Hewlett Packard
Enterprise

Performance addendum for HPE servers and WekaIO Matrix

Contents

Overview

The goal of this addendum to the "Architecture guide for HPE servers and WekaIO Matrix" is to assist readers with sizing their own WekaIO Matrix cluster by providing an example performance evaluation with result data. It is intended to be read with a general understanding of that document's content around Matrix architecture and HPE reference design.

While the architecture guide contains performance estimates for reference configurations, those are only read-focused numbers and intended as simple guidelines for solution sizing. A performance assessment write up helps inform readers who want more information while planning their Matrix cluster investment. This addendum aids solution conversation and sizing around bandwidth, IOPS, or latency requirements. It is also intended as a reference and comparison point for other Matrix performance data.

This testing and result data comes from a synthetic benchmark evaluation of WekaIO Matrix version 3.1.6 in a lab configuration very similar to the architecture guide's entry-level configuration. Synthetic benchmarks are a form of testing designed to have easily repeatable results. This makes it simpler to focus on one area of the system—in this case, storage performance—but may not fully reflect a user's experience using an application. The "real-world" benchmarks attempt to use a particular application or workloads emulating application usage to gauge solution performance. They're typically more complicated to understand relative to solution sizing and are outside the scope of this performance addendum.

Methodology

Data samples from testing must show reasonable latency and scale to demonstrate a performance bottleneck. This evaluation was not an attempt to achieve highest possible bandwidth or IOPS results but rather to characterize the cluster under test at more realistic performance scales.

Testing used the Flexible I/O Tester (fio) benchmark application (configuration documented in Appendix B). While there are a number of different benchmarking suites that can perform these types of tests, fio was chosen as a well-recognized synthetic benchmark that is easy to install and configure for anyone wanting to reproduce results.

The test runs cover commonly benchmarked I/O sizes (4 KB and 1 MB), with 100% read and write workloads. This not only provides more complete information about how well the given cluster can perform but also just as importantly demonstrates where the bandwidth and IOPS performance saturation points are for the cluster under test. Saturation, in this case, is defined as a point where a given performance metric is relatively level across increasing workload samples but shows a noticeable increase in latency.

Test configuration

This section covers high-level details of the cluster used to generate the data for this addendum (more complete hardware and software information documented in Appendix A.

This Matrix cluster differs in a few components from the configuration specified in the architecture guide due to availability. Rather than the reference Intel® Xeon® Gold 6126 CPUs and HPE InfiniBand 840QSFP28 adapters, the dedicated storage servers have Intel Xeon Gold 6134 CPUs and HPE InfiniBand 841QSFP28 adapters. The NVMe SSDs used are HPE 1.6TB NVMe x4 Lanes Mixed Use SSDs rather than 3.2 TB SSD.

Impact of hardware used versus reference configuration is as follows:

- **CPU:** The faster clock speed increases performance where I/O is bottlenecked on dedicated storage core processing.
- **InfiniBand adapter:** This particular adapter only supports InfiniBand and not 100Gb Ethernet. This variable should not impact test results relative to reference hardware.
- **NVMe SSDs:** 1 MB I/O write performance is expected to differ between 1.6 TB and 3.2 TB devices, as the write performance ratings are materially different (1.6 TB device is rated for max write throughput of 1400 MiB/s and the 3.2 TB device rated for up to 2000 MiB/s). Actual results data will, of course, require additional testing.

Test cluster diagram

The Matrix cluster under test is composed of two HPE Apollo 2000 Gen10 systems containing four HPE ProLiant XL170r Gen10 servers each (eight servers in total). Each server is configured as a dedicated Matrix storage server, with 12 cores per server dedicated to WekaIO exclusively. Of those 12 cores, four cores are used only as drive cores.

Twenty-five HPE ProLiant DL360 Gen9 servers are connected to the same EDR InfiniBand network as Matrix clients. A single core on each client is reserved for Matrix use as a front-end core. These load-generation clients were chosen solely for test availability and EDR connectivity, and do not represent any particular reference recommendation.

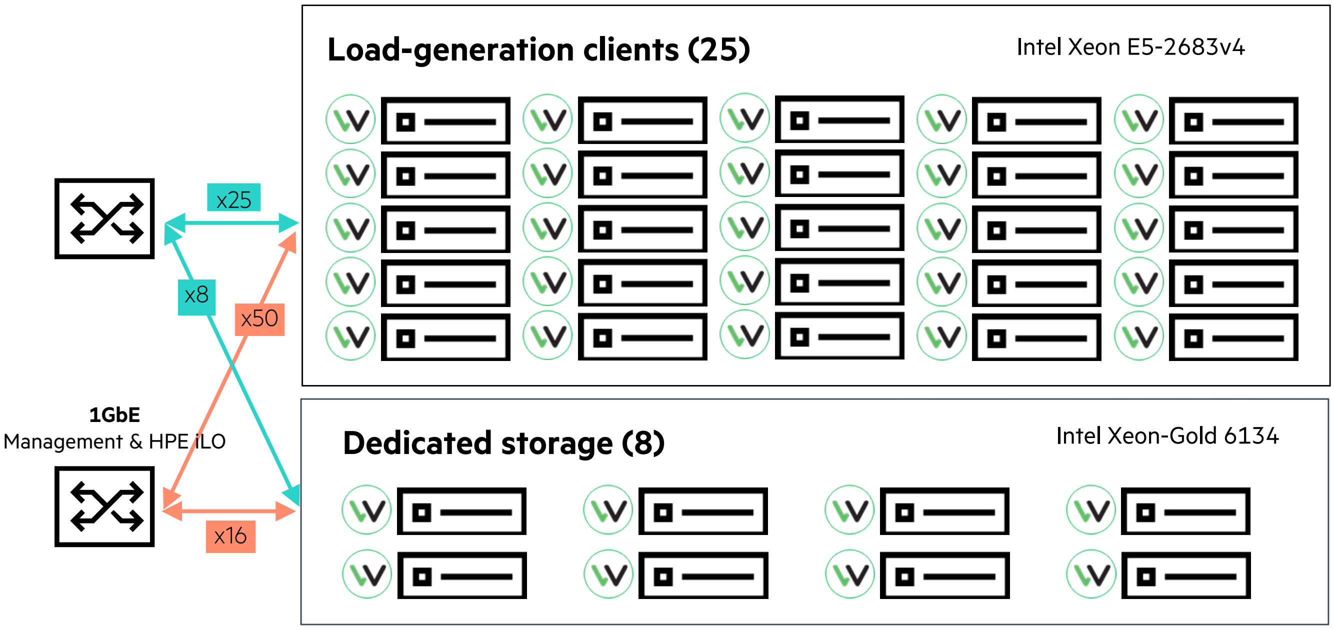

Figure 1. Matrix test cluster diagram

Performance analysis

Interpreting results

These tests are done by running fio in parallel across multiple load-generation clients, using a unique test directory per client. Directories are created on the test cluster's MatrixFS file system and accessed through the POSIX interface using the WekaIO Matrix kernel client.

Results summarize fio output and indicate mean quantity sampled over the length of the run (180 seconds) across all clients in the test. The key performance figures are bandwidth expressed in GB/s (1000^3 bytes/second), IOPS, and latency. I/Os are 4096 bytes (4K) or 1024^2 bytes (1M). Latency is in a quantity most appropriate for I/O type (milliseconds for 1M I/O, microseconds or usec for 4K I/O).

A unit of workload is a single fio job (process) running one outstanding I/O of the target size on that client. All clients run the same amount of work, so for example, 25 clients running 600 jobs indicates a workload of 24 jobs per client.

The performance analysis graph data shows the key performance metrics and cluster bottlenecks for each workload. Around each graph, a short discussion of the graph data is provided, along with how the key performance metrics for that scaling test might be improved upon by adding additional cluster hardware resources.

Key performance figures

The peak sample table entries, in the following table, describe the highest sampled values for cluster testing. It demonstrates testing that met or exceeded estimated performance for an entry-level cluster, as well as adding write workload data. It is not intended to be used as-is for solution sizing.

The last table line—saturation multiple—indicates the peak scale sample divided by single client peak sample. This is primarily useful in summarizing how quickly saturation was hit while scaling. Multiclient data shows conclusive performance saturation of the storage cluster in all cases but 4K reads.

Category	1M read bandwidth (GB/s)	1M write bandwidth (GB/s)	4K read IOPS	4K write IOPS
Single client peak sample	5.73	6.93	144,533	98,904
Peak scale sample	44.73	11.82	3,351,052	889,657
Saturation multiple	7.81	1.71	23.19	9.00

Extrapolating results

These key performance figures—and other performance data in this addendum—can be used as-is for a general level set of achievable performance when thinking about what your solution needs. They're also useful to indicate where more hardware will be required; if you can't realize the required performance even in an ideal case, then a cluster design clearly needs to increase the available resources.

It is not recommended to take these performance numbers as an exact sizing input to build other cluster BOMs or overall solutions.

- Synthetic benchmarking may not be a good representation of actual application workloads, so it's a much better practice to estimate some overhead until real-world application performance requirements are known. It's not atypical to buffer an additional 30% performance overhead of solution performance on a key metric—say, bandwidth—because real applications generate a mix of IO sizes, can have complex file layouts, and are not 100% read or write. The most efficient way to size the solution is, of course, a proof-of-concept test with the target application.
- Sizing hardware for performance on WekaIO Matrix is a balancing act between storage, CPU type, CPU core count, and network ports. Each choice is very material to cluster performance. While you may be able to extrapolate a performance number—say, linear percentage improvement by scaling CPU core frequency—it's possible that another bottleneck will be reached elsewhere before that ideal number is hit.

Near-term enhancements

Post Matrix 3.1.6., there will be software changes that will impact the test results as follows:

- Larger I/O performance should improve on reference flash devices. For the tests in this document, 1M writes are expected to show the most improvement—1M reads saturate on the single EDR port.
- Also coming is the ability to reserve more than one front-end core for Matrix clients. This is specifically for single clients with very heavy performance requirements and allows leveraging more than one EDR or 100GbE port with enough core resource allocation and bandwidth-based performance requirements.

Both features are targeted to be explored in upcoming performance evaluations.

Single client scaling

These graphs show single client scaling to effective performance saturation. Since the Matrix cluster and EDR port are capable of more performance than shown here, the bottleneck is the Matrix front-end client core.

If designing for clients with high-performance requirements, these results can be improved upon with hardware. On Matrix 3.1.6, use of faster and latest generation CPU cores are the best ways to improve client performance. Test samples using an additional HPE ProLiant XL170r Gen10 node with Intel Xeon Gold 6134 CPUs as a client showed significant double-digit percentage improvements over these load-generation servers—IOPS more so than bandwidth.

Post Matrix 3.1.6, the software will leverage more core reservation and network ports on the client.

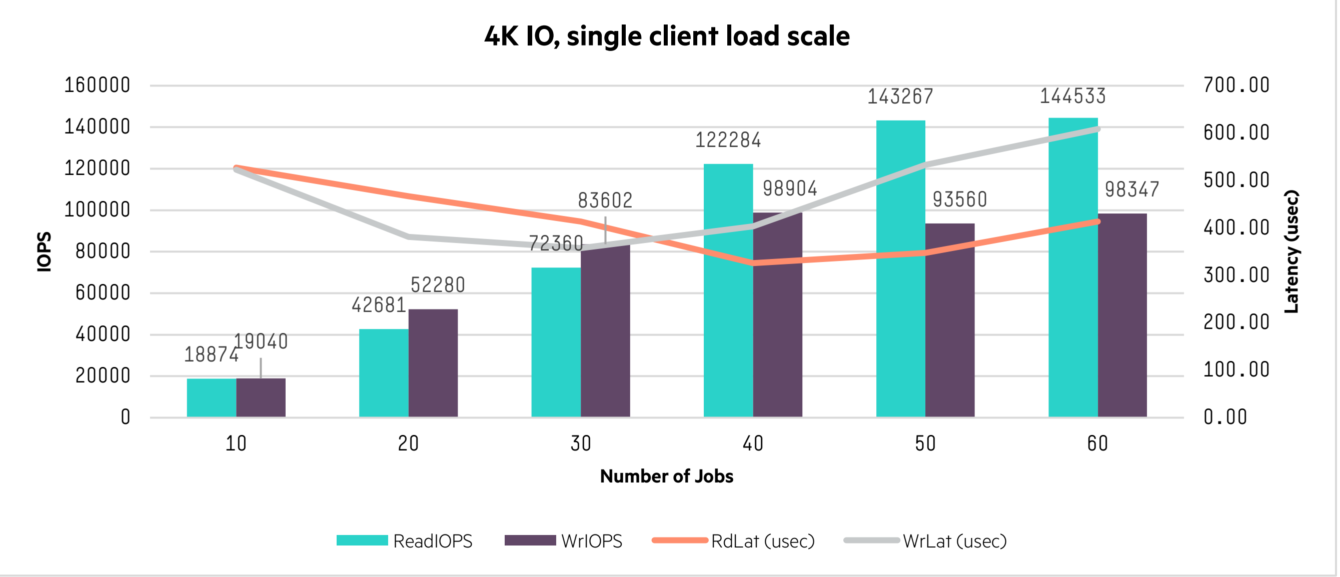

Figure 2. 4K IO, single client workload scaling

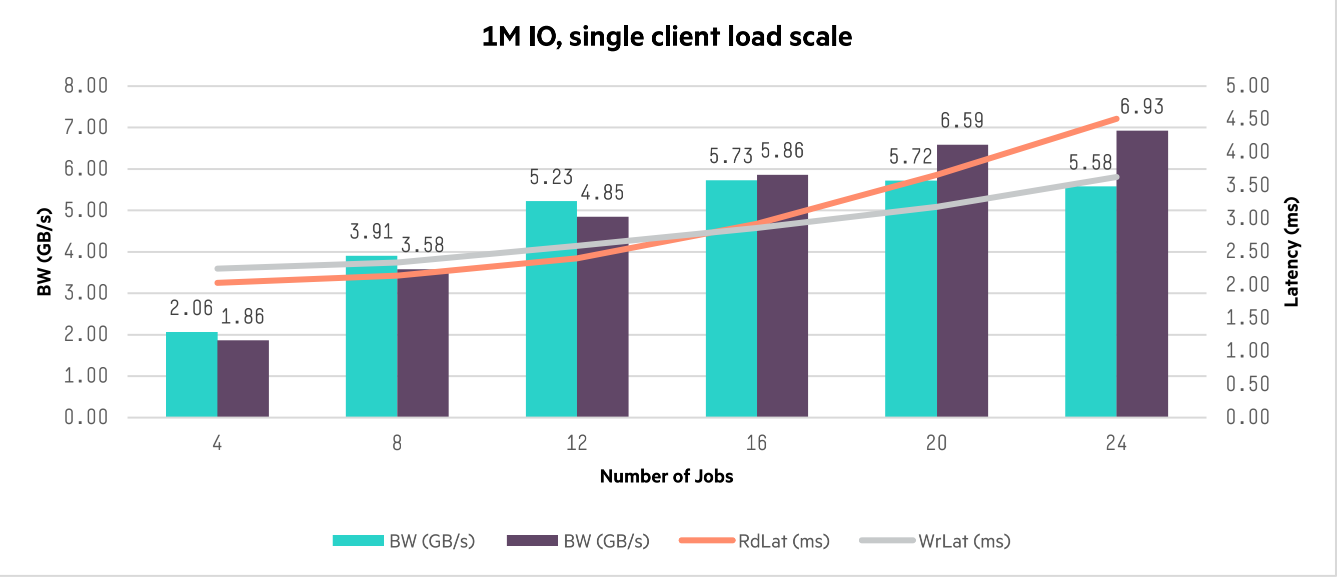

Figure 3. 1M IO, single client workload scaling

Multiclient scaling

Here, the approach is scaling load-generation client count until saturation at the Matrix storage nodes.

1M read I/O

In the case of the 1M read results, the cluster saturation point is the EDR ports on the dedicated storage servers. Reads on MatrixFS can result in additional cluster traffic to satisfy the I/O request, so the EDR ports are not serving all traffic directly to clients. Effectively half of the dedicated storage bandwidth can be consumed by cluster overhead on read I/O.

The 1M Read graph in Figure 4 shows significant total bandwidth gained between where scaling starts to level out versus that peak saturation point, but the impact on latency and the dramatic decrease versus linear scaling is clearly there. Increasing clients does increase the total cluster performance to a point, but well before we hit our total number of cluster clients.

Improving results seen here comes down to improving the ratio of networking ports to storage. This could have been done by adding another EDR card to the node. The cluster traffic overhead on reads means NVMe devices that can saturate a PCIe 3.0 x4 interface could actually utilize two x16 PCIe 3.0 network cards worth of read bandwidth. That, of course, doubles the storage-node port count and is really only a benefit in the bandwidth-driven read case.

Another way to improve that networking to storage ratio is to add more of the basic storage building block—more storage nodes. This also has the advantage of enabling for a wider, more efficient and performant MatrixDDP if the original design used less than 16 data chunks. Given that the bottleneck here is on the networking ports, the total amount of storage could be kept constant and still see a performance improvement.

1M write I/O

These 1M cluster write results are limited by the MatrixDDP configuration and the write capabilities of the drives. As with the 1M read case, there's some efficiency to be gained past the point where the scaling starts to taper off but writes saturate much more quickly.

To improve 1M write results, the recommended approach is a wider cluster with more storage server building blocks. A server with more storage and connectivity such as the HPE ProLiant DL360 Gen10 would also serve this end, but particularly for smaller clusters, it's better to design wider rather than deeper.

It's also possible to choose a MatrixDDP that balances cluster fault tolerance more toward performance if that's acceptable. Modifying MatrixDDP on the test cluster from 4+2 to 6+2 not only boosts write performance but also significantly impacts self-healing ability. As discussed in the architecture guide, a MatrixDDP less than the width of the cluster allows regenerated data chunks to be distributed to healthy cluster members, which is why our default recommendation for smaller clusters is still less than cluster width.

Here are some results that show performance scaling on the reference recommended 4+2 MatrixDDP.

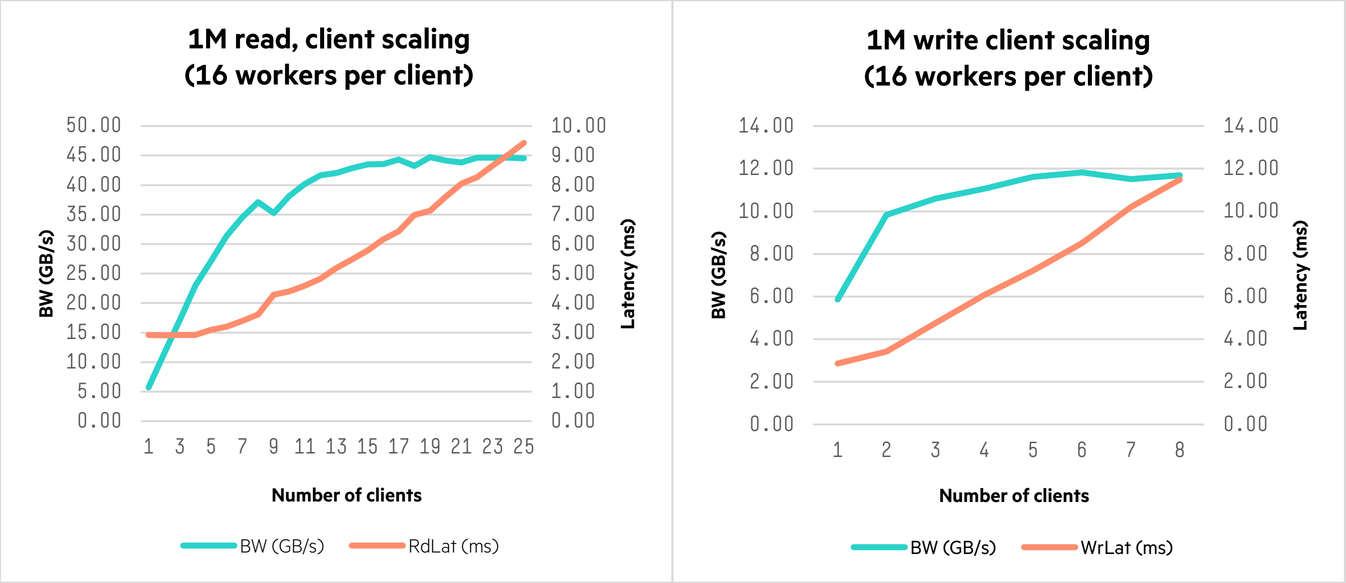

Figure 4. Multiple client scaling data, 1M IO

4K I/O

The peak client result samples show a comparison of IOPS and latency as client count and workload scaled rather than the fixed per client workload approach of the 1M testing.

4K read testing doesn't really characterize a cluster saturation point but rather the limits of our load-generation client pool. More detailed samples do show per client count saturation points that are consistent with this graph in Figure 5, and the performance curve here indicates a (fairly close to) linear scale all the way to the end.

To improve 4K read results more clients—or soon, more front-end cores—would be required to demonstrate storage node saturation. The assumption is that the next bottleneck seen would be storage-node reserved CPU core performance.

For 4K writes, the results show the limitation of storage node reserved CPU cores. That ~900K IOPS peak result is not a limitation of bandwidth, and it's far under a linear client scaling multiple at 25 clients. Also, there's no real improvement between 4+2 and a 6+2 MatrixDDP, so testing doesn't indicate a drive bottleneck like the 1M write data.

Only one storage core per drive can be dedicated and additional back-end cores (8) are already reserved for WekaIO. Beyond this, those resources are not I/O bottlenecks. Therefore, increasing cluster write IOPS will require more drives or drives with a higher write performance rating than the 1.6 TB MU used here. On the reference design using HPE ProLiant XL170r nodes with 4 NVMe drives per server, this means more nodes would have to be added to the cluster.

As with 1M IO, all 4K performance graphs show performance scaling on the reference recommended 4+2 MatrixDDP.

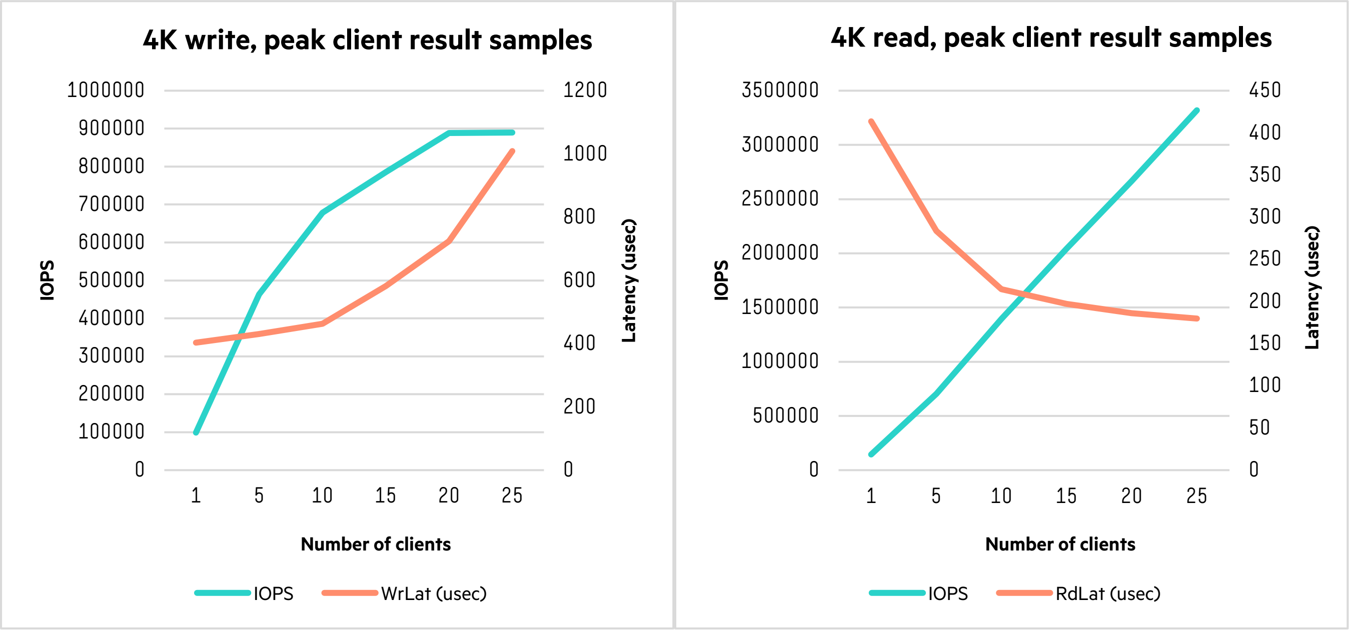

Figure 5. Multiple client scaling data, 4K IO

The more detailed data from 4K IOPS that contains workload scaling at given client counts reinforces the story from the IOPS peak sample graphs. In Figure 6 and 7, we've expanded the per-client data into total worker count, in this case, a worker being a single fio job executing a 4K IO. The total number of workers is evenly divided across clients.

Read data shows a comparison of samples at different worker counts as clients scale. Comparing the same amount of total work across different client counts highlights where the workload saturates. Figure 6 reinforces that the performance data is bound by client performance rather than the Matrix cluster, showing improved performance as client count increases but total workload is comparable. This is easiest to see in samples in the middle of the graph, where ~300 workers have significantly different IOPS results as sampled at 10–25 clients.

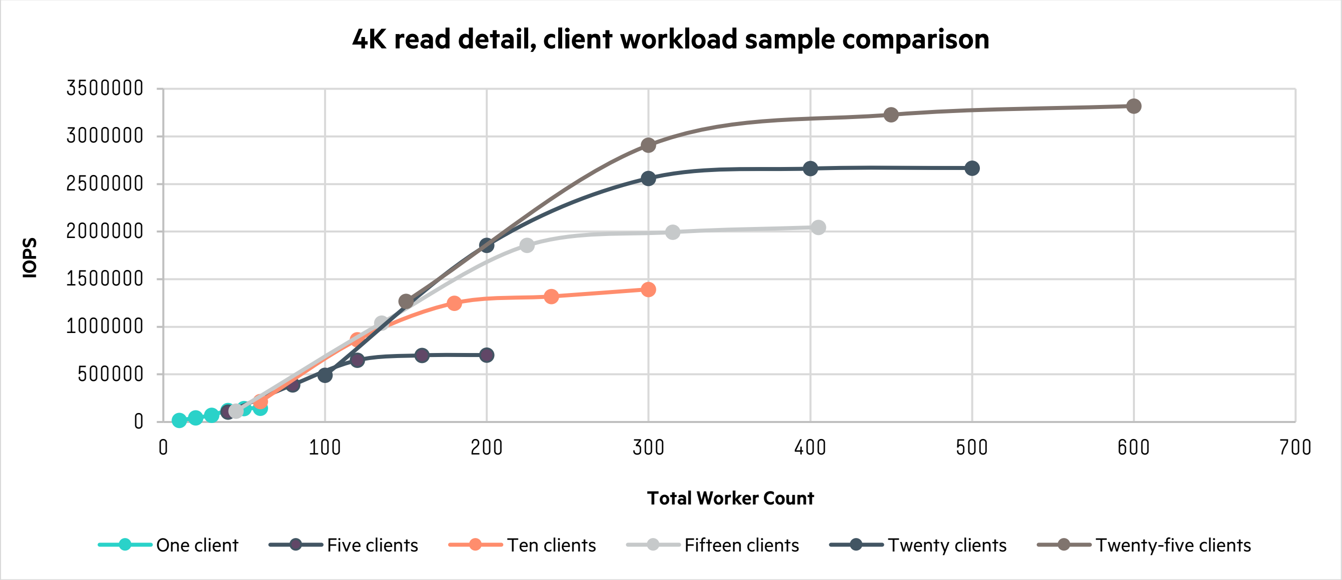

Figure 6. Detailed 4K read client scaling data

Figure 7 shows the back-end saturation by demonstrating sample convergence. Lower workload and client counts do show improvement while scaling, but from 15 clients and up the results track to pretty much the same curve.

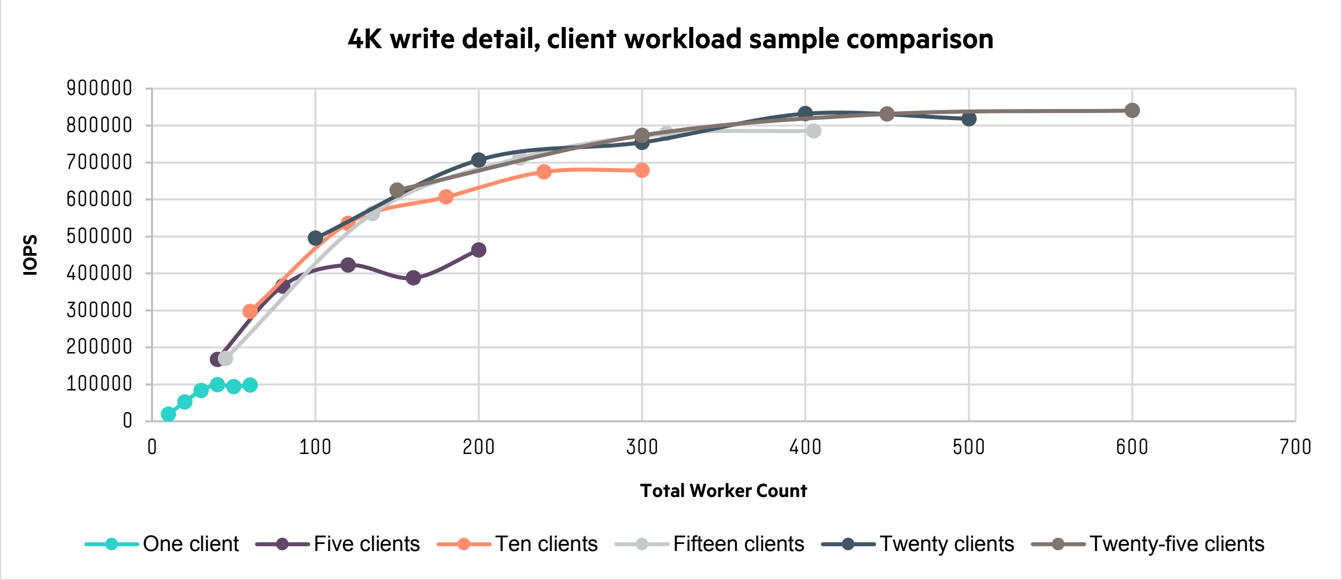

Figure 7. Detailed 4K write client scaling data

Summary

This document shows benchmarking results and process on WekaIO Matrix. Concrete examples help illustrate the sizing choices discussed in the architecture guide.

This (close to) reference design testing shows:

- The magnitude of the key performance figures in bandwidth, IOPS, and latency to set performance expectations.
- Where this particular hardware bottlenecks on those key performance tests, plus ways the design can be modified to achieve more performance.
- Enough data to reproduce for in-house vetting of results.

General guidelines

Increasing the number of building blocks (such as servers) or using higher performance drives is the simplest and most straightforward way to achieve better IOPS and bandwidth. When adding additional servers but keeping the drive count constant, the increased number of network ports, cores, and wider possible MatrixDDP also improves performance and usable capacity.

To scale a starting solution configuration on HPE reference platforms to higher levels of performance for a specific I/O metric and type, here are suggestions from the results:

- **Bandwidth-driven read requirements:** More network ports can also give the cluster more performance when cluster traffic and drive I/O are greater than available network bandwidth. For denser boxes such as the HPE ProLiant DL360 Gen10 server, PCIe x16 slots will be limited for achieving this—using high-performance NVMe a peak ratio would be two x4 NVMe devices per x16 EDR adapter.
- **IOPS-driven read requirements:** It’s typical to saturate on client resources first. Once the back-end does see IOPS saturation, more storage devices and more CPU core resources are the next steps.
- **Writes:** Generally, more total storage devices are needed, as write I/O tends to bottleneck on the drive capabilities. For write bandwidth, a peak ratio of four to five high-performance NVMe devices per x16 EDR adapter is ideal.

Upcoming work

Additional performance evaluation documents and solution sizing tool improvement are planned in a number of areas. This particular document’s data will be expanded and updated as new Matrix features and performance enhancements are released and testing is done.

Use the resource links at the end of this document to keep up to date with HPE and WekaIO’s ongoing effort to provide up-to-date reference configurations, sizing assistance, and general expertise for building your own cluster. For further questions, contact your HPE sales or account representative.

Appendix A: cluster configuration details

Software

Matrix storage systems

Category	Value
Operating system	CentOS 7.4
WekaIO MatrixFS version	3.1.6
MatrixDDP configuration	4+2
Core count	12 total cores
Core settings	4 dedicated storage cores, 8 unspecified
Filesystem size	24.82 TiB

Matrix client systems

Category	Value
Operating system	CentOS 7.4
WekaIO MatrixFS version	3.1.6
Core count	1 total core
Core settings	1 dedicated front-end core
Filesystem mount options	-o readcache

Hardware

Storage servers are two HPE Apollo 2000 Gen10 systems containing four HPE ProLiant XL170r Gen10 nodes used as WekaIO Matrix dedicated storage. Client servers are 25 HPE ProLiant DL360 Gen9 nodes used for test load generation. Quantities indicate total in configuration.

Matrix storage systems

Component name	Quantity	SKU
HPE ProLiant XL170r Gen10 1U Node Configure-to-order Server	8	867055-B21
HPE XL1x0r Gen10 Intel Xeon-Gold 6134 (3.2GHz/8-core/130W) FIO Processor Kit	8	874290-L21
HPE XL1x0r Gen10 Intel Xeon-Gold 6134 (3.2GHz/8-core/130W) Processor Kit	8	874290-B21
HPE 8GB (1x8GB) Single Rank x8 DDR4-2666 CAS-19-19-19 Registered SmartMemory Kit	96	815097-B21
HPE XL1x0r Gen10 Left Low Profile Riser Kit	8	874296-B21
HPE XL170r Gen10 16NVMe P2 Low Profile Riser Kit	8	874304-B21
HPE InfiniBand EDR/Ethernet 100Gb 1-port 841QSFP28 Adapter	8	872725-B21
HPE XL1x0r Gen10 M2 (NGFF) Riser Kit	8	874853-B21
HPE 240GB SATA 6G Mixed Use M.2 2280 3yr Wty Digitally Signed Firmware SSD	16	875488-B21
HPE Ethernet 1Gb 2-port 368FLR-T Media Module Adapter	8	866464-B21
HPE XL170r Gen10 S100i SATA Cable Kit	8	874305-B21
HPE 1.6TB NVMe x4 Lanes Mixed Use SFF (2.5in) SCN 3yr Wty Digitally Signed Firmware SSD	32	875597-B21
HPE Apollo r2800 24SFF-Flex Gen10 CTO Chassis	2	867159-B21
HPE r2800 Gen10 16SFF NVMe Backplane FIO Kit	2	874800-B21
HPE r2x00 Gen10 Redundant Fan Module Kit	2	874308-B21
HPE 1600W Flex Slot Platinum Hot Plug LH Power Supply Kit	4	830272-B21
HPE r2x00 Gen10 PSU Enablement Kit	2	880186-B21
HPE 2U Shelf-Mount Adjustable Rail Kit	2	822731-B21

Matrix client systems

Component name	Quantity	SKU
HPE ProLiant DL360 Gen9 4LFF CTO Server	25	755259-B21
HPE DL360 Gen9 Intel Xeon E5-2683v4 (2.1GHz/16-core/40MB/120W) FIO Processor Kit	25	818198-L21
HPE DL360 Gen9 Intel Xeon E5-2683v4 (2.1GHz/16-core/40MB/120W) Processor Kit	25	818198-B21
HPE 8GB (1x8GB) Single Rank x8 DDR4-2400 CAS-17-17-17 Registered SmartMemory Kit	200	805347-B21
HPE DL360 Gen9 Low Profile PCIe Slot CPU2 Kit	25	764642-B21
H240ar 12Gb 2-ports Int FIO Smart Host Bus Adapter	25	749976-B21
HPE DL360 Gen9 LFF Smart Array P440ar/H240ar SAS Cable	25	766211-B21
HPE 1TB 6G SATA 7.2K RPM LFF (3.5 inch) SC Midline 1yr HDD	25	861691-B21
HPE IB EDR/EN 100Gb 1P 840QSFP28 Adapter	25	825110-B21
HPE Ethernet 10Gb 2-port 560FLR-SFP+ Adapter	25	665243-B21
HPE 500W Flex Slot Platinum Hot Plug Power Supply Kit	25	720478-B21
HPE 1U LFF Gen9 Easy Install Rail Kit	25	789388-B21

Appendix B: test configuration details

Clients used fio 3.1.

Here is a sample configuration file from the test run (in this case, a job sample for 4K write). Parameters varied in this configuration file during testing are as follows:

- Blocksize (bs=) either 4096 or 1048576
- Number of jobs (numjobs=) scaled to various loads
- I/O pattern (readwrite=) either randwrite or randread
- Unique work directory created for each client (directory=)

Choice of ioengine from WekaIO feedback on typical best practices for their own testing with fio asynchronous I/O; performance differences between this and other I/O engines has not been quantified at this time.

```
[global]
  ioengine=posixaio
  name=randperftest
  bs=4096
  direct=1
  invalidate=1
  time_based=1
  ramp_time=15s
  runtime=180s
  readwrite=randwrite

  group_reporting
  filesize=209715200
  numjobs=30

[hostnamejob]
  iodepth=1
  nrfiles=1
  directory=/mnt/weka/clienttest/rg-matrixclient33
```

Resources

Product page

HPE HPC solutions

WekaIO Matrix product page

Learn more at
hpe.com/storage/wekaio

Sign up for updates

a00052112ENW, August 2018