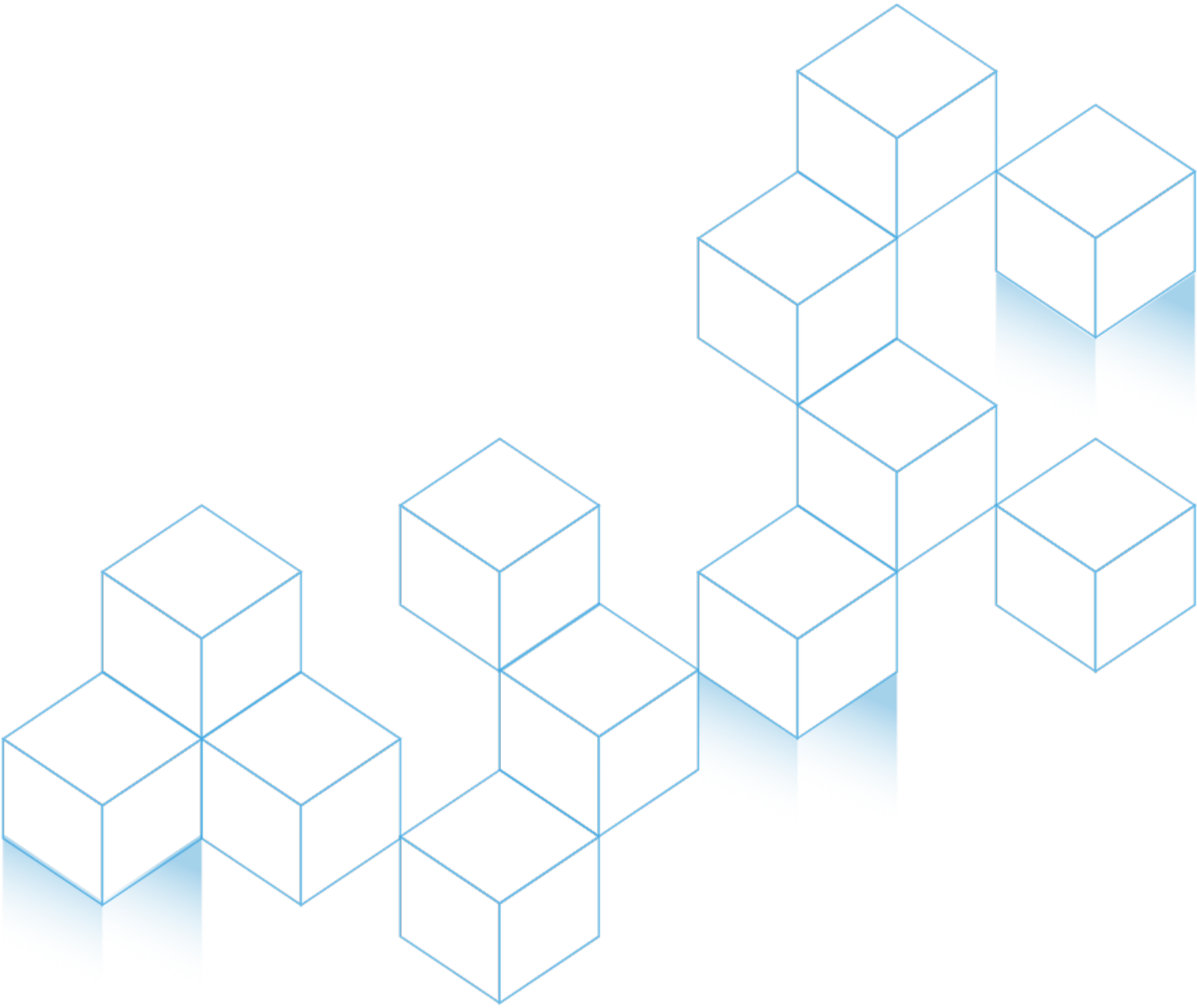
Skytap on Azure Storage Architecture and Performance Tuning for Skytap IBM Power Logical Partitions (LPARs)



September 2023

Introduction

This article discusses storage architecture for Skytap on Azure (SOA) in addition to common best practices for tuning IBM Power workloads that include IBM i, AIX, and Linux on Power workloads. One of the most overlooked variables in moving on-premises workloads to cloud-based solutions is the difference in storage architectures from traditional on-premises solutions.

Executive Summary

Skytap on Azure's storage architecture is designed to be cloud-native and multi-tenant from the ground up and carefully balances performance with high availability. SOA storage leverages native iSCSI for IBM Power and is built on a scaled-out architecture that leverages ZFS (formerly known as Zettabyte File System) for block-level storage. This architecture provides for unique capabilities that are truly unique to SOA.

Skytap has completed extensive performance testing and benchmarking on its infrastructure using standardized performance testing tools such as NDISK64 for AIX. For IBM i, we used in-house benchmarking tools that stress the Library structure of IBM i (QSYS.LIB) vs. testing the Integrated File System (IFS) which is found to deliver similar performance characteristics to AIX and Power Linux. Testing was done using a combination of 80% read and 20% write Input/Operations pers second (IOPS) that mirrors that of actual workloads we observe customers running in SOA today.

Our testing focused on understanding how to achieve the highest IOPS and data throughput while minimizing latency. The variables that have the biggest impact include block size, queue depth, and the number of disks and controllers assigned to an LPAR.

Skytap testing found:

- Queue Depth of 10 (AIX) is ideal for multi-disk workloads regardless of the number of disks, controllers, and block size.
- To achieve optimal performance, multiple disks and controllers are needed, but in most cases, only up to two are needed.
- For applications that use a smaller block size, and the number of IOPS is critical, 4K block sizes yielded the best performance.
- For applications that require high throughput, such as a data warehouse application, a 64K block size yielded the best performance.
- For applications that are sensitive to peaks in disk latency, a block size of 8K is ideal yielding both good throughput and high IOPS.

Storage Architecture in Skytap on Azure

Cloud-Based Power mission-critical workloads require highly available and performant storage. At the same time, these requirements must be balanced with those of a multi-tenant Cloud. Multi-Tenancy in Cloud provides greater economies of scale and workload portability than single tenant co-location and on-premises environments.

SOA provides this balance leveraging IBM's PowerVM Hypervisor with Skytap's own software defined storage platform. SOA provides a heterogeneous pool of Cloud-Based storage on self-encrypted SSDs that are made available to client LPARs using ZFS.

ZFS is unique because, unlike most other storage technologies, it combines both the roles of volume manager and the file system within SOA's storage subsystem. Therefore, it has complete knowledge of both the physical disks and volumes. SOAs implementation and use of ZFS provides capabilities that typical Enterprise and Cloud volume and file managers cannot achieve.

All of this is transparent to the LPAR as it sees standard SCSI disks made available to it from IBM PowerVM's Virtual Input Output Server (VIOS). Benefits of this architecture include:

- **Broad Filesystem Support:** Client LPARs can use any supported filesystem supported by its respective Operating System (OS) such as JFS, JFS2, Oracle ASM, Integrated File System (IFS), DB2, and others.
- **LPAR Portability:** LPARs don't have to "live" on a given Power Hosting Node (PHN), they can be moved to different PHNs at power cycle. If a PHN suffers a catastrophic failure, the LPAR can be restarted and automatically will be available on another PHN.
- **Shared Storage:** When disks are created as a disk set within SOA, that disk set can be made available to multiple LPARs within the same environment. This feature is commonly used to support Application and OS level clustering such as PowerHA, Oracle RAC, GPFS, and others.
- **Resiliency:** Physical double parity underpins powerful features of ZFS, called RAID-Z. Snapshots can be created of Power and x86 Filesystems on the fly, providing an instant backup of a single LPAR, multiple LPARs, or an entire Environment consisting of both LPARs and x86 Virtual Machines (VMs) in near real-time.
- **High Capability:** Skytap's use of ZFS and its Software Defined Storage (SDS) provides snapshots, copy-on-write, send-and-receive, hardware accelerated compression, and caching independent of the LPAR's filesystem.
- **Performance:** Depending on application type and configuration, several hundred thousand IOPS and more is made possible which is often a function of the capabilities of SOA's use of ZFS and its unique caching capabilities.
- **Caching:** Caching is an area of memory used to increase the performance of accessing data from storage. With caching in place, read and write requests from storage can be

provided much more quickly from memory vs. SSDs. Unlike standard Azure Premium Storage, Skytap on Azure caching is enabled for all disk volumes by default. Caching is discussed in greater detail in the next section.

Figure 1: Skytap on Azure Physical Storage Architecture for Power Workloads

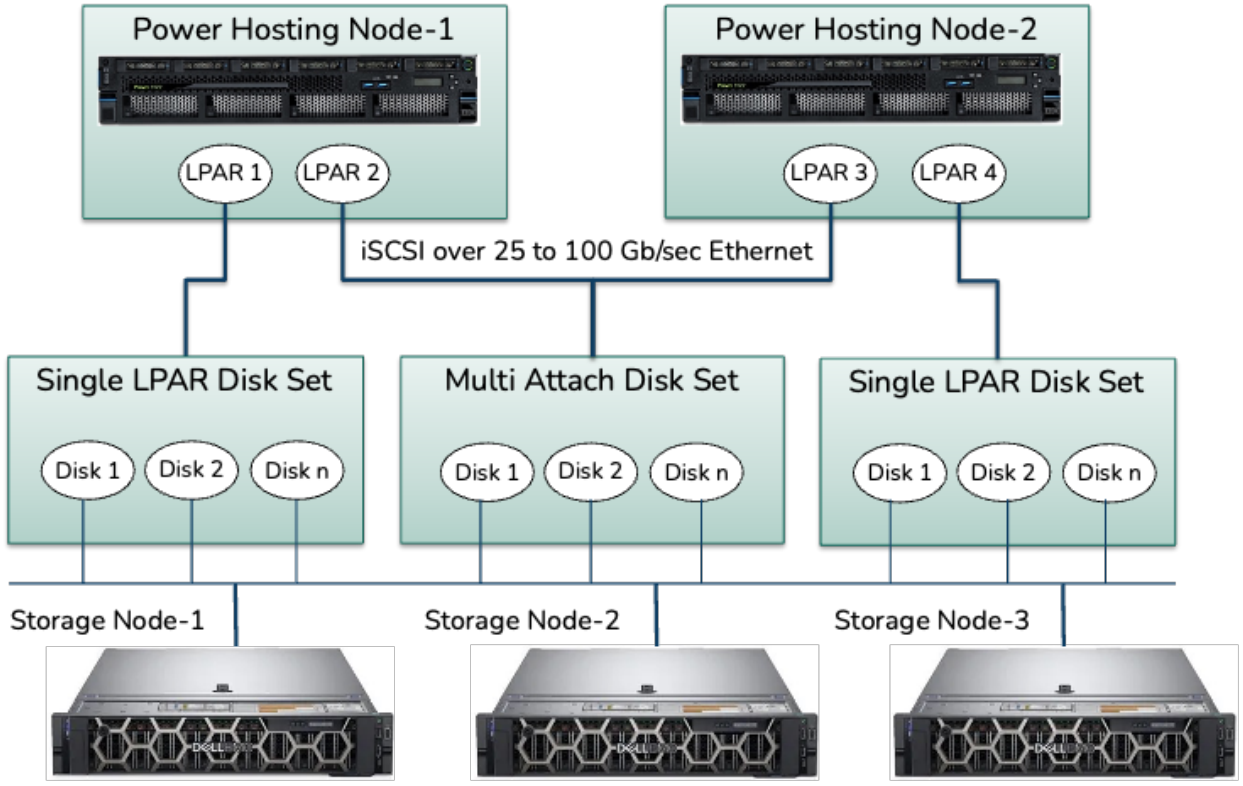
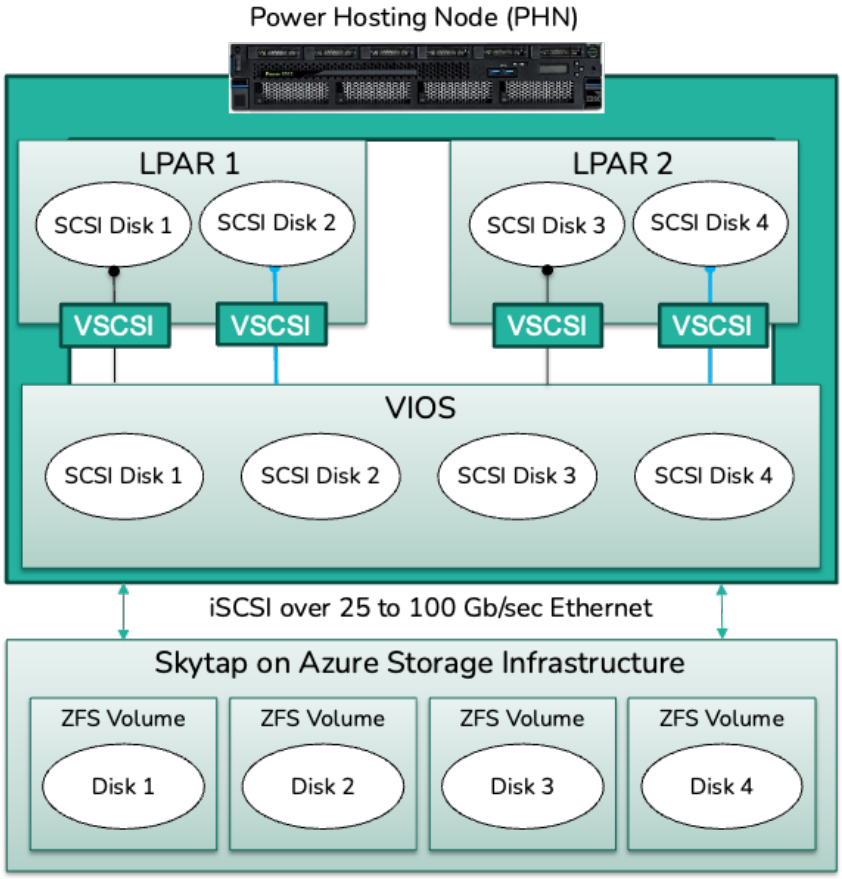


FIGURE 2: SKYTAP ON AZURE PHYSICAL TO VIRTUAL STORAGE ARCHITECTURE FOR POWER WORKLOADS



Caching

There are multiple layers of caching that occur within SOA:

- Client OS, ie. IBM i, AIX, Linux on Power
- VIOS
- Storage Node using ZFS caching
- Physical SSD

The most impactful caching comes from Skytap's implementation of ZFS and within the SN itself, as each is configured with a large memory cache that is used for Adaptive Replacement Cache (ARC).

ARC is a caching technology that combines the use of Least Recently Used (LRU) and Least Frequently Used (LFU) caches. The memory within a given SN is partitioned into two sections, one for each type. Furthermore, each of those caches maintains something called a "ghost list"

for each. When data is requested and is on the ghost list, but not in the respective cache, it is loaded into cache to increase the likelihood of a future hit. The result is a caching mechanism that is superior to most SANs and locally based storage subsystems.

Finally, unlike Azure Premium Storage for x86 workloads, caching is a part of all SOA storage, so all of this is done in the background and completely transparent.

Storage and Performance

There are several factors that determine how an application will perform from a storage perspective. This section describes the following variables that impact storage and application performance:

- CPU, Memory, and Network
- Disk IOPS (Input/Output Operations Per Second), throughput, latency

CPU, Memory, and Network

Before data reaches Skytap's storage subsystem, CPU, memory, and networking impact performance.

- **CPU**: If the processor is pegged, the LPAR is going to drive less storage IO and it may experience latency. This can be caused by setting an Entitled Capacity (EC) that is too low. Most applications take advantage of multi-threading that allows more parallel instructions to be processed. A multi-threaded application can handle storage latency more efficiently as when one thread is waiting for a response from storage, it can process another thread. To take advantage of multi-threading you need to ensure the appropriate number of Virtual CPUs are assigned. For IBM i LPARs, job priority will impact overall performance, so job priority needs to be tuned for each job and assigned to the correct system pool.
- **Memory**: Memory is critical for performance. An LPAR that is memory starved will generate IO from paging files that are detrimental to performance. Disk IO, like anything else, is a finite resource and using it in place of physical memory will bring any application to a halt.
- **Networking**: Workloads that are split across multiple systems require a fast network to take full advantage of CPU and Storage resources on a target system. Network performance and tuning within Skytap on Azure is discussed in TCP/IP Performance Tuning for Skytap on Azure IBM Power Logical Partitions (LPARs).

Disk IOPS (Input/Output Operations Per Second), throughput, latency

- **IOPS** are the number of requests that the client is sending to storage. An IO can be a read or write, and is either sequential or random. For x86 workloads, Skytap will set limits on each disk for a Virtual Machine and that is governed by the RAM you assign it. For Power workloads, there are no IOPS limits set by default. Although Skytap can apply IOPs limits at the disk level if the platform needs to govern excessive disk IO mitigating in cloud speak what is called a “noisy neighbor”.
- **Block Size** consists of several smaller pieces of data called bytes that are grouped together. Different applications, networks, disks all may use different block sizes. Older storage systems used a block size of 512 bytes, but today systems will use block sizes that can vary from 4K to 256K and some even larger. A block is the largest size of data that can be accessed in a single IO operation. When downstream systems use a block size that is smaller than the one that is received, it will be chunked into smaller sizes before the data is sent further down the IO chain.
- **Throughput, or bandwidth** is the amount of data that the client is sending to storage within an interval of time. Throughput is derived by taking this formula: IOPS x Block Size = Throughput. Applications such as data warehouses will use large block sizes, and they require a high amount of throughput.
- **Latency** is the amount of time it takes the client to receive a single IO request. Latency has a direct correlation to storage performance as the higher the latency, the lower the effective throughput of data delivered to the client. Latency also comes into play when we start discussing Queue Depth and the use of multi-tasking in your applications. As you tune block size and other aspects of your application, you must evaluate the latency created by those design decisions and make any necessary adjustments to ensure a performant application.
- **Multi-threading** is used to push more concurrent requests for compute and storage resources and when an application is multi-threaded and the LPAR is configured correctly, it can drive a higher throughput.
- **Queue Depth** is the number of outstanding IO requests an operating system can have for each Logical disk (LUN). Queue Depth and multi-threading are interrelated as a higher queue depth will allow a higher throughput up to a certain point. Setting a queue depth that is too high can result in overall latency that decreases performance.
- **The number of disks** on an LPAR allows for more data to be read and written at a given time and can provide a similar impact to overall performance as queue depth. The number of disks will often result in improved performance if the application workloads are evenly distributed across them.
- **The number of disk controllers** assigned to an LPAR allows a group of disks to be distributed and provides additional IO paths to virtual disks. Skytap allows up to 32 disks

per controller, and additional controllers are automatically added when 32+1 disks are added.[1]

Optimizing application performance for AIX and IBM i

The primary factors that impact application performance are block size, number of disks, number of controllers, multi-threading, and queue depth. Not all applications may allow you to control all aspects. For example, IBM i doesn't allow you to change queue depth and uses a default queue depth of 32. Oracle databases often are configured at a specific block size of 4K and it can't be changed unless you back up the database and restore it to a new one with the new block size.

Figure 3: Optimizing IOPS, throughput and latency for IBM Power workloads

Performance Factor	IOPS	Throughput	Latency
Block Size	Smaller block size yields higher IOPS	Larger block size will yield higher throughput	Larger block size will yield higher latency but higher throughput
Number of Disks	The higher number of disks, the higher the total IOPS	The higher number of disks, the higher the total throughput	Latency decreases with more disks up to a certain point
Number of Controllers	Adding additional disks with additional controllers will increase the total IOPS	Adding additional disks with additional controllers will increase the total throughput	Latency decreases with more disk controllers
Disk Striping	For IOPS to increase with multiple disks, workloads must be striped across those disks	For throughput to increase with multiple disks, workloads must be striped across those disks	Disk striping by itself does not increase latency to storage
Multi-threading	Multi-threaded workloads can drive higher IOPS when used in conjunction with additional disks and appropriate queue depth	Multi-threaded workloads can drive higher throughput when used in conjunction with additional disks and appropriate Queue Depth	Latency can increase with an application that has too many concurrent threads running
Queue Depth	Larger Queue Depth yields higher IOPS	Larger Queue Depth yields higher Throughput	Too low of a Queue Depth can cause high latency; however, a Queue Depth that is too high can increase latency as well

Understanding the IO requirements of your application is important and starts at the migration planning stage. SOA storage technology differs from on-premises. For example, on-premises Power workloads often used Fiber Channel based Storage Area Networks (SANs) vs. Skytap on Azure using iSCSI across a ZFS based storage subsystem using SSD JBODS (Just a bunch of disks).

- The biggest impact to overall throughput is block size. If an application requires large reads and writes of data, a larger block size should be used. However, if an application is sensitive to latency, a smaller block size is best. Block size is the biggest determining factor in overall throughput.
- AIX defaults to a Queue Depth of 3 per disk on iSCSI, which is often too low for Oracle Database Applications. Setting a Queue Depth higher will allow for higher concurrency, especially when Oracle ASM is used as a filesystem. Increasing Queue Depth between 8 to 12 will yield improved results, but setting it above this level may increase overall latency and become a zero-sum gain. For IBM i, queue depth cannot be changed and is set to 32 by the OS.
- Increasing the number of disks will spread IOs across multiple disks increasing both IOPS and throughput, but if the workload isn't distributed across those disks, "hot disks" can result, having a negative impact as to application performance. Furthermore, if it is decided to increase the queue depth of an LPAR, the same queue depth needs to be set for each disk. Since IBM i operates with a fixed queue depth, increasing disks is one of the most effective means of increasing IO concurrency.

Figure 4 below summarizes the best performing configurations by block size. These tests were done using NDISK64 running on AIX with data evenly distributed across all disks. A read/write ratio of 80%/20% was used as that is common in many databases. It is important to note that we look at the average **peak** latency as a relationship to block size instead of average latency. The reason is that average latency tends to be in the range of .2ms to .5ms, and it is the peak latencies that increase when the LPAR is under heavy IO load that causes application issues.

Figure 4: Skytap on Azure Power Storage Performance for IBM Power workloads by Block Size

Block Size	Controllers/Disks	Queue Depth	IOPS[2]	Throughput	Avg Peak Latency
4K	2/32	10	134,000	536 MB/sec	1.18ms
8K	2/32	10	76,400	610 MB/sec	1.36ms
16K	2/32	10	44,000	705 MB/sec	1.98ms
64K	2/32	10	19,200	1230 MB/sec	3.2ms
128K	2/32	10	6,800	872 MB/sec	5.2ms

Conclusion

In all cases, a Queue Depth of 10 provided the best performance, and it's important to note that the queue depth should be set to all disks. In the end, performance is ultimately going to be a factor of the workload, but from Skytap's synthetic tests we can conclude the following:

- Queue Depth of 10 (AIX) is ideal for multi-disk workloads regardless as the number of disks, controllers, and block size.
- To achieve optimal performance, multiple disks and controllers are needed, but too many can be detrimental.
- For applications that use a smaller block size, 4K block sizes yielded the best performance.
- For applications that require throughput, a 64K block size yielded the best performance.
- For applications that are sensitive to peaks in disk latency, a block size of 8K is ideal yielding both good throughput and high IOPS.

Summary

Skytap on Azure provides a robust storage architecture designed and optimized for cloud-based IBM Power workloads. When a user is looking to migrate workloads to the Cloud, an understanding of Skytap on Azure's storage architecture and performance characteristics is critical in the design and deployment of applications. Administrators and Architects must determine how to achieve the optimal Input/Output per second (IOPS) and data throughput requirements of their applications while minimizing latency with their decisions.

[1] Adding additional disks above 32 and increasing the number of Controllers may require a disk policy to be applied to your LPAR. This an be done by contacting Skytap support and support@skytap.com

[2] IOPS and Throughput results are reported from the client LPAR