Hype Cycle for Compute, 2023

Published 10 July 2023 - ID G00790907 - 83 min read

By Analyst(s): Tony Harvey, Jason Donham

Initiatives: I&O Platforms

Generative AI models will require new I/O technologies, such as Compute Express Link and function accelerator cards, to drive I/O performance. These will change how compute hardware is architected. I&O leaders can use this research to prioritize investment and optimize adoption of such technologies.

Additional Perspectives

- Summary Translation: Hype Cycle for Compute, 2023 (28 August 2023)

Strategic Planning Assumptions

By 2027, 75% of organizations will implement a data center infrastructure sustainability program to optimize costs and respond to pressure from stakeholders, up from less than 10% in 2023.

By 2027, 60% of new AI clusters will use function accelerator cards to manage and control network bandwidth, jitter and latency, up from 5% or less today.

By 2030, quantum computing as a service will be the predominant delivery mechanism for quantum computing technologies for over 75% of quantum computing users.

Analysis

What You Need to Know

The slowing of Moore's Law is driving a resurgence in hardware innovation. As CPU performance increases slow, accelerators, such as graphics processing units (GPUs) for generative artificial intelligence (AI) applications, are prompting changes in system and network architectures. These accelerators need interconnects like Compute Express Link (CXL) to accelerate access to memory and I/O. Additionally, the high-bandwidth, low-jitter, low tail latency networks needed by AI clusters will increase the need for function accelerator cards (FACs).

These new capabilities will all come with the cost of increased power consumption and cooling requirements. As a result, both direct-to-chip (DTC) liquid cooling and immersion cooling will become more prevalent, and server designs will need to adapt to these new technologies.

As energy costs increase and compute energy consumption grows, sustainability will become an even bigger issue. There will be a much stronger focus on delivering value from the energy consumed, and support for the circular economy will become more important.

For more information about how infrastructure and operations (I&O) leaders view the technologies aligned with this Hype Cycle, see 2021-2023 Emerging Technology Roadmap for Large Enterprises.

The Hype Cycle

This Hype Cycle describes the 25 most-hyped innovations in the compute market. For each technology, we define and analyze the benefit to enterprises, the current level of market penetration and the likely time it will take to reach the Plateau of Productivity. I&O leaders should use this research to determine whether and/or when to invest in these innovations.

New: Infrastructure orchestration is more than just infrastructure automation. It allows for the design, delivery and operation of services across on-premises, cloud and edge deployments. DTC liquid cooling and immersion cooling also appear on the Hype Cycle for the first time (see below).

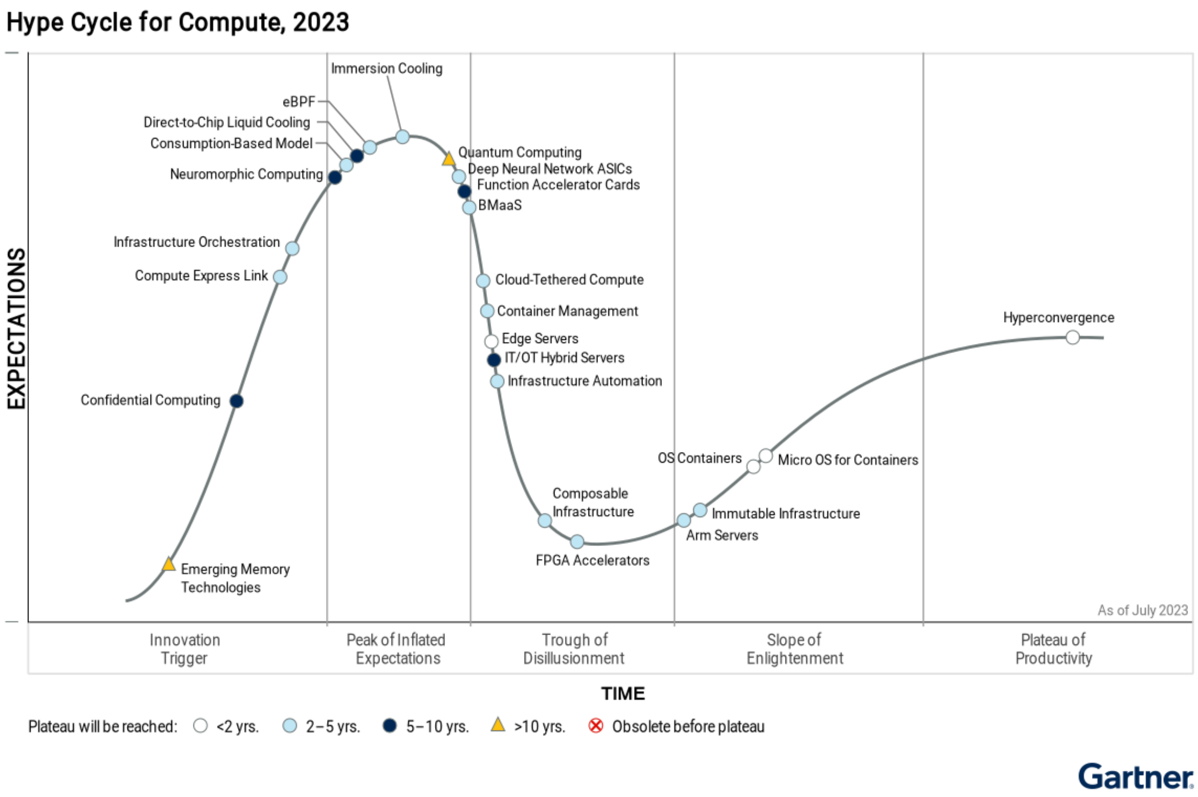
Peak Hype: DTC liquid cooling and immersion cooling are at the Peak of Inflated Expectations, with promises of sustainability and energy efficiency yet to match the reality of implementing these complex cooling systems. Quantum computing is still heavily hyped, but real-world use cases are in development.

Fast Movers: FACs (next-generation smart network interface cards/data processing units) and CXL are progressing rapidly, driven by the need for specialized accelerators to offload CPUs and the move beyond 100-Gbit Ethernet.

Slow Movers: Confidential computing has scarcely moved. The complexity of adapting applications in order to increase adoption remains a problem. The consumption-based model approach is also progressing slowly, as its purchasing complexity remains a barrier to users.

In the Trough: Composable infrastructure has fallen further into the Trough of Disillusionment as enterprises realize they are not using the composable features they have paid for. Arm servers have moved out of the trough as wider adoption occurs by hyperscalers and for AI and high-performance computing.

Figure 1: Hype Cycle for Compute, 2023



The Priority Matrix

The Priority Matrix maps the benefit rating for each technology or innovation against the amount of time the technology or innovation is likely to take to achieve mainstream adoption. This perspective can help users decide how to prioritize their server technology investments.

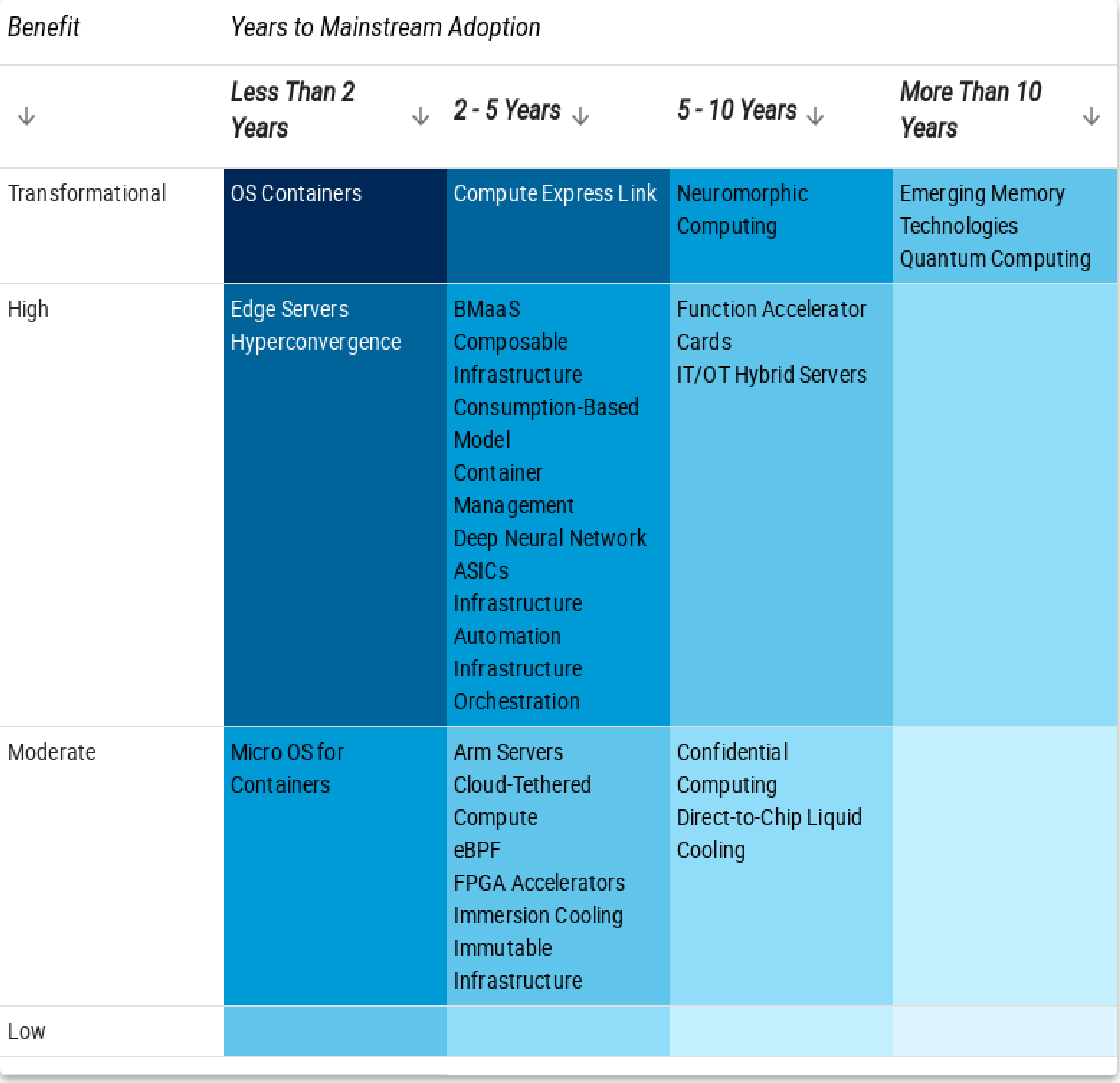
Transformational technologies, such as CXL, have the potential to affect how compute infrastructure is architected by disaggregating memory, allowing for memory tiering, and enabling new capabilities with cache-coherent AI and other accelerators. Quantum computing has huge potential to change how certain compute functions are performed, creating new possibilities to solve currently unsolvable problems. I&O leaders should start long-term planning for the impact of these technologies.

High-benefit entries such as bare metal as a service (BMaaS) and infrastructure orchestration offer new ways to deliver workloads. Additionally, FACs will enable networking and storage capabilities to be offloaded from CPUs and security functions to be enforced at the server level. Combined with moderate-impact technologies, such as Extended Berkeley Packet Filter (eBPF), this will enable new ways of using servers in high-performance networking environments.

Container management and, by implication, containers will continue to have significant benefits for application development and delivery, not only in the data center but out to the edge as well. This will encourage adoption of more moderately beneficial technologies such as micro operating systems (OSs) for containers.

Table 1: Priority Matrix for Compute, 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	OS Containers	Compute Express Link	Neuromorphic Computing	Emerging Memory Technologies Quantum Computing
High	Edge Servers Hyperconvergence	BMaaS Composable Infrastructure Consumption-Based Model Container Management Deep Neural Network ASICs Infrastructure Automation Infrastructure Orchestration	Function Accelerator Cards IT/OT Hybrid Servers	
Moderate	Micro OS for Containers	Arm Servers Cloud-Tethered Compute eBPF FPGA Accelerators Immersion Cooling Immutable Infrastructure	Confidential Computing Direct-to-Chip Liquid Cooling	
Low				

Source: Gartner (July 2023)

Off the Hype Cycle

With Intel stopping further development of Optane memory, persistent memory DIMMs no longer appear on the Hype Cycle.

Hardware-based security has reached maturity and has been removed from the Hype Cycle.

Serverless infrastructure has been removed to prevent duplication: It appears in Hype Cycle for Infrastructure Strategy, 2023.

On the Rise

Emerging Memory Technologies

Analysis By: Tony Harvey

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Emerging memory technologies, such as MRAM and ReRAM, are potential future replacements for DRAM in servers. They have the potential to provide higher density, lower power and persistence, but must show clear advantages over DRAM in order to succeed in the market.

Why This Is Important

Emerging memory technologies will be required to maintain future memory scalability growth. Both NAND flash memory and DRAM memory are reaching physical performance limits. The continued increase in flash memory layers (currently 232) demonstrates that scalability is no longer possible by shrinking cell sizes, and it will be increasingly difficult and costly to add more layers. NAND flash roadmaps show up to 1,000 layers, but 3D DRAM remains nascent with commercialization expected in three to six years.

Business Impact

Systems using emerging memory technologies enable scale-up computing systems to perform faster and handle larger analytics workloads. Alternatively, these systems can provide greater workload consolidation, reducing costs by shrinking the data center space and power required. In addition, a key impact of this technology will be to accelerate the adoption of in-memory computing architectures.

Drivers

There are three key technologies, although phase change memory is no longer being actively manufactured:

- Phase Change Memory — this technology is used in 3D XPoint memory from Intel and is no longer being actively manufactured.

- Magnetoresistive random-access memory (MRAM) — this is a type of nonvolatile memory (NVM) that uses the magnetic properties of controlled electrons to store data.
- Resistive random-access memory (ReRAM) — ReRAM is a type of nonvolatile memory in which the presence or absence of a conducting path within a dielectric between two electrodes is used to determine the stored state.

The advantages of these emerging memory technologies are:

- Resistance-based memory (RBM) technologies lend themselves well to scaling to finer process geometries, potentially leading to higher memory densities.
- RBM can accommodate multiple layers, another approach to increased density.
- As RBMs are persistent, they can support both system memory and storage applications.
- RBMs have lower power consumption and can simplify system design — they are persistent and do not need to be refreshed like DRAM.

Obstacles

In order to succeed when compared to DRAM or Flash, any emerging memory technology will need to overcome the following obstacles to move to market success:

- Performance must be similar to or faster than server DRAM.
- Cost per GB must be equal to, or significantly, lower than server DRAM.
- Power consumption must be equal to or lower than DRAM.
- System designers will need to be able to integrate DRAM and the Emerging Memory Technology in the same system with minimal changes required by the host.
- Use of the technology in a system must be transparent to applications.
- Must be available from multiple sources in modules that can be used interchangeably.

User Recommendations

- Engage with memory and server vendors by understanding their long-term plans for emerging memory in their server systems.
- Stress the need for application transparency and the need for a new technology to have significant advantage over DRAM to enable adoption by using the failure of Intel Optane.
- Ensure that vendors understand the need for a general purpose solution that, like DRAM, can be supplied by multiple vendors and used across a broad variety of CPUs and platforms.

Sample Vendors

Avalanche Technology; Everspin Technologies; Hewlett Packard Enterprise; IBM; KIOXIA; Micron Technology; Samsung Electronics; SK Hynix; Spin Memory; Western Digital

Gartner Recommended Reading

Emerging Tech Impact Radar: Compute and Storage

Emerging Technology Horizon for Semiconductors and Electronics, 2022

Emerging Tech Impact Radar: Semiconductor and Electronics Technologies

Emerging Tech: Compute Express Link Redefines Server Memory Architectures

2022 Strategic Roadmap for Compute Infrastructure

Confidential Computing

Analysis By: Mark Horvath, Bart Willemsen

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Confidential computing is a security mechanism that executes code in a hardware-based trusted execution environment (TEE), also called an enclave. Enclaves isolate and protect code and data from the host system (plus the host system's owners), and may also provide code integrity and attestation.

Why This Is Important

As privacy concerns and fines increase:

- Confidential computing combines a chip-level TEE with conventional key management and cryptographic protocols to enable unreadable computation. This enables a variety of projects where cooperation between different groups is critical, without sharing data or IP.
- The ongoing adoption of public cloud computing and the increased availability and viability of enclave technology allow data to be used in the cloud in a more trusted manner.
- Cross-border transfers are a complex, key component to many businesses, addressed directly by confidential computing.

Business Impact

Impacts include:

- Confidential computing may mitigate one of the major barriers to cloud adoption for highly regulated businesses, sensitive data workloads, or any organization concerned about unauthorized third-party access to data in use in the public cloud. This includes potential access by the infrastructure provider.
- Confidential computing allows a level of data confidentiality and privacy controls between competitors, data processors and data analysts that is very difficult to achieve with traditional cryptographic methods.

Drivers

- Cloud adoption is increasing alongside ongoing concerns regarding potential access to personal data by cloud service providers (CSPs).

- Global data residency restrictions are ongoing, with a need to segment content away from even the CSP with a level of independent assurance.

- Competitive concerns — not just around personal data, but also intellectual property — are spurring the adoption of confidential computing. This includes the need for confidentiality and protection against any third-party access, protection of the method of processing (including algorithmic functions) and protection of the data itself.

- Confidential computing has been mentioned as a viable protection mechanism by several authorities and standards bodies for these specific use cases. Correct implementation will help keep regulatory scrutiny at bay.

- Hyperscaler cloud providers are increasingly offering options that allow virtualized confidential computing, which allows apps to run without recoding or refactoring.

Obstacles

- Complexity of the tech and lack of trained staff or understanding of best implementation methods may hinder adoption and/or weaken deployment (e.g., key management/handling is done incorrectly, unaddressed side channel vulnerabilities).

- Trust is slow to build and quick to evaporate, especially when confidential computing is paired with occasional hardware vulnerabilities.

- Some forms of confidential computing are not usually plug-and-play, and are currently mostly reserved for high-risk use cases such as machine learning. Varying by vendor and technology, you may require a high level of effort but see only marginal security improvement over more pedestrian controls like Transport Layer Security (TLS), multifactor authentication (MFA), and customer-controlled key management services.

- Offerings directly from CSPs vary greatly in robustness, performance and reliability. Not all named confidential computing offers similar protection.

- Confidential computing that leverages cloud-native key management (KM) may be at risk of inadequate privacy because the CSP manages and has access to the keys. Therefore, using third-party KMaaS becomes more important.

- Confidential computing currently only integrates with the client's deployed technology. It is rare to find SaaS or BPaaS vendors offering integration to confidential computing — hence reducing protection choices.

User Recommendations

- Design (or duplicate) a sample application using one of the available abstraction mechanisms and deploy it into an instance with an enclave. Perform processing on datasets that represent the kinds and amounts of sensitive information you expect in real production workloads. This way, you can determine whether confidential computing affects application performance and seek ways to minimize negative results.
- Examine confidential computing for projects in which multiple parties, who might not necessarily trust each other, need to process (but not access) sensitive data in a way that all parties benefit from the common results. None of the parties should control the TEE in this scenario.
- Look for vendors that enable integration to a broader enterprise key management system and complementary encryption and PEC techniques.

Sample Vendors

Alibaba Cloud; Anjuna Security; Fortanix; Google; IBM; Intel; Microsoft; VectorZero

Gartner Recommended Reading

Three Critical Use Cases for Privacy-Enhancing Computation Techniques

Achieving Data Security Through Privacy-Enhanced Computation Techniques

Solution Criteria for Cloud Integrated IaaS and PaaS

Securing the Data and Advanced Analytics Pipeline

How to Make Cloud More Secure Than Your Own Data Center

Compute Express Link

Analysis By: Tony Harvey

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Compute Express Link (CXL) provides a set of protocols that run over the PCIe 5 or PCIe 6 buses. It is designed for the connection of CPUs, expansion memory, accelerators and I/O devices, both within the server chassis and between devices within a single rack. CXL technology maintains coherency between the CPU memory space and memory on attached accelerators or expanders, which allows resource sharing for higher performance, reduced software stack complexity and lowers overall system cost.

Why This Is Important

AI/ML and high core-count CPUs drive the need for more memory capacity and bandwidth. CXL enables increased memory capacity and bandwidth by allowing memory to be connected via the PCIe bus with minimal performance impact. Accelerators such as graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and data processing units (DPUs) need direct access to system memory to perform optimally. CXL provides necessary capabilities, such as cache-coherent shared memory access and memory semantics for I/O, allowing these capabilities.

Business Impact

- Enables new workloads to migrate to single and two socket systems due to increased memory capacity and bandwidth potentially reducing software licensing costs.
- Material performance benefits to HPC & AI/ML applications, enabling faster time to value.
- More effective usage of expensive accelerator devices as they can be shared across multiple systems.
- Reduced costs from higher utilization of expensive accelerators.
- New system designs with CXL 2.0, enabling future developments of tiered memory systems.

Drivers

- Industry momentum, driving CXL as the next generation of interconnect.
- The rise in workloads, such as AI/ML solid modeling, seismic analysis and advanced analytics, creating unprecedented demand for more memory, higher-performance I/O and accelerators.
- Need to better utilize expensive resources, such as DPU, FPGA and GPU accelerators by sharing across multiple systems.
- Tiered memory solutions that use a combination of CPU direct connected high-performance DRAM and CXL connected lower-cost DRAM that enables the use of existing low cost DDR4 memory with the latest CPUs.
- The power and cooling requirements of accelerators increase the need for system disaggregation which requires a cache coherent interconnect, such as CXL, that can expand beyond a single server.
- With CXL 2.0, the ability to disaggregate memory and CPU, enabling a truly composable system.
- Uptake in hyperscalers who can use their own CPU development to drive system disaggregation and memory expansion.

Obstacles

- Initial CPUs from AMD and Intel only support CXL 1.1. Support for CXL 2.0, with more capabilities for expanding memory, will follow in 2024.
- Requirements for new programming models have to be adopted and application redesign has to fully take advantage of the new capabilities. This will initially limit uptake to specific domains, such as AI/ML and HPC.
- Mainstream operating system and hypervisor vendors are only just starting to support CXL connected devices.

User Recommendations

- Identify the features in CXL that provide the most business benefit for your organization and get vendors to update you with their plans for support.
- Evaluate the use cases for 4-socket systems by looking at their needs for CPUs cores, memory capacity and bandwidth requirements and taking into account software licensing costs to identify if a single or dual-socket solution with CXL memory expansion would provide the necessary capabilities at a lower cost
- Work with vendors on their upcoming roadmaps for CXL by defining workloads and accelerators that will require CXL support.

Sample Vendors

AMD; Dell Technologies; Hewlett Packard Enterprise; Intel; Lenovo; Samsung Electronics

Gartner Recommended Reading

Emerging Tech: Compute Express Link Redefines Server Memory Architectures

Innovation Insight: Understand the Hype, Hope and Reality of Composable Infrastructure

2022 Strategic Roadmap for Compute Infrastructure

Emerging Technologies and Trends Impact Radar: Compute and Storage

Market Guide for Server Virtualization

Infrastructure Orchestration

Analysis By: Chris Saunderson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Infrastructure orchestration (IO) enables platform and I&O teams to design, deliver, operate and ensure orchestrated services across on-premises, cloud and edge deployments. IO enables templated service creation and management, spanning provisioning, Day-2 operations and integration with CI/CD, self-service portals, and API access to orchestrated services.

Why This Is Important

Infrastructure orchestration provides strategic workflow capabilities to drive life cycle delivery and ongoing maintenance of complex deployed infrastructure. These practices and tools enable agile, iterative automation delivery and execution of the processes required via self-service and API access. This investment improves the velocity and quality of infrastructure services, improves traceability and visibility of service delivery and reduces inconsistencies from manual activities.

Business Impact

Infrastructure orchestration drives consumer experience improvements of deploying and managing standardized infrastructure. I&O teams realize operational efficiencies through reduced manual efforts to deliver infrastructure, embedding security and compliance requirements into the delivered services, and offering cost optimization opportunities. I&O staff can transform their role into an automation-first focus and scale to meet increased business demands.

Drivers

- **Business agility:** Organizations must increase responsiveness to meet customer needs and adapt to market and technology changes. They must be able to deliver products that meet these changing demands and requirements quickly.

- **Cost optimization:** Infrastructure teams leverage orchestration to deliver scalable, reliable and secure platforms. This helps to improve delivery efficiency, reduce human work, and reduce downtime due to change failures.

- **Value extraction:** adoption of orchestration capabilities unlocks additional value from the automation tools already implemented, enabling incident response, request servicing and other tasks to be more richly automated and consumed.

- **DevOps:** Infrastructure orchestration is a key enabler of continuous software delivery, allowing the DevOps team to automate the provisioning and management of environments.

- **Infrastructure complexity:** Increasingly complex deployment topologies require greater automation to improve the consumability of infrastructure and the ongoing maintenance of deployments,

- **Security and compliance:** Increased automation enables the implementation of security and compliance controls through orchestration and avoids any audit failures. The end-to-end visibility and traceability of the provisioning and configuration can enable continuous compliance automation of the infrastructure.

Obstacles

- **Skill development:** Infrastructure orchestration practices and tools can be complex to implement and sustain, as they require skills beyond scripting to get maximum value. These tools leverage software engineering skills that can be challenging to find in I&O teams.

- **I&O operating models:** The organizational structure of many I&O teams is set up by domain specializations, making it hard to develop and deliver end-to-end services through orchestration. Perceptions of stability and reliability risks slow adoption.

- **Automation constraints:** To automate maintenance activities, a certain level of maturity needs to be reached within the organization. Orchestration requires that automated tasks be available to be able to realize maximum return on investment.

User Recommendations

- Identify and catalog use cases and constraints in your delivery workflows that are injecting delay into service delivery, especially for tasks that are executed manually.
- Benchmark existing service delivery execution time and quality problems to measure against to demonstrate improvement.
- Catalog operational tasks that are being executed manually today and are candidates to develop workflows to implement.
- Identify candidate orchestration platforms to execute proof of value testing with, ensuring that the candidates can be integrated into your existing operational environment.
- Monitor implementation to identify successes and opportunities for improvement and build a success story demonstrating velocity, quality, throughput and operational improvements.

Sample Vendors

Cloudsoft; Crossplane; Dell Technologies; env0; Itential; Morpheus Data; PagerDuty; Pliant; RackN; SpaceLift

Gartner Recommended Reading

Innovation Insight for Continuous Infrastructure Automation

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

Market Guide for Infrastructure Automation Tools

Market Guide for Continuous Compliance Automation Tools in DevOps

At the Peak

Consumption-Based Model

Analysis By: Jeff Vogel, Philip Dawson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

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

Business Impact

A consumption-based sourcing model and services strategy will:

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

Drivers

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

- The need for a more flexible cloud-like operating model for on-premises infrastructure.
- The massive growth of enterprise data that makes capacity planning difficult and upfront purchasing for three to five years of growth expensive and impractical.
- Prolonged procurement lead time increases due to persistent supply shortages.
- The need for an application-aware services delivery model.
- The preference for operating expenditure (opex) to capex with cloud-like benefits, while avoiding risks or costs associated with moving mission-critical workloads to the public cloud.
- The need for a more cost-effective, flexible and efficient sourcing strategy that aligns with business demands.
- The need to augment IT budget priorities to redirect investments to develop cloud-native platform skills that support business growth initiatives.
- The shift from exiting the life cycle management of infrastructure assets in the long term to freeing up IT resources.

Obstacles

A consumption-based sourcing model may:

- Be more expensive than capex financing.
- Be organizationally challenging to implement.
- Be unsuitable for IT operations that have a more stable and predictable growth and variability in forecast demand or lean toward sweating assets.
- Require minimum-usage commitment levels that can't be justified regardless of what is actually consumed.
- Require three- to five-year contracts with vendor-centric services.

- Lack the skills or culture alignment to shift from sourcing products to platform SLA services.
- Not take into account long-term supply chain price fluctuations during the contract period, when declining hardware costs or supply constraints are considered.
- Conflict with financial asset depreciation and amortization schedules or corporate balance sheet objectives.
- Conflict with established industry accounting standards and operational norms.
- Software licensing terms may be incompatible with the use of consumption based hardware.

User Recommendations

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

Sample Vendors

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

Gartner Recommended Reading

Market Guide for Consumption-Based Models for Data Center Infrastructure

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

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

Direct-to-Chip Liquid Cooling

Analysis By: Henrique Cecci

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Direct-to-chip (D2C) liquid cooling is a cooling technology that involves circulating coolant liquid over heat-generating components, such as CPUs, GPUs and memory modules, to draw off heat through cold plates or evaporation units. Compared to traditional air cooling, D2C liquid cooling is highly efficient in dissipating heat, potentially leading to increased efficiency, cost savings, improved reliability, space optimization and greater sustainability.

Why This Is Important

The latest CPUs and GPUs have much higher thermal density properties compared to older architectures. Moreover, server manufacturers are incorporating more CPUs and GPUs into each rack to keep up with the increasing demand for high-performance computing and AI applications. However, traditional air cooling systems are struggling to cope with the cooling requirements of these high-density racks in a sustainable and efficient manner but with D2C liquid cooling this is achievable.

Business Impact

Direct-to-chip (D2C) liquid cooling can bring multiple benefits to businesses, including improved energy efficiencies, additional cost savings, increased sustainability and improved reliability. By lowering operating temperatures, liquid cooling can boost computing systems' performance and energy efficiency, leading to significant cost savings in the long run. Additionally, it can optimize physical space utilization and enable the use of high-performance computing (HPC).

Drivers

- The growing demand for HPC and the increasing use of artificial intelligence (AI) and machine learning (ML) applications or high-density computing environments where traditional air cooling methods are insufficient.
- Requirements for lower energy consumption and associated costs.
- Data center space optimization and performance optimization.
- Lower noise level. Because liquid cooling systems don't rely on fans, they can operate more quietly than air-cooled systems.
- Sustainability and environment-friendly data center operations.

Obstacles

- High initial capital investment costs.
- Specialized expertise required to design and implement the system.
- Potential risks associated with leaks or system failures.
- Compatibility concerns with existing IT infrastructure.
- Regulatory requirements.
- Environmental impact concerns.
- Businesses may need time to realize the long-term cost savings of liquid cooling and overcome the early costs of using this technology.

User Recommendations

- Choose the right coolant by selecting one with good thermal conductivity, low viscosity and low electrical conductivity.
- Ensure proper flow rate. For example, maintaining the flow rate between 0.5 to 1.5 gallons per minute per kilowatt of heat.
- Maintain cleanliness and regularly replace the coolant, this is essential to maintain optimal performance and prevent any damage to the components.
- Take into account the design of the system, the cooling system should be designed to provide efficient cooling while minimizing the risk of leaks and damage to the electronics.

Sample Vendors

Asetek; Chilldyne; CoolIT Systems; Fujitsu; Huawei; Iceotope; JetCool Technologies; Rittal North America; Schneider Electric; STULZ

Gartner Recommended Reading

Market Guide for Servers

Emerging Tech Impact Radar: Compute and Storage

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

eBPF

Analysis By: Simon Richard

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Extended Berkeley Packet Filter (eBPF) is an enhancement to the Linux operating system kernel that allows specific instruction sets to run (sandboxed) inside the kernel. It allows companies to add features to Linux without changing kernel source code or requiring kernel modules.

Why This Is Important

eBPF increases the extensibility of Linux. It allows users to create hooks that are triggered by Linux kernel events. This offers a safer and simpler way to add capabilities, such as performance, security and visibility, in Linux. Technology vendors like ISVs and cloud providers use eBPF to avoid kernel-level modules, which carry inherent risks. eBPF is used in production at scale by hyperscalers, including AWS, Facebook and Netflix, and content delivery networks (CDN) such as Cloudflare.

Business Impact

eBPF improves observability, security and performance for applications. However, most enterprises will not use eBPF directly. Technology vendors do use eBPF as an underpinning technology in their products and services to improve the performance and safety of programs that run on Linux. eBPF allows extremely technically savvy organizations to safely and quickly make changes to Linux, compared to using alternative approaches, such as Linux kernel modules or upstreaming to the Linux distribution.

Drivers

- eBPF usage is driven by hyperscalers using it to deliver more efficient cloud offerings, as well as networking, monitoring and security vendors that use it in their products.

- Hyperscalers use eBPF to remediate kernel vulnerabilities without patching to address Day 0 vulnerabilities, and to more efficiently handle distributed denial of service (DDoS) attacks.

- Organizations are looking to accelerate the development speed of software that runs on Linux via avoidance of the requirement for upstream inclusion into the Linux distribution.

- Organizations are looking to improve the performance, security and monitoring capabilities of software running on Linux.

- eBPF is popular among technologically advanced companies, including technology vendors and hyperscalers, because it provides a standardized interface, supported kernel portability and requires less in-depth kernel programming knowledge.

- eBPF helps overcome scale and visibility limitations of iptables, which is the default networking stack in Linux. eBPF helps optimize and customize Linux network packet handling by processing them earlier in the cycle.

- Vendors are increasingly using eBPF in their career network infrastructure (CNI) software to improve performance, security and network visibility.

Obstacles

- While it is realistic for technology vendors and hyperscalers, most enterprises lack the expertise and skills necessary to build and integrate eBPF-based functions.
- Most enterprises do not have the awareness, need or risk tolerance to tackle Linux kernel challenges directly.
- Many older Linux kernels don't support eBPF, or only partially support the latest features.
- Security and system reliability concerns will severely limit what organizations are willing to deploy using eBPF, as poorly written eBPF programs can directly impact the operation of the Linux kernel.
- Integration challenges and backward compatibility with existing non-eBPF-enabled products.

User Recommendations

- Migrate to more modern platforms for organizations that are still using Linux distributions with limited or no eBPF support.
- Seek eBBF-based Kubernetes CNI solutions when scale, performance, visibility and security are top of mind.
- Use Linux variants that provide eBPF support to enable network performance, visibility and security products.
- Explore whether eBPF can meaningfully address the organization's performance or visibility challenges by supporting technologically advanced enterprises.
- Invest in eBPF to improve performance and visibility, to avoid falling behind competitors, for networking and network security vendors.

Sample Vendors

Aqua; Cloudflare; Isovalent; New Relic; Splunk; Sysdig; Tigera

Gartner Recommended Reading

Cool Vendors in Cloud Networking

Using Emerging Service Connectivity Technology to Optimize Microservice Application Networking

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

Quantum Computing

Analysis By: Chirag Dekate, Matthew Brisse

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Quantum computing is a type of nonclassical computing that operates on the quantum state of subatomic particles. These particles represent information as elements denoted as quantum bits (qubits). Qubits can be linked with other qubits, a property known as entanglement. Quantum algorithms manipulate linked qubits in their entangled state, a process that addresses problems with vast combinatorial complexity.

Why This Is Important

Quantum computing will not displace conventional computers. However, it will disrupt areas such as some classes of BQP (bounded-error, quantum, polynomial time) problem, quantum realistic simulations (used in material science, chemical simulations and drug discovery) and cryptography (security), where it will deliver results beyond what is feasible using classical techniques. Quantum computing could also advance the speed and/or quality of machine learning and optimization solutions.

Business Impact

With minimal investment required to investigate a broad range of quantum use cases, the potential rewards hugely outweigh the risks. Multiple use cases, such as optimization, run optimally on quantum computing system architectures. Also, the growing maturity of quantum ecosystems enables organizations to choose from a variety of quantum computing as a service (QCaaS) offerings. Enterprises need to plan for four key areas of impact: optimization, simulation, BQP and security.

Drivers

- Significant investments by governments, major corporations and startups amount, in aggregate, to more than $2 billion yearly.
- Enterprise and academic research teams have produced promising results for diverse use cases, including optimization and materials simulation, using current-generation noisy intermediate-scale quantum (NISQ) systems.
- Demonstrations of foundational quantum technology using electrons, ions, cold/neutral/helium atoms and photons are resulting in potential pathways to scalable quantum computing.
- The scale of superconducting gate-based quantum systems continues to increase, with some quantum computing vendors developing systems that scale to hundreds of qubits.
- Error correction algorithms and new methods such as error mitigation and error suppression are in development. These promise to make NISQ systems more usable.
- Managed service providers, including boutique quantum services companies, are partnering with enterprises to identify use cases and develop quantum algorithms.

Obstacles

- With few use cases guaranteeing an ROI, enterprises might deprioritize investments in quantum computing.
- Current, limited-scale qubit technology is too noisy and delivers returns of limited value.
- Standardization is lacking across programming, middleware and ecosystems.
- The market is highly fragmented, with over 600 startups operating in high-risk macroconditions. This exposes enterprises to innovation risk.
- Although small numbers of qubits can represent large amounts of data, quantum computers cannot convert large amounts of data to a quantum state, due to quantum RAM's immaturity.
- Unlike computing-on-silicon technology, there is no single physical computing stratum for quantum computing, and it is not possible to mix platforms at the quantum level. This results in a highly diverse range of potential platforms and in enterprises choosing platforms that might prove incompatible with future quantum computers.
- Enterprise leaders recognize that quantum computing will take more than 10 years to mature. This results in limited short-term investment.

User Recommendations

- Be frugal when it comes to investment in quantum computing. Focus on the problem you want to solve and ways to mature the quantum computing ecosystem. Quantum innovation is a long-term endeavor, so it is imperative to temper expectations.
- Create a pipeline for quantum computing talent by funding academic research projects that closely align with your use cases. When quantum computing becomes relevant to your organization, even a few quantum-capable employees will make a material difference.
- Plan for quantum-inspired classical optimization projects for skills development in areas such as warehouse routing, traffic routing, portfolio balancing and workforce planning.
- Plan for innovations in chemistry and materials science. Quantum computing has the potential to enable quantum-realistic simulations that could prove important in diverse fields, such as manufacturing, aerospace and defense.

Sample Vendors

Classiq; Google; IBM; Infleqtion; IonQ; IQM; PASQAL; Quandela; SandboxAQ; Zapata Computing

Gartner Recommended Reading

Cool Vendors in Quantum Computing

Infographic: How Use Cases Are Developed and Executed on a Quantum Computer

Preparing for the Quantum World With Crypto-Agility

Deep Neural Network ASICs

Analysis By: Alan Priestley

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

A deep neural network (DNN) application-specific integrated circuit (ASIC) is a purpose-specific chip designed to execute the computations utilized in a wide range of artificial intelligence (AI) applications. These chips can be deployed in either data center servers, edge computing systems or endpoint devices.

Why This Is Important

Many applications leverage DNN-based techniques to analyze captured data. These include object detection and classification in images and video streams, social media recommendation engines, autonomous vehicles, pharmaceutical analytics, and more recently, the large language models (LLMs) used in generative AI applications. To effectively execute many of these applications require the use of DNN ASICs optimized for specific workloads.

Business Impact

Leveraging DNN ASIC-based systems enables:

- Efficient analysis of high-volume complex datasets, such as videos, images and audio streams, enabling video analytics, object detection and classification, image recognition, LLM, and recommendation systems.
- Edge computers and endpoint devices capable of sophisticated local automated decision making and delivering enhanced user experience.
- Better performance and power efficiency, than solutions based on graphics processing units (GPUs) or general-purpose CPUs.

Drivers

- Increasing volume of complex unstructured data requires the use of processing techniques that leverage DNN models to analyze and enable business decisions to be made based on the data content.
- Executing DNN-based applications typically requires the use of computer systems that are capable of executing high volumes of highly parallel math operations.
- Many DNN models require training using large sets of known good data. GPUs can be used for this task but high-performance DNN ASICs designed for data center deployments can deliver a better solution to this problem.
- DNN ASICs can offer significantly better performance, at lower power, than many existing CPU- or GPU-based solutions available to execute DNN-based workloads.
- Often, trained DNN applications are deployed in locations, such as edge computing or endpoint devices, where power or form factor constraints prevent the use of many high-power AI devices. Many DNN ASICs are designed specifically for these deployments.

Obstacles

- Today, discrete GPUs are still the device of choice for many companies developing DNN-based AI applications.

- Most of the open-source software frameworks used by AI developers have native support for GPUs but require dedicated software tools and workflows to support DNN ASICs.

- Many companies developing DNN ASICs are startups, and while they often have the funding to develop a DNN ASIC, and supporting software, they lack the size to scale and grow their business, due to limited resources to support a broad range of AI developers.

- There is no standardization in DNN ASIC hardware design, with every vendor offering their own unique design and requiring specific software implementation to support each DNN ASIC.

- The large hyperscale cloud service providers are developing ASICs optimized for their specific DNN-based workloads, such as Google's Tensor Processing Units (TPUs) optimized for its TensorFlow-based applications and Amazon Web Services' (AWS') Trainium and Inferentia chips.

User Recommendations

- Use CPUs or cloud when DNN workloads are light enough to fit in conventional CPU-based infrastructure.

- Use GPUs or dedicated AI servers with DNN ASICs when DNN workloads would otherwise consume excessive server resources.

- Select DNN ASICs and vendors that offer or support the broadest set of DNN frameworks and toolsets.

- Specify edge computing and endpoint devices that integrate low-cost DNN ASICs to support edge inferencing and local decision making, in locations where power, form factor and communications cost are critical.

Sample Vendors

Amazon Web Services; Cerebras Systems; Google; Graphcore; Intel; NVIDIA; SambaNova Systems

Gartner Recommended Reading

Emerging Technologies: Tech Innovators in Neuromorphic Computing

Forecast: AI Semiconductors, Worldwide, 2021-2027

Forecast Analysis: AI Semiconductors, Worldwide

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Function Accelerator Cards

Analysis By: Anushree Verma

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Function accelerator cards (FACs) are a class of devices that have dedicated hardware accelerators with programmable processors to accelerate network, security and storage functions — known as DPUs/IPUs and/or SmartNICs. FACs improve data operations and services, server availability, and network performance and security, besides enabling connectivity to a network. They have onboard memory and peripheral interfaces, and can run independently.

Why This Is Important

FACs can improve server performance by up to 50%, via offloading functions such as virtual switching, security and application delivery controller (ADC). They can host dedicated network appliances, including firewalls. They can also improve security by placing security functions onto a securely booted, locked-down environment. Today, FACs are primarily adopted by hyperscalers and large cloud providers, and we estimate they will grow at a five-year CAGR of 65% through 2027.

Business Impact

FACs enable cost-efficient and energy-efficient data center environments, while improving performance. By offloading high overhead functions, they allow the server to host more workloads, which reduces the direct cost of additional servers and, in some cases, infrastructure software. In addition, they can facilitate data transmission between remote resources — primarily for HPC and artificial intelligence/machine learning (AI/ML) workloads.

Drivers

- Hyperscale cloud providers such as AWS, Microsoft Azure and Tencent, and other large cloud providers are using FACs today, and growing their implementation to achieve price/performance improvements.
- Vendors are aggressively marketing FACs, which are also referred to as data processing unit (DPU), infrastructure processing unit (IPU), SmartNICs, distributed services card (DSC) or programmable NICs.
- The rise of AI/ML workloads, solid modeling, seismic analysis and advanced analytics has created unprecedented demand on storage and network, resulting in latency and bandwidth issues.
- FACs can reduce the number of servers and hypervisor licenses by 10% to 30%, and may also decrease the number of application software licenses.
- Pulling security out of the server reduces the software-based surface area for attack.
- Telecommunication networks are moving toward virtualizing the network edge with 5G adoption, which leads to offloading 5G user plane function (UPF) and 5G network slicing to the FACs to achieve low latency and high throughput.
- FACs are increasingly bundled in high-performance solid-state storage systems to boost IOPS and minimize latency.
- FACs provide an alternative platform to host network appliances, such as firewalls and ADCs, with price/performance benefits in specific usage scenarios.
- Vendors with a large enterprise-installed base, including Hewlett Packard Enterprise (HPE) and VMware, have invested heavily in the technology and marketed it to organizations with specific usage scenarios until 2022. However, there has been a slowdown in the past few months.
- Increased consolidation in the market with AMD acquiring Xilinx and Pensando Systems, and Microsoft acquiring Fungible.

Obstacles

- Enterprises perceive FACs as a disruptive and dramatic departure from typical data center networking patterns, which limits adoption due to concerns over risk.
- There is confusion in the market due to vendors using different terminology, and providing different capabilities and architectures.
- Data plane programmability is high-risk, and has limited value and interest for enterprises.
- Hyperscale CSPs are able to justify the incremental price with the large-scale order and customization benefits they get by adopting FACs. However, enterprises are so far unable to do so, thereby hindering rapid adoption.
- Form factor and power consumption can impact rack, power and cooling budget, or occupy a full-size PCIe slot.
- Broadcom's pending acquisition of VMware creates uncertainty for potential buyers because Broadcom doesn't currently offer a FAC.

User Recommendations

- Use FACs for specific use cases, such as acceleration of NVMe-oF and AI/ML.
- Engage your existing data center infrastructure vendors on their plans for multivendor interoperability for offload on FACs, prior to your next server refresh.
- Investigate FACs to replace legacy components like physical firewalls and reduce the number of application licenses.
- Pilot FAC offerings to improve scale/security needs in the context of a large-scale data center network (1,000 switches), or to support extremely network sensitive workloads.
- Select FAC-based storage offerings, if you are an enterprise with applications that require microsecond latency performance when processing large datasets.
- Use a cross-functional team that includes networking, compute, storage and security personnel to evaluate FAC offerings.
- Focus on management and orchestration when evaluating FACs, as they are key differentiating factors.

Sample Vendors

AMD; Ethernity Networks; Intel; Kalray; Microsoft; Napatech; Nebulon; NVIDIA; Pliops; VMware

Gartner Recommended Reading

Emerging Technologies: Adoption Growth Insights — Function Accelerator Cards (Next-Gen SmartNICs, DPUs, IPUs)

Your Server Is Eating Your Network — Time to Rethink Data Center Network Architectures

Market Trends: Arm in the Data Center: Act Now to Develop Plans to Address This Shifting Market

BMaaS

Analysis By: Bob Gill, Philip Dawson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

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

Why This Is Important

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

Business Impact

- BMaaS offers the advantages of dedicated infrastructure (e.g., predictability, security and performance) with elasticity closer to IaaS than actual physical deployments.
- For example, it provides a cloudlike experience in a data center location better suited to customer needs for low network latency and data residency.
- BMaaS supplies a flexible integration platform at the nexus of public cloud access locations, such as colocation hubs or content delivery network (CDN) points of presence (POPs).

Drivers

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

Obstacles

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

User Recommendations

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

Sample Vendors

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

Gartner Recommended Reading

Break Down 3 Barriers to Cloud Migration

Sliding into the Trough

Cloud-Tethered Compute

Analysis By: Tony Harvey, David Wright

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Cloud-tethered compute is an approach to edge-in system management in which servers are designed to be deployed across a wide range of locations, but centrally administered from a vendor-provided console located in the public cloud. The cloud connection may be permanent or intermittent. It can deliver bare metal as a service (BMaaS), infrastructure as a service, PaaS or a combination of these solutions — typically, but not exclusively, in a subscription-based model.

Why This Is Important

Edge-in solutions and hybrid infrastructures that use both on-premises and cloud-based compute need to be managed from a single console that can deploy, update and monitor the infrastructure at scale. Cloud tethering provides an ideal way to do this, with a cloud-based management platform that provides easy connectivity and can scale as needed.

Business Impact

Cloud-tethered compute systems will affect businesses across IT, finance and procurement:

- IT teams will see a reduced need for local administration and maintenance, freeing them up for higher-value activities.
- New skills will be required in the business for contractual analysis, security and spend management for these systems.
- The IT and finance resource budget may cycle to a services-based delivery model.
- IT operations will be more flexible and aligned with business demands.

Drivers

- SaaS-based solutions are often easy to use and provide useful capabilities for managing devices at scale, especially when devices have intermittent connectivity.
- IT teams are being tasked with delivering differentiated IT services to the business. Avoiding local administration using a cloud-tethered compute system allows IT to focus on these higher-level services in a self-service automated fashion, without having to involve a traditional IT outsourcer.
- “Born in the cloud” companies that have no capability or desire to build on-premises solutions will find that cloud-tethered compute systems enable them to meet data sovereignty or latency requirements.
- There is a promise of “evergreen” technology refresh solutions that keep systems up-to-date with the latest technology, removing the need for IT to manage infrastructure refreshes.
- More realistic products that do not promise a complete cloud experience, but do promise a cloud-managed experience with access to cloud services.

Obstacles

- There is a risk of insufficient agility. Current three year agreements and fixed hardware investments are at odds with the dynamic and changing nature of the edge infrastructure markets.
- There are limitations on service availability, causing a misalignment with customer expectations. What services customers actually want and what the various providers are able to deliver have, to date, not matched up fully.
- There are differences in deployment models between IT-based solutions that deploy into data centers and cloud-tethered solutions that are being deployed into environments more traditionally associated with operational technologies.
- In many cases, supporting vendor expertise and maintenance of field solutions is new and untested.

User Recommendations

- Identify scenarios in which the tethered compute model provides clear business value versus a more-traditional IT solution.
- Ensure that field maintenance operations and SLAs are well-documented and understood.

- Use pilot programs to evaluate vendor capabilities and any necessary updates to I&O procedures, processes and skill sets.
- Organize a joint team that includes infrastructure and operations (I&O), operational technology (OT), vendor management and finance to evaluate all proposed cloud-tethered compute solutions.
- Assess the economics and requirements against a range of vendor solutions and consumption models. Each vendor will have very different capabilities.
- Ensure that contract terms and SLAs meet the requirements of the finance and IT teams, and that end-of-term options (or lack thereof) are fully understood.
- Clarify and document where the boundaries exist between the responsibilities of the supplier and those of the IT team. Elements such as data backup and application security are likely to be the end user's responsibility.

Sample Vendors

Avassa; EDJX; Hivecell; Microsoft; Pratexo; Spectro Cloud; Sunlight

Gartner Recommended Reading

Distributed Cloud: Does the Hype Live Up to Reality?

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

Market Guide For Edge Computing

Emerging Tech Impact Radar: Cloud Computing

Emerging Tech Impact Radar: Edge Computing

Container Management

Analysis By: Dennis Smith, Michael Warrilow

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

Gartner defines container management as offerings that enable the development and operation of containerized workloads. Delivery methods include cloud, managed service and software for containers running on-premises, in the public cloud and/or at the edge. Associated technologies include orchestration and scheduling, service discovery and registration, image registry, routing and networking, service catalog, management user interface, and APIs.

Why This Is Important

Container management automates the provisioning, operation and life cycle management of container images at scale. Centralized governance and security are used to manage container instances and associated resources. Container management supports the requirements of modern applications, including platform engineering, cloud management and continuous integration/continuous delivery (CI/CD) pipelines. Benefits include improved agility, elasticity and access to innovation.

Business Impact

Industry surveys and client interactions show that demand for containers continues to rise. This trend is due to application developers' and DevOps teams' preference for container runtimes, which use container packaging formats. Developers have progressed from leveraging containers on their desktops to needing environments that can run and operate containers at scale, introducing the need for container management.

Drivers

- The adoption of DevOps-based application development processes.
- The rise of cloud-native application architecture based on microservices.
- New system management approaches based on immutable infrastructure, which gives the ability to update systems frequently and reliably maintained in a "last known good state" rather than repeatedly patched.
- Cloud-based services built with replaceable and horizontally scalable components.
- A vibrant open-source ecosystem and competitive vendor market have culminated in a wide range of container management offerings. Many vendors enable management capabilities across hybrid cloud or multicloud environments. Container management software can run on-premises, in public infrastructure as a service (IaaS), or simultaneously in both.
- Container-related edge computing use cases have increased in industries that need to get compute and data closer to the activity (for example, telcos, manufacturing plants, etc.).
- AI/ML use cases have emerged over the past few years, leveraging the scalability capabilities of container orchestration.
- Cluster management tooling that enables the management of container nodes and clusters across different environments is increasingly in demand.
- All major public cloud service providers now offer on-premises container solutions.
- Independent software vendors (ISVs) are increasingly packaging their software for container management systems through container marketplaces.
- Some enterprises have scaled sophisticated deployments, and many more are planning container deployments. This trend is expected to increase as enterprises continue application modernization projects.

Obstacles

- More abstracted, serverless offerings may enable enterprises to forgo container management. These services embed container management in a manner that is transparent to the user.
- Third-party container management software faces huge competition in the container offerings from the public cloud providers, both with public cloud deployments and the extension of software to on-premises environments. These offerings are also challenged by ISVs that choose to craft open-source components with their software during the distribution process.
- Organizations that perform relatively little app development or make limited use of DevOps principles are served by SaaS, ISV and/or traditional application development packaging methods.

User Recommendations

- Determine if your organization is a good candidate for container management software adoption by weighing organizational goals of increased software velocity and immutable infrastructure, and its hybrid cloud requirements, against the effort required to operate third-party container management software.
- Leverage container management capabilities integrated into cloud IaaS and platform as a service (PaaS) providers' service offerings by experimenting with process and workflow changes that accommodate the incorporation of containers.
- Avoid using upstream open source (e.g., Kubernetes) directly unless the organization has adequate in-house expertise to support.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Google; IBM; Microsoft; Mirantis; Red Hat; SUSE; VMware

Gartner Recommended Reading

Market Guide for Container Management

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

IT/OT Hybrid Servers

Analysis By: Tony Harvey

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Information/operational technology (IT/OT) hybrid servers are edge devices that interface, collect and process data from OT systems that provide real-time control of physical systems and industrial processes. They are designed to operate with a higher resilience to shock, vibration, humidity and temperature than typical data center servers. Industrial communications interfaces — such as CAN bus, Modbus or Profinet protocols, as well as wireless or 5G technology — may also be included.

Why This Is Important

IT/OT hybrid servers allow the data created by OT systems to be processed in real time to optimize the process under control. Connections to IT networks allow hybrid IT/OT servers to collect and transmit data. This data can then be used for training AI/ML models to deliver further efficiencies and provide insight into manufacturing and production capacity and scheduling.

Business Impact

IT/OT hybrid servers help enterprises realize the potential of the large data pool that is generated by OT systems. The ability to use this data will generate new cost efficiencies and innovations in manufacturing and industrial control processes. Enterprises that successfully integrate IT/OT hybrid servers into their digital transformation strategy will lower their costs and deliver new services to market faster. Enterprises that do not adopt them, however, may find themselves left behind.

Drivers

- Businesses need real-time analysis and decision making based on capturing data that allows the optimization of industrial processes and assets to reduce costs and increase quality.

- By using near-real-time reporting of manufacturing, operations and production data, businesses will be able to provide more predictability in order cycles and a better usage of components.

- Equipment breakdowns can cause line stoppages, which drive manufacturing costs up. By enabling the collection and analysis of device monitoring, IT/OT servers enable predictive maintenance to prevent these issues.

- Regulatory and compliance requirements mandate that certain datasets should be processed and stored at edge locations, which requires the deployment of appropriate systems on-site. Further, latency and bandwidth limitations at these sites further stress the need for on-site systems.

- Organizations are collecting OT data to enable AI/ML training and digital-twin-model building.

- There is a need for specialized servers that can meet the environmental requirements for industrial sites.

Obstacles

- Industrial enterprises are cautious about the security risk of using IT and network connectivity systems in industrial process control, where failure could result in loss of life or significant property damage.
- IT and OT are separate groups with different cultures and different risk perceptions. The differences between these groups must be managed for any successful implementation.
- Businesses grapple with the complexity of defining what data must stay at the edge versus what data should be transmitted to and subsequently processed in the cloud.
- Budgeting for IT/OT hybrid servers can be difficult because there is an overlap between OT and IT systems.
- Management solutions designed to operate at large scale across a wide geographic range with highly variable connectivity characteristics are very immature.
- Standard IT equipment will not meet the harsh environmental requirements of industrial locations. Further, there could also be issues with electronic noise and interference.

User Recommendations

- Create an integrated IT/OT group that has full responsibility for these solutions, reducing the disconnects related to technology, management and budgeting.
- Reduce the risk of conflicts between the teams by aligning the IT & OT groups across architecture, governance, security and software management, and infrastructure, support and software acquisition.
- Develop a blended IT/OT culture that mixes the rigor and risk awareness of the OT engineering mindset with the flexibility and tolerance for change that is inherent in an IT mindset.
- Embed safety, security and risk training, foster awareness and include talent in hybrid IT/OT teams to ensure that systems are designed with safety and security in mind.
- Remove budget conflicts by defining upfront the budget sources for ongoing support, maintenance and dependencies across the entire combined IT/OT environment.

Sample Vendors

Dell Technologies; Hewlett Packard Enterprise; Lenovo; Schneider Electric

Gartner Recommended Reading

As IT and OT Converge, IT and Engineers Should Learn From Each Other

Survey Analysis: IT/OT Alignment and Integration

When Does a CIO Need to Be Involved in OT?

2022 Strategic Roadmap for IT/OT Alignment

How IT Standards Can Be Applied to OT

Infrastructure Automation

Analysis By: Chris Saunderson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Definition:

Infrastructure automation (IA) enables DevOps and infrastructure and operations (I&O) teams to deliver automated infrastructure services across on-premises and cloud environments. This includes the life cycle of services through creation, configuration, operation and retirement. These infrastructure services are then made available through platform delivery, self-service catalogs, direct invocation and API integrations.

Why This Is Important

IA delivers velocity, quality, efficiency and reliability, with scalable, declarative approaches for deploying and managing infrastructure. These tools integrate into delivery pipelines targeting deployment topologies that range from on-premises to the cloud, and enable infrastructure consumers to build what is needed when they need it. Once deployed, IA provides day-2 and beyond operational automation, and extends to provide policy compliance and enforcement capabilities.

Business Impact

Implementing and maturing IA services will enable:

- **Agility** — continuous infrastructure delivery and operations
- **Productivity** — version-controlled, declarative, repeatable, efficient deployments
- **Cost improvement** — reductions in manual effort expended via increased automation
- **Risk mitigation** — compliance driven by standardized configurations
- **Collaboration** — delivering environments that product teams need with security, cost and compliance requirements baked in.

Drivers

I&O leaders must automate delivery through tool and skills investments to mature beyond simple deployments. The target should be standardized platforms that deliver the systemic, transparent management of platform deployments. This same discipline must be applied to the operation of these deployed platforms, ensuring that efficient operations (including automated incident response) can be achieved. IA tools deliver the following key capabilities to support this maturation:

- Multicloud/hybrid cloud infrastructure delivery
- Support for immutable and programmable infrastructures
- Predictable delivery enabling automated operations
- Self-service and on-demand environment creation
- Integration into DevOps initiatives (continuous integration/delivery/deployment)
- Resource provisioning, including cost optimization capabilities
- Operational configuration management efficiencies
- Policy-based delivery and assessment/enforcement of deployments against internal and external policy requirements
- Enterprise-level framework to enable maturing of automation strategies

- Skills and practice development inside infrastructure teams, enabling agile and iterative development and sustaining of services

Obstacles

- The combination of tools needed to deliver IA capability can increase tool count and complexity.
- Software engineering skills and practices are required to get maximum value from tool investments.
- IA vendor capability expansion overlaps and confuses the tool landscape, resulting in over-investment.
- Steep learning curves can cause developers and administrators to revert to familiar scripting methods to deliver required capabilities.

User Recommendations

- Identify existing IA tools in use to catalog capabilities, identify use cases and document overlaps to aid decision making.
- Assess existing internal IT skills to incorporate training needs that more fully enable IA, especially for an automation architect role to coordinate standards development and implementation.
- Baseline how managed systems and tooling will be consumed (e.g., engineer, self-service catalog, API or on-demand).
- Integrate security and compliance requirements into scope for automation and delivery activities.
- Develop an IA tooling strategy that incorporates current needs and near-term roadmap evolution.

Sample Vendors

Amazon Web Services; HashiCorp; Microsoft; Perforce; Pliant; Progress; Pulumi; RackN; Upbound; VMware

Gartner Recommended Reading

Market Guide for Infrastructure Automation Tools

Innovation Insight for Continuous Infrastructure Automation

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

How to Start and Scale Your Platform Engineering Team

Composable Infrastructure

Analysis By: Tony Harvey, Paul Delory, Philip Dawson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

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

Business Impact

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

Drivers

- Compute Express Link (CXL) provides the necessary capabilities to disaggregate and pool memory and I/O as well as providing a standardized set of APIs to manage the disaggregated hardware.
- Hyperscale cloud vendors are moving toward composable designs utilizing CXL to increase hardware utilization and reduce the costs of stranded hardware.
- Test and development environments benefit from composability, where infrastructure with varying characteristics must be repeatedly deployed, deconstructed and redeployed.
- Multitenant environments benefit from composable infrastructure by allowing a pool of hardware to be dynamically configured, assigned, reconfigured and reassigned based on tenant requirements.

Obstacles

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

User Recommendations

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

Sample Vendors

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

Gartner Recommended Reading

Market Guide for Servers

Emerging Tech: Compute Express Link Redefines Server Memory Architectures

Emerging Tech Impact Radar: Compute and Storage

2022 Strategic Roadmap for Compute Infrastructure

FPGA Accelerators

Analysis By: Alan Priestley

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Field-programmable gate array (FPGA) accelerators are server-based, reconfigurable computing accelerators that deliver extremely high performance by enabling programmable hardware-level application and function acceleration.

Why This Is Important

AI workloads require processing of high volumes of massively parallel data. While a traditional CPU can handle this task, it is not very efficient. So, for many applications, it is better to use a chip designed specifically for this type of processing. While not originally designed for this task, FPGAs have large numbers of logic units that can be configured and interconnected to support the processing of highly parallel datasets, applying math operations to multiple data points in parallel.

Business Impact

FPGAs can deliver performance and power efficiency for a range of workloads:

- AI inference workloads that require energy-efficient, low-precision (8-bit and 16-bit) integer processing capabilities.

- Applications such as genome sequencing, real-time trading, video processing and AI inference in edge computing systems.

- 100 Gbps and faster networking solutions. FPGA SmartNICs can be used to offload the execution of network protocol stack from the CPU.

Drivers

- FPGAs feature a large array of programmable logic blocks, reconfigurable interconnects and memory subsystems that can be configured to accelerate specific algorithmic functions. This can be used to offload a range of specialized processing tasks from the main system processor.

- In data centers, FPGAs can be used in a range of use cases that require applying consistent processing operations to large volumes of data, such as high-frequency trading (HFT), hyperscale search, video analytics and DNA sequencing. For example, Microsoft is leveraging FPGAs for search analytics and networks, and Illumina's FPGA-based DRAGEN Bio-IT Platform enables high-performance genome-sequencing workflows.

- Major FPGA vendors, such as AMD (which acquired Xilinx) and Intel, along with a number of startups, such as Mipsology, are working to address FPGA programming challenges, with libraries and toolsets that enable FPGAs to be configured using software-centric programming models.

- Software frameworks (such as OpenCL) and programming environments (such as AMD Vitis and Intel's oneAPI) that lower the time and skills required to use FPGAs are enabling accelerated adoption of FPGAs.

- Major cloud service providers — such as Amazon Web Services (AWS), Baidu and Microsoft — offer FPGA-based instances that offer developers easier access to FPGA hardware.

- FPGAs can be used to implement programmable function accelerators or SmartNICs for use in high-performance networking products where devices with a mix of Arm processor cores and high-performance logic are well-suited.

Obstacles

- While a wide range of software applications — such as databases, security and encryption applications — could benefit from using an FPGA accelerator, very little commercial software is available that integrates FPGA support.

- Typically, FPGAs are configured using hardware programming languages such as register transfer level (RTL) and VHSIC Hardware Description Language (VHDL). These languages are complex to use and require hardware engineering and logic design skill sets, rather than software programming skills.

User Recommendations

- Identify application subsets that can be meaningfully impacted using FPGAs and where preconfigured solutions exist that can help dramatically transform key high-performance workloads (such as financial trading analytics and genome sequencing).

- Evaluate the maturity of software-centric programming toolsets as well as the costs associated with obtaining the necessary skill set and dealing with programming challenges.

- Leverage cloud-based services for provisioning FPGAs — such as Amazon EC2 F1 instances, Microsoft Azure, Baidu AI Cloud — to minimize risks.

Sample Vendors

Amazon Web Services; AMD; Baidu; Intel; Microsoft; Mipsology

Gartner Recommended Reading

Forecast: AI Semiconductors, Worldwide, 2021-2027

Forecast Analysis: AI Semiconductors, Worldwide

Emerging Tech Impact Radar: Artificial Intelligence

Climbing the Slope

Arm Servers

Analysis By: Alan Priestley

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Arm servers are built using microprocessors or systems-on-chip (SoCs) designed using processor cores based on the Arm instruction set architecture (ISA). Many of these processor designs leverage standard Arm IP. These IP-based designs enable vendors to customize processors for specific applications and workloads. Arm servers are being used by hyperscale cloud service providers and high-performing computing users to implement server infrastructure highly optimized for their workloads.

Why This Is Important

The use of Arm-based processor designs has long held the promise of more energy-efficient server designs, and hence lower operating costs for large-scale data centers. However, uptake has been limited as Arm processor core designs have lacked performance versus x86-based designs. Arm's latest Neoverse IP cores deliver competitive performance equivalent to current-generation x86 cores. This has made it viable to develop Arm-based processors optimized for use in servers.

Business Impact

Arm servers bring business benefits by:

- Lowering hardware and operational costs in targeted use cases compared to x86-based servers, especially for workload-specific appliances and open-source software applications.
- Acting as a competitive threat, influencing the x86 server processor vendors' product portfolios and pricing.

Drivers

- Arm's development of its server-optimized Neoverse IP core designs has enabled cloud service providers such as Amazon Web Services (AWS) and Alibaba to design their own Arm processors.
- Vendors like Ampere and NVIDIA are developing Arm-based processors for cloud service providers, high-performance computing and enterprise solutions.
- Growing use of high-level programming languages and the development of various microservice-based cloud-native applications are creating a set of applications and workloads that are not dependent upon the processor ISA.
- Performance improvements enabled by successive generations of processor designs mitigate the performance impact of using interpreted programming languages or just-in-time compilers, further breaking the dependency on processor ISA.
- The flexibility that the Arm ecosystem provides, in terms of developing processors optimized for a specific set of workloads, is creating increased interest in the use of Arm servers within data centers.
- Demand for web serving, caching, storage management and network connectivity products as well as high-performing computing workloads well-suited to the use of high-core-count Arm processor designs is also driving this technology.

Obstacles

- Many enterprise workloads are highly optimized for the x86 ISA and it may not be possible, or software vendors have no plans, to port to the Arm ISA.
- Many software tools that IT organizations use to manage their infrastructure and operations may not yet be available on the Arm ISA limiting the uptake of Arm servers in many traditional, on-premises data centers.
- The broad range of price and performance offered by x86 server processors and the reemergence of AMD server processors are both headwinds against the growth of Arm servers in enterprises.
- Intel's plans to offer x86 IP via its foundry service, will enable companies that might have designed their own Arm-based processors to develop their own custom x86 processors.
- The growth and increasing viability of the RISC-V processor.

User Recommendations

- Evaluate cloud deployment plans and the potential to use Arm server instances offered by major cloud service providers — often at attractive price points.
- Assess the use of Arm-based cloud instances where open-source and cloud-native development projects are planned.
- Investigate alternatives to x86 designs for future high-performance computing (HPC) deployments by evaluating OEM portfolios and roadmaps for Arm-based systems.
- Evaluate Arm-based systems in new use cases that are not dependent on the traditional installed base of applications and do not have interdependencies.
- Ensure that future developments are independent of the underlying processor ISA unless it is necessary for an optimized application to utilize specific processor features (e.g., Intel's AVX-512 and Deep Learning Boost [DL Boost] instructions for AI inference applications).
- Plan on using x86 servers as the primary architecture for on-premises infrastructure for the foreseeable future.

Sample Vendors

Amazon Web Services; Ampere; Arm; Marvell

Gartner Recommended Reading

Understanding the Opportunity for Arm-Based Servers

Market Trends: Arm in the Data Center: Act Now to Develop Plans to Address This Shifting Market

Emerging Tech Impact Radar: Compute and Storage

Immutable Infrastructure

Analysis By: Neil MacDonald, Tony Harvey

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

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

Business Impact

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

Drivers

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

- The GitOps deployment pattern, which emphasizes continuously synchronizing the running state to the software repository, has become an effective way to implement immutable infrastructure in Kubernetes-based, containerized environments.

- Infrastructure as code (IaC) tools (including first-party cloud provider IaC tools) have increasingly integrated configuration drift detection and correction, improving the practicality of implementing immutable infrastructure across an application's entire stack and environment.

- Interest in zero-trust and other advanced security postures where immutable infrastructure can be used to proactively regenerate workloads in production from a known good state (assuming compromise), a concept referred to as "systematic workload reprovisioning."

- For cloud-native application development projects, immutable infrastructure simplifies change management, supports faster and safer upgrades, reduces operational errors, improves security, and simplifies troubleshooting.

Obstacles

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

User Recommendations

- Reduce or eliminate configuration drift by establishing a policy that no software, including the OS, is ever patched in production. Updates must be made to individual components, versioned in a source-code-control repository, then redeployed.
- Prevent unauthorized change by turning off all administrative access to production compute resources. Examples of this might include not permitting Secure Shell or Remote Desktop Protocol access.
- Adopt immutable infrastructure principles with cloud-native applications first. Cloud-native workloads are more suitable than traditional on-premises workloads.
- Treat scripts, recipes and other codes used for infrastructure automation similar to the application source code itself, as this mandates good software engineering discipline.
- Include immutable infrastructure scripts, recipes, codes and images in your backup and ransomware recovery plans as they will be your primary source to rebuild your infrastructure after an infection.

Sample Vendors

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

Gartner Recommended Reading

Comparing DevOps Architecture to Automate Infrastructure and Operations for Software Development

2022 Strategic Roadmap for Compute Infrastructure

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

Innovation Insight for Continuous Infrastructure Automation

Market Guide for Cloud-Native Application Protection Platforms

Micro OS for Containers

Analysis By: Thomas Bittman

Benefit Rating: Moderate

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

A micro operating system (OS) for containers needs to be small and lightweight, and be designed specifically to support containers. Designed for clustered microservices architecture applications, it is often deployed in cloud environments. A micro OS for containers is intended for rapid deployment and horizontal scaling, with a typical image footprint ranging from 150MB to 500MB.

Why This Is Important

Modern, general-purpose OSs often have large footprints, are cumbersome to deploy and require relatively large hardware platforms. As containers, microservices and edge computing deployments develop a range of smaller form factors, more agile and clusterable micro OSs are required to support them.

Business Impact

Micro OSs will enable business agility through more rapid application development, easier and more efficient platform management, and high levels of horizontal scaling to meet business needs. Micro OSs will be core enablers to most new digital business applications — both in the cloud and at the edge.

Drivers

- Agile deployment models that require a small footprint software foundation
- Small, container-based solutions that don't require a rich general-purpose OS
- Microservices architectures and DevOps models that are designed for many small footprints
- Processing-constrained edge computing deployments using containers

Obstacles

- Micro OSs face decades of skills, process and application architectures centered on rich, general-purpose OSs and vendor business models surrounding existing OSs.
- Many micro OSs are being developed as subsets of existing OSs, but many new ones are also emerging. Not all of them will survive.
- Applications may need refactoring to operate with a micro OS.

User Recommendations

Micro OS technologies should be:

- Evaluated based on their technical maturity and footprint, and their interoperability with intended cloud providers and orchestration technologies.
- Assessed according to their feature sets, for example, too much, not enough or just right, and their fit with chosen container frameworks.
- Examined based on the viability of the vendor or the level of community support for open source.
- Rated according to the support and update technology provided by the vendor, which can be a subscription update service.

Sample Vendors

Amazon Web Services; Google; Microsoft; Red Hat; SUSE; VMware

Gartner Recommended Reading

Market Guide for Container Management

Prioritizing Security Controls for Enterprise Servers and End-User Endpoints

Designing and Operating DevOps Workflows to Deploy Containerized Applications With Kubernetes

OS Containers

Analysis By: Thomas Bittman, Philip Dawson

Benefit Rating: Transformational

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

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

Why This Is Important

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

Business Impact

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

Drivers

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

Obstacles

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

User Recommendations

Infrastructure and operations leaders responsible for data center infrastructure should:

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

Sample Vendors

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

Gartner Recommended Reading

Market Guide for Container Management

Prioritizing Security Controls for Enterprise Servers and End-User Endpoints

Designing and Operating DevOps Workflows to Deploy Containerized Applications With Kubernetes

Entering the Plateau

Hyperconvergence

Analysis By: Philip Dawson, Jeffrey Hewitt

Benefit Rating: High

Market Penetration: More than 50% of target audience

Maturity: Mature mainstream

Definition:

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

Why This Is Important

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

Business Impact

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

Drivers

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

- Hyperconvergence leads to lower operating costs, especially as it supports a greater share of the compute and storage requirements of the data center.
- Nutanix, an early innovator in hyperconverged integrated system (HCIS) hardware appliances, has largely shifted to a Hyper Converged Infrastructure (HCI) software revenue model and continues to increase its number of OEM relationships and partners.
- Larger clusters are now in use, and midsize organizations are considering hyperconvergence as the preferred alternative for on-premises infrastructure for block storage.
- Hyperconvergence vendors are achieving certification for more demanding workloads, including Oracle and SAP, and end users are beginning to consider hyperconvergence as an alternative to integrated infrastructure systems for some workloads.
- As more vendors support hybrid and public cloud deployments, hyperconvergence is a stepping stone toward public cloud agility as suppliers are expanding hybrid cloud deployment offerings for cloud-native applications.
- A number of niche hyperconvergence suppliers offer scale-down solutions to address the needs of remote office/branch office (ROBO) and edge environments.

Obstacles

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

User Recommendations

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

Sample Vendors

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

Gartner Recommended Reading

Market Guide for Full-Stack Hyperconverged Infrastructure Software

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

Market Guide for Integrated Systems

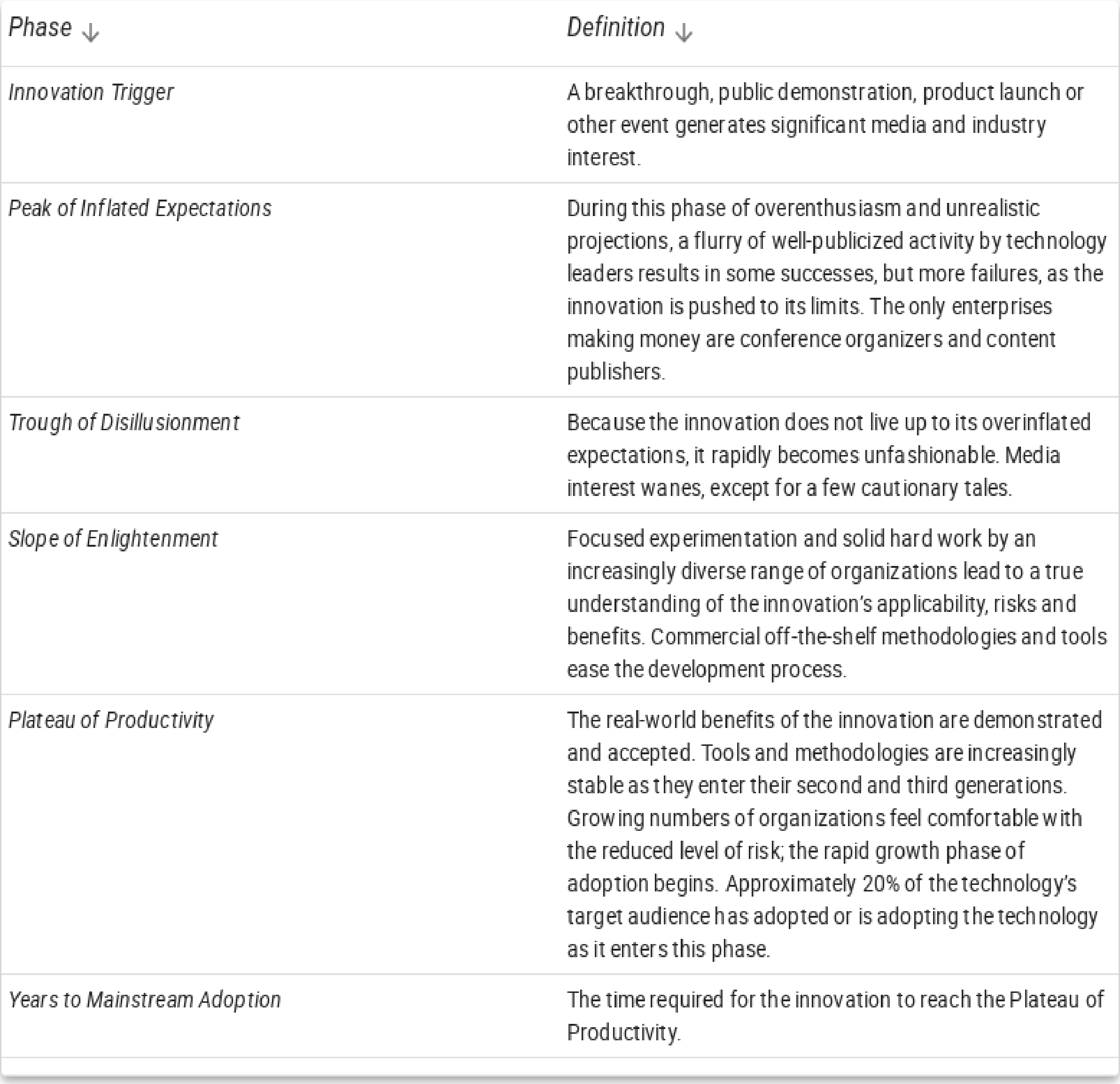
Appendixes

See the previous Hype Cycle: Hype Cycle for Compute, 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)

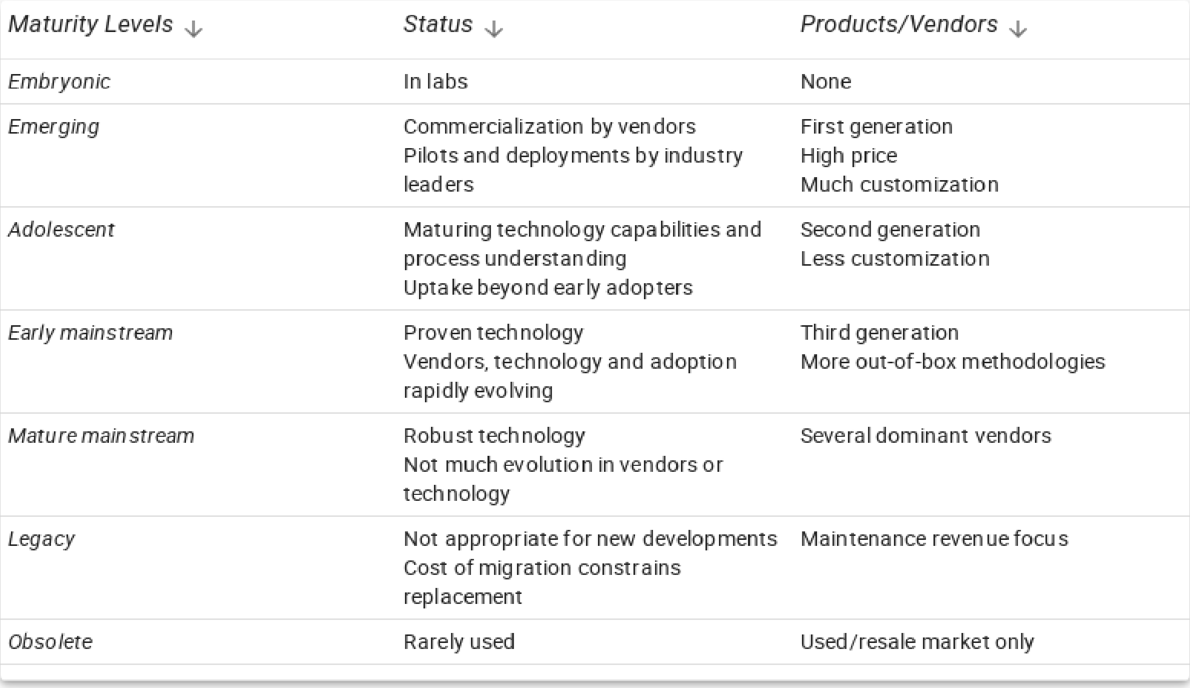
Table 3: Benefit Ratings

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 Compute, 2022 - 11 July 2022

Hype Cycle for Compute Infrastructure, 2021 - 22 July 2021

Hype Cycle for Compute Infrastructure, 2020 - 8 July 2020

Hype Cycle for Compute Infrastructure, 2019 - 26 July 2019

Hype Cycle for Compute Infrastructure, 2018 - 19 July 2018

Hype Cycle for Compute Infrastructure, 2017 - 21 July 2017

Hype Cycle for Compute Infrastructure, 2016 - 1 July 2016

Hype Cycle for Server Technologies, 2015 - 21 July 2015

Hype Cycle for Server Technologies, 2014 - 11 July 2014

Hype Cycle for Server Technologies, 2013 - 31 July 2013

Hype Cycle for Server Technologies, 2012 - 24 July 2012

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

3 Steps to Supercharge Your GTM Strategy and Programs Using Gartner Hype Cycles

Top Strategic Technology Trends for 2023

2022 Strategic Roadmap for Compute Infrastructure

Market Guide for Server Virtualization

Cool Vendors in Quantum Computing

Rethink Supercomputing for a Digital Era

Market Guide for Consumption-Based Models for Data Center Infrastructure

Table 1: Priority Matrix for Compute, 2023

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	OS Containers	Compute Express Link	Neuromorphic Computing	Emerging Memory Technologies Quantum Computing
High	Edge Servers Hyperconvergence	BMaaS Composable Infrastructure Consumption-Based Model Container Management Deep Neural Network ASICs Infrastructure Automation Infrastructure Orchestration	Function Accelerator Cards IT/OT Hybrid Servers	
Moderate	Micro OS for Containers	Arm Servers Cloud-Tethered Compute eBPF FPGA Accelerators Immersion Cooling Immutable Infrastructure	Confidential Computing Direct-to-Chip Liquid Cooling	
Low				

Source: Gartner (July 2023)

Table 2: Hype Cycle Phases

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

Phase ↓	*Definition* ↓

Source: Gartner (July 2023)

Table 3: Benefit Ratings

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

Source: Gartner (July 2023)

Table 4: Maturity Levels

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

Source: Gartner (July 2023)