IN&OUT

IT CONSULTING & ENGINEERING

IN&OUT AG

Positionspapier IBM Full Flash Storage für Kern Applikationen auf Power Systemen

Andreas Zallmann

Bereichsleiter Technology, In&Out AG

Version:	1.03
Datum:	1.6.2016
Klassifikation:	nicht klassifiziert

In&Out AG IT Consulting & Engineering

Seestrasse 353, CH-8038 Zürich
Phone +41 44 485 60 60
Fax +41 44 485 60 68

info@inout.ch, www.inout.ch

xw

Vorbemerkung

Die vorliegende Studie wurde im Auftrag von IBM ausgeführt. In&Out als unabhängiges Beratungsunternehmen versichert, dass IBM keinen Einfluss auf die Benchmarkresultate und auf die Inhalte dieser Studie genommen hat und diese unabhängig erstellt wurde. Es besteht keinerlei finanzielle Verbindung zwischen In&Out und IBM.

Einleitung

IBM verfügt mit dem FlashSystem 900 über ein sehr leistungsfähiges All-Flash Storage System im Portfolio.

Die In&Out AG verfügt über ausgewiesene jahrelange Erfahrung in Storage Performance Benchmarks und hat dazu das IO Performance Benchmark Tool IOgen™ entwickelt.

Im Rahmen eines Benchmarks konnte der Vorgänger des FS900, das FS840 vermessen werden. Das aktuelle FS900 Modell stand leider noch nicht zur Verfügung.

Abbildung 1- Flash System 900 (Quelle: IBM)

Management Summary

Besonders entscheidend für die Geschwindigkeit der Applikationen ist die Servicezeit oder Latenz der IO Requests. Diese liegt beim IBM FlashSystem bei ca. 0.2 ms für kleine Blockgrössen (gemessen am Server). Das ist Faktor 2 schneller als bei herkömmlichen Tiered Storage Systemen mit SSD Disks und Faktor 25 schneller als bei rein diskbasierten Storagesystemen. Bei einem Wechsel von Disk auf All Flash Storage sind bei storagelastigen Applikationen Performanceverbesserungen von Faktor 4 bis 20 durchaus realistisch. Die folgende Tabelle zeigt die gemessenen minimalen Latenzzeiten zum Storage für verschiedenen Blockgrössen und IO Pattern:

IO Pattern	100% Read 0% Write	70% Read 30% Write	0% Read 100% Write
1K Random	200 µs	200 µs	140 µs
4K Random	230 µs	215 µs	159 µs
8K Random	243 µs	226 µs	167 µs
64K Random	377 µs	442 µs	273 µs
256K Seq.	735 µs	1'236 µs	966 µs
1024K Seq.	2'287 µs	3'567 µs	2'024 µs

Tabelle 1 – Minimale Latency abhängig von Blockgrösse

Mit über 1'000'000 IOPS konnten wir mit dem FS840 einen neuen IO Rekord erreichen, den wir bisher auch auf Highend Storage Systemen noch nicht erreicht haben. Selbst unter dieser Höchstlast betrug die Servicezeit auf dem Server nur 0.5 ms. Bis zu 200'000 IOPS bleibt die Servicezeit bei konstant schnellen 0.2 ms.

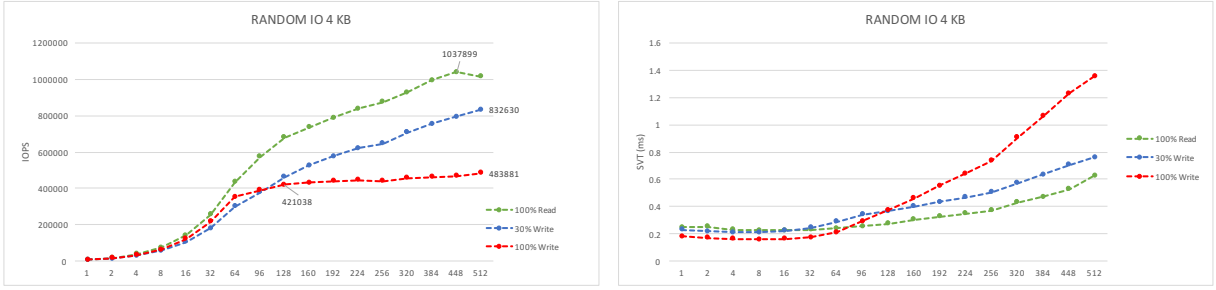

Abbildung 2 - Messergebnisse 4 KB Random

Diese Performancekennzahlen sind beeindruckend. Die IBM Angaben bezüglich Performance und Latency konnten somit in einem Real Life Benchmark bestätigt werden. Die Stabilität des IBM FlashSystem war im Benchmark uneingeschränkt gegeben, insgesamt wurden während der Tests ca. 100 Milliarden IOPS auf den Systemen durchgeführt.

Finanziell betrachtet kostet ein Full Flash Array pro TB in 2016 nur noch Faktor 3.5 mal mehr als ein herkömmliches Midrange Tiered Storage Array. Diese Kosten nähern sich zudem rasch an. Dabei ist insbesondere zu berücksichtigen,

dass Full Flash Systeme sehr häufig komprimiert betrieben werden, z.B. durch einen vorgeschalteten Storage Virtualisierungs-Layer wie IBM Sprectrum Virtualize (SVC) mit aktivierter Kompression (RtC - Real time Compression). Bei einer typischen Kompressionsrate von 1:3 bringt dies den Preis pro TB bei Flash Storage schon auf annähend das gleiche Niveau wie herkömmlichen Tiered Storage. Bei einem hohen IO Bedarf ist ein Flash System jedoch bereits um Faktoren günstiger.

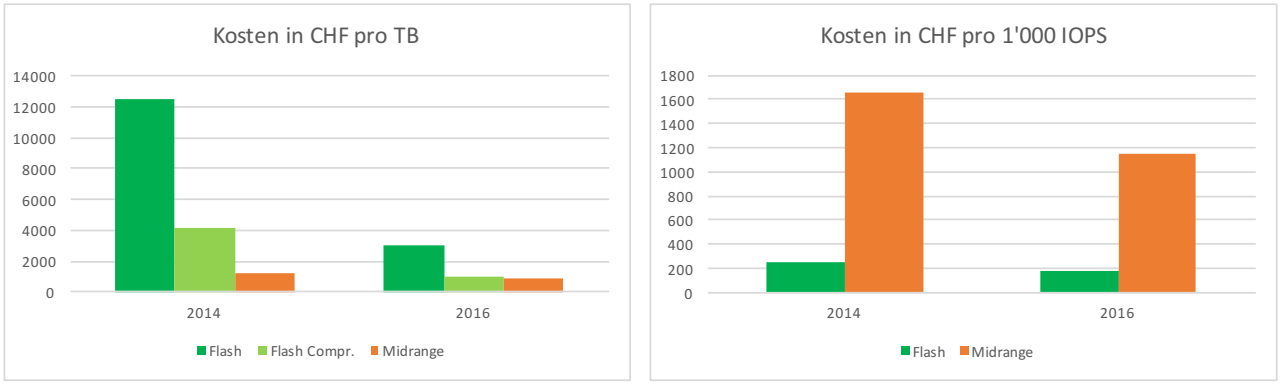

Abbildung 3 – Entwicklung Kostenvergleich

Flash Storage bietet auf der gleichen Stellfläche eine 6 Mal höhere Kapazität und beeindruckende 140 Mal höhere IOPS Leistung. Pro TB weist Flash Storage in etwa den gleichen Stromverbrauch wie ein Midrange Storage Array aus, bezogen auf die IOPS Leistung wird jedoch nur etwa 4% - 7% des Stroms verbraucht.

Fazit

Ist höchste Performance die oberste Prämisse, empfiehlt sich aufgrund der sehr geringen Latenzzeiten ein All-Flash Array. Sofern zusätzliche storagebasierte Funktionen benötigt werden, können All-Flash Systeme beispielsweise in Kombination mit einem Storage Virtualisierungslayer wie IBM Sprectrum Virtualize (SVC) zur Erweiterung der Funktionalität betrieben werden.

Geht es vor allem um Kapazität, weisen herkömmliche Tiered Midrangesysteme aktuell noch einen Kostenvorteil von Faktor 3-5 im Vergleich zu All-Flash Systemen auf, der allerdings zunehmend erodiert. Bei Einsatz von Datenkompression oder Deduplizierung auf Flash Systemen können die Kosten bereits heute auf ein vergleichbares Niveau gedrückt werden, allerdings auf Kosten eines teilweise reduzierten Performancevorteils der All-Flash Arrays.

Ausgangslage

Seit Mitte der 1990er Jahre sind flashbasierte SSD (Solid State Disks) verfügbar, die aber aufgrund der sehr hohen Kosten zunächst einem militärischen Einsatz vorbehalten waren. Frühe SSDs konnten hier primär den Vorteil der mechanischen Robustheit ausspielen.

Seit Mitte der 2000er Jahre und im grösseren Stil seit 2010 sind SSDs auch für den breiten Markt im Einsatz, vor allem in Notebooks (ebenfalls aufgrund der mechanischen Robustheit in Kombination mit der besseren Performance) und in leistungsfähigen Storagesystemen (hier vor allem aufgrund der sehr hohen Performance).

SSD Storage zeichnet sich vor allem durch seine geringe Zugriffszeit (< 0.5 ms) im Vergleich zu herkömmlichen Harddisks (5-10 ms) aus, dies entspricht einer Performancesteigerung um Faktor 10-20. Heute sind SSDs pro IO bereits kostengünstiger als herkömmliche Disks. Bei gleicher Kapazität sind SSDs jedoch immer noch etwa Faktor 10 teurer als reine Harddisks. Bezogen auf ganze Storagesysteme liegt dieser Wert eher bei Faktor 3-5.

Die Entwicklung der SSD basierten Storagesysteme ist sehr rasant und dabei den gesamten Storagemarkt komplett zu verändern. Während über die vergangenen Jahre vor allem die CPUs immer leistungsfähiger und der Storage dabei immer mehr zum Engpass wurde, kann jetzt durch schnelleren SSD Storage wieder ein gewisses Gleichgewicht hergestellt werden.

Auswirkungen Applikation

Herkömmliche Storagearrays weisen typische Latenzzeiten von ca. 5 ms für einen IO aus. Die Verarbeitung geht dank der immer schneller werdenden CPUs oft erheblich schneller von statten – im unten dargestellten Beispiel von IBM

wird eine Verarbeitungszeit von 0.2 ms angenommen. In diesem Szenario dauert eine Verarbeitung insgesamt 5.2 ms, nur 4% davon wird aktiv auf der CPU gerechnet.

Hingegen kann bei einem Full Flash Storage mit einer Antwortzeit von 0.2 ms (wie hier gemessen) die Verarbeitung insgesamt in 0.4 ms statt in 5.2 ms durchgeführt werden, dies entspricht einer Beschleunigung um Faktor 13. Die CPU Belastung steigt dabei von 4% auf 50%, wie in der folgende IBM Folie illustriert:

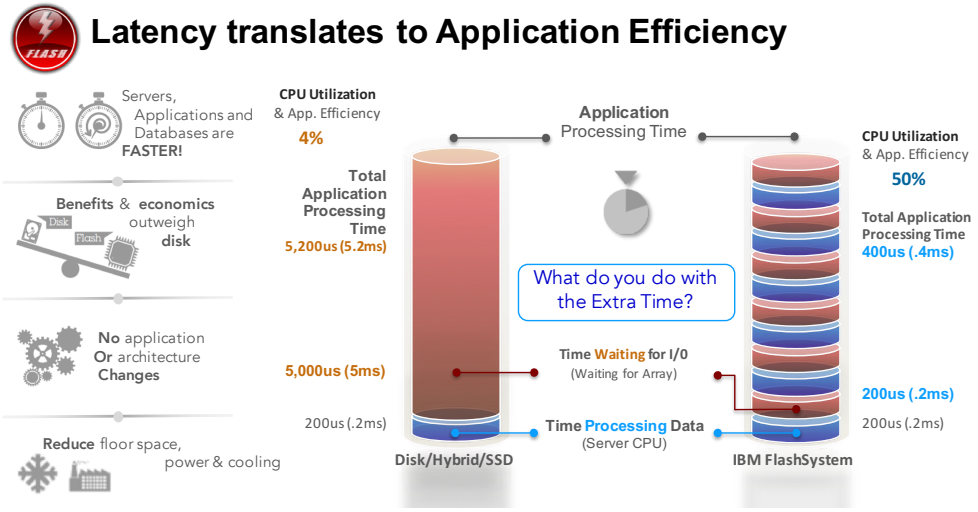

Abbildung 4 – CPU Auslastung in Abhängigkeit von IO Latency (Quelle: IBM)

Die mögliche Beschleunigung der Applikation durch Flashstorage hängt einerseits davon ab, wieviel schneller der Flashstorage ist und andererseits, wie hoch der IO Anteil ist. Dabei ist heutzutage ein Vergleich von All-Flash Storage mit 0.2 ms Latenz mit herkömmlichen Storage nicht realistisch. In der Regel sind performancekritische Anwendungen bereits auf Hybrid Storage mit einem SSD Tier. Diese Tiered Storage Systeme haben aber auch zum SSD Layer eine typische Latenzzeit von 0.4 – 0.5 ms, da der IO diverse Logikschichten durchläuft, z.B. Thin Provisioning, Striping, etc. Damit sind die Antwortzeiten im Vergleich zu All-Flash Storage etwa doppelt so hoch. Somit kann All-Flash Storage die IO Leistung nochmals verdoppeln. Beträgt der IO Anteil nur 10% ist dieser Effekt kaum messbar. Bei einem IO Anteil von über 50%, tritt jedoch eine signifikante und messbare Beschleunigung der Applikation ein.

Kundenbeispiel 1

Im ersten Kundenbeispiel hat ein sehr grosser Kunde im Lebensmittelbereich im November 2014 von IBM DS8700 Tiered Storage (SSD, FC, SATA) auf ein IBM FlashSystem 840 mit IBM Sprectrum Virtualize (SVC) gewechselt. Die durchschnittlichen Servicezeiten sind auf dem Storage von ca. 5 ms auf unter 0.5 ms gesunken, die Storagelatency hat sich somit um Faktor 10 verbessert.

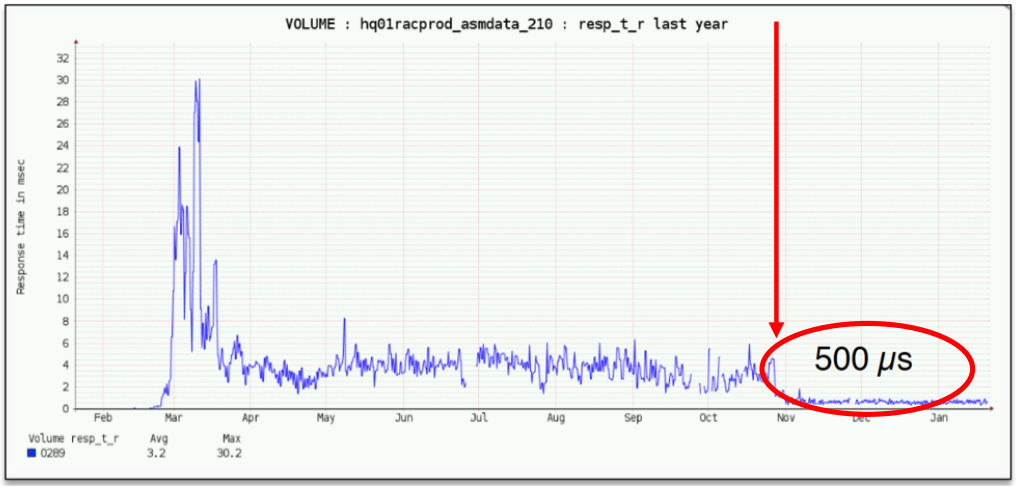

Abbildung 5 – Kundenbeispiel Migration DS8700 auf FS840

Kundenbeispiel 2

Im zweiten Kundenbeispiel hat ein weltweiter Marktführer im Einzelhandel eine kritische SAP Instanz von IBM DS5100 mit herkömmlichen Disks (ohne SSD) auf ein IBM FlashSystem 840 mit IBM Sprectrum Virtualize (SVC) migriert.

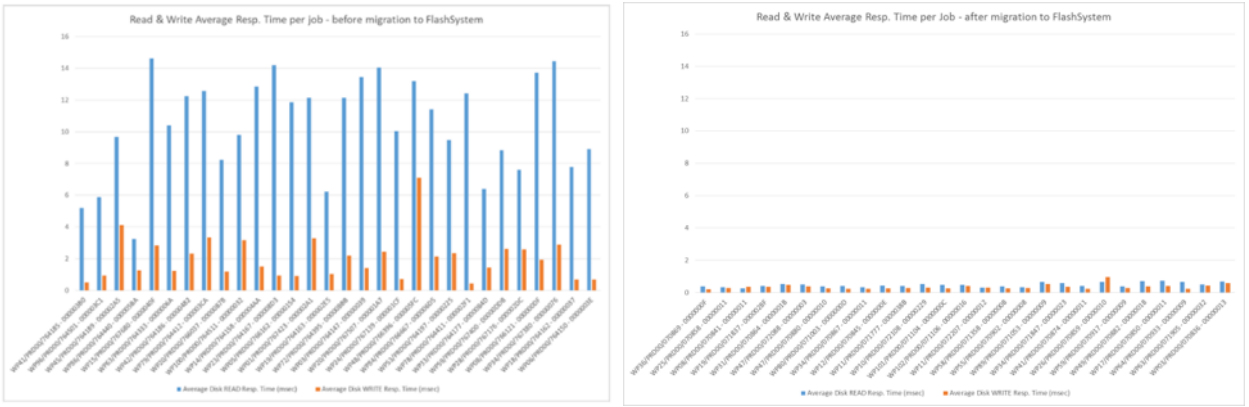

Abbildung 6 – Servicezeiten auf D 5100 Storage (links) und FS840 Full Flash Storage (rechts)

Vor der Migration lagen die durchschnittlichen Servicezeiten auf dem herkömmlichen DS 5100 Storage für Reads bei > 10 ms und für Writes bei ca. 2-4 ms. Nach Migration auf das FlashSystem 840 mit IBM Sprectrum Virtualize (SVC) lagen die Servicezeiten bei < 0.5ms, dies entspricht einer Beschleunigung von Faktor 4 bis 20.

Der Kunde hat bei Batchjobs im SAP Umfeld eine Beschleunigung von 50-70% festgestellt, d.h. die applikatorische Performance ist um Faktor 2 bis Faktor 4 gestiegen.

Beschleunigungspotential der Applikationen

Ermitteln sie die storagelastigen Applikationen, indem sie die Anzahl der IOPS und die durchschnittliche IO Servicezeit auf dem OS oder in der Datenbank (z.B. per Oracle AWR Reports) ermitteln. Das IBM Storage Team Schweiz unterstützt diese Analyse kostenfrei mit spezifischen Tools.

Prüfen Sie, ob ein All Flash System ausreichend ist, oder weitere Logikschichten notwendig sind, um beispielsweise storagebasierte Snapshots oder Spiegelungen auszuführen. In diesem Fall kommt beispielsweise das auf IBM Sprectrum Virtualize (SVC) basierte Produkt das IBM FlashSystemV9000 in Frage. Beachten sie, dass der Einsatz von zusätzlichen Funktionen auch zusätzliche Latency erzeugt, welche abhängig von den Funktionen und IO Load bei 0.1-0.7 ms liegen kann.

Ermitteln Sie das Beschleunigungspotential von Flash Storage (um welchen Faktor kann der IO beschleunigt werden und welchen Anteil hat die IO an der Gesamtleistung).

Beispiel 1: Der IO Anteil beträgt 40%, die durchschnittliche Servicezeit 5 ms auf einem diskbasiertem Storagesystem. Neu beträgt die Servicezeit auf Flash voraussichtlich 0.25 ms. Somit kann der 40% Anteil um Faktor 20 beschleunigt werden. Das Beschleunigungspotential der Applikation beträgt 38%.

Beispiel 2: Die Anwendung ist extrem IO kritisch, der IO Anteil beträgt 80%, befindet sich bereits auf einem Tiered Storage auf dem SSD Tier mit einer Servicezeit von 1 ms. Da bestimmte storagebasierte Funktionen notwendig sind, wird ein zusätzlicher SVC verwendet, die gesamte Latency kann auf 0.5 ms reduziert werden. Somit kann der 80% Anteil um Faktor 2 beschleunigt werden. Das Beschleunigungspotential der Applikation beträgt insgesamt 40%.

IBM All-Flash Storage

Die folgende Tabelle zeigt die technischen Spezifikationen des aktuellen IBM FlashSystem 900 und des hier gemessenen Vorgängers IBM FlashSystem 840:

All Flash Systems	FS840	FS900
Flash Typ	eMLC	IBM Enhanced MLC
Flash Module	1, 2, 4 TB	1.2, 2.9, 5.7 TB
Anzahl Module	Max. 12	Max. 12
Kapazität (Usable, Raid5)	Max. 40 TB / 37.5 TiB	Max. 57 TB / 51.8 TiB
Read Latency	135µs	155µs
Write Latency	90µs	90µs
100% Read	1'100'000 IOPS	1'100'000 IOPS
100% Write	600'000 IOPS	600'000 IOPS
70/30% Read/Write	750'000 IOPS	800'000 IOPS
Sequential Read	8 GBps	10 GBps
Sequential Write	4 GBps	4.5 GBps
Höheneinheiten	2U	2U
Leistungsaufnahme	625W	625W
Konnektivität	16 x 8 GBit FC oder 8 x 16 GBit FC, 16 x 10 GBit Eth., 4 x 40 GBit Infiniband	
Verschlüsselung	Optional AES-XTS 256-bit	

Tabelle 2 – Vergleich IBM Flash Storage Systeme (Angaben laut IBM)

Das IBM FlashSystem 840/IBM FlashSystem 900 bietet dabei reine Storagefunktionen (mit optionaler AES-XTS 256-bit „Data@Rest“ Verschlüsselung). Erweiterte Storagefunktionen wie Thin Provisioning, DeDup, Compression, Tiering, Mirroring oder FlashCopy etc. sind in den FS Produkten nicht verfügbar. Dies resultiert in einer extrem geringen Latency, da der IO keine weiteren Logikebenen durchläuft.

Für den Zugriff auf die Daten und das Flash Chip Management werden keine herkömmlichen CPUs verwendet, sondern besonders schnelle und hoch parallelisierbare FPGA (Field Programmable Gate Array) mit der Möglichkeit von hunderten Prozessen/Threads pro Prozessor (2-4 FPGAs sind pro FlashModule verbaut - dies ergibt bis zu 48 FPGAs Prozessoren pro IBM FlashSystem). Die zusätzlich pro FlashModule verbauten Power-PC CPUs sind für Out of Data Operations vorgesehen und werden zur Optimierung der gesamten Performance & Erweiterung der P/E (Program Erase Cycle) also die Erweiterung der Lebensdauer eingesetzt, darunter sind bekannte Flash spezifische Tasks wie Garbage Collection aber auch durch IBM erweitertes Error Handling, System Health, Statistics, etc., welche den Systembetrieb in Enterprise Storage Umgebungen über Jahre gewährleisten. Die Leistung der FGPAs ist mehr als ausreichend dimensioniert, eine typische Auslastungsmessung wie bei CPUs ist nicht gegeben.

Herkömmliche Full Flash Storage Arrays mit SSD Drives werden intern mit Fiber Channel, PCIe oder SAS angebunden. Die IBM FlashSystem hingegen haben einen reinen Hardware Datenpfad, der die Geschwindigkeit von Flash voll ausnutzen kann und dies wiederum resultiert in einer sehr niedrigen Latency. Zudem ermöglicht dies die Ansteuerung jeder einzelnen Flash-Chip Zelle - dies resultiert in einer um fast Faktor 10 verbesserten Lebensdauer. Damit kann heute für IBM FlashSystems 900 / V9000 unter Wartung bis zu 12 Jahre ein kostenfreier Ersatz von FlashModulen (komplette Einschübe), ohne jegliche Nutzungsauflagen durch IBM garantiert werden.

Zur Nutzung erweiterter Storagefunktionen koppelt IBM den Flash Storage mit IBM Sprectrum Virtualize (SVC). Das entsprechende Produkt heisst „IBM FlashSystem V9000“ und verfügt über Funktionen wie Synchron & Asynchron Mirroring, Tiering, Thin Provisioning, Flash Copy, RtC (Real time Compression), HyperSwap und die Möglichkeit andere Storage Systeme (aktuell über 250 Storage Systeme bekannter Hersteller) zu virtualisieren und auch entsprechend alle Funktionen anzubieten.

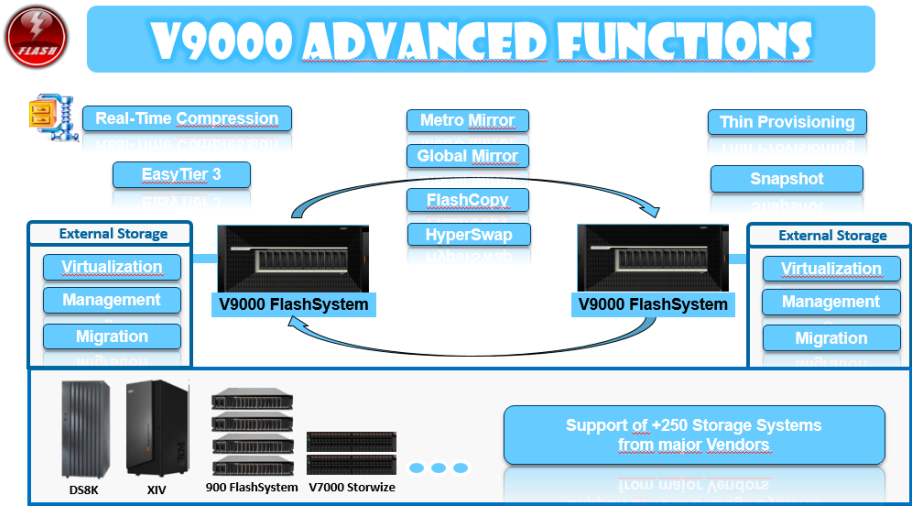

Abbildung 7 – Zusatzfunktionen IBM Sprectrum Virtualize (SVC). (Quelle: IBM)

IBM FlashSystem 900

Die FS900 Hardware besteht aus einem 2U hohen Gehäuse mit redundanten Power Supplies, Fans, Raid Controllern, Batterien und bis zu 12 FlashModule Einschüben mit jeweils maximal 5.7 TB Kapazität. Auch das Motherboard in den „Canisters“ist doppelt ausgelegt und wird im Cluster-Mode betrieben. Alle aktiven Komponenten sind im Betrieb austauschbar (Hot Swapable) und redundant vorhanden.

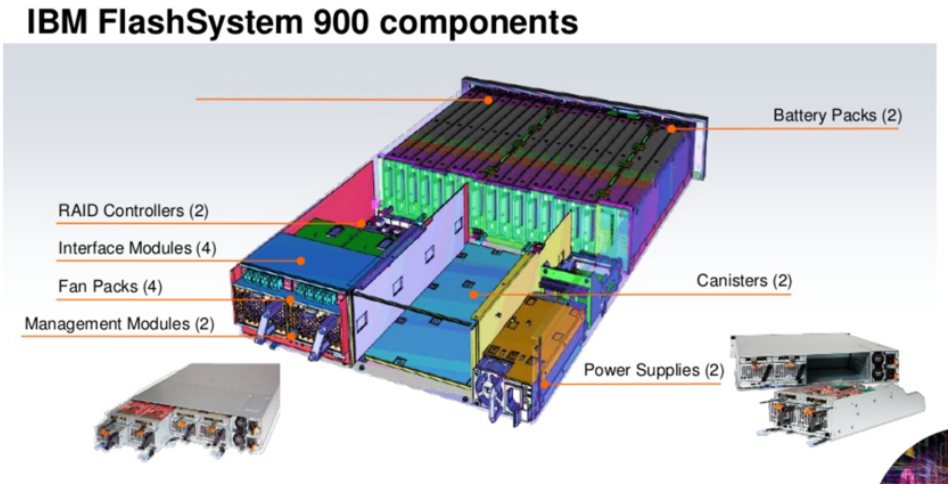

Abbildung 8 – Flash System 900 Hardware (Quelle: IBM)

Mit der Variable Stripe Raid (VSR) Technologie werden die Daten in einem einzelnen FlashModule über die darauf befindlichen Speicherchips verteilt, vergleichbar mit einem N+1 Raid-5 mit rotierender Parity. Bei Fehlern auf einem Speicherchip werden die Daten anhand der Parity wiederhergestellt und in einen Reservebereich verschoben.

Neben dem Variable Stripe Raid innerhalb der FlashModule, wird zusätzlich über die FlashModule selbst gestript (2D-RAID). Dabei werden bei Vollausbau von 12 FlashModulen 10 FlashModule für Daten, 1 FlashModule als Parity und 1 FlashModule als Hot Spare benutzt. Bei Ausfall eines vollständigen FlashModules wird das Hot Spare Modul benutzt und das defekte FlashModule kann während des Betriebs ausgetauscht werden.

Dabei ist VSR auf der Hardware per FPGA (Field Programmable Gate Array) implementiert und ist somit sehr schnell.

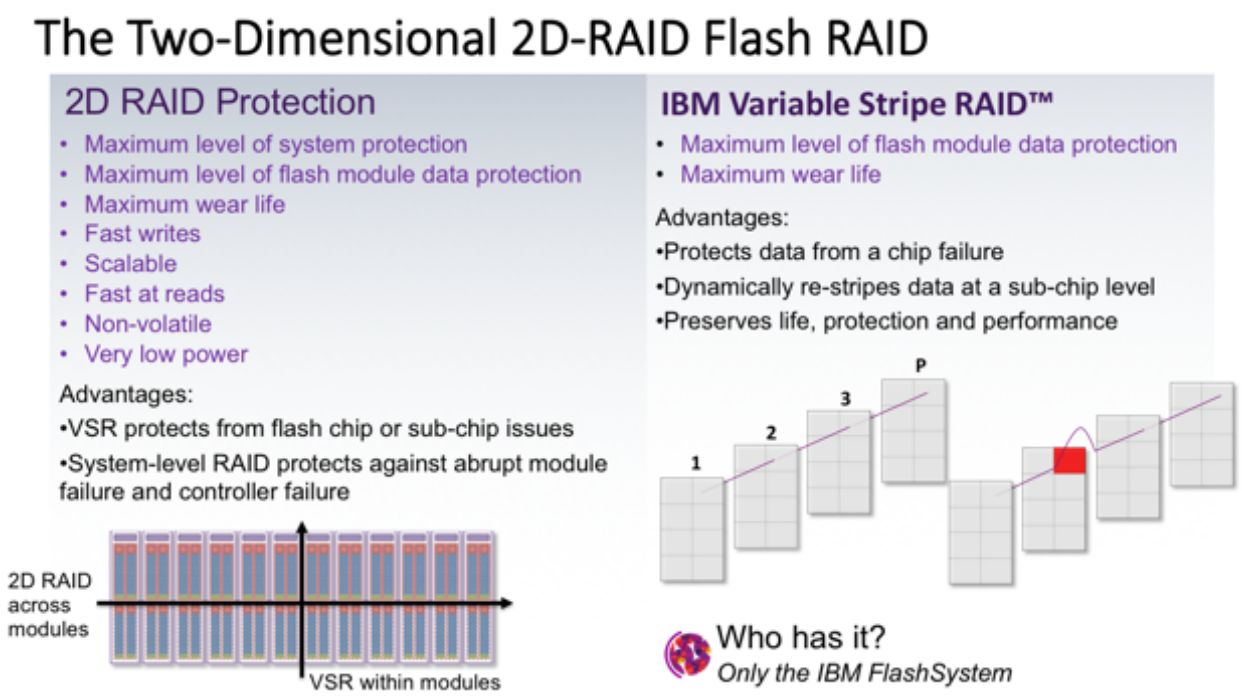

Abbildung 9 – Variable Striping Raid und 2D Flash Raid(Quelle: IBM)

Benchmark Setup

Der folgende Setup vom IBM FlashSystem 840 wurde gemessen:

- IBM FlashSystem 840 (Vorgänger des aktuellen Modells FlashSystem 900)
- 12 x 2 TB FlashModule
- RAID 5 (10+1+1) Konfiguration Raid über FlashModule und zusätzlicher Stripe innerhalb der FlashModule (s.o.)
- Usable Capacity bei RAID 5: 20 TB oder 18.75 TiB
- 8 x 16 GBit Fiber Channel Adapter
- 96 LUNs x 96 GiB = 9 TiB gemappt an IBM p760
- 2x24 LUNs x 8 GiB = 384 GiB gemappt an IBM S824

Folgendes Testsystem wurde eingesetzt mit insgesamt 96 LUNs und einer Kapazität von 9 TiB:

- IBM p760 (Power 7), 1 LPAR, 48 EC, 48 VP
- 48 Cores Power 7 @ 3.4 GHz
- SMT-1 (Kein Simultanous Multithreading[1])
- 8 Fiber Channel Ports @ 16 Gbit (Dual Port Cards), jeweils mit 8 LUNs, total 64 LUNs
- 8 Fiber Channel Ports @ 8 Gbit (Quad Port Cards) mit jeweils 4 LUNs, total 32 LUNs
- Direkt gemappte FC (NPIV)

Weiterhin wurde folgendes S824 System eingesetzt, da ein einzelnes p760 System alleine nicht die gesamte Performance des IBM FlashSystem 840 abrufen konnte:

- IBM S824 (Power 8), 2 LPARs je 1 EC und 12 VP
- 16 Cores Power 8 @ 4.157 GHz
- SMT-1 (Kein Simultanous Multithreading[1])
- 4 Fiber Channel Ports @ 8 GBit (Dual Port Cards)
- 2 VIOs mit je 2 Fiber Channel Ports per NPIV
- Je 24 LUNs x 8 GiB pro LPAR, 12 LUNs pro FC

[1] In diesem speziellen Fall mit einer sehr hohen Kernel Belastung durch die Read und Write Vorgänge wurden die besten Ergebnisse mit SMT-1 erreicht, normalerweise ist die SMT-4/8 (4/8-fach Multithreading) Einstellung zu empfehlen.

Benchmarking – 1'000'000 IOPS mit 0.5 ms Latency

Benchmarking Tool

Die Benchmarks wurde mit dem In&Out Benchmarking Tool IOgen™ 4.0 durchgeführt. IOgen kann verschiedene IO Profile vollautomatisch durchführen und ermittelt die Servicezeit auf dem Benchmarking Server. Diese ist höher als die Servicezeit auf dem Storage, da der IO auch noch zum Server übertragen werden muss. Für Applikationen ist zur Ermittlung des Beschleunigungspotentials der Wert auf dem Server massgebend.

IOgen 4.0 kann Benchmarks über verschiedene Server zeitlich synchronisieren und komplexe Profile zeitsynchron abfahren. IOgen ermittelt die Anzahl der IOs, die Bandbreite, die Latency und die CPU Belastung im Testzeitraum.

Zusätzlich wird jeder Test in verschiedene Intervalle unterteilt und neben der durchschnittlichen Performance auch die maximale und die minimale Performance gemessen und die Abweichung ermittelt. Damit kann die Testgüte und Aussagekraft der Messung beurteilt werden.

Benchmarking Profile

Für alle Benchmarks wurden drei verschiedene IO Profile gemessen (100% Read, 70% Read / 30% Write, 100% Write).

Die folgenden Blockgrössen wurden jeweils gemessen: 1 KB, 4 KB, 8 KB, 64 KB, 256 KB und 1 MB. Bei 1 KB bis 64 KB wurde ein Random IO Profil ausgeführt, bei 256 KB und 1 MB ein Sequential IO Profil.

Für jede Messung wurde folgende Parallelitäten gemessen: 1, 2, 4, 8, 16, 32, 64, 96, 128, 160, 180, 192, 224, 256, 320, 384, 448, 512.

Für jede Parallelität wurden drei Messintervalle von je 20 Sekunden gemessen, insgesamt 60 Sekunden.

Erläuterung zu den Grafiken

Pro Messung werden jeweils zwei Grafiken dargestellt:

- Linke Grafik: Anzahl der IOPS
- Rechte Grafik: Durchschnittliche Servicezeit in ms

Pro Grafik sind drei farbige Kurven dargestellt:

- 100% Read Profil (grün)
- 70% Read / 30% Write Profil (blau)
- 100% Write Profil (rot)

Auf der X-Achse sind die Parallelitäten von 1 bis 512 aufgetragen, auf der Y-Achse die jeweiligen Messwerte.

4 KB Random

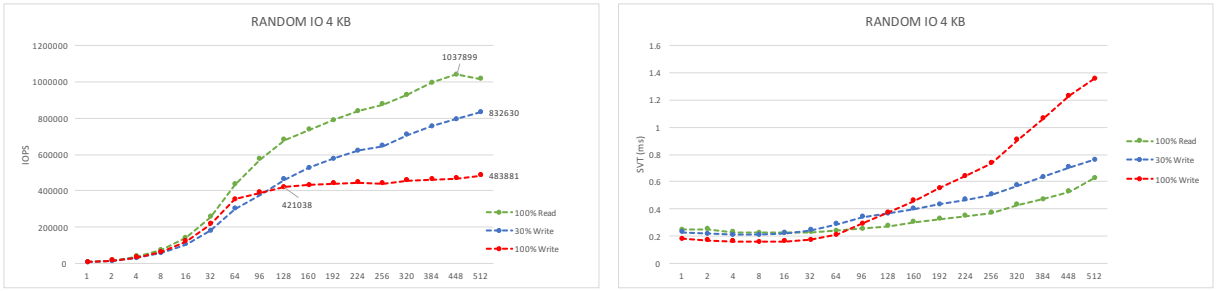

Abbildung 10 - Messergebnisse 4 KB Random

Bei einer Blocksize von 4 KB können mehr als 1 Million Random Reads und 480'000 Random Writes pro Sekunde erreicht werden.

Fast wichtiger als die enorm hohe Menge von Random IOPS ist die tiefe Latency, denn diese bestimmt die Geschwindigkeit der Applikationen. Die Servicezeiten liegen bis zu 200'000 IOPS bei ca. 0.2 ms und sind um Faktor 25 geringer als bei herkömmlichen Storagesystemen.

Jenseits der 200'000 IOPS steigen die Servicezeiten mit zunehmender Parallelität langsam an, bei Erreichen der Sättigung steigen die Servicezeiten deutlich an. Beachtlich ist, dass über 1 Mio. Random Reads mit einer Servicezeit

von gut 0.5 ms abgewickelt werden können. Die CPU Belastung der Server liegt bei 1'000'000 IOPS bei mehr als 90%. Mit mehr CPU Leistung auf den Servern könnte sogar noch etwas mehr Leistung abgerufen werden.

256 KB Sequential

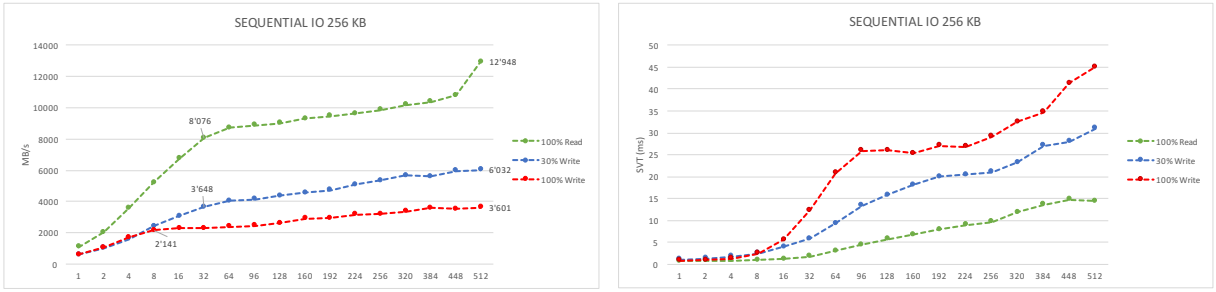

Abbildung 11 - Messergebnisse 256 KB Sequential

Bei 256 KB Blöcken wird eine Bandbreite von fast 13 GB/s beim Lesen und von knapp 4 GB/s beim Schreiben erreicht. Damit werden die Angaben von IBM sogar übertroffen.

Die Servicezeiten liegen aufgrund der grösseren Blocksize hier mit 0.7 ms deutlich höher, da die Übertragung zum Server signifikant länger dauert (alleine die Transferzeit eines 256 KB Blocks dauert über einen 16 Gbps FC Port bereits ca. 0.2 ms, bzw. ca. 0.4 ms über einen 8 Gbps FC Port). Ebenso steigen die Servicezeiten recht schnell an, da die Sättigung bereits mit einer Parallelität von 32 Prozessen erreicht wird. Die CPU Belastung spielt bei diesem Test keine signifikante Rolle.

Systemsicht

Auf der graphischen Konsole der FS 840 wurden kurzzeitig sogar Spitzenwerte von 1.15 Mio. IOPS ausgewiesen. Es hat sich dabei um einen Skalierbarkeitstest mit dem IBM Tool nstress64 gehandelt. Mit einem System p760 wurden 950'000 IOPS erreicht und zusammen mit zwei LPARs auf einer S824 die Maximalleistung von 1.15 Mio. IOPS. Daraufhin wurde der IOgen Test in dieser Kombination durchgeführt. IOgen hat im Vergleich zu nstress64 mehr Möglichkeiten aber auch einen etwas höheren CPU Verbrauch, deshalb wurden mit IOgen etwas geringere Werte gemessen.

Abbildung 12 – GUI auf FS 840 zeigt 1.15 Mio. IOPs

Performance

Die folgende Tabelle fasst die maximalen Performancekennzahlen und Antwortzeiten für die einzelnen Blocksizes und IO Patterns zusammen. Sofern bei geringerer Parallelität und geringeren Antwortzeiten bereits annähernd so gute Performancekennzahlen erreicht wurden, sind diese als zusätzlicher Wert kursiv dargestellt.

IO Pattern	100% Read 0% Write	70% Read 30% Write	0% Read 100% Write
1K Random IOPS	1'008'045 @0.5 ms	725'193 @ 1.0 ms	290'227 @ 3.0 ms *247'648 @ 0.4 ms*
4K Random IOPS	1'003'798 @ 0.55 ms **1'100'000 (IBM)**	823'630 @ 0.8ms	483'881 @ 1.4 ms *421'038 @ 0.4ms* **600'000 burst (IBM)**
8K Random IOPS	852'405@ 0.8 ms *724'893@ 0.4 ms*	533'513 @ 1.4ms *406'353@ 0.7ms*	324'283 @ 2.4 ms *260'508 @ 0.4 ms*
64K Random IOPS	155'441 @ 5 ms *134'534 @ 0.8 ms*	83'608 @ 10ms *63'594 @ 3 ms*	49'934 @ 10 ms *39'941 @ 0.7 ms*
256K Seq. MB/s	12.984 @ 15 ms *8.076 @ 3 ms* ***8'000 (IBM)***	6.032 @ 32 ms *3.648 @ 6 ms*	3.601 @ 45 ms *2.141 @ 3 ms* **4'000 (IBM)**
1024K Seq. MB/s	13.928 @ 40 ms 11.352 @ 20 ms	8.137 @ 78 ms 6.973 @ 18 ms	5.297 @ 105 ms 3.170 @ 18 ms

Tabelle 3 – Performance abhängig von Blockgrösse

Offensichtlich ist, dass bei kleinen Blocksizes die Maximalwerte bei sehr geringen Latenzen erreicht werden.

Die offiziellen Leistungskennzahlen von IBM sind in der Tabelle blau markiert eingefügt:

- Bei 4 KB Random Read konnte der offizielle IBM Wert von 1.1 Mio. IOPS in der Praxis mit 1.0 Mio. IOPS fast erreicht werden.
- Bei 4KB Random Write gibt IBM einen kurzfristigen (Burst Wert) von 600'000 IOPS an. Wir können immerhin knapp 500'000 IOPS erreichen und diesen Wert auch über mehrere Minuten halten.
- Bei 256 KB Sequential Read konnten wir den offiziellen Wert von 8 GB/s von IBM mit 13 GB/s sogar deutlich übertreffen.
- Bei 256 KB Sequential Write können wir mit 3.6 GB/s den IBM Wert von 4.0 GB/s fast erreichen.

Die IBM Angaben bezüglich Performance und Latency konnten somit in einem Real Life Benchmark bestätigt werden.

Latenz

Die folgende Tabelle zeigt die gemessenen minimalen Latenzzeiten zum Storage für verschiedene Blockgrössen und IO Pattern:

IO Pattern	100% Read 0% Write	70% Read 30% Write	0% Read 100% Write
1K Random	200 µs	200 µs	140 µs
4K Random	230 µs **130 µs (IBM)**	215 µs	159 µs **90 µs (IBM)**
8K Random	243 µs	226 µs	167 µs
64K Random	377 µs	442 µs	273 µs
256K Seq.	735 µs	1'236 µs	966 µs
1024K Seq.	2'287 µs	3'567 µs	2'024 µs

Tabelle 4 – Minimale Latency abhängig von Blockgrösse

Für kleine Blocksizes von 1 KB beträgt die Latency vom Server aus gemessen lediglich 200 µs (Lesen) respektive 140 µs (Schreiben). Mit zunehmender Blockgrösse steigen diese Werte entsprechend an. Selbst bei Blockgrössen von 256 KB sind die Latenzen immer noch um 1 ms.

Blau eingetragen sind auch die offiziellen IBM Zahlen. Diese sind im Storage gemessen worden, während wir hier die Latenzen über das SAN zum Server messen. Die angegeben Werte von IBM sind plausibel und realistisch.

Die Servicezeiten sind mit ca. 200 µs exzellent und selbst auf Highend Systemen mit SSD Storage kaum zu erreichen, da dort die IOs durch eine aufwändige Logik gehen und ausserdem der SSD Storage über ein SAS oder FC Interface angesprochen wird. In der Regel sind bei Tiered Storage Systemen auf SSD Disks am Server Latenzzeiten von 400 – 500 µs typisch, also in etwa das Doppelte wie beim All-Flash System FS 840 (siehe auch Abschnitt SPC-1 weiter unten).

Zusammenfassung Benchmark IBM FlashSystem 840

Besonders entscheidend für die Geschwindigkeit ist die Servicezeit oder Latenz der IO Requests. Diese liegt beim IBM FlashSystem bei ca. 0.2 ms für kleine Blockgrössen, im Vergleich zu 0.4 – 0.5 ms bei herkömmlichen Tiered Storage Systemen. Somit ist der All-Flash Storage Faktor 2 schneller als herkömmliche Tiered Storage Systeme mit SSD Disks und Faktor 25 schneller als rein diskbasierte Storagesysteme.

Die Gesamtzahl der IOs mit über 1 Million Random Reads pro Sekunde ist auf allerhöchstem Level und einzigartig für ein so kompaktes System. Um die notwendige CPU Leistung zur Erzeugung der IO Last bereitzustellen, kamen zwei Power Systeme mit total 64 Cores zum Einsatz. Dabei wurde die CPU Zeit nahezu ausschliesslich im Kernel des Betriebssystems verbracht, der Overhead durch das IOgen Tool betrug nur 5-10%.

Wie bei Flash Storage üblich, werden beim Schreiben signifikant weniger IOs erreicht als beim Lesen, dennoch liegt die Zahl auf sehr hohen Niveau.

Die Sequential IO Performance ist ebenfalls sehr gut, wenn man bedenkt, dass dies früher der grosse Schwachpunkt von SSD Storage war. Mit 12-13 GB pro Sekunde wird die theoretische Bandbreite der FC Anbindung von 8 x 16 GBit = 16 GBps fast erreicht.

Während unseres Testzeitraumes haben wir ca. 100 Milliarden IOs auf dem IBM FlashSystem 840 ausgeführt, ohne einen einzigen Fehler zu erhalten.

Kostenvergleich

Die Preise für Flashstorage sind pro Kapazität (TB) noch immer signifikant höher als bei herkömmlichen Disks. Allerdings ist der Preiszerfall im Flashstorage Bereich sehr stark (wie die folgende Tabelle zeigt)

Storagesystem	FS 840 (2014)	Tiered Midrange (2014)	FS 900 (2016)	Tiered Midrange (2016)
Kapazität	20 TB	200 TB	57 TB	200 TB
IOPS	1'000'000	150'000	1'000'000	150'000
Kosten	250'000 CHF	250'000 CHF	175'000 CHF	175'000 CHF
pro TB	12'500 CHF / 4'200 CHF*	1'250 CHF	3'070 CHF / 1'000 CHF*	875 CHF
pro 1'000 IOPS	250 CHF	1'660 CHF	175 CHF	1'150 CHF

Tabelle 5 – Kostenentwicklung Storagesysteme (Februar 2014 und Februar 2016) (= Preis mit vorgeschalteter Kompression von Faktor 1:3)*

Dabei ist insbesondere zu berücksichtigen, dass Full Flash Systeme sehr häufig komprimiert betrieben werden, z.B. durch einen vorgeschalteten IBM Sprectrum Virtualize (SVC) mit aktivierter Kompression. Bei einer typischen Kompressionsrate von 1:3 bringt dies den Preis pro TB bei Flash Storage schon 2016 auf annähend das gleiche Niveau wie herkömmlichen Tiered Storage.

In den folgenden Grafiken sind die Kostenentwicklungen der letzten 2 Jahre pro TB und pro 1'000 IOPS dargestellt.

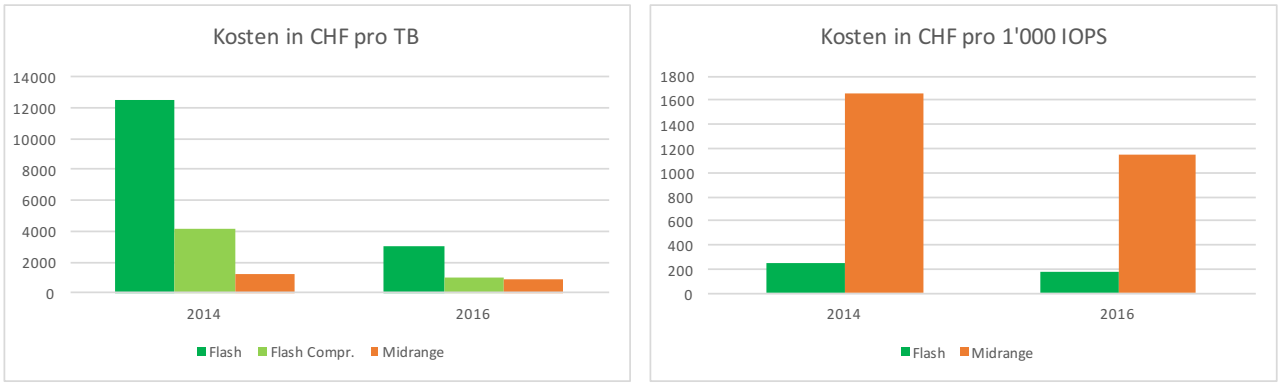

Abbildung 13 – Entwicklung Kostenvergleich

Erkennbar ist der Preiszerfall pro TB Kapazität im Flashbereich Faktoren höher als im herkömmlichen Storagebereich. Dieser Trend wird sich weiter fortsetzen.

Die Kosten pro 1'000 IOPS sind im Flash Storage bereits Faktoren günstiger als bei herkömmlichen Storage. Hier erwarten wir eine in etwa gleichbleibende Entwicklung, da auch in herkömmlichen Storagearrays immer mehr SSD Storage verbaut wird.

Footprint Vergleich

In der folgenden Tabelle wird der Footprint von Full Flash Storage und herkömmlichen Tiered Midrange Storage verglichen:

Storagesystem	FS 900	Tiered Midrange
Kapazität	57 TB	200 TB
IOPS	1'000'000	150'000
Platzbedarf	2U	42U
Strom (Maximum)	1.3 kW	5kW
Strom pro TB	23W	25W
Strom pro 1'000 IOPS	1.3W	33 W
Kapazität 1 Rack 42U	1'197 TB	200 TB
IOs 1 Rack 42U	21'000'000	150'000

Tabelle 6 – Footprintvergleich Storagesysteme (Februar 2016)

Trotz der höheren Kapazität beim herkömmlichen Storage ist die Platzeffizienz der Full Flash Systeme schon heute erheblich besser. So kann ein FS 900 bereits 57 TB Usable Storage auf einer Fläche von 2U anbieten, während für einen herkömmlichen Storage mit 200 TB Kapazität ein ganzes Rack mit 42U benötigt wird. Pro Standard 42U Rack kann somit bei Full Flash Arrays fast die 6-fache Kapazität (TB) und die 140-fache (!) IO Leistung bereitgestellt werden.

Der Stromverbrauch pro TB ist mit ca. 25W vergleichbar. Der Stromverbrauch pro 1'000 IOPS ist jedoch beim Full Flash Array um Faktor 25 geringer.

SPC-1 Vergleich

Das Storage Performance Council (SPC, http://www.storageperformance.org) ist eine gemeinnützige Organisation, die einen standardisierten Storage Performance Benchmark definiert hat, mit dem Hersteller unter definierten Bedingungen ihre Storagesysteme vermessen können. Der Storagebenchmark SPC-1 besteht aus 8 Profilen, die jeweils verschiedene Read/Write Ratios aufweisen. Es werden 4 KB Blöcke verwendet und die Read/Write Ratio beträgt zusammengezogen 70% - 30%.

Die folgende Tabelle vergleicht den FS 900 Storage mit dem aktuellen SPC-1 Rekordhalter Huawai:

Storagesystem	FS 900	SPC-1 Performance #1 Huawai OS 18'800 V3
Disks	12 x 5.7 TB eMLC Flash	512 x 400 GB SSD
Kapazität Usable[2]	57 TB	86 TB
IOPS SPC-1		
100% Load	440'011 @ 0.49 ms	3'010'007 @ 0.92 ms
50% Load	219'977 @ 0.3 ms	1'505.068 @ 0.55 ms
10% Load	43'986 @ 0.24 ms	300'986 @ 0.39 ms
Kosten	708'702 USD	2'370'764 USD
Platzbedarf	2U	2 Racks = 84U
SPC-1 Link	http://www.storageperformance.org/benchmark_results_files/SPC-1E/IBM/AE00008_IBM_FlashSystem-900/ae00008_IBM_FlashSystem-900_SPC-1E_full-disclosure-report-r1.pdf	http://www.storageperformance.org/benchmark_results_files/SPC-1/Huawei/A00163_Huawei_OceanStor-18800-V3/a00163_Huawei_OceanStor-18800-V3_SPC-1_executive-summary.pdf

Tabelle 7 – SPC-1 Vergleich

Für die Applikationsperformance entscheidend ist die Servicezeit (Latency) der IOs. Bei gleicher Auslastung weist das FS 900 die halbe Servicezeit der OS 18'800 auf, d.h. das FS 900 Array ist bei gleicher Auslastung für Applikationen nahezu doppelt so schnell wie das aktuell nach SPC-1 leistungsstärkste Highendstoragesystem.

[2] Storage Pool Capacity after data protection and hot space

Die maximale IO Performance von 3 Mio. IOPS gemäss SPC-1 für das Huawai OS 18'800 V3 System liesse sich alternativ auch mit 7 FS900 abdecken. Dies bei einer deutlich tieferen Latency (0.49 ms vs. 0.92 ms) und mit einem Fünftel des Platzbedarfs. Die im SPC-1 Benchmark Report angegebenen Kosten für das FS 900 von 708'702 USD sind Listenpreise, real dürfte hier von einem Preis von ca. 200'000 USD ausgegangen werden. Die Aussagekraft der Kostenangaben in den SPC-1 Benchmarks sind mit grosser Vorsicht zu geniessen, da nicht klar ausgewiesen wird, wieviel Rabatt auf den Listenpreis im Preis enthalten sind.

Real wird ein Storagesystem im Bereich von 10-50% seiner Maximallast betrieben. Die folgende Tabelle vergleicht entsprechende SPC-1 IOPS von 200'000 bis 800'000.

Storagesystem	FS 900	SPC-1 Performance #1 Huawai OS 18'800 V3
200'000 IOPS SPC-1	1 FS 900m 0.709 Mio. USD 219'977 IOPS @0.3 ms	1 OS 18'800 V3, 2.371 Mio. USD 300'986 IOPS @ 0.39 ms
400'000 IOPS SPC-1	2 x FS 900, 1.417 Mio. USD 439'954 IOPS @ 0.3 ms	1 OS 18'800 V3, 2.371 Mio. USD ca. 400'000 IOPS @ 0.45 ms
800'000 IOPS SPC-1	4 x FS 900, 2.835 Mio. USD 879'908 IOPS @ 0.3 ms	1 OS 18'800 V3, 2.371 Mio. USD ca. 800'000 IOPS @ 0.50 ms

Tabelle 8 – SPC-1 Vergleich

Bei einer Ziel IO Zahl von 200'000 IOPS SPC-1 könnte dies von einem FS 900 mit geringerer Latency und geringeren Kosten erbracht werden. Bei einer Verdopplung auf 400'000 IOPS SPC-1 müssten unter Einhaltung einer maximalen Auslastung von 50% bereits zwei FS 900 eingesetzt werden. Hier kann die Leistung von den FS 900 immer noch mit deutlich geringer Latency und geringeren Kosten erbracht werden. Bei 800'000 IOPS und Einsatz von 4 FS 900 sind die Kosten vergleichbar bei immer noch besserer Latency. Lastprofile mit mehr als 800'000 IOPS sind nicht sehr realistisch. Erkennbar ist auch, dass die SSD Kapazität pro TB im Highend Storage von Huawei 27'500 USD kostet und auf dem FS 900 70 nur 12'000 USD.

Fazit

Ist höchste Performance die oberste Prämisse, empfiehlt sich aufgrund der sehr geringen Latenzzeiten ein All-Flash Array. Dabei sind Funktionen wie Spiegelung oder Snapshots auf applikatorischer Ebene oder OS Ebene zu lösen. Auch hier gilt es den Performance Impact zu prüfen.

Sofern zusätzliche storagebasierte Funktionen benötigt werden, können All-Flash Systeme beispielsweise in Kombination mit IBM Sprectrum Virtualize (SVC) zur Erweiterung der Funktionalität betrieben werden. Dabei ist die zusätzliche Larency durch IBM Sprectrum Virtualize (SVC) zu beachten, wenn die Applikation nicht vom zusätzlichen RAM Cache dieser Lösung profitieren kann.

Geht es vor allem um Kapazität, weisen herkömmliche Tiered Midrange Storagesysteme aktuell noch einen Kostenvorteil von Faktor 3-5 im Vergleich zu All-Flash Systemen auf, der allerdings zunehmend erodiert. Bei Einsatz von Datenkompression oder Deduplizierung auf Flash Systemen können die Kosten allerdings bereits heute auf ein vergleichbares Niveau gedrückt werden, auf Kosten von etwas höheren Antwortzeiten gegenüber einem nativen All-Flash System.

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 Mitglied der Geschäftsleitung.

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 der In&Out Performance Benchmarking Tools IOgen™ (Storage IO Benchmarks), NETgen™ (Netzwerk Benchmarks) und CPUgen™ (CPU Benchmarks) und hat in den letzten Jahren sehr viele Kunden- und Hersteller-Benchmarks durchgeführt.