IDC MarketScape

IDC MarketScape: Worldwide Distributed Scale-Out File System 2022 Vendor Assessment

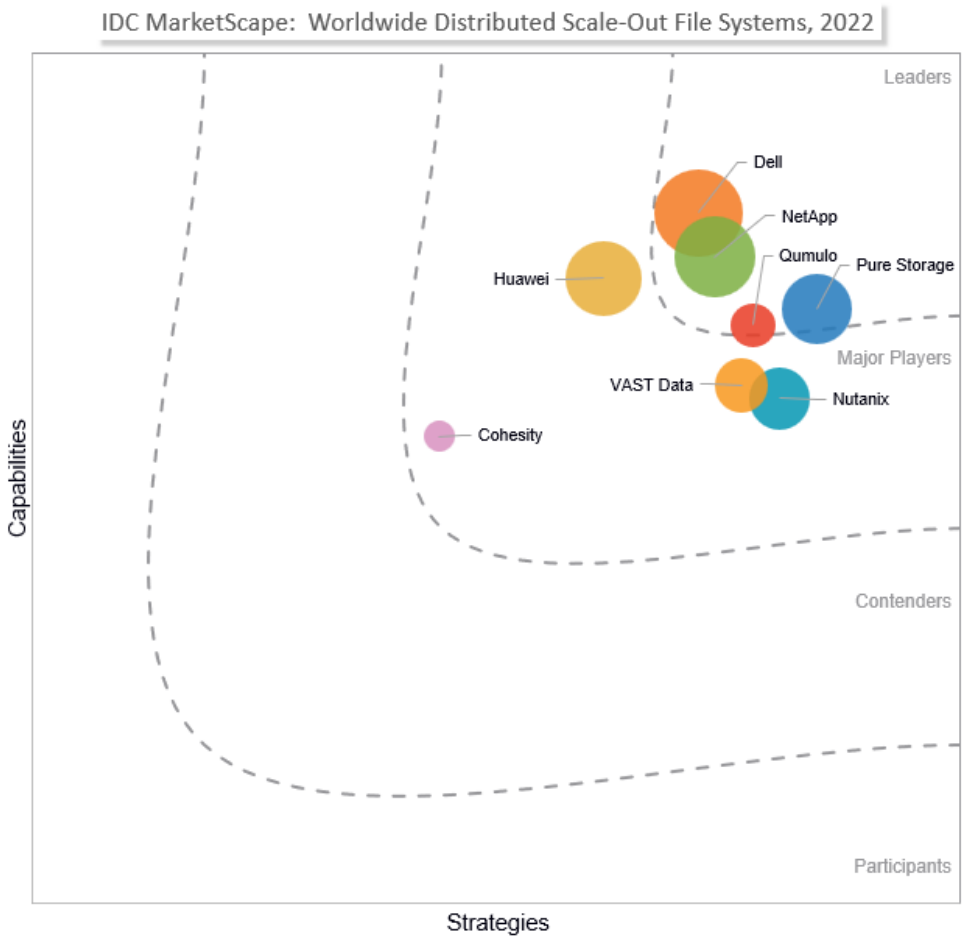
Eric Burgener

THIS IDC MARKETSCAPE EXCERPT FEATURES PURE STORAGE

IDC MARKETSCAPE FIGURE

FIGURE 1

IDC MarketScape Worldwide Distributed Scale-Out File System Vendor Assessment



Source: IDC, 2022

Please see the Appendix for detailed methodology, market definition, and scoring criteria.

April 2022, IDC #US49015322e

IN THIS EXCERPT

The content for this excerpt was taken directly IDC MarketScape: Worldwide Distributed Scale-Out File System 2022 Vendor Assessment (Doc # US49015322). All or parts of the following sections are included in this excerpt: IDC Opinion, IDC MarketScape Vendor Inclusion Criteria, Essential Guidance, Vendor Summary Profile, Appendix and Learn More. Also included is Figure 1.

IDC OPINION

Over the next five years, scale-out file systems will be widely deployed by enterprises looking to consolidate file-based workloads, improve file-based infrastructure efficiencies, and handle many of the performance and scalability requirements of modernized applications that are very data intensive. All of the products evaluated here will be able to do that very well for most enterprises, although there are some differences in top-end performance and scalability and ease of use between offerings — that is why Figure 1 has many of the vendors clustered closely together. What the reader should note, however, is that there can be significant differences between vendors in their architectures, product strategies, areas of focus, and software-defined flexibility that should be evaluated as purchase decisions are made.

The "Advice for Technology Buyers" section is probably the most important section to read for those who will be involved in making a purchase decision. This section introduces a number of strategic questions enterprises should ask themselves when determining what is most important in selecting a scale-out file system offering. As an example, all evaluated products can support a 1PB file system, but what each system looks like, how easy it is to manage and upgrade, how much it costs and, in general, how it gets there can be very different. There is no "best" offering in this market, but there are certain products that are better suited for certain workloads and will cater better to certain objectives like top-end performance and scalability, ease of use and management, lower energy and floorspace consumption, hybrid cloud capabilities, and how different access methods are supported.

Enterprises can expect a lot more innovation to occur in the scale-out file market going forward, driven primarily by the fact that 80% of the data that will be created over the next five years will be file and/or object based. If enterprises just need to simplify basic file sharing (home directories, etc.), there are a lot of very viable options (some of which are mentioned in the "Vendors to Watch" section). Modernized applications, particularly those using artificial intelligence (AI) or those which are very data intensive, will have additional demands that may not be well met by the simpler products, and that's where enterprises will need to turn to true distributed scale-out file system platforms.

IDC MARKETSCAPE VENDOR INCLUSION CRITERIA

This IDC study assesses the capabilities and business strategies of popular suppliers in the distributed scale-out file-based storage market segment. For a complete definition of distributed scale-out file systems (and a discussion of the new file-based storage taxonomy that IDC introduced in July 2021), see *Reclassifying File Storage — A New Approach for the Future of Digital Infrastructure* (IDC #US48051221, July 2021). This evaluation is based on a comprehensive framework and a set of parameters that gauge the success of a supplier in delivering a scale-out file-based storage solution to the enterprise market.

To be evaluated in this study, a vendor needs to have a scale-out file-based storage platform:

- **That conforms to IDC's taxonomy.** According to *Reclassifying File Storage — A New Approach for the Future of Digital Infrastructure* (IDC #US48051221, July 2021), assessed products need to meet the definition of a distributed scale-out file system platform or a clustered scale-up file system that is sold primarily against distributed scale-out file systems.
- **Whose intellectual property (IP) is fully owned by the vendor.** The vendor being assessed has developed the distributed scale-out file-based storage solution in-house or obtained the technology through acquisition.
- **That was generally available by September 2021 and generates at least $30 million in annual revenue.** This is to ensure that the vendor product has at least some level of maturity and market traction.

ADVICE FOR TECHNOLOGY BUYERS

Given that the vendors in this assessment are using widely varying product strategies, an important place to start the evaluation process for an enterprise is to understand which of the different approaches appeal to the enterprise and/or are a better fit for its needs. Do you like the idea of being able to manage block-, file-, and object-based workloads on the same storage system through a unified management interface? Do you prefer unified storage (which can avoid semantic loss issues but will use more storage capacity to provide multiprotocol access to the same data object) or multiprotocol access (which uses less storage capacity but where semantic loss may be an issue)? Are you a federal agency that requires FIPS 140-2 compliant encryption? Do you prefer a storage architecture built around server-based storage nodes or are you open to different architectures that may offer differentiators in certain environments? Six of the vendors assessed use server-based storage nodes (although some of them have some proprietary content), while two — NetApp and Pure — use different architectures.

Would you prefer to use traditional access methods like NFS and SMB but also have access to an intelligent client that offers significantly more parallelization if/when you might need it? Other vendors will tell you how they've extended the performance capabilities of NFS over TCP beyond the 2GBps limit per mount point with nconnect or features specific to their platform that still use the standard NFS client (for example) so you don't have to deploy an intelligent client. Do you require NDMP support? Are you interested in the idea of a cacheless architecture that can offer very high degrees of data concurrency or do more traditional cache-based architectures meet your needs just fine? Do you need POSIX compliance? POSIX really isn't the future, but there are hundreds of thousands of already deployed applications that use it.

Do you have a preference for an HCI-based architecture (like Cohesity or Nutanix) or a disaggregated storage approach? Do you want to buy your solution from a major OEM (Cisco sells Cohesity, Dell sells Nutanix, and HPE sells Qumulo) or would you prefer to buy it from the developing vendor directly (or a channel partner of theirs)? Do you like the idea of combining data protection and enterprise file sharing under a single system or not? While this is not an exhaustive list of questions, these are the kinds of questions an IT manager should ponder when evaluating scale-out file systems for enterprise workloads.

As with most enterprise workloads, high availability (HA) is important and enterprise file sharing is no exception. Solutions that have been around for a long time tend to have an extensive, proven feature set in this area. Understand your recovery point objectives (RPOs) and recovery time objectives (RTOs) for both local and disaster recovery, and match that with capabilities in the scale-out file

system offerings. Tunable erasure coding (EC) (so data durability and capacity utilization can be set differently for different workloads), snapshots, replication, a simple "snap to object" feature that makes it very easy to back up the entire namespace to an external object store, air-gap protection to defend against ransomware, and integration with third-party backup products like Commvault and Veritas, all these are features that can impact data protection workflows, availability, and recovery times.

Ease of management at scale is another differentiating area. There are many challenges in managing scale-out file system environments, and there has been a lot of employee interchange between the various scale-out file system players in the past 20 years. The challenges are well known at all vendors, but how they address them varies. If you have managed a scale-out file system before, what are your hot-button issues?

- Do you need absolutely the lowest latencies for random small file accesses or are sub-millisecond average response times good enough?
- Are you trying to consolidate workloads across your data stage pipelines that need both native and intelligent client-based access methods?
- Do you want to be able to rapidly create delta differentials for backup purposes without having to walk all the file trees?
- Do you want particularly low-capacity utilization of on-disk data protection options at your target level of durability because you have multiple petabytes of data under management?
- Do you need support for compression and/or deduplication because your data sets can benefit significantly from these technologies (or not, since much unstructured data does not compress and/or deduplicate very well)?
- Are disruptive upgrades and slow disruptive recovery in SMB environments a particular pain point?
- Are you particularly concerned about large capacity drive rebuild times or how easy and nondisruptive it is to expand the cluster by adding a new node?
- Are you concerned about how easy and efficient is it to use file quota management systems?

These (and many more) are all issues many scale-out file system administrators have struggled with.

The key to selecting a platform best suited for your requirements is to thoroughly understand your needs and preferences up front. The vendors assessed here all provide a range of performance, scalability, availability, and core functionality that meet the requirements for most enterprise file-based workloads, but among the eight vendors, there are very different ways to get there and very different emphases in their product designs. List what is most important to you, and map that to the vendor offerings. Doing that will require going beyond this document since we do not provide direct head-to-head comparisons between vendors. IDC has, however, published a number of technical reviews of different vendor offerings in separate research, discussing the benefits of the approaches they have taken.

VENDOR SUMMARY PROFILE

This section briefly explains IDC's key observations resulting in a vendor's position in the IDC MarketScape. While every vendor is evaluated against each of the criteria outlined in the Appendix, the description here provides a summary of the vendor's strengths and challenges.

Pure Storage

Pure Storage is positioned in the Leaders category in the 2022 IDC MarketScape for worldwide distributed scale-out file system.

Founded in 2009, Pure Storage is a large, publicly held enterprise storage vendor that sells only all-flash storage. The vendor changed the industry with its block-based FlashArray (which originally shipped back in 2012), and in 2016, it entered the unstructured storage market with its FlashBlade (which supports both file-based and object-based storage in the same system). FlashBlade has been very successful for Pure Storage, and had it been an independent business, it would have achieved unicorn status several years ago and actually crossed the $1 billion in lifetime sales line in June 2021. Even as the industry experienced a downturn during the pandemic years, Pure Storage was able to turn in steady revenue growth and has an installed base of 10,000 customers (across its entire enterprise storage portfolio).

FlashBlade is a unified (rather than a multiprotocol) storage platform supporting NFS, SMB, and S3 access methods and using an underlying key value store data organization method. It is fully hardware redundant, offers hot-plug replacement of all components, and delivers very high availability in production usage. It supports a broad range of data services, including compression, global EC (which can be spread across all blades across chasses in a cluster), replication, snapshots (including immutable SafeMode snapshots), audit logs, and 256-bit encryption. The system's efficient all-flash design requires less power and cooling and offers higher infrastructure density (for both performance and capacity) than many competitors. This allows smaller, more compact systems to meet customer requirements, also saving on datacenter floorspace. FlashBlade can be monitored and managed by Pure1 META, the vendor's AIOps hybrid cloud management platform.

Strengths

One of Pure Storage's original design tenets, reflected in both its FlashArray and its FlashBlade platforms, is that an all-flash array (AFA) overcomes so many of the performance issues in hybrid and HDD-based storage systems that defaults can be widely used when deploying systems (dual-parity RAID, always-on compression and encryption, etc.). This makes the FlashBlade extremely easy to use, a feature consistently noted by its end users. There are a very few settings which administrators can configure, but there are a very few times when an administrator may want those capabilities (unlike some other systems which often require sophisticated manual tuning expertise). FlashBlade's ease of use extends from initial deployment and storage provisioning to system expansion, upgrades, and failed component replacements.

The other strength of FlashBlade is its ability to deliver consistent performance at scale. FlashBlade uses a cacheless, scale-out architecture that requires significantly less external cabling than clusters built from server-based storage nodes. FlashBlade is a pluggable blade-based scale-out architecture. Its storage devices are called "blades," available in 17TB and 52TB capacities, and each chassis can accommodate up to 15 blades. The blades are a proprietary design — not off-the-shelf SSDs — and include both performance and capacity resources. The flash media on each blade is managed globally by Pure's Purity//FB storage operating system, and these devices have very different and better performance, endurance, and overprovisioning profiles than off-the-shelf SSDs. The internal backplane is based on Ethernet, all the blades are directly connected using the NVMe protocol, and as a result, the system delivers a higher degree of concurrency (an issue particularly important in dealing with large data analytics and densely consolidated storage workloads) than most competitors.

Contributing to FlashBlade's performance characteristics is its scale-out metadata architecture (based on a variable block metadata engine and distributed transaction database), a design which enables it to handle billions of files and objects with equally good performance for small and large files as well as random and sequential access.

CX has always been a focus for Pure Storage, and it has published the industry's only independently validated NPS for five years now. It is notable that its NPS has consistently been in the mid-80s over this period, in particular because for most enterprise storage vendors the quality of their CX tends to degrade as their company grows. Many factors contribute to Pure Storage's CX across the entire product life cycle, including its online sales quoting system, its self-service management interface, the infrastructure density of its arrays, the ease of use in managing its all-flash systems, the consistently high quality of its technical support, and its Evergreen Storage program.

Challenges

FlashBlade uses proprietary hardware. While the vendor makes good arguments about the benefits it offers to customers, that may be an issue for some enterprises. And while the performance it delivers across its access methods is very good, it does have a limited set of them. Many other distributed scale-out file storage platforms offer a broader array of access method options. Pure Storage is committed to native access methods and has no plans to introduce an intelligent client. While the FlashBlade architecture can handle high data ingest rates using NFS, SMB, and S3, it cannot compete with the "throughput to a single large file" performance of parallel scale-out file systems.

While FlashBlade can tier to external HDD-based storage using S3, it does not directly support HDDs. Although Pure Storage has features that lower the cost per gigabyte at the system level, it is not one of the less expensive distributed scale-out storage platforms to buy, but when it comes to all-flash systems, IDC strongly suggests that it is most important to look at overall total cost of ownership rather than just initial purchase price. Its cost profile makes it less suitable for colder storage workloads, although the advantages of its all-flash design are evident for workloads that have any kind of performance sensitivity, whether that is in supporting high degrees of data concurrency or very rapidly moving large data sets. Roughly 25% of FlashBlade's customers use it as a backup repository, citing its write ingest, infrastructure efficiency, and rapid restore advantages.

FlashBlade is not quite as "cloud friendly" as FlashArray, the vendor's dual-controller array. While FlashBlade supports file- and object-based replication to cloud-based targets, Purity//FB is not available in a software-defined version that can be run in the public cloud. And the platform lacks a deduplication feature, although deduplication does not provide much value for many unstructured data workloads. FlashBlade also does not support NVIDIA's GPUDirect Storage API, although it does offer a converged infrastructure stack offering (AI-Ready Infrastructure [AIRI]) with NVIDIA that includes NVIDIA DGX accelerated compute servers, FlashBlade storage, and Mellanox NVMe-oF networking — all under a single purchase SKU and with a single point of support contact with Pure Storage. The vendor's strategy with FlashBlade is to use industry-standard protocols as much as possible and argues that its high all-flash performance and high degree of data concurrency allow it to do an excellent job of keeping GPUs fed with data (without requiring the use of proprietary interfaces like GPUDirect Storage).

Consider Pure Storage When

FlashBlade excels at delivering high-performance, high infrastructure density and ease of use for unstructured data storage environments. Enterprises with FlashBlades also comment on the system's

ability to densely consolidate workloads with differing I/O profiles, a capability enabled by the high data concurrency it supports. Simultaneous use of FlashBlade as both a backup appliance with rapid restore and a platform for big data analytics projects is very common in its installed base. Top verticals generating revenue for FlashBlade include financial services, professional, technical and business services, government (FlashBlade is FIPS 140-2 compliant and has a thriving federal business), healthcare and life sciences, research, electronic design automation, and advertising, media, and entertainment.

APPENDIX

Reading an IDC MarketScape Graph

For the purposes of this analysis, IDC divided potential key measures for success into two primary categories: capabilities and strategies.

Positioning on the y-axis reflects the vendor's current capabilities and menu of services and how well aligned the vendor is to customer needs. The capabilities category focuses on the capabilities of the company and product today, here and now. Under this category, IDC analysts will look at how well a vendor is building/delivering capabilities that enable it to execute its chosen strategy in the market.

Positioning on the x-axis, or strategies axis, indicates how well the vendor's future strategy aligns with what customers will require in three to five years. The strategies category focuses on high-level decisions and underlying assumptions about offerings, customer segments, and business and go-to-market plans for the next three to five years.

The size of the individual vendor markers in the IDC MarketScape represents the market share of each individual vendor within the specific market segment being assessed, not the overall storage-related revenue of the vendor.

Several suppliers offer different file system offerings, although they do not all necessarily compete in the distributed scale-out file system segment. In cases where the vendor offers two scale-out file system types, IDC has worked with the vendor to select the product that most closely fits within the inclusion criteria of this study.

IDC MarketScape Methodology

IDC MarketScape criteria selection, weightings, and vendor scores represent well-researched IDC judgment about the market and specific vendors. IDC analysts tailor the range of standard characteristics by which vendors are measured through structured discussions, surveys, and interviews with market leaders, participants, and end users. Market weightings are based on user interviews, buyer surveys, and the input of IDC experts in each market. IDC analysts base individual vendor scores, and ultimately vendor positions on the IDC MarketScape, on detailed surveys and interviews with the vendors, publicly available information, and end-user experiences in an effort to provide an accurate and consistent assessment of each vendor's characteristics, behavior, and capability.

Market Definition

In July 2021, IDC introduced a new taxonomy for the file system market. There are four segments to the file system market: scale-up file storage, scale-up clusters, distributed scale-out file storage, and parallel scale-out file storage. The scale-up segment is small and shrinking in size, while all the growth

is being driven by scale-out products. Briefly, scale-out file systems distribute data across nodes while presenting a single data access namespace. There are some differences, however, in how data is distributed between scale-up clusters and scale-out file storage. In scale-up clusters, data is rarely ever distributed across nodes, and the throughput to a given file is limited to the bandwidth of the single node from which it is served. In scale-out clusters, data in a single file can be distributed across nodes, a design which can improve access performance, data concurrency, and recovery time.

Scale-up clusters and distributed scale-out file storage routinely compete for the same business in enterprises, and this vendor assessment includes vendors from both segments. For more detail on how each of these segments is defined, see *Reclassifying File Storage — A New Approach for the Future of Digital Infrastructure* (IDC #US48051221, July 2021).

Evolution in the Distributed Scale-Out File System Market

File system platforms have been widely used in the enterprise for file sharing. In the early 2000s, data under management grew, new types of file-sharing workloads emerged, and scale-out designs for file sharing began to become more popular. Distributed scale-out file systems became a mainstream alternative to the NetApp filers that dominated file sharing in the 1990s, and NetApp introduced a clustering capability to extend the scalability of its own offerings beginning with the release of "Clustered Data ONTAP" in the late 2000s.

Target workloads for these types of platforms included post-production and media streaming in the media and entertainment market, imaging and video, home directories, local and distributed file sharing, test and development, batch analytics, and backup/archive (although this latter workload was also a major target for many object-based storage vendors). Over most of the life of the scale-out file system market, two platforms were clearly huge players in the market: Dell, which had obtained the Isilon (now PowerScale) product with the acquisition of EMC in 2016, and NetApp, which has been focused on enterprise file-based storage since the company's founding in 1992.

Over the past 10 years, what IDC refers to as "second generation," distributed scale-out file systems were introduced by a number of mostly start-up vendors (Cohesity, Huawei, Nutanix, Pure Storage, Qumulo, and VAST Data). These newer platforms were characterized by more software-defined designs, focused on providing easier management at scale, improved storage infrastructure efficiencies, and in general being more "cloud-friendly." Some were specifically designed around newer storage technologies like NVMe, storage-class memory, and NVMe over Fabrics (NVMe-oF). Both Dell and NetApp have responded, and the distributed scale-out file system market is very different in 2022 from what it was in 2012.

Today, most vendors claim performance and scalability as differentiators, but in selecting a platform, enterprises should focus more on the different facets of performance and scalability that are important to their workloads as there are significant differences in these capabilities across the vendors reviewed in this study. Ease of use is another major differentiator between vendors. Selecting the right scale-out file-based storage platform demands that potential buyers look beyond high-level marketing messages proffered by vendors to understand which products best fit their unique storage I/O requirements.

Key Differences Among Vendors in Product Design Strategies

In today's digitally transforming world, enterprises are capturing, storing, protecting, and analyzing more data than ever before to drive better business insights in much more data-centric business models. To accommodate newer big data analytics workloads and increased scale, roughly 70% of

enterprises going through digital transformation also plan to modernize their server, storage, and/or data protection infrastructure by 2023. In doing so, they are looking for more deployment and purchasing flexibility, simplified management at multi-petabyte levels of scale, increased performance and availability, better affinity with a hybrid multicloud world, and improved infrastructure efficiencies that allow them to pursue denser storage workload consolidation (to narrow not only the number of storage platforms that must be supported but also the number of vendors).

There are several areas where the designs and product strategy focus of certain vendors diverge:

- **Software-defined storage.** "Software defined" is all about improving flexibility, whether that is the flexibility to deploy on different types of server hardware from different vendors, the ability to deploy the file system stack in the public cloud, or the ability to easily accommodate new storage devices and technologies over time. Software defined also tends to offer better technology refresh models. While most enterprises want to buy appliances that offer a single point of support contact and single SKU purchasing, they like the ability to select the hardware of their choice and have a vendor deliver it that way (usually through a channel partner that creates the combined hardware/software platform). Some vendors only deliver appliances on a single type of server hardware (i.e., their own), while others offer a variety of hardware options with their appliances.
- **Access methods.** Most of the assessed vendors are committed to using traditional access methods like NFS and SMB that do not require the installation of custom software on the client side. While there are certain performance scalability limitations to the use of native protocols, they do tend to meet most of the needs of enterprise workloads, and they are easy to deploy and manage. A number of the assessed vendors support options (like "nconnect" for NFS) to improve the scalability of native access methods. Parallel scale-out file systems (which are not being assessed in this document but will be assessed in a future IDC MarketScape document) use proprietary intelligent clients that support a parallelism that allows their throughput to a single large file to go significantly beyond where NFS and/or SMB can go, but a few legacy enterprise workloads can benefit from that. Many enterprises are deploying AI-driven big data analytics workloads as part of digital transformation, and certain stages of the AI data pipeline actually can benefit significantly from this increased throughput though. Several of the assessed vendors not only focus on native access methods (e.g., NFS and SMB) but also offer an intelligent client option that enterprises can deploy and use if/when needed.

  In general, the more access methods a scale-out file-based storage platform supports, the more options there are for denser workload consolidation. While NFS and SMB are the most popular file-based access methods, a number of vendors support other options as well like FTP, HTTP, and HDFS (although many HDFS workloads that are being modernized are moving to object-based storage). Many new applications being developed and deployed during digital transformation use Amazon's Simple Storage Service (S3) interface. Even though that is an object-based access method, an increasing number of vendors allow their file-based storage to be accessed over S3.
- **Cloud native.** Many new workloads are deployed in the public cloud, and enterprises are at the same time evaluating the disposition of their legacy on-premises workloads (rehost into a virtual machine [VM], refactor for cloud deployment, re-architect for optimized cloud deployment, replace [usually with a cloud-based version], or retire). More software-defined designs make this easier since a software-only product can also be deployed on web-scale infrastructure in a public cloud environment. Other features impacting "cloud-friendliness" include microservices design, container-based deployment, support for the Container Storage Interface (CSI, an interface that allows the storage system to provide persistent storage to

applications running in containers), APIs that support Kubernetes-based automation, and unified management consoles that provide comprehensive visibility into workloads that span on-premises and off-premises deployment models (e.g., an instance of a distributed scale-out file system running on hyperconverged infrastructure [HCI] on premises and also on web-scale infrastructure in a public cloud, with the two instances collaborating on a workflow).

Subscription-based licensing may also be viewed as a "cloudlike" capability, and most of the vendors offer this as an option (or as the only way to purchase their software). Several of the vendors (Cohesity, Nutanix, Qumulo, and VAST Data) started out as appliance vendors but have moved to a new software-only business model, making appliances available through channel partners. As long as customers still enjoy the ability to buy appliances that ease purchasing and deployment and have a single point of support contact for the file-based storage solution, they tend to be indifferent to this change, but it can have a major impact on vendor operations, margins, and revenue per employee.

- **Unified storage or multiprotocol access.** There are several different product strategies here. Two of the vendors (Huawei and Nutanix) can support block-, file-, and object-based storage on a single system, all managed from a unified interface. The software-defined nature of both of these platforms provides the flexibility to configure different storage pools within the system for different I/O profiles and access methods. One vendor supports both file- and object-based access methods (but not block) to data, although the data organization method is a key value store (Pure Storage). (Note that Huawei also supports this approach for file and object data, but it uses a separate volume-based data organization method for block.) If a data object can be natively written using either a file-based interface or an object-based interface to the key value store but to have it natively accessible by multiple interfaces requires multiple copies of the data, IDC refers to that as "unified storage."

  Other vendors use a file-based data organization method but support multiprotocol access to the same data through a variety of interfaces like NFS, SMB, HDFS, FTP, HTTP, NDMP, and S3. (IDC refers to this as "multiprotocol access.") With multiprotocol access, storage capacity is used more efficiently (both NFS and S3 access the same underlying data object), but the issue of semantic loss may arise. Semantic loss occurs when the interface through which data is accessed (e.g., NFS) does not support all the primitives of the interface used to initially write the data (e.g., S3) potentially limiting an application trying to access non-native data. This can be an issue for some applications but not for others.
- The ability to support multiple access methods to the same data on a single platform can make working with multistage data pipelines much easier. Data does not need to be copied over a network to another system, which must be managed separately and may require a different administrative skill set. Fewer systems can be purchased, and potentially fewer vendors can be managed. Software-defined flexibility can allow storage to be configured in a single system to meet a variety of different I/O profiles that may be required in different stages of a data pipeline. Sharing data can allow workload consolidation onto fewer platforms, but there are clearly caveats in doing so (risks to meeting performance and/or availability SLAs, security in sharing data across workloads, etc.).
- **On-disk data protection.** While file systems historically used replicas to protect data on disk, vendors have introduced interesting innovations that can provide better capacity utilization and higher durability and/or enable higher performance access to files. EC, which had historically been used in object-based storage platforms, is now available on a number of the vendor offerings evaluated in this assessment. EC distributes data more widely across devices and/or nodes, splitting it into data and parity bits. This approach makes better use of available storage capacity (than making full file copies) while offering the same or better data durability. Like replicas, EC can span geographical sites to provide site-level resiliency.

One vendor (VAST Data) has implemented an EC approach that offers significantly lower capacity utilization to meet data resiliency requirements for at-scale configurations (larger than 1PB) by distributing data more widely than any other vendor. Other vendors can use replicas for files below a certain size, and then transparently switch to EC as files get larger to save space. (Use of replicas for data protection may produce lower access latencies for small files, while more widely distributed EC can provide higher throughput for larger files.) And other vendors use a more RAID-like approach that operates at the block level, which allows it to recover just the missing data rather than full files for faster rebuilds. Each approach has its pros and cons, depending on customer preference for access latency, fast rebuild times, lower capacity utilization, and data durability.

- **Storage architecture.** All of the assessed systems but one (Pure Storage FlashBlade) use cache-based architectures. As write ingest scales, cache-based architectures can eventually hit a "write cliff" when the write cache is not being destaged to persistent storage as fast as it is filling up. When this occurs, write performance drops noticeably. The write cliff can be pushed out by adding nodes and more widely distributing the load (which all vendors do), creating larger high-speed caching tiers (which VAST Data has done by using a very large persistent storage-class memory-based layer as a write cache), and using other innovative software-based techniques to extend write performance. Cacheless architectures write directly to persistent storage, which is of course much larger than any caching tier (persistent or volatile) but often offers higher write latencies because persistent storage usually is slower than the memory media generally used in caching tiers. Pure Storage, which uses only solid state media in its FlashBlade, will explain why it thinks its approach is a better fit for file- and object-based workloads in the enterprise, and customers can decide for themselves which is best for their environment.
- **Data management strategies.** As rapid data growth continues, enterprises want features that allow them to implement more efficient data management strategies. Although data tiering (both within a system and to external targets) has been available for a long time, providing the visibility that enables effective data classification is really the right starting point for intelligent data management. While file usage metrics like frequency of access have long been used to determine data placement, new AI and machine learning (ML) technologies allow data placement to be better optimized in real time for performance as well as across tiers to reduce infrastructure and management costs. They can also identify data that can be safely deleted, ensuring that enterprises are only keeping data that has to be retained. While some vendors realized the importance of intelligent data placement early on, at this point, pretty much all of them are doing at least something in this area beyond just tracking frequency of access.
- **Deploying the scale-out file system.** All of the vendors require data to be migrated from third-party file systems into their own scale-out file system design. This is considered standard operating procedure to be able to take advantage of all the features of distributed scale-out file systems but is in contrast to certain file-based storage players (not evaluated in this vendor assessment) whose software layers on top of existing file systems to provide a unified namespace (requiring no data movement). One vendor in this assessment uses a very different strategy, enabling backup data to be converted into one or more scalable, shared access scale-out file systems. The strategy of this vendor (Cohesity) effectively combines data protection and enterprise file sharing into a single, centrally managed platform. Some enterprises find this very convenient, while others prefer to keep the two practices (data protection and file management) separate.

What enterprises can expect from half of the players in this market is that the scale-out file system software runs on commodity x86 hardware, and vendors generally support multiple hardware options

(e.g., Dell PowerEdge, HPE ProLiant, Lenovo, and/or Supermicro). Four of the vendors still prefer that customers buy appliances running their hardware of choice (Dell PowerScale, Huawei OceanStor Pacific, NetApp ONTAP, and Pure Storage FlashBlade), claiming that there are performance, availability, and/or other advantages that accrue. IDC expects that in the future most successful vendors in this space will provide a software-only version that can run on commodity x86 hardware, if for no other reason than to offer the opportunity to deploy their file system in public clouds (several vendors in this study in fact already do this). But it is true that many enterprises see advantages to using purpose built versus commodity hardware — Dell, Huawei, NetApp, and Pure Storage do this and are huge market players in scale-out file systems, although revenue growth rates are higher among the start-up players.

Other baseline expectations should include entry-level configurations that are highly available to require at least three nodes (although non-highly available configurations for edge deployments may be supported in a single VM), an ability to scale performance and capacity by adding nodes up to 100+ (in some cases quite a bit more), and an ability to mix node types (e.g., performance-intensive nodes that might be all-NVMe, hybrid nodes that can support a mix of SSD and HDD, and archive or capacity-intensive nodes that may be all HDDs). This ability to support mixed nodes enables most of these systems to offer a technology refresh model that is nondisruptive (just add the new node types as they become available) and can preserve existing investment (customers are not required to get rid of older nodes if they don't want to even as they add newer nodes).

LEARN MORE

Related Research

- *Qumulo Exhibiting Strong Momentum as It Serves the Evolving Unstructured Data Storage Needs of Enterprise Customers* (IDC #US48896622, March 2022)
- *VAST Data: A Technical Deep-Dive Look at a Compelling New Scale-Out Storage Architecture* (IDC #US48805222, February 2022)
- *Worldwide File- and Object-Based Storage Forecast, 2021-2025: New Enterprise Workloads Driving Strong Growth* (IDC #US48403021, December 2021)
- *Hyperconverged Infrastructure Adoption Trends — 3Q21: Building Block for Hybrid Cloud Infrastructure* (IDC #US48308121, October 2021)
- *Enterprise Workloads Resulting in Broader Adoption of Scale-Out File Storage Architectures* (IDC #US48305121, October 2021)
- *How to Compare Distributed Scale-Out File Storage Platforms for Use with Enterprise Workloads* (IDC #US48191621, September 2021)
- *Reclassifying File Storage — A New Approach for the Future of Digital Infrastructure* (IDC #US48051221, July 2021)

Synopsis

This IDC study represents a vendor assessment model called the IDC MarketScape. This study is a quantitative and qualitative assessment of the characteristics that assess a vendor's current and future success in the relevant market or market segment and provide a measure of the vendor's ability to become a leader or maintain leadership.

The distributed scale-out file system market segment, which is part of the file-based storage market, is an example of a large, maturing market that is still exhibiting low double-digit growth. This document assesses the capabilities and strategies of key vendors of scale-out file-based platforms. While seven of the assessed vendors have distributed scale-out file system designs, one of the vendors (NetApp) actually uses a scale-up cluster design but still meets the inclusion criteria of this vendor assessment study.

"While all evaluated vendors tout the performance, scalability, and ease of use of their file-based storage offerings, a closer look reveals important distinctions in how vendors define these metrics and build their products to achieve them," said Eric Burgener, research vice president, Infrastructure Systems, Platforms and Technologies Group, IDC. "To select the right product, enterprises need to understand the architectural differences between the different vendor approaches, understand the implications of those choices for their workloads, and then choose the product which best fits their requirements."

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

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com