Hype Cycle for Compute Infrastructure, 2021

Published 22 July 2021 - ID G00747396 - 88 min read

By Analyst(s): Tony Harvey

Initiatives: Data Center Infrastructure

Compute infrastructure options continue to grow beyond on-premises servers and cloud, as new solutions, such as bare metal as a service, come to market, and consumption-based sourcing models proliferate. I&O leaders must use this research to determine investment priorities and adoption timelines.

Additional Perspectives

- Summary Translation + Localization: Hype Cycle for Compute Infrastructure, 2021 (29 September2021)
- Invest Implications: Hype Cycle for Compute Infrastructure, 2021 (29 July 2021)

Analysis

What You Need to Know

Compute infrastructure can be located anywhere and must support operations for a broad variety of environments. A hybrid compute infrastructure consisting of on-premises cloud and edge is becoming the norm. I&O leaders focused on compute infrastructure should evaluate the technological innovations covered in this research to align compute strategies with the needs of each of these environments.

For more information about how peer 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

Users and apps are everywhere. It is not just employees who are remote; data can now be processed, transmitted and stored across its entire life cycle without ever touching a corporate-owned server, network or storage array. This has altered the nature of compute infrastructure in that the location of compute has become less important, and focus has shifted to determining the right delivery model for specific applications.

This research describes the 28 most-hyped innovations in compute infrastructure. For each technology, we define and analyze the value to enterprises, level of adoption and anticipated rate of future growth. I&O leaders should use this research to determine whether and/or when to invest in these innovations.

New: Bare metal as a service (BMaaS) and consumption-based sourcing represent shifts in the ways organizations consume and deliver compute infrastructure.

On the Rise: Next-generation interconnects, cloud-tethered compute and computational storage are moving up the slope and toward peak hype. Real-world implementations will be needed to take these over the peak.

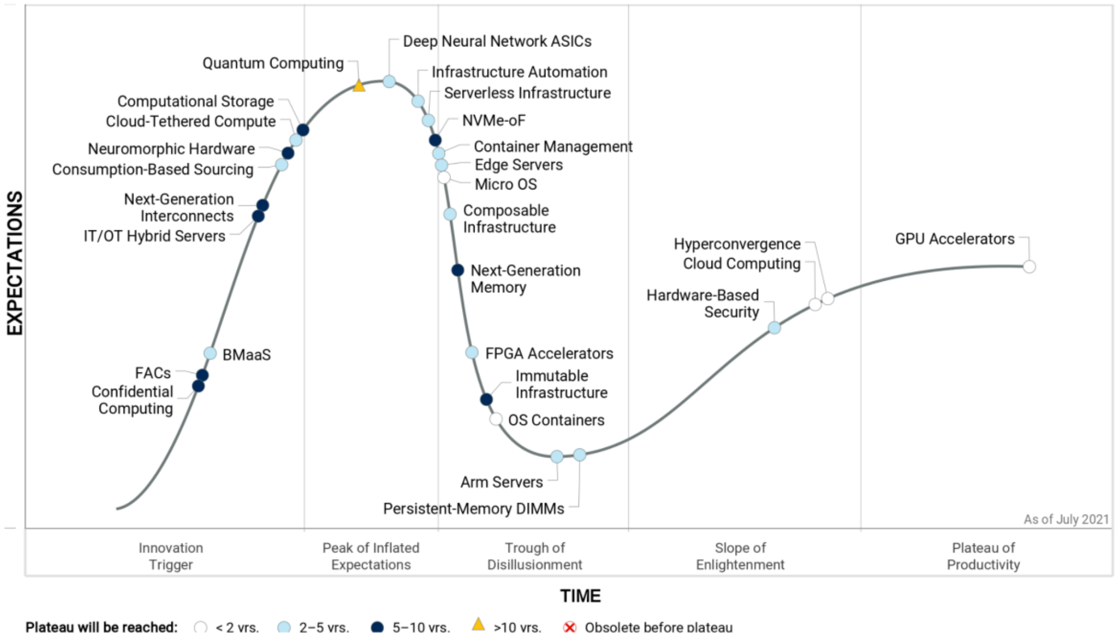
Peak Hype: Quantum computing once again remains hyped, but with few users. Deep neural network application-specific integrated circuits (ASICs) and infrastructure automation are moving more rapidly.

Fast Movers: Two of the three fastest movers, hardware-based security and immutable infrastructure, demonstrate the importance of security in this new environment. Hyperconvergence has made the transition to the Slope of Enlightenment.

In the Trough: Persistent memory dual in line memory modules (DIMMs) and ARM servers are solidly in the Trough of Disillusionment, where they suffer from the problem of requiring major changes to the codebase of OSs and applications.

Dropped: Server-side client graphics and in-memory computing have moved to mainstream adoption. dHCI is now being driven by only one vendor.

Figure 1: Hype Cycle for Compute Infrastructure, 2021



Source: Gartner (July 2021)

Downloadable graphic: Hype Cycle for Compute Infrastructure, 2021

The Priority Matrix

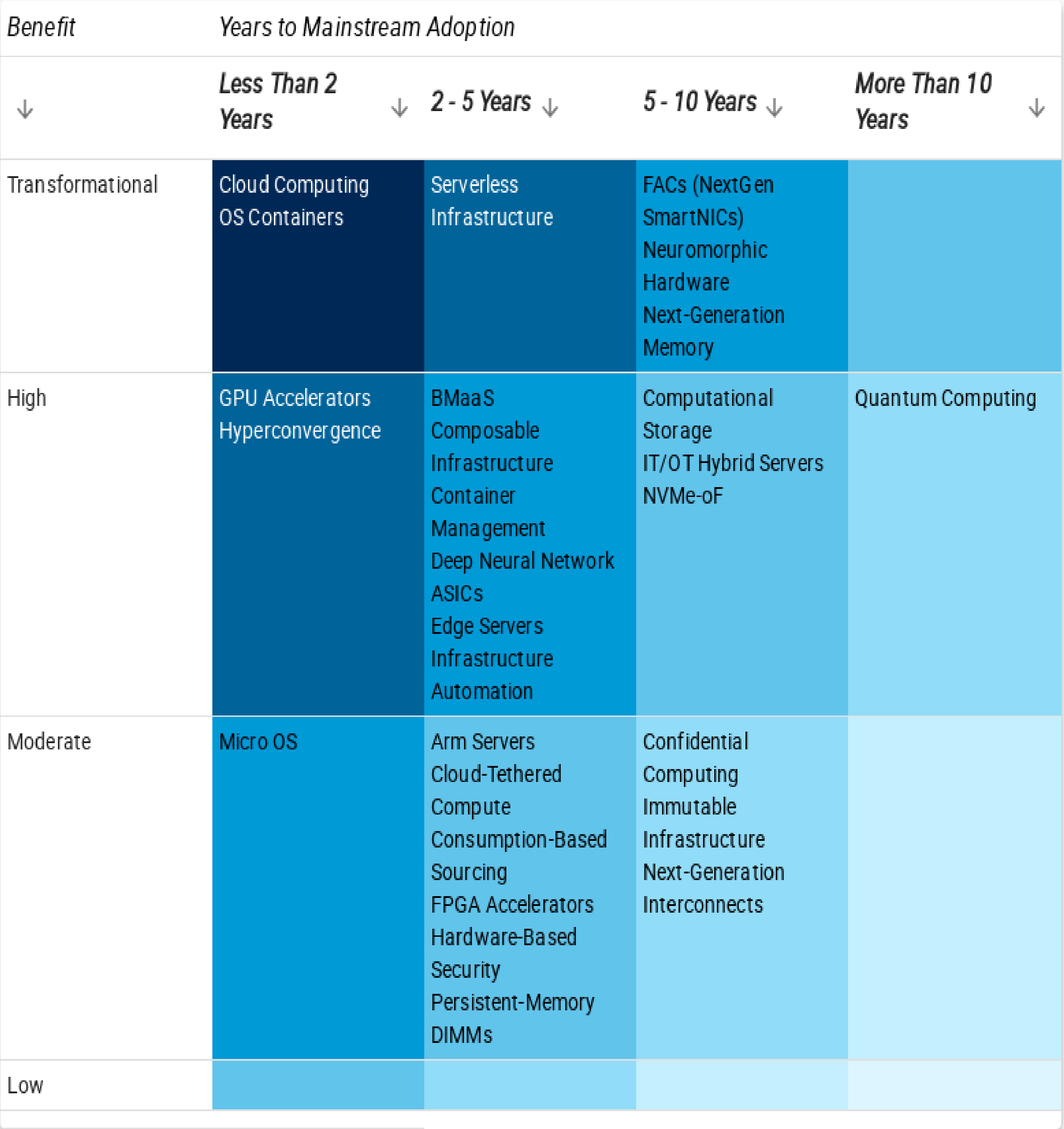
The Priority Matrix maps the benefit rating for each technology against the amount of time required to reach the beginning of mainstream adoption. This alternative perspective can help users determine how to prioritize their server technology investments.

Transformational technologies, such as OS containers and serverless infrastructures, affect compute infrastructure by eliminating the need for developers to manage hardware. To deliver a platform ready for these solutions, I&O leaders must adopt infrastructure automation and immutable infrastructure. Further out, I&O leaders should start to evaluate how function accelerator cards can increase performance and security for on-premises compute environments.

BMaaS is a new, high-impact technology that enables cloud-style delivery of compute closer to the edge. I&O leaders should evaluate BMaaS in colocation and edge environments, and look to deliver the same capabilities in their on-premises data centers. Nonvolatile memory express over fabrics (NVMe-oF) is becoming more widespread in storage arrays, offering significant performance increases for some workloads. In the longer term, evaluate use cases for computational storage to determine which could offer competitive advantage.

Table 1: Priority Matrix for Compute Infrastructure, 2020

(Enlarged table in Appendix)



Benefit	*Years to Mainstream Adoption*			
↓	*Less Than 2 Years* ↓	*2 - 5 Years* ↓	*5 - 10 Years* ↓	*More Than 10 Years* ↓
Transformational	Cloud Computing OS Containers	Serverless Infrastructure	FACs (NextGen SmartNICs) Neuromorphic Hardware Next-Generation Memory	
High	GPU Accelerators Hyperconvergence	BMaaS Composable Infrastructure Container Management Deep Neural Network ASICs Edge Servers Infrastructure Automation	Computational Storage IT/OT Hybrid Servers NVMe-oF	Quantum Computing
Moderate	Micro OS	Arm Servers Cloud-Tethered Compute Consumption-Based Sourcing FPGA Accelerators Hardware-Based Security Persistent-Memory DIMMs	Confidential Computing Immutable Infrastructure Next-Generation Interconnects	
Low				

Source: Gartner (July 2021)

On the Rise

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

- Confidential computing combines a chip-level TEE with conventional key management and cryptographic protocols to enable unreadable computation, allowing a variety of projects where cooperation is critical, without sharing data or IP.
- 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.
- Inquiry trends reveal increased anxiety among cloud-using organizations about CSP staff eavesdropping into or tampering with customer workloads. Confidential computing is an emerging mechanism that CSPs can offer to alleviate these concerns.

Business Impact

- Confidential computing may mitigate one of the major barriers to cloud adoption for highly regulated businesses or any organization concerned about unauthorized third-party access to data in use in the public cloud.
- Confidential computing allows a level of data privacy 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 CSPs.
- Global data residency restrictions are ongoing, with a need to keep access to content away from even the CSP.
- Competitive concerns — not just around personal data, but also intellectual property — are spurring adoption of confidential computing. This includes the need for confidentiality and protection against any third-party access.

Obstacles

- Complexity of the tech and lack of trained staff/understanding of best implementation methods may hinder adoption and/or weaken deployment (e.g., when key management/handling is done incorrectly, unaddressed side channel vulnerabilities).
- Trust is slow to build and quick to evaporate, especially when experimental technology like confidential computing is paired with occasional hardware vulnerabilities.
- Because the technology is new and novel, and it touches sensitive data, potential clients have a hard time identifying valid use cases for their business.
- Confidential computing can provide such protection now. Be mindful of the potential performance impacts and the extra cost. IaaS confidential computing instances (whether SGX-based or otherwise) will cost more to run.
- Confidential computing isn't usually plug-and-play, and should be reserved for the highest-risk use cases. Depending on the vendor, It can require a high level of effort but offers diminishing marginal security improvement over more pedestrian controls like TLS, MFA, and customer-controlled key management services.

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 to determine whether confidential computing affects application performance and to seek ways to minimize negative results.

- Review alternatives that achieve similar protection of sensitive data in use, such as multiparty encryption, data masking or tokenization. None of these require specific hardware.
- 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.

Sample Vendors

Alibaba Cloud; Fortanix; IBM; Intel; Microsoft; Private Machines

Gartner Recommended Reading

Top Strategic Technology Trends for 2021: Privacy-Enhancing Computation

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 Retain the Right Kinds of Control in the Cloud

How to Make Cloud More Secure Than Your Own Data Center

FACs (NextGen SmartNICs)

Analysis By: Anushree Verma

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

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 — deployed as an ASIC, FPGA or SoC. FACs improve data operations and services, server availability and bandwidth performance besides enabling connectivity to a network. While all FACs are essentially NICs, not all NICs/SmartNICs are FACs. They come with onboard memory and peripheral interfaces and run independently.

Why This Is Important

FACs can accelerate a whole variety of functions, like networking, security and storage functions, by offloading these functions from the server and dedicated appliances. For example, they improve server performance as much as 10% to 50% by offloading RAID, data compression, block management, encryption and session management for TLS.

FACs are primarily adopted by hyperscale CSPs like Microsoft, AWS and Baidu to improve server performance and enable bare metal provisioning of instances.

Business Impact

FACs enable more cost-efficient data center environments while improving performance. By offloading high overhead functions, they allow the server to host more workloads, hence the direct cost of additional servers and infrastructure software is saved as well in some cases. In addition, they can provide NVMe over Fabrics and GPUDirect storage to facilitate data transmission between remote resources — necessary for hybrid and cloud computing setups with off-site servers.

Drivers

FACs have started growing in popularity since 2019 with hyperscale offering services to customers who are using AWS Nitro and Microsoft AccelNet. However, in 2020, with the ecosystem support, it has gone further in popularity, moving ahead in terms of adoption and justifying its advancement in this Hype Cycle. In addition, vendors are actively and aggressively marketing FACs, which are also referred to as data Processing Unit (DPU), SmartNICs, Distributed Services Card (DSC) or Programmable NICs. Gartner estimates that FACs will grow to $1 billion in 2024 from less than $10 million in 2019, at an estimated rate of 155% CAGR during the time period. Some of the drivers for this transformative innovation include:

- 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 (fueled by COVID-19).

- Frequency and sophistication of cyberthreats make it imperative to have security functions offloaded and run in an independent, isolated manner in a secure way, specifically in bare metal use cases or microservices-based applications.
- 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.
- 2020 saw FACs getting ecosystem support with server virtualization vendors such as VMware porting its ESXi hypervisor to ARM ISA. This enables the cores on a FAC to offload the core hypervisor functions, adding new levels of hardware-enabled isolation and security within virtualized systems. Another benefit is network consolidation for less demanding workloads.

Obstacles

- Hyperscale CSPs are able to justify the incremental price with the large scale order and customization benefits they have by adopting FACs — their per month revenue saving is passed on to the customer. However, enterprises are yet unable to justify the incremental price, thereby hindering rapid adoption.
- This market is very price sensitive; enterprises need a clear use case and benefit given the higher price point, to drive SmartNICs and FACs.
- Silos in the organization (compute, storage and networking) create challenges in the combined benefit that FACs provide.
- Data plane programmability is high risk and has limited value and interest for enterprises.
- Form factor and power consumption are also an obstacle to adoption. While some FACs consume less than 45 watts and are smaller form factor, some FACs consume as much as 75 watts, blowing up rack, power and cooling budgets, and may have to occupy the only full-size PCIe slot in the server, preventing the installation of a GPU add-in card.

User Recommendations

- Bring in short-term value with FACs, but plan to eventually migrate more functions to drive long-term value.

- Engage with your vendor to not only generate incremental performance benefits, but — more importantly — replace legacy components like aggregation switches and reduce the number of application licenses. Most licenses are for per core, but eliminating the number of server and hypervisor licenses by 30% and the number of network security licenses can drive huge incremental benefits in terms of cost benefits and increasing performance.

- Evaluate FAC-based storage offerings if you are an enterprise with applications that require microsecond latency performance when processing large data sets.

- Engage with FAC vendors if you are a large DCN operator to address network scale/security needs. Enterprises should pilot FAC as an option to support extremely network sensitive workloads.

- Exercise careful evaluation according to use case; not all SmartNICs available in the market are FACs.

Sample Vendors

Broadcom; Ethernity Networks; Fungible; Nebulon; NVIDIA; Pensando; Pliops; VMware; XILINX

Gartner Recommended Reading

Market Trends: Function Accelerator Cards Disrupting Traditional Ethernet Adapter Market

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: Emerging

Definition:

Bare metal as a service (BMaaS) provides physical infrastructure such as compute, networking and storage via a cloud-like 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 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 can run workloads without restrictions based on hypervisor or OS compatibility, or performance concerns owing to other client workloads. This results in faster and more agile "physical" infrastructure deployment, efficiency in meeting elasticity and scalability demands, and broad geographic presence. BMaaS is also often selected over public cloud infrastructure to conform to legacy software licensing requirements based on permanent deploying onto fixed physical hosts (e.g., Oracle).

Business Impact

BMaaS offers:

- The advantages of dedicated infrastructure (predictability, security, performance) with the elasticity, scalability, and agility of IaaS.
- A cloud-like experience in a data center location better suited to customer needs for low network latency, data residency, etc.
- A flexible integration platform at the nexus of public cloud access locations, such as colocation hubs or content delivery network (CDN) POPs.

Drivers

BMaaS is not an entirely new concept, with offerings available in the early days of cloud from providers such as SoftLayer, later acquired by IBM. What's new is:

- The ability to act like a true 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.

- Cloud market consolidation has eliminated most of the early IaaS players. The surviving hyperscalers which come with extensive PaaS stacks don't suit users just looking for basic infrastructure, which BMaaS offers.
- Interest in cloud-native technologies as a path toward cloud independence and to reduce lock-in.

Drivers for bare metal as a service (BMaaS) include:

- Bare metal may solve the issue of physical workload location addressing the concerns that highly centralized offerings may pose, either due to latency concerns, enterprise control or data sovereignty and regulations.
- Bare metal offers the speed and agility of public cloud, with far greater control over workload and data placement.
- Bare metal is less costly than physical infrastructure, does not tie up capital and is faster to deploy.
- Bare metal offers broader geographic presence than hyperscale data centers.
- Bare metal provides a platform for multicloud integration.

Obstacles

- Complexity of another infrastructure environment.
- Customer must supply and configure much of the software, bearing the risk and cost of a greater portion of the full stack
- Requirements for unique or integration of multiple network offerings.
- Ease and flexibility of consumption may vary.
- Economics may vary by workload type, networking and storage services included.

User Recommendations

- Identify target candidates for BMaaS by evaluating IaaS-bound applications and determining whether the attributes of BMaaS better fit the circumstances.
- Build BMaaS into cloud assessment models by identifying those attributes that can only be addressed through the licensing compatibility, hypervisor independence, and location specificity of bare metal.
- Leverage bare metals' unique location benefits by identifying applications that require low latency or sovereignty through proximity to cloud on ramps.
- Select BMaaS for "cloud-native hosting" of legacy applications whose licensing terms are optimized for dedicated physical hosts.

Sample Vendors

Amazon Web Services (AWS); Cyxtera; Digital Realty Trust; Equinix, Oracle; Rackspace Technology

Gartner Recommended Reading

Break Down 3 Barriers to Cloud Migration

IT/OT Hybrid Servers

Analysis By: Tony Harvey

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

IT/OT hybrid servers are edge devices that are designed to interface, collect and process data from operational technology systems that provide real-time control of physical systems and industrial processes. They are designed to operate with higher resilience to shock, vibration, humidity and temperature than typical data center servers. Industrial communications interfaces — such as CAN bus, Modbus or Profinet, 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. By being connected to IT networks, IT/OT servers allow for the collected data to 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 pool of data 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 do not adopt IT/OT hybrid servers may find themselves left behind as enterprises that successfully integrate these systems into their digital transformation strategy will lower their costs and deliver new services to market faster.

Drivers

- Real-time analysis and decision making based on capturing data to optimize industrial processes
- Near-real-time reporting of, manufacturing and production data
- Device monitoring to enable predictive maintenance of industrial equipment to reduce line stoppages and downtime
- Use of Industrial IoT sensors that require data be processed and stored at edge locations
- Collection of OT data to enable AI/ML training and digital twin model building

Obstacles

- Caution from industrial enterprises about the use of IT systems in industrial process control, where failure could result in loss of life or significant property damage.
- Security risks of connecting industrial process control systems to the internet.
- The disconnect between IT and OT life cycles — where OT systems may last for 20 years or more, but IT systems typically have much shorter and faster life cycles.
- IT and OT are separate groups with different cultures and different perceptions of risk. The differences between these groups must be managed for any successful implementation.
- Complexities in defining what data needs to stay at the edge versus what should be delivered and processed in the cloud.

User Recommendations

CIOs looking to evaluate IT/OT hybrid servers must do so as part of an IT/OT integration program that should:

- Create an integrated IT/OT group that has full responsibility for these solutions.
- Align IT/OT in areas of 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 risk and security training, awareness and talent in hybrid IT/OT teams to ensure that systems are designed with security in mind.
- Optimize the costs for ongoing support, maintenance/updates and dependencies across the entire combined I/OT environment.

Sample Vendors

Dell EMC; Hewlett Packard Enterprise (HPE); Lenovo; Schneider Electric

Gartner Recommended Reading

Alternative Organizational Models for IT/OT Alignment

2020 Strategic Roadmap for IT/OT Alignment

When Does a CIO Need to Be Involved in OT?

Next-Generation Interconnects

Analysis By: Tony Harvey

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Next-generation interconnects are a computer system bus designed for short-range, cache-coherent connection of CPUs, memory, accelerators and I/O devices. These interconnects, typically defined by industry consortia, replace or build on PCI Express (PCIe) or proprietary interconnects from processor vendors. Examples include Compute Express Link (CXL), CCIX, Gen Z and Open Coherent Accelerator Processor Interface (OpenCAPI).

Why This Is Important

New high-bandwidth accelerators such as GPUs, FPGAs and DPUs need direct access to system memory to perform optimally. Next-generation interconnects provide the necessary capabilities, such as cache-coherent shared memory access and memory semantics for I/O, that will enable applications like AI and ML to produce new insights.

Business Impact

Business impacts will be limited until standardized silicon and software are available; however, as the industry coalesces on a standard, businesses will start to see:

- Material performance benefits to 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 of compute as a standardized solution drives competition in the market

Drivers

- The rise in workloads such as AI/ML solid modeling, seismic analysis and advanced analytics creating unprecedented demand for higher-performance I/O and accelerators
- Market stabilization around CXL for an inbox solution and Gen Z for external connectivity
- Need to better utilize expensive resources such as DPU, FPGA and GPU accelerators by sharing across multiple systems

Obstacles

- User and vendor confusion over multiple standards all offering similar capabilities.
- Availability being dependent on the release of CPUs supporting PCIe Gen5, which will not happen until late 2022 at the earliest.
- Requirement for new programming models to be adopted and application redesign to fully take advantage of the new capabilities. This will initially limit uptake to specific domains, such as AI/ML and HPC.

User Recommendations

- Avoid using CCIX, CXL or Gen Z support as a criterion for selecting hardware today.
- Ask vendors for roadmaps and assessments of remaining technical roadblocks.
- Verify that consortium members include vendors sufficient to implement the interconnects in all necessary components.
- Check which standards are embraced by hyperscalers, as their support will be a determining factor in the success and timeline of these standards.

Sample Vendors

AMD; Dell Technologies; Hewlett Packard Enterprise (HPE); Huawei; IBM; Intel; NVIDIA; Xilinx

Gartner Recommended Reading

Predicts 2020: Semiconductor Technology in 2030

Consumption-Based Sourcing

Analysis By: Jeff Vogel, Philip Dawson

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

A consumption-based sourcing model strategy for on-premises data center storage and compute infrastructure is an acquisition, deployment and support model that includes a cloud-like consumption and platform services model with a variable payment tied to measured use.

Why This Is Important

Server and storage vendors have launched or rebranded consumption-based offerings during the past two years, to provide cloud-inspired or alternatives to the public cloud. Program examples include Dell Technologies' Apex, HPE's GreenLake, Pure Storage's Pure as-a-Service and Cisco Plus. These programs align infrastructure costs with resource use, and to shift infrastructure spending from capital expenditures (capex) to cloudlike operating expenditure (opex).

Business Impact

A consumption-based sourcing strategy is a cloud-like managed service offering that:

- May shift maintenance and support costs' responsibility to vendors investing in artificial intelligence (AI) operations that automate operations
- Will preserve cash by avoiding upfront capex in exchange for strategic priorities
- Will shift IT and finance resource budget cycles to a services-based delivery model
- Will provide a more flexible and agile IT operations environment aligned with business demands

Drivers

Many 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 permanence of the cloud
- The need for a more flexible and agile IT operating model
- The massive growth in enterprise data
- The need for an application-aware services delivery model
- Preferences of 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 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

Obstacles

Consumption-based sourcing may:

- Be more expensive than a purchase.
- Be organizationally challenging to implement during long, protracted periods.
- Be unsuitable for IT operations that don't require flexibility to accommodate uncertain growth and variability in forecast demand.
- Require minimum-usage commitment levels, and do not scale down to zero — that is, minimum commitments require that users pay for most of the deployed infrastructure, regardless of what is actually consumed.
- Require three-to-five-year contracts with mandatory 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 metrics.
- Conflict with established industry accounting standards and operational norms.
- Conflict with consumption-based hardware, making software licensing terms impractical.

User Recommendations

By implementing a consumption-based strategy, IT leaders should:

- Adopt a cloud operating model as a platform strategy to shift to an IT-as-a-services-capable organization.
- Organize a joint team approach to include I&O, vendor management and finance to establish a strategic sourcing strategy and implement.
- Rightsize and align IT I&O to business demand.
- 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.
- Align and link consumption-based costs to specific usage at agreed on elastic utilization levels, 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 (HPE); IBM; Lenovo; NetApp

Gartner Recommended Reading

Enterprise Storage as a Service Is Transforming IT Operating Models

Enterprise Storage Is Growing Consumption-Based Pricing Revenue

Create Consumption-Based On-Premises Infrastructure Bundles for Midsize Enterprises

Key Considerations for CSOs Moving to a Consumption-Based Subscription Model

Neuromorphic Hardware

Analysis By: Alan Priestley

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Neuromorphic hardware comprises semiconductor devices inspired by neurobiological architectures. Neuromorphic processors feature non-von-Neumann architectures and implement spiking neural network execution models that are dramatically different from traditional processors. They are characterized by simple processing elements, but very high interconnectivity.

Why This Is Important

As of 2021, most AI development leverages parallel processing designs based on GPUs. These are high-performance, but at the same time high-power-consuming, devices that are not applicable in many deployments.

Neuromorphic systems utilize 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 and audio analytics.

Business Impact

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

- As of 2021, DNN algorithms require the use of high performance processing devices and vast amounts of data to train these systems, limiting scope of deployment.
- Neuromorphic devices 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

- Gartner expects that many different AI-related hardware technologies and architectures will come to market before a stable state of widespread, high-volume deployment is reached.
- Currently, much of the work on developing AI-based systems is focused on the use of DNNs that leverage highly parallel data-intensive processing techniques to simulate the operation of a biological brain.
- These AI developments typically use semiconductor devices, such as GPUs or custom designed ASICs. While meeting today's demands in terms of performance, the power consumption of many of these designs limits their ability to scale to meet the challenges of large scale AI implementations.
- Neuromorphic computing leverages the concept of spiking neural networks (SNNs) to model a biological brain.
- Different design approaches are being taken to implement neuromorphic computing designs — large scale devices for use in data centres, and smaller scale devices for edge computing and endpoint designs. Both these paths implement asynchronous designs that have the benefit of being extremely low power when compared to 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 deep neural networks and signal analysis at very low power.
- Neuromorphic systems can be simpler to train than DNNs, with the potential of in-situ training.

Obstacles

- **Accessibility:** As of 2021, GPUs are more accessible and easier to program than neuromorphic hardware. However, this could change when neuromorphic hardware and the supporting ecosystems mature.
- **Knowledge gaps:** Programming neuromorphic hardware 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.
- 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 hardware 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.

Sample Vendors

AnotherBrain; BrainChip; IBM; Intel; NeuroBlade; SynSense

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Emerging Technologies: Using Neuromorphic Neural Networks to Advance IoT Vision Projects

Forecast: AI Semiconductors, Worldwide, 2019-2025, 1Q21 Update

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

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 delivers bare metal as a service (BMaaS), infrastructure as a service (IaaS) and platform as a service (PaaS) in customer-controlled environments managed by the vendor via a network tether from a public or private cloud. The tether may need to be continuously connected for billing, or be able to operate disconnected with periodic connectivity. Updates are vendor-managed, removing responsibility for platform maintenance from the infrastructure and operations (I&O) team.

Why This Is Important

Cloud-tethered compute is compute infrastructure running on customers' premises or at edge locations that is delivered as a service, managed and maintained by the provider. It enables customers to deploy cloudlike services directly into their own environments without having to maintain the environment themselves.

Business Impact

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

- IT teams will see a reduced need for basic 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 will cycle to a services-based delivery model.
- IT operations will be more flexible and aligned with business demands

Drivers

- Need to connect to local data sources over low-latency networks
- Ability to deliver a cloudlike experience using on-premises infrastructure
- Need for IT teams to focus on higher-level business objectives, rather than IT operations and maintenance
- Enterprise need for hardware and IaaS solutions at the edge and on-premises
- Promise of “evergreen” technology refresh solutions that keep systems up-to-date with the latest technology

Obstacles

- Complex contracts that have large minimum usage requirements.
- Most cloud-tethered compute systems do not have a complete set of services that matches available public clouds.
- Public cloud vendor expertise in support of and maintenance of field solutions is new and untested.
- Large-scale edge deployments of tethered solutions are untested.

User Recommendations

- Identify scenarios in which the tethered compute model provides clear business value versus a more-traditional IT solution.
- Use cloud-tethered compute to enable IT teams to retire technical debt and focus on delivering value to the business.
- Organize a joint team that includes I&O, 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 IT team — elements such as data backup and application security are likely to be the end user's responsibility.

Sample Vendors

AWS Outposts; Google Anthos; Hivecell; Microsoft Azure Stack; Oracle; VMware

Gartner Recommended Reading

'Distributed Cloud' Fixes What 'Hybrid Cloud' Breaks

Prepare for AWS Outposts to Disrupt Your Hybrid Cloud Strategy

How to Bring the Public Cloud On-Premises With AWS Outposts, Azure Stack and Google Anthos

Best Practices for Tech CEOs to Manage Edge-to-Cloud Products

Top Emerging Trends in Cloud-Native Infrastructure

Computational Storage

Analysis By: Jeff Vogel, Julia Palmer

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Computational storage (CS) combines processing and storage to reduce performance inefficiencies and latency-sensitive application issues in the movement of data between storage and compute resources. CS offloads host processing from the main memory of the CPU to the storage device. CS involves more sophisticated processing capabilities located on the storage device. CS storage products employ greater processing power in the form of FPGA and ASICs with low-power CPU cores on the SSD.

Why This Is Important

CS is a new class of storage drive that provides power-efficient and consistent low-latency performance for latency-sensitive applications such as AI/ML, high-performance computing, immersive and mixed reality streaming, and high-frequency trading at the data center edge. Edge computing remains an opportunity, along with applications that favor distributed processing. CS may provide better memory management and energy savings over industry-standard solid-state media.

Business Impact

CS's low-power footprint vs. traditional SSD improves performance-per-watt ratio, decreasing power consumption costs for applications at the edge. Using a more powerful embedded compute engine with the SSD controller will increase storage efficiencies and lower overall application costs and total cost of ownership. CS drives have a domain-specific hardware engine that can effectively process compression, security and other critical drive functions, thus substantially reducing processing time.

Drivers

Computational storage brings computing power to storage to:

- Reduce performance inefficiencies and latency-sensitive issues in the movement of data between storage and compute resources.
- Reduce latency with edge workloads as data volumes increase, and movement of data becomes a bottleneck.
- Eliminate application performance issues that undermine real-time analysis when dataset sizes exceed memory.

- Remove bottlenecks in data-intensive applications such as AI/ML, database, high-performance computing, analytics, high-frequency trading, and immersive and mixed-reality streaming.
- Substantially improve performance, better memory management and energy savings over storage systems by embedding processing on the storage media.

Obstacles

CS market adoption is still largely in the development phase, with a handful of small vendors deploying proof-of-concept systems, and limited and small-volume production systems. CS system architecture:

- Is more complex and may require applications to be recompiled
- May require additional APIs, or for the host system to be aware of the services that are provided by the CS system
- Will require operational tasks to be offloaded into the SSD drives so host applications must be able to adequately communicate with CS storage drives
- Requires a standardized programming model and interface protocol, which are being actively developed by an industry standards body
- Requires a software framework that enables a host server application to interact with a CS device and is based on an open and standard framework that is currently being actively worked on
- Are relatively nascent and the value for what's being offered is difficult to justify in some cases

User Recommendations

- Explore potential benefits that can be gained from specific use cases, but should carefully weigh the cost vs. performance gains vs. opex savings and amount of work required to deploy. This is especially true where certain workloads that are very input/output-bound and would benefit the most from processing in storage.
- Take into account compression/decompression engines that enable the drive both to store more data per gigabyte of flash and to maintain high performance within a narrow band regardless of the read/write mix.
- Perform sufficient vendor due diligence as the market is largely small startups that may not be sufficiently funded or staffed to support large scale application demands.

Sample Vendors

Eideticom; NETINT Technologies; NGD Systems; Nyriad; Samsung Electronics;ScaleFlux

At the Peak

Quantum Computing

Analysis By: Martin Reynolds, Matthew Brisse, Chirag Dekate

Benefit Rating: High

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. The particles represent information as elements denoted as quantum bits (qubits). A qubit can represent all possible values of its two dimensions (superposition) until read. 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, but will disrupt areas such as organic chemistry (batteries), materials science and cryptography (security) where it will deliver results very hard to achieve with traditional methods. Quantum computing will also advance machine learning and optimization solutions in speed and/or quality.

Business Impact

For businesses where molecular structures such as drug discovery or materials science are important, quantum computers will enable designers to create tailored molecules with valuable and precise effects. These businesses are at risk of long-term disruption by patents developed using quantum computers. Quantum computers will also enable more efficient competitors in industries such as transportation and logistics where optimization can reduce costs and improve performance.

Drivers

- Significant investment by governments, major corporations and startups.
- Proven small-scale results from recently developed quantum systems.
- Demonstrations of foundational quantum technology using electrons, ions and photons.

- Frequent publication of novel algorithms based on quantum computing technology.
- Error correction algorithms and methods are under development.
- In-chamber control electronics are emerging to help scale qubit counts.
- Emerging control software that can abstract quantum hardware from quantum algorithms.

Obstacles

- Current demonstrated qubit technology is too noisy for error correction.
- External control electronics require too many signal lines to scale.
- High-value algorithms may have impractical readout times.
- Quantum computing may lose funding because compelling results remain far off.
- Lack of enterprise readiness to understand and utilize quantum computing.
- Lack of standardization across programming, middleware and assembler levels.
- Fragmented vendor landscape inhibiting customer adoption and engagement.

User Recommendations

- **Hire quantum capable individuals into other roles now:** When quantum computing becomes relevant to your organization, even a few quantum-capable employees will make a material difference to your organization's ability to make use of this tricky technology.
- **Plan for speed-critical quantum optimization projects for 2025 through 2028:** The first applications for quantum systems will be those where speed of optimization is more important than accuracy. For example, truck loading, where any solution has to be delivered before the truck is dispatched.
- **Plan for performance-critical quantum optimization for 2028 through 2030:** By 2028, optimization should be both faster than and superior to conventional technologies. If you use optimization, plan for quantum systems to deliver superior business results.

- **Plan for chemistry and materials science innovations from 2032 onwards:** By 2032, our model shows that quantum computers could deliver otherwise impossible innovations in materials science and biochemistry.

Sample Vendors

1QBit; D-Wave; Honeywell; IBM; IonQ; Microsoft; PsiQuantum; QC Ware; Rigetti Computing; Zapata Computing

Gartner Recommended Reading

Strategy Guide to Navigating the Quantum Computing Hype

Cool Vendors in Quantum Computing

Quantum Computing Planning for Technology General Managers

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 DNN computations utilized in a wide range of artificial intelligence applications. These chips can be deployed in either data center servers, edge computing systems or endpoint devices.

Why This Is Important

An increasing range of applications require the use of DNN-based AI techniques to analyze captured data. These include object detection and classification in images and video streams, natural language processing, social media recommendation engines, autonomous vehicles and pharmaceutical analytics. To effectively execute many of these applications requires the use of data center and edge computing systems, and endpoint devices that include 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, audio streams enabling video analytics, object detection and classification, image recognition, natural language processing 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 GPUs of general purpose CPUs.

Drivers

- Increasing volume of complex unstructured data requires the use of AI techniques that leverage DNN models to analyze and enable business decisions to be made based on the data content.
- Executing DNN-based AI 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 AI-based workloads.
- Often trained AI 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, 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 and 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, having 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, examples include Google's Tensor Processing Units (TPUs) optimized for its TensorFlow-based applications.

User Recommendations

Application and software engineering leaders planning an effective long-term strategy for the use of DNN-based applications and hardware must:

- 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, formfactor and communications cost are critical.

Sample Vendors

Amazon Web Services (AWS); Google; Graphcore; Groq; Intel; NVIDIA; SambaNova Systems; Syntiant

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Forecast: AI Semiconductors, Worldwide, 2019-2025, 1Q21 Update

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Predicts 2021: Artificial Intelligence Core Technologies

Infrastructure Automation

Analysis By: Chris Saunderson

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Definition:

Infrastructure automation (IA) allows DevOps and I&O teams to design and implement self-service, automated delivery services across on-premises and cloud environments. IA enables DevOps and I&O teams to manage the life cycle of services through creation, configuration, operation and retirement. These infrastructure services are then exposed via API integrations to complement broader DevOps toolchains, or consumed via a self-service catalog.

Why This Is Important

As a discipline, infrastructure automation evolved from the need to drive speed, quality and reliability with scalable approaches for deploying and managing systems. DevOps and I&O teams are using IA tools to automate delivery and configuration management of their IT infrastructure at scale and with greater reliability.

Business Impact

By enabling engineers and developers to automate the provisioning and configuration of infrastructure, organizations will realize:

- Agility improvements — continuous integration and delivery of infrastructure.
- Productivity gains — version controlled, faster, repeatable deployment.
- Cost improvements — reductions in manual effort through increased automation.
- Risk mitigation — standardized configurations across all elements driving compliance.
- Efficiency — reducing toil and enabling value-added work.

Drivers

I&O leaders must automate processes and leverage new tools to mature beyond simple deployments of standardized platforms and deliver the systemic, transparent management of platform deployments. IA tools deliver the following key capabilities to support this maturation:

- Multicloud/hybrid cloud infrastructure orchestration
- Support for immutable infrastructure
- Enable consumption of programmable infrastructure
- Self-service and on-demand environment creation
- Support DevOps initiatives (continuous integration/delivery/deployment)
- Resource provisioning
- Operational configuration management efficiencies
- Policy-based delivery and assessment/enforcement of deployments
- Enterprise-level framework to enable tooling strategy maturation

Obstacles

- Combination of tools needed to deliver IA capability can lead to an increase in tool count.
- Software engineering skills and practices are required to get maximum value from tool investments.
- Migration from point activities to infrastructure capability releases is a steep learning curve.
- IA vendor capability overlaps muddies tool landscape.
- Steep learning curves can lead to developers and administrators riveting to known scripting methods to deliver needed capabilities.

User Recommendations

- Identify existing IA tools in use to catalog capabilities and identify use cases.
- Assess existing internal IT skills to incorporate training needs to enable IA more fully.
- Baseline how managed systems and tooling will be consumed (engineer, self-service catalog, API or on-demand).
- Integrate security and compliance requirements into evaluation criteria.
- Develop an IA tooling strategy that incorporates current needs and near-term roadmap evolution (cloud adoption/proliferation).

Sample Vendors

Amazon Web Services; Chef; HashiCorp; Inedo; Microsoft; Pulumi; Puppet; Quali; VMware

Gartner Recommended Reading

Market Guide for Infrastructure Automation Tools

Innovation Insight for Continuous Infrastructure Automation

How to Build Agile Infrastructure Platforms That Enable Rapid Product Innovation

The Future of DevOps Toolchains Will Involve Maximizing Flow in IT Value Streams

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

How to Lead Digital Disruption With Programmable Infrastructure

Assessing HashiCorp Terraform for Provisioning Cloud Infrastructure

Serverless Infrastructure

Analysis By: Arun Chandrasekaran

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

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

Why This Is Important

Accelerating the development and delivery of software is a core imperative for I&O leaders. Not only do serverless technologies enable organizations to build and deliver software faster, but they also entail low operational overheads and an elastic pricing model. Cloud providers such as AWS, Microsoft Azure, Google Cloud, IBM and Oracle, CDN vendors such as CloudFlare, Akamai and Fastly, and OSS vendors are all innovating and making serverless products available for a broad set of use cases.

Business Impact

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

Drivers

In the past few years, “serverless infrastructure” as a term has evolved to include much more than function as a service (FaaS) products. Examples of FaaS products are AWS Lambda, Azure Functions, Google Cloud Functions and IBM Cloud Code Engine. Currently, it refers not only to a programming model such as FaaS, but also to an operational model where all provisioning, scaling, monitoring and configuration of the compute infrastructure are delegated to the platform. Examples of such architectures include serverless containers (such as AWS Fargate and Microsoft Azure container instances) and serverless databases (such as Amazon Aurora and Google Cloud Storage).

The key drivers of serverless infrastructure are:

- Operational simplicity — It obviates the need for infrastructure setup, configuration, provisioning and management.
- “Built-in” scalability — Infrastructure scaling is automated and elastic.
- Cost-efficiency — You only pay for infrastructure resources when the application code is running.
- Developer productivity and business agility — Abstracts infrastructure management and allows developers to focus on writing code and application design.

Obstacles

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

User Recommendations

Serverless infrastructure does not mean the end of traditional I&O roles. However, it will significantly change the way I&O roles operate. To adapt to serverless realities, I&O leaders must:

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

Gartner Recommended Reading

A CIO's Guide to Serverless Computing

A CTO's Guide to Top Practices for Open-Source Software

Compute Evolution: VMs, Containers, Serverless — Which to Use When?

NVMe-oF

Analysis By: Julia Palmer, Joseph Unsworth, Joe Skorupa

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Nonvolatile memory express over fabrics (NVMe-oF) is a network protocol that takes advantage of the parallel-access and low-latency features of NVMe Peripheral Component Interconnect Express (PCIe) devices. NVMe-oF tunnels the NVMe command to the remote subsystems. The specification defines a protocol interface and is designed to work with high-performance fabric technology, including remote directory memory access (RDMA) over Fibre Channel, InfiniBand or Ethernet with RoCEv2, iWARP or TCP.

Why This Is Important

NVMe-oF is a method of extending access to nonvolatile memory (NVM) devices, using the NVMe protocol to remote storage systems. This enables a front-end interface into storage systems, scaling out to large numbers of NVMe devices and extending the distance within a data center over which NVMe subsystems can be accessed. The goal of NVMe-oF is to significantly improve data center network latency and to provide comparable-to-local latency for remote NVMe devices.

Business Impact

- NVMe-oF offerings impact business use cases where low-latency application and workload requirements are critical to the bottom line.

- Though requiring potential infrastructure changes and upgrades, the clear benefits these technologies can provide will immediately attract high-performance computing customers who can quickly show a positive ROI.

- NVMe flash media with the combination of NVMe-oF protocols can deliver architectures that extend and enhance the capabilities of storage arrays.

Drivers

- NVMe is a storage protocol that is used within solid-state arrays and servers. It takes advantage of the latest NVM to address the needs of extremely-low-latency workloads and is now broadly deployed. However, the NVMe-oF data center storage protocol is still emerging and developing at different rates depending on the network encapsulation method.

- The NVMe-oF protocol can take advantage of high-speed networks and accelerate the adoption of next-generation storage architectures, such as disaggregated compute, scale-out software-defined storage, and hyperconverged and composable infrastructures, bringing super-low-latency application access to the mainstream enterprise.

- Unlike server-attached flash storage, shared accelerated NVMe and NVMe-oF can scale out to high capacity with high-availability features and be managed from a central location, serving dozens of compute clients.

- Most storage array vendors have already debuted at least one NVMe-oF capable product with nearly all vendors expected to do so within a year.

- In 2018, the NVMe standards body ratified NVMe/TCP as a new transport mechanism. In the future, it’s likely that TCP/IP will evolve to be a very important data center transport mechanism for NVMe-oF.

- The application layer also plays a role in driving adoption. Support for NVMe-oF in 2020 (both FC-NVMe and RDMA over converged Ethernet) by VMware vSphere v.7.0 will help drive more mainstream usage; however, Microsoft is yet to make a similar announcement.

Obstacles

- Implementation of end-to-end NVMe-oF infrastructure could require substantial changes to not only storage platforms but also networking and servers depending on the existing infrastructure.
- I&O leaders are struggling to justify ROI for end-to-end NVMe deployments as only a small percentage of workloads will benefit from such uplift.
- The cost and complexity have impeded the adoption of NVMe-oF solutions in mainstream enterprises.
- Software support for NVMe-oF is still nascent but as it will expand and mature, I&O leaders will have an additional choice of either deploying NVMe-oF with RDMA RoCEv2 or NVMe-oF over TCP/IP-based products to leverage latest Ethernet deployment, thereby easing the transition and providing investment protection.

User Recommendations

- Identify workloads where the scalability and performance of NVMe and NVMe-oF-based solutions justify the premium cost of such deployment. Target it for AI/ML, high-performance computing (HPC), in-memory databases or transaction processing.
- Users should investigate any other potential infrastructure bottlenecks, and consult existing suppliers on potential performance and TCO gains to justify the ROI.
- Identify a potential storage platform, network interface controller, host bus adapter and network fabric suppliers to verify that interoperability testing has been performed and that references are available.
- Verify the availability of NVMe-oF networks for HCI deployments to see performance improvement.

Sample Vendors

Dell Technologies; Excelero; Hitachi Vantara; IBM; Lightbits; NetApp; Pavilion Data; Pure Storage; Vast Data; WekaIO

Gartner Recommended Reading

Emerging Technology Horizon for Communications

2020 Strategic Roadmap for Storage

Prepare Your Storage and Data Management Strategy for the Impact of Artificial Intelligence Workloads

Critical Capabilities for Solid-State Arrays

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

Sliding into the Trough

Container Management

Analysis By: Dennis Smith

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

To manage containers at scale, container management provides capabilities such as container runtimes, container orchestration and scheduling, and resource management. Container management software brokers the communication between the continuous integration/continuous deployment (CI/CD) pipeline and the infrastructure via APIs, and aids in the life cycle management of containers.

Why This Is Important

Container runtimes simplify use of container functionality and enable integration with DevOps tooling and workflows. Productivity and/or agility benefits of containers include accelerating and simplifying the application life cycle, enabling workload portability between different environments and improving resource utilization efficiency. Container management makes it easier to achieve scalability and production readiness. It also optimizes the environment to meet business SLAs.

Business Impact

Gartner surveys and client interactions show that the demand for containers continues to rise. This is due to application developers' and DevOps teams' preference for container runtimes, which have introduced container packaging formats. Developers have quickly 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

- Container runtimes, frameworks and other management software provide capabilities such as packaging, placement and deployment, and fault tolerance (for example, clusters of nodes running the application).

- The emergence of de facto standards (for example, Kubernetes) and offerings from the public cloud providers have simplified deploying containers at scale. Many vendors enable management capabilities across hybrid cloud or multicloud environments by providing an abstraction layer across on-premises and public clouds. 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.).

- Data analytics use cases have emerged over the past few years, as have operational control planes that enable the management of container nodes and clusters.

- All major public cloud service providers now offer on-premises container solutions. Independent software vendors (ISVs) are starting to package their software for container management systems.

- Some enterprises have scaled sophisticated deployments, and many more have recently commenced container deployments or are planning to. This is expected to increase as enterprises restart application modernization projects postpandemic.

Obstacles

- 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 their 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.

- More abstracted, serverless offerings may enable enterprises to forgo container management. Among these services are Knative, AWS Lambda and Fargate, Azure Functions, and Google's Cloud Run. These services embed container management in a manner that is transparent to the user.

- 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 PaaS providers' service offerings by experimenting with process and workflow changes that accommodate the incorporation of containers.
- Avoid open-source deployments unless the organization has ample in-house expertise to support.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Microsoft; Mirantis; SUSE (Rancher Labs); Red Hat; 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 data, deliver content, and perform analytics and decision making close to data producers (e.g., sensors and cameras) and data consumers (e.g., people and actuators). They have broader and more general capabilities than gateway servers, but are less powerful or multitenant than micro data centers. Edge servers connect to enterprise or cloud data centers and gateways that are connected to many local endpoints.

Why This Is Important

As IoT and data produced by things grow at the edge, 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 a critical part of most enterprises' infrastructure topologies and will be critical for digital business strategies.

Drivers

- Increased requirement for low-latency computing
- Increased data production at the edge (video, sensors, etc.) and the relative low cost of computing versus bandwidth
- Increased number of near-real-time digital interactions between people and things at the edge

Obstacles

- Technologies that enable high-volume remote management with zero touch are immature.
- Diversity of edge interactions drives a diversity of compute requirements.
- Scale requirements at the edge can be very small or very large and demand can grow quickly.

User Recommendations

- Choose edge servers that can be deployed rapidly and are extensible in the field to match changing requirements.
- Evaluate edge servers for zero-touch remote management.
- Avoid lock-in where possible and plan for technology changes as the market rapidly evolves.
- Make security an upfront design requirement in any edge server deployment.
- Consider as-a-service options rather than acquiring hardware and software.

Sample Vendors

ADLINK; Cisco; Dell Technologies; Eurotech; Hewlett Packard Enterprise (HPE); Lenovo

Gartner Recommended Reading

2021 Strategic Roadmap for Edge Computing

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Predicts 2021: Cloud and Edge Infrastructure

Micro OS

Analysis By: Thomas Bittman

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

A micro operating system (micro OS) is an OS designed to be extremely small and lightweight. It is used most frequently with cloud computing, edge computing and containers. A micro OS is intended for rapid deployment and horizontal scaling. It is especially suited for granular workloads (such as microservices architectures) or for use with lightweight virtual machine (VM) appliances. The size of a micro OS ranges from 150MB to 500MB.

Why This Is Important

Modern operating systems can have large footprints, be cumbersome to deploy and require relatively large platforms. As containers, microservices and edge computing deployments develop a range of smaller form factors, and more agile micro OS platforms 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 agile scaling to business needs. Micro OSs will be a core enabler to most new digital business applications — both in the cloud and at the edge.

Drivers

- A range of edge hardware footprints — from very small and embedded, to broader edge servers
- Agile deployment models
- Container-based solutions
- Microservices
- Edge computing deployments

Obstacles

- Micro OSs are facing decades of skills, process and application architectures centered on rich, general-purpose operating systems — 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 will survive.

User Recommendations

Users should evaluate micro OS technologies based on:

- Technical maturity and size
- Feature set (i.e., too much, not enough, just right)
- The viability of the vendor or the level of community support for open source
- The support and update technology provided by the vendor (which can be a subscription update service)
- Interoperability with intended cloud providers and orchestration technologies
- The fit with chosen container frameworks

Sample Vendors

Canonical; Microsoft; Rancher Labs; Red Hat; SUSE; VMware

Gartner Recommended Reading

Assessing Infrastructure Requirements for Deploying Kubernetes

Composable Infrastructure

Analysis By: Philip Dawson

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

Composable infrastructure enables resources to be aggregated through software-defined, intelligent administration and limited automation, enabling infrastructure and operations (I&O) leaders to achieve higher resource utilization and faster application deployment. Although some blade-based server infrastructures include composable networking features, composable infrastructure describes a broader spectrum of capabilities that includes disaggregation of accelerator, memory and storage resources.

Business Impact

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. Composable infrastructure helps deliver next-generation agile infrastructure, where fast development and delivery mandate rapid and continuous integration. Increased utilization of high-cost resources, such as GPU accelerators and storage-class memory, can yield financial savings.

Drivers

- Current composable implementations are limited, in that resources are pooled or restricted to using hardware from a single vendor. We saw modest steps toward greater vendor collaboration in the 2020 through 2021 time frame — for example, an agreement between next-generation, fabric consortia Compute Express Link (CXL) and Gen-Z Consortium to cooperate on standards.

- Most use cases for composable infrastructure are in multitenant environments, in which composability enables the efficient sharing of pools of accelerators or storage. Another current use case is in test and development environments, where infrastructure with varying characteristics must be repeatedly deployed.

Obstacles

- A key step in the maturity timeline for composable infrastructure will be core technology that can disaggregate DRAM from compute and balance the use of persistent memory. This is competing with DRAM and uncertain adoption.

- Composable is often tied to hardware features and functions as an alternative to software-defined infrastructure (SDI). This increases lock-in to a mixture of chassis, form factors and management tools beyond blades. It is not as portable as SDI.

- 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 infrastructure must be resized and administered frequently, or when composability increases the use of high-cost components.
- Don’t replace existing infrastructure to obtain composable infrastructure unless you have sufficiently mature automation tools and skills to implement composable features and yield benefits.
- Verify that your infrastructure management software supports composable system APIs, or that you have the resources to write your own management tools.
- Don’t avoid infrastructure with composable features. Rather, don’t choose such infrastructure, because of those features, unless you are prepared to use them and they don’t overlap with any third-party toolsets.

Sample Vendors

Cisco; Dell Technologies; DriveScale; GigaIO; Hewlett Packard Enterprise (HPE); Intel; Liqid

Gartner Recommended Reading

Understand the Hype, Hope and Reality of Composable Infrastructure

Drive Administration, Application and Automation Capabilities of Infrastructure-Led Disruption

Decision Point for Data Center Infrastructure: Converged, Hyperconverged, Composable or Dedicated?

The Road to Intelligent Infrastructure and Beyond

Next-Generation Memory

Analysis By: Martin Reynolds, Chirag Dekate

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Next-generation memory (NGM) is a type of nonvolatile memory capable of displacing DRAM in servers. It will have the density and manufacturing cost of flash memory, but will be fast enough to augment, or even replace, DRAM.

Why This Is Important

Next-generation memory is a critical technology required to maintain memory scalability growth. Both flash memory and DRAM memory are showing signs of reaching physical performance limits. DRAM and flash memory are limited by the number of electrons that fit in a memory cell. The continued increase in semiconductor layers (currently 176) in flash memory demonstrates that scalability is no longer possible by shrinking cell sizes and it will be increasingly difficult and costly to scale in 3D.

Business Impact

Systems using NGM will enable a 5- to 10-times increase in fast, local storage capacity, enabling scale-up computing systems to perform faster or handle larger analytics workloads. Alternatively, these systems can provide greater consolidation, reducing costs by shrinking the data center space required. In addition, a key impact of this technology will be to accelerate the adoption of in-memory computing architectures.

Drivers

The current, most promising technology for next-generation memory is phase-change memory (PCM), which uses nanometer-scale electromechanical structures to persistently store data.

- PCM can be built at the smallest practical silicon geometries, providing scalability.
- PCM can accommodate multiple layers, another approach to increased density.
- PCM can support both system memory and storage applications.
- PCM can be written and erased in small blocks.

PCM has other potential applications:

- IBM is researching PCM for future neuromorphic AI processors.
- Mythic is shipping a field-programmable neuromorphic device.
- PCM could improve capabilities of FPGAs.

Obstacles

Despite a promising start, PCM has faded. Intel has not succeeded in driving it into mainstream systems, and Micron, Intel's partner, has abandoned the technology. These issues expose substantial, but not insurmountable, barriers to success.

- PCM density must remain and exceed at least 50% of the price per GB of server DRAM.
- PCM performance needs to become closer to DRAM.
- PCM memories must advance in vertical semiconductor scaling technology.
- Software systems must adaptively accommodate PCM and DRAM in the same system.

User Recommendations

Next-generation memory will enable servers with main memory of more than 6 TB. These servers bring a new price point to buyers for large-scale transaction processing and data analytics.

Users with large workload requirements should evaluate PCM for:

- Large-scale analytics workloads where in-memory performance drives substantial performance gains
- Transaction processing systems where memory lookup tables otherwise exceed DRAM capacity
- Virtualization environments with many virtual machines (VMs) with low, but frequent, utilization

Sample Vendors

Dell Technologies; Hewlett Packard Enterprise (HPE); Inspur; Intel; Lenovo; Supermicro

Gartner Recommended Reading

Top 10 Technologies That Will Drive the Future of Infrastructure and Operations

Predicts 2020: Semiconductor Technology in 2030

Determining the Data Center Opportunity Created for 3D XPoint Persistent Memory

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 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 that efficient and for many applications it is better to utilize 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 extreme performance and power efficiency for a growing number of workloads.

- They are well-suited for AI inference workloads as they excel in low-precision (eight bit and 16 bit) processing capabilities in energy-efficient footprints.
- The use of FPGA accelerators can provide significant benefit to applications such as genome sequencing, real-time trading, video processing and deep learning (inference).

Drivers

- FPGA accelerators feature a large array of programmable logic blocks, reconfigurable interconnects and memory subsystems that can be configured to accelerate specific algorithmic functions. This allows the offload of tasks from the main system processor.
- In the data center, 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.
- The inference portion of deep-learning workloads is one of the biggest growth opportunities for FPGAs in the data center.
- The major FPGA vendors (Intel and Xilinx), along with a number of startups such as Mipsology and Swarm64, are working to address the challenge of programming FPGAs 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 Xilinx Vitis and Intel's oneAPI that lower the time and skills required to use FPGAs are now enabling accelerated adoption of FPGAs.
- The major cloud service providers, such as AWS, Microsoft and Baidu, are now offering FPGA-based instances that enable developers easier access to FPGA hardware.

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 (e.g., financial trading analytics, genome sequencing, etc.).
- Evaluate the availability of FPGA-based hardware for use in data center server deployments, such as FPGA-based PCIe add-in cards.
- Outline costs associated with necessary skill set and programming challenges of FPGAs and the maturity of the software-centric programming toolsets.
- Leverage cloud-based services for provisioning FPGAs (e.g., AWS EC2 F1 instances, Microsoft Azure, Baidu cloud) to minimize risks.

Sample Vendors

Amazon Web Services; Baidu; Illumina; Intel; Microsoft; Mipsology; Swarm64; Xilinx

Gartner Recommended Reading

Emerging Technologies: Neuromorphic Computing Impacts Artificial Intelligence Solutions

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Forecast: AI Semiconductors, Worldwide, 2019-2025, 1Q21 Update

Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Immutable Infrastructure

Analysis By: Neil MacDonald, Tony Harvey

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Immutable infrastructure is a process pattern (not a technology) in which the system and application infrastructure, once deployed into production, is never updated in place. Instead, when changes are required, the infrastructure and applications are simply replaced from the development pipeline.

Why This Is Important

Immutable infrastructure ensures the system and application environment is accurately deployed and 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 and often applied 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, one — repair the application or image in the development pipeline and re-release into production. The result is an improved security posture with fewer vulnerabilities and 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 and will drive greater adoption.
- 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 modify or patch runtime systems.
- Although immutable infrastructure may appear simple, embracing it 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 application stacks have elements that are deployed in the form of virtual machine images. VM replacement is slower and requires greater coordination than other workload components such as containers.

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 the individual components, versioned in a source-code-control repository, then redeployed for consistency.
- Prevent unauthorized change by turning off all normal administrative access to production compute resources — for example, by not permitting SSH or RDP access.
- Adopt immutable infrastructure principles with cloud-native applications first. Cloud-native workloads are more suitable for immutable infrastructure architecture than traditional on-premises workloads.
- Treat scripts, recipes and other code used for infrastructure automation similarly to the application source code itself, which mandates good software engineering discipline.

Sample Vendors

Amazon Web Services; Ansible; Chef; Fugue; Google; HashiCorp; Microsoft; Puppet; SaltStack; Turbot

Gartner Recommended Reading

How to Make Cloud More Secure Than Your Own Data Center

Top 10 Technologies That Will Drive the Future of Infrastructure and Operations

Programmable Infrastructure Is Foundational to Infrastructure-Led Disruption

Adapting Vulnerability Management to the Modern IT World of Containers and DevOps

OS Containers

Analysis By: Thomas Bittman

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

OS containers are a shared OS virtualization technology that enables multiple applications to share an OS kernel without conflicting. This is enabled by what is usually called a “container daemon,” which 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 used for years to increase the density of lightly used workloads (e.g., Oracle Solaris Zones, Virtuozzo and FreeBSD jails), focused on 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 computing.

Business Impact

Container technologies are part of a development architecture that will help enterprises become more agile, with applications that can change quickly, scale rapidly to demand, and improve the density of capacity use. In production, containers will be used strategically for new applications designed for agile development, rather than existing, monolithic application architectures. However, for developer ease of use, containers will also be used as wrappers for traditional workloads.

Drivers

- Lightweight overhead for small applications (improving capacity utilization and density).
- Ease of use and reuse by application developers.
- Aligns to microservices architecture and agile development.

Obstacles

- Reliance on operating system for application isolation creates security concerns, especially in multitenant environments.
- Unlike with hypervisors, existing applications require redesign to take full advantage and leverage the benefits of containers.
- Container use is constrained by tools and operations immaturity and complexity — 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

- Use containers when security and manageability concerns are easily mitigated.
- Combine containers with 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; Oracle; Red Hat; Virtuozzo; VMware

Gartner Recommended Reading

Best Practices for Running Containers and Kubernetes in Production

Arm Servers

Analysis By: Alan Priestley, Martin Reynolds

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 in performance versus x86-based designs. Arm's new Neoverse IP cores deliver performance equivalent to, or better than, the 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 benefit 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

The adoption of Arm servers is being driven by:

- Arm ISA is the dominant architecture in a wide range of personal computing applications, such as PCs, tablets and smartphones, and high-volume consumer and industrial endpoint devices. Many of these are battery-powered applications, where Arm-based designs hit the power-performance "sweet spot."

- Arm's development of its Neoverse server-optimized core designs has enabled companies including Amazon Web Services (AWS), Ampere and Marvell to design customized Arm processors.

- 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.

- Better performance enabled by successive generations of processor designs mitigates 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 with 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

The adoption of Arm server has been limited by:

- Many enterprise workloads are being 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 Intel server processors and the reemergence of AMD server processors are both headwinds against the growth of Arm servers in enterprises.

User Recommendations

IT leaders must:

- 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.
- 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 (AWS); Ampere; Arm; Marvell

Gartner Recommended Reading

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

Top Strategic Technology Trends for 2021: Distributed Cloud

Data Center Infrastructure Primer for 2021

Persistent-Memory DIMMs

Analysis By: Alan Priestley

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Persistent-memory dual in-line memory modules (PM-DIMMs) are nonvolatile DIMMs that reside on the double data rate (DDR) DRAM memory channel. Unlike DRAM, PM-DIMMs are able to retain memory contents through a power failure. These devices integrate nonvolatile memory (either NAND flash or 3D XPoint) and a system controller chip and are also referred to as solid-state DIMMs.

Why This Is Important

DIMMs connect directly to a dedicated memory channel rather than a storage channel and do not face the data transfer bottlenecks of a traditional storage system. As a result, PM-DIMMs can achieve drastically lower latencies (at least 50% lower) than any existing solid-state storage solution and can be viable alternatives to DRAM memory, if the slower access speeds are acceptable.

Business Impact

This technology's impact on users will be overall improved system performance.

- The specific workloads expected to see early adoption are in the in-memory computing, virtualization, analytics, AI and HPC segments.
- Traditional storage subsystems will be impacted as applications are rearchitected to take advantage of large amounts of nonvolatile memory accessible as part of the main server system memory.

Drivers

- As workloads and datasets expand, there is an ever-growing demand for large memory arrays within servers.
- Modern processors can support up to 6TB of memory via their onboard memory controllers, and many systems are now deployed with over 1TB of memory installed.
- DRAM is the main memory technology in use today, driven by its fast read and write access times, but it is a volatile memory technology, and system power fails will result in loss of data stored in DRAM.
- The use of external storage for backup and recovery is standard practice, but recovering terabytes of data from external storage systems takes a significant time.
- PM-DIMMs offer a high-density nonvolatile memory solution that resides in the same memory channel as DRAM, with typically lower cost per bit than DRAM and higher memory densities.
- Being nonvolatile, PM-DIMMs can retain data through a power outage and enable faster system recovery.
- Memory technologies such as Intel's 3D XPoint have nearly replaced the use of NAND flash on PM-DIMMs, with the benefit that they can achieve the higher capacity of current DRAM DIMMs and at a lower cost per gigabyte.
- The second generation and later of Intel's Xeon Scalable processors integrate native support for 3D XPoint-based PM-DIMMs.

Obstacles

- Market adoption of PM-DIMMs has been hampered by the slow write performance and limited endurance of existing flash memory technologies.
- Use of any PM-DIMM requires a mix or all of the following — support by the host chipset, software optimization of the OS and applications, and optimization for the server hardware. Systems deploying PM-DIMMs will also require the installation of standard DRAM-based DIMMs to complement the PM-DIMMs. This is necessary to provide the operating system and applications with an area of memory capable of sustaining frequent high-speed write accesses.
- Greater adoption will require support across a wide range of server vendors, operating systems and applications.
- Use cases for persistent-memory technologies will need to spread beyond the extremely high-performance, high-bandwidth and ultralow-latency applications for which they are attracting most interest today.

User Recommendations

IT professionals should:

- Analyze their workloads to determine performance and memory capacity demands along with software vendor support for PM-DIMMs. Major software vendors, such as SAP and Oracle, have implemented support for PM-DIMMs, and some of the major cloud service providers are also leveraging PM-DIMMs in their cloud services.
- Assess the roadmaps of the major server and storage OEMs. When evaluating PM-DIMM deployments, consideration should also be given to total cost of ownership (TCO) comparisons between nonvolatile memory and DRAMs, especially with the dynamic pricing fluctuations in the DRAM market.
- Be aware that specific versions of servers and their firmware, applications, operating systems and drivers will be required to support PM-DIMMs. In addition, Intel's 3D XPoint-based PM-DIMMs are currently compatible only with servers that deploy the second generation or later Xeon Scalable processors.

Sample Vendors

Dell Technologies; Hewlett Packard Enterprise (HPE); Intel; NetApp; Oracle; SAP

Gartner Recommended Reading

Determining the Data Center Opportunity Created for 3D XPoint Persistent Memory

Predicts 2020: Semiconductor Technology in 2030

Vendor Rating: Intel

Forecast Analysis: NAND Flash, Worldwide

Climbing the Slope

Hardware-Based Security

Analysis By: Neil MacDonald, Tony Harvey

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

Hardware-based security uses chip-level techniques for the protection of critical security controls and processes in host systems independent of OS integrity. Typical control isolation includes encryption key handling, secrets protection, secure I/O, process isolation/monitoring and encrypted memory handling.

Why This Is Important

Adoption is increasing as hardware-based isolation capabilities are becoming standard in most hardware devices and cloud-based IaaS offerings, including emerging confidential computing offerings. These approaches strongly isolate parts of the system (and typically its security controls) from a breach of the application or OS. Interest in strong isolation techniques is rising in the face of ongoing disclosures of new types of side-channel attacks and requirements for cloud and data sovereignty.

Business Impact

If an OS is compromised, its security controls can be disabled and sensitive data in memory stolen; hardware-based security can prevent this. Hardware-based security can significantly reduce attack surfaces across computing devices, but these capabilities require support from operating system software and system management software. Upgrading to more recent versions of software and cloud providers, which use hardware-based security features, can materially increase system security.

Drivers

- The desire to extend trust from the hardware level of a system through the OS to applications and workloads, including containers that run above it. This root of trust needs a strong foundation in hardware.

- Software-based isolation of security controls is inevitably fallible and will be attacked, increasing interest in protection approaches rooted in hardware.
- The desire to use IaaS providers in potentially hostile parts of the world and protect these workloads from OS compromise or virtual machine and memory snapshotting is increasing.
- Most hardware platforms for servers and mobile devices, including Android and iOS devices, now include hardware-based isolation capabilities.
- Requirements for data sovereignty enabled by public cloud confidential computing offerings are driving demand for isolation approaches rooted in hardware.

Obstacles

- In public clouds, enterprises don't have access to the underlying hardware and must rely on hardware-based attestations provided by the CSP.
- Approaches to hardware-based confidential computing vary across microprocessor vendors, complicating application deployment using these techniques. No single approach covers all use cases. Abstraction layers, such as Asylo, may help but add another layer of complexity and are not widely adopted.
- Hardware-based security is strong, but may potentially still be broken by software flaws or side-channel attacks such as Spectre and Meltdown.

User Recommendations

- Patch and remain vigilant for unexpected breaches. For systems under direct enterprise control, implement a BIOS-level patching strategy to deal with exposures that require BIOS-level remediation.
- Make strong isolation of sensitive code and security controls a mandatory part of IT systems procurement, including IaaS.
- Evaluate the need for confidential computing capabilities only for the most critical applications in systems that move to public cloud infrastructure, to protect sensitive operations such as key management and sensitive intellectual property.
- Check for compatibility issues with third-party approaches that also use virtualization techniques, before activating Windows 10 virtualization-based security.

- Explore the use of hypervisor-based approaches with security rooted in hardware virtualization techniques as another way to achieve similar levels of strong isolation.
- Plan different strategies for different devices and server platforms as none of these mechanisms are interoperable.

Sample Vendors

Amazon Web Services (AWS); AMD; Apple; Bitdefender; Fortanix; Google; Hysolate; Intel; Microsoft; Samsung Electronics

Gartner Recommended Reading

Market Guide for Cloud Workload Protection Platforms

How to Make Cloud More Secure Than Your Own Data Center

Select the Right Key Management as a Service to Mitigate Data Security and Privacy Risks in the Cloud

Security Leaders Need to Do Seven Things to Deal With Spectre/Meltdown

Cloud Computing

Analysis By: David Smith

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

Cloud computing is a very visible and hyped topic and, even though it has passed the Trough of Disillusionment, it remains a major force in IT. Every IT vendor has a cloud strategy — although some strategies are better described as "cloud inspired." Users are unlikely to completely abandon on-premises models, but there is continued movement toward consuming more services from the public cloud. Much of the cloud focus is on benefits beyond cost savings, such as agility, speed and innovation.

Business Impact

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

Drivers

- Although the peak of hype has long passed, cloud still has more hype than most other technologies that are at or near the Peak of Inflated Expectations. Variations, such as private cloud computing, hybrid and distributed approaches, compound the hype and continue to drive cloud computing.
- Other drivers include cloud variations (such as hybrid IT and multicloud environments), which are now at the center of where the cloud hype currently is. Additionally, there are different types of cloud services (such as IaaS, PaaS and SaaS) at various stages of industry hype, as well as cloud-complementary approaches such as edge.
- New and advanced use cases for cloud introduce even more terms such as edge, distributed cloud, multicloud and cloud-native. These add to the overall cloud hype as well as the applicability of cloud to more and more scenarios, including enabling next-generation disruptions. We expect to see more such terms in the future as cloud evolves, which will continue to keep cloud hype high.
- An increasing number of tools, applications and platforms are only available in a cloud paradigm.
- New offerings specific for industries (industry clouds) and modifications meant to satisfy increasing sovereignty issues will continue to evolve.

Obstacles

- *Cloud computing* continues to be one of the most hyped terms in the history of IT. Its hype transcends the IT industry and has entered popular culture, which has further increased hype and confusion around the term. In fact, cloud computing hype is literally “off the charts,” as Gartner’s Hype Cycle does not measure amplitude of hype (meaning that a heavily hyped term such as *cloud computing* rises no higher on the Hype Cycle than anything else).
- The term “cloud” continues to be stretched and overused by a variety of vendors and users. In many cases, the true value is obscured or lost.
- The mistaken belief that cloud means that the service runs in some far away data center limits potential use cases. Hybrid and distributed cloud *are* cloud computing.
- There is the potential for global fragmentation of the broader cloud computing notion as governments or global regions attempt to create separate internet, web and cloud environments and standards.

User Recommendations

- Demand clarity from your vendors regarding cloud. Gartner’s definitions and descriptions (which align with other useful ones such as from National Institute of Standards and Technology [NIST]) of the attributes of cloud services can help with this.
- Examine specific usage scenarios and workloads, map your view of the cloud to that of potential providers and focus more on specifics than on general cloud ideas. Understanding the service models involved is key — especially an understanding of the shared responsibility model for security and other services.
- Beware of adopting cloud for the wrong reasons; it can lead to disastrous results. There are many myths surrounding cloud computing as a result of the pronounced hype.

Sample Vendors

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

Gartner Recommended Reading

Cloud and Edge Infrastructure Primer for 2021

The Cloud Strategy Cookbook, 2021

Revisiting the Top 10 Cloud Myths for 2020

Devise an Effective Cloud Computing Strategy by Addressing Five Key Areas

Hyperconvergence

Analysis By: Philip Dawson, Jeffrey Hewitt

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Definition:

Hyperconvergence combines storage, computing and networking into a single system in an effort to reduce data center complexity and increase scalability. Multiple servers can be clustered together to create pools of shared compute and storage resources, designed for convenient consumption. Sales models included appliances and reference architectures using certified servers or delivered as a service or in a public cloud.

Why This Is Important

IT leaders seeking solutions that obviate the requirement for proprietary, external controller-based storage that permit a single management interface across silos of infrastructure and that can be cost-effective for certain use cases (e.g., VDI, edge/IOT, hybrid cloud) will consider hyperconvergence as a viable option.

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.

Hyperconvergence is of particular value to midsize enterprises that can standardize on hyperconvergence and 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 HCI market is now trifurcating, centering around the data-center-led “hybrid cloud” cloud management use case, VDI use case and an “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.

- VMware vSAN utilization among VMware ESXi customers, and Microsoft Azure HCI and Storage Spaces Direct utilization among Microsoft Windows Server 2016 and 2019 Datacenter edition customers, are on the rise.

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

- Nutanix, an early innovator in HCIS appliances, has largely shifted to a software revenue model and continues to increase its number of OEM relationships and partners.

- 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 will also be a stepping stone toward public cloud agility. Meanwhile, suppliers are expanding hybrid cloud deployment offerings.

- A growing number of hyperconvergence suppliers are delivering scale-down solutions to address the needs of remote office/branch office (ROBO) and edge environments typically addressed by niche vendors.

Obstacles

- Applications designed for scale-up architectures (as opposed to scale-out) are unlikely to meet cost/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, networking is still split between hardware SDN and networking around SD-WAN driving edge deployments.
- For large organizations, hyperconverged deployments will remain another silo to manage.

User Recommendations

- Implement hyperconvergence for hybrid cloud infrastructure 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 DR 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.
- Choose HCI appliance options to enable scale-down optimization of resources for high-volume edge deployments.
- Ensure that clusters are sufficiently large to meet performance and availability requirements during single and double node failures, and require proof-of-concept to reveal any performance anomalies.

Sample Vendors

Cisco; Dell EMC; Microsoft; Nutanix; Pivot3; VMware

Gartner Recommended Reading

Magic Quadrant for Hyperconverged Infrastructure Software

Critical Capabilities for Hyperconverged Infrastructure Software

Toolkit: Sample RFP for Hyperconverged Infrastructure

The Road to Intelligent Infrastructure and Beyond

Entering the Plateau

GPU Accelerators

Analysis By: Alan Priestley, Martin Reynolds, Chirag Dekate

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Definition:

GPU-accelerated computing is the use of graphics processing units (GPUs) in servers to support the execution of the compute intensive workloads, used for the analysis of highly parallel datasets.

Why This Is Important

HPC and deep learning are essential to many digital business strategies. For this fast-growing workload, traditional enterprise ecosystems based on CPU-only approaches are not sufficient. They require the use of accelerator chips capable of applying a set of math operations, often floating point, to highly parallel datasets. GPUs meet this requirement in being programmable and able to efficiently apply math operations to large arrays of data points.

Business Impact

The use of of GPU accelerators benefits:

- HPC and deep learning applications that require the processing of large scale highly parallel datasets.
- Programmability challenges have been largely solved in GPU-accelerated computing by toolsets such as NVIDIA's CUDA.
- Cloud-hosted GPU environments enable easy testing and evaluation.

Drivers

- GPUs are highly parallel floating-point processors designed for graphics and visualization workloads, and are capable of delivering dramatic performance improvements over traditional CPUs.

- Over the last decade, NVIDIA and others have added programmable capability to GPUs, enabling software applications to access deep, fast-floating-point resources.
- GPU subsystems are actively deployed in two key markets: AI and high-performance computing (HPC).
- DNN-based AI technologies are maturing quickly, supported by open-source software frameworks from the large cloud providers. Today most of the DNN frameworks, including TensorFlow, Torch, Caffe, Apache MXNet and Microsoft Cognitive Toolkit, support GPU acceleration.
- The HPC market is a key segment requiring the use of GPU accelerators, where compute-intensive applications including molecular dynamics, computational fluid dynamics, financial modeling and geospatial applications can, in many cases, be dramatically accelerated.
- Although many application-specific integrated circuits (ASICs) are emerging that are capable of outperforming GPUs for these workloads, few offer the broad software ecosystem support and ease of programming of GPUs.

Obstacles

- Programming GPUs can be challenging due to parallelism requirements, execution order and code optimization. However, toolkits like NVIDIA's CUDA can dramatically lower the programming challenges.
- High performance GPU accelerators have very high power consumption, typically 350-400W, requiring the use of server designs capable of supporting large power supplies and high efficiency cooling solutions.
- Ensuring optimal GPU accelerator performance often requires they be deployed in servers with large memory subsystems and high performance networking to ensure maximum data throughput.

User Recommendations

- Carefully evaluate software and toolset maturity when selecting GPU compute platforms.
- Evaluate cloud versus on-premises solutions by assessing cost, bandwidth and data privacy.

- Avoid suboptimal, on-premises commitments by leveraging cloud deployments for initial proof of concepts and trials.

Sample Vendors

AMD; Intel; NVIDIA

Gartner Recommended Reading

An Action Plan for Growing AI-Accelerator-Enabled Server Revenue

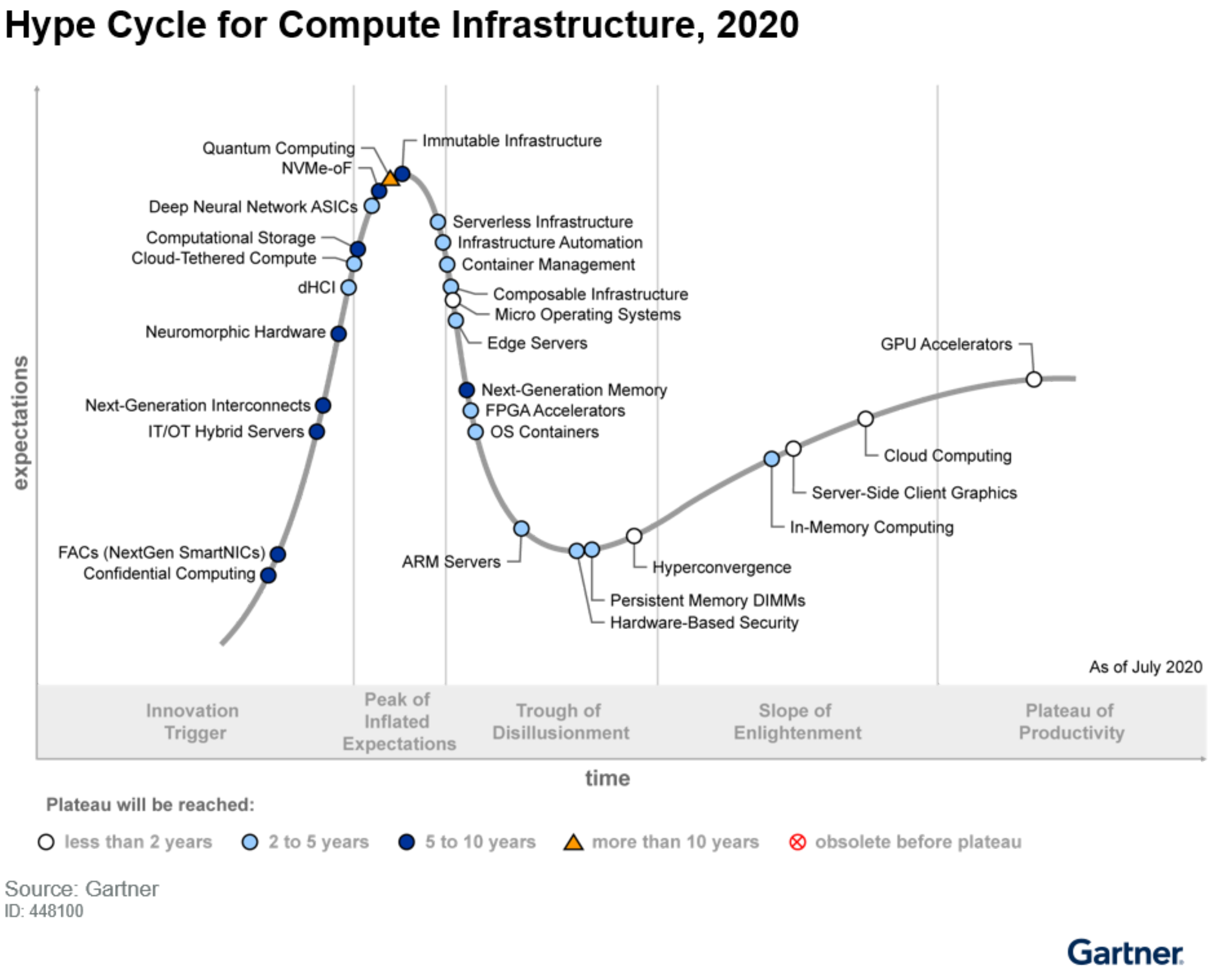
Emerging Technologies and Trends Impact Radar: Artificial Intelligence

Predicts 2021: Artificial Intelligence Core Technologies

Forecast Analysis: Discrete GPUs, Worldwide

Appendixes

Figure 2: Hype Cycle for Compute Infrastructure, 2020

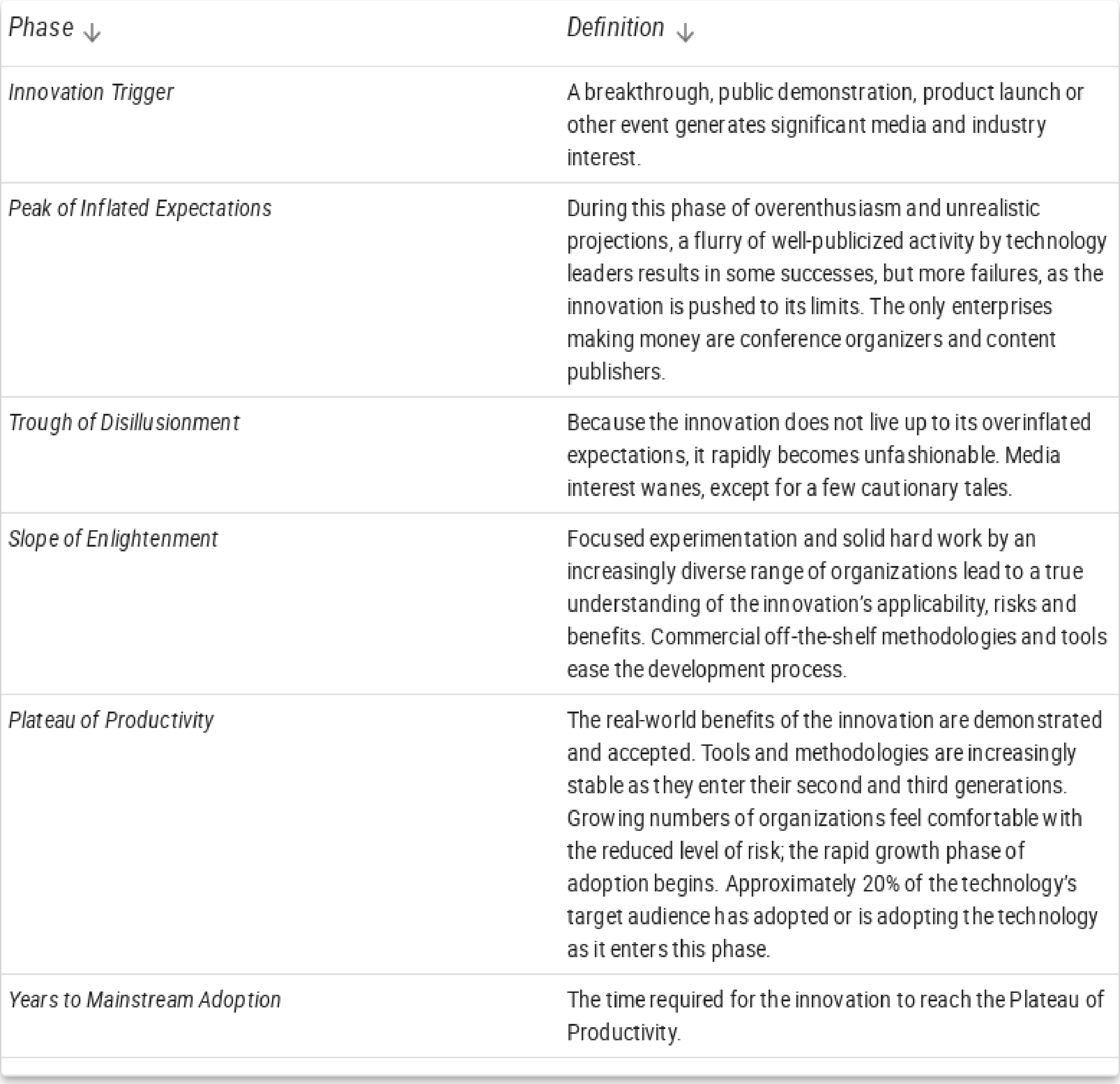


Source: Gartner (July 2020)

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 2021)

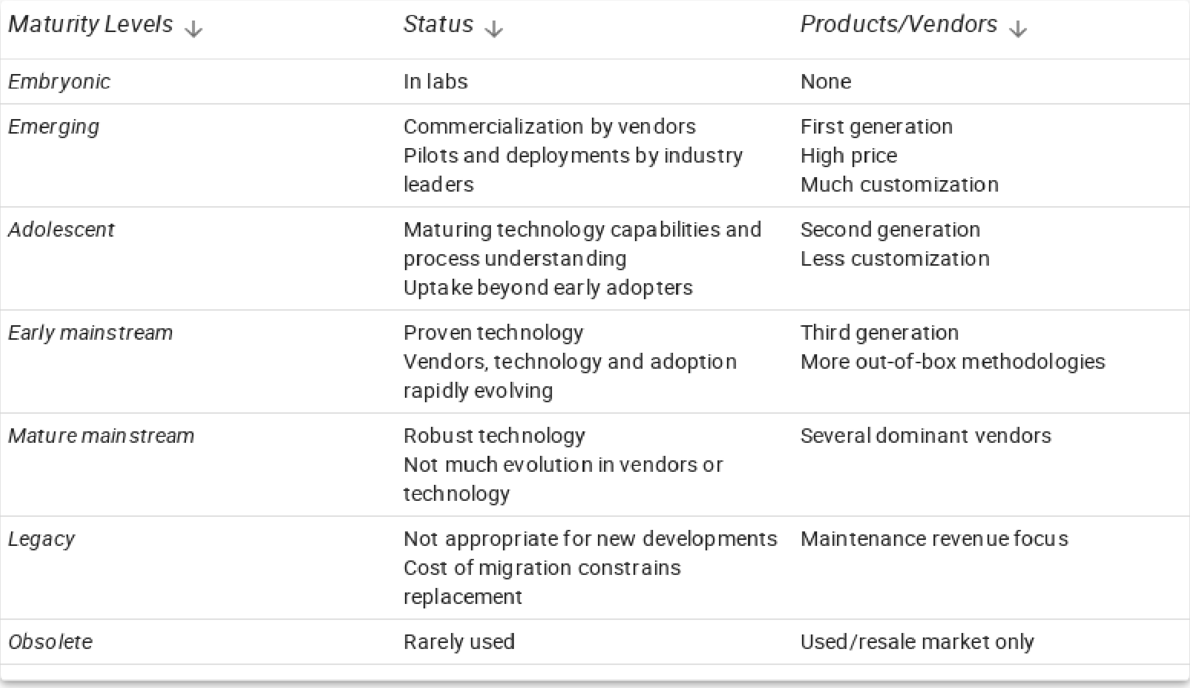
Table 3: Benefit Ratings

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 2021)

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 2021)

Evidence

2021 Hype Cycle Refinements

Gartner Hype Cycles produced in 2021 feature a number of refinements to the Hype Cycle methodology and presentation designed to make the Hype Cycle more accessible and useful. These refinements include:

Hype Cycle Graphic: The Hype Cycle graphic has been updated to provide increased clarity and differentiation between the various phases.

Interactive Hype Cycle: The Interactive Hype Cycle can be filtered to show results by time to plateau.

Priority Matrix: The Priority Matrix is now interactive. It can be used to navigate to various Innovation Profiles both within the Interactive Hype Cycle and the full document.

Innovation Profiles:

- The structure of the Innovation Profiles has been revised to provide a more consistent experience.
- Innovation Profile status data has been moved to the beginning of each profile.
- New or revised sections (Why This Is Important, Drivers, Obstacles) provide clearer, more-focused analysis.

Document Revision History

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 Author

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

2021-2023 Emerging Technology Roadmap for Large Enterprises

Understanding Gartner's Hype Cycles

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

2020 Strategic Roadmap for Compute Infrastructure

Table 1: Priority Matrix for Compute Infrastructure, 2020

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational	Cloud Computing OS Containers	Serverless Infrastructure	FACs (NextGen SmartNICs) Neuromorphic Hardware Next-Generation Memory	
High	GPU Accelerators Hyperconvergence	BMaaS Composable Infrastructure Container Management Deep Neural Network ASICs Edge Servers Infrastructure Automation	Computational Storage IT/OT Hybrid Servers NVMe-oF	Quantum Computing
Moderate	Micro OS	Arm Servers Cloud-Tethered Compute Consumption-Based Sourcing FPGA Accelerators Hardware-Based Security Persistent-Memory DIMMs	Confidential Computing Immutable Infrastructure Next-Generation Interconnects	
Low				

Source: Gartner (July 2021)

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 2021)

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 2021)

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 2021)