architectingit

Performance Benchmarking Volumez Data Infrastructure in the Public Cloud

EBOOK CATEGORY: TECHNICAL EVALUATION

Release 1.0
May 2024

Performance Benchmarking Volumez 2024 First Edition

Published by Brookend Limited.

Document reference number BRKSW0162-01.

No guarantees or warranties are provided regarding the accuracy, reliability or usability of any information contained within this document and readers are recommended to validate any statements or other representations made for validity.

Copyright © 2024 Brookend Ltd. All rights reserved. No portions of this document may be reproduced without the prior written consent of Brookend Ltd. Details are subject to change without notice. All brands and trademarks of the respective owners are recognised as such.

CONTENTS

How to Use this eBook

This eBook has been independently researched, written, and published by Architecting IT, a trading brand of Brookend Ltd, based in the UK. It contains graphs, tables, and associated data.

Please note, all content within this eBook, including graphics and charts, are copyright of the author and Brookend Limited, or their original owners. They may not be reproduced in part or whole, without express written permission. If you wish to licence this content, please contact us via email at vendorsupport@architecting.it.

Executive Summary

For many IT organisations, the public cloud is the primary location for the deployment of new applications. As the use of these platforms matures, customers want to ensure that infrastructure consumption is fully optimised. Cloud vendors provide native storage solutions but do not offer dynamic optimisation of capacity, throughput and IOPS across an aggregate of virtual instances or containers.

Volumez has developed a SaaS-based platform that orchestrates local cloud-based instance storage to create a virtual data infrastructure plane that provides NVMe storage to virtual instance or container-based applications. The Volumez solution aggregates storage from multiple virtual instances, turning it into a resilient data infrastructure where the customer can choose guaranteed delivery of latency, IOPS and throughput metrics per application instance using pre-defined or customisable policies.

In the public cloud, the customer pays for resources as an operational expense, so it is vital to ensure the capability of all virtual instances are fully exploited. This requirement also extends to distributed storage solutions deployed on public cloud infrastructure, including that from Volumez.

In this paper, we test the claims from Volumez that its distributed data plane technology can be used to efficiently aggregate ephemeral instance storage to deliver a resilient data infrastructure. We validate the capability to deliver sustained, guaranteed performance at sub-millisecond latencies, across a mix of workload types and data profiles.

The results show that the Volumez platform successfully abstracts the physical aspects of instance storage from the application requirements of IOPS, bandwidth and latency. In tandem, the storage and network resources are delivered with minimal overhead to form an optimized, balanced, and linearly scalable data plane -- taking advantage of the full performance capabilities of the physical infrastructure of the public cloud provider.

Introduction

Two decades ago, storage networking emerged as a new technology that enabled customers to pool and consolidate storage media resources into resilient storage appliances. The benefits to the business were both operational and financial. SANs (storage area networks) remove the fragmentation of storage deployed across hundreds or thousands of servers, while improving uptime and resiliency. This technology continues to be used today, even with virtual machines and containerised applications.

In many ways, the public cloud emulates much of the basic technology normally deployed on-premises. Virtual instances are analogous to virtual machines, we have virtual networking and of course, virtual storage to emulate the functionality of storage area networks.

Cloud service providers, such as Amazon Web Services (AWS) and Microsoft Azure (Azure) offer resilient networked storage that may be assigned directly to a virtual instance, either as a boot disk or for application storage. Solutions like AWS EBS (Elastic Block Store) and Azure Managed Disks enable the customer to dial in application-specific metrics, including IOPS and bandwidth per volume.

Unfortunately, while public cloud storage does a great job at providing a high quality of storage capability and performance, cloud storage is not as flexible as the SANs they replaced. In the on-premises world, customers benefit from thin provisioning and the shared nature of multi-tenant storage, such as high durability (typically six 9's) and the capability to dynamically optimise IOPS and bandwidth per volume, independent of the storage deployed.

In the public cloud, durability is charged at a premium, as is performance. In each case, the customer must choose a volume-specific assignment that doesn't benefit from the ability to pool resources. Efficiency gains (such as thin provisioning) are retained by the service provider. This scenario can lead to significant overprovisioning and cost, as each application deploys an expected "high watermark" of capacity and performance needs. Although the public cloud does allow some flexibility to change storage resources once deployed, scaling down, for example, isn't easy to achieve and not without potential outages.

The response from the storage industry has been to deploy virtual SANs in the public cloud using the same technology stack and storage controller architecture developed for on-premises SANs. While this lift-and-shift approach may seem excessive, with large-scale cloud deployments there are many potential benefits. In contrast, all public cloud service providers now offer instances with locally attached high-performance NVMe storage. Instance networking is generally a minimum of 10GbE, with 25GbE and 40GbE available. The latest Linux kernel enables native NVMe over Fabrics (NVMe-oF) support, specifically including NVMe/TCP. As a result, it is now possible to deploy a high-performance, scalable virtual distributed storage data plane based on public cloud instances, while optimising the use of any configured resources which the public cloud has to offer.

In this report, we will look at the Volumez cloud orchestration platform and briefly explain how it differs from the concept of a "SAN in the cloud". We validate the performance capabilities of the solution as a way to optimise public cloud block storage. The tests presented will validate Volumez' claims on guaranteed performance, throughput, latency, and scale. The testing has been performed independently, with Volumez providing a test environment.

Volumez

Volumez is a solution for deploying tailored data infrastructure in the public cloud. It operates as a cloud orchestration platform, transforming public cloud instances into a distributed storage data plane to be consumed by other virtual instance applications. Resources are dedicated to delivering storage capacity and performance, rather than mixing storage with application compute. The Volumez platform abstracts away the implementation details of cloud storage, replacing it with the capability to specify bandwidth, throughput (IOPS) and latency per storage volume, irrespective of volume capacity and based on dynamic policies set by the administrator (most public cloud storage aligns capacity and IOPS in a linear scale).

Storage is consumed by application virtual instances or containers as local NVMe devices that look identical to the block storage provided by the native platform but are delivered by NVMe-oF (in the tests performed in this analysis, the connectivity was NVMe/TCP). Volumez also supports Kubernetes, with a CSI plugin that enables volumes to be consumed by containerised applications.

While a full breakdown of the capabilities of the Volumez cloud orchestration platform are outside the scope of this report, there are some features that need to be discussed as they are relevant to the performance analysis work in this document.

- Volumez consumes ephemeral instance storage on cloud virtual instances to create a resilient, persistent storage data plane. Ephemeral local storage typically operates with latencies in tens of microseconds (much lower than other cloud block storage) but is lost when an instance is stopped, hibernated, or terminated (cloud platform dependent).
- The Volumez solution is orchestrated through a SaaS cloud portal, where the user specifies pools of nodes that provide storage resources to applications. The Volumez software provides the data resiliency, RAID rebuilds and data layout to optimise for host performance.
- A Volumez customer defines performance requirements through policies that specify the maximum throughput in terms of bandwidth and IOPS. Policies are then assigned to volumes, which in turn are mounted on application instances.

In this analysis, we look at the performance claims of the Volumez platform. In particular, we examine four aspects.

- The capability to deliver volume-specific performance for IOPS and bandwidth.
- The capability to deliver sustained performance at latency levels less than 1millisecond.
- The capability to scale performance to the demands of an application.
- The capability to deliver performance in Kubernetes environments.

In addition, we will examine the deterministic nature of the I/O delivered to applications. By this we mean how faithfully the platform delivers the dialled in IOPS and bandwidth settings, including the degree of deviation from the expected performance levels.

Performance Testing

In the public cloud, just as on-premises, storage I/O performance has a direct impact on application performance. Recent innovations in NAND flash media and new protocols such as NVMe have provided the building blocks to create the next generation of storage networking solutions. However, when measuring performance, the basics still apply. Persistent storage should deliver a high degree of efficiency, exploiting all of the available resources. This requirement is essential for public cloud deployments where the customer is charged for every aspect of the services delivered. Efficient resource usage is becoming increasingly harder to achieve for storage attached to single instances, making pooling the logical way forward.

Metrics

To validate the Volumez performance claims, we use three standard metrics that are widely used across the industry.

- **Latency** – a measure of the time taken to complete a single I/O operation, measured from the perspective of the application. Latency is usually quoted in milliseconds or microseconds. Lower figures are better. Latency directly affects the performance of applications where transactional consistency is important, such as structured databases.
- **IOPS** – the capability of a storage system to process I/O requests, or I/Os Per Second. Higher IOPS values are better. IOPS are important for streaming applications or those processing large volumes of data.
- **Bandwidth** – the amount of data a device or system can transfer over any given period of time, typically measured in MB/s (megabytes per second) or GB/s (gigabytes per second). Larger values are better. Bandwidth is important for applications that need to process large quantities of data, including multi-threaded tasks. Bandwidth is sometimes referred to as throughput.

In the tests performed in this analysis, we will measure IOPS and bandwidth of the Volumez solution, then optimise the configuration for latency.

Test Methodology

Architecting IT has developed a series of synthetic benchmarking tests based on typical workload profiles using the industry-standard *fio* tool. These have been divided into four groups.

- **Random Read** – small block size random reads, typical of OLTP and some AI/analytics tasks.
- **Random Write** - small block size random writes, typical of OLTP and some AI/analytics tasks.
- **Database** – mixed workload profile representing common applications such as databases where the I/O is a mixed percentage of read and write activity (generally 70/30 read/write split).
- **Balanced** – throughput-based profiles for sequential reading and writing of data, as seen in HPC applications.

Across each of these workload types, multiple tests were performed with a range of configuration parameters that varied I/O queue depth and block size. These are 4KB, 8KB, 64KB and 128KB block sizes, with four queue depth settings.

- QD=1 – measures the baseline latency expected from each I/O.
- QD=32 – measures the maximum throughput expected from each test.
- QD=1, 32 Jobs – measures baseline latency, but loads the available CPUs in the test environment.
- QD=80% - variable queue depth and job count to achieve 80% of maximum IOPS. Used to demonstrate consistent latency under load.

The script used to execute the tests runs each of the configurations sequentially with the following parameters.

- Ramp up time of 10 seconds, to enable the workload to stabilize.
- Run time of 120 seconds to ensure workloads are stable.
- Run at least 64 tasks in parallel to ensure sufficient workload is generated across the storage.

Test Bed Configuration

The test bed setup is deployed on AWS EC2 with the following specifications.

- RAID-1 storage configuration
- Storage Nodes – 16 (sixteen) i4i.4xlarge
- Application Node – 1 (one) c6in.32xlarge
- Test Volume – 1TiB

The storage nodes run the Volumez software, while the application node is used as the load generation virtual instance, running *fio*. Both storage and application nodes are connected through the standard client network offered by the cloud provider. Storage devices are exposed to the application node using NVMe/TCP.

The i4i nodes provide 16 vCPUs, 128GB of system memory, 25Gb/s networking and a single 3750GB Nitro NVMe SSD. The Volumez platform is used to aggregate the Nitro SSD volumes to deliver resilient storage, which would otherwise be lost when an i4i instance with Nitro SSDs is terminated.

Although the storage nodes deliver a high degree of performance capability, a performance policy is in place for the storage exported to the application node. This is configured with the following parameters.

- Write bandwidth – 6000MiB/s
- Read bandwidth – 12,000MiB/s
- Write IOPS – 1,200,000
- Read IOPS – 2,400,000
- Write Latency – 500µs
- Read Latency – 500µs

The aim of using a performance policy is to demonstrate the deterministic nature of the Volumez solution, to align with the demands of multiple applications using the same underlying storage data plane.

Test Results

This section of the document details the results of the performance testing, displayed as graphs, with additional commentary that explains the expected and actual outcomes.

Random Read

This test runs a set of 100% random read I/O across four block sizes and varying queue depths. These are queue depth=1 (QD1), queue depth=1 with 32 parallel tasks (QD1x32), queue depth=32 and queue depth set to reach 80% of maximum throughput (80PC).

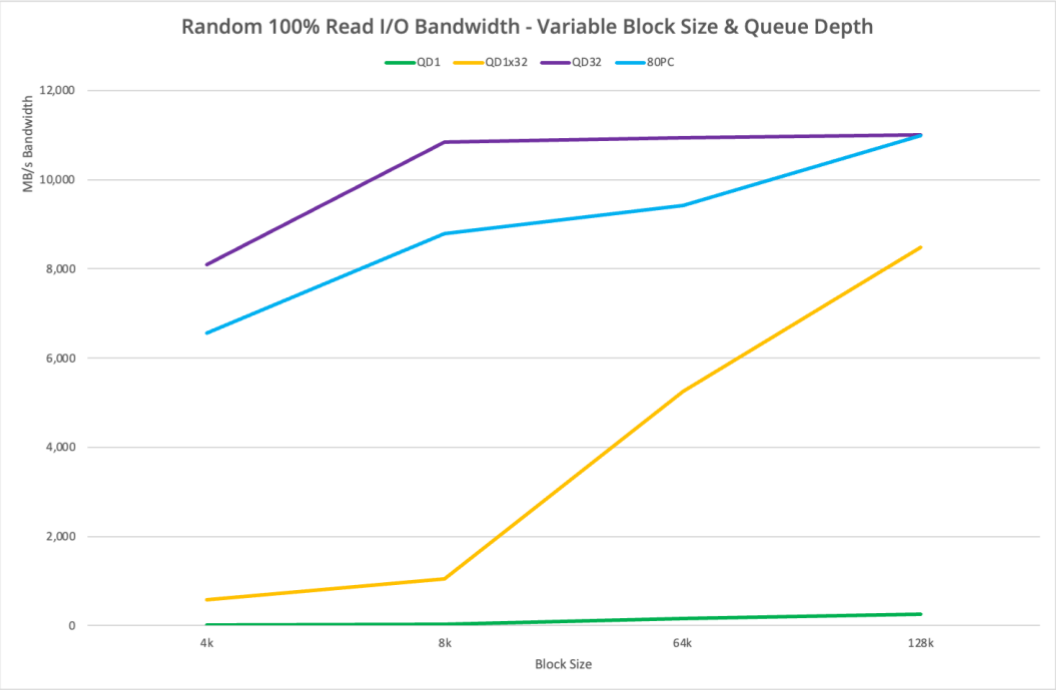

Figure 1 - Random 100% Read I/O (Bandwidth)

Looking first at Figure 1, as the block size increases, we should expect to see a corresponding increase in bandwidth when the queue depth fully loads the system. As we observe, at QD1, the pipeline of I/O isn't fully loaded, and the test environment isn't pushed to its limit. However, with QD1x32, the additional parallel tasks exploit the available resources and deliver greater throughput.

With a queue depth of 32 (QD32), the pipeline of I/O is "loaded" and driven to the maximum available throughput with a block size of 8KB and higher. The test has reached peak performance level and any increase in block size or queue depth would not see any improvement in throughput.

For the 80% test, we expect to see increasing throughput to a point where the maximum is reached. This occurs with a block size of 128KB. To understand the value of this test, we need to look at the latency results shown in Figure 2. Here we can see (logarithmic scale) that the QD1, QD1x32 and 80PC tests all deliver latency less than 1ms (1000 microseconds) and were measured in hundreds of microseconds. When running at 80% of the theoretical throughput maximum, latency was still maintained at less than 1ms. Compare this to the QD32 latency, where the latency increases as the test tries to push more I/O through than the backend storage can support.

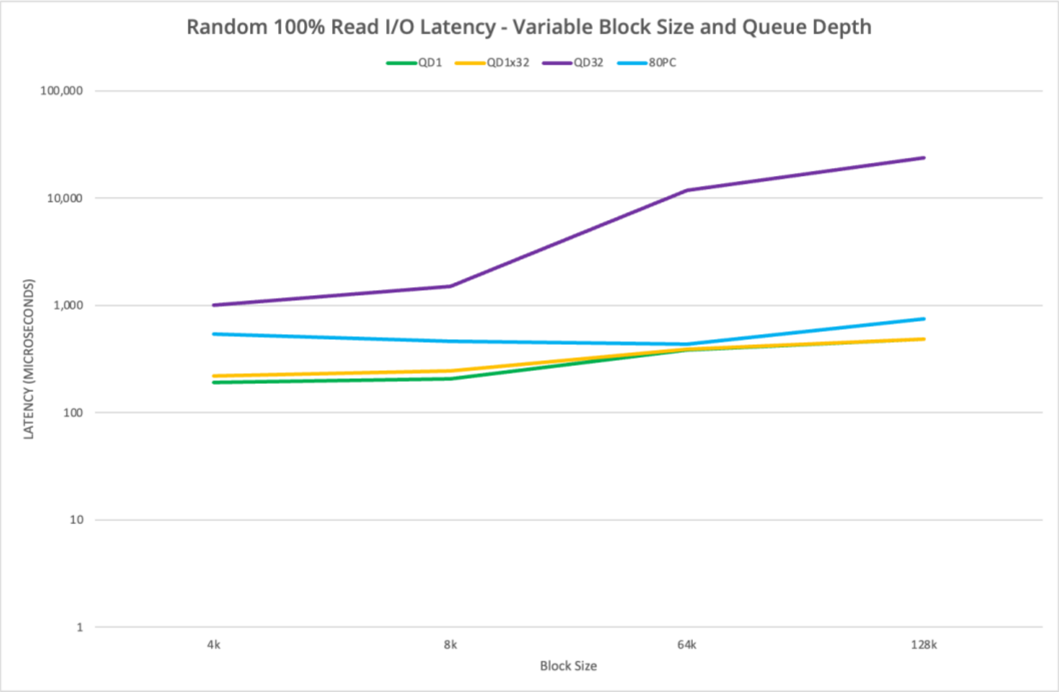

Figure 2 - Random 100% Read I/O (Latency)

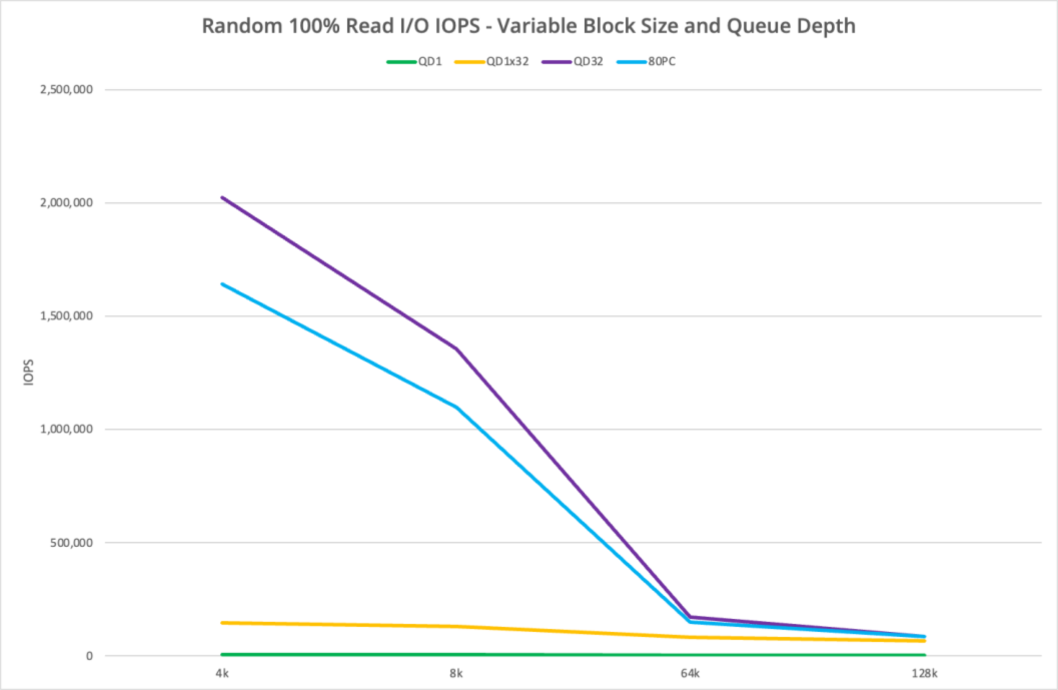

Figure 3 - Random 100% Read I/O (IOPS)

Finally, as we look at the IOPS, we expect to see a declining IOPS value as the block size increases. This is because there is a maximum throughput capability, which for the QD32 is the obvious example. The QD1 and QD1x32 tests aren't constrained by available resources, so (as expected) show a flat IOPS line. The 80% test, restricted by throughput, shows declining IOPS, but at a rate aligning to 80% of the maximum (QD32), until the optimum value for all the tests is reached at 128KB.

One final point to note. In many similar vendor tests, the throughput is stress tested using sequential I/O. In this test we were able to maximise the available resources with random I/O, demonstrating a high degree of parallelism in the Volumez solution.

Random Write

This test runs a set of 100% random write I/O across four block sizes and varying queue depths. These are queue depth=1 (QD1), queue depth=1 with 32 parallel tasks (QD1x32), queue depth=32 and queue depth set to reach 80% of maximum throughput (80PC).

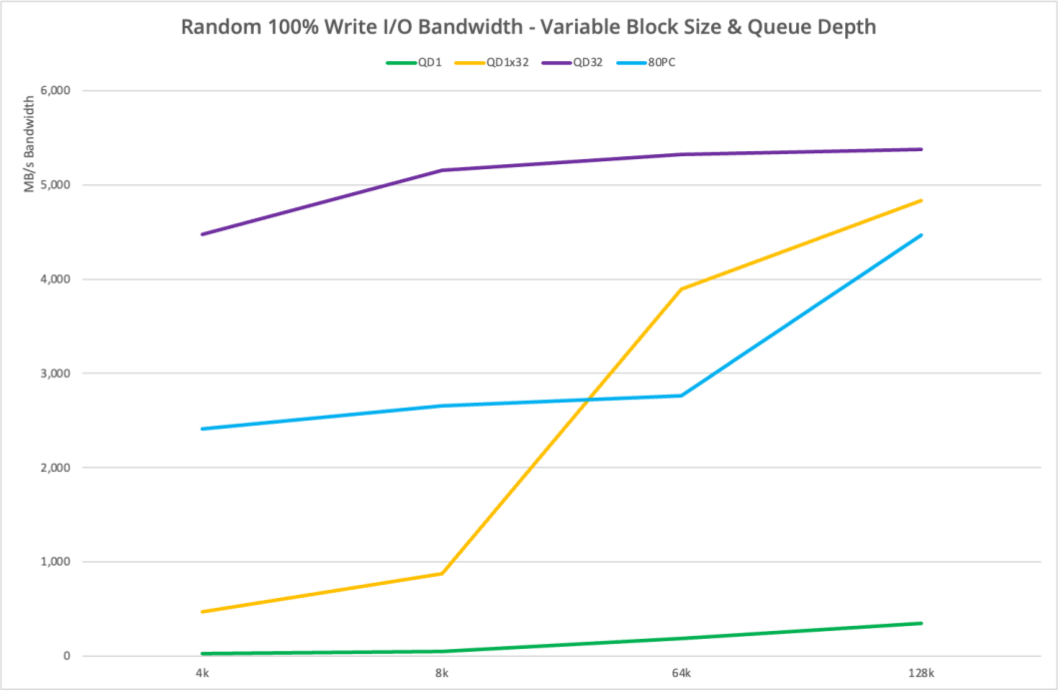

Figure 4 - Random 100% Write (Bandwidth)

In a similar trend to the random write I/O graphs, we should expect to see Increasing bandwidth from the tests that stress the infrastructure. Specifically, in Figure 4 we see that the QD32 and QD1x32 tests increase towards a maximum throughput of around 5000MB/s. The QD32 quickly reaches the available bandwidth set in the test at the 8KB block size. The 80% and 1xQD32 tests each similar maximums at 128KB.

Note that the QD1 test doesn't stress the available resources, as this test must wait for the completion of each I/O before submitting the next. However, looking at the impact on latency (Figure 5) we can see the QD1 test has consistent latency well below the 1ms threshold. Similarly, as we saw in the read I/O test, the latency figures for the 1xQD32 and 80% tests are well below the 1ms threshold.

The QD32 test, which will push the maximum throughput to the limit, starts to see increasing latency as the configured limit on throughput is reached at 8KB queue depth. The latency continues to increase (as expected) with larger block sizes, due to the aggregate limit on throughput available.

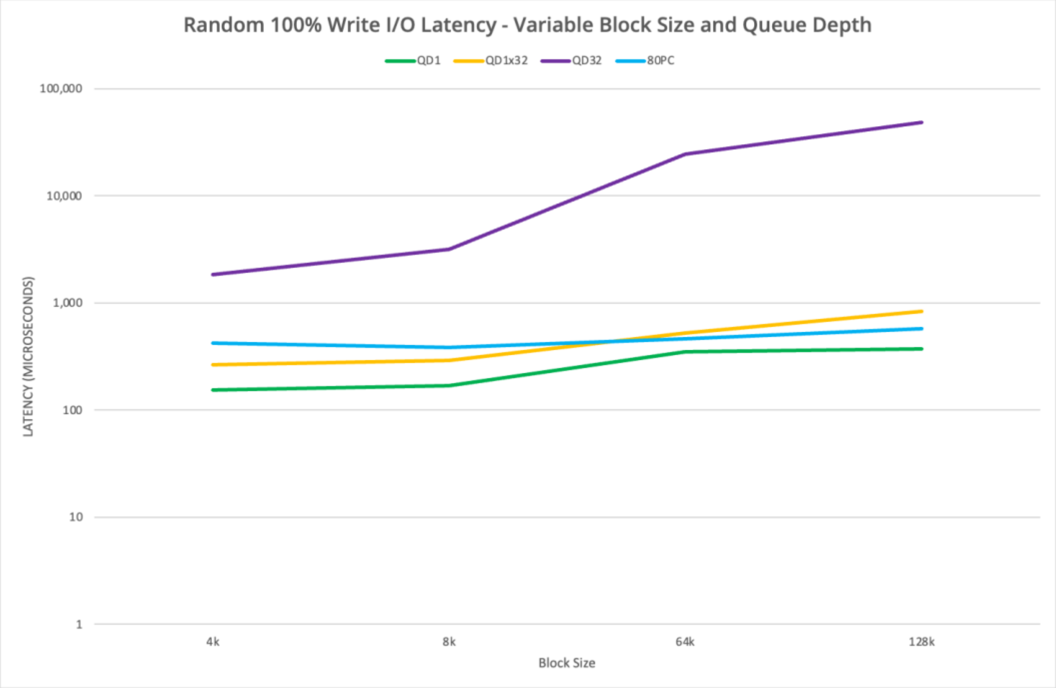

Figure 5 - Random 100% Write I/O (Latency)

Looking at the IOPS data shown in Figure 6, we expect to see declining IOPS from a maximum when the block size is at its smallest. We can see that this does indeed occur, with the IOPS count reducing as the block sizes increase correspondingly. This demonstrates that the test has reached the maximum throughput available for the QD32 and 80% tests, while the QD1 and QD1x32 tests are not constrained and therefore do not see the equivalent reduction in IOPS.

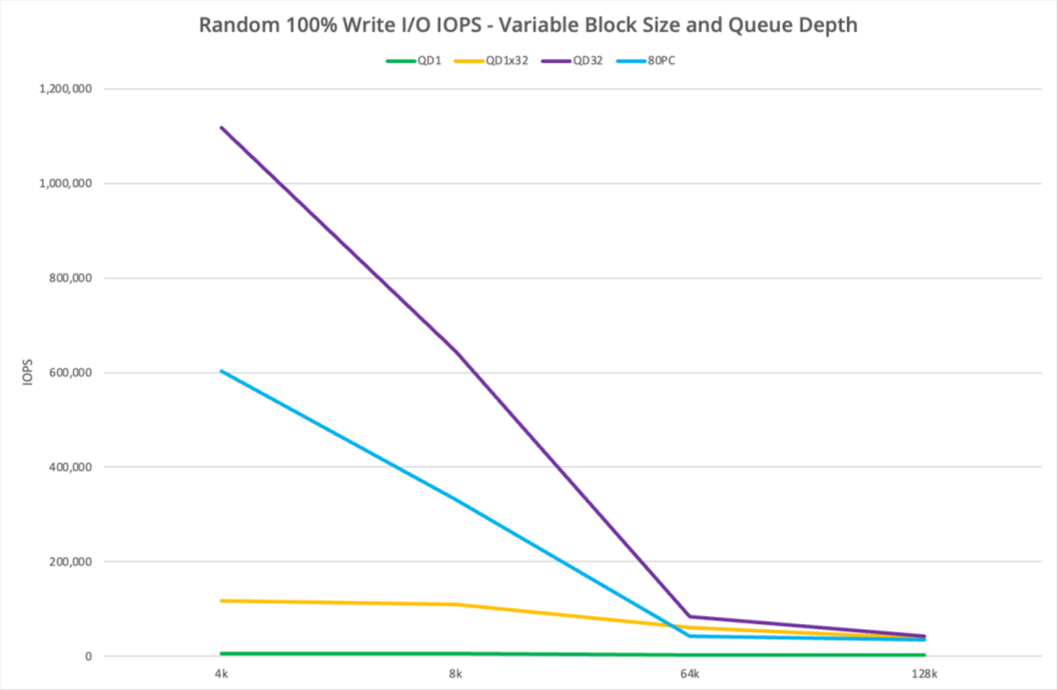

Figure 6 - Random 100% Write I/O (IOPS)

Database Mixed Read/Write

This test runs a set of mixed read/write (70%/30%) I/O across four block sizes and varying queue depths. These are queue depth=1 (QD1), queue depth=1 with 32 parallel tasks (QD1x32), queue depth=32 and queue depth set to reach 80% of maximum throughput (80PC). Read and Write results are shown separately and indicated with (R) and (W) respectively.

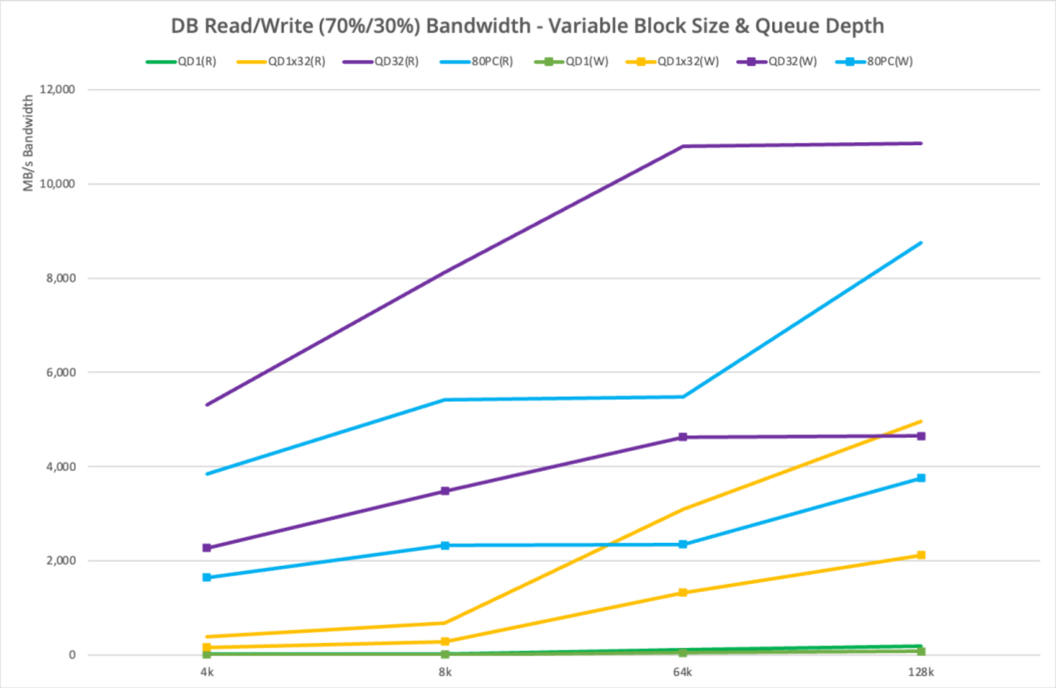

Figure 7 - Database Read/Write (70/30) Bandwidth

We expect the graphs to trend upwards towards the maximum configured bandwidth, as the block size increases. We also expect read I/O to show higher numbers than write I/O, partly because the test is configured to favour more reads than writes but also because NAND flash latency for writes is always higher than for reads.

The test shows the QD32 results trending upwards and reaching the configured maximum values, hitting the threshold at 64KB block size. The QD1 tests are unconstrained and don't load the “pipeline” of I/O sufficiently to see any substantial increase in bandwidth. The QD1x32 test shows an increase, while the 80% test also trends upwards for both read and write I/O.

It is important to note that this is a mixed workload test where we have separated out the read and write I/O for the purposes of the presentation. Despite this, the throughput capabilities are only marginally affected compared to the 100% read/write tests.

Looking at latency, the tests all performed at less than 1ms response time, except for the QD32 tests. We expect those to see an increase in value for latency as the tests are constrained by the throughput quality-of-service figures assigned in the Volumez platform configuration.

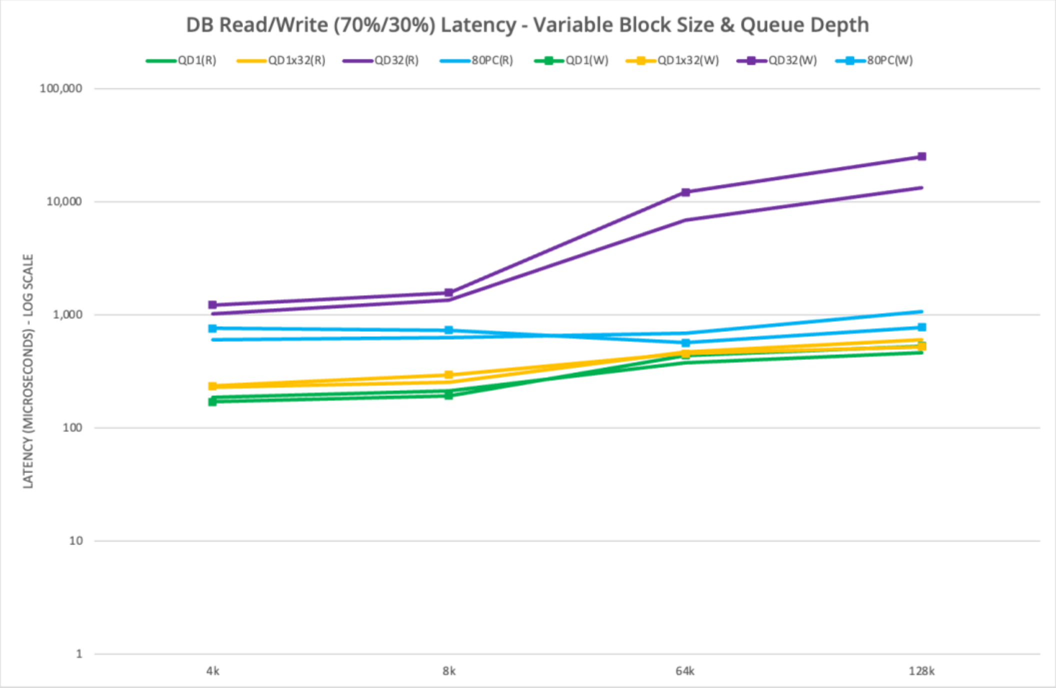

Figure 8 - Database Read/Write (70/30) Latency

As we look at the IOPS graph, we see a declining IOPS total as the block size increases, in line with what we've seen already in the previous 100% tests. However, the drop-off is not as pronounced as with the 100% read/write tests, because the maximum bandwidth wasn't reached until higher values of block size.

One final observation on this test shows that despite restricting the 80PC test to 80% of the bandwidth, the IOPS values are comparable to the unconstrained QD32 test.

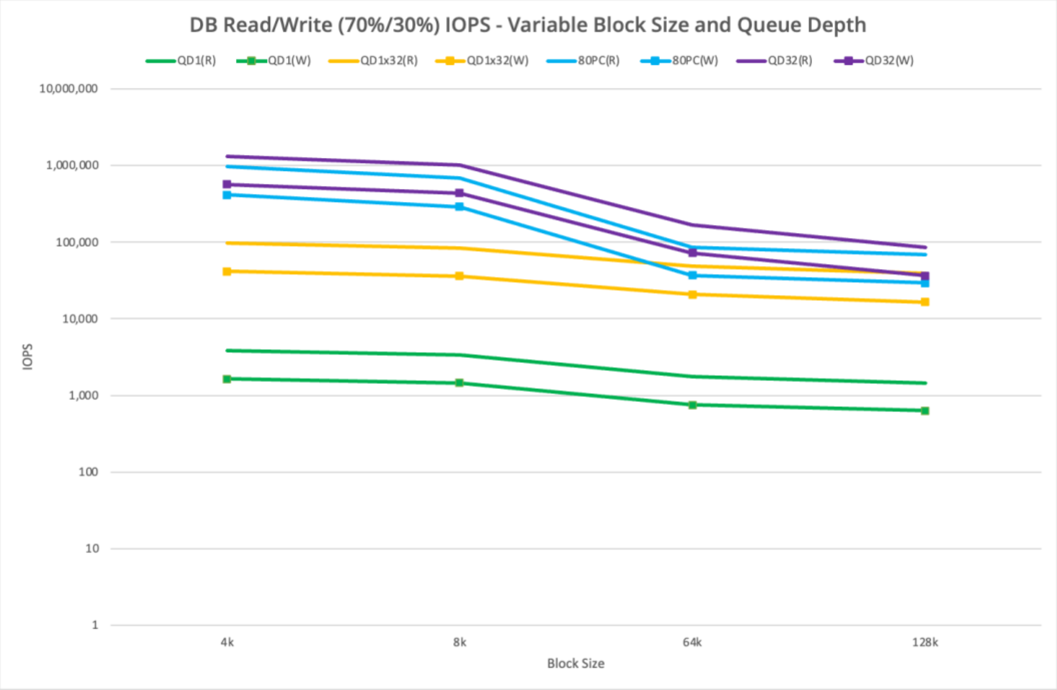

Figure 9 - Database Read/Write (70/30) IOPS

Balanced Mixed Read/Write

This test runs a set of mixed read/write (50%/50%) I/O across four block sizes and varying queue depths. These are queue depth=1 (QD1), queue depth=1 with 32 parallel tasks (QD1x32), queue depth=32 and queue depth set to reach 80% of maximum throughput (80PC). Read and Write results are shown separately and indicated with (R) and (W) respectively.

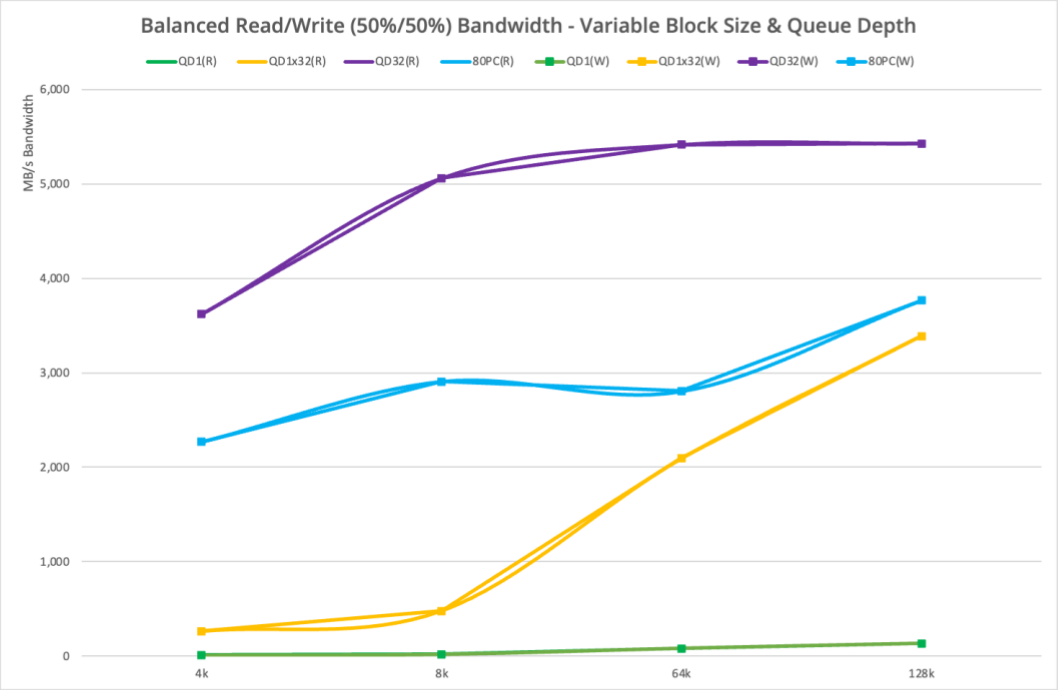

Figure 10 - Balanced Read/Write (50/50) Bandwidth

For the data shown in Figure 10, we've chosen to show the write I/O data as a smoothed curve with block markers, while the read I/O is the standard straight-line graph. This has been done because the data values were so similar that two straight-line graphs would not be clear enough to visualise separately.

We can see that at 50%/50% read/write, the QD32 test reaches the maximum constrained limit at 64KB. The 80% test and QD1x32 both trend towards the maximum and would have intersected at a larger block size. The QD1 data is unconstrained and doesn't generate enough pipeline work to increase throughput with increased block size.

Why does the QD32 test not show higher read performance compared to the 70/30 test? One reason could be the impact of equal reads and writes on the backend NAND flash media, which needs to aggregate and consolidate write I/O in cache. It could be that cache on the media itself is constrained for read caching, while some read activity could be delayed waiting for write I/O to complete.

Looking at the latency data shown in Figure 11, we see similar behaviour to the 70/30 test where the QD1, QD1x32 and 80PC tests follow similar latency figures for read and write I/O. The QD32 data differs as writes generally have a longer latency than reads on flash media.

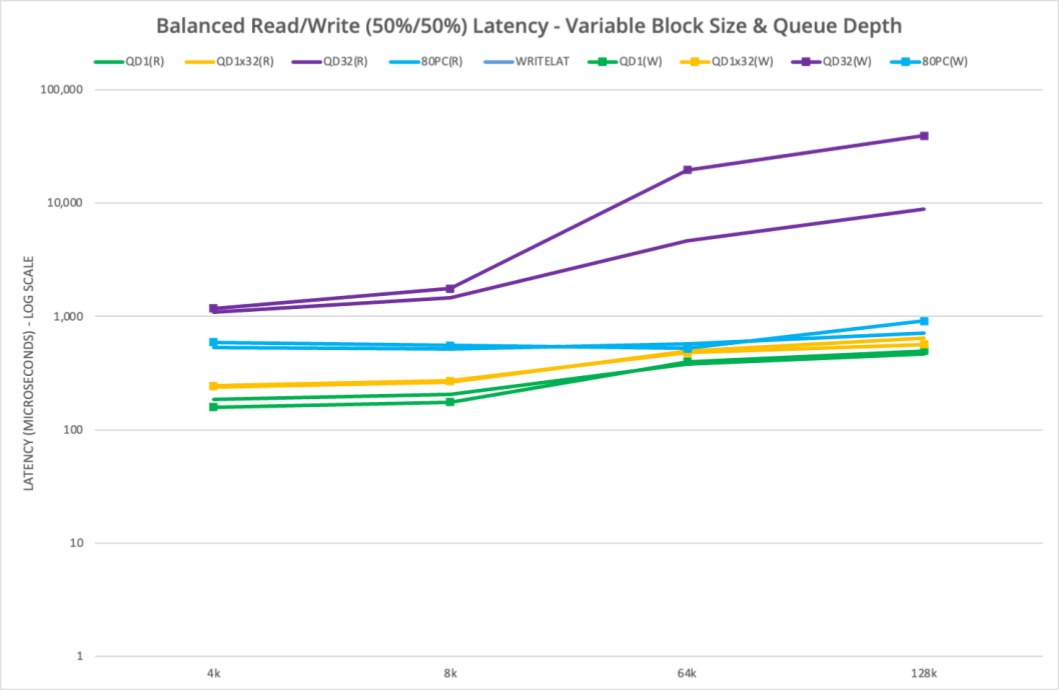

Figure 11 - Balanced Read/Write (50/50) Latency

In presenting the IOPS data, we have again chosen to show the write I/O using a smoothed line and the read I/O with straight lines. The data shows consistent performance for both read and write I/O, as we would expect from the throughput numbers. In addition, the IOPS decline in line with the increase in block size, because the maximum bandwidth capable is static.

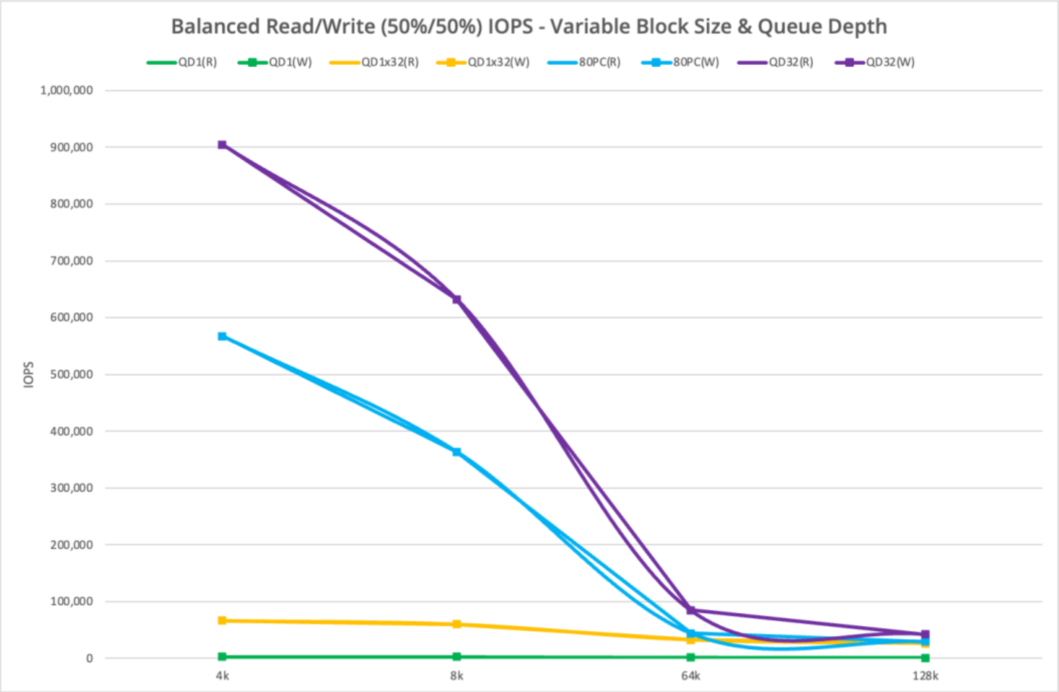

Figure 12 - Balanced Read/Write (50/50) IOPS

Observations

The data presented in the eight graphs shows bandwidth and latency for four tests, comprising random read, random write, database, and balanced profiles. The aim of these tests is two-fold. Firstly, to demonstrate that the Volumez platform can achieve the configured bandwidth and secondly that the I/O delivered is within acceptable latencies for modern NVMe-based storage.

Random Read – In this test (Figure 1 and Figure 2) the highest bandwidth is achieved with a queue depth of 32 and from 8KB I/O upwards. The first of the two graphs shows this peak reached with the 80% test at 128KB. In addition, the QD1 with 32 processes trends towards the maximum, demonstrating that I/O performance scales linearly. The latency graph, shown in a logarithmic scale for more clarity, shows each test achieving under 1 millisecond response time (with actual figures under 500 microseconds). Only the unconstrained queue depth of 132 increases latency with increased block size, due to the inevitable queuing of I/O.

Random Write – In this test (Figure 4 and Figure 5) the highest bandwidth (approximately 5GB/s) is achieved with a queue depth of 32 and from 8KB block size and upwards. The 80% test trends towards the maximum (although is deliberately constrained to highlight consistent latency), while the QD1 with 32 processes also trends towards the maximum. The latency graph shows all I/O except the QD32 test delivered sub-millisecond I/O responses, with the majority of the I/O delivered under 500 microseconds. Only the QD32 test exceeds the configured threshold due to the volume of I/O and block sizes used (this is a measure of back-end congestion on the media, not the Volumez software).

Database Read/Write – In these tests (Figure 7 and Figure 8) the graphs show data for both the read and write components separately. We've highlighted the write I/O component in the same bar colour but using square markers. In the Bandwidth test, the data shows throughput for read I/O reaching the 11GB/s maximum, while write I/O peaks at 5GB/s (both the configured values). The latency graph shows all I/O achieving sub-millisecond response times, with the QD1 tests both delivering latency under 600 microseconds.

Balanced Read/Write – In these tests (Figure 10 and Figure 11) the graphs show the I/O reaching a total value of approximately 11GB/s, with both read and write I/O performing to the same maximum values. We're drawn the read and write lines separately (write as curved lines) as the data is so closely aligned in terms of performance. The latency graphs show all of the tests (except QD32) deliver sub-millisecond performance, while the QD1 graphs deliver latency under 700 microseconds. Only the unconstrained QD32 test exceeds the latency QoS definitions, due to congestion at the backend media.

Conclusions

The aim of this paper was to test the claims made by Volumez with respect to its data infrastructure solution. The tests undertaken have been designed to demonstrate a mix of synthetic workloads that align to common data profiles in the enterprise.

We believe this data shows the consistent and deterministic nature of the Volumez platform, with I/O bandwidth and latency matching the expected configured values. Moreover, when backend media is congested, I/O latency remains constant, with sub-millisecond response times.

Why should public cloud users care about deterministic and predictable I/O performance? Ultimately the inefficient use of infrastructure resources, whether on-premises or the public cloud, results in additional costs. In the public cloud, customers pay through an operational expense model that amplifies any inefficiencies. That cost inefficiency is aggregated monthly. Any overpayment is unrecoverable the following month, by which time that overpayment, if not corrected, becomes liable again.

With traditional cloud storage, dynamic environments make the optimisation of resources a logistical and time-consuming challenge. Application I/O demand fluctuates with the demands of the business, however changing the performance resources assigned to individual virtual instances is a manual process. The typical approach taken by most businesses is to over-allocate and periodically review if assignments are excessively generous.

The Volumez solution delivers value for end users for three primary reasons. First, it aggregates resources into a pool from which abstracted metrics of IOPS and throughput can be assigned. Pooling consolidates fragmented demand and smooths the peaks and troughs across many virtual instances. The result is a much lower high watermark of demand than simply aggregating the requirements of each virtual instance application. Data infrastructure is thus tailored to the requirements of each workload, with policies differentiating and guaranteeing performance, which can be scaled linearly to continuously align with increasing workload demands.

Second, the solution can use high-performing instance storage, which is typically ephemeral. Volumez adds resiliency and predictability to the raw storage capacity. Pooling and aggregation also enable Volumez to deliver IOPS and bandwidth capabilities to a single instance that may not be available with existing storage options.

Third, performance metrics can be changed dynamically, independent of capacity or the application instance configuration. This means I/O can be scaled linearly, boosted or removed from a virtual instance without intervention. Through the use of policies, these changes can be implemented across tens or thousands of instances at a single click.

None of these capabilities just discussed would matter if performance from the Volumez platform did not deliver the predictability we have observed. End users need that reassurance. Our conclusion is that Volumez is more than capable of delivering the efficiency, scalability, and predictability that customers need for data intensive workloads in the public cloud.

More Information

Architecting IT is a brand name of Brookend Ltd, an independent consultancy working to explain technology and business value to the end customer. This is an independently written report, with data from testing performed in conjunction with Volumez.

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform. After working abroad, he co-founded an Internet-based music distribution company during the dotcom era, returning to consultancy in the new millennium.

Chris writes a popular blog at https://www.architecting.it/blog, hosts the Unpacked Network podcasts, attends many conferences and invitation-only events, and can be found providing regular industry contributions through LinkedIn other social media outlets.

Email: chris@architecting.it
LinkedIn:
https://www.linkedin.com/in/chrismevans/