IN&OUT AG

Hitachi Vantara VSP 5500 Performance Benchmark SUVA

Andreas Zallmann
CEO, In&Out AG

Brian Mathis
brian.mathis@suva.ch
Head of Server & Storage Platforms

«We are very pleased with the performance of the Hitachi Vantara VSP 5500 Storage Systems and are entering new performance dimensions, with which SUVA is very well positioned for the years to come.

Thanks to the valuable support of In&Out AG, we were able to carry out the Storage Evaluation in a very targeted and efficient manner and achieved an optimal result, also in financial terms.»

Version:	1.0.4 English
Date:	27.05.2020
Classification:	Public

In&Out AG IT Consulting & Engineering
Seestrasse 353, CH-8038 Zürich
Phone +41 44 485 60 60

info@inout.ch, www.inout.ch

Preliminary Remark

The Swiss Accident Insurance SUVA has re-evaluated its storage systems via a RFP and decided to purchase two Hitachi Vantara VSP 5500 systems. Within the scope of a proof of concept, the performance of the systems was tested according to the RFP specifications. The results presented here are based on the joint PoC between SUVA, Hitachi Vantara and In&Out.

SUVA

Suva is the largest accident insurance company in Switzerland. It employs around 4200 people at its headquarters in Lucerne, in 18 agencies throughout Switzerland and in the two rehabilitation clinics in Bellikon and Sion. For more than 100 years, Suva has ensured that work and leisure time in Switzerland are safer.

- Suva offers compulsory insurance cover for employed and unemployed persons against accidents and occupational diseases.
- Das Modell Suva vereint Prävention, Versicherung und Rehabilitation.
- The Suva model combines prevention, insurance and rehabilitation.

In&Out

In&Out AG from Zurich has been supporting its customers for years as an independent and manufacturer-neutral consulting company in the areas of IT infrastructure and data centers with a special focus on storage.

In addition to consulting services in the area of storage and support in storage RFPs, In&Out AG has many years of proven experience in storage performance benchmarks and has developed the benchmark tool IOgen™.

Hitachi Vantara & SUVA

Hitachi is firmly established worldwide in business-critical environments and has also been a reliable technology partner for SUVA for many years, responsible for data management. SUVA's requirements in terms of performance, reliability and disaster prevention are very high. This means that SUVA must use suitable and new technologies in IT. For this reason SUVA invited storage vendors to participate in the RFP for new storage and carried out a detailed and objective evaluation of the offerings.

Since the end of 2019, the latest flagship of the Hitachi Vantara block storage family, the VSP 5500, is available. This storage system supports up to 12 controllers (6 dual controller nodes) in a full mesh setup and can be equipped with NVMe or SAS SSD/FMD and HDD storage. The VSP5000 has already mastered or even surpassed the demanding objectives with flying colours at the medium expansion stage. From the comprehensive options that the VSP5000 offers in terms of technology composition, SUVA decided on an NVMe configuration with "Advanced Data Reduction". This enables SUVA to achieve an enormously high performance and storage density and also reduces costs. All further information about Hitachi VSP5000 are available under the following link.

https://www.hitachivantara.com/en-us/products/storage/all-flash-hybrid-flash-storage.html

Objective

Within the framework of a proof of concept, SUVA has verified the performance of the evaluated Hitachi Vantara VSP 5500 storage systems. In addition to SUVA specific workloads, typical random and sequential workloads were measured, which are common in the storage environment and are suitable for comparison.

Management Summary

Hitachi Vantara has a new top model in its portfolio, the VSP 5500, which sets new performance standards. Already with 8 controllers, it pushes into new dimensions of IO performance (summed over both systems):

- 10 Mio. IOPS Frontend Random Reads 8 KB
- 3.5 Mio. IOPS Frontend Random Writes 8 KB (50% synchronously mirrored, effectively 5.25 Mio. IOPS)
- 200 GB/s Frontend Sequential Read
- 50 GB/s Frontend Sequential Write (50% synchronously mirrored, effectively 75 GB/s)
- 4 Mio. IOPS Backend Random Reads 8 KB
- 750’000 IOPS Backend Random Writes 8 KB (50% synchronously mirrored, effectively 1.125 Mio. IOPS)
- 100 GB/s Backend Sequential Read
- 40 GB/s Backend Sequential Write (50% synchronously mirrored, effectively 60 GB/s)

The write performance is limited by the number of data center interconnects (ISLs) and could possibly be increased. In addition, the system is still equipped with a relatively small number of NVMe disks. Last but not least the system can be expanded with 50% more controller capacity.

With the NVMe backend the latency is on a very high performance level:

- 8 KB Read IOPS from Cache < 100µs
- 8 KB Read IOPS from the Backend < 300µs
- 8 KB Write IOPS into the Cache 300-350µs (Average at 50% synchronously mirrored Writes). Local Writes below 200µs, synchronously mirrored Writes below 500µs.

Test Setup

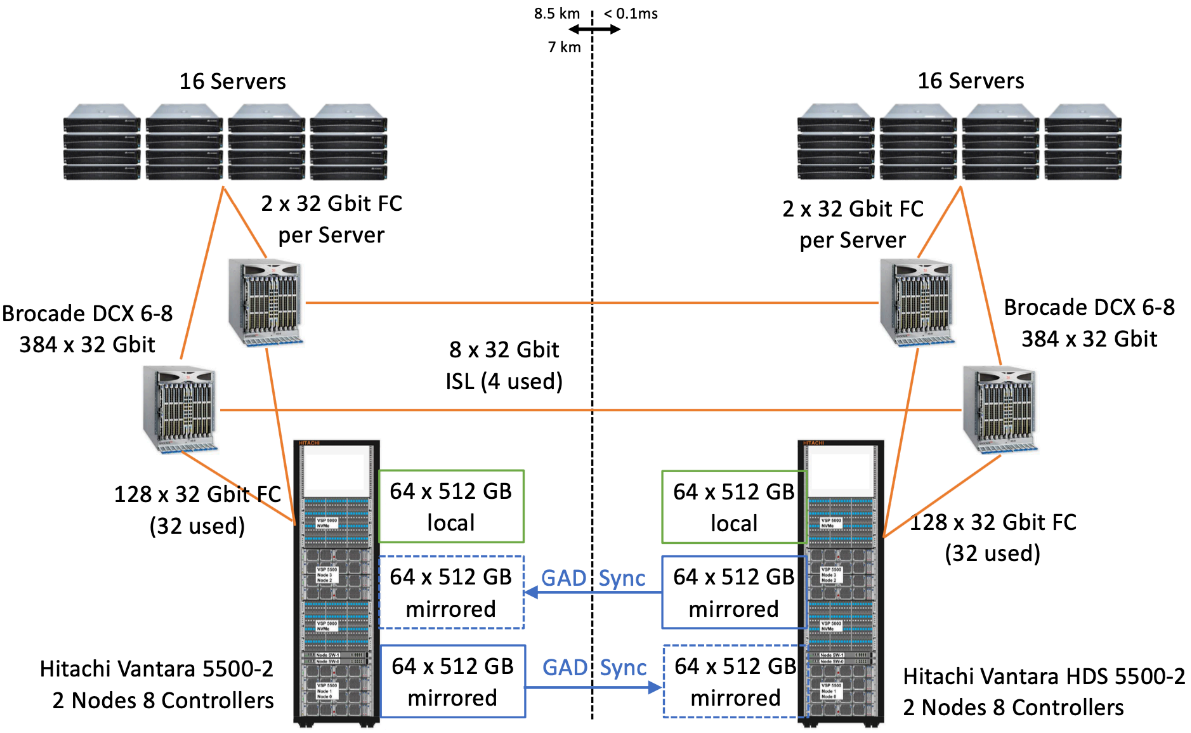


Figure 1 – Test Setup

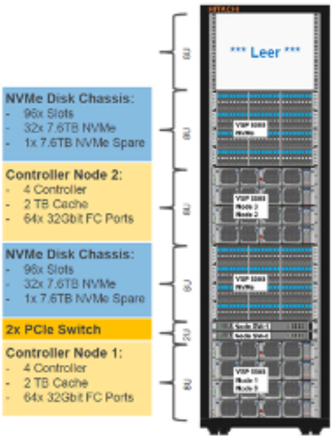
The storage systems are distributed over two locations. The RTT (Round Trip Time) between the locations is less than 0.1ms. Between the locations 4 ISL lines with 32 Gbit each (bidirectional) were used. The distance between the data centers (cable length) is 8.5 and 7 km, respectively.

VSP 5500 Configuration

Per location the following configuration was used:

	Per Location
Controller	8 (4N)
Cache	4'096 GB
FC	128 x 32 Gbit
- for Servers	32 x 32 Gbit
- for Remote Copy	4 x 32 Gbit
Disks	66 x 7.6 TB NVMe SSD, 2 of which Hotspares
	4 Raid Groups Raid 6 (14+2)
Usable Capacity	376 TiB
LUN Configuration	64 LUNs local x 512 GB = 32 TB
	64 LUNs sync. mirrored x 512 GB = 32 TB
	64 TB per Location, 128 TB in total
	50% sync. mirrored to remote location

Tabelle 1 – System Configuration per Location



It should be noted that the disk capacity was not yet fully installed for the proof of concept. In total, a usable capacity of 2 PiB will be used later.

Deduplication was configured as a Post Process and Inline Compression was enabled.

The configuration was chosen so that 50% of the capacity was synchronously mirrored to the other site, which corresponds to the requirement profile of SUVA. Thus 50% of all writes are synchronously mirrored.

With the 32 FC x 32 Gbit/s used, a theoretical maximum bandwidth of 1,024 GBit/s or 256 GB/s per site can be achieved. With the 4 FC x 32 Gbit used for remote copy (mirroring), 128 GBit or 16 GB/s (in both directions) can be transmitted.

Test Server

Per site 16 servers were in use, each connected to the SAN switches with 2 x 32 GBit FC. The bandwidth of the server connection per site was thus 1'024 GBit/s or 128 GB/s.

The servers were part of the RHEV (RedHat Enterprise Virtualization) farm, which was operated with other VMs during the tests. The following server types were in use:

- Huawei 2288 V5 with 2 Sockets Xeon Gold 6146 with 2 x 12 Cores, 3.2 Ghz, 768 GB RAM
- Lenovo ThinkSystem SR 650 with 2 Sockets Xeon Gold 6146 with 2 x 12 Cores, 3.2 Ghz, 768 GB RAM
- Lenovo SR650 G2 with 2 Sockets Xeon Gold 6246 with 2 x 12 Cores, 3.3 Ghz, 1.5 TB RAM

IOgen™

The I/O workloads were generated with IOgen™ 6.3.1 from In&Out AG. IOgen was installed directly on the RHEV Hypervisor of the 32 servers. The LUNs were accessed directly as raw devices (/dev/mapper/…).

IOportal™

The measurements of the I/O workloads acting on the storage systems as well as their performance (response times) and impact on the storage backends were carried out with IOportal™ of peaq GmbH. IOportal is the standard capacity and performance measurement tool for Hitachi storage systems and has been used by SUVA for many years.

Test Runs

The test runs were carried out by IOgen™ in parallel and synchronized to the millisecond on 32 servers. For each location, 16 servers were used, accessing their local storage system. The Writes were written 100% locally and 50% were synchronously mirrored to the other location.

The parallelism (number of worker processes) was slowly increased from 1 worker per server. All information regarding parallelism refers to the number of workers per server and must therefore be multiplied by 32 servers.

The following tests were performed:

- FE/BE RR – Random Read IOPS 8 KB Block Size
- FE/BE RW – Random Write IOPS 8 KB Block Size
- FE/BE SR – Sequential Read IOPS 1 MB Block Size
- FE/BE SW – Sequential Write IOPS 1 MB Block Size

Frontend (FE) tests were run against the frontend and the storage cache. Here only the first 1000 blocks were used per LUN, the cache hit ratio therefore quickly increased to 100%. In the backend (BE) tests, however, all blocks of the LUN are used. The cache hit ratio was close to 0%.

The same writing pattern was used for all write tests. 60% of the blocks were filled with random patterns and 40% of the blocks remained untouched implying high compression and deduplication factors. This distribution is typical for SUVA. With this pattern a dedup/compression (DeCo) factor of approx. 1.66:1 may be expected (100/60).

In the following analysis, frontend tests (solid lines) and backend tests (dashed lines) are displayed together.

For Random IOs the IOPS (IOs per second) and for Sequential IOs the throughput in GB/s (Gigabyte per second) are plotted. In addition, the latency measured on the servers is also always plotted in µs (microseconds) = 0.001 ms (milliseconds).

In our observations we distinguish between the maximum value (marked orange) and the optimum value (marked blue). The highest throughput is achieved for the maximum value, but the latency can already have increased very strongly due to the high load. The optimal value is considered the value where the throughput increases even more than the latency. Up to this point we are in the "healthy" range.

Random Read IOPS 8 KB

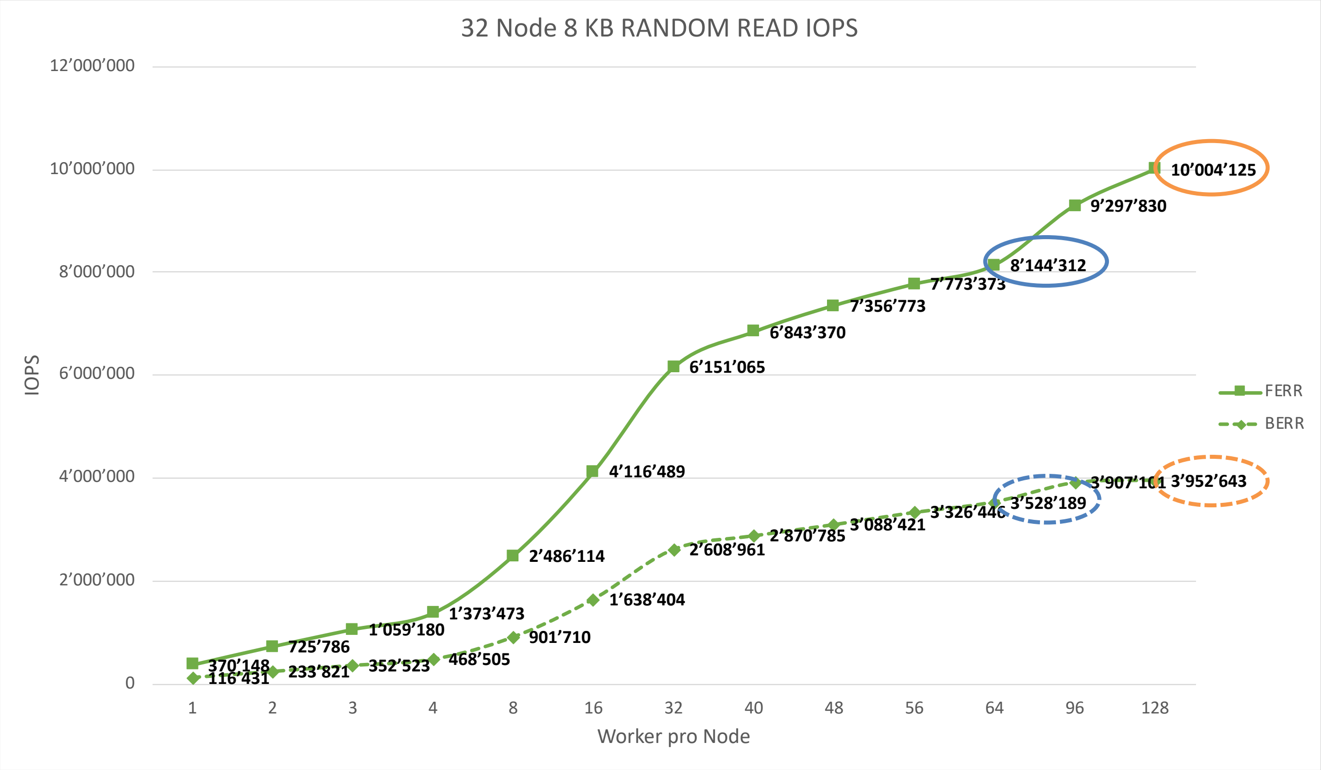


Figure 2 – Random Read Performance IOPS 8 KB

In the frontend, the two systems together can process more than 10 million frontend random read IOPS 8 KB (more than 5 million IOPS per system). In the back end, the systems can still deliver almost 4 million back end random read IOPS 8 KB in the "worst case" at 0% cache hit ratio, 2 million per system.

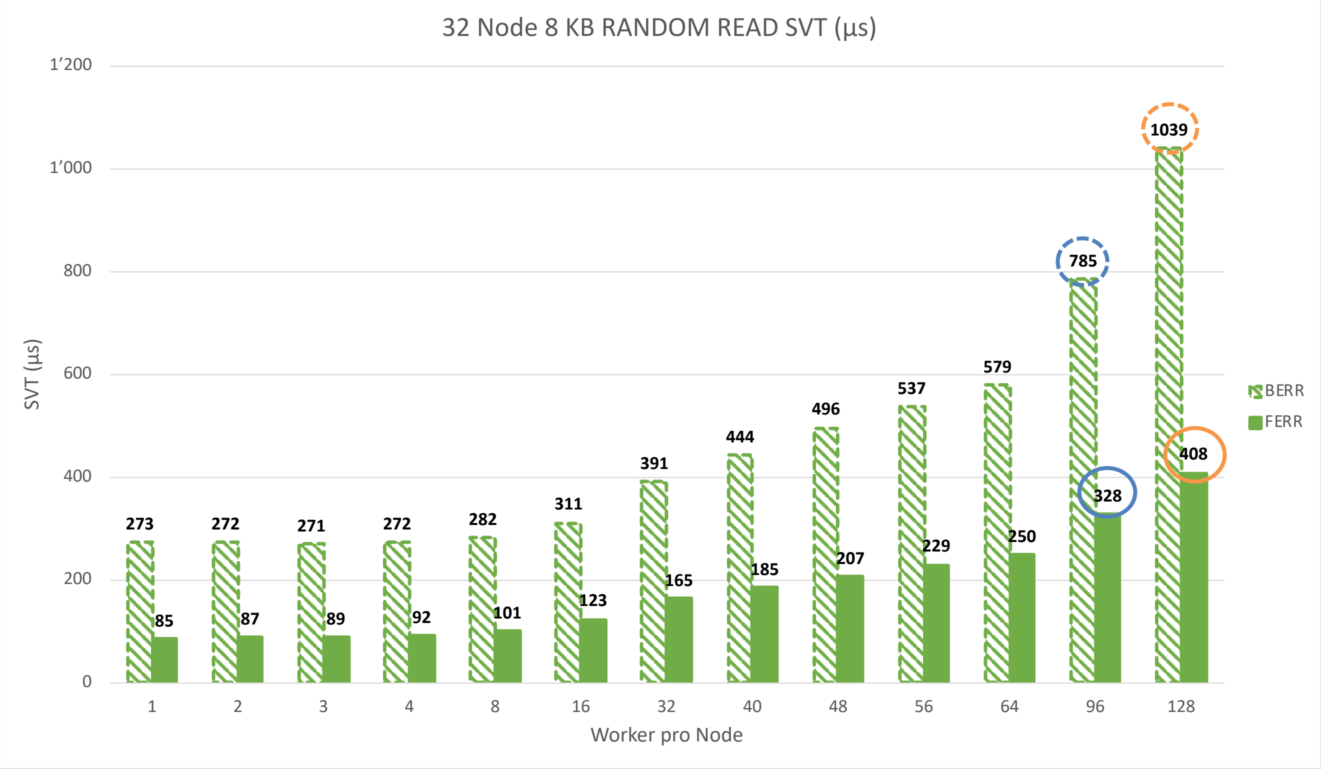


Figure 3 – Random Read Latency for 8 KB blocksize

Im Frontend beträgt die Latenz bei geringer Parallelität weniger als 100µs und steigt erst mit zunehmender Parallelität auf einen immer noch sehr guten Wert von 400µs (0.4 ms). Bei dieser sehr geringen Latenz von 400µs werden 10 Mio. IOPS verarbeitet. Der Optimalwert von über 8 Mio. Frontend Random Reads wird bei einer Latenz von 250µs erreicht.

In the frontend, the latency is less than 100µs at low parallelism and only increases with increasing parallelism to a still very good value of 400µs (0.4 ms). At this very low latency of 400µs, 10 million IOPS are processed. The optimum value of over 8 million frontend random reads is achieved at a latency of 250µs.

As an example of a front-end random read test, the following graphic shows the test process with 64 processes over 10 minutes with a 1 second resolution. The graphic shows that the IOPS and the service time remain absolutely constant.

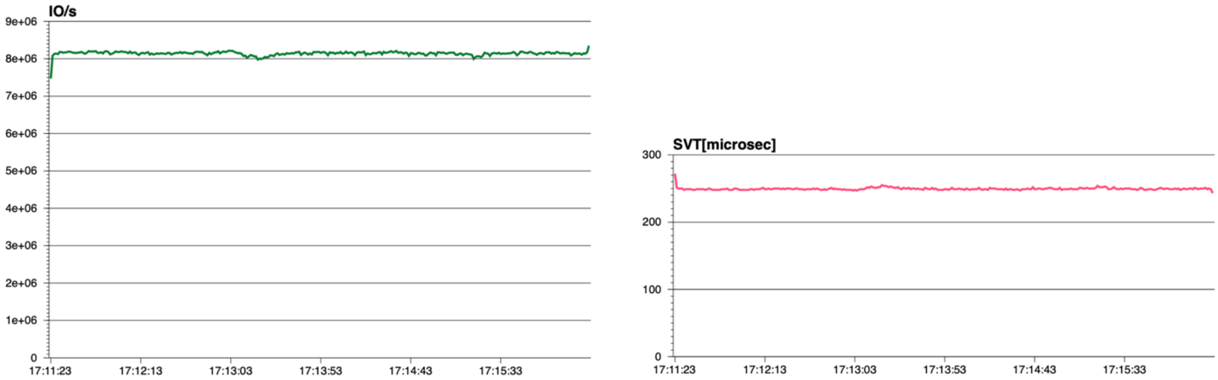


Figure 4 – Frontend Random Read Test (64 Processes)

The CPU load of the storage controllers reached values of over 70% in both frontend and backend tests.

Random Write IOPS 8 KB

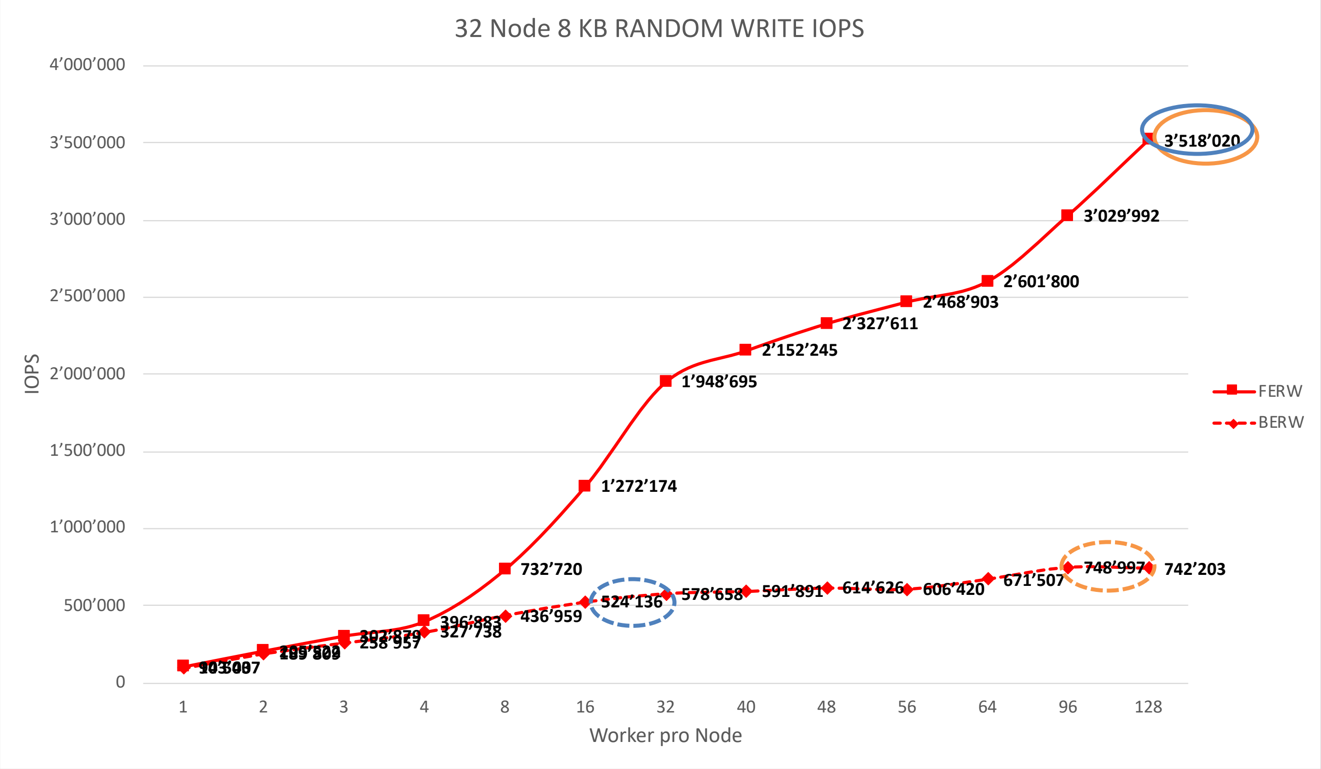


Figure 5 – Random Write Performance 8 KB

At the frontend, the two systems together processed more than 3.5 million frontend random write IOPS 8 KB (more than 1.75 million IOPS per system). Of the 3.5 million writes, 50% were local and 50% were synchronously mirrored to the other storage system. This means that 1.75 million writes were processed locally and an additional 875,000 writes were processed by the other side. This implies that 2.625 million writes were processed per storage system or 5.25 million front-end random write IOPS 8 KB were processed by both systems.

In the backend, the systems can process almost 750,000 Backend Random Write IOPS 8 KB in the "worst case" with 0% cache hit ratio, or 375,000 per system. Due to the mirroring effectively 1,125 million Backend Random Write IOPS were processed. The processing performance falls back to this rate when the write cache is completely filled, an extremely untypical scenario. The value corresponds to the downstaging rate of the NVMe SSDs. This could potentially be increased further with more NVMe SSDs.

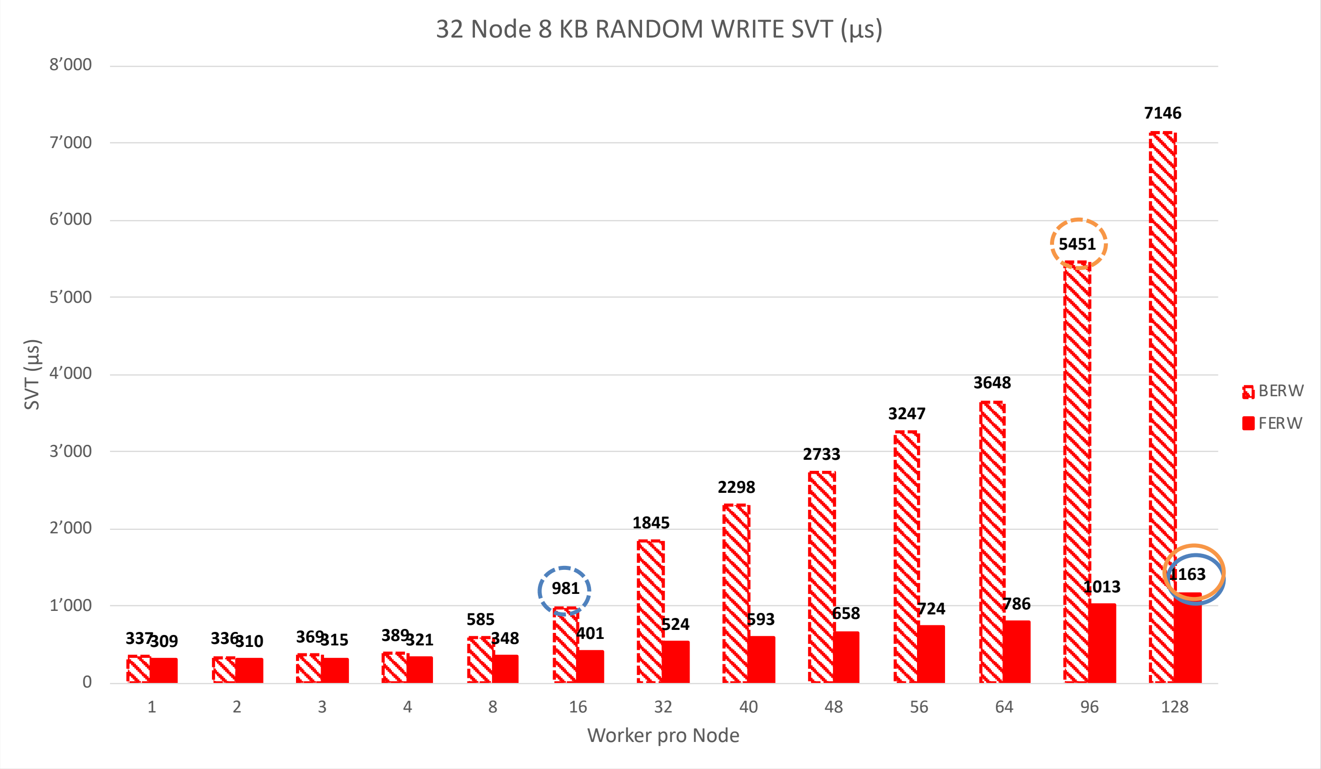


Figure 6 – Random Write Latency 8 KB

The frontend latency was around 309µs with low parallelism and remained very constant over a long period. At maximum parallelism the latency reached 1'163µs, with over 3.5 million random writes (effectively over 5 million random writes due to the mirroring).

It has to be taken into account that the 50% mirrored writes must be written to and acknowledged by the storage system at the remote site before the server receives the Write Acknowledge. With this background in mind the low latencies appear even more remarkable.

The backend latency was around 337µs for low parallelism. The latency diagram below shows the distribution of the latencies. 51.77% of the writes were processed with a latency of 100-250 µs. These are the non-mirrored writes. The other 50% mirrored writes are processed with a roundtrip time of about 100µs and a latency of more than 250µs due to the transport to the other location. Here 22.31% of the writes were executed with less than 500µs latency and 24.89% of the writes with less than 1'000µs latency.

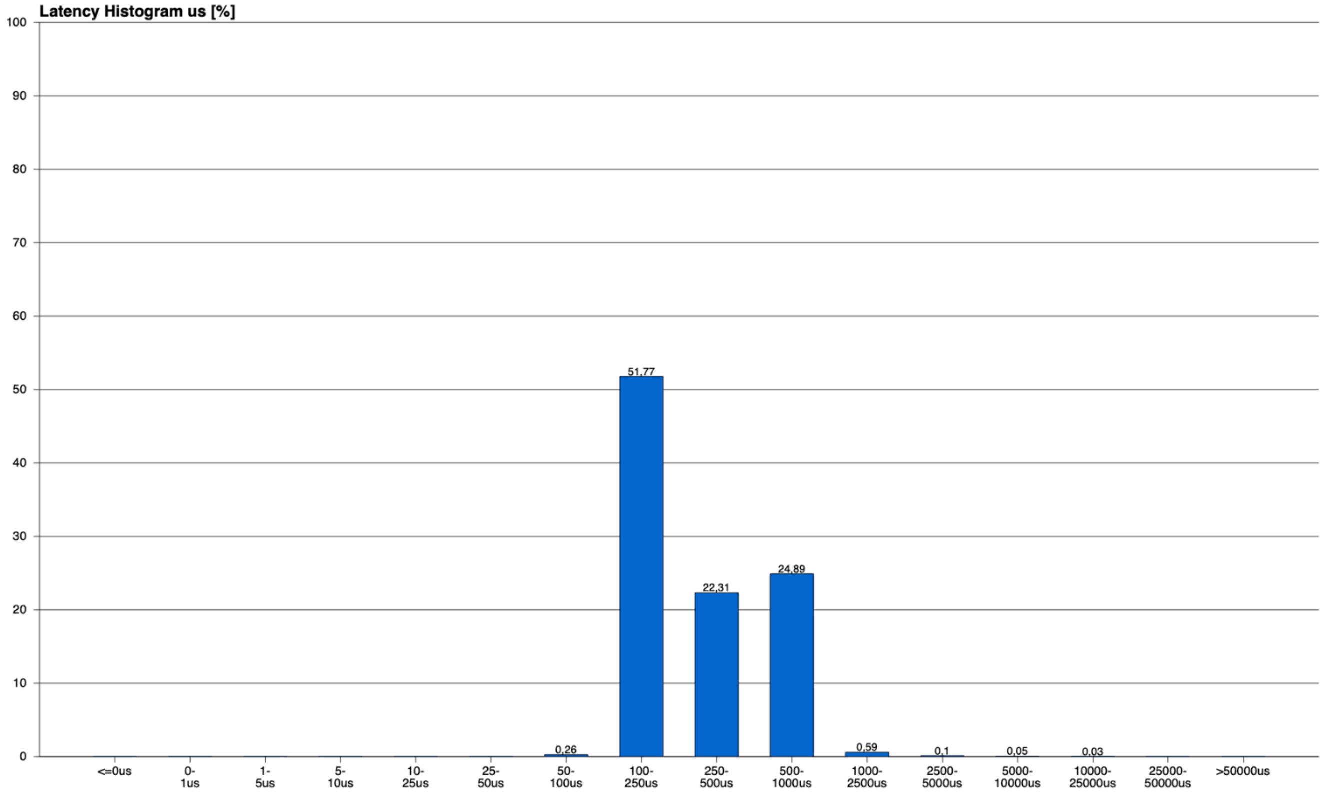


Figure 7 – Latency Histogram Backend Random Write (1 Process)

Up to 16 processes x 32 nodes (512 processes) the processing power increases stronger than the latency. The optimum value was found at 436,959 backend random writes 8 KB with a latency of 981µs. The maximum value of 750'000 backend random writes was reached with a significantly increased latency of 7 ms. The system was found to be no longer in the optimal range and could no more transfer the IOPS at optimal speed to destage to the SSDs.

As an example, the following graphic shows the test process for a backend random write test with 8 processes over 10 minutes with a 1 second resolution. At the beginning, the processing performance was higher and the service times remained lower, since the cache was not yet completely filled. The value then stabilized at a good 400'000 IOPS with a service time (latency) around 600 µs.

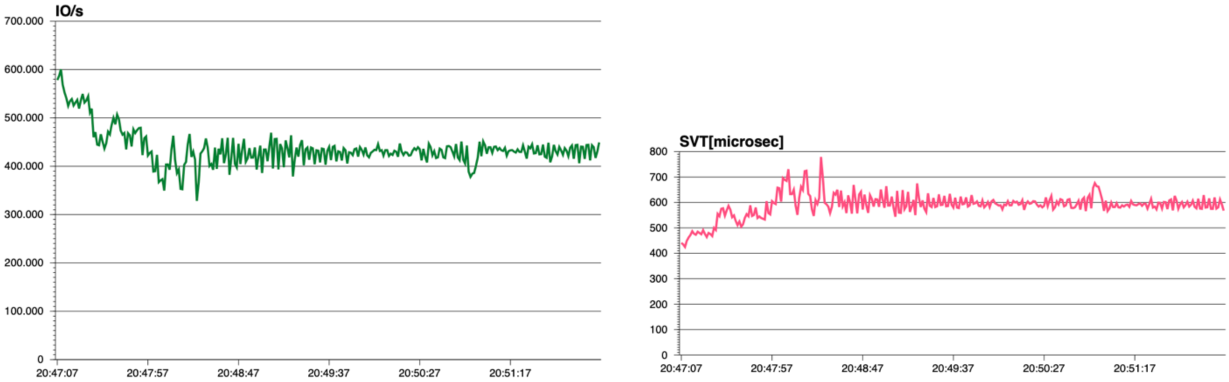


Figure 8 – Random Read Test (64 Processes)

The CPU load of the storage controllers was roughly 70% for the frontend tests and almost 90% for the backend tests.

Sequential Read 1 MB

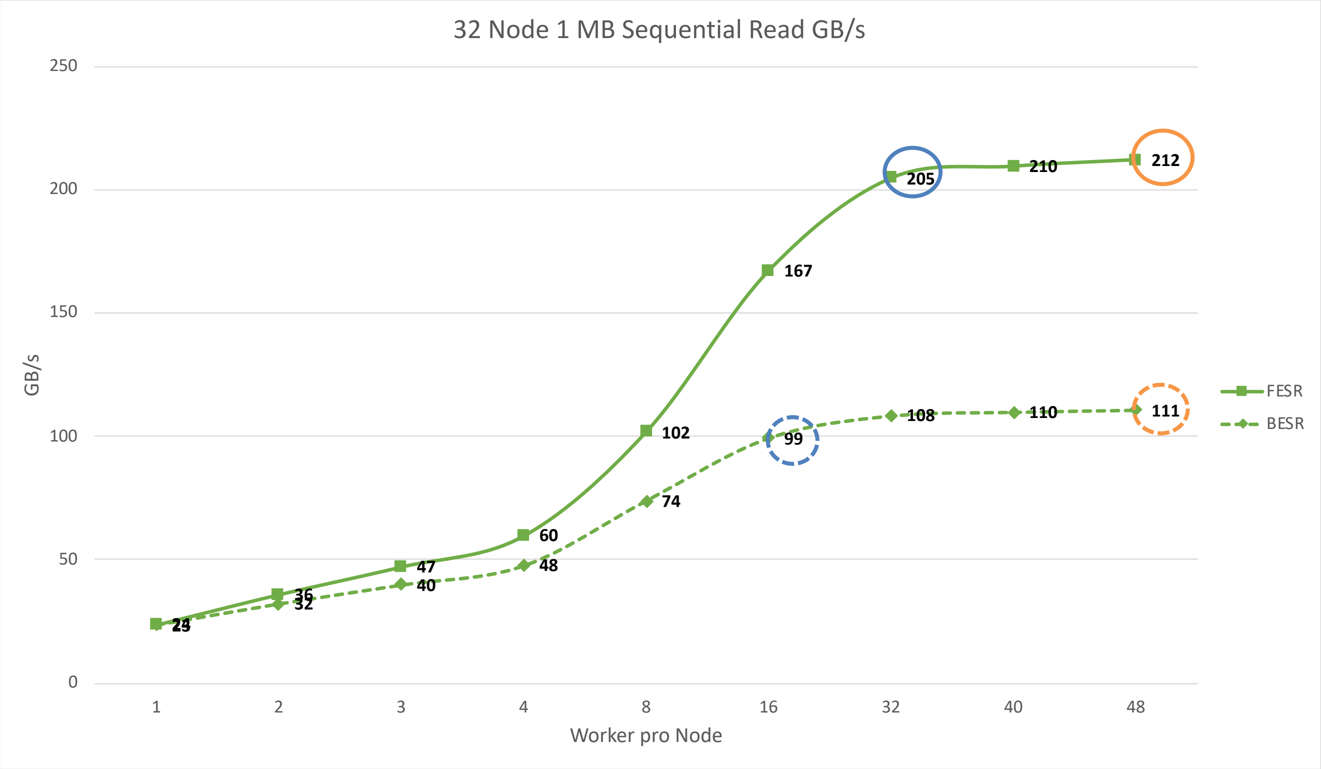


Figure 8 – Sequential Read Performance 128 KB

Combined, the two systems processed 212 GB/s in the frontend (which is more than 100 GB/s per system). The values are expressed in gigabytes per second, not in gigabits per second. In the back end, the systems could still process 111 GB/s in the "worst case" at 0% cache hit ratio, or 55 GB/s per system.

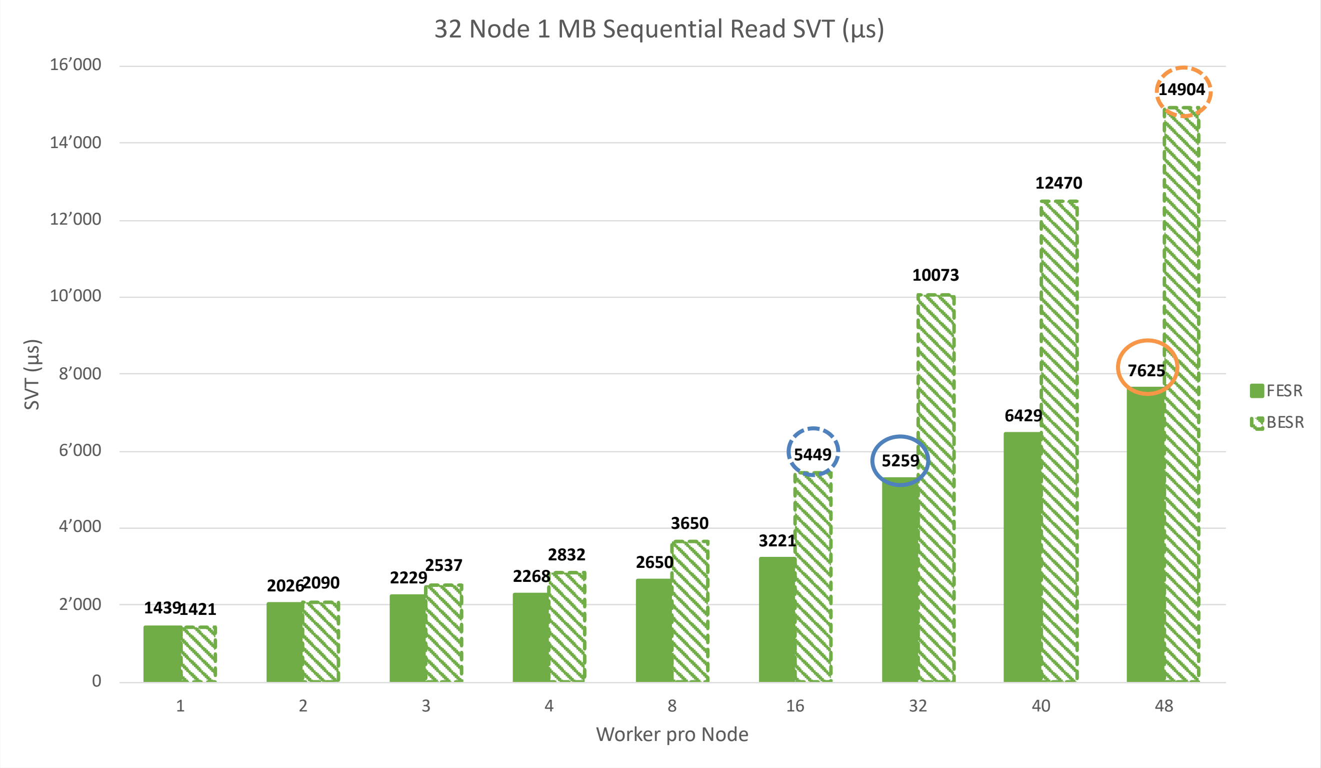


Figure 9 – Sequential Read Latency 1 MB

Both, the frontend and backend, exhibited a latency of about 1'400µs for reading a 1 MB block with low parallelism.

The frontend reached its optimal throughput of 205 GB/s with a latency of 5ms running 32 processes per server in parallel. Beyond that point, only the latency increased while throughput could only be improved insignificantly to 212 GB/s.

The backend reached its optimal throughput value of 99 GB/s with a latency of 5ms executing 16 processes per server in parallel. Again, beyond that point, only the latency increased while throughput increased only insignificantly to 111 GB/s.

As an example, the following figure shows the Frontend Sequential Read Test with 16 processes over 10 minutes with a resolution of 1 second. The figure shows that the processing performance remained constant at 160 GB/s with the service time of 3ms.

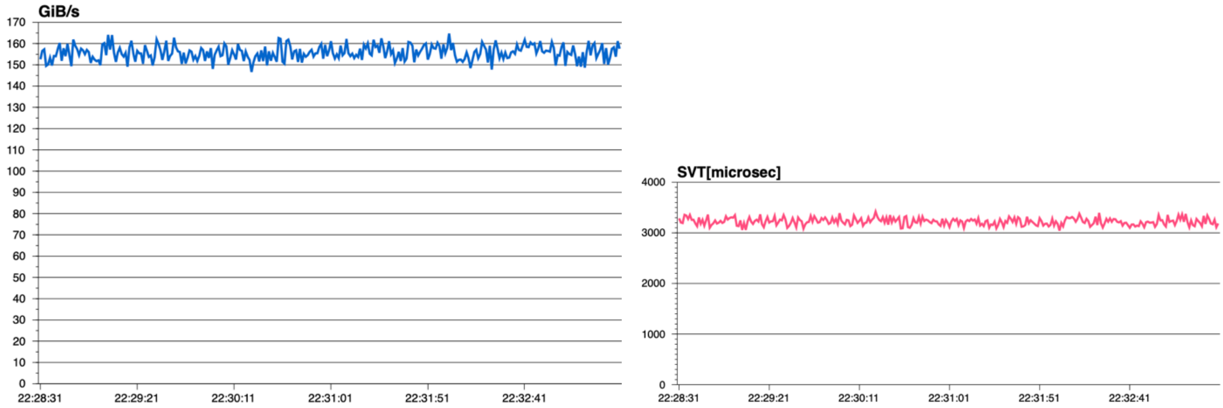


Figure 10 – Frontend Sequential Read Test (16 Processes)

The CPU load of the storage controllers for this test case was significantly lower than for the random 8 KB tests due to the much lower number of IOPS. CPU load reached a maximum of 50% for the frontend sequential read and a maximum of 65% for the backend sequential read tests.

Sequential Write 1 MB

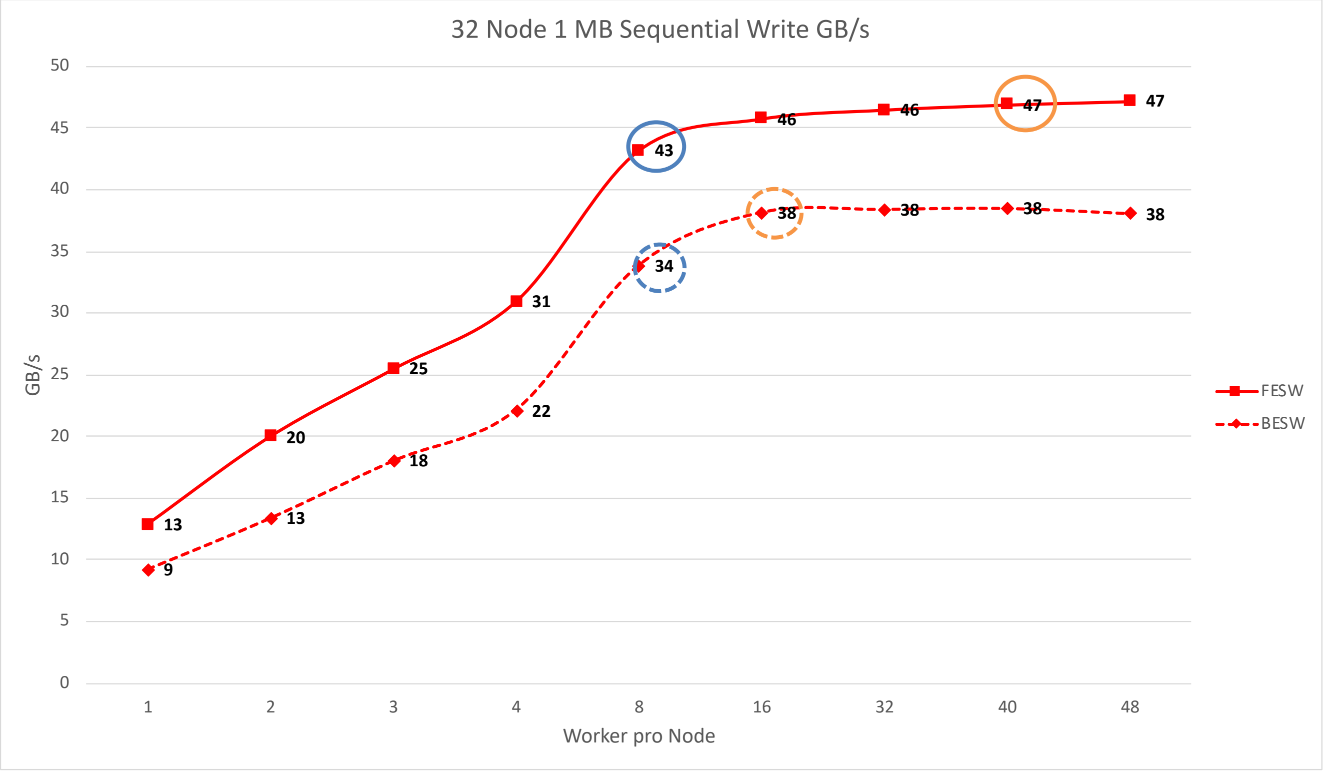


Figure 11 – Sequential Write Performance 1 MB

Combined the two systems processed almost 47 GB/s (approx. 23 GB/s per system). Of the 47 GB/s sequential writes, 50% were local and 50% were synchronously mirrored to the other storage system. This means that 23 GB/s were processed locally per storage and an additional 11.5 GB/s were processed by the other system. This means that 35 GB/s were processed per storage and 70 GB/s are processed combined by both systems.

The systems combined backend throughputs reached roughly 38 GB/s in the "worst case" scenario at a 0% cache hit ratio, or 19 GB/s per system. The backend value was found to be just a bit lower than the frontend value.

The write throughput during the tests was effectively only limited by the links between the two data centers, which could transmit up to 16 GB/s per direction with 4 links x 32 Gbit/s = 128 Gbit/s. The frontend transfer reached values of approx. 11.5 GB/s in each direction. The Interswitch Links (ISL) between the locations at those transfer values were utilized to over 70%. With more ISLs available, even higher write throughput values would be possible.

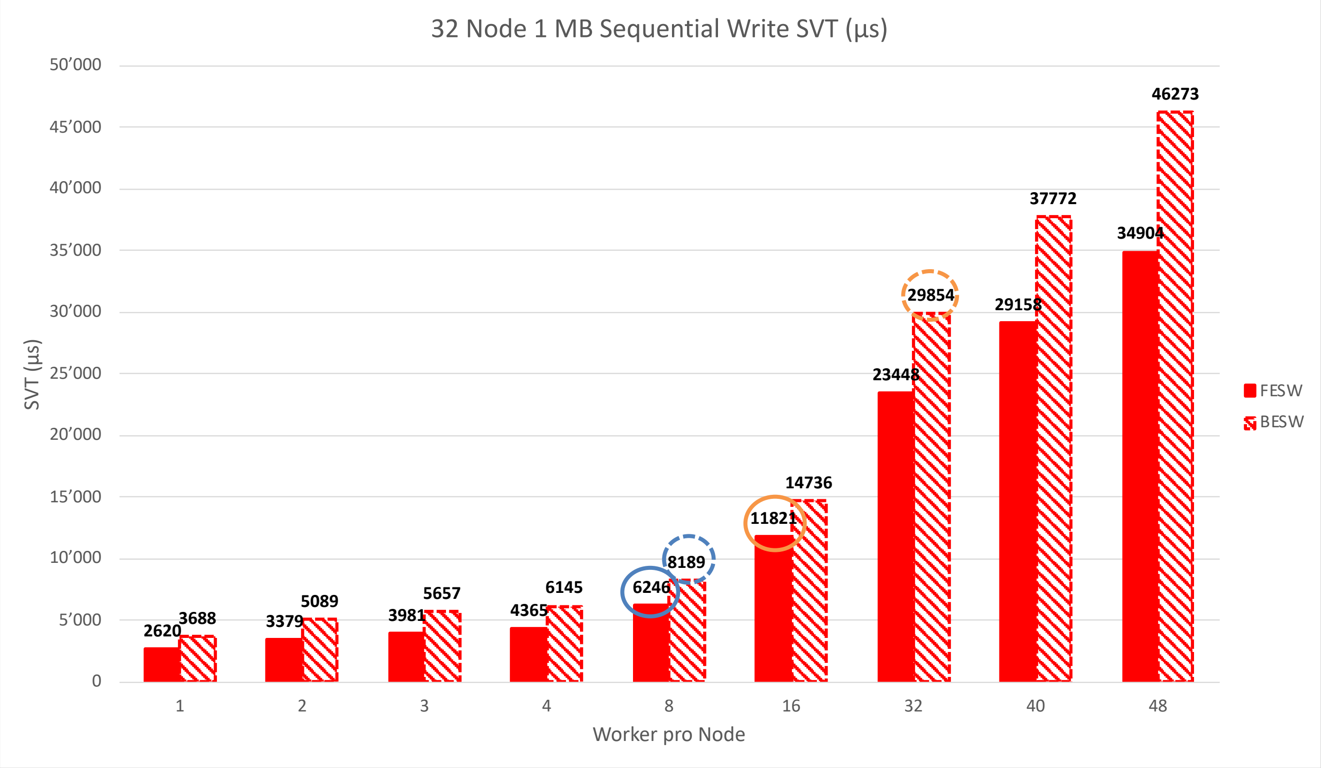


Figure 12 – Sequential Write Latency 1 MB

The frontend latency at low parallelism was found to be around 2.6ms, the corresponding backend latency around 3.7ms. The frontend processing speed increased rapidly to 200 GB/s with a latency of 6ms and while the backend throughput increased to values around100 GB/s with a latency of 8ms. Beyond these values, there is only an insignificant increase in throughput but a strong increase in latency, which is mainly due to the heavily used ISLs between the data centers.

The CPU load of the storage controllers reached up to 40% during the frontend tests and 65% during the back-end tests.

VSP 5500 I/O Analysis

With the help of the peaq IOportal the values reported by IOgen were compared with the key figures of the storage systems. The following graphic shows the key IO performance figures of one of the storage systems.

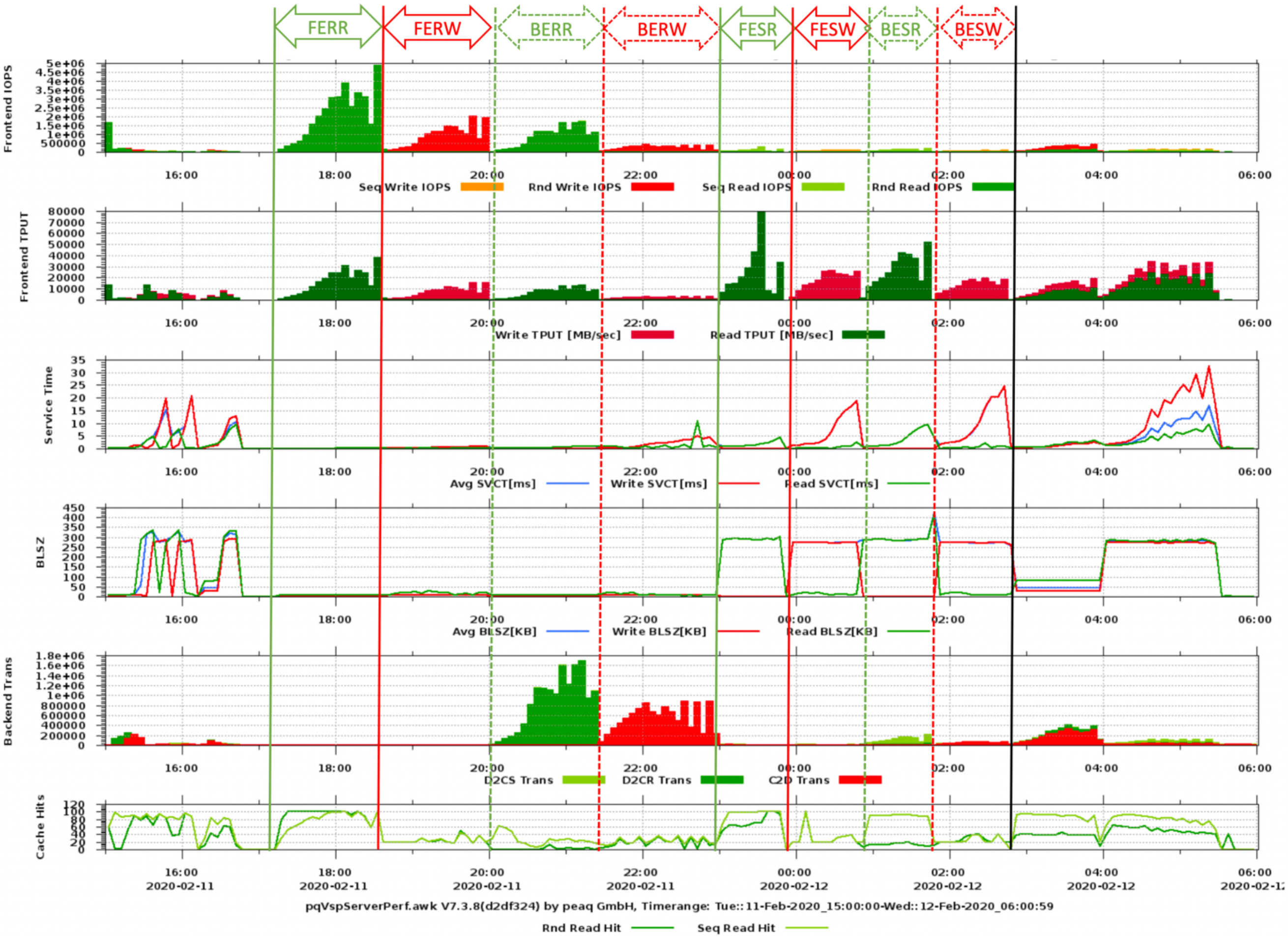


Figure 13 – Key Performance figures evaluated with the peaq IOportal

It can be seen the 5 million IOPS during the Frontend Random Read Test and the 1.75 million IOPS during the Backend Random Read Test. The IOPS for the Frontend Random Write Test correspond to 1.75 million IOPS and the Backend Random Write to 400'000 IOPS. Since these are only values for one system, they need to be doubled to get the combined results for both storage systems.

Furthermore, you can see that the block size (BLSZ) for the random tests was very small (8 KB) and that only for the backend tests transfers from the backend (green) or into the backend (red) took place.

For Frontend Sequential Reads you can see up to 80 GB/s throughput (peak values were about 100 GB/s per system or 200 GB/s in total, the IOportal graph shows 5 minute average values). Backend Sequential Reads throughput reached values of up to 50 GB/s per system (100 GB/s combined). The numbers for Frontend Sequential Writes also correspond to the measured numbers.

The IOportal confirms the used sequential test blocksize of 256 KB, the IOPS obviously being divided into slightly smaller chunks on the host level (block size of IOgen was originally set to 1 MB). Please note also, that during the frontend tests there were no observed backend transfers. Storage Backend activity only happened during backend testing. Overall, the values reported by IOgen and measured by IOportal correspond well for all the storage system key performance figures.

VSP 5500 Controller Utilisation

The following figure shows the CPU load of the individual controllers for each of the test phases on one of the storage systems.

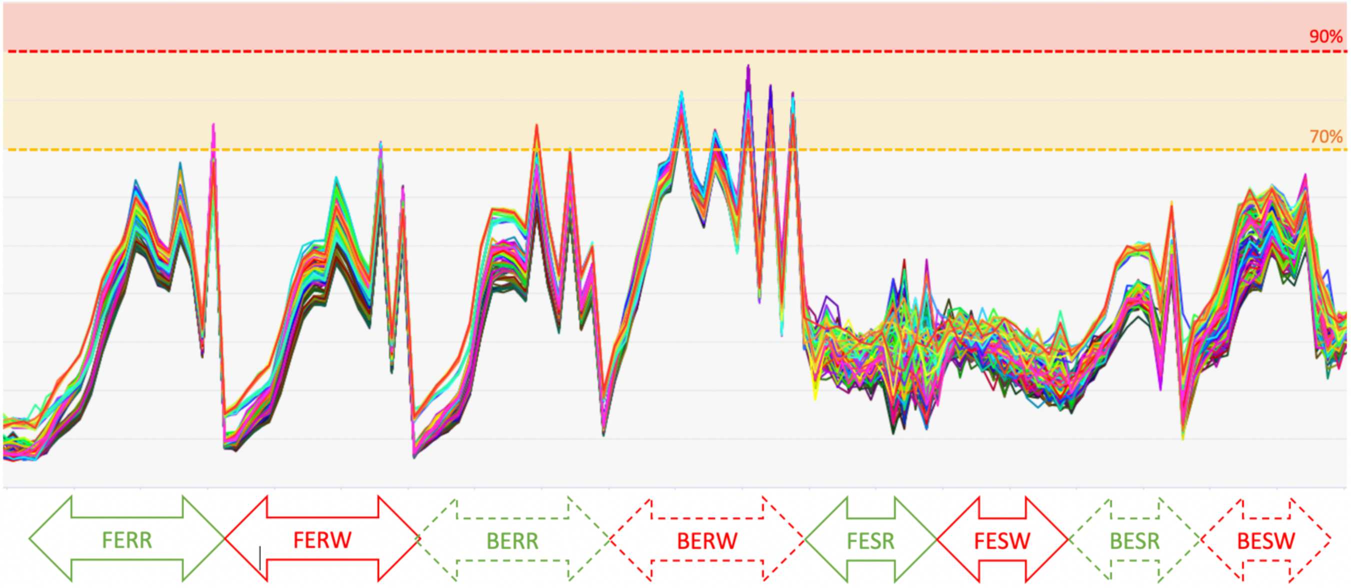


Figure 14 – Controller CPU utilisation during the tests

The storage system perfectly balanced the load of the individual controllers, all controllers being loaded at roughly the same level at all times: The controller loads reach 70 to 90% during the random tests with very high IOPS numbers at the peak. The highest CPU load was found during the backend random write (BERW) test. The controller load during the sequential IO tests with significantly larger block sizes was comparatively moderate (as one would expect) always remaining significantly below 70%.

VSP 5500 DeCo Rate

The combined deduplication and compression (DeCo) ratio was 1.85:1 after the tests were completed. This is even slightly above the expected value of 1.66:1.

Summary

The following table summarizes IOPS for Random IO, and throughput in GB/s for Sequential IO, together with the corresponding latency. The "Speed" column shows the value for one worker per node (32 workers in total), the "Optimium" column shows the optimal throughput (as long as the performance increases stronger than the latency) and the "Maximum" column shows the maximum throughput.

Test	Speed	Optimum	Maximum
8 KB Frontend Random Read	1 Worker 370'148 IOPS @ 85 µs	64 Worker 8'144'312 IOPS @ 250 µs	128 Worker 10'004'125 IOPS @ 408 µs
8 KB Backend Random Read	1 Worker 116'431 IOPS @ 273 µs	64 Worker 3'528'129 IOPS @ 571 µs	128 Worker 3'952'643 IOPS @ 1'039 µs
8 KB Frontend Random Write	1 Worker 97'411 IOPS @ 309 µs	128 Worker 3'518'020 IOPS @ 1'163 µs	128 Worker 3'518'020 IOPS @ 1'163 µs
8 KB Backend Random Write	1 Worker 95'543 IOPS @ 337 µs	16 Worker 524'136 IOPS @ 981 µs	128 Worker 748'997 IOPS @ 7'146 µs
1 MB Frontend Sequential Read	1 Process 23.5 GB/s @ 1'439 µs	32 Processes 205 GB/s @ 5'259 µs	48 Processes 212 GB/s @ 70625 µs
1 MB Backend Sequential Read	1 Process 23.7 GB/s @ 1'421 µs	16 Processes 99 GB/s @ 5'449 µs	48 Processes 111 GB/s @ 14'904 µs
1 MB Frontend Sequential Write	1 Process 13 GB/s @ 2'620 µs	8 Processes 43 GB/s @ 6'248 µs	40 Processes 48 GB/s @ 29'158 µs
1 MB Backend Sequential Write	1 Process 9 GB/s @ 3'688 µs	8 Processes 34 GB/s @ 8'189 µs	16 Processes 38 GB/s @ 14'763 µs

Table 2 – Key Performance Figures (Parallelism = Count Workers x 32 Servers)

The IOPS and GB/s values are combined values for both storage systems. The host-side writes are mirrored synchronously to the other storage at 50%. The effective number of writes to be processed is therefore 1.5 times higher.

Conclusion

The Hitachi Vantara VSP 5500 with 8 controllers pushes into new dimensions of IO performance (summed over both systems):

- 10 Mio. Frontend Random Read IOPS 8 KB
- 3.5 Mio. Frontend Random Write IOPS 8 KB (50% mirrored synchronously, effectively 5.25 Mio. IOPS)
- 200 GB/s Frontend Sequential Read
- 50 GB/s Frontend Sequential Write (50% mirrored synchronously, effectively 75 GB/s)
- 4 Mio. Backend Random Read IOPS 8 KB
- 750’000 Backend Random Write IOPS 8 KB (50% mirrored synchronously, effectively 1.125 Mio. IOPS)
- 100 GB/s Backend Sequential Read
- 40 GB/s Backend Sequential Write (50% mirrored synchronously, effectively 60 GB/s)

The write performance is limited by the number of data center interconnects (ISLs) and could possibly be increased. In addition, the system is still equipped with a relatively small number of NVMe disks. Last but not least the system can be expanded with 50% controller capacity.

Thanks to the NVMe backend the latency is at a very high performance level:

- 8 KB Reads from Cache < 100µs
- 8 KB Reads from Backend < 300µs
- 8 KB Writes to Cache 300-350µs (Average at 50% synchronously mirrored Writes).
  Local Writes below 200µs, synchronously mirrored Writes below 500µs.

With the VSP 5000 series Hitachi Vantara has a new top model in its portfolio that sets new performance standards. It remains very impressive that the performance was directly accessible without any major tuning.

Customer Voice

Brian Mathis
brian.mathis@suva.ch
Head of Server & Storage Platforms at SUVA

«We are very pleased with the performance of the Hitachi Vantara VSP 5500 Storage Systems and are entering new performance dimensions, with which SUVA is very well positioned for the years to come.

Thanks to the valuable support of In&Out AG, we were able to carry out the Storage Evaluation in a very targeted and efficient manner and achieved an optimal result, also in financial terms.»

About the Author

Andreas Zallmann,
andreas.zallmann@inout.ch
CEO In&Out AG,
Seestrasse 353, 8038 Zürich (Switzerland)
www.inout.ch

Andreas Zallmann studied computer science at the University of Karlsruhe and has been with In&Out AG since 2000. He is responsible for the Technology division and has been CEO of In&Out AG since 2016.
In&Out has years of practical experience in architecture, conception, benchmarking and tuning of storage and system platforms, especially for core applications for banks and insurance companies.

Andreas Zallmann is the developer of the In&Out Performance Benchmarking Tool IOgen™ and has carried out many customer and manufacturer benchmarks in recent years.

About peaq

The company peaq (www.peaq.ch) has been offering services and tools for many years that enable customers to get the most out of the exceptional performance of their Hitachi storage systems in a cost-effective manner. The peaq monitoring tool IOportal^{TM} optimizes the efficiency of the storage systems, while the rules-based provisioning tool SAM4H^{TM} provides fast and secure management of volumes and all sorts of simple to the most complex replications, especially GAD and 3DC.