TECHNOLOGY ASSESSMENT

Scale-Out File Systems on Object-Based Storage Platforms

Ashish Nadkarni

IN THIS EXCERPT

The content for this excerpt was taken directly from Scale-Out File Systems on Object-Based Storage Platforms (Doc# 258393). All or parts of the following sections are included in this excerpt: IDC Opinion, Situation Overview, Future Outlook, and Essential Guidance sections that relate specifically to Scality, and any figures and or tables relevant to Scality.

IDC OPINION

IDC estimates that by 2017, 79% of the storage capacity shipped into enterprises will be for storing unstructured data (see *Structured Versus Unstructured Data: The Balance of Power Continues to Shift,* IDC #247106, March 2014). Translating to roughly 106EB, this storage capacity directly correlates to a need to store and analyze an ever-growing set of mixed data types and has compelled suppliers to come up with innovative highly scalable solutions – one of which is object-based storage (OBS). Object-based platforms like the ones from Scality, Cleversafe, Exablox, DDN, EMC, NetApp, and others (to name a few) employ a flat tenant-account-container-object-based layout, which IDC considers to be one of the three key data organization models for any storage platform. Many such object-based platforms also employ features such as programmatic and customizable metadata access, policy management that can make use of various storage tiers, geodispersed data resiliency, and in-place analytics. However, a defining characteristic of such platforms is the ability to provide data access via RESTful APIs such as Amazon S3, OpenStack Swift, and CDMI – none of which are POSIX compliant. The pervasiveness of known file interfaces such as NFS and SMB, and also the flexibility of these platforms for use in various cloud and noncloud environments, has compelled some suppliers to create integrated file system overlays that provide the "best of both worlds." Suppliers like Scality, Exablox, and Coho Data lead when it comes to employing a distributed (scale-out) file system that integrates with an independent but underlying object-based platform. The benefits of this approach are:

- The distributed file system offers access via known POSIX-compliant file interfaces such as NFS and SMB. Overhead is reduced because the file system semantics are stored directly in the object store as metadata and the file system treats the object storage platform as a persistent data tier, in which files are stored as sharded objects.
- The distributed file system is scalable in tandem with the object store, removing the potential for a "sand clock like" constriction between the two. This is often the problem with cloud-enabled storage platforms (aka "cloud gateways") that treat the object store as an external disk tier but choose to keep the file system schematics elsewhere – partly because they employ unitary file systems like OpenZFS.

  Since the object storage platform is independent of the distributed file system, the supplier can choose to provide direct access to the platform via RESTful APIs – which then allow next-gen applications direct access to the storage system via APIs – and benefit from features like

policy management and programmatic metadata access. When used with a cloud orchestration stack like OpenStack, such platforms can truly become a unified platform that offers file, block, and object access via a single object-based data organization scheme. When used with industry-standard server-based hardware, this approach can truly deliver cloud-scale economics of a software-defined infrastructure. In fact, Ceph — an open source object platform — already delivers this benefit today. (Note: the Ceph file system is in early release phase and was used only in preproduction deployments at the time this document was written.)

IN THIS STUDY

This study discusses an approach that suppliers like Scality and Exablox are taking to delivering a "scale-out file system (SOFS) on object-based platforms" — an approach that seeks to deliver unified block, file and, of course, object access via a single object-based platform.

SITUATION OVERVIEW

Typical NAS solutions pose limitations because of their architecture. Such solutions employ unitary POSIX file systems as the persistence layer (upon which an additional distributed file system is built if the solution is indeed a scale-out solution). File systems pose limitations when it comes to persistence — in which files, directories, and file system metadata are stored in hierarchical structures:

- Access performance is dependent on the number of files, directories, and tree structures. The higher the number of files in a file system, the greater the risk of performance degradation.
- Maintaining consistency of file systems poses an overhead on the overall system and, often, issues such as fragmentation, journaling, and clients simultaneously operating on the same file system structures.
- The storage subsystem may pose additional limitations because of RAID, disk sizes, and other limitations posed by hardware/software.

IDC has long maintained that an object-based data organization scheme that uses a flat tenant-account-container-object-based layout — which along with block and file is one of the three principal data organization schemes — is more versatile than what it gets credit for. Instead of complex file system hierarchies, object storage systems provide simple key/value stores. This therefore provides a flat and virtually unbounded namespace based on key values. This approach seeks to solve the scaling limitations inherent in classical POSIX file systems. What started out as capacity-optimized storage for unstructured data is quickly gaining popularity in virtualized environments as a repository for virtual machines and container images. The use of flash in such platforms allows suppliers to provide intelligent performance optimization for frequently accessed data sets. IDC expects innovation in this space to accelerate as more and more suppliers move to an object-based storage data organization scheme.

A ding against object-based platforms, however, is that they lack POSIX-compliant file interfaces, making them incompatible in environments in which the application expects a POSIX file interface, like NFS and SMB. A Band-Aid solution that many employ is to use a cloud-enabled storage platform (aka "cloud gateway") that offers such an interface. However, the issue there is that the gateway itself is built on a unitary file system with all of the limitations discussed previously and only uses loosely coupled clustering that is external to the object storage itself (and therefore does not leverage any

clustering or scale-out technologies inherent to the object storage architecture). Furthermore, since the cloud gateway relies on RESTful APIs exposed by the object storage, the integration (and issues resulting from such integration) is limited to the capabilities of the API. For example, Amazon Web Services (AWS) S3 may only provide basic capabilities, but Swift or CDMI may provide more comprehensive capabilities. Nevertheless, the integration can run into issues. In fact, in a research survey conducted by IDC on Amazon Web Services usage trends, 29.6% of respondents said that issues with AWS S3 had to do with cloud gateway interoperability (see *AWS Storage Usage Trends,* IDC #255387, April 2015).

Elements of a "SOFS on OBS" Solution

Suppliers like Exablox, Scality, and Coho Data have taken a new approach. They have integrated a distributed file system into a key/value-based object store. This approach utilizes a fundamentally different persistence mechanism for storing data but provides both POSIX-compliant file system interfaces and, optionally, RESTful APIs. At first, this radically different approach sounds like too much complexity to address the problem of scaling. However, given the promise of providing a unified, hyperscalable, and software-defined storage solution for a variety of current generation applications that require POSIX interfaces and next-generation applications that are designed to work with APIs, this approach begins to make sense.

In this approach, there is direct integration between the scale-out file system and the object storage solution. While each supplier's architecture varies, such approaches have common elements:

- A "virtual" file system that presents a POSIX-compliant file system interface such as NFS and SMB (In essence, the solution provides a global [multiprotocol] namespace that is fully compatible with NFS v3/v4 and SMB 1.0/2.0/3.0 clients. In some cases, user space clients like FUSE are also supported.)
- An object store that provides a distributed, scale-out, and shared-nothing data persistence layer and a metadata database that is used to store the file system semantics such as directories, inodes, and hard and soft links (Some suppliers even make the metadata database programmatically accessible to provide an additional layer of intelligence.)
- Appropriate clustering, data distribution, and durability technologies that are different from typical RAID-based solutions (These include the use of techniques such as object replicas, erasure coding for data distribution, and clustering software that presents a tightly consistent or eventually consistent state at all times.)
- Technologies such as compression, deduplication, and encryption to ensure that the economics of an object store are maintained in spite of a layered approach
- Technologies such as snapshots and clones that allow the solution to be used in enterprise environments where data copies are essential for development/test/QA, analytics, and data protection
- From a manageability perspective, the solution be managed from "top down" as a NAS solution or in an à la carte manner as an object store – this depends on the use case and/or the environment in which it is deployed

Key benefits of this approach are:

- This solution immediately overcomes the typical limitations that are posed by a NAS solution. For example, file size limitations, limitations posed on the number of files, and limitations on the number of directory hierarchies.

- As a "unified" solution, this approach provides the flexibility required by organizations to scale performance and capacity independent of each other and in line with the "hyperscale" capabilities of the underlying object store. Certain technologies are compute intensive in nature, while others are more of a capacity hog. In either situation, the scaling is linear in nature.
- Given the software-defined nature of this solution, organizations can rely on industry-standard server hardware to grow the solution in an as-needed fashion. This allows organizations to move to a capex-friendly approach for storage.

Suppliers may additionally choose to make their solution discrete (in which case the storage system only runs storage workloads) or converged (in which case the storage system can run nonstorage workloads natively or via virtualized units like containers).

Supplier-Specific Implementations

Even though there is much commonality in the approach to delivering SOFS on OBS, there are several implementation-specific differences between solutions. For starters, the design of the object store itself differs from supplier to supplier. Since much of the characteristics of the overall solution, including scalability and resiliency, depend on the object store, this is a crucial distinction. The second implementation-specific difference is the choice of the supplier to allow direct access to the object store. In the case of making this direct access available, the supplier is simply trying to address a bigger market. Other suppliers may take a more reserved approach to avoid taking on the risk of spreading their resources thin.

Scality — SOFS on RING

Scality's Scale-Out File System is a virtual file system that presents POSIX semantics and is integrated directly into the Scality RING object store. Scality RING natively provides a distributed, scale-out, and shared-nothing database that is merged directly into the underlying object store. This database is referred to as MESA (Spanish for "table"). The database is part of the object store itself — with the database structures represented as distributed objects in the RING. The MESA database is used to represent file system structures used to represent a traditional file system. MESA itself is distributed across all of the RING's storage nodes, which make up the underlying scalable object store. As in the RING's object storage layer, an efficient peer-to-peer routing algorithm (known as CHORD) is used to distribute the MESA structures around the RING, in a shared-nothing manner.

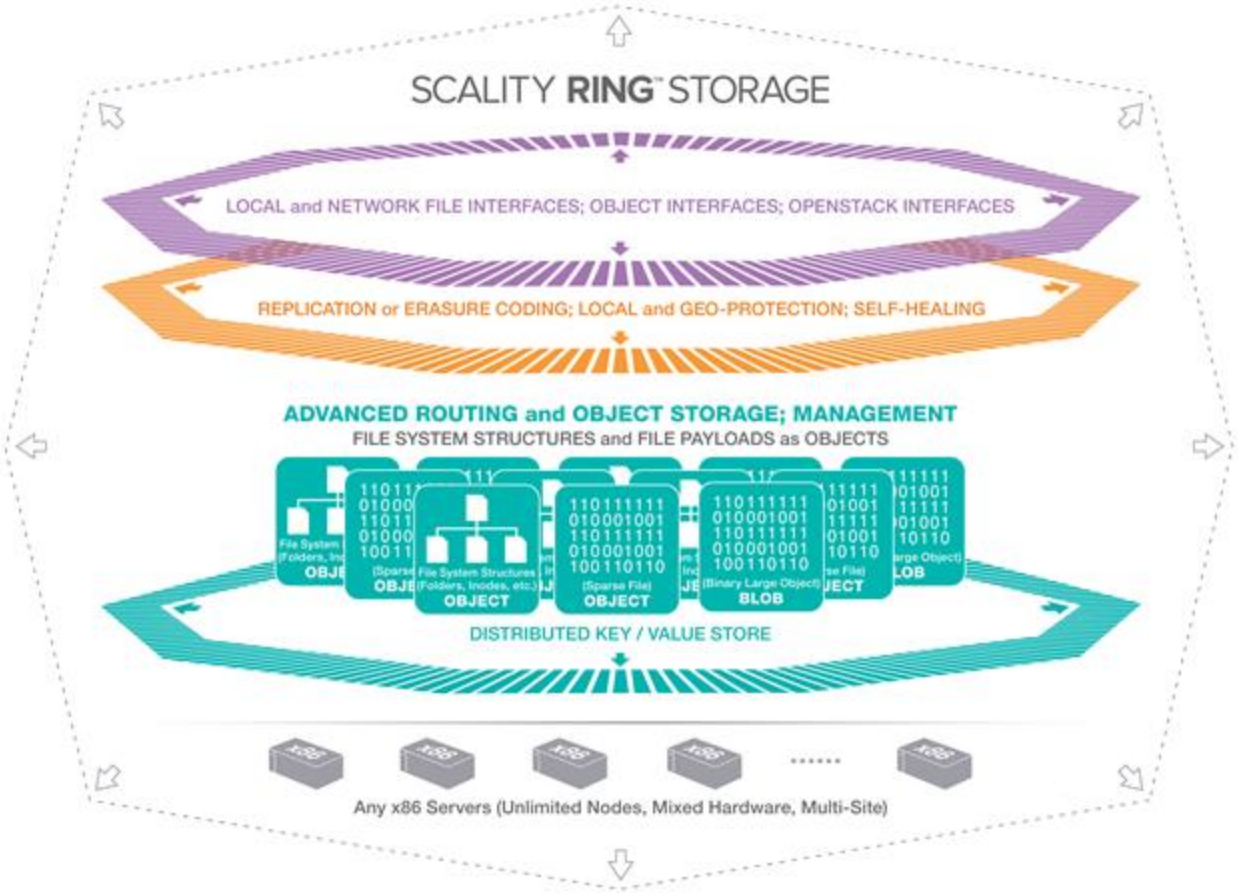
As with objects in the RING's object store, SOFS files have no size limitations, with multiterabyte files being stored in current production systems (and multipetabyte files tested). The RING can support up to 2^32 file systems, with support for billions of files per file system, without any need to preconfigure for capacity through thin provisioning of file system "volumes." File systems will utilize the RING's storage pool to expand as needed when files are created and updated.

SOFS works directly with the data protection and durability mechanisms present in the RING, including variable replication class of service (1–6 copies) and configurable, space-efficient erasure coding to provide protection against multiple, simultaneous component (disk, server) failures. Variable classes of service are also utilized to optimize the durability of critical structures such as the root file system chunk, directory chunks, and databases pages. In practice, all MESA database objects are stored with multiple copies for durability on a low-latency SSD-backed RING. The stripes, or actual data representing the files, are stored on a RING backed with lower-cost disks. To conserve space and increase durability, these stripes are stored using an erasure coded schema that fits the architecture and overhead requirements. The ratio of SSD-to-lower-cost HDD disks is typically 1:20.

SOFS file systems can be scaled out in capacity across as many storage nodes as needed to support application requirements and can be accessed by a scalable number of NFS, FUSE, SMB, or CDMI connectors to support application load requirements. Note that CDMI provides a unique capability of accessing SOFS data via standard RESTful protocols, thereby enabling namespace sharing between POSIX-based file system data and RESTful applications. Scality also builds its OpenStack persistent storage capabilities by leveraging SOFS for OpenStack Cinder data volumes and its planned OpenStack Manila file service support (in addition to existing OpenStack Swift and Glance capabilities). This enables a true "unified" file and object storage solution for OpenStack deployments (see Figure 1).

FIGURE 1

Scality — SOFS on RING Architecture



Source: Scality, 2015

FUTURE OUTLOOK

As the IT industry transforms itself from what IDC considers to be 2nd Platform applications to 3rd Platform applications, there will be a need for storage solutions that support both 2nd Platform access interfaces (file interfaces such as NFS and SMB) and 3rd Platform access interfaces (RESTful APIs). Since many of these two generational applications share the same data sets, it is imperative that suppliers look at unified storage solutions that offer the best of both worlds. However, suppliers can under the covers start to transform the storage architecture that underpins the data management paradigm for next-generation IT. Object stores that leverage a software-defined shared-nothing architecture, with a distributed programmatically accessible metadata database, are one such storage technology. By layering a virtual scale-out file system on top of a fundamentally different data organization and persistence layer (the object store), suppliers can provide a solid bridge for this 2nd to 3rd Platform application transformation — which IDC expects to be a multiyear journey lasting well into the next decade.

ESSENTIAL GUIDANCE

The sections that follow are adapted from *IDC MarketScape: Worldwide Object-Based Storage 2014 Vendor Assessment* (IDC #253055, December 2014).

Guidance for Buyers

All companies, small and large, grapple with data growth. As businesses become data driven to survive in the new economy, they will seek more data sources, collect more data, and look to analyze and store this data in a decentralized manner. In many cases, they will look to perform real-time analytics on this data as it is generated and where it gets captured. Many others will seek to create on-demand opex-driven cloud environments for internal and external consumption. Nontraditional use cases, especially for highly scalable and decentralized semistructured (machine generated) and unstructured data storage, will require storage platforms that provide POSIX and RESTful data access interfaces are set up for extreme scalability but allow this scaling to occur in a capex-friendly fashion. With a "SOFS on OBS" approach, suppliers are making their solutions friendly toward today's workloads but also support next-generation workloads.

Buyers should therefore look for the following key characteristics when evaluating OBS solutions:

- **Platform scalability:** Scalability is not just from a hardware perspective but also from throughput, I/O, file size, and file volume perspectives. A solution appropriate for a given environment will allow each dimension to scale independently. For SOFS on OBS solutions, the file system should scale linearly with the object store.
- **Data management:** Data layout and organization is an important piece as it may have performance, efficiency, and availability implications. Over time, as data grows, organizations will face the need to mine existing data for patterns that may build new business cases around new findings. A solution that supports advanced metadata, indexing, and analytics will be a key component of the infrastructure.
- **Storage efficiency:** The larger the data set and bigger the storage system, the greater the need of data management and reduction techniques (data deduplication, compression, thin provisioning, etc.). Data optimization technologies (automated data tiering) will also be essential. A solution appropriate for a given environment will allow many, if not all, of the previously mentioned features to be implemented and recalibrated without major disruptions.

- **Data resiliency:** Resiliency capabilities (like replication and erasure coding) and the granularity with which such capabilities can be applied (i.e., whether policies can be applied at an account, container, or object level) will be important considerations. Data resiliency should also be weighed against the platforms' CAP theorem profile.
- **Workload adjacency:** Several OBS platforms offer or are considering offering nonstorage workloads to run natively or via containers on the OBS platform. Since most OBS platforms are node based and use x86 platforms, they offer excellent workload adjacency for distributed and localizable workloads like Map/Reduce and hypervisors. This is an important consideration for decentralized storage and in situations where the data has a short shelf life.

In addition to the platform characteristics, buyers should look for the following supplier attributes:

- **Supplier's commitment to the platform now and for the future:** Strong road map, customer support and service, and overall track record on incorporating new features into the platform are some of the attributes buyers should look for in a supplier.
- **Partner ecosystem for applications and on-ramping:** The more comprehensive the ecosystem, the better placed the supplier in offering an end-to-end workload-optimized or use case-focused solution.

LEARN MORE

Related Research

- *Worldwide Storage and Device Management Software Market Shares, 2014: Opportunities from IT Industry Transitions* (IDC #257407, August 2015)
- *EMC Releases SPC Benchmark Data for VNX and VMAX3 Storage Arrays* (IDC #IcUS25821215, August 2015)
- *Qumulo: Start-Up Offers "Data Aware" Scale-Out Storage* (IDC #257047, July 2015)
- *IDC's Worldwide Storage Software Taxonomy, 2015* (IDC #256418, June 2015)
- *Evaluating Scale-Up and Scale-Out Architectural Differences for Primary Storage Environments* (IDC #256932, June 2015)
- *Portworx Launches Container-Aware Storage – The Era of Storage for Containerized Environments Is Here* (IDC #IcUS25710715, June 2015)
- *IDC's Worldwide Flash Storage Solutions in the Datacenter Taxonomy, 2015* (IDC #255995, May 2015)
- *Telco Giant Telefónica Deploys Caringo for Cloud Storage* (IDC #256226, May 2015)
- *Worldwide Enterprise Storage Systems Forecast, 2015-2019* (IDC #256302, May 2015)
- *Americas Storage Services Forecast, 2015-2019* (IDC #256324, May 2015)
- *Software-Defined Storage-Controller Software* (IDC #256455, May 2015)
- *Virtual SAN 6.0 Targets Primary Storage Workloads with All-Flash Configurations and Enterprise Data Services* (IDC #255092, April 2015)
- *IDC's Worldwide Storage for Virtualized Environments Taxonomy, 2015* (IDC #255270, April 2015)
- *AWS Storage Usage Trends* (IDC #255387, April 2015)
- *Worldwide Enterprise Storage Systems 2014-2018 Forecast Update* (IDC #254378, March 2015)

- *HGST Acquires Amplidata – Enters the Object Storage Market* (IDC #255082, March 2015)
- *IDC's Worldwide Storage for Big Data and Business Analytics Taxonomy, 2015* (IDC #254025, February 2015)
- *Nexenta's Success Mirrors That of the Burgeoning Software-Defined Storage Platform Market* (IDC #254077, February 2015)
- *IDC's Worldwide File- and Object-Based Storage Taxonomy, 2015* (IDC #254078, February 2015)
- *IBM Spectrum Storage: Software-Defined Transformation for IBM Storage* (IDC #254453, February 2015)
- *Worldwide Storage Services 2014-2018 Forecast: Regional Economic Bright Spots, Automation, and Cloud Transformations to Dominate* (IDC #252596, December 2014)
- *Scality Significantly Increases Market Opportunities by Adding a Commercial Market Focus* (IDC #252865, December 2014)
- *Sheepdog – The Next Ceph?* (IDC #252976, December 2014)
- *OpenStack in 2014: A Storage Deep Dive* (IDC #252996, December 2014)
- *IDC MarketScape: Worldwide Object-Based Storage 2014 Vendor Assessment* (IDC #253055, December 2014)
- *Market Analysis Perspective: Worldwide and U.S. Storage Industry, 2014* (IDC #253171, December 2014)
- *Market Analysis Perspective: Worldwide Storage and Data Management Services, 2014* (IDC #253230, December 2014)
- *Vendor and Sourcing Management: Aligning Sourcing and Location Strategy with Business Objectives* (IDC #253352, December 2014)
- *Worldwide Storage for Public and Private Cloud 2014-2018 Forecast* (IDC #252135, November 2014)
- *Cyber Group Deploys EMC ViPR for Next-Generation SaaS Application Infrastructure* (IDC #252419, November 2014)
- *U.S. SMB Storage Update: Shifting Technology Preferences as Cloud Storage Increases* (IDC #252544, November 2014)
- *Worldwide File- and Object-Based Storage 2014-2018 Forecast* (IDC #251626, October 2014)
- *Taxonomy Distinctions Between Storage Accounting in IDC's Enterprise Storage Systems and High-Performance Computing Research* (IDC #251935, October 2014)
- *IDC's Worldwide Flash Storage Solutions in the Datacenter Taxonomy, 2014* (IDC #250560, September 2014)
- *Worldwide Storage and Virtualized x86 Environments 2014-2018 Forecast* (IDC #250720, September 2014)
- *Worldwide Enterprise Storage Systems 2013 Vendor Shares: Adoption of Software-Based Storage and Cloud Continues* (IDC #251477, September 2014)
- *Worldwide Storage Market Overview, 2Q14* (IDC #251591, September 2014)
- *Worldwide Storage Software Market Update, 2Q14* (IDC #251642, September 2014)
- *Leading European Broadcaster Deploys MatrixStore from Object Matrix as Nearline Storage for Postproduction Video Content* (IDC #249808, August 2014)

- *Worldwide Storage and Device Management Software 2014-2018 Forecast and 2013 Vendor Shares: Future Impact from Software-Defined Storage* (IDC #250437, August 2014)
- *U.S. 2014 SMB Business Objectives, IT Spending Priorities, and Technology Attitudes: Increasing Focus on Performance Driving Cloud and Mobility Interest* (IDC #250441, August 2014)
- *Flash-Optimized Storage Architectures: Transforming Enterprise Storage* (IDC #WC20140821, August 2014)
- *IDC's Worldwide Software-Defined Storage Taxonomy, 2014* (IDC #247700, July 2014)
- *Systems, Platforms, and Enabling Software – A Comparison* (IDC #249730, July 2014)
- *Ethernet-Connected Drives: A New Era in Capacity-Optimized Storage Deployments?* (IDC #249778, July 2014)
- *IDC's Worldwide Storage Software Taxonomy, 2014* (IDC #249822, July 2014)
- *Enterprise Data Lake Platforms: Deep Storage for Big Data and Analytics* (IDC #250000, July 2014)
- *EMC Acquires TwinStrata – Plans to Offer Embedded Hybrid Cloud Functionality for Enterprise Storage Customers* (IDC #IcUS24980614, July 2014)
- *IBM Edge2014 – Turning Point for IBM's Storage Strategy* (IDC #249053, June 2014)
- *Worldwide Enterprise Storage Systems 2014-2018 Forecast: Alignment with the 3rd Platform Is the Next Must-Have* (IDC #248554, May 2014)
- *Structured Versus Unstructured Data: The Balance of Power Continues to Shift* (IDC #247106, March 2014)
- *Coho Data Announces DataStream – Web-Scale Storage Comes to Enterprise IT* (IDC #IcUS24727414, March 2014)
- *Worldwide Disk Storage Systems Market Update, 3Q13* (IDC #246584, February 2014)
- *IDC's Worldwide Cold Storage Ecosystem Taxonomy, 2014* (IDC #246732, February 2014)
- *Influence of Open Source in Storage Systems and Platforms – An Order of Magnitude Analysis* (IDC #246804, February 2014)
- *IDC's Worldwide Storage and the Cloud Taxonomy, 2014* (IDC #245595, January 2014)
- *IDC's Worldwide Storage and Big Data Taxonomy, 2014* (IDC #245938, January 2014)
- *Enterprise Architecture: Making Portfolio Management More Objective and Effective* (IDC #243946, October 2013)
- *IDC MarketScape: Worldwide Object-Based Storage 2013 Vendor Assessment* (IDC #244081, October 2013)
- *Flow: Design Objectives for Next-Generation Applications* (IDC #241643, June 2013)
- *IDC MarketScape: Worldwide Scale-Out File-Based Storage 2012 Vendor Analysis* (IDC #238923, December 2012)

Synopsis

This IDC study discusses an approach taken by some suppliers of scale-out file system on key/value object stores – an approach that seeks to deliver unified block, file and, of course, object access via a single object-based platform. As the IT industry transforms itself from what IDC considers to be 2nd Platform applications to 3rd Platform applications, there will be a need for storage solutions that support both 2nd Platform access interfaces (file interfaces such as NFS and SMB) and 3rd Platform access interfaces (RESTful APIs).

"Businesses can no longer afford to maintain separate storage infrastructure for their legacy (2nd Platform) and next-generation (3rd Platform) applications. The SOFS on OBS approach delivers the best of both worlds – it allows businesses to create a future-proof infrastructure that will help them seamlessly embrace innovation accelerators such as IoT, cognitive computing, and robotics built on the four pillars of mobile, social, cloud, and Big Data," said Ashish Nadkarni, research director at IDC.

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-insights-community.com
www.idc.com