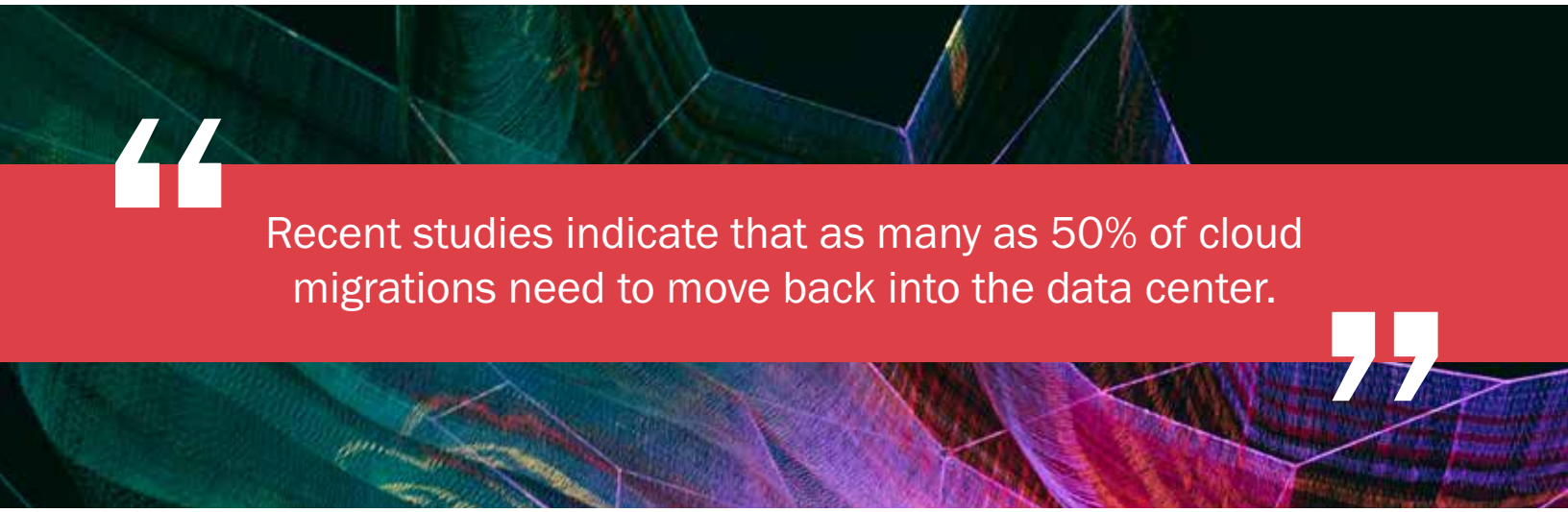
HOW TO PREVENT A CLOUD MIGRATION BOOMERANG

AVERE

Storage Switzerland, LLC

How to Prevent a Cloud Migration Boomerang

Recent studies indicate that as many as 50% of cloud migrations need to move back into the data center. While there are many reasons for this boomerang effect, lack of data consistency is one of the most common. The CAP theorem describes how distributed storage systems can only provide two of three elements: consistency, availability and partition tolerance. When considering applications for cloud migration, the CAP theorem must be applied to them to see if the cloud storage architecture is suitable for the needs of each application.



> Recent studies indicate that as many as 50% of cloud migrations need to move back into the data center.

Understanding the CAP Theorem

The CAP theorem focuses specifically on distributed data stores, which are the prevailing design used by cloud providers like Amazon Web Services, Google Cloud Platform and Microsoft Azure. Distributed data stores are built from a series of servers, called nodes. Each node has internal storage, which, when the nodes are clustered, becomes a global storage pool. As data is written to the store, it is distributed across the nodes either by replication of each object or the erasure coding of sub-segments of the object.

The CAP theorem suggests that a distributed data store can only provide two of these three elements: ***consistency, availability and partition tolerance.***

Partition tolerance *is the ability for the distributed store to continue to serve data in the event of a network failure, causing a drop or delay of an arbitrary number of messages between nodes.*

Since network failures are inevitable in any global cloud-based storage service, all systems have to provide partition tolerance. As a result, when considering a cloud storage solution, IT must choose between consistency and availability.

> When considering a cloud storage solution, IT must choose between consistency and availability.

Consistency *means the system must acknowledge every write before the application or corresponding process proceeds. The need for acknowledgement also applies to data protection.*

For example, a consistent data store must replicate the object to another node and verify completion before acknowledging the overall write. Providing a consistent model is expensive, especially in a distributed architecture since all nodes must all have the same data at any given point in time. Providing consistency, without impacting application performance in the distributed store, means high-speed intra-node networking, short distances between nodes and high-performance storage media. Even the nodes themselves are more expensive because then need to process data so much more quickly.

Despite the cost, there are times where a focus on consistency is an absolute must.

For example, a storage infrastructure supporting a financial institution needs to make sure that all nodes are in sync at all times. If a million-dollar withdrawal is made, and those nodes are not updated, serious problems can arise. More generally, applications that expect consistency can easily become corrupt if all data isn't always in sync throughout the storage infrastructure.

Availability Focus

Because of the cost associated with establishing a consistent distributed data store, most cloud providers focus on availability or an eventually consistent model.

An ***availability focus*** *means that when the initial write occurs, the acknowledgment back to the application is delayed until the data protection requirement is met, in the background, by replicating or completing the erasure coding of the data. Depending on location of the other nodes, the time to be consistent can range from seconds to several minutes. A high transaction environment may never actually get to a consistent state.*

There are numerous advantages to an availability focus, the primary one being a cost reduction.

The networking between nodes, the storage media in the nodes and the processing power of the nodes themselves don't need to perform anywhere near the capabilities of a distributed storage system with a consistency focus.

Additionally, an availability focus is ideal for data distribution. Whether for protection from a regional disaster or making sure data is geographically as close to users as possible, the availability focus enables that distribution without the need to provide performance for the originating user or application.

> An availability focus is ideal for data distribution.

Applying CAP to NAS and Object Storage (Cloud Storage)

On-premises applications considered for migration to the cloud often use network attached storage (NAS) today. These NAS systems traditionally have a consistency focus, even though much of the data on them would work just fine in a less consistent but more available design. To maintain consistency NAS systems typically offer scale-up functionality (i.e., all data access through a single node) or very limited scale-out via tightly coupled clusters and relatively small node counts. These NAS systems provide very low latency access to shared storage and as one would expect exacting consistency; they leverage read-after-write verification before application acknowledgment or traditional POSIX file system semantics.

NAS systems suffer the consequences of any other storage system focused on consistency:
high cost, limited scale and limited availability.

While most NAS systems do offer replication for disaster recovery, that replication is done asynchronously (eventually becoming consistent). The process is network performance sensitive, is almost always to an identical system and the secondary system eventually becomes a mirror image of the primary system.

The opportunity is that not all the data on the NAS requires the capabilities of a consistent architecture, in fact, most of the data on a NAS will work just fine on an architecture focused on availability. Even data sets on the NAS that do need consistency, often have a subset of the data that is better suited to an availability model.

*The most common type of availability focused storage system is **Object Storage**, a design used by all the cloud providers for their affordable storage tier. Object storage is low cost and scalable to thousands of nodes. It also is easy to distribute data across vast geographies, making the data accessible and resilient.*

Using CAP to Identify Cloud Compatible Applications

Given an understanding of CAP and the realization that the most cost-effective storage available from cloud providers is availability focused, the organization should be more easily able to determine which applications are best suited for use in the cloud.

The first step is to identify data sets where consistency is not an issue, a prime example being inactive or dormant unstructured data sets. Over 80% of the typical used NAS capacity is inactive data.

These entire data sets should move to more affordable distributed storage that focuses on availability. Moving this data to availability-focused storage not only drives down the cost of retaining this data, these types of storage architectures are better suited to their long-term retention. The challenge is not only identifying the inactive data and moving this data to the cloud, but also making sure that when the data is needed again that it is still available to users.

The next step is to determine if within active data sets there are subsets that are also dormant.

The problem is that identifying and moving these data subsets is even more difficult. The last step is to identify data that is very active and requires consistent representation throughout the storage architecture.

Migrating Consistency Sensitive Applications to the Cloud

While the three types of data can be manually identified and moved, the process is time-consuming and requires continual supervision by IT. A manual process, and even some that claim a level of automation, also means managing separate storage pods, one for active, consistency sensitive data, and one for inactive data that is less sensitive to consistency.

> A manual process, and even some that claim a level of automation, also means managing separate storage pods

An alternative is to leverage a cloud-aware file system that will automatically classify data and place it on the most appropriate type of storage based on its access pattern.

This solution could replace on-premises, high-performance NAS with a much smaller storage footprint that is designed to house consistency sensitive data. It should automatically move data between the on-premises and cloud storage with IT oversight but without IT intervention. At the same time, it should overlay the on-premises storage and cloud storage with a global file system so that users are always accessing data via the same path and protocol so access to the cloud is seamless.

The Consistent Cloud

The cloud-aware file system should also run natively in the cloud, which allows the organization to move applications seamlessly to the cloud without having to change them. The applications can run on the cloud providers' version of consistent storage, but the file system allows that investment to be minimal, as it will move data not needing consistency to the provider's less expensive object storage tier automatically.

When on-premises storage and cloud storage are combined with the cloud file system data and applications are free to move between on-premises and cloud, based on the requirements of the enterprise, cloud storage can be used as an archive, as a place to temporarily burst applications because of an unexpected peak or a place to permanently run the application.

> Cloud storage can be used as an archive, as a place to temporarily burst applications because of an unexpected peak or a place to permanently run the application.

Conclusion

The boomerang effect of most cloud initiatives is often the result of a mismatch between the performance requirements of the dataset and the cost savings goal of the organization. The gap between the two often leads the organization to either bring their migrated application back on-premises, hence the boomerang, or the organization has to make a much more sizable monetary investment in consistent cloud storage. A cloud file system bridges that chasm and enables the organization to place any application anywhere cost-effectively, striking the right balance between high-performance consistency and cost-effective availability.

About The Firm

Storage Switzerland is an analyst firm focused on the storage, virtualization and cloud marketplaces. Our goal is to educate IT Professionals on the various technologies and techniques available to help their applications scale further, perform better and be better protected. The results of this research can be found in the articles, videos, webinars, product analysis and case studies on our website ***storageswiss.com***

About Avere Systems

Avere helps enterprise IT organizations enable innovation with high-performance data storage access, and the flexibility to compute and store data where necessary to match business demands. Customers enjoy easy reach to cloud-based resources, without sacrificing the consistency, availability or security of enterprise data. A private company based in Pittsburgh, Pennsylvania, Avere is led by industry experts to support the demanding, mission-critical hybrid cloud systems of many of the world's most recognized companies and organizations.

Learn more at www.averesystems.com