

Designkriterien für die optimierte Hochverfügbarkeit eines IBM® RS/6000® SP® Systems mit SSA-Platten

Zusammenfassung:	Es werden die Möglichkeiten der "Single Points of Failure" einer IBM® RS/6000® SP® aufgezeigt und konzeptionelle Maßnahmen empfohlen, die Hochverfügbarkeit eines SP® Clusters und seiner SSA-Platten sicherzustellen bzw. zu optimieren
Autor:	Richard Handforth, UK richard@dedgug.freeseerve.co.uk
Ausgabe:	1.0
Datum:	30.09.1999
Referenz:	GE9002 (www.mannherz.de)

Die Inhalte dieser Seiten dürfen ohne vorherige schriftliche Zustimmung des Herausgebers nicht verändert und nicht öffentlich vorgeführt oder für kommerzielle Zwecke vervielfältigt werden. Dies gilt auch für eine Nutzung der Inhalte dieser Seiten auf anderen Web-Seiten oder vernetzten Rechnern.

Diese Seiten und die enthaltenen Informationen wurden nach bestem Wissen und Gewissen zusammengestellt, eine Rechtsverbindlichkeit kann hieraus nicht abgeleitet werden. Die Haftung für die Richtigkeit, Vollständigkeit und Verlässlichkeit von Informationen, Daten und Dokumenten wird auf die Fälle grob fahrlässigen oder vorsätzlichen Verhaltens beschränkt für jegliche Verluste oder Schäden (auch Folgeschäden), die dadurch entstehen, dass die Nutzerin oder der Nutzer auf Informationen vertraut, die er im Rahmen der Nutzung des Dienstes erhalten hat. Ebenfalls wird die Haftung auf die Fälle grob fahrlässigen oder vorsätzlichen Verhaltens beschränkt für jegliche Verluste oder Schäden (auch Folgeschäden), die der Nutzerin oder dem Nutzer auf andere Weise bei der Nutzung dieser Seiten entstehen (z. B. durch Herunterladen von Web-Seiten o. ä.).

Soweit wir vertraglich gegenüber bestimmten Kunden dazu verpflichtet sind, bestimmte Inhalte auf unseren Web-Seiten zur Verfügung zu halten, wird die Haftung auf die Fälle grob fahrlässigen oder vorsätzlichen Verhaltens beschränkt für jegliche Verluste oder Schäden (auch Folgeschäden), wobei dies für die leicht fahrlässige Verletzung vertragswesentlicher Pflichten nicht gilt.

Verantwortlich für den Inhalt ist die Autorin bzw. der Autor.



Inhaltsverzeichnis

1. Designkriterien für Festplatten- und Datenverfügbarkeit.....	3
1.1 Verfügbarkeit von SSA-Platten.....	3
1.2 Quorum und Spiegelung.....	3
1.2.1 Datenträgergruppe mit zwei Platten.....	4
1.2.2 Datenträgergruppe mit vier Platten.....	5
1.2.3 Datenträgergruppe mit sechs Platten.....	5
1.2.4 Datenträgergruppe mit acht Platten.....	5
1.2.5 Die Kombination von Quorum und Spiegelung.....	6
1.2.6 Spiegelung bei ungerader Plattenanzahl.....	6
1.2.7 Anordnung der Spiegelkopien.....	6
1.2.8 Plattenübergreifende logische Datenträger.....	7
1.3 Wiedereingliederung der Platten.....	7
1.3.1 Ersetzen ausgefallener Platten.....	7
1.3.2 Resynchronisierung veralteter Plattenpartitionen.....	7
2. "Single Points of Failure" (Spof) bei der SP® Hardware.....	8
2.1 Control Workstation.....	8
2.2 SP® Frame.....	8
2.3 SSA-Plattenschränke.....	9
2.4 SSA-Platten-Subsysteme.....	9
2.4.1 Netzteile.....	9
2.4.2 Signal-/Umgebungskarten des Subsystems.....	10
2.5 SSA-Platten-Schleifen.....	10
3. SPoFs in SP® Netzwerken.....	11
3.1 Internes Netzwerk.....	11
3.2 Client-Netzwerke.....	12
3.3 Switch-Netzwerke.....	12
4. Marken.....	14

1. Designkriterien für Festplatten- und Datenverfügbarkeit

1.1 Verfügbarkeit von SSA-Platten

Die Anordnung von gespiegelten Kopien logischer Datenträger (logical Volumes) auf Festplatten sowie die Frage, ob Quorum (Mehrheitsentscheid) ein- oder ausgeschaltet sein sollte, sind wichtige Designüberlegungen in einer HACMP-Umgebung. Schlechte Planung und mangelhaftes Design kann bedeuten, dass ein Festplattenausfall als „Single Point of Failure“ (SPoF) zu Ausfällen von Anwendungen, Datenkorruption und - in einigen Fällen - zu einem Systemabsturz führt. Normalerweise werden die Funktionen des ausgefallenen Knotens von dem automatisch übernehmenden System weitergeführt, wenn aber die Daten bereits verfälscht sind, kann die Anwendung deshalb weiterhin unbrauchbar sein. Außerdem muss der Ausfall einer Anwendung nicht notwendigerweise von HACMP erkannt werden, sodass eine automatische Übernahme auf einen anderen Knoten nicht erfolgt.

Alle SSA-Plattenkonfigurationen sollten über gespiegelte logische Datenträgerkopien verfügen, um ein Höchstmaß an Verfügbarkeit sicherzustellen. Dies wird jedoch aus Kostengründen nicht immer realisierbar sein. Wird keine Spiegelung verwendet, kann im Fall eines Ausfalls der Platte oder eines fehlerhaften Zugriffs auf das Plattensubsystem - aus welchen Gründen auch immer - die defekte Hardware nicht ohne eine längere Ausfallzeit ersetzt werden. In einigen Systemumgebungen mag dies tolerierbar sein, aber generell sind solche Konsequenzen für Anwendungen, die hochverfügbar sein müssen, nicht hinnehmbar.

Die Anordnung von Platten und deren Spiegelkopien innerhalb von Subsystemen und über deren Grenzen hinweg ebenso wie deren Verbindungen innerhalb der SSA-Schleifen sollte daher sorgfältig geplant werden, um ein Höchstmaß an Datenverfügbarkeit in einem Cluster sicherzustellen und Ausfallzeiten für die Wiedereingliederung ausgefallener Komponenten zu minimieren. Die Konfiguration von SSA-Platten muss derart erfolgen, dass - egal, welcher Knoten die Ressourcen innerhalb des Clusters steuert - die Hochverfügbarkeit der Platten und Daten sichergestellt ist, wenn bei der Unterbrechung einer Schleife eines der folgenden Kriterien zutrifft:

- Ein Knoten im Cluster kann auf alle Platten zugreifen, die sich in der von ihm verwalteten Ressourcengruppe befinden. Dies ist der Fall, wenn die Schleife unterbrochen wurde, aber alle Platten weiterhin über alternative Pfade erreichbar sind.
- Ein Knoten im Cluster kann auf einen vollständigen Satz von gespiegelten Platten innerhalb der Ressourcengruppe zugreifen. In diesem Fall können entweder einzelne Platten ausgefallen sein oder auf mehrere Platten kann wegen eines Adapter-, Subsystem- oder Knotenausfalls nicht zugegriffen werden.

In Clustern, deren Knoten mit nur einem einzigen SSA-Adapter ausgestattet sind, kann u.U. die Übernahme durch einen anderen Knoten erforderlich sein, damit die genannten Designvorgaben erreicht werden.

SSA-Schleifenausfälle entstehen innerhalb einer SSA-Konfiguration wenn

- ein Knoten,
- ein SSA-Adapter,
- ein SSA-Kabel,
- ein Platten-Subsystem,
- eine Platte innerhalb des Subsystems oder
- eine Signal-/Umgebungskarte ausfällt.

Der resultierende Plattenzugriff (oder sein Fehlschlagen) wird durch die Art und den Entstehungsort des Fehlers bestimmt. Die Konfiguration muss so geplant werden, dass die Konsequenzen dieser Effekte minimiert werden.

1.2 Quorum und Spiegelung

Wenn ein Datenträger aktiviert wird und Quorum eingeschaltet ist, müssen mehr als 50% der "Volume Group Descriptor Areas" (VGDA) auf den Platten ansprechbar sein, damit der Zugriff auf die Datenträgergruppe erhalten bleibt. Die VGDA's enthalten Informationen über die Datenträgergruppe sowie Zeitmarken, die zum Vergleich der Platten herangezogen werden, um festzustellen, ob logische Datenträger

nicht mehr aktuell (stale) sind.

Bei einer Datenträgergruppe, die nur aus einer Platte besteht, enthält diese zwei VGDA's. Bei einer Datenträgergruppe bestehend aus zwei Platten enthält die erste Platte zwei VGDA's und die zweite eine VGDA. Jede Datenträgergruppe mit drei oder mehr Platten enthält eine einzige VGDA auf jeder Platte.

Falls bei einer Datenträgergruppe mit zwei Platten diejenige mit den beiden VGDA's ausfällt, geht das Quorum verloren, die Datenträgergruppe wird deaktiviert und die Daten sind nicht länger verfügbar, selbst wenn die Spiegelung aktiviert war. Im Fall der Datenträgergruppen mit drei bzw. vier Platten darf maximal eine einzige Platte ausfallen, damit die Datenträgergruppe weiterhin aktiviert bleibt, da immer mehr als 50% der VGDA's verfügbar sein müssen.

Es ist möglich, das Quorum für eine Datenträgergruppe abzuschalten. Dies sollte jedoch nur dann erfolgen, wenn eine Spiegelung etabliert ist. Das bedeutet letztendlich, dass alle Platten bis auf eine ausfallen können und die Datenträgergruppe weiterhin aktiviert bleibt - trotzdem kann ein Teil der Daten, abhängig von der Anwendung, nicht nutzbar sein. Unter diesen Umständen muss dann auch mit der Möglichkeit der Datenkorrumpierung gerechnet werden.

Unter AIX® findet die Spiegelung auf der Ebene der logischen Datenträger unter Verwendung von einer bzw. zwei Spiegelkopien statt. Sämtliche Kopien sollten dabei auf verschiedenen Platten angelegt werden, dies ist jedoch nicht zwingend erforderlich. Alle Kopien können auch auf der gleichen Platte angeordnet sein, aber dieses Vorgehen entspricht nicht dem grundsätzlichen Ansatz der Spiegelung. Das Vorhandensein von drei Kopien erhöht die Sicherheit, dem stehen aber höhere Kosten gegenüber.

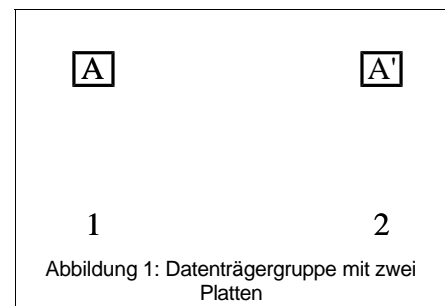
Idealerweise sollte in einer HACMP-Umgebung die Spiegelung und das Quorum **immer** aktiviert sein, um die Datenintegrität sicherzustellen. Unglücklicherweise bedeutet das für Datenträgergruppen mit zwei Festplatten (eine gespiegelt) eine mögliche Reduktion der Verfügbarkeit, falls bei aktivem Quorum die Platte mit den beiden VGDA's ausfällt.

Obwohl die Datenträgergruppe mit zwei Festplatten einen Spezialfall darstellt, ist in allen anderen Fällen ein Kompromiss zu finden zwischen den Anforderungen an die Verfügbarkeit und der potenziellen Möglichkeit der Datenkorrumpierung (für den Fall, dass Quorum abgeschaltet ist). Ganz besonders gilt dies, wenn eine gerade Anzahl von Platten gleichmäßig auf zwei Subsysteme verteilt sind und wo bereits der Ausfall eines Subsystems den Verlust des Quorums bedeutet.

Wenn das Quorum eingeschaltet ist, können die Platten über die SSA-Subsysteme wie im Folgenden beschrieben verteilt werden, damit die Verfügbarkeit einer Datenträgergruppe auch im Fall eines Subsystem-Ausfalls gegeben ist. Es ist zu beachten, dass die Spiegelung auf der Ebene der logischen Datenträger und nicht der physikalischen Platte ansetzt. In den folgenden Beispielen wird angenommen, dass alle logischen Datenträger auf jeder Platte an die gleiche Position auf der zugeordneten gespiegelten Platte dupliziert werden.

1.2.1 Datenträgergruppe mit zwei Platten

Es ist nicht möglich, zwei Platten über zwei Subsysteme anzuordnen und sicherzustellen, dass die Datenträgergruppe immer aktiv bleibt, wenn ein Subsystem ausfallen sollte und Quorum eingeschaltet ist. Es ist empfehlenswert, dass Quorum ausgeschaltet ist, weil dann die Datenträgergruppe aktiviert bleibt, egal welche Platte ausfallen sollte. Eine Datenkorrumpierung ist unwahrscheinlich, da im Fall eines zusätzlichen Ausfalls der zweiten Platte keinerlei Daten mehr gelesen oder geschrieben werden können (keine Platte verfügbar). In der gezeigten Situation ist es wichtig, dass die Systemuhren aller Knoten im Cluster korrekt eingestellt sind. Nur so kann - falls eine Übernahme durch HACMP vor dem Wiederausammenführen der Subsysteme erfolgte - nach der Wiederverfügbarkeit des Subsystems die Resynchronisation der logischen Datenträger erfolgreich sein. Sie erfolgt dann auf der Basis der korrekten aktuellen (zeitnahen) Kopie auf die inzwischen veraltete Kopie.



Eine Alternative zum Abschalten des Quorums ist, eine dritte "Quorum-Buster"-Platte, die sich in einem

dritten Subsystem befindet, der Datenträgergruppe hinzuzufügen. Dies erfordert zwar einen höheren Hardwareaufwand, bietet aber die Möglichkeit, die logischen Datenträger auf drei Platten zu verteilen (ein Beispiel für die Anordnung einer ungeraden Anzahl von Platten wird später gezeigt).

1.2.2 Datenträgergruppe mit vier Platten

Eine Datenträgergruppe aus vier Platten sollte über vier Subsysteme verteilt werden, damit sichergestellt ist, dass die Datenträgergruppe stets aktiv bleibt, egal welches Subsystem ausfallen sollte. Wenn keine vier Subsysteme zur Verfügung stehen, sollte eine "Quorum-Buster-Platte" hinzugefügt werden, in diesem Fall sollte eine 2-2-1-Anordnung gewählt werden.

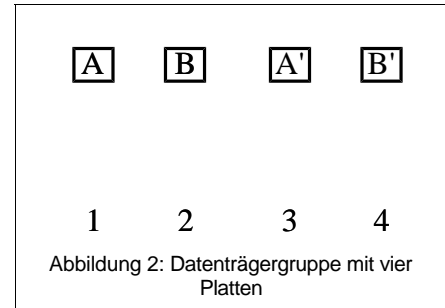


Abbildung 2: Datenträgergruppe mit vier Platten

1.2.3 Datenträgergruppe mit sechs Platten

Eine Datenträgergruppe mit sechs Platten muss auf drei Subsysteme verteilt werden, damit das Quorum bei Ausfall eines Subsystems erhalten bleibt.

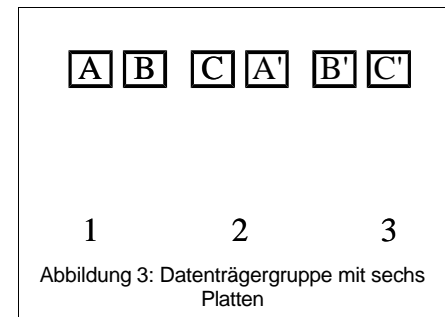


Abbildung 3: Datenträgergruppe mit sechs Platten

1.2.4 Datenträgergruppe mit acht Platten

Eine Datenträgergruppe aus acht Platten muss auf drei Subsysteme verteilt werden, so bleibt das Quorum bei Ausfall eines Subsystems erhalten.

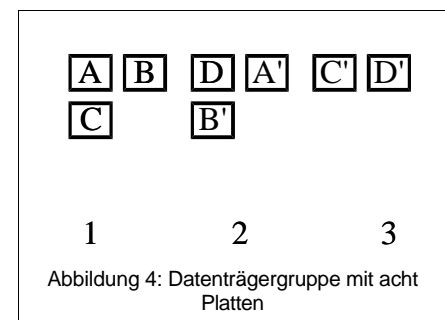


Abbildung 4: Datenträgergruppe mit acht Platten

1.2.5 Die Kombination von Quorum und Spiegelung

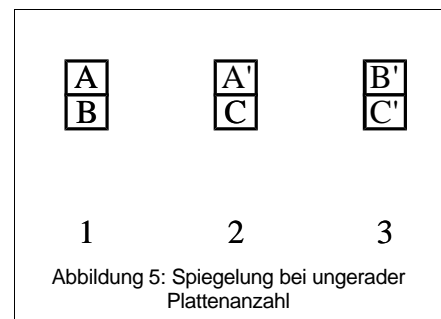
Die folgende Tabelle zeigt die idealen Anordnungen der Platten und der zugehörigen Spiegelplatten, sodass - mit Ausnahme der Lösung mit lediglich zwei Platten - Quorum eingeschaltet bleiben kann:

Platten gesamt	Platten in jedem Subsystem				Quorum
	1	2	3	4	
2	A	A'			aus
4	A	B	A'	B'	ein
6	A, B	C, A'	B', C'		ein
8	A, B, C	D, A', B'	C', D'		ein
10	A, B, C, D	E, A', B'	C', D', E'		ein
12	A, B, C, D	E, F, A', B'	C', D', E', F'		ein
14	A, B, C, D, E	F, G, A', B', C'	D', E', F', G'		ein
16	A, B, C, D, E, F	G, H, A', B', C'	D', E', F', G', H'		ein
18	A, B, C, D, E, F	G, H, I, A', B', C'	D', E', F', G', H', I'		ein
20	A, B, C, D, E, F, G	H, I, J, A', B', C', D'	D', E', F', G', H', I', J'		ein

Tabelle 1: Ideale Plattenanordnungen

1.2.6 Spiegelung bei ungerader Plattenanzahl

Die oben genannten Beispiele basierten auf einer geraden Anzahl von Platten, dabei ist es möglich, alle logischen Datenträger einer Platte auf die exakt gleiche Position der zugeordneten Spiegelplatte zu spiegeln. Es ist ebenso möglich, eine Spiegelung bei eingeschaltetem Quorum auch bei einer ungeraden Anzahl von Platten vorzunehmen. In diesem Fall muss die Platzierung der logischen Datenträger jedoch sorgfältig geplant werden, wie das Beispiel zeigt. Gegeben ist eine Anordnung aus drei logischen Datenträgern, die so auf Platten in anderen Subsystemen gespiegelt sind, dass das Quorum bei Ausfall eines Subsystems erhalten bleibt und gleichzeitig der Zugriff auf die logischen Datenträger weiterhin möglich ist. Dieses Designprinzip kann für jede Datenträgergruppe mit einer ungeraden Anzahl von Platten erweitert werden.



Die Verwendung einer ungeraden Anzahl von Platten (Ausnahme: nur eine einzelne Platte) hat den Vorteil, dass immer nur drei Subsysteme benötigt werden. Nachteilig ist dabei der größere Aufwand für Planung und Administration bei der Anordnung der logischen Datenträger.

1.2.7 Anordnung der Spiegelkopien

Um die größtmögliche Datenverfügbarkeit zu erreichen, müssen die Spiegelkopien in separaten Subsystemen angeordnet werden, und jedes Subsystem, das Spiegelkopien von logischen Datenträgern beherbergt, muss in einem anderen Plattenschrank untergebracht sein. Idealerweise sollten drei Subsysteme in der genannten Anordnung verwendet werden, dann kann auf die Daten auch zugegriffen werden (per Spiegelkopie, und das Quorum bleibt erhalten), wenn die Versorgungsspannung eines Plattenschanks ausfällt.

1.2.8 Plattenübergreifende logische Datenträger

In einer Datenträgergruppe mit einer geraden Anzahl von gespiegelten Platten können sich logische Datenträger auch über mehrere Platten erstrecken. Dies ist möglich, weil im Fall des Ausfalls einer einzelnen Platte (mit einem Teil des logischen Datenträgers) oder eines gesamten Subsystems jederzeit eine gespiegelte Kopie verfügbar ist.

Für Datenträgergruppe mit einer ungeraden Anzahl von Platten kann dieses Verfahren jedoch deutlich komplexer sein - insbesondere, wenn mehr als die Hälfte der Platten in der Datenträgergruppe verwendet werden sollen. In diesem Fall muss die Spiegelung auf der Ebene der physischen Partition erfolgen und es müssen Plattenzuordnungsdateien (Partition Map Files) verwendet werden. In diesen Fällen ist die Verteilung der Partitionen sehr sorgfältig zu planen, damit die Verfügbarkeit der logischen Datenträger auch dann garantiert ist, wenn eine Platte oder ein Subsystem ausfallen sollte. Dies sollte jedoch nur "der allerletzte Ausweg sein, da - neben der komplexen Administration - Auswirkungen auf die Leistung zu erwarten sind.

Bei Verwendung von fünf und mehr Platten kann sich ein logischer Datenträger jedoch durchaus über zwei (oder mehr) Platten erstrecken, ohne dabei Zuordnungsdateien einsetzen zu müssen - solange sie sich im selben Subsystem befinden.

1.3 Wiedereingliederung der Platten

Plattenspiegelung hilft, die Datenintegrität und -verfügbarkeit zu erhalten. Sollten Platten ausfallen oder logische Datenträger veraltet sein, ist jedoch einiger Aufwand erforderlich, um das System wieder in seinen Originalzustand zurückzusetzen.

1.3.1 Ersetzen ausgefallener Platten

SSA-Platten sind Hot-Plug-fähig, sie können also im laufenden Betrieb ausgetauscht werden, ohne dass das System heruntergefahren werden muss. Spiegelkopien von logischen Datenträgern müssen dann aber auf der neu eingesetzten Platte wiederhergestellt werden, um die Daten wieder zu synchronisieren.

Obwohl es äußerst unwahrscheinlich ist, können auch mehrere Platten gleichzeitig ausfallen. In diesem Fall kann das Ersetzen ziemlich zeitaufwendig sein und die Synchronisierung der Daten kann Einfluss auf aktuelle Benutzersitzungen haben, da die Antwortzeiten wegen der ein- und ausgabeintensiven Operationen ansteigen.

1.3.2 Resynchronisierung veralteter Plattenpartitionen

Immer wenn der Zugriff auf eine Platte scheiterte, ohne dass die Platte tatsächlich ausgefallen ist, führt das sehr wahrscheinlich zu einer Veraltung der logischen Datenträger. Die Anzahl der veralteten Partitionen hängt von dem Ausmaß der Ein-/Ausgabeaktivitäten (Lesen und Schreiben von Daten) ab, die während der Nichtverfügbarkeit der Platte weiterhin stattfinden. Wenn der Zugriff auf diese wieder möglich ist, müssen die veralteten logischen Datenträger wieder synchronisiert werden.

Es gibt zwei Möglichkeiten, die Daten auf den Platten zu resynchronisieren. Die verwendete Methode hängt davon ab, wie lange das System im laufenden Betrieb gehalten werden muss .

Die erste Methode - üblicherweise die schnellste, aber die Benutzer können dabei vorübergehend mit ihrer Anwendung nicht weiterarbeiten - besteht darin, die Anwendung herunterzufahren, die Datenträgergruppe zu inaktivieren und anschließend zu reaktivieren. Während der Reaktivierung werden die Daten automatisch resynchronisiert, die dafür erforderliche Zeit hängt von der Anzahl der betroffenen Platten ab sowie davon, wie veraltet die Daten der logischen Datenträger sind. Wenn die Versorgungsspannung eines ganzen Plattenschanks ausgefallen ist, kann diese Resynchronisierung einige Zeit in Anspruch nehmen. Nach erfolgter Resynchronisierung kann die Anwendung wieder gestartet werden.

Die zweite Methode erfordert mehr Zeitaufwand und beinhaltet das Entfernen aller betroffenen Platten aus der Datenträgergruppe, ihrem erneutes Hinzufügen und den erneuten Aufbau der logischen Datenträger und der Daten auf jeder Platte. Wenn Zuordnungsdateien die Anordnung der logischen Datenträger auf jeder Platte beschreiben, kann die Wiederherstellung der logischen Datenträger spürbar beschleunigt werden. Die

Resynchronisation der Daten kann auch hier einige Zeit in Anspruch nehmen, aber die einzige Auswirkung für die Benutzer sind längere Antwortzeiten (die Wiederherstellung kann aber auch erst in den Zeiten geringerer Systemauslastung vorgenommen werden).

2. "Single Points of Failure" (SPoF) bei der SP® Hardware

2.1 Control Workstation

Die Control Workstation selber stellt bereits einen SPoF dar, ist jedoch nicht so kritisch, wenn HACMP-Cluster betroffen sind. Sollte sie ausfallen, ist die Funktion der einzelnen Knoten nicht beeinträchtigt, aber die folgenden Beschränkungen bei der Systemadministration treten auf:

- Die Steuerung der SP® Hardware geht verloren.
- Der Systemdatenrepository ist nicht verfügbar.
- Konfigurationsänderungen sind nicht möglich.
- Softwareinstallationen von der Control Workstation aus sind nicht möglich.
- Im Fall eines Switch-Fehlers ist das Durchführen des Rücksetzens nicht möglich.
- Fehleraufzeichnungen von Alarmen durch die Knoten gehen verloren (die Informationen bleiben jedoch weiterhin auf den einzelnen Knoten gespeichert).
- Administrative Aufgaben unter Verwendung des PSSP sind nicht möglich.

Mit HACWS können diese möglichen Probleme jedoch ohne Benutzereingriff automatisch umgangen werden..

Alternativ kann eine Reservemaschine (Standby-Betrieb) verwendet werden, die mit einem `mksysb`-Systemabbild der ausgefallenen Control Workstation ausgestattet wurde und mit der SP® verbunden wird, sodass administrative und überwachende Aufgaben weiterhin durchgeführt werden können. In dieser Konfiguration sollten die folgenden Daten auf externen Datenträgergruppen abgespeichert sein:

- SP® Verwaltungsdaten
- AIX® Systemabbilder
- PSSP und die dazugehörigen Installationsdaten (Dateisystem `/spdata`)
- NIM-Konfigurationsdateien und weitere benötigte Installationspakete (`installp`-Datensätze)

Dies stellt sicher, dass stets die aktuellen Konfigurationsdaten auf dem Standby-System zur Verfügung stehen. Ein `mksysb`-Systemabbild des Vortags kann diese Aktualität u. U. nicht bereitstellen.

2.2 SP® Frame

Die Versorgungsspannung eines SP® Frames wird durch drei Netzteile (SEPBU) bereitgestellt, und zwar über ein einziges Netzkabel. Dies stellt damit einen SPoF dar - wird es herausgezogen, sind die Knoten wegen des Versorgungsspannungsausfalls außer Betrieb. Um die Hochverfügbarkeit sicherzustellen, sollten die Knoten über Schrankgrenzen hinweg konfiguriert werden, sodass eine automatische Übernahme auf die Knoten im anderen Frame angestoßen wird, wenn das Netzkabel eines Frames ausfallen sollte.

Obwohl auch die Frameüberwachung und die RS-232-Kabel von der Frameüberwachung zu den Knoten und der Control Workstation SPoFs darstellen, bedeutet ein Ausfall hier lediglich eine Einschränkung bei der Systemadministration. Je nach Fehler gilt dies für einzelne oder auch für mehrere Knoten. Der Zugriff der Clients und der normale Cluster-Betrieb ist jedoch dadurch nicht eingeschränkt.

Da jeder Frame mit redundanten Netzteilen ausgestattet ist, liegt bei diesen kein SPoF vor.

2.3 SSA-Plattenschränke

Jeder externe Plattenschrank verfügt über ein einziges Netzkabel und eine einzige interne Stromversorgungseinheit, beide dieser Komponenten sind damit SPoFs. Wären die logischen Datenträger auf den Platten in jedem Cluster schrankübergreifend gespiegelt - jeder verfügt dann über ein eigenes unabhängiges Netzteil - wäre dies für den Cluster in seiner Gesamtheit kein SPoF.

Sollte eine Stromversorgungseinheit ausfallen, oder das Netzkabel defekt sein oder herausgezogen werden, fällt die Versorgungsspannung für alle Subsysteme im betroffenen Schrank aus und alle logischen Datenträger auf jeder Platte veralten. Die Wiedereingliederung all dieser Platten in die zugehörigen Cluster und die Resynchronisierung der Daten aller logischen Datenträger hat Auswirkungen auf den Benutzerzugriff.

Es wird empfohlen, eine zweite Stromversorgungseinheit zu installieren, die von einem separaten Netzteil aus verkabelt und gespeist wird. Dann hat weder der Ausfall eines Netzkabels noch der Ausfall einer Stromversorgungseinheit direkte Auswirkungen auf den Betrieb des Clusters.

Sollte eine der doppelten Stromversorgungseinheiten ausfallen, werden entsprechende Nachrichten in die Fehlerprotokolle aller Knoten geschrieben, die mit den SSA-Platten des Schrank verbunden sind. Voraussetzung dafür ist, dass die Netzteile der SSA-Subsysteme auf die Stromversorgungseinheiten aufgeteilt wurde, wie die im Folgenden beschrieben wird.

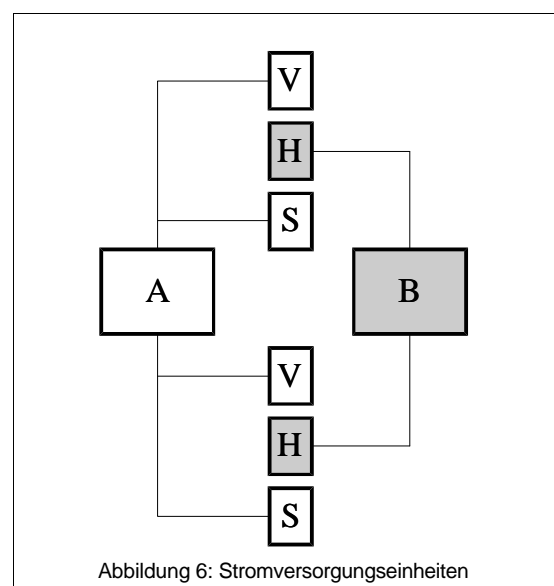
2.4 SSA-Platten-Subsysteme

2.4.1 Netzteile

Jedes SSA-Platten-Subsystem kann mit drei Netzteilen ausgestattet werden. Ein Netzteil versorgt die vorderen acht Festplatten, das zweite die hinteren acht Festplatten und das dritte dient als Reservenetzteil, falls eines der beiden anderen ausfällt (Standby). Das einzelne Netzteil ist kein SPoF, solange alle drei Einheiten installiert sind.

Das Netzkabel, das diesen Einheiten die Versorgungsspannung zuführt, stellt jedoch ein SPoF dar, weil ein Kabel alle drei Einheiten versorgt. Wenn wir den Cluster als Ganzes betrachten, muss das Kabel jedoch nicht unbedingt einen SPoF darstellen, wenn die Spiegelplatten in einem anderen Subsystem (und einem getrennten Plattenschrank) noch verfügbar sind. Sollte das Netzkabel ausfallen, oder die Stromversorgungseinheit eines Schrank ausfallen, gehen veralten alle Platten dieses Subsystems. Die danach erforderliche Wiedereingliederung in den Cluster und die Resynchronisierung der Daten, nachdem die Versorgung wieder sichergestellt ist, können Einfluss auf den Benutzerzugriff haben.

Neben der Empfehlung, jeden Plattenschrank mit einer zweiten Stromversorgungseinheit auszustatten, ist es ratsam, das einzelne Netzkabel zu den drei Netzteilen des Subsystems durch zwei Kabel zu ersetzen. Jedes beginnt bei einer einzelnen Stromversorgungseinheit, eines endet bei zwei Subsystemnetzteilen, das andere beim verbleibenden einzelnen Netzteil. Abbildung 6 zeigt eine mögliche Verkabelung. Dabei stellen A und B die beiden Stromversorgungseinheiten des Schrank dar, V ist das Subsystemnetzteil der vorderen acht Platten, H das Subsystemnetzteil der hinteren acht Platten und S ist das Reservenetzteil (Standby-Netzteil). Fällt A aus, dann sind nur die acht hinteren Platten jedes Subsystems weiterhin verfügbar. Bei Ausfall von B übernimmt das Reservenetzteil die Versorgung der acht hinteren Platten und das Subsystem bleibt verfügbar. Die Abbildung zeigt die Verwendung von nur zwei Subsystemen pro Schrank, aber das Prinzip kann auf alle Subsysteme im Schrank ausgeweitet werden.



2.4.2 Signal-/Umgehungskarten des Subsystems

Die Verkabelung eines Adapters verbindet den "J"-Stecker des Subsystems mit den internen SSA-Platten. Die "J"-Stecker sind paarig an Signal- oder Umgehungskarten angeschlossen, die an sich keinen SPoF darstellen (unter der Voraussetzung, dass ein Plattenzugriff weiterhin über alternative Schleifen über eine andere Umgehungskarte erfolgen kann oder Spiegelkopien über eine zweiten Schleife erreichbar sind).

Die Umgehungskarten in einem 7133-600 SSA-Subsystem enthalten "J"-Stecker, die nummeriert sind und so die erste Platte in einer Vierer-Anordnung (Bank) anzeigen, mit der sie verbunden sind. Die Paarbildung erfolgt dann als 1 und 16, 4 und 5, 8 und 9, 12 und 13. Die Umgehungskarten ermöglichen darüber hinaus zwei Betriebsarten.

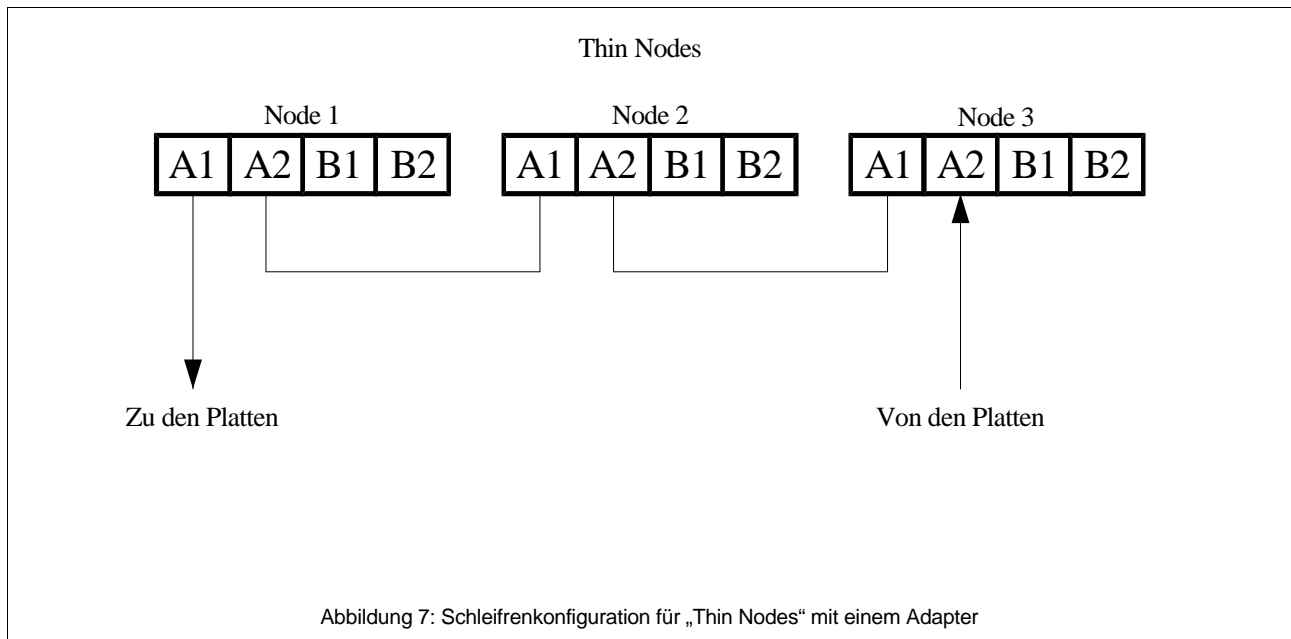
- Betriebsart "Bypass": Fällt ein Knoten aus, der mit beiden "J"-Steckern der Umgehungskarte verbunden ist, schließt die Karte die Schleife und verbindet so zwei Bänke zu je vier Platten. Diese beiden Bänke können Pseudoplatten enthalten, um die Lücken der unbenutzten Platteneinschübe zu füllen und damit eine Schleife ohne Unterbrechung sicherzustellen. Wenn jedoch mehr als drei Pseudoplatte hintereinander angeschlossen sind, wird die Schleife unterbrochen und die Platten sind nicht mehr verfügbar. Dies ist eine besondere Möglichkeit der "Bypass"-Betriebsart.
- Betriebsart "Forced Inline": Bei Ausfall eines Knotens, der an die "J"-Stecker einer einzigen "Bypass"-Karte angeschlossen ist, wird die Schleife unterbrochen (und bleibt unterbrochen), aber die Platten sind trotzdem weiterhin über die alternative Schleife erreichbar. Da mehrere Knoten mit Subsystemen in jedem Cluster verbunden sind, sollten alle Umgehungskarten jedes Subsystems auf die Betriebsart "Forced Inline" gesetzt sein.

2.5 SSA-Platten-Schleifen

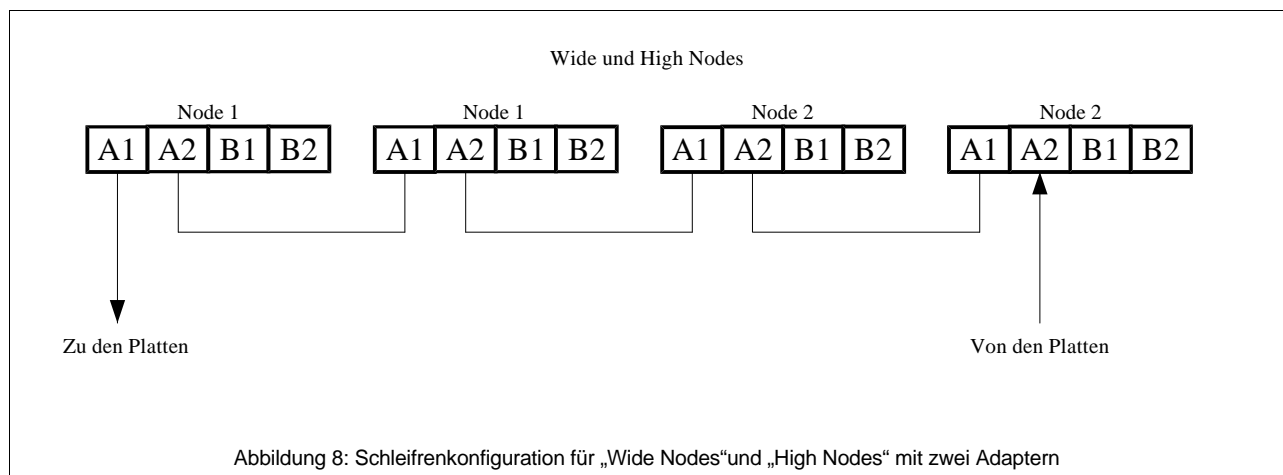
Alle externen Platten müssen mit allen Knoten verbunden werden, die im Fall einer Übernahme im Zugriff bleiben sollen. Sie müssen so verkabelt sein, dass im Fall eines Plattenausfalls oder SSA-Kabeldefekts - die Schleife ist dann unterbrochen - die Schleife kein SPoF darstellt und aktive Platten weiterhin über den Weg alternativer Schleifen erreichbar sind (oder zumindest die Spiegelkopien).

In Standardknoten (Thin Nodes) ist es u. U. nicht möglich, mehr als einen SSA-Adapter zu installieren. Fällt dieser eine Adapter aus, sind die Daten der Platte(n) nicht weiter verfügbar und es muss eine Übernahme durch einen anderen Knoten angestoßen werden. Zur Verbindung aller Knoten mit drei Subsystemen sind nur zwei Schleifen möglich, jede von ihnen muss so konfiguriert werden, dass eine hohe Verfügbarkeit garantiert ist und die volle Bandbreite der SSA-Adapter genutzt wird.

Eine Schleifen-Konfiguration für Standardknoten mit einem Adapter sollte wie folgt aussehen: Beide Schleifen, A und B, sind gleich, jede Schleife ist mit unterschiedlichen gespiegelten Platten verbunden. Cluster aus zwei bzw. drei Knoten werden ähnlich verkabelt.



Erweiterte Knoten (Wide Nodes und High Nodes) werden üblicherweise mit zwei SSA-Adaptoren ausgerüstet sein. Diese sind so zu verkabeln, dass bei Ausfall einer Platte oder eines SSA-Adapters alle Platten in der Schleife (bis auf die ausgefallene) über alternative Schleifenpfade erreichbar sind. Bei zwei SSA-Adaptoren sollte die Verkabelung der Erweiterten Knoten wie folgt vorgenommen werden: Beide Schleifen, A und B, sind gleich, aber sie sind mit verschiedenen Platten verbunden. Auch hier ist die Verkabelung bei Clustern aus zwei bzw. drei Knoten ähnlich durchzuführen.



3. SPoFs in SP® Netzwerken

3.1 Internes Netzwerk

Das interne Netzwerk der SP® stellt einen SPoF dar, der Ausfall eines einzigen Segments des Netzwerks bedeutet einen globalen Netzwerkausfall. Davon müssen die Anwendungen nicht unbedingt betroffen sein, aber eine Systemadministration ist nicht möglich. Außerdem können die Überwachungsdaemonen, die auf allen Knoten laufen und deren Systemzustand überwachen, nicht miteinander kommunizieren und den aktuellen Systemzustand ermitteln. HACMP wird den Fehler zwar bemerken, aber darauf nicht reagieren.

Das Netzwerk der SP® kann nicht redundant ausgelegt werden, aber die Auswirkungen eines Fehlers in diesem Netzwerk können minimiert werden, indem jeder Knoten direkt mit einem Router oder Switch verbunden wird. Ist jeder Knoten mit einer solchen Einheit direkt verbunden (und nicht alle Knoten über einen gemeinsamen Anschluss), bleiben Netzwerkfehler auf den einzelnen Knoten beschränkt, wenn ein einzelnes Netzkabel ausfallen sollte. In diesem Fall sollte allerdings auch ein Reserve-Switch bzw. -Router vorgesehen werden (Standby-Betrieb), damit nun nicht der Router/Switch selber ein SPoF wird.

3.2 Client-Netzwerke

Das Client-Netzwerk ermöglicht den Zugriff der Workstations auf Anwendungen, die auf den einzelnen Server-Knoten installiert sind. Dies ist in den meisten Fällen ein SPoF, da Reservernetzwerke für die Clients wegen der damit verbundenen Zusatzkosten relativ selten anzutreffen sind. Um die Auswirkungen eines Ausfalls zu minimieren, sollte jedes Client-Netzwerk so ausgelegt sein, dass es in Segmente - entweder durch Router, intelligente Switches, Hubs oder andere Mechanismen - aufgeteilt werden kann. Dadurch können mehrere Leitwege durch das Netzwerk zu den Cluster-Knoten bereitgestellt werden.

Wenn der mit dem Knoten direkt verbundene Netzwerkteil ausfällt, kann kein Client auf die Dienste des Knotens und seine Anwendungen zugreifen, sodass in diesem Fall eine Übernahme durch einen anderen Knoten angestoßen werden muss.

Befindet sich der übernehmende Knoten im gleichen Netzwerksegment, erkennt HACMP einen globalen Netzwerkfehler - eine Übernahme findet dann nicht statt. Um dies zu verhindern, sollten alle Knoten eines Clusters mit unterschiedlichen Netzwerksegmenten verbunden sein.

Fällt ein Netzwerksegment aus, das nicht unmittelbar mit einem Knoten verbunden ist, ist ein Zugriff der Clients auf den Knoten weiterhin möglich - vorausgesetzt, im Netzwerk steht ein alternativer Leitweg um das ausgefallene Segment herum bereit. Hier sollten die Clients über möglichst viele Einzelsegmente verteilt werden, damit die Auswirkungen eines einzelnen Segmentausfalls reduziert werden.

Ebenso ist bei der Auslegung der Client-Netzwerke zu berücksichtigen, welche Auswirkungen der Ausfall von primären bzw. sekundären Servern hat (wenn DNS oder NIS eingesetzt wird). DNS- und NIS-Primärserver und -Sekundärserver müssen auf unterschiedlichen (getrennten) Segmenten angeordnet werden, damit immer mindestens einer von ihnen bei einem Teilausfall des Netzwerks verfügbar ist.

3.3 Switch-Netzwerke

Da die SP® nur über ein Switch-Netzwerk verfügt, stellt der Switch selber einen SPoF dar. Der neuste Stand der Switch-Technologie verfügt über eingebaute Redundanzen und Wiederherstellbarkeit - deshalb sind Fehler hier eher lokaler als globaler Natur. Ein Totalausfall des Switches ist damit ein seltener Fall. Ist der Datenaustausch über den Switch als kritisch anzusehen, sollte allerdings über den Einsatz eines Reservernetzwerks mit hoher Geschwindigkeit, wie etwa FDDI, nachgedacht werden, das bei Totalausfall des Switches zum Einsatz kommt.

Um festzustellen, wie der Switch ausfallen könnte und um die Auswirkungen zu minimieren, müssen die einzelnen Netzwerkkomponenten betrachtet werden.

Das SP® Switch-Netzwerk besteht aus dem Switch-Board in jedem Frame, das diverse elektronische Bausteine enthält, Switch-Adaptern in den Knoten und den internen und externen Kabeln, die die einzelnen Komponenten miteinander verbinden. Die Switch-Boards erhalten ihre Stromversorgung von den einzelnen Netzteilen des Schanks, die über eine eingebaute Redundanz verfügen. Da Ausfälle des Switches in der Regel nur lokale Auswirkungen haben, führt ein Ausfall eines Switch-Kabels nur zum Ausfall eines Stranges, sodass eine Übernahme auf einen anderen Knoten erfolgen muss.

Jede SP® verfügt über einen Primär- und einen Reserveknoten für den sogenannten E-Primärknoten, das ist derjenige, der das gesamte Switch-Netzwerk initialisiert und wiederherstellt, nachdem Fehler erkannt wurden. Dieser Knoten stellt damit kein SPoF dar, aber es wird empfohlen, den Primär- und den Sekundärknoten in verschiedenen Schränken zu betreiben. Wird dies nicht berücksichtigt, ist bei einem Stromausfall eines Schanks mit einem globalen Switch-Netzwerkausfall zu rechnen.

In jedem Schrank befinden sich vier Switch-Bausteine, die die Kommunikation mit den Knoten abwickeln,

jeder Baustein bedient dabei vier Knoten. Fällt einer dieser Bausteine aus, ist die Verbindung zu den angeschlossenen vier Knoten unterbrochen. Hochverfügbare Plattensubsysteme müssen mindestens über zwei Kommunikationspfade verfügen, um gegen Knoten- und Adapterausfälle geschützt zu sein. Sind die mit dem Plattensubsystem verbundenen Knoten am selben Switch-Baustein angeschlossen, besteht die Chance, dass die Übertragung der Plattendaten über den Switch ausfällt. Platten-Subsystem sind gegen einen Ausfall eines einzelnen Switch-Bausteins geschützt, wenn die Platten mit Knoten in verschiedenen Frames verbunden werden (verschiedene Switch-Bausteine).

Der Switch verfügt über eine eingebaute Uhr, damit ein korrekter Datenempfang sichergestellt ist und alle Switch-Systeme mit derselben Zeitinformation betrieben werden können. Es befindet sich ein Haupt-Switch-Board in der SP®, das die korrekte Zeit an ggf. weitere untergeordnete Switch-Boards übermittelt. Üblicherweise befindet sich das Haupt-Switch-Board im ersten Frame. Auf dem Haupt-Switch-Board befindet sich der Hauptbaustein, der die Zeit an die anderen Bausteine weitergibt, welche wiederum die Verbindung zu den Knoten und weiteren Switch-Boards herstellen.

Der Switch ist mit einem Reserve-Hauptbaustein ausgestattet, um eine hohe Verfügbarkeit sicherzustellen. Fällt der Hauptbaustein aus, würde die SP® ohne diesen Reservebaustein die Switch-Zeitsynchronisierung verlieren, dies würde einen globalen Switch-Ausfall bedeuten. Die Wiederherstellung muss manuell erfolgen, die Konfiguration dieser Reservefunktion sollte auf einem untergeordneten Switch-Board in einem anderen Frame erfolgen, um diesen potenziellen SPoF zu vermeiden.

4. Marken

- Das  Logo ist eine in Deutschland eingetragene Marke von Mannherz EDV-Dienstleistungen
- AIX ist eine eingetragene Marke von IBM Corp. in den Vereinigten Staaten von Amerika und/oder anderen Ländern
- IBM ist eine eingetragene Marke von IBM Corp. in den Vereinigten Staaten von Amerika und/oder anderen Ländern
- SP ist eine eingetragene Marke von IBM Corp. In den Vereinigten Staaten von Amerika und/oder anderen Ländern

Alle anderen auf diesen Seiten erwähnten Produkt- bzw. Firmennamen sind Marken ihrer jeweiligen Eigentümer.