PROVIDING A UNIFIED DATA LAYER AT MEMORY SPEED FOR MODERN CLOUD ENvironments

White Paper
AGENDA

Abstract ........................................................................................................................................................................ 3
Cloud Adoption Continues to Advance ..................................................................................................................... 4
Data First Cloud Deployments with Alluxio ............................................................................................................... 5
An Open Cloud Solution for Multiple Cloud Vendors and Software Platform Providers ........................................... 6
Key Cloud Use Cases ............................................................................................................................................ 6/7
Conclusion ................................................................................................................................................................ 7
Virtualize any file or object store in a cloud or hybrid environment, connecting applications to heterogeneous storage with a standard interface. Access data with local performance, regardless of location or format, and without requiring multiple data copies, migration or ETL (extract, transform, load).

**ABSTRACT**

The cloud is rapidly becoming ubiquitous, with continued adoption focused on the flexibility and cost benefits of a utility infrastructure model. Enterprises are increasingly taking a “data first” view of infrastructure, which demands a new way of thinking in a world in which data is stored and accessed from multiple locations and providers. Performance and interoperability challenges, however, can present obstacles to cloud adoption and complicate data management. Techniques such as the use of data silos, ETL processes and multiple data copies, which are commonly employed to accommodate cloud limitations, often tend to offset the expected benefits of cloud infrastructure.

Alluxio offers a new way to enhance the benefits of cloud infrastructure without the performance limitations or interoperability challenges resulting from accessing disparate data sources in multiple, often remote, locations. Applications connect to a single virtual data layer that unifies storage, regardless of location or format, and without requiring the use of multiple data copies, data lakes or ETL processes. Data is cached locally in memory. This ensures the highest possible performance and avoids the usual latency penalties caused by remote access over the network. Alluxio provides a standard interface to applications and abstracts the storage interface. This means that no changes to applications or storage are required.

Alluxio works in any type of cloud environment, and was recently tested with the Open Telekom Cloud. Open Telekom Cloud is Deutsche Telekom’s OpenStack based cloud solution. It is powered by Huawei and provides trusted and reliable public cloud services to enterprise customers. Common use cases include analytics and Machine Learning (ML) applications in the cloud.
CLOUD ADOPTION CONTINUES TO ADVANCE

Cloud adoption has been one of the driving forces in digital transformation over the last decade. Forrester predicts that 2018 will mark a major milestone in that over 50% of all enterprises will use a public cloud platform. In the near future, with very few exceptions, virtually all applications will be run from a public or private cloud, or be deployed in a shared, virtualized cloud environment. While the industry has not yet completed the transformation, we are nearing the point where IT infrastructure will indeed be a utility.

Cloud infrastructure is a utility model, and one which provides the advantages of both flexibility and effectiveness.

The flexibility has several dimensions:
- Scale on demand
- Independently scale compute and storage resources
- Choose which services to consume based on what the vendor provides
- Choose which vendor to buy services from

Costs are minimized in several ways:
- By using lowest cost infrastructure components
- By paying for only what you use
- By exploiting cost advantages from scale of economies

As companies continue to adopt cloud computing, they are increasingly taking a ‘data first’ view of their infrastructure choices, in other words, the need to efficiently store, access and analyze data will be an important factor when making cloud infrastructure decisions.

For the foreseeable future, the focus will be on the cloud hybrid model, in which parts of the infrastructure is owned, and parts are provided by a third party cloud vendor. There will be multiple cloud providers. Data will need to move seamlessly across providers and geographies based on business objectives. Data sources will be shared across different applications and consumers, who will often have different business objectives. Data management needs to be addressed at all layers of the software stack. Resources, particularly compute and storage, are decoupled, placing demands on interoperability and APIs.

These realities create challenges for the current ways data is managed and can slow the progress of cloud adoption. The first obstacle is typically performance related. Data is not always going to be co-located with compute, with the resulting latency from remote access leading to slower performance. To compensate, companies often create multiple copies of data and employ time-consuming ETL processes to make data available at the right place and in the right format. These approaches offset the cost and flexibility advantages of cloud computing.

Multiple data sources and multiple sites also introduce complexity and interoperability challenges. Two cloud providers may very well require slightly different storage interfaces – even for similar object stores. It makes sense to store active data in higher performance storage, while storing less frequently accessed data in low cost, high capacity storage. Such a model, however, requires the data to migrate seamlessly when needed. Applications can require data from different storage systems with different formats. For example, a machine learning application might require data from an object store in the cloud as well as file data from another location. This problem is often addressed with ETL processes or by creating separate data lakes.
DATA FIRST CLOUD DEPLOYMENTS WITH ALLUXIO

Alluxio fits into existing cloud environments and acts as the data layer that virtualizes any connected file or object store. By caching active data in memory, applications achieve local data access performance levels, regardless of format or where the data is physically located. Applications connect through a single flexible interface without any changes to storage or applications, thereby eliminating interoperability issues. With this model, Alluxio is able to help companies realize the goal of using the lowest cost storage without experiencing any of the usual drawbacks resulting from slow remote performance and interoperability issues.

This model exploits the cost and flexibility advantages of cloud computing. Resources can be scaled on demand. And because compute and storage are decoupled, they can also be scaled independently. The flexibility is extended to physical infrastructure locations, such as datacenters located in multiple geographies, and the ability to deploy on multiple cloud vendors is ensured. Cost are controlled by paying for only what is consumed. The majority of data is stored in the lowest cost storage, usually object storage, with active data stored in the Alluxio memory layer for local performance.

Applications can access storage using any of the standard APIs supported by Alluxio, including, for example, S3 or a compatible version of Hadoop. Working data is cached in the Alluxio data layer and retrieved from persistent storage when requested by an application. Intelligent cache management ensures active data is available at local speed and ‘cold’ data is instantly accessible. Storage capacity is managed independently from compute and with no effect on performance. Over time, data can be migrated to the lowest cost storage tier that is available. Compute resources can ‘burst’ into the cloud and still have access to local data performance.
KEY CLOUD USE CASES

Remote Data Acceleration
In a global hybrid model, data can be anywhere. Performance issues from remote data access can limit cloud adoption. With Alluxio, working data can be stored locally to the application that needs it and accessed at memory speed. Long-term persistent storage can be optimized for cost and capacity independently of performance.

Analytics / Big Data
Insights result from matching the right analysis tools to the right data. Data lakes are a common tool to address analytics projects. This typically involves creating an independent storage silo dedicated to analytics. Data from multiple sources is migrated to a centralized data lake, usually through an ETL process. Multiple applications and users can share data lake resources. Data lakes are often deployed in Hadoop clusters with tightly coupled compute and storage. Capacity needs to be closely managed, and quality often suffers from ‘stale’ data.

Alluxio’s data layer supports all major cloud service providers, including Open Telekom Cloud. It also fits into both IaaS as well as PaaS cloud service models.

Open Telekom Cloud offers a complete range of horizontal and vertical cloud solutions. Alluxio provides a common data layer for Open Telekom Cloud that addresses cloud challenges such as performance, complexity and interoperability. A key advantage of Open Telekom Cloud is its easy integration of, and robust support for, standard cloud APIs. Customers can deploy Open Telekom Cloud together with Alluxio and seamlessly utilize any supported storage system or interface.

Alluxio supports the full functionality of Open Telekom Cloud services by providing comprehensive API support. While the list of available Open Telekom Cloud services is extensive, the following highlights several key services:

- Elastic Cloud Server (ECS)
- Auto Scaling (AS)
- Object Storage Service (OBS)
- Migration as a Service (MaasS)
- MapReduce Service (MRS)

Alluxio enables the creation of a virtual data lake in a hybrid cloud environment. Enterprises realize the benefits of low cost storage for large data sets shared across multiple applications without experiencing the usual performance and interoperability issues. The virtual data lake can be temporarily cached in memory, giving applications local performance. Only the working data set needs to be cached and low-cost systems can be used for long-term persistent storage. Data can be accessed in any file or object format and in any location, eliminating the need for multiple data copies or ETL processes.

Machine Learning / Artificial Intelligence
Machine Learning (ML) models improve with more data and repeated training. Large data sets and on-demand compute resources are well suited to cloud deployments. Performance is also critical, as ML demands multiple fast iterations to improve the model. Remote data access in the cloud can be too slow to practically deploy ML in the cloud. Data pipelines are also critical for ML applications, and repeated reads and writes to remote storage can be a significant performance limitation.
CONCLUSION

As more applications are being deployed in the cloud, new challenges are arising. Increasingly, a hybrid cloud infrastructure will become commonplace during a cloud transition phase, and for many companies the steady state of the future. As enterprises seek to maximize the value of the digital transformation, new strategies are needed to interact with data. Alluxio offers a new way to enhance the benefits of infrastructure like the Open Telekom Cloud without the performance limitations or interoperability challenges resulting from accessing disparate data sources in multiple, often remote, locations. This simplifies the transition process and protects today’s investment for the future.

Alluxio addresses these problems by caching working data sets in memory for ML applications in the cloud. Data scientists can tap into large and varied data sets for sophisticated model building. Using Alluxio, data can be shared between pipeline stages at memory speed. By reading and writing data in Alluxio, the data can stay in memory for the next stage of the pipeline, and this can greatly increase the performance. Alluxio also includes a Fast Durable Writes feature, which enables low latency and fault-tolerant writes.