Redundante Systeme

Hochverfügbare Systeme

# Was bedeutet „Redundanz“?

Redundanz (lateinisch redundantia) bedeutet Überfülle und im Sachzusammenhang das Vorhandensein von mehreren Komponenten (z.B. Server) um in Ausnahmefällen (z.B. defekte Hardware) den Betrieb des IT-Systems aufrecht zu erhalten.

# Wie wichtig sind redundante Systeme?

Für Unternehmen ist es sehr wichtig, dass IT-Systeme ihre Dienste zuverlässig und möglichst ohne Ausfälle anbieten. Daher sollten die Anforderungen an ein solches System gut analysiert und geplant werden. Die Komponenten (Hardware) sollten so stabil wie möglich ausgelegt sein, um eine hohe Zuverlässigkeit zu bieten und wenn es erforderlich ist redundant vorhanden sein. Das sind die Voraussetzungen für ein zuverlässiges bzw. hochverfügbares System.

# Was versteht man unter Hochverfügbarkeit?

Unter Hochverfügbarkeit versteht man die Wahrscheinlichkeit das ein System trotz Ausfall einer oder mehrerer Komponenten den Betrieb aufrecht halten kann.

Ein hochverfügbares System kann in sogenannte Verfügbarkeitsklassen eingeteilt werden, wofür es zwei bekannte Verfahren (**AEC** & 9er-System) gibt.

## Availability Environment Classification (AEC)

Die Harvard Research Group klassifiziert die Hochverfügbarkeit in der **AEC** mit sechs Klassen und ermöglichen dabei eine transparente und schnelle Zuordnung der Anforderungen an einzelne Komponenten oder ein ganzes System.

|  |  |  |
| --- | --- | --- |
| **HRG Klasse** | **Bezeichnung** | **Erklärung** |
| *AEC-0* | Conventional | Funktion darf unterbrochen werden und Datenintegrität ist nicht essentiell |
| *AEC-1* | Highly Reliable | Funktion darf zwar unterbrochen werden, die Datenintegrität muss jedoch immer gewährleistet sein |
| *AEC-2* | High Availability | Funktion darf zur Hauptbetriebszeit nur minimal oder innerhalb festgelegter Zeiten unterbrochen werden |
| *AEC-3* | Fault Resilient | Funktion darf maximal außerhalb festgelegter Betriebszeiten/ Hauptbetriebszeiten unterbrochen werden |
| *AEC-4* | Fault Tolerant | Funktion muss jederzeit aufrechterhalten werden  (24/7-Betrieb) |
| *AEC-5* | Disaster Tolerant | Funktion muss auch unter ungewöhnlichen Umständen  (z.B. im Katastrophenfall) gewährleistet sein |

## 9er System

|  |  |  |
| --- | --- | --- |
| **Verfügbarkeitsklasse** | **Verfügbarkeit** | **Maximale Ausfallzeit pro Jahr** |
| *2* | 99 % | 3 Tage 15 Stunden 36 Minuten |
| *3* | 99,9 % | 8 Stunden 46 Minuten |
| *4* | 99,99 % | 53 Minuten |
| *5* | 99,999 % | 5 Minuten |
| *6* | 99,9999 % | < 1 Minute |

Grundlage der Berechnung: 365 Tage und 24 Stunden pro Tag; die entsprechende maximale Ausfallzeit ist aufgerundet.

Im Regelfall spricht man ab Verfügbarkeitsklasse 3 oder höher von einem hochverfügbaren System.

# Wie wird Redundanz erreicht?

Um ein System redundant zu machen müssen mehrere Schichten betrachtet und gesichert werden.

## Schicht 1 - physikalische Redundanzen

In dieser Schicht werden physikalisch Komponenten eines Systems betrachtet. Um zusätzlichen Schutz zu erhalten kann man auf Georedundanz zurückgreifen (redundante Geräte an unterschiedlichen Orten). Die wesentlichen Bestandteile der physikalischen Redundanz sind:

* Stromversorgung (mehrere Netzteile, USV)
* Kühlung
* Leitungen (Stromkabel, Netzwerkkabel, Verbindungskabel)
* Hardware (Server, Switches, Festplatten (RAID-Systeme), Router, …)

### **Festplatten – RAID**

Ein RAID (Redundant Array of Independent Disks) zielt darauf ab, Daten sicher zu speichern, ohne dass diese verloren gehen. Dabei verwendet das RAID System meist ein redundantes Verfahren.

RAID ersetzt kein Backup. RAID hilft nur beim Ausfall einer Festplatte (nicht bei RAID 0). Wenn gleichzeitig eine zweite Festplatte ausfällt, dann hilft RAID nicht gegen Datenverlust (außer RAID 6). Deshalb gilt es bei einem Festplattenausfall schleunigst Ersatz zu beschaffen.

Offiziell gibt es 8 RAID-Level (0 bis 7), wobei nur die Level 0 bis 5 spezifiziert sind. und reichen aus, um die meisten Anforderungen abzudecken.

In einem RAID sollten grundsätzlich dauerbetriebstaugliche oder RAID-taugliche Festplatten (24x7) einsetzt werden. Solche Festplatten sind natürlich etwas teurer. Allerdings vibrieren sie weniger. Beim Betrieb mehrere Festplatten in einem Gehäuse können sich Festplatten sonst gegenseitig in Schwingung versetzen. Herkömmliche Festplatten mit hoher Speicherkapazität sind wegen der hohen Spur- und Datendichte sehr empfindlich gegen Vibrationen. Höhere Latenzzeiten und im schlimmsten Fall Schreib- oder Lesefehler sind die Folge. In einem RAID führt das zu Fehlerkorrekturen, die das System verlangsamen können.

Wenn ein RAID 1 oder 5 nach langjähriger Betriebszeit einen Festplatten-Ausfall erleidet, dann sollte man das defekte Laufwerk nicht sofort tauschen. Tauscht man die defekte Platte aus, dann kommt es nicht selten vor, dass nach einem stressigen Rebuild weitere Platten ausfallen. Dann muss man das ganze Spiel wiederholen. Im schlimmsten Fall fällt eine andere Festplatte während des Rebuilds aus. Dann sind die Daten definitiv futsch. Deshalb gilt, RAID ersetzt kein Backup!

Wenn also in einem RAID eine Festplatte ausfällt müssen zuerst die Daten auf den verbliebenen Festplatten gesichert und anschließend alle Festplatten des RAIDs ausgetauscht werden.

Bei einem RAID 6 kann man sich diese Vorgehensweise theoretisch sparen. RAID 6 kommt auch mit dem Ausfall einer zweiten Platte klar.

* Hardware-RAID (Der RAID-Controller nimmt die Verteilung der Daten und die Berechnung der Prüfsummen selber vor)
* Software-RAID (Einige RAID-Level können auch mit Hilfe von Software realisiert werden. Für die Verteilung der Daten und der Berechnung der Prüfsummen ist dann der Hauptprozessor zuständig.)
* Host-RAID (Eine Zwischenstufe zwischen Hardware- und Software-RAID ist Host-RAID. Dazu zählen der RAID-Chipsatz, der sich auf manchen Motherboards befindet und auch einige günstige RAID-Adapter.)
* RAID-0: Der RAID-Level 0 ist ein Festplatten-Verbund von zwei oder mehr Festplatten. Die Transferrate und Speicherkapazität der einzelnen Festplatten lassen sich einfach aufaddieren. Die 0 im RAID-Level steht für Null Daten-Redundanz. Die Daten werden nur abwechselnd auf zwei oder mehr Festplatten verteilt.
* RAID-1: Der RAID-Level 1 ist ein Festplatten-Verbund von zwei oder mehr Festplatten. Bei RAID 1 werden die Daten doppelt, also mindestens auf zwei Festplatten gespeichert. Man bezeichnet das als Datenspiegelung oder Mirroring. RAID 1 bietet so den bestmöglichen Schutz vor Datenverlust durch Festplattenausfall. Und beim Lesen von Daten hat man annähernd die doppelte Datentransferrate wie bei einem einzelnen Laufwerk.Aber, die Kapazität eines RAID-Laufwerks beträgt die Hälfte der eingesetzten Laufwerke. Man muss also doppelt so viel Geld ausgeben oder man hat nur halb so viel Speicherkapazität, wie bei einem Laufwerk. Doppeltes Laufwerk bedeutet auch doppelter Stromverbrauch und Kühlaufwand.
* RAID-5: Der RAID-Level 5 ist ein Festplatten-Verbund von drei oder mehr Festplatten mit besonderen Eigenschaften, die eine einzelne Festplatte nicht hat.Bei großen Datenmengen, die redundant gespeichert werden müssen, ist RAID 0 nicht akzeptabel und RAID 1 zu teuer, platzraubend und meistens überdimensioniert. Der RAID-Level 5 ist eine Weiterentwicklung aus den RAID-Leveln 3 und 4. Wie bei RAID 0 werden die Daten in Blöcke und zusätzlich Paritätsblöcke über alle Festplatten verteilt. Der RAID-Level 5 ist eine gute Kombination aus Datensicherheit und Speicherausnutzung. Bei 5 Festplatten beträgt die Speicherkapazität 80% von der Gesamtkapazität aller Festplatten. Beim Ausfall einer Festplatte geht die Datentransferrate zurück, weil die Daten aus der Parität berechnet werden müssen.
* RAID-6: Beim RAID-Level 6 wird wie bei RAID 5 ebenfalls mit verteilten Blöcken Paritätsblöcken gearbeitet. Zusätzlich wird ein Paritätslaufwerk verwendet, das über einen asynchronen Datenpfad und einen Cache verfügt. Bei RAID 6 werden also gleich zwei Paritätsblöcke gespeichert, so dass sich aus den Daten von verbliebenen Laufwerken die Daten rekonstruieren lassen. Während RAID 5 nur den Ausfall eines Laufwerks verkraftet, verträgt RAID 6 den Ausfall von zwei bei mindestens 4 Laufwerken.
* RAID-10: Das RAID 10 ist eine der besseren Varianten, sie vereint das RAID 0 und 1 miteinander. Dafür werden mindestens vier Festplatten benötigt. Die Daten werden immer auf ein Festplattenpaar unterschiedlich gespeichert (RAID 0). In einem Festplattenpaar werden die Daten dann aber wieder synchron gespeichert (RAID 1). Diese Variante überzeugt durch seine Geschwindigkeit, ist aber mit höheren Kosten verbunden.

## Schicht 2 – Datenübertragungsschicht

Auf der dieser Schicht gibt es in Abhängigkeit der eingesetzten Technologie verschiedene Verfahren, um Redundanzen zu ermöglichen. In der Regel bestehen diese immer aus zusätzlichen Leitungen bzw. physikalischen Übertragungswegen in einem Layer 2 Netzwerk (auch Broadcastdomain genannt).

Bei Ethernet wurde diese Redundanz erstmalig mit dem Spanning-Tree-Algorithmus möglich. Dazu wurden Weiterentwicklungen, wie **Rapid** **Spanning** **Tree** und **Spanning** **Forest**, geschaffen, um schnellere Konvergenz im Fehlerfall und eine Verwaltung von Spanning-Trees in einer **VLAN** Umgebung möglich zu machen.

### Link Aggregation IEEE 802.3ad

Sind zwei Switche durch mehrere physikalische Links verbunden, so kann dies aus reinem Redundanzzweck sein. In diesem Fall kann man eine redundante Port Lösung auf Layer 1 oder **Spanning** **Tree** als Protokoll auf Layer 2 einsetzen, um **Loops** und damit **Broadcast** **Storms** zu verhindern.

Eine bessere und gerne genutzte Möglichkeit ist es aber, diese physikalischen Links zu einem logischen Link zu bündeln und so gleichzeitig Redundanz und höhere Bandbreiten zu schaffen.

Während es zwar möglich ist, die Bündelung der physikalischen Links und den daraus resultierenden virtuellen Link statisch zu konfigurieren, ist es besser, ein entsprechendes Kontrollprotokoll zu nutzen. Durch das Kontrollprotokoll ist es möglich zu überprüfen, ob die zugehörigen Links auch sauber laufen und beide Switches darin übereinstimmen, welche Ports zu dem **virtuellen** **Link** gehören.

Frühere **Smarttrunk** Lösungen nutzten das so genannte **Huntgroup** **Protokoll**. Switches, die nach dem Standard **IEEE** **802**.**3ad** arbeiten, kommunizieren über **LACP** (Link Aggregation Control Protocol). Dabei werden **virtuelle** **Links** gebildet, die als **LAG** (Link Aggregation Group) bezeichnet werden. Wichtig ist, dass alle Links, die einem **virtuellen** **LAG** **Port** angehören, gleich konfiguriert sind. Sie müssen **Vollduplex** sein und die gleiche Geschwindigkeit haben. Dann kann dieser **virtuelle** **LAG** **Port** in der Konfiguration der Switches wie ein ganz normaler physikalischer Port genutzt werden – und damit einem **VLAN** angehören oder als **802**.**1q** **Trunk** definiert werden.

In einem **Chassis**-**basierten** **System** ist es möglich, einen **LAG** mit Ports verschiedener Module zu bilden, das Gleiche gilt für ein **Stack**-**basiertes** **System** mit mehreren Switches. Dies erhöht die Ausfallsicherheit, denn selbst falls ein ganzes Board ausfällt, ist weiter eine Verbindung zwischen den beiden **Chassis**-**basierten** **Switches** gegeben.

Der Standard **IEEE** **802**.**3ad** bietet Interoperabilität zwischen verschiedenen Herstellern. Zusätzlich sind einige Switches in den Standardeinstellungen schon so vorkonfiguriert, dass sie automatisch einen **LAG** bilden, wenn sie mit einer entsprechenden Gegenstelle verbunden werden. Sollte es aber nötig sein, einen Switch mit einer Gegenstelle zu verbinden, die kein **LACP** unterstützt, so kann man bei passenden Switches den **LAG** auch statisch konfigurieren. Zwar verzichtet man dann auf das **Kontrollprotokoll**, das vor Fehlern und Netzproblemen durch falsche Konfiguration schützt, kann aber trotzdem die Vorteile eines virtuellen, gebündelten Links nutzen.

Der Traffic kann über verschiedene Methoden auf die **physikalischen** **Links** verteilt werden. Möglichkeiten sind zum Beispiel durch ein einfaches **Round Robin Verfahren** gegeben oder basierend auf MAC oder IP Adressen der Pakete. Meist bietet die Untersuchung von Absende- und Ziel-IP eine ausgewogene Verteilung auf die physikalischen Links.

**IEEE** **802**.**3ad** und **Spanning** **Tree** schließen sich gegenseitig natürlich nicht aus. Während **802**.**3ad** **Link** **Aggregation** immer zwischen zwei direkt miteinander verbundenen Switches läuft, kann und sollte man weiterhin **Spanning** **Tree** nutzen, um Loops im gesamten Layer 2 geswitchten Netz zu verhindern. **Spanning** **Tree** betrachtet dann bei der Berechnung des aufspannenden Baumes den **logischen** **LAG** **Port** anstelle der zugehörigen **physikalischen** **Ports**; auch die Kosten für diesen **virtuellen** **LAG** **Port** entsprechen dabei der Summe der Bandbreiten aller zugehörigen physikalischen Ports.

### Spanning Tree IEEE 802.1d

Der **Spanning** **Tree** **Algorithmus** verhindert **Loops** auf der Datenübertragungsschicht. Die Notwendigkeit entsteht dadurch, dass Broadcasts in einem Ethernet-Netzwerk, in dem redundante Pfade vorhanden sind, unendlich lang kreisen und somit die verfügbare Bandbreite immer weiter verringern bis kein normaler Datenverkehr mehr möglich ist. Der heutzutage gebräuchliche **802**.**1d** Standard ging aus der von **Radia** **Perlman** für die **Firma** **DEC** entwickelten, ersten Implementierung des Algorithmus hervor.

Um **Loops** zu verhindern, tauschen die Netzwerkkomponenten (Switches) eines Layer 2 Netzwerks untereinander Nachrichten aus, die vom normalen Datenverkehr unterscheidbar sind. Anhand dieser Nachrichten wird dann eine der Komponenten zur **Wurzel** eines Baumes gewählt. Alle anderen Komponenten gliedern sich in diesen Baum ein. Pfade, die nicht innerhalb dieses Baums liegen (redundante Pfade) werden dabei ausgeschaltet.

Wenn ein neuer Switch oder Link hinzugefügt wird oder ein Switch ausfällt, wird dieser Baum neu berechnet. Solange dieser Baum aber nicht erneut vollständig aufgebaut ist, leiten die Switches nur die für den Aufbau des Baumes relevanten Nachrichten weiter. Das heißt, solange der Baum nicht vollständig wiederhergestellt ist, ist der normale Datenverkehr im Netzwerk unterbrochen. Die Neuberechnung des Baumes dauert typischerweise bis zu 60 Sekunden.

### Rapid Spanning Tree IEEE 802.1w

Der **802**.**1d** **Spanning** **Tree** **Algorithmus** wurde zu einer Zeit entwickelt, in der es ausreichend war, wenn sich das Netz nach einem Ausfall in einem Zeitraum von circa einer Minute erholte.

Heutzutage sind solche Ausfallzeiten nicht mehr akzeptabel. Daher wurde der Spanning Tree Standard zum **Rapid** **Spanning** **Tree** **Standard** (**RSTP**) weiterentwickelt, um schnellere Konvergenzzeiten zu ermöglichen. Prinzipiell wurde die Art und Weise der Berechnung des Baumes beibehalten. Die Nachrichten, die zwischen den Switches ausgetauscht werden, enthalten nun jedoch mehr Informationen. Außerdem wurde die Verarbeitung der Nachrichten verbessert. Die wichtigste Neuerung im **802**.**1w** Standard war die Möglichkeit, einen Port schneller in den **Forwarding**-**Modus** zu bringen, in dem normale Datenpakete ausgetauscht werden können. Im alten Standard wurden die Ports erst dann aktiviert, wenn der gesamte Baum konvergiert war.

Der neue Standard ist dagegen in der Lage sicherzustellen, dass ein Port früher reaktiviert werden kann. Hierfür können Endnutzerports als so genannte Edge Ports konfiguriert werden, die dann beim Aktivieren des Ports sofort aktiv werden. Außerdem können Ports im Fehlerfall so schneller auf einen alternativen Port Richtung Root umschalten, da **RSTP** Switches aktiv Rückmeldungen austauschen können. In entsprechend konfigurierten Netzwerken kann die Konvergenzzeit des Baums damit auf wenige hundert Millisekunden gesenkt werden.

### VLANs IEEE 802.1q

**VLANs** erlauben es, einen Switch mit einer Vielzahl physikalischer Ports logisch zu unterteilen, sodass bestimmte Portgruppen jeweils verschiedene **Broadcastdomains** bilden, obwohl sie physikalisch zum gleichen Switch gehören.

Der **IEEE** **802**.**1q** Standard erlaubt es, die Daten aus solchen logisch getrennten Portgruppen eindeutig (für die jeweilige Gruppe) zu kennzeichnen und sie in dieser Form zu einem anderen Switch zu transportieren. Der Link, über den diese markierten Pakete transportiert werden, wird in der Regel **Trunk** genannt. Dort können die Daten wieder den einzelnen Gruppen zugeordnet werden und, falls diese Gruppen auf diesem Switch ebenfalls existieren, zu den entsprechenden Ports geschickt werden. Man ist also in der Lage, eine logische **Broadcastdomain**-**Struktur** über eine physikalisch vorgegebene Struktur von untereinander verkabelten Switches zu legen. Dadurch können die Mitarbeiter einer Abteilung mit ihren Rechnern in derselben Broadcastdomain sein, obwohl die Abteilung auf verschiedene Gebäude verteilt ist.

### Multiple Spanning Trees IEEE 802.1s

**Multiple** **Spanning** **Trees** (**MST**) ist eine Ergänzung des **802**.**1q** Standards. Der **802**.**1w** **Rapid** **Spanning** **Tree** Standard wurde erweitert, um mehrere **Spanning** **Trees** zu unterstützen. Diese Ergänzung erlaubt sowohl schnelle **Konvergenz** als auch **Load** **Sharing** in einer **VLAN** Umgebung.

Mit **MST** wird es möglich, mehrere **Spanning** **Tree** **Instanzen** über **Trunks** hinweg aufzubauen. Dabei können in Gruppen zusammengefasste **VLANs** einzelnen **Spanning** **Tree** **Instanzen** zugeordnet werden. Die einzelnen Instanzen können dabei unabhängig voneinander verschiedene Topologien haben. Dafür werden die **Spanning** **Tree** **Parameter** wie **Root** **Priorität**, etc. für jede Instanz angepasst. Damit wird ein **Load** **Sharing** für unterschiedliche **VLAN** **Gruppen** über redundante Layer-2-Wege möglich. **MST** benutzt dabei eine modifizierte Variante des **Rapid** **Spanning** **Tree** **Protokolls**, genannt: **Multiple** **Spanning** **Tree** **Protocol** (**MSTP**) oder **IEEE** **802**.**1s**.

## Layer 3 – Netzwerkschicht

Auf der Netzwerkschicht wird Redundanz hauptsächlich durch intelligente **Routingprotokolle** wie **OSPF** und die Verbesserung der Erreichbarkeit des **Standardgateways** mit **VRRP** erreicht. Um dies sinnvoll zu ermöglichen, müssen schon auf der Netzwerkebene verschiedene Wege zum selben Ziel vorhanden sein.

### Open Shortest Path First

**Open** **Shortest** **Path** **First** (**OSPF**, **RFC** **2328**) ist ein hierarchisch aufgebautes Link State Routingprotokoll und hat sich als Standard für Interior-Gateway-Protokolle etabliert.

Im Gegensatz zu **BGP** (**Border** **Gateway** **Protokoll**), das als **Exterieur**-**Gateway**-**Protokoll** für das Routing zwischen autonomen Systemen genutzt wird, dienen **Interior**-**Gateway**-**Protokolle** dazu Routinginformationen innerhalb einer Organisation auszutauschen.

**Distance**-**Vector**-**Protokolle**, wie **RIP**, sind dafür mittlerweile meist ungeeignet, da diese sehr schlechte Konvergenzzeiten aufweisen – und Ausfallzeiten von bis zu mehreren Minuten sind in heutigen Netzwerken nicht mehr tolerierbar.

**Link**-**State**-**Protokolle** arbeiten eventgesteuert, die Informationen über Topologie Änderungen werden sofort durch das ganze Netz geflutet. Alle Router reagieren sofort darauf und berechnen bestehende Ersatzwege. Somit haben **Link**-**State**-**Protokolle** nur Konvergenzzeiten im Sekundenbereich. Diese schnellen Berechnungen von Ersatzwegen werden dadurch möglich, dass bei **Link**-**State**-**Protokollen** die gesamte Topologie eines Netzes allen Routern bekannt ist; gespeichert in der Topologie-Datenbank.

Da dadurch außerdem die periodische Verbreitung der gesamten Routinginformationen nicht mehr notwendig ist, arbeiten **Link**-**State**-**Protokolle** ressourcenschonender als **Distance**-**Vector**-**Protokolle**.

Der hierarchische Ansatz von **OSPF** macht dieses Routingprotokoll skalierbar, **OSPF** eignet sich daher auch für sehr große Netze. Die Hierarchie basiert auf einem zweistufigen Area Konzept. Es gibt eine zentrale **Backbone** **Area**, an die alle anderen Areas direkt angebunden sind. Durch Verfahren wie **Route** **Summarization** und das Definieren von Arealen als **Stub** **Area** oder **Not**-**So**-**Stubby** **Area** (**NSSA**, **RFC** **3101**) wird die Auswirkung von Topologie Änderungen auf das gesamte Netz minimiert.

**OSPF** ist im Moment das Standard Protokoll für Routing innerhalb des eigenen Netzwerks. Als offener Standard, der von allen Herstellern unterstützt wird, bietet es Kompatibilität zwischen allen Komponenten. Die schnellen Konvergenzzeiten, die **OSPF** als **Link**-**State**-**Protokoll** besitzt, sind für heutige Netze unverzichtbar.

Der hierarchische Ansatz unterstützt die Implementierung in Netzwerken jeder Größe. Netze mit **OSPF** als **Routingprotokoll** sind mit dem entsprechenden Netzwerk- und Adressdesign einfach zu erweitern. Als modernes Routingprotokoll unterstützt **OSPF** natürlich **VLSM** (**Variable** **Length** **Subnetmask**) und ermöglicht so Optimierungsverfahren wie **Route** **Summarization**.

Auf Grund seiner allgegenwärtigen Präsenz wurde **OSPF** auch als Routingprotokoll für das aufkommende IP Protokoll **IPv6** spezifiziert (**RFC** **2740**) und wird auch in Zukunft nicht aus den Netzwerken wegzudenken sein.

### Equal Cost Multi Path (ECMP)

Ein Paket kann in einem gerouteten Netzwerk über unterschiedliche, gleichwertige Pfade ans Ziel gelangen. Bei „Equal Cost Multi Path“ werden diese Pfade gleichzeitig zur Lastverteilung genutzt. Eine Redundanz wird hiermit jedoch nicht gewährleistet, dafür muss das darunterliegende Routingprotokoll sorgen.

Bei Verwendung von **ECMP** wählt der Router an dem sich der Pfad gabelt, unterschiedliche Next-Hops für die Pakete. Idealerweise sollten sich die Pakete natürlich gleichmäßig auf die beiden Pfade verteilen, was natürlich mit paketweisem Aufteilen am einfachsten möglich wäre. Dies ist aber in der Regel nicht sinnvoll, da es dann zu unterschiedlichen Laufzeiten und Paketreihenfolgen kommen kann. Es wird meist versucht, die Pakete flussbasiert, das heißt Sender-IP/Ziel-IP- oder Sender-IP+Port/Ziel-IP+Port-basiert aufzuteilen, da die zu einer Kommunikation gehörigen Pakete dann den gleichen Weg nehmen.

Prinzipiell kann **ECMP** in jedem gerouteten Netzwerk verwendet werden. In der Regel wird die maximale Anzahl der Pfade gleicher Qualität (in Einheiten der Metrik des entsprechenden Routingprotokolls) allerdings beschränkt.

### Virtual Router Redundancy Protocol

Das **Virtual** **Router** **Redundancy** **Protocol** (**VRRP**, **RFC** **2338**) dient dazu, dem Benutzer Redundanz bezüglich des Standardgateways zur Verfügung zu stellen. Während Router untereinander Routing Protokolle nutzen, um die aktuellsten Routing-Informationen auszutauschen und so bei einem Ausfall Ersatzwege zu lernen und zu nutzen, haben sehr viele Endclients statische Einträge für das Standardgateway. Was nun, wenn dieser Router ausfällt? Selbst wenn es einen Ersatzweg gibt, wie sollen die Clients darauf reagieren?

Die Lösung dazu liefert das **VRRP**-**Protokoll**. Die Grundidee besteht darin, einen virtuellen Router zu nutzen und diese IP Adresse als Standardgateway auf den Hosts zu konfigurieren.

Die physikalischen, redundanten Router kommunizieren dann über das **VRRP**-Protokoll und handeln aus, wer die Routing-Aufgabe des Standardgateways übernimmt. Dieser Router wird dann als **Master** bezeichnet, weitere **redundante** **Router** sind **Backup** **Router**. Fällt der Master aus, so wird das über **VRRP** erkannt und einer der **Backup** **Router** übernimmt die Aufgabe des **Masters**. Für den Client ist dies absolut transparent. Damit es keine Probleme bezüglich der **ARP**-**Einträge** gibt, nutzt der **virtuelle** **Router** eine zugehörige, für **VRRP** reservierte **MAC** **Adresse**.

So dient **VRRP** dazu, den **Single** **Point** **of** **Failure**, den die statische Konfiguration eines Standardgateways darstellt, zu eliminieren.

Durch die Konfiguration von zwei virtuellen Routern kann man eine zusätzliche Lastverteilung auf die redundanten Geräte erreichen. Man nutzt zwei **virtuelle** **IP** **Adressen**, jeder der Router ist dann für eine der beiden Adressen der Master, für die andere der **Backup** **Router**. Zusätzlich müssen die Clients dementsprechend entweder mit der Einen oder der anderen IP Adresse als **Standardgateway** konfiguriert werden.

Eine gleichmäßige Verteilung kann man zum Beispiel mit dem **Dynamic** **Host** **Configuration** **Protocol** (**DHCP**) erreichen. So werden beide Router als **Standardgateway** genutzt und man hat Lastverteilung in Kombination mit Redundanz. Weitere Informationen findet man im zugehörigen **RFC** **2338**.

Quellen

<https://www.recast-it.com/themen/verfuegbarkeitsklassen/>

<https://www.computerwoche.de/a/it-ganz-ohne-ausfaelle,3092390>

<https://www.globalsystem.ch/ratgeber/raid-systeme-erklaert/>

https://www.elektronik-kompendium.de/sites/com/1001011.htm

<http://www.ip-insider.de/redundanz-ausfallsicherheit-co-so-erreichen-sie-high-availability-a-274602/>