Hype Cycle for Edge Computing, 2023

Published 25 July 2023 - ID G00790891 - 108 min read

By Analyst(s): Bob Gill, Thomas Bittman, Philip Dawson

Initiatives: I&O Platforms

Edge computing success in the retail and industrial sectors reflects the evolution of IoT and distributed computing, while IT-driven use cases are developing. I&O leaders should use the technologies in this Hype Cycle to support both near-term implementations and still-emerging edge applications.

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

Analysis

What You Need to Know

Edge computing continues to reflect acceptance and maturation in OT-driven use cases, while much of general purpose IT-oriented adoption still remains shrouded in hype. From a market perspective, edge computing deployments continue to accelerate, driven by the trend toward digital transformation (as well as the need to collect, analyze and monetize data and interactions at the edge). Edge computing is increasingly seen as a front-end digital transformation complement to back-end cloud computing, NOT a replacement for cloud. They are synergistic. Edge orchestration, security and platform development all continue to progress, but the huge diversity in use cases and challenges for vendors to monetize that diversity result in slow progress.

Regardless, a large percentage of enterprises have begun deploying some form of edge computing solution, initially focused on one localized use case. Inevitably, enterprises that deploy edge computing for a single use case expand workloads rapidly. This tends to drive demand for more standard offerings in select industry verticals (such as retail), and more layered horizontal solutions across vertical industries, driven by rapidly maturing edge management and orchestration offerings.

The Hype Cycle

There are three important factors that have affected the edge computing trend in the last year:

Drivers Are Shifting: Low latency was an initial primary driver of edge computing. However, the increase in digital interactions at the edge, growing data volume at the edge and the resulting impact on bandwidth requirements are driving architectural considerations that favor moving even cloud-located processing elements to the edge. One aspect of edge computing that continues to accelerate is the inclusion of machine learning at the edge to not only facilitate data handling (i.e., “data thinning,” noise reduction), but to create closed-loop automation systems in operational technology (OT) environments, where data is not only collected and cataloged, but immediate action is driven by the results. This increased intelligence and resulting actions at the edge not only enable innovative new use cases, but are reflected in many vendors’ future roadmaps, such as model and fleet management bundled into edge management and orchestration offerings.

Diversity of Requirements Remains a Challenge: The majority of edge computing solutions continue to revolve around unique and explicit use cases including location-specific, context-specific, vertical-specific industry requirements, leaving the promise of a standardized cross-vertical platform still a future. On the IT-focused front, enterprises continue to focus on modernizing remote-office/branch-office solutions to support new edge workloads, many based on machine learning enhancements to general business functions.

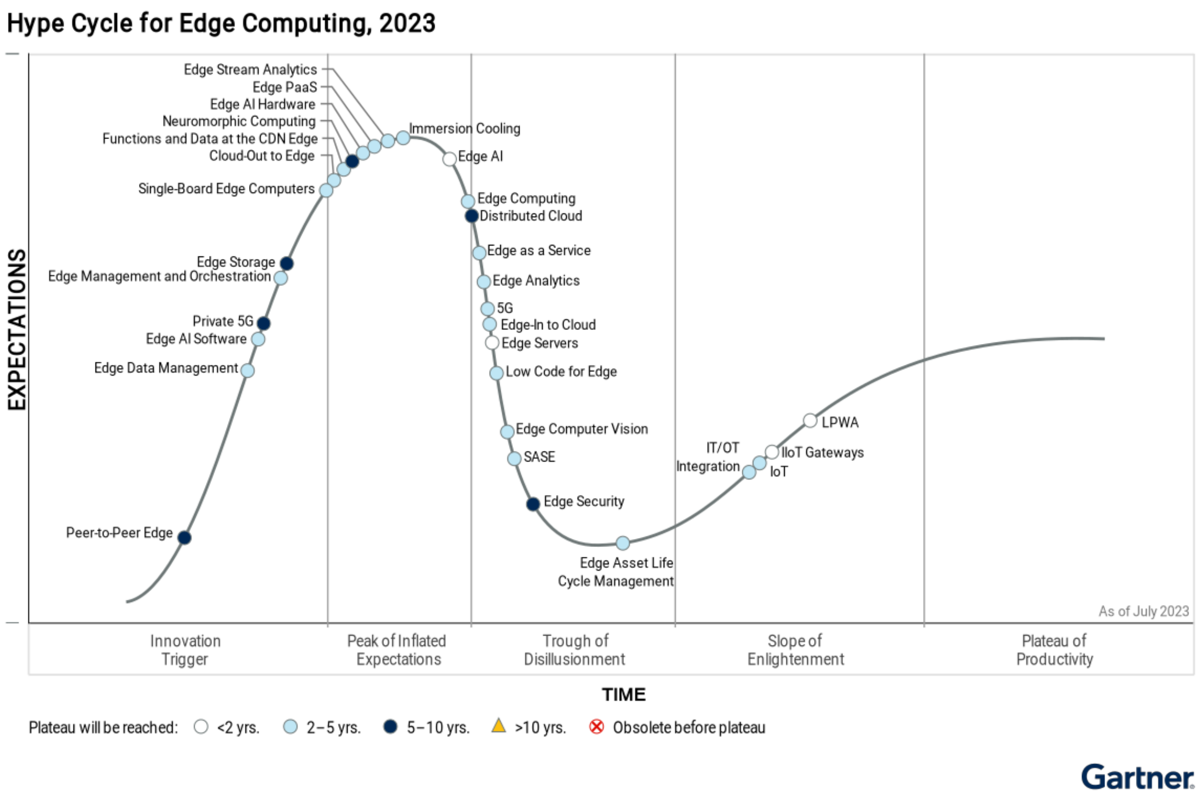
Enterprise Edge Moving From Tactical to Strategic: Gartner is seeing more inquiries about edge computing standards, extensibility and the creation of edge computing strategies.

Innovations that enable edge computing are evolving at different rates in the Hype Cycle:

- **Slow Movers:** Edge data management is often cited as a critical gap in deployment plans, particularly as edge implementations scale.
- **Moving Quickly:** Edge security, and in particular secure access service edge (SASE), have been developing quickly (primarily as a service), as well as intelligent technologies for the extreme edge, like single-board edge computers.
- **Peak Hype:** There’s significant market interest driving the hype for using CDN edge developer tools, edge AI and edge platform as a service (PaaS).

- **Very Early Innovations:** Peer-to-peer edge innovations continue to move slowly, emerging as the density of computing (and the data) at the edge grows (e.g., in factories).

Figure 1: Hype Cycle for Edge Computing, 2023



The Priority Matrix

The Priority Matrix provides perspective on the edge computing innovations that will have a bigger impact, and those that might take longer to fully mature.

- **Consider now:** Some core innovations are nearing maturity, including IoT in general, computing hardware like edge servers and industrial IoT gateways, hardware-based security, and integrated low-power wide-area (LPWA) network technologies. Also consider private 5G now as a higher speed wireless on-premises tool when the application mandates higher speed or greater numbers of devices (even though full mainstream adoption is more than five years away).

- **Evaluate in the near term:** A few transformation innovations will mature in the two- to five-year time frame, including edge AI, edge analytics and edge security. Edge security is still divided between the widespread application of existing security tools in OT-based implementations, SASE and the emerging workload-level security architectures, but each of these categories is maturing in parallel.
- **Monitor, but give them time:** A number of edge computing innovations will take more than five years to become mainstream, including neuromorphic computing, distributed cloud and peer-to-peer edge.

Table 1: Priority Matrix for Edge Computing, 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		Edge Computer Vision Edge Computing Functions and Data at the CDN Edge IoT SASE	Neuromorphic Computing	
High	Edge AI Edge Servers	5G Cloud-Out to Edge Edge AI Hardware Edge AI Software Edge Analytics Edge as a Service Edge Asset Life Cycle Management Edge Data Management Edge-In to Cloud Edge Management and Orchestration Edge Stream Analytics IT/OT Integration Low Code for Edge Single-Board Edge Computers	Distributed Cloud Private 5G	
Moderate	IIoT Gateways LPWA	Edge PaaS Immersion Cooling	Edge Security Edge Storage Peer-to-Peer Edge	
Low				

Source: Gartner (July 2023)

On the Rise

Peer-to-Peer Edge

Analysis By: Thomas Bittman

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Peer-to-peer edge computing enables distributed computing across an edge environment for resilience, workload orchestration, horizontal scaling, swarm learning, and interaction and cooperation between edge computing nodes using local or mesh networking as an enabling technology.

Why This Is Important

Much of the early focus on edge computing has been extending intelligent things to the cloud, and extending cloud capabilities closer to users and things at the edge. As more things connect at the edge, there will be a growth in capability, processing, interaction and decision making across things — creating systems of interaction at the edge. These systems can leverage each other for resilience, horizontal scaling and orchestration of work.

Business Impact

Peer-to-peer edge capability will enable enterprises with dense digital locations (especially factories) to maximize digital interactions and decision-making locally, increasing efficiency, lowering cost, and improving resilience and latency. It will enable specific types of use cases, such as swarm learning and collaborative immersive work.

Drivers

- The growth of complex systems and interactions at the edge.
- The importance of semiautonomy when the edge is disconnected.
- Unpredictable edge workload peaks requiring efficient orchestration.
- The need for resiliency of an edge fabric if a node goes down.

- Emergence of osmotic computing technologies.
- The need for local training, maintaining data locally and the growth of swarm learning.
- The growing maturity of blockchain technologies as an enabler.
- Mobility and coordination of endpoints (such as robot swarm decisions).
- Collaborative immersive computing.
- Mesh networking — allowing a remote device to communicate small packets of information via connections with other devices.

Obstacles

- Diversity of use cases and lack of standards will limit peer-to-peer edge computing to specific use cases, but will gradually generalize to broader use cases as standards evolve.
- Lack of skills necessary to establish and maintain a complex peer-to-peer architecture.
- Security challenges inherent in local networking.
- Complexity caused by rapidly changing nodes.

User Recommendations

I&O leaders should:

- Deploy custom solutions for peer-to-peer edge when the business case is very strong, and complex and lengthy development and rollout processes are worth the investment. Enterprises should do this carefully until broad solutions and standards appear.
- Evolve peer-to-peer edge in a stepwise manner: Basic connectivity and simple interactions first, then resilience and workload orchestration, and finally horizontal scaling and distributed processing.
- Align with CISOs for compliance regulations like NIST, GDPR and SIEM frameworks.

Sample Vendors

Alef Edge; Equinix; Fasetto; Hewlett Packard Enterprise (HPE); mimik; Rajant; Storj

Gartner Recommended Reading

Innovation Insight for Federated Machine Learning

Building an Edge Computing Strategy

Designing Blockchain Smart Contract Security and Access Control

Emerging Technologies: Critical Insights on Metaverse

Edge Data Management

Analysis By: Aaron Rosenbaum

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge data management comprises the capabilities and practices required to capture, organize, store, integrate and govern data outside of traditional data center and public cloud environments. An increasing number of digital business use cases, including those based on IoT solutions, will leverage data in edge environments. This expansion creates tremendous opportunities to optimize resources and drive real-time decisions and actions, but also brings complexity and governance challenges.

Why This Is Important

Valuable data is increasingly generated and used outside of traditional data centers and cloud environments. This data often has a shorter useful life span, requiring value to be captured near the place and time of its origin. This is the role of edge-computing environments deployed closer to assets in the physical world. Edge data management will both impact and enable IT leaders and their teams, requiring new capabilities and skills while also opening up new opportunities to deliver value.

Business Impact

Edge data management creates value in various ways:

- By distributing data management to the edge, data-centric solutions better support demand for local and real-time data.
- More solutions, such as for IoT use cases, must operate in disconnected (or intermittently connected and low-bandwidth) scenarios.
- It enables smarter physical assets and collections of assets, including remote management or autonomous behavior, via edge data.
- It addresses inconsistencies, protection, sovereignty and other governance issues arising from siloed edge environments.

Drivers

- **Extreme speed:** By placing data, data management capabilities and analytics workloads at optimal points ranging all the way out to endpoint devices, enterprises can enable more real-time use cases. In addition, the flexibility to move data management workloads up and down the continuum from centralized data centers or the cloud to edge devices will enable greater optimization of resources.
- **Data gravity:** Bandwidth costs and scenarios with limited or intermittent connectivity demand the ability to organize and process data closer to the edge.
- **Expanded scale and reach:** By using distributed computing resources, and spreading the load across the ecosystem, enterprises can broadly scale their capabilities and extend their impact into more areas of the business. These areas include use cases and outcomes traditionally managed only via operational technology (OT) teams, such as those managing equipment in industrial settings. Dedicated hardware for edge processing of data will continue to amplify these benefits.
- **Resiliency:** Pushing data management capabilities toward edge environments can also bring benefits in the form of greater fault tolerance and autonomous behavior. If edge environments do not require centralized resources, then issues with connectivity to, or unplanned downtime of, those centralized resources don’t disrupt processes that rely on local edge capabilities.

Obstacles

- **Management of distributed data architectures:** Data management has been largely based on principles of centralization — bringing data to central data stores (e.g., data warehouses), and then processing that data to create value. Edge environments break that model via distributed data architectures, raising complex choices about where to locate and aggregate data on the continuum of cloud/data center to edge. Determining the right balance of latency and consistency is one such choice.
- **Governance and security:** With the distribution and complexity of edge environments, data governance and security become challenging. Organizations should extend their governance practices and policies to address edge-resident data storage and processing capabilities, including disposal of ephemeral or nonvalue event data.
- **Organizational and skills considerations:** Many modern applications are being developed and deployed by OT teams lacking data management skills and oversight, or by IT teams lacking edge computing skills and experience.

User Recommendations

- Identify use cases where data management capabilities in edge environments can enable differentiated products and services by collaborating with OT and IT personnel working in edge locations.
- Expand the skill sets of IT and OT teams to include edge platforms and the technologies required to manage data and data-intensive workloads on them.
- Augment existing data management infrastructure to support edge deployment by partnering with product teams that are implementing IoT platforms and similar distributed computing architectures.
- Place a greater emphasis on end-to-end system design. Understanding the dependencies between all components of distributed data pipelines, analytics workloads and AI models will be crucial to success.
- Ensure safety and control by extending existing governance capabilities to edge data environments.

Sample Vendors

Couchbase; FairCom; IBM; Macrometa; Microsoft; MongoDB; ObjectBox; Xencia

Gartner Recommended Reading

Get Ready for Data Management at the Edge: Key Considerations and Actions

Building an Edge Computing Strategy

Forecast: Internet of Things, Endpoints and Communications, Worldwide, 2022-2032, 1Q23 Update

Forecast Analysis: Edge Hardware Infrastructure, Worldwide

Edge AI Software

Analysis By: Eric Goodness, Carlton Sapp

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge AI software refers to the use of systems for orchestrating, integrating, deploying and monitoring machine learning (ML) models across edge environments. While predominantly focused on inference, advanced systems may enable training at or near the edge based on local context data.

Why This Is Important

AI requires meticulous model versioning, A/B testing and optimized delivery for compute accelerators in the data center or edge. Edge AI software technologies and toolkits, such as containers and AI frameworks, enable enterprises to accelerate the orchestration, integration and deployment of AI in edge computing. As a subcategory of broader MLOps software, edge AI software emerged as purpose-built software focused on building and deploying native TinyML and integer quantization of models.

Business Impact

Using edge AI software technologies, enterprises can improve management of edge AI for outcomes such as:

- Use of streaming AI analytics with event-based architecture platforms and streaming data contexts.
- Manage versioning, roll out/roll back and monitoring of AI models deployed.
- Reduce communications cost with less data traffic across the edge and the cloud.
- Increase availability even when the edge is disconnected from the network.
- Latency-sensitive decision making for business-critical applications.

Drivers

- The need for mature solutions for smart video to monitor environments for much lower costs.
- IoT sensor data for acoustics, biometrics, temperature, humidity and vibration, among others, are basic features for solutions such as condition-based monitoring and predictive maintenance.
- For control patterns, supervisory deployment configurations can focus on executing highly regulated actions of IoT endpoints (IT and OT). For example, smart city deployments and intelligent industry 4.0 applications.
- For graph patterns, highly interconnected and relational IoT endpoints in distributed edge environments are commonly deployed in multiaccess edge computing applications to solve bandwidth optimization problems.
- For swarm patterns, deployment configurations can focus on decentralized edge environments or the collective actions of IoT endpoints. For example, logistics fleet optimization and management to enhance routes and extend the life of service vehicles.
- To adequately support these edge AI architectural patterns, organizations need to train AI models and optimize them for delivery (e.g., tuned for specific architecture, optimized for hardware or operating environment constraints, etc.), management (with roll back, and champion/challenger testing), monitoring (for changes in state or model drift, need for redeployment or retraining, etc.), and pairing with lower latency inference technologies at the edge, enabling faster decision making.

Obstacles

- Many of these applications are in trial phases, and broad adoption is a few years away. Some use cases with comparatively mature system stacks (e.g., surveillance cameras or video recognition systems) are closer to productization.
- Most use cases require a high degree of contextualization and customization. For most enterprises, edge AI software will offer limited capabilities in self-serve modalities. Maximizing value from edge AI software will require vendor or third-party professional services, driving higher costs and consequently creating adoption challenges.
- Edge AI is often deployed in operational environments — requiring coordination, cooperation and communication between IT and OT or engineering teams, which can be challenging in some organizations.

User Recommendations

- Deploy analytics closer to where the data is created and managed by curating AI edge aligned with relevant streaming data ecosystems.
- Architect middleware that enables the orchestration, integration and deployment of AI in edge systems.
- Augment IoT analytics architectures with edge AI architecture patterns to take advantage of data and model parallelism across IoT endpoints.
- Leverage model compression to push AI to brownfield assets (MPU and MCU, lower cortex) to create intelligence in aged assets. This approach has helped OEMs create premium service contracts on equipment that has been end-of-life/end-of-sales for years.
- Select edge AI software technologies that support the broadest set of relevant edge AI architectural patterns (e.g., swarm, control, graph).
- Organizations need to develop capabilities for better IT-OT collaboration and coordination for effective edge AI strategies.

Sample Vendors

Amazon Web Services (AWS); Boon Logic; FogHorn; IBM; Intel; Modzy; Neuton.AI; Sorba.AI

Private 5G

Analysis By: Sylvain Fabre

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

A private 5G network is based on 3rd Generation Partnership Project (3GPP) standard R15 or above to provide unified connectivity, optimized services and security for enterprises. A 5G private mobile network (PMN) is specific to the enterprise and used to interconnect people and things. Deployments can be entirely on-premises, with local breakout, or linked to a public cloud or local telco.

Why This Is Important

Multiple verticals will require 5G PMN deployments to realize the full effect of their digital transformation initiatives. Adopting new 5G standards earlier than communications service providers (CSPs) can offer on their public infrastructure can provide access to additional functionality. Distinct from the public network, private 5G supports voice, video, messaging, data and Internet of Things (IoT) with higher performance requirements. It can optimize cost or connectivity (for example, less expensive than Wi-Fi for large area coverage).

Business Impact

Private 5G enables transformational digital use cases for industry, especially in conjunction with other technologies, such as factory digital twin or edge AI for computer vision. 5G PMN can offer enterprises improved security, independence and enable efficiency gains; for example, complete 5G coverage in factories, with speeds over 1 Gbps, and support for edge and AI use cases with guaranteed performance levels.

Drivers

- Applicability and vertical specific integration are increasing. Beyond 3GPP, other bodies are now contributing, such as 5G Alliance for Connected Industries and Automation (5G-ACIA) or 5G Automotive Association (5GAA).
- Liberalization of the radio spectrum has opened up standard radio bands, often around 3.5 GHz, for use by private 5G networks.
- Requirement for full, reliable network coverage for machines, sensors and equipment, including indoor, outdoor, office and large industrial areas at lower cost than Wi-Fi.
- Performance profile for demanding industrial use cases, in particular when low-latency, high-bandwidth (especially uplink), and reliability are required and exceed the capabilities of the shared public infrastructure.
- Private 5G has another class of use cases, not focused on mobility initially, but requiring a high-performance backbone where wiring is complex and costly — such as in a factory deployment.
- Interest from telecommunications service providers (TSPs) that can offer 5G PMN to various verticals, such as I4.0 factory automation, mining, oil, utility and railroad companies. IoT providers, universities, stadiums and so on are thereby expanding into industries and generating new revenue.
- Alternative provider types beyond the CSPs, such as integrators, infrastructure vendors and hyperscalers, are driving new deployments and proofs of concept.
- Some enterprises deploy private networks because they want to run their network more independently, as their own infrastructure, with limited outside dependency. One example is long-term commitment from public network operators; also, data privacy can be a key concern, with data loss prevention security controls in place to ensure sensitive information does not leave the enterprise perimeter.
- Some defense and government clients have indicated a wish to have more control and visibility into the vendors involved in the mobile services provision, which can be an issue over a shared public network built and managed by a CSP.
- Low-latency applications using processing embedded in network infrastructure are logistically easier if the application, and infrastructure are owned by the same entity.

Obstacles

- Unclear business models and value justification vs. alternatives (e.g., 4G PMN).
- Perception that real value begins from 3GPP R16, and that maturity and availability of R16 solutions are still a work in progress; for example, with network slicing.
- Complex deployment and operation.
- Limited availability and cost of equipment designed to use the radio bands available for private network use.
- Module availability and pricing for R16 and up.
- Lack of outcome-based pricing models.
- Spectrum availability and/or cost in some countries.
- Perception of risk regarding timing and relevance of private 5G.
- Feedback from some industrial clients mentioned that the majority of their use cases could be serviced by a 4G private network, and/or NarrowBand-Internet of Things (NB-IoT) and other low-power wide-area networks (LPWA networks), such as LoRaWAN.

User Recommendations

- Differentiate from other providers, like large equipment vendors, systems integrators (SIs), resellers, smaller specialist network vendors and hyperscalers, by integrating PMN with other functions like supplier information management (SIM), IoT platforms, edge computing, design and managed services, and national roaming.
- Co-create networks by partnering with SIs and consultancies that have the required industry skills for design, deployment, and managed services engineering headcount and evaluation test bed environments. For example, build manufacturing 5G PMN with connectivity, security and AI capabilities.
- Design licensed and unlicensed/shared spectrum options where available.
- Supplement your engineering teams by working with IT service providers. Do not expect or plan on public 5G replacing WLAN in large portions of your environment. Instead, IT leaders should select private 5G for specialized use cases with large coverage areas and known application performance requirements.
- Identify use cases and their requirements to establish where 5G can be implemented — for example, in applications using HD wireless cameras.

Sample Vendors

AT&T; Celona; China Mobile; Ericsson; HPE (Athonet); Huawei; Nokia; T-Mobile; Verizon; Vodafone

Gartner Recommended Reading

Infographic: 5 Steps for Vendors to Scope and Run Successful POCs for Enterprise 5G PMNs

Market Guide for 4G and 5G Private Mobile Networks

3 Go-to-Market Strategies for Product Leaders in Private Mobile Networks

Research Roundup: How to Build Winning Propositions in 5G Private Mobile Networks

Quick Answer: What Metrics Can TSPs Consider for Their Private Mobile Network Solution Development?

Edge Management and Orchestration

Analysis By: Thomas Bittman

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge management and orchestration (EMO) provide edge infrastructure software, and software stack and application management and security in many remote, distributed deployments. A layered approach to EMO — previously delivered through monolithic application stacks — supports the assembly of functional, full-featured orchestration capabilities from subset-specific offerings.

Why This Is Important

EMO solutions enable zero-touch deployments, updates and remote management of widely distributed edge computing nodes — a core challenge with edge computing. An EMO solution can create a platform for enterprises to deploy a broad range of custom or packaged applications across thousands of edge computing deployments without on-location IT staff. It makes it possible to deploy flexible computing capability where it doesn't exist today.

Business Impact

Without EMO solutions, enterprises are locked in monolithic appliance stacks, forced to custom-build a platform or require on-location IT skills. EMO solutions enable enterprises to focus on the applications and data processing at the edge — instead of the infrastructure. Also, EMO enables the orchestration of cloud-native distributed apps into edge platform as a service (PaaS) environments.

Drivers

Digital transformation is pushing enterprises to create digital edges, which include IoT devices and edge computing solutions for real-time responses and data processing. A core challenge is the need for zero-touch, low-skill deployments at the edge with remote, centralized management at scale. That creates the need for EMO solutions to:

- Support new applications based on containers and existing applications using virtual machines and a hypervisor.
- Assist with orchestration across the edge and design tools to assist with deployments.

- Provide a core security foundation for a digital edge, which includes managing applications, data, network and access.
- Enable flexibility and extensibility because enterprises tend to deploy new applications and use cases over time, with varied requirements.

Obstacles

- Scaling from tens or hundreds of thousands of nodes and enabling effective monitoring and management at scale.
- Supporting the right edge footprint size needed, whether very small or large.
- Open enough to support a broad range of tools and technologies without sacrificing zero-touch management.
- Providing enough vertical domain capabilities without being limited to particular vertical industry domains.
- Targeting the appropriate edge decision-makers, whether IT, operations, line-of-business, developers or all of the mentioned.
- Competing architectural approaches and the need for standards and fewer choices — especially between distributed cloud and edge-native solutions.
- Partnerships and solution ecosystems matter deeply, especially to vertical industry domains, machine learning and hyperscale providers.

User Recommendations

- Choose an EMO solution that enables future extensibility, new workloads and new uses of edge data. Supporting or promoting standards matters greatly.
- Ensure that the EMO solution is effective at the volume and footprint scale you envision, in the geographies you require and with differences in connectivity.
- Evaluate the long-term viability of the EMO solution, understand the state of the partner ecosystem and ensure that the EMO solution integrates well with your chosen hyperscale provider.
- Fully evaluate the security capabilities of the EMO solution.
- Check references for real-world deployments that are similar to yours, with equivalent vertical industry requirements.

Sample Vendors

Avassa; Nearby Computing; Pratexo; Scale Computing; Spectro Cloud; Sunlight; Telco Systems; Veea; ZEDEDA

Edge Storage

Analysis By: Julia Palmer

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Edge storage enables the creation, analysis, processing and delivery of data services at, or close to, the location where the data is generated or consumed, rather than in a centralized environment. Edge storage can be deployed at regional or remote data centers, aggregation points, edge servers and edge gateways. Edge storage is not being delivered by a single technology, because it must be tailored to the specific edge computing use cases.

Why This Is Important

Infrastructure and operations (I&O) leaders are beginning the process of laying out a strategy for how they intend to manage data at the edge. Although I&O leaders embrace infrastructure as a service (IaaS) cloud providers, they also realize that a significant part of the infrastructure services will remain on-premises, and would require edge storage data services.

Business Impact

I&O leaders are changing up their storage strategies as edge computing becomes a critical part of an overall cloud-connected data center transformation. This is driving I&O leaders to change their role from the providers of infrastructure to providers of data services everywhere. With that in mind, I&O leaders are resetting their storage strategies and vendor selections for all deployments outside the public cloud infrastructure, focusing on specific data services for the edge.

Drivers

- As cloud migration continues, the infrastructure and operations team is transforming from the provider of data center infrastructure to the provider of the data services everywhere, focusing on select use cases that require storage at the edge. The four most popular use cases at the edge are distributed cloud/data center, data processing at the edge, content collaboration and access, and data ingest and streaming.
- Data services at the edge require many factors that preclude deployment in the public cloud: data gravity, cloud and bandwidth costs, application-specific data latency, and the effects on throughput of the speed of light. This is in addition to data security, data autonomy and data governance.
- Edge storage services are emerging for latency sensitive and bandwidth-intensive workloads that aren't ideally suited for the public cloud or the core data center. Examples of such workloads include real-time data processing, collaboration and the synchronization of massive amounts of data with online storage.

Obstacles

- I&O organizations struggle to determine what actions should be taken now, as opposed to taking a wait-and-see attitude and planning to optimize their IT operating models to mitigate risks and avoid pitfalls that may jeopardize efforts at the edge.
- The diversity of use cases, workloads, volume of data and unique infrastructure requirements at the edge introduces the potential for issues in system management, costs, security and resilience factors.
- Edge storage is not a single technology, because it needs to be tailored to the specific use cases.

User Recommendations

- Create edge storage platform initiatives by identifying edge-centric workloads, use cases and data service management methods.
- Choose an edge storage topology and platform approach by addressing unique workload requirements that are self-healing, software-defined and power-efficient, and can be elastically scaled up and down cost-effectively.
- Prepare for any new enterprise data center storage deployment to be edge-ready, by prioritizing requirements for the edge operating model and public cloud integration.
- Select edge storage products and technologies that focus on addressing key challenges, such as autonomous operations, centralized data management, performance density and data transfer optimization.

Sample Vendors

AWS; Dell; Hivecell; HPE; LINBIT; Microsoft Azure; Nutanix; SoftIron; StorMagic; VMware

Gartner Recommended Reading

Innovation Insight: Rethink Your Enterprise Storage and Cloud Data Services Strategies for the Edge Awakening

Competitive Landscape: Hyperscale Edge Solution Providers

Market Guide for Hybrid Cloud Storage

Predicts 2023: Edge Computing Delivery and Control Options Extend Functionality

At the Peak

Cloud-Out to Edge

Analysis By: Ed Anderson, Bob Gill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Cloud-out to edge describes an architectural construct where a centrally managed cloud environment, typically a hyperscale cloud, provides cloud service capabilities that are extended to edge environments. In a cloud-out to edge architecture, the cloud control plane, including security, identity and access management, governance, operations, programming models and interfaces, and other control elements, originate in the cloud and are then instantiated at the edge.

Why This Is Important

The move to public cloud drives centralization of operating processes, including the controls used to govern the environments. Cloud-out to edge is an architectural construct that supports the extension of public cloud control models to edge environments. Cloud-out to edge complements edge-in to cloud models. Cloud-out to edge is popular when organizations standardize IT operational control through centralized, public cloud environments.

Business Impact

IT environments are growing in complexity due to the expanding cloud use cases creating complexity, operational risks, and increased costs. Cloud-out to edge models extend cloud capabilities, including the cloud control plane, to other environments, including systems operating at the edge. Extending public cloud capabilities to edge environments can be a means to address the complexities of distributed, hybrid environments by unifying IT operations under a common operational framework.

Drivers

- Adoption of hyperscale public cloud services continues to increase. Gartner predicts continued growth in public cloud adoption with IT spending rates on public cloud services expected to grow almost 20% through 2027.

- Cloud operations have become critical for most organizations, driving increased investment in tools and skills in cloud management practices.
- Centralized, hyperscale cloud services are not well-suited for all application scenarios, particularly those better-suited to run at the edge. This creates an architectural divide between cloud and edge, which impacts operations, programming interfaces, security, identity and access, and application compatibility.
- Enterprise digitalization trends increasingly involve use cases and processes that operate in a distributed manner, driving the need for services deployed at edge locations.
- Distributed architectures, including edge computing and distributed cloud, can benefit from the distribution of cloud services from centralized environments to edge. Cloud-out to edge models can extend the cloud control plane to provide management, governance and oversight to edge environments.
- Standardizing technologies around cloud technologies and unifying operations using the cloud control plane can reduce complexity and help manage costs.

Obstacles

- Cloud-out to edge assumes a centralized cloud system, which may not exist in organizations that are still maturing their cloud strategy.
- Cloud-out to edge assumes the standardization of architecture, technologies and operational control using a centralized cloud service. This approach can increase dependence on a single cloud provider.
- Cloud-out to edge implementations may intersect with existing edge services, including operational technology, which may be managed outside the IT domain and funded by non-IT budgets.
- Multicloud strategies, which are common with most organizations, may conflict with the cloud-out to edge approach. Cloud-out to edge typically drives unification of technology, architecture and control to a single cloud environment, which may conflict with an organization’s desired multicloud approach.
- Cloud provider offerings purported to support cloud-out to edge implementations are still maturing, and often don’t deliver the full benefits of distributed cloud approaches.

User Recommendations

- Focus on the needs of use cases operating at the edge to determine whether an edge-in to cloud or a cloud-out to edge approach will work best.
- Establish a comprehensive cloud and edge strategy to guide cloud-out to edge and edge-in to cloud implementations. Let business value and operational benefits lead cloud and edge decisions.
- Build strong, centralized cloud operating capabilities before pursuing cloud-out to edge strategies. Cloud competencies will be critical to achieving success in cloud-out to edge implementations.
- Assess the risks and benefits of a cloud-out to edge approach, particularly if you have a stated multicloud strategy.
- Seek expert help from system integrators and managed service providers with expertise in both cloud and edge environments.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Google; IBM; Microsoft; Oracle; VMware

Gartner Recommended Reading

I&O Platforms Primer for 2023

Market Guide for Edge Computing

Quick Answer: How to Make the Right Choice Between Hyperconverged, Traditional and Distributed Cloud Infrastructure

Distributed Cloud: Does the Hype Live Up to Reality?

Emerging Tech: Hyperscale Edge Enables Integrated Edge Infrastructure and Platform Services

Single-Board Edge Computers

Analysis By: Tony Harvey

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Single-board edge computers are small, low-cost general-purpose systems that perform functions such as anomaly detection or AI inferencing (e.g., image recognition) at the edge. Based on a system-on-chip (SoC) solution, single-board computers are designed with the minimum capability to perform the tasks required. I/O interfaces will vary but at a minimum include a wired or wireless network interface.

Why This Is Important

Single-board computers have existed in the industrial space for some time. What is new are the capabilities of these devices, e.g., the ability to do real-time audio and video processing, and the need to connect these devices to IP-based networks. This introduces new capabilities, e.g., low-cost image recognition of dials and gauges so that they can be continuously and automatically monitored, as well as a new set of risks that come with connection to general-purpose networks.

Business Impact

Single-board computers provide an intermediate form factor that can be positioned at the edge, but are less expensive than standard servers, providing cost-effective edge processing in more locations. They offer lower cost intelligence for automation, monitoring and reporting of edge-located sensors and equipment that could not otherwise be managed due to cost constraints. This improves efficiency and enables better decision making, delivering both bottom-line and top-line benefits.

Drivers

- **Cost and scale:** Edge locations require devices in tens or hundreds of thousands locations to provide programmable real-time/low-latency responses to sensors and actuators. Single-board edge computers provide a low-cost, highly programmable solution.
- **Automated monitoring of unintelligent devices:** Many manufacturing and process control solutions are designed for human inspection. Single-board edge computers can provide low-cost computer vision and recognition solutions to automate monitoring for dials and gauges that require visual monitoring.
- **Speed to value:** Single-board edge computers use high-level programming languages that make it easy to prototype and deliver solutions. In addition, many are integrated into existing AI frameworks, e.g., NVIDIA VisionWorks for image recognition or eKuiper for streaming analytics.
- **Edge AI:** New AI-optimized inferencing chipsets and application-specific integrated circuits (ASICs) are being incorporated into single-board edge computers. Gartner's 2021 Emerging Tech Impact Radar: Edge AI identifies a range of AI-enabled edge technologies that are poised to broadly disrupt markets in the coming decade, including edge computer vision, neuromorphic computing, advanced swarm intelligence, the orbital edge and blockchain for IoT.

Obstacles

- A wide variety of manufacturers and a lack of platform standards may constrain deployment of edge applications to specific vendor platforms.
- Methods for managing and updating a very large number of distributed devices are relatively immature in the IT/OT space.
- Enabling remote automated control of systems that control industrial processes also opens up the potential for security breaches on systems where failures can result in injury or loss of life.

User Recommendations

- Evaluate the use of single-board edge computers for edge projects where a large number of low-cost devices will be required to provide capabilities, such as data processing, anomaly detection, image recognition, voice recognition or AI inferencing capabilities.
- Choose single-board edge computers that can be rolled out rapidly, without skilled staff on-site, that can easily be managed and updated in the field.
- Build security into the system and evaluate potential vendors for security across all areas, including physical, data storage, communications, management and updates.
- Use systems with ASICs that support existing Internet of Things (IoT) and artificial intelligence (AI) frameworks such as NVIDIA VisionWorks or, Amazon SageMaker Neo when selecting a single-board edge computer.

Sample Vendors

Arduino; Digilent; NVIDIA; Raspberry Pi Foundation; Texas Instruments; VersaLogic

Gartner Recommended Reading

Predicts 2023: Edge Computing Delivery and Control Options Extend Functionality

Infographic: Understanding Edge Computing

Emerging Technologies Impact Radar: Edge AI

Market Guide for Edge Computing

Functions and Data at the CDN Edge

Analysis By: Simon Richard

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Content delivery network (CDN) providers continue to augment their service runtime environment offerings at their edge nodes. Functions and data at the CDN edge provide either stateful- or stateless-based processing environments that support the execution of user-provided code. Technologies available at CDN nodes include virtual machines (VMs), containers, JavaScript or WebAssembly isolated instances, key-value stores, object storage and SQLite databases.

Why This Is Important

CDN providers have been on a long, slow path to manage their investments in commodity networks, compute and services. Serverless and microservices architecture has gained traction as a way to build cloud-native applications and enable continuous delivery. As developers venture into geographically distributed apps requiring low latency and high performance, edge environments that support stateful processing are becoming another valuable offering. In some cases, CDN can replace a centralized PaaS.

Business Impact

Adoption of CDN and edge-based serverless architecture will:

- Lead to increased scalability, reduced costs and faster time to market.
- Help achieve the true benefits of cloud-native operations and create a more consistent and manageable environment for cloud operations.
- Help enable new edge-optimized use cases.
- Require adjustments to the organization’s practices and strategies around security, application development, infrastructure and operations, mandating closer coordination.

Drivers

- As many enterprises deploy applications and workloads that cannot tolerate latency delay (4K video, gaming and trading), consumers of CDNs will begin to look beyond content delivery and engage in locating processing and data storage for their apps in the CDN edge.
- Recent addition of key-value pairs, object storage and the emerging SQLlite databases adds state to the function offerings.
- Gartner sees typical edge use cases, including geographic fencing, A/B testing, personalization and stage migrations.
- Emerging PaaS solutions are using CDN serverless functions and data offerings to host their application components, and in some cases, the entire application.
- Vendors now support multiple languages, including JavaScript and Wasm.

Obstacles

Obstacles for adoption could include:

- Disparate technology stacks that might fail to deliver performance and low latency.
- Regional compliance and privacy considerations that preclude code or data from existing outside a specific location. This could also be a driver in some cases.
- A lack of observability, traceability, and testing and debugging best practices, and a lack of API and container registries at edge platform offerings, which will limit security, governance and DevOps support for APIs and services.
- Preset limits around compute or storage that limit use cases’ applicability.

User Recommendations

- Evaluate CDN developer edge services for use cases of extending cloud-native applications, microservices implementations or service integrations to achieve improved productivity and cost efficiency.
- Optimize or augment your application architecture by determining which component should run in the client side (including browser, mobile, IoT devices, etc.), the CDN edge or the origin.
- Avoid CDN edge delivery if the project requires fine-grained control over application infrastructure operations, or where cost estimates are excessive.
- Choose a CDN provider that has both edge functionality and developer integrations. Ensure that your provider shares developer tooling roadmaps and beta product invitations.

Sample Vendors

Akamai; Amazon Web Services (AWS); Cloudflare; Fastly; Fly.io; Vercel

Gartner Recommended Reading

CDN Providers Must Deliver Platforms, Not Just Content

Competitive Landscape: CDN and Edge Services

Neuromorphic Computing

Analysis By: Alan Priestley

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Neuromorphic computing is a technology that provides a mechanism to more accurately model the operation of a biological brain using digital or analog processing techniques. These designs typically use spiking neural networks (SNNs), rather than the deep neural networks (DNNs) of the current generations of AI technologies, feature non-von Neumann architectures and are characterized by simple processing elements, but very high interconnectivity.

Why This Is Important

Currently, most AI development leverages parallel processing designs based on GPUs. These are high-performance, but high-power-consuming, devices that are not applicable in many deployments. Neuromorphic computing utilizes asynchronous, event-based designs that have the potential to offer extremely low power operation. This makes them uniquely suitable for edge and endpoint devices, where their ability to support object and pattern recognition can enable image, audio and sensor analytics.

Business Impact

AI techniques are rapidly evolving, enabled by radically new computing designs.

- Today’s deep neural network (DNN) algorithms require the use of high-performance processing devices and vast amounts of data to train these systems, limiting scope of deployment.
- Neuromorphic computing designs can be implemented using low-power devices, bringing the potential to drive the reach of AI techniques out to the edge of the network, accelerating key tasks such as image and sound recognition.

Drivers

- Different design approaches are being taken to implement neuromorphic computing designs — large-scale devices for use in data centers, and smaller-scale devices for edge computing and endpoint designs. Both these paths leverage spiking neural networks (SNNs) to implement asynchronous designs that have the benefit of being extremely low power when compared with current DNN-based designs.
- Semiconductor vendors are developing chips that utilize SNNs to implement AI-based solutions.
- Neuromorphic computing architectures have the potential to deliver extreme performance for use cases such as DNNs and signal analysis at very low power.
- Neuromorphic systems can be trained using smaller datasets than DNNs, with the potential of in situ training.

Obstacles

- Accessibility: GPUs are more accessible and easier to program than neuromorphic computing. However, this could change when neuromorphic computing and the supporting ecosystems mature.
- Knowledge gaps: Programming neuromorphic computing will require new programming models, tools and training methodologies.
- Scalability: The complexity of interconnection challenges the ability of semiconductor manufacturers to create viable neuromorphic devices.
- Integration: Significant advances in architecture and implementation are required to compete with other DNN-based architectures. Rapid developments in DNN architectures may slow advances in neuromorphic computing, but there are likely to be major leaps forward in the next decade.

User Recommendations

- Prepare for future utilization as neuromorphic architectures have the potential to become viable over the next five years.
- Create a roadmap plan by identifying key applications that could benefit from neuromorphic computing.
- Partner with key industry leaders in neuromorphic computing to develop proof-of-concept projects.
- Identify new skill sets required to be nurtured for successful development of neuromorphic initiatives, and establish a set of business outcomes/expected value to set management's long-term expectations.

Sample Vendors

AnotherBrain; Applied Brain Research; BrainChip; GrAi Matter Labs; Intel; Natural Intelligence; SynSense

Gartner Recommended Reading

Emerging Technologies: Tech Innovators in Neuromorphic Computing

Emerging Technologies: Top Use Cases for Neuromorphic Computing

Forecast: AI Semiconductors, Worldwide, 2021-2027

Emerging Tech Impact Radar: Artificial Intelligence

Edge AI Hardware

Analysis By: Alan Priestley

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Edge artificial intelligence (AI) hardware comprises a wide range of systems, add-in boards and chips designed and optimized to execute deep neural network (DNN)-based data analytics applications within edge hardware infrastructure deployments.

Why This Is Important

Many edge computing and endpoint deployments utilize applications that leverage DNN-based algorithms to analyze complex datasets close to their point of capture (such as video, images, audio or sensor data). The processing required by these applications is often beyond that which can be easily supported by traditional microprocessors and microcontrollers. Consequently, a wide range of new edge AI hardware designs optimized specifically for these tasks are being developed.

Business Impact

Edge AI hardware brings significant business benefit in designs that:

- Require sophisticated AI or machine learning (ML)-based applications to be executed in low-power edge computing and endpoint devices.
- Need a local analysis of captured data.
- Demand low-latency decision making.
- Have limited, intermittent or costly connectivity to central data centers.
- Face governance or regulatory restrictions on the type of data that can be captured and stored.

Drivers

Edge AI hardware enables data analysis to be undertaken at, or close to, its point of capture.

- A key element of many system designs is being able to interpret this data and make decisions on the data content in a timely and efficient manner. However, data volume and complexity captured by edge computing and endpoint devices are increasing.

- For many deployments, it may not be practical to transfer this data to the cloud or a remote data center and utilize hosted analytics services. This is especially the case where low latency, interactivity, autonomy, privacy and security are required, or communication services may be intermittent or expensive.
- Edge AI hardware can be utilized in many deployments that would previously have required transferring the data to the cloud or remote data center in order to analyze. Example use cases include video surveillance, facial/gesture recognition, factory automation, voice response/control and monitoring a wide range of sensor data.
- Developments by semiconductor vendors are delivering chips optimized to efficiently execute DNN-based analytics. Managing this within constrained power and form factor has made it viable to deploy AI-based analytics into a wide range of edge computing and endpoint applications.
- Edge AI hardware complements the processing capabilities deployed in edge computing and endpoint devices freeing resources for local decision making and process control.

Obstacles

- To fit within the power and price constraints, edge computing and endpoint device deployments require optimized AI hardware and have the potential to impact overall analytics performance.
- The range of deployments where this equipment can be utilized may be limited, necessitating the use of cloud or data-center-based service to implement the desired analytics functions.
- Many DNN-based applications are developed within data centers using high-performance graphics processing unit (GPU)-based systems. This may limit the ability of these applications to be deployed on edge AI hardware.
- Even if the edge AI hardware is GPU-based, applications may still require optimization before they can be deployed in lower-performance non-data-center locations to ensure latency/responsiveness targets are met.

User Recommendations

- Determine the need to leverage edge AI hardware by assessing the types and complexity of data to be analyzed.
- Simplify development workflow by leveraging cloud-based services and toolsets.
- Assess the use of edge AI hardware for local analytics by taking into account the sophistication of remote decision making that can be implemented using edge AI hardware versus data-center-based analytics services.
- Assess the cost, availability and reliability of data communication to remote locations.
- Examine the regulatory and governance impacts of where data is stored and analyzed.

Sample Vendors

Advantech; Dell Technologies; Hailo; Hewlett Packard Enterprise (HPE); Intel; NVIDIA

Gartner Recommended Reading

Emerging Technologies: Tech Innovators in Neuromorphic Computing

Forecast: AI Semiconductors, Worldwide, 2021-2027

Emerging Tech Impact Radar: Edge AI

Emerging Tech Impact Radar: Edge Computing

Edge PaaS

Analysis By: David Wright

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Edge PaaS is a type of cloud-oriented application platform that is purpose-built for capacity-constrained environments at the network edge. Edge PaaS platforms are designed for highly distributed, latency-sensitive and edge-aware use cases, such as interactive gaming, network function virtualization, local device management and real-time analytics.

Why This Is Important

Edge PaaS makes distributed "edge cloud applications" viable for the enterprise by offering:

- Cloud-native runtime environments (e.g., containers or WebAssembly) for easy app deployment at the network edge.
- Optimized performance on edge hardware, enabling compute-intensive processing and analytics to execute closer to edge data sources.
- Cloud-based administration combined with node-level fault tolerance, giving IT ops teams centralized control with built-in distributed resilience.

Business Impact

Edge PaaS will bring the benefits of the cloud to IT environments outside the data center:

- Enterprises can implement applications at the network edge using standard cloud DevOps tools and practices.
- SaaS providers can deploy multitenant edge services that meet enterprise use cases in areas such as AR/VR, device control and real-time analytics.
- Corporate web developers can "escape the browser" by running low-latency edge JavaScript functions at their CDN providers' network points of presence.

Drivers

- The rise of 5G networking and distributed cloud computing is driving demand for "edge cloud applications" that are built with cloud-native tools but deployed at the network edge.
- Edge infrastructure products now support a wider variety of application use cases. Gartner's current analysis of the Edge Infrastructure market forecasts high growth in areas such as distributed business processing, device control, personal monitors and immersive experiences (see Forecast Analysis: Edge Hardware Infrastructure, Worldwide).
- Edge-ready hardware acceleration. Resource-efficient chipsets and function accelerator cards (FACs) are being packaged into edge servers, making it more economical to do processor-intensive work like encryption and AI inferencing at the edge.
- Kubernetes is moving to the edge. Cloud Native Computing Foundation (CNCF)-compatible initiatives, such as MicroK8S, K3S and KubeEdge will introduce the edge to a vast number of container-oriented developers and platform engineers.
- Serverless edge Functions as a Service (Edge FaaS). The ability to compile and execute JavaScript web functions at a CDN PoP (rather than in the browser or on a remote server) will drive the development of mobile and web-based user interfaces optimized for low-latency use cases.
- Distributed cloud computing. Public cloud providers are offering distributed infrastructure solutions capable of running edge cloud applications in a customer data center using locally deployed cloud servers managed by the provider.
- The gradual emergence of Web3. The metaverse, distributed ledgers and the Interplanetary File System (IPFS) are spurring development of a next generation of globally decentralized, peer-to-peer e-commerce and financial systems.
- The rise of Edge AI. Gartner's 2022 Emerging Tech Impact Radar: Edge AI identifies a range of AI-enabled edge technologies that are poised to broadly disrupt markets in the coming decade.

Obstacles

- Lack of coherent edge strategies in the enterprise. Many organizations have been slow to recognize the value of deploying services at the network edge and have yet to formulate a plan.
- Incomplete toolsets. While many standard DevOps tools can be used with Edge PaaS platforms, they lack integrated support for their "edge-native" capabilities.
- Insufficient flexibility. Many Edge PaaS offerings today do not include out-of-the box connectors to legacy enterprise systems or public cloud services, making integration more difficult.
- Potential fracturing of the Kubernetes standard at the edge.
- Standard CNCF-compliant K8S is not optimal for use cases where worker machines in a cluster are dispersed across an unreliable network, as many control plane services needed by nodes require continuous access to a cluster server. While several CNCF incubation projects exist to address this, a market leading alternative has not yet emerged.

User Recommendations

- Work with business stakeholders in your organization to examine their unmet needs at the network edge, and map Edge PaaS platform capabilities to these unmet needs.
- Distinguish opportunities for container-based Edge PaaS applications from lightweight "edge web" needs that would be best addressed by serverless Edge FaaS, and from IoT use cases best addressed by IoT-specific platforms.
- Many Edge PaaS offerings in the market are immature today. Select potential Edge PaaS or FaaS vendors for pilot projects based on a balanced mix of current capabilities and committed roadmap features.
- When building early Edge PaaS and FaaS prototypes, evaluate how well they integrate with your existing DevOps teams, tools and processes to ensure compatibility with your broader cloud strategy.
- Formulate an initial edge strategy around emerging disruptive technologies like Web3 and Edge AI by first determining how these innovations are likely to affect your customers and your market.

Sample Vendors

Akamai; Avassa; Cloudflare; EDJX; F5; Fastly; Section; StarlingX; Taubyte

Edge Stream Analytics

Analysis By: Paul DeBeasi

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge stream analytics is a technology that drastically reduces the time from data ingestion to business insight, compared with batch-oriented architectures. Prior to the emergence of edge stream analytics, data from industrial devices was first stored in repositories such as data lakes and historians, and then analyzed by business intelligence (BI) applications. Edge stream analytics inverts the traditional model of store first and then analyze, to analyze first and then store.

Why This Is Important

Digital transformation is distributing enterprise computing to the edge where customers, employees and enterprise assets are located. Gartner predicts that, by 2025, more than 50% of enterprise-managed data will be created and processed outside the data center or cloud. Edge stream analytics is a specific type of edge analytics that processes streams of data that emanate from physical and digital sources (e.g., connected products, industrial devices, retail systems).

Business Impact

Edge stream analytics provides situational awareness to improve operational decision making. For example, it can improve overall equipment effectiveness (OEE) by processing unstructured, high-volume, high-velocity, diverse, time-series data from industrial equipment with low latency. Edge stream analytics can operate when the edge is disconnected, thus supporting use cases that require the ability to operate autonomously from other edge, cloud or data center locations.

Drivers

- Proliferation of Industrie 4.0 initiatives that require a data-driven approach to manufacturing.
- Convergence of operational technology (OT) with information technology (IT).
- Use cases that require autonomous edge operation such as nuclear plants.
- Increasing need to produce real-time, data-driven business insight.
- Greater availability of stream analytics platforms designed for small footprints.
- Expanding deployment of edge computing technology.

Obstacles

- Lack of technical skills.
- Integration complexity with edge operational technology.
- Product diversity (e.g., conventional BI tools, stream platforms, open-source projects).
- Lack of edge standards and reference architectures.

User Recommendations

- Define specific use cases with measurable business outcomes that reflect the impact on processes, activities and decisions. Edge stream analytics initiatives fail when use cases are vague and desired outcomes are not measured.
- Develop an edge stream analytics proof of concept on a cloud platform before doing so at the edge. This approach will shorten your learning time because cloud stream analytics products are more mature than edge stream analytics products.
- Invest in training on how to develop, integrate and operate event stream analytics. Include the operation, engineering and IT teams.
- Define your service-level requirements and capacity constraints. Your design must have the capacity to ingest and process the streams.

Sample Vendors

Amazon Web Services; Confluent; Microsoft; SAP; SAS; Software AG

Gartner Recommended Reading

Innovation Insight for Streaming Data in Motion: The Collision of Messaging, Analytics and DBMS

5 Essential Practices for Real-Time Analytics

Essential Patterns for Event-Driven and Streaming Architectures

Design IoT Stream Analytics From Edge to Platform

Immersion Cooling

Analysis By: Jeffrey Hewitt, Philip Dawson

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Immersion cooling is a type of data center server cooling system that immerses server boards in a nonconductive heat transfer liquid, typically built using an immersion container in a dense, closed system. Immersion systems deliver well-above-average power efficiency, enabling compute systems to run at high performance while requiring less floor space.

Why This Is Important

Immersion cooling shifts power from cooling to computing, potentially doubling the compute density in power-constrained locations. It allows servers to operate in constrained environments such as 5G network control nodes and Internet of Things (IoT) edge servers, and significantly increases the efficiency of enterprise servers by reducing the need for air cooling.

Business Impact

Immersion cooling systems enable enterprises to deliver on sustainability and deploy higher levels of compute capability to strategic locations than is possible with conventional air-cooled racks. Key applications include data centers in facilities with limited space, factory automation, edge data centers and data centers in remote or unattended locations. Immersion cooling is well suited to the small remote data centers that will support 5G mmWave deployments.

Drivers

- **Immersion cooled systems are smaller, quieter and more efficient than traditional rack systems.** Their initial value will most likely come from outside the data center, where they enable higher compute density at higher energy efficiency and lower noise. Although the capital cost of the system is typically higher because of the mechanical and cooling infrastructure involved, there are environments where these systems outperform any alternative.

- **Immersion cooling can recover expensive server space and power costs.** For an enterprise constrained to operate servers in expensive spaces, there is often a floor power budget that covers both equipment and cooling. An immersion cooling system could improve power efficiency by 40%, theoretically, enabling 67% more energy for computing within the same power budget. These systems may also recover floor space, and are so quiet that they need no sound baffling.

- **Immersion cooling enables edge servers to operate in otherwise hostile locations.** Medium-scale edge computing nodes or wireless telecom nodes often operate under the thermal, spatial and power constraints of a remote server bunker, pole or closet. Shipboard or truck-based mobile data centers also benefit from these space and power efficiencies. For certain GPU-centric small-scale supercomputing tasks, these systems represent a practical on-premises solution. Isolation of the components also facilitates their use in locations with high levels of particulate pollutants like dust.

Obstacles

- **Data centers must be replumbed for immersion cooling.** Immersion cooling requires redundant plumbing for the warm water loop. It is most efficient when used with a passive heat exchanger.
- **Cooling fluids require special handling.** Immersion systems use vegetable oil or fluorocarbons. Oil-based systems require that staff handle and bag oil-coated boards for repair or replacement. This requires skills normally not possessed by IT administrators. Fluorocarbon systems operate with robotic, sealed pods as all fluids have to be recovered and contained.
- **Nonstandard compact server motherboards deliver the best economics.** To achieve the floor space reductions that liquid cooling offers, most systems require smaller motherboards as many systems are horizontal, rather than vertical. Vertical systems can use standard motherboards and are well suited to environments such as 5G closets, where the vertical form factor is a better fit.

User Recommendations

- **Evaluate immersion cooling for environments where power and space are expensive.** Immersion cooled systems are significantly smaller, quieter and more energy efficient than traditional data center racks. These systems will prove cost-effective in space- and power-constrained environments.
- **Fully comprehend how to communicate the use of vegetable oil in environments.** Despite marketing terms, the cooling fluids are primarily vegetable-derived and present manageable fire and mess risks. Understanding how to present and defuse these issues will be important in review meetings.
- **Plan to use immersion cooling for larger edge server deployments.** Edge systems inevitably face power, cooling and space constraints. Immersion cooling significantly eases cooling design by eliminating the need to pass air over components. Immersion cooling also improves reliability through lower overall operating temperatures, and exclusion of oxygen from contacts.

Sample Vendors

Asperitas; Green Revolution Cooling; Iceotope; Immersion Systems; QCooling; Submer; TMGcore

Gartner Recommended Reading

Predicts 2023: Edge Computing Delivery and Control Options Extend Functionality

Unlock the Business Benefits of Sustainable IT Infrastructure

Edge AI

Analysis By: Eric Goodness

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge AI refers to the use of AI techniques embedded in non-IT products, IoT endpoints, gateways and edge servers. It spans use cases for consumer, commercial and industrial applications, such as autonomous vehicles, enhanced capabilities of medical diagnostics and streaming video analytics. While predominantly focused on AI inference, more sophisticated systems may include a local training capability to provide optimization of the AI models at the edge.

Why This Is Important

Many edge computing use cases are latency-sensitive and data-intensive, and require an increasing amount of autonomy for local decision making. This creates a need for AI-based applications in a wide range of edge computing and endpoint solutions. Examples include real-time analysis of edge data for predictive maintenance and industrial control, inferences and decision support where connectivity is unreliable, or video analytics for real-time interpretation of video.

Business Impact

The business benefits of deploying edge AI include:

- Real-time data analysis and decision intelligence
- Improved operational efficiency, such as manufacturing visual inspection systems that identify defects, wasted motion, waiting, and over- or underproduction

- Enhanced customer experience, through feedback from AI embedded within products
- Connectivity cost reduction, with less data traffic between the edge and the cloud
- Persistent functions and solution availability, irrespective of network connectivity
- Reduced storage demand, as only prioritized data is passed on to core systems
- Preserved data privacy at the endpoint

Drivers

Overall, edge AI has benefited from improvements in the capabilities of AI. This includes:

- The maturation of machine learning operationalization (MLOps) and ModelOps tools and processes support ease of use across a broader set of features that span the broader MLOps functions. Initially, many companies came to market with a narrowcast focus on model compression.
- The improved performance of combined ML techniques and an associated increase in data availability (such as time-series data from industrial assets).

Business demand for new and improved outcomes solely achievable from the use of AI at the edge, which include:

- Reducing full-time equivalents with vision-based solutions used for surveillance or inspections.
- Improving manufacturing production quality by automating various processes.
- Optimizing operational processes across industries.
- New approaches to customer experience, such as personalization on mobile devices or changes in retail from edge-based smart check-out points of sale.

Additional drivers include:

- **Increasing number of users upgrading legacy systems and infrastructure in "brownfield" environments.** By using MLOps platforms, AI software can be hosted within an edge computer or a gateway (aggregation point) or embedded within a product with the requisite compute resources. An example of this is AI software deployed (TinyML) deployed to automotive or agricultural equipment to enhance asset monitoring and maintenance.

- **More manufacturers embedding AI in the endpoint as an element of product servitization.** In this architecture, the IoT endpoints, such as in automobiles, home appliances or commercial building infrastructure, are capable of running AI models to interpret data captured by the endpoint and drive some of the endpoints' functions. In this case, the AI is trained and updated on a central system and deployed to the IoT endpoint. Examples of the use of embedded (edge) AI are medical wearables, automated guided vehicles and other robotic products that possess some levels of intelligence and autonomy.

- **Rising demand for R&D in training decentralized AI models at the edge for adaptive AI.** These emerging solutions are driven by explicit needs such as privacy preservation or the requirement for machines and processes to run in disconnected (from the cloud) scenarios. Such models enable faster response to changes in the environment, and provide benefits in use cases such as responding to a rapidly evolving threat landscape in security operations.

Obstacles

- Edge AI is constrained by the application and design limitations of the equipment deployed; this includes form factor, power budget, data volume, decision latency, location and security requirements.
- Systems deploying AI techniques can be nondeterministic. This will impact applicability in certain use cases, especially where safety and security requirements are important.
- The autonomy of edge AI-enabled solutions, built on some ML and deep learning techniques, often presents questions of trust, especially where the inferences are not readily interpretable or explainable. As adaptive AI solutions increase, these issues will increase if initially identical models deployed to equivalent endpoints subsequently begin to evolve diverging behaviors.
- The lack of quality and sufficient data for training is a universal challenge across AI usage.
- Deep learning in neural networks is a compute-intensive task, often requiring the use of high-performance chips with corresponding high-power budgets. This can limit deployment locations, especially where small form factors and lower-power requirements are paramount.

User Recommendations

- Determine whether the use of edge AI provides adequate cost-benefit improvements, or whether traditional centralized data analytics and AI methodologies are adequate and scalable.
- Evaluate when to consider AI at the edge versus a centralized solution. Good candidates for edge AI are applications that have high communications costs, are sensitive to latency, require real-time responses or ingest high volumes of data at the edge.
- Assess the different technologies available to support edge AI and the viability of the vendors offering them. Many potential vendors are startups that may have interesting products but limited support capabilities.
- Use edge gateways and servers as the aggregation and filtering points to perform most of the edge AI and analytics functions. Make an exception for compute-intensive endpoints, where AI-based analytics can be performed on the devices themselves.

Sample Vendors

Akira AI; Edge Impulse; Falkonry; Imagimob; Litmus; MicroAI; Modzy; Octonion Group; Palantir

Gartner Recommended Reading

Building a Digital Future: Emergent AI Trends

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Edge Technologies Offer Strong Area of Opportunity — Adopter Survey Findings

Emerging Tech Impact Radar: Edge AI

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

Sliding into the Trough

Edge as a Service

Analysis By: Bob Gill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge as a service (EaaS) describes a model in which some or all edge software and/or hardware is offered via provider-owned and operated assets, requiring little to no ownership of infrastructure on the part of the customer.

Why This Is Important

A lack of standards and the diversity of edge devices and workloads make edge implementation challenging, while the potential for selecting a technical “dead end” is high. EaaS is a delivery model for edge computing in which a vendor (e.g., a systems integrator, independent software vendor or cloud provider) offers some or all of the infrastructure required to deliver edge-based applications, shielding the customer from technical complexity and market volatility.

Business Impact

Many enterprises choose to procure edge services managed by a third party, as:

- Most enterprises do not possess the skills or experience to build and manage complex, distributed systems that incorporate diverse end systems and software stacks.
- EaaS delivers and maintains prebuilt solutions with a primary focus on meeting SLAs that are based on business outcomes.
- EaaS simplifies adoption of complex edge initiatives, while lessening complexity, risk of obsolescence and technical debt incurred.

Drivers

- Organizations looking to deploy edge computing are finding that the crowded and rapidly shifting technology space is making platform selection difficult.
- Enterprises looking to limit risk find EaaS lowers the barrier to entry. Rather than making technology choices, they can contract for a business outcome (for example, retail store operations with explicit SLAs) and be insulated from the infrastructure that the provider uses to deliver the solution.
- EaaS solves for a lack of skills and experience in building and operating edge computing solutions.
- EaaS helps avoid incurring technical debt and reduces exposure to technology obsolescence.
- EaaS Lowers the barrier to implementation of edge computing and enables growth.

Obstacles

- Edge use cases are so individual that providers may not be in a position to solve all enterprise requirements economically.
- Ongoing operations at scale may be more costly than if the enterprise operated the infrastructure at a high degree of efficiency and automation.
- Placing all responsibility in the hands of the provider naturally drives vendor lock-in, and some EaaS offerings may not support future edge requirements.
- Sourcing a solution externally may limit integration with other internal applications and systems.

User Recommendations

- Evaluate edge as a service offerings by creating a build-versus-buy model for edge deployment and operations.
- Reduce initial cost outlays and pressure on accurate configuration sizing by positioning the edge capabilities as a more elastic service, rather than an explicit hardware configuration and purchase.
- Weigh enterprise needs for customization and differentiation against realistic assessments of in-house technical expertise, the organization’s stance on “opex versus capex,” and the breadth of the solution (a targeted, specific application set versus a more general, distributed infrastructure platform).

- Examine EaaS as a means to speed time to market for many use cases by lowering the initial cost outlays, technical hurdles, operational expertise required and “platform risk” present in such a nascent market.

Gartner Recommended Reading

Market Guide for Edge Computing

Edge Analytics

Analysis By: Peter Krensky

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Analytics is the discipline that applies logic (e.g., “rules”) and mathematics (“algorithms”) to data to provide insights that drive organization strategy and decision making. “Edge” analytics means that the analytics are executed in distributed devices, servers or gateways located outside of data centers and public cloud infrastructure closer to where the data and decisions of interest are created and executed.

Why This Is Important

Gartner client inquiries about the impact of edge on data and analytics continue to increase. With a growing relevance, by 2025, more than 50% of enterprise-managed data will be created and processed outside the data center or cloud. Demand for real-time decision making closer to where the data of interest is created and stored is one of many drivers for edge analytics.

Business Impact

The origins of edge analytics offerings were primarily in the support of decentralized deployments for device-isolated insights. However, connectivity advances, demands for cross-device analytics and innovations surrounding IoT have dramatically increased the scale and complexity of edge analytics use cases. Real-time event analytics and decision making, autonomous behavior of assets, and fault-tolerant applications hold tremendous potential value for enterprises in many industries.

Drivers

- Advantages of edge analytics include faster response times, reduced network bottlenecks, data filtering, reliability, increased access to data and reduced communications costs.
- Data sovereignty and governance issues related to sensitive/regulated data can constrain D&A teams from adopting centralized/cloud-based environments — moving data outside its originating geography can violate sovereignty regulations. By locating analytics in edge environments, the data remains in the originating locations, increasing the likelihood of compliance.
- The increase of distributed cloud and hyperconverged solutions from public cloud providers, including Amazon Web Services (AWS Outposts), Microsoft (Azure Stack Hub) and Google Cloud (Anthos), are further decentralizing previously cloud-restricted workloads. This perimeter expansion of the cloud brings compute and storage closer to the edge — creating new possibilities for edge-centric analytic workloads.
- 5G networks continue to grow in relevancy and, combined with mobile edge computing, will increase edge analytics use cases — particularly for latency-sensitive deployments.
- More analytics solutions, such as those supporting IoT use cases, need to operate in disconnected (or intermittently connected) scenarios. By bringing more powerful analytics capabilities to edge environments, these solutions need not rely on centralized data centers or cloud resources. As demand grows for “smarter” physical assets in many industries, supporting autonomous behavior will be a common requirement.

Obstacles

- Some of the disadvantages of edge analytics include increased complexity, lack of cross-device analytics, overhead of device maintenance and technical currency demands.
- Architectural design and development best practices for traditional or cloud-resident analytics typically assume or prioritize data/analytics centrality and do not carry over directly for edge analytics use cases.
- Vendor choices include two extremes in terms of provider scale — with early and unknown startups competing head-to-head with global megavendors. This drives a mix of platform/protocol standards and complicates going concern considerations for prospective buyers.
- Edge analytics can increase the complexity of enterprise standards and governance (data privacy, security, etc.), which has the potential to delay overall value realization objectives.

User Recommendations

Analytics leaders should consider edge analytics across the following five imperatives:

- Provide analytic insights for individual devices, assets or a larger distributed site even in the midst of disconnection from cloud or data center infrastructure and resources (e.g., driverless cars).
- Provide data sovereignty. Many regulations or data privacy laws require data be kept in the location of origin or the organization deems the transfer of data to introduce too many security vulnerabilities.
- Adapt to scenarios where network connectivity does not have the ability to support desired latency or stability requirements.
- Address scenarios where cross-device interdependencies serving as part of a larger system require edge-resident analytics.
- Redesign analytic strategies where it costs too much to upload the full volume of generated data and where there is no benefit to moving device-level data to a central location for aggregated analysis.

Sample Vendors

Amazon Web Services; Arundo; CloudPlugs; FogHorn; Microsoft; PTC; Samsara; TIBCO Software

Gartner Recommended Reading

Market Guide for Edge Computing

Innovation Insight for Edge AI

The Edge of the Edge Overview

Emerging Technologies Impact Radar: Edge AI

5G

Analysis By: Sylvain Fabre

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

5G is the fifth generation cellular technology standard by the 3rd Generation Partnership Project (3GPP). The standard targets maximum downlink and uplink throughputs of 20 Gbps and 10 Gbps, respectively. Latency is as low as 4 milliseconds in a mobile scenario and can be as low as 1 millisecond in ultra reliable low-latency communication scenarios, down to centimeter-level location accuracy indoors, and massive IoT scalability. New system architecture includes core slicing and wireless edge.

Why This Is Important

5G supports the 4th industrial revolution and IoT. Its fast and reliable real-time data transfer will benefit many industries. 5G supports eMBB, URLLC and MIoT — vital for enterprise transformation. 3GPP 5G standards releases deliver incremental functionality in: R15, extreme mobile broadband; R16, industrial IoT (massive IoT, slicing and security) — latest commercially available release; R17, MIMO enhancements, sidelink, DSS, IIoT/URLLC, bands up to 71GHz, nonterrestrial networks; and RedCap R18 is under definition with a planned freeze date in 1Q24.

Business Impact

- 5G enables three main technology deployments; each supports distinct new services for multiple industries and use cases of digital transformation, and possibly new business models (such as latency as a service). These are enhanced mobile broadband (eMBB) for HD video, mMTC for large IoT deployments, and URLLC for high-availability and very low-latency use cases, such as remote vehicle operations.
- Promising applications for 5G use include fixed wireless access, IoT support and private mobile networks.

Drivers

- Over 249 operators have rolled out 5G (see GSA), 30% of public mobile networks, and some form of 5G capability is penetrating lower cost smartphones in vendors' portfolios (with over nine versions of the technology depending on the band and the 3GPP release).
- Gartner estimates that 5G-capable handset penetration in 2025 will reach 54% worldwide, and 78% in Western Europe, with 5G-capable handset share of sales reaching 80% in 2023 in Western Europe from 51% in 2021. North America share will rise to close to 87%.
- 5G capability is starting to deliver value in emerging always-on wearables use cases.
- Increased data usage per user and device requires a more efficient infrastructure.
- Requirements from industrial users value 5G lower latency from ultra reliable and low-latency communications (URLLC) and expect 5G to outperform rivals in this area.
- Demand continues for massive machine-type communications (mMTC) to support scenarios of very dense deployments up to the 5G target of one million connected sensors per square kilometer. While diverse networks can offer adequate and cost-effective alternatives to 5G for many use cases (e.g., LPWA, NB-IoT, LoRa, Wi-SUN), overall total cost of ownership (TCO) and future proofness may not be as good.
- Availability has increased for industry-specific spectrum options (e.g., CBRS).
- Competitive pressures continue, for example, if one CSP launches 5G in the market others usually have to follow or risk losing market share — this includes both public as well as private 5G offerings.

Obstacles

- Issues with availability and cost of spectrum, in particular for industrial private networks, occur in some countries.
- Security concerns arise when using 5G in critical industrial scenarios.
- Availability and pricing of networks and modules for R16 and beyond solutions.
- Upgrade to 5G SA (stand-alone) core is needed for more advanced R16 releases (such as slicing), and commit to the continuous evolution of 5G releases over R17, R18 and beyond.
- Cost of radio network upgrades for 5G coverage and availability may require additional sites.
- Use of higher frequencies and massive capacity requires denser deployments with higher frequency reuse, which could raise network costs.
- Uncertainty exists about use cases and business models that may drive 5G for many CSPs, enterprises, and technology and service providers (TSPs).
- Feedback from some industrial clients mentioned that the majority of their use cases could be serviced by a 4G private network, Wi-Fi and/or NB-IoT, and other LPWA such as LoRa.

User Recommendations

- Enable R16 and above 5G for enterprise connectivity for mobile, nomadic and FWA secondary/tertiary use cases for branch location redundancy, as long as 5G is not the primary link for high-volume or mission-critical sites and unless there are no other options.
- Provide clear SLAs for network performance by testing installation quality for sufficient and consistent signal strength, signal-to-noise ratio, video experience, throughput and coverage for branch locations.
- Ensure backward compatibility to 4G devices and networks, so 5G devices can fall back to 4G infrastructure.
- Focus on architecture readiness — such as SDN, NFV, CSP edge computing and distributed cloud architectures, and end-to-end security — in preparation for 5G.
- Build an ecosystem of partners to target industry verticals more effectively with 5G before your competition.

Sample Vendors

Ericsson; Huawei; Mavenir; Nokia; Qualcomm; Rakuten Symphony; Samsung Electronics; ZTE

Gartner Recommended Reading

Emerging Tech: 5G mmWave at a Crossroads

Infographic: 5 Steps for Vendors to Scope and Run Successful POCs for Enterprise 5G PMNs

Invest Implications: Magic Quadrant for 5G Network Infrastructure for Communications Service Providers

Market Guide for 4G and 5G Private Mobile Networks

Quick Answer: What Vendor Product Leaders Need to Know About MWC Barcelona 2023

Edge-In to Cloud

Analysis By: Bob Gill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge-in to cloud describes an architecture where edge applications, servers and gateways make use of cloud-derived technologies and connect as needed to public cloud services — but they are deployed and can operate independently from the public cloud. Rather than using the programming models, or the IAM capability of a hyperscale cloud platform ("cloud out"), the application is centered around its role at the edge first, and its need for hyperscale cloud application services second.

Why This Is Important

Many edge solutions are designed to operate mostly independently of the cloud while maintaining connections as needed with one or more hyperscale cloud providers. These "edge-in to cloud" architectures leverage some cloud services, but use edge-specific functionality to support edge requirements, rather than pushing public cloud architecture to the edge as a complete platform. This is critical where cloud independence, multicloud core or predominantly "brownfield" infrastructures are present.

Business Impact

Making the decision whether to adopt a "cloud-out" or an "edge-in" model is fundamental to the set of services available, and the extensibility of the platform providing edge computing in the enterprise. This decision affects:

- Cloud independence
- Application and platform availability
- Technology stack availability
- Evaluations of whether an edge-centric approach may better align with a brownfield modernization effort

Drivers

- While the hyperscale cloud providers continue to grow their portfolio of cloud-to-edge solutions, enterprises report that such offerings are still functionally incomplete, not optimized for specific solutions, and poorly supported. Enterprise architects we speak to indicate that they want to avoid single-cloud vendor solutions and must reduce the chance for lock-in.
- Edge-centric approaches can provide solutions that are more tuned to edge requirements.
- They provide cloud independence for enterprises that are planning to be multicloud.
- They provide greater flexibility when dealing with legacy brownfield environments.
- They come with a greater likelihood of compatibility with existing vertically focused devices, application software and systems integrators.

Obstacles

- Edge-in architectures require implementation expertise beyond a normal IT skill set, leading to cost overruns and mistakes in implementation.
- The lack of “off the shelf” edge-in solutions is more likely to create a long-term dependence on a systems integrator or value-added reseller (VAR).
- Some edge-in platforms are generic and do not offer an application ecosystem, while others may be too focused on a specific vertical subset and the investment cannot be recouped across a range of use cases.
- Many edge-in platform vendors are early stage startups, providing increased risk due to inability to support projects at a very large scale. These vendors may lack the staying power to survive in a rapidly changing marketplace.
- Absence of common standards or frameworks for implementation will result in highly customized solutions that will be harder to extend or replace over time.

User Recommendations

- Use overall requirements to determine whether an application is ideally an edge extension of a cloud-based application, or an edge application that may use some cloud resources on the back end.
- Determine whether the edge and Internet of Things (IoT) portfolio will be used in conjunction with a single hyperscale cloud provider, or be multicloud.
- When adopting a newer edge-in platform, consider your need for long-term support and have a backup plan in place if your platform vendor does not thrive or is purchased by another vendor.
- Ensure architectural compatibility with your systems integrator (SI) or independent software vendor (ISV), as these solution- or vertical-specific providers deliver the actual business value of the effort.
- Ensure alignment with any edge-in or cloud-out preference they have.

Sample Vendors

LF Edge; Mirantis; Platform9; Rancher

Gartner Recommended Reading

Market Guide for Edge Computing

Hype Cycle for Edge Computing, 2022

Cool Vendors in Edge Computing

Infographic: Understanding Edge Computing

Edge Servers

Analysis By: Thomas Bittman

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Early mainstream

Definition:

Edge servers collect and deliver data, and perform analytics and inference close to IoT data producers (e.g., sensors and cameras) and data consumers (e.g., people and IoT actuators). They are often ruggedized for deployment outside of data centers, have broader and more general capabilities than gateway servers, but are less powerful than micro data centers.

Why This Is Important

As IoT and data produced by things grow at the edge, and as varied use cases at the edge increase, computing power is needed to aggregate and correlate this data, and turn many connected things into smart systems. Edge servers that can handle harsh environmental conditions and power limitations, with zero-touch remote management, will fill that requirement.

Business Impact

Edge servers improve the bottom line through increased plant automation, predictive maintenance, better efficiency and quality control. They improve the top line by enabling faster decision making for opportunities, more business interactions and better customer experiences. Whether owned by enterprises or acquired as a service, edge servers will become an important part of most enterprises' infrastructure topologies and digital business strategies.

Drivers

- Growing requirement for compute in locations where responses must be low-latency, in real-time or must continue in the event of an internet failure.
- Increasing data production at the edge (video, sensors, etc.) and the relative low cost of computing versus bandwidth.
- Increasing number of near-real-time digital interactions between people and things at the edge.
- Growing variety of use cases at the edge.

Obstacles

- Software that enables high-volume remote management with zero touch is immature.
- Scale requirements at the edge can be very small or very large and demand can grow quickly.
- Existing operational technology (OT) requirements, practices, ownership and culture.
- Large numbers of devices, widely geographically dispersed in remote locations can cause physical deployment and maintenance issues.

User Recommendations

- Choose edge servers that can be deployed rapidly and are easily flexible and extensible to match changing requirements.
- Evaluate edge servers for zero-touch remote management.
- Avoid hardware lock-in where possible, putting focus on applications, software platforms, management frameworks.
- Make security an upfront design requirement in any edge server deployment.
- Select as-a-service options rather than acquiring hardware and software to reduce capital expenses and to pay based on usage.

Sample Vendors

ADLINK Technology; Cisco Systems; Dell Technologies; Eurotech; Hewlett Packard Enterprise; Lenovo

Gartner Recommended Reading

Building an Edge Computing Strategy

Market Guide for Servers

Market Guide for Edge Computing

Low Code for Edge

Analysis By: Bob Gill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Low-code solutions for the edge allow simplified development of edge applications through the assembly of prebuilt blocks of code into a working application. The interface may offer a visual “drag and drop” approach, where I/O, functions and processes are represented by visual elements positioned and linked on a “canvas.” Code blocks, or “nodes,” may be created or modified through the use of lower-level languages for extensibility, while prebuilt apps or frameworks can assist the new user.

Why This Is Important

Edge computing applications are often complex, multiplatform, distributed applications made even more difficult by the diversity of devices at the edge and a lack of established standards. Edge-specific low-code solutions provide prebuilt and configured “chunks” of code, which are dragged, dropped and linked together based on business logic or data flows to quickly create functioning edge applications without knowledge of low-level interfaces or structures.

Business Impact

Low-code development for edge applications can provide simplicity, speed and agility to the development of edge applications. They:

- Break complex applications into prebuilt and preintegrated templates and “applets.”
- Enable business unit and operational users to create their own applications.
- Speed time to deployment of apps.
- Break the development bottleneck of waiting for IT and provide a self-service model.

Drivers

- The complexity of writing and maintaining distributed applications in rapidly changing operational settings.
- The need to mask the complexity of diverse devices, data formats and upstream targets.
- The complexity of integrating remote distributed applications with existing back-end processes and systems, often “owned” by other business units.
- IT-driven development may not possess the business-context and process expertise of the OT-based systems experts.
- The need to provide operational users the ability to exercise the infrastructure directly.
- The ability to continue to innovate after the SI’s initial building phase.

Obstacles

The current level of granularity and performance of resulting applications may uncover that low-code tools face the following challenges:

- There is a lack of standards or clear leaders.
- Low code’s coarse-grained approach may limit the extensibility of the solution.
- There is a lack of training in app development “hygiene” among nontechnical staff.
- Deeper evaluation of complex apps is needed, especially long-term scalability and total cost of ownership.
- Putting such powerful tools into the hands of business users may introduce governance issues, particularly surrounding the use of enterprise data.

User Recommendations

- Detail who will be using these tools and to what end. Governance is still required.
- Evaluate offerings that tend to specialize in the platforms you are implementing.
- Involve IT development for assistance in creating new application modules or scripting to modify existing modules.
- Plan to use low code as a user tool for data analysis and prototyping, but look to IT development or systems integrators to move prototypes to enterprisewide, scalable performant systems.

Sample Vendors

Node-RED; Tulip; Vantiq; Virtuoso

Edge Computer Vision

Analysis By: Sandeep Unni, Bob Gill

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Edge computer vision (CV) refers to applying artificial intelligence (AI) on data streams for real-time inference of operational and business intelligence pushed from core compute services to low-resource edge hardware. It includes numerous CV techniques like image, action and pattern recognition; object counting and classification; and facial authentication that run embedded within a non-IT asset, such as a device endpoint, a gateway device or on a local edge server.

Why This Is Important

CV is being deployed in a wide range of industries and applications such as surveillance, security, factory inspection systems, facial recognition, retail store experiences and autonomous vehicles. This creates a massive volume of data that must be analyzed and used for decision making. For many applications, the data analysis must be automated, and it is time-critical, requiring processing either close to or at the point and time of data capture.

Business Impact

Edge CV enables quick and consistent visual decision making at scale, and its business impacts include:

- Significant productivity increase from process automation of routine visual tasks.
- Improved safety by alerting to dangerous situations and ensuring use of protective equipment.
- Operational efficiency from inventory availability, reduced theft and other digitalization strategies in a retail store.
- Reduced asset downtime via preventative maintenance solutions.
- Higher production output through automated assembly automation.

Drivers

- Video and image usage at the edge has advanced significantly from simple surveillance and monitoring to include operational tasks, such as classifying objects for parts picking in manufacturing or smart check-out in a retail store.
- Edge processing solves network congestion, computing costs, privacy and security concerns, and the analysis delays of a disconnected or remote environment such as oil rigs or factories. Deployment of edge CV systems and devices is being further improved by developments in visual processing and graphics processing units and availability of synthetic data to bridge data availability and training.
- Enhancements in model compression and optimization techniques will enable larger models and workloads to run on resource-constrained edge hardware.
- Enterprise use of high-resolution cameras, with consequent higher data volumes, drives the need for high-performance edge CV systems to lower bandwidth and latency constraints.

Obstacles

- Integration of the edge CV applications with back-end infrastructure and software stack may need additional development effort.
- Deployments of CV capabilities on the camera endpoint may require the use of new hardware and technology due to power and form factor constraints.
- Lack of domain expertise and skills in adopter organizations slows down market adoption and momentum, although low-code/no-code offerings are being offered.
- Operational AI systems are important to ensure the life cycle management of models, continued update and scalable use, and maintaining model optimization.
- Hardware compute constraints and model size limitations in the industry today are impeding deployments at scale, though model compression and optimization techniques as well as advances in edge hardware (such as AI chips) are emerging.
- Many vendor solutions in the market are still use-case specific, limiting ability for adopter organizations to handle multivendor, multisolution scenarios.

User Recommendations

- Assess potential use cases for edge CV in your organization against measurable business value metrics they can provide.
- Evaluate applications that require real-time responses, generate significant data, have privacy or security limitations, or unreliable cloud connectivity as ideal for edge CV solutions.
- Tackle CV use cases in your organization where lack of sufficient training data has prevented CV deployments, by seeking out emerging generative AI capabilities in vendor roadmaps such as for synthetic data generation to aid faster model training.
- Track emerging developments and vendor landscape to ensure continuous improvement of your edge AI strategy.
- Work with your solution provider to determine optimal deployment footprint whether on the edge, or in a central data center on-premises or cloud.
- Select specialized edge CV offerings (focused by industry or use case) over generic solutions due to improved model performance and accuracy, as well as integration support.

Sample Vendors

Amazon Web Services; AMD (Xilinx); Chooch AI; Hikvision; Intel; Leda Technology; NVIDIA; Pilot AI

Gartner Recommended Reading

Emerging Technologies: Computer Vision Is Advancing to Be Smarter, More Actionable and on the Edge

Emerging Tech Impact Radar: Edge AI

Market Guide for Edge Computing

Innovation Insight for Edge AI

5 Top Practices of Successful Edge Computing Implementers

SASE

Analysis By: Neil MacDonald, Andrew Lerner

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Secure access service edge (SASE) delivers converged network and security capabilities, including SD-WAN, SWG, CASB, firewall and zero trust network access (ZTNA). SASE supports branch office, remote worker and on-premises secure access use cases. SASE is primarily delivered as a service and enables zero trust access based on the identity of the device or entity, combined with real-time context and security and compliance policies.

Why This Is Important

SASE is a key enabler of modern digital business transformation, including work from anywhere and the adoption of edge computing and cloud-delivered applications. It increases visibility, agility, performance, resilience and security. SASE also dramatically simplifies the delivery and operation of critical network and security services mainly via a cloud-delivered model. SASE reduces the number of vendors required for secure access to one or two explicitly partnered vendors.

Business Impact

SASE enables:

- Digital business use cases (such as branch office transformation and hybrid workforce enablement) with increased ease of use while reducing costs and complexity via vendor consolidation and dedicated circuit offload.
- Infrastructure and operations and security teams to deliver a rich set of networking and network security services in a consistent and integrated manner to support the needs of digital business transformation, edge computing and work from anywhere.

Drivers

- Digital business transformation including the adoption of cloud-based services by mobile workforces, edge computing and business continuity plans that must include a flexible, anywhere, anytime, secure, identity-based logical perimeter model of SASE.
- The need to flexibly support digital business transformation efforts with a zero trust security architecture while managing complexity is a significant factor for the adoption of SASE, primarily delivered as a cloud-based service.
- For IT, SASE can reduce the deployment time for new users, locations, applications and devices.
- For information security, SASE enables a single way to set policy enforcement consistently across all types of access — internet, web applications and private applications, reducing the attack surface and shortening remediation times.
- Enterprise desire to simplify network and network security deployments via the reduction of policy engines and management consoles.

Obstacles

- **Organizational silos, existing investments and skills gaps**: A full SASE implementation requires a coordinated and cohesive approach across security and networking teams, which is challenging given refresh/renewal cycles, silos and existing staff expertise.
- **Organizational bias and regulatory requirements for on-premises deployment**: Some customers have an aversion to the cloud and want to maintain control.
- **Global coverage:** SASE depends upon cloud delivery, and a vendor's cloud footprint may prevent deployments in certain geographies, such as China, Africa, South America and the Middle East.
- **SASE maturity:** SASE capabilities vary widely. Sensitive data visibility and control is often a high-priority capability, but it is difficult for many SASE vendors to address. While your preferred single vendor may lack the capabilities you require, two-vendor partnerships can be a viable approach.

User Recommendations

- Involve the security architect and network architect when evaluating offerings and roadmaps from the incumbent and emerging vendors to ensure an integrated approach.
- Leverage WAN, firewall, VPN hardware refresh cycles or software-defined WAN (SD-WAN) deployments to update network and network security architectures.
- Explore single-vendor SASE, dual-vendor SASE and managed SASE options when investing, but avoid deploying SASE with more than two vendors, regardless of vendor marketing for all core services to minimize complexity and improve performance.
- Use vendor combinations — when selecting a dual-vendor solution — that have explicit integration including turnkey automation and visibility, and ideally management and data plane integration.
- Combine branch office and remote access in a single implementation to ensure consistent policies and minimize the number of vendors required.
- Leverage branch office transformation and dedicated circuit offload projects to adopt SASE.

Sample Vendors

Cato Networks; Cisco Systems; Cloudflare; Forcepoint; Fortinet; Juniper Networks; Netskope; Palo Alto Networks; Versa Networks; Zscaler

Gartner Recommended Reading

2022 Strategic Roadmap for SASE Convergence

Market Guide for Single-Vendor SASE

The Future of Network Security Is in the Cloud

Magic Quadrant for SD-WAN

Magic Quadrant for Security Service Edge

Edge Security

Analysis By: Neil MacDonald

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Edge security offerings protect the integrity of the hardware, operating system, application platform, application workload, its data and its network communications at distributed edge computing locations.

Why This Is Important

Like any IT system, edge computing locations are targets for attack, including data theft, data poisoning, denial of service and placement of malware (for example, Bitcoin mining). Edge security combines the requirements of data center computing security with the scale, remoteness and heterogeneity of mobile and Internet of Things (IoT) computing security.

Business Impact

For edge computing strategies to succeed, the integrity of edge computing workloads and data must be protected. Without protection, edge computing capabilities could be rendered inaccessible and the data could be stolen, copied or tampered with, and adjacent edge nodes and devices could be attacked. Edge attacks can lead to potentially catastrophic results that could threaten health and safety if monitoring and control signals are lost, tampered with or spoofed.

Drivers

- Driven by network constraints and costs as well as privacy and compliance requirements, organizations are expanding workloads to the edge — and security needs to be a part of this.
- Sensitive data and intellectual property will be stored at — and transmitted from — the edge, so must be protected.
- New security solutions and approaches are needed that securely support intermittent access and protect from physical tampering and theft.
- The standardization of edge computing software stacks around containers and Kubernetes will allow cloud-native security concepts to be modified for the edge.
- Enabling remote access/automated control of systems that control industrial processes where failures can result in injury or loss of life requires a high level of security.
- Edge hardware upgrades and replacement cycles provide opportunities for edge security improvements.

Obstacles

- The historic diversity of hardware and application platforms makes it difficult to develop a single-edge security strategy.
- The market for edge security is emerging with a large number of overlapping offerings.
- Most edge hardware devices were not designed with adequate hardware security controls.
- Security wasn't a high priority during the procurement of most edge computing offerings.
- Most edge computing platforms won't have the capacity for a standard security stack.
- Complete edge security protection strategies must address the entire stack — hardware, OS, application platform, application, data and network security and their life cycle.

User Recommendations

- Design protection that treats the network as compromised, hostile and intermittent.
- Restrict all access (including admins) to/from the edge using a zero trust network access (ZTNA), secure access service edge (SASE) or security service edge (SSE) offering.
- Make tamper-resistance and hardware life cycle management a part of security control evaluation — assuming that hardware will be attacked, tampered with, stolen, or destroyed.
- Include at least mandatory data encryption, network security, workload integrity, application control, memory protection, behavioral monitoring and intrusion detection/prevention in an edge protection strategy and product evaluation.
- Favor offerings that are centrally managed — ideally cloud-based — and provide for tightly controlled administrative access.
- Require the use of identity-based policy management for edge equipment — ideally preprovisioned, certificate-based and stored in hardware.
- Require vendors to support Linux containers and container-based security offerings, which are expected to be widely used for distributed edge computing architectures.

Sample Vendors

Amazon Web Services (AWS); Appgate; Dell Technologies; Fortinet; Hewlett Packard Enterprise (HPE); Johnson Controls (Tempered Networks); Red Hat; StorMagic; VMware; ZEDEDA

Gartner Recommended Reading

Market Guide for Single-Vendor SASE

2022 Strategic Roadmap for SASE Convergence

Building an Edge Computing Strategy

Edge Asset Life Cycle Management

Analysis By: Thomas Bittman

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Edge asset life cycle management includes onboarding, monitoring, change management and decommissioning of edge computing systems. Edge computing systems can be intelligent gateways, edge servers and workstations deployed in remote offices or sites, or Internet of Things (IoT) systems, such as embedded devices or IoT servers. They may also be part of an operational technology (OT) vendor system.

Why This Is Important

Edge asset life cycle management software provides better control and visibility to the edge infrastructure. This can include edge servers, intelligent gateways or embedded devices. Edge asset life cycle management software ensures that infrastructure life cycle management policies that are applied in the data center or the cloud can be similarly applied to edge environments.

Business Impact

As enterprises build distributed IT architectures, infrastructure and operations (I&O) leaders must ensure they can centrally manage or support the business unit in managing the health and life cycle of these systems, regardless of location. System management software is crucial and when organizations begin to scale their edge projects and deploy large numbers of endpoints, gateways and edge servers, centralized management of these distributed systems is imperative. This ensures better visibility and control of the environment.

Drivers

- Edge environments, such as remote offices, are often managed by data center system management software, which is less capable of handling the scale and distribution of edge assets.
- Gateways provided by OT vendors typically lack consistent life cycle management capabilities.
- System management for IoT-centric edge environments is still nascent. In this case, IoT asset management software is usually provided by the hardware vendors as software as a service (SaaS) offerings.
- Cloud vendors and emerging asset management vendors now offer IoT platforms that support remote management of IoT gateways and embedded devices.

Obstacles

- Edge life cycle management requires agreement between the responsible operations teams (IT and OT) on how asset health is to be managed, and the types of asset updates that will be allowed.
- Although it is relatively easy to update remote assets through wired networks, over-the-air (OTA) updates via cellular networks are often challenging and may result in inconsistent results.
- Updating is subject to device availability, which may be restricted to production shutdown periods for OT systems.
- Updating low-power-constrained devices operating on low-power wide-area networks (LPWANs) is a challenge. In some cases, the cost of OTA updates on constrained devices may be higher than the cost of replacing the asset.
- Vendor-led asset management tools become untenable as heterogeneous edge assets grow in scale and volume.

User Recommendations

- Leverage separate edge life cycle platforms when the ownership of edge assets (e.g., the operations team versus IT infrastructure and operations teams) and policies for edge life cycle management are distinctly different.
- Choose platforms that support easier packaging and delivery of software updates (such as containers) to the edge system to manage IoT assets.
- Synchronize work with planned maintenance shutdowns to manage OT-related edge devices.
- Consolidate IoT functionality — analytics, visualization, application and device life cycle management — onto a common IoT platform.
- Ensure that updates are delivered in a secure manner and the base OS/firmware of the asset has a rollback option.
- Focus on the usability aspect of these platforms, which is equally important, particularly when the enterprise begins to scale from a small number of assets to thousands.

Sample Vendors

Amazon Web Services (AWS); CloudPlugs; Infiot; ioTium; Microsoft; SECO

Gartner Recommended Reading

Magic Quadrant for Global Industrial IoT Platform

Emerging Technologies and Trends Impact Radar: Internet of Things for Industrial Manufacturing

Climbing the Slope

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

IT/OT Integration

Analysis By: Kristian Steenstrup

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Information technology/operational technology (IT/OT) integration is the end state sought by organizations (most commonly, asset-intensive organizations), where IT and OT are treated as a seamless entity with cohesive authority and responsibility. In this state, there is an integrated process and information flow. Integration includes infrastructure, software, processes and potential resources.

Why This Is Important

The IT content of OT has grown exponentially and their integration allows for unprecedented efficiency gains. Yet, for most organizations, IT and OT are managed by separate groups with different approaches to managing technology. Integration can be initiated by IT departments. However, operational business units may also seek integration when trying to solve other challenges, such as dealing with cybersecurity, rising support costs, safety concerns, disaster recovery or software administration.

Business Impact

Opportunities and benefits from transparency, and an integrated value chain based on data, come from integrating the systems. As IT and OT platforms and technologies converge through increasing the use of IT architecture within OT, a successful digital business manages both IT and OT together. There is a shared responsibility, even though direct reporting lines may not shift. Data can be shared, and process flows become continuous and coherent, with minimal interruptions.

Drivers

- With IT/OT integration for asset-intensive digital businesses, organizations will be much more capable of managing, securing and exploiting data, information, and processes.
- IT/OT integration results in integrated systems, processes and teams of people, as technology domains with different areas of authority and responsibility come together.
- A common driver is for better reliability and maintenance strategies through more direct access to condition, and the use of on-premises data and SaaS solutions for plants and equipment.
- Integrated operational intelligence will provide better production management, quality control and responses to events in the supply chain, and more efficient production processes. The result will be a more agile and responsive organization.
- Digital twins, digital threads, product as a service and equipment as a service require remote Internet of Things (IoT) and OT data collection, and hence integration of IT and OT domains.
- The data from OT systems will be the fuel for better decision making in areas such as operations (adjusting and responding to production events), energy consumption, environmental sustainability, material consumption, and product quality, safety and reliability.
- A single data ownership and governance can be set up, resulting in clear end-to-end accountability for data owners.

Obstacles

- A lack of common governance structures due to a siloed approach to managing technology in the past has to be overcome.
- Without incentives, this will not change because historically, IT and OT had little contact and have different reporting lines.
- Completely integrated approaches to IT and OT are difficult to achieve because of the deeply rooted tradition in many businesses, where engineers and operations staff have been the “exclusive owners and operators” of OT.
- Many companies have disparate standards of technology in IT and OT, and even different standards for documenting the technologies, making initial planning difficult.
- A common data model spanning IT and OT rarely exists.
- Risk appetite across IT and OT, which is currently diverse, may have to be aligned.
- With the increased number of attacks in OT that have originated in IT, most stakeholders in OT will be cautious about “opening the door to ransomware” when integrating IT and OT.

User Recommendations

- Evaluate the IT/OT integration challenges and benefits in your specific company, and individual business units within the company.
- Achieve consensus across groups and with senior management, and create an alignment activity first to manage governance and standards. Sustainable integration needs well-planned IT/OT alignment.
- Add a more integrated approach to technology progressively. This integration should extend at least to data exchange and platform maintenance, with particular attention paid to communications, cybersecurity, and enterprise architecture. In some companies, that commonality will lead to an organization no longer delineated between IT and OT.
- Balance increased complexity and risk on the one hand, versus the potential benefits from better production management and more efficient production processes.
- Initiate IT/OT alignment discussions to arrive at common standards for platforms, security and architecture.

Sample Vendors

Accenture; Cisco; Eurotech; NTT DATA; PTC; Rockwell Automation

Gartner Recommended Reading

Quick Answer: What Are IT/OT Alignment and IT/OT Integration?

Manufacturing Insight: How to Position Hybrid IT/OT Offerings

How IT Standards Can Be Applied to OT

Survey Analysis: IT/OT Alignment and Integration

When Does a CIO Need to Be Involved in OT?

IIoT Gateways

Analysis By: Thomas Bittman

Benefit Rating: Moderate

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

An industrial Internet of Things (IIoT) gateway is a bridge between a field network and the IoT platform or, sometimes, a business application. IIoT gateways are data aggregation points for field devices or wireless networks. They provide local storage and some compute capabilities, as well as user interfaces for data processing and system management.

Why This Is Important

IIoT gateways provide protocol translation and optimization as well as data normalization, and they can be a critical point of data aggregation for industrial IoT projects. These gateways enable efficient transmission of data from endpoints to the IoT platform by significantly reducing bandwidth costs by normalizing and filtering most data at the edge.

Business Impact

IIoT gateways are viewed as a critical component of any industrial IoT project as these gateways have transitioned from systems that aggregate, translate, filter and forward data, to systems that can provide localized and near-real-time insights. Manufacturing, automotive and utility sectors are increasingly deploying IIoT gateways as part of the Industrie 4.0 project initiatives. IIoT gateways also play a critical role in building management and smart city initiatives.

Drivers

- The growth of IoT devices and data production at the edge requires gateways for connectivity and data processing — especially where low latency is a requirement.
- The variety of data formats and connectivity methods requires normalization and protocol translation and conversion.
- While more general-purpose edge servers will take on deeper analysis and inference roles, the larger demand for light data processing and local aggregation of IoT interactions are best handled by gateways.

Obstacles

- Most industrial gateway vendors offer in-house-developed gateway management utilities, with limited integration to Tier 1 IoT platforms, complicating deployment, change management and scalability.
- A myriad of edge network protocols and rapidly evolving cellular network technology warrant a modular gateway architecture. However, most IIoT gateways do not offer this, demanding frequent reinvestment and system redesign.
- Gateways in industrial settings are deployed in physically challenging and mission-critical environments and, therefore, need to meet stringent safety and security requirements.
- Users face challenges when technically integrating with existing PLCs and other legacy equipment, as well as organizational, cultural and process challenges working across the traditional IT/OT divide.

User Recommendations

- Create a roadmap for the current and future role of gateways in the IoT architecture, and invest in a combination of low-end IIoT gateways and intelligent IoT gateways to address various use cases.
- Standardize on gateways that can be managed and programmed by preferred IoT platforms and can provide software development kits for custom integration and management.
- Extend the usable life of IIoT gateways and reduce the need for new investments by choosing gateways that are modular and can accommodate new peripherals.
- Select IIoT gateways that are certified for international, as well as country- and industry-specific safety standards, and provide an adequate level of overall system security.

Sample Vendors

Advantech; Aspen Technology (AspenTech); Cisco; ClearBlade; Dell Technologies; Eurotech; Hewlett Packard Enterprise (HPE); Lantronix

Gartner Recommended Reading

Magic Quadrant for Global Industrial IoT Platforms

Emerging Technologies and Trends Impact Radar: Internet of Things for Industrial Manufacturing

2022 Strategic Roadmap for Edge (IoT) Networking

LPWA

Analysis By: Tim Zimmerman, Aapo Markkanen

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Low-power wide-area (LPWA) network technologies are a set of wireless wide-area network solutions designed for applications that are characterized by their long signal range, low cost and low throughput. Communications are typically over 5 miles/10 kilometers and involve small amounts of data at throughput of 0.3 Kbps to 50 Kbps between low-powered devices, such as sensors that have a battery life of several years.

Why This Is Important

Smart cities, utilities, agriculture, oil and gas, healthcare, and mining have use cases where devices in remote or hard-to-reach areas have a small amount of data that needs to be transmitted to an application. Sensors with integrated LPWA radios, which may be battery-operated or connected to power, are used in solutions that communicate data over vast distances.

Business Impact

LPWA network technologies allow low data sensors to communicate status or change in status, enabling organizations to act on or react to information that can improve operational efficiency, reduce cost, and reduce risk and liability. LPWA solutions also allow the elimination of human labor needed to collect the data from disparate sensors such as meters, pipes or machinery, while operating with relatively low power, often in remote environments. Similarly, the labor costs associated with battery replacements can be reduced considerably.

Drivers

- LPWA solutions provide the ability to collect information from widely disparate locations.
- LPWA solutions offer cost savings from applications such as metering or sensing, where personnel currently manually collect data from an end device, or service it in remote or challenging locations.
- LPWA can be used to replace existing 2G/3G deployments due to 2G/3G sunset in many countries.
- Third Generation Partnership Project (3GPP) Release 17 (Rel 17) defines technical specifications for integrated, supplemental coverage services (SCS) 5G over satellite. The combination of these technologies with 3GPP LPWA will accelerate transformation in remote areas for use cases like agriculture, mining, and oil and gas.

Obstacles

- The cost of proprietary radio infrastructure to communicate to the remote tags such as sensors or meters.
- Lack of interoperability between proprietary tags.
- Immaturity of standards for eSIM-related use cases.
- Some LPWA are non-IP based networks, which can limit the types of applications as well as the network tools that are used for monitoring and security.
- Lack of education, which has often paired LPWA with 5G technologies and therefore it will only be available when 5G is available.

User Recommendations

- Evaluate LPWA for sensor communications including smart metering, asset management or remote sensors used for monitoring applications such as agriculture, construction and utilities, if you are an organization with many geographically dispersed sensors.
- Require service providers to build end-to-end solutions including hardware, connectivity, security, protocol conversion and applications, including integration.

Sample Vendors

Actility; AT&T; Deutsche Telekom; Kerlink; Myriota; Orange; SAT4M2M; Semtech; Soracom; Verizon Communications

Gartner Recommended Reading

Magic Quadrant for Managed IoT Connectivity Services, Worldwide

Critical Capabilities for Managed IoT Connectivity Services, Worldwide

Market Guide for IoT Mobile Virtual Network Enablers

Appendixes

See the previous Hype Cycle: Hype Cycle for Edge Computing, 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 Edge Computing, 2022 - 12 July 2022

Hype Cycle for Edge Computing, 2021 - 4 August 2021

Hype Cycle for Edge Computing, 2020 - 31 July 2020

Hype Cycle for Edge Computing, 2019 - 9 August 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

Market Guide for Edge Computing

Hype Cycle for Edge Computing, 2022

Cool Vendors in Edge Computing

Table 1: Priority Matrix for Edge Computing, 2023

Benefit	Years to Mainstream Adoption			
↓	Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational		Edge Computer Vision Edge Computing Functions and Data at the CDN Edge IoT SASE	Neuromorphic Computing	

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
High	Edge AI Edge Servers	5G Cloud-Out to Edge Edge AI Hardware Edge AI Software Edge Analytics Edge as a Service Edge Asset Life Cycle Management Edge Data Management Edge-In to Cloud Edge Management and Orchestration Edge Stream Analytics IT/OT Integration Low Code for Edge Single-Board Edge Computers	Distributed Cloud Private 5G	
Moderate	IIoT Gateways LPWA	Edge PaaS Immersion Cooling	Edge Security Edge Storage Peer-to-Peer Edge	
Low				

Source: Gartner (July 2023)

Gartner

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)