Hype Cycle for Compute Infrastructure, 2020

Published: 8 July 2020 **ID:** G00448100

Analyst(s): Tony Harvey, Daniel Bowers, Chirag Dekate

While AI, cloud and security remain hyped, COVID-19 instantly changed the priorities for I&O. For I&O leaders, this means urgently supporting the new imperatives around remote work and cost reduction while continuing to enable cloud, security and AI.

Table of Contents

List of Tables

List of Figures

Analysis

What You Need to Know

This document was revised on 2 October 2020. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

Businesses were already challenging I&O to provide increased agility, rapid deployment of new applications and better performance. The COVID-19 pandemic has only increased the urgency, accelerating demand for compute innovations that can scale rapidly and deliver more, more cost-effectively. Gartner has identified the most relevant innovations in the compute space for I&O leaders to evaluate and deliver on these challenges.

For more information about how peer I&O leaders view the technologies aligned with this Hype Cycle, please see "2020-2022 Emerging Technology Roadmap for Large Enterprises."

The Hype Cycle

Compute infrastructure includes technologies used inside the data center and in off-premises locations, including cloud and edge.

Cloud computing continues to transform I&O, and the rapid response required by the COVID-19 pandemic is driving more changes. Delivering IT services remotely will become the new normal for a significantly larger subset of end users. Strategic initiatives around migrating to cloud, remote work and automation will continue, and there will be a renewed focus on cost optimization.

Trends in compute infrastructure covered in this year's Hype Cycle include:

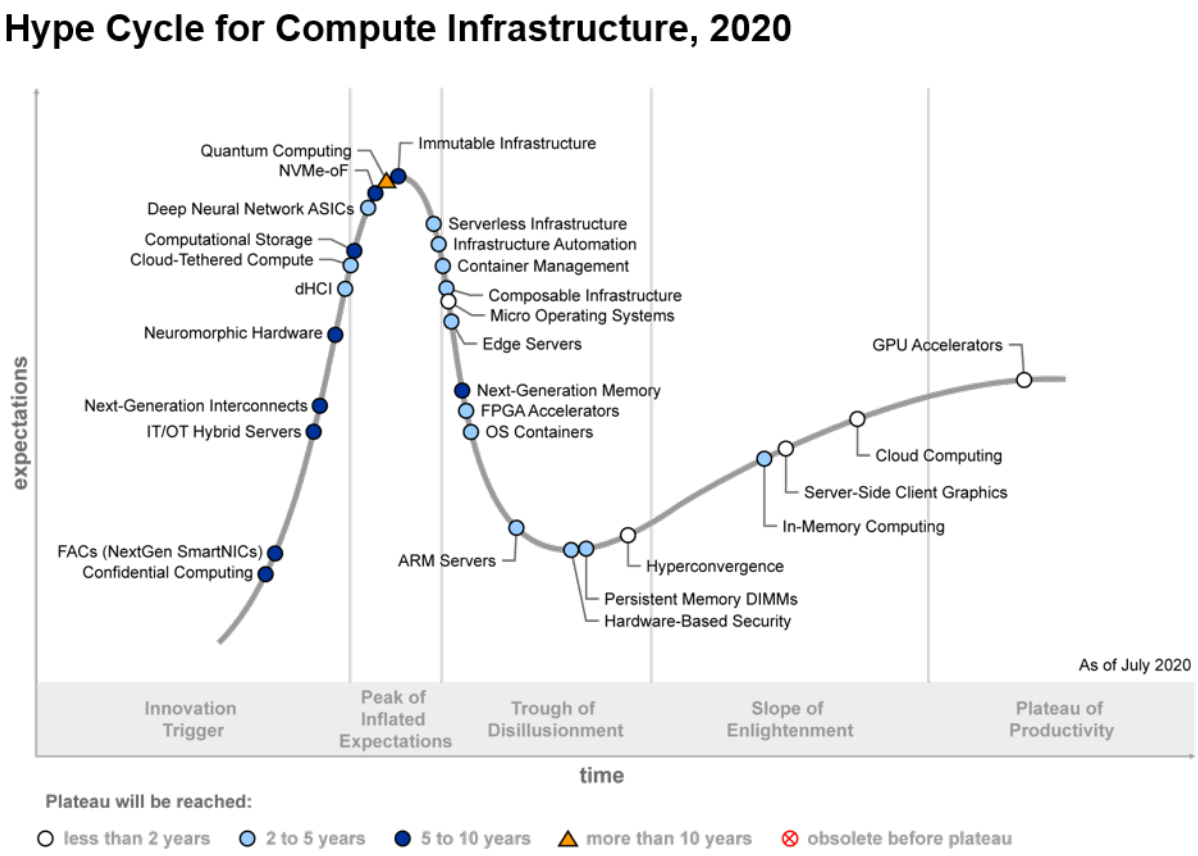
- **Agility and Security:** Confidential computing, immutable and serverless infrastructure, and containers show the continued need to deliver an agile and secure infrastructure.
- **Edge and Remote:** Edge computing once meant IoT. COVID-19 demonstrated that edge computing is where your users are. Innovation points are cloud tethered compute, edge servers and server-side client graphics.
- **AI and Automation:** Neuromorphic hardware, infrastructure automation, deep neural network ASICs and GPU accelerators will grow in importance as AI, machine learning and automation become embedded into the day-to-day operations of I&O.

While hyperconverged has accelerated up the curve toward the Slope of Enlightenment, driven by the need for on-premises infrastructure to become more automated and cloudlike, other technologies such as persistent memory DIMMs remain in the Trough of Disillusionment. Though some promised benefits may not have been delivered, there is value when used correctly, and further developments could push these technologies forward.

At the Peak of Inflated Expectations, quantum computing remains hyped but with few practical use cases, while immutable infrastructure and NVMe-oF are likely to move more rapidly around the curve.

New technologies appearing include function accelerator cards, dHCI and computational storage. Whether these technologies mature, become niche or fade into obscurity will depend on whether they can deliver business value in a rapidly changing environment.

Figure 1. Hype Cycle for Compute Infrastructure, 2020



Source: Gartner
ID: 448100

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.

Cloud computing will continue to transform the compute industry as it progresses toward the Plateau of Productivity. Over the next five years, IMC will become more widely used, and containers and serverless infrastructure will continue to deliver more flexible and adaptable platforms for application developers. Further out, I&O leaders should start to evaluate how function accelerator cards can be applied to increase performance for specialized applications.

High-impact technologies like infrastructure automation will start to drive increased productivity in early adopters, leading to more hype, while NVMe-oF will become more widespread in storage arrays and offer significant performance increases for some workloads. In the longer term, use cases for computational storage should be evaluated for potential competitive advantage.

Figure 2. Priority Matrix for Compute Infrastructure, 2020

Priority Matrix for Compute Infrastructure, 2020

benefit	years to mainstream adoption			
	less than two years	two to five years	five to 10 years	more than 10 years
transformational	Cloud Computing	In-Memory Computing OS Containers Serverless Infrastructure	FACs (NextGen SmartNICs) Neuromorphic Hardware Next-Generation Memory	
high	GPU Accelerators Hyperconvergence Server-Side Client Graphics	Composable Infrastructure Container Management Deep Neural Network ASICs Edge Servers Infrastructure Automation	Computational Storage IT/OT Hybrid Servers NVMe-oF	Quantum Computing
moderate	Micro Operating Systems	Cloud-Tethered Compute dHCI FPGA Accelerators Hardware-Based Security Persistent Memory DIMMs	Confidential Computing Immutable Infrastructure Next-Generation Interconnects	
low		ARM Servers		

As of July 2020

Source: Gartner
ID: 448100

Off the Hype Cycle

To provide readers with clearer, more focused research that supports their analysis and planning, we have included only those innovation profiles most strongly linked to the compute infrastructure Hype Cycle and its theme.

Removed from this Hype Cycle:

- OpenPower has not successfully expanded beyond a small number of vendors and has not been adopted as a mainstream solution
- Redfish is widely implemented across server vendors and has moved to mainstream adoption

Moved to other Hype Cycles:

- ML augmented DCs has moved to the Hype Cycle for infrastructure strategies
- Software-defined infrastructure has been moved to the Hype Cycle for infrastructure strategies
- Edge supercomputing has moved to the Hype Cycle for edge computing

These innovation profiles were renamed or consolidated:

- Scale out/scale up in memory computing have been consolidated into a single technology named in-memory computing
- NVMe and NVMe-oF have been renamed NVMe-oF
- Tethered compute was renamed cloud-tethered compute

On the Rise

Confidential Computing

Analysis By: Steve Riley

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.

Position and Adoption Speed Justification: In most cases, confidential computing implementations make use of Intel's Software Guard eXtensions (SGX) or ARM TrustZone. This architecture and set of instructions allow developers to isolate and execute code in private regions of memory unavailable to any process outside those regions, including processes at higher privilege levels. To take advantage of these enclaves, developers must specifically code for them (unless they add a software intermediary). AMD offers a similar technology called Secure Encrypted Virtualization (SEV); however, it doesn't shield code from the platform owner and doesn't offer integrity protection.

While confidential computing can be deployed on-premises, it is in the context of public cloud that most organizations express interest. 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 could offer their customers to alleviate these concerns. In a cloud provider's implementation, it would rely on a combination of hardware and software that attempts to prevent provider and unapproved third-party access to data in use.

However, the major cloud providers aren't rapidly adopting SGX or making it broadly available. Microsoft Azure offers one type of SGX instance in some regions, IBM offers SGX instances in Cloud Kubernetes Service bare metal hosts in most regions, and Alibaba supports SGX on bare-metal instances only. Intel's apparent lack of urgency at making SGX available in data center-class Xeon processors exacerbates the situation. Google has yet to announce its form of confidential computing. Amazon Web Services follows a different path: AWS Nitro Enclaves creates a separate instance that contains a single encrypted communications path to a parent instance and is hardware-isolated such that the parent instance has no access to the enclave's memory or compute. Gartner surmises that CSPs are reluctant to cede any part of their infrastructure to opaque elements over which they have little to no control.

Various emerging abstraction layers can eliminate the requirement for developers to code explicitly for a particular CPU's TEE. Asylo (by Google) and Open Enclave (by Microsoft) are available now. Carrying the abstraction layer notion even further, Fortanix Runtime Encryption removes the need to modify code, allowing existing applications to take advantage of enclaves. The Confidential Computing Consortium, a Linux Foundation project, aims to organize these efforts without imposing strict standards.

User Advice:

- Confidential computing isn't plug-and-play and should be reserved for the highest-risk use cases. It represents a high level of effort but offers diminishing marginal security improvement over more pedestrian controls like TLS, MFA, and customer-controlled key management services.
- Allocate time in the calendars of cloud application developers and cloud security architects to learn about the available confidential computing options and to experiment with the technology. Not all TEEs offer the same security guarantees or the same requirements for integration with existing and new code. In addition, confidential computing doesn't protect from vulnerable application code, and may contain its own exploitable vulnerabilities.
- 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.

Business Impact: Confidential computing potentially removes the remaining barrier to cloud adoption for highly regulated businesses or any organization concerned about unauthorized third-party access to data in use in the public cloud. It's likely that auditors and regulators will demand,

for certain data types, increased protection including high barriers to provider and government access. 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 is a marginally useful but technically challenging mechanism appropriate only for the most sensitive of use cases. Protecting data in use shifts trust to the CPU's ability to enforce enclave boundaries. However, hardware may not always be trustworthy, as the Spectre and Meltdown vulnerabilities demonstrated. Speculative execution is exhibiting many unexpected side effects and could be used to conduct side-channel attacks that reveal SGX-protected secrets. Furthermore, SGX (and other TEEs) are themselves interesting to attackers and are likely to attract their attention.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Alibaba Cloud; Fortanix; IBM Cloud; Intel; Microsoft Azure; Private Machines

Recommended Reading: "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

Definition: Function accelerator cards (FACs) are a class of network interface hardware that help improve and accelerate server availability, bandwidth performance and data transport efficiency in a network, besides enabling connectivity to a network. While all FACs are essentially NICs, not all NICs/SmartNICs are FACs. It comes with an in-built processor, onboard memory and peripheral interfaces, and is deployed either as an ASIC, an FPGA or an SoC and, hence, should be programmable in most cases.

Position and Adoption Speed Justification: FACs have started growing in popularity since end of 2019, and these have just started shipping in volumes from early 2020. They accelerate a variety of network-specific capabilities including network services, security and storage functions. FACs will play a critical role in modern cloud architecture and software-defined networking. Traditional NICs typically depend on servers to run complex networking stacks, such as encryption for security or

load balancing, etc., besides the basic network processing tasks. This takes away the processing capacity from VMs as the CPU time required to process each packet adds to the latency per packet before transferring it to a NIC. This results in latency issues. By offloading a whole host of network functions, services including security and storage from the server, FACs, improve server availability, bandwidth performance, and data transport efficiency besides enabling connectivity to a network.

However, these are still being adopted by hyperscale or other cloud service providers for various use cases now such as Amazon Web Services (AWS) Nitro System. They can be very useful at the edge where there are varying workload scenarios. Enterprises have minimal adoption until now. Given the benefits, FACs is carving out a separate market for itself, which will cannibalize the existing NIC market as well. By 2024, we estimate that one in three 10G+ and higher NICs shipped will be a FAC. FACs are estimated to grow at about 119% compound annual growth rate (CAGR; 2024-2019) while enterprises who have just started adopting it is estimated to grow at 115% CAGR (2024-2019).

User Advice: Discuss FAC options with network vendors or procure it directly and customize if you operate at a massive scale in order to optimize your server availability. This can be done through techniques such as load balancing to split workload across multiple physical or virtual servers and offloading of different server functions such as security (SSL), compression and virtualization.

It can be deployed either as a stand-alone device or integrated onto the motherboard as well.

However, not all SmartNICs available in the market are FACs and lead to server optimization so careful evaluation according to the use case is imperative.

Business Impact: FACs help improve:

- Server availability and thus optimize resource utilization (CPU cores, memory, storage and network) with lower latency and jitter.
- Flexibility due to programmability of FACs for a nondisruptive upgrade or adaptability for varying workloads whenever required. This will thus provide new, more agile systems which lead to overall network scalability and efficiency.

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Broadcom; Ethernity Networks; Mellanox Technologies; Pensando Systems

Recommended Reading: “Market Trends: Function Accelerator Cards Disrupting Traditional Ethernet Adapter Market”

IT/OT Hybrid Servers

Analysis By: Tony Harvey

Definition: IT/OT hybrid servers are designed to interface, collect and process data from operational technology systems that provide real-time control of physical systems and industrial process. These servers are placed in operational environments such as factories and mines and are designed to operate with higher resilience to shock, vibration, humidity and temperature than typical data center servers. They will also include industrial communications interfaces such as CAN bus, Modbus or Profinet and may include wireless or 5G technology.

Position and Adoption Speed Justification: IT/OT hybrid servers became available in 2018, but there has yet to be any major adoption of these systems. Although the servers appear to provide the necessary functionality and IoT is pushing systems further out toward the edge, there are still many barriers to adoption of IT/OT hybrid solutions. Industrial enterprises are cautious about the use of IT systems in industrial process control, where failure could result in loss of life or significant property damage and IT/OT hybrid server vendors need to demonstrate that they are secure and reliable before there is widespread adoption.

User Advice: IT/OT hybrid servers represent an opportunity and a risk for industrial enterprises. The potential benefits of using the data being generated by machines and industrial processes to drive cost efficiencies and deliver new solutions are enormous, but the security risks of allowing remote access to systems that, if interfered with, could result in loss of life, significant property or environmental damage and financial impacts must be resolved. Many OT systems were never designed to be connected to an open network such as the internet and have little to no security features.

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

- 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.
- Rationalize the costs for ongoing support, maintenance/updates and dependencies (networks, hardware, OSs, software and people) in a combined IT and OT environment.

Business Impact: IT/OT hybrid servers can help the enterprise 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. As use of IT/OT hybrids expands, the requirements for security management, cross-service coordination between IT and OT, data sharing and service-level management will grow to where most larger enterprises will need to create an integrated IT/OT group that has full responsibility for these solutions. Smaller enterprises may need to work with specialist IT/OT consultancies to assist with the integration in the early stages.

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.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Dell EMC; Hewlett Packard Enterprise; Lenovo; Schneider Electric

Recommended Reading: "Industrial Enterprise Customers Want to Accelerate IT-OT Convergence"

"IT/OT Convergence and Implications"

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

Next-Generation Interconnects

Analysis By: Daniel Bowers; Tony Harvey

Definition: Next-generation interconnects are a type of computer system bus designed for short-range, cache-coherent connection of data center processors, memory, accelerators, and I/O devices. These fabrics replace or build upon PCI Express (PCIe) or proprietary interconnects from processor vendors. Consortia formed by semiconductor vendors and hyperscalers to develop these interconnects include Cache Coherent Interconnect for Accelerators (CCIX), Compute Express Link (CXL), Gen-Z and Open Coherent Accelerator Processor Interface (OpenCAPI).

Position and Adoption Speed Justification: Nearly all I/O devices interconnect to the main system microprocessor via industry-standard PCIe interfaces, and interconnects between nodes use standardized network fabrics such as Ethernet and InfiniBand. However, emerging use cases including artificial intelligence and composable infrastructure require connections with higher bandwidth and cache-coherent shared memory space.

CCIX, CXL, Gen-Z, and OpenCAPI are open, consortia-driven initiatives, with specifications still being finalized, and hardware in the development or prototype stage. These fabrics promise greater flexibility for interconnecting systems and components. Future composable infrastructure is especially dependent on such fabrics to enable disaggregation of memory from processors.

Although OpenCAPI has been implemented in IBM's POWER9 processor, with compatible FPGA and flash storage devices available from companies such as Molex, it has not been widely adopted by other vendors. The market appears to be consolidating around CXL for connections local to the CPU, with all major CPU vendors (AMD, ARM, IBM, Intel) part of the CXL consortium alongside accelerator vendors and several hyperscalers. The CXL and Gen-Z consortia have signed a memorandum of understanding. Gen-Z has shifted focus onto shared pool of memory systems across longer distances compared to CXL's shorter-range CPU focus. CCIX has made its base

specification available publicly and signed an MOU with the ARM-focused Green Computing Consortium on common standards.

User Advice: Avoid using CCIX, CXL, or Gen-Z support as a criterion for selecting hardware today. While CXL is likely to emerge as the winner of these three, technical challenges remain, and new silicon must be released before production-ready products are available. Ask vendors for roadmaps and assessments of remaining technical roadblocks and verify that consortium members include vendors sufficient to implement the interconnects in all necessary components. Monitor which standard(s) are embraced by hyperscalers as their support will be a determining factor in the success and timeline of these standards.

OpenCAPI, as implemented in IBM Power Systems and OpenPOWER servers, is viable today for use cases involving OpenCAPI-compatible, specialized network interface adapters and customized FPGA accelerator cards from companies like Alpha Data, Mellanox Technologies and Molex.

Business Impact: Standards and consortia will evolve, as component vendors and hyperscalers try to influence the standards to enhance their own capabilities. Interest in next-generation interconnects will rise with vendor promises, but business impacts will be limited until standardized silicon and software are available.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

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

Recommended Reading: “Artificial Intelligence Creates New Semiconductor Business Opportunities”

“Forecast Analysis: Data Center Workload Accelerators, Worldwide”

“Predicts 2020: Semiconductor Technology in 2030”

Neuromorphic Hardware

Analysis By: Alan Priestley

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.

Position and Adoption Speed Justification: Neuromorphic systems continue to be at a prototype stage. IBM’s TrueNorth and exploratory work on multilevel phase change memory technologies, the European Union’s Human Brain Project (SpiNNaker and BrianScaleS), and BrainChip’s Spiking

Neuron Adaptive Processor technology are examples of neuromorphic hardware. Intel has developed a research chip, Loihi, and a range of servers and boards leveraging this chip to address a range of AI workloads: Loihi offers a higher degree of connectivity than competing implementations. Intel has also started training practitioners using its Loihi-based systems, as an early step to future adoption.

There are three major barriers to the deployment of neuromorphic hardware:

- Accessibility: Today GPUs are more accessible and easier to program than neuromorphic hardware; however, this could change when neuromorphic chips (NC) 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.

At the moment, these projects are not on the mainstream path for deep neural networks (DNNs), but that could change with a surprise breakthrough in programming techniques, however, ongoing working in developing chips for neuromorphic computing continues, for this reason we have moved neuromorphic hardware closer to the peak.

User Advice: Neuromorphic computing architectures leverage spiking neural networks and have the potential to deliver extreme performance for use cases such as deep neural networks and signal analysis at very low power. Neuromorphic systems are also simpler to train than DNNs, with the potential of in-situ training. Furthermore, neuromorphic architectures can enable native support for graph analytics. Most of the neuromorphic architectures today are not ready for mainstream adoption. However, these architectures have the potential to become viable over the next five years. I&O leaders can prepare for neuromorphic computing architectures by:

- Creating a roadmap plan by identifying key applications that could benefit from neuromorphic computing.
- Partnering with key industry leaders in neuromorphic computing to develop proof of concept projects.
- Identifying new skill sets required to be nurtured for successful development of neuromorphic initiatives.

Business Impact: Rapid developments in DNN architectures may slow advances in neuromorphic hardware but NC holds the promise of enabling extremely lower power AI development. There are likely to be major leaps forward in hardware in the next decade, if not from neuromorphic hardware, then from other radically new hardware designs.

Neuromorphic systems promise of lower power operation makes them uniquely suitable for edge and endpoint devices, where their ability to support object and pattern recognition can support image and audio analytics.

We are in the midst of an extremely rapid evolution cycle, enabled by radically new hardware designs, suddenly practical DNN algorithms and huge amounts of big data used to train these systems. Neuromorphic devices have the potential to drive the reach of AI techniques further to the edge of the network, and potentially accelerate key tasks such as image and sound recognition inside the network. They will require significant advances in architecture and implementation to compete with other DNN-based architectures.

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: BrainChip; IBM; Intel

Recommended Reading: "Emerging Technology Analysis: Neuromorphic Computing"

"Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

"Forecast Analysis: Data Center Workload Accelerators, Worldwide"

"Product Managers Developing AI Chips Must Clearly Identify Target Markets"

"5 Questions a Product Manager Must Ask When Creating an AI-Enabled Edge Product Strategy"

dHCI

Analysis By: Tony Harvey; Julia Palmer

Definition: Distributed HCI is a three-tier storage architecture using separate compute and storage nodes. Storage can be either scale-out external controller-based storage or dedicated SDS. dHCI offers the simplified management model of HCI while allowing compute and storage to be expanded independently.

Position and Adoption Speed Justification: External controller-based (ECB) and software-defined storage (SDS) vendors have introduced distributed HCI (dHCI) products to target customers who want the ease-of-use of hyperconverged systems, but also need asymmetric scaling of compute and storage. dHCI can also support bare metal workloads, and can provide more predictable latency and higher storage throughput than hyperconverged.

Datrium introduced the first dHCI solution (DVX) in 2016. Other vendors followed, including NetApp HCI in 2017 and HPE Nimble Storage dHCI in 2019. dHCI is gaining traction and enabling growth in integrated infrastructure systems (IIS) as more vendors enter the market.

User Advice: dHCI solutions offer many of the advantages of ECB storage solutions combined with the simplicity and VM-level storage provisioning capabilities of hyperconverged solutions. I&O leaders should evaluate dHCI solutions when they have workloads that:

- Require a mix of different server sizes and configurations.
- Consume large amounts of storage capacity.
- Have unbalanced compute and storage growth requirements.
- Demand extremely high transaction rates or throughput.
- Require predictable latency.

When considering implementing a dHCI solution, I&O leaders should:

- Identify specific workloads or initiatives where a dHCI system would be suitable.
- Implement jointly by server, storage, and virtualization teams, as skills and project alignment from all three is required.
- Deploy first as a proof of concept to ensure performance, availability, automation, and ease-of-use expectations are met.

Business Impact: dHCI can provide increased agility, reduced service costs as well as increase the likelihood of meeting service levels. I&O leaders who successfully implement a dHCI solution will view dHCI as a strategic investment that enables an automated, agile architecture that delivers the flexibility and scalability required for modern business.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Datrium; HPE; NetApp

Recommended Reading: "Market Share Analysis: Data Center Hardware Integrated Systems, Worldwide, 4Q19 Update"

"How I&O Leaders Should Leverage New dHCI Solutions"

Cloud Tethered Compute

Analysis By: Tony Harvey; David Wright

Definition: Cloud-tethered compute is a model where MaaS, IaaS, or PaaS are delivered in a customer-controlled environment but managed by the vendor via a network tether from a public or private cloud. The system may require the tether to be continuously connected, for billing, or be able to operate disconnected with periodic connectivity. Updates are managed by the vendor, removing the responsibility of maintenance of the platform from the I&O team.

Position and Adoption Speed Justification: Cloud-tethered compute solutions are starting to become more common. Initially, they were developed by public cloud vendors to provide public

cloud services such as IaaS and PaaS in client-controlled environments, for example, Azure Stack Hub and AWS Outposts. Other vendors have started to develop products in this space. For example, companies like Hivecell deliver an edge-focused solution; Dell EMC delivers a VMware-based private cloud with VMware Cloud on Dell EMC; and other vendors like HPE and Lenovo deliver metal as a service and other capabilities through consumption-based models. Current adoption is relatively low, although the potential for growth due to data sovereignty, connectivity and latency issues at the edge is significant.

User Advice: Cloud-tethered compute systems are relatively new to the market and may be missing key capabilities; and some entrants do not have clear SLAs and the commercial terms are not yet fully developed. Be careful when evaluating tethered compute systems to ensure that they provide the features that the development teams require and that you clearly understand the SLAs provided, what happens at the end of the contract, and how upgrades and expansions midterm are handled.

Key areas to consider when evaluating cloud-tethered compute systems:

- Provided as a service: Some or all elements are provided as a service and remain the property of the vendor. You must have a clear understanding of who owns which element and what are the consequences of a service shutdown.
- Connectivity: Does the system require a permanent connection for the tether, or can it operate in disconnected mode? What are the limitations when operating in disconnected mode?
- Response times and hardware maintenance services: Users more used to a Tier 1 vendor service contract for on-premises maintenance may struggle to adapt to a service model that is based on next-business-day, whole-unit replacement or lower.
- Contract terms: Understand what will happen at the end of the term and how any midterm changes, upgrades or additions will affect the termination date and costs.
- Security: Cloud-tethered compute systems typically use a “shared security model,” where some responsibilities belong to the customer and some to the vendor. Although the security of the vendor may be excellent, the customer’s security team must be engaged to ensure all security risks are mutually addressed.
- Variable pricing: Similar to cloud services, monthly or other recurring charges for a cloud-tethered system could vary based on usage, making costs and budgeting less predictable than traditional infrastructure.
- Data sovereignty: Although the data is held in a location under control of the user, the system is not under user control. At a minimum, any data being used on these systems should be encrypted at rest. In addition, cloud-tethered compute vendors should be able deliver secure data deletion and even storage device retention services to ensure that sensitive data does not leave the control of the user.

Business Impact: Cloud-tethered compute represents a new form of computing in which the customer’s IT gives responsibility for and control over some of its data center assets to a cloud provider. It reduces the need to perform many of the routine tasks traditionally performed by the IT staff, especially as related to server maintenance. It will, however, create new needs for skills in

contractual analysis, security and spend management for these systems. Application- and data-level security and backup will still remain an I&O responsibility. In many cases, the I&O function may welcome the removal of basic duties related to system maintenance, which will enable it to focus on delivering higher-level services.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: AWS Outposts; Google Anthos; Hivecell; Microsoft Azure Stack; Oracle; VMware

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

Definition: Computational storage (CS) combines processing and storage media to allow applications to run on the storage media, offloading host processing from the main memory of the CPU. The aim is to reduce data movement by processing data directly within 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 ASICs and low-power CPU cores on the SSD.

Position and Adoption Speed Justification: 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. CS-based systems may include data compression, encryption and redundant array of independent disks (RAID) management. As data volumes increase, movement of data becomes a bottleneck. As storage sizes normally vastly exceed memory, the data has to be read in from the storage media. This impedes application performance, undermining real-time analysis for most datasets. The principle that storage is separate from processing remains a core tenet of most enterprise IT systems. Data-intensive applications such as AI/ML, high-performance computing, analytics, high-frequency trading and immersive and mixed-reality streaming stand to benefit the most by removing the bottleneck. Edge computing remains an opportunity, along with applications that favor distributed processing. Furthermore, putting processing on the storage media may provide substantial performance gains, better memory

management and energy savings over storage systems comprising industry-standard solid-state media.

User Advice: The CS market is still in the development phase with a handful of vendors fielding POC, limited and small-volume production systems. Early use cases for CS are rapidly emerging, including machine learning processing, real-time data analytics, high-frequency trading, multimedia production, and higher-performance computing. Applications that work across multiple CS nodes will perform best and offer the greatest benefits. CS system architecture is more complex, 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.

I&O leaders are advised to explore possible benefits that can be gained from specific use cases, but should carefully weigh the cost vs. performance gains. This is especially true where certain workloads are very input/output-bound and would benefit the most from processing in storage. The segment is led by small startups, so I&O leaders are advised to perform sufficient due diligence. Also, monitor ongoing Storage Networking Industry Association (SNIA) CS Technical Work Group (TWG) developments where standards and interoperability work is being actively pursued with over 20 companies participating.

Business Impact: There is both a cost and time factor involved in shuffling terabytes of data around. CS can provide material performance benefits to data-intensive applications, especially in edge computing. Combined with its low power footprint, CS increases the performance-per-watt ratio, therein decreasing power consumption costs for applications at the edge. CS provides a programmable platform for value-added storage services such as erasure coding and database analytics functions, substantially reducing application and server/compute costs. CS is complementary to container workloads. Utilizing more powerful compute into solid-state media controllers will increase storage efficiencies and lower overall application costs, allowing the application to directly access the NAND flash chips inside the CS drives. A material increase in bandwidth between application and the solid-state media device/controller is achieved by taking advantage of the common flash interconnect channel standard. However, the challenge facing computational storage vendors is how to broaden the target audience beyond a few highly advanced customers with the resources and capabilities to perform significant internal application development and testing with clear ROI advantages.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Eideticom; NETINT Technologies; NGD Systems; Nyriad; Samsung; ScaleFlux

Recommended Reading: “Prepare Your Storage and Data Management Strategy for the Impact of Artificial Intelligence Workloads”

“Market Share Analysis: Solid-State Drives, Worldwide, 2018”

At the Peak

Deep Neural Network ASICs

Analysis By: Alan Priestley

Definition: A deep neural network (DNN) application-specific integrated circuit (ASIC) is a purpose-specific processor designed to execute the DNN computations utilized in a wide range of artificial intelligence applications.

Position and Adoption Speed Justification: DNN ASICS are being used in a divers set of data center, edge and endpoint applications, some examples include object detection and classification in images and video streams, natural language processing, social media recommendation engines, autonomous vehicles and pharmaceutical analytics.

There are two major phases of DNN-based AI application development:

- Training: Large volumes of known data are processed by the DNN ASIC. These operations are data throughput intensive and typically require the use of floating point math.
- Inferencing: New or unknown data is analyzed by the DNN. These tasks are latency dependent and can utilize integer math.

A majority of training and inferencing tasks currently use GPUs, but the use of DNN ASICs can deliver significantly higher performance and lower power consumption than CPUs or GPUs when executing DNNs.

While it is possible for the same DNN ASIC to be used for both training and inference tasks, devices are being developed that are optimized for a specific task and often for a specific class of DNN. The training phase typically takes place in a data center and leverage large scale designs, typically high power chips, optimized for data throughput. Having developed and trained a DNN-based AI application it is typically deployed on a DNN ASIC optimized for inference operations. These chips may be used in data center deployments but often will be utilized in edge or endpoint systems where there may be constrained formfactors and power resources, requiring highly optimized chip designs.

Many companies have announced plans to DNN ASICs for both training and inference ranging from traditional semiconductor vendors to startups. The large hyperscale cloud service providers are also developing ASICs optimized for their specific DNN-based workloads, examples include Google's tensor processing units (TPUs) optimized for its TensorFlow-based applications.

User Advice: The benefits of DNN ASICs in processing the highly parallel operations required for today's AI-based applications are significant. However, widespread use of DNN ASICs will require the standardization of neural network architectures and support across diverse DNN software development frameworks. Plan an effective long-term DNN strategy comprising DNN ASICs by choosing ASICs and vendors that offer or support the broadest set of DNN frameworks to deliver business value faster. General purpose CPU vendors are also adding new instructions to their CPUs

to support DNN-based workloads and these should also be evaluated when assessing the use of ASICs to accelerate these DNN-based applications.

Business Impact: Hardware acceleration will enable DNN-based workloads to address more opportunities in a business through improved cost and performance. Use cases that can benefit from DNNs include video analytics and object detection, image recognition, natural language processing and recommendation systems.

IT leaders deploying deep neural network applications should include DNN ASICs in their planning portfolio. We expect this market to mature quickly, possibly within the three-year depreciation horizon of new systems.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon; Baidu; Google; Graphcore; Hailo; Huawei; Intel; SambaNova Systems; Syntiant

Recommended Reading: "Forecast Analysis: Data Center Workload Accelerators, Worldwide"

"Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

NVMe-oF

Analysis By: Julia Palmer; Joseph Unsworth; Joe Skorupa

Definition: Nonvolatile memory express over fabrics (NVMe-oF) is a network protocols that is taking advantage of the parallel-access and low-latency features of NVMe PCIe devices. NVMe-oF enables tunneling the NVMe command set over additional transports beyond PCIe over various networked interfaces to the remote subsystems across a data center network. The specification defines a common protocol interface and is designed to work with high-performance fabric technology including RDMA over Fibre Channel, InfiniBand or Ethernet with RoCEv2, iWARP or TCP.

Position and Adoption Speed Justification: NVMe is a storage protocol that is being used within solid-state arrays and servers. It takes advantage of the latest nonvolatile memory to address the needs of extreme-low-latency workloads and is now broadly deployed. However, NVMe-oF, which requires a storage network, is still emerging and developing at different rates depending on the network encapsulation method. Today, many NVMe-oF offerings that use fifth generation and/or sixth generation Fibre Channel (FC-NVMe) are available, but adoption of NVMe-oF within 25/50/100 Gigabit Ethernet is slower. In November 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 an important data center transport for NVMe-oF. The NVMe-oF protocol can take advantage of high-speed networks and will 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 during 2021.

User Advice: Buyers should clearly identify workload where the scalability and performance of NVMe-based solutions and NVMe-oF justify the premium cost of an end-to-end NVMe deployment. There are a select variety of highly performant workloads that can utilize the technology, such as AI/ML, high-performance computing (HPC), in-memory databases or transaction processing. Next, identify appropriate potential storage platform, NIC/HBA and network fabric suppliers to verify that interoperability testing has been performed and that reference customers are available. Should buyers not desire the immediate performance gains and associated costs, then they should investigate how simply and nondisruptively their existing products migration path to ensure investment protection for the future.

Most storage vendors already offer solid-state arrays with internal NVMe storage, and during the next 12 months an increasing number of infrastructure vendors will offer support of NVMe-oF connectivity to the compute hosts. HCI vendors will deliver NVMe storage in an integrated offering during the next 12 to 18 months, but customers need to verify the availability of NVMe-oF networks between HCI nodes to see significant performance improvement. Similarly, when customers require NVMe-oF storage networks that encompass switches, host bus adapters (HBAs), and OS kernel drivers, IT infrastructure modernization will be required.

This potential requirements of making major changes to the data center storage networking and servers are slowing down the adoption of NVMe-oF solutions in mainstream enterprises. However, due to the better interoperability and availability of NVMe-oF over Fibre Channel within the next two years, I&O leaders implementing NVMe-oF will likely choose to deploy it within an existing Fibre Channel SAN infrastructure. Investment protection for customers with existing fifth-generation or sixth-generation FC SANs is compelling because customers can implement new fast NVMe storage arrays and connect via NVMe-oF to servers while using the same media. Therefore, old and new storage, network switches and host bus adaptors can run together in the same FC-based storage network (SAN), with SCSI and NVMe storage separated by zones, as long as compatible fifth- or sixth-generation FC equipment is used. Furthermore, as support for NVMe-oF expands and matures, 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.

Business Impact: Today, NVMe SSDs and NVMe-oF offerings can have a dramatic impact on business use cases where low-latency requirements are critical to the bottom line. Though requiring potential infrastructure enhancements, the clear benefits these technologies can provide will immediately attract high-performance computing customers who can quickly show a positive ROI. Designed for all low-latency workloads where performance is a business differentiator, NVMe SSD with NVMe-oF will deliver architectures that extend and enhance the capabilities of modern general-purpose solid-state arrays. Most workloads will not need the multimillion IOPS performance that

these new technologies offer, but most customers are demanding the lower, consistent response times provided by NVMe-based systems.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Dell EMC; Excelero; Hitachi Vantara; IBM; Kaminario; Lightbits; NetApp; Pavilion Data Systems; Pure Storage; StorCentric

Recommended Reading: "Top 10 Technologies That Will Drive the Future of Infrastructure and Operations"

"2019 Strategic Roadmap for Storage"

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

"Critical Capabilities for Solid-State Arrays"

Quantum Computing

Analysis By: Martin Reynolds; Matthew Brisse; Chirag Dekate

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 can address problems with vast combinatorial complexity.

Position and Adoption Speed Justification: Quantum computers are not general-purpose computers. Rather, they are accelerators capable of running a limited number of algorithms with orders of magnitude of speedup over conventional computers. These problems fall into a broad category of optimization, where a traditional algorithm would take impossibly long to find a solution. Quantum computers are superior for problems with small input and output, but enormous combinatorial complexity. Quantum computers will scale using a mix of existing technology, and a combination of new algorithms and conventional computing.

Hardware based on quantum technology is unconventional, complex and leading-edge. Current qubits are good enough to demonstrate the potential of quantum computing, but quantum systems face challenges in scale, noise and connectivity that require yet unknown breakthroughs to offer business value.

The technology continues to attract significant funding, and a great deal of research is underway at many university and corporate labs. Most practitioners are working on gate-model quantum

computers, which sequence the qubits through operations that prepare input data and create a solution. In such systems, input data is embodied in the program steps.

Quantum annealing is a style of quantum computing that uses entanglement and superposition in a way more akin to analog computing. D-Wave's quantum annealers, with thousands of qubits, are demonstrating practical solutions, but also face scaling and noise challenges, and do not yet deliver a business advantage.

An alternative solution, Fujitsu's Digital Annealer, offers the equivalent of 8,000 slow, but high-quality, qubits in a rack mount package. Digital Annealers may solve problems otherwise assigned to quantum computers, potentially delaying the point at which quantum computers offer business advantage.

User Advice: In the few known applications, quantum computers could operate exponentially faster than conventional computers. Quantum computers will act as accelerators to problems preprocessed and postprocessed by digital computers.

Early applications will likely be in the class of optimization, such as routing or portfolio optimization. Later applications will address organic chemistry, drug discovery, materials science and code breaking (as prime number factoring).

Quantum computers will eventually compromise today's cryptographic key exchange protocols. Quantum-safe cryptographic algorithms are in the final stages of the standardization process, and should be a midterm strategic initiative for organizations where data must be protected over decades.

If a practical quantum computer offering appears, check its usefulness across the range of applications that you require. It will probably be dedicated to a specific application, and this is likely to be too narrow to justify a purchase. For those customers interested in quantum computing, Gartner recommends the use of quantum as a service (QaaS). QaaS providers such as IBM, Rigetti Computing, Xanadu, D-Wave and others are offering application development and test tools that work with remote hardware.

Business Impact: Quantum computing could have a huge effect, especially in areas such as optimization, machine learning, cryptography, drug discovery and organic chemistry. Although outside the planning horizon of most enterprises, quantum computing could have strategic impacts in key businesses or operations.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: 1QBit; Alibaba Cloud; D-Wave; Google; IBM; Microsoft; QC Ware; QinetiQ; Rigetti Computing; Zapata Computing

Recommended Reading: "Strategy Guide to Navigating the Quantum Computing Hype"

"Emerging Technology Analysis: Act Now on Quantum-Safe Encryption or Risk Losing Deals"

"Quantum Computing Planning for Technology General Managers"

Immutable Infrastructure

Analysis By: Steve Riley

Definition: Immutable infrastructure is not a technology capability, rather it is a process pattern in which the system and application infrastructure, once instantiated, is never updated in place. Instead, when changes are required, the infrastructure is simply replaced. Immutable infrastructure could encompass the entire application stack, with in-versioned templates provisioned via APIs, which are most commonly available in cloud IaaS and PaaS.

Position and Adoption Speed Justification: Immutable infrastructure is typically used by organizations that take a DevOps approach to managing cloud IaaS or PaaS; however, it can be used in any environment that supports infrastructure as code. It represents a significant change in process for traditional infrastructure and operations groups. It may manifest as:

- Native cloud capabilities, such as Amazon Web Services (AWS) CloudFormation or Microsoft Azure Resource Manager templates
- Cloud management platforms, such as Flexera
- Software tools, such as HashiCorp's Terraform
- The customer's own automation scripts

Some or all of an application stack will be instantiated in the form of virtual machine images or containers, combined with continuous configuration automation tools that run after initial boot. Containers can be quickly replaced during runtime, while VM replacement is slower and requires greater coordination among other workload components. Containers improve the practicality of implementing immutable infrastructure and will drive greater adoption.

User Advice: Immutable infrastructure ensures that 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. Cloud-native workloads are more suitable for immutable infrastructure architecture than traditional on-premises workloads. And, because redundancy may be required by CSP terms of service to receive service-level agreement relief, workloads designed with an immutable infrastructure approach lend themselves to easier replication.

The application stack for immutable infrastructure is typically composed of layered components, each of which should be independently versioned and replaceable. The base OS for the master image may be updated using traditional patching tools, or automatically or manually updated.

Automation is then used to bundle components into artifacts suitable for atomic deployment, including VM images, container images, storage objects, network connections, and other necessary resources. The scripts, recipes, and other code used for this purpose should be treated similarly to the application source code itself, which mandates good software engineering discipline.

Some organizations that use immutable infrastructure reprovision only when a change is necessary. Others automatically refresh the infrastructure at frequent intervals (known as systematic workload reprovisioning) to eliminate configuration drift, to update components in which vulnerabilities were discovered, or to possibly eliminate advanced persistent threats. Frequent refresh is only practical in environments with fast and reliable provisioning; thus, it benefits strongly from containers. Integrate with a ticketing system so that refreshes can be initiated and tracked to completion.

The use of immutable infrastructure requires strict operational discipline. IT administrators should eliminate the habit of making one-off or ad hoc modifications to avoid configuration drift. Updates must be made to the individual components, versioned in a source-code-control repository, then redeployed so that everything is entirely consistent. No software, including the OS, is ever patched in production. Organizations that use immutable infrastructure may turn off all normal administrative access to instantiated compute resources — for example, not permitting SSH or RDP access. IT leaders should set a hard date for when all new workloads will use immutable infrastructure if technically feasible; deadlines can be effective motivators of behavior change.

None of the vendors listed in this innovation profile sell a product called “immutable infrastructure.” Rather, they offer one or more elements that help to establish an immutable infrastructure style. Expect to purchase multiple tools.

Business Impact: Taking an immutable approach to server and compute instance management simplifies automated problem resolution by reducing the options for corrective action to, essentially, one. This is to destroy and recreate the compute instance from a source image containing updated software or configuration that addresses the problem. 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. In other words, getting to that single corrective action is not without effort. Treating infrastructure immutably is an excellent test of the completeness of your automation framework and the confidence of your platform. The immutable approach is a management paradigm, not a technology capability. The long-term outcome is one in which the workload defines the infrastructure, which is the opposite of traditional scenarios.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

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

Recommended Reading: "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"

"Solution Path for Infrastructure Automation"

"How to Make Cloud More Secure Than Your Own Data Center"

Serverless Infrastructure

Analysis By: Arun Chandrasekaran

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 consisting of all the necessary underlying resources (specifically, the compute, storage, networking and language execution environment) required to execute an application or service are automatically provisioned and operated.

Position and Adoption Speed Justification: The term "serverless" is a misnomer, but serverless computing does transform how compute and associated resources are provisioned, operated and consumed. The most prominent manifestation of serverless computing is serverless functions or fPaaS. With fPaaS, application code is packaged into fine-grained units called "functions," with the execution of these functions delivered as a managed service. The key benefits of serverless fPaaS are:

- Operational simplicity — It obviates the need for infrastructure setup, configuration, provisioning and management.
- "Built-in" scalability — In serverless functions, infrastructure scaling is automated and elastic, which makes it very appealing for unpredictable, spiky workloads.
- Cost-efficiency — In public cloud-based serverless environments, you only pay for infrastructure resources when the application code is running, which exemplifies the "pay as you go" model of the cloud.
- Developer productivity and business agility — Serverless architectures allow developers to focus on what they should be doing — writing code and focusing on application design.

Serverless delivery of IT services has gained broad notice after Amazon popularized its Amazon Web Services (AWS) Lambda function platform as a service (fPaaS). Although some associate the notion of serverless exclusively with fPaaS, the significance of serverless, as demonstrated by the leading vendors (including Amazon, Google and Microsoft), extends beyond functions to an operational model where all provisioning, scaling, monitoring and configuration of the compute infrastructure are delegated to the platform. Examples of such services include AWS Fargate, Amazon Simple Queue Service (SQS), Amazon Athena, Microsoft Azure Container Instances (ACI)

and Google Cloud Run, to name a few. Hence, fPaaS is no longer the only form of serverless platform services.

User Advice: Serverless infrastructure does not spell the end of traditional I&O roles. However, it will significantly change the way I&O roles operate. Although perhaps counterintuitive, serverless does require operations but, instead of managing physical infrastructures, I&O leaders increasingly will have to adapt to new serverless realities by:

- Including the cost implications of event-driven application architectures and the pricing models of different vendors to ensure cost governance and budget control when planning for serverless deployments by considering API gateway, network egress and other costs.
- Revising 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.
- Rethinking 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 fPaaS in the role of service provider.

Business Impact: New application architectures, such as microservice patterns, are enabling unique competitive differentiation for companies that can rapidly scale their applications with the continuous deployment of software features, a high level of resiliency and more automation. Serverless infrastructure, implemented on-premises or off-premises, enables applications to be built quickly and deployed at a large scale. As such, it is suitable for any customer or web-facing activity in which speed of response and dynamic scalability are concerns. For variable workloads, serverless can be economical, compared with alternatives, due to its ability to provision and consume infrastructure resources only when they're needed. On-premises implementations are uncommon today due to data integration and scalability challenges.

To reap the benefits of serverless, organizations must invest time upfront to build a proof of concept (POC) to validate assumptions about the application design, code, scalability, performance and total cost of ownership.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Alibaba Cloud; Amazon Web Services; Cloudflare; Google Cloud Platform; IBM; Iguazio; Microsoft Azure; Oracle; Red Hat; VMware

Recommended Reading: "A CIO's Guide to Serverless Computing"

"2019 Strategic Roadmap for Compute Infrastructure"

"Evolution of Virtualization: VMs, Containers, Serverless — Which to Use When?"

Infrastructure Automation

Analysis By: Chris Saunderson

Definition: Infrastructure automation (IA) tools allow DevOps and I&O teams to design and implement self-service, automated delivery services across on-premises and cloud environments. IA tools enable 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 an administration console.

Position and Adoption Speed Justification: 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.

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
- Support for programmable infrastructure
- Self-service and on-demand environment creation
- Resource provisioning
- Configuration management

IA tools have become increasingly similar in the breadth of their configuration management content and enterprise capabilities. IA vendors are developing greater knowledge of configuration artifacts and state, activity patterns, roles, and policy. Vendors are leveraging these insights to prevent misconfigurations, resolve problems and provide more advanced deployment and optimization capabilities.

As IA tools are increasingly accepted by development and I&O groups, organizations are looking to replace their tribal implementations with an enterprisewide IA tool strategy.

User Advice: Because IA tools provide a programmatic framework, the costs associated with them extend beyond just the licensing cost (or the lack thereof), so enterprises should include professional services and training requirements in cost evaluations. In particular, most I&O organizations should expect to invest in training because not all infrastructure administrators have the skills needed to use these tools successfully. IA tools have a learning curve, and it is tempting for developers and administrators to revert to known scripting methods to complete specific tasks. DevOps and IT operations leaders who want to maximize the value of IA tool investments must ensure that their organizations' culture can embrace IA tools strategically.

Use the following criteria to determine which IA vendor and product is appropriate:

- Internal IT skills
- Ecosystem surrounding IA tools
- Method for interacting with managed systems
- Security and compliance capabilities
- Authentication and authorization support
- Alignment to other tools within operating environment
- Orchestration functionality
- Scalability
- Platform and infrastructure content support

Business Impact: By enabling infrastructure administrators and developers to automate the deployment and configuration of settings and software in a programmatic way, organizations across all verticals stand to realize:

- Agility improvements — By enabling continuous integration and delivery concepts to IT infrastructure management.
- Productivity gains — Via faster deployment and repeatable, version-controlled configuration of infrastructure.
- Cost-reduction improvements — Via significant reductions in required manual interactions by highly skilled and high-cost staff by automating "day 2" operational tasks. Licensing cost reductions may also be achieved.
- Risk mitigation — Compliance improves via the consistent use of standardized, documented processes and configurations across physical and virtual infrastructures.

IA tools can drive efficiencies in operational configuration management, as well as provide a flexible framework for managing the infrastructure of DevOps initiatives. They achieve this by integrating with other toolchain components — continuous integration (CI) and application release orchestration — in support of continuous delivery.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services (AWS); Chef; HashiCorp; Inedo; Microsoft Azure; Pulumi; Puppet; Quali; VMware

Recommended Reading: "Market Guide for Infrastructure Automation Tools"

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

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

"How to Lead Digital Disruption With Programmable Infrastructure"

"Assessing HashiCorp Terraform for Provisioning Cloud Infrastructure"

Container Management

Analysis By: Dennis Smith

Definition: Container management supports the management of containers at scale. This category of software includes container runtimes, container orchestration and scheduling, resource management and other container management capabilities. 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. It can also be used to more efficiently package COTS applications.

Position and Adoption Speed Justification: Gartner surveys show that the demand for containers continues to rise. This is likely due to the growing adoption of container runtimes, which have introduced common container packaging formats that are more easily consumable by, and useful to, application developers and those with a DevOps approach to IT operations. Container runtimes, frameworks and other management software have increased the utility of containers by providing capabilities such as packaging, placement and deployment, and fault tolerance (e.g., cluster of nodes running the application). The emergence of de facto standards (e.g., Kubernetes) and offerings from the public cloud providers are also driving adoption. Container management integrates these various elements to simplify deploying containers at scale. Many vendors enable the 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 for that purpose.

The most common use of containers is focused specifically on Linux environments, and management software follows accordingly; however, there has been a gradual adoption of Windows containers. Container-related edge computing use cases have also increased, along with deployments involving bare-metal servers and the emergence of operational control planes that support containers and VMs.

Among the functionalities that container management systems provide are orchestration and scheduling, monitoring and logging, security and governance, registry management, and links to CI/CD processes. Among the vendor offerings are hybrid container management software, public cloud IaaS solutions specifically designed to run containers and PaaS frameworks that have incorporated integration with container management software. All major public cloud service providers are now deploying on-premises container solutions.

There is a high degree of interest in, and awareness of, containers within global organizations. Though many enterprises are planning or have recently commenced container deployments, few

have containerized a significant portion of their application workloads. Additionally, there is significant grassroots adoption from individual developers who use containers with increasing frequency in development and testing — particularly for Linux. Container management software has progressed from an early-adopter technology to adolescent, where it remains.

User Advice: Organizations should begin exploring container technology as a means for packaging and deploying applications and their runtime environments. Depending on the environment, container management tools are often deployed complementarily with continuous configuration management tools. As container integration is added to existing DevOps tools and to the service offerings of cloud IaaS and PaaS providers, DevOps-oriented organizations should experiment with altering their processes and workflows to incorporate containers. An organization may be a good candidate if it meets the following criteria:

- It's DevOps-oriented or aspires to become DevOps-oriented.
- It has high-volume, scale-out applications with a willingness to adopt microservices architecture, or has large-scale batch workloads.
- It has aspirational goals of increased software velocity and immutable infrastructure.
- It intends to use an API to automate deployment, rather than obtaining infrastructure through a self-service portal.

Organizations must also factor in their desire for hybrid and/or multicloud deployments into vendor selection, as many vendors offer container management software that can be deployed in different cloud environments.

Business Impact: Container runtimes make it easier to take advantage of container functionality, including providing integration with DevOps tooling and workflows. Containers provide productivity and/or agility benefits, including the ability to accelerate and simplify the application life cycle, enabling workload portability between different environments and improving resource utilization efficiency and more. Container management software simplifies the art of achieving scalability and production readiness and optimizes the environment to meet business SLAs.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Google Cloud Platform; IBM; Microsoft Azure; Mirantis; Rancher Labs; Red Hat; VMware

Recommended Reading: "Best Practices for Running Containers and Kubernetes in Production"

"Market Guide for Container Management"

"Best Practices to Enable Continuous Delivery With Containers and DevOps"

Sliding Into the Trough

Composable Infrastructure

Analysis By: Daniel Bowers; Philip Dawson

Definition: Composable infrastructure creates physical systems from shared pools of resources using an API. The exemplary implementation uses disaggregated banks of processors, memory, storage devices and other resources, all connected by a fabric. However, composable infrastructure can also aggregate or subdivide resources in traditional servers or storage arrays.

Position and Adoption Speed Justification: Servers, storage and fabrics are traditionally deployed as discrete products with predefined capacities. Individual devices, or set amounts of resources from individual devices, are connected together manually and dedicated to specific applications. Composable infrastructure allows resources to be aggregated through software-defined intelligent automation, enabling infrastructure and operations leaders to achieve higher resource utilization and faster application deployment. Although some blade-based server infrastructures have long included composable networking features, composable infrastructure describes a broader spectrum of capabilities including disaggregation of accelerator, memory and storage resources.

Current 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 2020; for example, an agreement between next-generation fabric consortia, Compute Express Link (CXL) and Gen-Z Consortium, to cooperate on standards. A key step in the maturity timeline for composable infrastructure will be technology that can disaggregate DRAM from compute.

User Advice: The deployment of composable infrastructure is appropriate where infrastructure must be resized frequently, or where composability increases the utilization of high-cost components. The majority of current use cases are in multitenant environments where composability allows 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.

Don’t replace existing infrastructure to obtain composable infrastructure unless you have sufficient mature automation tools and skills to implement composable features. Verify that your infrastructure management software supports composable system APIs, or that you have the resources to write your own management tools. However, don’t avoid infrastructure with composable features. Rather, don’t choose such infrastructure *because* of those features unless you are prepared to use them.

Business Impact: 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 in multitenant environments. However, a proliferation of vendor-specific APIs and the lack of off-the-shelf software for managing composable systems are headwinds to widespread adoption.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Dell Technologies; DriveScale; GigaIO; Hewlett Packard Enterprise (HPE); Intel; Liqid; Western Digital

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"

Micro Operating Systems

Analysis By: Thomas Bittman

Definition: A micro operating system (micro OS) is an OS designed to be extremely small and lightweight, used most frequently with cloud computing, edge computing and containers. It 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.

Position and Adoption Speed Justification: Most micro OSs are being developed as subsets of existing, mature OSs that were originally designed to be general-purpose, rich in function and relatively large (normally 5GB to 10GB). The idea of "just enough OS" emerged with VMs and virtual appliances, especially for packaging and supporting a small application. However, what's driving micro OSs is the growth of cloud computing, container technologies, DevOps, edge computing and microservices architecture (even smaller application) requirements. As use of container technologies has increased, there have been a number of major OS vendors that have announced micro OSs, and a few important acquisitions (e.g., Red Hat acquiring CoreOS). As the adoption of container technologies and the use of microservices architectures grow for new applications, micro OSs will be used as the OS architecture of choice. Because micro OSs are essentially minimal subsets of mature OSs, they should mature relatively quickly as they are adopted; and their use will grow rapidly, as container technologies are adopted. In addition, more digital business applications will require granularity, ongoing changes, agility, rapid scaling and rapid provisioning, driving usage of micro OSs.

Note that there are smaller OSs that are less general-purpose, typically designed for real-time systems or as embedded OSs that usually have a footprint in the 4KB to 30KB range.

User Advice: 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

Business Impact: The size of the OS is a prime determinant in the agility and speed with which a container (or VM) can be provisioned or duplicated (for scaling). The complexity of the OS can also inhibit rapid application development models and affect management. Micro OSs will be used by developers building and maintaining applications in an agile development process to increase their provisioning speed (for agile development) and to improve the speed of automated horizontal scaling. The end result will be faster updates and scaling for applications that are often going to be consumer-facing or more real time. Micro OSs will be a core enabler to most new digital business applications — both in the cloud and at the edge.

Benefit Rating: Moderate

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Canonical; Microsoft; Rancher; Red Hat; SUSE; VMware

Recommended Reading: "Assessing Infrastructure Requirements for Deploying Kubernetes"

Edge Servers

Analysis By: Thomas Bittman

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 capability 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.

Position and Adoption Speed Justification: Edge serving has been in existence for decades in one form or another. For example, edge servers are used by content delivery networks (CDNs) to perform content caching, outbound delivery and low-level, end-user interactions. Industrial programmable logic controllers (PLMs) can read sensors and take simple actions when certain triggers are met. However, such traditional edge servers lack the performance to do sophisticated analysis, and often lack the capacity to handle large volumes of data. For some use cases, edge servers are mature. For others, edge servers are still evolving.

Edge serving is evolving rapidly, driven by the explosion of data being generated at the edge by the industrial Internet of Things (IoT) and immersive experiences. There will not be enough compute

power, storage capacity or network bandwidth in the cloud or central data storage facilities to handle the volumes of data that will be generated by these devices in the future. The need to locally handle large volumes of data, make real-time decisions and execute machine learning algorithms is increasing. Hence, the edge server must become ever-more powerful, and will morph into micro or even macro edge data centers that can be deployable to almost any remote location.

These edge computing devices will need to gather the data, perform network protocol translation and preprocess the data before passing only the valuable data to an enterprise or cloud data center. They will also perform intense analytics processing and storage that has traditionally been performed in the cloud or at a more central location. They must also be manageable at extreme scale, operate in diverse environments that might lack traditional infrastructure, and be able to provide robust and modern digital and physical security capabilities.

Gartner predicts that demand for edge servers in support of industrial IoT and immersive experiences (e.g., augmented or mixed reality in the workplace, school or mall) will grow dramatically through 2021. By year-end 2021, more than 50% of large enterprises will deploy at least one edge-computing use case to support IoT or immersive experiences, versus fewer than 5% in 2019. By 2022, more than 50% of enterprise-generated data will be created and processed outside the data center or the cloud.

User Advice: Edge server requirements will be unique for each situation; however, the ability to evolve as requirements evolve is important. Because the data being collected and the actions that need to be performed are evolving rapidly, choose edge servers that can be deployed rapidly and are extensible in the field to match changing requirements. Expect this market to evolve rapidly, avoid lock-in where possible and plan for technology changes in a few years in many cases. Edge servers are only as good as the data analytics that they support, so give preference to IoT frameworks that are mature and have broad support. Security requirements must be an upfront design goal in any edge server deployment.

Business Impact: Edge servers will become a critical part of most enterprises' infrastructure topologies — either owned by the enterprise or leveraged as-a-service through cloud or telecom providers. The center of gravity of digital business will expand from central processing in the enterprise and the cloud to real-time processing and engagement at the edge. Edge servers will be important enablers of new digital business capabilities.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Early mainstream

Sample Vendors: ADLINK; Cisco; Dell; Eurotech; Hewlett Packard Enterprise (HPE); Lenovo

Recommended Reading: "Why and How I&O Should Lead Edge Computing"

"How to Overcome Four Major Challenges in Edge Computing"

"Exploring the Edge: 12 Frontiers of Edge Computing"

Next-Generation Memory

Analysis By: Martin Reynolds; Chirag Dekate

Definition: Next-generation memory (NGM) is a type of nonvolatile memory capable of replacing 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.

Position and Adoption Speed Justification: Phase-change memory (PCM) is shipping in storage devices, and Intel's server platforms, introduced in 2019, support it as part of main memory. In the long term, Intel's Optane PCM could replace DRAM and flash memory in many systems, including as embedded memory, based on higher density and lower cost. Micron Technology codeveloped the technology with Intel and offers it under the QuantX brand. However, Intel servers only support Intel-branded products.

Other technologies exist, but only spin-transfer torque memory might approach the density and performance requirements for main memory.

Over the last year, PCM has not advanced in density, and DRAM prices have dropped. Samsung now offers a small-block flash product, which improves performance of solid state drives at the expense of slightly lower density. This product, with likely significantly lower cost, constricts the opportunity for NGM.

These developments cause us to move NGM closer to the trough.

User Advice: Next-generation memory will enable servers with main memory of more than 6 TB, and up to 60 TB of fast, nonvolatile storage. The nonvolatile storage will be significantly faster than any SSD today. In 2019, server vendors released eight-socket systems supporting 32 TB of main memory.

These servers bring a new price point to buyers for large-scale transaction processing and data analytics. Users should evaluate these products for in-memory database or analytics workloads that require larger capacity than available with DRAM, or workloads requiring high-speed persistent memory that can survive server power loss.

Although the current technology is uneconomical for mainstream servers, it is viable for systems that benefit from very large main memory and the consequent performance advantages. Furthermore, as NGM is denser than DRAM, it may represent the only path to large memory subsystems.

These systems are potentially transformational in businesses that have large-scale look-up, matching and analysis challenges, including banking, transportation and logistics. The combination of 36TB of lower-cost memory and eight advanced CPUs in a 2U chassis represents a compelling advantage in price, performance and rack space.

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 to this, a key impact of this technology will be to accelerate adoption of in-memory computing architectures. However, the economics and performance of this memory are not yet suitable for mainstream applications.

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Dell; Hewlett Packard Enterprise; Inspur; Intel; Lenovo; Supermicro

Recommended Reading: "Top 10 Technologies That Will Drive the Future of Infrastructure and Operations"

"Predicts 2020: Semiconductor Technology in 2030"

FPGA Accelerators

Analysis By: Alan Priestley

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.

Position and Adoption Speed Justification: 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 FPGAs to offload 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.

FPGAs are typically configured using hardware programming languages, such as register transfer level (RTL) and VHSIC Hardware Description Language (VHDL), that are very complex to use; this has held back widespread adoption. However, major FPGA vendors (Intel and Xilinx), along with a number of startups such as Mipsology and Swarm64, are working to address this with libraries and toolsets that enable FPGAs to be configured using software-centric programming models.

Adoption is also becoming easier, helped by frameworks such as OpenCL that lower the time and skills required to use FPGAs. Workloads like deep learning (inference) and easier access to development platforms, exemplified by Amazon Web Services' FPGA-enabled instance types, are also driving adoption of FPGAs within the data center.

Today, the biggest growth opportunity for FPGAs in the data center is in the inference portion of deep-learning workloads. Given the evolving nature of this new use case and the maturing of the surrounding software ecosystem, FPGA accelerators have been positioned pretrough.

User Advice: FPGA accelerators can enable dramatic performance improvements within significantly smaller energy consumption footprints than comparable commodity technologies. I&O leaders need to evaluate applicability of FPGA accelerators by:

- Identifying application subsets that can be meaningfully impacted using FPGAs.
- Evaluating the availability of FPGA-based hardware for use in data center server deployments, such as FPGA-based PCIe add-in cards.
- Outlining costs associated with necessary skill set and programming challenges of FPGAs and the maturity of the software-centric programming toolsets.
- Leveraging cloud-based FPGA services to accelerate development.

I&O leaders should use FPGA accelerators when:

- Preconfigured solutions exist that can help dramatically transform key workloads (e.g., financial trading analytics, genome sequencing, etc.).
- Algorithms will evolve requiring frequent updates at the silicon level that can be utilized by broader applications.

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 (8 bit and 16 bit) processing capabilities in energy-efficient footprints. While programmability continues to be a major challenge, limiting broader adoption of FPGAs, I&O leaders should evaluate FPGA-based solutions for genome sequencing, real-time trading, video processing and deep learning (inference). I&O leaders can further insulate against risks by utilizing cloud-based infrastructures for provisioning FPGAs (e.g., Amazon EC2 F1 instances, Microsoft Azure, Baidu cloud).

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Baidu; Intel; Microsoft Azure; Xilinx

Recommended Reading: “Forecast Database, AI Neural Network Processing Semiconductors, 1Q20”

“Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide”

“Forecast Analysis: Data Center Workload Accelerators, Worldwide”

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

OS Containers

Analysis By: Thomas Bittman

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.

Position and Adoption Speed Justification: 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. Docker has become a popular container packaging format for developers, and micro OSs are being developed by most major OS vendors. A container/cluster ecosystem is maturing that includes open source, such as Kubernetes.

User Advice: Several container technologies are targeting agile development and cloud deployment, including Linux containers (LXC), Docker's libcontainer, Apache Mesos, Microsoft Windows Containers and VMware vSphere Integrated Containers. Several container technologies designed specifically for consolidation and density are mature (e.g., Oracle Solaris Zones and Virtuozzo); however, all container technologies will significantly improve resource utilization and densities — the difference will be in the richness of management tools available for them. Development teams focused on agile development and microservices can leverage container technologies.

OS containers do not replace hypervisor-based virtualization technologies — such as VMware or Kernel-based Virtual Machine (KVM) — because they essentially benefit different kinds of applications and target different problems. Containers focus on application design and developer requirements, and hypervisors focus on capacity management and infrastructure and operations (I&O) professionals. Security and manageability concerns are often mitigated by deploying containers within VM architectures (either layered or explicitly integrated, such as the Microsoft and VMware offerings). Although VMs are inherently "heavier," most of the "weight" comes with the size of the OS itself. Balancing application needs, developer needs and operational needs will determine whether to run containers on bare metal, on VMs or integrated with VMs.

Business Impact: Container technologies are a part of a development architecture that will help enterprises become more agile overall, 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 containers and agile development, rather than existing, monolithic application architectures. However, for developer ease of use, containers will also be used as wrappers for traditional workloads, pushing enterprises to develop their container management expertise quickly.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Canonical; Docker; Mesosphere; Microsoft; Oracle; Red Hat; Virtuozzo; VMware

Recommended Reading: "Best Practices for Running Containers and Kubernetes in Production"

ARM Servers

Analysis By: Martin Reynolds; Alan Priestley; Daniel Bowers

Definition: ARM servers are built using microprocessors based on the ARM instruction set architecture. ARM was originally designed as lightweight IP-based RISC processor core that traded single-thread performance for a small silicon area. Processor designs typically have low power consumption and a rich set of on-chip features such as networking and cryptographic acceleration. Servers built with ARM processors are offered both as appliances targeting specific workloads and for general-purpose workloads.

Position and Adoption Speed Justification: Processors based on the ARM instruction set architecture (ISA) are the dominant architecture in a wide range of personal computing applications, such as 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." To address server architectures, where the balance shifts to performance over power efficiency, ARM offers server-optimized core designs to silicon vendors. Companies including Amazon Web Services (AWS) and Huawei have also designed customized ARM cores with specialized features.

The ecosystem for ARM servers has developed slowly. Production-level support from Linux OS vendors including Red Hat and SUSE is now available, but application support is mostly limited to open-source software. ARM servers are well-suited for web serving, caching, storage management and network connectivity products. AWS has recently announced the availability of Amazon EC2 compute instances based on the second generation of its ARM-based Graviton processor. However, the broad range of price and performance of Intel server processors and the reemergence of AMD server processors are both headwinds against the growth of ARM servers.

An unusual aspect of the ARM server business is the pace at which chip vendors enter and leave the market. Currently Ampere, Marvell and Huawei are the main suppliers of ARM-based server processors. New ARM processor vendors including NUVIA have appeared, while others including AMD and Qualcomm have withdrawn plans for ARM processors for general-purpose server use. The hyperscaler cloud providers including Amazon, Google, Facebook and Microsoft all have activity in ARM servers. AWS's Graviton processor line, which is both sold via Amazon EC2 instances and used internally by Amazon for its infrastructure, could represent a turning point for ARM server maturity.

User Advice: Data center managers should plan on x86 servers as the primary architecture for the foreseeable future. Although major suppliers continue to bring ARM servers to market, these products are often experimental or targeted for very specific markets, such as HPE Apollo servers targeting HPC, and Fujitsu's A64FX processor for supercomputers. ARM technology can be

effective when packaged in a network or storage appliance where there are fewer software integration challenges.

Business Impact: ARM servers can bring lower hardware costs in targeted use cases compared to x86-based servers, especially in workload-specific appliances with modest compute requirements and in certain open-source software applications. The competitive threat of ARM servers can also influence the x86 server processor vendors' product portfolios and pricing. However, the slow pace of end-user adoption and the dominance of x86 in data center compute mean that ARM servers are unlikely to materially change the server landscape before 2025.

Benefit Rating: Low

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Ampere; Bamboo Systems; Fujitsu; Huawei; Marvell

Recommended Reading: "Market Trends: Emerging Opportunities for Semiconductor Vendors at the Hyperscale Cloud Service Providers"

"Market Trends: Who Sells Servers for Hyperscale Data Centers"

"Top 10 Strategic Technology Trends for 2020: Empowered Edge"

Hardware-Based Security

Analysis By: Neil MacDonald; Tony Harvey

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 monitoring and unencrypted memory handling.

Position and Adoption Speed Justification: Adoption is increasing and becoming mainstream, as hardware-based isolation capabilities are becoming standard in most hardware devices and cloud-based IaaS 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 has risen in the face of ongoing disclosures of new types of side-channel attacks. Another driver is the desire to use IaaS providers in potentially hostile parts of the world and protect these workloads from virtual machine and memory snapshotting. However, disillusionment remains as methods vary wildly among vendors, and some strong isolation capabilities such as Intel Software Guard Extensions (SGX) require applications to be rewritten and are incompatible with techniques used by AMD. Abstraction layers, such as Asylo, may help but add another layer of complexity and are not widely adopted.

Multiple implementations are appearing across vendors, OSs and chipsets:

- Samsung's Knox security hypervisor, where a supervisory process monitors the OS kernel for aberrant behavior. The supervisory process runs at a higher privilege level than the OS and cannot be compromised.
- Intel SGX provides a new privilege level for running code, which can be set up in a user-level process but excluded from operating system or hypervisor access. Multiple public cloud providers now support SGX including Alibaba Group, IBM and Microsoft.
- AMD has a similar set of technologies for protecting memory against physical and system software-based attacks: Secure Memory Encryption (SME), Transparent Secure Memory Encryption (TSME) and Secure Encrypted Virtualization (SEV).
- In 2018, Intel introduced chip-level Threat Detection Technology (TDT) that was further improved in 2019 and is now supported by multiple security offerings, including Microsoft Windows Defender.
- Microsoft uses hardware-based virtualization features in Windows 10 and Windows Server 2016 to create a protected code execution space for monitoring the OS and providing security features with Device Guard and Credential Guard.
- VMware built AppDefense — a way to monitor and protect applications from the hypervisor layer, outside of the workload, protected by virtualization-enabled hardware.
- Apple has developed and shipped its iOS Secure Enclave processor to protect sensitive operations and monitor kernel integrity.
- Google has developed a custom chip, Titan, for hardware-based root of trust and is deployed throughout their data centers. This chip binds a strong identity to each server, verifies the integrity of firmware and software, and creates a nonrepudiable audit trail of all changes to each machine.
- Amazon Web Services (AWS) uses a custom-designed Nitro Controller on its Nitro-based systems that includes a special micro-controller for security isolation and integrity measurements called the Nitro Security Chip. This becomes the foundation for its confidential computing offering called Nitro Enclaves, although this isolation uses the Nitro Hypervisor.

User Advice:

- Hardware-based security is strong, but may potentially still be broken by software flaws, or side-channel attacks such as Spectre and Meltdown. Patch and remain vigilant for unexpected breaches.
- Most systems will include hardware-rooted isolation and integrity capabilities by default. Make strong isolation of sensitive code and security controls a mandatory part of IT systems procurement, including IaaS.
- For systems under direct enterprise control, implement a BIOS-level patching strategy to deal with exposures that require BIOS-level remediation.

- For systems that move to public cloud infrastructure, evaluate the need for confidential computing capabilities only for the most critical applications to protect sensitive operations such as key management and sensitive intellectual property.
- Although the SGX approach is compatible with hypervisors, there may be unanticipated interactions. For example, it may not be possible to snapshot, suspend and restore a partition with a protected process. Understand limitations of hypervisor-based functionality before implementing SGX.
- Before activating Windows 10 virtualization-based security, check for compatibility issues with third-party approaches that also use virtualization techniques.
- Hypervisor-based approaches with security rooted in hardware virtualization techniques are another way to achieve similar levels of strong isolation (for example, Hysolate and Bitdefender have offerings that use this approach).
- None of these mechanisms are interoperable, so plan different strategies for different devices and server platforms.

Business Impact: If an operating system 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 require that operating system software and system management software be able to make use of it. Upgrading to most recent versions of software that can use hardware features and using cloud systems with advanced security, can materially increase system security.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon Web Services (AWS); Apple; Bitdefender; Fortanix; Google; Hysolate; Intel; Microsoft; Samsung Electronics; VMware

Recommended Reading: “Market Guide for Cloud Workload Protection Platforms”

“How to Make Cloud More Secure Than Your Own Data Center”

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

“How to Mitigate Firmware Security Risks in Data Centers, and Public and Private Clouds”

“Key Management as a Service Exposes Different Risks to Data in Public Clouds”

Persistent Memory DIMMs

Analysis By: Alan Priestley

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

Position and Adoption Speed Justification: 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 and reliability are acceptable.

The markets adoption of PM-DIMMs has been hampered by the slow write performance and limited endurance of existing flash memory technologies. However, the introduction of the 3D XPoint nonvolatile memory technology, from Intel and Micron, brought substantial performance and reliability gains over traditional flash memory. Intel's introduction of its Optane DC persistent memory DIMMs enables the use of PM-DIMM technology in data center servers (albeit those based on the latest generations of Intel's Xeon Scalable Processors).

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 will be necessary to provide the operating system and applications with an area of memory capable of sustaining frequent high-speed write accesses. To achieve greater adoption, support will be required across a wide range of server vendors, operating systems and applications. In addition, use cases for persistent memory technologies will need to spread beyond the extremely high-performance, high-bandwidth and ultra-low-latency applications for which they are attracting most interest today.

This technology has faced a number of challenges and has not yet reached maturity. Memory technologies such as 3D XPoint will replace the use of NAND flash on PM-DIMMs; however, these devices have not long been commercially available. The benefit of 3D-XPoint-based PM-DIMMs is that they can achieve higher capacity of current DRAM DIMMs and at a lower $/GB. While Intel has added support for this technology in its 2nd Gen Xeon Scalable Processor (SP), to utilize the persistence and exploit the performance requires ecosystem support, especially from software vendors. For this reason, this technology is currently in the trough.

User Advice: IT professionals should analyze their workloads to determine software vendor support for PM-DIMMs and the performance and memory capacity demands. Major software vendors, such as SAP and Oracle, have recently implemented support for PM-DIMMs and Google has also announced the intent to leverage PM-DIMMs in its cloud services. Since this technology is still nascent, users must assess the roadmaps of the major server and storage OEMs along with those of the SSD appliance vendors that will be launching DIMM-based storage systems. When evaluating PM-DIMM deployments consideration should also be given to TCO comparisons between nonvolatile memory and DRAMs, especially with the dynamic pricing fluctuations in the DRAM market.

In their evaluations, IT professionals should 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 only currently compatible with servers that deploy the second generation or later Xeon SPs.

Business Impact: This technology's impact on users will improve overall system performance. The specific workloads expected to see early adoption are in the in-memory computing, virtualization, analytics, AI and HPC segments. There may also be an impact on traditional storage subsystems as applications are rearchitected to take advantage of large amounts of nonvolatile memory accessible as part of the main server system memory.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Dell EMC; Formulus Black; Google; Hewlett Packard Enterprise; Huawei; Intel; Lenovo; NetApp; Oracle; SAP

Recommended Reading: "Determining the Data Center Opportunity Created for 3D XPoint Persistent Memory"

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

"Predicts 2020: Semiconductor Technology in 2030"

"Vendor Rating: Intel"

"Forecast Analysis: NAND Flash, Worldwide, 1Q20 Update"

Hyperconvergence

Analysis By: Philip Dawson; Jeffrey Hewitt

Definition: Hyperconvergence is scale-out software-integrated infrastructure designed for IT leaders seeking operational simplification. Hyperconvergence provides a building block approach to compute, network and storage on standard hardware under unified management. Hyperconvergence vendors build appliances using off-the-shelf infrastructure, engage with system vendors that package software as an appliance, or sell software for use in a reference architecture or certified server. Hyperconvergence may also be delivered as a service or in a public cloud.

Position and Adoption Speed Justification: Hyperconvergence solutions are maturing and adoption is increasing as organizations seek management simplicity. VMware vSAN utilization among VMware ESXi customers, and Storage Spaces Direct utilization among Microsoft Windows Server 2016 and 2019 Datacenter edition customers are on the rise. Nutanix, an early innovator in HCIS appliances, has largely shifted to a software revenue model and continues to increase the number of OEM relationships. 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. Meanwhile, suppliers are expanding hybrid cloud deployment offerings. 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. Meanwhile, 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.

User Advice: IT leaders should implement hyperconvergence when agility, modular growth and management simplicity are of greatest importance. The acquisition cost of hyperconvergence may be higher and the resource utilization rate lower than for three-tier architectures, but management efficiency is often superior. Hyperconvergence requires alignment of compute and storage refresh cycles, consolidation of budgets, operations and capacity planning roles, and retraining for organizations still operating separate silos of compute, storage and networking. Adopt for mission-critical workloads, only after developing knowledge with lower-risk deployments, such as test and development. Workload-specific proofs of concept are an important step in meeting the performance needs of applications. Consider the impact on DR and networking. Test 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. Consider nonappliance options to enable scale-down optimization of resources for high-volume edge deployments. In product evaluations, consider the ability to independently scale storage and compute, retraining costs, and the ability to avoid additional operating system, application, database software and hypervisor license costs. In large deployments, plan for centralized management of multiple smaller clusters. For data center deployments, ensure that clusters are sufficiently large to meet performance and availability requirements during single and double node failures. While servers are perceived as commodities, they differ greatly in terms of power, cooling and floor space requirements, and performance, so evaluate hyperconvergence software on a variety of hardware platforms for lowest total cost of ownership and best performance.

Business Impact: The business impact of hyperconvergence is greatest in dynamic organizations with short business planning cycles and long IT planning cycles. Hyperconvergence enables IT leaders to be responsive to new business requirements in a modular, small-increment fashion, avoiding the big-increment upgrades typically found in three-tier infrastructure architectures. Hyperconvergence provides simplified management that decreases the pressure to hire hard-to-find specialists. It will, over time, lead to lower operating costs, especially as hyperconvergence supports a greater share of the compute and storage requirements of the data center. For large organizations, hyperconverged deployments will remain another silo to manage. 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. As more vendors support public cloud deployments, hyperconvergence will also be a stepping stone toward public cloud agility.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Sample Vendors: Cisco Systems; Dell Technologies; Hewlett Packard Enterprise (HPE); Huawei; Microsoft; Nutanix; Pivot3; Red Hat; Scale Computing; VMware

Recommended Reading: "Magic Quadrant for Hyperconverged Infrastructure"

"Critical Capabilities for Hyperconverged Infrastructure"

"Toolkit: Sample RFP for Hyperconverged Infrastructure"

"The Road to Intelligent Infrastructure and Beyond"

"Use Hyperconverged Infrastructure to Free Staff for Public Cloud Management"

Climbing the Slope

In-Memory Computing

Analysis By: Philip Dawson; Massimo Pezzini

Definition: In-memory computing (IMC) is an application architecture style which assumes that all the data required by applications for processing is located in the main memory of their computing platforms. In IMC-style applications, a persistent, nonvolatile data store (hard drive or solid-state drive) is used to permanently store in-memory data for recovery purposes, to manage overflow situations, to manage the information life cycle and to transport data to other locations, but is not used as the primary location for the application data.

Position and Adoption Speed Justification: IMC delivers dramatic improvements in performance, scalability and analytic sophistication over traditional architectures. IMC is enabled by a range of rapidly converging software technologies and advancements in hardware architectures.

The emergence of persistent memory in the form of 3D XPoint, Intel Optane and NVM Express (NVMe) has sparked renewed interest in IMC. Meanwhile, adoption has grown, driven by several factors including:

- Tens of thousands of SAP clients (of any size) have adopted the SAP HANA in-memory DBMS either to retrofit traditional workloads, such as SAP ERP and SAP Business Warehouse, or to power new applications such as SAP S/4HANA.
- Packaged application providers (such as Microsoft and Oracle) delivered IMC-enabled add-ons to their products.
- Traditional DBMS platforms from almost every vendor include IMC capabilities as standard features, or as options.
- Increasing availability of open-source IMC technologies that make adoption affordable for midsize organizations and startups.

- IMC technologies are increasingly embedded in mainstream software products and cloud services (for example, applications, business process management tools, application platforms and integration platforms).
- A growing number of providers offering cloud-based renditions of their IMC-enabling, or IMC-enabled, products.
- Convergence of several IMC technology streams into proto-IMC platforms.
- The emergence of a new generation of memory technologies that will combine the low data access latency of DRAM and the nonvolatility and affordable cost of flash memory.

The already notable use of IMC across vertical sectors, geographies and business sizes continues to expand. This is stimulated by the growing number of digital business initiatives, the mounting interest around real-time analytics and emerging use cases such as IoT, omnichannel (e-commerce or banking), event-driven decisions, hyperscale architectures and hybrid transactional/analytical processing (HTAP).

Factors that will be obstacles to even-faster adoption include:

- Market and technology fragmentation
- Cloud services enabling alternate IMC adoption at low cost and risk
- Security and IT operations challenges
- Disappointing results when just lifting and shifting traditional applications on top of IMC-enabling technologies
- Long development and maturity cycles for new IMC-enabled versions of packaged applications

User Advice: I&O and application leaders in charge of modernizing their architectures to support strategic initiatives must identify which of the following approaches is the best path for their IMC adoption:

- Developing new custom (or purchasing new packaged) groundbreaking applications, natively based on IMC design principles.
- Replatforming or reengineering traditional applications for IMC technologies. These approaches are less invasive than the IMC-native style, but usually lead to incremental benefits.

Application leaders should also:

- Monitor their strategic vendor roadmaps to identify how IMC impacts their investment plans
- Make sure that the particular variants of IMC that are offered are suited to their needs

IMC skills and best practices are rapidly becoming commonplace. Nonetheless, it is still advisable for mainstream organizations to incrementally become familiar with IMC architectures by successfully deploying a few IMC applications, based on the less disruptive approaches, before embarking on more ambitious, native IMC-style projects.

Business Impact: IMC has a long-term, disruptive impact by radically changing users' expectations, application design principles, product architectures and vendor strategies. IMC opens up a number of opportunities for digital innovation, such as hyperscale applications, HTAP, omnichannel customer experience, API economy, situation awareness, event-driven IT, and self-service and real-time analytics.

IMC-style applications drive transformational business benefits by enabling IT leaders to:

- Deliver orders of magnitude faster, analytical, transaction processing and HTAP applications.
- Support hyperscale business models serving very large numbers of globally distributed, mobile-enabled users or smart machines interacting in real time.
- Provide deeper and greater real-time business insights and situation awareness, via event stream processing.
- Radically reduce the time to value of new applications through near-interactive development processes.
- Improve price performance and resilience of complex systems.

Organizations leveraging IMC are better positioned to build defensible business differentiation than those sticking with traditional architectures. Organizations that fail to endorse IMC risk falling behind in the race for leadership in the digital era.

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: GigaSpaces; IBM; Intel; Microsoft; Oracle; Qlik; SAP; Software AG; Tableau; Workday

Recommended Reading: "Predicts 2019: In-Memory Computing at a Turning Point, Driven by Emerging Persistent-Memory Innovation"

"Forecast: In-Memory Computing Platforms (AIM Components), Worldwide, 2017-2022"

"SAP S/4HANA Research Roundup"

Server-Side Client Graphics

Analysis By: Nathan Hill; Philip Dawson

Definition: Server-side client graphics are used to speed up the execution of compute-intensive applications and to improve the rendering of screen graphics, especially for virtual desktop infrastructure (VDI) workloads. Graphics processing unit (GPU) graphic cards are installed on

servers and dedicated to individual users or shared among user sessions running in physical or virtual machines (VMs).

Position and Adoption Speed Justification: Server GPU platforms targeted at improving the remote client computing user experience (for example, with server-based computing [SBC] and VDI) help address remaining multimedia limitations with centralized computing, while maintaining desktop densities through CPU offloading. The use of GPU-based desktops for knowledge workers is now an option. The technology is integrated at the hypervisor layer for shared-user platforms, although SBC can share a GPU across multiple sessions in a physical deployment. The technology also enables a whole new set of design, high-end 3D modeling, engineering and architecture use cases with an increasing number of production deployments. There is increasing use of GPUs on cloud server platforms to enhance desktop as a service (DaaS) especially for architecture, engineering and construction (AEC) use cases.

User Advice: Configuring graphic cards on servers can restrict choices (for example, forcing the use of systems large enough to accommodate a graphics card internally). GPUs add cost and increase power consumption. Balance the advantages of a dedicated pool of GPU-configured servers against the flexibility and other benefits of including GPUs in a virtualized resource pool that handles a broad range of use cases.

High-performance workstation use cases, especially designers working with complex 3D models and graphic-intensive workloads, should be assessed first to see if the server-side client graphic model can improve performance by positioning compute close to the data. The best way to do this is by reducing model load times and avoiding data distribution and synchronization challenges. This architecture can also increase accessibility, collaboration and software license utilization, as well as enable access to a remote talent pool that would be harder to source when restricted to geographically local staff.

Consider GPU-based DaaS where an operating expenditure (opex) model is preferred and where consumption can be flexed up and down. This delivers increased agility, even if the total cost of ownership (TCO) may be higher over the long term than on-premises options.

Business Impact: Server-side client graphics expand the use case for remote client desktop and application delivery to graphically intensive workloads. Remote delivery models have generally focused on task and process workers and have struggled to penetrate the dominance of the traditional distributed PC model for more demanding users. This technology enables richer content delivery for multimedia and 3D graphics.

Knowledge worker use cases will continue to be cost-sensitive, especially when compared with distributed delivery options. At the higher end, designer use cases are likely to drive a high penetration where the business case is particularly strong in protecting intellectual property, improving storage performance with large data models and/or increasing accessibility.

Industries with significant computer-aided design needs, such as manufacturing, have been early adopters in the technology life cycle. However, this technology should now be considered cross-industry (including healthcare and higher education). I&O leaders are increasingly looking for GPU-based workspace options from cloud service providers.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: AMD; Intel; NVIDIA

Recommended Reading: "Forecast Analysis: Discrete GPUs, Worldwide"

"Market Guide for Desktop as a Service"

Cloud Computing

Analysis By: David Smith

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

Position and Adoption Speed Justification: Cloud computing is a very visible and hyped technology and has passed the Trough of Disillusionment. Cloud computing 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 cloud and enabling capabilities not easily accessible elsewhere. Much of the cloud focus is on agility, speed and other benefits beyond cost savings.

"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 had the effect of increasing 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).

Although the peak of hype has long since passed, cloud still has more hype than many other technologies that are at or near the Peak of Inflated Expectations. Variations, such as private cloud computing and hybrid approaches, compound the hype and reinforce the conclusion that one profile on a Hype Cycle cannot adequately represent all that is cloud computing. Some cloud variations (such as hybrid IT and now multicloud environments) are now at the center of where the cloud hype currently is. And, of course, there are different types of cloud services such as IaaS, PaaS and SaaS, each at various stages of industry hype.

New and advanced use cases for cloud introduce even more terms such as 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.

User Advice: User organizations must demand clarity from their vendors around cloud. Gartner's definitions and descriptions (which align with other useful ones such as NIST) of the attributes of

cloud services can help with this. Users should look at specific usage scenarios and workloads, map their 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 the need to understand the shared responsibility model for security.

Vendor organizations should focus their cloud strategies on more specific scenarios and unify them into high-level messages that encompass the breadth of their offerings. Differentiation in hybrid cloud strategies must be articulated. This will be challenging, as all are "talking the talk," but many are taking advantage of the even broader leeway afforded by the term. "Cloudwashing" should be minimized. Gartner's Cloud Spectrum can be helpful.

Adopting cloud for the wrong reasons can lead to disastrous results. There are many myths surrounding cloud computing as a result of the hype (see "Revisiting the Top 10 Cloud Myths for 2020" for details and advice).

Business Impact: The cloud computing model is changing the way the IT industry looks at user and vendor relationships. Vendors must become providers, or partner with service providers, to deliver technologies indirectly to users. User organizations will watch portfolios of owned technologies decline as their service portfolios grow.

Potential benefits of cloud include cost savings and capabilities related to the flexible and dynamic usage models of cloud (including concepts that go by names such as "agility," "time to market" and "innovation"). Organizations should formulate cloud strategies that align business needs with those potential benefits. Agility is the driving factor for organizations embracing cloud most of the time.

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Amazon; Google; IBM; Microsoft; Oracle; Red Hat; Salesforce; SAP

Recommended Reading: "Cloud Computing Primer for 2020"

"The Cloud Strategy Cookbook, 2019"

"Revisiting the Top 10 Cloud Myths for 2020"

"Four Types of Cloud Computing Define a Spectrum of Cloud Value"

Entering the Plateau

GPU Accelerators

Analysis By: Alan Priestley; Martin Reynolds; Chirag Dekate

Definition: GPU-accelerated computing is the use of a graphics processing unit (GPU) to accelerate highly parallel compute-intensive portions of the workloads in conjunction with a CPU.

Position and Adoption Speed Justification: GPUs are highly parallel floating-point processors designed for graphics and visualization workloads. Over the last decade, NVIDIA and others have added programmable capability to GPUs, enabling software applications to access deep, fast-floating-point resources. A number of GPU designs also host very high-bandwidth memory subsystems. For many highly parallel, repetitive, compute-intensive applications, GPUs deliver dramatic performance improvements over traditional CPUs. GPU subsystems are actively deployed in two key markets: high-performance computing (HPC) and AI.

In HPC, compute-intensive applications including molecular dynamics, computational fluid dynamics, financial modeling and geospatial applications can, in many cases, can be dramatically accelerated using GPUs. Programming GPUs can be challenging because execution order and code optimization are critical. However, toolkits like NVIDIA's CUDA can dramatically lower the programming challenges.

In AI, DNN 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. Although many ASICs are emerging, few offer broad ecosystem support. Ease of programming GPUs, using tools such as CuDNN, and broad ecosystem support continue to be distinct differentiators and advantages.

GPU computing has moved forward on the Slope of Enlightenment primarily due to maturity of system stacks resulting in easier adoption.

User Advice: GPU-accelerated computing can deliver extreme performance for highly parallel compute-intensive workloads in HPC, DNN training and inferencing. GPU computing is also available as a cloud service and may be economical for applications where utilization is low but urgency of completion is high.

Leverage GPU-based solutions to accelerate compatible applications by:

- Selecting GPU compute platforms that offer the most mature software stack.
- Optimizing infrastructure costs by evaluating cloud-hosted GPU environments for proof of concept (POC) and prototype phases.

Use GPU accelerators when applications require extreme performance and have high degrees of compute parallelism (example: many high-performance computing and deep learning applications).

Business Impact: HPC and deep learning are essential to many digital business strategies. For this fast-growing workload, traditional enterprise ecosystems based on CPU-only approaches will not suffice. Leverage mature GPU technologies for select HPC applications and deep learning infrastructures. Programmability challenges have been largely solved in GPU-accelerated computing by toolsets such as CUDA. I&O leaders can minimize risk by using cloud-hosted GPU environments for testing and evaluation.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Mature mainstream

Sample Vendors: AMD; Intel; NVIDIA

Recommended Reading: "Top 10 Technologies That Will Drive the Future of Infrastructure and Operations"

"Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

"Forecast Analysis: Data Center Workload Accelerators, Worldwide"

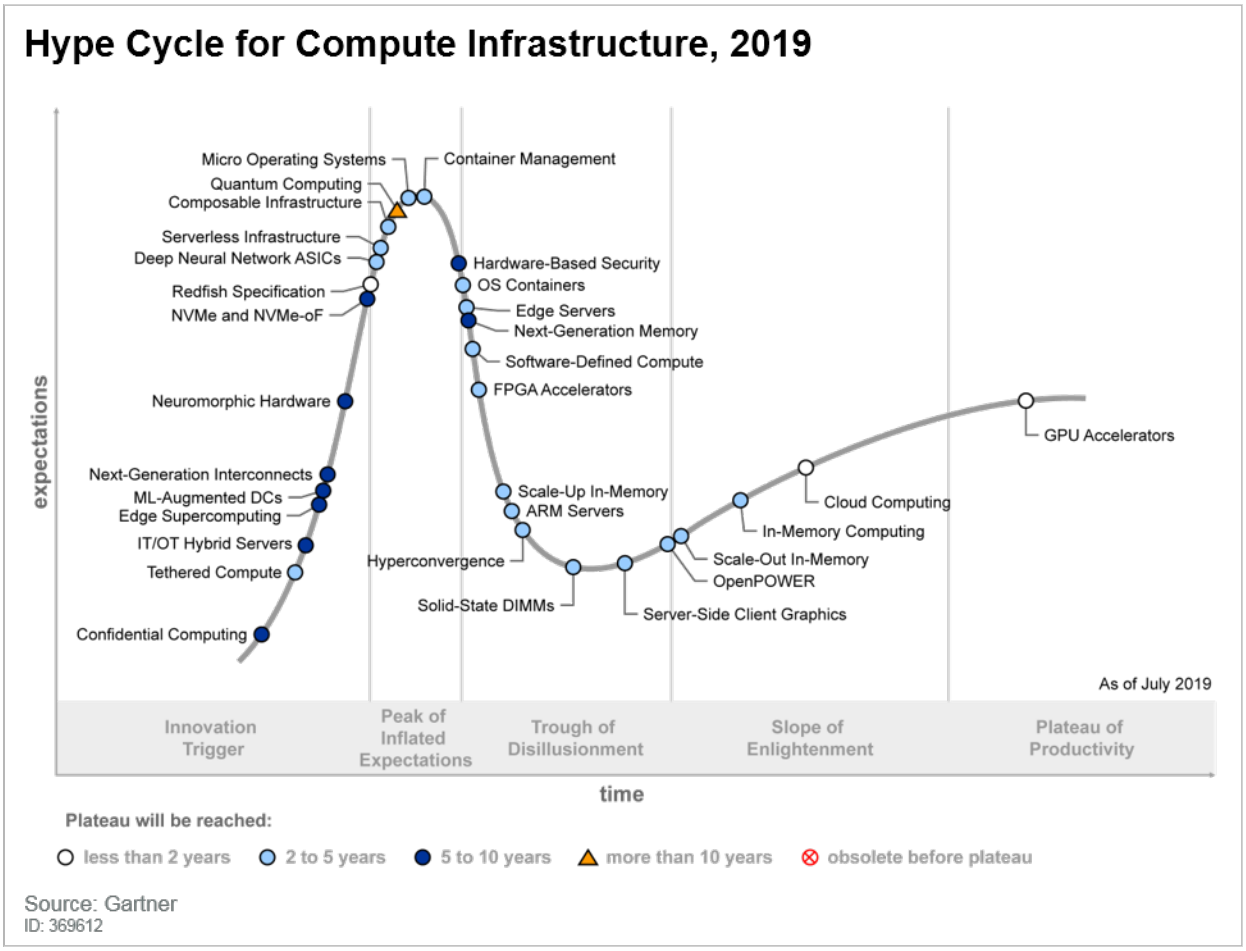
"Forecast Analysis: Discrete GPUs, Worldwide"

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

"Predicts 2019: Artificial Intelligence Core Technologies"

Appendixes

Figure 3. Hype Cycle for Compute Infrastructure, 2019



Source: Gartner
ID: 369612

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 1. Hype Cycle Phases

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant press 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 technology is pushed to its limits. The only enterprises making money are conference organizers and magazine publishers.
Trough of Disillusionment	Because the technology 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 technology'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 technology 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 technology to reach the Plateau of Productivity.

Source: Gartner (July 2020)

Table 2. 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 2020)

Table 3. Maturity Levels

Maturity Level	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 2020)

Gartner Recommended Reading

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

Understanding Gartner's Hype Cycles

2019 Strategic Roadmap for Compute Infrastructure

Market Guide for Compute Platforms

Market Guide for Server Virtualization

Top 10 Strategic Technology Trends for 2019

Evidence

The sources used to arrive at the descriptions, placements and maturities included:

- Analyst conversations with Gartner clients about technology adoptions and plans
- Briefings from technology vendors about current and planned compute technologies
- Gartner market share and other quantitative research data

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp