

Einbindung von Servern in das RZ-Netz

von Dr.-Ing. Behrooz Moayeri



Die meisten Fehler in Rechenzentrumsnetzen sind nach Erfahrungen unserer Troubleshooting Teams auf Mängel in der Koordination zwischen der Konfiguration der Netzkomponenten und der netzbezogenen Konfiguration von Servern zurückzuführen. Dabei können manche solche Fehler Auswirkungen weit über den einzelnen Server haben, an dem die Konfiguration vorgenommen wurde. Ganze Subnetze oder sogar ganze RZ-Netze können von einem solchen Fehler betroffen sein.

Der vorliegende Beitrag ist eine Zusammenfassung der Erfahrungen mit solchen Fehlern und der Erkenntnisse aus den Projekten, in denen es um die abgestimmte Konfiguration der RZ-Netzkomponenten und der Server geht. Generell ist eine engere Koordination der Disziplinen erforderlich, die aus historischen Gründen insbesondere in großen Unternehmen unterschiedlichen Instanzen zugeordnet sind: Kompetenz-Center für Server und für das Netz.

Auf dem Rechenzentrum Infrastruktur-Redesign Forum 2010 der ComConsult Akademie werden aktuelle Fragestellungen im Zusammenhang mit RZ-Netzen von Referenten mit jahrelangen Erfahrungen in diesem Bereich behandelt. Unter anderem wird diskutiert, wie Fehler in Rechenzentren auch durch ein robustes Design vermieden oder zumindest in ihrer Auswirkung eingeschränkt werden können.

weiter auf nächster Seite

Einbindung von Servern in das RZ-Netz

Fortsetzung von Seite 1



Dr.-Ing. Behrooz Moayeri hat viele Großprojekte mit dem Schwerpunkt RZ-Redesign geleitet. Er gehört der Geschäftsleitung der ComConsult Beratung und Planung GmbH an und betätigt sich als Berater, Autor und Seminarleiter.

Zunehmende Vielfalt von Servern in Rechenzentren

In den Rechenzentren sind vielfältige Typen von Servern im Einsatz (siehe Abbildung 1).

Nur wenige Rechenzentren werden auf grüner Wiese gebaut und ausschließlich mit neuer Serverhardware ausgestattet. Die meisten Rechenzentren sind historisch gewachsen. Obwohl Serverhardware nach 3-4 Jahren als veraltet gilt, gibt es viele Gründe, auch ältere Serverhardware weiter zu betreiben, vor allem motiviert durch die Vermeidung von Risiken bei der Umstellung einer Anwendung auf eine neue Plattform, wenn die Anwendung die hohe Leistung neuer Hardware nicht unbedingt braucht.

So ist in den meisten RZs eine Vielfalt von Servern mit unterschiedlicher Technologie im Einsatz, die weiter anzubinden sind:

- Ältere Server, die teilweise nachträglich mit Netzadaptern ausgestattet wurden,
- Rack Server mit eingebauten Netzadaptern,
- Blade Enclosures mit diversen Typen von Connectivity-Modulen und
- Virtuelle Server, die über virtuelle Switches (Funktionalität in den Hosts) mit dem Netz verbunden sind.

Unabhängig vom Servertyp haben sich einige Verfahren und Mechanismen wie Adapter Teaming und Lastverteilung in den letzten Jahren bei allen Herstellern ähnlich entwickelt. Wir gehen im Folgenden zunächst auf diese Verfahren ein, bevor besondere Aspekte bei Blade- und virtuellen Servern diskutiert werden.

Adapter Teaming

Mit Adapter Teaming wird erreicht, dass alle Ausfallszenarien bis auf den Ausfall des gesamten Servers durch automatische Redundanzmechanismen abgedeckt werden. Auch bei Einsatz von Serverclustern und hochverfügbaren virtualisierten Umgebungen bevorzugen viele RZ-Betreiber die redundante Anbindung jedes physikalischen Servers an das Netz. Diese Präferenz wird in der Regel wie folgt begründet:

- Server-Cluster-Mechanismen können den redundanten Anschluss jedes einzelnen Servers nicht voll ersetzen. Failover-Mechanismen bei Server-Clustern greifen wesentlich langsamer als Redundanzmechanismen bei Adapter Teaming. Wenn auf Adapter Teaming verzichtet wird, führt der Ausfall der Netzverbindung eines Server-Cluster-Knotens zum Failover von einem zum anderen Server mit deutlich längerer Ausfallzeit als bei einer reinen Adapterumschaltung.
- Server, die nicht per Server Cluster abgesichert werden, können immerhin von Adapter Teaming profitieren.
- Mit Adapter Teaming können die Adap-

ter so eingestellt werden, dass sie regelmäßig Multicasts senden, so dass Pakete an die Adressen der Server nicht geflutet werden.

Das grundsätzliche Netzkonzept für Adapter Teaming geht aus der Abbildung 2 hervor.

Dieses Netzschema ist Grundlage der heute meistens angewandten Verfahren für den redundanten Netzanschluss eines einzelnen Servers. Dieses Konzept sieht vor, dass beide Netzadapter des Servers an das selbe IP-Subnetz (und somit an die selbe Layer-2-Broadcast-Domäne, d.h. auch das selbe virtuelle LAN – VLAN) angeschlossen werden.

Zwischen den beiden Ports, die für den Anschluss des Servers verwendet werden, muss eine Layer-2-Verbindung bestehen. Im dargestellten Netz erfolgt dies mithilfe zweier Layer-2-Switches, die über eine direkte Verbindung miteinander verbunden sind. Beide Layer-2-Switches befinden sich in derselben Layer-2-Broadcast-Domäne, auf die das IP-Subnetz A abgebildet ist. Sehr wohl können an dieses IP-Subnetz verschiedene Layer-3-Switches angeschlossen werden, wie in der Abbildung 2 auch dargestellt ist.

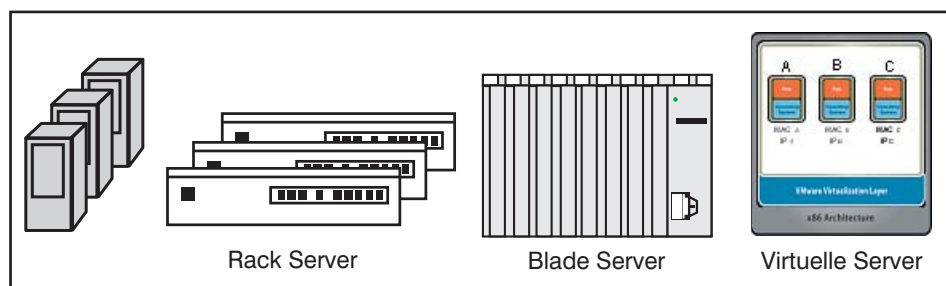


Abbildung 1: Vielfalt der Servertypen

Einbindung von Servern in das RZ-Netz

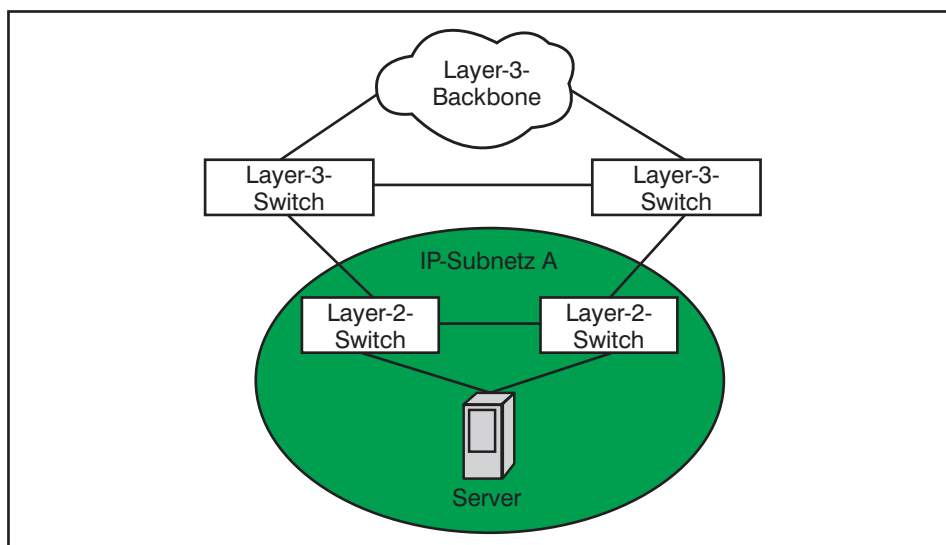


Abbildung 2: Redundanter Serveranschluss in einem IP-Subnetz

Beispiel HP Network Fault Tolerance

Exemplarisch für alle anderen ähnlichen Mechanismen sei hier Network Fault Tolerance (NFT) von HP erwähnt. Das Verfahren ist in der Abbildung 3 dargestellt.

Im Normalfall ist die Network Interface Card (NIC) mit der Nummer 1 der aktive Netz-adapter des Servers. Fällt der Layer-2-Switch, an den dieser Adapter angeschlossen ist, oder die Verbindung des Adapters zu diesem Switch aus, wird dieser Ausfall vom Teaming-Treiber bemerkt, der den zweiten physikalischen Adapter (NIC 2) aktiviert. Eine Layer-2-Verbindung zwischen den beiden Layer-2-Switches ist notwendig. Beide Adapter befinden sich in derselben Broadcast-Domäne bzw. VLAN.

Bei NFT und anderen ähnlichen Teaming-Modi ist immer nur eine Karte aktiv (Primary Adapter), und die anderen Karten sind standby. Eine Lastverteilung findet nicht statt. Alle Adapter im Team verwenden dieselbe IP-Adresse. Die Switches bemerken nichts von NFT und bedürfen keiner speziellen Einstellung.

Eine Erweiterung von NFT ist der Mechanismus „NFT mit Preference Order“. Hierbei wird der Primäre Adapter administrativ gesetzt. Ohne eine solche Reihenfolge bleibt zum Beispiel nach einem Fail-over der Adapter, der die Funktion des aktiven NIC übernommen hat, weiterhin aktiv, während bei Aktivierung der Preference Order die Wiederverfügbarkeit des administrativ gesetzten Adapters eine Umschaltung zurück zu diesem Adapter verursacht.

Transmit Load Balancing

HP Transmit Load Balancing (TLB) bietet eine Lastverteilung beim Senden vom Server. Alle Adapter in einem Team melden sich mit derselben IP-Adresse im Netz. Alle Adapter, primäre und standby-NICs, senden, aber nur der primäre Adapter empfängt. Es ist keine spezielle Konfiguration an den Switches notwendig. Auch hier müssen sich alle Netzkarten in der gleichen Broadcast-Domäne bzw. im gleichen VLAN befinden.

Switch-Assisted Load Balancing (SLB) bietet eine Lastverteilung in Send- und Empfangsrichtung. Am Switch muss eine Link Aggregation Group (LAG) konfiguriert oder mit dynamischen Mechanismen zwischen Switch und Adaptern ausgehandelt werden.

Zusammenfassend kann exemplarisch

anhand der HP-Namenskonventionen zwischen folgenden Teaming-Modi unterschieden werden:

- Network Fault Tolerance (NFT):
 - Ein Adapter ist der primäre Adapter.
 - Alle anderen Adapter im Team sind standby.
 - Der Anschluss der Adapter an verschiedene Switches ist möglich.
- NFT mit Preference Order: entspricht NFT, wobei der primäre Adapter administrativ gesetzt ist
- Transmit Load Balancing (TLB):
 - Der primäre Adapter sendet und empfängt.
 - Andere Adapter senden nur.
 - Anschluss der Adapter an verschiedene Switches ist möglich.

- TLB mit Preference Order: entspricht TLB, wobei der primäre Adapter administrativ gesetzt ist.

- Switch-Assisted Load Balancing (SLB)
 - Link Aggregation zwischen Server und Switch wird genutzt.
 - Der Anschluss der Adapter erfolgt an denselben Switch (oder einen Switch-Cluster, der Multi-Chassis Link Aggregation unterstützt).

Basic und Advanced Teaming

Bei HP wird zwischen Basis- und Zusatz-Funktionen von Adapter Teaming unterschieden. Unter die Basisfunktionen fallen die folgenden:

- NFT
- NFT with Preference
- TLB
- TLB with Preference
- SLB

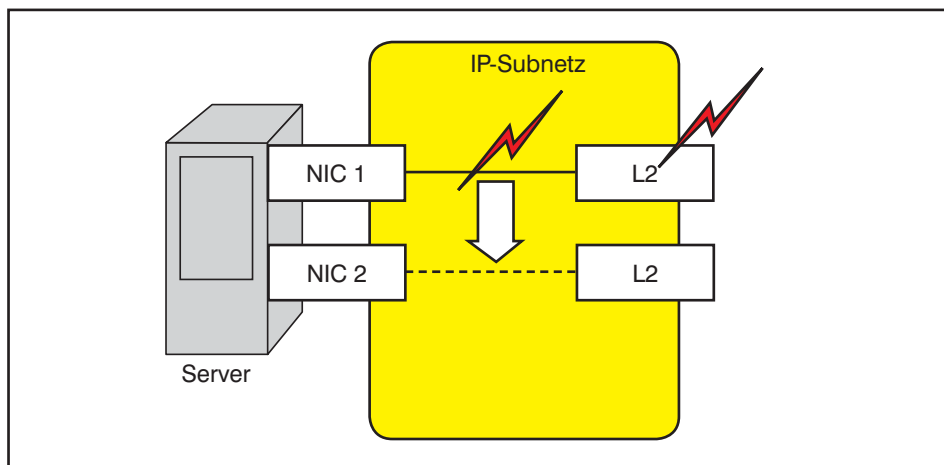


Abbildung 3: Network Fault Tolerance (NFT) von HP

Einbindung von Servern in das RZ-Netz

- 802.3ad Dynamic: Link Aggregation Control Protocol (LACP) zwischen Server und Switch
- Link Loss Detection
- Transmit Validation Heartbeats
- Receive Validation Heartbeats (wird von SLB nicht unterstützt)

Mit Basic Teaming kann eine Netzsegmentierung im übergeordneten Netz (hinter dem Upstream Switch) nicht optimal behandelt werden. Sind zwei Adapter mit den Switches S1 und S2 verbunden, und sind beide Switches noch aktiv, jedoch ohne Layer-2-Verbindung untereinander, bestimmt der Teaming-Treiber den aktiven Adapter ohne Rücksicht darauf, über welchen Teil der in zwei Segmente zerfallenen Broadcast-Domäne der Server tatsächlich mit seinen Zielen kommunizieren kann.

U.a. um dieses Problem zu beheben, bietet HP daher die Advanced Teaming Features an, die folgende Funktionen umfassen:

- Active Path:
 - Ein externer Echo Node muss erreicht werden.
 - Wird der Echo Node von einem Adapter nicht erreicht, geht der Teaming-Treiber davon aus, dass der Adapter (temporär) aus dem Team zu nehmen ist.
- Fast Path:
 - Spanning Tree Bridge Protocol Data Units (BPDUs) müssen empfangen werden.
 - Sonst kann Fail-over ausgelöst werden.
- Dual Channel („Team of Teams“)
 - Jeweils mehrere Adapter gehören einem Team an, wobei SLB in diesem Team möglich ist.
 - Zwei Teams werden gebildet, wobei die beiden Teams mit verschiedenen Switches verbunden sein können.
- Dynamic Dual Channel: entspricht Dual Channel mit LACP.

Ein Beispiel für Active Path ist in der Abbildung 4 dargestellt. Erreicht der Teaming-Treiber über einen Adapter des Teams den sogenannten Echo Node (hier zum Beispiel den Default Router) nicht, wird der Adapter durch einen anderen im Team als aktiver Adapter ersetzt.

Welcher Teaming-Modus ist besser?

Es stellt sich die Frage, ob Lastverteilungsfunktionen von Adapter-Teaming-Lösungen genutzt werden sollten. Der Autor ist der Auffassung, dass der Aktiv-Standby-Modus

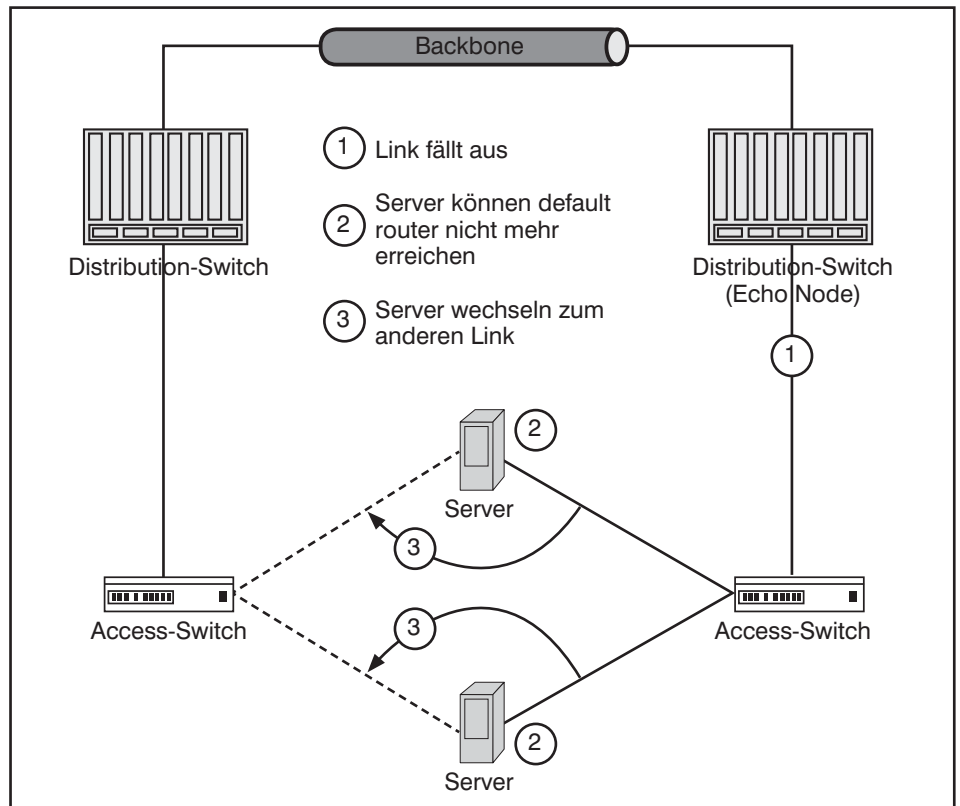


Abbildung 4: Active Path

zu bevorzugen ist. Diese Auffassung basiert auf folgenden Argumenten:

- Der Aktiv-Standby-Modus wird von allen Herstellern unterstützt. Die Netzkonfiguration muss nicht von Fall zu Fall anders erfolgen.
- Reine Fault Tolerance sorgt dafür, dass der Verkehr des Servers immer einen bestimmten Weg im Netz nimmt, der für beide Richtungen genutzt wird. Asymmetrische Verkehrsleitung, welche mit der Gefahr von Flutung im Netz verbunden ist, wird vermieden.
- Für Fehlersuche und Analyse ist es essenziell wichtig, den gesamten Verkehr eines Servers über einen Port zu führen.
- 10Gigabit Ethernet stellt bisher für keinen Server einen Engpass dar (und der Standard IEEE 802.3ba für 40Gigabit/100Gigabit Ethernet wurde im Juni verabschiedet). Somit entfällt die Notwendigkeit von Adapter Load Balancing, das mit einer komplexen und fehleranfälligen Netzkonfiguration einher geht.

Ferner steht man vor der Wahl verschiedener Konfigurationen dafür, wie der aktive Adapter bestimmt wird. Der Mechanismus HP Active Path ist natürlich verlockend,

weil damit sogar alle anderen Layer-2-Redundanzmechanismen im RZ überflüssig werden. Das Layer-2-Netz im RZ kann ohne Redundanz arbeiten, und es reicht, wenn die Server mit verschiedenen Switches in diesem Layer-2-Netz verbunden werden. Active Path ist jedoch nicht bei allen Teaming-Varianten und auf allen Servern möglich.

Weil nicht alle Server einen Mechanismus wie Active Path unterstützen, muss das Netz ausschließen, dass zwei Adapter im Team von der funktionierenden eigenen Netzverbindung ausgehen, jedoch mit zwei disjunkten Layer-2-Netzen verbunden sind. Fazit: Layer-2-Redundanzmechanismen werden im RZ-Netz (leider) gebraucht.

Anbindung von Blade-Servern

Bei vielen Unternehmen werden Blade Server eingesetzt. Die Varianten des Netzanschlusses von Blade-Servern sind wie folgt:

- Pass-through-Modul: Die internen Ethernet-Verbindungen der Server werden 1:1 nach außen verlängert. Diese Variante stellt kein Problem dar, da es sich um eine reine physikalische Verlängerung der externen Verbindungen bis

Einbindung von Servern in das RZ-Netz

zu den Netzadaptern der Blade-Server handelt. In diesem Fall sind die Blade-Server wie externe physikalische Server zu behandeln.

- Switch-Modul: Layer-2- oder Layer-3-Switches werden als Einschübe in den Blade Enclosures eingesetzt. Interne Ethernet-Verbindungen zu den Servern werden vom Switch mit externen Ports verbunden.
- HP VirtualConnect und vergleichbare Lösungen anderer Hersteller: Die Netzadapter der Server werden virtualisiert. Nach außen stellt sich das Connectivity-Modul wie ein Pass-through-Modul dar. Eine interne Querverbindung von zwei externen Ports des Connectivity-Moduls kann unterbunden werden.

Im Zusammenhang mit Blade Switches wird die Frage immer wieder diskutiert, ob auf diesen Spanning Tree zu nutzen oder zu deaktivieren ist. Es kommt immer wieder zu inkompatiblen Spanning-Tree-Konfigurationen zwischen Blade- und externen Switches. Außerdem wird Spanning Tree nicht gerade stabiler, wenn die Zahl der am Algorithmus teilnehmenden Switch-Instanzen steigt. Es reicht ein defekter oder falsch konfigurierter Blade-Switch, um eine ganze Layer-2-Domäne mit Spanning Tree zu destabilisieren.

Daher bevorzugt der Autor Verfahren wie HP VirtualConnect (auch andere Hersteller unterstützen solche Verfahren). Virtual Connect ist eine Layer-2-Komponente, die folgende Ports per Bridging miteinander verbinden kann:

- Downlink-Port mit Downlink-Port
- Downlink-Port mit Uplink-Port

Die interne Querverbindung zwischen Uplink-Ports wird verhindert, sodass ein VirtualConnect-Modul immer so konfiguriert wird, dass eine Schleifenbildung über das Modul ausgeschlossen ist. Da VirtualConnect nicht am Spanning-Tree-Protokoll teilnimmt, gilt es aus der Sicht des Upstream-Switches als Endgerät und nicht als Switch. Daher kann der entsprechende Port im Upstream-Switch so konfiguriert werden, dass die Spanning-Tree-Initialisierung nicht vollständig durchlaufen wird (Fast Start bzw. PortFast, je nach Hersteller).

Auf VirtualConnect-Modulen kann der Modus SmartLink verwendet werden, damit bei Ausfall aller Uplinks auch alle Downlinks deaktiviert werden. Die Nutzung von SmartLink ist in der Abbildung 5 dargestellt.

stellt. Ein Blade-Server mit Adapter Teaming reagiert auf den Link-Ausfall mit dem Fail-over zu einem anderen Adapter, der mit einem anderen VirtualConnect-Modul verbunden sein kann. Mehrere VirtualConnect-Module können in einer Blade Enclosure eingesetzt werden.

Virtuelle Switches

In Virtualisierungsumgebungen wie zum Beispiel auf VMware ESX Hosts werden virtuelle Switches eingesetzt, welche die Virtuellen Maschinen (VM) über den Netzadapter des Host-Systems mit externen Switches verbinden. Im Falle VMware ist der vSwitch (bzw. Standard-Switch) in den meisten Umgebungen im Einsatz. Dessen Arbeitsweise entspricht dem VirtualConnect-Ansatz bei HP Blade Enclosures. Der VMware-Standard-Switch arbeitet wie folgt:

- Es erfolgt kein Lernen von MAC-Adressen, sondern eine Zuordnung von MAC-Adressen anhand der Registrierung der VM-NICs.

• Es erfolgt keine Teilnahme am Spanning Tree im Netz, d.h. kein Senden und keine Verarbeitung von Bridge Protocol Data Units (BPDUs), daher sind die entsprechenden Ports auf externen Switches so zu konfigurieren wie Ports für ganz normale Endgeräte.

• Virtuelle Switches können nicht miteinander verbunden werden.

• Innerhalb eines virtuellen Switches ist keine Schleife möglich.

Der Standard Switch unter VMware hat vor allem den Nachteil, dass er auf jedem ESX Host separat konfiguriert werden muss. Erst der Distributed Switch bzw. der in VMware integrierbare Cisco Nexus 1000V ermöglicht eine zentrale Konfiguration von Switches auf mehreren ESX-Servern. Die Arbeitsweise wie oben beschrieben kann beibehalten werden.

Servercluster

Ein Server Cluster ist eine Gruppe eigen-

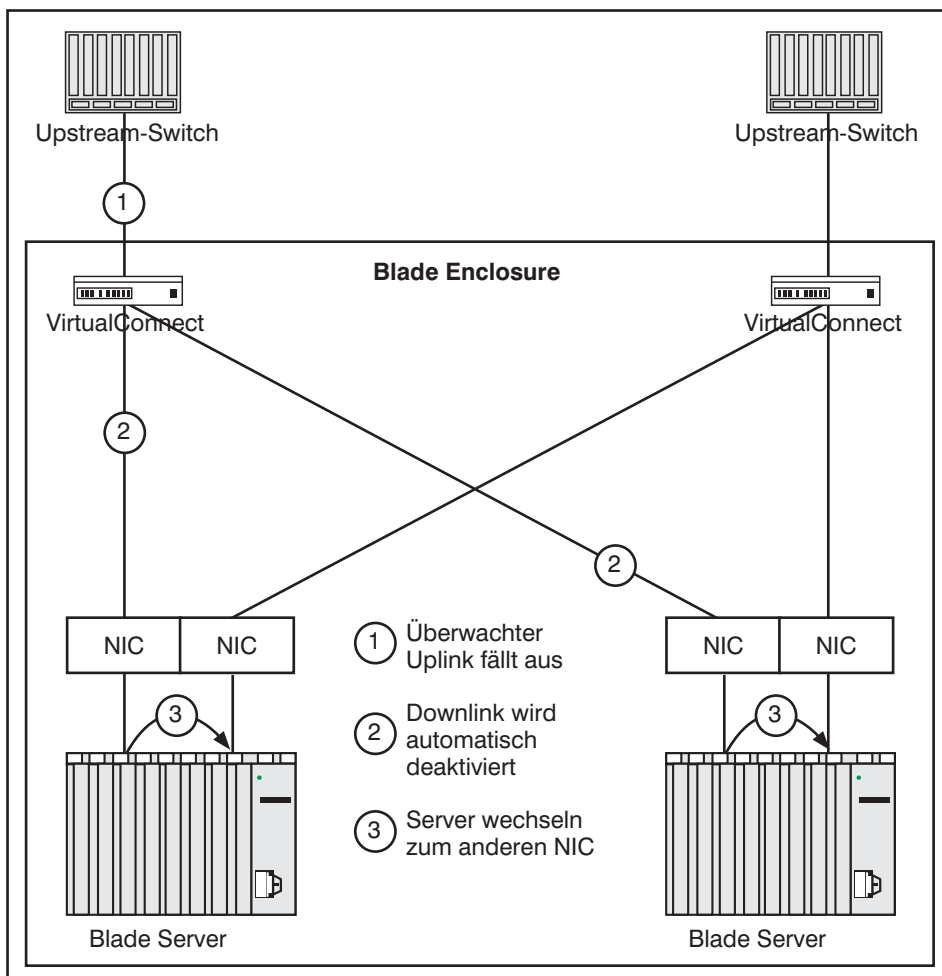


Abbildung 5: VirtualConnect SmartLink

Einbindung von Servern in das RZ-Netz

ständiger Rechner, die zur Realisierung gemeinsamer Applikationen oder Dienste zusammenarbeiten und sich auch nach außen hin (aus der Sicht der Clients) als eine Einheit darstellen. Fail-over-Mechanismen sorgen dafür, dass bei Ausfall eines Servers oder der Netzanbindung eines Servers die für eine Applikation oder einen Dienst erforderlichen Ressourcen auf einem anderen Knoten im Cluster zur Verfügung gestellt werden.

Im Zusammenhang mit reinen Clustermechanismen ohne Lastverteilung sind Fehlkonfigurationen mit Auswirkungen auf das gesamte Netz selten. Problematisch sind jedoch einige Lastverteilungsmodi bei Serverclustern, vor allem Microsoft Network Load Balancing (NLB). Mehrere identische Instanzen derselben Applikation müssen dafür gleichzeitig auf verschiedenen Servern lauffähig sein. NLB verteilt die Last auf verschiedene Server. Alle Knoten im Cluster erscheinen im Netz unter derselben (oder denselben) virtuellen Cluster-IP-Adresse(n).

Damit NLB korrekt funktioniert, muss jedes Paket, das an die virtuelle IP-Adresse gerichtet ist, jeden Knoten im Cluster erreichen. Jeder Knoten entscheidet dann für sich, welche Pakete er akzeptiert und bearbeitet. Hier entsteht ein Problem in geschichteten Netzen. Ein Layer-2-Switch lernt anhand des Source-Adress-Feldes im MAC-Header die Zuordnung einer Adresse zu genau einem physikalischen Port. NLB kann unter diesen Bedingungen nicht arbeiten, da die Forderung, dass jeder Knoten im NLB-Cluster jedes an den Cluster gerichtete Paket empfängt, nicht erfüllt wird.

Man kann NLB so konfigurieren, dass eine Maskierung der Source-MAC-Adresse durch eine Adresse erfolgt, die der Cluster nicht für den Empfang von Paketen verwendet. Zum Empfang wird eine MAC-Adresse genutzt (und per ARP anderen Stationen mitgeteilt), die nie im Source-Feld eines Paketes auftaucht und daher den Switches unbekannt bleibt. Pakete mit Zieladressen, die einer Brücke gemäß der Spezifikation IEEE 802.1D unbekannt sind, werden gemäß dieser Spezifikation auf alle Ports derselben Broadcast-Domäne verteilt, d. h. die Pakete an den NLB-Cluster werden geflutet und erreichen alle NLB-Knoten.

Die Flutung eines Paketstroms auf alle Ports einer Broadcast-Domäne hebt die lasttrennende Wirkung von Layer-2-Switching auf. Es ist z. B. denkbar, dass Ports mit der Bitrate 100 Mbit/s von einem Paketstrom, der eigentlich nur zwei Ports mit

1 Gbit/s als Empfänger betrifft und durch Flutung auf alle anderen Ports übertragen wird, regelrecht blockiert werden. Außerdem verwendet die Architektur mancher Layer-2-Switches intern für die Übertragung von Unicasts (Pakete an eine einzige Adresse, bestimmt für einen einzigen Port) andere, schnellere Pfade und Mechanismen als solche, die zur Übertragung von Broadcasts, Multicasts und Paketen mit unbekannter Zieladresse genutzt werden. Oft kann die Verteilung von Broadcasts, Multicasts und Paketen unbekannter Zieladresse nur mit wenig Performance erfolgen und belastet zudem die Switches stark.

Dabei ist es meist unerheblich, ob für NLB eine Flutung von Paketen mit unbekanntem Ziel erfolgt oder, wie von Microsoft auch unterstützt, Multicasts zum Erreichen des Clusters genutzt werden. Die massive Übertragung von Multicasts und Paketen mit unbekanntem Ziel stößt fast immer an die Grenzen der Leistungsfähigkeit des Layer-2-Switches für Punkt-zu-Mehrpunkt-Übertragung.

Aus diesen Gründen ist der Einsatz von NLB nur als eine nicht skalierbare Lösung für kleine Server-Bereiche geeignet, nicht jedoch für ein mittleres oder großes RZ. Load Balancing sollte so realisiert werden, dass die Pakete an den Lastverteilungscluster vom Netz nicht „vervielfältigt“ werden müssen.

Fazit

Bei der Konfiguration der Netzanbindung von Servern kann man einiges falsch machen. Leider sind dann die Auswirkungen solcher Fehler nicht immer auf einen Server beschränkt, sondern können große Bereiche im RZ lahmlegen. Daher ist eine sorgfältig durchdachte, zwischen den Verantwortlichen für das Netz und für die Server abgestimmte Konfiguration der netzbezogenen Konfiguration von Servern erforderlich. Dies kann am besten durch die Etablierung von Standardkonfigurationen im Unternehmen erreicht werden.

Jetzt Leser werden**Der Netzwerk Insider**

Der Netzwerk Insider erscheint 12 Mal im Jahr im PDF-Format und informiert Sie per eMail über die Hintergründe aktueller Netzwerk-Technologien. Jeden Monat werden zwei Themen gewählt, über die in ausführlicher Form topaktuelle Insider-Informationen gegeben werden. Der Netzwerk-Insider vertritt die Sichtweise von Technologie-Anwendern und bewertet Produkte und Technologien im Sinne der wirtschaftlichen und erfolgreichen Umsetzbarkeit in der täglichen Praxis. Durch seine strenge wirtschaftliche Unabhängigkeit (keine Hersteller-Anzeigen) kann er es sich leisten, Schwachstellen und Nachteile offen anzusprechen. Der Netzwerk-Insider ist bekannt für seine kritische, herstellerneutrale und fundierte Technologie-Bewertung.

Hier können Sie sich zum Netzwerk Insider kostenlos und ohne jede Verpflichtung registrieren lassen:

<http://www.comconsult-akademie.de/de/Registrierung.php>