



Workshop

Solaris / SAM-QFS Host Anbindung ST6x40/ST25x0

Ralf Werner

Senior Storage Consultant

Sun Microsystems GmbH

Storage Group

+49 173 6506175

Einführung

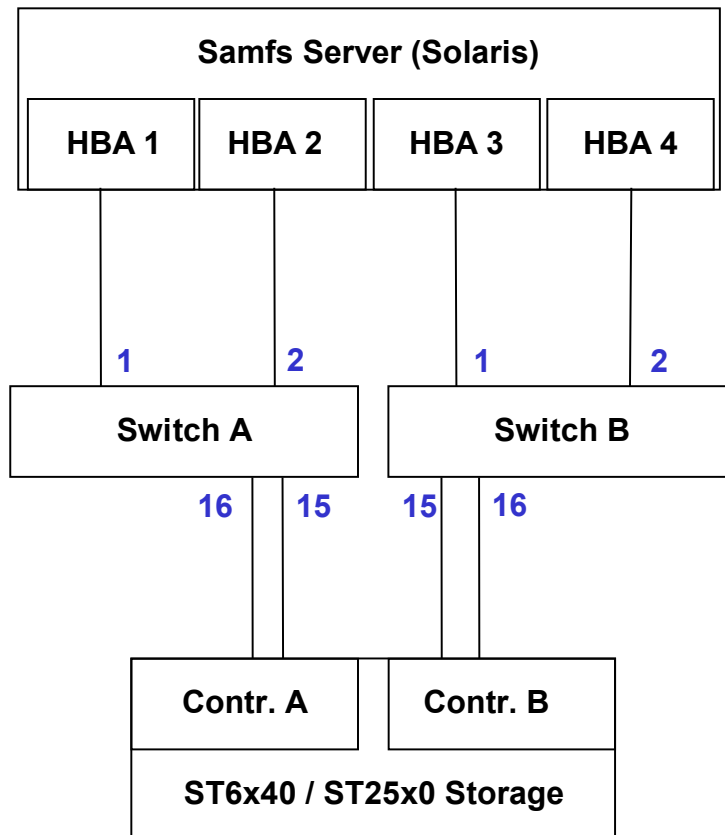
- **Die Konfigurationshinweise in dieser Präsentation gelten sowohl für ST6x40 als auch für ST25x0 Systeme**
 - > Die relevanten Eigenschaften sind bei beiden Systemen identisch implementiert
 - > Cache Mirroring, IO-Alignment, Stripe Sets, ...
- **Im wesentlichen zielt diese Präsentation darauf ab, größere Konfigurationen möglichst optimal zu implementieren**
 - > Full-Stripe-Write Konfiguration, um das Cache Mirroring Limit der Storage Controller zu umgehen
 - > Ohne Full-Stripe-Writes sind die folgenden maximalen Durchsatzwerte zu erwarten
 - > ST6540: ca. 400MB/s bis 440MB/s (Write)
 - > ST6140: ca. 280MB/s bis 310MB/s (Write)
 - > Mit Full-Stripe-Writes sind die folgenden maximalen Durchsatzwerte zu erwarten
 - > ST6540: ca. 1100 MB/s bis 1200 MB/s (Write)
 - > ST6140: ca. 600 MB/s bis 700 MB/s (Write)
 - > Werte sind keine zugesicherten Eigenschaften
 - > Werte abhängig von Blockgröße, Anzahl Festplatten, Anzahl parallele Streams, ...

Einführung

- **Für kleinere Konfigurationen mit wenigen Volumes kann es ggf. vorteilhafter sein, kein Full-Stripe-Write zu nutzen, wenn der notwendige Gesamtdurchsatz unter den oben genannten Cache Mirroring Limits der Controller liegt**
 - > Kann den Durchsatz für ein einzelnes Volume erhöhen, der mit Write Cache höher sein kann, als ohne Write Cache (Full-Stripe-Write)

Zoning

- **WWPN- oder Port-Zoning**
 - > Switch nicht auf “Durchzug” stellen
 - > Dringend empfohlen: “Single-Initiator-Zoning”



Zonen

Switch A

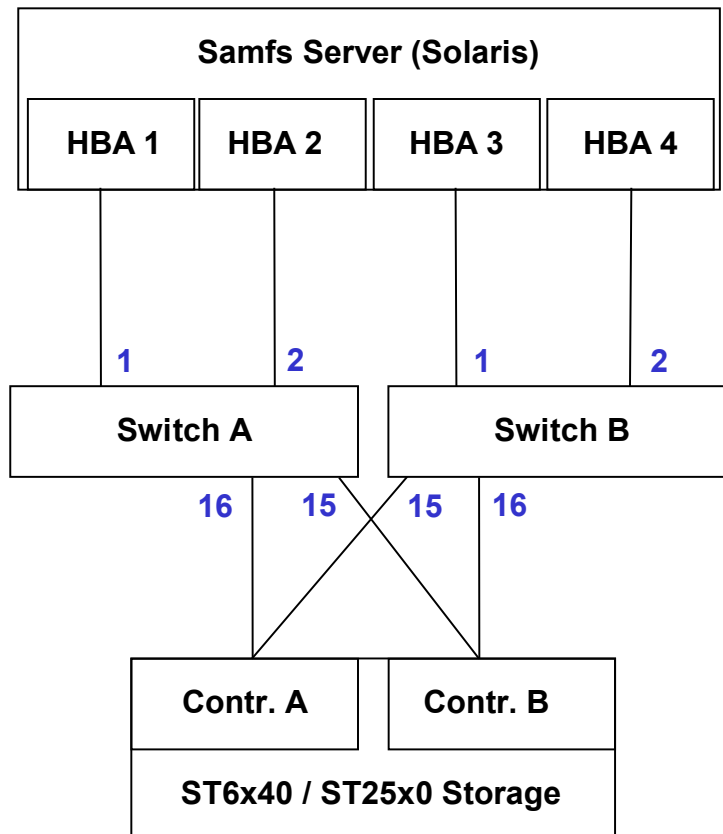
- > Zone 1: Port 1 & Port 16
- > Zone 4: Port 2 & Port 15

Switch B

- > Zone 1: Port 1 & Port 16
- > Zone 4: Port 2 & Port 15

Zoning

- **Optional: Zusätzliche “Kreuzverkabelung” zwischen Switch und Storage**
 - > Jedes Volume über alle vier HBA's erreichbar
 - > Round-Robin Load-Balancing über alle vier HBA's zu einem Controller
 - > Kein Volume Transfer auf anderen Controller bei Switch Ausfall



Zonen

Switch A

- > Zone 1: Port 1 & Port 16
- > Zone 2: Port 2 & Port 16
- > Zone 3: Port 1 & Port 15
- > Zone 4: Port 2 & Port 15

Switch B

- > Zone 1: Port 1 & Port 16
- > Zone 2: Port 2 & Port 16
- > Zone 3: Port 1 & Port 15
- > Zone 4: Port 2 & Port 15

Multipathing

Solaris 10 Sparc und x86

- **Solaris 10 Sparc und x86**
 - > Alle Treiber im Betriebssystem enthalten
 - > Default auf Sparc Systemen: Multipathing ist deaktiviert
 - > Default auf x86 Systemen: Multipathing ist aktiviert

- **Aktivierung und Konfiguration des Multipathing**

- > Mit einzelnen Kommando: *stmsboot -e*

```
# stmsboot -e
```

```
WARNING: This operation will require a reboot.
```

```
Do you want to continue ? [y/n] (default: y) <enter>
```

- > Reboot des Servers notwendig

Maximum IO Size

Solaris 10 Sparc und x86

- **Maximale IO Size auf Solaris sollte angepasst werden**
 - > Verhindert, das große IO's in kleinere "zerhackt" werden
 - > Default Wert für Maximale IO Size ist 128kB auf Sparc bzw. 56kB auf x86
- **Maximale IO Size auf 8 MB einstellen**
 - > Eintrag Solaris Sparc und x86 in /etc/system:
set maxphys = 0x800000
 - Wert kann überprüft werden mit: *echo "maxphys/D" | mdb -k*
 - > Eintrag Solaris Sparc in /kernel/drv/ssd.conf:
ssd_max_xfer_size = 0x800000;
 - > Eintrag Solaris x86 in /kernel/drv/sd.conf:
sd_max_xfer_size = 0x800000;

Filesystem

SAM-QFS

- **Filesystem Typ “ma”**
 - > Ist in der Lage Nutzdaten und Metadaten auf getrennten Volumes zu halten
 - > Volumes für Nutzdaten
 - > Volumes vom Equipment Typ “mr”, “md” oder “mXYZ”
 - > Striping über mehrere Volumes möglich
 - > Volumes für Metainformationen
 - > Volumes vom Equipment Typ “mm”
 - > Kleine IO Blockgröße (DAU Size) für diese Volume (4k oder 16k DAU Size)
 - > Striping der Metainformationen über mehrere Volumes möglich
 - > Gute Basis um hohen Durchsatz zu erreichen
 - > Equipment Typ “mr” oder “mXYZ” lassen große IO Blockgrößen (DAU Size) zu
 - > Getrennte Nutzdaten und Metadaten
 - > Shared SAN Filesystem mit QFS Lizenz
 - > Mehrere Server können via FC gleiches Volume nutzen
 - > z.B. Single Writer/Shared Reader, Shared Reader/Writer

Filesystem

SAM-QFS

- **Filesystem Typ “ms”**
 - > Nutzdaten und Metadaten in einem Filesystem
 - > Volumes nur vom Equipment Typ “md” (kein “mr” und “mXYZ”)
 - > Striping über mehrere Volumes möglich
 - > Keine gute Basis um hohen Durchsatz zu erreichen
 - > Equipment Typ “md” läßt nur relativ kleine IO Blockgrößen zu
 - > Equipment Typ “md” hat keine konstante IO Blockgröße (Schlecht für Full-Stripe-Write)

SAM-QFS Parameter

Equipment Typ

- “md” - Dual Allocation
 - > Es wird erst mit 4k IO Blockgrößen auf die Volumes geschrieben
 - > Bis die Dateigröße 32k erreicht hat
 - > Danach wird mit IO Blockgrößen der eingestellten DAU Size geschrieben
 - > Nur DAU Size von 16k, 32k und 64k möglich
 - > Keine gute Basis für hohen Durchsatz
 - > Full-Stripe-Write nur für eine Blockgröße je Volume möglich
 - > Nur kleine IO-Blockgrößen möglich
- “mr” -
 - > Es wird nur mit konstanter IO Blockgröße (DAU Size) geschrieben
 - > DAU Size läßt sich flexibel einstellen (16k bis mehrere MB)
 - > Gute Basis für hohen Durchsatz
 - > Full-Stripe-Write für die Volumes möglich
 - > Große IO-Blockgrößen möglich

SAM-QFS Parameter

Equipment Typ

- “mXYZ” - Stripe Groups
 - > Volumes werden zu Stripe Groups zusammen gefasst
 - > In großen Umgebungen (viele Volumes) Separierung von Streams
 - > Kann konkurrierende Zugriffe auf Volumes reduzieren
 - > Durchsatz für einzelnen Stream durch grÖÙe der Stripe Group limitiert
 - > Es wird nur mit konstanter IO BlockgrÖÙe (DAU Size) geschrieben
 - > DAU Size lÄÙt sich flexibel einstellen (16k bis mehrere MB)
 - > Gute Basis für hohen Durchsatz
 - > Full-Stripe-Write für die Volumes möglich
 - > GrÖÙe IO-BlockgrÖÙen möglich
 - > Trennung von Streams in Stripe Groups

SAM-QFS Parameter

Disk Allocation Unit (DAU) Size

- **DAU Size für Datenbereiche**
 - > Gibt die IO Blockgröße an, mit der geschrieben wird
 - > DAU Size ist auch die Mindestkapazität, die für eine einzelne Datei allokiert wird
 - > Bei kleinen Dateien wird ggf. durch große DAU Size Kapazität verschwendet
 - > DAU Size sollte/muß identisch mit der Stripe Width sein
 - > Für Full-Stripe Writes, damit Cache Mirroring der ST6x40 Controller nicht den limitierenden Faktor für den Durchsatz darstellt
 - > Größere DAU Size führt zu größeren IO's auf den Festplatten
 - > Mit großen IO's kann höherer Durchsatz je Festplatte erzielt werden als mit kleinen IO's

SAM-QFS Parameter

Stripe Size

- **Stripe Size = 0** → **Komplette Datei wird auf ein einzelnes Volume geschrieben**
 - > Gilt für Equipment Type “mr” und “md”
 - > Bei Equipment Type “mXYZ” erfolgt Striping über alle Volumes innerhalb einer Stripe Group
 - > Bietet sich an, wenn viele Streams (Dateien) parallel geschrieben werden
 - > Durchsatz je Stream durch das Volume limitiert
- **Stripe Size = 1** → **Datei wird über alle Volumes gestriped**
 - > Equipment Type “mr” und “md”: 1 DAU je Volume bei
 - > Bei Equipment Type “mXYZ” erfolgt Striping über Stripe Groups (1 DAU je Stripe Group)
 - > Bietet sich an, wenn sehr wenige Streams (Dateien) parallel geschrieben werden
 - > Sehr hoher Durchsatz, limitiert durch Anzahl Festplatten und Controller Limit
- **Stripe Size = 2** → **Datei wird über alle Volumes gestriped.**
 - > Equipment Type “mr” und “md”: 2 DAU's je Volume bei
 - > Bei Equipment Type “mXYZ” erfolgt Striping über Stripe Groups (2 DAU's je Stripe Group)
 - > Bietet sich an, wenn nicht mit Full-Stripe-Write geschrieben werden
 - > Einzelner Stream (oder zumindest wenige) kann hohen Durchsatz erreichen, limitiert durch Anzahl Festplatten und Controller Limit
- **Stripe Size größer 2 verhält sich analog zu Stripe Size = 2**

Welche Parameter für hohen Durchsatz?

- **Angaben basieren auf Erfahrungen, mit denen gute Resultate erzielt wurden**
- **Verwendete Parameter aber immer abhängig von den individuellen Anforderungen und der Gesamtkonfiguration**
- **Es gibt keine universelle Konfiguration, die bei allen Installationen die leistungsfähigste ist**

▶ **Ausprobieren!**

- **Filesystem: SAM-QFS**
 - > Trennung von Metadaten und Nutzdaten
 - > Es kann “mr” und “mXYZ” Equipment Typ verwendet werden
 - > Größere DAU Size von “mr” und “mXYZ” führt zu größeren IO's auf den Festplatten
 - Erhöht Durchsatz je Festplatte
 - > Ermöglicht Full-Stripe-Write Konfiguration

Welche Parameter für hohen Durchsatz?

- **Equipment Typ: “mr” oder “mXYZ”**
 - > Abhängig von Anforderungen und Hardware Ausstattung (z.B. Anzahl Festplatten)
 - > “mr”
 - > Eher bei kleineren Umgebungen (weniger Volumes)
 - > “mXYZ”
 - > Eher bei sehr großen Umgebungen
 - > Wenn mit einem Teil der vorhandenen Festplatten schon das Durchsatzlimit des Controller erreicht werden kann
 - > Um Durchsatz je Stream zu begrenzen (auf Leistung der Stripe Group)
- **DAU Size: Groß (1024k oder 2048k)**
 - > Abhängig von Dateigrößen, die geschrieben werden
 - > Wenn Dateien üblicherweise sehr groß, dann DAU auf 2048k einstellen
 - > DAU Size größer 2048K bei ST6x40 mit Full-Stripe-Write nicht hilfreich
 - Überschreitet Large IO Size der ST6x40 Systeme
 - > Große IO's auf den Festplatten
 - > Erhöht üblicherweise Durchsatz je Festplatte

Welche Parameter für hohen Durchsatz?

- **Stripe: 0 oder 1**
 - > Stripe=0
 - > Bei Equipment Typ “mXYZ”
 - > Eher bei viele Streams mit kleinen Dateien bei Equipment Typ “mr”
 - > Stripe=1
 - > Eher bei wenige Streams mit großen Dateien bei Equipment Typ “mr”

ST6x40 Einstellungen

Vdisks

- **Optimierung für hohen Durchsatz (MBPS)**
 - > Festplatten einer Vdisk sollten unter Berücksichtigung der optimalen Leistung in den Speichersystemen ausgewählt werden (an Loop-Switch Contention denken)
- **RAID-5 hat sich bewährt**
 - > Bestes Verhältnis zwischen Leistung und Kapazität
- **Für Full-Stripe-Write Nutzung nur RAID-5 Vdisks mit 5 oder 9 Festplatten verwenden**
 - > 4 oder 8 Nettoplatten je Vdisk ergeben bei jeder wählbaren Segment Size der Volumes jeweils eine Stripe Width mit Base2
 - > Stripe Width von 64k, 128k, 256k, 512k, 1024k, 2048k, 4096k, ...
 - > Keine "krummen" Werte wie beispielsweise 112k (7*16k), 896k (7*128k)
 - > Bessere Kapazitätsausnutzung bei 9 Festplatten je Vdisk
 - > Vdisk mit 9 Festplatten empfiehlt sich, wenn auch der Durchsatz auf einer einzelnen LUN möglichst hoch sein soll
 - > Erhöht den Durchsatz eines Streams, wenn Samfs Stripe Size = 0 verwendet wird
 - > Samfs Stripe Size = 0 bewirkt, dass eine komplette Datei auf ein Volume geschrieben wird. Nächste Datei wird das auf nächstes Volume geschrieben

ST6x40 Einstellungen

Volumes

- **Optimierung für hohen Durchsatz (MBPS)**
- **Möglichst nur eine LUN je Vdisk**
 - > Verringert die Wahrscheinlichkeit das konkurrierende IO's auftreten
- **Große Segment Size für Volumes wählen**
 - > Erhöht Durchsatz je Disk (und dadurch auch je Vdisk)
 - > Gute Werte für Segment Size sind 256k bis 512k
 - > Auch andere Werte möglich
 - > Abhängig von individueller Anforderung und Konfiguration
 - > Große Segment Size erfordert große QFS DAU Size für hohen Durchsatz
- **Cache Mirroring für Volumes aktivieren (Standardeinstellung)**
 - > Stellt Datenintegrität sicher
 - > Bei Full-Stripe-Writes schreibt Controller automatisch direkt auf Festplatten ohne den Cache zu nutzen
 - > Deaktivieren des Cache Mirrorings erhöht Durchsatz, wenn keine Full-Stripe-Writes geschrieben werden, kann jedoch zu Datenverlust führen

Solaris Disks mit EFI Label Slices

- **Für QFS / SAM-QFS / SAM-FS wird die Verwendung von EFI Labels empfohlen**
 - > Vereinfacht das IO-Alignment, da der Beginn einer Partition über Eingabe des Start-Sektors erfolgen kann und nicht auch einer Cylindergrenze liegen muss
 - > Bei VTOC-Label muss Start-Sektor der Partition auf Cylindergrenze liegen
- **EFI Label reservieren bei Standardformatierung die ersten 34 Sektoren**
 - > 34 Sektoren entsprechen 17 KB
 - > Der Anfang einer Slice liegt somit ohne IO-Alignment bei keiner möglichen Segment Size der ST6x40 oder ST25x0 Systeme am Anfang eines Segments
 - > Dies führt zu Misalignment mit der Segment Size des Volumes
 - > Sektor für den Start der Slices muss angepasst werden
 - > Eine Slice sollte an einem Stripe Set des Volumes beginnen
 - > Ideal für Full Stripe Writes
 - > Ideal für hohen Durchsatz in großen Umgebungen (viele Volumes)

IO-Alignment

- **Partition muss an einem Stripe Set beginnen**
- **Samfs DAU Size muß der Stripe Width des Volumes entsprechen**
 - > $\text{Stripe Width (Stripe Set Size)} = \text{Segment Size} * \text{Netto-Festplatten der Vdisk}$
- **Samfs Stripe Size nicht relevant für IO-Alignment**
 - > Stripe Size kann nach Bedarf abweichen

Anzahl Disks je Vdisk	RAID-Level	ST6x40 LUN Segment Size	ST6x40 Cache Block Size	QFS DAU Size	Partition Start Sektor für IO-Alignment
5	5	16k	16k	64k	128
5	5	32k	16k	128	256
5	5	64k	16k	256k	512
5	5	128k	16k	512k	1024
5	5	256k	16k	1024k	2048
5	5	512k	16k	2048k	4096
9	5	16k	16k	128k	256
9	5	32k	16k	256k	512
9	5	64k	16k	512k	1024
9	5	128k	16k	1024k	2048
9	5	256k	16k	2048k	4096
9	5	512k	16k	4096k	8192

- **Start Sektor 8192 passt immer bei Raid-5 (4+1) oder Raid-5 (8+1)**
 - > Unabhängig von verwendeter Segment Size

Solaris Disks mit EFI Label

Beispiel: Default Partitionierung mit Mis-aligned IO's

- **Beispiel:**
 - > Stripe Width des Volumes 512k
 - > Segment Size 64k bei RAID-5 (8+1)
 - > Segment Size 128k bei RAID-5 (4+1)

- **format Kommando**

Current partition table (original):

Total disk sectors available: 2576924638 + 16384 (reserved sectors)

IO's um 34 Sektoren "verschoben". Bei 512k IO werden 495K auf den ersten Stripe und 17k auf den nächsten Stripe geschrieben.

Part	Tag	Flag	First Sector	Size	Last Sector
0	root	wm	34	1.20TB	2576924636
1	unassigned	wm	0	0	0
2	unassigned	wm	0	0	0
3	unassigned	wm	0	0	0
4	unassigned	wm	0	0	0
5	unassigned	wm	0	0	0
6	unassigned	wm	0	0	0
8	reserved	wm	2576924638	8.00MB	2576941021

Solaris Disks mit EFI Label

Beispiel: Partition anlegen (Solaris x86)

format -e <device>

fdisk (Frage nach 100% Solaris Partition mit "no" beantworten)

1 (Neue Partition anlegen)

f (Partition vom Typ EFI)

5 (Label auf Disk schreiben und Submenü verlassen)

label

1 (Partition vom Typ EFI)

partition (Jetzt kann die Partition Table nach belieben angelegt werden.
Es können jetzt auch Sektoren für den Start der Partition gewählt werden und nicht nur Cylinder)

label (Entgültigen Label schreiben)

q

q

Solaris Disks mit EFI Label

Beispiel: Partitionierung mit IO-Alignment

- **Beispiel:**
 - > Stripe Width des Volumes 512k
 - > Segment Size 64k bei RAID-5 (8+1)
 - > Segment Size 128k bei RAID-5 (4+1)

- **format Kommando**

Current partition table (original):

Total disk sectors available: 2576924638 + 16384 (reserved sectors)

IO beginnt an einem Stripe Set. Die 512k des IO werden auf einen Stripe geschrieben. (Full-Stripe-Write)

Part	Tag	Flag	First Sector	Size	Last Sector
0	root	wm	1024	1.20TB	2576924636
1	unassigned	wm	0	0	0
2	unassigned	wm	0	0	0
3	unassigned	wm	0	0	0
4	unassigned	wm	0	0	0
5	unassigned	wm	0	0	0
6	unassigned	wm	0	0	0
8	reserved	wm	2576924638	8.00MB	2576941021

Mount Parameter

- **Write Behind**
 - > Guter Startwert ist: Write Behind Wert = 2 * DAU Size
 - > Nicht relevant, wenn Direct IO verwendet wird
 - ▶ **Ausprobieren!**
- **Direct IO**
 - > Gute Erfahrung mit Direct IO
 - > Als Mount-Parameter forcedirectio verwenden
 - > Kann Durchsatz erhöhen
 - ▶ **Ausprobieren!**

Meßwerte von Konfigurationen

- Gemessen mit vdbench
- Werte sind keine zugesicherten Eigenschaften

Modell	Drives	Anzahl Vdisks	Disks je Vdisk	RAID-Level	Seg. Size	DAU Size	Streams	FS	Durchsatz MB/s	Bemerkung
1 * ST6540	300GB/10k FC	1	5	5	64k	256k	1	SAM-QFS	52	Full-Stripe-Write
1 * ST6540	300GB/10k FC	28	5	5	64k	256k	8	SAM-QFS	1330	Full-Stripe-Write
3 * ST6540	300GB/10k FC	84	5	5	64k	256k	8	SAM-QFS	3440	Full-Stripe-Write



Q & A

Danke!



THE MATERIAL CONTAINED WITHIN THIS PRESENTATION MAY NOT BE ALTERED OR DUPLICATED IN ANY WAY WITHOUT THE EXPRESS AUTHORISATION OF THE AUTHOR.

MENTION OF THIRD-PARTY PRODUCTS IS FOR INFORMATION PURPOSES ONLY AND SUN ACCEPTS NO LIABILITY FOR THEIR SELECTION OR PERFORMANCE.