Die Grid-Architektur mit DataCore SANsymphony-V

- Mehr als nur ein Cluster-System -

INHALT

Vorwort 1

Grundlagen SANsymphony-V 1

Grundlagen Cluster 2

Vergleich: Grid-Architektur vs. Cluster-Lösung 2

SANsymphony-V: In Verbindung mit Applikations-Clustern..... 3

SANsymphony-V: Intergrid Kommunikations Design..... 3

SANsymphony-V: Ausfallszenarien und Verhaltensweisen beim Verlust von Intergrid Kommunikations Verbindungen..... 4

Erweiterte Redundanz-Szenarien für SANsymphony-V: Intergrid Kommunikations-Verbindungen durch einen dritten Standort..... 5

Zusammenfassung 6

Vorwort

Hochverfügbarkeit und Ausfallsicherheit von Servern, Applikation sowie Speichersystemen gehören nach wie vor zu den wichtigsten Anforderungen in modernen IT-Infrastrukturen. Für diese Aufgabenstellung werden typischerweise Cluster-Lösungen empfohlen. Neben dem Cluster existiert aber auch ein zweiter Lösungsansatz: Das Grid - ein anderes Konzept mit gleicher Zielsetzung. Auch wenn die beiden Begriffe „Cluster" und „Grid" im Markt häufig als Synonym verwendet werden, ergeben sich bei einem Blick „unter die Haube", kleine aber für den Anwender zum Teil doch sehr relevante Unterschiede. Dieses White Paper fasst am Beispiel von DataCore SANsymphony-V eine Reihe dieser wichtigen Details zusammen und dient dadurch Projektverantwortlichen als konzeptionelle Entscheidungshilfe.

Grundlagen SANsymphony-V

DataCore SANsymphony-V repräsentiert eine "aktiv-aktiv" Grid-Architektur und ist dadurch kein Cluster-System. Die Knoten präsentieren gespiegelte Disks als „aktiv-aktiv" - Speicher-Laufwerke. Dies bedeutet, dass die LUN[1] des Applikationsservers (vDisk[2] in SANsymphony-V) nicht nur auf einem SANsymphony-V Server bereitgestellt wird, sondern über beide SANsymphony-V Server gleichzeitig angesprochen werden kann (read und write). Eine von SANsymphony-V gespiegelte Disk ist vollkommen redundant verfügbar, sowohl auf Controller-, Pfad- als auch auf Datenspeicherungsebene. Dieser Grid-Ansatz ist im Gegensatz zu einer Cluster-Lösung grundsätzlich nicht von einem Split-Brain betroffen und benötigt daher auch keinen Tie-Breaker[3] , d.h. keine Quorum-Instanz[4] und keine Witness[5] . Im Folgenden wird diesbezüglich nur noch der Begriff „Witness" verwendet.

Definition Split Brain:
http://de.wikipedia.org/wiki/Split_Brain_%28Informatik%29

1 LUN: Logical Unit Number. Bezeichnet in diesem Kontext eine Festplatte, die aus einem zentralen Storage zur Verfügung gestellt wird.
2 vDisk: Virtual Disk. Die aus SANsymphony-V an den Applikation-Server bereitgestellten Speicherbereiche werden auf SANsymphony-V Ebene als vDisk bezeichnet.
3 Tie Breaker (deutsch: Knoten-Löser) verhindert in einem Cluster die Split-Brain-Situation, in dem er als dritte Instanz entscheidet welche Instanz aktiv sein soll.
4 Siehe Tie-Breaker. Identische Funktion, anderer Name
5 Witness (deutsch Zeuge): Siehe Tie-Breaker. Identische Funktion, anderer Name. Alternative Bezeichnungen sind: Quorum, Failover-Manager

Grundlagen Cluster

Verglichen mit einer Grid-Architektur, steht in einer Cluster-Umgebung eine Ressource üblicherweise nur jeweils auf einem der Knoten aktiv zur Verfügung. Der oder die verbleibenden Knoten haben die Aufgabe diesen Service neu zu starten, sobald der derzeit aktive Knoten ein Problem hat. Typischerweise kommt bei Cluster-Systemen ein „Voting-Algorithmus"[6] zum Einsatz, mit dem der Cluster entscheidet, ob eine Ressource aktiv geschaltet wird oder ob sich der Cluster in einem Zustand der Partitionierung befindet (auch Split-Brain genannt). Die Ressource wird nur dann aktiv genommen, wenn der Cluster eine Mehrheit der Stimmen hat. Aus diesem Grund muss eine ungerade Anzahl von Stimmen im Cluster zur Verfügung stehen, um eine Mehrheit zu erreichen bzw. zu garantieren.

Bei einem typischen Zwei-Raum-Design befindet man sich bei einer Cluster-Architektur in folgendem Dilemma:

- Beim Ausfall der Seite mit der Mehrheit der Cluster-Knoten kann kein Failover der Dienste stattfinden, da die verbleibende Seite keine Mehrheit im Cluster mehr erreicht. Aus diesem Grund wird die Anzahl der Knoten in einem verteilt auf zwei Standorten operierenden (so genannten stretched oder verteiltem) Cluster, typischerweise gleichmäßig aufgeteilt.

Damit ergibt sich jedoch eine mögliche Patt-Situation in mehrerlei Hinsicht:

- Es besteht keine Mehrheit mehr für die Cluster-Knoten in einer Site.
- Der Ausfall einer Site ist von der kompletten Trennung der Verbindung zwischen den Sites nicht zu unterscheiden - dieses Szenario wird dann als "Split Brain" bezeichnet.

Um diese Patt-Situation aufzulösen, kommt bei Cluster-Systemen ein Tie-Breaker (auch Quorum, oder Witness genannt) zum Einsatz. Dabei handelt es sich um eine Ressource, die von einem der Cluster-Knoten gehalten wird. Dieser Knoten besitzt demnach zwei Stimmen innerhalb des Clusters und kann somit die Mehrheit der Site in der er steht erreichen - d.h. die Überhang-Stimme entscheidet, welche der zwei Seiten aktiv bleibt und/oder ob ein Failover stattfindet.

Sofern die Witness im primären oder sekundären Standort untergebracht wurde, ist dies aber komplett nutzlos. Beim Ausfall des gewählten Standortes kann auch der Redundanz-Standort nicht online gehen, da er die Mehrheit im Cluster nicht erreicht. Daher empfehlen die meisten Hersteller für die Platzierung einer Witness eine dritte Lokation. Bei manchen Lösungen wird diese sogar vorausgesetzt. Diese dritte Lokation muss von der primären und sekundären Lokation aus erreichbar sein. Ohne diese dritte Site kann kein automatischer Site-Failover stattfinden.

Vergleich: Grid-Architektur vs. Cluster-Lösung

Normalerweise sind auch sogenannte "aktiv-aktiv"-Cluster pro Ressource, Service oder Dienst ein „aktiv-passiv"-System, denn auch wenn alle Knoten individuelle und voneinander unabhängige Dienste zur Verfügung stellen können, so ist ein Dienst trotzdem nur über einen Knoten ansprechbar. Für diesen (aktiven) Knoten stellen die anderen Cluster-Knoten einen Standby-Knoten dar. Sie werden erst dann aktiv, wenn dies nötig werden sollte. Es gibt Cluster, die aus Sicht der Applikation als „echte aktiv-aktiv"-Cluster agieren. Hierzu gehört z.B. der Microsoft Scale Out File Service (SOFS). Doch technisch gesehen ist hier faktisch ebenfalls nur ein Knoten für zentrale Aufgaben zuständig, in diesem Fall der sogenannte Koordinator-Knoten. Die Rolle des Koordinators muss bei einem Ausfall dieses Knotens ebenfalls geschwenkt werden, was auch diesen Cluster aus Sicht der eigentlichen Cluster-Ressource zu einem „aktiv-passiv"-System macht. Unterstützende Technologien lassen hierbei den eigentlich „aktiv-passiv" Operationsmodus nach außen hin als „aktiv-aktiv" erscheinen. In dem Beispiel des SOFS ist hierfür das SMB3 Protokoll zuständig, welche bei einem Failover der Koordinator-Funktion für eine reibungslose Kommunikation sorgt.

DataCore SANsymphony-V arbeitet als Grid. Grid-Knoten sind der Architektur nach voneinander unabhängig. Da im Grid keine zentrale Instanz existiert und alle Knoten für sich autonom agieren, kann in einem SANsymphony-V Grid kein "Split Brain" entstehen. Für eine gespiegelte vDisk agiert SANsymphony-V als „aktiv-aktiv"-System und steht damit im Gegensatz zu dem oben beschriebenen „aktiv-passiv"-Verfahren jederzeit zur Verfügung.

6 Auch als Cluster-Votes bezeichnet. Je nach Implementierung erhält jeder Teilnehmer im Cluster eine Stimme. Die Seite mit der Mehrheit der Stimmen „überlebt" bzw. ist verantwortlich für die Erbringung der Dienste.

Bei einem Cluster kann die Situation aufkommen, dass ein Cluster-Knoten die falsche Entscheidung trifft, einen Dienst online zu nehmen. Meistens passiert dies, wenn der Cluster-Knoten der Meinung ist, er sei der letzte aktive Cluster-Knoten eines Clusters. Die Witness-Instanz hilft dem Cluster-Knoten zu entscheiden, ob dies tatsächlich der Fall ist.

Entsprechend der Cluster-Definition verhalten sich Cluster-Knoten im Normalfall folgendermaßen:

1. Die Cluster-Knoten prüfen auf ihre Partner-Knoten.
2. Sind diese nicht erreichbar, dann wird die Witness geprüft.
3. Ist die Witness ebenfalls nicht erreichbar, muss der Knoten abgeschottet sein und er wird alle Dienste beenden, so dass diese von Partner-Knoten online gebracht werden können.
4. Ist die Witness jedoch erreichbar, muss der Knoten die Annahme treffen der letzte verbleibende Knoten zu sein. In dem Fall wird der Cluster-Knoten alle Dienste online bringen bzw. starten.

Innerhalb eines SANsymphony-V Grids muss nichts "aktiv geschaltet werden", da beide Systeme eines Spiegels zum gleichen Zeitpunkt aktiv sind. Es ist eine echte „aktiv-aktiv" - Lösung.

Das bedeutet auch, dass keine Instanz[7] benötigt wird um die Entscheidung zu treffen, ob die Ressource nun online geschaltet wird soll oder nicht. Daraus folgt, dass das System auch keine falsche Entscheidung treffen kann. Alle Instanzen haben sich in einem aktiven Zustand befunden.

SANsymphony-V: In Verbindung mit Applikations-Clustern

In der IT-Architektur stehen oberhalb des SANsymphony-V Grids häufig Cluster-Systeme. Diese benötigen ihre Witness um im Fehlerfall eine korrekte Entscheidung zu treffen. SANsymphony-V als darunter liegendes Storage-Grid, wird Dank der Grid-Architektur auch bei einem vollständigen Kommunikationsverlust zwischen den beteiligten DataCore Servern weiterhin die LUNs präsentieren. Dabei toleriert SANsymphony-V auch vom darüber liegenden Applikations-Cluster eventuell falsch getroffene Entscheidungen.

Falls ein weiterer Cluster-Knoten fälschlicherweise den Dienst startet, obwohl der Dienst noch auf einem anderen Knoten in der anderen Site aktiv ist, schützt SANsymphony-V die Applikation. Sobald dieser Zustand erkannt wird, werden die zugehörigen Teile eines Spiegels auf "double inaccessible[8]" geschaltet. Damit wird eine Dateninkonsistenz bei einem Wieder-Aufspiegeln verhindert.

In diesem Zustand kann der Nutzer entscheiden, welche der beiden Seiten online genommen werden sollen. Es besteht sogar die Möglichkeit, beide Seiten unabhängig voneinander online zu bringen. Hierfür wird der Spiegel aufgetrennt und in separate vDisks verwandelt, die an die Applikationsserver zur Prüfung herangereicht werden können. In diesem Fall steht auch der DataCore Support mit erweiterten Analyse-Möglichkeiten zur Verfügung.

SANsymphony-V: Intergrid[9] Kommunikations Design

Tritt ein vollständiger Kommunikations-Verlust zwischen den beteiligten SANsymphony-V Knoten ein, impliziert dies üblicherweise eine mangelhafte / fehlerhafte Umsetzung des Hochverfügbarkeits-Designs auf SANsymphony-V Ebene.

DataCore empfiehlt redundant geführte Spiegel-Strecken, um eine Unterbrechung der „in-band" Kommunikation zu verhindern. Diese Spiegel-Strecken werden von der SANsymphony-V Block-Virtualisierung für den eigentlichen Datenabgleich verwendet.

Des Weiteren muss eine dritte, eine so genannte „out-of-band"-Verbindung, existieren. Diese muss von den beiden erstgenannten Verbindungen vollkommen unabhängig sein. Da es sich bei der „out-of-band" Verbindung um eine Standard-TCP-Verbindung mit wenig Datenverkehr für Konfigurations-Daten handelt, sind dafür auch schwache Leitungen oder geroutete Netzwerke ausreichend. Grundsätzlich wird jedoch eine Layer-2 Kopplung für die „out-of-band" Verbindung empfohlen, um mögliche Beeinflussung durch zusätzliche Komponenten auszuschließen.

7 Bedeutet: Kein Tie-Breaker, kein Witness, kein Quorum, kein Failover-Manager
8 Nicht zugreifbar
9 Intergrid bezeichnet Verbindungen innerhalb des Grids.

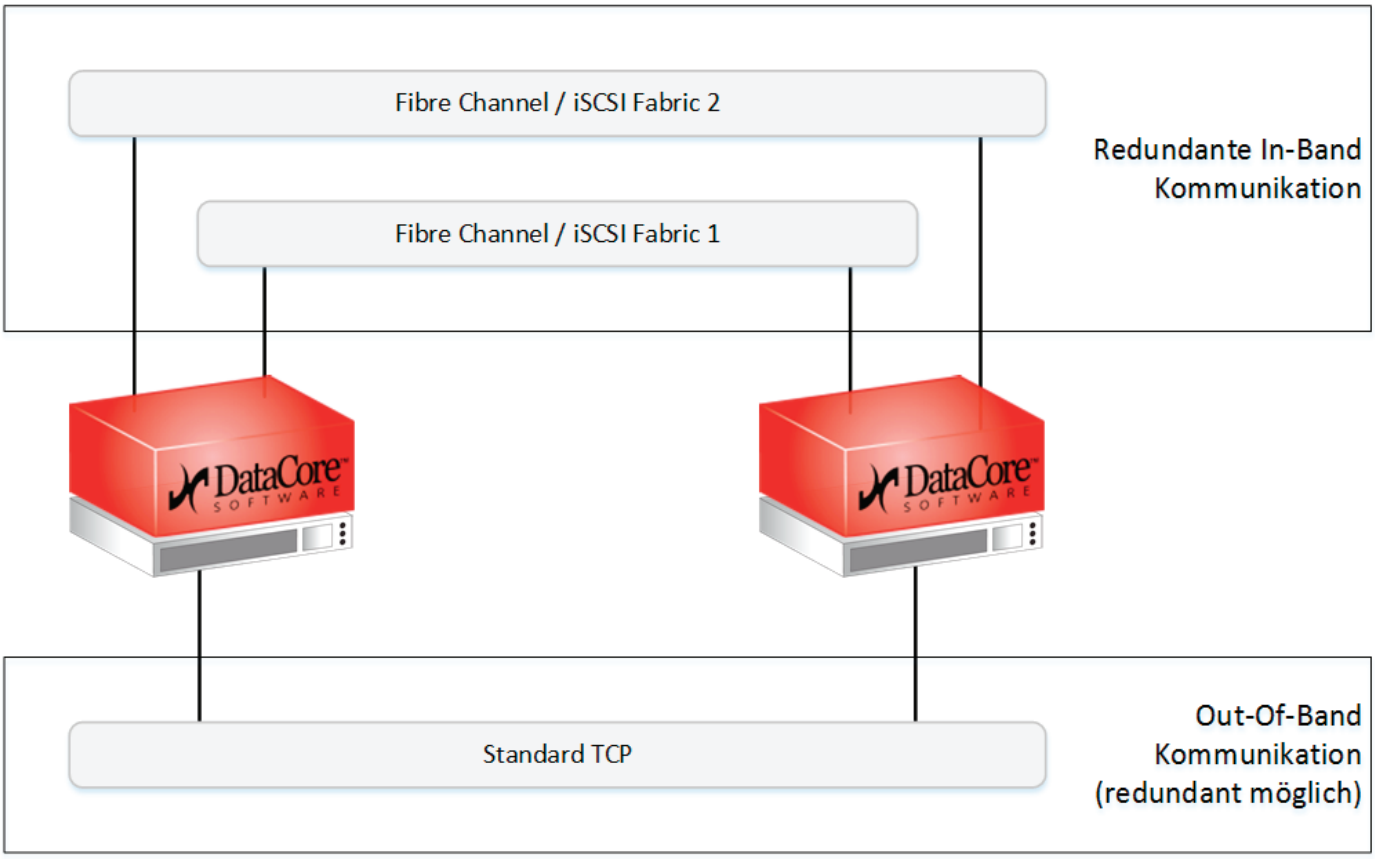

Abbildung 1: SANsymphony-V besitzt drei voneinander unabhängige Kommunikationsverbindungen

Sofern alle drei Verbindungen über unterschiedliche Leitungswege laufen, ist es mit diesem Design möglich, eine komplette Verbindungs-Trennung zwischen den SANsymphony-V Knoten zu verhindern.

SANsymphony-V: Ausfallszenarien und Verhaltensweisen beim Verlust von Intergrid Kommunikations-Verbindungen

Sollte es dennoch zu Ausfällen kommen, reagiert das System wie folgt:

- Bei dem Ausfall einer "in-band" Kommunikations-Leitung, wird transparent auf die verbleibende Redundanz-Leitung geschwenkt. Sind die Bandbreiten ausreichend, werden der angeschlossene Server und die darauf laufende Applikation keinerlei Veränderung bemerken. Die Festplatten sind nach wie vor voll redundant zugreifbar.
- Falls die Redundanz-Leitung der „in-band" Kommunikation ebenfalls/zeitgleich ausfällt, kann SANsymphony-V immer noch über die „out-of-band" Kommunikations-Strecke eigenständig erkennen, welche Seite die aktuelle Seite eines Spiegels ist. In dem Fall wird die nicht mehr aktuelle Seite der gespiegelten Festplatte auf "inaccessible" geschaltet. Zu diesem Zeitpunkt befindet sich der Spiegel in dem Zustand „out of sync".
- Falls nun auch die „out-of-band" Verbindung getrennt wird, bleiben die SANsymphony-V Knoten für sich online und präsentieren weiterhin ihre zu dem Zeitpunkt aktiv gewesenen vDisks.

Wenn nun der Link zwischen den SANsymphony-V Servern wiederhergestellt wird sowie der darüber liegende Cluster fälschlicherweise auf beiden Seiten die Dienste online genommen und damit auf beide Seiten des Storages geschrieben hat, schaltet das System auf "double inaccessible ". SANsymphony-V wird eigenständig versuchen, durch eine Block-Analyse diesen Zustand aufzulösen. Typischerweise obliegt in diesem Falle die Entscheidung jedoch dem Benutzer. Dabei unterstützt der DataCore Support. Es empfiehlt sich in diesem Fall einen „Severity 1 Incident [10]" zu eröffnen.

10 Höchste Dringlichkeits-Stufe

Erweiterte Redundanz-Szenarien für SANsymphony-V: Intergrid Kommunikations-Verbindungen durch einen dritten Standort

Sofern eine dritte Site existiert, kann DataCore SANsymphony-V diese für eine zusätzlich redundante Wegeführung nutzen. Es wird aber keine aktive Instanz in dieser dritten Lokation benötigt. Technisch gesehen ergibt ein dritter Standort für SANsymphony-V Vorteile: Wenn alle Sites miteinander verbunden sind, ergibt dies ein Dreieck (oder auch ein Ring).

Innerhalb des Dreiecks kann SANsymphony-V sowohl über den direkten Weg miteinander kommunizieren als auch über den indirekten Weg. Dies ergibt eine zusätzliche Ebene der Redundanz.

- Dabei übernimmt im Fibre Channel (FC) Umfeld das interne Fabric[11] -Routing das Umschalten auf die Redundanz-Strecke. Im Standard-Betrieb wird von der Fabric der direkte Weg bevorzugt (Berechnung über Hops / Bandbreiten).
- Bei Ethernet-Verbindungen kann das Umschalten per Spanning Tree[12] , Rapid Spanning Tree[13] oder Multi-Instance Spanning Tree[14] erfolgen. Gegebenenfalls kann die Redundanz auch über Stacking-Methoden hergestellt werden.

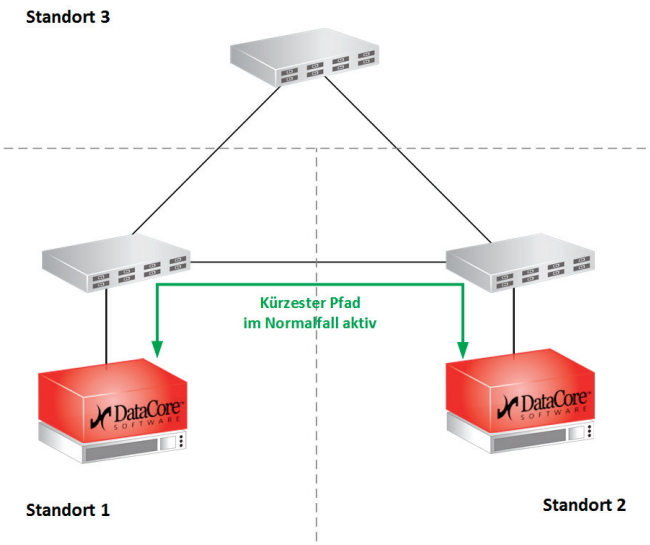

Abbildung 2: Normalbetrieb

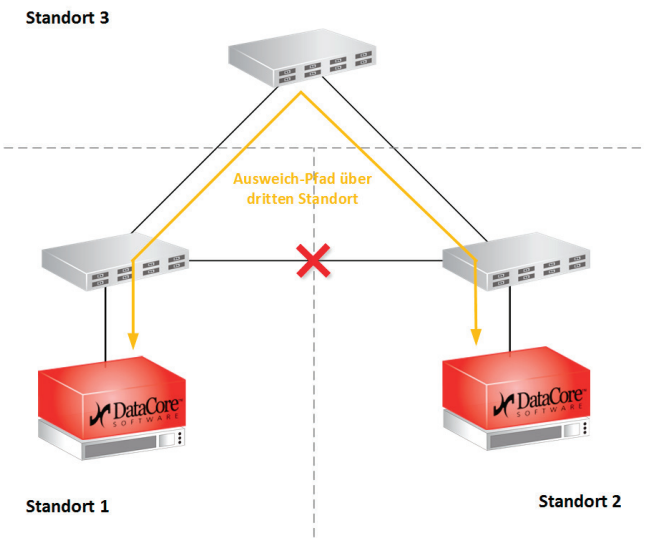

Abbildung 3: Bei Ausfall der direkt-Verbindung wird über dritten Standort umgeleitet

Die beiden Abbildungen zeigen lediglich eine von drei Verbindungen, die über den dritten Standort geführt werden. Ist keine dritte Site vorhanden, ist unbedingt auf die redundante Weg-Führung zwischen den Standorten zu achten.

11 Eine Fabric bezeichnet den Zusammenschluss / Verbund von mehreren FC-Switches zu einer logischen Instanz. Für eine redundante „in-band" Kommunikation sind demnach redundante Fabrics nötig.

12 STP: Spanning Tree Protocol – Ein Protokoll um Netzwerk-Kurzschlüsse ("Loops") zu verhindern. Hierzu wird einer der Links in dem Switch-Verbund geblockt. Anhand des STP Protokolls werden Veränderungen in der Topologie erkannt und etwaige, derzeit geblockte Pfade wieder für Kommunikation frei gegeben.

13 RSTP: Rapid Spanning Tree Protocol. Erweiterung des STP Protokolls. Erlaubt eine schnellere Erkennung von Topologie-Änderungen.

14 MSTP: Multiple Instance Spanning Protocol. Erweiterung des STP-Protokolls. Erlaubt multiple Instanzen des Spanning-Tree Protokolls innerhalb eines Switches / Switch-Verbundes um komplexere Topologien aufbauen zu können.

Zusammenfassung

DataCore SANsymphony-V ist eine Grid-Architektur und kein Cluster-System. Durch diese Grid-Architektur agiert das System als echte „aktiv-aktiv“ Lösung und somit wird auch kein Witness benötigt.

Auf eine redundante Wege-Führung von Kabeln zwischen den Rechenzentren in einer verteilten SANsymphony-V Installation ist zwingend zu achten, um die maximale Redundanz der Lösung zu erreichen.

Eine dritte Site wird nicht benötigt, kann aber die Möglichkeit für zusätzliche Redundanzen auf Ebene der „in-band“ und „out-of-band“ Kommunikation ermöglichen.

Durch die Option einer dritten Site ergibt sich der Vorteil, dass für jede Installation individuell entschieden werden kann, ob die „in-band“, „out-of-band“ oder sogar beide Kommunikations-Strecken auch über diese verfügbar sein sollen.

Somit können die Redundanzen der Intergrid Verbindungen nach Belieben, Bedarf und / oder Budget aufgebaut und dimensioniert werden.

0915