Solution Showcase

IBM: Parallel File System Performance for Analytics

Date: May 2016 **Author:** Scott Sinclair, Senior Analyst

Abstract: The expectations that businesses have for their unstructured digital content are escalating. Big data and business intelligence applications unlock value previously inaccessible from unstructured content, enabling businesses to reach new levels of competitiveness and profitability. Access to real-time insights, however, requires increased performance as cold data archives see greater activity. Unstructured storage systems can no longer simply focus on cost-effective capacity scaling. IBM, a leader in enterprise IT solutions, has optimized its Elastic Storage Server product line to combine the cost-effective deployment flexibility of software-defined storage (SDS) with the performance scaling of a parallel file system.

Overview

The idea that data has value is not new. Businesses have been leveraging the digital information contained within IT storage systems for years. The extent to which businesses are able to unlock that value, however, is increasing. Big data solutions, such as Hadoop, are architected to better access the insights previously hidden in unstructured digital content. In a recent ESG research study, business intelligence and data analytics was the second most-cited response among initiatives identified by organizations as the most important IT priorities of 2016.[1]

As businesses seek to better enable timely decision making, IT organizations look to real-time analytics. In a separate ESG research study focused on big data, IT organizations were asked what requirements were driving them to evaluate new business intelligence/analytics solutions. The most commonly identified response was that the organization was moving toward more real-time analytics.[2] The demand for real-time insights has introduced a renewed emphasis on performance for unstructured data storage solutions, in addition to, rather than in spite of, the standing requirement of cost-effective capacity scaling. In response, high-performance-oriented parallel file system architectures, often associated with high-performance compute solutions, have found an increased interest from the enterprise.

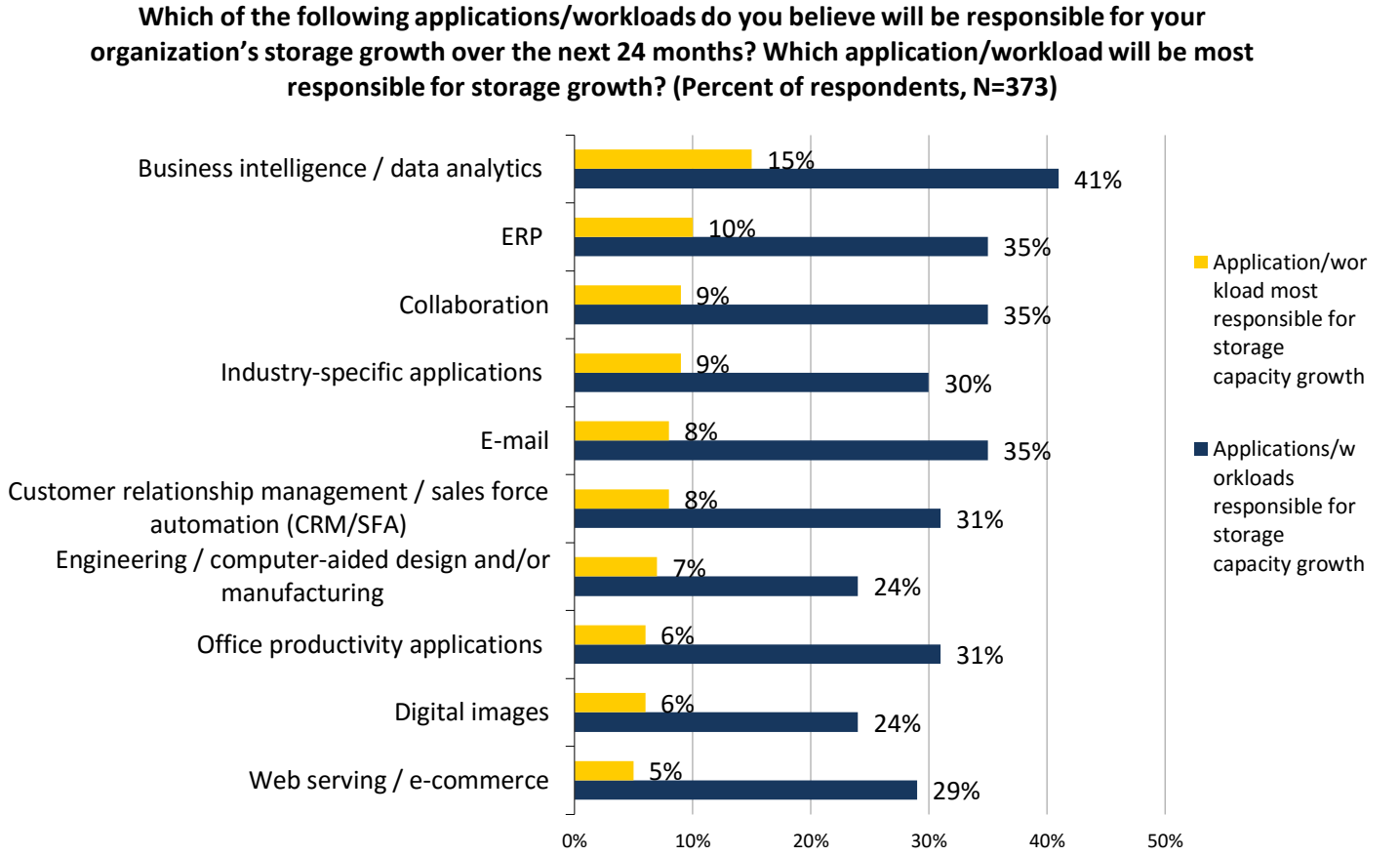
IBM's Elastic Storage Server (ESS) leverages the company's Spectrum Scale parallel file system to offer a solution designed to scale to high performance and massive capacities while reducing capital costs by leveraging erasure coding technology for storage efficiency. The net result targets efficient performance scaling with enterprise-level manageability and resiliency.

[1] Source: ESG Research Report, *2016 IT Spending Intentions Survey*, February 2016.
[2] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends*, January 2015

The Demand for Greater Performance in File Workloads

ESG recently conducted a couple of research studies that both revealed an increased emphasis on business intelligence and data analytics solutions. Additionally, in a separate study focused specifically on the storage industry, business intelligence and data analytics workloads was also the most commonly identified workload category expected to drive storage growth over the next 24 months. These responses outpaced commonly high growth applications such as ERP, collaboration, and email. This study was conducted by polling 373 storage experts across a wide range of IT organizations.[3] These data points reveal that not only is data analytics a key initiative, but that it also has a significant impact on data storage environments.

FIGURE 1. Top Ten Applications/Workloads Most Responsible for Storage Growth



Source: Enterprise Strategy Group, 2016

These points add to the earlier identified trend highlighting an increased demand for real-time analytics. Demand for immediate insights significantly increases the transactional performance requirements placed on the storage infrastructure, and pushes them well beyond the performance commonly expected from longer-term archiving or simple file serving. To achieve lower latency and a higher rate of transactions, storage implementations for big data or business intelligence workloads have in some cases leveraged internal server storage or DAS-based storage. These solutions, however, often are limited in data protection and resiliency features needed to meet the standards required by enterprise environments. As such, these storage solutions typically house either a secondary or tertiary copy of the content, increasing data storage and management costs. Additionally, since these DAS architectures are limited in scale, the data that is leveraged in the analysis is often a subset of the content pool. Limiting the data potentially limits the quality of the

[3] Source: ESG Research Report, *2015 Data Storage Market Trends*, October 2015.

results. External storage systems provide greater scalability while also offering enterprise-level resiliency features. When considering storage solutions that offer both performance and capacity scalability, scale-out NAS or scale-out file system solutions are frequently at the forefront of the investigation. While these terms are regularly used interchangeably, there is a critical difference that impacts performance efficiency, which becomes a factor when architecting business intelligence and data analytics environments.

Scale-out NAS Versus Parallel File System

The concept of scale-out storage was introduced as a means to extend storage scalability beyond the confines of the traditional NAS filer silo. Traditional storage architectures are difficult to manage, protect, and maintain at scale. In response, scale-out NAS and scale-out file systems emerged and offer similar benefits. Both present a single pool of storage across multiple storage devices. These devices scale in a parallel fashion, adding incremental performance as capacity is added. In terms of storage features and functionality, different vendors offer different feature sets, but there is typically nothing inherent in the architecture of either scale-out NAS or scale-out file systems that would limit specific functionality.

Despite the similarities, there can be a distinct difference in performance efficiency. In scale-out NAS architectures, the capacity of multiple storage nodes is consolidated and presented as a single storage pool. When files are stored, however, individual files are individually dispersed across the multiple nodes. Often with these solutions, an individual client can access only one file through only one node at a time. This architecture creates bottlenecks. A parallel file system, as an alternative, uses distributed locking to synchronize access to data and metadata on a shared disk. This protects the consistency of file system structures in the presence of concurrent updates from multiple nodes and provides single system image semantics without a centralized server handling all metadata updates. In these instances, a single client can leverage the performance of multiple scale-out systems rather than being limited to a single device. This capability is critical for environments with larger files, allowing a single large file to be spread across multiple nodes. Different technology solutions, however, offer different capabilities despite architecture similarities. When evaluating solutions, the ability to spread content across multiple nodes is a key capability to look for when seeking efficient performance scalability.

Another consideration when evaluating scale-out file systems or NAS architectures is the deployment method. Is the solution SDS-based and offered as software or is the solution delivered as a collection of integrated appliances? Each deployment option introduces a different set of limitations and advantages. IBM's Elastic Storage Server endeavors to combine the benefits of both SDS and integrated appliances, leveraging IBM's Spectrum Scale SDS technology, while resolving some of the challenges of SDS deployments.

IBM Elastic Storage Server: An Integrated SDS Architecture

The definition of SDS has been blurred as different storage providers leverage the term to mean different things. Yet, there is a general consensus that includes the ability to abstract storage software from the storage hardware. In the case of IBM's ESS solution, the Spectrum Scale technology delivers a parallel file system as software, which ESS validates and integrates onto a cost-effective server hardware platform. While a fully abstracted SDS solution provides benefits in hardware choice, this flexibility is not ideal for every environment. Procuring storage software and hardware separately shifts much of the integration and validation effort to the IT organization. These solutions offer the flexibility but increase the responsibility of validating the resulting solution.

IBM ESS Feature Overview

- Integrated and validated SDS-based storage server
- Parallel file system architecture
- Global name space and multi-site read/write access
- Multiple protocol support: NFS, CIFS, S3, Swift, and HDFS for in-place analytics
- Support for two- or three-fault-tolerant erasure coding as well as three-way and four-way replication

By delivering Spectrum Scale as an integrated and validated system, IBM's ESS offers many of the cost-effective benefits of SDS, while also providing the solution validation. The inherent SDS design allows IBM to leverage performance innovations in server hardware quickly, without requiring the IT organizations to procure individual components and then piece them together. The result is a parallel file system architecture that is delivered as an integrated and validated package. In addition, IBM's ESS offers a number of incremental storage benefits, including:

- Multiple protocols for a wide variety of workloads: Built on a parallel file system architecture, IBM's ESS offers multi-protocol support, including NFS, CIFS, S3, Swift, and HDFS connectivity. ESS offers the ability for a single solution to serve multiple applications simultaneously without restricting the applications to a specific set of protocols. Support for HDFS allows running analytics applications on the data without having to copy data to a separate storage silo, allowing for in-place analytics. Without the ability to support multiple protocols and data types, the complexities and costs of managing storage silos remain, even if those silos can scale out.

- Global collaboration empowerment: Intelligent caching of data at remote sites ensures that data is available with local read/write performance across geographically distributed sites using Active File Management (AFM).

- Information lifecycle management: Using storage policies transparent to end-users, data can be compressed, encrypted, or migrated to the tape or cloud to help cut costs; data can also be migrated to high-performance media tiers, including server cache, based on a heat map of data to lower latency and improve performance.

- High availability and data protection: ESS includes the core data management and protection capabilities required for enterprise storage environments. ESS leverages IBM's Spectrum Scale RAID support, which offers highly reliable two-fault-tolerant and three-fault-tolerant erasure coding, as well as three-way and four-way replication.

- Flexible deployment methodology: By leveraging Spectrum Scale's SDS architecture, ESS is able to provide extended levels of hardware flexibility. For example, ESS implements data protection functionality as software and the solution can leverage both DAS-based and JBOD hardware. Additionally, ESS can be added to existing Spectrum Scale clusters running on above third-party storage hardware for increased flexibility.

Despite the benefits of an integrated solution, some may continue to argue for a fully abstracted SDS offering where the hardware is procured separately for greater deployment flexibility. For those environments, IBM provides its Spectrum Scale SDS solution. IBM's ESS storage targets a slightly different space, and extends IBM's storage portfolio to those organizations that seek cost-optimized and efficient scale-out storage, yet trust IBM to optimize the hardware.

The Bigger Truth

The rise of big data and business intelligence workloads has placed a renewed emphasis on unstructured data and the storage systems that house that data. Whether to deliver new and differentiated products or services, or serve their customers in more effective and meaningful ways, businesses continually endeavor to achieve a competitive advantage. Locked inside the digital content collected from multiple applications and workloads are the insights that can enable that business opportunity. Unlocking this potential, however, requires greater capabilities from the supporting IT and storage infrastructure. Unstructured storage solutions that overemphasize high-capacity scaling or cost per capacity at the detriment of performance are at a significant disadvantage in analytics environments. IBM's ESS is designed to offer flexible and efficient performance scalability combined with the cost-effectiveness of SDS. The result is an architecture that can serve the business's need for analytics while simultaneously answering its need to control costs.

www.esg-global.com contact@esg-global.com P. 508.482.0188