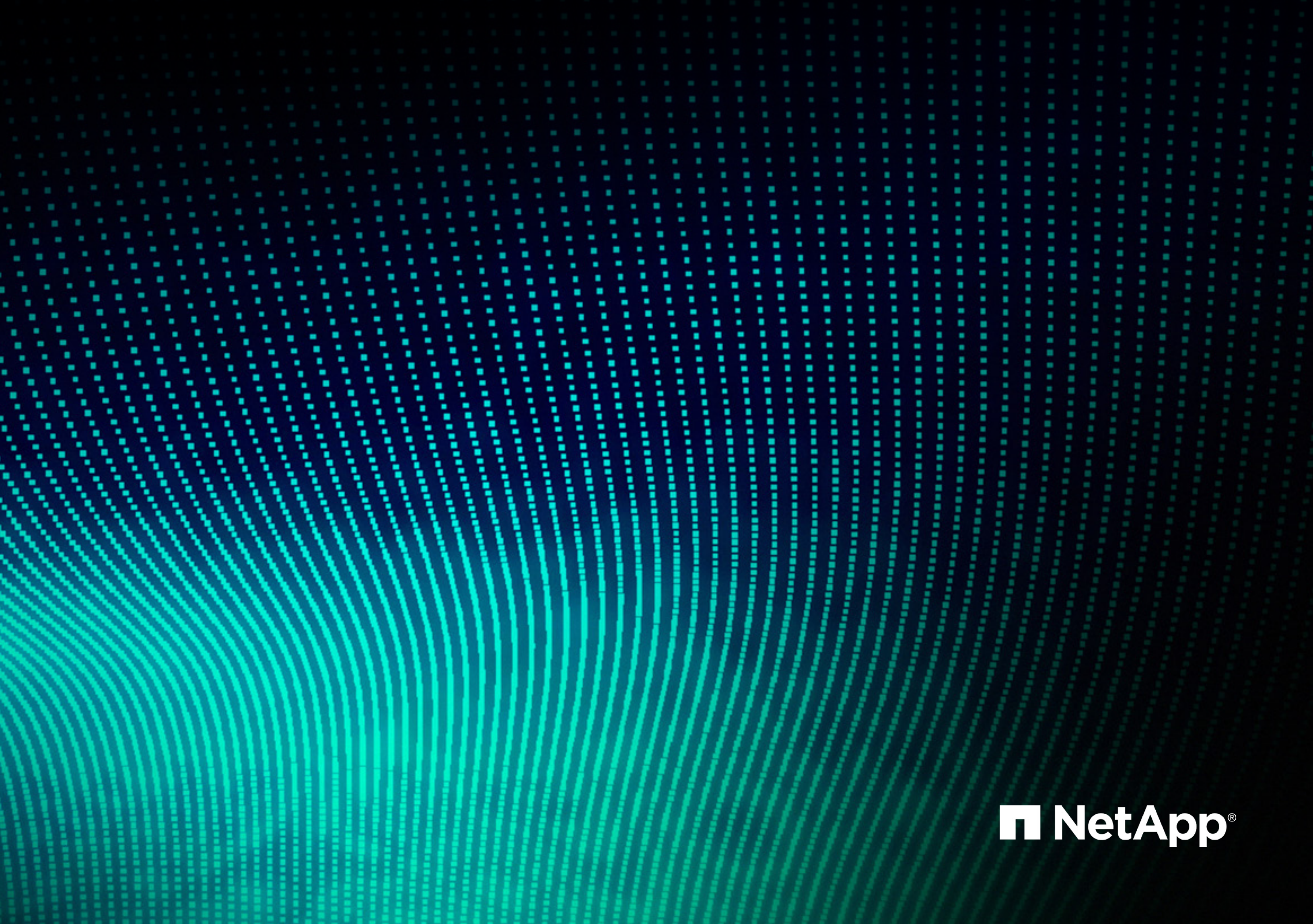
How to Realize Your File Requirements in Azure

A deep dive into how to solve challenges that moving to the cloud presents for file services, enterprise database applications, high-performance computing and analytics.



Contents

Prelude

As companies make the decision to transform IT operations by moving more of their applications[1] to the cloud, they are often confronted with barriers. Some are simple annoyances that we hate but learn to cope with, while others are far more critical to getting your applications into the cloud.

One of these core struggles revolves around a seemingly simple topic – File Services. Though file services were once thought to *not* be a requirement in the cloud, enterprises quickly started to encounter this problem, from simple shared directories of data to more complex issues with enterprise applications, and even high-performance computing environments.

In this paper, we open with an overview of a new file service offering in Azure, and then examine in-depth the challenges IT departments face when addressing cloud mandates across:

- file services-based applications,
- enterprise database environments,
- high-performance computing workloads,
- and data analytics.

We highlight considerations for choosing Azure NetApp Files as the solution for data infrastructure challenges with these application environments and point you in the right direction to get started.

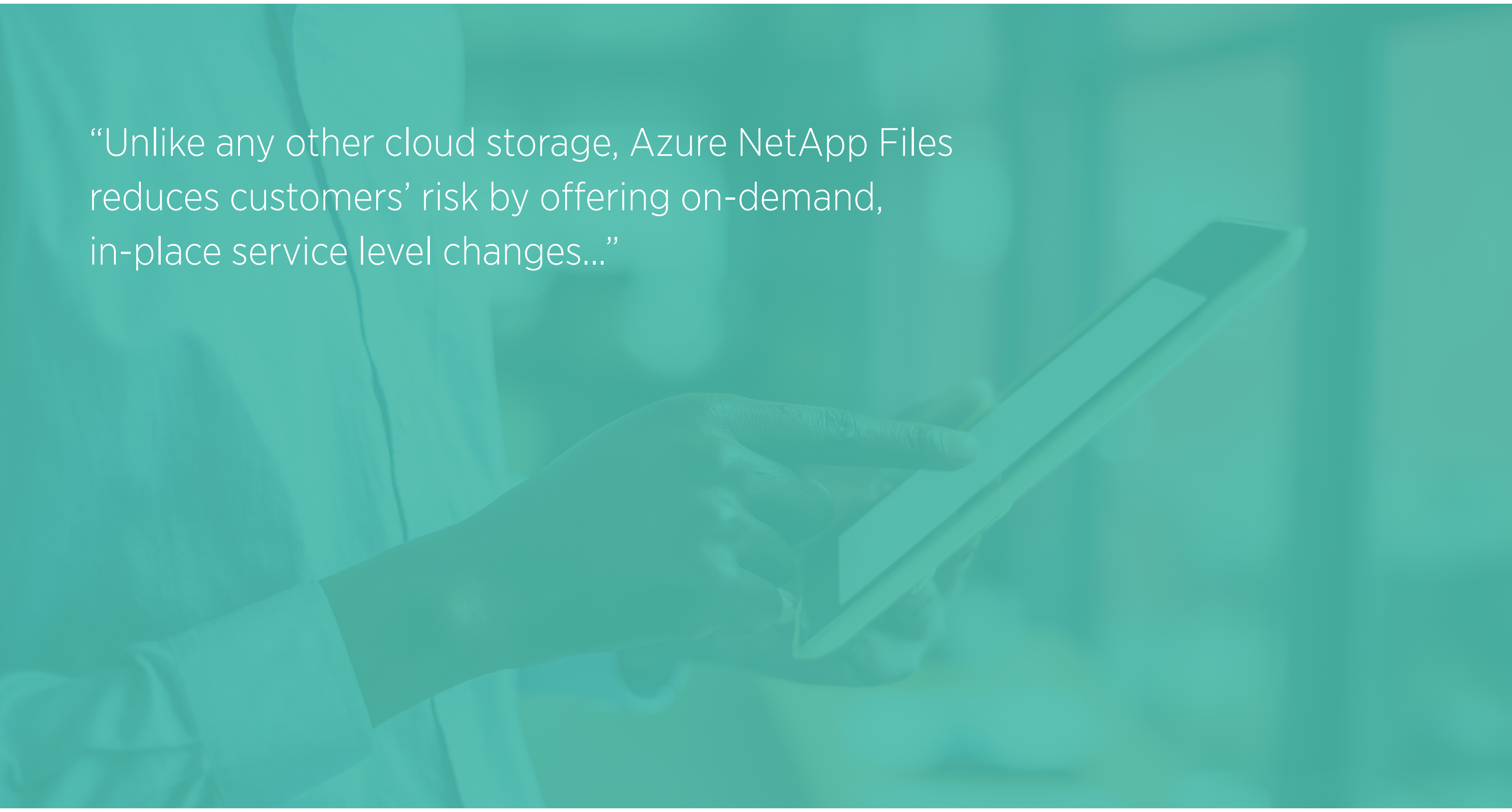
An Introduction to Azure NetApp Files

Azure NetApp Files is a Microsoft Azure service that offers a fully managed, highly available solution for provisioning file services in Azure. NetApp was selected by Microsoft to build this solution because of our industry-leading reputation and years of experience in building data management systems for the cloud, providing data protection, scalability, and advanced storage management. The service is used like any other cloud storage service, allowing seamless administration with all your cloud infrastructure. Azure NetApp Files is sometimes described as a first-party service, which simply means that while Azure NetApp Files is built on NetApp Technology, it is sold and supported by Microsoft.

The service brings support for both Linux and Windows files, with three service levels, such that customers can choose the protocol and performance that best match their application requirements. Unlike any other cloud storage, Azure NetApp Files reduces customers' risk by offering on-demand, in-place service level changes, so the service level can be changed in seconds without requiring a time-consuming data copy.

Azure NetApp Files offers three performance tiers:

- Standard service level:
  16MB of throughput per TB
- Premium service level:
  64MB of throughput per TB
- Ultra service level:
  128MB of throughput per TB



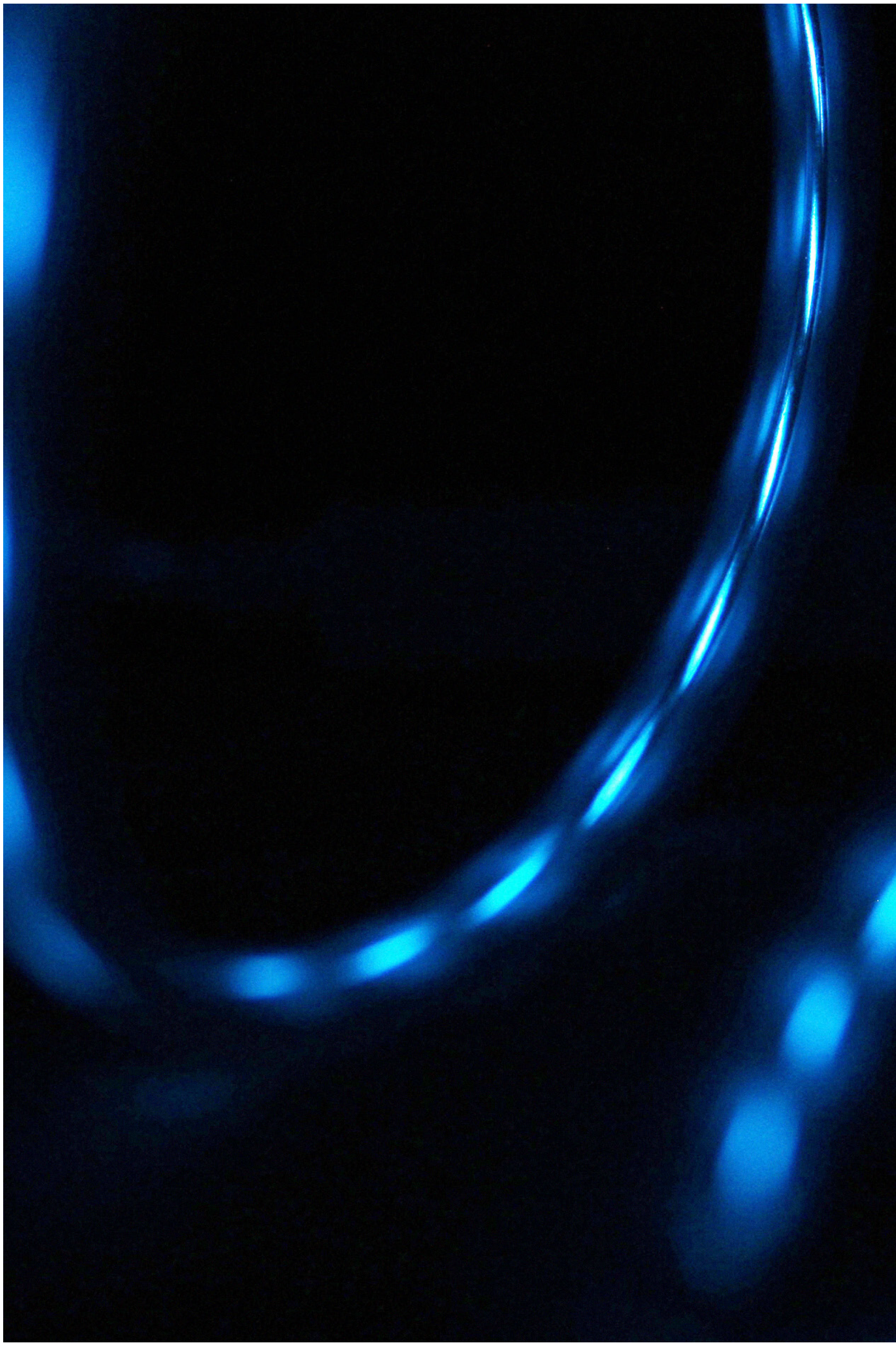
> "Unlike any other cloud storage, Azure NetApp Files reduces customers' risk by offering on-demand, in-place service level changes..."

Included with the service is NetApp's Snapshot™ technology, which allows customers to take a highly efficient, point-in-time snapshot of their data volume without having added capacity, cost, or impact on application performance. No longer does a cost trade-off need to be made with how regularly snapshots are taken.

ADDITIONAL AZURE NetApp FILES CAPABILITIES

Benefits:

- Built on highly available NetApp technology that ensures your data is available when applications need it most
- Simple Azure-native service interface on the Azure Portal means there is nothing to purchase – you consume Azure NetApp files as part of your Azure agreement with Microsoft
- Linux NFS v3, v4.1 and Windows SMB support
- Maximums
  - Volume size 100TB
  - File Size 16TB
- Secure, always-on encryption-at-rest
- Built on Azure service infrastructure with full Azure CLI support



File Services – Challenges in the Cloud

When it comes to shared storage in the cloud, enterprises – even those operating exclusively in the cloud – are struggling to deploy mission-critical workloads to Azure. In addition to the obvious challenge of running a major workload in the cloud, there is a need to spin-up file-based storage – whether in small amounts or for large-scale, high-performance file applications – while still providing availability guarantees and data management functions to ensure robust data across multiple data formats and operating systems.

According to research from IDC, Linux owns 68% of the OS market share globally[2], and it's no surprise that Linux is dominating in Azure. Enterprises love to build applications on Linux. The system is robust, flexible, proven, and, as a "free" open source technology, does not require permissions or command licensing fees.

Files systems in the cloud must provide the following:

Agile Data Performance. With a wide variation in performance for file applications, from simple shared files to files in critical database environments, it is important to have file services that support multiple SLAs but can scale to high-performance access and throughput for peak operability.

Data Compatibility. An enterprise's file services need to be compatible with all their host data formats and operating systems. Linux environments are the de facto standard when it comes to enterprise applications, but often access comes from both Linux and Windows clients. Having a robust, multi-protocol file service solution is a requirement to moving any of these applications to Azure.

High Data Availability. When it comes to running an enterprise file service, any disruption in normal operation can negatively impact the business. Whether an outage is caused by a disaster or through an update process, it is essential to ensure the availability of the file share, with zero downtime and no data loss.

Robust and Protected Data. File shares require companies to comply with industry-specific data security and data protection regulations, and because file shares are often the most important part of a business's operation, the ability to automatically create copies is mandatory to meet the stringent RPO, RTO, and backup requirements for most industries.

Data Security. When moving outside the data center, security becomes a major concern for data anywhere it resides, whether in flight or at rest. It's imperative that the file system is in the user's control at all times. Ensuring data security at all levels is crucial to preventing losses before they happen. Role-based access and encrypted data can keep data more secure.

File Services with Azure NetApp Files

On top of the performance, with Azure NetApp Files, users no longer have to worry about storage management. The service takes care of all the setup, configuration, updates, performance, and service levels. The service is built on the NetApp technology that has served enterprise customers for more than 25 years – faster and better than ever.

NetApp Snapshot Technology

For data protection, no solution supports a file service more efficiently than NetApp Snapshot technology. Snapshot copies give Azure NetApp Files the ability to provide companies with point-in-time backups that can be created instantly and do not add to their data footprint in Azure, which can save both time and money while ensuring that important data protection goals are maintained. You also get powerful, high-speed data copies based on any one of these Snapshot copies that can be used for a number of purposes, such as building test environments and restoring systems in the event of accidental or malicious data loss events.

Fully managed service. Data management is handled completely by Microsoft, not the customer.

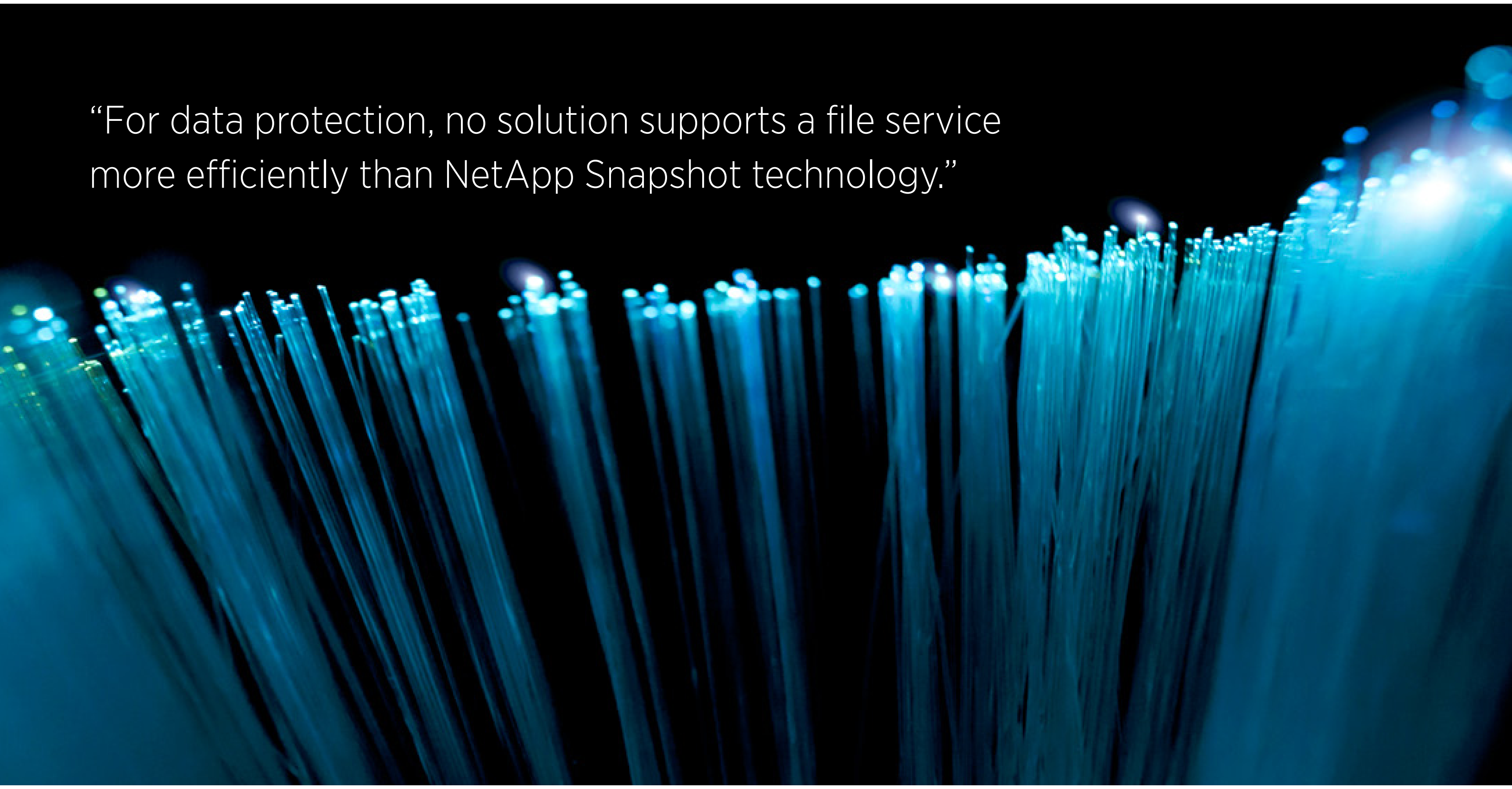
Scalability and performance. Spin up to 100TB at times of extreme high-performance storage needs in just seconds.

Compatibility. Support for SMB, NFSv3, and NFSv4.1 file shares, giving shared file access across Linux, UNIX, and Windows operating systems for greater host-client data operability.

Integration. Complete integration with file directory metadata, keeping domain credentials, access and authentication, and group memberships, including full compatibility with Microsoft Active Directory.

Data protection. Data corruption or loss can be prevented with efficient, automatic data Snapshot copies.

Automation. Schedule tasks directly via the Azure CLI to meet file-share demands with automation and orchestration capabilities.

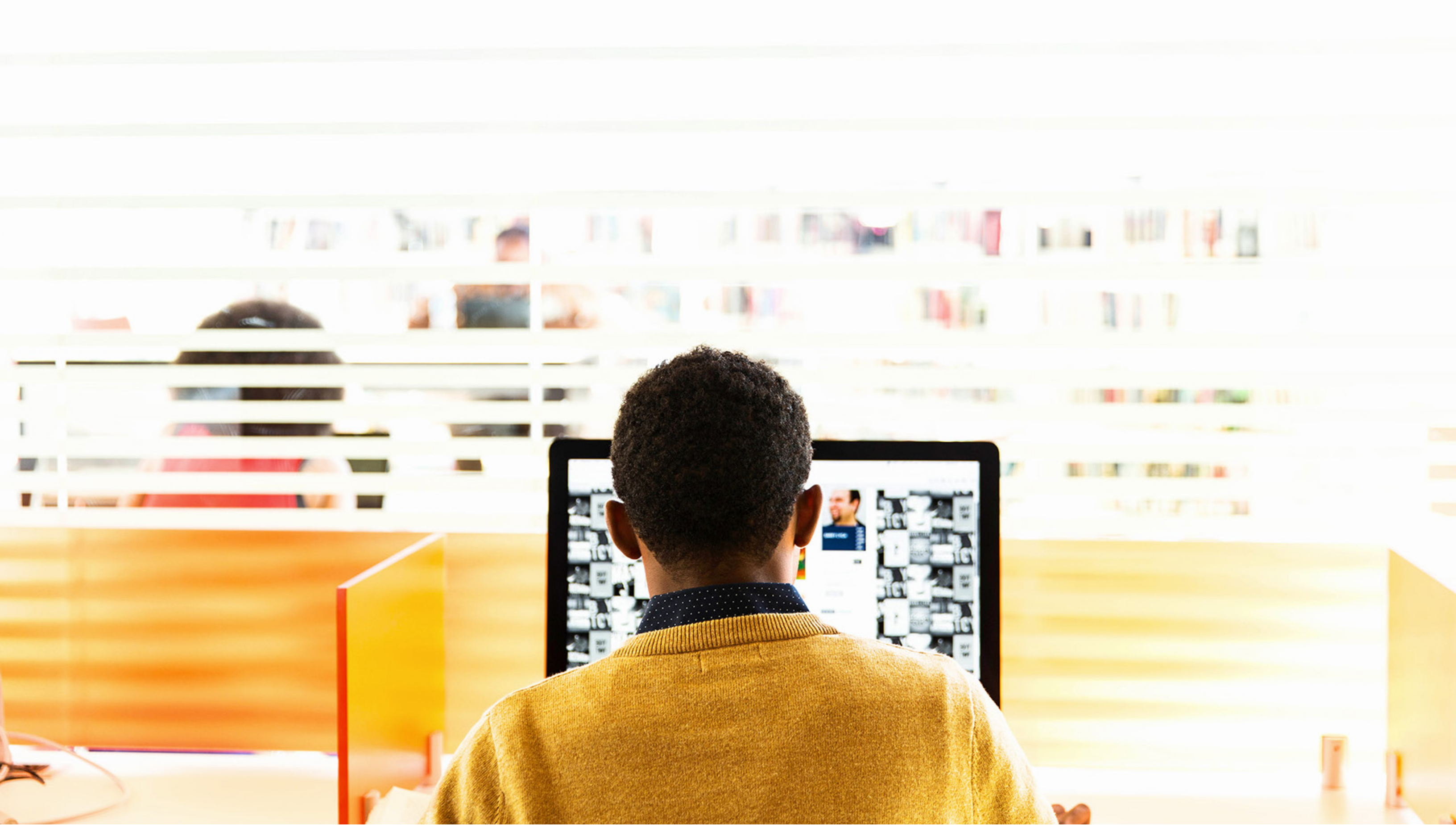


> “For data protection, no solution supports a file service more efficiently than NetApp Snapshot technology.”

Enterprise Database Applications – Challenges in the Cloud

In order to meet mandates to get core application environments to the cloud, there comes the need to have a level of performance that is often difficult to achieve with today's highly flexible cloud resources. Where compute scales with technology, having highly performant data access is a serious challenge.

Robust, high-performing, and scalable storage in the cloud is essential for deploying enterprise applications, which often have key performance requirements of some form of database. Reliable storage and retrieval of data across hundreds or thousands of concurrent client connections are paramount. Large organizations rely on their databases to be the permanent systems of record for all business transactions, which means that the data they contain must survive localized server and disk failures, as well as any site-wide failures that necessitate disaster recovery. As data volumes grow, database administrators require the flexibility to quickly scale up both the size and performance of database storage volumes to meet demand.



The following list summarizes the most crucial features that the storage environment must support:

Reliability. Because the storage environment is an integral part of the database platform, it must be available for sustained access by database servers, without fail. If access to the storage is interrupted, database operations might come to a halt, potentially causing a major disruption for all dependent applications and systems.

Durability. When a database system is ACID-(atomicity, consistency, isolation, and durability) compliant, it means the database system can guarantee that when a database transaction has been committed, the data is durable; that is, it will survive a failure. Storage environments must ensure adequate data redundancy to protect against failures.

Performance. Because almost all database operations involve reading or writing data, I/O performance determines the speed at which a database can operate. Ideally, the storage environment provides the ability to allocate storage of varying performance. This gives database administrators the flexibility to match storage capacity to performance requirements in order to stay cost effective.

Security. Organizations with strict requirements for data storage require features such as encrypted transport and data encryption at rest. These features are necessary to protect sensitive data, such as personal, financial, and healthcare information.

Data Protection. Due to the criticality of data for an enterprise application, there is a need to keep that data protected. Leveraging the capabilities of storage infrastructure to take a snapshot copy and then restore these copies is a requirement for backup and restore operations to prevent inadvertent data loss and data corruption.

Rapid Test/Development Access. A rapid copy, or clone, of data is often required to perform testing, such as for a database upgrade or an application update and deployment. Being able to clone existing volumes and quickly create temporary, writable, and up-to-date copies of large production databases, without any adverse effects to the live environment, is a big win for DevOps engineers and database administrators. Software developers with access to database test environments are able to accelerate the development and testing of new application features, ensuring faster time-to-market (TTM).

Enterprise Database Applications with Azure NetApp Files

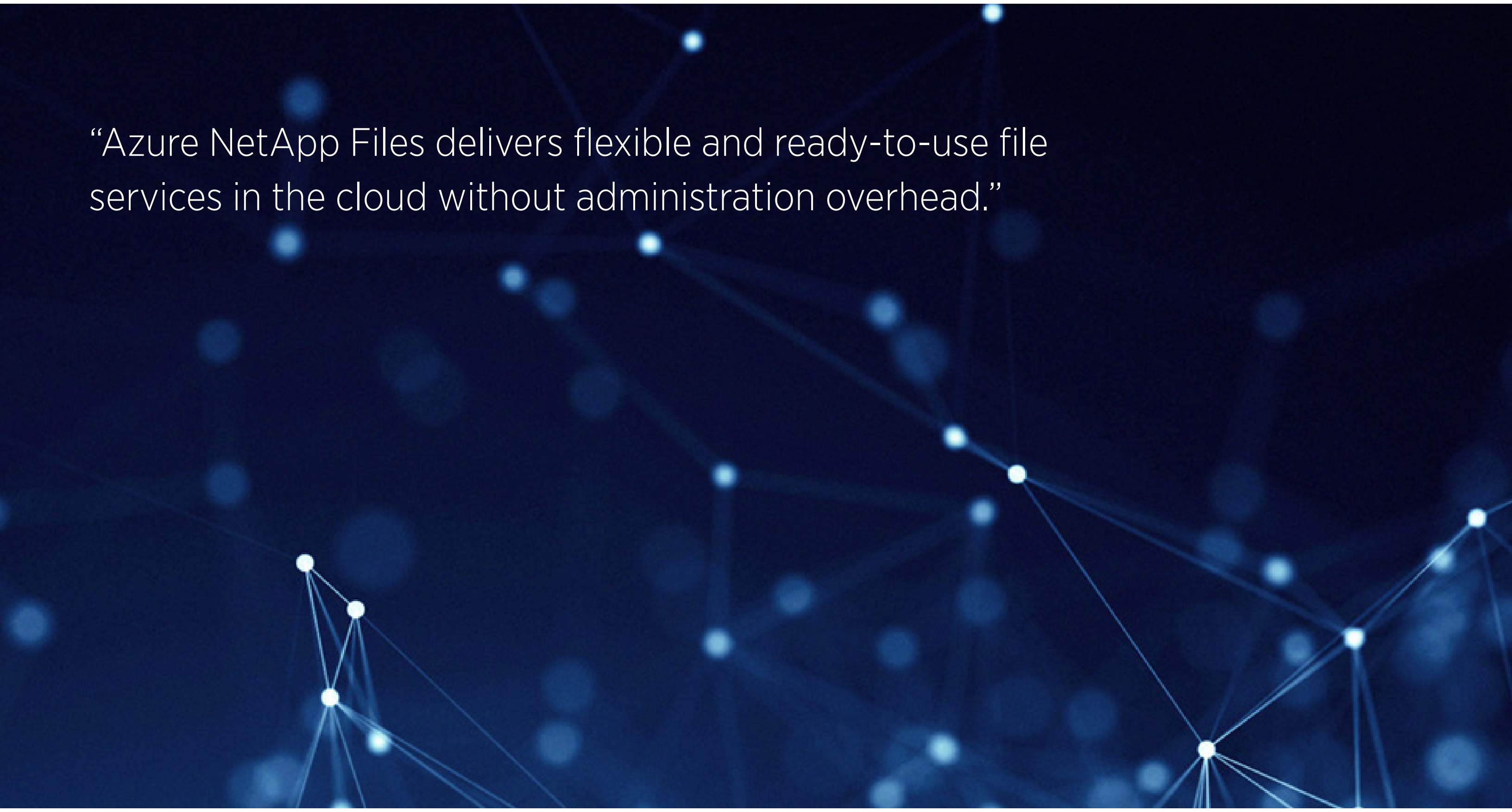
Azure NetApp Files delivers flexible and ready-to-use file services in the cloud without administration overhead. With Azure NetApp Files, users can simply set up a data volume in seconds with the availability and performance their business-critical applications require. As well as providing a powerful solution for creating and managing cloud file services, Azure NetApp Files offers an additional set of storage management features that simplify many common database administrative tasks.

Enterprise Applications can realize the following benefits:

Snapshot copy and restore. Unlike any other cloud snapshot capabilities, NetApp Snapshot technology offers instant creation of a snapshot copy of a volume without impacting the performance of the application. As well, the snapshot itself is extremely efficient and takes no additional capacity. The volume can then be instantly restored back to the point in time when the copy was created, whenever it is needed. To create a consistent Snapshot copy, NetApp recommends that users first quiesce their database system to ensure that in-flight I/O operations have been completed.

Storage cloning. Azure NetApp Files can quickly create writable copies, or clones, of existing data volumes. This is especially useful for database administrators who need the flexibility to rapidly create database test environments without the overhead of manually copying large volumes of data.

Replication and data synchronization. Leverage NetApp replication and synchronization services that can read data from any dataset, whether cloud based or on-premises, and incrementally synchronize them with your volume. Synchronization can also be performed in the opposite direction, out of the cloud volumes, to other locations.



> “Azure NetApp Files delivers flexible and ready-to-use file services in the cloud without administration overhead.”

Using Azure NetApp Files makes it easier to manage cloud storage deployments by providing the following tools and support to cloud architects and database administrators:

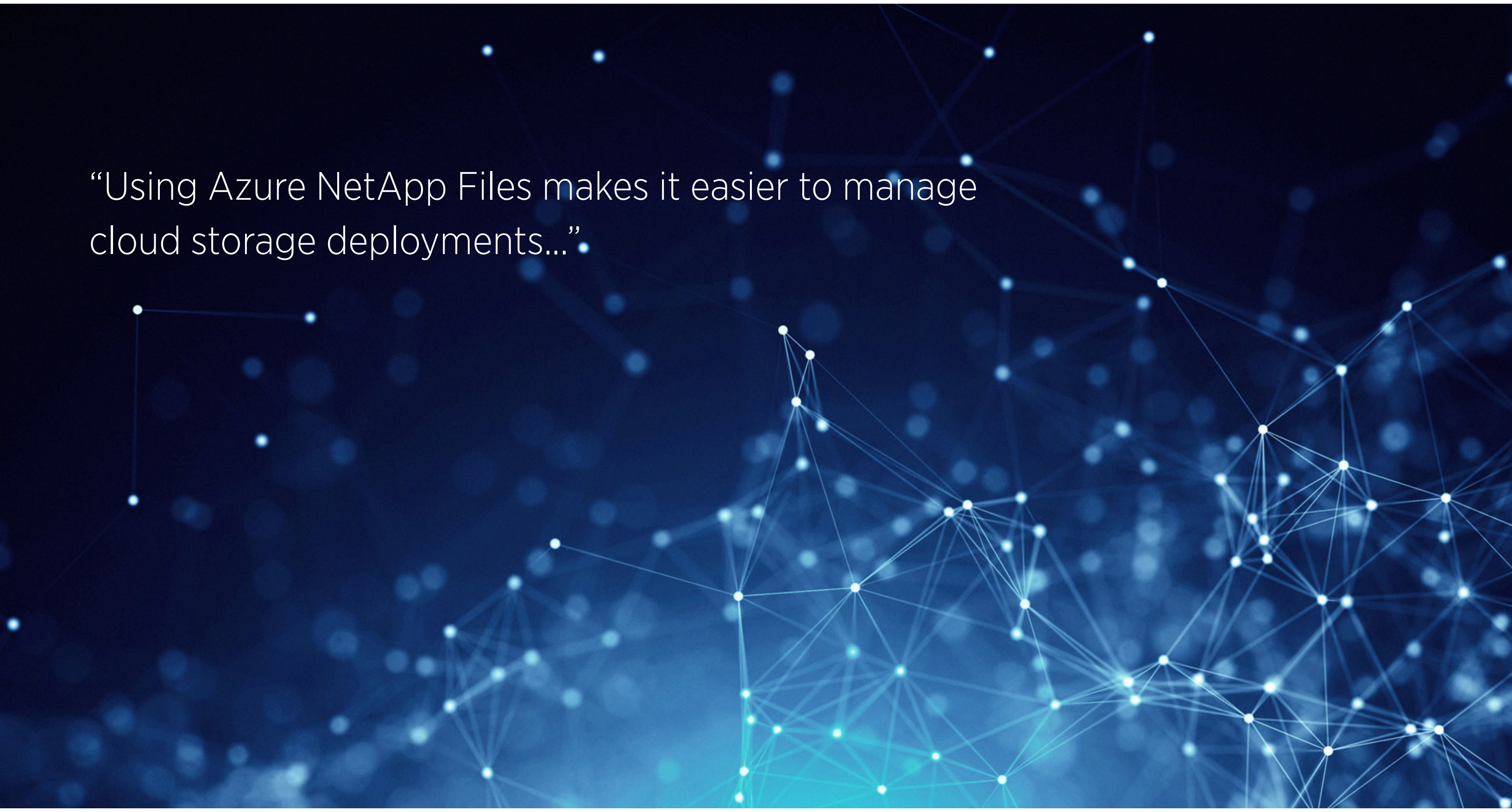
Support for all major database platforms. Reliable database storage for all major database systems, including SAP, Oracle, PostgreSQL, MySQL, MongoDB, and Microsoft SQL Server (over SMB). Oracle Direct NFS is able to open multiple, parallel client sessions to NFS shares to further increase I/O performance and scalability.

Configurable Service Levels. Storage is allocated in accordance with the service level defined when a volume is created. The service level can then be changed on demand to best suit the user's needs. This allows the performance of a volume to be controlled and storage pools of varying size and performance to be made available to a database system. Users can control cloud storage costs by allocating faster storage only where necessary.

Robust data protection. Azure NetApp Files is a cloud-native service for allocating storage in the cloud and therefore requires no user management of underlying resources. Because volumes are highly available, customers can be sure that their data will be durable and online when they need it. They can also use built-in replication capabilities to set up secondary regional copies of their data for enhanced protection.

Data encryption. All data is encrypted at-rest, and users can encrypt data both at-rest and in-transit by using a VPN. For NFS and SMB connections, users can encrypt data-in-transit from the database server to the storage volume. This provides transparent protection from malicious attempts to access the data.

Scalability. Volumes can be expanded on-the-fly, whenever more storage is required, without compromising performance or data protection. Database administrators can also reduce the size of the volume as needed to use their resources better. They can easily allocate storage for new databases or existing databases that are growing in size, without the need to manage any of the underlying physical infrastructure. This substantially reduces the administrative overhead of reacting to a change in database storage requirements.



"Using Azure NetApp Files makes it easier to manage cloud storage deployments..."

High-Performance Computing - Challenges in the Cloud

The cloud is effectively based on an agile pool of infrastructure resources – namely compute resources – that allows working environments to expand and contract as needed regardless of the size of the task at hand. By extension, one would expect high-performance computing (HPC) environments to be the perfect workloads to move to the cloud because compute is endless.

The struggle is that it never really is that simple. While cloud providers continue to increase the capabilities of the compute nodes, every application has a reliance on data. To get workloads, such as Oil & Gas, Genomic sequencing, or even Electronic Design Automation (Semiconductor design) applications to perform in the cloud, the data that these environments reside in must have not only the speed to keep up with compute, but also the resilience and availability to ensure that these critical workloads can get up and running and processed in the most timely fashion.

The following items highlight the requirements to moving these high-performance computing workloads to the cloud:

High Data Availability. High-performance computing workloads often ask a lot of the data layer in order for the computations to be completed as rapidly as possible, and often these runs can be long – in some cases in the order of days. This means that a compute run cannot afford to have a disruption for any reason, especially due to the dataset not being available. It is imperative that the dataset have the highest availability for the applications to minimize costs.

Simple, intuitive interface.
HPC workloads are regularly part of a business operation and the specialists that run these environments are looking for solutions that don't require storage-centric knowledge. A goal of the cloud is to make resources simple and easy to consume and storage should be no different.

Shared Access to data. HPC applications are often highly parallelized, where each compute node in a cluster needs access to data. Shared access to the dataset amongst a large set of compute clients allows for a broader distribution of the workload and, therefore, a faster completion time for the computational task.

Scalability. Datasets grow and contract depending on the task at hand, and the cloud infrastructure must grow and contract regularly to meet the needs of the computation. This is required for both the total capacity of the dataset and the performance needed for computation.

Reliability. Because the storage environment is an integral part of the database platform, it must be available for sustained access by database servers, without fail. If access to the storage is interrupted, database operations might come to a halt, potentially causing a major disruption for all dependent applications and systems.

High-Performance Computing with Azure NetApp Files

Azure NetApp Files looks like a service that was custom-built for a high-performance computing environment, which is exactly true. The service levels and the base feature capabilities are in place to address the critical needs of many HPC datasets.

HPC environments realize the following enterprise benefits:

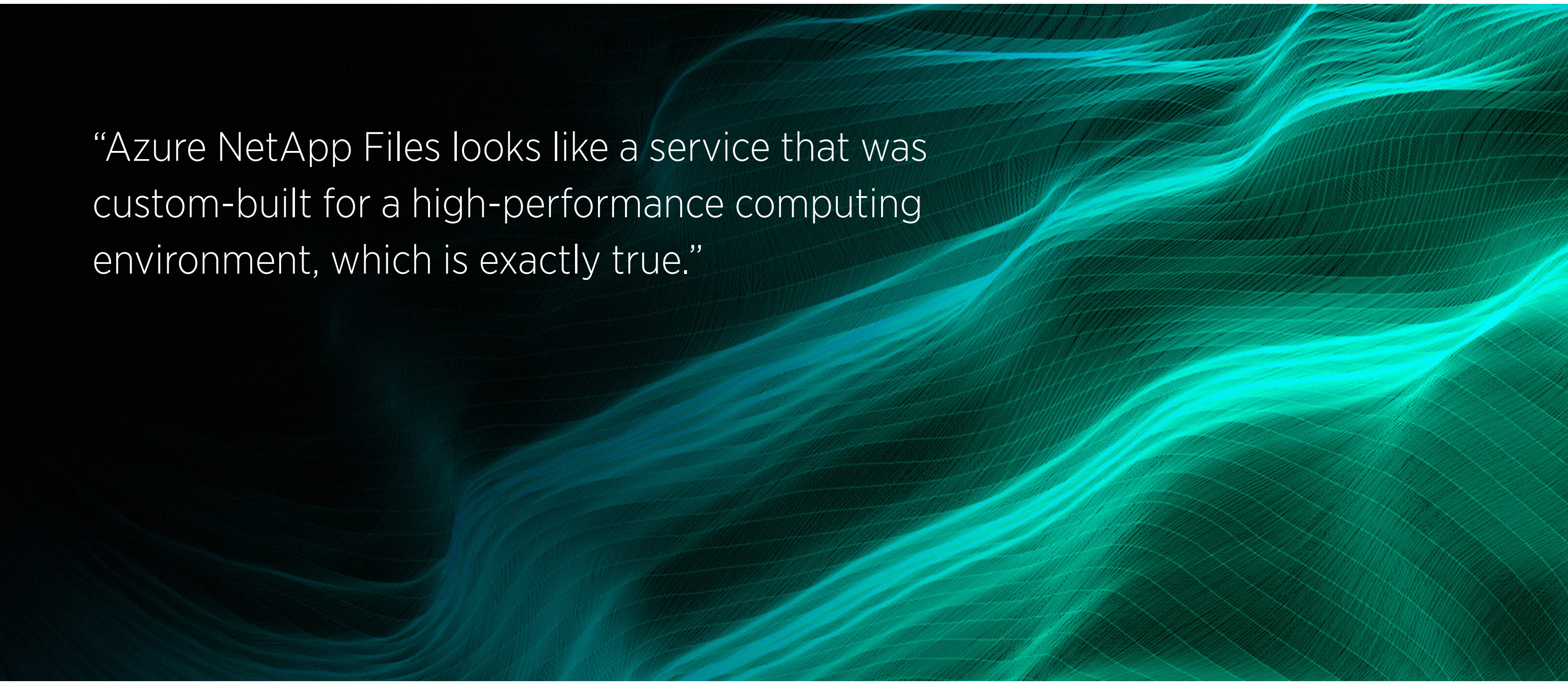
Unprecedented Cloud Performance. The Ultra service level provides performance for even the most stringent applications. The Azure NetApp Files performance operates at sub-millisecond latencies – unprecedented across any cloud file service – enabling companies to move and operate their HPC workloads in the cloud as if those applications were on-premises, in a way that would never have been possible before.

Large-scale Shared Access. Azure NetApp Files offers wide-scale access to both Linux and Windows file shares. The highly parallelized HPC architectures can grow their compute node count and effectively reduce the computational time for a given task and save overall costs to the business by having shorter development cycles or quicker data response times.

Configurable Service Levels. Storage is allocated in accordance with the service level defined when a volume is created. Azure NetApp Files uniquely allows the service level to be changed on demand to best suit the performance requirements of the computational task. This allows users/administrators agile control of performance – and effectively costs – by allocating faster storage only when necessary for the task. As a core design principle of the service, when an application needs to reach a given service level, Azure NetApp Files delivers consistent performance, at all times, without variation.

Highly Available Service. Azure NetApp Files is a cloud-native service build on highly available NetApp technology directly with the Azure data centers. It ensures that even the most critical applications always have access to the required data. A service is only as good as the infrastructure from which it operates and, with Azure NetApp Files, you can be confident that the service will be available when your HPC applications require it most.

Scalability. HPC workloads are always running on variable datasets and Azure NetApp Files makes it simple and easy, through the Azure portal or via Azure CLI, to adjust the capacity of a given volume to expand and contract on-the-fly. This substantially reduces the administrative overhead of reacting to a change in dataset requirements.



> “Azure NetApp Files looks like a service that was custom-built for a high-performance computing environment, which is exactly true.”

Data Analytics – Challenges in the Cloud

With data being the lifeblood of a company, looking for ways to utilize enterprise data for better business outcomes has driven the growth of data analytics, while the flexibility and scale of compute in the cloud has brought the two of these forces together.

Data analytics requires processing large volumes of data from multiple sources. This data must be stored on a fault-tolerant platform that can sustain high levels of performance to facilitate access by analytics solutions such as Apache Hadoop and Apache Spark. Moving the source data to the cloud enables services such as Azure HDInsight to process the data with nearly unlimited horizontal scalability. It also reduces the complexity of setting up an analytics compute cluster, removing the barrier to entry for many organizations.

Data in large organizations grows organically, normally spreading across various databases and other repositories. One of the first tasks in setting up a data analytics platform is to consolidate all relevant data into a single repository, or data lake. This repository can then be accessed by a cluster of compute nodes that apply different algorithms to the data in an attempt to find patterns and gain insights.

Data lakes are simultaneously accessed by hundreds or thousands of compute nodes, which requires the host storage service to guarantee scalable and predictable I/O performance.

This can be difficult to achieve with a custom file server solution built on cloud compute and storage resources, because managing the capacity and performance characteristics of the underlying disks becomes more and more challenging as the deployment grows.

Another major challenge for creating a centralized data repository is keeping the data synchronized with the source data after the initial baseline copy is created. As the source data continues to change, updates must be applied efficiently to the analytics data repository.

Once the raw source data is synchronized with a data lake, a certain amount of preprocessing may be required to optimize the data for processing by downstream analytics engines. Data engineers require independent copies of the data to work with in order to develop these data transformation routines. Due to the huge volumes of information in a data lake, it's very difficult to maintain multiple, up-to-date test copies of the data.

With the data lake ready to serve out data, a compute cluster can be used to target the repository and execute analytics workloads. HDInsight is a framework for building distributed data processing solutions that natively supports cluster scheduling, management, and Map Reduce operations. The foundation is a clustered file service for horizontally scaling out data storage with fault tolerance. Each compute can operate on its local data, as well as on the data in the rest of the cluster. In this way, a cluster is able to support a wide range of other data processing platforms.

Creating a data lake in the cloud means that it is no longer necessary to set up a Hadoop cluster manually. Cloud services provide ready-to-use solutions for working with Hadoop and data warehousing, respectively.

Data Analytics with Azure NetApp Files

The successful establishment of data lakes in the cloud opens up a world of possibilities for data analysis. Azure NetApp Files is part of an easy-to-use, robust, high-performance platform with the precise feature set required to create and support data lake environments.

As described in the previous sections, Azure NetApp Files offers many advantages for data operations. Here is a summary of how Azure NetApp Files benefits data analytics solutions:

Robust infrastructure. Azure NetApp files, built on NetApp technology, brings decades of experience to the infrastructure the service is built upon. The service's highly available infrastructure with dedicated connections to the Azure compute environment means that the infrastructure is fully optimized and readily available.

High I/O performance. Processing large volumes of data, as is typical in analytics environments, requires consistent, high-performance I/O systems to ensure that data is readily available to compute resources. With Azure NetApp Files' three service levels that can be changed on demand, the data lake performance can be tailored to the analytics engine requirements.

Scalability. Azure NetApp Files scales data access to a level that is not possible with other shared file services. As analytics clusters grow in size, the storage environments they depend on must continue to provide predictable high performance. This can be especially difficult to achieve with custom-built NAS solutions.

Faster results. Analytics environments usually require temporary working copies of the data to perform preprocessing operations. Such an environment is required, for example, when testing data transformations that enrich the source data. By using Snapshot and the cloning technology that come with Azure NetApp Files, writable cloned volumes can be created in a very short time. And multiple clones of the same source volume can be created concurrently.

Data consolidation. Data can be seamlessly synchronized to and from multiple data sources. Data can be consolidated from cloud-based environments, on-premises systems, and even across cloud platforms. Consolidating data from multiple sources into Azure NetApp Files brings consistent performance with enterprise data protection.

Multiple Service Levels. Environments have the flexibility to have their data performance and cost optimized to match the need with multiple service levels. Unlike with other data services, choosing the wrong performance is of little concern – with the ability to change on demand, making a change can be done in just a few clicks.

With Azure NetApp Files, NetApp brings to bear decades of experience in building enterprise NAS solutions. This means that the service easily scales to meet the most demanding conditions, providing concurrent access to thousands of client hosts and applications. Scalability to this degree is a challenging requirement for large-scale environments and is impossible to achieve with custom-built NAS solutions.

"Azure NetApp Files is part of an easy-to-use, robust, high-performance platform..."

Conclusion and Next Steps

Database systems are complex enterprise applications that depend heavily on the I/O systems they use. For the best results, storage services must combine performance, data protection, scalability, security, and flexibility into a single solution.

High-performance, scalable and highly available shared file storage is crucial to delivering a data analytics platform. The ability to effectively manage data from multiple source systems can be another major obstacle. Azure NetApp Files provides cloud-based file service solutions that address the major challenges in creating a repository for data analytics workloads, and can be used with custom-built Apache Hadoop clusters or public cloud analytics services.

Azure NetApp Files has been purpose-built to deliver the highest levels of I/O performance and scalability. End users simply input the size of storage volume they need, choose the appropriate service level for their performance requirements, and NetApp takes care of the rest. This removes the significant burden on organizations to manage in-house NAS solutions.

The synchronization capabilities of Azure NetApp Files allow data from multiple systems to be consolidated into a single storage volume. Data can also be synchronized out of Azure NetApp Files to provide integration with other external systems. Volume cloning adds to the ability to manage and work with large volumes of data.

Deciding whether Azure NetApp Files is right for you is simple because it is a service you can easily run from your Azure portal. For instructions on how to get started, visit the Microsoft Azure NetApp Files web page.

Resources

[1] https://www.statista.com/statistics/475768/cloud-applications-market-cagr-by-segment/

[2] https://www.idc.com/getdoc.jsp?containerId=US43753318

About NetApp
NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organizations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation, and optimize their operations.
For more information, visit www.netapp.com. #DataDriven