G00276070

Market Guide for Open-Source Storage

Published: 23 November 2015

Analyst(s): Arun Chandrasekaran, Julia Palmer

The shift to bimodal IT in enterprises is making open-source storage relevant for several Mode 2 workloads. Infrastructure and operations leaders should use a decision matrix based on costs, risks and available internal skills to determine whether and where open-source storage should be implemented.

Key Findings

- High acquisition, maintenance and exit costs for proprietary storage platforms are affecting infrastructure and operations budgets and hindering the funding of new projects.
- The primary benefits of open-source software are freedom, innovation and flexibility, with costs as an ancillary reward.
- There are different business models associated with open-source storage, and confusion regarding their relative strengths and weaknesses continues to be a problem.
- Infrastructure and operations leaders are concerned that open-source storage is difficult to manage and that its long-term viability is uncertain.
- Open-source storage adoption requires a dedicated team of experts, with an agile DevOps approach.

Recommendations

- Use a decision matrix based on costs, risks and available internal skills to determine whether, when and where to deploy open-source storage.
- Identify workloads that can be run on commodity hardware with decoupled storage software; identify workloads that cannot be handled in a cost-effective manner by proprietary storage arrays.
- Procure vendor-provided maintenance and support services for open-source storage products that are deployed in production environments.
- Focus on hardware design and choose cost-effective reference architectures that have been certified by the vendors and for which support is delivered in an integrated manner.

- Ensure the success of open-source storage by aligning I&O closely with application owners and business units and by fostering community involvement.

Market Definition

Open-source storage refers to core storage software that is used to create a storage array in which the software is abstracted from the underlying hardware, and where the source code is made available to the public through a free license. Similar to proprietary storage, open-source storage software supports block (storage area network [SAN]), file (network-attached storage [NAS]) or object protocols. It can be deployed in scale-up or scale-out environments. Recent innovations in multicore processors and increasing CPU core density, combined with an innovative open-source ecosystem, make open-source storage attractive for an increasing set of workloads and as a viable alternative to proprietary storage.

Market Direction

Open-source storage has been in existence for several decades. This software, once used largely for "skunkworks" projects and specific use cases (e.g., file sharing), is now a growing contender in most cloud, big data and other Mode 2 implementations. Due to the industry standardization of storage hardware, this market has evolved to the point where customers can choose proven and compatible hardware components to build their own storage and choose their preferred open-source storage software to provide the higher-layer storage management and data services. Alternatively, for customers that prefer a more turnkey appliance, open-source vendors are closely working with hardware original equipment manufacturers (OEMs) and original design manufacturers (ODMs) to offer reference architectures and tightly integrated appliances.

Despite open source being the de facto approach among the large cloud providers (where the ethos is "We will try open source first, or we will build it ourselves"), enterprises are still cautious about the adoption of open-source storage. This is due to several perceived and real risks, including deployment complexity and lack of worldwide 24/7 support. Gartner believes that open-source storage software is now ready for more general deployment. Infrastructure and operations (I&O) leaders should use this research to help them understand the key players and projects in the open-source storage software market.

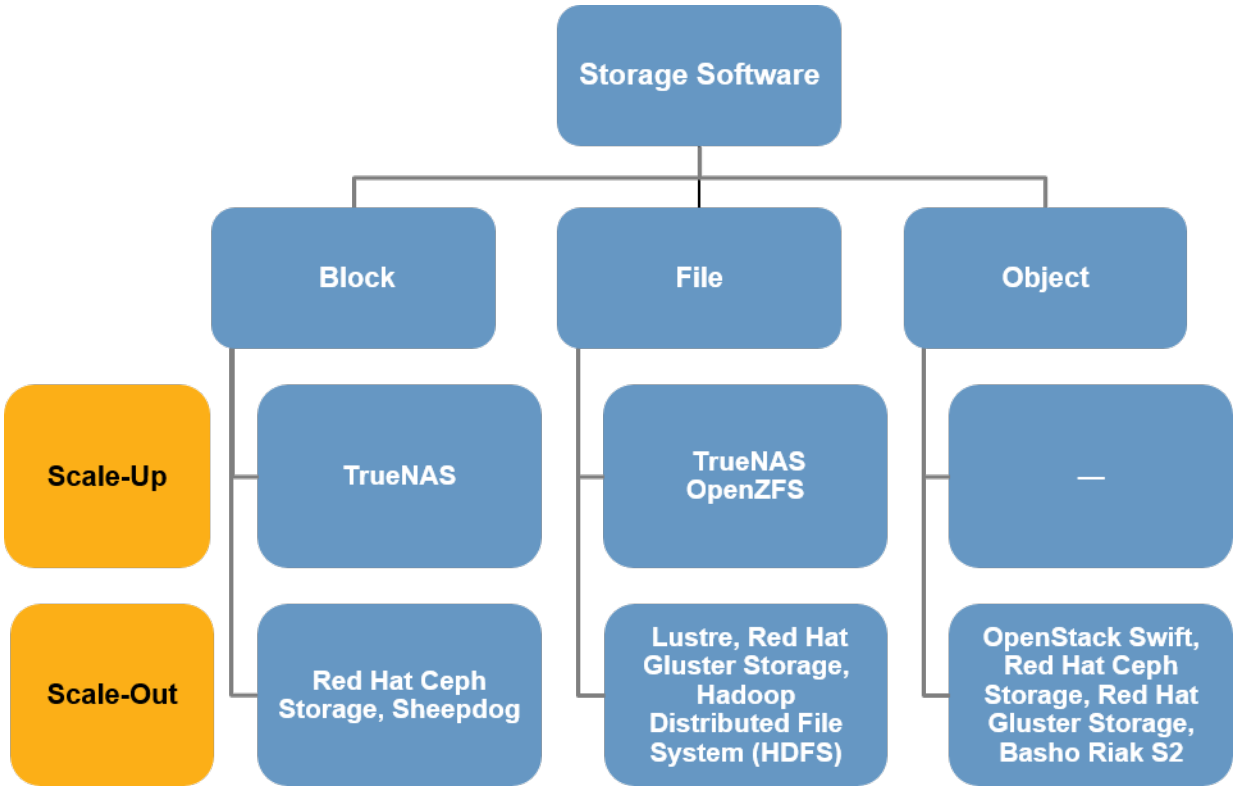
Most recently, I&O leaders have begun considering open-source storage software to support an even broader range of use cases, including private cloud infrastructure as a service (IaaS), big data analytics and online transaction processing (OLTP) environments. These use cases are made possible by the fact that leading open-source storage solutions cover the gamut of storage needs, including block, file and object-based access, as well as offering multiprotocol support (see Figure 1). This broader use — the inflection point at which open-source storage has shifted from niche to more-general applicability — also stems from several advances in the market:

- **Community Evolution —** Open-source infrastructure frameworks (e.g., Hadoop and OpenStack) are tightly packaged infrastructure software projects that are backed by the Apache Software Foundation, with a growing pool of active committers.
- **"Opencore" Models —** To ease adoption, several startups and large vendors are packaging open-source storage software with custom software, as well as orchestration and management tools (all of which are usually proprietary). This business model is contentious within the open-source communities, due to the overlay of proprietary code. Although this may cause proprietary, commercial software to be wrapped under the veneer of open-source APIs, the trade-off is that customers gain flexibility in evaluating options that are most suited to their environments.
- **OEM Support —** Vendors including Dell, Fujitsu, HPE, IBM, Quanta Computer, SanDisk and Super Micro Computer offer packaged, open-source software/hardware appliances, along with reference architectures. This enables I&O leaders to address a range of usage scenarios.

Market Analysis

As cloud computing, big data analytics and information archiving push the capacity, pricing and performance frontiers of traditional storage, there has been a renewed interest in open-source software solutions as a means to achieve high scalability in capacity and performance at lower acquisition costs. The rise of open-source platforms, such as Apache Hadoop and OpenStack, which are backed by a large, innovative community of developers and vendors, as well as the entry of disruptive vendors, including Red Hat (Ceph and Gluster) and Intel (Lustre), is enabling enterprises to seriously consider open-source storage. Several vendors are bringing products to market, based on the popular OpenZFS project. Figure 1 provides an overview of the common, open-source storage options.

Figure 1. Open-Source Storage Software Products/Projects



Source: Gartner (November 2015)

In addition to the representative projects listed in this research, the following vendors/projects may be a good fit for specific use cases:

- **Lustre —** Enterprises looking for a high-performance, distributed file system should put Lustre on their shortlists. Lustre is a proven parallel file system in supercomputing environments; however, it can be complex to manage, and it lacks file replication features. Lustre is backed by Intel, which offers management tools and different editions of the software, including an edition that runs on hyperscale clouds, such as Amazon Web Services (AWS) and Azure.
- **TrueNAS —** TrueNAS leverages the popular FreeNAS community project, which is based on OpenZFS. iXsystems has created a packaged unified storage array based on OpenZFS, and the vendor offers high-availability options and maintenance and support for the product. It is available as a hybrid and an all-flash array, supporting file and block protocols (e.g., AFP, iSCSI, NFS and SMB). The product supports all common hypervisors, including ESXi, Hyper-V, KVM and XenServer. iXsystems, which is headquartered in San Jose, California, offers different tiered support options for small businesses and enterprises.
- **Others —** Other open-source, parallel file systems include BeeGFS, which is popular in high-performance computing (HPC) academic environments in Western Europe (most notably, Germany), as well as XtreemFS, MogileFS and OrangeFS. Sheepdog, which is being incubated at NTT, is a distributed object and block store that is starting to see adoption in OpenStack

environments. Basho Technologies has open sourced its object storage product named Riak S2, which comes in various configurations targeting developers, enterprises and service providers. Nexenta uses a hybrid development model that combines open source with proprietary code. NexentaStor is Nexenta's flagship unified file and block storage solution. NexentaStor's community edition is based on open-source kernel, is limited to 18TB raw capacity and is available for home use, whereas the enterprise product includes additional closed and proprietary pieces. The community edition cannot be deployed for production use cases.

Representative Projects

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Table 1 analyzes the popular open-source storage projects.

Table 1. Comparison of Popular Open-Source Storage Projects

Open-Source Storage Project	Ceph	HDFS	Gluster (GlusterFS)	Swift
Storage Technology	Block and Object	File	File and Object	Object
Backing	Red Hat, SUSE	Hadoop distribution vendors, such as Cloudera, Hortonworks, IBM and Pivotal	Red Hat	SwiftStack and other OpenStack distribution vendors, such as Canonical, Cisco, HPE, IBM, Mirantis, Red Hat and SUSE
Licenses	GNU's Not Unix (GNU) Lesser General Public License (LGPL)	Apache License 2.0 (GNU General Public License [GPL] v.3 compatible)	GNU GPL v.3	Apache License 2.0 (GNU GPL v.3 compatible)
Community	Ceph	Hadoop	Gluster	Swift
Key OEM Partners	Dell, Fujitsu, Super Micro, SanDisk	Cisco, Dell, Fujitsu, HPE, Lenovo, Super Micro	Cisco, HPE, Lenovo, Super Micro	Cisco, Dell, HPE, Seagate, Super Micro
Popular Use Cases	Virtual machine (VM) storage, content distribution, archiving	Big data analytics, archiving	Archiving, backup, content distribution, home directories	Archiving, content distribution

Source: Gartner (November 2015)

Hadoop Distributed File System

A subproject in the Apache Hadoop project, Hadoop Distributed File System (HDFS) is an open-source storage project, developed by volunteers under the Apache Software Foundation. HDFS benefits from dedicated contributors from organizations such as Cloudera, Facebook, Hortonworks, IBM, LinkedIn and Yahoo. Several leading vendors, including Cisco, HPE, IBM, Microsoft, Oracle, Pivotal (EMC), SAP and Teradata, offer their own distributions of HDFS or have formed partnerships with key distribution vendors, such as Cloudera and Hortonworks, to bring joint solutions to market. MapR offers a forked-code, proprietary version of HDFS called MapR-FS.

Strengths:

- **Seamless handling of big data and built-in fault tolerance —** HDFS's architecture is built for big data analytics through the seamless handling of large files and distributed storage. HDFS relies on a shared-nothing architecture with direct-attached storage/just a bunch of disks (DAS/JBODs), yet it exhibits a high tolerance for hardware failure of data nodes due to software replication. With HDFS, hardware failure is treated as the rule, rather than the exception.
- **Hadoop 2.0 enhancements —** A key goal of the 2.0 release was to transform Hadoop from a batch processing system into a multiapplication system by overhauling MapReduce with a dedicated resource manager — Yet Another Resource Negotiator (YARN). Several other enhancements were made to HDFS: It now has namespace federation, which enables multiple Name Nodes (each addressing a single namespace) to act in a federated manner. These federated namespaces support workload isolation, enabling multitenant Hadoop clusters. Other significant enhancements have included support for NFS and snapshots for improved availability, as well as quotas and archival tiering.

Challenges:

- **HDFS's architectural limitations —** HDFS is not a general-purpose file system, and it has primarily been architected to support big data batch processing and long-term, low-cost data retention. At this point, HDFS is not mature enough to be used for non-Hadoop workloads, due to its append-only nature, although there is rising interest in using HDFS for such non-Hadoop use cases as file sharing and backup. Also, with the rise of Apache Spark (see Note 1), and given that Spark doesn't always need HDFS underneath, the alternatives to HDFS may increase with the growing adoption of Spark.
- **Backup and disaster recovery —** Backup and disaster recovery continue to be an Achilles heel for large Hadoop clusters, due to a lack of independent software vendor (ISV) support and the absence of remote replication capabilities. HDFS's snapshot capabilities can capture only files that are closed when the snapshot is taken. Third-party providers, such as WANdisco, have capabilities to bridge this gap.

Community: Hadoop

Key OEM Partners: Cisco, Dell, Fujitsu, HPE, Lenovo and Super Micro

Popular Use Cases: Big data analytics and archiving

OpenStack Swift

Started in 2010, OpenStack is an open-source project available under the Apache License 2.0. The OpenStack Foundation promotes the development and distribution of various OpenStack projects, including Swift, which is a distributed object storage software. Swift is an important foundation of OpenStack, has been available since the first release and is closely integrated with other OpenStack projects. Recent enhancements to Swift include the introduction of erasure codes and bulk upload capabilities. Swift can run on standard x86 servers and uses a standard API to interact with applications and other storage services. For customers looking for an easier-to-deploy product and vendor support (in addition to community support), SwiftStack would be a viable option.

SwiftStack provides a custom dashboard, runtime functions, such as authentication integration; load balancing, monitoring and reporting tools; maintenance; and SLA-backed support. Customers that are looking for a unified storage product can consider alternatives, such as Ceph, that are compatible with the Swift API and offer more than object storage.

Strengths:

- **Massive scalability —** Swift can easily scale to multi-PB environments through its ring architecture, which maps object storage data to physical devices.
- **Popularity of API —** Several cloud providers provide public cloud storage services based on the Swift project, enabling easier hybrid cloud environments. In addition, the Swift API is rapidly gaining support among ISVs (second only to the Amazon S3 API), making it easier for customers to integrate with their on-premises applications and infrastructure.

Challenges:

- **Performance limitations —** End-user organizations need to work closely with the OEMs to optimize and tune the hardware to extract the most optimal performance from OpenStack Swift. Moreover, Swift's throughput performance increases as the object size increases, which means a great deal of tuning is required in small-file/object environments.
- **Weak security and compliance features —** Features such as native encryption and write once, read many (WORM), which are important for the archiving use case in regulated industries, are still missing in Swift.

Community: Swift

Key OEM Partners: Cisco, Dell, HPE, Super Micro

Popular Use Cases: Archiving and content distribution

Ceph

Ceph is an open-source, scale-out unified-storage project. In April 2014, Red Hat acquired Inktank to complement its Gluster acquisition and to bolster its OpenStack portfolio. OpenStack

deployments are clearly the sweet spot for Ceph, because Ceph offers tight integration with the various OpenStack projects. The Red Hat Ceph Storage v.1.3 software, released earlier this year, delivered better stability and scalability features, as well as performance improvement tweaks. With this release, Ceph has improved software quality from code quality assurance to documentation, as well as improved operational efficiency by adding features such as multicluster management. SUSE also offers maintenance and support for Ceph as a silo storage product and as a part of its OpenStack distribution (SUSE Cloud).

Strengths:

- **Active community —** Ceph has a thriving community of hundreds of individual contributors, as well as strategic corporate contributors, such as SanDisk, Intel, CERN, Mellanox Technologies and Yahoo (Apart from Red Hat). Ceph has a rapid software release cadence, and its contributors produce more than 1,000 commits per month. The Ceph community has recently established an advisory board with external members from leading ISVs and hardware OEMs to assist the community in enhancing its strategic roadmap.
- **Scalability and ecosystem support —** Ceph is a highly scalable, distributed unified storage project with several multi-PB deployments. The product offers integration with other open-source community initiatives, such as OpenStack, Kernel-based Virtual Machine (KVM) and CloudStack. Ceph has tight integration with OpenStack block storage (via Cinder and Glance), while providing support for OpenStack Swift and Amazon Simple Storage Service (S3) APIs.

Challenges:

- **Immature file system —** Although Ceph has adequate block and object storage capabilities, the distributed Posix file system, CephFS, is not ready for production. Ceph lacks support for Windows clients, VMware ESXi and Microsoft Hyper-V hypervisors.
- **Need for detailed instrumentation —** Large Ceph deployments require a major effort from enterprise users, requiring careful instrumentation to optimize and fine-tune cluster performance and reliability for hardware and the software stack.

Community: Ceph

Key OEM Partners: Cisco, Dell, Fujitsu, Super Micro, SanDisk

Popular Use Cases: VM storage, content distribution and archiving

Gluster

Red Hat acquired Gluster in 4Q11, and the product is now called Red Hat Gluster Storage. It is a scale-out, multiprotocol (NFS, RESTful APIs, Server Message Block [SMB]), open-source storage software solution, with petabyte-scale capacity and improved snapshot and replication capabilities. Following Red Hat's acquisition of Inktank, the Red Hat Gluster Storage product now primarily targets file-protocol-based use cases. Version 3.1 of Red Hat Gluster Storage released in June 2015 delivered erasure coding, silent data corruption detection, automated tiering with support for flash, and easier snapshot scheduling. Red Hat Gluster Storage now has support for NFSv4, SMB3 and

SSL-based network encryption. However, Red Hat needs to address the growing product overlap between Ceph and Gluster and better articulate Gluster's unique value proposition and use cases.

Strengths:

- **Deployment flexibility —** Red Hat Gluster Storage is a preintegrated software product consisting of Red Hat Enterprise Linux (RHEL), GlusterFS and the extensible file system (XFS). Red Hat Gluster Storage can be run on most industry-standard x86 servers with Ethernet or InfiniBand support and offers NFS, SMB and HDFS interfaces for enterprise use. For on-premises deployments, Red Hat Gluster Storage is installed on bare-metal hardware, or it can be installed in KVM or VMware hypervisors to pool storage resources. For public cloud environments, Red Hat Gluster Storage is supported in AWS and Microsoft Azure, where it aggregates native block storage to offer Posix compatible scale-out file services.
- **Vendor support and partnerships —** Red Hat Gluster Storage benefits from Red Hat's complementary open-source community projects, technical support capabilities (community, standard and premium support options) and close relationships with ISVs, such as Commvault, Hortonworks, ownCloud and Splunk.

Challenges:

- **Feature gaps —** GlusterFS lacks some capabilities that enterprise IT buyers want from a file system, such as native data reduction features.
- **Lack of prepacked solutions —** Reference architectures for Red Hat Gluster Storage are not easily available, when compared with the other products profiled in this research.

Community: Gluster

Key OEM Partners: Cisco, Dell, HPE, Lenovo and Super Micro

Popular Use Cases: Archiving, backup, content distribution and large home directories

Market Recommendations

Although open-source storage offers considerable benefits, enterprises must accurately gauge the benefits, costs and risks associated with it, and they must develop a clear long-term plan. With the emerging maturity of open-source storage solutions, enterprise IT buyers should not overlook their value proposition. Choose the right use case for open-source storage, based on its maturity of the project, depth of community and/or support, and the criticality of the environment in which it is deployed.

Gartner Recommended Reading

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

"IT Leaders Can Benefit From Disruptive Innovation in the Storage Industry"

"The Issues and Costs of Building Your Own Storage, Compared to Buying Storage or Using Cloud Storage"

"Hype Cycle for Storage Technologies, 2015"

"Should I Use Open Source in My Infrastructure?"

Evidence

This research is based on end-user inquiries handled by the author in 2014 and 2015. Gartner data center surveys in 2014 reveal that 20% of enterprise customers had either deployed or were actively investigating open-source storage. A detailed briefing was conducted with all major vendors profiled in this research.

Note 1 Apache Spark

Apache Spark is a large-scale, in-memory data-processing engine.

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