MARKET NOTE

SwiftStack Stakes Its Play in the AI/ML Market

Amita Potnis

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: SwiftStack Stakes Its Play in the AI/ML Market

On March 12, 2019, SwiftStack announced its multicloud AI/ML Data Management solution. This announcement is significant for SwiftStack as it is backed by customers using this solution. The announcement also marks an internal shift for the company to support demanding new workloads.

Key Takeaways

- Cognitive/AI applications and workloads will continue to shape the storage industry, with the central theme "scale, performance, intelligence, and self-management."
- One of the biggest challenges in the AI/ML segment, specially as it relates to autonomous cars, is the amount of data that is ingested from various endpoints, and the need to store it close to the compute resources while maintaining the ability to move it from and to various locations such as public cloud or on-premises datacenter. SwiftStack's reference architecture that includes Cisco servers and switches, NVIDIA's DGX servers for acceleration, and the company's object-based storage solution that includes 1space (enabling policy-based data placement) is poised to address these challenges.
- SwiftStack's recent entry in the AI/ML space comes with proven backing from two customers that include NVIDIA DRIVE, an autonomous vehicle development platform that uses SwiftStack's object-based storage offering for its DRIVE infrastructure. Currently, the SwiftStack-supported NVIDIA infrastructure supports 1PB of data per car per week with 30 cars under testing.

Source: IDC, 2019

May 2019, IDC #US45026519

IN THIS MARKET NOTE

On March 12, 2019, SwiftStack announced its multicloud artificial intelligence (AI)/machine learning (ML) data management solution. This announcement is significant for SwiftStack as it is backed by customers using this solution. The announcement also marks an internal shift for the company to newer workloads while serving as a proof point for object-based storage.

IDC'S POINT OF VIEW

Artificial Intelligence Overview

AI, ML, and continual deep learning (DL) technologies are poised to transform how consumers and enterprises operate and gain insights. As data becomes core and invaluable to the new digital economy, managing it across edge to core to cloud, analyzing in near real time, and affecting outcomes by learning from and acting on data is becoming increasingly important. Successful organizations leverage AI/ML/DL to deliver meaningful predictions, improving processes within industries such as healthcare and effective decision making.

AI has been around for decades, but because of the pervasiveness of data, seemingly infinite scalability of cloud computing, availability of AI accelerators, and sophistication of the ML and DL algorithms, AI has grabbed the center stage of business intelligence. IDC predicts that by 2019, 40% of DX initiatives will use AI services; by 2021, 75% of commercial enterprise apps will use AI, over 90% of consumers will interact with customer support bots, and over 50% of new industrial robots will leverage AI. AI solutions will continue to see significant corporate investment over the next several years.

As AI makes its way into mainstream digital economy, many organizations find themselves in initial proof-of-concept (POC) stage with only a few in full production stage. Regardless of where organizations are in their stage of adopting AI, building, testing, optimizing, training, inferencing, and maintaining accuracy of models is always top of mind. Any ML and DL algorithm needs huge quantities of training data, and AI effectiveness depends heavily on high-quality, diverse, and dynamic data inputs. Data management of these data sets is complex and challenging. AI is not only reshaping business processes, but it is driving a need to reconsider the underlying infrastructure as well. In the age of digital economy, a key consideration is to manage the data from edge to core to cloud, analyze it in near real time, learn from it, and then act on it to affect outcomes. IoT, mobile devices, big data, machine learning, and cognitive/AI all combine to continually sense and collectively learn from an environment. IDC's *Worldwide Storage for Cognitive/AI Workloads Forecast, 2018-2022* (IDC #US43707918, April 2018) forecasts that software, services, and hardware spending on AI and ML will grow from $12 billion in 2017 to $57.6 billion by 2021.

While AI bears the promise of social impact through examples such as autonomous cars or diagnosis or treatment of rare diseases based on genomic research, it faces several infrastructure challenges and questions around the role of compute and storage resources in an AI infrastructure solution. The reality is that the ideal AI infrastructure is an optimal combination of compute and storage.

Generally, when thinking of AI, most organizations primarily focus on GPU-based parallel processing compute resources for training and inferencing. All-flash arrays (AFA) offerings are largely thought of as optimal storage tiers to support the massive parallelism of the GPUs. As per IDC's *Cognitive, ML, and AI Workloads Infrastructure Market Survey* conducted in January 2018 (n = 405, 1,000+ U.S. employees and 500+ Canadian employees), today, traditional SAN/NAS is largely used for on-

premises run of AI/ML/DL workloads because of their existing deployment footprint and earlier stages of AI adoption, but with the need to scale dynamically, store large volumes of data at relatively low cost, and support high-performance, software-defined storage, hyperconverged infrastructure, and all-flash arrays with newer memory technologies will gain adoption, aligned with the individual offering-specific advantages and the data pipeline stage of AI deployment.

Recent years have seen several solutions geared toward AI workloads that boast a well-balanced solution based on innovative storage architectures that provide performance in terms of availability, capacity, throughput, latency, and IOPs. Some of the primary requirements of any AI storage infrastructure are:

- **Scalability and cost efficiency.** IDC's 2018 survey of 405 IT respondents and decision makers who had completed an AI project in North America indicates that massive data volumes and associated quality and management issues are key AI deployment challenges. The scale of data generated by AI/ML workloads is making customers and vendors consider software-defined object-based storage as a viable storage alternative.
- **Parallel architecture.** GPU compute layer provides massive parallelism, essential for AI, ML, and DL workflows. However, unless the storage layer is able to match with similar parallelism, expensive GPU cycles get poorly utilized, delaying the experiment time to value. At petabyte scale, the storage layer is expected to deliver hundreds of gigabytes of throughput.
- **Data durability and locality.** The storage layer needs to complement the compute layer, by enabling data locality, as well as provide extreme durability at petabyte scale, not just within the datacenter but across geographies. Traditional data protection and durability schemes like RAID and backups fall short at petabyte scale.
- **Reliability and availability.** IDC's 2018 *Server and Storage Infrastructure Availability Survey* indicates that 47% of 358 respondents agree that there is significant difference between the availability features on different storage platforms and they use it for making enterprise storage choices.
- **Hybrid cloud/multicloud data management.** Data is increasingly distributed across on-premises, colocation, and public cloud environments. For this data visibility, access, control, and single-pane-of-glass management, tools across all deployment locations is a must.

What Is SwiftStack?

Over the past nine years, SwiftStack has established itself as a provider of unified object- and file-based data storage and management platform comprised of SwiftStack Storage and 1space. SwiftStack platform is a software-defined storage offering that runs on commodity server hardware. The company claims that its offering simplifies on-premises infrastructure, enables scale and utilization level of data, and extends seamlessly to the public cloud.

The product was initially brought to market as an object-based storage software offering with support for Swift and S3 APIs. The release of SwiftStack 6 in December 2017 enabled built-in file access (SMB/NFS), which is integrated into the core of SwiftStack. This technology has contributed to the open source community in a project called ProxyFS. SwiftStack employs a shared-nothing architecture and can be scaled as needed based on projected workloads. Core to SwiftStack is the ability to scale a cluster with a single namespace across multiple geographic regions. SwiftStack integrates 1space, a centralized policy and data placement functionality that creates a single namespace across private and public cloud storage services from Amazon, Google, and so forth. The solution supports metadata storing, tagging using middleware, and indexing leveraging Elasticsearch along with new search

capabilities provided by SwiftStack client making it easy for end users to find and access data. SwiftStack also provides advanced remote cluster monitoring capabilities, leveraged by its professional consulting team.

SwiftStack claims that it has production customers that span on-premises as well as public cloud, many of whom have deployed 10PB and greater, proving the product's ability to scale. Over the years, SwiftStack boasts over a hundred customers across verticals such as life sciences, media and entertainment, video surveillance, backup and recovery and, most recently, AI and ML.

Many storage vendors have come to market with solutions that are geared toward AI/ML workloads. One thing in common across most of these storage vendors is that the solution almost always incorporates a file-based AFA storage offering running alongside GPU-based compute resources. SwiftStack's AI/ML solution on the other hand uses object-based storage software supported with a file system (ProxyFS) deployed on disk-based Cisco storage servers for throughput and NVIDIA DGX-1 servers for compute. SwiftStack has made bold claims that its solution along with rich metadata tagging and search provided by SwiftStack middleware and the SwiftStack Client will help a customer's AI/ML edge-core-cloud strategy.

Why SwiftStack in AI/ML?

SwiftStack recently announced reference architectures with strategic technology partners for the AI/ML workloads. This solution comprises Cisco UCS S3260 storage servers, Cisco 100GbE network switches, and NVIDIA DGX-1 or NVIDIA DGX station GPU servers for compute. This combination of technology from Cisco, NVIDIA, and SwiftStack is targeted to address all key requirements of AI/ML workflows such as massive scalability, scalability, throughput, low latency, extensibility to the cloud, tagging and search, multi-protocol support, and cost efficiency.

SwiftStack's parallel architecture is transformative for distributed AI, ML, and DL workflows. Traditional storage architectures were designed for contention-based workloads and fall short of performance, scale, and value. SwiftStack can achieve high throughput and concurrency by adding distributed commodity servers and, with enough spindles and pipes, can effectively feed the massively parallel GPU layer. SwiftStack reference architecture is designed for upward of 100GBps of throughput at 15PB scale.

SwiftStack supports both replicas as well as erasure coding for data durability. Multiregion erasure coding can support 4:2, 8:3, and 15:3 policies for 10 x 9s, 14 x 9s, and 12 x 9s durability, respectively. The GPU compute layer has enough memory and flash to make latency and data locality a nonissue.

SwiftStack boasts entry in the AI/ML space with two customers in the autonomous cars industry. One of SwiftStack's customer, NVIDIA, leverages it within its AI/ML hybrid cloud infrastructure stack purpose built for optimal compute and storage resources, allowing data to be stored closer to compute resources. NVIDIA is enabling the automotive market with its NVIDIA DRIVE — Autonomous Vehicle Development Platforms, and using SwiftStack for its DRIVE infrastructure. To begin, NVIDIA collects multiple petabytes of data from cars, which is then pushed into the public cloud for backup. Data is then replicated locally to purpose built on-premises AI/ML infrastructure using the S3 API via a high-bandwidth interconnect for training, testing, and so forth. SwiftStack is deployed on 12 storage nodes amounting to 1.2PB of storage per node (15PB/rack) with 9 DGX servers amounting to 9 petaflops of compute power. SwiftStack serves as an integral part of the entire process, especially when housing ever-growing data sets that amount to over 1PB of data per week per car while testing approximately 30 cars simultaneously. Data housed on SwiftStack is enriched with a labeling UI supported by over a thousand human data labelers who label tens of million objects per month.

SwiftStack's AI/ML Technology and Platform Partnerships

In addition to proven customer success in the autonomous cars space, SwiftStack also offers its capabilities via technology ecosystem with Valohai and other ISV partners. Valohai offers a DL PaaS that automates machine orchestration, version control, and pipeline coordination for data science teams. Valohai integrates with SwiftStack with a choice of S3 or Swift protocols and users can scale models to hundreds of CPUs or GPUs on-premises or in cloud.

SwiftStack has value-added reseller (VAR) partnerships with GPL Technologies, SHI, World Wide Technology, Presidio, and others, which provides customers with the flexibility, technology, and economics to build purpose-built solutions to match their use cases.

The Bottom Line

Cognitive/AI applications and workloads will continue to shape the storage industry, with the central theme "scale, performance, intelligence, and self-management." Naturally, storage suppliers are aggressively building or need to build AI workload-centric products and solutions. Support for multiformat storage infrastructure is also crucial, given that machine learning, cognitive computing, and other forms of AI must pull both structured and unstructured data from multiple sources that rely on modern APIs such as S3 API, Swift API, NFS, and SMB. The new infrastructure will have to push speed, agility, and scale to entirely new levels if the enterprise hopes to draw meaningful results from all that number crunching.

SwiftStack with its latest AI/ML reference architecture, proven customer success, and technology partnerships is positioned for successful deployments and customer acquisition.

LEARN MORE

- *Worldwide File-Based Storage Forecast, 2018-2022: Storage by Deployment Location* (IDC #US44457018, December 2018)
- *Enterprises to Adopt Cloud-Native Applications in Next 12 Months: Drivers Include Security, Costs, Big Data AI/ML Initiatives* (IDC #US44448818, November 2018)
- *Data Management: Success with a Method to the Madness* (IDC #US44415618, November 2018)
- *Red Hat Acquires NooBaa – Makes a Shift from Storage to Hybrid Cloud Data Management* (IDC #IcUS44484218, November 2018)
- *Worldwide Composite Media Workloads (Compute and Storage) Infrastructure Forecast, 2018–2022* (IDC #US44281818, October 2018)

Synopsis

This IDC Market Note discusses SwiftStack's multicloud AI/ML data management solution.

"Infrastructure challenges are the primary inhibitor for broader adoption of AI/ML workflows," said Amita Potnis, research director at IDC's Infrastructure Systems, Platforms and Technologies Group. "SwiftStack's multicloud data management solution is the first of its kind in the industry and effectively handles storage I/O challenges faced by edge-to-core-to-cloud, large-scale AI/ML data pipelines."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com