service Mesh		

Source: Gartner (July 2023)

On the Rise

Off-Grid Power

Analysis By: Tony Harvey, Autumn Stanish, Jason Donham, Philip Dawson

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Off-grid power supply and delivery is a growing option for high-power-demanding data centers and can contribute to reducing the carbon footprint of on-premises data centers. It can be used to reduce the load on already overstressed power grids as well as provide higher reliability in areas where power delivery can be unreliable.

Why This Is Important

High-capacity data centers now require megawatts (MW) of power, with some of the larger data centers' campuses requiring hundreds of MW or more (see Google's announcement on sustainability). This is putting inordinate pressure on power generation and distribution networks, and global power prices have grown by 22% since February 2022. Organizations that need to run on-premises data centers must look for additional power capacity and locations from a network that simply cannot continue to provide them.

Business Impact

The lack of power availability can halt or pause data center expansion. This will delay the implementation of new IT capabilities as well as increase costs for data center alternatives, such as colocation space and cloud services. By investing in off-grid power, businesses mitigate the risks of lack of power, grid reliability issues or unpredictable volatility in the wholesale energy markets.

Drivers

- Energy costs and requirements are growing exponentially for all data centers, as there is more demand for their services.

- Distribution networks for electricity are already overstretched and the move to electrification, for cooking, heating and transport, is only exacerbating the issue. There have already been moratoriums on new data center builds in some countries, and this is expected to continue (see Dutch Call a Halt to New Massive Data Centres, While Rules Are Worked Out, DutchNews.nl).

- Sustainable energy sources, such as solar, tidal or wind, are unpredictable and must be backed up with a continuous power source.

- The new concept of small modular reactors (SMRs) offers between 5 MW and 1,000 MW of power. An SMR could be used to power a data center or a number of data centers. The first SMRs are in production in Russia and China (see IAEA Ups Support for SMRs, Nuclear Engineering International). While other SMRs have been licensed by their local nuclear regulators, they do not yet have a go-live date.

Obstacles

- A large-scale capital investment is required for setting up an SMR, large-scale wind farm or solar array, however, data centers typically have a 15- to 20-year life span, allowing a long-term view of ROI.

- There are security and planning permission impacts in running an SMR, with additional security controls and deterrence needed, that would not normally be required except for Tier 4 data centers.

- Nuclear SMR reactors, while sustainable, have a PR problem and will generate political and planning opposition.

- Local or central government licensing must be considered, with some countries averse to nuclear power due to political pressure, or to large-scale wind and solar farms due to local resident resistance.

User Recommendations

- Review your data center expansion plan by assessing any issues related to grid capacity, reliability, and/or energy price volatility in current and future locations.

- Prepare for off-grid generation by investigating what investments in solar, wind, SMR, fuel cell, heat or multisource power generation could be used to power your current and planned data center facilities.

- Ensure that ROI calculations include the value of off-grid generation by evaluating a data center build project for a 10- to 15-year investment period, and include the possibility of incentives for sustainable power or off-grid capability.

Gartner Recommended Reading

The Road to a Net Zero Data Center

Quick Answer: How Do I Assess On-Site Renewable Energy Options?

Forecast Analysis: Data Center Sites, Worldwide

Toolkit for Estimating Data Center Build and Modernization Costs by Tier Level

Strategies to Plan for GHG Emissions Reduction

Everything as a Service

Analysis By: Jason Donham, Philip Dawson

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Everything as a service (XaaS) as a Hype Cycle innovation encapsulates the whole umbrella services delivery program that drives agility, operational efficiency, ease of management and cost optimization through strategic migration of services and projects into XaaS.

Why This Is Important

Organizations can free themselves from the commitment of a long-term investment in infrastructure by adopting the XaaS model. IT leaders have never had more data center infrastructure deployment choices than they do today. Traditional on-premises infrastructure has given way to the hybrid platform and cloud operating model, which in turn has increased complexity and resulted in myriad architecture choices. XaaS drives and packages the portfolio of services.

Business Impact

Business objectives are the driver behind any move toward the XaaS model. Agility, operational efficiency, ease of management and cost optimization are four objectives that are directly addressed by strategic migration of services to XaaS. In-house talent and service providers will assist in the design of a horizontal services model to meet the needs of the organization.

Drivers

- The gradual shifting of the control plane from on-premises to cloud-based, along with increasing migration to SaaS, is driving a reduction in traditional data center infrastructure footprints.

- XaaS is the bundling together of cloud computing services to create an economical, easy-to-consume product for hybrid cloud environments. It is the result of the disaggregation of traditional IT projects into components, domains and services that are then bundled together for consumption.

- Organizations are seeking to establish new predictable cost and consumption models. This is in contrast to vendors who are rapidly moving toward XaaS in an effort to provide as many services as possible to customers.

- Another vendor-offered XaaS solution to the supply chain issues has been the use of consumption-based pricing models. In these models, capacity is installed and managed by the vendor and then paid for when actually consumed — unlike traditional procurement models.

Obstacles

- XaaS only temporarily fills the skills gap in organizations where advanced technology is adopted and IT administrators do not have the skills to operate the systems efficiently.

- Long-term commitments to enterprise-owned data center capacity may limit flexibility. Moreover, migrating business-critical workloads to SaaS- and XaaS-based offerings will result in large, incremental reductions in the requirement for application modernization and transformation.

- Hardware vendors who are not unified in their approaches and offerings are attempting to layer cloudlike platform solutions onto their infrastructure. Application and platform vendors are offering distributed platforms that focus on the integration of infrastructure as a service (IaaS), platform as a service (PaaS) and/or SaaS.

- Migrating applications into a services model is complex. Some workloads are cloud-ready or currently running in a public cloud, so applications need to be redefined in order to operate in a new services infrastructure framework.

User Recommendations

- Migrate legacy applications not suited for cloud migration to on-premises consumption-based infrastructure. This brings cloudlike XaaS operating model benefits to legacy workloads that are otherwise incompatible with public cloud computing.

- Implement a mix of XaaS and on-premises consumption solutions. These solutions are billed together in an opex model as consumption-based services.

- Move beyond the traditional limitations of purchase, ownership and depreciation by expanding the range of possible XaaS consumption options for DC infrastructure replacement.

- Prepare for increasing variability and spend as a result of XaaS and consumption-based costs for virtual infrastructure.

- Adopt The Gartner Framework for Public Cloud Financial Management best practices and integrate with existing expense management tooling.

Sample Vendors

Amazon Web Services (AWS); Cisco; Dell; Google; Hewlett Packard Enterprise (HPE); IBM; Lenovo; Microsoft; Oracle

Gartner Recommended Reading

Predicts 2023: XaaS Is Transforming Data Center Infrastructure

How Do I Plan for Migrating My Data Center Infrastructure Into an XaaS Model?

Beyond FinOps: The Gartner Framework for Public Cloud Financial Management

Augmented Reality in Data Centers

Analysis By: Henrique Cecci

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Augmented reality (AR) technology has several applications in data centers, particularly in managing, securing, maintaining and expanding data center infrastructure, troubleshooting, and staff training.

Why This Is Important

AR in data centers provides an innovative way to manage, operate, and maintain data center infrastructure. It may help infrastructure and operations (I&O) teams work more efficiently and effectively. With AR, data center operators can visualize and manipulate virtual objects in real time, improving the accuracy and speed of tasks such as troubleshooting, maintenance and repairs, and can also enhance staff training and education programs.

Business Impact

AR in data centers can deliver improved productivity, cost savings, enhanced safety, improved training and knowledge transfer, faster issue resolution, and reduced downtime. AR provides technicians with real-time data visualization, remote guidance, and interactive 3D models to facilitate complex tasks and maintenance processes. It can eliminate the need for physical manuals and reduce the time needed for troubleshooting and repairs.

Drivers

Key drivers for using AR in data centers include:

- **Improved efficiency:** AR enhances the accuracy and speed of data center maintenance and repairs, reducing downtime and increasing uptime.

- **Cost savings:** AR data center technicians identify issues more quickly and accurately, reducing the need for additional labor and minimizing the risk of equipment damage.

- **Improved safety:** AR provides technicians with real-time information about potential hazards, allowing them to take proactive measures to mitigate risks and avoid accidents.

- **Enhanced training:** AR is used to provide immersive and interactive training experiences for data center technicians, enabling them to develop new skills and knowledge more quickly and effectively.

- **Improved customer experience:** AR helps data center operators provide a more engaging and informative experience for customers, improving satisfaction and retention.

Obstacles

Typical obstacles for using AR in data centers include:

- **Technology complexity:** Implementing and maintaining an AR system is technically challenging, requiring a combination of hardware, software and connectivity.
- **Integration:** Integrating AR with existing data center management tools like existing data center infrastructure management (DCIM) or IT service management (ITSM) tools and systems is challenging and complex.
- **Cost:** The cost of implementing AR solutions is high, especially for small and medium-sized data centers.
- **Training:** AR technology requires specialized skills and training.
- **Security:** Data center security is crucial, and AR potentially introduces new security risks that must be addressed.
- **Adoption:** The adoption level of AR technology in data centers is still relatively low due to a lack of awareness and understanding of its benefits.

User Recommendations

- Clearly define your goals and objectives, for example, improving maintenance, reducing downtime, or enhancing employee training.
- Start with a small pilot project to test the effectiveness of AR technology before scaling up to larger and/or production deployments. Focus on integration and compatibility with existing I&O tools and systems like ITSM, DCIM or others.
- Ensure AR adoption complies with your organization's security policies, as well as with industry regulations and standards. Train staff in standardized processes and necessary skills.
- Continuously evaluate the effectiveness of AR in your data center operations and identify areas for improvement to maximize the benefits and ROI.

Sample Vendors

Axonom; Dell EMC; Inceptum; Nlyte; Schneider Electric; Vertiv

Gartner Recommended Reading

Emerging Technologies: Tech Innovators in Augmented Reality — AR Cloud

Emerging Technologies: Tech Innovators in Augmented Reality — Spatial Web

Emerging Technologies: Tech Innovators in Augmented Reality — Augmentation and Spatial Interaction Layer

Net-Zero Data Centers

Analysis By: Simon Mingay, Philip Dawson, Autumn Stanish, Matthew Brisse

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Emerging

Definition:

Net-zero data centers aim to achieve a net-zero climate target by minimizing their direct and indirect greenhouse gas (GHG) emissions, and offsetting the balance appropriately. They should be able to demonstrate world-class energy and water efficiency, circularity practices, and asset utilization.

Why This Is Important

Enterprises and data center/cloud service providers are under mounting pressure from customers, investors, regulators and employees to reduce GHG emissions, increase energy efficiency, and establish a credible sustainability narrative. Pressure is growing on infrastructure and operations (I&O) leaders to adopt methods to increase transparency and performance, in order to make their data center operations efficient and environmentally sustainable.

Business Impact

IT leaders are confronted by a significant, unrelenting, year-on-year growth in compute and storage capacity, with spiraling energy consumption and associated GHG emissions. For enterprises with energy and GHG reduction targets, this is unsustainable. To meet cost targets, organizations must make radical improvements in data center, and I&O efficiency and emissions. Net-zero data centers will be essential infrastructure for all enterprises with ambitious or net-zero GHG targets.

Drivers

- Alignment of I&O with the organization’s energy and GHG targets.
- External stakeholders’ (specifically customers, investors and regulators) net-zero expectations.
- The need to mitigate the liability associated with costs of energy and GHG emissions.
- The need to build resilience in the face of increased contention for electrical and renewable supply capacity.
- Increased emphasis on cost reduction in IT systems.
- Increased consumption of cloud services, moving workloads out of the enterprise data center.
- Data center consolidation, caused by office consolidation following the COVID-19 pandemic, end of economical life of data centers and refocus of business operations.
- Customers demanding GHG footprint transparency from data center and cloud service providers.

Obstacles

- Unrelenting and significant growth in compute and storage capacities.
- Lack of a strong business case in the absence of ambitious enterprise GHG reduction goals.
- Lack of availability and cost of renewable energy, along with lack of capital to invest in power purchase agreements.
- Imprecise, complex and costly measurement, management, and mitigation of Scope 3 GHG emissions.
- The immaturity of circular economy practices and services.
- Costs of transition to more efficient cooling and HVAC systems, using technology such as immersion and free air cooling.
- Reducing water consumption.
- Lack of cost-effective low-carbon alternatives to diesel generators.

User Recommendations

- Secure a long-term supply of renewable energy.
- Measure the success of the data center sustainability program in broader enterprise sustainability initiatives by creating realistic KPIs for GHG emissions and water consumption.
- Conduct an audit of the data center's GHG emissions, waste and water consumption to understand its footprint. Liaise with the enterprise's sustainability, energy management, real estate and legal teams to build the business case and support the program.
- Reduce the data center's emissions footprint by investing in energy efficiency measures, GHG, and water and waste reduction.
- Follow through the full green value chain and do not consider the cloud as a legitimate offset of responsibility.
- Support the development of global industry/university consortia to focus on innovative solutions and standards for low-power computation, and data communication and storage.

Gartner Recommended Reading

Strategies to Plan for GHG Emissions Reduction

Building a Low-Carbon Energy Strategy

Ignition Guide to Building a Net-Zero Greenhouse Gas (GHG) Emissions Roadmap

Maverick Research: Net Zero Will Stall Tech Growth and Innovation

Toolkit for Estimating Data Center Build and Modernization Costs by Tier Level

Cloud Sustainability

Analysis By: Ed Anderson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Cloud sustainability is the use of cloud services to achieve sustainability benefits within economic, environmental and social systems. As such, cloud sustainability refers to both the sustainable operation and delivery of cloud services by a cloud service provider, as well as the consumption and use of cloud services by organizations and individuals to achieve sustainability outcomes.

Why This Is Important

Cloud sustainability is a key digital technology supporting organizations in their use of technology to achieve their sustainability ambitions. Cloud computing models are well-suited to deliver sustainability benefits because of their ability to operate at scale using a shared services model, which results in efficient use of computing resources. Hyperscale cloud data centers can be physically located near renewable energy sources further extending their potential to lessen environmental impact.

Business Impact

Increasing attention and focus on environmental and social issues is motivating organizations to improve their sustainability posture. Pressure from customers, investors, partners, regulators, employees, and the public at large is motivating organizations to establish sustainability goals and to demonstrate sustainability outcomes. Cloud computing has great potential to improve sustainability outcomes through efficient operations and the delivery of cloud-based technology innovations.

Drivers

- Sustainability is a rising imperative for organizations across all industries and in all countries and regions around the world. Although sustainability encompasses environmental, social and economic factors, environmental sustainability receives the most attention.
- Corporate climate and decarbonization commitments are typically cascaded to individual business functions, including IT. Consequently, IT organizations are looking at all possible ways to implement such strategies, including cloud sustainability initiatives.
- Market data shows that customers, investors, regulators, citizens and employees increasingly value organizations with demonstrable commitments to sustainability.
- Sustainability investments correlate with operational efficiency. Most organizations operating in an increasingly sustainable fashion also recognize other benefits such as reduced spending on energy, reductions in waste and improvements in water use.
- Cloud providers, being among the world’s largest data center operators, show strong commitments to cloud sustainability and are making demonstrable progress toward delivering sustainable cloud service offerings.
- Regulatory and legislative mandates for sustainability are increasingly common across regions and industries. The use of cloud services and other digital technologies will help organizations comply with future regulatory reporting requirements.

Obstacles

- Sustainability definitions, metrics and reporting standards are inconsistent, varying by region and industry. Defining, tracking and reporting sustainability performance is complex for most organizations.

- Cloud providers claim to have made great strides in offering sustainable cloud solutions, but these claims are often difficult to verify and contribute to potential "greenwashing." The lack of sustainability reporting standards makes it difficult to interpret and validate provider claims.

- Achieving cloud sustainability outcomes is a shared responsibility between the cloud provider and the customer. Cloud providers must demonstrate sustainable cloud operations, and cloud consumers must employ sustainability practices in their use of cloud services.

- Renewable energy is a key enabler of cloud sustainability and yet there is insufficient capacity to generate and store the energy required to meet the needs of the world's cloud service offerings.

User Recommendations

- Establish internal sustainability goals including specific metrics and sustainability outcomes by doing a materiality assessment to determine which sustainability outcomes are most important to your organization.

- Determine the role cloud sustainability will play in the achievement of sustainability outcomes. Build internal credibility for cloud sustainability by ensuring that the sustainability benefits of specific cloud service offerings are independently validated.

- Engage relevant executives and other internal stakeholders proactively that are tasked with creating and achieving sustainability goals. Establish credible metrics for measuring and reporting cloud sustainability outcomes.

- Look to cloud providers and other experts, including IT service providers, for best practices in operating and consulting cloud services in a sustainable manner.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Google; IBM; Microsoft; Oracle; Salesforce; SAP; Scaleway; VMware

Gartner Recommended Reading

Executive Leadership: Sustainability Primer for 2023

Quick Answer: How Green Are Public Cloud Providers?

Build an Environmental Cloud Sustainability Strategy

Make Sure Technology Helps More Than Hurts Sustainability

Sustainability: A Customer Priority and Provider Imperative

Distributed Integrated Infrastructure

Analysis By: Philip Dawson

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Distributed integrated infrastructure is general-purpose data center infrastructure on-premises that has applications and solutions integrated, marketed, and supported between vendors and partners. These can either be traditional infrastructure vendors or cloud providers' infrastructure.

Why This Is Important

Cloud infrastructure delivery and infrastructure as a service (IaaS) are now commonplace in a majority of clients. On-premises distributed integrated infrastructure is architected, funded, packaged and controlled or managed by the infrastructure or cloud provider. This allows the data center's cloud infrastructure to be newly delivered and experienced.

Business Impact

This on-premises distributed integrated infrastructure has been developed through a bimodal origin and approach:

- On-premises cloud infrastructure is delivered by cloud providers for governance and sovereignty, or other location issues such as network latency.

- On-premises cloud infrastructure is packaged by systems vendors for their infrastructure offerings and consumption models as IaaS-like to inhibit the overall move to the cloud and capture any fallout projects from the cloud.

Drivers

- Distributed integrated infrastructure is based on IaaS offerings that are aligned to an infrastructure control plane or management console that is aligned to the infrastructure provider. This may include infrastructure packaging, refresh, measurement and chargeback as a cloudlike infrastructure on-premises or in a data center.

- Distributed integrated infrastructure is delivered from either systems vendors or hyperscalers (who also package distributed integrated platforms). This type of consumption model is not new and similar to previous capacity-on-demand programs but has more cloudlike elasticity — primarily for the infrastructure.

- Distributed integrated infrastructure evolved from the virtualization of compute, storage and network delivered through integrated systems, software-defined infrastructure and hyperconvergence. This has become the foundation and management of everything as a service (XaaS) as an infrastructure platform and architecture for distributed integrated infrastructure.

- Distributed integrated infrastructure allows life cycles to be aligned between projects and services. It also allows technical debt to be reduced and multiple platforms to be delivered across a common cloudlike infrastructure on-premises.

- Third-party solutions and software platform as a service (PaaS) offerings are integrated on top of the infrastructure; either through vendor partnerships or channel partner integration. These PaaS offerings are good for existing tactical lift and shift and cloud migration.

Obstacles

- Common infrastructure for cloud delivery models makes sense for standardization, but alignment to a vendor's offering can create cloud silos tied to a vendor's technology that is difficult to integrate into hybrid or multicloud environments.
- Sticking to a single vendor while repackaging on-premises projects to cloudlike delivery leads to increased lock-in and potential issues around forecast consumption and infrastructure delivery quotas and compliance.
- Multiple third-party solutions sitting on top of a commonly distributed infrastructure platform can lead to platform contention and suboptimal dependence on third-party integration or channel partner service skills issues. Also, be aware of the XaaS and PaaS vendor's long-term cloud strategic roadmap.
- Software platform silos on common infrastructure are still silos, but technical debt overall is smoothed and reduced but investments rise as an offset.

User Recommendations

- Examine the realistic benefits of a commonly shared modernized infrastructure that distributed integrated infrastructure delivers — don't get bogged down with taxonomy.
- Prioritize the integration with third-party software platforms when selecting these solutions so as to minimize the risk of infrastructure vendor lock-in.
- Contrast the strength of distributed integrated infrastructure with third-party software platforms; looking for true multivendor partnerships, not just channel partner delivery for integration and support.

Sample Vendors

Amazon Web Services (AWS); Dell Technologies; Google; Hewlett Packard Enterprise (HPE); IBM; Microsoft; VMware

Gartner Recommended Reading

Predicts 2023: XaaS Is Transforming Data Center Infrastructure

Rationalizing Applications and Infrastructure for Cloud Delivery

How Do I Plan for Migrating My Data Center Infrastructure Into an XaaS Model?

Operational AI Systems

Analysis By: Chirag Dekate, Soyeb Barot, Sumit Agarwal

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Operational AI systems (OAISys) enable orchestration, automation and scaling of production-ready and enterprise-grade AI, comprising ML, DNNs and Generative AI. OAISys integrates DataOps, ModelOps, MLOps and deployment services to deliver enterprise-grade governance, including reusability, reproducibility, release management, lineage, risk and compliance management, and security. It also unifies development, delivery (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Why This Is Important

OAISys can help enterprises:

- Standardize, govern and automate AI engineering and deployment technologies, and accelerate productization of AI.
- Eliminate system integration friction and impedance mismatch across DataOps, ModelOps, MLOps, deployment and governance platforms.
- Scale AI initiatives by enabling orchestration across hybrid, multicloud, edge AI or IoT.
- Enable discoverable, composable and reusable AI artifacts (data catalogs, feature stores, model stores) across the enterprise context.

Business Impact

OAISys deliver production AI systems that:

- Systemize analytics and AI engineering technologies, including ModelOps and MLOps platforms.
- Integrate existing data, analytics and DSML platforms.

- Utilize reusability components including feature and model stores, monitoring, experiment management, model performance and lineage tracking.
- Homogenize governance including compliance, risk, security, and cost across deployment (hybrid, multicloud, IoT) and operational (streaming, batch) contexts.

Drivers

- Enable business stakeholders to leverage AI as a service that is customized to their enterprise context.
- IT leaders need to deliver, manage and govern AI models within enterprise applications deployed across multiple contexts and jurisdictions (hybrid, multicloud, edge AI and IoT).
- Traditional siloed approaches of data management and AI engineering create integration challenges across the data ingest, processing, model engineering and deployment.
- OAISys enables enterprises to standardize and automate development, management, deployment, maintenance and governance technologies to deliver comprehensive, flexible and composed end-to-end AI systems.
- It helps align and automate the data, AI model deployment and governance pipelines.
- Operationalization and automation platforms are a core part of how early enterprise AI pioneers scale productization of AI by leveraging existing data, analytics and governance frameworks.
- Standardizing data pipelines, including DataOps toolchains, creating reusability components such as data catalogs and ETL registries, monitoring, security, access control and lineage tracking.
- The enterprise OAISys enables unification of two core contexts: deployment context across hybrid, multicloud, edge AI and IoT, and operational context across batch and streaming processing modes that commonly occur as enterprises train and deploy production models.

Obstacles

- Enterprises with low data and AI maturity levels will find OAISys intimidating to build, deliver and support.
- OAISys requires integration of full-featured solutions with select tools that address portfolio gaps with minimal overlap. These include capability gaps around feature stores, model stores, governance capabilities and more.
- OAISys requires a high degree of cloud maturity, or the ability to integrate data and model pipelines across deployment contexts. The potential complexity and costs may be a deterrent for organizations just starting their AI initiatives.
- Enterprises seeking to deliver OAISys often seek “unicorn” experts and service providers to productize AI. Fully featured vendor solutions that enable OAISys are hard to come by, and enterprises often have to build and support these environments on their own.

User Recommendations

- Focus AI engineering activities to deliver business context customized operational AI systems.
- Rationalize data and analytic environment and leverage current (simplified subset of) investments in data management, DSML, ModelOps and MLOps tools to build OAISys.
- Leverage cloud service provider environments as foundational environments to build OAISys along with rationalizing your data, analytics and AI portfolios as you migrate to the cloud.
- Avoid building patchwork OAISys that integrate piecemeal functionality from scratch (and add another layer of tool sprawl). Utilize point solutions sparingly and surgically to plug feature/capability gaps in fully featured DataOps, MLOps and ModelOps tools.
- Actively leverage your existing data management, DSML, MLOps and ModelOps platforms as building blocks, rather than starting from scratch.

Sample Vendors

Amazon Web Services; Dataiku; DataRobot; Domino Data Lab; Google; HPE Ezmeral Software; IBM; Iguazio; Microsoft; ModelOp

Gartner Recommended Reading

2023 Planning Guide for Analytics and Artificial Intelligence

Emerging Tech Impact Radar: Data and Analytics

Quick Answer: How Should CXOs Structure AI Operating Models?

Distributed Integrated Platforms

Analysis By: Philip Dawson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Distributed integrated platforms are infrastructure and application platforms deployed in data centers that have infrastructure platforms and solutions or applications integrated, marketed and supported by a primary vendor or provider.

Why This Is Important

Cloud platform delivery is now commonplace in a majority of clients and this is driving a competitive response with server vendors delivering software platform integration as a packaged platform as a service (PaaS). Distributed integrated platforms in data centers are architected, funded, packaged and controlled or managed by the application or cloud platform provider and their partners. This allows new application and cloud infrastructure experience for data center delivery.

Business Impact

Distributed integrated platforms combine application and data layers of integrated infrastructure. This packaging and consumption is similar to integrated systems with more cloudlike elasticity — for both the infrastructure and application across everything as a service (XaaS). These offerings use a control plane or management console aligned with the application platform provider. This may include infrastructure and software platform or application packaging, measurement and chargeback.

Drivers

- On-premises distributed integrated platforms are different from distributed integrated infrastructure as they are dedicated solutions in a similar way to that of traditional integrated systems and integrated platforms whereas integrated infrastructure is for general-purpose applications integrated separately.

- The on-premises element can alleviate the data proximity, data residency and location-based network issues for providers, the control plane audit and compliance capabilities, trade risk and responsibility for remote governance by the application platform vendor. This control shift to the application provider is inhibitive to the overall adoption of a remotely managed solution in an on-premises environment to see both sides of the trade-off.

- Adoption is also more likely where data center staff lack deep integration support skills or where I&O teams wish to focus solely on application delivery rather than infrastructure services.

- Distributed integrated platforms take the packaging of data platforms and/or application platforms on top of the distributed infrastructure. This allows for improved management and operational efficiency as part of the overall integration with the application vendor on their own infrastructure.

- Distributed integrated platforms allow life cycles to be aligned between infrastructure projects and applications services and technical debt to be reduced across a single platform. This is delivered across a common on-premises cloud platform.

Obstacles

- While the primary PaaS stack and application are more integrated with distributed integrated platforms, third-party solutions and software PaaS offerings are more difficult to integrate on top of the infrastructure and platform. Third-party support can be exposed for application integration and transformation is at the provider's pace not the user's.

- Common infrastructure for cloud delivery models makes sense, but alignment to a vendor's offering can create cloud silos tied to vendors that are difficult to integrate into hybrid or multicloud environments, but technical debt overall is suboptimized and reduced.

- While repackaging on-premises platforms to cloudlike delivery, avoid the same vendor increased lock-in and potential issues around forecast consumption, licensing, and platform delivery quotas and compliance.

- This type of integrated platform does reduce technical debt overall but investment and costs increase to maintain the current nature of cloud platforms.

User Recommendations

- Look at the realistic benefits of a common, shared, modernized infrastructure and application platforms that distributed integrated platforms deliver instead of getting bogged down with taxonomy.

- Prioritize the integration with third-party applications and software platforms. Don't overinvest in a single vendor's infrastructure and application management and tooling.

- Evaluate the strength of distributed integrated platforms and look for true partnerships and integration in geographic and vertical routes to market.

- Manage a reduction of technical debt by investing in cloud and consumption models absorbing costs as part of XaaS and platform delivery.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Microsoft; Oracle

Gartner Recommended Reading

Quick Answer: How Can I Optimize the Use of Programmable Platforms for Effective Software Delivery?

Rationalizing Applications and Infrastructure for Cloud Delivery

How Do I Plan for Migrating My Data Center Infrastructure Into an XaaS Model?

Intelligent Platforms

Analysis By: Philip Dawson, Nathan Hill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Intelligent platforms provide the administration composability of infrastructure and programmable API functions with automated infrastructure intelligence. They integrate compute, storage and networking assets with some or the entire application software stack, creating dedicated workload architectures. Intelligent platform vendors also include components such as application intimacy, management tools, OSs, and virtualization bought and/or consumed as a service.

Why This Is Important

Intelligent platform solutions are differentiated against integrated system or hyperconverged infrastructure (HCI) solutions with a separate software stack purchase tied to the hardware. Pricing strategies vary greatly throughout the integrated software stack solutions as part of the shift to consumption-based infrastructure delivery. Intelligent platforms also integrate applications and business logic as bundles and partnerships.

Business Impact

Intelligent platforms optimize:

- Delivery of workload performance or application manageability that crosses over from hardware that promises lower operational costs and increased IT agility via automated, pooled resources.
- Automation and machine learning of complete stacks, hardware administration and software programmability on top of integrated systems.

- They are stand-alone running proprietary workloads that rarely compete with each other as the software stacks set the hardware options.

Drivers

- The intelligent platform market is influenced by multiple aspects of resilience and availability across on-premises, hosting or colocation and cloud locations driving composable, programmable and intelligent functions.
- Intelligent platforms are integrated as everything as a service (XaaS), with automation and management, and differ from integrated stack systems, which are hardware-integrated dedicated appliances.
- Multiple vendors are driving the market for intelligent platforms around integrated systems, HCI, cloud and virtualization. Intelligent platforms are built from a software perspective on top of HCI rather than a traditional integrated stack system that is built as a hardware appliance around hyperconverged integrated systems (HCIS).
- Vendors such as Microsoft, Nutanix and VMware are promoting valid intelligent platform software, and the market momentum around HCI software in the cloud now creates a market for multiple hardware vendors to build software management and integration services.

Obstacles

- Hybrid and multicloud strategies may not integrate well with integrated platforms, continuing the silo mentality of cloudlike delivery.
- Other platform as a service (PaaS) momentum is being integrated from packaged vendors such as SAP and Oracle, which are bundling integrated stack systems and distributed cloud infrastructure with application platform and database management system (DBMS) software. Here, the intelligence is with the PaaS software, not the intelligent infrastructure.
- An intelligent platform provides balanced XaaS workload performance, application optimization and integration, but this comes at the expense of greater vendor dependency, and inflexibility for future application customization and workload requirements.

User Recommendations

- Select infrastructure software management frameworks for overlay management as well as links to cloud infrastructure. Do not implement hardware-dependent or locked-in intelligent platform frameworks and adapters.
- Define successful intelligent platform implementations by assessing data center stakeholders and other vested interests (for example, procurement) with other lines of business responsible for agreeing with SLAs.
- Automate the infrastructure requirements for cloud management platforms (CMPs) through the use of intelligent platforms as you deliver XaaS through infrastructure platforms.

Sample Vendors

CU Coding; DataDirect Networks (DDN); Dell Technologies; Hewlett Packard Enterprise; Microsoft; Nutanix; Oracle; VMware

Gartner Recommended Reading

How to Evolve Your Physical Data Center to a Modern Operating Model

Quick Answer: How Can I Optimize the Use of Programmable Platforms for Effective Software Delivery?

How Do I Plan for Migrating My Data Center Infrastructure Into an XaaS Model?

Container-VM Convergence

Analysis By: Michael Warrilow, Tony Iams

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Container-virtual machine (VM) convergence refers to the fusion of hypervisor and operating system (OS)-based virtualization technologies. By integrating and optimizing the most desirable features of containers and virtual machines, container-VM convergence delivers improved workload isolation and greater infrastructure utilization. Underlying technologies include optimized virtual machine monitors (hypervisors) integrated with container runtimes.

Why This Is Important

Containers appeal to the need for modern, agile infrastructure, whereas virtual machines are a foundational element of data center infrastructure. Container-VM convergence promises the “best of both worlds” and new options for future infrastructure needs. Using innovation from public cloud providers and open-source communities, container-VM convergence creates opportunities to improve infrastructure elasticity, scalability and efficiency, while supporting modern application development.

Business Impact

Container-VM convergence enables greater infrastructure agility without sacrificing security. By infusing a cloud-inspired approach, it enables I&O teams to manage hypervisor and OS-based virtualization together. It also supports initiatives to increase developer and operational productivity. By supporting a gradual transition to cloudlike infrastructure for both on- and off-premises, it preserves existing infrastructure investments and reduces additional spend over the medium term.

Drivers

- Increasing interest is being driven by disruption in the server virtualization market and the desire for increased competitive negotiation power. This has resulted from the market dominance of a small number of server virtualization providers with few alternatives for on-premises requirements. Container-VM convergence has the potential to do this and further disrupt the competitive landscape before 2025.

- I&O teams struggle to meet the demands for cloudlike infrastructure in noncloud environments. A major factor for this deficiency is that traditional IT infrastructure is limited in the ability to deliver modern requirements, such as immutable infrastructure, agile delivery and API-driven management. Container-VM convergence provides an infrastructure technology to bridge these needs.

- CIOs require technologies such as containers to satisfy digital business requirements. This is because they are investing in areas such as digital business transformation, API integration and legacy application modernization. Containers are becoming a foundational element required to leverage many emerging technologies (cloud-native infrastructure, AI/ML and edge computing).

- Container-VM convergence can support current and future I&O requirements, and help balance rising financial pressures that will increasingly require I&O to refocus on IT cost optimization. If existing infrastructure assets are maintained for longer than was originally anticipated, container-VM convergence can reduce resulting pressure to continue supporting digital transformation. This is achieved by avoiding investment in new tools, skills and processes.

- The financial outlay is limited to upgrading existing infrastructure software; however, the benefit will be achievable across traditional and future requirements.

Obstacles

- Practices and technologies related to container-VM convergence favor Linux. For non-Linux enterprise workloads, a different operation/production support model must be maintained.
- Successful adoption of new operational approaches requires cultural change. This may constrain the uptake of container-VM convergence, leading to friction with application developers, business-led IT projects and/or circumvention of existing I&O capabilities. In turn, this will lead to compliance issues in regulated, high-security environments.
- Similarly, financially constrained organizations may struggle to see the benefit of additional investment over and above existing sunk costs.
- Container-VM convergence is not an established technology in enterprise environments.
- Container-VM convergence is yet to establish itself as being suited to the complexity of on-premises legacy environments.

User Recommendations

- Plan for container-VM convergence to deliver modern, cloud-inspired infrastructure supporting enterprise developers, while maintaining operational management and security for production environments.
- Monitor developments in enterprise adoption of open-source projects related to container-VM convergence, including Kata Containers, gVisor and Firecracker.
- Establish and maintain consensus on the role of virtual infrastructure in data center, distributed cloud, public cloud and edge — recognizing that various virtualization technologies may be necessary.
- As a precursor to selecting container-VM convergence, conduct or utilize a skills inventory. An I&O skills roadmap will help to verify existing and future skills needed to support emerging technology trends like this. Where skills deficiencies are identified, adopt container-VM convergence in a more gradual fashion that supports future requirements without detracting or unnecessarily complicating existing service-level requirements.

Sample Vendors

Alibaba Cloud; Amazon Web Services (AWS); Fly; Google; Microsoft; Red Hat; SUSE; VMware

Gartner Recommended Reading

Market Guide for Server Virtualization

Market Guide for Container Management

Revirtualization

Analysis By: Philip Dawson, Michael Warrilow

Benefit Rating: Low

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Revirtualization is the introduction of a replacement hypervisor-based virtualization technology and associated software for management and resilience for live migration and host-based recovery. It is initiated through change of delivery models (i.e., cloud or edge) or through broader vendor divestiture, merger and acquisition.

Why This Is Important

Anecdotal evidence from Gartner clients indicates that, traditionally, it has not been common for them to "revirtualize." We are seeing more everything as a service (XaaS) and hybrid cloud delivery from private, public and cloud on-premises. As a result, organizations are increasingly moving to an alternate cloud-provider-tied hypervisor, built on Open Source. Clients are also looking at ways of defining the scope of extending enterprise agreements virtualization vendors as they move to subscription models.

Business Impact

Client feedback indicates revirtualization offers limited improvement in total cost of ownership, immature administrative and management tooling, increased operational burden, and concerns about suitability to meet enterprise-scale requirements. There are a lot of transitional costs and investments without related transformational and business benefits. However, this element of risk is introduced to offset exposure with increased audit and contractual issues with private cloud virtualization providers.

Drivers

- Revirtualization often occurs as part of migration, transition and transformation to cloud delivery models for on-premises or hosted environments.

- Revirtualization also can be introduced due to commercial audit and license issues or vendor merger, acquisition or divestiture. We have seen a large base of private cloud customers questioning the impact and status of Dell's divestiture of VMware and the subsequent ongoing acquisition by Broadcom.

- Hypervisors are easily substituted with open-source commercial alternatives. This is often tied to a software-defined infrastructure (SDI) and storage and network upsell as a packaged hyperconverged alternative for private cloud deployments.

- Our research indicates that over half of hyperconverged infrastructure (HCI) vendor instances ship with the same vendor's hypervisor fueling revirtualization.

- Extending an existing cloud workload placement policy to virtual workloads benefits the evaluation of revirtualization alternatives.

- Modern application architectures designed for agility, scalability and elasticity can benefit from the agility of containers. This has raised the expectations of agility and versatility for a future server virtualization platform and revirtualization alternatives.

- Revising infrastructure environments to leverage automation in infrastructure as a service plus platform as a service (IaaS+PaaS) — without refactoring applications — can be significantly more cost-effective (and is often referred to as lift and optimize).

Obstacles

- Server virtualization within an enterprise requires much more than just a hypervisor. It encompasses a variety of operational activities and integration across a broad ecosystem of infrastructure and applications.
- Enterprises have already found it challenging to replace incumbent hypervisors and server virtualization platforms for their on-premises private cloud workloads. This also has been apparent with storage virtualization and hyperconvergence.
- Application modernization cannot be bypassed by revirtualization. Public cloud lift and shift alone is suboptimal for most traditional workloads. Server virtualization platforms have typically been procured as perpetual licenses with annual software maintenance. Such purchases are treated as capital expenditure (capex) under common accounting practices. This is in contrast to most revirtualization subscription-based licensing, which is treated as operational expenditure (opex).

User Recommendations

- Review revirtualization and the strategic, long-term role of server virtualization for infrastructure and application modernization.
- Identify any available material used for evaluating relevant alternatives (HCI, public cloud and/or container adoption) to avoid rework.
- Take advantage of Gartner and other external resources.
- Create a scoping revirtualization task force by identifying sufficient cross-disciplinary technical expertise (network, storage, compute, application, security, architecture, IT operations, procurement and compliance).
- Communicate regularly with key stakeholders and senior IT leadership.
- Focus on achieving business-defined revirtualization objectives, reducing technical debt and controlling vendor lock-in. Price is not the only factor to be considered.
- Incorporate preferred revirtualization option(s) within a formal, architecture-driven workload placement policy covering both existing and new workloads.

Sample Vendors

Citrix; Microsoft; Nutanix; Oracle; VMware

Gartner Recommended Reading

Market Guide for Server Virtualization

Market Guide for Full-Stack Hyperconverged Infrastructure Software

Quick Answer: How Should VMware Customers Prepare for the Broadcom Acquisition?

Quick Answer: What Would Be Required for a Large-Scale Migration From VMware's Server Virtualization Platform?

Managing Your Dependence on VMware: Identifying Contingencies for vSphere

At the Peak

Programmable Platform

Analysis By: Philip Dawson, Bill Blosen

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Programmable platforms are API-driven for delivery of applications in a cloud model by using and applying methods and tooling from the software development area to management of IT infrastructure and platform concepts. It includes resilient platform architectures and agile techniques.

Why This Is Important

Software engineering and infrastructure and operations (I&O) have coexisted but been separated by their differing toolsets and APIs. Modern digital businesses need their software engineering and I&O teams to deliver a cohesive platform that encompasses both application and infrastructure delivery. Programmable infrastructure delivers the underlying technical capability that enables this integration.

Business Impact

Greater value and agility (rather than cost optimization) is achieved via programmable platforms' ability to drive adaptive application delivery. Programmable infrastructure, API provisioning and automated processes allow faster responses to new business demands, driving service quality and freeing application delivery staff and administration staff from infrastructure functions. Programmable platforms enable a sustainable and highly responsive IT infrastructure service to the business.

Drivers

- Software architects and engineers are moving to modular distributed applications built on containerization, control, data and service architecture. In essence, they use the pattern of separating the application front end from the business logic back end.
- API layers are being adopted through self-service capabilities and organizing the programmable platform APIs into paved roads.

- Software engineers are reducing the cognitive load of using APIs and programmable platforms, improving developer experience, driving productivity and retention of key talent, and also improving adherence to architectural and security guardrails.

- I&O teams are moving workloads and application delivery to cloud infrastructure and platforms as in anything-as-a-service (XaaS) models. In essence, they have embraced programmable infrastructure, that is, applying software development methods, APIs and tooling to manage the control and data planes around I&O services.

- The incumbent architecture of programmable platforms is deploying modular building blocks. Updates to modules are automatically rolled out to any platform that is currently using that module rather than updating each platform discreetly, leveraging efficiency, scale and management dependencies.

- Platform engineering principles guide the programmable platform buildout using templates, APIs and automation to simplify the usage and adoption of programmable infrastructure. The platforms being built must respond to the developers pain points and ease adoption of the most used functions. Agile product ownership is the best model to build feedback loops between the developer programmable platform teams and the platform's users and communities.

Obstacles

- The boundaries between programmable platforms and infrastructure are at best emerging. I&O teams must be conscious of how their architectures and deployments interface with the programmable infrastructure at the control plane and data plane.

- The architecture and order of APIs across and up the stack and across layers need planning and integration, increasing lock-in. I&O and software teams conceptualize these layers as part of one programmable platform with four distinct layers: the application presentation front end, the business logic back-end functions, the control or management, and data tier repository. Moreover, programmable platforms are restricted in both topology and maturity, which drives clients toward platform as a service (PaaS) and cloud delivery.

- Programmable platform governance is limited by the lack of standardization of the APIs. In balance, established mature APIs drive engagement or interface between the layers managed through well-defined and structured APIs.

User Recommendations

- Use platform engineering principles to improve developer experience and ease the cognitive load of programmable platforms. Before considering programmable platforms, balance the aversion to vendor specific services lock-in which is prevalent in PaaS and XaaS.
- Deliver platform engineering principles by providing developer platforms to abstract and address the complexity of the APIs.
- Design programmable platform control plane APIs to not only monitor and manage consumption and provision, but also provide governance and compliance guardrails.
- Use APIs for SLAs, chargeback and consumption models led by the drive of standardization and automation of cloud delivery with programmable platforms.

Sample Vendors

Avesha; CU Coding; Microsoft; Oracle; SAP; Silk

Gartner Recommended Reading

Quick Answer: How Can I Optimize the Use of Programmable Platforms for Effective Software Delivery?

Adopt Platform Engineering to Improve the Developer Experience

A Software Engineering Leader's Guide to Improving Developer Experience

Consumption-Based Model

Analysis By: Jeff Vogel, Philip Dawson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

A consumption-based sourcing model strategy for hybrid cloud on-premises data center storage, compute and networking infrastructure is an acquisition, deployment and support model that includes a cloud-like pay-for-use and platform services model optimized for predictable usage.

Why This Is Important

The consumption-based model provides IT operations with an on-premises cloud-like operating model for storage, compute and networking. It eliminates capital expenditure (capex) financing, simplifies capacity planning and optimizes asset usage to actual workload use, effectively aligning asset costs-to-value. It has brought a whole new way of procurement sourcing and asset consumption, with pay-as-you-use and as-a-service platforms becoming the preferred deployment methodology for storage and compute.

Business Impact

A consumption-based sourcing model and services strategy will:

- Shift responsibility for maintenance and support costs to vendors investing in AI for IT operations (AIOps) to automate IT administration.
- Preserve cash by avoiding upfront capex in exchange for strategic priorities.
- Shift IT and finance resource budget cycles to a services-based platform delivery model.
- Provide more flexible and agile IT operations aligned with business demands.

Drivers

Infrastructure and operations (I&O) leaders are embracing cloud-native hardware and software consumption models as a strategy to replace owned, on-premises infrastructure and to lower data center operations' costs. This trend is driven by:

- The need for a more flexible cloud-like operating model for on-premises infrastructure.
- The massive growth of enterprise data that makes capacity planning difficult and upfront purchasing for three to five years of growth expensive and impractical.
- Prolonged procurement lead time increases due to persistent supply shortages.
- The need for an application-aware services delivery model.
- The preference for operating expenditure (opex) to capex with cloud-like benefits, while avoiding risks or costs associated with moving mission-critical workloads to the public cloud.

- The need for a more cost-effective, flexible and efficient sourcing strategy that aligns with business demands.
- The need to augment IT budget priorities to redirect investments to develop cloud-native platform skills that support business growth initiatives.
- The shift from exiting the life cycle management of infrastructure assets in the long term to freeing up IT resources.

Obstacles

A consumption-based sourcing model may:

- Be more expensive than capex financing.
- Be organizationally challenging to implement.
- Be unsuitable for IT operations that have a more stable and predictable growth and variability in forecast demand or lean toward sweating assets.
- Require minimum-usage commitment levels that can't be justified regardless of what is actually consumed.
- Require three- to five-year contracts with vendor-centric services.
- Lack the skills or culture alignment to shift from sourcing products to platform SLA services.
- Not take into account long-term supply chain price fluctuations during the contract period, when declining hardware costs or supply constraints are considered.
- Conflict with financial asset depreciation and amortization schedules or corporate balance sheet objectives.
- Conflict with established industry accounting standards and operational norms.
- Software licensing terms may be incompatible with the use of consumption based hardware.

User Recommendations

- Adopt a cloud operating model as a platform services strategy to shift to ITOps-as-a-service to increase productivity and flexibility.
- Organize and implement a joint team approach to include I&O, vendor management and finance to establish a strategic sourcing strategy.
- Rightsize and align IT I&O resources to a consumption-based platform model to free up resources to focus on business priorities.
- Assess the economics and requirements against a range of vendor consumption programs before committing.
- Ensure that contract terms match financial requirements, accounting for capex versus opex, and that contracts include appropriate end-of-term options, such as book value buyout.
- Address licensing options and term constraints as they pertain to usage.
- Link consumption-based costs to specific usage level requirements along with remediation terms to enforce minimum levels.
- Retire legacy technical debt and onerous support fees, and modernize systems and processes.

Sample Vendors

Cisco; Dell Technologies; Hewlett Packard Enterprise; IBM; Lenovo; NetApp; Pure Storage

Gartner Recommended Reading

Market Guide for Consumption-Based Models for Data Center Infrastructure

Competitive Landscape: Consumption-Based Model for On-Premises Infrastructure

Quick Answer: How Can I Use Storage as a Service to Reduce IT Spend?

Generative AI

Analysis By: Svetlana Sicular, Brian Burke

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Generative AI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound business impacts, including on content discovery, creation, authenticity and regulations; automation of human work; and customer and employee experiences.

Why This Is Important

Generative AI exploration is accelerating, thanks to the popularity of Stable Diffusion, Midjourney, ChatGPT and large language models. End-user organizations in most industries aggressively experiment with generative AI. Technology vendors form generative AI groups to prioritize delivery of generative-AI-enabled applications and tools. Numerous startups have emerged in 2023 to innovate with generative AI, and we expect this to grow. Some governments are evaluating the impacts of generative AI and preparing to introduce regulations.

Business Impact

Most technology products and services will incorporate generative AI capabilities in the next 12 months, introducing conversational ways of creating and communicating with technologies, leading to their democratization. Generative AI will progress rapidly in industry verticals, scientific discovery and technology commercialization. Sadly, it will also become a security and societal threat when used for nefarious purposes. Responsible AI, trust and security will be necessary for safe exploitation of generative AI.

Drivers

- The hype around generative AI is accelerating. Currently, ChatGPT is the most hyped technology. It relies on generative foundation models, also called “transformers.”

- New foundation models and their new versions, sizes and capabilities are rapidly coming to market. Transformers keep making an impact on language, images, molecular design and computer code generation. They can combine concepts, attributes and styles, creating original images, video and art from a text description or translating audio to different voices and languages.

- Generative adversarial networks, variational autoencoders, autoregressive models and zero-/one-/few-shot learning have been rapidly improving generative modeling while reducing the need for training data.

- Machine learning (ML) and natural language processing platforms are adding generative AI capabilities for reusability of generative models, making them accessible to AI teams.

- Industry applications of generative AI are growing. In healthcare, generative AI creates medical images that depict disease development. In consumer goods, it generates catalogs. In e-commerce, it helps customers “try on” makeup and outfits. In manufacturing, quality inspection uses synthetic data. In semiconductors, generative AI accelerates chip design. Life sciences companies apply generative AI to speed up drug development. Generative AI helps innovate product development through digital twins. It helps create new materials targeting specific properties to optimize catalysts, agrochemicals, fragrances and flavors.

- Generative AI reaches creative work in marketing, design, music, architecture and content. Content creation and improvement in text, images, video and sound enable personalized copywriting, noise cancellation and visual effects in videoconferencing.

- Synthetic data draws enterprises’ attention by helping to augment scarce data, mitigate bias or preserve data privacy. It boosts the accuracy of brain tumor surgery.

- Generative AI will disrupt software coding. Combined with development automation techniques, it can automate up to 30% of the programmers’ work.

Obstacles

- Democratization of generative AI uncovers new ethical and societal concerns. Government regulations may hinder generative AI research. Governments are currently soliciting input on AI safety measures.
- Hallucinations, factual errors, bias, a black-box nature and inexperience with a full AI life cycle preclude the use of generative AI for critical use cases.
- Reproducing generative AI results and finding references for information produced by general-purpose LLMs will be challenging in the near term.
- Low awareness of generative AI among security professionals causes incidents that could undermine generative AI adoption.
- Some vendors will use generative AI terminology to sell subpar “generative AI” solutions.
- Generative AI can be used for many nefarious purposes. Full and accurate detection of generated content, such as deepfakes, will remain challenging or impossible.
- The compute resources for training large, general-purpose foundation models are heavy and not affordable to most enterprises.
- Sustainability concerns about high energy consumption for training generative models are rising.

User Recommendations

- Identify initial use cases where you can improve your solutions with generative AI by relying on purchased capabilities or partnering with specialists. Consult vendor roadmaps to avoid developing similar solutions in-house.
- Pilot ML-powered coding assistants, with an eye toward fast rollouts, to maximize developer productivity.
- Use synthetic data to accelerate the development cycle and lessen regulatory concerns.
- Quantify the advantages and limitations of generative AI. Supply generative AI guidelines, as it requires skills, funds and caution. Weigh technical capabilities with ethical factors. Beware of subpar offerings that exploit the current hype.
- Mitigate generative AI risks by working with legal, security and fraud experts. Technical, institutional and political interventions will be necessary to fight AI's adversarial impacts. Start with data security guidelines.
- Optimize the cost and efficiency of AI solutions by employing composite AI approaches to combine generative AI with other AI techniques.

Sample Vendors

Adobe; Amazon; Anthropic; Google; Grammarly; Hugging Face; Huma.AI; Microsoft; OpenAI; Schrödinger

Gartner Recommended Reading

Innovation Insight for Generative AI

Emerging Tech Roundup: ChatGPT Hype Fuels Urgency for Advancing Conversational AI and Generative AI

Emerging Tech: Venture Capital Growth Insights for Generative AI

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

ChatGPT Research Highlights

Multicloud

Analysis By: David Smith

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Multicloud computing is the use of multiple public cloud providers to provide the same general class of IT solution, workload, application or use case. It is much more common in infrastructure as a service (IaaS) and converged IaaS/platform as a service (PaaS) scenarios than SaaS. While multi-SaaS environments are possible, these would typically be stovepiped situations.

Why This Is Important

Multicloud has the potential to lower the risk of cloud provider lock-in, can provide best-of-breed capabilities for specific use cases and can provide service resilience and migration opportunities, in addition to the core cloud benefits of agility, scalability and elasticity. It also may be used to obtain public cloud services in different geographic locations for global companies.

Business Impact

Multicloud provides agility and can also provide a basis to lower cloud provider lock-in and increase workload migration opportunities. However, multicloud can also create additional complexity and, therefore, cost increases. Also, many organizations find that a multicloud environment is unavoidable for most.

Drivers

- Many organizations end up in a multicloud environment through acquisitions and mergers. Unintended multicloud environments can be rationalized into a purposeful multicloud strategy.

- Enterprises typically start with one provider and focus first on costs, but, over time, become concerned about lock-in. Thus, the first use of multicloud is often based on procurement issues to encourage competition, or as result of a merger and acquisition.

- As multiple cloud providers are in use, the need to manage and govern those services becomes important. Eventually, some enterprises adopt multicloud architectures. This approach relies on architectural principles and portability solutions, and can potentially enable even cloudbursting and other dynamic placement efforts.

- Many deliberate multicloud strategies are designed to take advantage of differentiated capabilities within the same general class (e.g., IaaS) from multiple cloud providers while applications run in a single cloud provider stack. Some applications may have a multicloud architecture themselves.

- The hype around multicloud is driving adoption, as providers often use this industry buzz term to justify why their offerings should be considered when another cloud service already exists.

Obstacles

- Multicloud is often confused with hybrid cloud. The reality is that multicloud and hybrid cloud often coexist in a multi-hybrid cloud environment that spans multiple public cloud providers, as well as between public and private implementations.

- Multicloud is not a practical solution for improving availability and enhancing disaster recovery or business continuity, as these goals are more effectively achieved in other ways within a provider's ecosystem.

- Multicloud environments are complex and often result in cost increases. Effort and cost are more often required to secure and manage multiple cloud environments. Organizations need to invest in the right skills to manage and deal with more complex integration solutions.

User Recommendations

- Ensure your multicloud strategy is coordinated with your overall cloud strategy. When embracing multicloud approaches, account for the tools, skills, processes and other resources to ensure you will achieve the right outcomes.
- Establish security, management, governance guidelines and standards to manage cloud service sprawl and increasing costs, and develop criteria for deciding placement of services.
- Focus on coordination and strategy across the enterprise to identify the types of services needed to deliver the benefits of a multicloud environment. Be prepared to incur additional expenses on training and skill development across roles, including engineers and operators.
- Do not just shift vendor lock-in to a cloud management platform (CMP) and/or a cloud service brokerage (CSB), even though they may enable governance and optimizations in a multicloud environment.

Gartner Recommended Reading

The Cloud Strategy Cookbook, 2023

A Multicloud Strategy Is Complex and Costly, but Improves Flexibility

A CTO's Guide to Multicloud Computing

BMaaS

Analysis By: Bob Gill, Philip Dawson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Bare metal as a service (BMaaS) supplies physical infrastructure (e.g., compute, networking and storage) via a cloudlike consumption model. BMaaS differs from infrastructure as a service (IaaS) in that the provider offers physical infrastructure dedicated to a specific user at the individual host level, and users provide all of the software installed into it. A provisioning layer coordinates requests for specific infrastructure combinations to discrete equipment in the provider's data center.

Why This Is Important

BMaaS runs workloads without hypervisor or OS compatibility restrictions on workload performance. This improves performance (no sharing/overhead), security (no sharing) and uptime (nothing else brings the system down). BMaaS is often chosen over virtual public cloud infrastructures to conform to legacy software licensing needs, based on permanent deployment onto fixed physical hosts. BMaaS isn't new, but it is gaining momentum augmenting, rather than replacing, on-premises equipment.

Business Impact

- BMaaS offers the advantages of dedicated infrastructure (e.g., predictability, security and performance) with elasticity closer to IaaS than actual physical deployments.

- For example, it provides a cloudlike experience in a data center location better suited to customer needs for low network latency and data residency.

- BMaaS supplies a flexible integration platform at the nexus of public cloud access locations, such as colocation hubs or content delivery network (CDN) points of presence (POPs).

Drivers

- Include the capability to act like a public cloud, rather than a dedicated hosting environment — programmable automation, elastic scalability down to the individual host level, and pay-as-you-go (PAYG) economics and consumption models.
- There is interest in cloud-native technologies as a path toward cloud independence that reduces lock-in.
- Bare metal may solve the issue of physical workload location, addressing the concerns that highly centralized offerings may pose, due to latency concerns, enterprise control, or data sovereignty and regulations.
- Bare metal offers the speed and agility of the public cloud, with far greater control over workload and data placement.
- The noncontinuous use of bare metal can be less costly than physical infrastructure; it does not tie up capital expenditure (capex) and is faster to deploy operationally as operating expenditure (opex).

Obstacles

- Adding another infrastructure environment increases complexity.
- Customer or service providers must supply and configure much of the software, bearing the risk and cost of a greater portion of the full stack.
- Unique network offerings may be required or multiple offerings may need to be integrated.
- Ease and flexibility of consumption may vary, especially up from infrastructure into application delivery.
- Economics may vary by application delivery, workload type, networking and included storage services.

User Recommendations

- Build BMaaS into cloud assessment models by identifying the attributes that can be addressed only through the software licensing compatibility, hypervisor independence and the location specificity of bare metal.
- Leverage bare metal's unique location benefits by identifying applications that require low latency or sovereignty through proximity to cloud onramps.
- Select BMaaS for "cloud-native hosting" of legacy applications, with licensing terms optimized for dedicated physical hosts.

Sample Vendors

Amazon Web Services; Cyxtera; Digital Realty Trust; Equinix; Oracle; Rackspace Technology

Gartner Recommended Reading

Break Down 3 Barriers to Cloud Migration

Distributed Cloud

Analysis By: David Smith, Daryl Plummer, Milind Govekar, David Cearley

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Distributed cloud refers to the distribution of cloud services to different physical locations, while operation, governance, updates and evolution of the services are the responsibility of the originating cloud provider. Distributed cloud computing is a style of cloud computing where the location of cloud services is a critical component of the model.

Why This Is Important

Distributed cloud enables organizations to use consistent cloud-based services wherever needed, while the cloud service provider retains the responsibility of managing the technology, implementation and evolution of the capabilities. It gives organizations the flexibility to support use cases that will benefit from cloud services, regardless of their dependence on specific locations. Organizations can use distributed cloud to reimagine use cases where cloud computing is not currently feasible.

Business Impact

A major notion of the distributed cloud concept is that the provider is responsible for all aspects of delivery and manages the distributed capabilities “as a service.” This restores cloud value propositions that are broken when customers are responsible for a part of the delivery, as is true in private and some hybrid cloud scenarios. The cloud provider must take responsibility for how the overall system is managed. Otherwise, the value proposition of distributed cloud is compromised.

Drivers

- Historically, location has not been relevant to cloud computing definitions. In fact, the variations on cloud (e.g., public, private, hybrid) exist because location can vary.

- Distributed cloud supports both tethered and untethered operations of cloud services from the cloud provider, “distributed” out to specific and varied physical locations. This enables an important characteristic of distributed cloud operation — low-latency compute where the compute operations for the cloud services are closer to those that need the capabilities. This can deliver major improvements in performance and reduce the risk of global network-related outages.

- Data sovereignty and other regulatory issues may require services be delivered from locations beyond the data centers of the public cloud service provider.

- Perceived and real security and privacy concerns with off-premises applications and infrastructure drive some consumers to prefer on-premises solutions.

- Latency needs of IoT/edge applications require services to be located close to the edge.

- Distributed cloud is still a single-cloud provider, and the managed cloud assets are still part of the cloud provider’s portfolio.

- Disconnected operations can be supported with distributed services that can operate independently.

Obstacles

- Customers can't abandon existing technologies in favor of complete and immediate migration to the public cloud, due to sunk costs, latency requirements, regulatory requirements, and the need for integration.
- Different approaches to distributed cloud have different value propositions (e.g., portability, software, appliance). Customers need to maintain visibility back to original goals.
- Distributed services are a relatively small subset of the centralized services, will take time to expand, and will likely never reach 100% parity with public cloud.
- Distributed cloud in your data center will have limits to scale and elasticity, which do not exist with the centralized public cloud. More advanced approaches like distributed cloud embedded in networking or telecom equipment — or delivered as metro area services — are very immature.

User Recommendations

- Overcome the fear of a single franchise controlling the public cloud and on-premises cloud estates, and consider targeted use of distributed cloud.
- Identify scenarios where distributed cloud use-case requirements can be met by evolution of a hybrid cloud model and where the requirements are substantially different. Prefer distributed cloud over building a hybrid cloud. Use the distributed cloud model to prepare for the next generation of cloud computing by targeting location-dependent use cases.
- View vendor claims of the scope of services available and their functional parity with public cloud services skeptically, and demand specific details and data to back up the claims.
- Temper concern about vendor revenue recognition and reporting. As with many capabilities that are thought of as more feature than product, revenue recognition and reporting by vendors are only one indicator of success.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Microsoft; Oracle

Gartner Recommended Reading

The Cloud Strategy Cookbook, 2023

Comparing On-Premises Public Cloud Appliances: AWS Outposts, Microsoft Azure Stack Hub and Google Distributed Cloud Edge

Distributed Cloud: Does the Hype Live Up to Reality?

Edge Computing

Analysis By: Bob Gill, Philip Dawson

Benefit Rating: Transformational

Market Penetration: More than 50% of target audience

Maturity: Early mainstream

Definition:

Edge computing describes a distributed computing topology in which data storage and processing are placed in optimal locations relative to the location of data creation and use. Edge computing locates data and workloads to optimize for latency, bandwidth, autonomy and regulatory/security considerations. Edge-computing locations extend along a continuum between the absolute edge, where physical sensors and digital systems converge, to the “core,” usually the cloud or a centralized data center.

Why This Is Important

Edge computing has quickly become the decentralized complement to the largely centralized implementation of hyperscale public cloud. Edge computing solves many pressing issues, such as sovereignty, unacceptable latency and bandwidth requirements, given the massive increase in data produced at the edge. The edge-computing topology enables the specifics of Internet of Things (IoT), digital business and managed distributed IT solutions.

Business Impact

Edge computing improves efficiency, cost control, and security and resilience through processing closer to where the data is generated or acted upon, fostering business opportunities and growth (e.g., customer experience and new real-time business interactions). Earliest implementations succeeded in enterprises that rely on operational technology (OT) systems and data outside core IT, such as the retail and industrial sectors.

Drivers

- Growth of hyperscale cloud adoption has exposed the limits of extreme centralization. Latency, bandwidth requirements, the need for autonomy and data sovereignty or location requirements may be optimized by placing workloads and data closer to the edge, rather than centralizing in a hyperscale data center.
- Data growth from interactive applications and systems at the edge often cannot be economically funneled into the cloud.
- Applications supporting customer engagement and analysis favor local processing for speed and autonomy.
- IoT is evolving from simply reporting device status to using edge-located intelligence to act upon such status, bringing the benefits of automation and the creation of immediately responsive closed loop systems.
- Edge computing's inherent decoupling of application front ends and back ends provides a perfect means of fostering innovation and enhanced ways to do business. For example, using technologies such as machine learning and industrial sensors to perform new tasks at locations where business and operational events take place, or at the point of interaction with a retail customer, can drive significant business value.

Obstacles

- The diversity of devices, software controls and application types all amplify complexity issues.

- Widespread edge topology and explicit application and networking architectures for edge computing are not yet common outside vertical applications, such as retail and manufacturing.

- Edge success in industrial IoT applications and enhancing customer experience in retail are well-understood, but many enterprises still have difficulty understanding the benefits, use cases and ROI of edge computing.

- A lack of broadly accepted standards slows development and deployment time, creating lock-in concern for many enterprise users.

- Edge physical infrastructure is mature, but distributed application management and orchestration challenges are still beyond most vendor-supplied component management offerings. The tasks of securing, maintaining and updating the physical infrastructure, software and data require improvement before management and orchestration can mature.

User Recommendations

IT leaders responsible for cloud and edge infrastructure should:

- Create and follow an enterprise edge strategy by focusing first on business benefit and holistic systems, not simply focusing on technical solutions or products.

- Position edge computing as an ongoing, enterprisewide journey toward distributed computing, not simply individual isolated projects.

- Establish a modular, extensible edge architecture through the use of emerging edge frameworks and design sets.

- Accelerate time to benefit and derisk technical decisions through the use of vertically aligned systems integrators and independent software vendors that can implement and manage the full orchestration stack from top to bottom.

- Evaluate “edge-as-a-service” deployment options, which deliver business-outcome-based solutions that adhere to specific SLAs while shifting deployment, complexity and obsolescence risk to the provider.

Gartner Recommended Reading

Market Guide for Edge Computing

5 Top Practices of Successful Edge Computing Implementers

Intelligent Infrastructure

Analysis By: Philip Dawson, Nathan Hill

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Emerging

Definition:

Intelligent infrastructure is built from simple, repeatable infrastructure building block components, integrated and managed in a standardized, automated manner. It optimizes infrastructure resources for application consumption through infrastructure machine learning (ML) and tuning as software overlays through an automated software intelligence plane.

Why This Is Important

Intelligent infrastructure encapsulates generative AI and ML into the infrastructure configuration. Building on the capabilities of virtualization, it adds the dynamic hardware composition capability of a composable infrastructure to deliver a hardware configuration that is optimized for a specific application. Intelligent infrastructure additionally adds or feeds the generative AI/ML automation functions to the intelligence plane.

Business Impact

Intelligent infrastructure is an innovation in delivering automated optimized systems for application delivery. It builds on earlier innovations, including converged, hyperconverged, software-defined and composable infrastructures, helping deliver hybrid cloud-like infrastructure on-premises or with a provider. It feeds off the application API-led programmable infrastructure that tunes infrastructure through system calls and requests, which improves application and infrastructure integration.

Drivers

- IT leaders now recognize that cloud infrastructure, cloud platforms and cloud-native applications drive the overall composable, programmable and intelligent infrastructure journey.
- Cloud delivery and edge expansion are fueling the standardization of infrastructure, design and architecture and the expansion of the three areas to edge and Internet of Things (IoT) locations beyond remote offices/branch offices (ROBOs).
- Adding generative AI and automation on top of this infrastructure composition capability ensures that infrastructure is always optimized for the application load.
- In intelligent infrastructure, the "control plane" is enhanced with automation driven by infrastructure analytics ML, to become an automated "intelligence plane."

Obstacles

- The intelligence plane automates infrastructure and workload provisioning to application consumption. Intelligent infrastructure should not be tied to hardware features, but rather software functions.
- As with software-defined and composable infrastructures, traditional system vendors often tie intelligent infrastructure to hardware-related features, which can propel lock-in.
- Cloud management platforms are used as overlays for cloud migrations. Intelligent infrastructure has to adapt to hybrid cloud and multicloud delivery, delivering client value whether on-premises, with a provider or public cloud through anything as a service (XaaS).

User Recommendations

- Select infrastructure solutions based on their ability to meet the current business requirements while still offering the flexibility to exploit the integration and automation of intelligent infrastructure innovations to be delivered over the next five years.
- Increase agility and business alignment by integrating application, asset management and sourcing information into the infrastructure intelligence and control planes as a drive to platform- and infrastructure-driven consumption models.
- Prepare for the evolution of application delivery and workload provisioning by incorporating intelligence/ML infrastructure functions with intelligent fabrics into your future system requirements.

Sample Vendors

Cisco; CU Coding; Hewlett Packard Enterprise; IBM; Intel; Microsoft; Tintri; VMware

Gartner Recommended Reading

How to Evolve Your Physical Data Center to a Modern Operating Model

Market Guide for Servers

Quick Answer: How Can I Optimize the Use of Programmable Platforms for Effective Software Delivery?

Sliding into the Trough

Site Reliability Engineering

Analysis By: George Spafford, Daniel Betts

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Site reliability engineering (SRE) is a collection of systems and software engineering principles used to design and operate scalable resilient systems. Site reliability engineers work with the customer or product owner to understand operational requirements and define service-level objectives (SLOs). Site reliability engineers work with product or platform teams to design and continuously improve systems that meet defined SLOs.

Why This Is Important

SRE emphasizes the engineering disciplines that lead to resilience; but individual organizations implement SRE in widely varying ways such as a defined role or a set of practices. SRE teams can serve as an operations function, and nearly all such teams have a strong emphasis on blameless root cause analysis. This is to decrease the probability and/or impact of future events and to enable organizational learning, continual improvement and reductions in unplanned work.

Business Impact

The SRE approach to improving reliability and resilience is intended for products and platforms that need to deliver customer value at speed at scale while managing risk. The two primary use cases are to improve the reliability of existing products/platforms or to create new products or platforms that need reliability from the start.

Drivers

- Clients are under pressure to meet customer requirements for reliability while scaling their digital services and are looking for guidance to help them.

- While Google originated what became known as SRE and continued to evolve it, practitioners are developing and sharing new practices as well. Potential practitioners looking for pragmatic guidance to improve the reliability of their systems have a rich body of knowledge they can leverage that works well with agile and DevOps.

- Organizations are adopting highly skilled automation practices (usually DevOps), and usage of infrastructure-as-code capabilities (which usually requires a cloud platform) to deliver digital business products reliably.

- The most common use case based on inquiry calls with clients is to leverage SRE concepts to improve the reliability of existing systems that are not meeting customer requirements for availability, performance or are proving difficult to scale.

Obstacles

- Insufficient internal marketing to understand what agile, DevOps or product teams need or would value and then explaining how the value SRE can deliver will justify the costs and risks incurred. Without marketing its benefits, SRE adoption tends to be less certain or slower. The SRE concept by itself is insufficient — people must continuously believe it is worthwhile.

- Finding SRE candidates who have the right mix of development, operations and people skills is a big challenge for clients. Impacts on initial adoption and scaling efforts as well.

- Rebranding of a traditional operations team without changing to adopt SRE practices, only SRE in name.

- Clients have voiced problems with product owners who overly focus on functional requirements and not nonfunctional requirements thus slowing improvements and support of SRE within the organization.

User Recommendations

- Leverage practices pragmatically based on need. Don’t feel that you must implement SRE exactly the way Google does it, learn what works for you.

- Detect an opportunity to begin that is politically friendly, will demonstrate sufficient value and has an acceptable risk profile.

- Start small, focus, learn, improve, and demonstrate value — do not try to change everything at once.

- Work with the customer or product owner to define clear, obtainable SLOs based on their needs.
- Implement monitoring and improve observability to objectively report on actual performance relative to the SLOs.
- Product owners must be accountable for functional and non-functional requirements of their products.
- Instill collaborative working between site reliability engineers, developers and other stakeholders to help them learn how to design, build and evolve their products to meet SLOs.
- Create a community, implement effective organizational learning practices and evolve SRE practices.

Sample Vendors

Atlassian; Blameless; Datadog; Dynatrace; New Relic; OpsRamp; PagerDuty; Splunk

Programmable Infrastructure

Analysis By: Philip Dawson, Nathan Hill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Programmable infrastructure is the concept of using and applying methods and tooling from the software development area to management of IT infrastructure. This includes, but is not limited to, APIs, immutability, resilient architectures and agile techniques.

Why This Is Important

Programmable infrastructure ensures optimal resource utilization, while driving cost efficiencies. A continuous delivery approach requires continuous insight and the ability to automate application responses. Moving to an API-driven infrastructure is the key first necessary step to enabling anti-fragile and sustainable automation through programmatic techniques.

Business Impact

Greater value (rather than cost reduction) is achieved via programmable infrastructure's ability to drive adaptive automation — responding faster to new business infrastructure demands, driving service quality and freeing staff from manual operations. Programmable infrastructure reduces technical debt with investment and enables a sustainable and highly responsive IT infrastructure service to the business.

Drivers

- Programmable infrastructure strategies are applied to private cloud, hybrid cloud and infrastructure platforms as well as public cloud. Demand for programmable infrastructure grows as heterogeneous infrastructure strategies are embraced.
- Programmable infrastructure is needed to manage the life cycle of infrastructure delivery from provisioning, resizing and reallocation to reclamation, and in the case of external resources, manage elasticity and the termination of consumption.
- Programmable infrastructure is needed to optimize and reduce the dependency on the infrastructure life cycle. More importantly, it enables the desired (performance, cost, speed) infrastructure provisioning and orchestration in line with business demands.

Obstacles

- The ongoing cost of refreshing API-enabled infrastructure components on-premises after initial implementation adds financial pressure to organizations.
- Applying automation to existing monolithic infrastructure components fails due to the lack of platform agility and vendor lock-in.
- While APIs enable integration across different infrastructure platforms, the lack of open APIs/API compatibility across vendor platforms creates a siloed mentality.
- The implementation of programmable infrastructure is hampered by the early adoption of it within infrastructure and operations (I&O), and the shortage of skilled software engineering resources to comprehensively exploit it (especially in web technologies such as HTTP and JSON to develop these APIs).

User Recommendations

- Deploy a programmable infrastructure to further abstract application from infrastructure delivery and pursue an agile digital business outcome.
- Implement a programmable infrastructure by investing in infrastructure automation tools and continuous delivery (example vendors for these markets are listed below, but no single vendor or platform can enable an organizationwide programmable infrastructure strategy) leading to API-led programmable platforms.
- Invest in infrastructure and DevOps, and modernize legacy IT architectures to implement an API-driven infrastructure.
- Examine reusable programmable infrastructure building blocks leveraging programmable infrastructure strategy built on repeatable and available skills from providers.

Sample Vendors

Amazon Web Services; CU Coding; Google; IBM; Microsoft; Oracle; Quality Technology Services; RackN; Tencent; VMware

Gartner Recommended Reading

Market Guide for Servers

Predicts 2023: XaaS Is Transforming Data Center Infrastructure

Quick Answer: How Can I Optimize the Use of Programmable Platforms for Effective Software Delivery?

Serverless Infrastructure

Analysis By: Jeffrey Hewitt

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Serverless infrastructure is a model of IT service delivery in which the underlying enabling resources are used as an opaque, unlimited, shared pool that is continuously available without advance provisioning and priced in units of the consumed IT service. The runtime environment (that is, the compute, storage, networking and language execution environment) required to execute an application or service is automatically provisioned and operated.

Why This Is Important

Accelerating the development and delivery of software is a core imperative for IT leaders. Not only do serverless technologies enable organizations to build and deliver software faster, but they also offer low operational overheads, resource scaling as needed and an elastic pricing model. Cloud providers and open-source software vendors are all innovating and making serverless products available for a broad set of use cases.

Business Impact

Serverless technologies enable organizations to build cloud-native applications with newer application architectures — such as microservices, which can usher in higher degrees of resiliency, elasticity and agility for digital workloads. Serverless technologies also enable consumption of platform services by developers and business users, with the infrastructure provisioning and life cycle management abstracted away from the consumer.

Drivers

- **Evolution:** In the past few years, the term "serverless infrastructure" has evolved to include much more than function as a service (FaaS) products. Currently, it refers not only to a programming model such as FaaS, but also to an operational model where all provisioning, scaling, monitoring and configuration of compute infrastructure are delegated to the platform. Examples of such architectures include serverless containers and serverless databases.

- **Operational simplicity:** Serverless infrastructure obviates the need for IT departments to perform infrastructure setup, configuration, provisioning and management.

- **"Built-in" scalability:** Infrastructure scaling is automated and elastic.

- **Cost-efficiency:** You only pay for infrastructure resources when they are needed to support requested transactions.

- **Developer productivity and business agility:** Serverless infrastructure abstracts infrastructure architecture and allows developers to focus on writing code and designing applications.

Obstacles

- **Vendor lock-in:** As with most cloud functionality, the leading serverless implementations are proprietary to specific cloud providers. If an application has to move from one cloud platform to another, it will have to be significantly reengineered.
- **Low degree of control:** The managed service model and the runtime virtualization of serverless technologies bestow huge benefits, but at the cost of little or no control over the service. The environment is a "black box" that must be used as it is.
- **Skills gap:** Serverless operations require a major shift in skills and best practices, with much more code and API-oriented service delivery.

User Recommendations

- Ensure cost governance and budget control by evaluating the cost implications of event-driven application architectures and the pricing models of different vendors. Be aware of API gateway, network egress and other costs.
- Revise data classification policies and controls to account for the fact that objects in a content store can now represent code as well as data.
- Rethink IT operations from infrastructure management to application governance, with an emphasis on ensuring that security, monitoring, debugging requirements and application SLAs are being met. In cases where on-premises deployment is merited, IT teams can support FaaS in the role of service provider.

Sample Vendors

Alibaba; Amazon Web Services; Apache Software Foundation; DigitalOcean; Google; IBM; Knative; Microsoft; OpenFaaS; Oracle

Gartner Recommended Reading

A CTO's Guide to Serverless Computing

When to Use Serverless Computing to Optimize Cloud Costs?

Balancing Opex and Capex

Analysis By: Kevin Ji, Michael Warrilow

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Capital expenditure (capex) is the traditional model of acquiring IT assets to support business operations. The rapid adoption of cloud is leading to an acceleration in the rise of operational expenditure (opex) spend. Cloud is not a cost-saving solution unless infrastructure funding is optimized. An appropriate balance of capex and opex is required based on suitable criteria to avoid the risks and demonstrate the value of investment.

Why This Is Important

Enterprise infrastructure and applications are increasingly cloud-based, escalating the shift away from more traditional capital-based models of IT expenditure. Furthermore, the existing legacy infrastructure needs to be refreshed periodically as planned. Increasing cloud spend is shifting the balance between capex and opex, causing budgetary and cost management pressures in many organizations, which may be exacerbated further due to the impact of economic headwinds.

Business Impact

Changes to funding models affect IT operations, projects and budgets, such as by increasing the risk of greater variability of spending. IT finance teams are being forced to update their policies, processes and procedures, and to anticipate and plan for changes to funding models and business justification. Meanwhile, IT procurement teams responsible for contract management are observing increases in pay-as-you-go (PAYG) and pay-as-you-use (PAYU) contracts.

Drivers

- Cloud spending can drive opex up by enabling new business initiatives to be explored more easily under PAYG or PAYU models, such as through minimum viable products and campaign-based workloads. The shift to cloud exemplifies this acceleration, as spending on public cloud services is increasing in terms of its share of IT spending and forecast models.

- Under a public cloud service model, the customer agrees to pay a service fee in exchange for the rights to use the cloud provider's services over a specified term or volume. In almost all cases, cloud expenditure is treated as opex under applicable accounting standards.

- Traditional capex-based purchases of on-premises hardware are in transition. To compete with public cloud, many hardware providers are encouraging the adoption of consumption-based hardware. Increased adoption of distributed cloud would add to this trend.

- On-premises infrastructure software providers are offering attractive initial transition plans to encourage clients to switch to subscription-based models, adding to the scale of the challenge.

Obstacles

- The large-scale migration of workloads to the cloud has seen an increase in IT spending beyond the original budget. This is particularly true right after the initial migration, when proper governance and workload optimization have yet to be implemented.

- Those responsible for IT budgeting have tended to participate in justifying change in IT investments. However, many organizations take a siloed approach to cloud cost management, when responsibility should be dispersed across the organization.

- Different application architectures through cloud migration may drive different cloud resource consumption with spending, such as application lift-and-shift model compared with rearchitecting through PaaS services. This creates complexity in communication between teams to align with the right architecture design on cost impact.

- Many organizations only project cloud costs three years ahead. However, a five-year projection is better at predicting ROI, as it often captures the on-premises hardware refresh cycle.

User Recommendations

- Establish a three- to five-year transition plan to demonstrate the budgeting portfolio switch from capex to opex across both cloud and on-premises environments.
- Compare different cloud workload architectures to find a suitable migration path that optimizes long-term operating costs.
- Identify capex that will be impacted by the conversion to opex spend as asset life cycles prompt further scrutiny into the benefits of upfront purchasing.
- Track and accurately account for continuing increases in cloud expenditures by working with architecture and finance teams, as well as application owners, anticipating greater scrutiny from CFOs and auditors into increased, variable opex.
- Avoid justifying the change in balance with cost savings. Focus primarily on the business value of improved flexibility and agility, access to innovation and a reduced need to maintain and support hardware and software.

Gartner Recommended Reading

How Cloud Adoption Will Increase Opex Budgets

Market Impact: Cloud Shift — 2022 Through 2025

Realize Cost Savings After Migrating to the Cloud

How Executive Leaders Should Justify Moving to the Cloud

Composable Infrastructure

Analysis By: Tony Harvey, Paul Delory, Philip Dawson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Composable infrastructure uses an API to create physical systems from shared pools of resources. The implementation connects disaggregated banks of processors, memory, storage devices and other resources by a hardware fabric. However, composable infrastructure software can also aggregate or subdivide resources in traditional servers or storage.

Why This Is Important

Servers, storage and fabrics are traditionally deployed as discrete products with predefined capacities. Individual devices, or resources, are connected manually and dedicated to specific applications, making the system inflexible and expensive to change and scale. Composable infrastructure replaces this with a pool of components that can be dynamically assigned as needed, increasing agility, easing capacity planning and reducing costs.

Business Impact

Stranded hardware resources that are underutilized represent significant costs in IT. The composable infrastructure enables hardware resources to be aggregated from a pool of components via APIs to dynamically match the infrastructure to the needs of the workload. This increases component utilization, reduces hardware overprovisioning, decreases costs, and improves IT responsiveness to the business's requirements.

Drivers

- Compute Express Link (CXL) provides the necessary capabilities to disaggregate and pool memory and I/O as well as providing a standardized set of APIs to manage the disaggregated hardware.

- Hyperscale cloud vendors are moving toward composable designs utilizing CXL to increase hardware utilization and reduce the costs of stranded hardware.

- Test and development environments benefit from composability, where infrastructure with varying characteristics must be repeatedly deployed, deconstructed and redeployed.

- Multitenant environments benefit from composable infrastructure by allowing a pool of hardware to be dynamically configured, assigned, reconfigured and reassigned based on tenant requirements.

Obstacles

- Current composable implementations are limited in that pooled resources are restricted to using hardware from a single vendor.
- Existing composable infrastructures are limited to just composing storage and I/O, limiting the use cases.
- A proliferation of vendor-specific APIs and a lack of off-the-shelf software for managing composable systems are also headwinds to widespread adoption.

User Recommendations

- Deploy composable infrastructure when the workload or use case demands that infrastructure must be resized and administered frequently or when composability increases the use of packaged standardized high-cost components.
- Replace existing infrastructure to obtain composable infrastructure only if you have sufficiently mature automation tools and skills to implement composable features and yield financial or business benefits.
- Verify that your infrastructure management software supports composable system APIs or that you have the resources and skill sets to write your own management tools.

Sample Vendors

Cisco; Dell Technologies; GigaIO; Hewlett Packard Enterprise; Intel; Liqid; Western Digital

Gartner Recommended Reading

Market Guide for Servers

Emerging Tech: Compute Express Link Redefines Server Memory Architectures

Emerging Tech Impact Radar: Compute and Storage

2022 Strategic Roadmap for Compute Infrastructure

Service Mesh

Analysis By: Anne Thomas

Benefit Rating: Low

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

A service mesh is a distributed computing middleware that manages communications between application services — typically within managed container systems. It provides lightweight mediation for service-to-service communications and supports functions such as authentication, authorization, encryption, service discovery, request routing, load balancing, self-healing recovery and service instrumentation.

Why This Is Important

A service mesh is lightweight middleware for managing and monitoring service-to-service (east-west) communications — especially among microservices running in container management systems, such as Kubernetes. It provides visibility into service interactions, enabling proactive operations and faster diagnostics. It automates complex communication concerns, thereby improving developer productivity and ensuring that certain standards and policies are enforced consistently across applications.

Business Impact

Service mesh is one of many management technologies that provide software infrastructure for distributed applications. Service meshes are most often used with services deployed in container management systems, such as Kubernetes. This type of middleware, along with other management and security middleware, helps provide a stable environment that supports "Day 2" operations of containerized workloads. However, the technology is complex and often unnecessary for smaller deployments.

Drivers

- Microservices and containers: Service mesh adoption is closely aligned with microservices architectures and container management systems like Kubernetes. Service mesh supports useful functionality in ephemeral environments, such as dynamic service discovery and mutual Transport Layer Security (mTLS) between services.

- Observability: As microservice deployments scale and grow more complex, DevOps teams need better ways to track operations, anticipate problems and trace errors. Service mesh automatically instruments the services and feeds logs to visualization dashboards.

- Resilience: A service mesh implements the various communication stability patterns (including retries, circuit breakers and bulkheads) that enable applications to be more self-healing.

- Bundled feature: Many container management systems now include a service mesh, inspiring DevOps teams to use it. The hyperscale cloud vendors provide a service mesh that is also integrated with their other cloud-native services.

- Federation: Independent vendors such as Buoyant, greymatter.io, HashiCorp, Kong and Solo provide service meshes that support multiple environments.

Obstacles

- Not necessary: Service mesh technology can be useful when deploying microservices in Kubernetes, but it's never required.

- Complexity: It's complex to use and administer, and there are increasing discussions on why not to use a service mesh in technology discussion groups and social media.

- Redundant functionality: Users are confused by the overlap in functionality among service meshes, ingress controllers, API gateways and other API proxies. Management and interoperability among these technologies is still nascent within the vendor community.

- Overhead: Service mesh technology consumes resources and typically adds overhead to the interactions it manages. Some vendors now support alternate architectures, such as a shared-agent model to reduce overhead, but this solution reduces the observability benefits.

- Competition with "free": Independent service mesh solutions face challenges from the availability of platform-integrated service meshes from the major cloud and platform providers.

User Recommendations

- Determine whether the value you might get from a service mesh in terms of improved security or observability is worth the increase in complexity and administration of the service mesh. A service mesh becomes more valuable as the number of service-to-service (east-west) interactions increases.

- Favor the service meshes that come integrated with your container management system unless you have a requirement to support a federated model.

- Reduce cross-team friction by assigning service mesh ownership to a cross-functional platform engineering team that solicits input and collaborates with networking, security and development teams.

- Accelerate knowledge transfer and consistent application of security policies by collaborating with I&O and security teams that manage existing API gateways and application delivery controllers.

Sample Vendors

Amazon Web Services; Ambient Mesh; Buoyant; Google; HashiCorp; Istio; Kong; Microsoft; Solo.io

How a Service Mesh Fits Into Your API Mediation Strategy

Climbing the Slope

Immutable Infrastructure

Analysis By: Neil MacDonald, Tony Harvey

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Immutable infrastructure is a process pattern (not a technology) in which the system and application infrastructure, once deployed, are never updated in place. Instead, when changes are required, the infrastructure and applications are simply updated and redeployed through the CI/CD pipeline.

Why This Is Important

Immutable infrastructure ensures the system and application environment, once deployed, remains in a predictable, known-good-configuration state. It simplifies change management, supports faster and safer upgrades, reduces operational errors, improves security, and simplifies troubleshooting. It also enables rapid replication of environments for disaster recovery, geographic redundancy or testing. This approach is easier to adopt with cloud-native applications.

Business Impact

Taking an immutable approach to workload and application management simplifies automated problem resolution by reducing the options for corrective action to, essentially, just one — repair the application or image in the development pipeline and rerelease. The result is an improved security posture and a reduced attack surface with fewer vulnerabilities and a faster time to remediate when new issues are identified.

Drivers

- Linux containers and Kubernetes are being widely adopted. Containers improve the practicality of implementing immutable infrastructure due to their lightweight nature, which supports rapid deployment and replacement.

- The GitOps deployment pattern, which emphasizes continuously synchronizing the running state to the software repository, has become an effective way to implement immutable infrastructure in Kubernetes-based, containerized environments.
- Infrastructure as code (IaC) tools (including first-party cloud provider IaC tools) have increasingly integrated configuration drift detection and correction, improving the practicality of implementing immutable infrastructure across an application's entire stack and environment.
- Interest in zero-trust and other advanced security postures where immutable infrastructure can be used to proactively regenerate workloads in production from a known good state (assuming compromise), a concept referred to as "systematic workload reprovisioning."
- For cloud-native application development projects, immutable infrastructure simplifies change management, supports faster and safer upgrades, reduces operational errors, improves security, and simplifies troubleshooting.

Obstacles

- The use of immutable infrastructure requires a strict operational discipline that many organizations haven't yet achieved, or have achieved for only a subset of applications.
- IT administrators are reluctant to give up the ability to directly modify or patch runtime systems.
- Applying the immutable infrastructure pattern is most easily done for stateless components. Stateful components, especially data stores, represent special cases that must be handled with care.
- Implementing immutable infrastructure requires a mature automation framework, up-to-date blueprints and bills of materials, and confidence in your ability to arbitrarily recreate components without negative effects on user experience or loss of state.
- Many enterprise applications are stateful applications deployed on virtual machines. These applications are oftentimes commercial off-the-shelf and are not designed for fully automated installation when redeployed.

User Recommendations

- Reduce or eliminate configuration drift by establishing a policy that no software, including the OS, is ever patched in production. Updates must be made to individual components, versioned in a source-code-control repository, then redeployed.

- Prevent unauthorized change by turning off all administrative access to production compute resources. Examples of this might include not permitting Secure Shell or Remote Desktop Protocol access.

- Adopt immutable infrastructure principles with cloud-native applications first. Cloud-native workloads are more suitable than traditional on-premises workloads.

- Treat scripts, recipes and other codes used for infrastructure automation similar to the application source code itself, as this mandates good software engineering discipline.

- Include immutable infrastructure scripts, recipes, codes and images in your backup and ransomware recovery plans as they will be your primary source to rebuild your infrastructure after an infection.

Sample Vendors

Amazon Web Services; Google; HashiCorp; Microsoft; Perforce; Progress; Red Hat; Snyk; Turbot; VMware

Gartner Recommended Reading

Comparing DevOps Architecture to Automate Infrastructure and Operations for Software Development

2022 Strategic Roadmap for Compute Infrastructure

To Automate Your Automation, Apply Agile and DevOps Practices to Infrastructure and Operations

Innovation Insight for Continuous Infrastructure Automation

Market Guide for Cloud-Native Application Protection Platforms

Software-Defined Infrastructure

Analysis By: Philip Dawson

Benefit Rating: Low

Market Penetration: 20% to 50% of target audience

Maturity: Obsolete

Definition:

Software-defined infrastructure (SDI) enables abstraction of the physical infrastructure, with its services exposed via APIs enabling greater levels of automation, policy-based orchestration and reuse. SDI includes software-defined data center, network, storage, compute and SD edge infrastructure.

Why This Is Important

Software-defined is the further abstraction of software from hardware. It enables businesses to be more agile and flexible by enabling programmatic control of the infrastructure through software interfaces. SDI combines compute (SDC), network (SDN) and storage (SDS), but SDI also extends to non-data-center infrastructure, with the use of either software-defined monitoring devices or machines.

Business Impact

While data center SDI is embedded in other data center initiatives, such as cloud and hyperconverged infrastructure, SDI is now focused on key verticals operating in multiple edge locations, such as retail, manufacturing, retail banking, distribution and utilities. It also continues to extend Internet of Things (IoT), non-data-center SDI and SDS storage initiatives for new IT and software-defined WAN (SD-WAN) operations and functions.

Drivers

- SDI data center infrastructure is well-covered with compute (SDC), network (SDN, now obsolete), edge (SD-WAN), and storage (SDS), but SDI also extends to non-data-center infrastructure with the use of monitoring devices or machines that are software-defined.
- SDI reaches beyond and between software-defined data centers (SDDCs), leveraging SDI benefits and features for new multimode applications and edge and/or IoT endpoints.
- In 2023, SDI's continued presence of hype is enabled through the use of sensors and adapters that are abstracted through software, stretching SDI to the edge, IoT and operational technology (such as retail point of sale [POS]), rather than traditional, IT-driven SDI through a data center or cloud.
- Key verticals operating in multiple, geographically distributed locations, such as retail, manufacturing, retail banking, distribution and utilities, are extending IoT and non-data-center SDI initiatives for new edge and IoT operations and functions.

Obstacles

- SDI is now tied to extending data center vendor technology, not interoperability.
- SDI overlaps other integrated systems taxonomy, like hyperconvergence, as it drives cloud to data center and edge adoption.
- SDI continues releasing vendor-specific silo technology (not heterogeneous and service-driven) and, hence, it continues to be obsolete as multivendor interoperability standards and technology silos persist, limiting SDI integration between vendors.
- SD-WAN segmentation is driving SDI to the edge and is architecturally different from SDN, which is focused more on data center infrastructure convergence.

User Recommendations

- Include the integration and measurement of non-data-center edge infrastructure, as SDI initiatives roll out tied to SD-WAN and edge initiatives.
- Focus on core IT SDI for compute, network, storage and facilities, but expand the impact of SDI on IoT, edge computing, remote office/branch office (ROBO) and other operational technologies.
- Anticipate SDI to be tied to a specific vendor or technology silo, such as SDS storage and SD-WAN network hardware or virtualization software. Be cautious not to commit to a vendor's SDI without realizing the specific area of lock-in.

Sample Vendors

IBM (Red Hat); Intel; Microsoft; VMware; Wipro Enterprises

Gartner Recommended Reading

Predicts 2023: XaaS Is Transforming Data Center Infrastructure

How Do I Plan for Migrating My Data Center Infrastructure Into an XaaS Model?

IoT

Analysis By: Alfonso Velosa, Scot Kim, Emil Berthelsen

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

The Internet of Things (IoT) is an enabler and accelerator for digital transformation of enterprises. IoT enables enterprises to improve business processes and enhance decisions with asset information, via embedded technology, to communicate and sense/interact with their internal states or external environment. IoT solutions span assets, IT and OT systems, communications, applications, data and analytics, and AI. Enterprises use IoT-enabled solutions to develop new revenue and operating models.

Why This Is Important

IoT is an enabler and accelerator for composable business initiatives. Most enterprises lack granular data or information about their assets' or products' state, adding cost and inefficiency to their processes. For operators of assets such as airports or utilities, adding IoT capabilities provides new understanding of the asset, enabling them to optimize operations. For OEMs, IoT insights into how products are used help improve new designs and to enable product-as-a-service models.

Business Impact

IoT will impact enterprise business operations, customer engagement, competitive position and product strategies by enabling:

- Process optimization: This covers the spectrum from costs to operations, while improving asset health and conserving resources.
- New revenue strategies: This includes generating revenue via improved products, services and data monetization.
- Safety and compliance focus: This includes documenting alignment to regulations, preventing dangerous breakdowns and employee injury.

Drivers

- Enterprises globally are increasingly implementing business objectives that use IoT solutions at large scale, internally and across ecosystems.

- Early majority enterprises, from hospitals to manufacturers to building management companies, are using IoT-enabled solutions to engage customers and optimize operations, reflecting business requirements across the spectrum of low-to-high asset-intensive industries.

- OEMs increasingly add IoT-enabled capabilities to improve value, meet competitive pressure, drive differentiation and add new revenue streams.

- Leading-edge enterprises are using IoT to drive transformative strategies. For example, product as a service or SLA-compliance asset uptime or new digital twin or data management value propositions.

- Shorter payback time frames (six to 18 months on average) make IoT-enabled business projects attractive.

- Technology and service providers have realigned their go-to-market strategy to highlight value to enterprise customers. This includes IoT enabled applications, AI solutions optimized for IoT, and engaging partners to provide training and culture change.

Obstacles

- Many enterprise leaders underestimate the political capital required to support IoT projects, since these are really business transformation projects that require engagement with business and frontline OT workers, to change culture and processes.
- There is a lack of a cross-functional enterprise center of excellence to focus on developing best practices for IoT-enabled business projects and sharing them, driving IT-OT alignment and allocating budgets, personnel and resources.
- The lack of standards inhibits the ability of enterprises to deploy large-scale IoT solutions that involve multiple vendors, from sensors, gateways and communications, to implementation, integration and analysis. This adds cost increases and schedule delays to IoT-enabled projects and programs.
- The cluttered market of IoT hardware and software vendors often lacks the vertical domain understanding to quantify business solutions to the core business, operations and IT stakeholders.

User Recommendations

- Forge IoT centers of excellence across business units and cross-functional business, engineering and IT stakeholders for business transformation objectives. Invest time and effort on culture change, such as incentives to foster cross-organizational collaboration around desired IoT-enabled business outcomes and IT/OT/ET alignment. Use it to drive enterprisewide best practices and objectives.
- Ensure teams focus on IT and operational architectures to address key technology complexity, security and integration challenges, and start building roadmaps for long-term composability and multiyear deployments.
- Plan to implement a multivendor approach for IoT platforms, analytics and applications when implementing multiple use cases across different business units or countries.
- Establish accountability, participation, predictability and transparency policies for IoT to address sponsorship, budgets, digital ethics, data ownership and rights to monetize IoT data.

Sample Vendors

AT&T; Endowance Solutions; Falkonry; GE Digital; Hexagon; Litmus; Oracle; Samsara; Siemens; Wiliot

Gartner Recommended Reading

Magic Quadrant for Global Industrial IoT Platforms

Magic Quadrant for Indoor Location Services

Infographic: IoT Use-Case Prism for Sustainability and ESG

Toolkit: 5 Digital Twin/IoT Project Success Drivers

Important and Compelling Innovations for Commercial IoT Use Cases

OS Containers

Analysis By: Thomas Bittman, Philip Dawson

Benefit Rating: Transformational

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

OS containers are a shared OS virtualization technology that enables multiple applications to share an OS kernel without conflicting. A “container daemon” provides logical isolation of processes. This enables several applications to share an OS kernel while maintaining their own copies of specific OS libraries.

Why This Is Important

Containers were previously used to increase the density of lightly used workloads, for improved infrastructure management. Now containers are focused on developer requirements for agile development, rapid provisioning and real-time horizontal scaling, especially for microservices architecture applications and cloud-native computing.

Business Impact

Container technologies are part of a development architecture that helps enterprises become more agile, with applications that can change quickly, and scale rapidly to demand. In production, containers will often be used for new applications designed for agile development. However, for developer ease of use, containers will also be used as wrappers for traditional, monolithic workloads.

Drivers

- Lightweight overhead for small applications (improving capacity utilization and density)
- Portability — containers package up the code and its dependencies making it easier to migrate workloads reliably and predictably
- Ease of use and reuse by application developers
- Alignment with microservices architecture and agile development

Obstacles

- Reliance on the OS for application isolation can create security concerns, especially in multitenant environments.
- Containers are not direct replacements for hypervisors and, unlike with hypervisors, existing applications require redesign to take full advantage of the benefits of containers.
- Container use is constrained by the immaturity and complexity of tools and operations, especially in security, monitoring, data management and networking.
- Developing the right operational model for Kubernetes deployments is difficult, and requires organizational evolution and new skills.

User Recommendations

Infrastructure and operations leaders responsible for data center infrastructure should:

- Use containers when security and manageability concerns are easily mitigated.
- Combine containers with virtual machines (VMs) to separate developer concerns from capacity management, and when the performance overhead of VMs is an acceptable trade-off.

Sample Vendors

Canonical; Docker; Microsoft; Mirantis; Oracle; Red Hat; Virtuozzo; VMware

Gartner Recommended Reading

Market Guide for Container Management

Prioritizing Security Controls for Enterprise Servers and End-User Endpoints

Designing and Operating DevOps Workflows to Deploy Containerized Applications With Kubernetes

SD-WAN

Analysis By: Jonathan Forest

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Early mainstream

Definition:

Software-defined wide-area network (SD-WAN) products are primarily used to connect branch offices. They provide dynamic path selection, based on business or application policy, routing, centralized orchestration of policy and management of appliances, and virtual private network (VPN) and zero-touch configuration. SD-WAN products are WAN transport/carrier-agnostic and create secure paths across physical WAN connections.

Why This Is Important

SD-WAN products optimize site availability, performance, cost and agility for enterprise WANs, and are aligned with the broader shift of applications to public cloud workloads. Client demand remains high, and we estimate that approximately 120,000 customers have deployed SD-WAN products in their production networks.

Business Impact

SD-WAN products create simpler, more cost-effective branch-office WANs that support modern application and cloud architectures. They are much easier to deploy, offer more agility and deliver better performance than traditional, router-based or next-generation firewall (NGFW) solutions. Other benefits include easier operational management at the WAN edge, simpler connectivity to the cloud, better application-specific performance and greater branch availability than traditional routers.

Drivers

- Digitalization and cloud adoption are driving more applications from private data centers to the public internet, including public clouds (SaaS and IaaS). The desire to have direct connectivity to applications (without backhauling to a data center) in order to minimize latency continues to drive SD-WAN adoption.
- The renewal of first-generation SD-WAN service contracts provides an opportunity for organizations to reevaluate vendor offerings.
- With path selection functionality and manageability, SD-WAN technology can deliver better performance and availability of applications than static policy-based routing.
- Organizations are increasingly investing in secure access service edge (SASE) offerings, which drives SD-WAN deployments.
- The expiration of Multiprotocol Label Switching (MPLS) contracts makes organizations rethink their WAN architectures.
- Organizations worry that they may lag behind in the adoption of SD-WAN technology.

Obstacles

- Some vendors are promoting the myth that enterprise WANs and SD-WANs are no longer required. Some vendors claim that internet access and cloud-delivered security are all that is needed.
- The fact that some employees of organizations are working remotely reduces the demand for SD-WAN solutions.
- A lack of cloud adoption in certain industries and geographies reduces the benefits of SD-WAN solutions.
- Enterprises that backhaul traffic within a metropolitan area do not benefit from a local internet breakout to reduce latency when connecting to the cloud.

User Recommendations

- Involve security teams and cloud teams in the vendor selection process.
- Compare on-premises security with cloud-delivered security to determine the best security architecture.
- Confirm SD-WAN cloud onramp capabilities by involving cloud teams and validating orchestration with cloud service providers such as Microsoft (Azure) and Amazon Web Services (AWS), and carrier-neutral facilities such as Equinix. Confirm virtual image compatibility with various cloud platforms, for cloud-first organizations.
- Shortlist and execute a proof of concept based on desired functionality by focusing on ease of use, application performance, cloud connectivity and security functionality use cases, as opposed to perceived leadership in the market or incumbency.

Sample Vendors

Cisco; Fortinet; Hewlett Packard Enterprise (HPE); Huawei; Palo Alto Networks; Versa Networks; VMware

Gartner Recommended Reading

Magic Quadrant for Network Services, Global

Critical Capabilities for Managed Network Services

Magic Quadrant for SD-WAN

Critical Capabilities for SD-WAN

Magic Quadrant for Managed Network Services

Entering the Plateau

Cloud Computing

Analysis By: David Smith

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Cloud computing is a style of computing in which scalable, elastic, IT-enabled capabilities are delivered as a service using internet technologies.

Why This Is Important

Elements of cloud computing evolved in the late 1990s, and the concept gained significant attention around 2006. Since then, cloud computing has evolved and become a mainstream force that influences all aspects of IT. It is becoming the default computing model. Although it is now climbing the Slope of Enlightenment, it continues to be a visible, hyped and evolving trend that gives rise to new emerging trends and technologies. The hype has moved to various subelements such as hybrid cloud, sovereign cloud and cloud-native.

Business Impact

The primary potential benefits of cloud computing are cost savings, efficiency, agility, speed and innovation. To gain the greatest value, organizations should formulate cloud strategies that align these benefits with business needs. Cloud computing is changing the way the IT industry looks at user-vendor relationships. Vendors must become or partner with service providers to deliver technologies as a service to users.

Drivers

- Drivers include cloud variations (such as hybrid IT and multicloud environments), which are now at the center of where the cloud hype currently is. Additionally, there are different types of cloud services (such as IaaS, PaaS and SaaS) at various stages of industry hype, as well as cloud-complementary approaches such as edge.

- New and advanced use cases for cloud introduce even more terms such as edge, distributed cloud, multicloud and cloud-native. These add to the overall cloud hype as well as the applicability of cloud to more and more scenarios, including enabling next-generation disruptions. We expect to see more such terms in the future as cloud evolves, which will continue to keep cloud hype high.

- An increasing number of tools, applications and platforms are only available in a cloud paradigm.

- New offerings specific for industries (industry clouds) and modifications meant to satisfy increasing sovereignty issues will continue to evolve.

Obstacles

- Cloud computing continues to be one of the most hyped terms in the history of IT. In fact, cloud computing hype is literally “off the charts,” as Gartner’s Hype Cycle does not measure amplitude of hype (meaning that heavily hyped terms such as cloud computing rise no higher on the Hype Cycle than anything else). Today, much of that hype is in aspects of cloud such as multicloud and cloud-native.

- The term “cloud” continues to be stretched and overused by a variety of vendors and users. In many cases, the true value is obscured or lost.

- There is the potential for global fragmentation of the broader cloud computing notion as governments or global regions attempt to create separate internet, web and cloud environments and standards. Risk and uncertainty of cost and budget overruns refrain the application of cloud computing.

User Recommendations

- Demand clarity from your vendors regarding the cloud. Gartner’s definitions and descriptions (which align with other useful ones such as from the National Institute of Standards and Technology [NIST]) of the attributes of cloud services can help with this.

- Examine specific usage scenarios and workloads, map your view of the cloud to that of potential providers and focus more on specifics than on general cloud ideas. Understanding the service models involved is key — especially an understanding of the shared responsibility model for security and other services.

- Beware of adopting cloud for the wrong reasons; it can lead to disastrous results. There are many myths surrounding cloud computing as a result of the pronounced hype.

- Build a business-driven cloud strategy that lays out your overall vision for cloud computing. Augment the strategy with detailed implementation plans or roadmaps. Focus primary attention on public cloud with other approaches (private, distributed, sovereign) as secondary considerations.

Sample Vendors

Alibaba Cloud; Amazon Web Services (AWS); Google; IBM; Microsoft; Oracle; Salesforce

Gartner Recommended Reading

The Cloud Strategy Cookbook, 2023

The Future of Cloud Computing in 2027: From Technology to Business Innovation

Private Cloud Computing

Analysis By: Thomas Bittman

Benefit Rating: Moderate

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

Private cloud computing is a form of cloud computing used by only one organization, or one that ensures an organization is completely isolated from others. As a form of cloud computing, it has full self-service, full automation behind self-service, and usage metering. It does not have to be on-premises, or owned or managed by the enterprise.

Why This Is Important

Cloud services offer many benefits, but isolation may also be required for security or regulatory reasons. Private cloud offers complete isolation, while providing the convenience and ease of use of cloud services. Private and public cloud are at opposite ends of the isolation spectrum. Public cloud providers have offered virtual private cloud, dedicated instances, dedicated hosts and distributed cloud offerings. Thus, there are a variety of isolation choices between public and private cloud.

Business Impact

Organizations that build a private cloud service are emulating public cloud computing providers to acquire similar benefits — mainly agility for new cloud-native applications, and business value and growth. When the goals are IT modernization or efficiency for existing applications, cloud-inspired deployments — that are more customized and less automated — are more appropriate.

Drivers

- **Regulatory or privacy reasons:** Private cloud may be useful where data or applications cannot reside on a public cloud, or need to reside in a specific location or on-premises.

- **Unique cloud service requirements:** Due to specific enterprise requirements (or support of existing applications), public cloud providers may not offer specific capabilities needed by the enterprise.

- **Evolution of virtualization:** Private cloud can be seen as a natural evolution of an existing virtualized environment, virtualizing all infrastructure and providing a service interface.

- **Standardization:** Specific private cloud services can be used to drive users to more standard offerings, further reducing costs and increasing automation.

- **Platform-level services:** Cloud services can provide rapid deployment, agile change, rapid scaling and innovation at the platform level.

Obstacles

- Building a custom private cloud can be very costly and complex. Moreover, most deployments that are called “private cloud” actually do not have cloud characteristics.
- Building an initial self-service offering is one thing, but maintaining and adding new, innovative features is usually untenable for enterprises.
- Private clouds being built to support existing applications require significant nonstandardization, which reduces automation potential.
- Users often have little motivation to move to a more standard model, unless they make fundamental business changes, such as adopting usage-based pricing for services.

User Recommendations

- Rule out public cloud offerings — including forms of distributed cloud — before investigating private cloud.
- Evaluate third-party hosting options and avoid building your own, if possible.
- Choose cloud-inspired technologies rather than true cloud, if IT efficiency or modernization for existing applications is the goal.
- Focus on business and application needs first, and let that determine the cloud service offerings.
- Focus on services that fit the cloud model (standard, high-volume and self-service); those that require agility and horizontal scalability; and usages that might be short-lived.
- Build or buy private cloud services with the potential to interoperate with, integrate with, or migrate to public cloud services in the future. Develop an exit strategy.
- Manage the scope of work. Start small and expand based on the business case.
- Build expertise in managing multicloud by going beyond just the private cloud.

Sample Vendors

HPE; IBM; VMware

Gartner Recommended Reading

Quick Answer: How Do I Obtain Isolated Private Cloud Services?

Differentiate Hosted Private Cloud Offerings Using These 7 Dimensions

Hyperconvergence

Analysis By: Philip Dawson, Jeffrey Hewitt

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

Hyperconvergence combines storage, computing and networking into a single system that reduces data center complexity and increases scalability. Multiple servers can be clustered together to create pools of shared compute and storage resources (or nodes), designed for convenient consumption. Delivery models include physical and virtual appliances, reference architectures, as a service or public cloud.

Why This Is Important

Infrastructure and operations (I&O) leaders seeking a cost-effective solution with a single management interface that excludes proprietary, external hardware controller-based storage should consider hyperconvergence as a viable option. Possible use cases include virtual desktop infrastructure (VDI), edge/Internet of Things (IoT), hybrid cloud and cloud-native.

Business Impact

Hyperconvergence enables IT leaders to be responsive to new business requirements in a modular, small-increment fashion, avoiding the large-increment upgrades typically found in three-tier infrastructure architectures. It is of particular value to midsize enterprises that can standardize on hyperconvergence and to the remote sites of large organizations that need cloudlike management efficiency with on-premises edge infrastructure.

Drivers

- Hyperconvergence provides simplified management that decreases the pressure to hire hard-to-find specialists. Adoption is greatest in dynamic organizations with short business planning cycles and long IT planning cycles tied to hybrid cloud delivery. The hyperconverged infrastructure (HCI) market is now trifurcating, focusing on the data-center-led “hybrid cloud” management use case with cloud-native applications, the VDI use case and the “edge/IoT” remote management use case.

- Hyperconvergence leads to lower operating costs, especially as it supports a greater share of the compute and storage requirements of the data center.

- Nutanix, an early innovator in hyperconverged integrated system (HCIS) hardware appliances, has largely shifted to a Hyper Converged Infrastructure (HCI) software revenue model and continues to increase its number of OEM relationships and partners.

- Larger clusters are now in use, and midsize organizations are considering hyperconvergence as the preferred alternative for on-premises infrastructure for block storage.

- Hyperconvergence vendors are achieving certification for more demanding workloads, including Oracle and SAP, and end users are beginning to consider hyperconvergence as an alternative to integrated infrastructure systems for some workloads.

- As more vendors support hybrid and public cloud deployments, hyperconvergence is a stepping stone toward public cloud agility as suppliers are expanding hybrid cloud deployment offerings for cloud-native applications.

- A number of niche hyperconvergence suppliers offer scale-down solutions to address the needs of remote office/branch office (ROBO) and edge environments.

Obstacles

- Applications designed for scale-up architectures (as opposed to scale-out ones) are unlikely to meet cost or performance expectations when deployed on hyperconverged infrastructure.
- The acquisition cost of hyperconvergence may be higher, and the resource utilization rate lower than for three-tier architectures.
- While HCI has somewhat matured from a hypervisor compute and storage function, software defined in networking is split between the obsolete software-defined networking (SDN) and networking around software-defined WAN (SD-WAN), driving edge deployments.
- For large organizations, hyperconverged deployments will remain another silo to manage.

User Recommendations

- Implement hyperconvergence for hybrid cloud infrastructure and cloud-native applications when agility, modular growth and management simplicity are of greatest importance.
- Establish that hyperconvergence requires alignment of compute, network and storage refresh cycles; consolidation of budgets; operations and capacity planning roles; and retraining for organizations still operating separate silos.
- Test the impact on disaster recovery and networking under a variety of failure scenarios, as solutions vary greatly in performance under failure, their time to return to a fully protected state and the number of failures they can tolerate.
- Ensure that clusters are sufficiently large to meet performance and availability requirements during single and double node failures, and require proofs of concept to reveal any performance anomalies.

Sample Vendors

Cisco; Dell; Microsoft; Nutanix; Sangfor; Scale Computing; StorMagic; VMware

Gartner Recommended Reading

Market Guide for Full-Stack Hyperconverged Infrastructure Software

Gartner Peer Insights 'Voice of the Customer': Hyperconverged Infrastructure Software

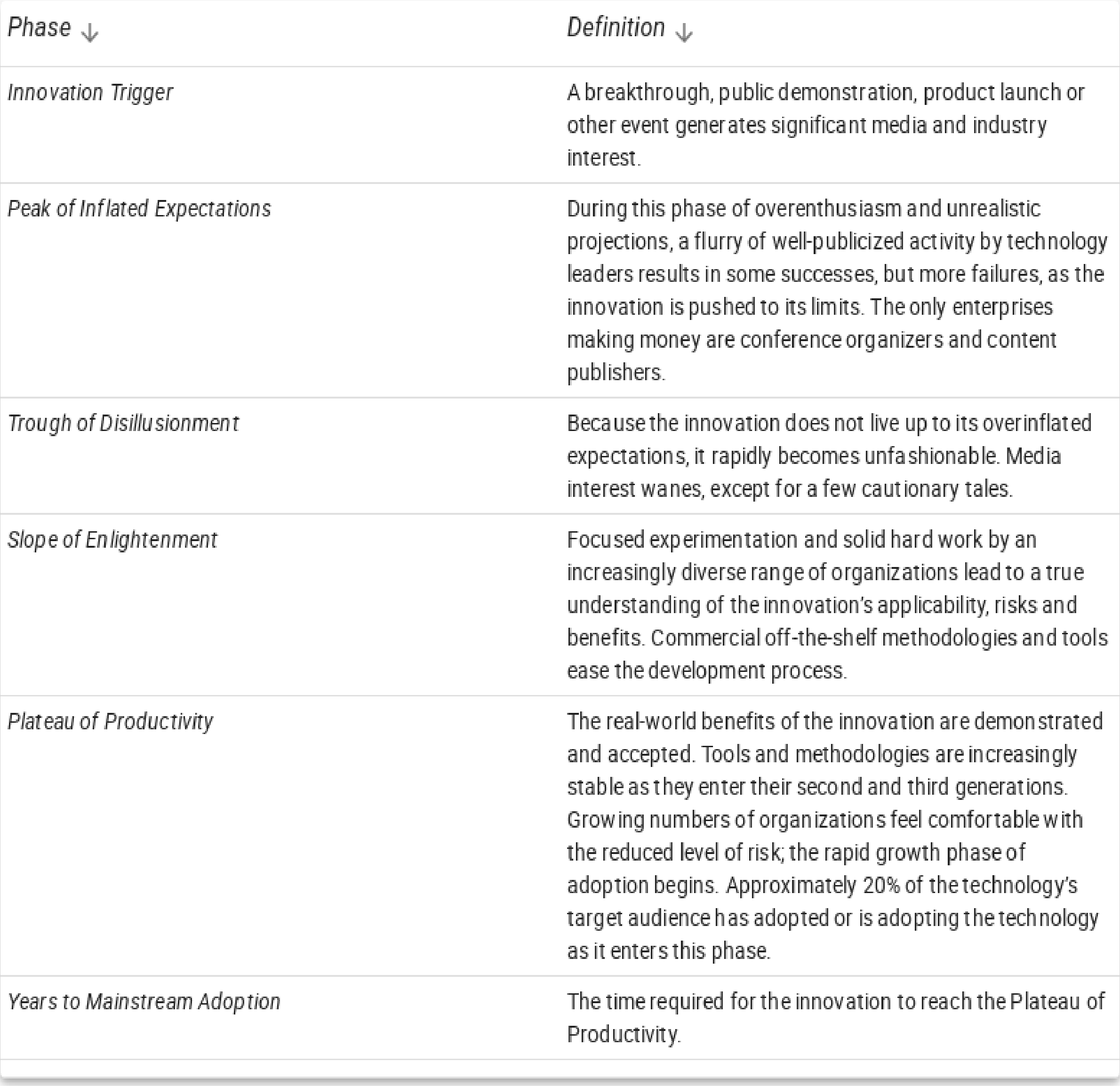
Appendixes

See the previous Hype Cycle: Hype Cycle for Infrastructure Strategy, 2022

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

(Enlarged table in Appendix)



Phase ↓	Definition ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (July 2023)

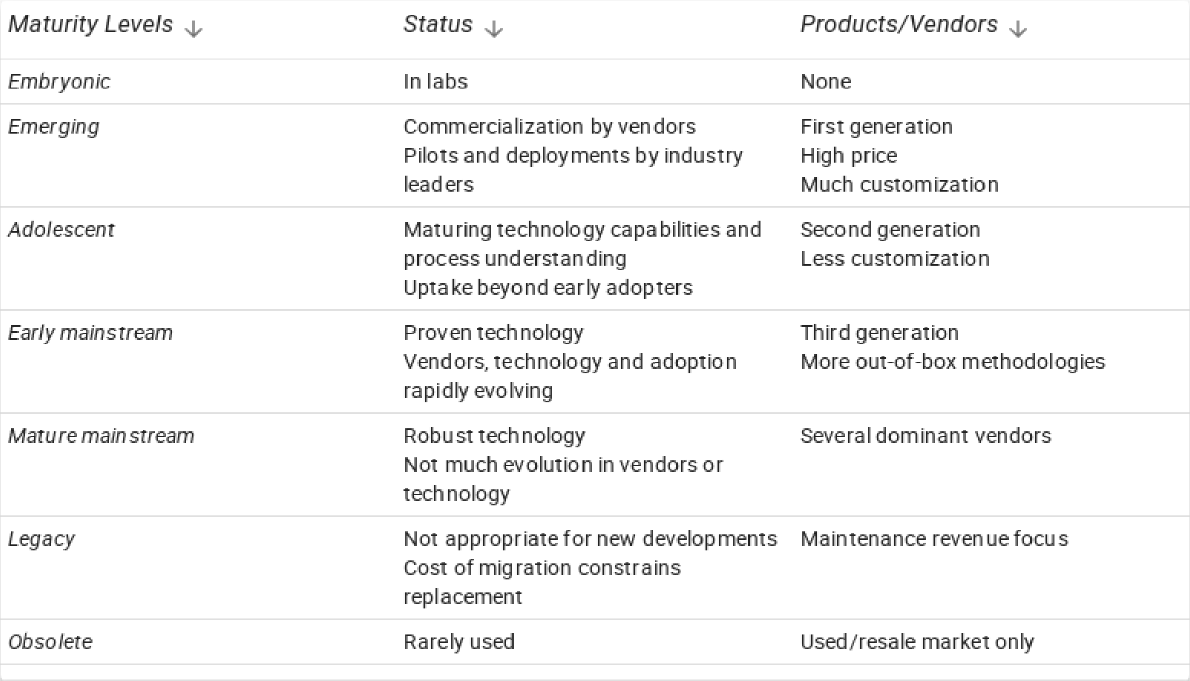
Table 3: Benefit Ratings

Benefit Rating ↓	*Definition* ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2023)

Table 4: Maturity Levels

(Enlarged table in Appendix)



Maturity Levels ↓	*Status* ↓	*Products/Vendors* ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)

Document Revision History

Hype Cycle for Infrastructure Strategy, 2022 - 11 July 2022

Hype Cycle for Infrastructure Strategies, 2021 - 21 July 2021

Hype Cycle for Infrastructure Strategies, 2020 - 8 July 2020

Hype Cycle for Infrastructure Strategies, 2019 - 31 July 2019

Recommended by the Authors

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

Understanding Gartner's Hype Cycles

Tool: Create Your Own Hype Cycle With Gartner's Hype Cycle Builder

Break Down 3 Barriers to Cloud Migration

Rationalizing Applications and Infrastructure for Cloud Delivery

Predicts 2023: XaaS Is Transforming Data Center Infrastructure

Table 1: Priority Matrix for Infrastructure Strategy, 2023

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	Cloud Computing OS Containers Serverless Infrastructure	Edge Computing Generative AI IoT Site Reliability Engineering		
High	Balancing Opex and Capex Hyperconvergence SD-WAN	BMaaS Cloud Sustainability Composable Infrastructure Consumption-Based Model Distributed Integrated Platforms Intelligent Infrastructure Multicloud Programmable Infrastructure	Container-VM Convergence Distributed Cloud Everything as a Service Intelligent Platforms Net-Zero Data Centers Off-Grid Power Operational AI Systems Programmable Platform	
Moderate	Private Cloud Computing	Augmented Reality in Data Centers Distributed Integrated Infrastructure Immutable Infrastructure		

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Low		Revirtualization Service Mesh		

Source: Gartner (July 2023)

Table 2: Hype Cycle Phases

Phase ↓	*Definition* ↓
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Phase ↓	*Definition* ↓

Source: Gartner (July 2023)

Table 3: Benefit Ratings

Benefit Rating ↓	*Definition* ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2023)

Table 4: Maturity Levels

Maturity Levels ↓	*Status* ↓	*Products/Vendors* ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (July 2023)