IN&OUT AG

Huawei OceanStor Dorado 8000 V6 Benchmark

Andreas Zallmann
CEO, In&Out AG

Version:	1.1
Datum:	10.08.2020
Klassifikation:	Öffentlich

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

info@inout.ch, www.inout.ch

Vorbemerkung

Das vorliegende Whitepaper wurde im Auftrag der Firma Huawei unabhängig und neutral von der In&Out AG erstellt. Die Testumgebung wurde von Huawei Schweiz bereitgestellt.

Huawei

Huawei wurde 1987 gegründet und ist ein weltweit führender Anbieter von Informations- und Kommunikationstechnologie (ICT) Infrastruktur und intelligenten Geräten mit aktuell 194'000 Mitarbeitern. Seit einigen Jahren bietet Huawei Enterprise Storage Lösungen an und hat sich damit insbesondere in Europa etablieren und Marktanteile erobern können.

Huawei wird seit 2016 im Gartners Leader Quadranten für General Purpose Storage aufgeführt. Ab 2019 hat Gartner die Magic Quadrants für General Purpose und All Flash Storage in dem Magic Quadrant «Primary Storage» zusammengeführt. Auch hier wird Huawei im Leaders Quadranten geführt.

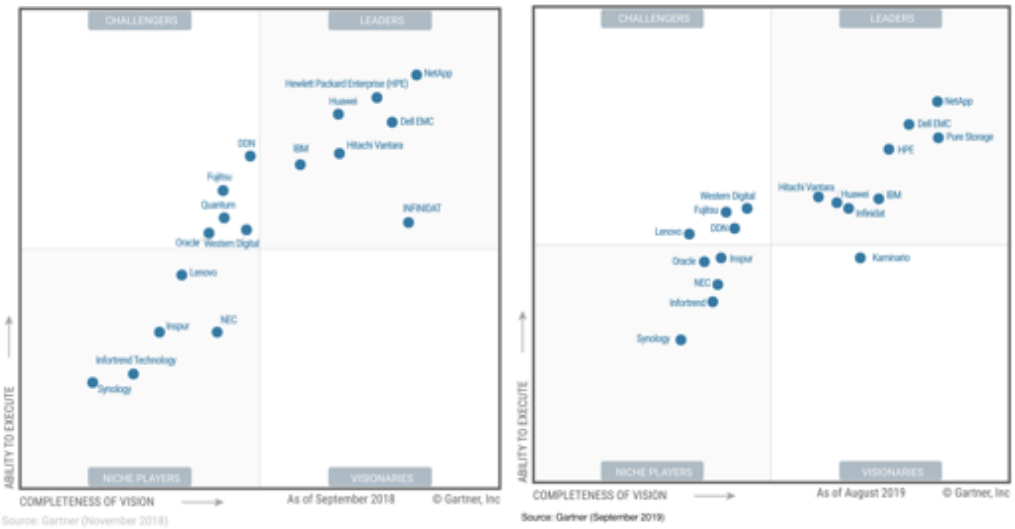


Abbildung 1 – Magic Quadrant «General Purpose Storage 2018» (links) und «Primary Storage 2019» (rechts)

Huawei OceanStor Dorado 8000 V6

Seit Anfang 2020 ist die neueste Generation der Allflash Produktreihe «Oceanstor Dorado V6» auf dem Markt. Die Dorado Systeme verfügen über eine «SmartMatrix» Full-Mesh Architektur, bei der voller Zugriff auf die Storage Objekte selbst bei Ausfall von 3 von 4 Controller eines Enclosures erhalten bleibt.

Die Dorado 8000 V6 ist dabei das zweitleistungsstärkste System, das mit bis zu 16 Controllern ausgestattet werden kann. Seit dieser Generation werden als Controller CPUs ARM Kunpeng 920 Prozessoren in 7nm Technologie mit 128 Cores und 2.6 GHz verwendet. Die Dorado 8000 V6 ist in einer SAS (Serial Attached SCSI) Variante und in einer state-of-the-Art NVMe Variante (Non-volatible Memory Express) erhältlich. Der maximale Ausbau der V8000 Systeme beträgt bis zu 3'200 SAS oder 800 NVMe SSDs und bis zu 16 TB Cache Memory. Im vorliegenden Test wurde die NVMe Variante getestet.

Abbildung 2 – Dorado 8000

In&Out AG

Die In&Out AG aus Zürich begleitet ihre Kunden als unabhängiges und herstellerneutrales Beratungsunternehmen seit Jahren in den Bereichen IT Infrastruktur und Datacenter mit besonderem Fokus auf Storage.

Neben Beratungsleistungen im Bereich Storage und Begleitung in Storage Ausschreibungen (auch GATT/WTO) verfügt In&Out über ausgewiesene jahrelange Storage Performance Erfahrung und hat das Benchmark Tool IOgen™ entwickelt.

Zielsetzung

Der Benchmark der neuen Huawei OceanStor Dorado 8000 V6 Systeme hatte zum Ziel, die Leistungsfähigkeit der Systeme zu zeigen, aber auch die hohe Ausfalltoleranz. So wurden während dem Test von den verfügbaren 8 Controllern in den beiden Storagesystemen bis zu 7 Controller entfernt. Dabei konnten die vorhandenen gespiegelten Storage-LUNs vollumfänglich genutzt werden.

Management Summary

Ein lokales 4 Controller Dorado 8000 V6 System kann folgende Leistungskennzahlen erreichen:

- 750'000 8KB Frontend Random Reads oder Writes,
  kombiniert Read und Write sogar 1.2 Mio. IOPS mit Latenzen im Teillastbereich von ca. 200 µs
- 650’000 8KB Backend Random IOPS mit Read Latenzen von 500 µs und Write Latenzen von 200 µs
- 6 GB/s Sequential Reads oder Writes im Frontend wie im Backend, im Frontend Read/Write sogar über 12 GB/s. Während diese Werte mit guten Latenzen auch im Backend erreicht werden, wird beim bidirektionalen Read/Write lediglich ein Durchsatz von 6 GB/s pro Storage erreicht. Die Latenz liegt bei moderater Last bei unter 500 µs.

Es ist aufgrund der Zahlen anzunehmen, dass ein vollausgebautes Dorado 8000 V6 System ca. 3 Mio. 8 KB Random Reads oder Writes verarbeitet oder bidirektional fast 5 Mio. 8 KB Random IOPS. Beim Durchsatz wären ca. 24 GB/s zu erwarten.

Dabei bleiben dank der SmartMatrix Architektur selbst bei Ausfall von drei der vier Controllern alle Storagepfade online und selbst **ein einzelner Controller** erreicht immer noch beeindruckende Werte:

- 500'000 Frontend IOPS
- 250’000 Backend IOPS
- 5 GB/s Durchsatz Frontend pro Richtung
- 2 GB/s Durchsatz Backend Read und 5 GB/S Durchsatz Backend Write, Read und Write 1.6 GB/s

Bei dem typischen Anwendungsfall der synchronen Spiegelung zeigt sich beim Lesen keinerlei Unterschiede zur ungespiegelten Performance. Bei **vollständiger Spiegelung** über 2 Storagesysteme werden folgende Schreib-Werte erreicht:

- 1 Mio. Frontend und Backend Random Reads 8 KB gespiegelt
- 800’000 Frontend und 560'000 Backend Random Writes 8 KB gespiegelt
- 1.4 Mio Frontend Random Reads/Writes und 750’000 Backend Random Reads/Writes 8 KB gespiegelt
- 9 GB/s Frontend und Backend Sequential Read gespiegelt
- 11 GB/s Frontend/Backend Sequential Write Durchsatz gespiegelt
- 20 GB/s Frontend Sequential Read/Write Durchsatz und 7.5 GB/s Backend Sequential Read/Write Durchsatz

Der Impact bei Ausfall von einzelnen Controllern in einem Storagesystem war für sequentielle und grosse IOPS vergleichsweise moderat. Bei kleinen zufälligen IOPS vor allem im Backend schlägt der Ausfall eines einzelnen Controllers bereits signifikant zu Buche, da dann bei zufällig verteilen IOPS der Zugriff auf die Disks hinter einem Controllerpaar zum Engpass werden kann. Dieser Effekt ist aber natürlich unvermeidlich, sofern die Controllerleistung der kritische Faktor ist.

Beeindruckend war, dass in jeder Situation (sogar bei Ausfall von 7 der 8 Controller) die LUNs immer online verfügbar blieben und bei Ausfall von 3 von 4 Controllern in einem System auch immer alle Pfade vom Server zum Storage online blieben.

Neben den Performancetests wurden noch verschiedene von Huawei vorgeschlagene funktionale Tests mit VMware ESX und Oracle erfolgreich durchgeführt.

Die Stabilität und das Verhalten der Systeme war in den Tests einwandfrei. Wir konnten keine Ausfälle oder unerklärlichen Performanceschwankungen feststellen. Die Bedienung der Systeme war selbst für Benutzer, die mit dem System überhaupt nicht vertraut waren, einfach und intuitiv möglich.

Huawei hat hier ein neues Top Modell im Portfolio, dass bereits mit 4 Controllern beeindruckende Performancekennzahlen bei tiefen Latenzen erreicht.

Testsetup

Getestet werden zwei Huawei Dorado 8000 V6 Storage Systeme mit einem Controller Enclosure und 4 Controllern, sowie einem NVMe Disk Enclosure mit 22 NVMe Disks.

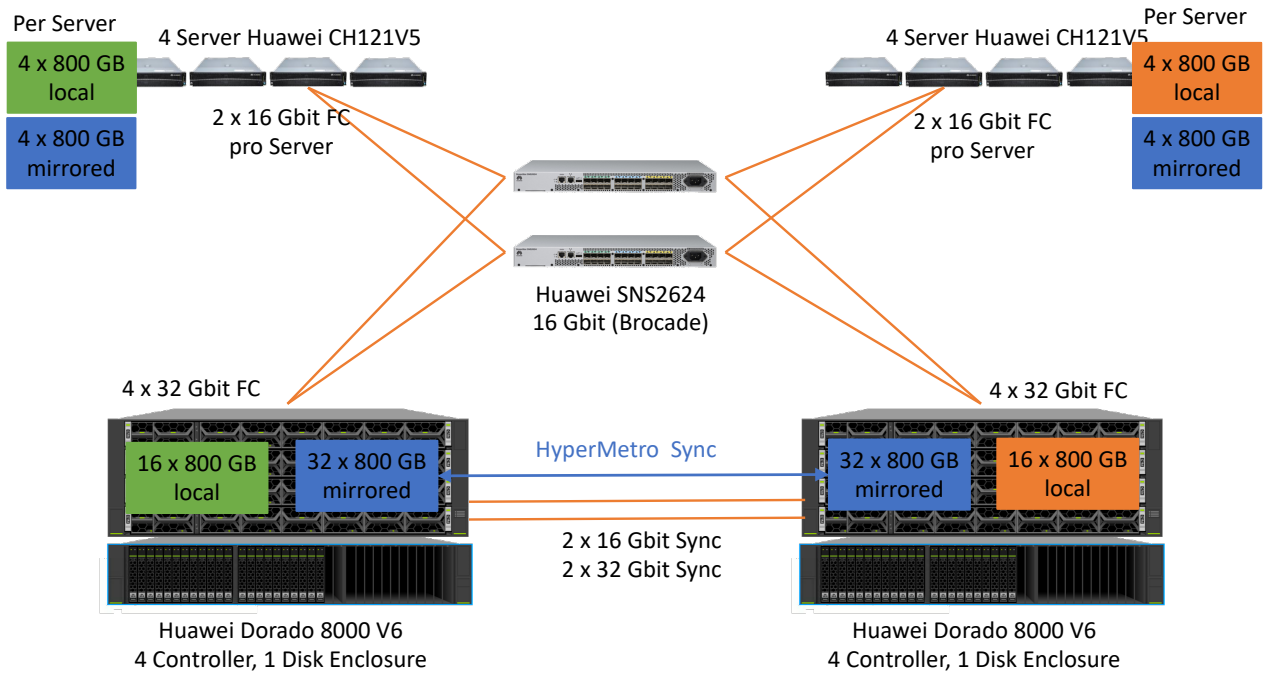


Abbildung 3 – Testsetup

Die beiden Storagesysteme sind mit 4 x 32 Gbit FC an zwei Huawei SNS2624 SAN Switches verbunden, die jedoch nur mit 16 Gbit FC Adaptern bestückt sind. Somit ist die nutzbare Geschwindigkeit 4 x 16 GBit = 8 GB/s pro Storagesystem. Die Systeme sind zur Spiegelung der Daten mit 2 x 32 Gbit FC und 2 x 16 Gbit FC direkt miteinander verbunden.

Pro Storagesystem sind 16 lokale (ungespiegelte) LUNs mit je 800 GB konfiguriert und 32 gespiegelte Devices mit je 800 GB. An jeden Server werden je 4 lokale als auch 4 gespiegelte LUNs gemappt. Jeder Server sieht die lokalen LUNs je viermal vom entsprechenden Storage und die gespiegelten LUNs je viermal von beiden Storagesystemen. Somit sind pro Server 4 x 4 = 16 lokale LUNs und 2 x 4 x 4 = 32 gespiegelte LUNs sichtbar.

Huawei OceanStor Dorado 8000 V6 Konfiguration

Die beiden Dorado 8000 V6 wurden in der folgenden Konfiguration genutzt:

	Getestete Konfiguration	Maximale Konfiguration
Storage	Huawei OceanStor Dorado V6 6.0.0.SPH7	
Firmware	7.60.03.011	
Controller	4 Controller (A-D)	16 Controller
Cache	1'024 GB insgesamt Read und Write	16'384 GB total (1'024 GB pro Controller)
FC	8 x 32 Gbit, effektiv 16 Gbit (4 für Server, 4 für Storagespiegelung)	448 x 32 Gbit (112 IO Slots oder 28 pro Controller Enclosure mit je 4 x 32 Gbit Ports)
Disks	1 NVMe Disk Enclosure mit 22 NVMe SSD x 3.492 TB, 76 TB Rohkapazität	Bis zu 23 NVMe Disks Enclosures mit bis zu 3'200 SAS oder NVMe SSD
Raid-Konfiguration	Virtual Raid 2.0+ mit Raid-6 Dualprotection und verteiltem Hotspare (Policy Low), 60 TB Nutzkapazität Die Disk Enclosures werden über vier 100 Gbit/s RDMA Ports mit den Controllern verbunden und verfügen über eigene ARM CPUs und Memory für das Offloading bestimmter Aktivitäten.	
LUN Konfiguration	16 LUNs lokal (ungespiegelt) x 800 GB = 12.8 TB 32 LUNs synchron gespiegelt auf zweites System x 800 GB = 25.6 TB Effektiv genutzt 38.4 TB, Initialisiert und beschrieben mit zufälligen Mustern	

Tabelle 1 – Systemkonfiguration je Storage

Inline Deduplication und inline compression sind aktiviert die Disk encryption ist jedoch nicht aktiviert.

```
developer:/>show lun_workload_type general id=0

  ID                    : 0
  Name                  : Default
  IO Size               : 8KB
  Compression Enabled   : --
  Dedup Enabled         : --
  Type                  : Reserved
```

Test Server

Bei den Servern handelt es sich um Huawei Blade Server CH121V5 mit 2 Sockets Xeon Gold 5120 mit 2 x 14 Cores bei 2.2 Ghz Taktfrequenz und 320 bis 448 GB Memory. Pro Blade Server sind 2 x 16 GBit FC konfiguriert, somit ist eine theoretische Bandbreite von 4 GB/s pro Server und insgesamt von 32 GB/s (parallel lesen und schreiben) möglich. Die Blades befinden sich in einem Huawei Blade Chassis E9000.

Auf den Servern war jeweils Red Hat Enterprise Linux (RHEL) 7.6 installiert.

IOgen™

Die Messungen erfolgen mit IOgen™ 6.3.1 der In&Out AG. IOgen wurde auf RHEL 7.6 installiert. Das integrierte Multipathing von RedHat wurde nicht benutzt, sondern IOgen hat direkt auf den mehrfach sichtbaren Raw Devices /devs/sdx gearbeitet und damit automatisch alle verfügbaren Kanäle verwendet.

Testläufe

Die Testläufe wurden durch IOgen™ parallel und zeitlich auf die Millisekunde synchronisiert auf bis zu 8 Servern durchgeführt.

Es wurden Tests auf lokalen LUNs (ungespiegelt) und auf den gespiegelten LUNs durchgeführt. In einem Mixed Szenario wurden die lokalen und die gespiegelten LUNs mit gleicher Wahrscheinlichkeit von je 50% benutzt.

Die Haupttestszenarien sind:

I. Test Lokal: Lokale LUNs von einem Storage mit 1, 2, 4 Servern und von beide Storages mit 8 Servern
II. Test Verfügbarkeit Lokal: Lokale LUNs von einem Storage mit 4 Servern bei Ausfall von 1,2, oder 3 Controllern
III. Test Spiegelung: Gespiegelte, Gemische und Lokale LUNs von beiden Storages und je 8 Server
IV. Test Verfügbarkeit Spiegelung: Gespiegelte LUNs von beiden Storages bei Ausfall von 1,2,3 oder 7 Controllern

Folgende Tests wurden durchgeführt:

a. Frontend Random 8 KB
b. Backend Random 8 KB
c. Frontend Sequential 128 KB
d. Backend Sequential 128 KB

Für alle Tests werden reine **Lesetests** (Read), reine **Schreibtests** (Write) und gemischte **Lese/Schreibtests** (Read/Write) durchgeführt, bei denen 50% gelesen und 50% geschrieben wird. Bei allen Schreibvorgängen wird ein «pseudozufälliges» Schreibmuster verwendet, welches kaum komprimierbar ist.

Frontend (FE) Tests werden gegen das Frontend und den Storage Cache gefahren. Hier werden pro LUN immer nur die ersten 1000 Blöcke verwendet, die Cache Hit Ratio beträgt deshalb annähernd 100%. Bei den ***Backend (BE)*** Tests werden hingegen alle Blöcke des LUNs verwendet, die Cache Hit Ratio liegt bei nahe 0%.

Für Random IOs werden die IOPS (IOs pro Sekunde) und für Sequential IOs der Durchsatz in MB/s (Megabyte pro Sekunde) ermittelt und jeweils für alle Tests die auf dem Server gemessene Latenz (Latency) in µs (Microsekunden) = 0.001 ms. In allen Grafiken wird die IO Leistung (IOPS oder MB/s) links und die Latenz rechts dargestellt. Es werden jeweils immer die Grafiken für Read, Write und Read/Write untereinander aufgeführt. Auf der X-Achse ist jeweils die Parallelität (Anzahl Worker Prozesse pro Server) aufgetragen.

I. Test Lokal

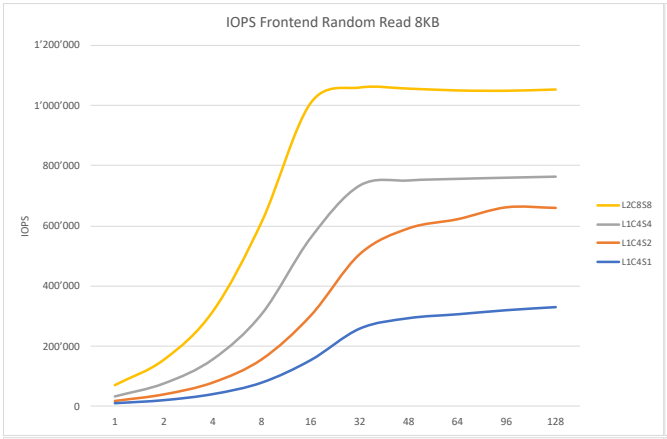
Bei den lokalen Tests werden die lokalen, ungespiegelten LUNs verwendet. Dabei wurden folgende vier Tests durchgeführt:

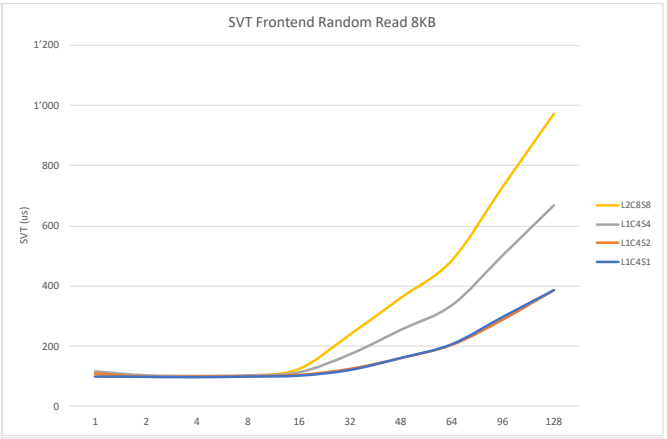
- L1C4S1 (blau) – Lokaler Test mit 1 Storage, 4 Controllern und 1 Server (theoretische Bandbreite ca. 4 GB/s)
- L1C4S2 (orange) – Lokaler Test mit 1 Storage, 4 Controllern und 2 Servern (theoretische Bandbreite ca. 8 GB/s)
- L1C4S4 (grau) – Lokaler Test mit 1 Storage, 4 Controllern und 4 Servern (theoretische Bandbreite ca. 8 GB/s)
- L2C8S8 (gelb) – Lokaler Test mit 2 Storages, 8 Controllern und 8 Servern (theoretische Bandbreite ca. 16 GB/s)

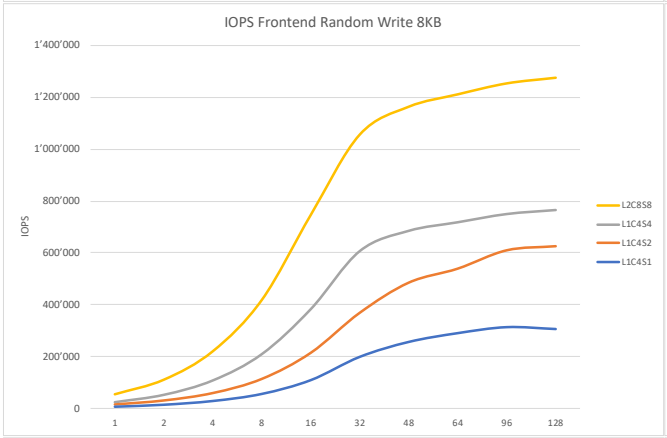
Die beiden letzten Tests sind durch die maximale Storageanbindung bei 8 GB/s pro Storage limitiert (4 x 16 Gbit am SAN Switch).

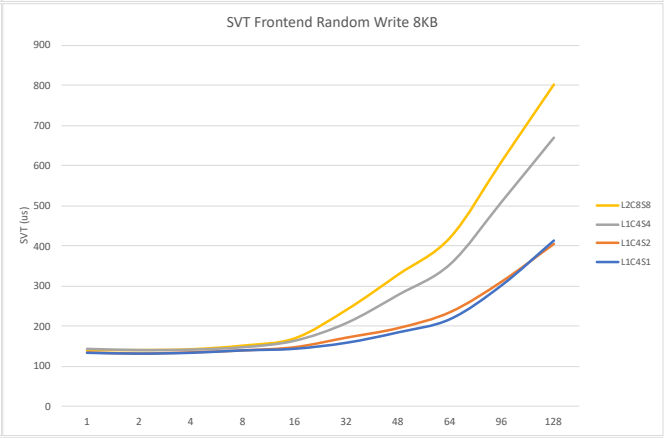
In dieser Testreihe fokussieren wir auf den Optimalwert, bei dem der Durchsatz stärker steigt als die Latenz. Bis dahin bewegen wir uns sozusagen im «gesunden» Bereich. Der Optimalwert für IOPS und die MB/s wird jeweils fett dargestellt, zusammen mit der bei diesen Werten gemessen Latenz am Server. Darunter wird jeweils der Maximalwert (die Latenz ist für diesen Wert aufgrund der hohen Last nicht mehr aussagekräftig) und die minimale Latenz bei geringer Last dargestellt.

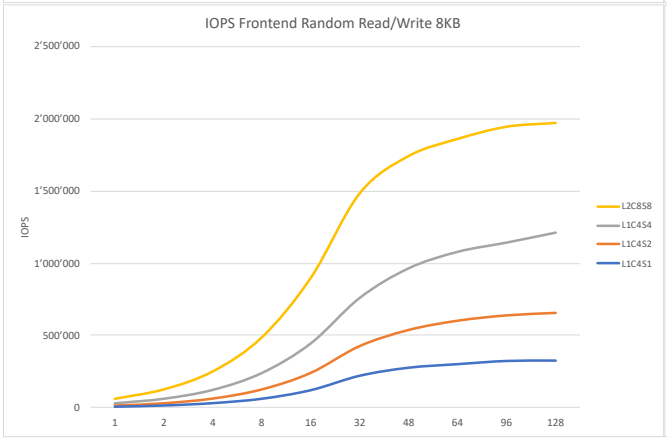
Ia. Lokal Frontend Random 8 KB

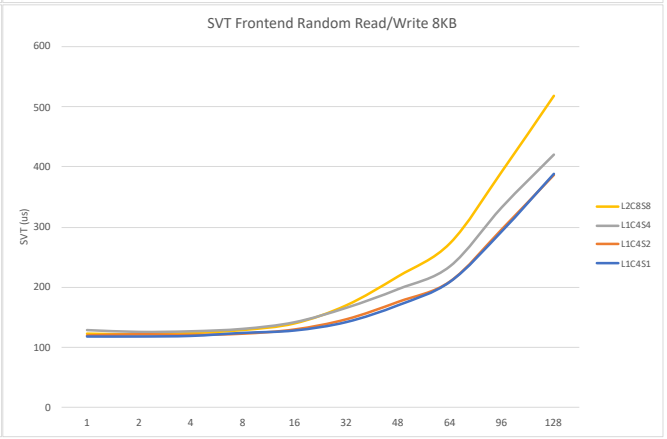












Die Cached Frontend IOPS sind pro Server mit 2 FC Adaptern 16 Gbit auf 300'000 IOPS (blau) beschränkt, mit zwei Servern steigt diese Zahl auf gut 600'000 IOPS (orange) und bei vier Servern wird mit 760'000 IOPS (grau) das Limit der getesteten Storagekonfiguration erreicht. Beim Read/Write Szenario können bidirektional 1'200'000 IOPS erreicht werden. Mit zwei Storages und acht Servern können mehr als 1 Mio. IOPS, bidirektional fast 2 Mio. IOPS (gelb) erreicht werden. Die Latenz liegt bei geringer Last bei exzellenten 100 µs und steigt dann zum Optimalwert auf ca. 200-300 µs.

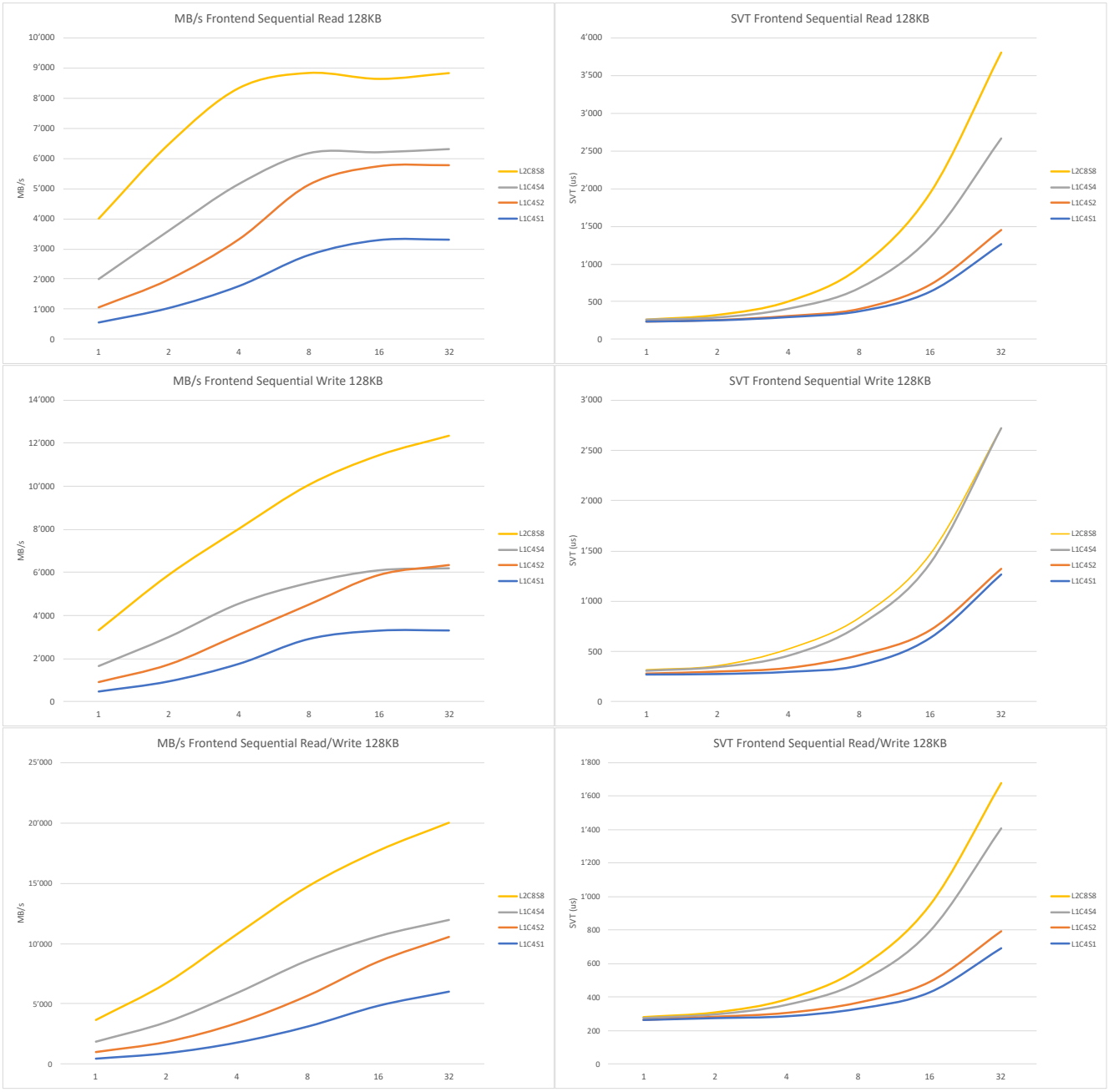
Test	Read	Write	Read/Write
L1C4S1 (blau)	**304 kIOPS @ 208µs**	**291 kIOPS @ 217µs**	**303 kIOPS @ 209µs**
1 Storage, 4 Cntr, 1 Server	328 kIOPS Max, 98µs Min	315 kIOPS Max, 131µs Min	327 kIOPS Max, 118µs Min
L1C4S2 (orange)	**622 kIOPS @ 204µs**	**540 kIOPS @ 235µs**	**604 kIOPS @ 210µs**
1 Storage, 4 Cntr, 2 Server	661 kIOPS Max, 98µs Min	627 kIOPS Max, 131µs Min	658 kIOPS Max, 120µs Min
L1C4S4 (grau)	**756 kIOPS @ 336µs**	**716 kIOPS @ 355µs**	**1'079 kIOPS @ 235µs**
1 Storage, 4 Cntr, 4 Server	764 kIOPS Max, 100µs Min	763 kIOPS Max, 142µs Min	1'212 kIOPS Max, 126µs Min
L2C8S (gelb)	**1'051 kIOPS @ 485µs**	**1'253 kIOPS @ 611 µs**	**1'743 kIOPS @ 218µs**
2 Storage, 8 Cntr, 8 Server	1'057 kIOPS Max, 99 µs Min	1'275 kIOPS Max, 140µs Min	1'972 kIOPS Max, 122µs Min

Ib. Lokal Backend Random 8 KB

Da die Latenz beim Lesen aus dem Backend mit ca. 600 µs deutlich höher als bei Cache Reads mit 200 µs ist, ist der Read Durchsatz von 220'000 IOPS mit 1 Server etwas geringer als beim Frontend mit 300'000 IOPS. Mit einer grösseren Anzahl Servern gleicht sich diese Zahl immer mehr an. Die Write Performance und Latenz ist hingegen vergleichbar, da der Write immer in den Cache erfolgt.

Test	Read	Write	Read/Write
L1C4S1 (blau)	**220 kIOPS @ 581µs**	**300 kIOPS @ 317µs**	**267 kIOPS @ 477µs**
1 Storage, 4 Cntr, 1 Server	**220** kIOPS Max, 558µs Min	310 kIOPS Max, 135µs Min	267 kIOPS Max, 352µs Min
L1C4S2 (orange)	**392 kIOPS @ 652µs**	**536 kIOPS @ 356µs**	**444 kIOPS @ 575µs**
1 Storage, 4 Cntr, 2 Server	392 kIOPS Max, 577µs Min	536 kIOPS Max, 139µs Min	444 kIOPS Max, 398µs Min
L1C4S4 (grau)	**634 kIOPS @ 802µs**	**638 kIOPS @ 800µs**	**621 kIOPS @ 824µs**
1 Storage, 4 Cntr, 4 Server	634 kIOPS Max, 630µs Min	638 kIOPS Max, 143µs Min	**621** kIOPS Max, 409µs Min
L2C8S (gelb)	**1'020 kIOPS @ 751µs**	**1'033 kIOPS @ 370 µs**	**1'125 kIOPS @ 682µs**
2 Storage, 8 Cntr, 8 Server	1'047 kIOPS Max, 595 µs Min	1'112 kIOPS Max, 141µs Min	1'253 kIOPS Max, 392µs Min

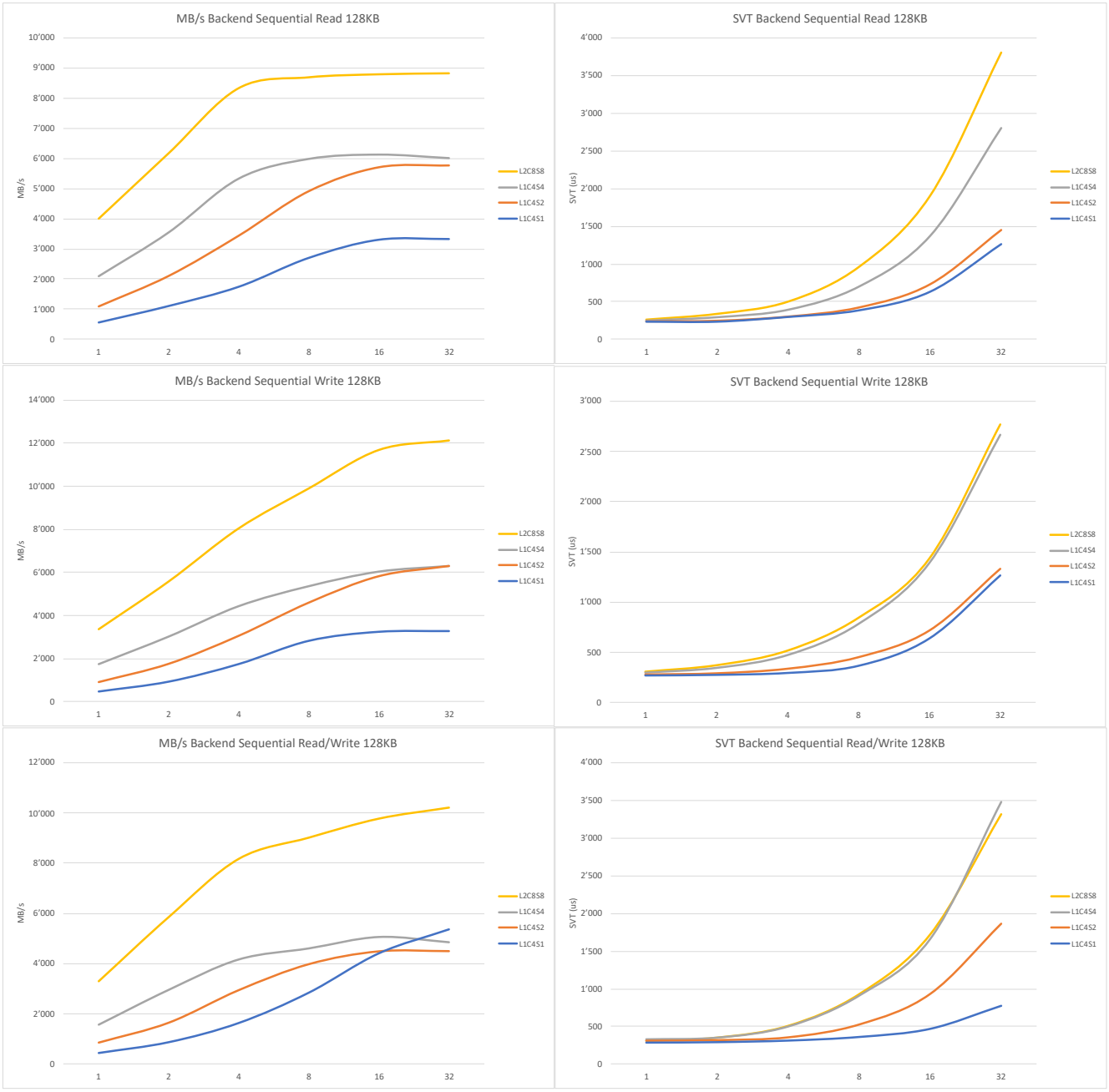
Ic. Lokal Frontend Sequential 128 KB



Beim Lesen und Schreiben im Frontend von einem Server kann von der theoretischen Bandbreite von 4 GB/s mit bis zu 3.3 GB/s mehr als 80% erreicht werden. Bei bidirektionalen Lesen und Schreiben können über 6 GB/s erreicht werden. Bei zwei Servern verdoppeln sich diese Durchsätze fast, bei 4 Servern wird der maximale Durchsatz pro Storage bei gut 6 GB/s oder 12 GB/s bidirektional erreicht. Bei Einsatz beider Storagesyteme können die Werte auf 9 GB/s Read, 12 GB/s Write und 20 GB/s Read/Write nochmals massiv gesteigert werden. Die Latenz liegt bei moderater Last bei unter 500 µs. Dabei werden bis zu 75% der theoretisch verfügbaren Bandbreite genutzt.

Test	Read	Write	Read/Write
L1C4S1 (blau)	**2'800 MBs @ 373µs**	**2'908 MBs @ 359µs**	**4'880 MBs @ 428µs**
1 Storage, 4 Cntr, 1 Server	3'314 MBs Max, 237µs Min	3'303 MBs Max, 269µs Min	6'052 MBs Max, 263µs Min
L1C4S2 (orange)	**5'117 MBs @ 408µs**	**5'882 MBs @ 712µs**	**10'579 MBs @ 792µs**
1 Storage, 4 Cntr, 2 Server	5'761 MBs Max, 242µs Min	6'341 MBs Max, 279µs Min	10'579 MBs Max, 265µs Min
L1C4S4 (grau)	**6'166 MBs @ 679µs**	**5'501 MBs @ 761µs**	**10'575 MBs @ 792µs**
1 Storage, 4 Cntr, 4 Server	6'301 MBs Max, 261µs Min	6'175 MBs Max, 311µs Min	11'921 MBs Max, 275µs Min
L2C8S (gelb)	**8'331 MBs @ 949µs**	**12'335 MBs @ 2'723µs**	**17'685 MBs @ 948µs**
2 Storage, 8 Cntr, 8 Server	8'831 MBs Max, 261µs Min	12'335 MBs Max, 315µs Min	19'999 MBs Max, 281µs Min

Id. Lokal Backend Sequential 128 KB



Der Durchsatz im Backend ist sowohl beim Lesen wie beim Schreiben vergleichbar mit dem Frontend Durchsatz. Beim Lesen können vermutlich dank Prefetching auch sehr gute Latenzen erreicht werden, die nur wenig über den gecachten Latenzen liegen. Auffällig ist allerdings, dass beim gleichzeitigen Lesen und Schreiben der Durchsatz im Vergleich zu den Frontend Zahlen deutlich geringer ist.

Test	Read	Write	Read/Write
L1C4S1 (blau)	**2'688 MBs @ 389µs**	**2'847 MBs @ 367µs**	**4'417 MBs @ 474µs**
1 Storage, 4 Cntr, 1 Server	3'312 MBs Max, 238µs Min	3'300 MBs Max, 271µs Min	5'371 MBs Max, 291µs Min
L1C4S2 (orange)	**4'918 MBs @ 425µs**	**5'834 MBs @ 718µs**	**3'982 MBs @ 526µs**
1 Storage, 4 Cntr, 2 Server	5'711 MBs Max, 240µs Min	6'307 MBs Max, 283µs Min	4'502 MBs Max, 310µs Min
L1C4S4 (grau)	**5'969 MBs @ 701µs**	**5'376 MBs @ 779µs**	**4'612 MBs @ 910µs**
1 Storage, 4 Cntr, 4 Server	6'112 MBs Max, 249µs Min	6'313 MBs Max, 296µs Min	5'066 MBs Max, 331µs Min
L2C8S (gelb)	**8'328 MBs @ 502µs**	**9'896 MBs @ 846µs**	**9'010 MBs @ 931µs**
2 Storage, 8 Cntr, 8 Server	8'819 MBs Max, 262µs Min	12'123 MBs Max, 311µs Min	10'201 MBs Max, 315µs Min

I. Zusammenfassung

Ein lokales 4 Controller Dorado 8000 V6 System kann folgende Leistungskennzahlen erreichen:

- 750'000 8KB Frontend Random Reads oder Writes,
  kombiniert Read und Write sogar 1.2 Mio. IOPS mit Latenzen im Teillastbereich von ca. 200 µs
- 650’000 8KB Backend Random Reads, Writes oder kombiniert Read/Write mit Read Latenzen von 500 µs und Write Latenzen von 200 µs
- 6 GB/s Sequential Reads oder Writes im Frontend wie im Backend, im Frontend Read/Write sogar über 12 GB/s. Bidirektionaler Read/Write Durchsatz von 6 GB/s. Die Latenz liegt bei moderater Last bei unter 500 µs.

II. Test Verfügbarkeit lokal

Die getesteten Systeme verfügen über 4 Controller in einem Controller Enclosure. Theoretisch können bis zu 4 Enclosures mit 16 Controllern zu einem Storagesystem zusammengeschlossen werden. Dabei bilden aber immer 4 Controller eine Einheit, die gemeinsamen Zugriff auf alle NVMe Disks hat.

Wir haben deshalb im laufenden Betrieb zunächst einen Controller dann zwei und schliesslich drei Controller entfernt und den Einfluss auf die Performance gemessen.


L1C4S4 (grau), Baseline


L1C3S4 (blau), 3 Controller


L1C2S4 (grün), 2 Controller


L1C1S4 (orange), 1 Controller

Abbildung 4 – Tests II - Darstellung Huawei Device Manager

Im Einzelnen wurden die folgenden Tests durchgeführt:

- L1C4S4 (grau) – Lokaler Test mit 1 Storage, 4 Controllern und 4 Servern, **Baseline**
- L1C4S4 (blau) – Lokaler Test mit 1 Storage, 3 Controllern und 4 Servern
- L1C4S4 (grün) – Lokaler Test mit 1 Storage, 2 Controllern und 4 Servern
- L2C8S4 (orange) – Lokaler Test mit 1 Storage, 1 Controller und 4 Servern

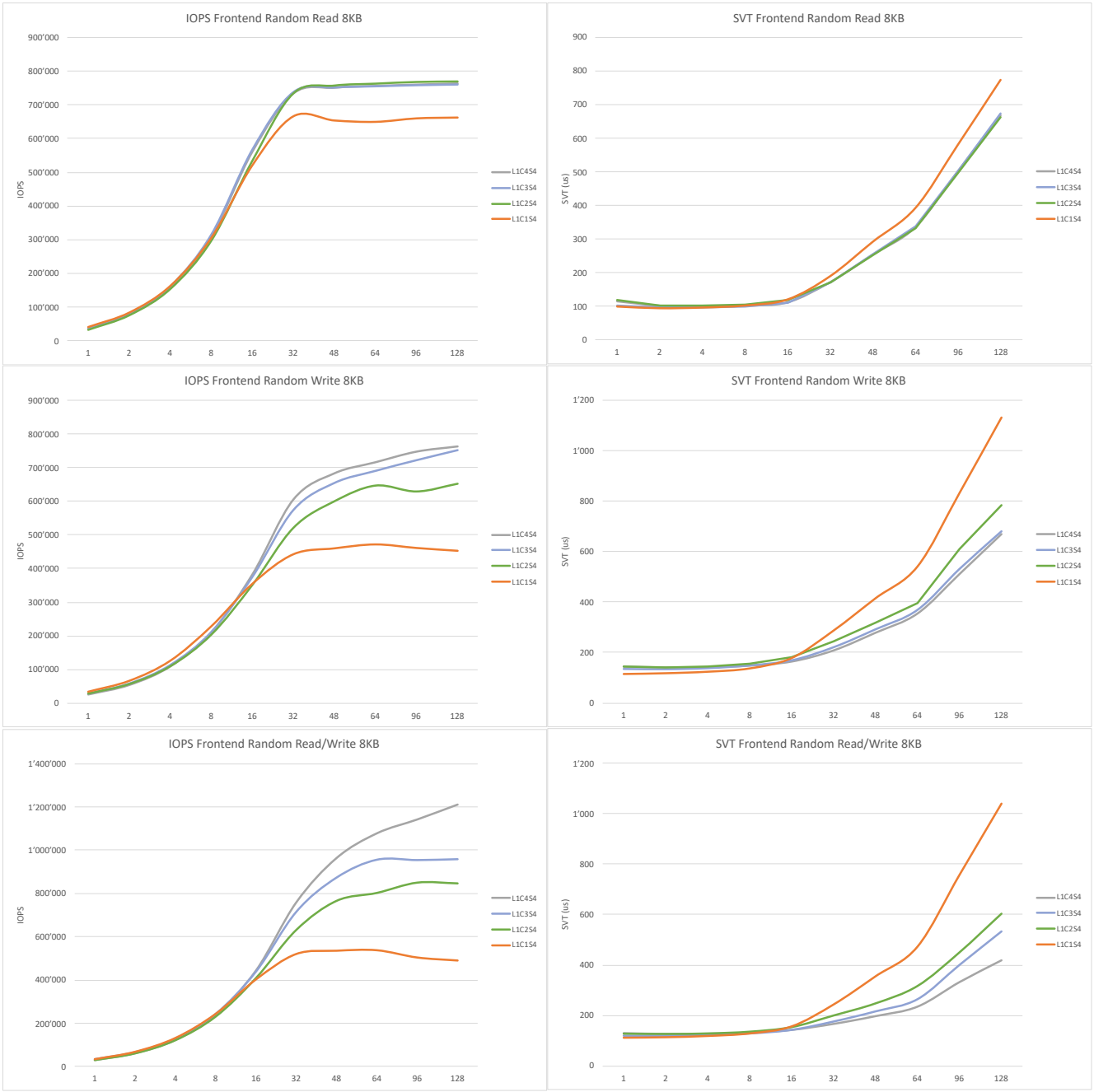
Der Test L1C4S4 (grau) dient dabei als Baseline für die Performance mit allen vier Controllern und wurde schon im vorangehenden Testrun I. beschrieben.

Bemerkenswert ist, dass in der Tat bei Ausfall / Entfernen von drei der vier Controller immer noch alle LUNs auf allen FC Pfaden für den Server sichtbar sind. Es geht also auch in diesem Fall keiner der Pfade zum Server offline.

```
mpathe (36a400e2100b8c87c02ae6f1700000000) dm-7 HUAWEI  ,XSG1
size=801G features='0' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=0 status=active
  |- 15:0:0:1 sdb  8:16   active undef running
  |- 16:0:4:1 sdz  65:144 active undef running
  |- 15:0:3:1 sdr  65:16  active undef running
  `- 16:0:7:1 sdap 66:144 active undef running
```

In der Tabelle wird für dieses Testszenario jeweils der Höchstwert pro Test angegeben und der resultierte Impact durch Ausfall der Controller in Prozent. Dabei ist ein gleichzeitiger Ausfall von mehr als einem Controller in einem Enclosure ein sehr unwahrscheinliches Ereignis.

IIa. Lokal Frontend Random 8 KB



Es ist klar erkennbar, dass die Frontend Reads selbst mit einem Controller fast vollständig abgewickelt werden können. Hier ist der Workload am geringsten, da weder ein Zugriff auf das Storage Backend noch die Kalkulation einer Checksum notwendig ist.

Beides ist beim Schreiben hingegen notwendig und somit ist dort auch bei gemischten Schreib/Lese Vorgängen eine Reduktion der IOPS festzustellen. Bei Halbierung der Controller beträgt die Leistungreduktion maximal 30%, bei nur noch einem statt vier Controllern wird immer noch mehr als die Hälfte der IOPS erreicht (anstelle eines erwarteten Rückgangs von 75%).

Bemerkenswert ist, dass bei geringer Last kein Unterschied in der Latenz festzustellen ist, diese steigt dann aber mit zunehmender Last mit weniger Controllern schneller an.

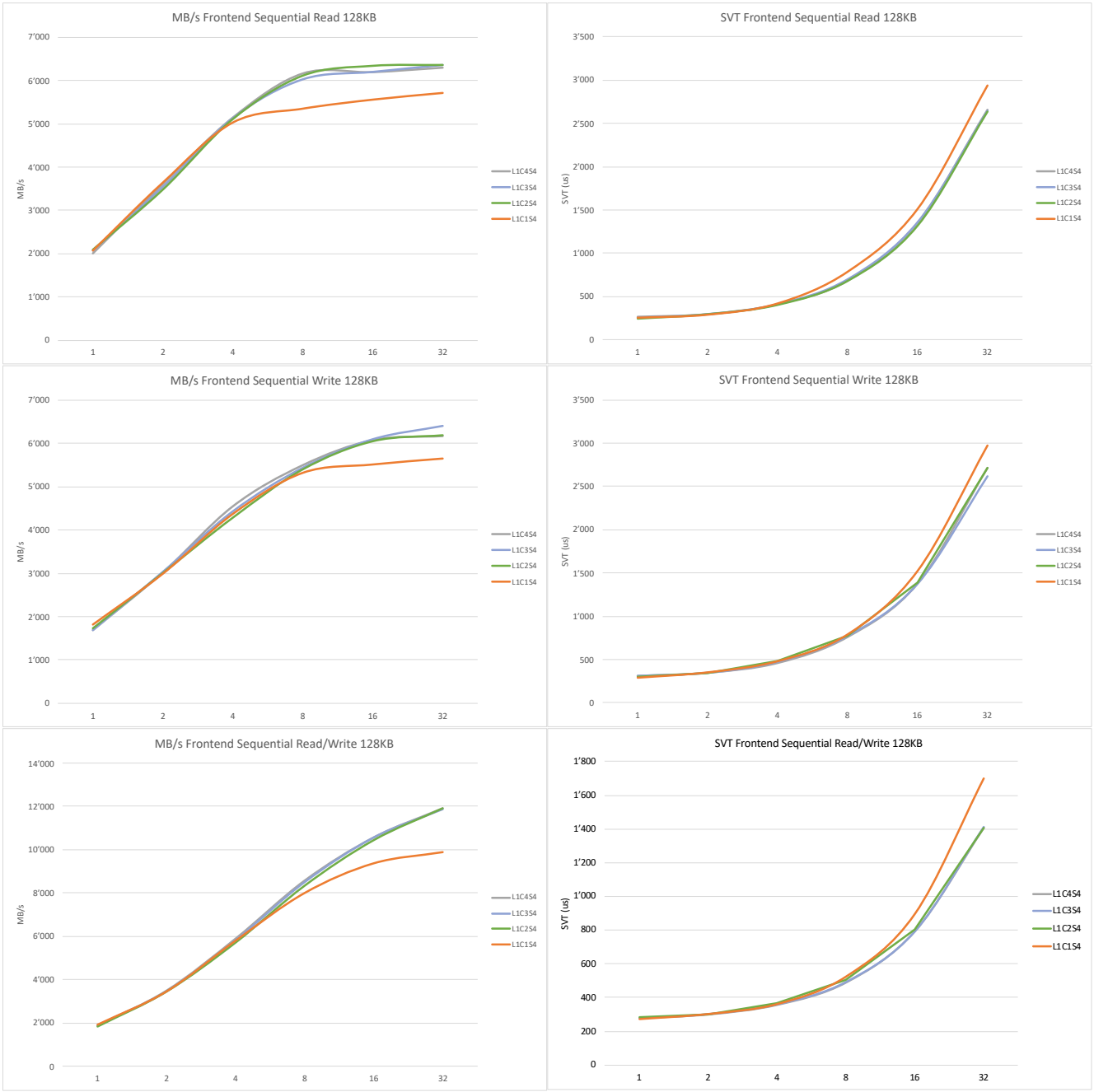
Test	Read	Write	Read/Write
L1C4S4 (grau) 1 Storage, 4 Cntr, 4 Server	763 kIOPS Max	763 kIOPS Max	1'212 kIOPS Max
L1C3S4 (blau) -25% 1 Storage, 3 Cntr, 4 Server	759 kIOPS Max **-0%**	751 kIOPS Max **-2%**	960 kIOPS Max **-21%**
L1C2S4 (grün) -50% 1 Storage, 2 Cntr, 4 Server	769 kIOPS Max **-0%**	651 kIOPS Max **-15%**	850 kIOPS Max **-30%**
L1C4S4 (orange) -75% 1 Storage, 1 Cntr, 4 Server	662 kIOPS Max **-13%**	470 kIOPS Max **-38%**	539 kIOPS Max **-46%**

IIb. Lokal Backend Random 8 KB

Bei Backend IO ist der Impact eines Controllerausfalls gravierender. Da jeweils zwei Controller beim Backend-Zugriff ein Paar bilden, wird das um einen Controller reduzierte Paar zum Engpass bei den Backend IOPS. Der Performanceimpact bei Ausfall von einem Controller in einem Paar oder von zwei Controllern in beiden Paaren ist ähnlich hoch. Eine erneute deutliche Reduktion ist messbar, wenn nur noch ein Controller im Einsatz ist. Der Impact bei Ausfall von 3 der 4 Controllern liegt bei maximal 61% anstelle eines erwarteten Rückgangs von 75%. Bei Backend Reads ist der Ausfall eines Controllers auch bei geringer Last mit einer leichten Erhöhung der Latenz einhergehend. Beim sehr unwahrscheinlichen Szenario mit nur 1 verbleibenden Controller (orange) ist im Mixed Workload (siehe Grafik unten rechts) eine signifikante Erhöhung der Latenz feststellbar.

Test	Read	Write	Read/Write
L1C4S4 (grau) 1 Storage, 4 Cntr, 4 Server	634 kIOPS Max	638 kIOPS Max	621 kIOPS Max
L1C3S4 (blau) -25% 1 Storage, 3 Cntr, 4 Server	508 kIOPS Max **-20%**	508 kIOPS Max **-20%**	510 kIOPS Max **-18%**
L1C2S4 (grün) -50% 1 Storage, 2 Cntr, 4 Server	460 kIOPS Max **-27%**	478 kIOPS Max **-25%**	453 kIOPS Max **-27%**
L1C4S4 (orange) -75% 1 Storage, 1 Cntr, 4 Server	245 kIOPS Max **-61%**	394 kIOPS Max **-38%**	222 kIOPS Max **-54%**

IIc. Lokal Frontend Sequential 128 KB

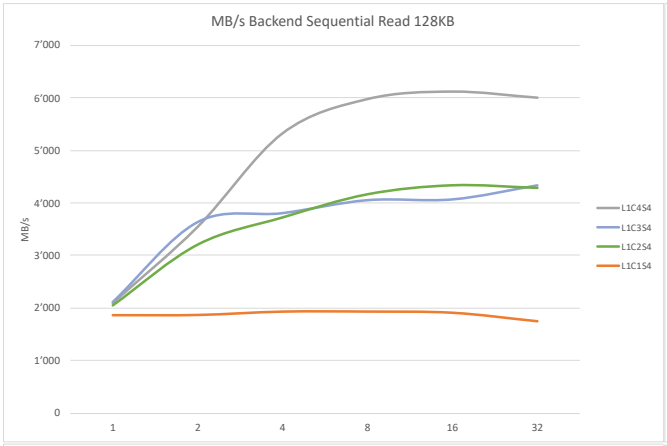


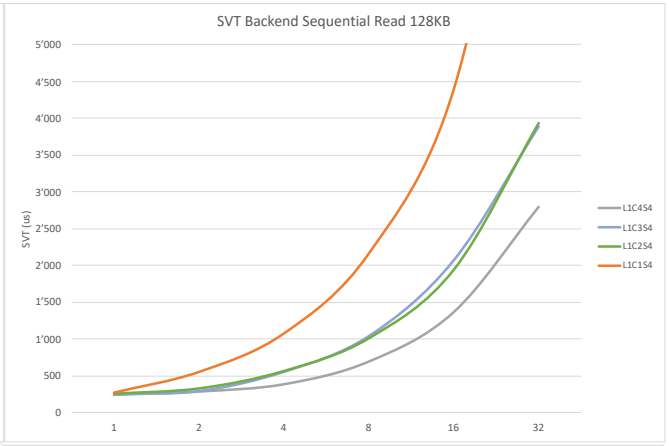
Da bei diesem Workload bei einer Blocksize von 128 KB die Anzahl der IOPS wesentlich geringer ist, kann der Sequential Workload auch bei Ausfall von bis zu zwei Controllern problemlos und ohne Performanceimpact abgewickelt werden. Erst bei Ausfall von drei Controllern zeigt sich ein leichter Impact von maximal 17%. Bei Frontend IOPS werden immer die gleichen Blöcke überschrieben und die Controller müssen keine Checksums berechnen.

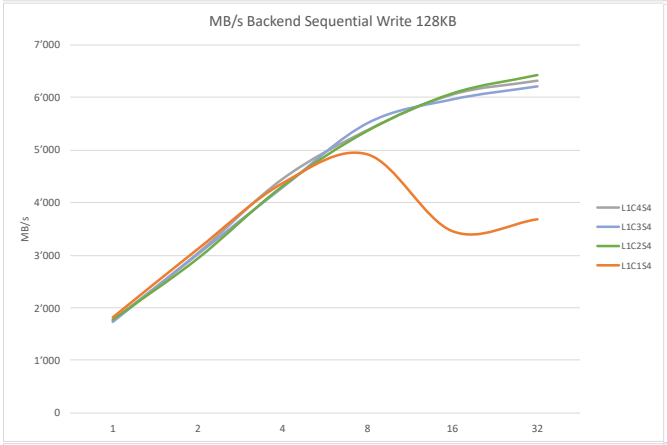
Bemerkenswert ist, dass kaum Unterschiede in der Latenz festzustellen ist.

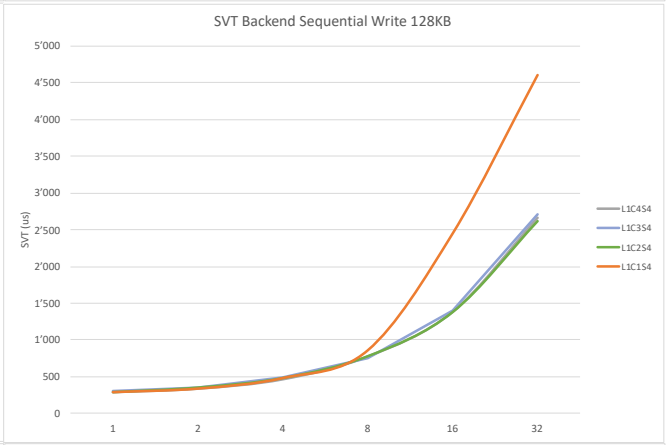
Test	Read	Write	Read/Write
L1C4S4 (grau) 1 Storage, 4 Cntr, 4 Server	6'301 MBs Max	6'175 MBs Max	11'920 MBs Max
L1C3S4 (blau) -25% 1 Storage, 3 Cntr, 4 Server	6'350 MBs Max **-0%**	6'407 MBs Max **-0%**	11'862 MBs Max. **-0%**
L1C2S4 (grün) -50% 1 Storage, 2 Cntr, 4 Server	6'360 MBs Max **-0%**	6'194 MBs Max **-0%**	11'906 MBs Max. **-0%**
L1C4S4 (orange) -75% 1 Storage, 1 Cntr, 4 Server	5'707 MBs Max **-9%**	50654 MBs Max **-8%**	9'869 MBs Max **-17%**

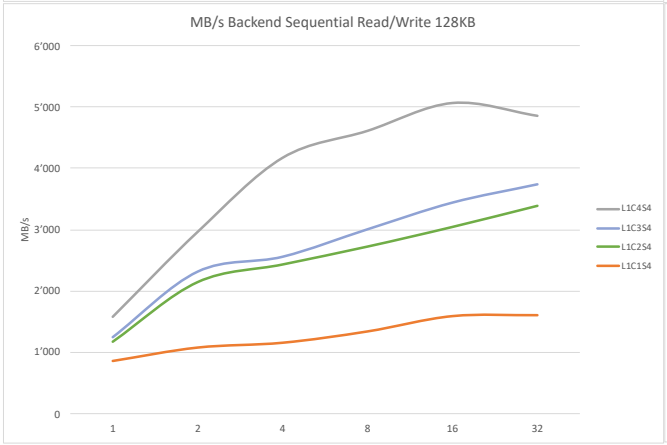
IId. Lokal Backend Sequential 128 KB

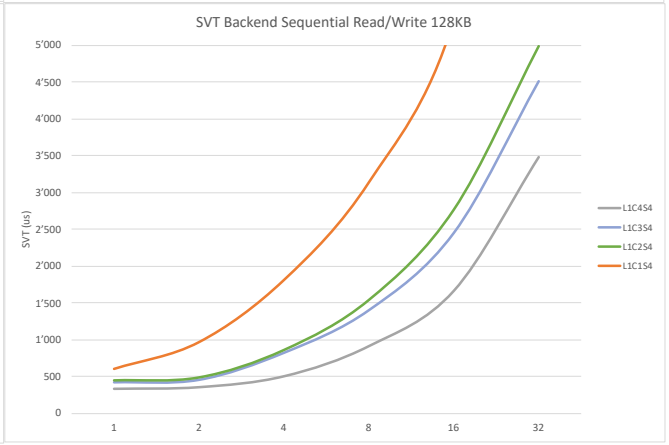












Bei Backend IO ist der Impact eines Controllerausfalls gravierender. Da jeweils zwei Controller beim Backend-Zugriff ein Paar bilden, wird das um einen Controller reduzierte Paar zum Engpass bei den Backend IOPS. Der Performanceimpact bei Ausfall von einem Controller in einem Paar oder von zwei Controllern in beiden Paaren ist deshalb mit bis zu 33% sehr ähnlich. Eine deutliche Reduktion ist messbar, wenn nur noch ein Controller im Einsatz ist. Der Impact bei Ausfall von 3 der 4 Controllern liegt bis maximal 68% (statt der theoretisch erwarteten 75%).

Die Latenzen sind zumindest bei geringer Last vergleichbar, steigen aber dann aufgrund der früheren Überlastung bei weniger Controllern schneller an.

Test	Read	Write	Read/Write
L1C4S4 (grau) 1 Storage, 4 Cntr, 4 Server	6'112 MBs Max	6'313 MBs Max	5'067 MBs Max
L1C3S4 (blau) -25% 1 Storage, 3 Cntr, 4 Server	4'334 MBs Max **-29%**	6'211 MBs Max **-2%**	3'737 MBs Max**. -26%**
L1C2S4 (grün) -50% 1 Storage, 2 Cntr, 4 Server	4'331 MBs Max **-29%**	6'423 MBs Max **-0%**	3'397 MBs Max**. -33%**
L1C4S4 (orange) -75% 1 Storage, 1 Cntr, 4 Server	1'937 MBs Max **-68%**	4'924 MBs Max **-22%**	1'606 MBs Max**. -68%**

II. Zusammenfassung

Bei Ausfall von drei der vier Controller eines Controller Enclosures bleiben alle LUNs auf allen Pfaden gegenüber dem Server weiterhin sichtbar und zugreifbar. Der Performanceimpact hängt davon ab, wie stark die Controller durch das Testszenario belastet werden und ob der Controller überhaupt einen Engpass darstellt oder beispielsweise die Bandbreite vom Server.

Im Worst Case, wenn der Controller der Engpass war, kann eine Reduktion der Bandbreite bis zu 68% beobachtet werden, wenn 3 von 4 Controllern und damit 75% der Leistung ausfallen. Bei vielen Szenarien ist es aber deutlich weniger. Im Frontend (Cached IOPS) beträgt der Impact nur 46% für Random IOs oder 17% bei Sequential IOs.

Insgesamt erreicht **ein einzelner Controller** bei Ausfall von drei Controllern immer noch beeindruckende Werte:

- 500'000 Frontend IOPS
- 250’000 Backend IOPS
- 5 GB/s Durchsatz Frontend pro Richtung
- 2 GB/s Durchsatz Backend Read und 5 GB/S Durchsatz Backend Write, Read und Write 1.6 GB/s

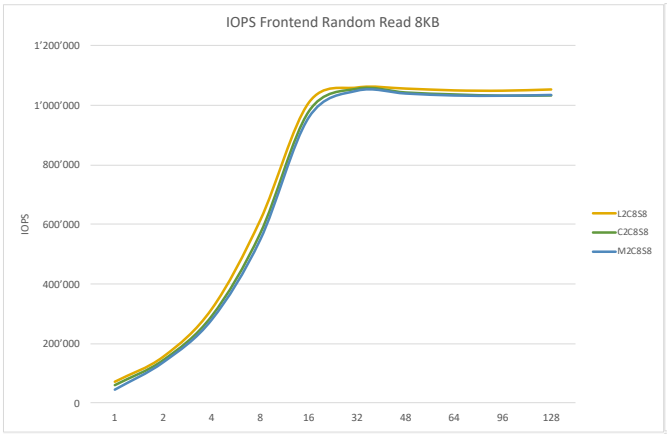
III. Test Gespiegelt

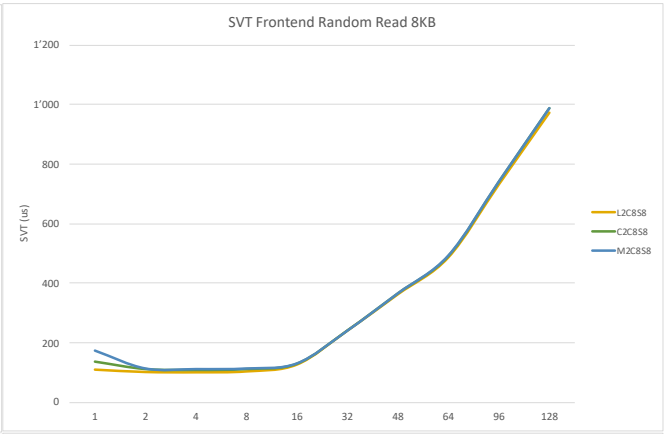
Bei dieser Testreihe wird die Auswirkung der synchronen Storage Spiegelung gemessen. Ausgehend von der Baseline L2C8S8, also einem Test mit 8 Servern auf beiden Storagesystemen mit insgesamt 8 Controllern auf ausschliesslich lokalen Disks, wird der gleiche Tests auf ausschliesslich gespiegelten Disks wiederholt und mit je 50% gespiegelten und 50% ungespiegelten Disks.

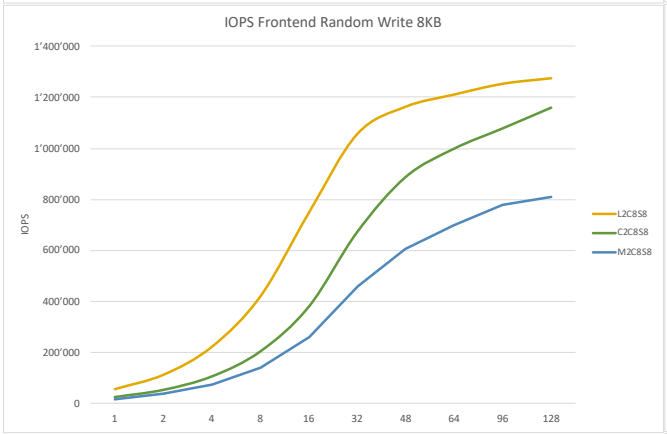
- L2C8S8 (gelb) – Lokaler Test mit 2 Storages, 8 Controllern und 8 Servern (Baseline ungespiegelt)
- C2C8S8 (grün) – Gespiegelt und lokaler Test kombiniert mit 2 Storages, 8 Controllern und 8 Servern (50% gespiegelt, 50% lokal)
- M2C8S8 (blau) – Gespiegelter Test mit 2 Storages, 8 Controllern und 8 Servern (voll gespiegelt)

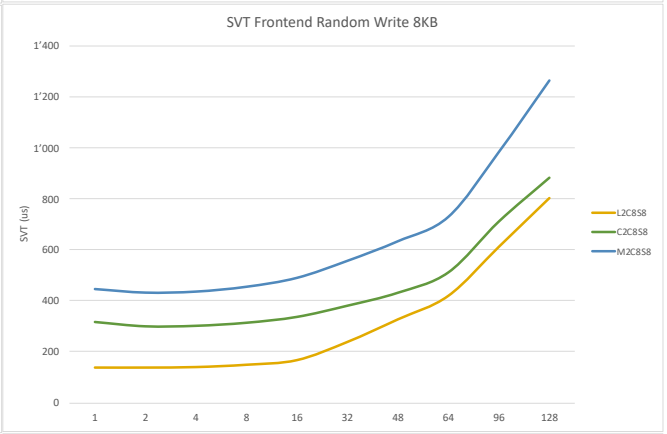
Grundsätzlich ist bei diesen Tests bei reinen Leseoperationen kein signifikanter Impact zu erwarten. Beim Schreiben und abgeschwächt beim kombinierten Schreiben und Lesen muss jedoch jeder Schreibauftrag auf beiden Storagesystemen statt nur auf einem Storagesystem durchgeführt werden. Es ist deshalb mindestens eine Halbierung der Write IOPS zu erwarten, wenn nicht ein Engpass ausserhalb des Storage zum Tragen kommt. Ebenso steigt natürlich die Latenz bei Schreibaufträgen an, da der Schreibauftrag nicht nur lokal im Cache abgelegt wird, sondern auch zum zweiten Storagesystem transportiert und von diesem bestätigt werden muss, bevor der Schreibauftrag dem Server bestätigt wird.

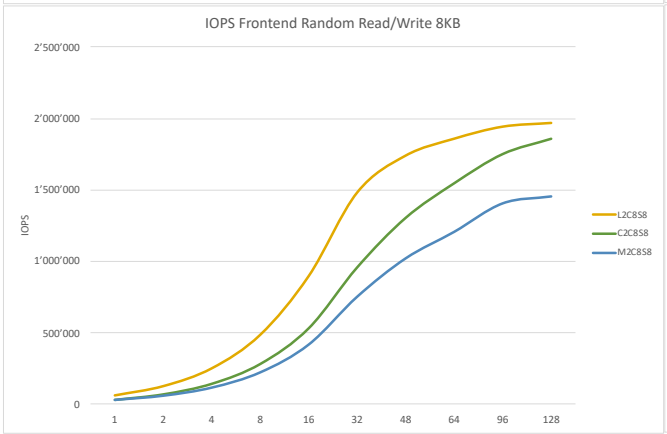
IIIa. Mirrored Frontend Random 8 KB

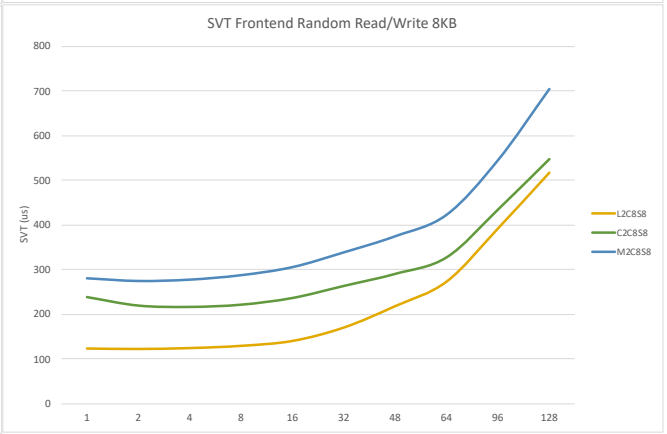










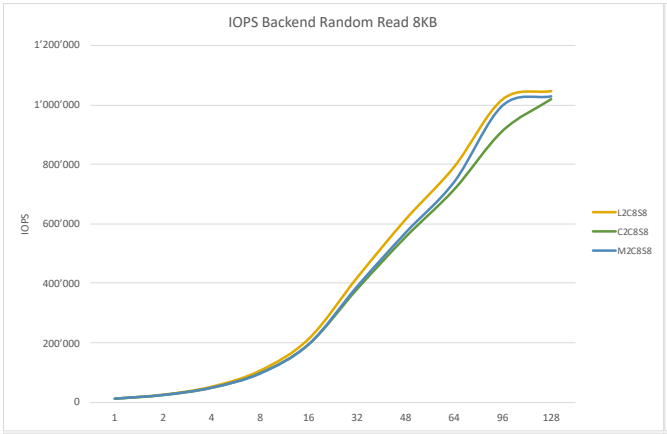


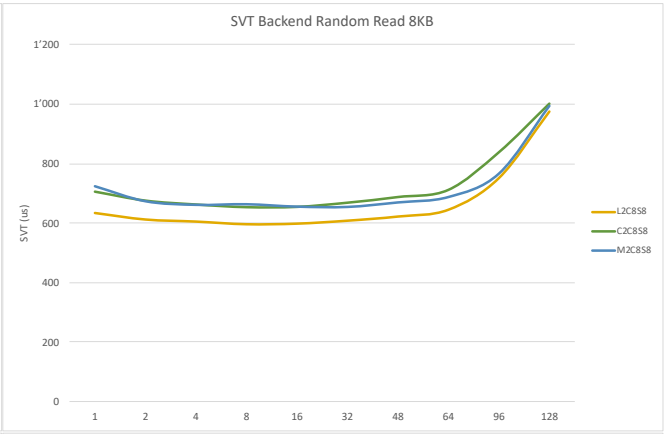
Wie erwartet sind die reinen Lesoperationen völlig unabhängig von der Spiegelung der Daten. Bei reinen Schreibaufträgen sinkt der Durchsatz um 37% deutlich weniger als die erwarteten 50%, der Mixed Workload mit Read und Write Aufträgen hat einen Performance Impact von 27%. Werden nur 50% der IOPS auf gespiegelten LUNs durchgeführt (C2C8S8) ist der Impact weniger als 10%.

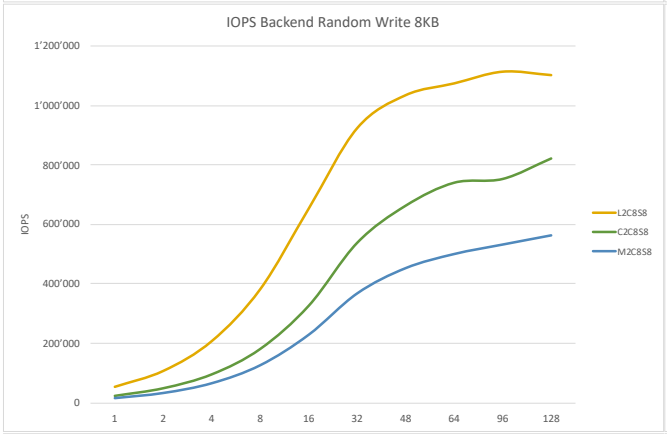
Beim Lesen ist die Servicezeit absolut identisch, beim Schreiben liegt die Latenz statt bei 140 µs beim lokalen Schreibauftrag bei 430 µs beim gespiegelten Schreibauftrag. Die synchrone Spiegelung führt also zu einer zusätzlichen Latenz von 290 µs. Dies stellt einen sehr geringen Impact für eine synchrone Spiegelung dar.

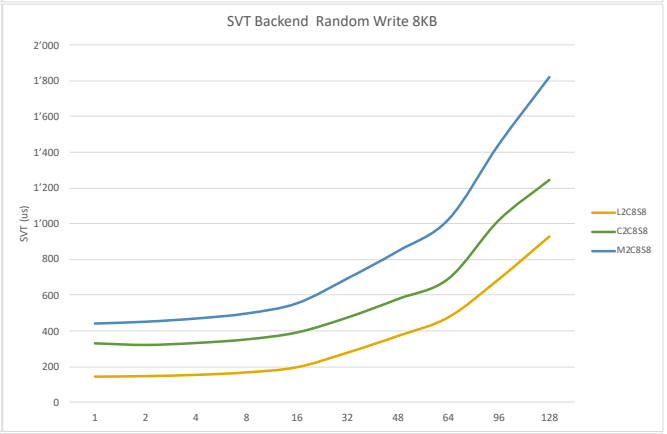
Test	Read	Write	Read/Write
L2C8S8 (gelb) 0% gespiegelt	1’054 kIOPS Max	1’275 kIOPS Max	1’972 kIOPS Max
C2C8S8 (grün) 50% gespiegelt	1’045 kIOPS Max **-1%**	1’159 kIOPS Max **-9%**	1’861 kIOPS Max **-6%**
M2C8S8 (blau) 100% gespiegelt	1’051 kIOPS Max **-1%**	809 kIOPS Max **-37%**	1’454 kIOPS Max **-27%**

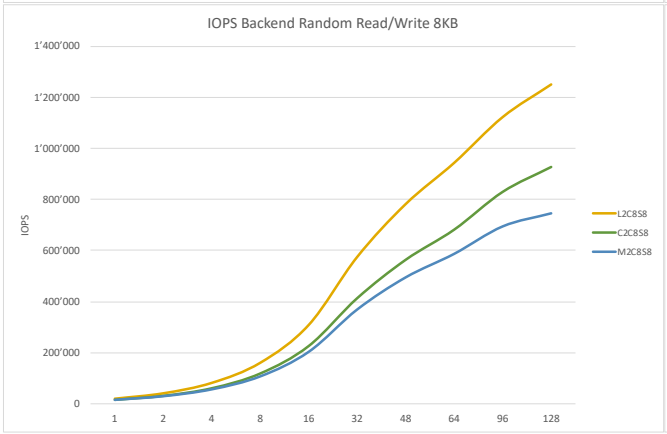
IIIb. Mirrored Backend Random 8 KB

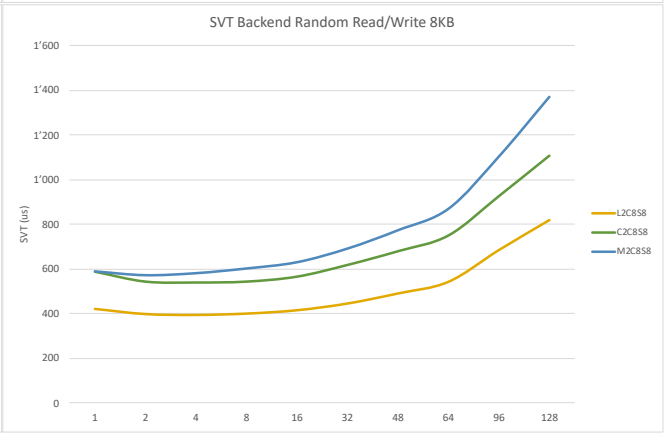












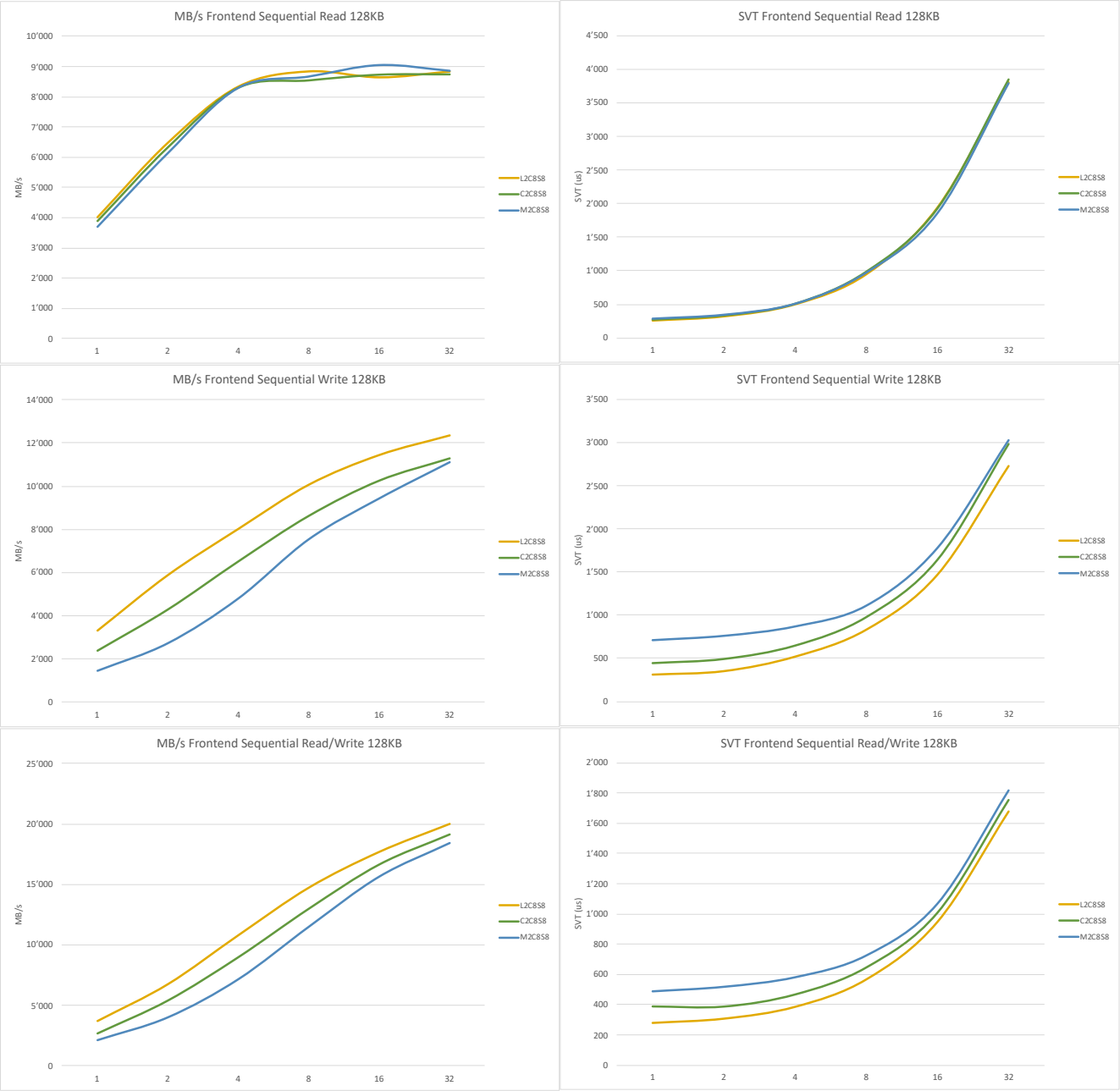
Wie erwartet sind die reinen Leseoperationen auch im Backend unabhängig von der Spiegelung der Daten. Bei reinen Schreibaufträgen sinkt der Durchsatz wie erwartet um 49%, der Mixed Workload mit Read und Write Aufträgen hat einen Performance Impact von 40%, hier müssen die Writes natürlich zweifach auf beiden Storages erfolgen.

Werden nur 50% der IOPS auf gespiegelten LUNs durchgeführt (C2C8S8) ist der Impact bei 25% entsprechend halb so gross, da nur die Hälfte der LUNs gespiegelt sind.

Beim Lesen ist die Servicezeit auf gespiegelten LUNs im Backend geringfügig höher, beim Schreiben liegt die Latenz statt bei 140 µs beim lokalen Schreibauftrag bei 440 µs beim gespiegelten Schreibauftrag. Die synchrone Spiegelung führt also zu einer zusätzlichen Latenz von 300 µs. Dies stellt einen sehr geringen Impact für eine synchrone Spiegelung dar.

Test	Read	Write	Read/Write
L2C8S8 (gelb) 0% gespiegelt	1'048 kIOPS Max	1'112 kIOPS Max	1'253 kIOPS Max
C2C8S8 (grün) 50% gespiegelt	1'021 kIOPS Max **-3%**	823 kIOPS Max **-26%**	926 kIOPS Max **-26%**
M2C8S8 (blau) 100% gespiegelt	1'029 kIOPS Max **-2%**	564 kIOPS Max **-49%**	746 kIOPS Max **-40%**

IIIc. Mirrored Frontend Sequential 128 KB

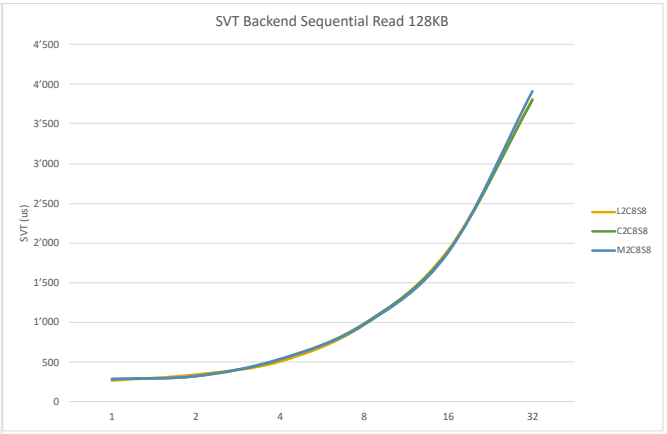


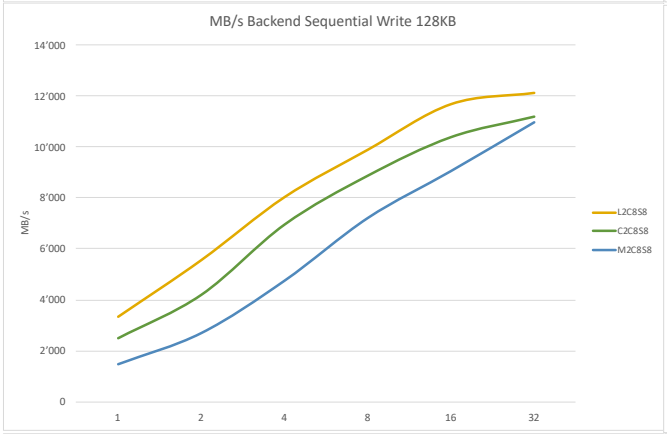
Wie erwartet sind die reinen Leseoperationen völlig unabhängig von der Spiegelung der Daten. Da die Controller mit den relativ wenigen, grossen Schreibaufträgen nicht am Limit sind, kann auch die doppelte Anzahl Writes bei den gespiegelten LUNs ohne signifikante Leistungseinbusse abgewickelt werden, somit zeigen sich bei (teilweise) gespiegelten LUNs keine grösseren Performanceeinbussen (maximal 11%).

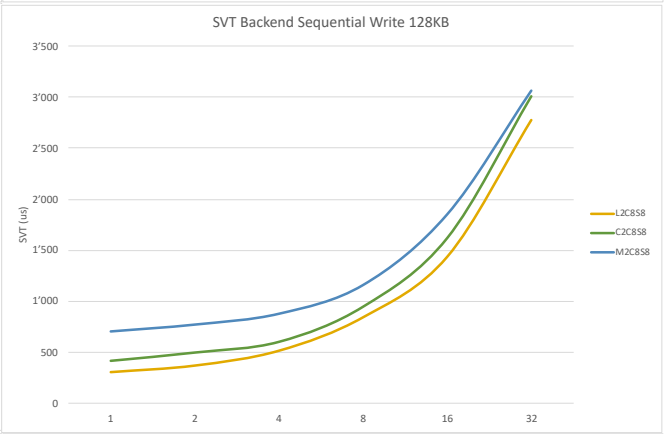
Allerdings steigt bei den Schreibvorgängen die Latenz von 315 µs beim lokalen Schreiben auf 712 µs bei den gespiegelten Writes, eine Steigerung um 397 µs für die synchrone Spiegelung, die erfreulicherweise für einen 128 KB Block nur ca. 100 µs höher liegt als bei einem 8 KB Block.

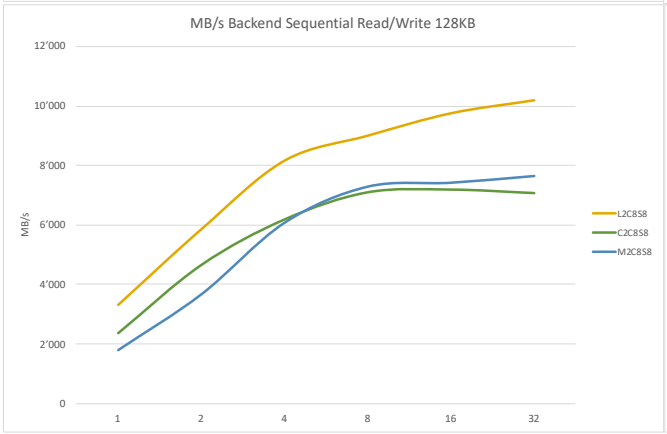
Test	Read	Write	Read/Write
L2C8S8 (gelb) 0% gespiegelt	8'826 MBs Max	12'428 MBs Max	19'999 MBs Max
C2C8S8 (grün) 50% gespiegelt	8'724 kIOPS Max **-1%**	11'269 kIOPS Max **-9%**	19'116 kIOPS Max **-4%**
M2C8S8 (blau) 100% gespiegelt	9'048 kIOPS Max **-0%**	11'115 kIOPS Max **-11%**	18'437 kIOPS Max **-8%**

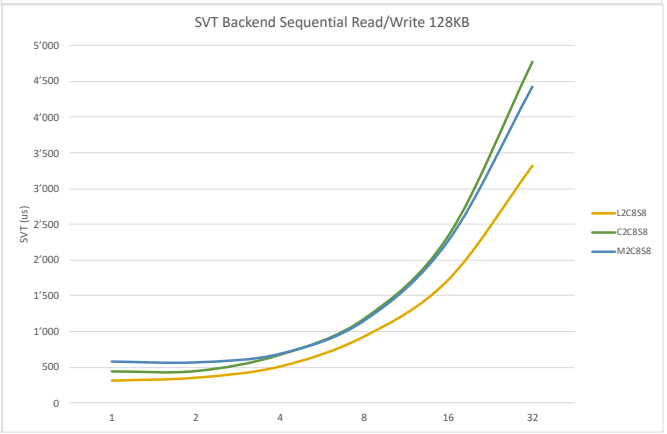
IIId. Mirrored Backend Sequential 128 KB











Wie erwartet sind die reinen Leseoperationen auch im Backend völlig unabhängig von der Spiegelung der Daten. Da die Controller mit den relativ wenigen, grossen Schreibaufträgen nicht am Limit sind, kann auch die doppelte Anzahl Writes bei den gespiegelten LUNs mit sehr moderaten Leistungseinbussen von maximal 29% abgewickelt werden.

Allerdings steigt bei den Schreibvorgängen die Latenz von 311 µs beim lokalen Schreiben auf 708 µs bei den gespiegelten Writes, eine Steigerung um 397 µs für die synchrone Spiegelung, die erfreulicherweise für einen 128 KB Block nur ca. 100 µs höher liegt als bei einem 8 KB Block.

Test	Read	Write	Read/Write
L2C8S8 (gelb) 0% gespiegelt	8'820 MBs Max	12'123 MBs Max	10'202 MBs Max
C2C8S8 (grün) 50% gespiegelt	8'841 MBs Max **-0%**	11'175 MBs Max **-8%**	7'201 MBs Max **-29%**
M2C8S8 (blau) 100% gespiegelt	8'919 MBs Max **-0%**	10'967 MBs Max **-10%**	7'627 MBs Max **-25%**

III. Zusammenfassung

Die synchrone Datenspiegelung und die damit doppelten Schreibaufträge wirken sich, wie erwartet, beim Lesen gar nicht aus. Bei sequentiellen Schreiben (ob im Frontend oder im Backend) können ebenfalls nur geringe Leistungseinbussen durch die Spiegelung festgestellt werden, der Durchsatz sinkt um maximal 29%.

Bei Random Writes ist der Impact am deutlichsten, da hier die Controller effektiv das Limit darstellen und somit zwangsläufig bei einer Verdoppelung der Schreibtaufträge eine Reduktion des Durchsatzes erfolgt. Die Reduktion beträgt im Frontend maximal 37% und im Backend wie erwartet 50%.

Die Latenz beim Lesen ist unverändert und beim Schreiben werden 300 µs bei 8 KB und 400 µs bei 128 KB Blöcken zusätzliche Latenz für die synchrone Spiegelung notwendig.

Insgesamt erreichen wir bei **vollständiger Spiegelung** über 2 Storagesysteme folgende Werte:

- 1 Mio. Frontend und Backend Random Reads 8 KB gespiegelt,
- 800’000 Frontend und 560'000 Backend Random Writes 8 KB gespiegelt
- 1.4 Mio Frontend Random Reads/Writes und 750’000 Backend Random Reads/Writes 8 KB gespiegelt
- 9 GB/s Frontend und Backend Sequential Read gespiegelt
- 11 GB/s Frontend/Backend Sequential Write Durchsatz gespiegelt
- 20 GB/s Frontend Sequential Read/Write Durchsatz und 7.5 GB/s Backend Sequential Read/Write Durchsatz

Es ist anzunehmen, dass ein vollausgebautes Dorado 8000 V6 System die vierfache Last verarbeiten kann.

IV. Test Verfügbarkeit gespiegelt

In einem gespiegelten Setup mit zwei Storagesystemen mit je 4 Controllern wurden schrittweise in einem Storagesystem 1, 2 und dann 3 Controller entfernt. Im letzten Schritt wurde das zweite System mit 4 Controllern zusätzlich abgeschaltet, sodass nur noch einer von 8 Controllern aktiv war.

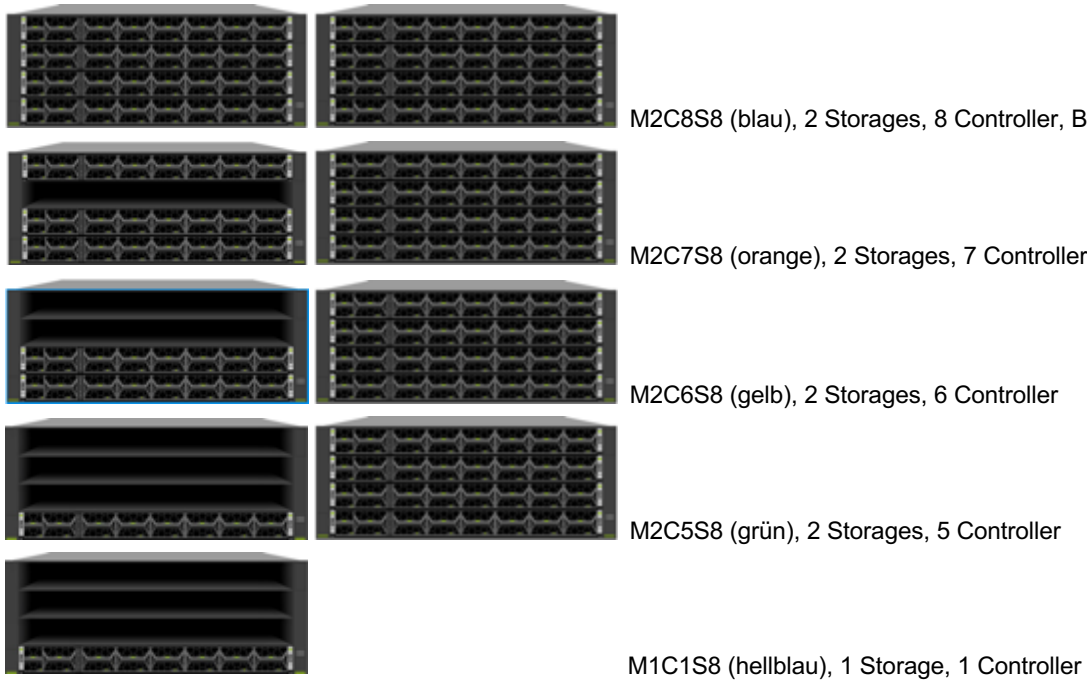


Abbildung 5 – Tests IV - from Huawei Device Manager

Im Einzelnen wurden die folgenden Tests durchgeführt:

- M2C8S8 (blau) – Gespiegelter Test mit 2 Storages, 8 Controllern und 8 Servern, **Baseline**
- M2C7S8 (orange) – Gespiegelter Test mit 2 Storages, 7 Controllern und 8 Servern
- M2C6S8 (gelb) – Gespiegelter Test mit 2 Storages, 6 Controllern und 8 Servern
- M2C5S8 (gelb) – Gespiegelter Test mit 2 Storages, 5 Controllern und 8 Servern
- M1C1S8 (gelb) – Gespiegelter Test mit 1 Storage, 1 Controllern und 8 Servern

Der Test M2C8S8 (blau) dient dabei als Baseline für die Performance mit allen acht Controllern über beide Storages und wurde schon im vorangehenden Testrun I. und II. beschrieben.

Bemerkenswert ist, dass in der Tat bei Ausfall / Entfernen von drei der vier Controller eines Storages immer noch alle LUNs für den Server auf allen FC Pfaden sichtbar sind. Es ging keiner der Pfade zum Server offline. Bei Abschalten des zweiten Storages sieht der Server dann das gespiegelte LUN nicht mehr über alle 8 Pfade, sondern nur noch über 4 Pfade auf dem verbleibenden Storage mit nur einem Controller.

In den Tabellen wird für dieses Testszenario jeweils der Höchstwert pro Test angegeben und der resultierte Impact durch Ausfall der Controller in Prozent.

IVa. Mirrored Frontend Random 8 KB

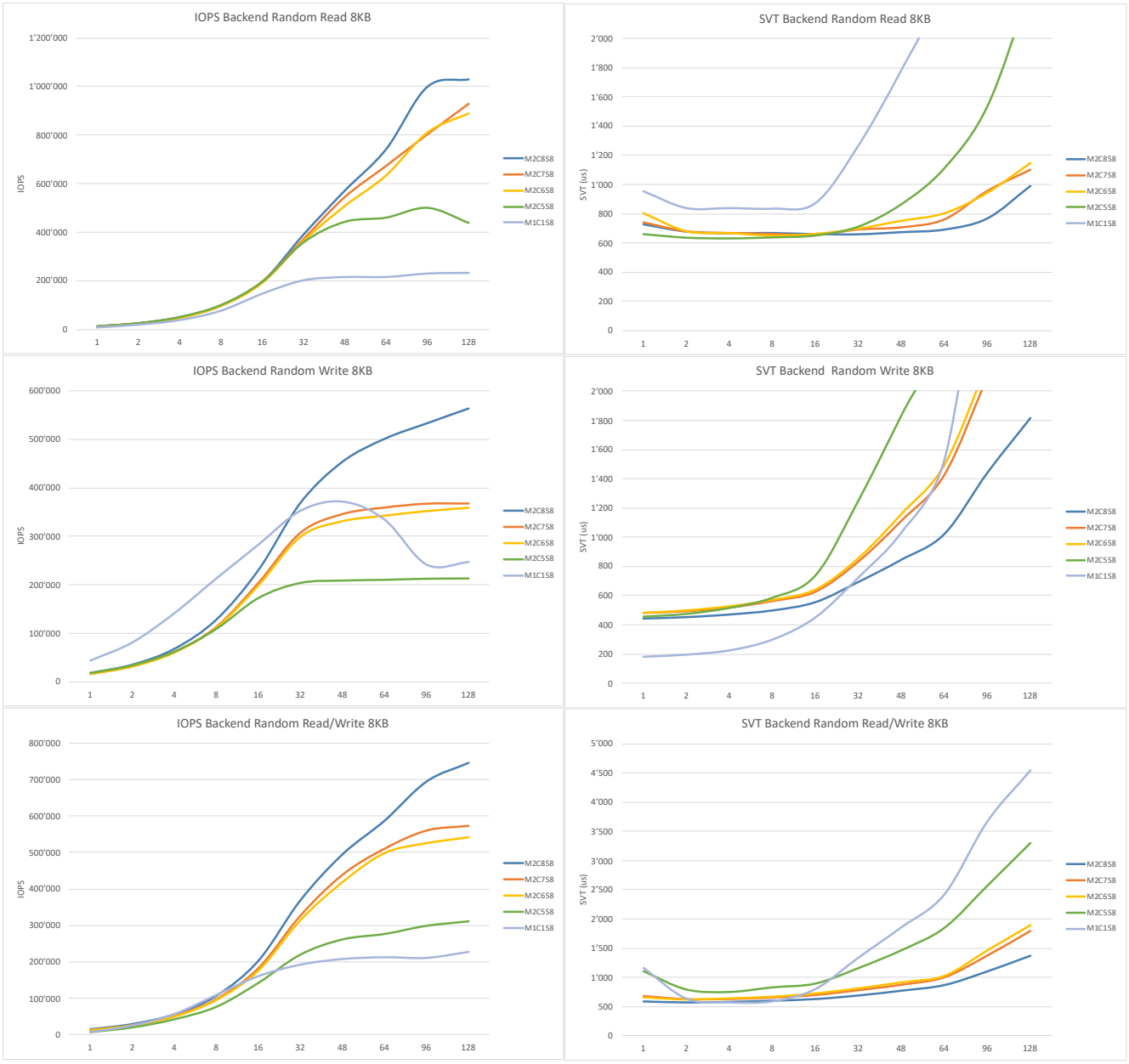
Verteilt über zwei Storagesysteme können selbst mit nur einem Controller in einem Storage bis zu 1 Mio. Frontend Reads abgewickelt werden. Hier ist der Workload am geringsten, da weder ein Zugriff auf das Storage Backend notwendig ist, noch die Kalkulation einer Checksum. Beides ist beim Schreiben hingegen notwendig und bei Ausfall von zwei und drei Controllern ist eine Leistungsreduktion von bis zu 50% zu erkennen. Bei Ausfall eines Controllers eines Paars stellt dieses «halbierte» Paar den Engpass dar und das Storagesystem kommt auf die gleiche Performance wie bei Ausfall von je einem Controllern in beiden Paaren (gelb und orange). Wenn auch der dritte Controller ausfällt (grün), reduziert sich die Leistung um dann 71% bei Reduktion der Controllerleistung um 75%.

Wenn nur ein Storage im Einsatz ist, reduziert sich beim Lesen der Durchsatz um 39% der Leseleistung von beiden Arrays, beim Schreiben hingegen ist der Impact geringer, da in diesem Fall keine Spiegelung der Daten auf den anderen Storage erfolgt.

Die Latenzen bleiben in allen Szenarien gleich, steigen aber abhängig von der Lastsituation unterschiedlich schnell an. Bei Einsatz von nur einem Storage sind die Schreibaufträge ohne synchrone Spiegelung natürlich sehr viel schneller (M1C1S8).

Test	Read	Write	Read/Write
M2C8S8 (blau) 2 Storages, 8 Cntr., 8 Server **Baseline**	1’051 kIOPS Max	809 kIOPS Max	1’454 kIOPS
M2C7S8 (orange) 2 Storages, 7 Cntr., 8 Server	1’044 kIOPS Max **-1%**	431 kIOPS Max **-47%**	786 IOPS Max **-46%**
M2C6S8 (gelb) 2 Storages, 6 Cntr., 8 Server	1’040 kIOPS Max **-1%**	427 kIOPS Max **-47%**	783 kIOPS Max **-46%**
M2C5S8 (grün) 2 Storages, 5 Cntr., 8 Server	1’045 kIOPS Max **-0%**	233 kIOPS Max **-71%**	429 kIOPS Max **-71%**
M1C1S8 (hellblau) 1 Storage, 1 Cntr., 8 Server	645 kIOPS **-39%**	494 kIOPS **-39%**	611 kIOPS **-58%**

IVb. Mirrored Backend Random 8 KB

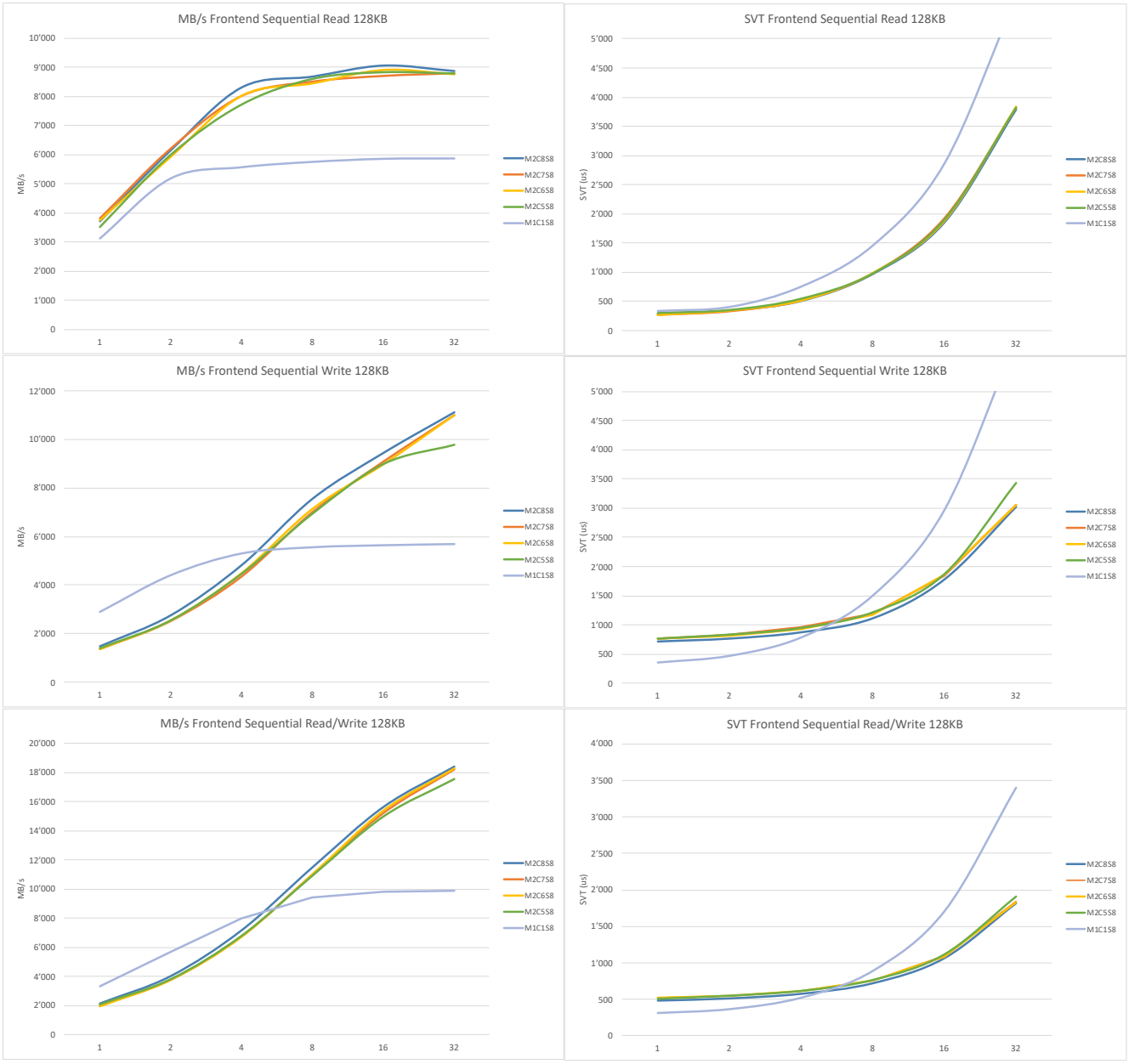


Bei Backend IO ist der Impact eines Controllerausfalls gravierender. Da jeweils zwei Controller beim Backend Zugriff ein Paar bilden, wird das um einen Controller reduzierte Paar zum Engpass bei Backend IOPS. Der Performanceimpact bei Ausfall von einem Controller in einem Paar oder von zwei Controllern in beiden Paaren ist deshalb sehr ähnlich (gelb und orange). Eine erneute deutliche Reduktion ist messbar, wenn dann nur noch ein Controller im Einsatz ist (grün). Der Impact bei Ausfall von 3 der 4 Controllern liegt bei maximal 2%.

Fällt der zweite Storage und damit die Spiegelung weg, so steigt die Anzahl der verarbeiteten Writes wieder leicht, während aber die Read Performance bei 1/8 der Controller um 77% reduziert ist.

Test	Read	Write	Read/Write
M2C8S8 (blau) 2 Storages, 8 Cntr., 8 Server **Baseline**	1'029 kIOPS Max	564 kIOPS Max	746 kIOPS Max
M2C7S8 (orange) 2 Storages, 7 Cntr., 8 Server	927 kIOPS Max **-10%**	367 kIOPS Max **-35%**	573 kIOPS Max **-23%**
M2C6S8 (gelb) 2 Storages, 6 Cntr., 8 Server	890 kIOPS Max **-14%**	358 kIOPS Max **-37%**	542 kIOPS Max **-27%**
M2C5S8 (grün) 2 Storages, 5 Cntr., 8 Server	500 kIOPS Max **-51%**	214 kIOPS Max **-62%**	312 kIOPS Max **-58**
M1C1S8 (hellblau) 1 Storage, 1 Cntr., 8 Server	232 kIOPS Max **-77%**	372 kIOPS Max **-34%**	226 kIOPS Max **-70%**

IVc. Mirrored Frontend Sequential 128 KB

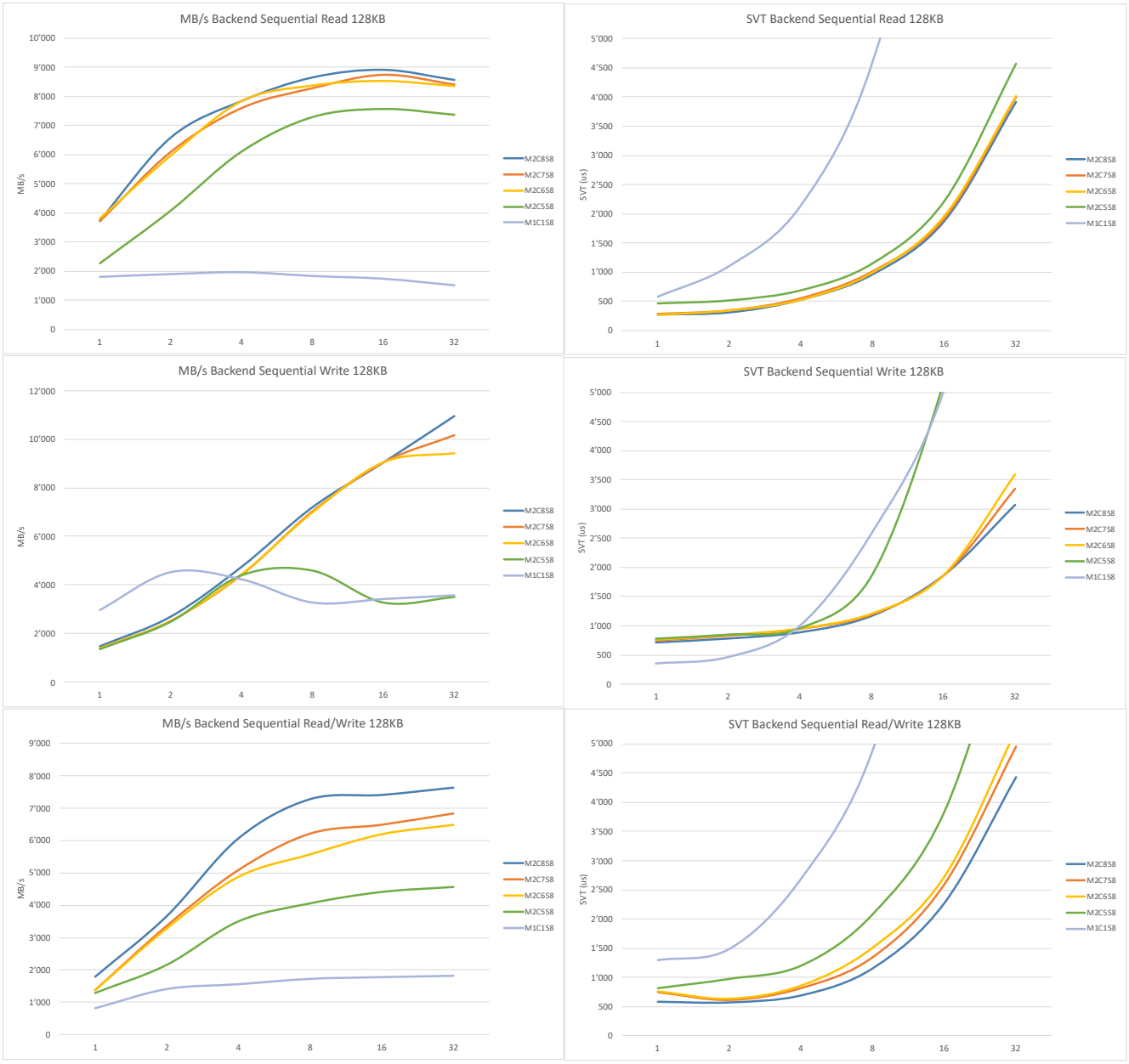


Da bei diesem Workload bei einer Blocksize von 128 KB die Anzahl der IOPS wesentlich geringer ist, kann der Sequential Workload auch bei Ausfall von drei der vier Controllern problemlos und mit sehr geringem Performanceimpact abgewickelt werden. Bei Frontend IOPS werden immer die gleichen Blöcke überschrieben und die Controller müssen keine Checksums berechnen.

Bemerkenswert ist, dass kaum Unterschiede in der Latenz festzustellen ist. Bei Einsatz von nur einem Storage ist der Schreibauftrag ohne die synchrone Spiegelung natürlich sehr viel schneller (M1C1S8).

Test	Read	Write	Read/Write
M2C8S8 (blau) 2 Storages, 8 Cntr., 8 Server **Baseline**	9'048 kIOPS Max	11'115 kIOPS Max	18'437 kIOPS Max
M2C7S8 (orange) 2 Storages, 7 Cntr., 8 Server	8'797 MBs Max **-3%**	11'002 MBs Max **-1%**	18'195 MBs Max **-1%**
M2C6S8 (gelb) 2 Storages, 6 Cntr., 8 Server	8'921 MBs Max **-0%**	11'010 MBs Max **-0%**	18'292 MBs Max**. -1%**
M2C5S8 (grün) 2 Storages, 5 Cntr., 8 Server	8'824 MBs Max **-2%**	9'769 MBs Max **-12%**	17'592 MBs Max**. -5%**
M1C1S8 (hellblau) 1 Storage, 1 Cntr., 8 Server	5'880 MBs Max **-35%**	5'695 MBs Max **-49%**	9'873 MBs Max **-46%**

IVd. Mirrored Backend Sequential 128 KB



Bei Backend IO ist der Impact eines Controllerausfalls gravierender. Dennoch ist bei Ausfall von einem oder gar 2 Controllern nur ein überraschend geringer Impact feststellbar, die verbleibenden Controller können offensichtlich die grossen Blöcke noch performant verarbeiten. Deutlich wird dann der Impact vor allem beim Schreiben bei Ausfall von 3 von 4 Controllern (grün) in einem Storage und dann insbesondere bei Abschalten des anderen Storages (hellblau).

Die Leselatenzen sind bei eingeschränkter Zahl von Controllern bereits etwas höher. Beim Schreiben sind die Unterschiede etwas geringer aufgrund des Write Caches, besonders schnell ist das reine Schreiben bei nur einem verbleibenden Storage-system (M1C1S8).

Test	Read	Write	Read/Write
M2C8S8 (blau) 2 Storages, 8 Cntr., 8 Server **Baseline**	8'919 MBs Max	10'967 MBs Max	7'627 MBs Max
M2C7S8 (orange) 2 Storages, 7 Cntr., 8 Server	8'752 MBs Max **-2%**	10'159 MBs Max **-7%**	6'831 MBs Max **-10%**
M2C6S8 (gelb) 2 Storages, 6 Cntr., 8 Server	8'543 MBs Max **-4%**	9'427 MBs Max **-14%**	6'470 MBs Max**. -15%**
M2C5S8 (grün) 2 Storages, 5 Cntr., 8 Server	7'567 MBs Max **-15%**	4'604 MBs Max **-58%**	4'564 MBs Max**. -40%**
M1C1S8 (hellblau) 1 Storage, 1 Cntr., 8 Server	1'952 MBs Max **-78%**	3'580 MBs Max **-67%**	1'825 MBs Max**. -76%**

IV. Zusammenfassung

Bei Ausfall von drei der vier Controller eines Controller Enclosures bleiben alle LUNs auf allen Pfaden gegenüber dem Server weiterhin sichtbar und zugreifbar. Der Performanceimpact hängt davon ab, wie stark die Controller durch das Testszenario belastet werden und ob der Controller überhaupt einen Engpass darstellt oder beispielsweise die Bandbreite vom Server.

Im Worst Case, wenn der Controller der Engpass war, kann logischerweise eine Reduktion der Bandbreite bis zu 71% beobachtet werden, wenn 3 von 4 Controllern ausfallen. Bei vielen Szenarien ist es aber deutlich weniger.

Bei einer synchronen Spiegelung zwischen zwei Storagesystemem werden bei Ausfall von 3 der 4 Controllers eines Storagesystems immer noch beeindruckende Werte erreicht:

- 1 Mio. Frontend und 500'000 Backend Random Reads 8 KB gespiegelt
- 230'000 Frontend Random Writes und 215’000 Backend Random Writes
- 430’000 Frontend Random Read/Write Sequential und 310’000 Backend Random Read/Writes
- 9 GB/s Frontend Sequential Reads und 8 GB/s Backend Sequential Reads
- 5.5 GB/s Frontend Sequential Reads und 4.5 Backend Sequential Writes
- 10 GB/S Frontend Sequential Read/Write und 2 GB/s Backend Sequential Read/Write

V. Funktionale Tests

Auf Vorschlag von Huawei haben wir ausserdem diverse funktionale Tests für VMware ESX und Oracle durchgeführt, die wir hier kurz zusammenfassen. Die Ergebnisse der Tests sind bei In&Out detailliert dokumentiert und können bei Bedarf zur Verfügung gestellt werden.

Va. Funktionale Tests VMware ESX 6.7

Basierend auf VMware ESX 6.7 wurden folgende funktionale Tests erfolgreich durchgeführt:

- T01 – Mapping der maximalen LUN-ID von 1023 an ein System, Rescan auf dem Server
  Löschen aller LUNs und Rescan auf dem Server
  Erzeugen von LUNs mit der gleiche LUN-ID und Rescan auf dem Server
- T02 – Erzeugen einer VMDK Harddisk und einer RDM Harddisk und Rescan auf dem Server
- T03 – Clonen einer VM per vCenter
  a) ohne Hardwareunterstützung des Storages (VAAI disabled)
  b) mit Hardwareunterstützung des Storages (VAAI enabled)
- T04 – Migration einer VM auf einen anderen Datastore per vCenter
  a) ohne Hardwareunterstützung des Storages (VAAI disabled)
  b) mit Hardwareunterstützung des Storages (VAAI enabled)
- T05 – Thin provisioning / Space reclamaition
  a) Erzeugen thin provisioning LUNs auf einem VMFS6
  b) Schreiben zufälliger Daten, die effektive Grösse des LUNs auf dem Storage nimmt entsprechend zu
  c) Entfernen der LUN auf dem VMFS, nach ca. 5 Minuten wird der Storageplatz freigegeben
- T06 – Thick provisioning
  a) Erzeugen thick provisioning LUNs auf einem VMFS6 ohne Hardwareunterstützung des Storages (VAAI disabled)
  b) Erzeugen thick provisioning LUNs auf einem VMFS6 mit Hardwareunterstützung des Storages (VAAI enbabled)
- T07 – Erstellen eines VM Templates von einer bestehenden VM, Erzeugen und starten von 10 VMs vom Template

Weitere Tests mit Ausfällen einzelner Komponenten wie Controllern, Links zwischen Storagesystemen etc. konnten aus Zeitgründen nicht mehr durchgeführt werden. Grössenteils wurden diese aber bereits im Rahmen der Performancetests auf Bare Metal Systemen erfolgreich getestet.

Vb. Funktionale Tests Oracle

Basierend auf Oracle 18.3 und einer Grid Infrastructure Umgebung wurden folgende Tests erfolgreich durchgeführt:

- T01 – Erstellen eines Storage-Snapshots einer aktiven Oracle Datenbank, weitere Aktivitäten auf der Datenbank sowie einem Rollback auf den Storage-Snapshots.
- T02 – Remote asynchrone Replikation mit Storage Snapshots auf der Remote Site zum Erstellen einer Datenbank-Kopie zu einem bestimmten Zeitpunkt.
- T03 – Auswirkungen des Ausfalls von Storage Controllers auf die Verfügbarkeit der Datenbank
- T04 – Migration von einem lokalen Storage System zu einer durch HyperMetro-Replikation ausfallsicheren Storage-Konfiguration ohne Unterbruch der Datenbanken / Applikationen
- T05 – Ausfall eines Storage-Systems in einer HyperMetro-Cluster-Konfiguration und der Auswirkungen auf die Oracle-Datenbank.

Alle Tests konnten erfolgreich durchgeführt und abgeschlossen werden. Die Datenbank konnte die durch einen Lastgenerator erzeugten Lese- und Schreibtransaktionen ohne Unterbruch weiter verarbeiten.

Fazit

Ein lokales 4 Controller Dorado 8000 V6 System kann folgende Leistungskennzahlen erreichen:

- 750'000 8KB Frontend Random Reads oder Writes,
  kombiniert Read und Write sogar 1.2 Mio. IOPS mit Latenzen im Teillastbereich von ca. 200 µs
- 650’000 8KB Backend Random IOPS mit Read Latenzen von 500 µs und Write Latenzen von 200 µs
- 6 GB/s Sequential Reads oder Writes im Frontend wie im Backend, im Frontend Read/Write sogar über 12 GB/s. Während diese Werte mit guten Latenzen auch im Backend erreicht werden, wird beim bidirektionalen Read/Write lediglich ein Durchsatz von 6 GB/s pro Storage erreicht. Die Latenz liegt bei moderater Last bei unter 500 µs.

Es ist aufgrund der Zahlen anzunehmen, dass ein vollausgebautes Dorado 8000 V6 System ca. 3 Mio. 8 KB Random Reads oder Writes verarbeitet oder bidirektional fast 5 Mio. 8 KB Random IOPS. Beim Durchsatz wären ca. 24 GB/s zu erwarten.

Dabei bleiben dank der SmartMatrix Architektur selbst bei Ausfall von drei der vier Controllern alle Storagepfade online und selbst **ein einzelner Controller** erreicht immer noch beeindruckende Werte:

- 500'000 Frontend IOPS
- 250’000 Backend IOPS
- 5 GB/s Durchsatz Frontend pro Richtung
- 2 GB/s Durchsatz Backend Read und 5 GB/S Durchsatz Backend Write, Read und Write 1.6 GB/s

Bei dem typischen Anwendungsfall der synchronen Spiegelung zeigt sich beim Lesen keinerlei Unterschiede zur ungespiegelten Performance. Bei **vollständiger Spiegelung** über 2 Storagesysteme werden folgende Schreib-Werte erreicht:

- 1 Mio. Frontend und Backend Random Reads 8 KB gespiegelt
- 800’000 Frontend und 560'000 Backend Random Writes 8 KB gespiegelt
- 1.4 Mio Frontend Random Reads/Writes und 750’000 Backend Random Reads/Writes 8 KB gespiegelt
- 9 GB/s Frontend und Backend Sequential Read gespiegelt
- 11 GB/s Frontend/Backend Sequential Durchsatz Write gespiegelt
- 20 GB/s Frontend Sequential Durchsatz Read/Write und 7.5 GB/s Backend Sequential Durchsatz Read/Write

Der Impact bei Ausfall von einzelnen Controllern in einem Storagesystem war für sequentielle und grosse IOPS vergleichsweise moderat. Bei kleinen zufälligen IOPS vor allem im Backend schlägt der Ausfall eines einzelnen Controllers bereits signifikant zu Buche, da dann bei zufällig verteilen IOPS der Zugriff auf die Disks hinter einem Controllerpaar zum Engpass werden kann. Dieser Effekt ist aber natürlich unvermeidlich, sofern die Controllerleistung der kritische Faktor ist.

Beeindruckend war, dass in jeder Situation (sogar bei Ausfall von 7 der 8 Controller) die LUNs immer online verfügbar blieben und bei Ausfall von 3 von 4 Controllern in einem System auch immer alle Pfade vom Server zum Storage online blieben.

Die Stabilität und das Verhalten der Systeme war im Test einwandfrei. Wir konnten keine Ausfälle oder unerklärlichen Performanceschwankungen feststellen. Die Bedienung der Systeme war selbst für Benutzer, die mit dem System überhaupt nicht vertraut waren, einfach und intuitiv möglich.

Huawei hat hier ein neues Top Modell im Portfolio, dass bereits mit 4 Controllern beeindruckende Performancekennzahlen bei tiefen Latenzen erreicht.

Über den Autor

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

Andreas Zallmann hat Informatik an der Universität Karlsruhe studiert und ist seit dem Jahr 2000 bei der In&Out AG. Er ist verantwortlich für den Geschäftsbereich Technology und seit 2016 CEO der In&Out AG.
Die In&Out verfügt über jahrelange Praxis-Erfahrung in Architektur, Konzeption, Benchmarking und Tuning von Storage- und Systemplattformen insbesondere für Core Applikationen für Banken und Versicherungen.

Andreas Zallmann ist der Entwickler des In&Out Performance Benchmarking Tool IOgen™ und hat in den letzten Jahren sehr viele Kunden- und Hersteller-Benchmarks durchgeführt.