

Sonderdruck für Gordion

Netzwerk-Redesign der Kölner Verkehrs-Betriebe AG

Redundante Rechenzentren

Ein Umzug des Rechenzentrums lag den IT-Mitarbeitern der Kölner Verkehrs-Betriebe AG (KVB) schon lange am Herzen, bot doch der bereits 1964 bezogene alte Verwaltungsbau einfach zu wenig Platz. Zudem hegte man den Wunsch nach einem zweiten Rechenzentrum, das auch Sicherheitsaspekten, räumlich verteilt, Rechnung trägt und zudem wichtige Backup-Funktionen übernehmen sollte. Die Entscheidung zum Bau eines neuen Verwaltungsgebäudes, dem „Westforum“, bot die Gelegenheit zur Erweiterung.

Da das alte Verwaltungsgebäude kernsaniert und damit das darin enthaltene RZ umziehen musste, packte man die Gelegenheit beim Schopf. Die KVB beschloss, das alte Rechenzentrum aufzulösen und gleichzeitig zwei neue RZs aufzubauen. Es definierten sich insbesondere drei primäre Aufgaben für die EDV: erstens der Aufbau zweier neuer Rechenzentren, zweitens der Umzug des alten RZs in die zwei neuen Rechenzentren mit minimaler Ausfallzeit und drittens die Ausstattung und Integration des neuen Verwaltungsgebäudes (Westforum). Im Zuge des Verwaltungsneubaus wurde entschieden, dass das bestehende, alte Ver-

waltungsgebäude (VGB) kernsaniert werden soll. Da im alten VGB das bisherige Rechenzentrum (RZ) unterbracht war, stand de facto der Aufbau von zwei neuen Rechenzentren (RZ-A und RZ-B) sowie der Umzug des alten RZs bevor. Die beiden Rechenzentren sind räumlich verteilt und liegen daher insbesondere in getrennten Brandabschnitten. Die Verbindung zwischen den RZs ist mit zweimal 144-Adern-9/125-µm Monomode-Glasfasern realisiert, die auf verschiedenen Wegen verlegt sind. Das Rechenzentrum-B ist hierbei nicht, wie man vermuten könnte, ein reines Backup-RZ. Vielmehr ist die

Produktion auf beide Rechenzentren verteilt. Insbesondere sind die Daten über zwei Festplatten-Arrays in einem SAN gespiegelt. Daran angeschlossen sind acht Hochverfügbarkeits-Serversysteme/Cluster. Diese sind verteilt in beiden Rechenzentren installiert. Beide Rechenzentren sind also produktiv und halten für das jeweils andere RZ Backup-Funktionen bereit.

Der Umzug des alten Rechenzentrums in die beiden neuen RZs fand während des Betriebs statt, wobei ein Nahverkehrsunternehmen naturgemäß 24 Stunden sieben Tage in der Woche arbeitet. Das übliche Ausweichen auf Wochenenden stellte demnach keine Lösung dar. Oberstes Ziel war eine Minimierung der Beeinträchtigung des Betriebs bei gleichzeitiger wirtschaftlicher Berücksichtigung des Budgets. Zusätzliche Netzwerkkomponenten hätten die Abwicklung des Umzugs sicher beschleunigt. Die Wirtschaftlichkeit solcher Maßnahmen war jedoch fragwürdig, zumal man nach dem Umzug keine Verwendung für solche „Überbrückungskomponenten“ gehabt hätte. So blieb die Wahl zwischen Leihkomponenten oder geeignetem Umzugsdesign. Man entschied sich für letztere Variante. Dabei kamen vier Phasen heraus, in denen der RZ-Umzug vollzogen wurde, inklusive der Einbindung des Westforums. Für den Aufbau der beiden neuen Rechenzentren sowie zur Integration des neuen Verwaltungsgebäudes „Westforum“ wurde die bisherige Struktur der aktiven Kom-

ponenten schrittweise erweitert. Das Ziel war vor allem eine sanfte Migration mit minimaler Ausfallzeit. Bereits vor einigen Jahren hatten sich die Kölner Verkehrs-Betriebe für aktive Komponenten von Extreme Networks entschieden, die den Core im bisherigen Rechenzentrum definierten (zweimal Black Diamond-6808). Zudem waren auch bereits zahlreiche Edge-Devices, also Etagen-Switches, von Extreme Networks im Einsatz (diverse „Summits“). Im Hinblick auf den Investitionsschutz und zudem hervorragende Erfahrungen mit diesen Komponenten lag es nahe, die zusätzlich benötigten aktiven Komponenten ebenfalls auf Basis von Extreme Networks auszuwählen. Für das Projekt wurden folgende Switch-Komponenten des Herstellers sowie ein zentraler WAN-Router von Cisco im Netzwerk-(Re-)Design berücksichtigt:

Core-Switches, Backbone/Sekundär-Ebene:

Black Diamond 6808: Switched Gigabit Ethernet (1000Base-SX/LX, Layer 3). Sechs Multilayer Chassis Switches Black Diamond 6808 von Extreme Networks (10-Slot-Chassis, Dual Load Sharing Switch Fabric, passive Backplane). 128 GBit/s non-blocking Layer2/3-Switching. Link Aggregation nach IEEE 802.3ad, Frame Tagging (802.1Q), Unterstützung von Policy Based QoS mit Bandbreitenlimitierung und CoS (802.1p, Diffserv). Skalierbar auf 10/100/1000-MBit/s und 10-GbE-Applikationen. 96 Millionen Pakete pro Sekunde. Vollduplex mit Wire-speed Switching/Routing, inklusive Wire-speed Filtering bis zur TCP/UDP-Port-Ebene.

Server-Switches:

Black Diamond 6808 und Alpine 3808: Switched 10/10/1000 MBit/s. Zur primären Anbindung der Server wurde das Black-Diamond-System eingesetzt. Zur redundanten Serveranbindung steht in jedem RZ das kleinere und daher preiswertere Alpine 3808-System bereit, das in weiten Teilen auf denselben Features wie der Black Diamond basiert. Der Alpine bietet jedoch I/O-Module

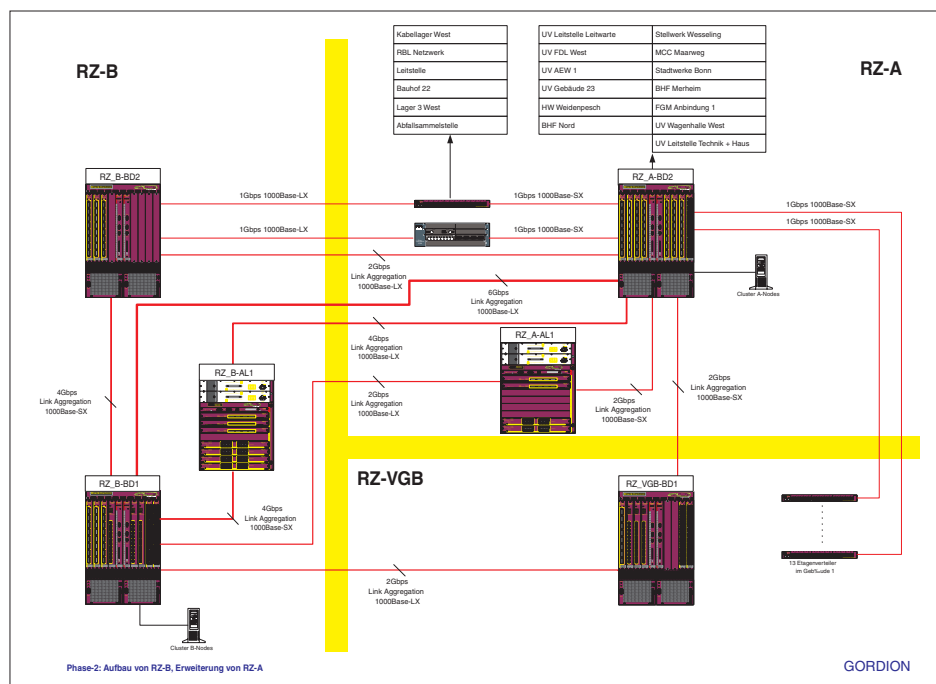


Bild 1. Aufbau von Rechenzentrum-B und die Erweiterung von RZ-A in Phase 2

mit geringerer Port-Dichte sowie eine kleinere Switch-Fabric (64 GBit/s).

Etagen-Switches, Tertiärebene und Cluster-Heartbeat:

Summit48si, Summit200: Switched Fast Ethernet (48 x 10/100Base-TX) mit 2 x GBIC-based 1000BaseX-Uplink. Der Summit48si verfügt über dieselbe Switching-Firmware wie der Black Diamond und das Alpine-System. Das Gerät unterstützt ebenfalls auf allen Ports Wirespeed Layer-3 und besitzt eine non-blocking Switch Fabric. Der Summit200-48 unterstützt ebenfalls 48x 10/100Base-TX sowie 2x 1000BaseT beziehungsweise alternativ 2x mini-GBIC-based 1000BaseX-Uplinks. Das Feature-Set ist eine Teilmenge des „i“-Feature-Sets vom Black Diamond, Alpine und Summit-„i“, die Switch-Fabric leistet 13,6 GBit/s non-blocking.

WAN-Router: Cisco 7206 VXR: Zur Anbindung von Außenstellen per 2-MBit/s-Fest- und 64- bis 128-kBit/s-Wählverbindung.

Rahmenbedingungen des Netzwerkdesigns

Das entwickelte Netzwerkdesign sowie die Vorgehensweise der Umsetzung basiert auf

folgenden Zielen respektive Rahmenbedingungen:

- sanfte Migration ohne radikale Änderungen (Minimierung der Downtime),
- redundantes Netzwerk-Redesign ohne Single Point of Failure im Core-Bereich,
- TCP/IP als alleiniges Netzwerkprotokoll,
- Integration der vorhandenen Netzwerksysteme (Investitionsschutz),
- automatische Netzrekonfiguration bei Ausfällen,
- redundante Serveranbindungen,
- Realisation eines neuen IP-Adresskonzepts,
- Ablösung des historisch bedingten, flachen IP-Adresskonzepts

und Migration in eine neue Struktur, insbesondere mit getrennten IP-Adressbereichen für Server und Unterverteiler. Der Beginn der Planungen für das Redesign lag im April 2003. Die Realisation begann im November 2003. Zuerst wurde das Rechenzentrum-A fertiggestellt (Februar 2004) und anschließend das Rechenzentrum-B (März 2004). Das Westforum ist Ende Mai 2004 integriert worden. Der Aufbau der zwei neuen Rechenzentren, der Umzug des vorhandenen RZs sowie die Integration des neuen Verwaltungsgebäudes (Westforum) wurde in vier Phasen un-

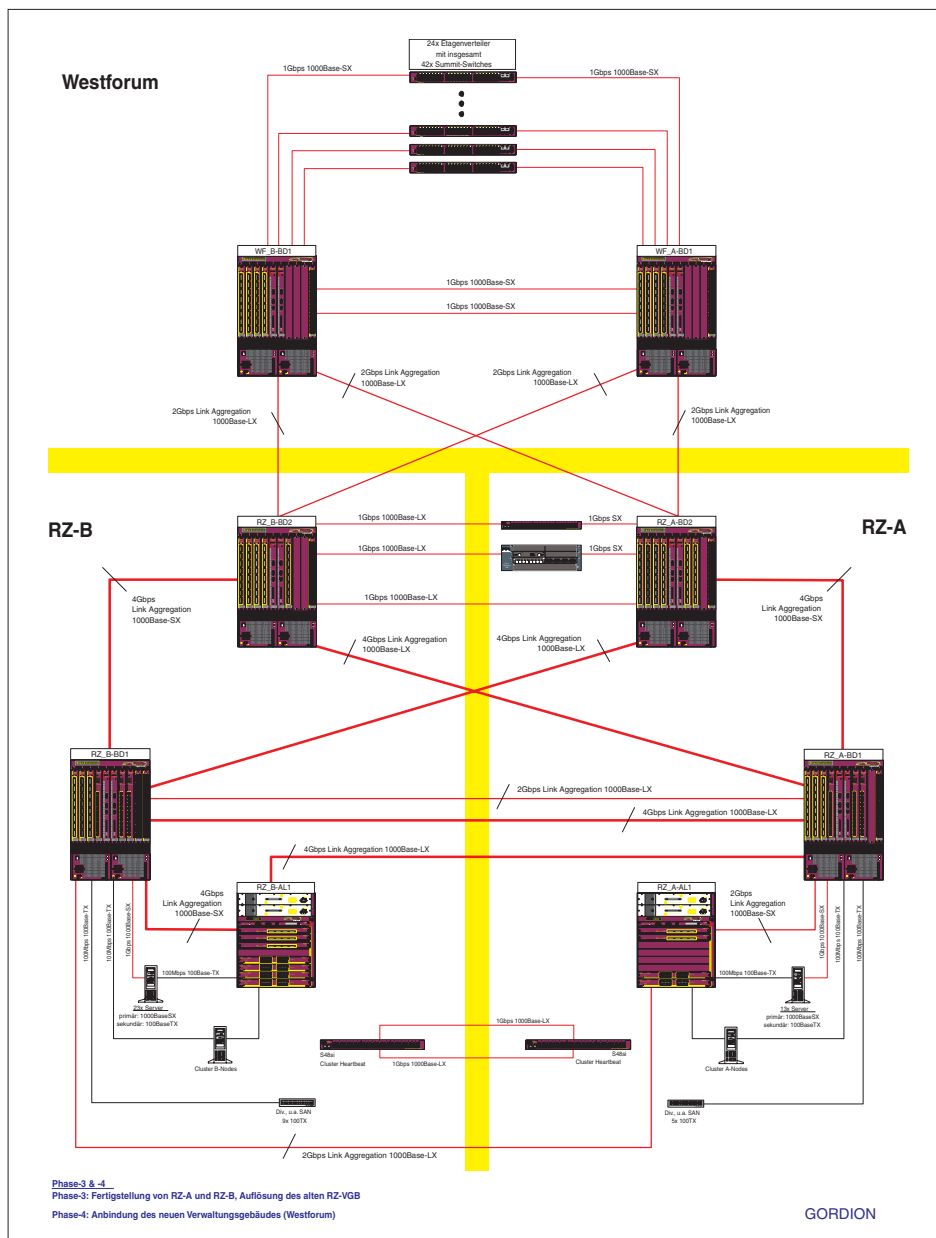


Bild2. Fertigstellung von RZ-A und RZ-B sowie die Auflösung des alten RZs-VGB in Phase 3

terteilt, welche nachfolgend skizziert sind. Diese vier Phasen waren durch die Zielsetzung einer kostenoptimierten Umsetzung bestimmt. Hierzu zählt primär der Verzicht auf „Überbrückungskomponenten“. Also vor allem Switch-Systeme, die „nur“ den Umzug erleichtern, jedoch im Endausbau nicht mehr benötigt werden.

Im ursprünglichen Rechenzentrum des alten Verwaltungsgebäudes „VGB“ standen zwei redundante Core-Switches (Extreme Networks, Black Diamond 6808) an denen Etagen und Server angebunden waren. Zudem war noch ein Xylan-Omni-S/R-Switch im Einsatz, der zur Anbin-

dung der weiteren KVB-Infrastruktur diente.

Phase 1: Aufbau von Rechenzentrum-A

Der Aufbau von RZ-A begann mit der Hinzunahme eines neuen Core-Switches (Black Diamond 6808), zur Aufnahme der vorhandenen Etagenverbindungen im alten Gebäude „VGB“ sowie zur Anbindung der Server. Die Außenstellen wurden unter anderem mittels eines Cisco-WAN-Routers 7206VXR an das neue RZ-A angebunden (Fast Ethernet TX, 64 kBit-/2MBit-Verbin-

dung). Weitere Außenstellen sind per Fast Ethernet FX über einen Summit48si (und Medienkonverter) angebunden. Die Gigabit-Anbindung einiger Außenstellen erfolgt über den Switch BD2 im RZ-A.

Innerhalb von Phase 1 wurden die notwendigen VLANs neu auf dem Black Diamond in RZ-A aufgesetzt (RZ_A-BD2). Des Weiteren sind die vorhandenen Server-VLANs zu diesem Switch verlängert worden. Damit war die Voraussetzung für die Anbindung des neuen RZ-A geschaffen.

Zur Steuerung dieser redundanten Topologie werden verschiedene Mechanismen eingesetzt. Hierbei ergaben sich folgende Anforderungen:

Erstens: die zwingende Unterdrückung von Netzwerkschleifen zur Verhinderung von Broadcast Storms (eine Schleife ermöglicht im Gigabit Ethernet das Kreisen von mehreren 10.000 Broadcast-Frames pro Sekunde und kann so die Netzwerkkommunikation blockieren).

Zweitens: die Bereitstellung eines Default-Gateways für die angeschlossenen Client-/Server-PCs, inklusive redundantem Gateway. Die redundante Instanz wird auf einem zweiten Switch konfiguriert und ermöglicht die Ausfallsicherheit beziehungsweise die Erreichbarkeit des Gateways im Falle einer Störung des Primär-(Layer-3-)Switches. Zwingend zu beachten ist jedoch, dass maximal eine Instanz gleichzeitig aktiv ist. Andernfalls wären zwei Routing-Interfaces mit der gleichen IP-Adresse aktiv und somit Netzstörungen vorprogrammiert. Die Steuerung der redundanten Instanz sowie die Unterdrückung von Netzwerkschleifen sind durch ein geeignetes Protokoll geregelt. Im Netzdesign der KVB übernimmt diese Aufgabe das Extreme Standby Router Protocol (ESRP).

Zur Bekanntgabe der aktiven Router-Interfaces in einer redundanten Konfiguration sind statische Routing-Einträge nicht geeignet. Zur dynamischen Änderung der Routing-Tabellen bei einer Störung ist ein dynamisches Routing-Protokoll notwendig. Im Netzwerk der KVB ist hierzu das OSPF-Protokoll im Einsatz (Open Shortest Path First).

Phase 2: Aufbau von Rechenzentrum-B

In Phase 2 wurde das neue Rechenzentrum-B mit zwei Black Diamond 6808 sowie einem Alpine 3808 bestückt. Zudem erhielt Rechenzentrum-A einen zusätzlichen Alpine 3808. Die Alpines dienen hierbei zur Aufnahme von redundanten Serveranschlüssen (10/100/1000 MBit/s). Insgesamt sind 75 Server im Einsatz. Mit Phase-2 wurde die komplette Räumung von VGB-Rechenzentrum vorbereitet (Bild 1). Die Server blieben zunächst noch am Core-Switch „RZ-VGB BD1“ im alten Rechenzentrum angeschlossen. Während der Phase 2 wurde dann die Layer-3-Forwarding-Funktion vom alten RZ auf die neuen Rechenzentren A und B übertragen (Rollenumkehr). Zu Beginn der Phase-2 wurden die bestehenden Server-VLANs auf den Switches RZ-B_BD1 und RZ-A_BD2 für die Übernahme der Default-Gateway-Funktion eingerichtet. Der Switch RZ-VGB_BD1 diente anschließend nur noch zur Übertragung der Serverdaten zum jeweiligen Switch mit aktivem Interface (Default-Gateway). Im Hinblick auf die IP-Adressänderung wurden zwei neue Server-VLANs eingerichtet. Beide VLANs sind über zwei Switches pro RZ (Black Diamond und Alpine) erreichbar. Damit ist ein redundanter Netzanschluss der Server möglich. Zudem wurden in Phase 2 redundante Anschlüsse für die Anbindung der Außenstellen und Unterverteiler auf RZ-B_BD2 erstellt. Abschließend sind die Server aus dem alten RZ-VGB in die beiden neuen RZs migriert worden.

Phase 3: Fertigstellung von RZ-A und RZ-B, Auflösung des alten RZ

Der noch vorhandene Black Diamond 6808 im RZ-VGB wird in Rechenzentrum-A überführt und das RZ-VGB aufgelöst. Mithilfe des nun „freien“ Black Diamond wird eine symmetrische Konfiguration von RZ-A und RZ-B realisiert (Bild 3 unten). Die BD1-Switches in RZ-A und RZ-B sind die primären Server-Switches. Sie übernehmen das Layer-3-Switching (IP-Routing) für die insgesamt vier Server-VLANs.

Alle Server-VLANs können in beiden RZs genutzt werden. Von den vier Server-VLANs sind zwei VLANs so genannte Legacy-Netze, also Netze mit alten IP-Netzadressen. Diese werden nach der Umstellung der entsprechenden Server auf die neue IP-Adressstruktur aufgelöst. Die Alpine-Switches bieten die redundanten Netzanschlüsse für die Server. Jeweils vier Server-VLANs sind in beiden RZs vorhanden und können parallel genutzt werden. Jeweils zwei Server-VLANs pro RZ sind aktiv, es wurde also eine 2:2-Lastverteilung vorgenommen. Das heißt, zwei VLANs arbeiten mit aktiver Layer-2/Layer-3-Forwarding-Funktion (Master), und zwei VLANs arbeiten nur mit Layer-2-Forwarding-Funktion (Slave) in Richtung des anderen RZs. Netzwerkschleifen werden durch die selektive Forwarding-Funktion des Slaves unterdrückt.

Die beiden BD2-Switches in RZ-A und RZ-B bilden die Standortverteiler-Switches. Die Gebäude des KVB-Geländes sowie die Außenstellen werden über diese Switches redundant angeschlossen. Die Server-Switches und Standortverteiler sind über geroutete IP-Verbindungen mit jeweils viermal 1-GbE per Link-Aggregation verbunden. Diese Vorgehensweise bietet Ausfallsicherheit sowohl für Port-Ausfälle als auch bei Störungen von Switches oder Modulen: Solange ein BD-1 und ein BD-2 funktioniert, ist die Netzverfügbarkeit gesichert beziehungsweise kann die Verfügbarkeit bei einem Ausfall des primären Standort-Switches (BD2-RZA) durch Umpatchen bestimmter Verbindungen wieder hergestellt werden.

Phase 4: Integration des neuen Verwaltungsgebäudes (Westforum)

Im Westforum dienen zwei weitere Black Diamonds zur redundanten Anbindung von insgesamt 24 Etagenverteilern, bestehend aus 42 Switches vom Typ „Summit48si und 200“. Phase 4 diente zur Integration des neuen Verwaltungsgebäudes in das KVB-Netz. Hierzu wurden zwei Black Diamond-Switches als zentrale Switches, zur Anbindung der insgesamt 24 Unterverteiler im Westforum definiert. Die BDs sind redundant konfiguriert und in zwei (verteil-

ten) Räumlichkeiten installiert. Jeder der Unterverteiler besitzt jeweils eine LWL-Verbindung zu beiden Core-Switches. Für jeden Unterverteiler ist ein VLAN mit eindeutiger IP-Adresse konfiguriert. Das Extreme Standby Router Protocol (ESRP) steuert auch in diesem Szenario die Redundanzen. Zur optimalen Ausnutzung der beiden zentralen Switches wird eine Hälfte der Unterverteiler über den Switch WF-A_BD1 versorgt, die andere nutzt den zweiten Switch WF-B_BD1.

Bei Ausfall eines Core-Switches, eines Moduls oder einer Verbindung (Port) wird automatisch die redundante Instanz auf dem anderen Black Diamond aktiviert. Netzwerktechnisch ist das Westforum über mehrere geroutete Verbindungen an die Standort-Switches des KVB-Netzes (jeweils BD2 in RZ-A und B) angeschlossen. Jede Verbindung besteht aus zwei GbE-Ports, die mittels Link-Aggregation zusammengeschaltet sind. Die Zentral-Switches im Westforum besitzen Anschlüsse zu beiden Standort-Switches. Somit wird die Möglichkeit eines Totalausfalls der Netzanbindung im Westforum auf ein Minimum reduziert.

Fazit

“Das Vier-Phasen-Konzept für die KVB hat sich bewährt“, resümiert Gerd Brabender, IT-Bereichsleiter und Projektverantwortlicher der KVB. Die avisierte sanfte Migration in zwei neue Rechenzentren sowie die Integration eines neuen Verwaltungsgebäudes wurden mit minimaler Downtime erreicht. Der Netzwerk-Core enthält keinen Single Point of Failure, Störungen werden durch eine automatische Netzrekonfiguration entschärft. TCP/IP ist nun alleiniges Netzwerkprotokoll, und zudem wurde die historisch bedingte, flache IP-Adressstruktur abgelöst.

Oliver Lindlar/mw

Oliver Lindlar ist Dipl.-Wirtschaftsinformatiker und arbeitet im Projektmanagement bei Gordion.

■ Info: Gordion Network Consulting
Tel.: 02241/49040
Web: www.gordion.de