INSTITUT FÜR INFORMATIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

Diplomarbeit

Mandantenfähiges Netz- und Sicherheitskonzept für den Betrieb virtueller Infrastrukturen am Beispiel von VMware im Münchner Wissenschaftsnetz

Bernhard Schmidt

INSTITUT FÜR INFORMATIK
DER LUDWIG–MAXIMILIANS–UNIVERSITÄT MÜNCHEN

Diplomarbeit

Mandantenfähiges Netz- und Sicherheitskonzept für den Betrieb virtueller Infrastrukturen am Beispiel von VMware im Münchner Wissenschaftsnetz

Bernhard Schmidt

Aufgabensteller: PD Dr. Helmut Reiser

Betreuer: PD Dr. Wolfgang Hommel
Bastian Kemmler

Abgabetermin: 25. April 2013

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 25. April 2013

. . . . . .

(Unterschrift des Kandidaten)

Die Virtualisierung ist aus den heutigen IT-Umgebungen nicht mehr wegzudenken. Richtig eingesetzt ermöglicht sie zum Einen beim Betrieb eigener Dienste meistens hohe Effizienzgewinne und Kosteneinsparungen bei gleichzeitigen Vereinfachungen, und zum Anderen das sehr schnelle und einfache Bereitstellen von Rechenleistung für Kunden. Sie müssen nicht mehr aufwändig physische Server auswählen, bestellen, aufbauen, installieren, betreiben, abbauen und verschrotten, sondern können bei einem Dienstleister ihrer Wahl mit wenigen Mausclicks je nach Bedarf virtuelle Server bestellen. Diese Möglichkeiten werden unter dem Begriff Infrastructure-as-a-Service (IaaS) zusammengefasst.

Das Zusammenfassen vieler Dienste in einer Infrastruktur bringt jedoch auch Sicherheitsprobleme mit sich, sowohl durch die zusätzliche Komplexität des Hypervisors als auch durch den unbedachten Einsatz gewohnter Konfigurationen, die den Anforderungen, die in einer mandantenfähigen Infrastruktur entstehen, nicht ausreichend gewachsen sind.

Diese Arbeit befasst sich mit den Sicherheitsproblemen, die durch den Einsatz von Virtualisierungsumgebungen entstehen. Diese sind auch im eigenen Einsatz relevant, jedoch beim Bereitstellen von IaaS-Diensten an Kunden deutlich ausgeprägter. Hierbei müssen spezifische Abtrennungen zwischen den einzelnen Mandanten beachtet sowie auf deren spezielle Anforderungen eingegangen werden, um eine Attraktivität der Plattform für Kundengruppen sicherzustellen. Gleichzeitig muss jedoch ausgeschlossen werden, dass durch das Hosting unterschiedlicher Mandanten auf der gleichen Plattform neue Sicherheitslücken entstehen.

Nach einer Beschreibung des am Leibniz-Rechenzentrum vorgefundenen Szenarios enthält die Arbeit eine Analyse von möglichen Angriffen. Der Schwerpunkt liegt hierbei auf den netzbasierten Angriffen, da diese eine Basis für viele weitergehende Attacken bilden und eine vollständige Betrachtung sämtlicher Angriffsvektoren den Rahmen dieser Arbeit sprengen würde.

Bei der Evaluierung von verschiedenen Produkten und strukturellen Maßnahmen zur Problemlösung stellte sich heraus, dass keine Variante derzeit in gewachsenen virtuellen Umgebungen, bei denen Teile der Infrastruktur als vorhanden angesehen werden müssen, ausreichend Schutz bieten kann. Daher wird basierend auf mehreren Standardprodukten und -protokollen wie VXLAN und Proxy-ARP eine neuartige Struktur skizziert, welche nicht nur die aufgezählten Sicherheitsprobleme lösen kann, sondern den Kunden auch zusätzliche Dienste bereitstellen kann. Ein weiterer Vorteil ist, dass nur wenige nicht-invasive Änderungen an der Infrastruktur nötig sind.

Eine Proof-of-Concept-Implementierung der Lösung, basierend auf der im Linux-Kernel integrierten VXLAN-Funktionalität und Netfilter, zeigt, dass der vorgeschlagene Weg auch in der Praxis funktioniert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Seit dem Aufkommen von Virtualisierungstechnologien wie VMware und Xen boomt das Geschäft mit dem Verkauf von Rechenzeit in Form von virtuellen Servern. Neben den Firmen, die Virtualisierung für den Betrieb ihrer eigenen Infrastruktur verwenden sind Anbieter für Clouddienstleistungen wie Amazon EC2 und Microsoft Azure aus dem Boden geschossen, die virtuelle Server in Millionenstückzahlen für ihre Kunden betreiben. Diese Angebote gewinnen aufgrund der deutlich besseren Flexibilität mittlerweile auch Kunden, die früher mit Angeboten wie Shared Webhosting zufrieden waren.

Es wird jedoch oft übersehen, dass der Betrieb eines kompletten Serversystems deutlich höhere Anforderungen an die Administration stellt. So dürfen sowohl bei der Konfiguration als auch bei der täglichen Pflege die Sicherheitsimplikationen nicht aus den Augen gelassen werden. Das Bewusstsein für diese Problematik fehlt vielen Kunden, weswegen Einbrüche in schlecht abgesicherte Serversysteme zur Tagesordnung gehören. Daraus folgt, dass sich der Anbieter nicht nur regelmäßig im Umgang mit kompromittierten Kundenmaschinen bewähren muss, sondern gleichzeitig sowohl präventiv als auch reaktiv eine Ausbreitung einer Kompromittierung auf seine anderen Kunden verhindern muss.

Die vorliegende Arbeit behandelt exemplarisch ein Netz- und Sicherheitskonzept am Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, welches Server-Hosting auf einer virtuellen Infrastruktur für seine Kunden anbietet. Die vorgestellten Probleme und Lösungen gelten jedoch auch gleichartig für andere Anbieter von Hostinglösungen.

1.1 Umgebung

Das Leibniz-Rechenzentrum betreibt eine große Infrastruktur zur Virtualisierung von Servern auf der Basis von VMware. Nach der zuerst erfolgten Migration eigener physischer Server aus Gründen der Redundanz und der Energieersparnis („Green-IT“) wird seit einigen Jahren verstärkt dazu übergangen, den Kunden diese Infrastruktur zum Erfüllen ihrer eigenen Aufgaben anzubieten (Infrastructure-as-a-Service). Der Kunde kann dabei zwischen einem zentral durch die Administratoren des LRZ gewarteten System („attended hosting“ - Windows 2008 R2 oder SLES 11) oder der Bereitstellung der rohen Rechenkapazität zum Betrieb einer eigenen Installation („unattended hosting“) wählen. Beiden Varianten ist gemeinsam, dass der Kunde administrative (root) Rechte auf seinem Server erhält, da auch die attended Varianten nur eine sichere Initialkonfiguration und automatische Sicherheitsupdates von installierten Distributionspaketen beinhalten. Die Installation und Konfiguration der Anwendung unterliegt immer dem Benutzer selbst.

Diese Aktivitäten haben positiven Nutzen für alle Beteiligten, da sich mit der Rezentralisierung von Diensten Synergieeffekte einstellen und der Kunde hohe Rechenleistungen für seine Anwendungen erwerben kann, ohne seine lokale Infrastruktur damit zu belasten (Strom, Klima, Netzanbindung). Diese Vorteile sind allerdings aus sicherheitstechnischer Sicht auch durchaus Nachteile, wenn viele Systeme mit guter Netzanbindung und hoher Leistung an zentraler Stelle zusammenstehen. Diese üben einen besonders hohen Reiz auf potentielle Angreifer aus und können ihrerseits wieder für weitergehende Angriffe verwendet werden.

1.2 Gefahr durch Angriffe

Das nachfolgend beschriebene Beispiel eines realen Sicherheitsvorfalls soll diese Gefahr verdeutlichen. Ein Kunde benutzt einen virtuellen Server mit Systemwartung („attended“) unter SLES 11 zum Betrieb seines Dienstes. Da das LRZ keine Unterstützung bei der Installation von Anwendungen geben kann und nicht

alle Applikationen im Lieferumfang von SLES enthalten sind, installierte er neben diversen anderen Softwarepaketen auch eine veraltete JBoss-Version. Durch eine in früheren Versionen von JBoss enthaltene Sicherheitslücke im vom Netz erreichbaren Management-Frontend „JMX-Console“ konnte der Angreifer eine JSP-basierte Shell-Backdoor installieren und einen Botnet-Client aus dem Netz nachladen und starten. Dieser Botnet-Client wurde erst eine Woche nach der Kompromittierung entdeckt, da er durch SSH-Scans auf Ziele außerhalb des Münchner Wissenschaftsnetzes im routinemäßigen Monitoring am Internetübergang auffiel.

Dieses Beispiel zeigt exemplarisch eine Vielzahl der Probleme, die beim Angebot einer Server-Plattform an externe Kunden auftreten können. Während in der Distribution enthaltene Softwarepakete während des Wartungszeitraums der Distribution Sicherheitsupdates erfahren und diese automatisch eingespielt werden können, müssen Kunden oft spezifische Programme oder Programmversionen von Hand einspielen, die in der Distribution nicht enthalten sind. Diese können von automatischen Tools nicht mehr erfasst werden und werden oft auch nicht mehr durch den Nutzer von Hand aktualisiert. Selbst wenn die Software an sich keine Sicherheitslücken aufweist, so muss sie oft durch den Benutzer konfiguriert werden. Auch dadurch können Sicherheitslücken (zum Beispiel schwache Passwörter oder unvollständige Zugriffslisten) entstehen.

Um diese Operationen am System durchführen zu können ist es im Allgemeinen notwendig, administrative Rechte am System zu erhalten. Diese Rechte erlauben bei Kompromittierung ein Aushebeln von Sicherheitsfunktionen wie Firewalls/Paketfiltern, TCP-Wrapper, Sicherheitsupdates, und periodischen Scans. In diesem Fall hätte allerdings beispielsweise auch eine Firewall keinen Effekt gezeigt, da der Einfallsvektor über den mit Absicht nach außen offenen Dienst erfolgte. Ein weiteres Problem ist, dass der Betreiber der Infrastruktur oft keine Information darüber erhält, welche Dienste durch das System des Kunden erbracht werden sollen. Dies macht es sehr schwer, durch eine Anomalieerkennungen von außen (zum Beispiel durch ein netzbasiertes IDS oder eine Verhaltensanalyse) eine Kompromittierung des Kundensystems zu erkennen. Derartige Analysen sind nur mit einer sehr hohen Erkennungsschwelle möglich (um zeitaufwändige Fehlalarme zu verhindern), was es einem Angreifer oft erlaubt der Erkennung zu entgehen. Nicht zuletzt muss bedacht werden, das die Kompromittierung einer Kundenmaschine unter Umständen die Kompromittierung weiterer Systeme erleichtert oder erst ermöglicht. So werden im Münchner Wissenschaftsnetz Systeme oft ohne weitere Sicherheitsmaßnahmen mit privaten IPv4-Adressen an das Netz gehängt und auf die implizite Firewall im NAT-Gateway (Secomat) vertraut. Dies schützt nur vor Angriffen von außen, schlägt aber fehl wenn Angriffe durch eine gehackte Maschine von innerhalb des Münchner Wissenschaftsnetzes ausgehen [LRZ 12e].

1.3 Aufgabenstellung und Ziel der Arbeit

Im Rahmen dieser Arbeit soll ein umfassendes Sicherheitskonzept für virtuelle Infrastrukturen entstehen, welches zwar primär auf die Bedingungen am LRZ zugeschnitten ist, jedoch zu großen Teilen auch in anderen Installationen verwendet werden kann. Nach einer umfassenden Analyse des aktuellen Zustands und der möglichen Angriffsvektoren werden Technologien vorgestellt, welche die benannten Sicherheitsprobleme beheben können. Hierbei kommen einerseits Maßnahmen zum Einsatz, die eine Kompromittierung verhindern oder zumindest erschweren – beispielhaft kann man hier das Patchmanagement oder Paketfilter nennen – andererseits allerdings auch Methoden zur Früherkennung und zur Isolierung der Kunden untereinander, um im Fall eines erfolgreichen Angriffs keine Probleme mit anderen Kunden zu bekommen. In diese Kategorie können verschiedene Trennungsmechanismen auf der Netzschicht fallen. Ziel ist es auch, nicht nur bekannte Angriffe zu verhindern, sondern durch strukturelle Änderungen auch bisher unbekannte Angriffe frühzeitig eindämmen zu können.

Netzbasierte Angriffe laufen im Gegensatz zu den im Allgemeinen nicht vorhersehbaren und schnell geschlossenen Sicherheitslücken in der Virtualisierungssoftware sehr häufig nach dem gleichen Muster ab. Diese werden außerdem gerne als Grundlage für weitere Angriffe innerhalb der selben Infrastruktur benutzt. Daher liegt der Fokus der Arbeit bei der Aufstellung und Implementierung eines umfassenden Sicherheitskonzepts gegen Angriffe in der Netzumgebung. Während jedoch im Hosting für den eigenen Bedarf diese Probleme noch durch hohe Standards in der Administration abgemildert werden können, ist in mandantenfähigen Umgebungen mit vielen verschiedenen Betreuern ein konsistentes Sicherheitsniveau nicht mehr zu gewährleisten. Dort wird daher das Verhindern eines Übergreifens von Kompromittierungen eine wichtige Grundvoraussetzung für einen sicheren Betrieb.

1.4 Struktur der Arbeit

Die vorliegende Arbeit gliedert sich grob in drei Bereiche. Nach einer Einführung in die aktuell betriebene Infrastruktur werden in Kapitel 2 neben den Schutzzielen auch die nicht-funktionalen Anforderungen an eine einsetzbare Lösung formuliert.

Im Kapitel 3 werden, mit einem Fokus auf netzbasierte Angriffe, denkbare Angriffsvektoren auf die Infrastruktur der Virtualisierungsplattform erläutert und mögliche Gegenmaßnahmen diskutiert. Aus diesen exemplarischen Angriffen ergeben sich in Kapitel 4 die funktionalen Anforderungen, gegen die einige denkbare Konzepte evaluiert werden.

In den seltensten Fällen können jedoch alle bekannten Sicherheitsmaßnahmen durchgeführt werden, da sich Konzepte zum Teil gegenseitig ausschließen oder in den verwendeten Komponenten nicht eingesetzt werden können. In Kapitel 5 wird daher aus den Mechanismen der vorangegangenen Kapitel eine Systemzusammenstellung vorgeschlagen, die im Rahmen der Möglichkeiten eine möglichst große Bandbreite an Fragestellungen löst und dabei tatsächlich praktisch einsetzbar ist. Im Folgenden wird diese Zusammenstellung nicht nur in Textform, sondern auch in einer prototypischen Implementierung dargestellt. Ein Überblick über weitere Vorschläge zur Erhöhung der Sicherheit, die nicht im Fokus dieser Arbeit liegen, ist ebenfalls Teil dieses Kapitels.

Ein Ausblick auf zukünftiges Verbesserungspotential in der Infrastruktur rundet die Arbeit in Kapitel 7 ab.

2 Szenario

„Das Leibniz-Rechenzentrum (LRZ) ist gemeinsames Rechenzentrum der Ludwig-Maximilians-Universität München, der Technischen Universität München sowie der Bayerischen Akademie der Wissenschaften; es bedient auch die Fachhochschule München und die Fachhochschule Weihenstephan. Zusätzlich betreibt das LRZ Hochleistungsrechensysteme für alle bayerischen Hochschulen, sowie einen Bundeshöchstleistungsrechner, der der wissenschaftlichen Forschung an den deutschen Hochschulen zur Verfügung steht.“ [LRZ 12d] Es bietet neben dem Münchner Wissenschaftsnetz (MWN), welches die meisten Gebäude der teilnehmenden Einrichtungen verbindet, eine große Anzahl von zentralen Diensten wie Mail, Webserver oder Dateiablagedienste. Diese stehen dem Kundenkreis zum Teil unentgeldlich, zum Teil in kostenpflichtigen Angeboten zur Verfügung [LRZ 12c].

2.1 Ausgangszustand

Der Ausgangszustand der in dieser Arbeit primär behandelte Infrastruktur des Leibniz-Rechenzentrums wird in den folgenden Unterpunkten detailliert dargestellt. Zur Vermeidung von Missverständnissen im Virtualisierungsbereich wird die folgende Nomenklatur verwendet:

Host
: Physischer Server, Plattform für den Hypervisor

Container
: Hypervisor-Seite eines Gasts (Konfiguration der virtuellen Maschine)

Virtuelle Maschine
: Betriebssystem-Seite eines Gasts

Gast
: Oberbegriff für Container und Virtuelle Maschine

Insbesondere muss hier auf die Unterscheidung zwischen einem physischen (VMware-)Host und dem virtuellen Host geachtet werden, sowie auf den Unterschied zwischen der Hypervisor-seitigen Konfiguration des Containers (beispielsweise RAM oder der Anbindung einer virtuellen Netzwerkkarte an einen virtuellen Switch) und der Gast-seitigen Konfiguration (beispielsweise der IP-Adresse oder der lokalen Firewall) geachtet werden.

2.1.1 Serverhardware (Hosts)

Die Virtualisierungsinfrastruktur des LRZ besteht aus fünf Bladecentern („Enclosures“) der Firma Hewlett-Packard vom Typ BladeSystem c7000 [HP 12a], die in zwei unterschiedlichen Räumen im Rechenzentrum Garching („Rechnerwürfel“) aufgebaut sind. Diese Räume liegen auch in unterschiedlichen Brandabschnitten, die die vollständige Abschaltung der Stromversorgung im Brandfall regeln. Dadurch stehen auch im Katastrophenfall noch Ressourcen zur Verfügung, um den Betrieb der kritischen virtuellen Systeme aufrecht erhalten zu können. Die Bladecenter enthalten jeweils sechs Netzteile, die auf drei Phasen verteilt sind. Alle Phasen werden als sogenannter EV4-Strom von einer ganzen Kette unterbrechungsfreier Stromversorgungen gespeist. EV4 ist die Bezeichnung der höchsten Stromklasse im Rechenzentrum des LRZ. Diese beinhalten als primäre Absicherung einen statischen Part, in dem die Spannungsversorgung für wenige Minuten durch Batterien aufrecht erhalten werden kann. Diese werden wiederum von einer dynamischen, auf kinetischer Rotationsenergie basierenden Stromversorgung gestützt, um Schwankungen im Stromnetz ausgleichen zu können. Bei einem Komplettausfall der externen Stromversorgung springt ein Notstromdiesel ein.

Die Bladecenter bieten jeweils Platz für 16 Einschübe und sind am LRZ voll mit Einschüben vom Typ Hewlett Packard BL490c G6 belegt. Jedes dieser Blades verfügt über zwei Prozessoren (Intel Xeon E5540, Nehalem-Architektur) mit jeweils vier physischen Rechenkernen, sowie über 96GB Arbeitsspeicher. Durch Hyperthreading werden dem Virtualisierer daher 16 Prozessoren pro Blade angeboten. Die Einschübe verfügen außerdem über jeweils zwei 10Gbit/s-Anbindungen an die Midplane des Bladecenters und verbinden sich dort an zwei sogenannte Flex-10 Module [HP 12c], die für die Anbindung an die Außenwelt sorgen. Das Flex-10 Modul läuft dabei im sogenannten „tunneled VLAN“ Modus, bei dem alle 4096 in IEEE 802.11q definierten VLANs verwendet werden können.

2.1.2 Netzinfrastruktur

Die Flex-10 Module der Bladecenter sind mit jeweils 2*10 Gbit/s-Ethernet an zwei modulare Switches vom Typ Hewlett-Packard Procurve 5412zl („VMware-Switches“) angebunden und nutzen dabei die Technik LACP (IEEE 802.3ad) zur Bündelung der Bandbreiten. Diese Funktionalität wird direkt in den Flex-10 Modulen erbracht, da der in VMware integrierte Switch („distributed vSwitch“) auch in der aktuellen Version noch keine Bündelung von Links zur Bandbreitenerhöhung unterstützt.

Die VMware-Switches besitzen zu jeweils einem Switch des Typs Hewlett-Packard Procurve 8212zl („Zentralswitch“) ebenfalls einen 2*10Gbit/s-Trunk nach dem LACP-Standard. Die Zentralswitches und die VMware-Switches sind untereinander ebenfalls mit 20Gbit/s verbunden.

Die bisher erwähnte Switchinfrastruktur arbeitet rein auf der Schicht 2 des ISO/OSI-Referenzmodells und verwendet bis auf Filter gegen unerwünschte Spanning-Tree-Partner (*BPDU Filter*) keine Sicherheitsmechanismen.

Für das Routing auf Schicht 3 sowie als zentraler Verteilerpunkt auf Schicht 2 kommen zwei zu einem *Virtual Switching System (VSS)* zusammengeschalteten *Cisco Catalyst 6509* mit *Supervisor Engine 720 (PFC3CXL)* zum Einsatz. Dabei agieren zwei getrennte Chassis, die ebenfalls wieder in unterschiedlichen Brandabschnitten lokalisiert sind, in allen Belangen wie ein System mit der doppelten Anzahl an Linecards [CiscoVSS]. Da alle relevanten Switches an beide Hälften der VSS angeschlossen sind kann selbst beim vollständigen Ausfall einer Hälfte, zum Beispiel beim Ausbruch eines Feuers, der Betrieb aufrecht erhalten werden. Einzig die verfügbare Bandbreite ist hier eingeschränkt. Es ist geplant, dieses System Anfang 2013 durch ein redundantes System basierend auf zwei Cisco Nexus 7010 im vPC-Modus zu ersetzen, welche ähnliche Fähigkeiten haben.

Zur Netzsegmentierung werden VLANs nach dem IEEE 802.1q-Standard eingesetzt. Diese werden historisch dazu verwendet, Netze mit unterschiedlichen Kommunikationsprofilen und Sicherheitsanforderungen zu trennen. So gibt es existierende, mit der bestehenden physischen Infrastruktur geteilte VLANs für vom Internet erreichbare Server, für aus dem MWN erreichbare Server oder aus dem LRZ erreichbare Server. Eine Trennung nach internen Nutzergruppen (Abteilungen oder Gruppen) findet zunehmend bei Neueinrichtungen statt. Bei virtuellen Servern für externe Kunden kommen Sammel-VLANs für eine Trennung von *attended* und *unattended* VMs sowie eine Trennung zwischen weltweit erreichbaren öffentlichen und privaten, MWN-weit erreichbaren IPv4-Adressen. Damit stehen folgende VLANs zur Verfügung:

- VLAN 1780 – LRZ-intern, VMware-Server, Test, LRZ-weit
- VLAN 1781 – LRZ-intern, VMware-Server, Test, MWN-weit
- VLAN 1782 – LRZ-intern, VMware-Server, Test, weltweit
- VLAN 1783 – LRZ-intern, VMware-Server, Produktion, LRZ-weit
- VLAN 1784 – LRZ-intern, VMware-Server, Produktion, MWN-weit
- VLAN 1785 – LRZ-intern, VMware-Server, Produktion, weltweit

- VLAN 1786 – externe Kunden, VMware-Server, attended MWN-weit
- VLAN 1787 – externe Kunden, VMware-Server, attended, weltweit
- VLAN 1788 – externe Kunden, VMware-Server, unattended, MWN-weit

- VLAN 1789 – externe Kunden, VMware-Server, unattended, weltweit

Des weiteren werden administrative Funktionen der VMware-Infrastruktur ebenfalls über seperate VLANs bedient. Hierzu gehört die Anbindung an das Storage-Netz, das Management der VMware-Hosts und das Netz für die Verschiebung von virtuellen Maschinen (vMotion).

Zur Schleifenvermeidung wird das **Multiple Spanning Tree Protocol (MSTP)** nach IEEE 802.1s verwendet, wobei nur eine MSTP-Instanz mit allen VLANs zum Einsatz kommt. Daher ist die Spanning-Tree-Topologie für alle VLANs gleich. Im Normalfall fungiert das Cisco VSS-Pärchen dabei als Root-Bridge, so dass sowohl der Trunk zwischen den beiden Zentralswitches als auch der Trunk zwischen den beiden VMware-Switches durch Spanning-Tree abgeschaltet wird (*blocking*).

Sicherheitsmechanismen

In den Cisco-Routern kommen Anti-Spoofing-Filter (Unicast Reverse Path Filter, BCP 38) zum Einsatz, um ein Fälschen der Absenderadresse zu verhindern. Dieser Schutz agiert jedoch prinzipbedingt nur auf VLAN-Ebene, es ist weiterhin möglich Adressen von Systemen im gleichen Subnetz zu fälschen. Außerdem werden auf einigen Subnetzen statische oder reflexive (d.h. selbstlernende, ähnlich dem Prinzip des stateful Packetfilters agierende) Filter eingesetzt, die jedoch vom Administrator der im Netz gehosteten Server nicht selbst verändert werden können. Ansonsten kommen keine präventiven Sicherheitsmechanismen zum Einsatz.

Einen optionalen zusätzlichen Sicherheitsmechanismus stellt das Produkt der „virtuelle Firewalls“ auf Basis einer Cisco ASA 5580 Appliance dar. Diese dient als Subnetzfirewall zwischen dem MWN und dem Netz des Kunden, welches dafür als seperates VLAN ausgeführt sein muss. Die Kunden erhalten dabei einen für sie dedizierten virtuellen Kontext auf der physischen Firewall, der nur den Verkehr ihres VLANs behandelt. Sie können darauf Regelsätze selbst konfigurieren und auch andere Konfigurationsänderungen durchführen.

2.1.3 Hypervisor

Als Virtualisierungsinfrastruktur kommt VMware ESXi in der aktuellen Version 5.0 zum Einsatz. Dieser stellt gemäß der Definition von Robert P. Goldberg [Gold 72] einen Type-1-Hypervisor (*Bare-Metal-Hypervisor*) dar, bei dem der Hypervisor ohne dazwischenliegendes Betriebssystem direkt auf der Hardware aufsetzt. Mit der Ausnahme von KVM, welches als Prozess auf einem Linux-Kernel aufsetzt und diesem die Hardware-Ansteuerung überlässt, entsprechen alle aktuellen Virtualisierungslösungen in großen Umgebungen diesem Typ.

Die 80 Hosts sind zu verschiedenen Clustern zusammengefasst, die von einem dedizierten Managementserver (vCenter Server) verwaltet werden. Die genaue Anzahl und Aufteilung ändert sich bei Bedarf, es existieren jedoch mindestens dedizierte Cluster für den Produktivbetrieb von LRZ-eigenen VMs und Kunden-VMs (jeweils in der Ausprägung *attended* und *unattended*, siehe 2.1.5), einige Test-Cluster sowie ein dediziertes System für den Betrieb der Exchange-Server. Außerdem wird für ein großes externes Projekt ein dedizierter Cluster betrieben, um die sich aus diesem Projekt ergebenden Einschränkungen im täglichen Betriebsablauf auf die dort benutzte Infrastruktur zu beschränken.

Neben den manuellen oder halbautomatischen Konfigurationsarbeiten, die im Allgemeinen durch die Windows-Software vSphere Client vorgenommen werden, kümmert sich der vCenter Server automatisch nach vorgegebenen Regelsätzen um die Hochverfügbarkeit und Lastverteilung der virtuellen Maschinen, indem diese zwischen den Hosts verschoben werden (vMotion, vMotion DRS). Sofern ein Host ungeplant ausfällt werden dabei beendete VMs automatisch auf einem anderen Host wieder gestartet (vMotion HA).

Zusätzlich gibt es eine Test-Infrastruktur basierend auf älteren Rack-Servern des Typs Sun Fire X4150, in dem größere Änderungen vorab getestet werden können. Auf dieser Infrastruktur kommen mehrere Cluster mit zum Teil dedizierten vSphere Servern zum Einsatz.

Innerhalb der Cluster sind die virtuellen Maschinen in einem Baum organisiert, der die Organisationsstruktur (Abteilungen, Gruppen und Projekte) des LRZ wiederspiegelt. Externe Kunden müssen Projektverträge für das Hosting von virtuellen Maschinen abschließen. Alle Container eines Projekts sind in einem Ordner gebündelt.

Berechtigungen werden überwiegend auf Ordner (und damit alle darin enthaltenen virtuellen Maschinen), in Ausnahmefällen direkt auf die einzelnen Maschinen vergeben. Die normalen Administratoren der VMs erhalten die Berechtigungen eines so genannten Users, welche die folgenden Funktionen umfasst.

- Ein-/Ausschalten, Reset
- Verbinden auf die virtuelle Konsole

LRZ-interne Benutzer und ausgewählte Kunden erhalten zusätzlich die Berechtigung, Snapshots ihrer Container anzulegen, einzuspielen oder zu löschen. Sie erhalten jedoch keine Rechte, die Konfiguration des Containers zu verändern.

Ein Zugriff auf die internen Netze des Virtualisierers, insbesondere auf das Managementnetz, ist über eine Firewall abgesichert und nur von bestimmten Systemen aus möglich. Da einige Operationen wie das Verbinden auf die virtuelle Konsole und das Mounten von virtuellen Datenträgern eine direkte Verbindung vom Client zum VMware-Host benötigen [VMsec], sind hierfür jedoch Ausnahmen definiert, die zum Teil den Zugriff auf bestimmte Ports für Teile des MWN erlauben.

2.1.4 Storage

Die Datenhaltung für die Virtualisierungsinfrastruktur obliegt mehreren Filern des Herstellers NetApp. Ein sogenanntes Metrocluster, ein redundantes System aus wiederum auf zwei Brandabschnitte aufteilten Filerköpfen vom Typ FAS3170 mit dazugehörigen Festplattenshelfs, bildet das Rückgrat für die Speicherung von produktiven VMs. Zusätzlich existiert ein zweites NetApp-System zum Betrieb von Test- und Staging-VMs. Zur Anbindung an VMware wird das NFS-Protokoll verwendet. Alle betriebenen VMware-Hosts sind in der Lage, auf alle bereitgestellten Datastores (NFS-Mounts) zuzugreifen. Eine weitergehende Limitierung findet nicht statt.

Da sich die NetApp-Filer in einem dedizierten VLAN befinden und nur von den VMware-Hosts erreicht werden können, sind sie vor Angriffen über das Netz vergleichsweise gut geschützt. Sie werden daher im Rahmen dieses Sicherheitskonzepts nur am Rande betrachtet. Es ist jedoch nicht ausgeschlossen, über einen Exploit Zugriff auf den Hypervisor, durch diesen auf den Datastore und damit auf die unverschlüsselten Festplattenabbilder anderer Container zuzugreifen [VMDK-Vulnerability]. Derartige Sicherheitslücken lassen sich nicht vermeiden und nur durch rigerose Einschränkung von Zugriffsrechten und eine weitgehende Trennung von Nutzerdaten entschärfen.

2.1.5 Gast-Betriebssysteme

Den internen und externen Kunden stehen mehrere Betriebssysteme zur Auswahl.

Attended Hosting - SuSE Linux

Die am häufigsten nachgefragte Version eines virtuellen Servers ist ein SuSE Linux Enterprise Server (SLES). Zum Zeitpunkt dieser Arbeit war die Version 11 Servicepack 2 aktuell. Diese auch im LRZ verwendete Serverdistriubtion bietet von Haus aus einen bis zu zehnjährigen Supportzeitraum durch den Hersteller, während der Sicherheitsaktualisierungen und, im begrenzten Maße, neue Funktionen bereitgestellt werden. Zusätzlich werden durch die im LRZ eingesetzten Managementscripten initiale Sicherheitsparameter gesetzt und eine kontinuierliche Überwachung des Servers sichergestellt.

Die folgenden Funktionalitäten sind derzeit in einer Attended SLES-Installation im Auslieferungszustand aktiviert:

- TCP-Wrapper (*/etc/hosts.allow*) für alle Dienste die diese Funktionalität unterstützen, insbesondere für den SSH-Daemon
- SuSE Firewall-Paket
- Automatische Sicherheitsaktualisierungen von Distributionspaketen

- Statusmails über verschiedene Systemparameter an den Kunden
- Inventarisierung im Managementsystem „LRZmonitor“
- Shell-Zugriff durch die Linux-Gruppe

Über die sichere Konfiguration der im Auslieferungszustand aktivierten Dienste hinaus werden keine Sicherheitseinstellungen vorgenommen. Einzelne Details wie die Festlegung des Wartungstages für Sicherheitsaktualisierungen oder die Ausnahme von bestimmten Paketen sind durch den Nutzer konfigurierbar. Abgesehen davon unterliegt es der Verantwortung des Administrators, seine selbst installierten Anwendungen auf einem aktuellen Stand zu halten und für eine sichere Konfiguration zu sorgen.

Attended Hosting - Windows

Ein weiteres Standbein ist das Angebot von virtuellen Windows-Servern auf Basis von Windows Server 2003, 2008 und 2008 R2. Hierbei kommen die am LRZ üblichen Sicherheitsmaßnahmen für Windows-Rechner zum Einsatz, die die folgenden Punkte umfassen:

- Integration ins Active Directory
- Gruppenrichtlinien
- automatische Windows-Updates, bei Bedarf mit automatischem Reboot [LRZ 12f]
- Client für Microsoft System Center
- Sophos Virenscanner

Ebenso wie bei den Linux-Servern unterliegt der sichere Betrieb von installierter Anwendungssoftware ausschließlich dem Nutzer des Servers.

Unattended Hosting

In die letzte Kategorie fallen Server, deren Betriebssystem nicht durch das LRZ unterstützt wird oder bei denen die LRZ-üblichen Sicherheitsmaßnahmen abgeschaltet sind. Neben nicht unterstützten Linux-Distributionen wie Red Hat Enterprise Linux (und deren Derivaten CentOS und Scientific Linux) können darunter auch andere Betriebssysteme, beispielsweise Solaris oder Vertreter der BSD-Welt verstanden werden. Eine weitere Gruppe sind wie schon genannt die Vertreter der Attended Hosting Varianten, bei denen beispielsweise automatische Sicherheitsaktualisierungen auf expliziten Kundenwunsch abgeschaltet wurden.

Da beim Unattended Hosting kein Zugriff durch das LRZ mehr stattfinden und auch der Stand der Sicherheitsaktualisierungen nicht mehr festgestellt werden kann, müssen Unattended VMs in einem Sicherheitskonzept als generell unvertrauenswürdig eingestuft und weitgehend isoliert werden.

2.1.6 Kundenkreis

In der Einführungsphase wurden hauptsächlich physische Server von LRZ-Diensten virtualisiert oder neue LRZ-Dienste von Anfang an auf virtuellen Servern aufgebaut. Hierbei kommt die seit Jahren erprobte Arbeitsteilung zwischen der für das Betriebssystem verantwortlichen (Linux- bzw. Windows)-Systemgruppe und dem Dienstbetreiber zur Geltung, die die Kompetenzen für den sicheren Betrieb einer Teilkomponente bei der Gruppe bündelt, die am meisten Erfahrung damit hat.

Im Zuge der bereits in der Einleitung angesprochenen Rezentralisierung und der verstärkten Aufstellung des LRZ als Anbieter von Rechenkapazitäten werden nun virtuelle Maschinen auch den am MWN angeschlossenen Institutionen und Lehrstühle angeboten [LRZ 12b]. Durch den dadurch massiv erweiterten, sehr heterogenen Kundenkreis entstehen neue Herausforderungen. So setzen viele Institute seit Jahren Serversysteme ein, die vom LRZ nicht offiziell unterstützt werden. In diesen Fällen bleibt nur entweder ein Betrieb als Unattended VM, oder die Umstellung auf eine neue Distribution, mit der keine Erfahrungen vorliegen.

Eine Sicherheitspolicy, die den Anforderungen aller Kunden Rechnung trägt ist die große Herausforderung dieser Arbeit. Die Virtualisierungsinfrastruktur ist hierbei nicht der erste Dienst, der spezifisch mandantenfähig angeboten werden muss. Ein anderes Beispiel sind die bereits im Kapitel 2.1.2 erwähnten virtuellen Firewalls auf der Basis einer Cisco ASA. Die dortigen Bedingungen unterscheiden sich jedoch von denen der vorliegenden Virtualisierungsumgebung, da hier aus Skalierungsgründen die einzelnen Mandanten in einer gemeinsamen Netzstruktur angebunden werden.

2.1.7 Netzmonitoring, Incident-Behandlung

Die in dieser Arbeit hauptsächlich diskutierten Maßnahmen sind überwiegend im Bereich der Prävention angesiedelt. Das bedeutet, dass sie Angriffe erschweren und Auswirkungen von erfolgreichen Angriffen limitieren sollen. Da selbst bei perfekten Voraussetzungen Sicherheitsprobleme nicht immer zu vermeiden sind, existieren auch einige reaktive Maßnahmen, die eine Kompromittierung erkennen und zur manuellen Behandlung melden sollen.

Das MWN betreibt neben einem funktionsfähigen Abuse-Desk für die Meldung von auffälligen Maschinen (beispielsweise ausgehende SSH-Scans) mehrere automatische Mechanismen zur Erkennung von abnormalem Kommunikationsverhalten [LRZ 12a]. Diese überwachen am Übergang ins Internet („X-WiN-Übergang“) Parameter wie Verbindungsanzahl, Paketanzahl oder Datenvolumen und schlagen bei Abweichungen vom Standardwert Alarm. Die möglichen Reaktionen reichen von einer automatischen Sperrung am oben genannten Übergang bis zur Auslösung von Security Incident-Prozessen.

2.2 Nicht-funktionale Anforderungen

An Lösungen und Sicherheitskomponenten, die in virtualisierten Hostingumgebungen zum Einsatz kommen, werden verschiedene Anforderungen gestellt. Eine Nicht-Erfüllung erschwert den Einsatz einer Technologie zum Teil oder verhindert ihn sogar. Die genaue Zusammenstellung und Gewichtung dieser Anforderungen ist abhängig von der Infrastruktur und dem Einsatzgebiet und wird hier am konkreten Beispiel des Leibniz-Rechenzentrums dargestellt. In anderen Umgebungen können sich auch noch weitere Anforderungen ergeben.

In diesem Unterkapitel werden die nicht-funktionalen Anforderungen (NF) an das System formuliert. Die funktionalen Anforderungen und die daraus resultierenden Testszenarien sind im Kapitel 4.1 dokumentiert.

NF1 – Sicherheit

Eine grundlegende Anforderung an alle Komponenten ist die Sicherheit. Dies mag bei der Evaluierung von Sicherheitsmechanismen paradox erscheinen, wird jedoch häufig übersehen und führt zur Einführung zusätzlicher Sicherheitslücken bei der Lösung bestehender. So müssen zusätzliche Komponenten sicher gegen missbräuchliche Nutzung konfigurierbar sein und dürfen keine von entfernten Angreifern ausnutzbaren Sicherheitslücken haben. Da dies bei komplexen IT-Systemen nicht zu garantieren ist, muss zur Evaluierung auf eine Abschätzung auf Erfahrungswerten und ein Sicherheitskonzept geachtet werden, welches potentiellen Angreifern wenigstens rudimentäre Sperren in den Weg legt.

Diese Anforderung gilt als erfüllt, wenn die Komponente durch einen sicheren, verschlüsselten und authentifizierten Weg verwaltet werden kann und die üblichen Rollenkonzepte zur Administration möglich sind.

NF2 – Komplexität

Eine sich aus der Sicherheit direkt ergebende wichtige Anforderung ist die der (beschränkten) Komplexität. Ein System, dessen zumindest grundlegende Funktionsweise selbst vom Administrator nicht mehr verstanden wird („Black-Box“), kann auch bei einem durch einen Angreifer verursachten Fehlverhalten nicht trivial identifiziert werden und bleibt unter Umständen daher länger unentdeckt. Diese Anforderung besteht selbstverständlich nicht nur an jeden einzelnen Mechanismus, sondern auch an das Gesamtsystem.

NF3 – Mandantenfähigkeit

Bereits im Titel der Arbeit wird die Anforderung der Mandantenfähigkeit genannt. Dieser Begriff beschreibt die Fähigkeit, auf der gleichen Infrastruktur mehrere Kunden bedienen zu können, ohne dass diese gegenseitigen Einblick oder gar Änderungsmöglichkeiten in ihre Konfiguration haben. Beim Angebot von Dienstleistungen an externe Kunden ist dies, im Gegensatz zu den internen Virtualisierungsumgebungen von Firmen („Private Clouds“), dringend nötig. Andernfalls müssten die Sicherheitseinstellungen der gesamten Infrastruktur auf die maximalen Anforderungen beziehungsweise minimalen Fähigkeiten aller Kunden beschränkt werden. Ein wichtiger Unterpunkt der Mandantenfähigkeit ist die Möglichkeit für den Kunden, die Einstellungen seines bestellten Produkts eigenständig über Self-Service zu modifizieren. Im Gegensatz zur Änderung durch einen Service Request, der in den meisten Fällen den manuellen Eingriff einer berechtigten Person des Betreibers auslöst, kann der Kunde über ein Self-Service-Portal Änderungen zu jeder Tages- und Nachtzeit vornehmen. Diese können dann in vielen Fällen automatisch in die Infrastruktur eingepflegt werden und stehen sofort zur Verfügung. Neben dem verringerten Aufwand auf der Betreiberseite (und damit schlussendlich einer Kostenersparnis) sind bei Verfügbarkeit eines Self-Service-Portals die Kunden eher bereit, standardmäßige Einschränkungen durch eine Sicherheitsrichtlinie hinzunehmen, wenn sie diese Bedarf schnell deaktivieren können.

Diese Anforderung gilt als erfüllt, wenn alle für den Endkunden relevanten Einstellungen der Sicherheitskomponente für diesen spezifisch festgelegt werden können. Dabei ist es zunächst unerheblich, ob diese Einstellungen nur vom Betreiber der Plattform oder auch vom Kunden selbst vorgenommen werden können. Sofern jedoch in der Praxis häufiger Einstellungen vorgenommen werden müssen ist die Möglichkeit eines Self-Service-Portals vonnötigen.

Bei Maßnahmen, die ohne Ausnahme und ohne Konfiguration für alle Kunden gelten sollen, entfällt diese Anforderung.

NF4 – Benutzbarkeit

Aus den beiden vorgenannten Anforderungen schließt sich auch die Anforderung der Benutzbarkeit. Wenn der Kunde selbst Änderungen an seinem Produkt vornehmen kann, so muss er auch in der Lage sein, die Folgen seiner Änderung zu verstehen. Bei einem großen Personenkreis mit unklarem Wissensstand kommt man daher nicht umhin, möglichst einfache Formulierungen zu wählen, sinnvolle Standardvorgaben zu machen und komplexe Entscheidungen nicht dem unbedarften Nutzer zu überlassen. Gleichzeitig sind jedoch auch Benutzer mit einem sehr hohen Wissensstandard im Nutzerkreis, die eine Einschränkung der Konfigurationsmöglichkeiten nicht hinnehmen können oder wollen. Eine Abwägung zwischen diesen beiden Extremen oder aber eine unterschiedliche Möglichkeiten je nach Wissensstand sind hier nötig. Eine weitere Möglichkeit dieses Problem zu umgehen ist der Einsatz von generischen Sicherheitsmechanismen, die keine Konfiguration durch den Benutzer erfordern.

NF5 – Interoperabilität

Ein wesentlicher Vorteil von standardisierten Protokollen ist die Interoperabilität. Sie erlaubt, Komponenten verschiedener Hersteller zusammenzuschließen, einzelne Teile davon auszutauschen und zu aktualisieren und dabei keine Rückschritte vornehmen zu müssen. Diese Interoperabilität von standardkonformen Komponenten ist eine wesentliche Voraussetzung für eine zukunftssichere Lösung. Leider hinkt die Standardisierung den tatsächlichen Anforderungen oft um mehrere Jahre hinterher (oder die Anforderungen sind so speziell, dass eine Standardisierung nicht möglich ist), in denen die Hersteller proprietäre Lösungen auf den Markt bringen. Diese benötigen häufig nicht nur den Einsatz einer spezifischen Komponente, sondern auch den Einsatz und gegebenenfalls Ersatz anderer Komponenten in der Infrastruktur durch Produkte des Herstellers. Am Ende führen die vorgenannten Abhängigkeiten häufig dazu, dass die Infrastruktur nicht mehr auf die aktuellen Anforderungen angepasst werden kann, da die Änderung automatisch einige andere Komponenten außer Betrieb nehmen würde.

Maßnahmen, die diese Anforderung erfüllen, dürfen keine herstellerspezifischen Erweiterungen in anderen Komponenten erfordern.

NF6 – Kompatibilität

Während der Einsatz von Sicherheitsmechanismen bei Neubauten „auf der grünen Wiese“ und mit Spezialisierung auf einen bestimmten Anwendungszweck wesentlich einfacher einzusetzen sind muss am LRZ auf eine Kompatibilität mit den bestehenden Anwendungen, Diensten und Komponenten geachtet werden. Maßnahmen, die zum Zweck der Sicherheit bei Housingkunden den Betrieb der LRZ-eigenen Dienste erheblich einschränken oder gar unmöglich machen können nicht eingesetzt werden.

NF7 – Skalierbarkeit

Last but not least muss eine gewählte Sicherheitslösung auch auf die Vielzahl von Kunden beziehungsweise Mandanten skalierbar sein. Der beste mandantenfähige Sicherheitsmechanismus nützt nichts, wenn er so viele Ressourcen verschlingt, dass im Extremfall nur noch ein Mandant auf der Infrastruktur bedient werden kann. Es gibt in einer komplexen Infrastruktur sehr viele Ressourcen, die nicht in unendlicher Anzahl verfügbar sind (beispielhaft seien hier 802.1q VLAN-IDs oder IPv4-Subnetze genannt) und eine unüberwindbare Hürde für einen Ausbau der Dienstleistung darstellen können.

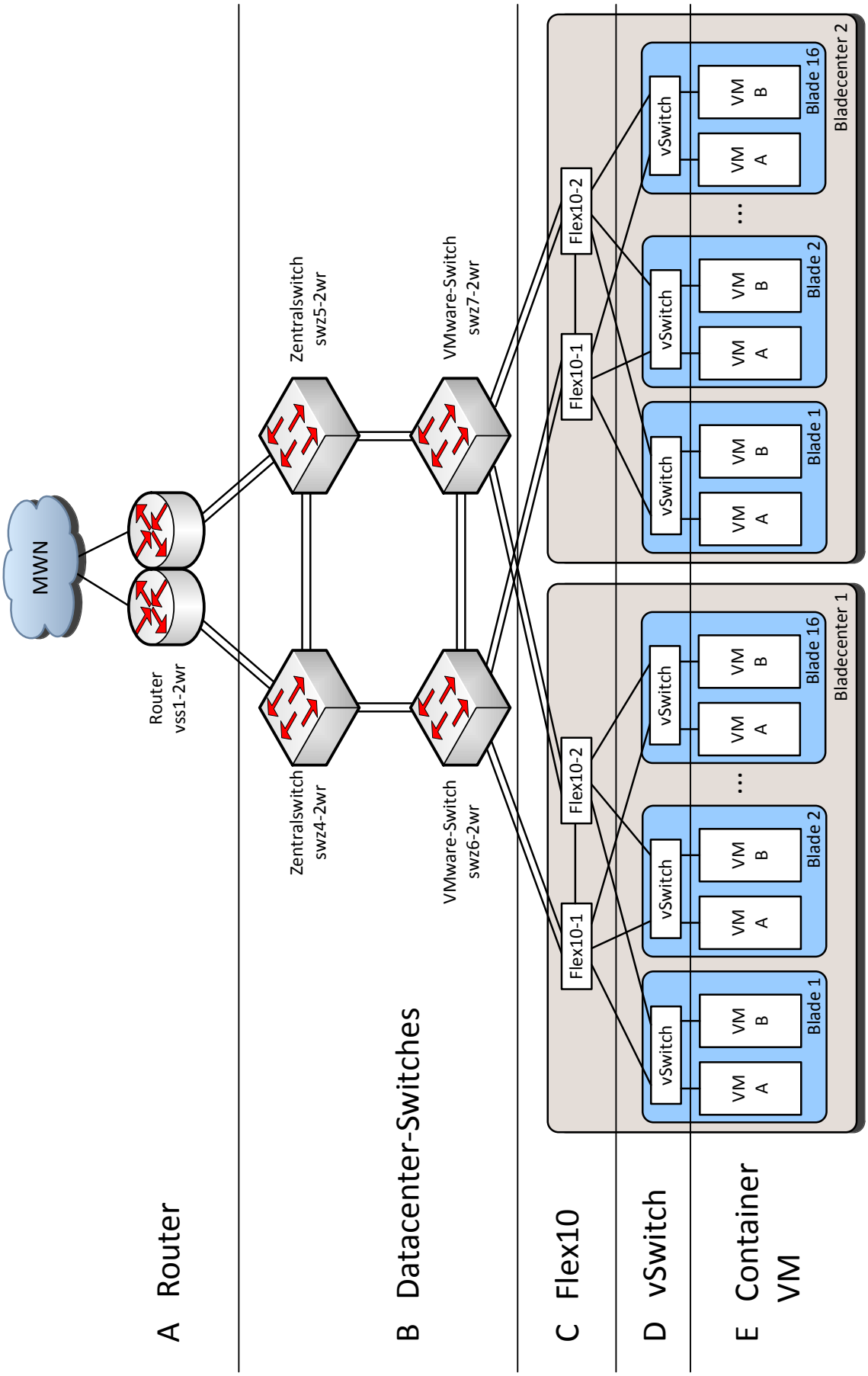


Abbildung 2.1: Netztopologie

3 Technische Grundlagen und Angriffe

Angriffe auf Computersysteme lassen sich zuallererst in **gezielte** und **ungezielte** Angriffe einteilen. Unter gezielten Angriffen versteht man die Ausrichtung der Angriffsvektoren am Zielsystem, in den meisten Fällen auch unter Ausnutzung von öffentlich verfügbaren oder auf anderen Wegen erhaltenen Informationen (zum Beispiel *Social Engineering*). Der Zweck ist dabei häufig direkt mit dem Zielsystem oder dessen Betreiber verknüpft und kann dabei Elemente wie Datenausspähung oder gar Datenveränderung beinhalten.

Im Gegensatz dazu werden bei ungezielten Angriffen oftmals (alt-)bekannte Sicherheitslücken bei vielen erreichbaren Systemen ausgenützt. Der Angreifer hat dabei selten ein Interesse am konkreten Zielsystem, sondern will dieses nur als Ausgangspunkt für weitere Angriffe, für den Versand von Spam oder zur Ausnutzung der Rechenkapazität missbrauchen.

Während sich die ungezielten Angriffe oft durch Maßnahmen verhindern oder zumindest erschweren lassen, die zu den Grundlagen des sicheren Systemmanagements zählen (wie regelmäßige Sicherheitsupdates und sichere Passwörter), bieten die gezielten Angriffe dem fähigen Angreifer eine wesentlich breitere Angriffsfläche. Hierbei werden auch oft Schwachstellen in verschiedenen Komponenten kombiniert, um schlussendlich zu einem erfolgreichen Angriff zu kommen.

Ein oft benutzter Angriffsvektor ist die Kompromittierung eines (unter Umständen schlechter gesicherten) Drittsystems, welches sich entweder aufgrund von expliziter Konfiguration oder aber aufgrund der Netztopologie als Ausgangspunkt für weitere Angriffe eignet. Dies kann beispielsweise durch den Betrieb im gleichen VLAN (und damit Zugriff auf eine gemeinsame Broadcast-Domain) gegeben sein, wie es in der aktuellen virtuellen Infrastruktur (siehe Kapitel 2.1.2) geschieht.

Die in den folgenden Unterkapiteln beschriebenen Angriffsszenarien sind die technischen Basis für weiterführende Angriffe wie

- **Ausspähung**, wobei der Datenaustausch zwischen zwei Systemen belauscht aber nicht verändert werden kann
- **Spoofing**, bei dem die Identität eines Kommunikationsteilnehmers angenommen wird
- **Man-in-the-middle** als Unterart der beiden vorangegangenen Angriffe, bei dem der Datenaustausch zwischen zwei Systemen (unbemerkt) verändert wrd
- **Denial-of-Service**, wobei der Dienstbetrieb bis zur Nichtverfügbarkeit gestört wird.

3.1 Data Link Layer (Schicht 2)

Die Schicht 2 im OSI-Schichtenmodell ist für eine zuverlässige, weitgehend fehlerfreie Übertragung zwischen zwei benachbarten Systemen zuständig. Verbreitete Vertreter dieser Protokolle sind Ethernet (IEEE 802.3[IEEE Std 802.3-2008]) und das damit eng verwandte Wireless LAN (IEEE 802.11), HDLC oder PPP auf Punkt-zu-Punkt-Verbindungen sowie Infiniband und Myrinet im Höchstleistungsrechnen.

Bei einem geteilten Medium regelt die Sicherungsschicht die Kollisionsvermeidung und -erkennung beim Zugriff. Während diese Aufgabe auch heute noch eine große Rolle spielt, zum Beispiel beim Wireless-LAN Standard IEEE 802.11, kommen in Rechenzentrumsnetzen schon seit über 10 Jahren nur noch kollisionsfreie Punkt-zu-Punkt-Verbindungen wie Switched Ethernet zur Anwendung. Im Rahmen dieser Arbeit werden daher nur die Sicherheitsfaktoren eines auf IEEE 802.3 basierenden Netzes betrachtet.

MAC-Adressen bestehen aus 48-Bit, bei denen die beiden niederwertigsten Bits (LSB = *Least Significant Bits*) im ersten Byte eine spezielle Bedeutung haben. Das niedrigste Bit beschreibt, ob es sich bei der angegebenen Adresse um eine Broad- oder Multicast-Adresse handelt (b1=1), mit der mehrere Systeme gleichzeitig angesprochen werden können, oder um eine Unicast MAC-Adresse (b1=0), die spezifisch einem Endsystem zugeordnet ist. Das zweitniedrigste Bit beschreibt, ob die ersten 3 Bytes der Adresse dem Hersteller durch die IEEE zur kollisionsfreien Vergabe zugeteilt wurden (b2=0) oder ohne eine eindeutige Zuweisung generiert wurden (b2=1).

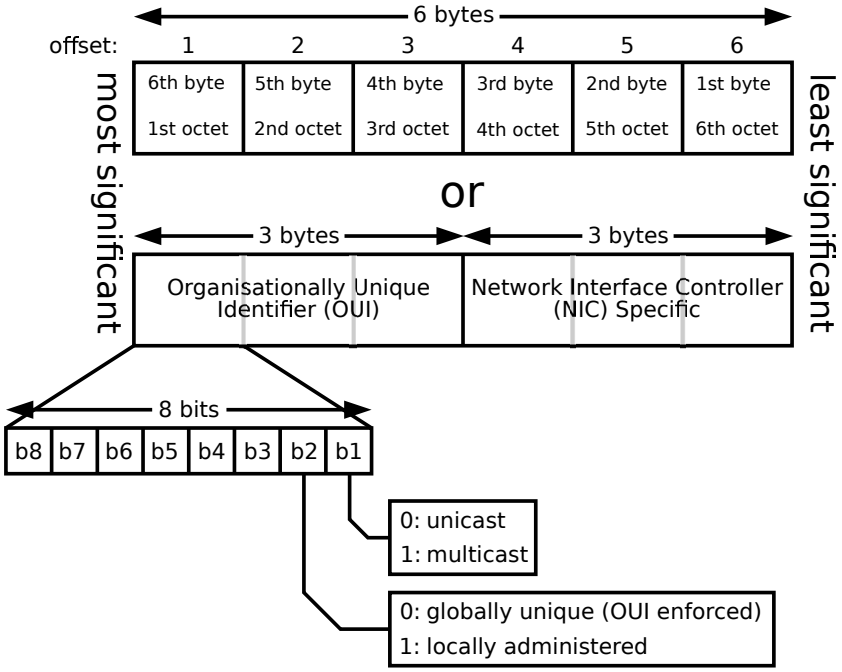


Abbildung 3.1: Multicast- und u/l-Bit im Ethernet Header (Quelle: Wikipedia-Benutzer kju [inkj 08])

Jedes Gerät, welches an einem IEEE 802-basierten Netz teilnehmen will, benötigt eine zumindest innerhalb dieses Netzes eindeutige Unicast-Adresse der Schicht 2 (MAC-Adresse). Zu diesem Zweck werden von der IEEE Blöcke (OUI Assignments) von jeweils 3 Byte Größe an Hardwarehersteller vergeben, welche diese dann ohne Mehrfachverwendung auf ihre produzierten Geräte verteilen. Damit ist sichergestellt, dass jede MAC-Adresse weltweit nur einmal vergeben ist. Diese Methodik funktioniert jedoch in Virtualisierungsumgebungen nicht mehr, da bei diesen die virtuelle Netzwerkkarten der Gastsysteme im Allgemeinen keinen Bezug zu einem physischen Gerät haben können und daher auch keine global eindeutige MAC-Adresse mehr zugewiesen bekommen können. Obwohl der Standard durch die Nutzung des u/l-Bits auch eine lokale zufällige Vergabe vorsieht, haben es alle bekannten Hersteller von Virtualisierungssystemen vorgezogen, einen eigenen globalen 24 Bit-Block bei der IEEE zu beantragen und eine kollisionsfreie Vergabe innerhalb dieses Blocks in ihren Managementsystemen sicherzustellen. Eine Liste der herstellerspezifischen OUI-Blöcke ist in Tabelle 3.1 zu finden.

Firma	MAC-Prefix
VMware	00:50:56
Microsoft Hyper-V	00:15:C0
Xen	00:16:3E

Tabelle 3.1: MAC-Adressräume der Hypervisor-Hersteller

3.1.1 Wegefindung

Der IEEE 802.3 (Ethernet)-Header besteht aus einem sechs Byte langen Feld für die Zieladresse, einem sechs Byte langen Feld für die Quelladressse sowie einem zwei Byte langen Feld für das enthaltene Protokoll (Ethertype).

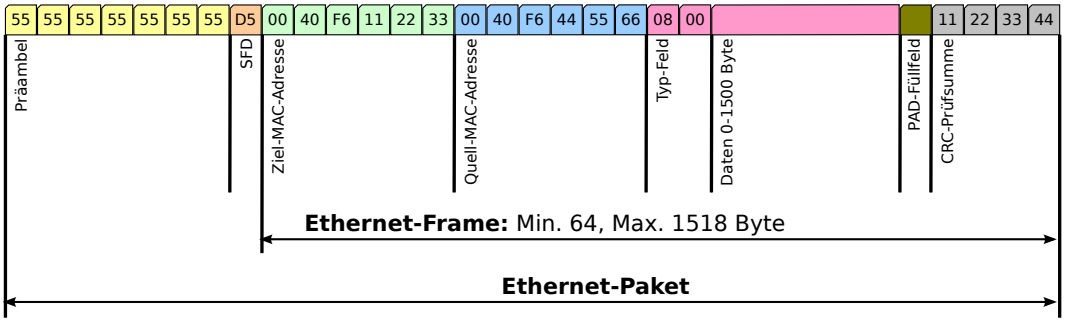


Abbildung 3.2: Der 802.3 Ethernetheader (Vorlage: Wikipedia-Benutzer Bluepoke [Blue 09])

Die Quelladresse ist dabei immer die Unicast MAC-Adresse des sendenden Systems, während die Zieladresse eine Unicast-MAC, eine Multicast-MAC oder die Broadcast-MAC sein kann. In den Anfängen des Ethernets auf einem Shared Medium wurde ein ins Netz gesendetes Ethernet-Frame von allen teilnehmenden Stationen empfangen. Da dies offensichtliche Geschwindigkeitsengpässe im LAN hervorruft wurden *MAC-Bridges* (kurz: Bridge) eingeführt; die heute in Massen verwendeten Ethernet-Switches sind vom Standard her Multiport-Bridges. Diese Geräte halten intern eine Tabelle vor, in der eine Zuordnung aller bekannten Unicast-MAC-Adressen zu einem physischen Port der Bridge vorgenommen wird. Verkehr an eine bekannte Unicast-Adresse wird dadurch nur noch auf einen Port gesendet. Die Zuweisungstabelle kann statisch konfiguriert werden, jedoch sind üblicherweise alle Switches *transparente Bridges*, das heißt selbstlernend. Hierbei werden Frames mit unbekannter Zieladresse an alle Ports weitergeleitet (Flooding). Beim Empfang eines Frames wird die Quelladresse und der empfangende Port in der internen Weiterleitungstabelle aktualisiert, so dass der Switch jederzeit eine aktuelle Sicht auf den Pfad zu einem bestimmten Teilnehmer hat.

Ein Seiteneffekt einer Bridge ist eine erhöhte Sicherheit, da (Unicast-)Frames, die zwischen zwei Teilnehmern ausgetauscht werden, bei einer korrekten Weiterleitungstabelle nicht auf dem Port eines Dritten gesendet und daher auch nicht beobachtet werden können. Dies macht Man-in-the-Middle-Angriffe deutlich schwieriger.

MAC Spoofing

Beim MAC-Spoofing sendet ein Angreifer ein Frame mit der (gefälschten) Absenderadresse eines anderen Teilnehmers auf dem gleichen Segment. Aufgrund der vorher beschriebenen sofortigen Aktualisierung der Weiterleitungstabelle im Switch zeigt nun auf allen Switches, die dieses Frame empfangen, der Eintrag für den gefälschten Eintrag in Richtung des Angreifers. Dadurch werden Pakete an diese Adresse nicht mehr an den legitimen Eigentümer, sondern an den Angreifer gesendet. Dieser kann die Pakete nun mitlesen, unterdrücken oder verändert weiterschicken. Geschieht dies beispielsweise mit der MAC-Adresse des Routers, so kann der gesendete Verkehr des gesamten Netzsegments überwacht werden.

Gegenmaßnahmen für dieses Problem basieren auf einer Einschränkung des Programmierung der gesamten Weiterleitungstabelle. Eine Einschränkung auf eine, fest eingestellte MAC-Adresse ist jedoch in einigen Einsatzgebieten zu restriktiv. Hierzu gehören insbesondere auf der Schicht 2 arbeitende VPN- und Firewalllösungen.

Promiscuous Mode

Der Promiscuous Mode bezeichnet einen Modus, in dem eine Bridge (Switch) durch eine Konfiguration oder einen Software-Fehler in einen Repeater (Hub) umgewandelt wird, der den vollständigen Verkehr unabhängig

von den Zieladressen auf jedem Port versendet. Dadurch ist es einem Angreifer möglich, jeglichen Verkehr auf dem Netzwerksegment zu belauschen, ohne dass andere Teilnehmer davon etwas erfahren.

Die Nutzung des Promiscuous Mode in Switches ist im Allgemeinen auf manuelle Eingriffe des Systemadministrators, beispielsweise zum Debugging, beschränkt. Es existieren jedoch immer wieder Umstände, mit denen einzelne Systeme temporär in den Promiscuous Mode geschaltet werden können. Ein oft praktizierter Angriff zur Herstellung dieses Modus ist das sogenannte **MAC Flooding**, bei dem ein System sehr viele verschiedene, zufällig generierte Quelladressen verwendet. Da die Switchinfrastruktur jede im Netz benutzte MAC-Adresse speichert und nur über eine beschränkte Speicherkapazität verfügt (im Allgemeinen zwischen 8000 und 16000 Einträgen), kann es dadurch zu einem Überlauf der Weiterleitungstabelle kommen. Viele Systeme schalten in diesem Fall in den Promiscuous Mode, um die Kommunikation zumindest rudimentär aufrecht erhalten zu können.

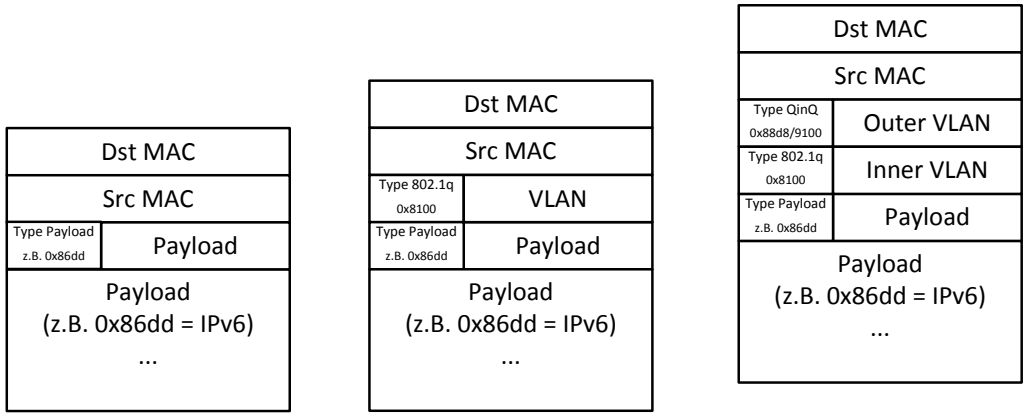
Dieser Angriff ist nicht möglich, wenn geeignete Gegenmaßnahmen gegen MAC Spoofing (siehe der vorherige Punkt) getroffen wurden.

3.1.2 VLAN Tagging

Als zusätzliche Form der Segmentierung dient das VLAN-Tagging gemäß IEEE 802.1Q, bei dem ein physisches Netz in bis zu 4096 virtuelle Netze unterteilt wird.

Dies geschieht durch das Voranstellen eines 32-Bit breiten Headers, der unter anderem eine 12-Bit breite VLAN-ID beinhaltet. Pakete werden nur auf Switchports versendet, die Teilnehmer eines bestimmten VLANs sind. Dabei wird zwischen einem sogenannten Access-Port (untagged, Edge-Port, Netzrand), der einem einzigen VLAN zugewiesen ist, und einem tagged Port (Trunk) unterschieden. Auf Access-Ports wird dabei vor dem Senden das VLAN-Tag entfernt (POP-Operation) und beim Empfang hinzugefügt (PUSH-Operation). Für das am Access-Port angebundene Gerät ist das VLAN daher nicht sichtbar und transparent. Im Gegensatz dazu werden auf Trunk-Ports die VLAN-Tags nicht entfernt, so dass das Gerät auf der Gegenseite diese Trennung weiter durchsetzen kann. Jedoch kann auch auf Trunk-Ports ein sogenanntes natives VLAN definiert werden, bei dem das Tag vor der Übertragung entfernt wird.

Technisch gesehen bilden VLANs damit seperate Broadcast-Domains, in denen nur Teilnehmer des entsprechenden VLANs miteinander kommunizieren können.



(a) kein VLAN-Tag

(b) 802.1Q VLAN-Tag

(c) doppeltes VLAN-Tag 802.1ad („Q-in-Q“)

Abbildung 3.3: Headerstrukturen im Ethernet

Diese Konfigurationsmöglichkeiten können bei einem sogenannten **Double VLAN-Tagging**-Angriff missbraucht werden. Es ist technisch möglich und gerade im Providerumfeld häufig anzutreffen, mehrere Ebenen der VLAN-Kapselung anzuwenden (IEEE 802.1ad, auch unter 802.1QinQ bekannt). Dies dient dazu, den Adressraum

von nur 4096 VLANs zu erweitern (durch doppeltes Tagging beispielsweise auf $4096 * 4096 = 16,7$ Millionen) oder eine logische Trennung zwischen dem VLAN des Carriers und den VLANs des Nutzers zu erreichen). Korrekt eingesetzt hat diese Technik keine Sicherheitsnachteile. Für einen Angreifer kann es allerdings möglich sein, an einem Access-Port ein bereits mit einem VLAN-Tag versehenes Frame zu versenden oder an einen Trunk-Port ein Frame mit zwei VLAN-Tags. Wenn dieses Frame auf einem Trunk-Port gesendet wird, dessen natives VLAN dem des ersten VLAN-Tags entspricht, so wird der Switch beim Senden des Pakets dieses erste Tag entfernen. Für den Empfänger sieht es nun so aus, als wäre das Paket in dem VLAN des zweiten Tags empfangen worden. Dieser Mechanismus ermöglicht Angreifern, die Barrieren zu überspringen und Pakete in beliebige VLANs zu verschicken, wie in Abbildung 3.5 zu sehen ist.

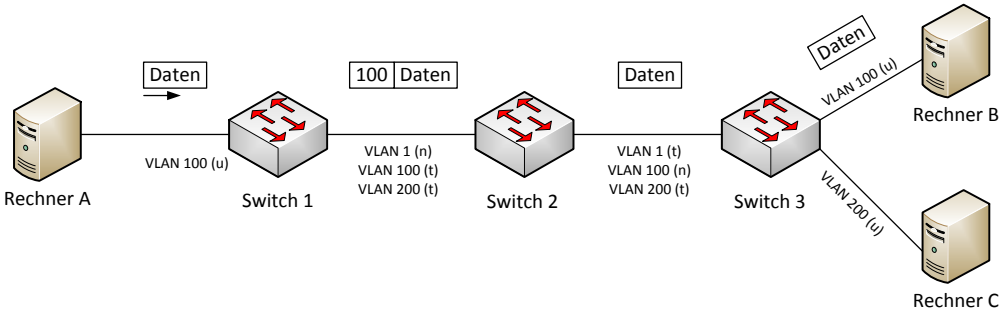


Abbildung 3.4: Normalzustand 802.1Q VLAN-Tagging

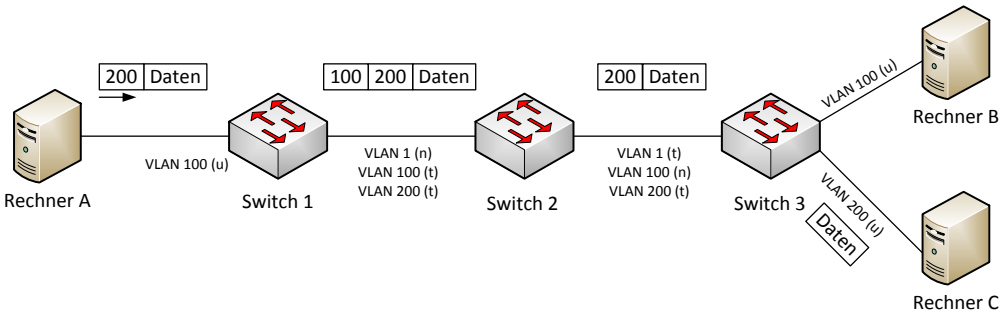


Abbildung 3.5: Angriff durch doppelte VLAN-Tags

Als Gegenmaßnahme müssen Pakete mit VLAN-Tags an Access-Ports herausgefiltert werden. Dies ist im Allgemeinen leicht über den im Header hinterlegten Typ des Ethernet-Frames möglich, was jedoch gerade in den Anfängen nicht in allen Implementierungen erfolgte [AKO+ 05].

3.1.3 Spanning Tree

Da im Ethernet-Protokoll Broadcast-, Multicast- und Unknown-Unicast-Pakete (siehe Kapitel 3.1.1) an alle Ports eines VLANs geschickt werden kommt es bei einer nicht-schleifen-freien Topologie (beispielsweise einem Ring) zu einem sogenannten *Packet Storm*, bei dem ein Paket unendlich lange im Kreis geschickt wird. Wenn mehr als eine Schleife existiert kommt es sogar zu einer exponentiellen Vermehrung der Pakete, die dadurch die gesamte verfügbare Bandbreite belegen und die Nutzdaten verdrängen.

Um diesen Fehlerzustand zu vermeiden kommt in den meisten Netzen das *Spanning-Tree* Protokoll gemäß IEEE 802.1d[IEEE Std 802.1d-2004] oder seine Nachfolger *Rapid Spanning-Tree* (RSTP, IEEE 802.1w) und *Multiple Spanning-Tree* (MSTP, IEEE 802.1s) zum Einsatz. Ihnen ist gemein, dass sie aus einer nicht-schleifenfreien Topologie durch die Abschaltung von redundanten Verbindungen eine schleifenfreie Topologie, den sogenann-

ten Spanning-Tree, erzeugen. Dadurch werden ausgehend von einer (konfigurierten oder automatisch gewählten) Wurzel, der sogenannten Root-Bridge, die kürzesten Verbindungen beibehalten und Daten auf nicht in dieser Menge enthaltenen Verbindungen verworfen. Beim Ausfall einer Verbindung oder eines Systems wird dieser Prozess erneut durchlaufen.

Zur Berechnung der Topologie versenden Spanning-Tree fähige Switches an ihre Nachbarn *Bridge Protocol Data Units* (BPDU), in der sie ihre aktuelle Konfiguration und ihre Distanz zum Wurzelknoten mitteilen. Gelingt es einem Angreifer eine gefälschte BPDU einzuschleusen, so kann er die Netzinfrastruktur zu einer von ihm gewünschten Wegewahl zwingen oder den Netzverkehr sogar unterbrechen.

Eine weitere Gefahr ist die Fälschung einer *Topology Change Notification* (TCN), die im Spanning-Tree Protokoll bei einer Topologieänderung versendet wird. Sie sorgt dafür, dass alle beteiligten Switches ihre Weiterleitungstabelle löschen und neu aufbauen, um geänderte Wege sofort finden zu können. Bei einer gelöschten Weiterleitungstabelle behandeln die Switches alle Frames als Unknown-Unicast und leiten sie an alle Ports weiter. Dadurch kann der Angreifer kurzzeitig nicht für ihn bestimmten Verkehr mitschneiden.

Zur Verhinderung dieses Angriffs müssen Spanning-Tree BPDUs von nicht-autorisierten Ports geblockt werden.

3.1.4 Flooding

Ein weiterer Angriff besteht im Versenden von Broadcast- oder Multicast-Frames mit einer hohen Datenrate, dem sogenannten *Flooding*. Zum einen werden diese Frames auf jedem Switchport in dem entsprechenden VLAN versendet, wodurch unnötig Bandbreite belegt wird, zum anderen werden Broadcast-Pakete von allen Netzwerkkarten empfangen und an den Netzwerkstack des Betriebssystems zur Klassifikation weitergegeben. Dies bringt eine erhöhte CPU-Last mit sich, die zur vollständigen Belegung aller Ressourcen führen kann.

3.2 Address Resolution (Schicht 2.5)

Zur Übertragung auf einem Schicht 2-Protokoll müssen Adressen des Network Layer (zum Beispiel IPv4-Adressen) mit einem geeigneten Mechanismus auf die Adressen des Data Link Layers (zum Beispiel MAC-Adressen bei Ethernet) abgebildet werden. Sie werden prinzipiell bei allen Data Link Layern benötigt, die der BMA-Topologie (Broadcast, Multi Access) entsprechen. Bei Punkt-zu-Punkt Verbindungen (beispielsweise HDLC oder PPP) werden sie nicht benötigt, bei NBMA-Topologien (Non-Broadcast, Multi Access) wie ATM und Frame-Relay werden externe Klassifikationen benötigt. Da diese im Rechenzentrum jedoch kaum noch vorkommen werden sie im Rahmen dieser Arbeit nicht behandelt.

Während diese Protokolle streng genommen medienspezifisch auf dem Data Link Layer arbeiten, verbinden sie Informationen der Schichten 2 und 3. Sie werden deshalb in der Praxis auch oft als „Schicht 2.5“ bezeichnet. Zu dieser Familie gehören neben dem *Address Resolution Protocol* (ARP), welches eine IPv4-Adresse auf eine Ethernet-Adresse abbildet, auch der etwas generischere *Neighbor Discovery*-Teil von IPv6. Dieser wird, da er im Gegensatz zu ARP als eigenständigem Schicht 3-Protokoll, als ICMPv6-Anwendung auf Schicht 4 implementiert. Dem besseren Verständnis halber wird diese Funktion jedoch auch in diesem Kapitel behandelt.

Die Funktionsweise dieser Protokolle ist ähnlich: Soll ein Paket an eine Layer3-Adresse geschickt werden, welche sich gemäß Routingtabelle in einem direkt angeschlossenen Netzsegment befindet (on-link, *directly connected*), so wird die Adresse des Network Layers in einer lokalen Tabelle (ARP-Tabelle (IPv4), Neighbor-Tabelle (IPv6)) gesucht. Ist kein Eintrag vorhanden, so wird das auslösende Paket zurückgehalten und eine Suchanfrage in das Netzsegment geschickt (ARP Query, Neighbor Solicitation). Dieser Versand geschieht über Broad- oder Multicast, um alle angeschlossenen Systeme zu erreichen. Das System, auf dem die gesuchte Layer3-Adresse konfiguriert ist, meldet sich an den Absender zurück (ARP Reply, Neighbor Advertisement). Die erhaltenen Informationen werden nun in die ARP- bzw. Neighbor-Tabelle eingetragen und das Paket an die gelernte Adresse auf dem Data Link Layer verschickt.

Die erlernte Zuordnung wird für eine konfigurierbare Zeit in der Tabelle vorgehalten und danach, je nach Implementation, entweder sofort entfernt oder durch einen erneuten Lernvorgang erneuert.

3.2.1 ARP/NA spoofing

Da die Suchanfragen, wie bereits beschrieben, im Allgemeinen an eine Broad- oder Multicast-Adresse versendet und daher von jedem System im gleichen Netzsegment empfangen werden ist ein möglicher Angriffsvektor, die Anfrage schneller mit der eigenen MAC-Adresse zu beantworten als das eigentliche Zielsystem. Da diese Protokolle mit Ausnahme des kaum eingesetzten *SEcure Neighbor Discovery* (SEND[RFC3971]) über keinerlei Authentifizierung verfügen können durch diesen Angriff falsche Einträge in die Adresstabellen eingefügt werden. Dadurch kann der Verkehr an das Zielsystem umgeleitet und mitgeschnitten oder verändert werden.

Angriffe dieser Art dienen üblicherweise der Vorbereitung für Man-in-the-Middle-Attacken. Sie sind bei einer traditionellen Trennung zwischen Layer2- und Layer3-Infrastruktur schwierig zu unterbinden, da die Layer2-Geräte keine Informationen über erlaubte IP-Adressen erhalten.

IPv4

ARP-Anfragen werden in allen bekannten Implementationen initial beim ersten Paket und periodisch nach Ablauf eines Timers (bei Cisco-Routern vier Stunden) mit einem Broadcast-Paket im gesamten Netzsegment gestellt. Sie können daher von allen Systemen in einer Broadcast-Domain empfangen. Da, wie bereits erwähnt, keine Authentifizierung dieser Pakete erfolgt ist es für den Angreifer ein leichtes, die Anfrage mit seiner eigenen MAC-Adresse zu beantworten. Ist er dabei schneller als der legitime Inhaber der Adresse wird die Antwort in der ARP-Tabelle gespeichert und alle Pakete an diese IP-Adresse an den Angreifer gesendet.

IPv6

IPv6 enthält, abgesehen von dem kaum eingesetzten SEND-Protokoll, ebenfalls keine Authentifizierung. Jedoch wurden, allerdings hauptsächlich aus Performance-Gründen, bei IPv6 mehrere Mechanismen eingeführt die diesen Angriff zumindest erschweren.

IPv6 Neighbor Solicitations werden nicht mehr an die Broadcast-Adresse versendet, sondern an eine von 2^24-Multicast-Gruppen, die aus den letzten 24-Bit der gesuchten IPv6-Adresse gebildet werden. Da diese Gruppen standardkonformes IPv6-Multicast darstellen, werden die Neighbor Solicitations beim Einsatz von MLD-Snooping (*Multicast Listener Discovery*) nur an die Ports gesendet, an denen ein Client vorher eine entsprechende Anfrage (*MLD Response*) für die Gruppe geschickt hat. Es ist daher (beim Einsatz von MLD-Snooping) nicht möglich, generell alle Neighbor Solicitations im Netzsegment zu erhalten.

Bei einem gezielten Angriff ist es jedoch problemlos möglich, die Multicast-Gruppe zur Adresse des angegriffenen Systems zu abonnieren und dadurch die Nachrichten zu erhalten. Da jede Gruppe 2^40 IPv6-Adressen enthält ist eine Überwachung der MLD-Snooping-Tabellen auf den Switches und eine entsprechende Alarmierung bei doppelter Verwendung nicht ohne weiteres zu bewerkstelligen.

Eine weitere Neuerung bei IPv6 ist die Einführung der *Neighbor Unreachability Detection*, der nach Ablauf der (standardmäßig nur im Minutenbereich liegenden) Gültigkeit die bisher bekannte Zuordnung erst durch eine Unicast-Nachricht erneuert. Erst wenn auf diese Anfrage keine Antwort erfolgt wird das Zielsystem mit einem Multicast-Paket gesucht. Da Unicast-Nachrichten im Normalfall nicht an allen Ports gesendet werden (siehe Kapitel 3.1.1) ist dieser Mechanismus vor einem Angreifer versteckt und kann daher auch nicht beeinflusst werden.

3.2.2 gratuitous ARP/unsolicited NA

Es gibt jedoch legitime Gründe, einen bereits bekannten Eintrag in der ARP/Neighbor-Tabelle vorzeitig zu ersetzen. Dies kommt beispielsweise häufig bei Cluster- oder Hochverfügbarkeitssystemen vor. Bei diesen werden eine oder mehrere IP-Adressen, auf denen ein Dienst betrieben wird, bei einem Hardware-Ausfall auf ein redundantes System geschwenkt. Da nicht alle Implementationen dabei die MAC-Adresse des lokalen Interfaces wechseln können oder dies durch die Netzkomponenten als möglicher Angriff sogar aktiv unterbunden

wird (siehe Kapitel 3.1.1) muss dabei die ARP/Neighbor-Tabelle auf allen Systemen im lokalen Netzsegment „auf Zuruf“ angepasst werden.

IPv4

Bei IPv4 dient zu diesem Zweck ein sogenanntes „Gratuitous ARP“ Paket. Dieses ähnelt stark einer herkömmlichen ARP-Antwort (ARP Reply), wird jedoch ohne vorherige Anforderung (ARP Query) an die Broadcast-Adresse verschickt, von allen Systemen im Subnetz empfangen und in die lokale ARP-Tabelle eingetragen ([RFC2002] Section 4.6).

IPv6

Bei IPv6 wird ein derartiges Update als „Unsolicited Neighbor Advertisement “ ([RFC4861] Section 7.2.6) bezeichnet. Im Gegensatz zu normalen Neighbor Advertisement-Nachrichten, die per Unicast an den anfragenden Knoten zurückgeschickt werden, laufen diese als Multicast-Paket auf der ALL-NODES Multicast-Gruppe (ff02::1).

Im Gegensatz zu IPv4 ist dieser Mechanismus bei IPv6 explizit als „nicht zuverlässig“ bezeichnet. Beim Umzug einer Adresse auf einen anderen Node kommt als Rückfallebene die bereits erwähnte *Neighbor Unreachability Detection* zum Einsatz.

3.3 Network Layer (Schicht 3) und höher

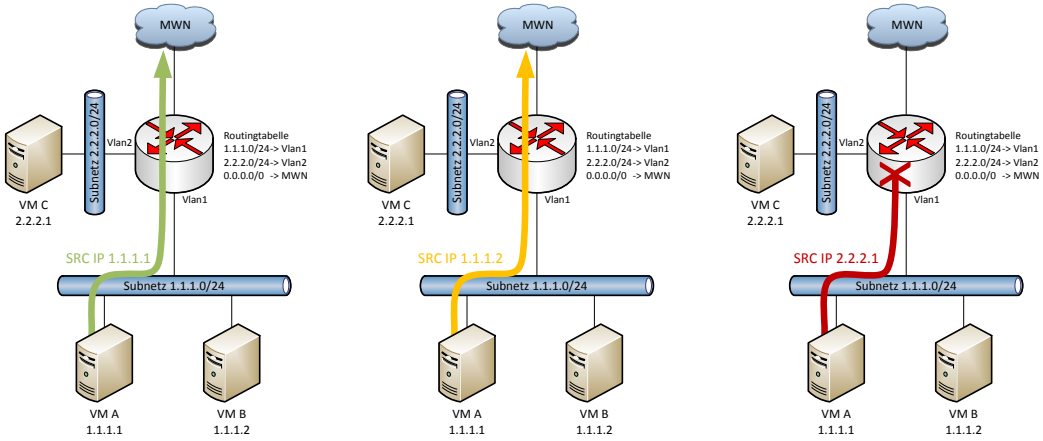
3.3.1 IP spoofing

Eine oft missbrauchte Eigenschaft der paketorientierten Datenverbindung ist die fehlende Ende-zu-Ende-Integrität der Pakete. Dies betrifft nicht nur den Inhalt (Payload), sondern auch die Steuerinformationen (Header). Ein Beispiel dafür ist die Fälschung der Absender-Adresse (Spoofing). Im Unterschied zu ARP/ND-Spoofing (siehe Kapitel 3.2.1) werden hierbei nur Pakete mit einer gefälschten Absenderadresse versendet, jedoch keine Maßnahmen getroffen um Antworten an diese Adresse auch wieder empfangen zu können.

Neben Denial-of-Service-Angriffen durch Überlastung, bei denen durch einen gefälschten Absender eine Filterung und Nachverfolgung durch den angegriffenen Teilnehmer erschwert wird, werden Spoofing-Angriffe auch häufig im Zusammenhang mit Protokollen eingesetzt, die auf dem verbindungslosen UDP basieren. Diesen fehlt, im Gegensatz zum datenstromorientierten TCP, eine zumindest rudimentäre Verifizierung des Kommunikationspartners durch den *TCP 3-Way Handshake*. Ein Beispiel für eine derartige Anwendung sind dynamische DNS-Updates [RFC2136].

Eine andere, in den vergangenen Monaten sehr häufig zu erlebende Nutzung des IP-Spoofings ist die sogenannte „DNS Amplification Attack“. Während bei einem herkömmlichen DoS-Angriff die volle Bandbreite durch die Angreifer bereit gestellt werden muss (sei es durch die direkte Nutzung oder durch Botnetze), werden bei einer DNS Amplification Attack DNS-Anfragen mit der Quelladdresse des angegriffenen Systems an legitime, gut angebundene Nameserver verschickt. Diese Nameserver senden eine DNS-Antwort, die zum Teil um den Faktor 40 größer ist als die Anfrage, an die (gefälschte) Absender-Adresse. Dadurch müssen die Angreifer selbst nur einen Bruchteil der Bandbreite vorhalten, die Verstärkung („Amplification“) geschieht durch missbrauchte Server. Der grundlegende Mechanismus dieser Angriffe ist seit vielen Jahren bekannt [DNS Amplification], jedoch ist erst seit Kurzem eine starke Zunahme zu verzeichnen. Der Grund liegt vermutlich in der zunehmenden Verbreitung von DNSSEC und EDNS0, welches zu deutlich größeren Antwortpaketen in signierten Zonen führt.

IP Spoofing-Angriffe können generell in zwei Klassen unterteilt werden. Bei der Nutzung von völlig fremden Quell-IP-Adressen („off-link“-Adressen) werden Adressen verwendet, die zu anderen Providern gehören, nicht zugewiesen („unassigned“) oder nicht geroutet sind („unrouted“). Diese Pakete sind für einen Serviceprovider ausgehend sehr leicht zu erkennen, da Pakete mit diesen Quelladressen niemals aus der eigenen Infrastruktur



(a) Kein Spoofing (b) „on-link“ IP-Spoofing (c) „off-link“ IP-Spoofing

Abbildung 3.6: Funktionsweise von uRPF

kommen dürfen und daher verworfen werden können. Die Empfehlung BCP 38 [BCP 38] befasst sich mit diesem Thema und empfiehlt, auf allen Schnittstellen zu Kundensystemen fremde Quelladressen automatisch zu blockieren. Die meisten Routerhersteller haben unter dem Begriff *uRPF* (Unicast Reverse Path Forwarding) einen Mechanismus entwickelt, der diese Anti-Spoofing-Filter automatisiert. Dabei wird beim Empfang eines IP-Pakets auf einer mit uRPF geschützten Routerschnittstelle ein Lookup der Quelladresse in der Routingtabelle vorgenommen. Nur wenn ein Paket an die Absenderadresse (also beispielsweise das Antwortpaket) auf der gleichen Schnittstelle herausgeschickt werden würde (das Routing also symmetrisch ist), wird das Paket akzeptiert. Wie aus dieser Beschreibung bereits hervorgeht ist diese Technik daher nur bei symmetrischen Verkehrspfaden einsetzbar. Dies ist im Allgemeinen am Rand von Rechenzentrumsnetzen der Fall, nicht jedoch bei asymetrischen Anbindungen (zum Beispiel Satelliten) oder in Backbone-Netzen. Die Beschreibung verdeutlicht auch, dass ein Einsatz dieser Technologie nur in der direkten Nähe des Absenders sinnvoll ist.

Wesentlich schwieriger ist es, die missbräuchliche Nutzung einer fremden IP-Adresse aus dem lokalen Subnetz („on-link“) zu verhindern. Der klassische uRPF-Mechanismus prüft nur die Informationen der Schicht 3 (Sourceadresse und eingehendes Routerinterface), aber korreliert diese nicht mit den Informationen der ARP- bzw. Neighbor-Tabelle. Selbst wenn die Netzinfrastruktur einen Spoofing-Schutz für die darunterliegenden Schichten implementiert, so ist der Versand mit einer nicht zugehörigen MAC-Adresse ohne weiteres möglich. Die Abbildung 3.6 zeigt die unterschiedliche Behandlung. Während im Fall des „off-link“ Spoofings die Quelladresse eindeutig aufgrund der Routingtabelle als nicht dem Quellinterface zugehörig erkennbar und damit filterbar ist, sind Spoofing-Versuche von IP-Adressen im gleichen Subnetz nicht durch diesen Mechanismus zu erkennen.

Erweiterungen an diesem Mechanismus werden derzeit in der *Source Address Validation Improvements Working Group* (Savi WG) der IETF diskutiert und sind zum Teil von Herstellern auch proprietär implementiert. Sie basieren dabei üblicherweise darauf, den Layer2-Komponenten Informationen über valide Schicht 2/3-Mappings bereitzustellen (zum Beispiel über DHCP-Snooping oder SEND) und diese die ungültige Kombinationen bereits am Netzrand filtern zu lassen.

3.3.2 Redirect

Sowohl IPv4 (RFC 792, ICMPv4 Type 5) als auch IPv6 (RFC 2461, ICMPv6 Type 137 Code 0) sehen den Versand einer „ICMP Redirect“ Nachricht vor, um Pakete von einem Host an einen anderen Router umzuleiten. Dies passiert üblicherweise, wenn darauf folgende Pakete an das gleiche Ziel dadurch einen kürzeren Weg nehmen können. Die Abbildung 3.7 zeigt ein Beispiel der Änderung.

- Host A sendet ein Paket an Host B. Da der Zielhost nicht im lokalen Subnetz (definiert durch die eigene

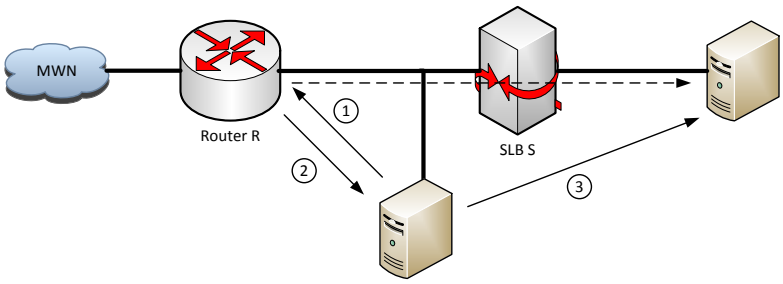


Abbildung 3.7: ICMP Redirect

IP-Adresse und die Subnetzmaske) liegt wird das Paket an das Gateway, Router R gesendet

- Router R hat eine lokale Route für das Zielnetz, in dem Host B liegt. Sie zeigt auf einen Host im gleichen Subnetz, in dem auch A liegt und demzufolge auch direkt senden könnte. Der Router sendet sowohl das Paket gemäß seiner Routinginformation an den SLB als auch ein ICMP-Redirect an Host A, der einen kürzeren Weg zu Host B über den SLB beschreibt.
- Host A übernimmt diese Information (Route zu Host B über SLB) in seine lokale Routingtabelle und sendet folgende Pakete direkt.

Aufgrund von Sicherheitsbedenken, der ausreichenden Kapazität auf Routern und der oftmals praktizierten strikten Trennung von Transportnetzen und Netzen für Server kommt dieser Mechanismus nur noch selten legitim zum Einsatz.

ICMP Redirect-Nachrichten werden jedoch weiterhin von den meisten Betriebssystemen im Auslieferungszustand akzeptiert und können dadurch einem Angreifer erlauben, Netzverkehr für ein bestimmtes Ziel auf sich umzuleiten und damit einen Man-in-the-Middle-Angriff durchzuführen. In dieser Nachricht müssen 28 Byte (IPv4) bzw. bis zu 1200 Byte (IPv6) eines Originalheaders enthalten sein, allerdings kann der Angreifer bereits über einen anderen Angriff Kenntnis eines versendeten Pakets erhalten oder selbst ein Paket mit einer voraussagbaren Antwort verschickt haben.

3.3.3 Rogue DHCP

Einige Betreiber sind aus verschiedenen Gründen dazu übergegangen, auch Server mit dem DHCP-Protokoll (Dynamic Host Configuration Protocol, RFC 2131) zu konfigurieren. Hierbei können auch feste IP-Adressen vergeben werden. Als Vorteil gilt, dass Änderungen an Netzparametern wie dem Gateway, der Netzmaske oder den DNS-Servern zentral vorgegeben werden können, was insbesondere bei strikter Trennung in der Administration die Abstimmung erleichtert. Zum anderen ist DHCP in vielen Netzen für Booten vom Netz (PXE) sowieso nötig und stellt daher keinen Zusatzaufwand dar.

Wie bei nahezu allen Protokollen, die älter als wenige Jahre sind, wurde auch beim Design von DHCP auf Sicherheitsmechanismen weitgehend verzichtet. DHCP-Anfragen werden ohne kryptographische Absicherung als Broadcast auf dem Port 67/UDP versendet und können prinzipiell von allen Teilnehmern im gleichen Netzsegment beantwortet werden. Dabei können Sie dem Anfragenden falsche Informationen übermitteln und damit auf eine einfache Weise in den Netzverkehr eingreifen.

3.3.4 ICMPv6 Rogue RA

Einen wesentlich gravierenden Einfluss auf die Verbindungen haben die sogenannten „Rogue Routeradvertisements“ im IPv6-Protokoll. IPv6 spezifiziert zusätzlich zur bekannten statischen Adressvergabe oder der Vergabe mittels DHCP auch die „Stateless Address Autoconfiguration“ (SLAAC). Ein wesentlicher Bestandteil dieses Mechanismus sind die Router Advertisements (ICMPv6 Type 134, RFC 4861). Durch diese kann

ein Gerät seine Rolle als Default-Gateway und ein /64-Prefix signalisieren, in welchem sich Systeme automatisch eine IPv6-Adresse basierend auf ihrer Hardware-Adresse (bei Ethernet üblicherweise modified EUI-64) konfigurieren können.

Ein Angreifer kann sich innerhalb eines Netzsegments durch den Versand eines Router Advertisements selbst zum Default-Gateway definieren und dadurch einen Man-in-the-Middle Angriff durchführen oder die Pakete verwerfen (Denial-of-Service). Dieser Angriff ist gravierender als ein fremder DHCP-Server, da dafür nur ein Paket an eine Multicast-Gruppe gesendet werden muss und die Einstellung sofort auf allen Knoten aktiv ist, die dieses Paket verarbeiten. Daher muss der Versand eines Router Advertisements durch ein nicht-autorisiertes System verhindert werden [RFC6105].

3.3.5 Fragmente und Extension Header

Fragmente sind eine Möglichkeit, große IPv4/IPv6-Pakete auf mehrere Frames der Schicht 2 aufzuspalten („fragmentieren“) und vom Zielsystem wieder zusammensetzen zu lassen („reassemblieren“). Diese Fragmentierung hat mehrere Sicherheitsimplikationen, die bereits seit mehr als zehn Jahren bekannt sind [PtNe 98].

Zum einen ist der Header der Transportschicht (Schicht 4, üblicherweise TCP/UDP/ICMP) nur im ersten Fragment enthalten, nicht jedoch in den folgenden Fragmenten. Ein Sicherheitsgateway, welches aufgrund von diesen Informationen (beispielsweise Ports in einem Paketfilter) eine Entscheidung treffen muss, kann dies daher nur durch Speicherung und Reassemblierung aller Fragmente tun. Dies übersteigt Speicher und Rechenleistung in vielen Geräten, die deswegen in vielen Fällen alle Fragmente passieren lassen.

Zum zweiten besteht insbesondere bei IPv6 das Problem der überlappenden Fragmente. Fragmente definieren einen „Offset“ (die Position im reassemblierten Paket) und die Daten, die an dieser Stelle hinterlegt werden sollen. Da nicht definiert ist, ob ein Fragment nur genau am Ende des vorherigen Fragments beginnen darf, ist es möglich durch überlappende Fragmente einen Datenbereich im Paket mehrfach zu bestimmen. Dabei ist es von der Implementation des empfangenden Stacks abhängig, welche Version der Daten verwendet wird. Diese Technik kann verwendet werden, Angriffsdaten vor Entdeckung zu verstecken.

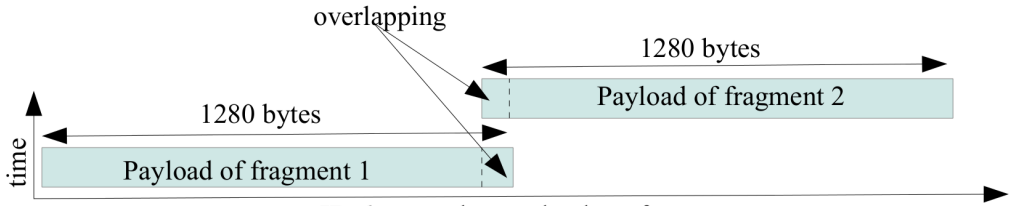


Abbildung 3.8: Überlappende Fragmente (Quelle: Antonios Atlasis[Atla 12])

Diese Methode kann beispielsweise verwendet werden, um Sicherheitsmechanismen in Netzkomponenten zu umgehen, die nicht in der Lage sind, Pakete vor der Überprüfung wieder zusammenzusetzen [IPv6-Frag-Bypass].

3.4 Applikationsspezifische Angriffe

Der weitaus häufigste Grund für erfolgreiche Angriffe ist jedoch ein Problem in einer Anwendung. Diese können beispielsweise durch eine unsichere Konfiguration verursacht werden, wie die Nutzung von Standardpasswörtern oder die fehlende Absicherung von eigentlich internen Diensten. Ein Beispiel für einen derartigen Angriff, der zur Kompromittierung des gesamten Servers führte, ist in Kapitel 1 beschrieben. Ein anderes Einfallstor können Sicherheitslücken in gewollt nach außen angebotenen Diensten sein.

Während Sicherheitslücken in vom Betriebssystem beziehungsweise der Distribution mitgelieferten Applikationen durch regelmäßige Aktualisierungen zumindest eingeschränkt abgesichert werden können, sind dem LRZ als Infrastrukturbetreiber bei der unsicheren Konfiguration von Diensten oder dem eigenverantwortlichen Nutzen veralteter Softwareversionen durch den Kunden die Hände gebunden. Hier kann es neben einem

möglichst sicheren Auslieferungszustand nur darum gehen, dem Kunden einen möglichst einfachen Weg zur selbstständigen Absicherung zu bieten.

Eine generelle Möglichkeit dies zu tun ist die Bereitstellung von einfachen Methoden, die Verbindungen zum betriebenen virtuellen System von außen zu kontrollieren. Im Allgemeinen wird dies durch eine Firewall oder einen Paketfilter erreicht. Wenn dieser im Auslieferungszustand alle Verbindungen von außen auf Dienste auf dem Gast unterbindet, so ist der versehentliche Start einer global erreichbaren Applikation nicht mehr möglich, der Kunde muss zusätzlich den entsprechenden Kommunikationsport explizit freischalten.

3.5 Hypervisor und Cross-Channel-Angriffe

Neben dem offensichtlichen Angriffsvektor über die im Allgemeinen gewollte externe Kommunikation des virtuellen Servers über die Netzanbindung existieren in Virtualisierungsumgebungen noch spezifische Angriffspunkte, die durch die Virtualisierung entstehen.

Dazu gehören zum einen Angriffe direkt auf den Hypervisor, die üblicherweise das Ziel einer Rechteausweitung (*Privilege Escalation*) zur Ausführung von Code in einem höheren Kontext oder aber auch das Abgreifen von geschützten Informationen (*Information Leakage*) verfolgen. Dazu werden Programmierfehler und mangelhafte Validierung in den Schnittstellen des Hypervisors ausgenutzt. Ein bekanntes Beispiel für einen derartigen Angriff ist der bereits zitierte Angriff einer Gruppe von Sicherheitsforschern beim IT-Sicherheitsdienstleister ERNW, bei dem durch den Import eines manipulierten VMware-Images beliebige Dateien und Geräte auf dem Host gelesen werden können [VMDK-Vulnerability]. Dies umfasst sowohl die lokalen Konfigurationsdateien des ESXi-Hosts (wie beispielsweise über Umwege die Datei `/etc/passwd`) als auch die Konfiguration und Plattenabbilder aller anderen virtuellen Maschinen auf einem eingebundenen Datastore. Von den dort gespeicherten Dateisystemen können auch sensible Informationen wie SSL-Schlüsselpaare im Allgemeinen problemlos extrahiert werden.

Eine andere Angriffsklasse sind Cross- oder Side-Channel-Angriffe. Hierbei versucht ein Angreifer, Informationen über den internen Zustand eines Systems herauszufinden, indem das Verhalten bei bestimmten Abfragen analysiert wird. Dabei handelt es sich oft um subtile zeitliche Änderungen in der Antwort, die beispielsweise verraten an welcher Stelle die Prüfung eines Eingabeblocks fehlgeschlagen ist. Auch Änderungen in der Stromaufnahme sind oft eine Quelle für statistische Angriffe. Diese Angriffe sind oft erfolgreich, Schlüsselmaterial von speziell geschützter Hardware wie Smartcards [DKL$^+$ 00] oder Fremdsystemen [Koch 96][Bern 05] zu extrahieren.

Die heutige Server-Virtualisierung als quasi-parallele Ausführung mehrerer, auch den unterschiedlichsten Mandanten zuordneten Maschinen auf einer geteilten (*shared*) Hardware-Plattform ermöglicht weitere Angriffe, die üblicherweise die Ausspähung privater Informationen einer anderen virtuellen Maschine auf dem gleichen Host zum Ziel hat. Der Angriff kann dabei vom legitimen Benutzer ausgehen, aber auch von einem über andere Wege kompromitierten System. So demonstrierten amerikanische Sicherheitsforscher einen erfolgreichen Angriff über CPU-Timingeffekte auf eine ElGamal-basierte Verschlüsselung, die auf einer anderen virtuellen Maschine auf dem gleichen Host lief [ZJRR 12]. Ein anderer demonstrierter Angriff basierte auf den Timingeffekten der RAM-Deduplikation, die optional auf Xen und KVM-basierten Virtualisierungssystemen zur Verfügung stehen [SIYA 11]. Auch VMware ESXi kann bei Knappheit des physikalischen Arbeitsspeichers unter dem Stichwort *Transparent Page Sharing* (TPS) eine entsprechende Deduplikation einsetzen [VMtps].

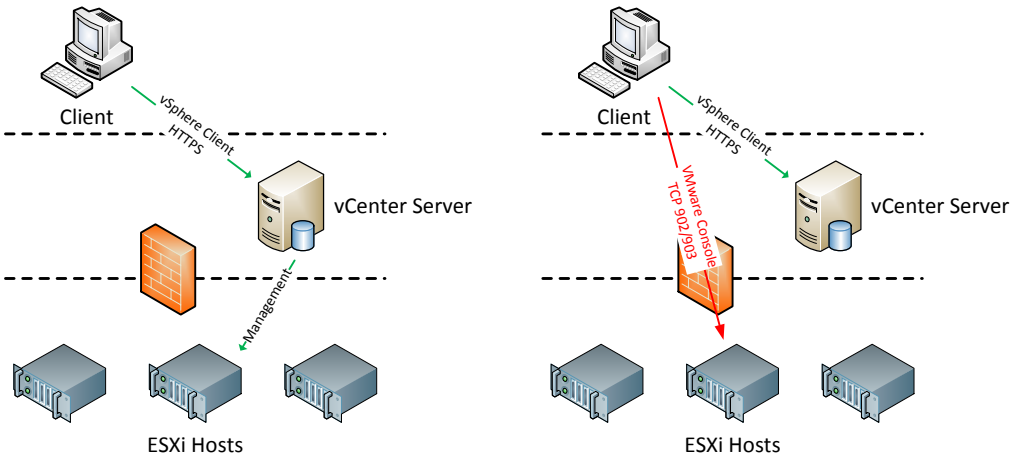
Gegen beide Arten von Angriffen ist Abhilfe nur schwer zu schaffen. Exploits des Hypervisors basieren in den meisten Fällen auf gewollt bereitgestellte oder gar benötigten Schnittstellen, die nicht so einfach beschränkt werden können. Wie für alle Softwareprodukte gilt die Empfehlung, die einschlägigen Mailinglisten (zum Beispiel *Full-Disclosure*[1]) sowie die Veröffentlichungen des Herstellers zu verfolgen und verfügbare Aktualisierungen zeitnah einzuspielen. Noch unbekannte Sicherheitslücken (sogenannte *Zero-Day-Exploits*) können im Allgemeinen nicht verhindert werden. Hierbei gilt ebenso wie für die virtuellen Server, dass die Sicherheitsarchitektur nicht nur präventiv eine Kompromittierung erschweren sollte, sondern nach einem erfolgreichen

[1] `http://seclists.org/fulldisclosure/`

Angriff die weitere Infrastruktur nicht schutzlos dem kompromittierten System ausgeliefert sein sollte. Hierzu dienen insbesondere interne Einschränkungen der Berechtigungen und Kommunikationswege.

Gegen Side-Channel-Angriffe sind ebenfalls kaum generische Lösungen verfügbar. Änderungen an Algorithmen wie dem CPU-Scheduler wirken nur bei ganz spezifischen Angriffen. Die einzig sichere Lösung ist eine Vermeidung des Betriebs zweiter virtueller Maschinen auf einem Host [RTSS 09], was dem eigentlichen Zweck der Virtualisierung entgegen spricht. Eine Trennung „wichtiger“ virtueller Server ist in diesem Kontext ebenfalls nicht sinnvoll, da dieses Attribut zwangsläufig subjektiv von jedem Kunden für seine eigenen virtuellen Server vergeben wird und nicht für eine sinnvolle Unterscheidung benutzbar ist.

In einem von VMware bereitgestellten *Hardening Guide* [VMhard] werden viele Einstelloptionen einerseits und strukturelle Maßnahmen andererseits diskutiert um den Hypervisor und das umliegende Managementsystem abzusichern. Zum großen Teil sind diese, insbesondere was die Struktur und Absicherung der Managementschnittstellen angeht, offensichtlich, seit Jahren in allen großen Systemen empfohlen und in der LRZ-Infrastruktur bereits umgesetzt (siehe Kapitel 2.1). Beispiele dafür sind Punkte wie die Trennung des Management-, Storage- und vMotion-Netzes in seperaten VLANs (*isolate-mgmt-network-vlan*, *isolate-storage-network-vlan*, *isolate-vmotion-network-vlan*) und einem Zugriff auf diese Netze nur durch freigegebene Maschinen durch eine Firewall hindurch (*restrict-mgmt-network-access-gateway*, *restrict-mgmt-network-access-gateway*). Diese Firewall-Regeln müssen jedoch allen Nutzern des vSphere-Clients (und damit auch Kunden) einen Zugang auf bestimmten Ports im VMware-Managementnetz erlauben, da für den Aufbau einer virtuellen Konsole eine direkte Verbindung zwischen dem vSphere-Client und dem verwendeten ESXi-Host aufgebaut wird, wie in der Grafik 3.9 zu sehen ist.



(a) Zugriff Managementaufgaben

(b) Zugriff Konsole

Abbildung 3.9: Kommunikationswege des vSphere Clients

Eine Lösung für dieses Problem existiert noch nicht.

Sofern sich aus der Evaluierung des *Hardening Guides* oder aus davon unabhängigen Überlegungen Empfehlungen für die Konfiguration der VMware-Infrastruktur am LRZ ergeben werden diese in Kapitel 6.4 vorgestellt.

4 Evaluation

Im folgenden Kapitel werden verschiedene Möglichkeiten zur Absicherung der bisherigen Infrastruktur evaluiert. Der Fokus liegt dabei, wie schon im Kapitel 1.3 erläutert, auf der Verhinderung von netzbasierten Angriffen, da Attacken auf das Netz als unterste Schicht in vielen Umgebungen sehr erfolgreich ohne spezifische Anpassungen durchgeführt werden können. Sie werden daher oft als Vorbereitung verwendet, um beispielsweise nach der Kompromittierung eines schlecht konfigurierten Servers weitere Angriffe auf besser gesicherte Server (unter Umständen eines anderen Mandanten) im gleichen Netzsegment durchzuführen. Diese müssen unbedingt verhindert werden, um ein Übergreifen zu verhindern.

Nach einer Beschreibung der Testprozeduren, die die wichtigsten bekannten Angriffsszenarien aus dem Kapitel 3 widerspiegeln, werden zum einen im Kapitel 4.2 der in VMware integrierte Distributed vSwitch und zwei Alternativprodukte anhand den Testprozeduren getestet und der nicht-funktionalen Anforderungen aus Kapitel 2.2 bewertet. Zum anderen werden, anhand der gleichen Testmethodik, tiefergreifende strukturelle Änderungen am bisherigen Systemaufbau auf ihre Eignung geprüft.

Als Quintessenz aus den beschriebenen Angriffen lässt sich zusammenfassen, dass eine sichere Netzinfrastruktur die Verbreitung von nicht der Policy entsprechenden Paketen frühzeitig unterbinden muss, so dass diese kein unter Umständen anfälliges System erreichen können.

4.1 Funktionale Anforderungen und Testprozeduren

Alle praktischen Tests wurden in einer Testinfrastruktur durchgeführt. Diese war bis zu den physischen Switches (B) deckungsgleich mit dem Produktivnetz der Virtualisierungsinfrastruktur am LRZ (Abbildung 2.1). Die beiden VMware-Hosts bestanden aus zwei Sun Fire X4200, die mit jeweils einer 10-Gigabit-Ethernet-Leitung an die beiden VMware-Switches angeschlossen wurden. Die Funktionalität der in der normalen Infrastruktur zwischengeschalteten Flex-10 Module (C) konnte in diesem Kontext nicht evaluiert werden, da ein Test dieser Module die Außerbetriebnahme eines ganzen Bladecenters mit 16 Hosts bedurft hätte. Sofern die Konfiguration der Flex-10 Module Auswirkungen auf die Testergebnisse haben könnte, wird dies in der Folge erwähnt.

Die sich ergebende Netztopologie ist in Abbildung 4.1 zu finden.

Die beiden Hosts wurden mit dem auch produktiv verwendeten kommerziellen Hypervisor VMware ESXi 5.0 betrieben. Die Konfiguration mit getrennten VLANs für Management, die Speicheranbindung und vMotion wurde von der produktiven Infrastruktur übernommen. Die benutzten VLANs und die zugehörigen IP-Adressbereiche sind in Tabelle 4.1 dokumentiert.

VLAN-ID	Nutzung	IPv4-Subnetz	IPv6-Subnetz
73	Nutzerverkehr	10.156.252.0/25	2001:4CA0:0:10D::/64
69	Management vMotion	10.156.252.128/26	n/a
70	Storage	10.156.252.192/26	n/a

Tabelle 4.1: Verwendete VLANs und IP-Adressen im Testaufbau

Auf der Testinfrastruktur werden mehrere virtuelle Maschinen, basierend auf Debian 7.0 (Codename Wheezy), bereitgestellt, die je nach Test als Angreifer oder angegriffene Systeme fungieren können. Sie befinden sich alle in einem gemeinsamen VLAN 73 und gemeinsamen IPv4/IPv6-Subnetz. Besonderes Augenmerk verdient die Platzierung der Container auf den beiden Hosts. Hierbei sollen sowohl Angriffe innerhalb eines Hosts (der Verkehr wird dabei rein innerhalb des VMware vSwitch behandelt) als auch über Hostgrenzen hinweg durch

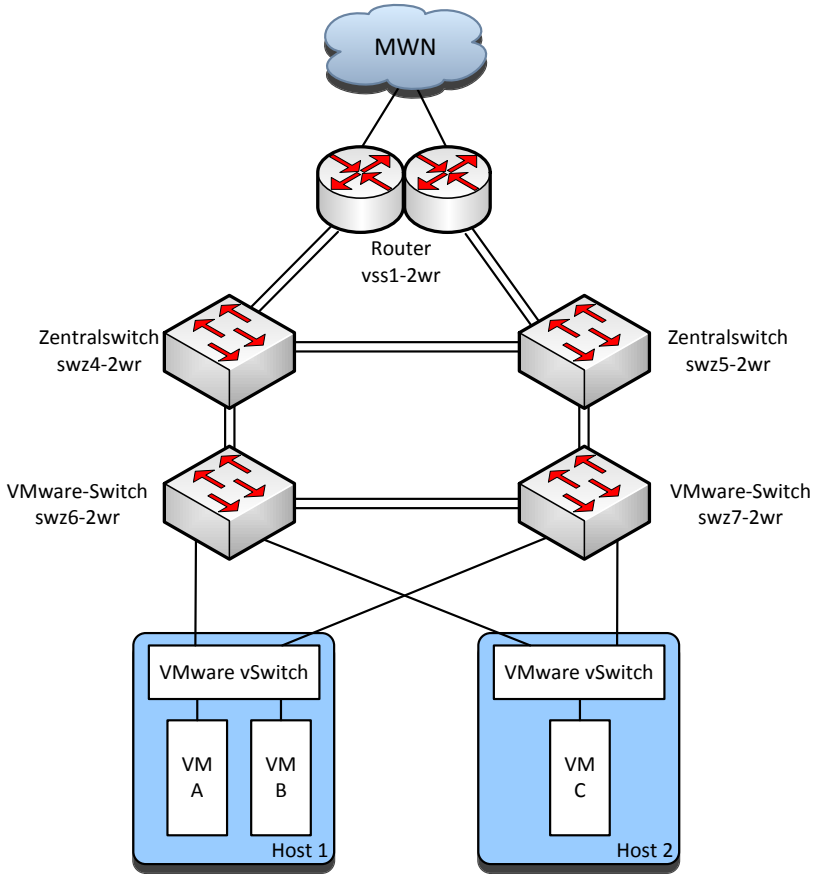


Abbildung 4.1: Netztologie Testaufbau

die physische Infrastruktur geprüft werden. Es wird daher ein angegriffenes System (VM B) zusammen mit dem Angreifer (VM A) auf dem Host 1 platziert, und eine weitere virtuelle Maschine (VM C) als weiteres angegriffenes System auf dem Host 2.

Die Konfiguration der virtuellen Maschinen ist in Tabelle 4.2 zu finden. Als IPv6-Adresse wird nur die globale, durch SLAAC aus der MAC-Adresse gebildete Adresse aufgeführt. Die Link-Local-Adressen und durch Privacy Extensions gebildeten temporären Adressen sind hier nicht gelistet.

Bezeichnung	Hostname	MAC-Adresse	IPv4-Adresse	IPv6-Adresse
A	lxbscDA-01	00:50:56:8e:36:41	10.156.252.1	2001:4ca0:0:10d:250:56ff:fe8e:3641
B	lxbscDA-02	00:50:56:8e:36:42	10.156.252.2	2001:4ca0:0:10d:250:56ff:fe8e:3642
C	lxbscDA-03	00:50:56:8e:36:43	10.156.252.3	2001:4ca0:0:10d:250:56ff:fe8e:3643

Tabelle 4.2: IP- und MAC-Adressen der Test-VMs

4.1.1 Data Link Layer (Schicht 2)

L2-1 – Verhinderung von MAC-Spoofing

Die Grundlage mehrerer Angriffsvarianten ist der Versand von Paketen mit einer gefälschten Absendeadresse auf Schicht 2 (siehe beispielsweise das MAC-Spoofing in Kapitel 3.1.1). Eine sichere Infrastruktur verhindert

daher den Versand von Ethernet-Frames mit MAC-Adressen, die nicht dem jeweiligen Container zugeordnet sind. Dabei darf ausschließlich die Konfiguration des Hypervisors von Belang sein, da der Angreifer unter Umständen Superuser-Berechtigungen innerhalb der virtuellen Maschine erlangen kann.

Da es in Einzelfällen, zum Beispiel beim Einsatz von Bridging innerhalb des Containers, eine legitime Anwendung für den Versand von Ethernet-Frames mit fremden MAC-Adressen gibt, sollte dieser Filter im Hinblick auf die Mandantenfähigkeit für einzelne Container abschaltbar sein.

Testprozedur

- Auf VM A
  - Senden eines Frames mit gefälschter Absenderadresse `00:11:22:33:44:55` mit dem Tool *packETH* (Abbildung 4.2)
  - Setzen der MAC-Adresse `FE:FF:FF:FF:FF:FF` mittels `ip link set eth0 mac FE:FF:FF:FF:FF:FF`
  - Übertragung eines Frames
- `tcpdump -n -s 0 -i eth0 ether host FE:FF:FF:FF:FF:FF` auf VM B und VM C
- `show mac address-table vlan 73` auf Switches

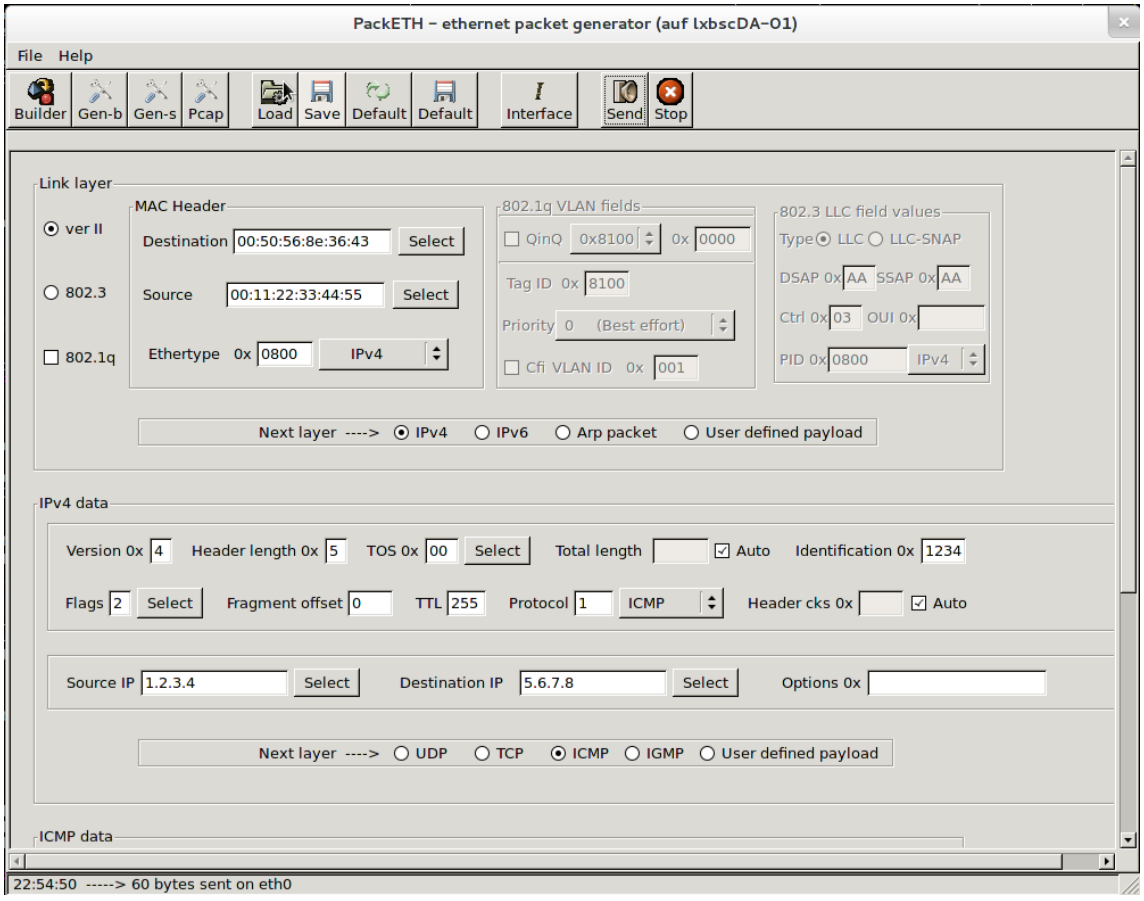


Abbildung 4.2: packETH Testsetup MAC-Spoofing

Bewertung

- Frame mit Source-MAC `FE:FF:FF:FF:FF:FF` darf nicht auf VM B und VM C empfangen werden
- MAC `FE:FF:FF:FF:FF:FF` darf nicht in Switch-Tabelle auftauchen

L2-2 – Versand von Kontrollpaketen

Ein anderer, für den stabilen Betrieb der Infrastruktur ausgesprochen gefährlicher Angriff ist der Versand von unautorisierten Paketen mit Managementprotokollen. Für einen sicheren Betrieb muss daher die Infrastruktur den Versand der folgenden Pakete aus Containern verhindern.

- Spanning-Tree BPDU
- 802.1Q VLAN-Tags
- CDP und LLDP

Insbesondere bei BPDUs ist eine transparente Weitergabe explizit nicht wünschenswert, da der Empfang von fremden BPDUs auch tief in der Netzinfrastruktur Folgen haben kann. Besonders gefährlich werden Spanning Tree BPDUs, wenn sie innerhalb eines 802.1Q VLANs transportiert werden, wie es auch beim Cisco-proprietären *Per-VLAN Rapid Spanning-Tree* (PVRSTP) der Fall ist. Dadurch können sie unter Umständen durch viele Switchebenen unangetastet verteilt werden, bevor sie aufgrund eines Konfigurationsfehlers oder Firmware-Bugs eine Wirkung zeigen.

Testprozedur

Aufgrund der Gefahr für die Betriebssicherheit des Produktivnetzes dürfen diese Tests nur in einer getrennten Infrastruktur durchgeführt werden. Hierbei wird anstatt des Uplink-Switches ein Notebook mit dem Programm Wireshark verwendet

- Auf VM A
  - Versand von LLDP-Paketen mit dem Paket `lldpd`
  - Versand von STP-BPDUs mit einer Linux-Bridge, konfiguriert mit `brctl stp br0 on`
  - Versand von mit 802.1Q Header versehenen Paketen mittels `packETH`
- `wireshark` auf dem Uplink-Port des VMware-Hosts
- `tcpdump` auf VM B und C

Bewertung

- Weder auf VMs noch auf dem Uplink-Port dürfen Kontrollpakete sichtbar sein

L2-3 – Flooding („Storm-Control“)

Eine weitere Schwachstelle sind hohe Paketraten, die auf der Ethernet-Schicht als Broad- oder Multicast abgebildet werden.

Broadcast-Pakete werden nicht nur auf allen Switchports im gleichen VLAN versendet (und belegen dort Bandbreite), sondern werden auch von jedem System empfangen und verarbeitet. Dies kann starke Belastungen hervorrufen, insbesondere wenn durch Virtualisierung auf einem Host viele Gast-VMs die Pakete jeweils seperat verarbeiten müssen. Zu den wenigen legitimen Anwendungen für Broadcast-Pakete gehört ARP zur Adressauflösung. Hierbei sind jedoch auch in Ausnahmefällen nur wenige Pakete pro Sekunde zu erwarten. Eine sichere Konfiguration erlaubt daher, die Rate von durch die Gast-VM versendeten Broadcast-Pakete auf einen niedrigen Wert festzusetzen. Eine mandantenfähige Konfiguration ist hierbei nicht nötig.

Ähnlich stellt sich die Situation bei Multicast-Paketen dar. Auch diese werden (bei Abwesenheit von Mechanismen wie IGMP/MLD-Snooping) auf allen Switchports versendet und belegen dort Bandbreite, auch wenn gar kein Empfänger für diesen Datenstrom dort angeschlossen ist. Sie werden im Allgemeinen bereits durch den Hardware-Filter der Netzwerkkarte gefiltert, wenn keine lokale Anwendung an dieser spezifischen Gruppe Interesse angemeldet hat (Join). Allerdings gilt zu beachten, dass im Virtualisierungsumfeld sowohl der virtuelle Switch des Hypervisors als auch die Netzwerkkarte des Containers durch den Prozessor emuliert wird und daher durchaus Belastungsspitzen auftreten können. Neben lokalen Anwendungen wie *IPv6 Neighbor Discovery*, welches ähnlich zu ARP mit sehr geringen Paketraten arbeitet, kann es jedoch gerade im Multimedia-

und Streamingbereich legitime Anwendungen von auch hochbitratigem Multicast-Verkehr geben. Daher sollte es hier möglich sein, für jeden Container Ausnahmen von der Beschränkung zu konfigurieren.

Testprozedur

Aufgrund der Gefahr für die Betriebssicherheit des Produktivnetzes dürfen diese Tests nur in einer getrennten Infrastruktur durchgeführt werden. Hierbei wird anstatt des Uplink-Switches ein Notebook mit dem Programm Wireshark verwendet

- Auf VM A
  - Versand von großen Broadcast-Pings mit `ping -b -f -s 1000 255.255.255.255`
  - Versand von großen Multicast-Pings mit `ping -b -f -s 1000 239.1.1.1`
- `wireshark` auf dem Uplink-Port des VMware-Hosts
- `tcpdump` auf VM B und C

Bewertung

- Die Datenrate der Broadcast-Pakete muss auf weniger als 1 Mbps beschränkt sein
- Die Datenrate der Multicast-Pakete muss standardmäßig auf weniger als 1 Mbps beschränkt, aber pro VM oder VLAN konfigurierbar sein (Anforderung der Mandantenfähigkeit)

L2-4 – Versand unbekannter Ethertypes

Jedes Ethernet-Frame enthält ein 16-Bit Feld (Ethertype), welches das in den Nutzdaten enthaltene Protokoll beschreibt. Diese werden durch die IEEE öffentlich festgelegt [IEEE-Ethertype]. Eine Auswahl wichtiger Werte ist in Tabelle 4.3 zu finden.

Ethertype	Protokoll
0x0800	IPv4
0x0806	ARP
0x8100	IEEE 802.1Q VLAN-Tag
0x86DD	IPv6
0x88CC	LLDP

Tabelle 4.3: Gängige Ethertype-Werte

In den heutigen Rechenzentren werden, von sehr wenigen Ausnahmen abgesehen, nur noch IPv4 und IPv6 verwendet. Die Nutzung anderer Protokolle wie NetBEUI, IPX/SPX oder AppleTalk ist stark rückläufig. Sie sind entweder gar nicht erst zum Transport über Router geeignet (NetBEUI) oder eine entsprechende Konfiguration ist absichtlich nicht vorgenommen (IPX). Eine Benutzung innerhalb des Münchner Wissenschaftsnetzes ist also ausgeschlossen.

Gleichzeitig besitzen viele Kontrollpakete, wie sie in **L2-2 – Versand von Kontrollpaketen** beschrieben werden, dedizierte Ethertypes. Eine Möglichkeit, die Sicherheit der Infrastruktur zu erhöhen kann also sein, eine Liste der erlaubten Ethertypes zu konfigurieren, die von einer virtuellen Maschine versendet werden können.

Testprozedur

- Auf VM A
  - Versand eines Broadcast-Pakets mit dem Ethertype 0x1234 durch das Paket `packETH`
- `tcpdump` auf VM B und C

Bewertung

- Eine Whitelist für die Ethertypes 0x0800 (IPv4), 0x0806 (ARP) und 0x86DD (IPv6) muss konfigurierbar sein

- Paket mit Ethertype 0x1234 darf nicht auf VM B und VM C sichtbar sein

4.1.2 Address Resolution (Schicht 2.5)

L25-1 – Verhindern von ARP/ND-Spoofing

Zur Verhinderung von Angriffen auf ARP beziehungsweise das ihm eng verwandte *IPv6 Neighbor Discovery*, wie es im Kapitel 3.2.1 beschrieben wurde, müssen ARP-Antworten (oder Neighbor Advertisements) mit falschen IP/MAC-Kombinationen verworfen werden, ehe sie einen anderen Teilnehmer im Netz erreichen können.

Diese Anforderung setzt einen Paradigmenwechsel voraus. Gemäß dem ISO/OSI-Schichtenmodell muss ein Switch als Gerät der Schicht 2 keine Informationen über das Protokoll oder gar den Kommunikationsinhalt der übergeordneten Schichten besitzen. Gleichzeitig müssen aber ARP/ND-Antworten über ihre Adresse im Schicht 3-Protokoll möglichst nah am Verursacher verifiziert und gegebenenfalls verworfen werden. Aus diesem Grund besitzen viele Komponenten heute die Möglichkeit, bestimmte Protokolle des Network Layers mitzulesen und Informationen daraus zu extrahieren. Ein Beispiel dafür sind neben IGMP-Snooping für die Optimierung von IPv4-Multicast-Verkehr auch DHCP-Snooping, um gültige IP/MAC-Kombinationen für Sicherheitsfunktionen zu lernen.

Da DHCP in Servernetzen nur selten zur Anwendung kommt sollten andere Methoden vorhanden sein, um gültige IP/MAC-Kombinationen zu konfigurieren. Im IPv6-Bereich kommt erschwerend dazu, dass bei der oft eingesetzten *Stateless Address Autoconfiguration* (SLAAC) keine zentrale Adresszuweisung mehr erfolgt, die konfigurierten Adressen sind daher erst bei der ersten Verwendung bekannt und können nicht aus einer authoritativen Quelle bezogen werden. Zusätzliche Probleme machen die IPv6 Privacy Extensions [RFC4941], bei denen ein Endsystem in kurzen, periodischen Abständen zufällige Adressen generiert und diese für Verbindungen verwendet. Die Nutzung von SLAAC in Servernetzen ist jedoch umstritten und kann im Umfeld der LRZ-Virtualisierungsinfrastruktur für Kunden unter Umständen vermieden werden.

Testprozedur

- Periodische *Echo Requests* auf IPv4/IPv6-Adresse von VM B
- Auf VM A
  - `tcpdump -n -s 0 -i eth0 host <IPv4-B> or host <IPv6-B>`
  - Versand eines Promiscuous ARP-Pakets für IPv4-Adresse von VM B mit `arping -U -I eth0 <IPB>`
  - Versand eines Unsolicited ND-Pakets für IPv6-Adresse von VM B durch Hinzufügen der Adresse
- `tcpdump` und Ausgabe von `ip neigh` auf VM C
- Beobachten der Neighbor-Tabelle am Default-Gateway

Bewertung

- Pakete an VM B dürfen nicht auf VM A empfangen werden
- Gefälschte Pakete dürfen nicht auf anderen Systemen empfangen werden, falsche Einträge (IP B, MAC A) dürfen nicht in der Neighbor-Tabelle von VM C sichtbar sein
- Welche Methoden stehen zur Konfiguration von erlaubten IP-Adressen zur Verfügung? Was ist die maximale Anzahl von Adressen pro Container?

Wenn die gefälschten ARP-Pakete nur auf dem Standardgateway sichtbar sind gilt der Test als *bedingt bestanden*, da auf dem zentralen Gerät Maßnahmen gegen die Akzeptanz gefälschter ARP-Einträge leichter durchgeführt werden können.

4.1.3 Network Layer (Schicht 3)

L3-1 – Verhindern von IP-Spoofing

Wie bereits im Kapitel 3.3.1 genannt können IP-Adressen zum einen innerhalb des Subnetzes („on-link“) und zum anderen in einem globalen Fokus („off-link“) gefälscht werden.

Fremde IP-Adressen außerhalb des lokalen Subnetzes können durch die hohe Verbreitung von Anti-Spoofing-Filtern auf Routern (uRPF[BCP 38]) nur zum Angriff auf Systeme innerhalb des gleichen VLANs verwendet werden. Sofern nicht gleichzeitig durch andere Angriffe eine Man-in-the-Middle-Situation hergestellt wird sind diese Angriffe selten kritisch. Gefährlicher ist hingegen das Fälschen einer fremden Adresse aus dem gleichen Subnetz, da diese die üblichen Anti-Spoofing-Filter auf dem Router überwinden und weltweit versendet werden können.

Das Verhindern von IP-Spoofing hat sehr ähnliche Voraussetzungen wie **L25-1 – ARP-Spoofing**, da auch hier ein Gerät der Schicht 2 Informationen über gültige IP/MAC-Kombinationen vorhalten muss, um empfangene Pakete zu validieren.

Testprozedur

- Auf VM A
  - Setzen einer fremden IPv4- und IPv6-Adresse im gleichen Subnetz
  - Versand von ICMPv4/ICMPv6 Echo Requests auf VM B, VM C und ein Ziel außerhalb des Subnetzes
  - Setzen einer fremden IPv4- und IPv6-Adresse außerhalb des lokalen Subnetzes
  - Versand von ICMPv4/ICMPv6 Echo Requests auf VM B, VM C und ein Ziel außerhalb des Subnetzes
- `tcpdump` auf VM B, VM C und Ziel außerhalb des Subnetzes

Bewertung

- Pakete mit gefälschtem Absender dürfen auf keinem Ziel sichtbar sein
- Die eingesetzten Filter müssen mit der Existenz von mehreren IPv4/IPv6-Adressen pro Container kooperieren können, ohne ihre Sicherheitswirkung zu verlieren

L3-2 – Rogue DHCP

Testprozedur

- Anbindung des Testnetzes an die produktive LRZ-DHCP-Infrastruktur mittels DHCP-Relay
- Auf VM A
  - Installation des ISC dhcpd im IPv4-Modus
  - Installation des ISC dhcpd im IPv6-Modus
- `dhclient` auf VM B und VM C

Bewertung

- DHCP-Clientsoftware auf VM B und VM C darf nur authorisierte Antwort vom LRZ DHCP-Server erhalten

L3-3 – RA Guard

Testprozedur

- Auf VM A
  - Versand eines *Unsolicited Router Advertisements* durch `radvd`
- `tcpdump` auf VM B und VM C

Bewertung

- VM B und VM C dürfen das Paket nicht empfangen

Sofern nur das Default-Gateway dieses Paket empfangen kann (durch das Kommando `show ipv6 routers` zu überprüfen) gilt der Test als bestanden, da Router Advertisements als Konfigurationshilfsmittel für Endpunkte (Hosts) von Routern generell nicht benutzt werden.

4.1.4 L4 – Paketfilter und Firewall

Dieser Punkt evaluiert die Verfügbarkeit von Filtern auf den Schichten 3 und 4, also bei IP- und UDP/TCP-Verbindungen. Diese werden im Allgemeinen durch Firewalls oder Paketfilter bereitgestellt.

Zur vollen Erfüllung dieses Testpunkts muss es möglich sein, **Verbindungen** (dies impliziert eine stateful Firewall) von externen IP-Adressen auf spezifische Dienste (Ports) der VM zu unterbinden.

Stehen nur Paketfilter ohne Verbindungsanalyse (stateless) zur Verfügung, ist dieser Punkt teilweise erfüllt.

Diese Konfiguration muss durch den Kunden selbst veränderbar sein, da Änderungen häufig vorgenommen werden müssen.

4.2 Produktevaluation Hypervisor-Switch

Zunächst werden neben dem in VMware standardmäßig integrierten vSwitch (in der Konfigurationsvariante *Distributed vSwitch*, wie er am LRZ eingesetzt wird, das Konkurrenzprodukt *Nexus 1000V* der Firma Cisco Systems evaluiert. Dies geschieht unter der Prämisse, dass das in Kapitel 2.1 beschriebene Szenario (und insbesondere die Platzierung von virtuellen Maschinen unterschiedlicher Kunden in Sammel-VLANs gemäß dem Unterkapitel 2.1.2) beibehalten werden soll. Hierbei sollen keine umfassenden Umbaumaßnahmen am Netz nötig sein.

Neben den beiden genannten Produkten ist auf dem Markt nur noch der DVS 5000V [IBM 11] der Firma IBM verfügbar. Dieser kann jedoch in der bestehenden Infrastruktur nicht eingesetzt und auch nicht getestet werden, da dieser die Sicherungsaufgaben durch das 802.1Qbg-Protokoll an den physischen Switch übergibt und selbst nur die Ethernet-Frames durchleitet. Da dieser Standard von keiner im LRZ eingesetzten Netzkomponente unterstützt wird, kann dieses Produkt nicht evaluiert werden. Die öffentlich verfügbaren Informationen sind bei weitem nicht detailliert genug, um die Eignung für den Einsatz in der Virtualisierungsinfrastruktur des LRZ einzuschätzen.

Es kann davon ausgegangen werden, dass in den nächster Zeit weitere Alternativprodukte auf den Markt kommen werden. Insbesondere die aufkommende OpenFlow-Thematik zur herstellerübergreifenden Anbindung zwischen *Control Plane* (Routingprotokolle, Management) und *Forwarding Engine* (Datenverkehr) wird dabei vermutlich einen deutlichen Zugewinn an Flexibilität erlauben, die sich viele Hersteller zu Nutze machen werden. Ein Ausblick auf diese Technologien ist im Kapitel 7.2 zu finden.

4.2.1 VMware dvSwitch

Der virtuelle Switch (*vSwitch*) von VMware ESXi ermöglicht klassischerweise als Hypervisor-Switch die Kommunikation von Containern auf der einen Seite mit den physischen Uplink-Interfaces der Hosts auf der

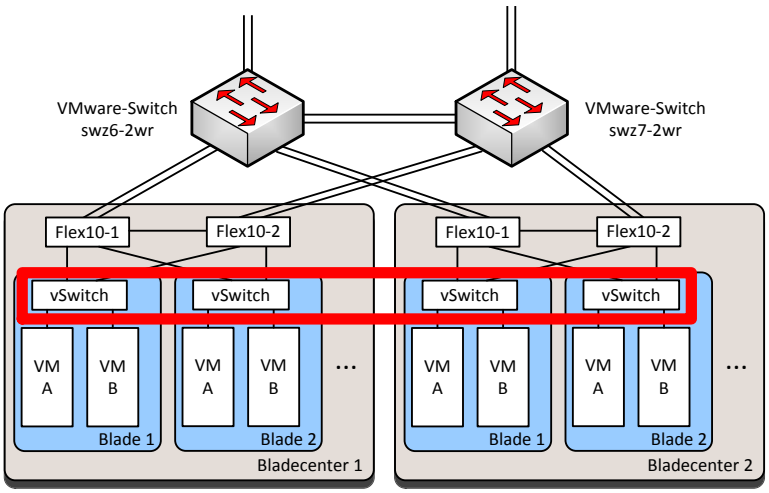


Abbildung 4.3: Hypervisor-Switch

anderen Seite, kann jedoch auch rein intern verwendet werden. Der vSwitch ist in mehrere Portgruppen unterteilt, die einem bestimmten Konfigurationsprofil entsprechen. Eine Portgruppe entspricht im Allgemeinen einem externen VLAN. Eine Portgruppe enthält wiederum (virtuelle) Ports, die mit den virtuellen Netzwerkkarten der Container verbunden sind.

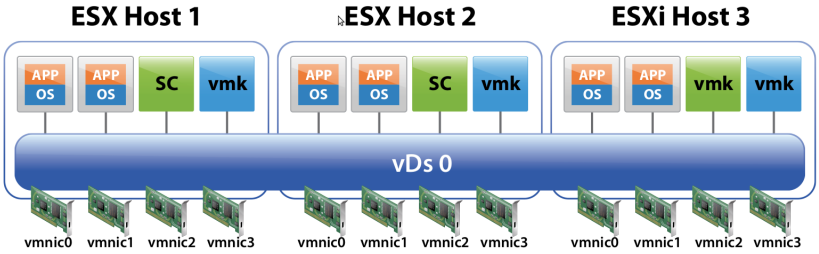


Abbildung 4.4: Konzept eines VMware dvSwitch Quelle: VMware Inc. [VMdvS]

Der vSwitch unterstützt in der vorliegenden Version von ESXi 5.0 die folgenden Features:

- VLAN-Trunk auf dem Uplink und Zuweisung eines ungetaggten Access-VLANs für Container (= Portgruppe)
- Empfang und Senden von *Cisco Discovery Protocol*-Paketen zur Identifizierung des verbundenen Geräts auf den Uplink-Interfaces
- Traffic-Shaping für durchschnittliche Bandbreite und Spitzenbandbreite eines Ports eines Containers
- Redundanz und Lastverteilung bei mehreren Uplink-Interfaces
- Einschränkung von Layer 2-Änderungen
  - Promiscous Mode – Empfang von Paketen an fremde Unicast-MACs
  - MAC-Adressänderungen – Setzen einer benutzerdefinierten MAC-Adresse auf die Netzschnittstelle innerhalb des Containers
  - Gefälschte Übertragungen – Senden von Paketen mit MAC-Adressen, die nicht der im Container gesetzten MAC-Adresse entsprechen

- Konfigurierbarkeit aller zuletzt genannten Parameter pro Portgruppe oder pro Container

Am LRZ kommt der vSwitch in der Konfigurationsvariante *dvSwitch – Distributed vSwitch* zum Einsatz. Dieser verfügt nicht über mehr Funktionen, sondern erleichtert die Verwaltung über mehrere Hosts hinweg durch eine zentrale Konfiguration. Auf die Sicherheitsarchitektur hat dies jedoch keinen Einfluss.

Wie bereits an der Aufzählung der Fähigkeiten zu sehen ist, verfügt der vSwitch kaum über Fähigkeiten, die über die Schicht 2 hinausgehen. Insbesondere verfügt er über keine Sicherheitsmechanismen, für die er auf Informationen der Schicht 3 (zum Beispiel IP-Adressen) zugreifen müsste.

Der Konfigurationsparameter „Promiscous Mode“ definiert, ob die Portgruppe des vSwitch bei einer Anforderung des Promiscous Mode durch das Betriebssystem (beispielsweise durch `ip link set ... promisc on` oder durch tcpdump) in den Promiscous Modus gesetzt wird und alle Pakete in dieser Portgruppe, auch wenn sie nicht an die MAC-Adresse des Containers geschickt werden, auf dem Gast sichtbar sind. Dies betrifft jedoch selbstverständlich nur Pakete, die bereits aus anderen Gründen über diesen lokalen vSwitch geschickt werden. Dies kann bei Broad- und Multicast-Paketen der Fall sein oder bei Frames, die an einen anderen Gast auf dem gleichen Host geschickt werden. Sie ermöglichen jedoch, auch im *dvSwitch*, kein Abhören von Paketen, die nur Gastsysteme auf anderen Hosts betreffen. Das Setzen des Promiscuous Mode durch einen Container wird nicht protokolliert.

Wird der Konfigurationsparameter für „MAC-Adressänderungen“ auf „Ablehnen“ gesetzt, so werden Pakete nach dem Umsetzen der MAC-Adresse mit `ip link set ...` beim Versand am Eingangsport der VM am vSwitch verworfen. Der Zähler „Verworfen - Ingress-Pakete“ zählt dabei die verworfenen Pakete. In diesen Zähler fließen jedoch auch andere verworfene Pakete, die durch Überlast oder andere Sicherheitseinstellungen entstehen, ein, so dass er nicht zur Überwachung herangezogen werden kann. Eine dedizierte Möglichkeit einen Logeintrag zu generieren bietet VMware nicht an.

Der Konfigurationsparameter „Gefälschte Übertragungen“ regelt unabhängig davon den Versand von Paketen von MAC-Adressen, die nicht der Netzschnittstelle zugeordnet sind (sei es durch VMware selbst oder durch eine MAC-Adressänderung des Nutzers (wenn erlaubt)). Verworfene Pakete werden nicht protokolliert und tauchen, im Gegensatz zu MAC-Adressänderungen, auch nicht in den Statistiken auf.

Für einen sicheren Betrieb müssen alle drei Parameter auf „Ablehnen“ gesetzt werden, wie in Abbildung 4.5 gezeigt wird.

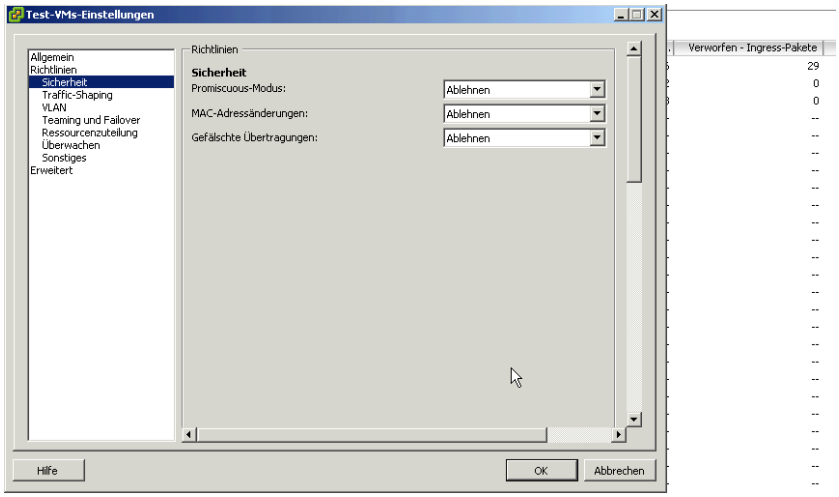


Abbildung 4.5: Empfohlene Einstellungen auf VMware Portgruppen

Alle Parameter können bei Bedarf sowohl auf der gesamten Portgruppe als auch auf den den einzelnen Ports gesetzt werden. Sie gehen jedoch verloren, wenn der Container von der Portgruppe gelöst wird. Dies kann beispielsweise beim Verschieben in eine andere Portgruppe, wechseln des virtuellen Netzadapters oder beim Löschen und Neuimportieren einer VM passieren.

Unter Berücksichtigung der empfohlenen Konfiguration ergibt sich damit die folgende Bewertung bezüglich den funktionalen Anforderungen.

L2-1 – Verhinderung von MAC-Spoofing
Voll erfüllt, wenn „MAC-Adressänderungen“ und „Gefälschte Übertragungen“ deaktiviert sind.

L2-2 – Versand von Kontrollpaketen
Voll erfüllt, alle Kontrollpakete werden unterbunden, verworfene Pakete werden aber nicht nicht protokolliert.

L2-3 – Flooding („Storm-Control“)
Nicht erfüllt, eine Einschränkung ist nur bezogen auf die Gesamtbandbreite einer VM möglich, aber nicht dediziert für Broad- und Multicastpakete

L2-4 – Versand unbekannter Ethertypes
Nicht erfüllt, eine Einschränkung ist nicht möglich

L25-1 – Verhinderung von ARP-Spoofing
Nicht erfüllt, keine Layer3-Funktionalität

L3-1 – IP-Spoofing
Nicht erfüllt, keine Layer3-Funktionalität

L3-2 – Rogue DHCP
Nicht erfüllt, keine Layer3-Funktionalität

L3-3 – RA Guard
Nicht erfüllt, keine Layer3-Funktionalität

L4 – Paketfilter und Firewall
Nicht erfüllt, keine Layer3/4-Funktionalität

4.2.2 Cisco Nexus 1000V

Das Produkt Nexus 1000V der Firma Cisco Systems ist ein Ersatz für den in VMware integrierten vSwitch. Es besteht aus den Komponenten VEM (Virtual Ethernet Module), welches als tatsächlicher Ersatz des vSwitch in den Hypervisor auf jedem Host geladen wird, und einer übergeordneten Managementinstanz namens VSM (Virtual Supervisor Module). Die VEM verarbeiten dabei den tatsächlichen Datenverkehr, während die (gegebenenfalls redundanten) VSMs nur Änderungen an Topologie und Konfiguration verarbeiten. Sie entsprechen damit konzeptionell der bei großen Netzkomponenten praktizierten Trennung von *Control Plane* (Routingprotokolle, Management) und *Forwarding Engine* (Datenverkehr). Zusammen ermöglichen diese Komponenten ein zentrales Management von verteilten Switching-Instanzen ähnlich dem *Distributed vSwitch*.

Der Cisco Nexus 1000V unterstützt in der zum Zeitpunkt der Arbeit aktuellen Version 4.2(1)SV2 unter anderem die folgenden für diese Arbeit relevanten Funktionen [N1V-Datasheet]:

- Port Security
- IP Source Guard, Dynamic ARP Inspection, DHCP Snooping
- VXLAN
- Private VLAN
- IGMP Snooping
- BPDU-Filter
- Ingress- und Egress ACLs
  - Layer2: Source-MAC, Destination-MAC, Ethertype, VLAN
  - Layer3/4: Source-IPv4, Destination-IPv4, L4-Protokoll, Source-Port, Destination-Port

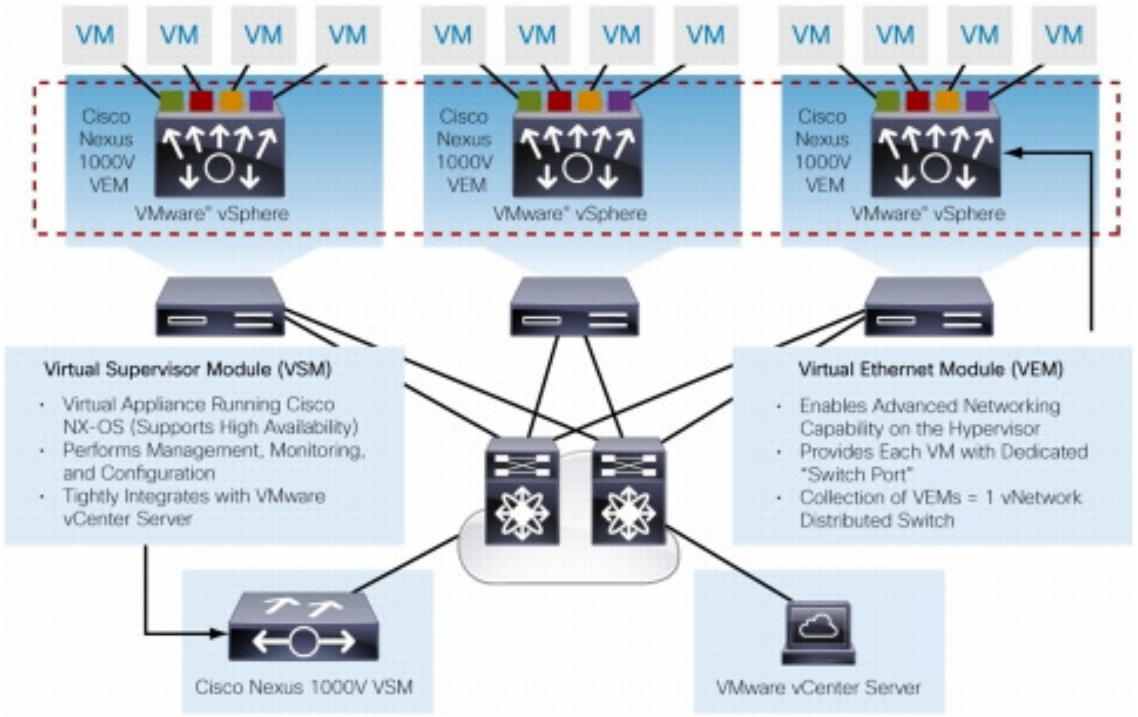


Abbildung 4.6: Nexus 1000V Architektur, Quelle: Cisco Systems [N1V-Datasheet]

Eine im Jahr 2012 unerwartete Lücke besteht im Bereich IPv6, welcher durch das Produkt in der aktuellen Version noch gar nicht unterstützt wird. Eine entsprechende Unterstützung ist auf der Cisco-internen Roadmap erst für die im zweiten Quartal des Jahres 2013 erwartete Version mit dem Codenamen „Dao“ zu erwarten.

Der Nexus 1000V basiert laut den Werbeaussagen von Cisco auf dem Betriebssystem NX-OS, welches auf einem Linux-Kernel aufsetzt [N1V-Datasheet]. Dieses Betriebssystem wird unter anderem auch in den physischen Hochleistungsgeräten Nexus 7000 verwendet, die für den zukünftigen Betrieb im Münchner Wissenschaftsnetz ab dem Jahr 2013 eingesetzt werden. Ob jedoch tatsächlich auf diesen sehr unterschiedlichen Plattformen (auf dem Nexus 7000 werden die meisten zeit- und durchsatzkritischen Funktionen durch spezialisierte Hardware erbracht, während im Nexus 1000V alle Funktionen durch die CPU erbracht werden müssen) die gleiche Softwarebasis läuft oder nur ein ähnliches Bedienkonzept durch die gewohnte Cisco-CLI vorliegt ist ohne tieferen Einblick in die Softwareentwicklung nicht festzustellen.

Die Konfiguration des Nexus 1000V entspricht konzeptionell einem physischen Cisco-Switch, an dessen Ports jeweils eine einzelne virtuelle Netzwerkkarte und damit ein einziger Gast angebunden sind. Wie alle Cisco-Switches sind auf dem Nexus 1000V im Auslieferungszustand nahezu keine Sicherheitsfunktionen aktiv, mit Ausnahme von L2-2 sind alle beschriebenen Testszenarien durchführbar. Er bietet jedoch weitreichende Konfigurationsmöglichkeiten zur Absicherung des Verkehrs. Hierzu werden auf oberster Ebene sogenannte *Port Profile* definiert, die in etwa mit den aus dem VMware vSwitch bekannten Portgruppen vergleichbar sind. Neben der Definition des zugewiesenen VLANs können hier Sicherheitsmechanismen wie Filter (ACLs) und MAC-Adresslimitierungen (Port Security) definiert werden, die automatisch für alle Container mit diesem Port Profil angewendet werden. Dem VMware-Administrator stehen dann im vSphere-Client die definierten Port Profile als Portgruppen zur Verfügung.

Aufgrund der umfangreichen Konfigurationsmöglichkeiten wird im Folgenden detailliert auf die Fähigkeiten des Nexus 1000V zur Erfüllung der Anforderungen eingegangen.

L2-1 – Verhinderung von MAC-Spoofing

Der Nexus 1000V unterstützt die auf Cisco-Switches übliche Konfiguration der Port-Security. Er erlaubt dabei die Beschränkung der Anzahl von gelernten MAC-Adressen hinter einem Port (zwischen 1 und 1025) und eine Festlegung von Alterungs-Parametern, um unbenutzte MAC-Adressen wieder aus der Liste zu entfernen. Zusätzlich stehen drei verschiedene Aktionen beim Verstoß gegen diese Grenzen bereit („shutdown“ deaktiviert den virtuellen Switchport, „protect“ und „restrict“ verwerfen das verstoßende Paket mit unterschiedlichen Arten des Protokollierens). Außerdem steht eine sogenannte „sticky“ Einstellung bereit, bei der eine einmal gelernte MAC-Adresse auch über den Reboot einer virtuellen Maschine hinaus in der Konfiguration hinterlegt wird.

Bei all diesen Einstellungen lässt der 1000V jedoch die Fähigkeit des in VMware integrierten vSwitch schmerzlich vermissen, die in der Konfiguration des Containers hinterlegte MAC-Adresse als einzige erlaubte MAC-Adresse zu definieren. Die Konfiguration, die am nähesten die vSwitch-Fähigkeiten emuliert lautet

```
port-profile type vethernet PORTGRUPPE
  switchport port-security violation restrict
  switchport port-security maximum 1
  switchport port-security mac-address sticky
  switchport port-security
```

Dadurch wird die Quell-Adresse des ersten Pakets, welches beim erstmaligen Einschalten einer virtuellen Maschine von dieser versendet wird, gelernt und dauerhaft („sticky“) in die Konfiguration eingetragen. Ein Senden von anderen Quell-Adressen ist nicht mehr möglich, da diese bereits am Switchport verworfen werden. Hierbei wird sowohl ein Logeintrag generiert als auch ein entsprechender Zähler inkrementiert, der über die Kommandozeile und SNMP abgefragt werden kann.

```
VSM %ETH-PORT-SEC-2-ETH_PORT_SEC_SECURITY_VIOLATION_MAX_MAC_VLAN: Port Vethernet5
        moved to RESTRICT state as host 0011.2233.4455 is trying to access the
        port in vlan 73
VSM# show port-security interface vethernet 5
Port Security              : Enabled
Port Status                : Secure UP
Violation Mode             : Restrict
Aging Time                 : 0 mins
Aging Type                 : Absolute
Maximum MAC Addresses      : 1
Total MAC Addresses        : 1
Configured MAC Addresses   : 0
Sticky MAC Addresses       : 1
Security violation count   : 100
```

Die Konfiguration eines vEthernet-Interfaces, und damit die gespeicherte MAC-Adresse, unterliegt jedoch ähnlichen Einschränkungen wie die Konfiguration eines spezifischen Ports einer Portgruppe im vSwitch. Sie ist nur lose mit dem Container verbunden und kann bei bestimmten Aktionen wie dem Exportieren/Importieren oder Migrieren des Containers zurückgesetzt werden. Gelingt es dem Angreifer nun, als erstes Paket nach dem Neustart ein Paket mit einer gespooften MAC-Adresse zu verschicken, so kann er diese Adresse uneingeschränkt nutzen.

L2-2 – Versand von Kontrollpaketen

Der Nexus 1000V unterstützt wie sein VMware-Pendant kein Spanning-Tree, sondern sorgt intern für eine Schleifenfreiheit, indem Pakete von einem physischen Uplink-Port niemals auf einem anderen physischen Uplink-Port versendet werden. Spanning-Tree Kontrollpakete werden in beiden Richtungen verworfen.

Normale Access-Ports (siehe Kapitel 3.1.2) von VMs akzeptieren keine mit einem einfachen 802.1Q-Tag versehenen Pakete. Bei den Tests im Rahmen dieser Diplomarbeit wurde jedoch eine potentielle Sicherheitslücke gefunden. Wird ein Paket mit einem 802.1Q-Tag und der VLAN-ID 0 versendet, so wird dieses Paket dennoch akzeptiert, das VLAN-Tag entfernt und das Paket regulär weitergeschickt. Dies geschieht zur Unterstützung

von Priorisierungsinformationen gemäß dem IEEE 802.1p Standard, der den 802.1Q-Header benutzt und bei Nichtverwendung des VLAN-Taggings die VLAN-ID auf 0 setzt. Leider werden jedoch die Daten, die nach dem Entfernen des 802.1Q-Headers übrig bleiben, nicht mehr regulär geprüft. So ist es insbesondere möglich, durch die Verknüpfung von zwei 802.1Q-Headern ein gesendetes Paket mit einem beliebiges VLAN-Tag zu versehen.

Das folgende Beispiel soll anhand eines Wireshark-Traces das Problem verdeutlichen. Die VM A sendet ein Ethernet-Frame, das aus zwei hintereinander geketteten 802.1Q-Headern besteht. Die VLAN-ID des äußeren Tags ist 0, was dem ungetaggten, das heisst nativen VLAN entspricht. Die VLAN-ID des inneren Tags ist 73.

```
Frame 1: 122 bytes on wire (976 bits), 122 bytes captured (976 bits)
Ethernet II, Src: (00:50:56:8e:36:41), Dst: (00:50:56:8e:36:42)
    Destination: Vmware_8e:36:42 (00:50:56:8e:36:42)
    Source: Vmware_8e:36:41 (00:50:56:8e:36:41)
    Type: 802.1Q Virtual LAN (0x8100)
802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 0
    000. .... .... .... = Priority: Best Effort (default) (0)
    ...0 .... .... .... = CFI: Canonical (0)
    .... 0000 0000 0000 = ID: 0
    Type: 802.1Q Virtual LAN (0x8100)
802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 73
    000. .... .... .... = Priority: Best Effort (default) (0)
    ...0 .... .... .... = CFI: Canonical (0)
    .... 0000 0100 1001 = ID: 73
    Type: Unknown (0x1234)
Data (100 bytes)
```

Empfängt der Nexus 1000V dieses Frame, so wird das äußere VLAN-Tag entfernt und das Frame weiter an den Empfänger geschickt. Hierbei überprüft der Nexus 1000V jedoch nicht, ob das Paket valide Informationen beinhaltet. So ist in diesem Beispielpaket ein 802.1Q-getaggtes Frame enthalten, welches weder von lxbscDA-01 empfangen noch an lxbscDA-02 gesendet werden dürfte, da 802.1Q-Frames auf Accessports nicht erlaubt sind. Die VM B empfängt jedoch das folgende Frame.

```
Frame 1: 118 bytes on wire (944 bits), 118 bytes captured (944 bits)
Ethernet II, Src: (00:50:56:8e:36:41), Dst: (00:50:56:8e:36:42)
    Destination: Vmware_8e:36:42 (00:50:56:8e:36:42)
    Source: Vmware_8e:36:41 (00:50:56:8e:36:41)
    Type: 802.1Q Virtual LAN (0x8100)
802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 73
    000. .... .... .... = Priority: Best Effort (default) (0)
    ...0 .... .... .... = CFI: Canonical (0)
    .... 0000 0100 1001 = ID: 73
    Type: Unknown (0x1234)
Data (100 bytes)
```

Bisher konnte dieses Problem nur zwischen zwei VMs auf dem gleichen Host in der gleichen Portgruppe nachgewiesen werden. Es scheint dadurch nur möglich zu sein, ein getaggtes Frame an eine andere VM im gleichen VLAN zu schicken. Dies sollte nur in den wenigsten Fällen Konsequenzen haben, da die VMs im regulären Betrieb keine VLANs benutzen können und daher auch diese nicht konfiguriert haben. Da unbekannte VLANs im Allgemeinen von den im Container laufenden Betriebssystemen verworfen werden besteht hier keine große Gefahr. Es scheint bislang nicht möglich zu sein, ein Paket tatsächlich in ein fremdes VLAN zu senden und damit eine 12 Jahre alte Sicherheitslücke [Tayl 00] neu zu implementieren. Dennoch wirft dieses Problem kein gutes Licht auf die Implementierung des Nexus 1000V. Ein entsprechender Hinweis wurde an Cisco-Mitarbeiter versendet und löste am 12.11.2012 einen Supportfall aus (SR623813473), der jedoch bis zum Abschluss der Arbeit nicht beantwortet wurde.

Eine möglicher Workaround ist eine explizite Konfiguration von erlaubten Ethertype-Werten, wie sie im Punkt *L2-4 Versand unbekannter Ethertypes* gezeigt wird.

L2-3 – Flooding („Storm-Control“)

Verkehr von und zu einer virtuellen Maschine kann auf dem Nexus 1000V unter anderem anhand der folgenden Kriterien einer QoS-Klasse zugeordnet werden (*class-map*):

- CoS-Feld im 802.1Q-Header (802.1p)
- IPv4 Precedence-Headerfeld (Quality-of-Service Markierung)
- IPv4 DSCP-Headerfeld (Quality-of-Service Markierung)
- IPv4- und MAC-Accessliste
- Paketlänge

Diese QoS-Klassen können in einer *policy-map* einer Bandbreitenbeschränkung unterworfen werden. Die kleinste konfigurierbare Beschränkung beträgt 250 kBit/s. Pakete, die nicht den Richtlinien entsprechen, können verworfen oder mit einer entsprechenden QoS-Markierung (DSCP, Precedence, 802.1p) versehen werden.

Diese Fähigkeiten können dazu benutzt werden, einen effektiven Schutz gegen den exzessiven Versand von Broad- und Multicast-Paketen zu implementieren. Hierzu wird eine MAC-Accessliste angelegt, die alle Broad- und Multicast-Pakete anhand des entsprechenden Bits im Ethernet-Header (siehe Kapitel 3.1).

```
mac access-list ACL-BROADCAST
  10 permit any 0100.0000.0000 feff.ffff.ffff
```

Die ACL *ACL-BROADCAST* wird nun der QoS-Klasse *CLASS-BROADCAST* zugeordnet.

```
class-map type qos match-any CLASS-BROADCAST
  match access-group name ACL-BROADCAST
```

In der Policy-Map *LIMIT-BROADCAST* kann jetzt die QoS-Klasse *CLASS-BROADCAST* mit einem Policer versehen werden, der den Verkehr auf 250 kBit/s beschränkt.

```
policy-map type qos LIMIT-BROADCAST
  class CLASS-BROADCAST
    police cir 250 kbps bc 200 ms conform transmit violate drop
```

Zum Schluss muss die Policy-Map noch jeder Portgruppe zugeordnet werden, für deren Verbindungen sie gelten soll.

```
port-profile type vethernet Testuser
[...]
  service-policy input LIMIT-BROADCAST
```

Nun ist die Policy-Map jedem Container zugeordnet und beschränkt deren Broad- und Multicast-Verkehr. Verstöße gegen das Limit werden individuell pro Port protokolliert.

```
VSM# sh policy-map interface

Global statistics status :   enabled

Vethernet2

  Service-policy (qos) input:   LIMIT-BROADCAST
    policy statistics status:   enabled

    Class-map (qos):   CLASS-BROADCAST (match-any)
      127556 packets
      Match: access-group ACL-BROADCAST
        127556 packets
      police cir 250 kbps bc 200 ms
        conformed 17982 bytes, 54 bps action: transmit
        violated 135388185 bytes, 0 bps action: drop
```

L2-4 – Versand unbekannter Ethertypes

Eine Möglichkeit, den Versand von Paketen mit unerwünschten Protokollen zu unterbinden sind Filter auf Schicht 2, die nur bekannte und gewünschte Protokolle im Ethernet-Header erlauben. Dazu dienen auf dem Nexus 1000V die sogenannten *MAC ACLs*, wie sie auch schon bei der QoS-Konfiguration verwendet wurden.

Eine Access-Liste, die nur IPv4, ARP und IPv6 erlaubt kann schnell definiert werden.

```
mac access-list LIMIT-PROTOCOLS
  statistics per-entry
  10 permit any any ip      ! IPv4
  20 permit any any 0x806   ! ARP
  30 permit any any 0x86dd  ! IPv6
  40 deny any any
```

Auch dieser Filter kann wieder pro Portgruppe oder auch pro Port aktiviert werden.

```
port-profile type vethernet Testuser
  mac port access-group LIMIT-PROTOCOLS in
```

Beim Auslesen der globalen Statistik fällt auf, dass der Zähler für die IPv4-Regel immer bei Null Treffern steht, obwohl IPv4-Verkehr im Testnetz erfolgt.

```
VSM# sh mac access-lists LIMIT-PROTOCOLS

MAC access list LIMIT-PROTOCOLS
        statistics per-entry
        10 permit any any ip [match=0]
        20 permit any any 0x806 [match=4]
        30 permit any any 0x86dd [match=21840]
        40 deny any any [match=31]
```

Weitere Untersuchungen zeigen, dass IPv4-Verkehr nicht von der MAC ACL behandelt wird. Weder ein Löschen der entsprechenden Zeile in der ACL noch das explizite Verbieten von IPv4 zeigt eine messbare Wirkung, IPv4-Verkehr ist für die virtuelle Maschine weiterhin möglich. Dieses Verhalten ist gewollt und in der Dokumentation beschrieben und dient vermutlich dazu, Geschwindigkeitsprobleme und Verwirrung des Administrators durch die doppelte Filterung eines einzigen Pakets zu vermeiden. Eine Änderung dieses Verhaltens ist nicht möglich, was leider das geplante Nutzen dieser Fähigkeiten zur Simulation von Private-VLAN unmöglich macht (siehe auch Kapitel 6).

Ein Verbot anderer Protokolle wirkt hingegen sofort. Auch können mit dieser Konfiguration die in L2-2 beschriebenen Angriffe mit doppeltem VLAN-Tagging nicht mehr durchgeführt werden.

L25-1 – Verhinderung von ARP-Spoofing und L3-1 – IP-Spoofing

Die Funktionalitäten zur Verhinderung von ARP- und IP-Spoofing sind auf Cisco-Geräten eng miteinander gekoppelt und laufen dort unter den Begriffen

- DAI („Dynamic ARP Inspection“) für ARP
- IPSG („IP Source Guard“) für IP

Sie stehen erst mit der kostenpflichtigen *Advanced*-Lizenz zur Verfügung.

Beide Funktionen benötigen zunächst eine Liste von erlaubten VLAN-MAC-IP-Interface-Zuordnungen, welche bei Cisco als „Bindings“ bezeichnet werden. Das Lernen dieser Zuordnungen erfolgen in der Regel über die Analyse von ausgetauschten DHCP-Nachrichten („DHCP-Snooping“). Hierzu müssen alle Maschinen in einem VLAN als DHCP-Client konfiguriert sein sowie auf dem Nexus 1000V das entsprechende Feature für das VLAN aktiviert sein. Außerdem müssen die Ports, von denen korrekte DHCP-Antworten kommen können, als *vertrauenswürdig* („trusted“) markiert werden. Dies ist standardmäßig bei den physischen Uplinks der Fall, eine Anpassung ist daher nur nötig wenn virtualisierte DHCP-Server betrieben werden.

```
svs switch edition advanced
feature dhcp

ip dhcp snooping
ip dhcp snooping vlan 73
```

Sobald die Adresszuweisung durch DHCP erfolgt ist ist ein entsprechender Eintrag im Switch hinterlegt.

```
VSM# sh ip dhcp snooping binding
MacAddress         IpAddress        LeaseSec  Type        VLAN  Interface
-----------------  ---------------  --------  ----------  ----  -------------
00:50:56:8e:36:41  10.156.252.14    3582       dhcp-snoop  73    Vethernet2
```

Die dadurch erlernte Tabelle kann nun von *Dynamic ARP Inspection* verwendet werden, die von virtuellen Maschinen ausgehenden (dies bedeutet, aus Sicht des Nexus 1000V von virtuellen Maschinen eingehenden) ARP-Pakete zu analysieren und die Inhalte gegen die Tabelle valider *Bindings* zu überprüfen.

Die Aktivierung dieser Funktionalität erfolgt pro VLAN in der globalen Konfiguration.

```
ip arp inspection vlan 73
ip arp inspection validate src-mac dst-mac ip
```

Zusätzlich können global noch weitere Konsistenzprüfungen aktiviert werden. Diese umfassen

- **src-mac** vergleicht die (durch Port-Security validierte) Quell-MAC-Adresse des Ethernet-Frames mit der Quell-MAC-Adresse im ARP-Paket
- **dst-mac** vergleicht die Ziel-MAC-Adresse des Ethernet-Frames mit der Ziel-MAC-Adresse im ARP-Paket
- **ip** unterdrückt alle ARP-Pakete, die spezielle IPv4-Adressen wie Broad- oder Multicast-Adressen enthalten

Pro VLAN steht eine Statistik bereit, die erlaubte und unterdrückte ARP-Pakete auflistet. Sie zeigt jedoch keine individuellen Verstöße pro Container an.

```
VSM# sh ip arp inspection statistics

Vlan : 73
-----------
ARP Req Forwarded  = 45
ARP Res Forwarded  = 2
ARP Req Dropped    = 9
ARP Res Dropped    = 21
DHCP Drops         = 30
DHCP Permits       = 2
SMAC Fails-ARP Req = 0
SMAC Fails-ARP Res = 0
DMAC Fails-ARP Res = 0
IP Fails-ARP Req   = 0
IP Fails-ARP Res   = 0
```

Die zweite Funktionalität, die auf der DHCP-Snooping-Datenbank aufsetzt, ist *IP Source Guard*. Hierbei wird die Quell-IPv4-Adresse jedes IP-Pakets, welches durch eine VM verschickt wird, validiert. Die Konfiguration erfolgt in diesem Fall nicht für jedes VLAN, sondern für jede Portgruppe.

```
port-profile type vethernet Testuser
 ip verify source dhcp-snooping-vlan
```

Statistiken über erlaubte und unterdrückte Pakete stehen nicht zur Verfügung, weder global noch für jeden Container.

Eine Unterstützung von IPv6 ist nicht vorhanden. Eine entsprechende Erweiterung ist für das erste Quartal 2013 angekündigt, es bleibt jedoch abzuwarten, ob diese ebenso wie bei IPv4 die Unterstützung von DHCP

benötigt. Die Nutzung von DHCPv6 zur Adressvergabe (*stateful DHCPv6*) ist nur sehr selten anzutreffen, wesentlich üblicher sind statische IPv6-Adressen oder die Nutzung von SLAAC. Diese implizieren auch die Nutzung von mehreren, zum Teil (bei Privacy Extensions) wechselnden IPv6-Adressen gleichzeitig. Ob die für IPv4 vorhandenen Sicherheitsmechanismen daher ohne weiteres auf IPv6 übertragen werden können ist sehr fraglich.

L3-2 – Rogue DHCP

Bei aktiviertem DHCP-Snooping werden DHCP-Antworten von Servern hinter nicht explizit oder implizit konfigurierten *trusted*-Ports unterdrückt. Eine andere Möglichkeit stellt die Nutzung von IPv4-Accesslisten dar, die Pakete vom Port 67/UDP (DHCP-Server) filtern.

In beiden Fällen werden fremde DHCP-Server wirkungsvoll unterbunden.

Eine Unterstützung von IPv6 ist nicht vorhanden.

L3-3 – RA Guard

Eine Unterstützung für IPv6 ist nicht vorhanden.

L4 – Paketfilter und Firewall

Der Nexus 1000V unterstützt prinzipiell durch IP-Accesslisten sowohl eingehend als auch ausgehend eine Filterung anhand der folgenden Kriterien:

- Layer 4-Protokoll
- TCP- und UDP-Ports
- ICMP Typ und Code
- IGMP Typ
- IP Precedence-Feld
- DSCP-Feld
- TCP-Flags (SYN, ACK, FIN, PSH, RST, URG)

Die Filter sind ausschließlich stateless und können daher nur auf Paketebene, aber nicht auf Verbindungsebene filtern.

Eine Konfigurationsmöglichkeit für den Kunden ist nicht vorgesehen. Es gelten die gleichen Einschränkungen wie bei L2-1, die ACL kann entweder einer Portgruppe zugewiesen werden und gilt daher für alle virtuellen Maschinen der gleichen Portgruppe, oder direkt einem spezifischen vEthernet, welches aber beim Abschalten oder Umkonfigurieren des Containers gelöscht werden kann.

Die Möglichkeiten sind daher nur bedingt geeignet.

Fazit

Das Gesamturteil über den Cisco Nexus 1000V ist sehr zwiespältig. Auf der positiven Seite ist zu vermerken, dass selbst die kostenlose Essential-Edition viele Fähigkeiten mitbringt, die im Standard VMware-Switch aus Sicht des Netzbetriebs vermisst werden. So ermöglicht er durch DHCP-Snooping, ARP-Inspection und IP-Source-Guard, den Standardfall eines mit DHCP angebundenen Gastsystems einfach und sicher zu konfigurieren, während die Unterstützung für IPv4- und MAC-Accesslisten es erlaubt, den Verkehr noch weiter zu reglementieren. Die Bedienung erfolgt ähnlich zu den bekannten Cisco-Geräten. Über die Port-Profile können

die Standardarbeiten bei der Administration von virtuellen Maschinen, nämlich die Zuweisung zu einem bestehenden VLAN/Portgruppe, durch die VMware-Administratoren in der grafischen Oberfläche erledigt werden, während die Feinheiten der Konfiguration vor ihnen versteckt werden kann.

Dem stehen jedoch auch starke Nachteile gegenüber. Die Nicht-Unterstützung von IPv6 in einem Produkt, welches erst vor wenigen Jahren neu auf den Markt gekommen ist und im Jahr 2012 mehrere große Aktualisierungen erfahren hat, ist nur als lächerlich zu bezeichnen. Durch die sehr starke Kopplung zwischen DHCP-Snooping und anderen Sicherheitsmechanismen bleibt es abzuwarten, ob die äquivalenten Fähigkeiten auch für IPv6 vollständig zur Verfügung stehen werden. Sollte es dort ebenfalls eine Abhängigkeit zu stateful DHCPv6 geben wäre dieser Mechanismus nicht sinnvoll einsetzbar.

Ein zweiter sehr erschreckender Punkt ist die fehlende Kopplung der im vCenter Server hinterlegten MAC-Adresse mit den Einstellungen für Portsecurity im Nexus 1000V. Obwohl dieser die Einstellungen offensichtlich auslesen kann scheint es nicht möglich zu sein, diese Informationen adäquat zu nutzen um ein MAC-Spoofing durch den Gast unmöglich zu machen. Stattdessen werden Konzepte aus der Welt der physischen Server, bei denen es zwar keine genormte Quelle für diese Informationen aber dafür eine feste Verbindung zwischen Switchport und angeschlossenem Gerät gab, missbraucht, um wenigstens ein wahlfreies Wechseln der MAC-Adresse im Gast zu verhindern.

Das Fehlverhalten bei doppelten 802.1q-Tags ist zwar vermutlich auf den meisten Gastsystemen mangels VLAN-Konfiguration nicht ausnutzbar, hinterlässt aber einen schlechten Nachgeschmack was die generelle Qualität der Softwareimplementierung angeht.

4.3 Strukturevaluation

Neben der im vorherigen Kapitel diskutierten Option, den im Hypervisor integrierten Switch durch ein Produkt mit stärkerer Fokussierung auf die Sicherheit zu ersetzen, um damit sowohl die bereits bestehende physische Infrastruktur unverändert weiterzubetreiben als auch wie geplant die virtuellen Server verschiedener Kunden innerhalb von Sammel-VLANs zu platzieren, gibt es auch andere Möglichkeiten.

4.3.1 dedizierte VLANs pro Kunde

Sieht man von direkten Angriffen, die auf die Ausnutzung von Sicherheitslücken in der Netzinfrastruktur zielen (insbesondere L2-2 und L2-3) ab, können die meisten Angriffe auf andere Systeme nur innerhalb der gleichen Broadcast-Domain durchgeführt werden. Sie können daher nicht über Routergrenzen hinweg übertragen werden, sondern nur benachbarte Systeme im gleichen VLAN angreifen. Wird die Anzahl von Rechnern in einem Subnetz reduziert, so reduziert sich gleichzeitig auch die Anzahl der durch die Kompromittierung eines Systems gefährdeten Maschinen.

Im besten Fall befinden sich in einem VLAN nur noch zwei Systeme: ein virtueller Server und ein Routerinterface für die Verbindung nach außen. Diese Konfiguration erfüllt die funktionalen Anforderungen wie folgt

L2-1 – Verhinderung von MAC-Spoofing
: Voll erfüllt, da sich kein System im Segment befindet dessen Verkehr ein Angreifer übernehmen könnte

L2-2 – Versand von Kontrollpaketen
: Nicht erfüllt, Angriffe beziehen sich auf Netzkomponenten, die immer noch alle durch die VM gesendeten Pakete empfangen können

L2-3 – Flooding („Storm-Control“)
: Nicht erfüllt, keine Einschränkung der Senderrate (aber unter Umständen limitierte Reichweite

L2-4 – Versand unbekannter Ethertypes
: Grundlegend erfüllt, Router muss unbekannte Ethertypes ignorieren

L25-1 – Verhinderung von ARP-Spoofing
: Voll erfüllt, kein System im Segment dessen Verkehr oder Adresse ein Angreifer übernehmen könnte

L3-1 – IP-Spoofing
Voll erfüllt, alle on-link Adressen gehören zum virtuellen Server, off-link Adressen werden durch uRPF am Router geblockt

L3-2 – Rogue DHCP
Voll erfüllt, kein anderer DHCP-Client im gleichen Segment

L3-3 – RA Guard
Voll erfüllt, kein anderer IPv6-Host im gleichen Segment

L4 – Paketfilter und Firewall
Nicht erfüllt, keine Einschränkung

Bezüglich den Angriffen L2-1, L25-1 und L3-1 ist zu sagen, dass der Angreifer hier sehr wohl die MAC- oder IP-Adresse des Routers spoofen könnte. Er würde damit jedoch, abgesehen von einem eventuellen Kommunikationsverlust des Angreifers selbst, keinen Effekt erzielen. Diese Konfiguration bietet damit eine maximale Sicherheit gegen alle bekannten und unbekannten Angriffe, die sich auf den weitgehend unauthentisierten Layern auf und unter Schicht 3 abspielen.

Leider haben dedizierte VLANs pro System einige Skalierungsprobleme, die sie für den Betrieb im großen Maßstab nur beschränkt brauchbar machen.

Als Erstes stehen für den Bereich der 802.1Q VLAN-IDs nur 12 Bit, also 4096 eindeutige Identifikationsnummern zur Verfügung (siehe Kapitel 3.1.2). In großen Campusnetzen wie dem Münchner Wissenschaftsnetz sind bereits heute mehr als 2000 VLANs belegt. Die VLAN-IDs können unter Umständen mehrfach verwendet werden, sofern keine Verbindung auf Schicht 2 zwischen den einzelnen Instanzen vorliegt. Dies erhöht aber den Verwaltungsaufwand enorm, da nicht nur eine freie VLAN-ID gesucht werden, sondern auch auf eine Layer 2-freie Verbindung zwischen den Einsatzorten dieses VLANs geachtet werden muss. Eine vergessene Konfiguration kann dabei dazu führen, dass zwei getrennte Netze zusammmengeschaltet werden. Da die eingesetzten Router des Münchner Wissenschaftsnetzes (Cisco Catalyst 6500) intern ebenfalls als Switch mit einer angehängten Layer 3-Forwarding Engine arbeiten, sind diese bei der VLAN-Zuteilung zu beachten. Eine mögliche Lösung aus Providernetzen, bei denen mehrere VLANs ineinander gekapselt werden (IEEE 802.1ad, „QinQ“), ist mit der aktuell vorhandenen Infrastruktur weder auf Routerseite noch im Bereich der virtuellen Infrastruktur möglich.

Ein anderer Nachteil ist im Bereich der IP-Adressierung zu finden, insbesondere im IPv4-Protokoll. Abgesehen von Spezialkonstrukten, die nur in eng spezifizierten Umgebungen möglich sind, muss jedem VLAN mindestens ein vollständiges, eindeutiges und überlappungsfreies Subnetz zugewiesen werden. Diese stehen in festen, aus Zweierpotenzen abgeleiteten Größen zur Verfügung, wobei die niedrigste Adresse für die Netzadresse und die höchste Adresse für die Broadcastadresse reserviert sind. Von den verbleibenden Adressen benötigt der Router des Providers mindestens eine, beim Einsatz eines *First Hop Redundancy Protocol* wie VRRP oder HSRP jedoch bis zu drei Adressen (Tabelle 4.4). Eine Übersicht der üblichen Subnetzgrößen und der darin nutzbaren Adressen ist in Tabelle 4.5 zu finden.

IP-Adresse	Verwendung
10.0.0.0	Netzadresse
10.0.0.1	VM 1
10.0.0.2	VM 2
10.0.0.3	VM 3
10.0.0.4	Datacenter Router 1
10.0.0.5	Datacenter Router 2
10.0.0.6	Default Gateway (FHRP)

Tabelle 4.4: Nutzbare Adressen am Beispiel 10.0.0.0/29

Wie darin zu sehen sind je nach Konfiguration 3-5 Adressen pro Subnetz nicht für den Dienst – den Betrieb von virtuellen Servern für Kunden – nutzbar. Da die Subnetze einem VLAN/Kunden fest zugeordnet sind und IPv4-Adressen nur noch in einem sehr beschränkten Maß neu zur Verfügung stehen [RIPE-IPv4], müssen die Netze möglichst klein und auf die Anforderung des Kunden angepasst vergeben werden. Der prozentuale

Präfixlänge	Anzahl Adressen	Nutzbare Adressen insgesamt	Nutzbare Adressen bei 1 IP für Router	Nutzbare Adressen bei 3 IPs für Router
/30	4	2	1	-
/29	8	6	5	3
/28	16	14	13	11
/27	32	30	29	27
/26	64	62	61	59
/25	128	126	125	123
/24	256	254	253	251

Tabelle 4.5: Subnetzgrößen und nutzbare Adressen

Verlust wird allerdings umso größer, je kleiner das Subnetz ist. Außerdem muss, wenn der Bedarf über den zugewiesenen Adressraum hinaus wächst, das Netz vergrößert werden. Dies ist im Allgemeinen nicht ohne den Wechsel der Adressen (*Renumbering*) auf allen bestehenden Systemen des Kunden möglich.

Derartige Probleme existieren bei IPv6 nicht mehr, da die empfohlene Größe für Servernetze Platz für 2^{64} Adressen beinhaltet ([RFC5375] 3.) und diese in einer ausreichenden Anzahl zur Verfügung stehen.

Nicht zuletzt ist im konkreten Beispiel des LRZ der Overhead negativ zu bemerken, der durch die manuelle Einrichtung von VLANs auf den beteiligten Komponenten entsteht. Insbesondere auf den reinen Layer2-Komponenten könnte hier jedoch einfach eine Automatisierung stattfinden.

Aufgrund der beschriebenen Skalierungsprobleme ist es daher Best-Practice bei vielen Providern und auch im Bereich der Campusanbindungen im MWN, alle Systeme eines Kunden in ein dediziertes VLAN zu legen. Dadurch können zwar weiterhin kompromittierte Rechner Angriffe auf andere Systeme durchführen, aber nur auf diejenigen des gleichen Kunden. Dadurch ist es immerhin nicht mehr möglich, schlecht administrierte Systeme des einen Kunden als Zwischenstation zum Angriff auf besser geschützte Systeme eines anderen Kunden zu verwenden.

VXLAN

Der von Mitarbeitern der Hersteller Arista, Broadcom, Cisco, VMware, Citrix und Red Hat maßgeblich entwickelte IETF-Entwurf *VXLAN* ist ein Tunnelungsmechanismus, um Layer2-Segmente (VLANs) über Layer3 hinweg zu transportieren. VXLAN wird hauptsächlich aus drei Gründen entwickelt [VXLAN]:

Limitierungen von geswitchtem Ethernet

Ethernet ist aufgrund dem Fehlen einer TTL und dem Fluten von Paketen an Broad-, Multi- oder unbekannte Unicast-Adressen (Kapitel 3.1) nicht von selbst in der Lage, mit redundanten Pfaden umzugehen. Zur Vermeidung von Schleifen und den dadurch hervorgerufenen Paketstürmen kommt daher üblicherweise das Spanning-Tree-Protokoll zum Einsatz. Dieses erstellt aus einer beinahe beliebig geformten physischen Topologie durch Abschaltung beziehungsweise Blockierung von Verbindungen eine schleifenfreie Topologie. Die Bandbreite der deaktivierten Verbindungen ist damit nicht nutzbar, außerdem führt es zu Umwegen im Netzverkehr.

Aufgrund dieser Einschränkungen kommen in modernen Netzen im WAN-Bereich nahezu ausschließlich geroutete Verbindungen mit den Protokollen IPv4, IPv6 und MPLS zum Einsatz. In Rechenzentren werden hingegen noch häufig geswitchte Netze verwendet, wobei auch hier ein Umdenken festzustellen ist.

Der Einsatz eines gerouteten Netzes widerspricht jedoch einer anderen Anforderung, die sich aus großen Virtualisierungsumgebungen ergibt. Container sollen ohne Neustart und sogar ohne einen merkbaren Ausfall zwischen Hosts verschoben werden können. Um hierbei ihre Netzverbindung nicht zu verlieren muss am neuen Standort das gleiche VLAN verfügbar sein. Dies ist in vergleichsweise kleinen Umgebungen wie dem aktuellen Ausbaustand des LRZ mit zwei zentralen Switchen noch möglich, würde aber spätestens an der Anforderung des Migrierens in ein anderes Rechenzentrum zu großen Problemen

geführen, da dafür die Netze auf Ethernet-Ebene gekoppelt werden müssten und durch Loops entstehende Instabilitäten beide Rechenzentren in Gefahr bringen könnten.

Begrenzte Anzahl von 802.1q-VLAN-Tags

Wie bereits in Kapitel 3.1.2 beschrieben stehen für die logische Segmentierung 4096 mögliche VLAN-Tags zur Verfügung. Diese Anzahl ist im Betrieb schnell erreicht, insbesondere da viele Geräte nur eine deutlich begrenztere Anzahl von gleichzeitig konfigurierten VLANs unterstützen (beispielsweise können auf den derzeit als Zentral- und VMware-Switches eingesetzten HP Procurve-Switches maximal 2048 VLANs verwendet werden).

Ein anderer Nachteil beim Einsatz von VLANs ist, dass die benötigten VLANs auf allen Switches, die dauerhaft oder auch erst im Fehlerfall (durch den Einsatz von Spanning-Tree) aktiv auf dem Pfad zwischen allen möglichen Sendern und Empfängern liegen, konfiguriert werden muss. Dies kann in großen Umgebungen schnell aufwändig und fehleranfällig werden.

Begrenztes Fassungsvermögen von MAC-Tabellen

Wie ebenfalls bereits in Kapitel 3.1) beschrieben müssen potentiell alle an einem VLAN teilnehmenden Switches alle in diesem VLAN vorkommenden MAC-Adressen im Speicher halten. Die maximale Anzahl beträgt bei heutigen Ethernet-Switches je nach Hersteller und Modell zwischen 4000 und 64000 Adressen.

Die in der Virtualisierungsinfrastruktur des LRZ eingesetzten Systeme haben laut den jeweiligen Datenblättern das folgende Fassungsvermögen:

- *Cisco 6500 VSS*: 110000 Einträge
- *HP Procurve 8400zl und 5400zl*: 64000 Einträge
- *HP Flex-10*: 8192 Einträge
- *VMware ESXi 5.1 vSwitch*: unbekannt

Im Rechenzentrum des LRZ sind bereits über 10000 MAC-Adressen im Einsatz. Dieser Umstand zeigt, dass die Flex-10 Module bei der Nutzung aller VLANs in der Virtualisierungsinfrastruktur bereits heute nicht mehr ausreichend Platz bieten würden.

VXLAN löst diese Probleme durch eine Verpackung der Ethernet-Frames in einem auf IP aufbauenden UDP-Tunnel. Diese können ohne weitere Probleme auch über größere Strecken und alle routingfähigen Netze übertragen werden, auch wenn diese kein Ethernet unterstützen (beispielsweise SONET/SDH).

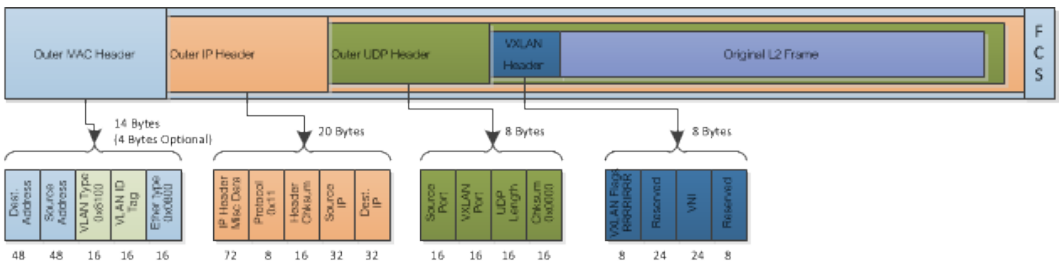


Abbildung 4.7: VXLAN-Paketstruktur, Quelle Kamau Wanguhu [Wang 11]

Das Ein- und Auspacken geschieht auf der Rolle des VTEP (*VXLAN Tunnel End Point*), welcher Ethernet-Frames in ein spezifisches VXLAN-Segment (vergleichbar mit einem VLAN) schicken beziehungsweise daraus empfangen kann. Das VXLAN-Segment ist durch den 24-Bit breiten VNI (*VXLAN Network Identifier*) festgelegt, wodurch mehr als 16 Millionen Segmente möglich sind.

Für den Betrieb des VTEP wird auf diesem die MAC-Adresstabelle erweitert, so dass neben der MAC-Adresse und dem (lokalen) Switchport auch die IP-Adresse des VTEP hinterlegt werden kann, hinter dem die MAC-Adresse erreichbar ist. Ein Frame an eine bekannte entfernte MAC-Adresse wird in einem UDP-Paket an den Ziel-VTEP gesendet, welches in jedem normalen Netz geroutet wird (Abbildung 4.8).

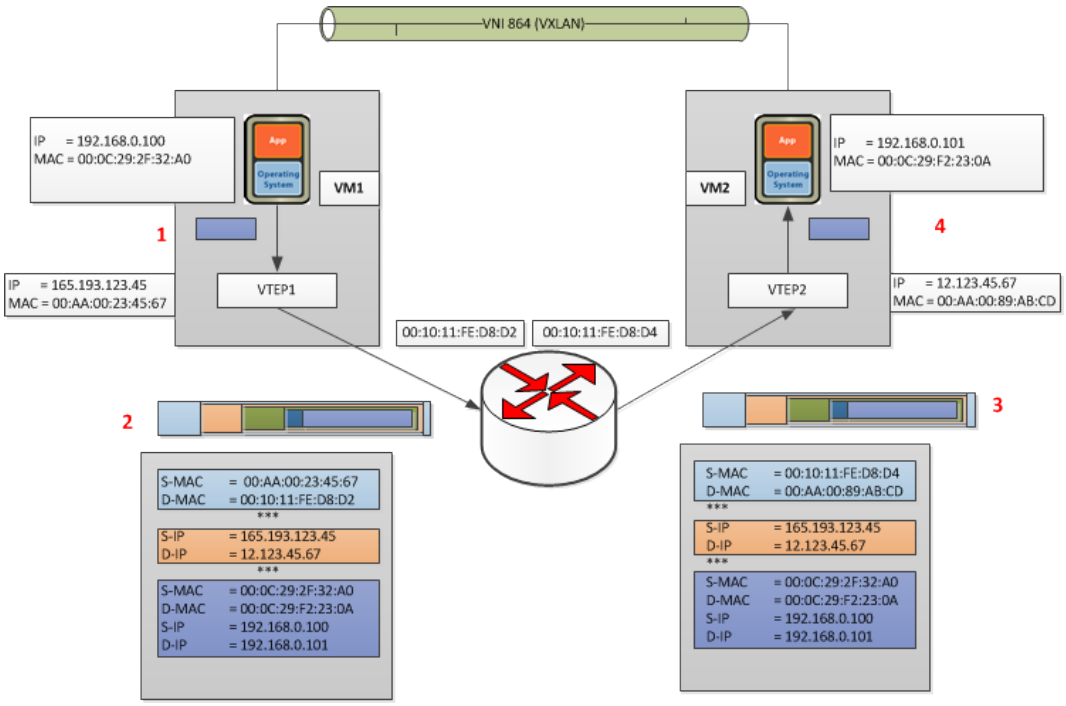


Abbildung 4.8: VXLAN-Kommunikation, Quelle Kamau Wanguhu [Wang 11]

Sofern die Ziel-MAC-Adresse noch nicht in der auf dem VTEP geführten Tabelle bekannt ist („Unknown Unicast“) oder aber eine Multi- oder Broadcastadresse, wird das Frame an eine konfigurierbare IP-Multicastgruppe versendet. Diese Gruppe kann für alle VNIs unterschiedlich oder auch gleich sein. Beim Transport über eine geroutete Layer3-Verbindung muss diese Multicast-Routing unterstützen, was bei allen kommerziell genutzten Plattformen der Fall ist.

Die MAC-Adresstabelle auf dem VTEP wird sowohl beim Empfang von Unicast- als auch Multicast-Paketen befüllt.

Viele Hersteller bieten bereits heute VXLAN-Gateways zur Umsetzung zwischen 802.1q-VLANs und VXLAN-Segmenten an. Der im Kapitel 4.2.2 vorgestellte Cisco Nexus 1000V unterstützt diese Funktionalität und wurde erfolgreich getestet. Der VMware Distributed vSwitch unterstützt VXLAN erst ab der Version VMware ESXi 5.1, was im LRZ noch nicht im Einsatz ist und daher im Rahmen dieser Arbeit nicht evaluiert wurde. Linux unterstützt ab Kernelversion 3.7 ebenfalls VXLAN-Tunnel.

Wie bereits zu Beginn genannt existieren neben VXLAN noch andere Protokolle, die Lösungen für diese Problemstellungen bieten wollen. Technisch bis auf die Details sehr ähnlich ist der konkurrierende Ansatz NVGRE, der von den Firmen Dell, Intel und Microsoft propagiert wird. Im Gegensatz zu VXLAN gibt es jedoch bisher kaum kommerzielle Implementierungen, so dass die Lebenszeit dieses Protokolls begrenzt sein dürfte.

Ein anderes artverwandtes Protokoll ist OTV (*Overlay Transport Virtualization* [OTV]) von Cisco. Es wird insbesondere zur Verbindung zwischen Rechenzentren eingesetzt und ist beispielsweise in der Nexus 7000-Serie verfügbar. Unbekannte Ziel-MAC-Adressen werden nicht transportiert, dafür existieren explizite Kommunikationswege, um gelernte MAC-Adressen in allen Standorten bekannt zu machen. Cisco bewirbt das noch von keinem anderen Hersteller implementierte OTV als Alternative zum standardisierten, aber vergleichsweise komplexen VPLS [Cisco-OTV-VPLS].

Ein weiterer, bereits von der IETF verabschiedeter Standard ist TRILL (*Transparent Interconnection of Lots of Links*, [RFC6325] und [RFC6326]). Er bietet vor allem eine Lösung für die zu Beginn angesprochene Probleme mit Spanning-Tree und Netzschleifen, indem zwischen TRILL-fähigen Switches mit dem Link-State-Routingprotokoll IS-IS der kürzeste Weg zwischen dem ersten (Ingress) und letzten (egress) TRILL-Switch berechnet wird (Dijkstra-Algorithmus). Ferner werden die Auswirkungen von Schleifen durch einen Hop

Count und Filtermechanismen begrenzt. Eine ähnliche Methodik verfolgt der konkurrierende IEEE 802.1aq Standard[vdP 12]. Die Verfügbarkeit dieser Protokolle in Produkten ist nur langsam gegeben, viele Hersteller bieten jedoch bereits seit längerem artverwandte aber proprietäre Lösungen an (Cisco FabricPath, HP Meshing).

Das letzte in diesem Zusammenhang relevante Protokoll ist MPLS (*Multi Protocol Label Switching* [RFC3031]). Es ist ein seperates Layer3-Protokoll für die Data-Plane, benötigt aber für Management und Control-Plane IPv4. Es basiert auf dem aus ATM und Frame-Relay bekannten Konzept, Pakete mit einer auf einer Verbindung für ein Ziel eindeutigen Markierung zu versehen (Label). Auf jedem Zwischenstop wird das Label dann durch die entsprechende Markierung auf dem nächsten Link ersetzt (Label Swap). Da die Tabelle vergleichsweise klein gehalten werden kann wurde diese Methodik früher häufig dazu eingesetzt, schnelle Router mit wenig Arbeitsspeicher ausschließlich für diesen Verkehr zu verwenden und damit Geld zu sparen. Dieser Grund wird heutzutage jedoch nur noch selten zur Entscheidung herangezogen. Auf MPLS haben sich jedoch einige andere Technologien entwickelt, die in modernen Netzen nicht mehr wegzudenken sind. Dazu gehören separierte Layer3-VPNs auf einer gemeinsamen Infrastruktur, transparente Ethernet-Verbindungen über ein MPLS-Netz (EoMPLS, Ethernet over MPLS) und das darauf aufbauende VPLS (*Virtual Private Line Service*). Dadurch können ebenfalls Layer2-Netze über ein nahezu beliebig geformtes Layer3-Netz aufgebaut werden. Der Nachteil von MPLS ist, dass alle Layer3-Geräte zwischen zwei MPLS-Endpunkten ebenfalls MPLS unterstützen müssen, um die Swap-Operation durchführen zu können.

Der Nachteil aller genannten Alternativen abgesehen von NVGRE ist die fehlende Unterstützung direkt in einem Hypervisor. Daher muss bei allen Technologien noch vor der Virtualisierungsinfrastruktur auf herkömmliche VLANs umgesetzt werden, was die Skalierungsvorteile zunichte macht.

4.3.2 private VLAN

Private VLANs sind eine ursprünglich von Cisco entwickelte und mittlerweile durch die IETF standardisierte [RFC5517] Methode, Sicherheitsprobleme auf den Schichten 2 und 2.5 durch Segmentierung zu verhindern. Im Gegensatz zu einer Trennung durch VLANs geschieht dies jedoch nicht vollständig, sondern durch die Verhinderung von Kommunikation zwischen bestimmten Ports im gleichen VLAN. Der Begriff „VLAN“ ist hier irreführend; die Beschreibung im Standard und die Konfiguration der Marktführer basieren zwar auf speziell markierten VLANs, der gleiche Effekt lässt sich jedoch auch anders darstellen.

Die Funktionalität von Private VLANs gemäß Standard erfolgt durch die Einführung dreier neuer Rollen, die einem Switchport zugewiesen werden können.

- Isolated Ports
- Community Ports (mit Community n)
- Promiscuous Ports

Verkehr zwischen den einzelnen Rollen wird gemäß Tabelle 4.6 eingeschränkt.

	Isolated	Promiscuous	Community 1	Community 2
Isolated	✗	✓	✗	✗
Promiscuous	✓	✓	✓	✓
Community 1	✗	✓	✓	✗
Community 2	✗	✓	✗	✓

Tabelle 4.6: Verkehrsbeziehungen von PVLAN

Im Rahmen dieser Arbeit werden nur Isolated und Promiscuous Ports betrachtet. Community Ports bieten nur eine Möglichkeit, die Einschränkungen selektiv für bestimmte Paarungen wieder aufzuweichen, bieten jedoch keine zusätzliche Sicherheit.

In einem einfachen Beispiel ist die Funktionsweise von Private VLAN schnell erklärt. An einem Switch ist am Promiscuous Port 1 ein Router (das Default-Gateway) angeschlossen ist und an den Isolated Ports 2 und 3 jeweils ein Host (Abbildung 4.9).

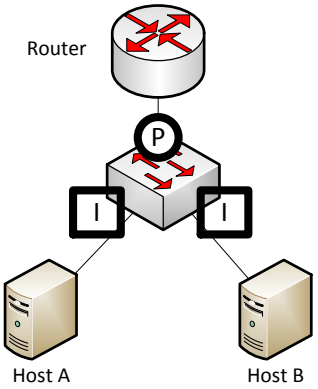


Abbildung 4.9: Einfaches Beispiel Private VLAN

Gemäß der Tabelle 4.6 können damit beide Hosts in beiden Richtungen mit dem Router kommunizieren, jedoch nicht untereinander. Einige Angriffe, beispielsweise Rogue DHCP und Rogue RA, werden durch diese Technologie vollständig verhindert, da für diese ein Angreifer (auf Host A) ein Paket direkt an den Zielhost (Host B) schicken müsste. Andere werden erschwert oder deutlich eingeschränkt. So kann Host A weiterhin ein ARP-Paket mit der IP-Adresse des Hosts B an den Router schicken und damit den Eintrag in der Neighbor-Tabelle überschreiben (ARP-Spoofing), er kann jedoch nicht ein ARP-Paket mit der IP-Adresse des Routers an Host B senden und dessen Eintrag für den Default-Router überschreiben.

Generell gilt mit Private VLAN, dass Angriffe damit nicht mehr auf jedem Host abgewehrt werden müssen, sondern eine entsprechende Konfiguration und Filterung auf den Geräten an Promiscuous Ports ausreicht. Diese sind im Allgemeinen nur in deutlich geringerer Anzahl vorhanden (1-2 Router pro Netzsegment) und unterstehen der Administration des Netzbetreibers.

Um trotz Private VLANs eine (gefilterte) Kommunikation zwischen den Hosts im gleichen Subnetz zu ermöglichen wird Private VLAN oft mit Proxy-ARP kombiniert. Hierbei antwortet der Router stellvertretend auf alle ARP-Anfragen, selbst wenn die Zieladresse im gleichen Netzsegment lokalisiert ist. Der ankommende Nutzdaten-Verkehr wird auf der gleichen Schnittstelle empfangen und gesendet. Im Endeffekt verarbeitet damit der Router alle Pakete zwischen Host A und Host B, obwohl diese im gleichen Subnetz liegen. Ein Proxy-ARP (beziehungsweise Proxy-NDP) für **alle** IPv6-Adressen ist in IPv6 nicht ohne weiteres möglich, da dafür 2^{24} Multicast-Gruppen abonniert und verarbeitet werden müssten. Bei IPv6 können aufgrund der Trennung der Subnetzmaske von grundlegenden Netzfunktionalitäten wie dem Standardgateway jedoch andere Lösungen gefunden werden, beispielsweise /128-Präfixlängen und Link-Local Gateways. Eine genauere Spezifikation der benötigten Einstellungen ist in Kapitel 6 beschrieben.

Auf dem Papier erfüllt das Konzept der privaten VLANs daher, für sich allein gesehen, die Anforderungen wie folgt:

L2-1 – Verhinderung von MAC-Spoofing
: Nicht erfüllt

L2-2 – Versand von Kontrollpaketen
: Nicht erfüllt, Angriffe beziehen sich auf Netzkomponenten, die immer noch alle durch die VM gesendeten Pakete empfangen können

L2-3 – Flooding („Storm-Control“)
: Nicht erfüllt, keine Einschränkung der Senderrate (aber unter Umständen limitierte Reichweite)

L2-4 – Versand unbekannter Ethertypes
: Grundlegend erfüllt, Router muss unbekannte Ethertypes ignorieren

L25-1 – Verhinderung von ARP-Spoofing

Teilweise erfüllt, kein System kann den Eintrag des Routers gegenüber einem anderen System fälschen, aber den Eintrag eines anderen Systems gegenüber dem Router

L3-1 – IP-Spoofing
Teilweise erfüllt, IP-Adresse des Routers nicht fälschbar, „off-link“ Spoofing zum Angriff gegen Systeme im gleichen VLAN nicht möglich

L3-2 – Rogue DHCP
Voll erfüllt, kein anderer DHCP-Client im gleichen Segment

L3-3 – RA Guard
Voll erfüllt, kein anderer IPv6-Host im gleichen Segment

L4 – Paketfilter und Firewall
Nicht erfüllt, keine Einschränkungen

Private VLANs haben jedoch auch einige Anforderungen und Nachteile, die im Kontext großer Netze wie dem Leibniz-Rechenzentrum zu Tage treten. Im Allgemeinen existieren zwischen dem Router und dem Teilnehmer eine ganze Reihe von Geräten mit Switchfunktion. Im vorliegenden Szenario (siehe Kapitel 2.1) liegen zwischen dem Router und einer beliebigen virtuellen Maschine mindestens vier Switches:

- HP Procurve 8412zl („Zentralswitch“)
- HP Procurve 5406zl („VMware-Switch“)
- HP Flex10-Modul
- VMware Distributed vSwitch

Da diese Technologie keine Angriffspakete erkennt und unterdrückt, sondern das Verhalten von Switches modifiziert, müssen alle beteiligten Geräte dieses Feature unterstützen. Andernfalls ist der Schutz nur partiell je nach Platzierung des Containers vorhanden. Ein Beispiel für dieses Problem zeigt Abbildung 4.10, in der ein Angreifer auf VM A die VM C erreichen kann, VM B jedoch nicht.

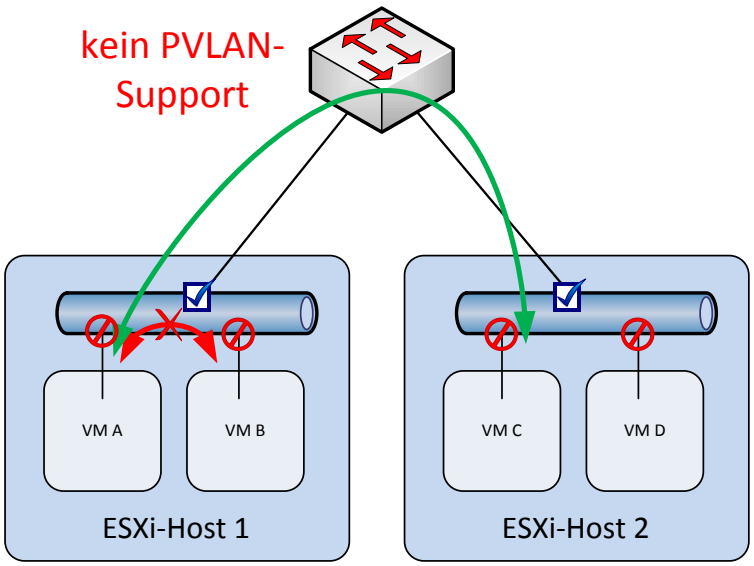


Abbildung 4.10: Unvollständiger PVLAN-Support

Eine Analyse der Datenblätter der beteiligten Systeme zeigt, dass alle aktuell verwendeten Produkte eine derartige Konfiguration zumindest partiell unterstützen. Allerdings treten hier mehrere Einschränkungen auf, die einen Einsatz ausschließen:

- HP Procurve Switches bieten eine Konfiguration namens *Source-Port Filters*, bei der die Kommunikation zwischen einzelnen Switchports komplett verhindert werden kann ([HP 12b] Seite 590ff). Diese Filter gelten jedoch global für alle VLANs auf einem Port und können daher in der LRZ-Umgebung nicht eingesetzt werden.
- HP Flex10-Module erlauben die Konfiguration eines *Private Networks*, bei dem die Kommunikation zwischen Server-Blades unterbunden werden kann ([HP 10] Seite 88ff). Diese Einstellung ist jedoch, ähnlich zu den Procurve Switches, nur auf der Basis physischer Ports möglich und gilt daher für alle VLANs, auch für Management, vMotion und Storage.

Zuletzt ist auch noch zu bedenken, dass die Nutzung von Private VLANs mit der Nutzung des Redundanzmechanismus Spanning-Tree (siehe Kapitel 2.1.2) kollidiert. Dies liegt darin begründet, dass die Rollen *Promiscuous Port* beziehungsweise *Isolated Port* einer physischen Schnittstelle zugeordnet werden müssen. Spanning-Tree verändert jedoch die Netztopologie im Falle der Störung einer Verbindung oder eines Geräts. Dies kann dazu führen, dass ein Isolated-Port auf einmal zu einem Promiscuous-Port werden muss (und umgekehrt). Die nachfolgenden Abbildungen verdeutlichen dies am Beispiel der LRZ-Infrastruktur. Im Normalbetrieb (Abbildung 4.11) ist die Verbindung zwischen den beiden „VMware-Switches“ durch Spanning-Tree blockiert, die Konfiguration der Rolle daher irrelevant. Im Fall eines Fehlers in der Verbindung zwischen dem rechten „Zentralswitch“ und dem rechten „VMware-Switch“ wird dieser Link aktiv und benötigt daher die in Abbildung 4.12 Konfiguration der Rollen. Bei Wiedererreichen des Normalbetriebs wird dieser Link wieder inaktiv. Sollte nun jedoch der Link zwischen den linken Switches ausfallen (siehe Abbildung 4.13), so müssen die Rollen vertauscht werden.

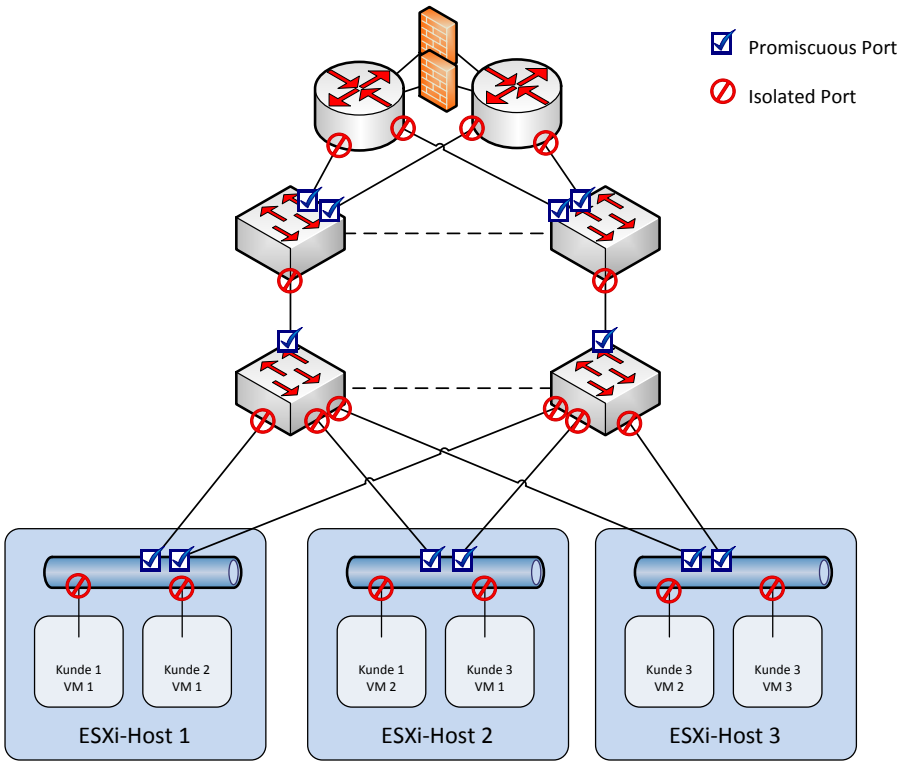


Abbildung 4.11: Private VLAN + STP - Normalbetrieb

Da im Allgemeinen ein Ausfall nicht vorhersehbar ist, würde die Konfiguration der Private VLAN-Funktionalität eine große Gefahr für den Betrieb der Infrastruktur darstellen. Eine Alternative wäre die Konfiguration aller redundanten Verbindungen als *Promiscuous Port*. Dies würde zwar die Redundanz ermöglichen, allerdings im Fehlerfall die Sicherheitsmechanismen außer Kraft setzen.

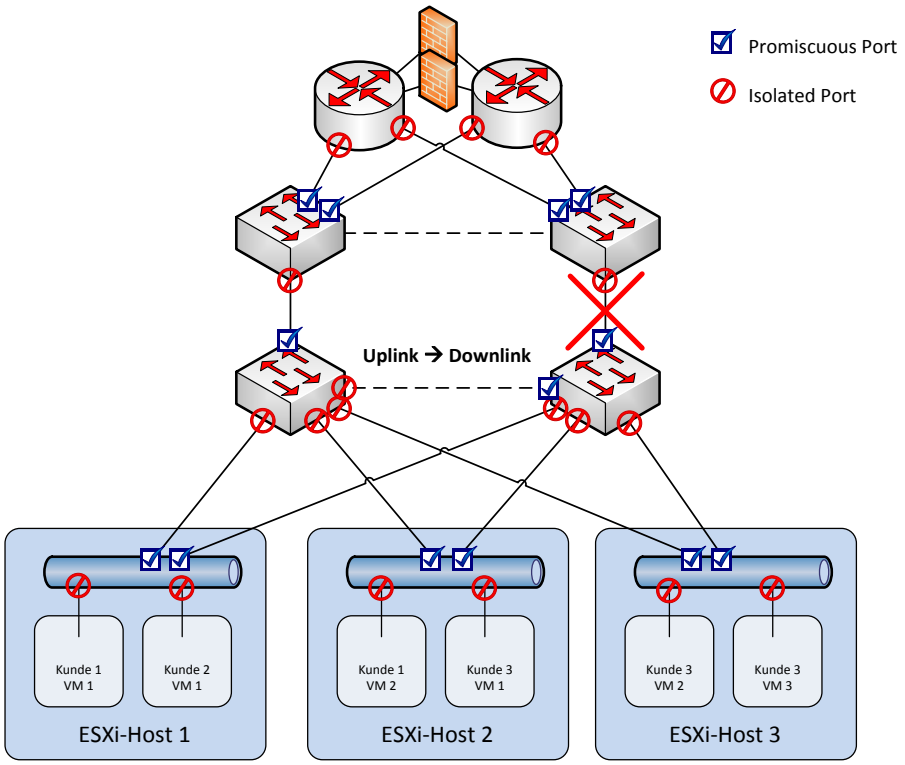


Abbildung 4.12: Private VLAN + STP - Ausfall Rechts

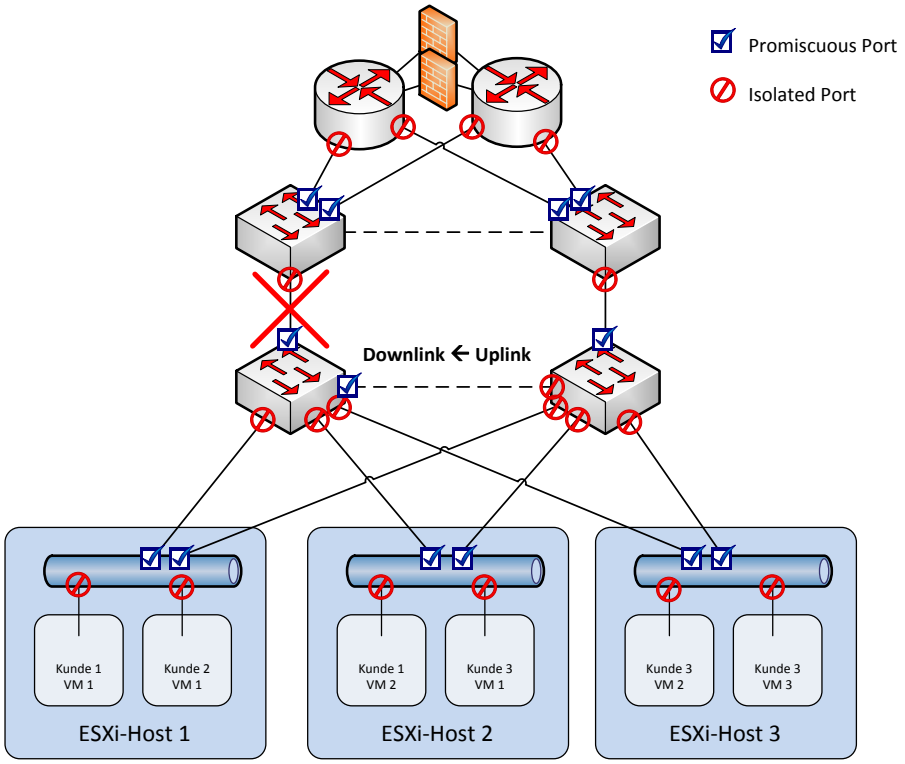


Abbildung 4.13: Private VLAN + STP - Ausfall Links

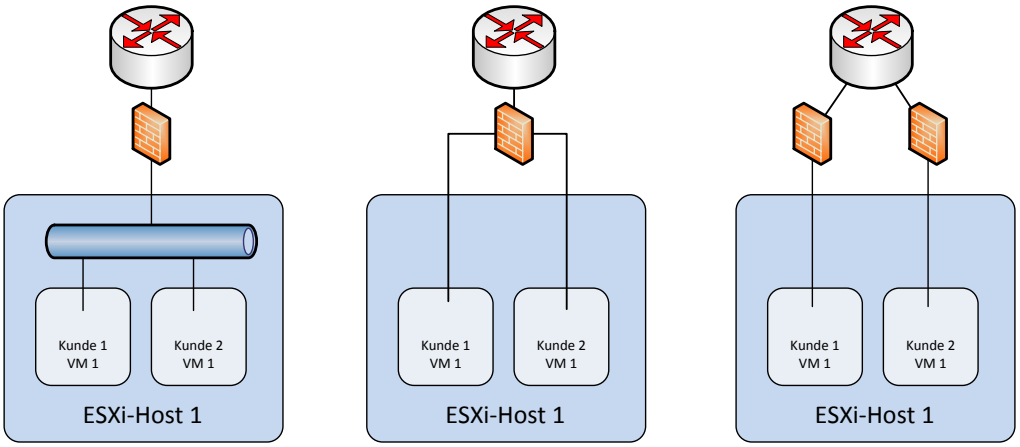
4.3.3 Subnetz-Firewall

Eine Erweiterung des VLAN-Konzepts stellen Subnetz-Firewalls dar, die Verkehr beim Überschreiten der Subnetzgrenze gemäß konfigurierbaren Filterlisten analysieren und bearbeiten können. Je nach Platzierung der Firewall und der zu schützenden Systeme unterscheidet man mehrere Szenarien, die in der Abbildung 4.14 grafisch dargestellt sind. Bei der Diskussion der Vor- und Nachteile muss nun erneut zwischen den prinzipiellen Eigenschaften und den Möglichkeiten der am LRZ vorhandenen Hardware unterschieden werden.

Im einfachsten Fall wird durch den Betreiber der Infrastruktur eine zentrale Firewall bereitgestellt, die als Standardgateway zwischen dem globalen Netz und einem zentralen Sammel-VLAN mit VMs aller Kundengruppen agiert. Der Regelsatz dieser zentralen Firewall wird dabei durch den Infrastruktur-Betreiber definiert und ist für die Kunden nicht oder nur durch manuelle Interaktion mit dem Betreiber änderbar. Die Firewall kann dabei jedoch nur Verkehr filtern, der durch sie hindurch geroutet wird, der Verkehr innerhalb des gleichen VLANs (und damit auch zwischen den virtuellen Maschinen unterschiedlicher Kunden) unterliegt den normalen, bereits in den vorherigen Kapiteln definierten Einschränkungen.

Zur weiteren Trennung kann der Betreiber nun analog zu Kapitel 4.3.1 VLANs zur Trennung der Kunden untereinander einsetzen und diese auf einer zentralen Firewall anbinden. Er erreicht dadurch ein Sicherheitsniveau, das mindestens dem der dedizierten VLANs entspricht, und kann zusätzlich noch seinem Kunden eine Firewall anbieten die den Verkehr seines Subnetzes von und nach außen filtert. Die Regelverwaltung unterliegt allerdings erneut dem Betreiber der zentralen Firewall, dem die Kundenwünsche in einer geeigneten Form übermittelt werden müssen. Diese Lösung teilt sich mit den dedizierten VLANs ebenso die schlechten Skalierungseigenschaften bezüglich dem Ressourcenverbrauch bei VLANs und IP-Adressen.

Eine vollständige Mandantenfähigkeit erreicht erst eine weitere Ausbaustufe, bei der jedem Kunden nicht nur ein dediziertes VLAN zur Verfügung gestellt wird, sondern auch eine dedizierte Firewall mit eigenem Regelsatz. Durch die hierbei erfolgte 1:1 Zuordnung eines Kunden zu einer Firewall ist es möglich, die Verwaltung des Regelsatzes durch eine geeignete Schnittstelle an den Kunden zu delegieren. Die Skalierungseigenschaften verschlechtern sich jedoch nochmalig, da bei mindestens gleichbleibend hohem Verbrauch von VLANs und IP-Adressen der Netzverkehr zwischen zwei Kunden zweimal gefiltert werden muss.



(a) Zentrale Firewall, 1 VLAN

(b) Zentrale Firewall, mehrere VLANs

(c) dezentrale Firewalls

Abbildung 4.14: Firewall-Szenarien

Aufgrund der Anforderung der Mandantenfähigkeit ist in der vorliegenden Infrastruktur nur die Lösung mit dezentralen Firewalls sinnvoll. Das LRZ bietet sowohl seinen internen Gruppen als auch den Kunden im MWN bereits ein Produkt „virtuelle Firewall“ (Kapitel 2.1.2) auf Basis von Cisco ASA und FWSM an, bei denen ein physisches Gerät durch Konfigurationsoptionen in mehrere virtuelle Einheiten, die sogenannten *Security Contexte*, unterteilt wird. Diese Contexte sind jeweils für mehrere konfigurierte VLANs zuständig und beinhalten

eine eigene Regel- und Benutzerverwaltung, die es ermöglicht die Konfiguration des Regelsatzes komplett an einen Benutzer auszulagern. Der Kunde wird damit in die Lage versetzt, selbstständig und ohne manuelle Interaktion mit seinem Dienstanbieter Regeln zu definieren und einen Einblick in den Status seines Netzes zu nehmen. Die Anforderung **L4 – Paketfilter und Firewall** ist daher voll erfüllt.

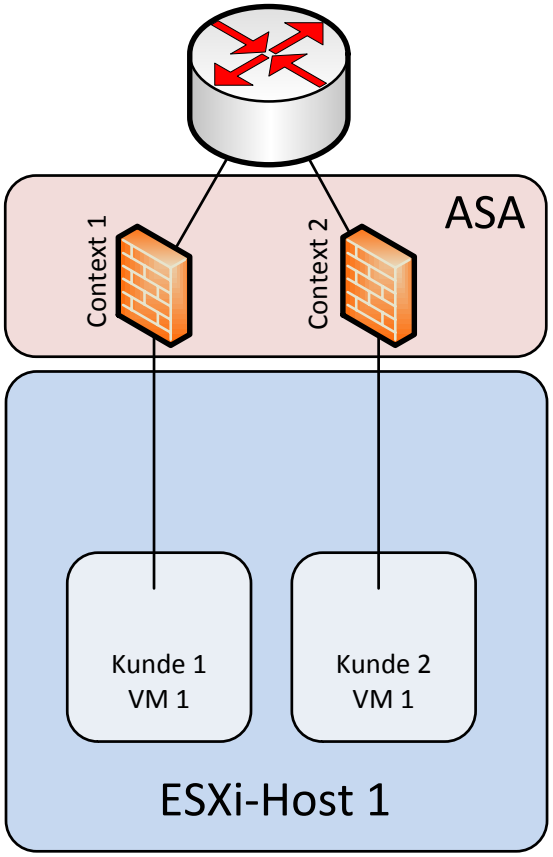


Abbildung 4.15: Virtuelle Firewall am LRZ

Diese Lösung ist jedoch, bedingt durch die reale Implementation, noch ressourcenintensiver als in der Theorie. Da die verwendeten ASA-Firewalls pro Context ein dediziertes, das heißt nicht von einem anderen Kontext verwendetes, VLAN mit einem dedizierten Transportnetz benötigen, erhöht sich der Verbrauch von VLANs und IP-Adressen nochmal deutlich. Die virtuellen Contexte müssen, wie es bei kommerzieller Hardware üblich ist, selbstverständlich lizenziert werden. Die maximale Anzahl an Contexten ist limitiert und beträgt selbst für die aktuellen Topmodelle nur 250 virtuelle Firewalls.

Ein weiterer Nachteil der aktuellen Firewalllösung ist, dass diese kein *Local Proxy-ARP* für den Einsatz von Private VLAN unterstützen, so dass eine Kombination dieser Sicherheitsmechanismen nicht möglich ist.

Da die virtuellen Firewalls nur einen Zusatznutzen in Form von vom Kunden konfigurierbaren Paketfiltern bieten und ansonsten die Nutzung eines dedizierten VLANs implizieren erfüllt diese Lösung die Anforderungen für L2, L25 und L3 in gleicher Weise wie dedizierte VLANs gemäß Kapitel 4.3.1. Die Teilnehmer eines VLANs sind daher von außen vergleichsweise gut gegen Angriffe schützbar, jedoch ihren Nachbar-VMs im gleichen VLAN schutzlos ausgeliefert.

4.3.4 zentral provisionierte virtuelle Firewall (Shared)

Eine mögliche Alternativvariante zur einer dezentralen Firewall ist eine zentrale Firewall, die durch eine abgesetzte Konfigurationsplattform mandantenfähig gemacht wird. Hierbei erhalten die Kunden keinen direkten Zugriff auf die Firewall, sondern können durch ein Selfservice-Portal die ihre Systeme betreffenden Regeln verwalten. Ein Provisionierungssystem sammelt die Regeln aller Kunden einer Firewall und generiert daraus einen allgemeinen Regelsatz, der dann in die Firewall geladen wird.

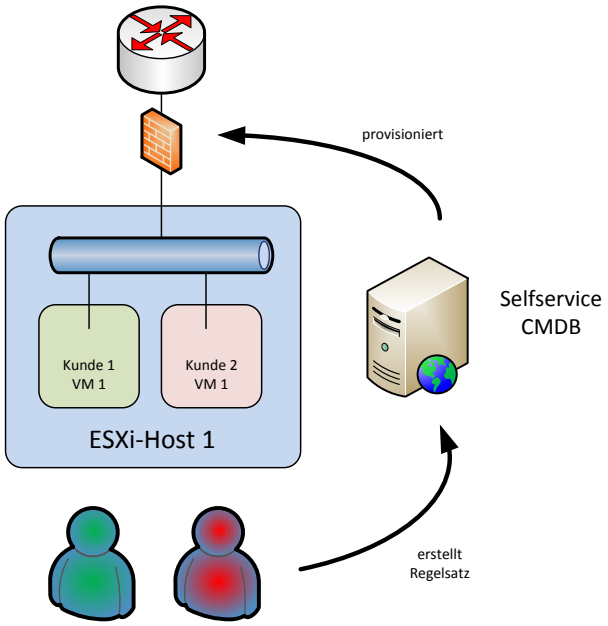


Abbildung 4.16: zentral provisionierte Firewall

Als Vorteile dieser Lösung kann neben der Ressourcenersparnis auch gelten, dass die durch den Kunden hinterlegten Regelsätze unabhängig von der verwendeten Firewallplattform sein können und daher beim Austausch der Hardware angepasst werden können. Nachteilig wirkt sich dann jedoch aus, dass das verwendete Regelwerk nicht alle Fähigkeiten der verwendeten Firewall ausschöpfen kann, sondern nur die wichtigsten.

4.3.5 Firewall auf dem Gastsystem

Mit einer Firewall auf dem Gastsystem ist das Konzept gemeint, welches im Allgemeinen als „hostbasierte Firewall“ beschrieben wird. Dieser Begriff soll jedoch im Rahmen dieser Diplomarbeit nicht verwendet werden, da der Begriff „Host“ hier für den physischen Virtualisierungsserver steht (siehe Kapitel 2.1 für die Definition).

Unter einer Firewall beziehungsweise einem Paketfilter im Gastsystem versteht man eine Funktionalität im Kernel der virtuellen Maschine, welche ein- und ausgehende Pakete durch einen Regelsatz filtert. Übliche Implementationen sind *iptables/netfilter* für Linux, die *Windows Firewall* für Windows sowie *pf* für BSD-basierte Systeme.

Sowohl für Linux als auch für BSD-basierte Systeme stehen viele Projekte zur Verfügung, welche die vergleichsweise komplexe Konfiguration der Filtersysteme benutzerfreundlich kapseln. Zu diesen Projekten gehört beispielsweise die auf SuSE-Systemen standardmäßig installierte *SuSEfirewall2*, welche dem Benutzer einfache Kommandos zur Freischaltung von Diensten bereitstellt. So ist beispielsweise die Öffnung des TCP-Ports 80 zur Kommunikation mit HTTP, beispielsweise für einen Webserver, mit dem folgenden Kommando durchführbar:

```
SERVER# SuSEfirewall2 open DMZ TCP 80
```

Wenn jedoch kompliziertere Regelsätze benötigt werden stößt die SuSE-Firewall im Bezug auf die Einstellungsmöglichkeiten an ihre Grenzen. Hier kann der Nutzer dann (neben der selbstverständlich möglichen vollständig manuellen Konfiguration) aus verschiedenen Programmen wählen. Vergleichsweise nah an der tatsächlichen Konfiguration und damit eher eine Hilfe für den erfahrenen Administrator sind zum einen das Projekt *ferm* („for easy rule making“), zum anderen die aus einer studentischen Arbeit am LRZ entstandene *LRZ-Firewall* [Mül 10]. Beide Projekte können jedoch auch so konfiguriert werden, dass einfache Portfreischaltungen nach Vorbild der SuSE-Firewall mit einem einfachen Kommando durchführbar sind. Eine deutlich weitreichendere Abstraktion bieten Projekte wie *Shorewall*.

Ein Vorteil von diesen Systemen gegenüber einer externen Firewall ist der mögliche Zugriff auf die Prozessverwaltung, welcher weitergehende Filterungsmöglichkeiten erlaubt. So ist es möglich Regelsätze zu erstellen, die von der Benutzerkennung oder dem Prozess, der die Verbindung aufbauen oder annehmen will, abhängen. Diese Informationen werden nicht im IP-Paket übertragen und stehen daher einer externen Firewall prinzipiell nicht zur Verfügung.

Der Regelsatz ist üblicherweise durch Programme und Nutzer mit Administratorberechtigungen auf dem Host änderbar. Dies ist im Tagesgeschäft oft eine Vereinfachung, ermöglicht einem Angreifer jedoch auch nach einer erfolgreichen Kompromittierung die Firewall selbst zu deaktivieren.

Ein weiterer Nachteil ist die Platzierung des Filters auf der zu schützenden virtuellen Maschine. Bei Firewalls in großen Umgebungen wird oft durch die Firewall eine Trennung in Sicherheitszonen vorgenommen, zwischen denen unterschiedlich starke Vertrauensbeziehungen herrschen. So kann auf der Firewall beispielsweise die Verbindung vom externen Interface blockieren, obwohl die Quell-IP-Adresse des Pakets eigentlich authorisiert wäre. Ein Spoofingschutz verhindert jedoch, ähnlich zu uRPF auf Routern, dass interne IP-Adressen als Quelle auf dem externen Interface verwendet werden können. Bei Filterung auf dem zu schützenden System gibt es jedoch im Allgemeinen nur eine Schnittstelle, nämlich die externe Verbindung. Auf dieser können sowohl legitime als auch gefälschte Informationen eintreffen.

4.4 Sonstige Alternativen

An dieser Stelle sollen noch kurz mehrere Ideen angesprochen werden, welche in Zukunft bei der Betrachtung dieser Frage eine Rolle spielen könnten. Sie sind jedoch noch im Entwicklungsstadium und vermutlich erst in mehreren Jahren einsatzbereit.

4.4.1 OpenFlow/SDN

Ein mittelfristig bedeutsames Thema ist das sogenannte „Software Defined Networking“ (SDN). Dieser Ansatz beschreibt hauptsächlich eine Trennung zwischen der Data-Plane (die Pakete zwischen Switch- und Routerports weiterleitet) und der Control-Plane (zuständig Management und Wegefindung). Diese Trennung wird schon seit Jahren von allen Herstellern in großen Routern praktiziert, jedoch sind hier im Allgemeinen die Control-Plane und die Data-Plane im gleichen Gehäuse zu finden. Ein Beispiel für diese Konfiguration ist der Cisco 6500, dessen MSFC-Submodul die Control-Plane darstellt und die zentrale PFC-Komponente zusammen mit den abgesetzten DFCs in den Schnittstellenkarten die Data-Plane implementiert. Dieses System ist theoretisch in der Lage, bei einem Reset der Control-Plane weiterhin Pakete zu routen. Auch einige proprietäre Ansätze mit einer externen Control-Plane existieren auf dem Markt, zum Beispiel in Form der Juniper Q-Fabric.

SDN zielt jedoch zusätzlich auf eine herstellerunabhängige Schnittstelle zwischen Hardware (Data-Plane) und Software (Control-Plane) ab. Dies ermöglicht theoretisch einen unabhängigen Produktzyklus von Hardware und Softwarekomponenten. Ein bekannter Vertreter ist der *OpenFlow*-Standard, welcher von vielen Herstellern unterstützt wird. OpenFlow definiert das Konzept eines Paketstroms, der durch Filteroperationen auf ein oder mehrere Kriterien definiert wird. Mögliche Kriterien („Flow Match Fields“) in der aktuellen Version 1.2 des Standards sind:

- Layer1-Informationen: Physischer Switchport, logischer Switchport
- Layer2-Informationen: Source-MAC, Destination-MAC, Ethertype, VLAN-Tag
- Layer3/4-Informationen: IPv4/IPv6-Adressen, Protokoll, Ports, QoS-Werte, MPLS-Labels

Trifft ein Paket am Switch ein, werden diese Kriterien über eine dedizierte Ethernet-Verbindung an einen oder mehrere OpenFlow-Controller geschickt. In diesen werden die enthaltenen Informationen analysiert und eine zugehörige Anweisung („Action“) oder auch eine Liste von Anweisungen an den Switch zurückgeschickt. Mögliche Anweisungen sind:

- **Output**: Senden des Pakets an einen oder mehrere definierte Ports
- **Set-Queue**: Setzen einer Warteschlange, zum Beispiel im Bereich QoS
- **Drop**: Verwerfen des Pakets
- **Push-Tag/Pop-Tag**: Hinzufügen oder Entfernen von Markierungen (MPLS oder VLAN)
- **Set-Field**: Überschreiben eines Felds im Paket
- **Change-TTL**: Setzen, Verringern oder Kopieren von TTL-Werten

Damit nicht für jedes Paket eine Anfrage an den Controller gestellt werden muss können durch den Controller Kombinationen aus Kriterien und Aktionen in die Hardware einprogrammiert werden. Alle Pakete, die auf einen bestehenden Eintrag passen, werden ohne erneute Anfrage direkt verarbeitet. Ein klassischer Layer2-Switch würde beispielsweise nur auf die VLAN-ID und die Ziel-MAC-Adresse filtern, während ein Layer3-Router zur Routingentscheidung mindestens die IPv4/IPv6-Ziel-Adresse benötigt. Es können jedoch auch komplexe Filter und Operationen verwendet werden, welche in klassischen Layer2- und Layer3-Architekturen nicht möglich sind. OpenFlow-Controller können außerdem als zentrale Komponenten eine Gesamtsicht über die Netztopologie, die darin verwendeten und verfügbaren Bandbreiten und sonstige Informationen haben. Dadurch ist es theoretisch möglich, die Nachteile von geswitchten Netzen zu umgehen.

Große Teile des OpenFlow-Standards sind jedoch optional und werden nicht von jeder Hardware unterstützt. So implementiert der HP Procurve 5400zl in aktuellen Firmware-Versionen einen OpenFlow-Switch gemäß der Version 1.0 des Standards, unterstützt jedoch nur sieben Kriterien in Hardware (Source-IPv4, Destination-IPv4, L4-Protokoll, L4-Sourceport, L4-Destinationport, VLAN und Interface). Die zur freien Programmierung eines Switches interessanten MAC-Adressen werden nicht in Hardware behandelt, sondern belasten den Prozessor.

Weitere Nachteile der OpenFlow-Technologie sind die Abhängigkeit von den zentralen Controllern, die oft über dedizierte Leitungen an jeden OpenFlow-Switch angebunden sind, und die generell steigende Komplexität des Netzes. Auch ist der Speicherplatz für die in Hardware behandelten Flow-Einträge begrenzt, wodurch im Realbetrieb schnell eine Überlastung der Controller auftreten kann.

Das auf SDN und OpenFlow-basierte Lösungen für Xen- und KVM-Virtualisierungsumgebungen spezialisierte Unternehmen *Nicira* wurde im Juli 2012 von VMware übernommen[VMwa 12]. Bisher sind allerdings noch keine Produkte aus dieser Neuaquisition entstanden.

4.4.2 IEEE 802.1Qbg und 802.1Qbh

Einen völlig anderen Ansatz nutzen im Gegensatz dazu die IEEE-Protokolle 802.1Qbg und 802.1Qbh. Sie verlagern die Zuständigkeit für Sicherheit und damit die Komplexität nicht in den Hypervisor, sondern entfernen dessen Intelligenz im Gegensatz dazu vollständig aus dem Datenpfad. Verkehr von virtuellen Maschinen auf dem gleichen Host (im gleichen VLAN) wird nicht über den virtuellen Switch des Hypervisors – im Umfeld dieses Standards auch als *Virtual Ethernet Bridge (VEB)* bezeichnet – lokal und damit innterhalb des Hypervisors gebridged, sondern im sogenannten *Virtual Ethernet Port Aggregator (VEPA)*-Modus direkt an einen 802.1Qbg-fähigen Uplink-Switch geschickt (Abbildung 4.17). Dieser kann den Verkehr messen, überwachen oder filtern. Ist die Ziel-VM auf dem gleichen Host wie die Quelle muss den Verkehr unter Umständen (konträr

zur ursprünglichen Definition einer Ethernet-Bridge, welche dieses Verhalten explizit verbietet) über den gleichen Link wieder an den Quell-Host zurückgeschickt (*Hairpin Switching*) werden, weshalb der VEPA-Modus vom Switch direkt unterstützt werden muss.

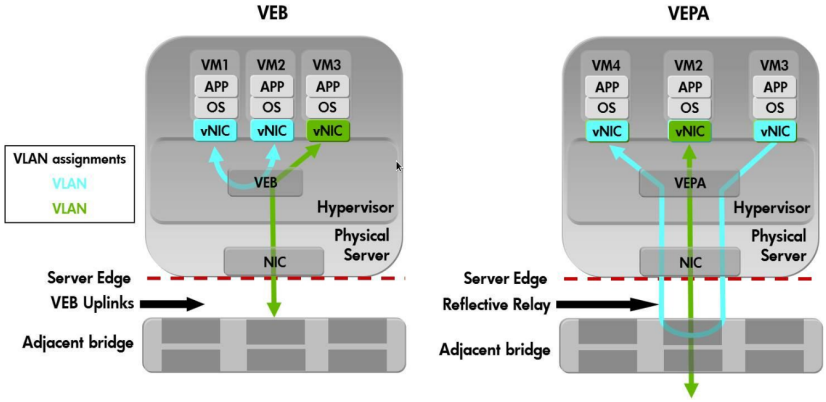


Abbildung 4.17: Vergleich Verkehrsfluss VEB und VEPA, Quelle HP [HP 11]

Eine Erweiterung ist das sogenannte *Multi-Channel VEPA*, bei welchem der Verkehr nicht nur über den physischen Uplink-Switch umgeleitet, sondern auch noch pro virtueller Maschine entsprechend markiert wird. Dies wird durch die Nutzung doppelter VLAN-Tags (QinQ) erreicht und ermöglicht noch feingranularere Filter auf dem Switch.

4.5 Fazit

Die Ergebnisse der Einzeltests sind in der Tabelle 4.7 zusammengefasst.

Produkt	L2-1	L2-2	L2-3	L2-4	L25-1	L3-1	L3-2	L3-3	L4
VMware vSwitch	++	++	-	-	-	-	-	-	-
Cisco Nexus 1000V	o	- (Bug)	++	++	+ (IPv6: -)	+ (IPv6: -)	+ (IPv6: -)	-	o
dedizierte VLANs[1]	+/-	n/a	o	+/-	+/-	+/-	+/-	+/-	-
private VLAN	-	-	-	o	o	o	+	+	-
virtuelle Firewall[2]	(+/-)	n/a	(o)	(+/-)	(+/-)	(+/-)	(+/-)	(+/-)	++
shared Firewall	-	-	-	-	-	-	-	-	++
Firewall Gastsystem	-	-	-	-	-	-	-	-	+

Tabelle 4.7: Testresultate

Zusammenfassend muss leider festgestellt werden, dass keine Lösung alle Anforderungen erfüllt, die an einen sicheren Betrieb gestellt werden.

Der Standard-vSwitch von VMware bietet auf den Ebenen 2.5 und aufwärts keinerlei Sicherheitsfunktionalitäten mehr. Das Alternativprodukt Cisco Nexus 1000V bringt einen signifikanten Mehraufwand im Betrieb mit sich, verhindert aber nur einen Teil der Angriffe wirksam. Selbst diese Sicherheitsfunktionen stehen zum Teil nur bei der Nutzung von DHCP zur Konfiguration der virtuellen Maschinen zur Verfügung, was bisher noch nicht der Fall ist. Die Implementierung der Sicherheitsrichtlinie durch Konfiguration und gegebenenfalls Austausch des virtuellen Switches ist daher nicht möglich.

Die beste Bewertung bezüglich der funktionalen Anforderungen liegt bei der Nutzung von sehr kleinen, im Idealfall nur mit einem Gast versehenen dedizierten VLANs vor, welche die Schwächen der virtuellen Switches egalisiert. Aufgrund der beschränkten Ressourcen (VLAN-IDs, IPv4-Adressen) ist dies jedoch nicht für alle Kunden möglich.

[1] in anderem VLAN / im gleichen VLAN
[2] Bewertung in Klammern vom impliziten dedizierten VLAN

Die Nutzung von Paketfiltern hingegen ist die einzige Möglichkeit, Angriffe auf Applikationen über das Netz zu verhindern. Sie bieten für sich allein gestellt jedoch keinerlei Sicherheit vor lokalen Spoofing-Angriffen auf den unteren Schichten.

4.5.1 Bewertung der nicht-funktionalen Anforderungen

Die nicht-funktionalen Anforderungen, die im Kapitel 2.2 definiert wurden, können nicht quantitativ, sondern nur qualitativ bewertet werden.

NF1 – Sicherheit

Mit Ausnahme des Produkts Cisco 1000V (Kapitel 4.2.2), welches in der aktuellen Version einen erschreckenden Fehler in der Behandlung von 802.1Q-VLAN-Paketen aufweist, erfüllen alle Produkte dieses Kriterium. Es werden keine prinzipbedingten zusätzlichen Angriffsvektoren geöffnet. Zusätzliche Programme und Managementschnittstellen, wie sie beispielsweise bei der Implementierung eines Selfservice-Portals zur zentralen Provisionierung einer Firewall (Kapitel 4.3.4) benötigt werden, können jedoch eine Sicherheitslücke darstellen und müssen daher sicher ausgelegt werden. Hier muss auch eine Beeinflussung der anderen Mandanten verhindert werden.

NF2 – Komplexität

Auch in diesem Punkt schneidet das System Cisco Nexus 1000V etwas schlechter ab, da der zusätzliche Betrieb von virtualisierten, redundanten Managementsystemen (VSM) einen nicht zu verachtenden Zusatzaufwand darstellt. Die Kommunikation zwischen den Managementsystemen und den virtuellen Switchmodulen auf den VMware-Hosts (VEM) ist nicht immer leicht in der Handhabung. Die Komplexität sowohl in der Konfiguration als auch in den Fähigkeiten ist gegenüber den herkömmlichen virtuellen Switches stark erhöht, so dass hier (Cisco-)Fachwissen zur Administration und Fehlersuche nötig ist. Ein weiterer Nachteil ist die fast zwangsläufige Nutzung von DHCP zur Adressvergabe.

Eine ähnliche leichte Abwertung erhält erneut ebenfalls die zentral provisionierte Firewall, da hier zwangsläufig viele Verarbeitungsstufen zwischen dem Selfservice-Portal und dem tatsächlich ausgerollten gemeinsamen Regelsatz liegen. Konsistenzprüfungen müssen erfolgen und Fehler an die Benutzer zurückgemeldet werden.

NF3 – Mandantenfähigkeit

Die Lösungen unter Einsatz des VMware dvSwitch (4.2.1) sowie dedizierte oder private VLANs haben keine Parameter, welche im Regelfall durch den Kunden beeinflusst werden sollten. Sie sind daher von dieser Anforderung nicht betroffen.

Der Cisco Nexus 1000V bietet, genau wie die zentral provisionierte Firewall, einen Zusatznutzen zur Filterung des Datenverkehrs. Diese Filter sind prinzipiell für jeden Mandanten seperat einstellbar, im Normalfall aber nur für den Administrator erreichbar. Wenn diese Möglichkeit den Kunden angeboten werden soll, muss also ein Selfservice-Portal und ein entsprechender Mechanismus zur Provisionierung implementiert werden.

Die Firewall auf dem Gastsystem kann ebenso direkt vom Kunden verändert werden wie ein dedizierter Subnetz-Firewall-Kontext (Kapitel 4.3.3 (c)).

NF4 – Benutzbarkeit

Bei der Nutzung sowohl des VMware dvSwitch als auch des Alternativprodukts Cisco Nexus 1000V bemerkt der Nutzer keine Einschränkungen. Auch dedizierte oder private VLANs sind für den Kunden transparent und geben ihm keine zusätzlichen Eingriffsmöglichkeiten.

Die Konfiguration der Firewall hingegen muss aufgrund der erwarteten Häufigkeit und Dringlichkeit von Änderungen durch den Benutzer selbst durchgeführt werden können. Hierzu stehen je nach Variante und Implementation eine Java-Oberfläche (Cisco ASDM für Cisco ASA-Firewalls), eine mit dem Browser bedienbare Web-GUI (beispielsweise im LRZ-Serviceportal oder pfSense), native graphische Betriebssystemschnittstellen (Windows Firewall), Kommandozeilenbefehle (SuSE-Firewall, shorewall) oder Textdateien zur Konfiguration (ferm, LRZ-Firewall) zur Verfügung. Während die Konfigurationen bei der Verwendung aller Fähigkeiten sehr komplex werden kann, sind die Basisaufgaben wie die Freischaltung eines einzelnen Ports für Verbindungen von außen auch für unerfahrene Benutzer leicht möglich.

NF5 – Interoperabilität

Mit Ausnahme der starken (VMware dvSwitch) beziehungsweise schwachen (Cisco Nexus 1000V, dieser existiert auch für die Konkurrenz-Produkte Hyper-V und Xen) Bindung an den VMware Hypervisor entsprechen alle hier vorgestellten Lösungen veröffentlichten Standards und sind damit herstellerübergreifend interoperabel.

NF6 – Kompatibilität

Die einzige hier vorgestellte Lösung, welche eine großflächige Veränderung der Topologie und damit eine Beeinflussung der bestehenden LRZ-Dienste mit sich bringt besteht im Einsatz von privaten VLANs. Durch die Einschränkungen der aktuell eingesetzten HP-Switches, welche einen Private-VLAN-ähnlichen Betriebsmodus nur auf Portebene bieten, müssen entweder Parallelstrukturen aufgebaut (eine Darstellung der nötigen Änderungen ist in Kapitel 5 zu finden) oder alle LRZ-Dienste auf einen Schlag umgestellt werden.

Alle anderen Lösungen können so konfiguriert werden, dass die Sicherheitsmechanismen nur für die externen Kunden und/oder nur für Neuinstallationen aktiv werden.

NF7 – Skalierbarkeit

In diesem Bewertungspunkt schneiden zwei der untersuchten Lösungen deutlich schlechter ab als die anderen.

Dedizierte VLANs erreichen ihr Sicherheitsniveau durch eine möglichst kleine Anzahl von Rechnern (in diesem Fall virtuellen Maschinen), die sich innerhalb des gleichen VLANs befinden. Im Idealfall, in dem ein dediziertes VLAN alle anderen Sicherheitsmechanismen schlägt, befindet sich nur eine einzige VM in einem VLAN. Im Gegenzug heisst dies, dass pro virtueller Maschine ein VLAN benötigt wird. Es stehen nur insgesamt 4096 VLANs zur Verfügung, von denen im MWN bereits mehr als 2000 in Benutzung sind. Es ist prinzipiell zwar möglich, die VLAN-IDs mehrfach zu verwenden, allerdings kann dies im Fall einer Fehlkonfiguration zum „Kurzschluss“ zwischen zwei Broadcast-Domains und damit zu Sicherheitslücken und Störungen führen.

Eine weitere knappe Ressource, die in jedem VLAN benötigt wird, sind die IP-Adressen. Wie in Tabelle 4.5 dargestellt steigt der Overhead bei der Zuweisung von IPv4-Netzen an, je kleiner das Subnetz ist. Bei voller Redundanz und gleichzeitig maximaler Sicherheit müssen für jede virtuelle Maschine acht IPv4-Adressen reserviert werden. Der Ressourcenverbrauch ist also bei dieser Lösung indirekt proportional zum gewünschten Sicherheitsniveau. Dieses kann auch nachträglich nicht mehr ohne weiteres geändert werden, da sowohl beim Zusammenlegen als auch beim Trennen von Servern in einem VLAN zumindest bei einem Teil der Systeme die IP-Adressen geändert werden müssen.

Wird statt den herkömmlichen 802.1Q-basierten VLANs das VXLAN-Protokoll verwendet, so stehen ausreichend eindeutige VLANs zur Verfügung.

Ebenfalls schlecht skalierend ist die Lösung der Subnetz-Firewalls, sofern die bereits im LRZ eingesetzte „virtuelle Firewall“ basierend auf Cisco ASA-Kontexten verwendet wird. Von diesen stehen selbst im derzeit größten Modell nur 250 zur Verfügung. Zusätzlich benötigt diese Variante mindestens zwei dedizierte VLANs pro Kunde, mit den zuvor beschriebenen Skalierungshemmnissen.

4.5.2 Gesamtfazit

Zur Gesamtbewertung der vorgeschlagenen Lösungen werden die erreichten Bewertungen in den funktionalen und nicht-funktionalen Anforderungen in ein Punktesystem übersetzt. Hierbei ergibt sich die Schwierigkeit, dass die Sicherheit von Firewallstrukturen von der Sicherheit der zugrunde liegendne Netz-Infrastruktur abhängt und daher für sich allein nicht bewertet werden kann. Es werden daher zunächst die vier zur Verfügung stehenden eigenständigen Strukturen

- VMware dvSwitch
- Cisco Nexus 1000V
- dedizierte VLANs
- private VLAN

bewertet, indem je nach erreichter Bewertung der funktionalen Anforderungen Punkte vergeben werden.

- ++ (übererfüllt) ergibt +2 Punkte
- + (erfüllt) ergibt +1 Punkt
- o (ausreichend) oder n/a (nicht bewertbar) ergibt 0 Punkte
- - (nicht erfüllt) ergibt -2 Punkte

Bei den nicht-funktionalen Anforderungen werden für herausragend schlechtes Abschneiden pro Anforderung bis zu zwei Minus- oder Pluspunkte vergeben.

Dieses Schema ergibt das in der Tabelle 4.8 angegebene Resultat:

Anforderung	dvSwitch	1000V	dedizierte VLAN	private VLAN
L2-1	2	0	1	-1
L2-2	2	-1	0	-1
L2-3	-1	2	0	-1
L2-4	-1	2	1	0
L25-1	-1	0	1	0
L3-1	-1	0	1	0
L3-2	-1	0	1	1
L3-3	-1	0	1	1
L4	-1	0	-1	-1
$\sum$ Funktional	-3	3	5	2
NF 1	0	-1	0	0
NF 2	0	-1	0	0
NF 3	0	1	0	0
NF 4	0	0	0	0
NF 5	0	-1	0	0
NF 6	1	0	0	-2
NF 7	0	0	-2	0
$\sum$ Nicht-Funktional	-3	3	5	2
$\sum$ **Gesamt**	-2	1	3	-4

Tabelle 4.8: Gesamtresultat

Die Firewall-Lösungen können untereinander ebenfalls verglichen und bewertet werden, die Gewichtung und damit die Rangfolge hängt jedoch von den individuellen Anforderungen des jeweiligen Kunden ab. Eine allgemeingültige Platzierung ist daher nicht möglich.

5 Lösungen im Detail

Wie man an der Bewertung im vorangegangenen Kapitel sehen kann, gibt es keine Technologie, die für sich alleine gestellt alle Anforderungen erfüllen kann.

Der in VMware integrierte vSwitch unterstützt nur minimale Sicherheitsvorkehrungen und ist für sich allein gestellt nicht geeignet, eine Absicherung in einem geteilten Netz mit vielen virtuellen Maschinen herzustellen. Die kostenlose Essential-Edition des Cisco Nexus 1000V bietet hier einige zusätzliche Möglichkeiten, zeigt jedoch besorgniserregend Schwächen bei der Verhinderung von MAC-Spoofing und enthält außerdem ein Problem beim VLAN-Tagging, welches zumindest ein schlechtes Licht auf die Implementierung wirft.

Die einzige aus Sicherheitssicht annehmbare Lösung für alle Probleme der Schichten 2-3 ist das dedizierte VLAN für jede einzelne Maschine. Dies ist jedoch aus Gründen der Ressourcenknappheit mit den bekannten Mitteln utopisch. Es müssen daher auch in Zukunft mehrere Gastsysteme in einem VLAN zusammengeschaltet werden, die gegenseitig wieder über die meisten Angriffe verwundbar sind.

Die Nutzung des Private-VLAN-Konzepts innerhalb dieser VLANs würde zumindest einige Probleme entzerren und zusätzlichen Schutz bieten. Diese Funktionalität steht jedoch in der aktuellen LRZ-Infrastruktur nicht ohne Weiteres zur Verfügung. Zum einen unterstützen die Flex10-Module Private-VLANs nur global für alle VLANs in einem *vNet*. Dieses wird definiert durch eine Menge an Uplink-Ports zu Switches und Downlink-Ports zu Blades, wobei diese Ressourcen jeweils exklusiv belegt sind. Eine Nutzung wäre daher nur möglich, wenn dedizierte Uplink-Schnittstellen für VLANs mit Private VLAN bereitgestellt werden. Zum anderen unterstützen die verwendeten HP-Switches dieses Feature nicht. Da diese eine direkte Verbindung zwischen den Hosts herstellen wird das Sicherheitskonzept der Private VLANs ausgehebelt, wie in Kapitel 4.3.2 beschrieben wird.

Eine lange verfolgte Idee war die Emulation von Private-VLAN durch MAC-Accesslisten auf dem Nexus 1000V. Private VLANs basieren darauf, dass die direkte Kommunikation zwischen den Gastsystemen unterbunden wird. Im Standard muss dies auf jedem einzelnen Switch passieren, da jeder Switch für sich allein diese Entscheidung treffen muss. In einer hypothetischen Umgebung von einem VLAN, welches ausschließlich einen Router und eine beliebige Anzahl von virtuellen Maschinen beinhaltet, gibt es jedoch noch ein anderes Kriterium für die Herkunft des Pakets, welches von Switches sogar transparent durchgeleitet wird. Dies ist die Quell-MAC-Adresse, welche bei den virtuellen Maschinen immer aus dem Bereich des Herstellers kommt (siehe Tabelle 3.1). Wenn dieser Verkehr am Nexus 1000V zu virtuellen Maschinen hin verworfen wird, ist eine dem Private VLAN äquivalente Separierung erreicht.

Eine mögliche Konfiguration am Beispiel einer VMware-Infrastruktur (Prefix: 00:50:56) könnte folgendermaßen aussehen:

```
mac access-list DROP-FROM-VM
  10 deny 0050.5600.0000 0000.00ff.ffff any
  20 permit any any

port-profile type vethernet Testuser
  mac port access-group DROP-FROM-VM out
```

Leider zeigte sich bei den Untersuchungen des Nexus 1000V (Kapitel 4.2.2 Punkt L2-3), dass diese MAC-ACLs nur Verkehr beeinflussen, der nicht IPv4-Verkehr ist. IPv4-Verkehr wird durch seperate IPv4-ACLs behandelt, wobei bei diesen nicht auf die MAC-Adresse gefiltert werden kann. Dies bedeutet, dass eine Private VLAN-Simulation durch ACLs für IPv6- und ARP-Verkehr aktuell möglich ist, jedoch nicht für IPv4. Es steht zu erwarten, dass IPv6-Verkehr ab der im ersten Quartal 2013 zu erwartenden Unterstützung von IPv6-ACLs ebenfalls nicht mehr durch MAC-ACLs behandelt wird. Diese Lösung scheidet daher leider aus.

Alle auf Firewalls basierenden Sicherheitsmechanismen helfen ausschließlich beim Schutz vor Angriffen auf Layer 4 und höher. Sie können Angriffe auf versehentlich erreichbare Dienste verhindern oder den Zugriff auf authorisierte IP-Adressen beschränken. Sie helfen jedoch nichts bei Sicherheitslücken oder schlechter Konfiguration (beispielsweise Standardpasswörter) von absichtlich weltweit erreichbaren Diensten. Auch können sie nicht gegen Angriffe auf den Schichten 2 und 3 innerhalb des gleichen VLANs schützen und sind daher nur als Zusatzschutz zu gebrauchen.

Von den in Kapitel 4 untersuchten Varianten bleiben daher bei realistischer Betrachtung nur zwei LÖsungen übrig, die im gegebenen Szenario einsetzbar sind und einen Fortschritt im Bezug auf die Sicherheit bieten. Diese werden im Folgenden genauer untersucht.

5.1 Konventionelle Lösung - Sicherheitsfunktionen im virtuellen Switch

Eine erste denkbare Lösung basiert auf seit Jahren erprobten Standardverfahren, wie sie auch im Bereich des physischen Serverhostings benutzt werden. Aufgrund der Vielzahl der Komponenten im Bereich der Virtualisierung ist es zwar schwieriger diese einzusetzen, jedoch mit einigen Einschränkungen möglich.

Trotz allen Einschränkungen und Nachteilen ist für ein geteiltes VLAN mit vielen einfachen virtuellen Maschinen das Konzept der Private VLANs als Basis nötig. Andernfalls sind die Sicherheitsprobleme nur schwer in den Griff zu bekommen. Außerdem müssen die zugesicherten Fähigkeiten der virtuellen Server stark eingeschränkt werden, damit die Sicherheitsmechanismen funktionieren können. Auch wenn derzeit noch nicht alle diese Richtlinien von der Netzseite her erzwungen werden können, so wird dies jedoch hoffentlich im Rahmen der Weiterentwicklung möglich sein.

5.1.1 Produktdefinition Virtueller Server

Das Standardprodukt sollte die folgenden Einschränkungen umfassen:

- Nutzung einer Netzwerkkarte mit einer MAC-Adresse aus dem VMware-Bereich
- Nutzung einer IPv4-Adresse, bezogen per DHCPv4
- Nutzung einer IPv6-Adresse, bezogen per stateful DHCPv6
- Kommunikation nur mit IPv4 und IPv6
- Kein Multicast- oder Broadcastverkehr zwischen den teilnehmenden Systemen

Die Nutzung von stateful DHCPv6 ist aktuell noch nicht mit allen Betriebssystemen problemlos machbar. Dennoch ist anzunehmen, dass die in der Zukunft bereitstehenden IPv6-Sicherheitsfunktionen des Nexus 1000V stateful DHCPv6 als einzige sinnvolle Quelle von validen MAC-IPv6-Port-Zuordnungen unterstützen werden.

Im Umkehrschluss schließt dieser Funktionsumfang eine Nutzung von mehreren IP-Adressen oder wechselnden MAC-Adressen aus. Auch eine Nutzung von Hochverfügbarkeits-Clustern, die IP-Adressen zwischen mehreren Rechnern verschieben (*Heartbeat*, *carp*) oder gar mit Multicast arbeiten (*Microsoft NLB*) scheidet hier aus. Eine Kombination mit physischen Servern ist ebenfalls nicht gestattet.

Diese definierten Einschränkungen haben das Ziel, diese Kunden bei Verfügbarkeit von tragfähigen Private-VLAN-Strukturen und entsprechenden Sicherheitsmechanismen für die Verifikation von IP- und MAC-Adressen ohne Rückfrage und ohne unerwartete Änderungen umstellen zu können.

Kunden, die mit diesen Einschränkungen nicht einverstanden sind, muss ein Angebot für ein dediziertes VLAN gemacht werden, in dem sie ihre virtuellen Maschinen betreiben können. Dieses VLAN kann jeweils nur von einem einzigen Projekt genutzt werden. Sicherheitsmaßnahmen werden in Absprache mit dem Kunden gelockert. Dieses VLAN kann bei Bedarf sowohl von virtuellen als auch von physischen Servern verwendet werden. Das Angebot sollte aufgrund der begrenzten Ressourcen von VLAN-IDs kostenpflichtig sein und die

entstehenden Kosten decken, wobei sowohl Einrichtungskosten als auch monatliche Kosten anfallen sollten. Auch eine Kombination mit einem Kontext der virtuellen Firewall ist hier denkbar.

5.1.2 Technische Implementierung

Wie bereits mehrfach angesprochen unterstützen die in den Bladecentern verbauten Flex10-Module einen Private-VLAN Modus nur global für alle VLANs einer vNet-Gruppe vor. Daher müssen die vorhandenen 2*2*10GE Verbindungen in zwei 2*10GE Gruppen aufgeteilt werden, von denen eine Gruppe mit Private VLANs konfiguriert wird und die andere Gruppe mit Standardkonfiguration den restlichen Verkehr bedient. Dieser umfasst neben den alten Netzen, bei denen eine Einführung von Private VLANs nur mit großen Umstellungen möglich ist, Management, vMotion und Storage.

Da die bisherige Anbindung an das Rechenzentrumsnetz durch die „VMware-Switches“ nicht Private-VLAN fähig ist und dies wegen dem Einsatz von Spanning-Tree auch nicht sinnvoll erscheint, ist ein Ersatz dieser Infrastruktur zumindest für Uplinks der Private-VLAN-Gruppe nötig. Zu diesem Zweck wird vorgeschlagen, das ab Anfang 2013 durch eine neue Routerplattform ersetzte Hausrouterpaar (VSS) direkt zur Anbindung von VMware einzusetzen. Da dieser Private-VLAN unterstützt kann dieser Sicherheitsmechanismus eingesetzt werden. Die resultierende Netztopologie ist in Abbildung 5.1 zu sehen.

Die zweite Uplink-Gruppe mit den regulären VLANs kann je nach Bedarf entweder an den bisherigen „VMware-Switches“ verbleiben, oder ebenfalls auf die VSS geschwenkt werden. Dies hätte den Vorteil, dass dieser Verkehr auf Schicht 3 direkt geroutet oder aber ohne den Einsatz von Spanning-Tree an das Rechenzentrumsnetz abgegeben werden könnte. Auch unterstützt die VSS eine chassisübergreifende Bündelung der Links durch das LACP-Protokoll, wodurch die verfügbare Bandbreite wesentlich besser ausgenutzt wird.

An Stelle des regulären vSwitch kann, auch wenn er sicherheitsmäßig nicht alle Hoffnungen erfüllt, die kostenpflichtige Advanced-Edition des Cisco Nexus 1000V eingesetzt werden. Die freie Essential-Edition verhindert kaum Angriffe auf Layer 2.5 und Layer 3, so dass ein Migrationsaufwand nicht gerechtfertigt erscheint. DHCP-Snooping, DAI und IPSG hingegen bieten einen handfesten Sicherheitsvorteil gegenüber dem Standard-Switch.

Da der Nexus 1000V von der Konfiguration her deutlich mehr Möglichkeiten bietet als der integrierte vSwitch stellt sich die Frage der Zuständigkeit bei der Administration. Auch wenn im Namen der Begriff „Switch“ vorkommt, erstrecken sich die Konfigurationsoptionen von Layer 2 über Layer 3 (Sicherheitsfunktionen) bis weit in die VMware-spezifische Welt hinein. Es wird daher davon abgeraten, die Administration fest einer bestimmten Gruppe zuzuordnen. Stattdessen sollte die Verwaltung kollegial und gleichberechtigt ohne feste Zuständigkeiten durch das Fachpersonal der drei betroffenen Abteilungen erfolgen. Dies funktioniert, von wenigen Ausnahmen abgesehen, im Allgemeinen im LRZ sehr gut, da niemand ohne Not Änderungen an Systemen durchführt, über deren Auswirkungen Unklarheit besteht. Im Gegenzug erhöht es die Qualität der Lösungen enorm, wenn die beteiligten Personen auch Erfahrungen außerhalb ihres Standardgebietes sammeln und damit einen Überblick über die Gesamtstruktur erhalten.

Der Einsatz des DHCP-Snoopings als Quelle für Sicherheitsinformationen impliziert die Nutzung von DHCP zur Konfiguration des Gasts. Betriebssysteme, die DHCP nicht oder in nicht ausreichendem Maße unterstützen, können auch auf unattended VMs gemäß der Produktspezifikation nicht eingesetzt werden, da ohne einer gültige DHCP-Adresse kein Netzzugriff möglich ist. Ebenso ist eine Nutzung von mehreren IP-Adressen nicht möglich. Diese Sonderfälle müssen auf das kostenpflichtige Produkt eines dedizierten VLANs verwiesen werden.

Die IPv6-Anbindung der virtuellen Server ist noch nicht abschließend definierbar, da Sicherheitsfunktionen des Nexus 1000V für IPv6 erst im Laufe des Jahres 2013 zur Verfügung stehen werden. Basierend auf den bereits für die Cisco 7600 Routerserie implementierten Fähigkeiten ist davon auszugehen, dass eine volle Sicherheit gegen Spoofing nur unter der Nutzung von stateful DHCPv6 zur Adressvergabe, ähnlich zum aus IPv4 bekannten DHCP-Snooping, möglich sein wird. Bei der Nutzung anderer Methoden wie SLAAC oder statischer Adressierung können mangels einer authoritativen Quelle korrekter Adresszuweisungen nur Bindungen aus dem Verkehr erlernt und gegen Spoofing geschützt werden, ein Spoofing von unbenutzten oder über einen längeren Zeitraum nicht aktiven Adressen aber möglich ist.

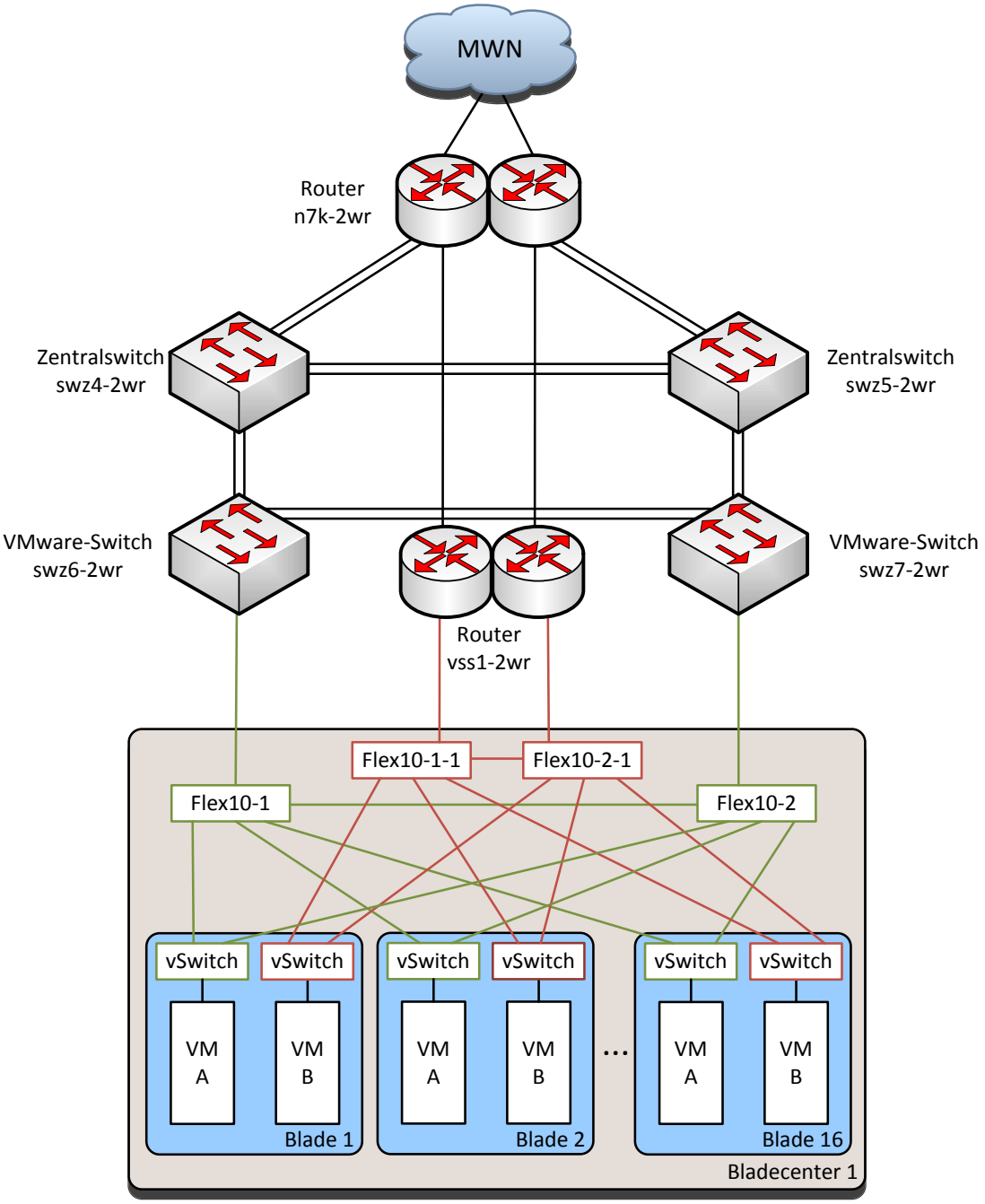


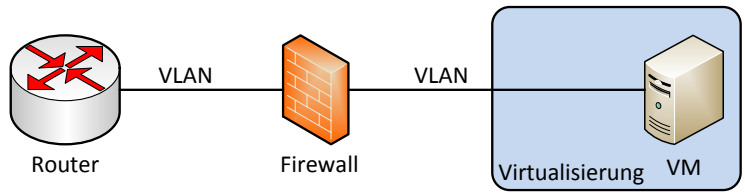
Abbildung 5.1: Netztopologie mit Private VLAN

Es wird empfohlen, die in Kapitel 2.1.2 definierten Sammel-VLANs für externe Kunden und dabei insbesondere die Trennung zwischen *attended* und *unattended* beizubehalten. Die MWN-weite Erreichbarkeit wird in IPv4 wie bisher über die Zuweisung einer privaten IPv4-Adresse erreicht. Für IPv6 ist im Münchner Wissenschaftsnetz auf absehbare Zeit nicht an eine Nutzung von IPv6-Netzen aus dem nicht weltweit gerouteten ULA-Bereich gedacht, so dass auch diese Netze globale IPv6-Adressen erhalten werden. Die genaue Absicherung gegen Zugriffe von außen ist noch in der Diskussion.

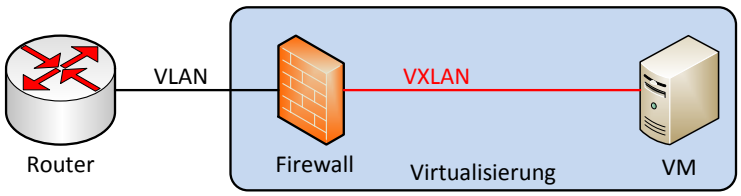
Neben der Wahl der IPv4-Adresse finden auch auf dedizierten VLANs keine Filterung über statische Acces-

slisten am Router statt, da diese nicht über eine mandantenfähige Lösung verwaltet werden können. Aufgrund der begrenzten Hardware-Ressourcen und der durch die Verwaltung und insbesondere Änderung großer Accesslisten auftretenden Prozessorlast wird ebenfalls davon abgeraten, analog zum Vorschlag einer geteilten Firewall in Kapitel 4.3.4 eine zentrale Accessliste durch ein mandantenfähiges Portal zu generieren und auf den Router zu laden, zumal diese Filter nur statische Paketfilter sind und damit ohne Nebenwirkungen keinen ausreichenden Schutz bieten können.

Sofern die Nutzung einer netzseitigen Firewall gewünscht wird, kann diese durch einen Kontext auf der virtuellen Firewall bereitgestellt werden. Hierbei werden jedoch zwingend zwei dedizierte VLANs (outside/extern und inside/intern) mit entsprechenden Kosten benötigt. Aufgrund der begrenzten Anzahl der möglichen Kontexte auf einer Firewall und den hohen Kosten für ein derartiges System ist auch hier auf eine entsprechende Preisgestaltung zu achten. Eine andere Möglichkeit ist der eigenständige Betrieb einer Firewalllösung durch den Kunden in einer virtuellen Maschine, die als Gateway fungiert. Neben Open-Source Produkten wie *pfSense* oder diversen mehr oder weniger mächtigen Frontends für das Linux-Paketfiltersystem *netfilter* stehen auch diverse kommerzielle Lösungen wie *Cisco ASA1000V*, *Sophos UTM* und *Stonesoft StoneGate* als fertige virtuelle Appliances bereit. Bei einer derartigen Lösung kann das interne VLAN sogar als VXLAN ausgeführt werden, was 802.1q-VLAN-Tags einspart und die Administration erleichtert (Abbildung 5.2).



(a) VLAN/VLAN mit physischer Firewall



(b) VLAN/VXLAN mit virtualisierter Firewall

Abbildung 5.2: Virtualisierte Firewall und VXLAN

5.1.3 Bewertung

Kombiniert werden die funktionalen Anforderungen analog zu Kapitel 4.2.2 folgendermaßen bewertet.

L2-1 – Verhinderung von MAC-Spoofing
Eingeschränkt erfüllt aufgrund den beschriebenen Einschränkungen des Nexus 1000V

L2-2 – Versand von Kontrollpaketen
Voll erfüllt, wenn Versand unbekannter Ethertypes unterbunden wird

L2-3 – Flooding („Storm-Control“)
Voll erfüllt

L2-4 – Versand unbekannter Ethertypes
Voll erfüllt

L25-1 – Verhinderung von ARP-Spoofing
IPv4: Voll erfüllt, wenn DHCP-Snooping und DAI verwendet werden kann

IPv6: nicht erfüllt

L3-1 – IP-Spoofing
IPv4: voll erfüllt, wenn DHCP-Snooping und IP Source Guard verwendet werden kann
IPv6: nicht erfüllt

L3-2 – Rogue DHCP
IPv4: Voll erfüllt, wenn DHCP-Snooping verwendet werden kann
IPv6: nicht erfüllt

L3-3 – RA Guard
Nicht erfüllt, keine IPv6-Filterfunktionalität

L4 – Paketfilter und Firewall
IPv4: Eingeschränkt erfüllt, Konfiguration prinzipiell möglich aber stateless und nicht mandantenfähig
IPv6: nicht erfüllt, keine IPv6-Filterfunktionalität

5.2 Innovative Lösung - VXLAN-basierte dedizierte VLANs

Die zweite mögliche Variante bricht hingegen mit den bekannten Konzepten und nimmt einen anderen Ansatz. Aus Sicht der grundlegenden logischen Netztopologie entspricht diese Konfiguration am ehesten der Nutzung eines VLANs pro virtueller Maschine. Diese Lösung ist, sofern die beteiligten Netzkomponenten keine durch durchlaufenden Verkehr ausnutzbaren Sicherheitslücken aufweisen, gegen alle in diesem Dokument beschriebenen Spoofing-Angriffe immun und daher aus Sicht der Sicherheit das absolute Optimum.

5.2.1 Produktdefinition Virtueller Server

Im Gegensatz zur vorgenannten Lösung gibt es hier kaum technische Anforderungen an die Konfiguration der virtuellen Maschine. Um den Vergleich zu ermöglichen, sollen jedoch hier die gleichen Kriterien angewandt werden. Das Standardprodukt sollte daher die folgenden Einschränkungen umfassen:

- Nutzung einer Netzwerkkarte mit einer MAC-Adresse aus dem VMware-Bereich
- Nutzung einer oder mehrerer IPv4-Adressen oder Subnetze, statisch oder mit DHCPv4 konfiguriert
- Nutzung eines oder mehrerer IPv6-Adressen aus einem Subnetz (/64), statisch, mit SLAAC oder DHCPv6 konfiguriert
- Kommunikation nur mit IPv4 und IPv6
- Kein Multicast- oder Broadcastverkehr zwischen den teilnehmenden Systemen

Die Nutzung mehrerer IP-Adressen ist technisch kein Problem, sollte jedoch nur als ein kostenpflichtiges Standardprodukt verfügbar sein.

Eine Nutzung von Hochverfügbarkeits-Clustern, die IP-Adressen zwischen mehreren Rechnern verschieben (*Heartbeat*, *carp*) oder gar mit Multicast arbeiten (*Microsoft NLB*) scheidet auch hier aus. Eine Kombination mit physischen Servern ist technisch nicht möglich, da dort kein Hypervisor mit VTEP-Funktionalität Verwendung findet. Die Nutzung eines, auch kommerziell verfügbaren, VXLAN-Gateways zur Wandlung zwischen dem VXLAN-Segment und einem 802.1q-VLAN ist hier im Allgemeinen mit zu hohem Aufwand verbunden und bringt keinerlei Vorteile gegenüber der durchgehenden Nutzung eines VLANs auch für die virtuellen Server, so dass in diesem Fall ein herkömmliches geswitchtes VLAN für virtuelle und physische Server Verwendung finden sollte.

Es ist möglich, Kunden ein VXLAN-Segment für mehrere virtuelle Maschinen zur Verfügung zu stellen, wenn die Nutzung von HA oder Clusterkommunikation gewünscht wird. Auch dies ist jedoch eine Sonderkonfiguration, die nur zu entsprechenden Konditionen angeboten werden sollte.

5.2.2 Technische Implementierung

Bei der Nutzung eines VLANs für jede virtuelle Maschine ist jedoch der Vorrat an möglichen 802.1q-VLAN-Tags (4096) schnell erschöpft, zumal bereits etwa 50 Prozent der möglichen Tags im Münchner Wissenschaftsnetz in Benutzung ist. Der oft in Providernetzen benutzte Ausweg durch doppelte VLAN-Tags (Q-in-Q) wird von der Routerplattform im Münchner Wissenschaftsnetz nicht unterstützt. Eine Nutzung der VXLAN-Technologie bringt zunächst keine Verbesserung, da das entsprechende Netz zunächst physisch als VLAN zu einem entsprechenden Gateway übertragen werden muss und dabei die Ressourcen bereits verbraucht werden (Abbildung 5.3a). Die Nutzung eines einzelnen VLANs auf der physischen Seite und mehrerer VXLAN-Segmente auf der Virtualisierungsseite (Abbildung 5.3b) implizit die Funktionalität eines Layer2-Switches im Gateway, mit allen bereits aus den vorherigen Kapiteln bekannten Anforderungen an den Schutz vor Spoofing und anderen Angriffen im lokalen Netzsegment.

Es wird daher eine Lösung benötigt, in der das VXLAN-Gateway die Routingfunktionalität auf Layer 3 übernimmt und damit die einzelnen, jeweils nur eine VM umfassenden VXLAN-Segmente auf Layer 2 voneinander trennt. Eine naheliegende Lösung wäre der Betrieb eines routenden Gateways ähnlich einer Firewall aus dem vorhergehenden Kapitel in einer virtuellen Maschine, welche eine virtuelle Netzwerkkarte in jedem benötigten VXLAN-Segment erhält (Abbildung 5.4). Aus Sicht des Gateways ist damit mit jeder Karte nur ein System verbunden. Diese Lösung skaliert jedoch nicht, da VMware ESXi nur maximal 10 virtuelle Netzwerkkarten pro Container unterstützt [VMmax].

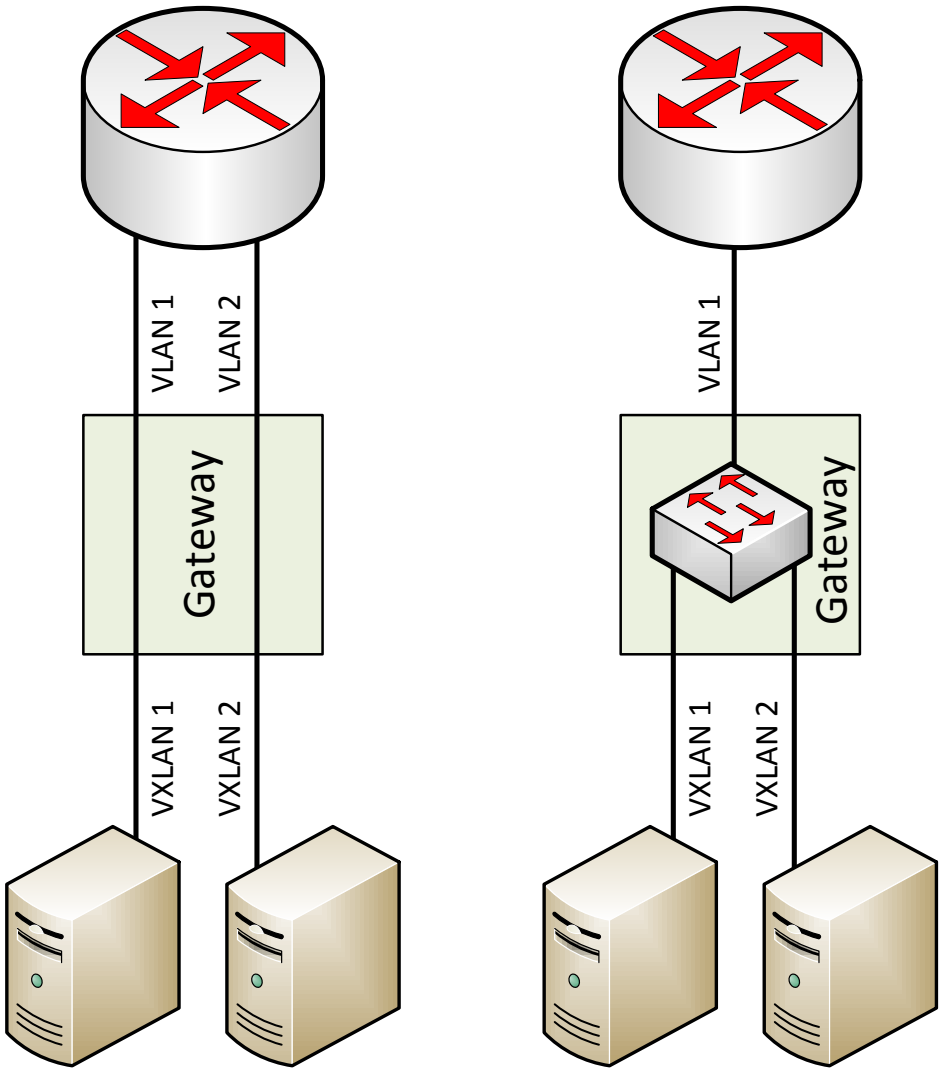
Aus diesem Grund muss die VXLAN-Unterstützung im benutzten Gateway direkt enthalten sein, so dass die benötigte Encapsulierung und Decapsulierung ohne Umwege direkt geschehen kann (Abbildung 5.5). Bisher haben zwei Hersteller Systeme angekündigt, die diese Funktionalität implementieren (F5 BigIP und Brocade ACS). Beide Systeme sind jedoch eher als Serverloadbalancer mit einer Anbindung der Backend-Server über VXLAN gedacht. Die meisten Hersteller verstehen unter VXLAN-Gateways jedoch auf Schicht 2 arbeitende 1:1-Umsetzer zwischen einem 802.1q-Tag und einem VXLAN-Segment.

Eine mögliche Alternative stellt Linux dar, welches beginnend mit der Kernelversion 3.7 direkte VXLAN-Unterstützung bietet und damit sehr viele VXLAN-Segmente direkt anbinden kann. Neben einem reinen Routing, welches schon an sich alle Angriffe auf Schicht 2 und 2.5 über Segmentgrenzen hinweg unterbindet, unterstützt der Linux-Kernel durch seinen integrierten Paketfilter *netfilter* auch alle anderen benötigten Sicherheitsfunktionen wie Spoofing-Filter. Theoretisch wäre es sogar denkbar, über das *Connection Tracking* eine vollständige Firewall mit eigenen Regelsätzen für jedes Segment zu erstellen. Hierbei sollte allerdings darauf geachtet werden, dieses zentrale System nicht mit unnötigen Aufgaben zu belasten.

Ein Problem dieser Lösung ist jedoch der IPv4-Adressverbrauch, wie er bereits in Kapitel 4.3.1 beschrieben wurde. Jede virtuelle Maschine ist in ihrem eigenen VLAN und benötigt für einen Betrieb nach herkömmlichen Regeln ein dediziertes Subnetz, welches neben einer Adresse für die virtuelle Maschine eine Adresse für das Gateway sowie Netzadresse und Broadcastadresse beinhalten muss. Es werden also mindestens vier IPv4-Adressen pro virtueller Maschine verbraucht, von denen drei Adressen nicht nutzbar sind.

Auch dieses Problem ist jedoch durch etwas Kreativität lösbar. Bei *Proxy-ARP* antwortet ein Gerät (im Allgemeinen ein Router) stellvertretend auf ARP-Anfragen für bestimmte IP-Adressen, die sich gemäß Netzadresse und Netzmaske im lokalen Subnetz befinden sollten, dies jedoch aus unterschiedlichen Gründen nicht der Fall ist. Dadurch wird ein Eintrag in der ARP-Tabelle hinzugefügt und das Frame an die MAC-Adresse des Routers adressiert, der das Paket dann nach Belieben routen kann. Proxy-ARP sorgt hier also dafür, dass ein Paket, welches gemäß der Konfiguration direkt an ein Nachbarsystem im gleichen Layer2-Segment geschickt werden sollte, über einen beliebigen Layer3-Weg geroutet werden kann. Diese Technologie kann hier verwendet werden, um trotz der seperaten VLANs keinen Mehrverbrauch von Adressen zu verursachen.

Proxy ARP beziehungsweise Proxy NDP für alle Adressen ist in IPv6 nicht möglich, da ein Proxy der entsprechenden *solicited-node multicast address* Multicast-Gruppe beitreten muss ([RFC4861] 7.2.8). Von diesen existieren 2^{24}, was jede Netzkomponente überfordern würde. Zum Glück ist dies jedoch auch nicht nötig. Zum einen existieren genug IPv6-Netze, um jeder virtuellen Maschine ein dediziertes /64-Netz zuweisen zu können. Zum anderen ist es durch die in IPv6 praktizierte Trennung von Nutzverkehr auf globalen Adressen auf der einen Seite, und Kontrollverkehr auf Link-Local-Adressen auf der anderen Seite möglich, die Default-Route auf eine in allen Subnetzen konstante Adresse zeigen zu lassen.



(a) 1:1-Zuordnung (b) 1:n-Zuordnung

Abbildung 5.3: Zuordnungen zwischen VLAN-Tag und VXLAN-Segment

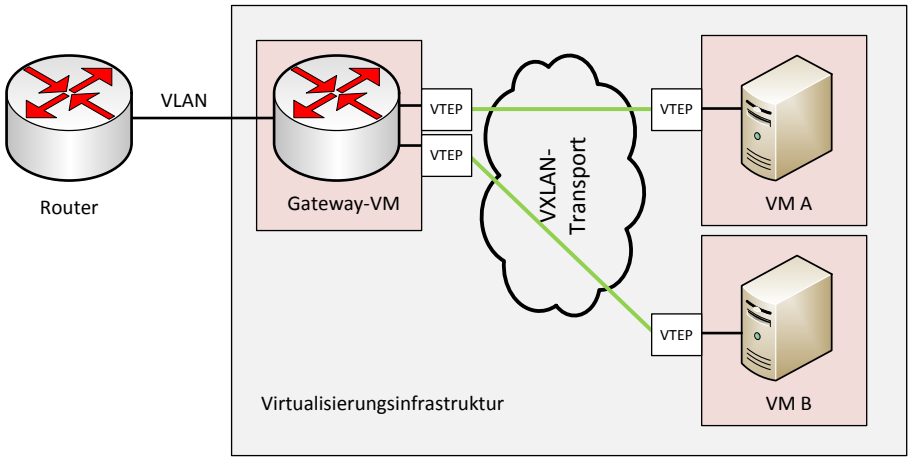


Abbildung 5.4: virtuelles VXLAN-Gateway

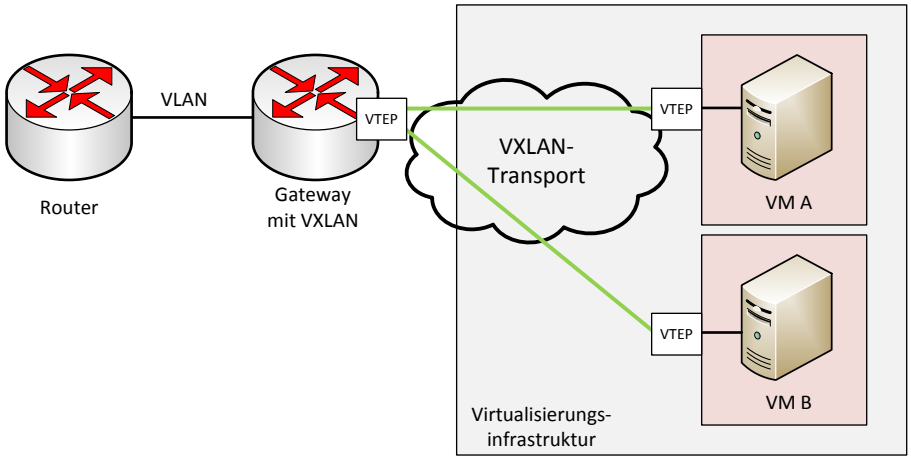


Abbildung 5.5: VXLAN-Gateway mit direktem VXLAN-Support

Als großer Vorteil kann auch gelten, dass die IP-Adressen vollständig unabhängig von der benutzten Netztopologie sind. Zwei IPv4-Adressen innerhalb eines logischen Subnetzes stehen nicht mehr in Verbindung zueinander als IPv4-Adressen aus völlig anderen Netzbereichen. Selbst die Adresse des Default-Gateways wird nur noch benötigt, um den Gast-Betriebssystemen eine normale Konfiguration vorzuspielen. Sie muss noch nicht einmal auf dem benutzten VXLAN-Gateway konfiguriert sein. Es können auch beliebig viele IPv4- oder IPv6-Adressen und -Netze zu den virtuellen Maschinen geroutet werden, ohne die Sicherheit zu gefährden.

Ein kleiner Nachteil dieses Konzepts ist, dass DHCP-Funktionalität für die virtuelle Maschine nicht mehr ohne weiteres zur Verfügung steht. Diese wird beim Angebot virtueller Server am LRZ zur Zeit nicht verwendet, könnte jedoch in Zukunft dazu dienen, Betriebssysteminstallationen vom Netz zu starten. Der Grund für diese Einschränkung liegt in der Spezifikation des DHCP-Relay-Agents, welcher die über Broadcast empfangenen DHCPDISCOVER-Nachrichten in Unicast umwandelt und zu einem oder mehreren DHCP-Servern schickt. Zur Identifikation des Subnetzes, in dem sich der Client befindet, setzt der Relay-Agent das Feld *giaddr* in der DHCP-Nachricht auf die eigene Adresse im Subnetz ([RFC2131] Section 4.1). In diesem Fall besitzt jedoch der VXLAN-Router keine eigene Adresse auf dem VXLAN-Interface. Dieses Konzept ist bei Routern als *unnumbered Ethernet* bekannt. Einige Hersteller wie Cisco ermöglichen dort über Umwege eine Nutzung von DHCP, indem sie die Adresse einer anderen, mehrfach verwendeten Schnittstelle eintragen und die Ant-

worten entsprechend zuordnen [Cisco-DHCPu]. Dem DHCP-Server gegenüber wird hierbei ein klassisches Subnetz mit mehreren Teilnehmern simuliert. Eine in diesem Fall sogar zweckdienlichere Methode wäre, den DHCP-Server unter Beachtung der IP-Adresszuweisungen aus der CMDB selbst auf dem VXLAN-Router zu implementieren und dabei die Information über das eingehende Interface direkt auszuwerten.

5.2.3 Bewertung

Die Bewertung unterscheidet sich abhängig davon, ob als unterliegender virtueller Switch der Standard-vSwitch von VMware oder der Cisco Nexus 1000V eingesetzt wird. Da der VMware-vSwitch in der im LRZ eingesetzten Version 5.0 noch kein VXLAN unterstützt konnte diese Funktionalität nicht getestet werden. Es wird daher hier davon ausgegangen, dass der vSwitch der Version 5.1 keine Rückschritte gegenüber der Version 5.0 machen wird.

L2-1 – Verhinderung von MAC-Spoofing
vSwitch: voll erfüllt
1000V: Eingeschränkt erfüllt aufgrund den beschriebenen Einschränkungen des Nexus 1000V

L2-2 – Versand von Kontrollpaketen
vSwitch: voll erfüllt
1000V: Voll erfüllt, wenn Versand unbekannter Ethertypes unterbunden wird

L2-3 – Flooding („Storm-Control“)
vSwitch: nicht erfüllt
1000V: Voll erfüllt

L2-4 – Versand unbekannter Ethertypes
vSwitch: nicht erfüllt
1000V: Voll erfüllt

L25-1 – Verhinderung von ARP-Spoofing
Voll erfüllt, kein System im Segment dessen Verkehr oder Adresse ein Angreifer übernehmen könnte

L3-1 – IP-Spoofing
Voll erfüllt, alle on-link Adressen gehören zum virtuellen Server, off-link Adressen werden durch uRPF am VXLAN-Gateway geblockt

L3-2 – Rogue DHCP
Voll erfüllt, kein anderer DHCP-Client im gleichen Segment

L3-3 – RA Guard
Voll erfüllt, kein anderer IPv6-Host im gleichen Segment

L4 – Paketfilter und Firewall
Voll erfüllbar, wenn Filterregeln auf dem VXLAN-Gateway definiert werden

In dieser Konfiguration werden keine Sicherheitsfunktionen des Cisco Nexus 1000V benötigt, weswegen die kostenlose Essential-Edition ausreichend wäre. Er bietet mehr Funktionen als der in VMware integrierte vSwitch, ist allerdings auch aufwändiger zu administrieren und benötigt eine Zusammenarbeit zwischen den Spezialisten der einzelnen Abteilungen. Die Zusatzfunktionalität zur Einschränkung der Broadcast-Bandbreiten wiegt den nicht sonderlich engagierte Cisco-Support bezüglich des Sicherheitsproblems beim doppelten VLAN-Tagging und die deutlich aufwändigere und gleichzeitig unsicherere Konfiguration gegen gefälschte MAC-Adressen nicht auf. Sofern die Funktionalität des Nexus 1000V daher nicht an anderer Stelle gebraucht wird wird empfohlen, weiterhin den VMware vSwitch einzusetzen.

Basierend auf dem in Kapitel 4.5.2 aufgestellten Schema erhält die vorgeschlagene Lösung die in der Tabelle 5.1 aufgeführte Gesamtbewertung.

[1] Bewertung nach Datenblatt

	Nexus 1000V		ESXi 5.1 dvSwitch[1]	
Anforderung	Bewertung	Punkte	Bewertung	Punkte
L2-1	+	1	++	2
L2-2	-	-1	++	2
L2-3	++	2	o	0
L2-4	++	2	+	1
L25-1	+	1	+	1
L3-1	++	2	++	2
L3-2	++	2	++	2
L3-3	++	2	++	2
L4	o	0	o	0
$\sum$ Funktional		11		12
NF 1		-1		0
NF 2		-1		-1
NF 3		0		0
NF 4		0		0
NF 5		-1		0
NF 6		0		0
NF 7		0		0
$\sum$ Nicht-Funktional		-3		-1
$\sum$ **Gesamt**		8		11

Tabelle 5.1: Bewertung der VXLAN-basierten Lösung

5.3 Fazit

Die als zweite Lösung vorgestellte VXLAN-basierende Konfiguration schlägt in jeder Hinsicht die Lösungsvariante 1. Während bei dieser einfachen Variante eine von Haus aus unsichere und gegen interne Angriffe nicht gehärtete Topologie durch aufwändige Filter und Heuristiken gegen bekannte Angriffe geschützt werden soll, ist diese Segmentierung durch den Einsatz von vielen VXLAN-Segmenten in der Lösung 2 inhärent gegeben. Allein diese Segmentierung verhindert alle bekannten und zukünftigen Angriffe auf den Ebenen 2 und 2.5.

Der einzige Wermutstropfen besteht in der Nichtverfügbarkeit kommerzieller Systeme für die Rolle des VXLAN-Gateways. Die in der vorgeschlagenen Lösung verwendeten Funktionalitäten ist jedoch eine Kombination aus den Fähigkeiten der meisten am Markt vertretenen Router verbunden mit den Fähigkeiten einfachter VLAN-VXLAN-Gateways. Es kann daher davon ausgegangen werden, dass entsprechende kommerzielle Angebote bald zur Verfügung stehen.

Andererseits ermöglichen die Linux-basierten VXLAN-Gateways zusätzliche Funktionen wie Firewalling und einfaches Debugging mit Standard-Tools wie ping, tcpdump und wireshark, und können horizontal durch das Hinzufügen weiterer Gateways skaliert werden. Der Betrieb eines Linux-basierenden Routers (um nichts anderes handelt es sich bei diesem Gateway) sollte für ein wissenschaftliches Rechenzentrum auch ohne Herstellersupport möglich sein. Durch die selbstständige Implementierung des Konfigurationssystems kann dieses an die bestehenden CMDBs des Leibniz-Rechenzentrums angebunden werden und eine vollständige Automatisierung der Netzschnittstelle bieten.

Für die Nutzer der virtuellen Maschinen ist diese Netzkonfiguration völlig transparent. Die Konfiguration und auch das Verhalten entspricht aus Nutzersicht dem gewohnten Betrieb eines Servers in einem herkömmlichen Netz.

6 Proof of Concept, Migration und Empfehlungen

Nachdem bereits mehrere Ideen zur Lösung des vorgegebenen Problems im Rahmen dieser Diplomarbeit an Fehlern, fehlenden Implementierungen oder sogar dokumentierten Ausnahmen gescheitert sind (als Beispiel sei hier die Private VLAN-Emulation durch MAC-Accesslisten aus Kapitel 5 genannt), wurde die in allen Punkten überlegen bewertete innovative Variante praktisch implementiert. Hierzu wurde die in Kapitel 4.1 aufgebaute Testumgebung benutzt.

6.1 Implementation

Für die beispielhafte Implementation wurde zunächst auf der bestehenden LRZ-Infrastruktur ein auf Debian basierendes VXLAN-Gateway (lxbscDA-VXLAN) installiert. Dieses System erhielt eine Schnittstelle in das VLAN 70, welches als dediziertes Storage-Netz gerade nicht in Verwendung war, und eine Schnittstelle in das Management-Netz (VLAN 69). Die VMware-Testinfrastruktur konnte für den Betrieb dieses Gateways nicht verwendet werden, da zumindest für den Nexus 1000V VXLAN-Tunnel mit einer lokalen virtuellen Maschine nicht möglich sind, sondern zwingend über eine physische Schnittstelle transportiert werden müssen.

Zur Unterstützung von VXLAN wurde ein Kernel mit der Version 3.7-rc6 und eine aktueller Entwicklungsversion der iproute2-Utilities (git HEAD, Commit df5574d06) installiert. Das IPv4-Netz im VLAN 70 wurde halbiert (10.156.252.192/26, siehe Tabelle 4.1), die erste Hälfte als Transportnetz zwischen dem Hausrouter des LRZ und dem VXLAN-Gateway verwendet und die zweite Hälfte zur Nutzung mit virtuellen Maschinen vorgesehen. Zu diesem Zweck wurde dieses Netz auf eine IPv4-Adresse im Transportnetz geroutet, die dem VXLAN-Gateway zugewiesen wurde.

Im VLAN 69 (Management) wurde ein neues privates IPv4-Subnetz zum Transport zwischen den drei VTEPs, dem VXLAN-Gateway und den beiden VMware-Hosts der Testumgebung, angelegt.

Hostname	**IPv4-Adresse**
ESXi Host 1	10.0.0.1/24
ESXi Host 2	10.0.0.2/24
lxbscDA-VXLAN	10.0.0.250/24

Tabelle 6.1: VXLAN-IPv4-Adressen

Auf den ESXi-Hosts wurde auf der Nexus 1000V-Komponente eine neue Portgruppe mit der vxlan-Fähigkeit definiert und über den vSphere-Client auf jedem Host ein vmknic-Adapter in dieser Portgruppe angelegt.

```
port-profile type vethernet VXLAN
  vmware port-group
  switchport mode access
  switchport access vlan 69
  capability vxlan
  no shutdown
  state enabled
```

Nachdem die entsprechende Kommunikation zwischen dem VXLAN-Gateway und den ESXi-Hosts verifiziert wurde konnten die entsprechenden Definitionen der VXLAN-Segmente vorgenommen werden. Hierzu wurden die folgenden Parameter festgelegt:

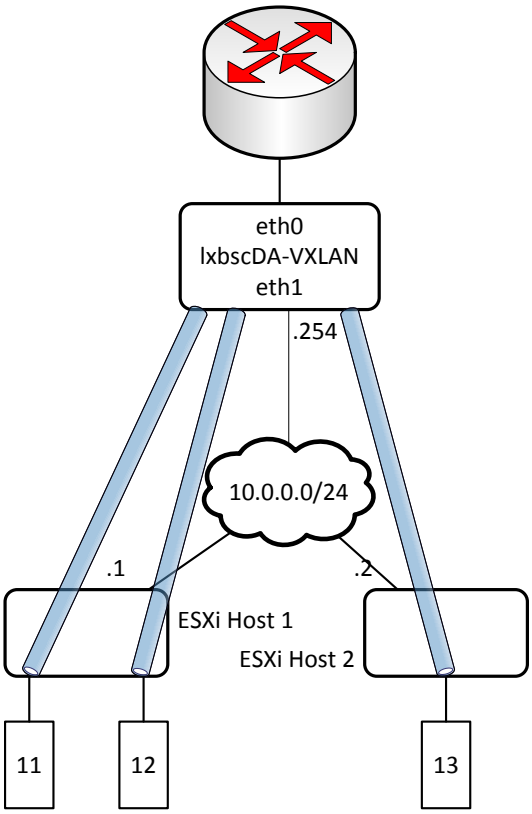


Abbildung 6.1: physische VXLAN-Topologie

Hostname	VXLAN-ID	Multicast-Gruppe	IPv4-Adresse	IPv6-Subnetz
lxbscDA-11	11111	239.0.0.42	10.156.252.227	2001:4CA0:0:FF11::/64
lxbscDA-12	22222	239.0.0.42	10.156.252.228	2001:4CA0:0:FF12::/64
lxbscDA-13	33333	239.0.0.42	10.156.252.229 + .230	2001:4CA0:0:FF13::/64

Tabelle 6.2: Virtuelle Maschinen mit zugehörigen VXLAN-IDs und IP-Adressen

Anhand dieser Parameter konnten nun die Portprofile auf der Seite des Cisco Nexus 1000V definiert werden. Hierzu muss zuerst für jedes VXLAN-Segment eine „bridge-domain“ definiert werden und diese dann in einer Portgruppe referenziert werden. Um eine Isolation des Verkehrs gemäß dem Konzept zu gewährleisten ist eine Portgruppe und eine bridge-domain für jede virtuelle Maschine nötig.

```
bridge-domain lxbscDA-11
  segment id 11111
  group 239.0.0.42
!
port-profile type vethernet VXLAN-lxbscDA-11
  vmware port-group
  switchport mode access
  switchport access bridge-domain lxbscDA-11
  no shutdown
  state enabled
```

Nun muss noch das VXLAN-Gateway konfiguriert werden. Hierbei sind die folgenden Arbeitsschritte nötig.

- Definition eines neuen vxlan-Interfaces mit der VXLAN-ID, Gruppe und der Schnittstelle im VXLAN-

Transportnetz

  - `ip link add vx-lxbscDA-11 type vxlan id 11111 group 239.0.0.42 dev eth1`

- Setzen der MTU
  - `ip link set vx-lxbscDA-11 mtu 1500`
- Aktivieren der Schnittstelle
  - `ip link set vx-lxbscDA-11 up`
- Aktivieren von IPv4 Proxy ARP
  - `sysctl -w net.ipv4.conf.vx-lxbscDA-11.proxy_arp=1`
- Hinzufügen der statischen IPv6 Link-Local-Adresse für das Standard-Gateway
  - `ip addr add FE80::1/64 dev vx-lxbscDA-11`
- Routen aller benötigten IPv4- und IPv6-Adressen
  - `ip route add 10.156.252.227/32 dev vx-lxbscDA-11`
  - `ip route add 2001:4CA0:0:FF11::/64 dev vx-lxbscDA-11`

Im Rahmen des Tests wurde festgestellt, dass die Konfiguration der physischen VXLAN-Schnittstelle durch den Kernel nicht übernommen wurde und Multicast-Kommunikation daher fehlschlug [Schm 12]. Ein entsprechender Bugfix wurde in der Zwischenzeit von Yan Burman entwickelt [Burm 12] und wird in den Kernel 3.8 einfließen. Ein Workaround bestand in der Definition einer Hostroute für die VXLAN-Gruppe.

Zur Vereinfachung der Konfiguration wurde ein Bash-Script geschrieben, welches die Konfiguration zeilenweise aus einer Textdatei ausliest und die Schnittstellen entsprechend konfiguriert.

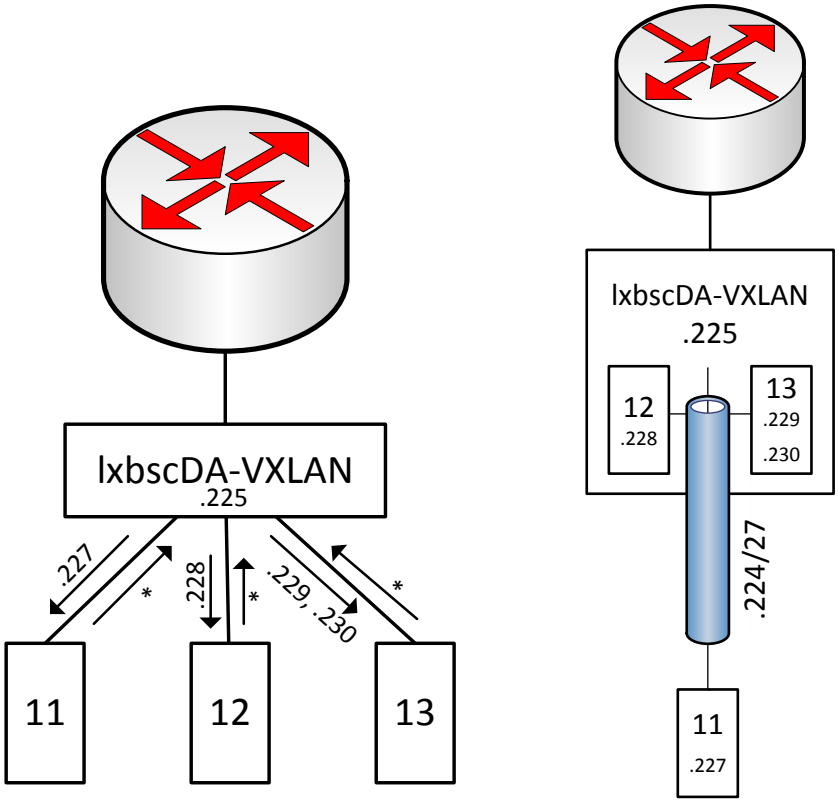
```
#!/bin/sh
cat vxlan-config | while read NAME ID MCAST ROUTES; do
        INT=vx-$NAME
        ip link add $INT type vxlan id $ID group $MCAST dev eth0
        ip link set $INT mtu 1500
        ip link set $INT up
        sysctl -w net.ipv4.conf.$INT.proxy_arp=1
        ip addr add fe80::1/64 dev $INT
        for ROUTE in $ROUTES; do
                ip route add $ROUTE dev $INT
        done
done
```

Die Konfigurationsdatei der Testumgebung sieht daher folgendermaßen aus:

```
lxbscDA-11 11111 239.0.0.42 10.156.252.227 2001:4ca0:0:ff11::/64
lxbscDA-12 22222 239.0.0.42 10.156.252.228 2001:4ca0:0:ff12::/64
lxbscDA-13 33333 239.0.0.42 10.156.252.229 10.156.252.230 2001:4ca0:0:ff13::/64
```

In einem realen Betrieb sollte diese Konfiguration selbstverständlich aus einer Datenbank generiert werden und den vollen Lebenszyklus einer virtuellen Maschine inklusive Löschung abbilden können.

Wird nun auf dem Gast die Netzwerkschnittstelle aktiviert und eine der zugewiesen Adressen konfiguriert, ist eine Verbindung möglich. Die Konfiguration auf der Seite des Gastsystems unterscheidet sich hierbei nicht von der seit Jahrzehnten bekannten Konfiguration einer IP-Adresse, Netzmaske und Default-Gateway. Der tatsächliche Wert der Netzmaske ist weitgehend irrelevant, da alle von der virtuellen Maschine ausgehenden ARP-Anfragen gleichartig vom VXLAN-Gateway beantwortet werden. Auch der Wert des Default-Gateways ist nicht von technischer Bedeutung, wobei diese IPv4-Adresse auf den meisten Betriebssystemen im gleichen (durch IPv4-Adresse und Netzmaske definierten) Subnetz liegen muss. Neben der statischen Konfiguration wäre auch die Adresszuweisung über DHCPv4 möglich, wenn auf dem VXLAN-Gateway ein entsprechender DHCP-Server oder ein DHCP-Relay agieren würde. Die Abbildung 6.2 zeigt den Unterschied. Aus der Sicht



(a) real (b) virtuell

Abbildung 6.2: Netztopologie im Kundennetz

der virtuellen Maschine befinden sich die anderen IP-Adressen im konfigurierten IPv4-Subnetz im gleichen Netzsegment, an der MAC-Adresse als Alias auf dem VXLAN-Gateway erkennbar. In der Realität befinden sich jedoch nur einzelne Gastsysteme in einem abgetrennten Subnetz, die gesamte Kommunikation läuft über das VXLAN-Gateway.

Als Beispiel wird nun lxbscDA-11 konfiguriert, wie wenn er sich mit anderen Rechnern zusammen in einem geteilten Subnetz 10.156.252.224/27 mit dem Gateway 10.156.252.225 befinden würde. Im Syntax der Debian-Konfigurationsdatei */etc/network/interfaces* sieht diese Konfiguration folgendermaßen aus:

```
iface eth0 inet static
        address 10.156.252.227/27
        gateway 10.156.252.225
```

Gemäß dieser Kombination von Adresse und Netzmaske müssten nun die Adressen 10.156.252.228 (lxbscDA-12) und 10.156.252.229 (lxbscDA-13) mit ARP im gleichen Netzsegment direkt aufzulösen sein. Eine ARP-Auflösung und ein ICMP Echo Request funktionieren auch problemlos. Beim näherer Betrachtung zeigt sich, dass in der ARP-Tabelle alle IPv4-Adressen die gleiche MAC-Adresse haben.

```
10.156.252.225 dev eth0 lladdr 1a:26:f2:ce:1b:e7 REACHABLE
10.156.252.226 dev eth0 lladdr 1a:26:f2:ce:1b:e7 REACHABLE
10.156.252.228 dev eth0 lladdr 1a:26:f2:ce:1b:e7 REACHABLE
10.156.252.229 dev eth0 lladdr 1a:26:f2:ce:1b:e7 REACHABLE
```

Dies liegt wie bereits eingangs beschrieben daran, dass das VXLAN-Gateway alle ARP-Anfragen mit seiner eigenen Adresse beantwortet. Während der Netzwerkstack also gemäß seinen Einstellungen den Verkehr an

lxbscDA-12 direkt schickt, wird der Verkehr tatsächlich an das VXLAN-Gateway geschickt und dort auf IPv4-Ebene geroutet. Ein Traceroute bestätigt diesen Verdacht.

```
root@lxbscDA-11:~# traceroute lxbscDA-12
traceroute to 10.156.252.228 (10.156.252.228), 30 hops max, 60 byte packets
 1  10.0.0.250 (10.0.0.250)  0.308 ms  0.329 ms  0.276 ms
 2  lxbscDA-12 (10.156.252.228)  0.841 ms  0.929 ms  0.983 ms
```

IPv6 wird ebenfalls auf die althergebrachte Art und Weise konfiguriert, in diesem Fall steht jedoch jeder virtuellen Maschine ein dediziertes /64-Netz zur Verfügung. In einem Debian-System könnte die Konfiguration wie folgt definiert sein:

```
iface eth0 inet6 static
        address 2001:4ca0:0:ff11::1/64
        gateway fe80::1
```

Im Gegensatz zu IPv4 ist die Definition des Gateways auf die spezielle Link-Local-Adresse hier wichtig, da ein Proxy-Neighbor-Discovery für wahlfreie Adressen aufgrund der Vielzahl an dafür nötigen *Solicited-Node*-Multicast-Gruppen ([RFC4291] 2.7.1) nicht standardkonform möglich ist Dafür stehen jeder virtuellen Maschine auch 2^{64} Adressen zur wahlfreien Verwendung bereit. Auch der IPv6-Verkehr wird über das VXLAN-Gateway geroutet, wie der folgende Traceroute zeigt.

```
root@lxbscDA-11:~# traceroute6 -q1 lxbscDA-12
traceroute to lxbscDA-12 (2001:4ca0:0:ff12::1), 30 hops max, 80 byte packets
 1  fe80::1826:f2ff:fece:1be7%eth0 (fe80::1826:f2ff:fece:1be7%eth0)  0.387 ms
 2  lxbscDA-12 (2001:4ca0:0:ff12::1)  0.998 ms
```

Durch das Routing des Verkehrs aller virtuellen Maschinen, auch untereinander, auf einem frei konfigurierbaren VXLAN-Gateway ergibt sich nun eine Sicherheitscharakteristik, die dem der seperaten VLANs sehr ähnlich ist. Spoofing-Angriffe auf Schicht 2 und 2.5 sind nun nicht mehr möglich, da sich kein angreifbares System im gleichen VLAN bzw. VXLAN-Segment befindet. Ein off-link Spoofing auf Schicht 3 kann im VXLAN-Gateway verhindert werden, da dieses anhand der eingehenden vxlan-Schnittstelle trivial die Adresse auf Gültigkeit prüfen kann. Dazu kann beispielsweise das ab Kernel 3.3 enthaltene *RPFILTER*-Modul von iptables verwendet werden. Es implementiert einen Filter, der analog zur uRPF-Prüfung die Quelladresse eingehender Pakete gegen die Routingtabelle validiert.

```
root@lxbscDA-VXLAN:~# iptables -t raw -L PREROUTING -n -v
Chain PREROUTING (policy ACCEPT 17721 packets, 5393K bytes)
 pkts bytes target     prot opt in     out     source               destination
    2   168 LOG        all  --  vx+    *       0.0.0.0/0            0.0.0.0/0
rpfilter invert LOG flags 0 level 4 prefix "SPOOFING "
    2   168 DROP       all  --  vx+    *       0.0.0.0/0            0.0.0.0/0
rpfilter invert
```

Für den Umgang mit auffälligen Paketen stehen alle Fähigkeiten des netfilter-Frameworks zur Verfügung. Hierzu gehört neben dem Protokollieren auch Umleitungsfunktionen, um derartige Pakete transparent auf ein IDS oder einen Honeypot umzuleiten.

6.2 Migration

Zum praktischen Einsatz der vorgeschlagenen Lösung im Produktivbetrieb müssen zunächst organisatorische Fragen geklärt werden.

6.2.1 Organisatorische Aspekte

Im Gegensatz zu den bisher ausgeübten Trennung zwischen dem Netzbetrieb, der nur in großen und unregelmäßigen Abständen durch manuelle Konfiguration ein neues VLAN mit einem ganzen IP-Adressbereich

für mehrere hundert virtuelle Server bereit stellt, und der Serveradministration, welche die einzelnen virtuellen Maschinen auf der Infrastruktur bereitstellt und konfiguriert, bedarf diese Lösung engerer Kooperation zwischen den Abteilungen. Wie auch schon bei früheren Teilaspekten genannt wäre eine gemeinsame Administration dieses Systems mit allen beteiligten Abteilungen empfehlenswert.

Dies liegt zum einen in der Rolle der VXLAN-Gateways, welche als IP-Router und Sicherheitsgateway die Schnittstelle für den Netzbetrieb gegenüber dem Kunden darstellt. Zum anderen jedoch für jede neue virtuelle Maschine ein VXLAN-Interface mit den entsprechenden Routingeinträgen angelegt werden. Dies ist nur mit einer vollständigen Automatisierung und dem Bezug der benötigten Informationen (VXLAN-ID, IP-Adressen, eventuell Filterregeln) aus der gemeinsamen Konfigurationsdatenbank (CMDB) möglich. Eine ausreichende Definition dieser Schnittstellen muss vor dem Aufbau der Infrastruktur vorgenommen werden.

Auch im Bereich der Virtualisierungsinfrastruktur sind Automatisierungen nötig, um das Anlegen einer neuen Portgruppe (mit dedizierter VXLAN-ID) für jeden neuen virtuellen Server zu automatisieren.

6.2.2 Auswirkungen auf den Dienstleistungskatalog

Wie bereits in Kapitel 5 festgestellt müssen einige Einschränkungen für das Produkt „Virtueller Server“ festgelegt werden. Diese resultieren zum Teil aus echten Einschränkungen und Anforderungen der verwendeten Lösung, zum Teil aus der Formalisierung von Sicherheitsmechanismen und zum Teil aus der Vermeidung von Sonderfällen und damit dem erhöhten Administrationsaufwand. Das Ziel ist, mehr als 90% der Kunden durch ein Standardprodukt abdecken zu können.

Für das Standardprodukt sollten die folgenden Einschränkungen im Dienstleistungskatalog oder den Vertragsbedingungen hinterlegt werden:

Nutzung einer Netzwerkkarte mit einer MAC-Adresse aus dem VMware-Bereich
: Zur Verhinderung von MAC-Spoofing muss der Versand von Frames mit einer fremden MAC-Adresse unterbunden werden. Dies ist bei VMware nur möglich, wenn die MAC-Adresse durch den vCenter Server aus dem für VMware vorgesehenen Bereich vergeben wurde.

Nutzung einer IPv4-Adresse, statisch konfiguriert
: Es kann prinzipiell eine beliebige Anzahl von IPv4-Adressen auf dem VXLAN-Segment konfiguriert werden. Dies sollte jedoch ein kostenpflichtiges Zusatzprodukt darstellen.

Nutzung mehrerer IPv6-Adressen aus einem Subnetz (/64), statisch/SLAAC konfiguriert
: Jedem VXLAN-Segment steht ein /64-IPv6-Subnetz zur Verfügung, aus dem beliebig viele Adressen verwendet werden können. Eine Beschränkung nur auf wenige IPv6-Adressen ist technisch nicht sinnvoll möglich, da dies Seiteneffekte bei der Nutzung von Privacy Extensions hat

Kommunikation nur mit IPv4 und IPv6
: Das VXLAN-Gateway übernimmt die Funktionalität eines IPv4- und IPv6-Routers auf der Schicht 3, Frames anderer Protokolle können nicht geroutet und daher nicht verwendet werden.

Kein Multicast- oder Broadcastverkehr zwischen VMs
: Broadcast-Verkehr wird nicht über Netzsegment-Grenzen (Router) hinweg transportiert, eine Kommunikation mit anderen VMs über Broadcast ist daher nicht möglich. Geroutetes Multicast könnte technisch mit dem PIM-Protokoll über das VXLAN-Gateway transportiert werden, jedoch sind viele Multicast-verwendenden Protokolle (beispielsweise zur Clusterkommunikation) nicht in der Lage, geroutetes Multicast zu verwenden.

Keine netzbasierte Firewall

Obwohl diese Einschränkungen erst nach der Migration auf die VXLAN-basierte Lösung aktiv werden, müssen sie zeitnah dem Kunden kommuniziert werden um Vertragsbestandteil zu werden. Trotz der Einschränkungen sollte das Produkt für die meisten Einsatzzwecke ausreichend sein, nur Hochverfügbarkeitslösungen, die IP-Adressen zwischen den Serverknoten verschieben oder Multicast zur Kommunikation einsetzen, können mit diesem Basisprodukt nicht betrieben werden.

Für eventuelle Sonderwünsche können Zusatzprodukte definiert werden. Diese sollten, da sie limitierte Ressourcen verbrauchen oder den Administrationsaufwand erhöhen, mit periodisch wiederkehrenden Kosten verbunden sein, um eine sinnlose Nutzung zu vermeiden.

Mögliche Zusatzprodukte

Die folgende, nicht endgültige Liste von Zusatzprodukten ist in diesem Zusammenhang sinnvoll und sollte daher mit im Dienstleistungskatalog verankert werden.

mehrere IPv4-Adressen, geroutetes IPv4/IPv6-Subnetz
: Zusätzliche IPv4-Adressen sowie extra IPv4- und IPv6-Subnetze können mit der geplanten Technologie einer einzelnen VM zugewiesen werden. Diese Subnetze sollten aus einem gemeinsamen Pool kommen, um die Routingtabelle im Kernnetz durch Aggregation entlasten zu können.

Dediziertes VXLAN-Segment
: Zur direkten Kommunikation, auch über Broad- und Multicast, zwischen eigenen virtuellen Maschinen kann ein dediziertes VXLAN-Segment bereitgestellt werden. Da die im MWN verwendete Routerplattform nativ kein VXLAN unterstützt ist eine Routeranbindung nicht möglich. Diese VXLAN-Segmente können beispielsweise zur internen Kommunikation zwischen Knoten eines Hochverfügbarkeitsclusters oder zur Anbindung eigener VMs an eine unter Eigenregie betriebene virtuelle Subnetzfirewall benutzt werden.

Dediziertes VLAN-Segment
: Analog zum VXLAN-Segment ermöglicht diese Option die Kommunikation zwischen den eigenen Systemen, hier ist jedoch optional auch einen im LRZ gehosteten physischen Server oder eine LRZ-Firewall möglich. Hier kann ebenfalls ein eigener VLAN-Anschluss an den Hausrouter angeboten werden, wobei Größe des angebotenen Subnetzes einer Kostenstaffelung unterworfen sein sollte.

6.2.3 Migration

Eine Migration auf die vorgeschlagene Lösung kann nur durchgeführt werden, wenn auf der Seite der Virtualisierungsinfrastruktur ein VXLAN-fähiger virtueller Switch vorhanden ist. Dies erfordert entweder ein turnusmäßiges Softwareupgrade auf VMware ESXi 5.1, oder die Installation des Cisco Nexus 1000V. Aufgrund der Komplexität des Nexus 1000V wird bei der Skizzierung des Migrationspfades von einer Nutzung von ESXi 5.1 ausgegangen.

Die VXLAN-Router können aufgebaut werden, ohne einen Einfluss auf die bestehende Infrastruktur zu nehmen, da sie nicht in den physischen Weg zur Virtualisierungsinfrastruktur eingebracht werden müssen. Es ist ausreichend, dass sie ein gemeinsames VLAN mit den ESXi-Hosts besitzen, welches über die bestehende Infrastruktur geswitched wird. Das Management-VLAN wäre ein Kandidat, aus Sicherheitsgründen und für die Isolierung des Verkehrs ist es jedoch empfehlenswert ein neues VLAN für diese Zwecke zu verwenden. Ein Einsatz von geroutetem Multicast ist beim Betrieb innerhalb des LRZ-Rechnergebäudes noch nicht vorteilhaft und würde nur die Komplexität erhöhen.

Je nach der benötigten Verfügbarkeit müssen Redundanz und Skalierbarkeit geplant, implementiert und getestet werden. Aufgrund der zwangsläufig zu Beginn auftretenden Ungereimtheiten ist es empfehlenswert, einen Testbetrieb mit wenigen virtuellen Maschinen zu beginnen und dabei das geplante System durch manuell herbeigeführte Ausfälle und Störungen zu testen.

Bei der Nutzung eines neuen, für die Nutzung des VXLAN-Routers dedizierten IP-Adressbereich können neue virtuelle Maschinen in diesem Adressbereich sofort eingesetzt werden, ohne weitere Umstellungsarbeiten am Netz vorzunehmen. Sofern jedoch ein bestehendes Netz (zum Beispiel die bisher definierten Kundennetze in Kapitel 2.1.2) verwendet werden, so muss zunächst das gesamte VLAN auf das VXLAN-Gateway umgestellt werden. Alle in diesem VLAN befindlichen virtuellen Server müssen dabei zeitgleich auf die neue, VXLAN-basierte Portgruppe umgestellt werden. Dies bringt im Idealfall eine Ausfallzeit von wenigen Sekunden mit sich, die Benutzer müssen keine Änderungen vornehmen. Diese virtuellen Maschinen können nun, wie auch in einem herkömmlichen VLAN, ungehindert miteinander kommunizieren. Nach dieser Umstellung ist es

möglich, die virtuellen Maschinen einzeln und nacheinander jeweils in ein dediziertes VXLAN zu migrieren, um die vollen Sicherheitsvorteile dieser Struktur zu nutzen. Auch bei dieser finalen Änderung ist keine Interaktion oder Absprache mit den Benutzern erforderlich, da auch hierbei im Idealfall die Netzverbindung nur für wenige Sekunden unterbrochen wird. Aus Kundensicht befindet sich sein System auch nach Abschluss der Migration noch mit hunderten fremden Servern im gleichen Subnetz.

6.3 Skalierbarkeit und Verfügbarkeit

Der reine Routingdurchsatz von Linux-basierten Routern ist durch mehrere Optimierungszyklen im Kernel und durch den Einsatz von Techniken zur Multicore-Verarbeitung wie *Receive Packet Steering* (RPS) und mehreren Queues bereits seit mehreren Jahren auf deutlich mehr als 10Gbit/s angestiegen [Brou 09]. Auch unter dem Einsatz von Filtern sind auf Standardkomponenten schon Datenraten jenseits der 10Gbit/s erreicht worden, wobei hier das Optimierungspotential bei der Anordnung der Regeln zum Tragen kommt. Nicht zuletzt basiert bereits die Secomat-Infrastruktur, die am LRZ den Verkehr von privaten IPv4-Adressen ins Internet filtert, auf Standardkomponenten und ist täglich mit Datenraten weit über einem Gigabit pro Sekunde belastet.

Messungen des japanischen Forschers Naoto Matsumoto mit dem auch im Testaufbau verwendeten Release Candidate zufolge ist der zusätzliche Overhead durch die Nutzung von VXLAN auf etwa 10% zu beziffern [Mats 12], bei einem absoluten Durchsatz von über 20 Gbit/s. Diese Zahlen wurden zwar unter der Nutzung von Infiniband-Hardware gemessen, eine signifikante Erhöhung des Overheads ist mit 10-Gigabit-Ethernet jedoch nicht zu erwarten.

Die Anzahl der Routingeinträge und die Anzahl der (virtuellen) Netzwerkschnittstellen hat in aktuellen Kernelversionen auch nur noch marginalen Einfluss auf die Geschwindigkeit. Bereits 2006 wurden von den Linux-Entwicklern Änderungen in den Kernel eingebracht, welche die Nutzung mehrerer tausend Schnittstellen ermöglichen [Hemm 06]. Beim IPv6-Tunnelbroker SixXS[1] sind Tunnelserver mit über 4000 aktiven Tunneln in Betrieb [2].

Sollte jedoch wider erwarten dennoch ein Geschwindigkeitsproblem auftreten, so kann das Gesamtsystem durch die Hinzunahme weiterer Gateway-Systeme linear skaliert werden, da die Gateway-Funktionalität für die einzelnen virtuellen Maschinen für den Nutzer transparent zwischen den VXLAN-Routern verschoben werden kann. Auf dem gleichen Weg ist auch eine automatische Redundanz möglich, durch die der Ausfall eines Systems binnen weniger Sekunden abgefangen werden kann.

6.4 Sonstige Empfehlungen

Die am LRZ bereits praktizierte Trennung zwischen LRZ-eigenen und im Kundenauftrag betriebenen virtuellen Servern auf unterschiedlichen ESXi-Clustern und in unterschiedlichen VLANs ist dabei sinnvoll und muss unbedingt beibehalten werden. Kunden sind untereinander damit zwar durch Side-Channel-Angriffe, wie sie in Kapitel 3.5 beschrieben werden, angreifbar, sind aber von LRZ-Infrastruktur rudimentär getrennt. Auch eine Abtrennung von Testsystemen wird hier bereits durchgeführt.

Leider besteht diese Trennung nur auf der Ebene der ESXi-Hosts. Bei einer erfolgreichen Kompromittierung des Hypervisors, wie er bereits demonstriert wurde [VMDK-Vulnerability], kann dieser auf alle *Datastores* (Speicherbereiche) zugreifen und dort lesende oder sogar schreibende Veränderungen vornehmen. Auch besteht auf der Ebene der Managementnetze keine Trennung zwischen dem LRZ-eigenen VMware-Cluster und dem Cluster für Kundenmaschinen. Diese Trennung sollte zumindest auf Storage-Ebene umbedingt eingeführt werden. Ein Zwischenschritt zur Verringerung der Angriffsfläche wäre die Beschränkung des NFS-Zugriffs auf den Speichersystemen auf die dem jeweiligen Cluster zugeordneten Hosts.

Eine weitergehende Trennung in verschiedene Kundengruppen oder gar noch kundenspezifische Sicherheitsklassen ist jedoch nicht unbedingt sinnvoll, da die Art eines Angriffs vorher nicht bekannt ist und je nach

[1] `http://www.sixxs.net`
[2] `http://www.sixxs.net/pops/netcologne/`

Angriffsvektor verschiedene Prioritäten gelten. So ist eine VM, welche allein auf einem Host läuft (und damit die Skalierungsvorteile der Virtualisierung zunichte macht) zwar gegen Side-Channel-Angriffe durch andere VMs auf dem gleichen Host gefeit, gleichzeitig aber durch den weitgehenden Ausschluss des Schedulers anfälliger gegen Timing-Analysen in Netzprotokollen.

Für den Zugriff von außen auf die ESXi-Hosts durch das Managementnetz, der für die Verbindung mit dem vSphere-Client nötig ist (vergleiche Kapitel 2.1.2 und 3.5), existiert derzeit noch keine technische Lösung. Ein Proxyserver für das proprietäre Protokoll zwischen Client und ESXi-Host würde das Problem lösen, steht jedoch zur Zeit noch nicht zur Verfügung. Hier muss durch die Firma VMware eine entsprechende Software entwickelt und zur Verfügung gestellt werden.

Bei der Bereitstellung der virtuellen Maschinen sind nur wenige Änderungen sinnvoll. So wäre es unter Umständen sinnvoll, auf neuen Systemen die Passwortauthentifizierung über SSH generell zu deaktivieren und vom Kunden bei der Einrichtung seinen öffentlichen SSH-Schlüssel zu erfragen. Diese Einschränkung kann allerdings vermutlich nach der Ersteinrichtung nicht mehr erzwungen werden, da es viele Gründe geben kann eine Passwortauthentifizierung einzusetzen.

7 Zusammenfassung und Ausblick

In diesem Kapitel werden die in dieser Arbeit evaluierten Probleme und Lösungen noch einmal reflektiert und zusammengefasst. Zuletzt folgt ein Ausblick auf weitere Forschungsfelder und Verbesserungsmöglichkeiten, die im Rahmen der vorliegenden Arbeit nicht oder nur am Rande besprochen wurden.

7.1 Zusammenfassung

Durch die zunehmende Verbreitung von Virtualisierungsplattformen steigen die Anforderungen an die Sicherheit der zugrundeliegenden Infrastruktur an. Dies ist insbesondere der Fall, wenn diese Dienstleistungen nicht nur zur Konsolidierung des eigenen IT-Betriebs verwendet, sondern auch an externe Kunden verkauft werden. Durch die auf viele Gruppen verteilte Administration sind zentrale Sicherheitsrichtlinien, die eine Kompromittierung verhindern sollen, nicht mehr so einfach durchsetzbar. Gleichzeitig steigt mit der Verfügbarkeit von günstigen virtuellen Servern die Anzahl der angreifbaren Systeme an, da nun auch für kleine Arbeitsgruppen die schnelle Einrichtung eines eigenen Servers lohnend erscheint. Diese haben jedoch oft wenig Erfahrung im sicheren Betrieb und verlieren gerade im universitären Umfeld durch die Stellenpolitik oft in kurzen Zeitabständen große Teile ihres Wissens. Um so wichtiger ist es, dass gerade in virtualisierten Umgebungen nicht nur der Schutz vor Angriffen (*Protection*), sondern insbesondere das Verhindern eines Übergriffs eines kompromittierten Systems auf andere Mandanten (*Containment*) eine wichtige Rolle spielen.

Im Groben lassen sich Angriffe in diesem Kontext auf drei Klassen aufteilen:

1. **Angriffe auf das (virtuelle Gast-)System** selbst, welche üblicherweise durch einen Sicherheitslücke oder schlechte Konfiguration von außen erfolgen. Diese Angriffe sind heutzutage gang und gäbe und unterscheiden nicht zwischen einem physischen Server, einem virtuellen Server auf der eigenen Infrastruktur oder einem virtuellen Server auf einer fremden Infrastruktur. Das Ziel ist im Allgemeinen, geschützte Daten dieses Servers zu stehlen oder die Ressourcen des angegriffenen Servers für eigene Zwecke zu nutzen. Als Anbieter eines Infrastruktur-Dienstes, wie es das LRZ mit der Bereitstellung virtueller Server tut, sind diese Angriffe nicht zu verhindern. Es besteht höchstens die Möglichkeit, dem Betreiber des Gastsystems (dem Kunden) Hilfsmittel zur Hand zu geben, um sein System möglichst sicher zu konfigurieren.

2. **Angriffe auf die Virtualisierungsplattform** basieren auf einem spezifischen Angriff auf den Hypervisor und dessen Infrastruktur, der die Trennung zwischen den verschiedenen Gastsystemen durchsetzt. Diese Angriffe sind im Allgemeinen hoch plattform- und versionsabhängig und versuchen, über Fehler in der Implementierung des Hypervisors aus dessen Sicherheitskonzept auszubrechen. Diese Lücken sind oft nur kurzzeitig ausnutzbar bis das Problem behoben wird. Das Ziel der meisten Angriffe dieser Art ist, schutzwürdige Daten von *anderen* Systemen in der gleichen Infrastruktur auszuspähen. Diese Lücken sind daher virtualisierungsspezifisch.

3. **Angriffe auf die Netzinfrastruktur** sind eigentlich ein Problem, welches schon seit Jahrzehnten auch mit physischen Servern bekannt ist. Sie sind oft Vorboten eines Angriffs auf einen Server (zum Beispiel durch *Man-in-the-Middle*-Angriffe zum Ausspionieren und Fälschen von Daten), werden aber auch nach einer erfolgreichen Kompromittierung zum Verbergen benutzt. In rein physischen Netzen existieren schon seit einigen Jahren Produkte, die adäquate Sicherheitsmechanismen gegen diese Art von Angriffen mitbringen. Sie sind jedoch in den virtuellen Infrastrukturen wieder ein größeres Thema geworden, da zum einen die Masse der Systeme den Einsatz bekannter Lösungen immer mehr erschwert, und zum anderen Komponenten von Herstellern ohne Erfahrung in der Netzabsicherung eingesetzt werden.

Während also die Angriffe der ersten Klasse nicht wirksam verhindert werden können (insbesondere wenn ein Kunde selbst für seine Absicherung zuständig ist) und Angriffe der zweiten Klasse meistens kein strukturelles Problem sondern eine behebbare Sicherheitslücke sind, existieren für die meisten Angriffe der dritten Klasse bereits seit Jahrzehnten Methoden zur Absicherung. Sie müssen jedoch in der gesamten Infrastruktur (und damit oft Hersteller- und Gruppenübergreifend) eingeführt werden und bringen oft andere Einschränkungen mit sich. Es sieht so aus, als würden sich die Hersteller dieser Systeme immer noch mehr auf die werbewirksamen Geschwindigkeitsversprechen konzentrieren, als die grundlegenden Sicherheitsprobleme eines derartigen „Massenhostings“ in den Griff zu bekommen. Daher lag der Fokus in der Ausarbeitung dieser Arbeit auf den netzbasierten Angriffen. Hier kann mit der Absicherung gegen nur wenige, bereits bekannte Angriffe eine hohe Schutzwirkung erzielt werden.

Im Rahmen dieser Arbeit wurden dabei mehrere Technologien und Produkte evaluiert, die in der Lage sind, jeweils einige Sicherheitsprobleme einzudämmen. Es existiert jedoch derzeit auf dem Markt keine Lösung, die alle Probleme lösen kann, skalierbar ist und innerhalb des bestehenden Rechenzentrumsnetzes, welches nicht ausschließlich für die Virtualisierung von Kundenservern neu aufgebaut wurde, eingesetzt werden kann. Es wurde daher eine neue umfassende Lösung entwickelt, die einen hohen Schutz bietet. Die Grundfunktionen (wie VLAN/VXLAN-Trennung, Proxy-ARP, Filter, ...) sind für sich genommen bekannt, werden jedoch hier in einer neuen, zum Teil nicht vorher existierenden Zusammenstellung eingesetzt. Man kann jedoch davon ausgehen, dass die ersten kommerziellen Implementierungen, welche ein ähnliches Konzept zur Trennung basierend auf VXLAN-Tunneln verfolgen, im Laufe des Jahres 2013 zur Verfügung stehen. Eine auf Open-Source-Komponenten basierende Implementierung, die im Rahmen dieser Arbeit entstand, konnte die meisten Sicherheitsprobleme der dritten Angriffsklasse lösen und ist, auch das ist ein großer Vorteil gegenüber den herkömmlichen Lösungen, trivial auf die bestehende Infrastruktur des LRZ anzuwenden. Sie stellt keinen erstzunehmenden Zusatzaufwand dar, erfordert nicht den aufwändigen Austausch bestehender Komponenten und kann nachträglich, in einzelnen Schritten, ausgerollt werden.

7.2 Ausblick

Basierend auf dem hochskalierbaren System zur Netztrennung können den Kunden in Zukunft auch weitere Dienste angeboten werden, ohne die Sicherheit anderer Mandanten zu kompromittieren. Mögliche Zusatzdienste sind die Anbindung an den Server-Load-Balancer (SLB), VPN-Tunnel oder die Nutzung des neu im MWN eingeführten MPLS zur abgetrennten Übertragung zwischen Kundennetz und virtuellem Server (L3VPN).

Bevor das vorgestellte System in den Produktivbetrieb überführt werden kann ist jedoch noch etwas Entwicklungsarbeit zu leisten. Das wichtigste Projekt ist die Automatisierung, bei der die Konfiguration des VXLAN-Gateways vollständig aus den Konfigurationsdaten (zum Beispiel der CMDB) der konfigurierten virtuellen Maschinen erstellt wird. Zwei weitere Anforderungen, die lose miteinander gekoppelt sind, sind zum einen eine Hochverfügbarkeit durch die Bereitstellung eines Failover-Partners, zum anderen die horizontale Skalierbarkeit durch Hinzunahme weiterer Gateway-Systeme. Diese beide Anforderungen können beispielsweise mit einem Heartbeat-Cluster ähnlich zum bereits eingesetzten *Secomat*-System erfüllt werden.

Ein weiterer Punkt, der in einer späteren Arbeit aufgegriffen werden könnte, ist die Konfiguration einer zentral provisionierten Firewall (siehe Kapitel 4.3.4). Sofern die hier entwickelte Lösung dauerhaft eingesetzt werden soll können diese Regeln direkt als Netfilter-Regeln auf dem VXLAN-Gateway implementiert werden. Wenn diese Konfigurationsschnittstelle mandantenfähig ausgeführt wird, können dedizierte Firewalls im Virtualisierungsumfeld weitgehend ersetzt werden.

Wie bereits in Kapitel 6 ausgeführt können versuchte Angriffe trivial durch die Sicherheitsmechanismen nicht nur verhindert, sondern auch mitprotokolliert werden. Diese Meldungen sind für sich genommen jedoch wenig hilfreich, sondern müssen in das LRZ-weite *Security Information and Event Management*-System (SIEM) eingespeist werden.

Die Diskussion der Angriffe in Kapitel 3 hat gezeigt, dass die meisten Sicherheitslücken durch das veraltete Konzept eines großen Netzsegments (VLAN), in dem klassisches Layer2-Switching erfolgt, ermöglicht werden. Dieses Problem wird durch die vorgeschlagene Lösung entschärft, in der diese Netzsegmente sehr klein gehalten und damit sicher gemacht werden. Die Virtualisierungsplattform implementiert jedoch weiter-

hin einen Switch. In Zukunft könnte der virtuelle Switch im Hypervisor durch einen virtuellen Router ersetzt werden, welcher die routende Funktionalität des hier vorgestellten VXLAN-Gateways zwischen den virtuellen Maschinen übernimmt. Der Hypervisor kann dann mit einem klassischen dynamischen Routingprotokoll wie RIP, OSPF oder BGP die Adressen der gerade auf ihm laufenden virtuellen Maschinen bekannt geben, so dass auch Migrationen zwischen verschiedenen Hosts in Sekundenbruchteilen abgeschlossen sind. Allerdings müsste in diesem Fall der Hypervisor auch alle Sicherheitsfunktionen eines Routers implementieren. Hierzu gehören dann insbesondere auch Spoofing-Filter und Firewall-Regeln.

Literaturverzeichnis

[AKO+ 05] ALTUNBASAK, HAYRIYE, SVEN KRASSER, HENRYL. OWEN, JOCHEN GRIMMINGER, HANS-PETER HUTH und JOACHIM SOKOL: Securing Layer 2 in Local Area Networks. In: LORENZ, PASCAL und PETRE DINI (Herausgeber): Networking - ICN 2005, Band 3421 der Reihe Lecture Notes in Computer Science, Seiten 699–706. Springer Berlin Heidelberg, 2005, http://dx.doi.org/10.1007/978-3-540-31957-3_79 .

[Atla 12] ATLASIS, ANTONIOS: Attacking IPv6 Implementation Using Fragmentation. In: BlackHat Europe 2012, Amsterdam, 2012. , http://media.blackhat.com/bh-eu-12/Atlasis/bh-eu-12-Atlasis-Attacking_IPv6-WP.pdf .

[BCP 38] FERGUSON, P. und D. SENIE: Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing, 2000, http://tools.ietf.org/html/bcp38 .

[Bern 05] BERNSTEIN, DANIEL J.: Cache-timing attacks on AES. Technischer Bericht, 2005.

[Blue 09] BLUEPOKE: Ethernetpaket.svg, February 2009, http://de.wikipedia.org/w/index.php?title=Datei:Ethernetpaket.svg .

[Brou 09] BROUER, JESPER DANGAARD: 10Gbit/s Bi-Directional Routing on standard hardware running Linux, September 2009, http://vger.kernel.org/netconf2009_slides/LinuxCon2009_JesperDangaardBrouer_final.pdf .

[Burm 12] BURMAN, YAN: net/vxlan: Use the underlying device index when joining/leaving multicast groups, December 2012, http://patchwork.ozlabs.org/patch/207664/ .

[Cisco-DHCPu] SYSTEMS, CISCO: Cisco IOS DHCP Relay Agent Support for Unnumbered Interfaces, http://www.cisco.com/en/US/docs/ios/12_1t/12_1t2/feature/guide/dt_dhcpu.html .

[Cisco-OTV-VPLS] SYSTEMS, CISCO: Technology Comparison: Cisco Overlay Transport Virtualization and Virtual Private LAN Service as Enablers of LAN Extensions, 2012, http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-574984.html .

[CiscoVSS] SYSTEMS, CISCO: Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440, 2010, http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/white_paper_c11_429338.pdf .

[DKL+ 00] DHEM, JEAN-FRANÇOIS, FRANÇOIS KOEUNE, PHILIPPE-ALEXANDRE LEROUX, PATRICK MESTRÉ, JEAN-JACQUES QUISQUATER und JEAN-LOUIS WILLEMS: A Practical Implementation of the Timing Attack. In: QUISQUATER, JEAN-JACQUES und BRUCE SCHNEIER (Herausgeber): Smart Card Research and Applications, Band 1820 der Reihe Lecture Notes in Computer Science, Seiten 167–182. Springer Berlin Heidelberg, 2000, http://dx.doi.org/10.1007/10721064_15 .

[DNS Amplification] VAUGHN, RANDAL und GADI EVRON: DNS Amplification Attacks, March 2006, http://www.isotf.org/news/DNS-Amplification-Attacks.pdf .

[Gold 72] GOLDBERG, ROBERT P.: Architectural Principles for Virtual Computer Systems. Doktorarbeit, Harvard University, Cambridge, MA, 1972, http://www.dtic.mil/cgi-bin/GetTRDoc?AD=AD772809&Location=U2&doc=GetTRDoc.pdf .

[Hemm 06] HEMMINGER, STEPHEN: lkml: Re: Thousands of interfaces, October 2006, http://marc.info/?l=linux-kernel&m=116231905124278&w=4 .

[HP 10] HEWLETT-PACKARD: HP Virtual Connect Ethernet Cookbook: Single and Multi Enclosure Domain (Stacked) Scenarios, 2010, http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01990371/c01990371.pdf .

[HP 11] HEWLETT-PACKARD: ISS Technology Focus, Volume 10, Number 2, 2011, http://h20000.www2.hp.com/bc/docs/support/SupportManual/c02877995/c02877995.pdf .

[HP 12a] HEWLETT-PACKARD: HP BladeSystem c7000 Enclosure, Mai 2012, http://h18004.www1.hp.com/products/quickspecs/12810_div/12810_div.pdf .

[HP 12b] HEWLETT-PACKARD: HP Switch Software Access Security Guide, June 2012, http://bizsupport1.austin.hp.com/bc/docs/support/SupportManual/c03389646/c03389646.pdf .

[HP 12c] HEWLETT-PACKARD: HP Virtual Connect Flex-10 10Gb Ethernet Module for c-Class BladeSystem, 2012, http://h18004.www1.hp.com/products/quickspecs/13127_div/13127_div.html .

[IBM 11] IBM: IBM Distributed Virtual Switch 5000V, October 2011, http://public.dhe.ibm.com/common/ssi/ecm/en/qcd03011usen/QCD03011USEN.PDF .

[IEEE-Ethertype] AUTHORITY, IEEE REGISTRATION: ETHERTYPE PUBLIC LISTING, 2012, http://standards.ieee.org/develop/regauth/ethertype/eth.txt .

[IEEE Std 802.1d-2004] IEEE Standard for Information Technology — Telecommunications and Information Exchange Between Systems — Local and Metropolitan Area Networks — Media Access Control (MAC) Bridges, 2004, http://standards.ieee.org/getieee802/download/802.1D-2004.pdf .

[IEEE Std 802.3-2008] IEEE Standard for Information Technology — Telecommunications and Information Exchange Between Systems — Local and Metropolitan Area Networks — Specific Requirements — Part 3: Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, 2008, http://standards.ieee.org/about/get/802/802.3.html .

[inkj 08] INDUCTIVELOAD und KJU: MAC-48 Address.svg, October 2008, http://en.wikipedia.org/wiki/File:MAC-48_Address.svg .

[IPv6-Frag-Bypass] HEUSE, MARC: Full Disclosure: Bypassing Cisco's ICMPv6 Router Advertisement Guard feature, May 2011, http://seclists.org/fulldisclosure/2011/May/446 .

[Koch 96] KOCHER, PAULC.: Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In: KOBLITZ, NEAL (Herausgeber): Advances in Cryptology — CRYPTO '96, Band 1109 der Reihe Lecture Notes in Computer Science, Seiten 104–113. Springer Berlin Heidelberg, 1996, http://dx.doi.org/10.1007/3-540-68697-5_9 .

[LRZ 12a] LRZ: Beschränkungen und Monitoring im Münchner Wissenschaftsnetz, Oktober 2012, http://www.lrz.de/services/netz/einschraenkung/ .

[LRZ 12b] LRZ: Kostenpflichtige Dienstleistungen des LRZ im Überblick, 2012, http://www.lrz.de/wir/regelwerk/dienstleistungskatalog.pdf .

[LRZ 12c] LRZ: LRZ: Unsere Servicepalette, 2012, http://www.lrz.de/services/ .

[LRZ 12d] LRZ: LRZ: Wir über uns, 2012, http://www.lrz.de/wir/ .

[LRZ 12e] LRZ: Sichere Institutsnetze mit Unterstützung des LRZ, 2012, http://www.lrz.de/services/security/sicherheitspakete-lrz/ .

[LRZ 12f] LRZ: Windows Server Update Service (WSUS) am LRZ für Microsoft Produkte, 2012, http://www.lrz.de/services/security/mwnsus/ .

[Mats 12] MATSUMOTO, NAOTO: A First Look At VXLAN over Infiniband Network On Linux 3.7-rc7 & iproute2, November 2012, http://www.slideshare.net/naotomatsumoto/a-first-look-at-xvlan-over-infiniband-network-on-linux-37rc7 .

[Mül 10] MÜLLER, ALEXANDER: Linux-basierte Personal Firewall für den Einsatz im LRZ, November 2010, http://www.nm.ifi.lmu.de/pub/Fopras/muel10/ .

[N1V-Datasheet] SYSTEMS, CISCO: Cisco Nexus 1000V Series Switches Data Sheet, September 2012, http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/data_sheet_c78-492971.html .

[OTV] H. GROVER, D. RAO, D. FARINACCI und V. MORENO: IETF Draft: Overlay Transport Virtualization, July 2011, http://tools.ietf.org/html/draft-hasmit-otv-03 .

[PtNe 98] PTACEK, THOMAS H. und TIMOTHY N. NEWSHAM: Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection. Technischer Bericht, Secure Networks, Inc., Suite 330, 1201 5th Street S.W, Calgary, Alberta, Canada, T2R-0Y6, 1998, http://citeseer.ist.psu.edu/ptacek98insertion.html .

[RFC2002] PERKINS, C.: RFC 2002 IP Mobility Support, October 1996, http://tools.ietf.org/html/rfc2002 .

[RFC2131] DROMS, R.: RFC 2131 Dynamic Host Configuration Protocol, March 1997, http://tools.ietf.org/html/rfc2131.html .

[RFC2136] P. VIXIE, S. THOMSON, Y. REKHTER und J. BOUND: RFC 2136 Dynamic Updates in the Domain Name System (DNS UPDATE), April 1997, http://tools.ietf.org/html/rfc2136 .

[RFC3031] E. ROSEN, A. VISWANATHAN, R. CALLON: RFC 3031 Multiprotocol Label Switching Architecture, January 2001, http://tools.ietf.org/html/rfc3031 .

[RFC3756] P. NIKANDER, J. KEMPF und E. NORDMARK: RFC 3756 IPv6 Neighbor Discovery (ND) Trust Models and Threats, May 2004, http://tools.ietf.org/html/rfc3756 .

[RFC3971] J. ARKKO, J. KEMPF, B. ZILL und P. NIKANDER: RFC 3971 SEcure Neighbor Discovery (SEND), March 2005, http://tools.ietf.org/html/rfc3971 .

[RFC4291] HINDEN, R. und S. DEERING: IP Version 6 Addressing Architecture, February 2006, http://tools.ietf.org/html/rfc4291 .

[RFC4861] T. NARTEN, E. NORDMARK, W. SIMPSON und H. SOLIMAN: RFC 4861 Neighbor Discovery for IP version 6 (IPv6), September 2007, http://tools.ietf.org/html/rfc4861 .

[RFC4941] T. NARTEN, R. DRAVES und S. KRISHNAN: RFC 4941 Privacy Extensions for Stateless Address Autoconfiguration in IPv6, September 2007, http://tools.ietf.org/html/rfc4941 .

[RFC5375] VELDE, ET AL G. VAN DE: RFC 5375 IPv6 Unicast Address Assignment Considerations, December 2008, http://tools.ietf.org/html/rfc5375 .

[RFC5517] HOMCHAUDHURI, S. und M. FOSCHIANO: RFC 5517 Cisco Systems' Private VLANs: Scalable Security in a Multi-Client Environment, February 2010, http://tools.ietf.org/html/rfc5517 .

[RFC6105] E. LEVY-ABEGNOLI, ET AL: RFC 6105 IPv6 Router Advertisement Guard, 2011, http://tools.ietf.org/html/rfc6105 .

[RFC6325] R. PERLMAN, ET AL: RFC 6325 Routing Bridges (RBridges): Base Protocol Specification, July 2011, http://tools.ietf.org/html/rfc6325 .

[RFC6326] D. EASTLAKE 3RD, ET AL: RFC 6326 Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS, July 2011, http://tools.ietf.org/html/rfc6326 .

[RFC826] PLUMMER, D. C.: Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware, November 1982, http://tools.ietf.org/html/rfc826 .

[RIPE-IPv4] NCC, RIPE: RIPE NCC Begins to Allocate IPv4 Address Space From the Last /8, September 2012, https://www.ripe.net/internet-coordination/news/announcements/ripe-ncc-begins-to-allocate-ipv4-address-space-from-the-last-8 .

[RTSS 09] RISTENPART, THOMAS, ERAN TROMER, HOVAV SHACHAM und STEFAN SAVAGE: Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In: Proceedings of the 16th ACM conference on Computer and communications security, CCS '09, Seiten 199–212, New York, NY, USA, 2009. ACM, http://doi.acm.org/10.1145/1653662.1653687 .

[Schm 12] SCHMIDT, BERNHARD: linux-net: VXLAN multicast receive not working, November 2012, http://markmail.org/message/65e6x4mllykoeejd .

[SIYA 11] SUZAKI, KUNIYASU, KENGO IIJIMA, TOSHIKI YAGI und CYRILLE ARTHO: Memory deduplication as a threat to the guest OS. In: Proceedings of the Fourth European Workshop on System Security, EUROSEC '11, Seiten 1:1–1:6, New York, NY, USA, 2011. ACM, http://doi.acm.org/10.1145/1972551.1972552 .

[Tayl 00] TAYLOR, DAVID: Intrusion Detection FAQ: Are there Vulnerabilites in VLAN Implementations? VLAN Security Test Report, July 2000, http://www.sans.org/security-resources/idfaq/vlan.php .

[vdP 12] POL, RONALD VAN DER: TRILL and IEEE 802.1aq Overview, April 2012, https://noc.sara.nl/nrg/publications/TRILL-SPB.pdf .

[VMDK-Vulnerability] MATTHIAS LUFT, DANIEL MENDE, ENNO REY und PASCAL TURBING: VMDK Has Left the Building – Attacking Cloud Infrastructures by Malicious VMDK Files, 2012, http://www.insinuator.net/2012/05/vmdk-has-left-the-building/ .

[VMdvS] VMWARE: VMware vNetwork Distributed Switch: Migration and Configuration, 2012, http://www.vmware.com/files/pdf/vsphere-vnetwork-ds-migration-configuration-wp.pdf .

[VMhard] VMWARE: vSphere 5.0 Hardening Guide, 2012, http://communities.vmware.com/docs/DOC-19605 .

[VMmax] VMWARE: Configuration Maximums VMware vSphere 5.0, 2011, http://www.vmware.com/pdf/vsphere5/r50/vsphere-50-configuration-maximums.pdf .

[VMsec] VMWARE: vSphere-Sicherheit ESXi 5.0, 2012, http://pubs.vmware.com/vsphere-50/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-50-security-guide.pdf .

[VMtps] VMWARE: Transparent Page Sharing (TPS) in hardware MMU systems, 2011, http://kb.vmware.com/kb/1021095 .

[VMwa 12] VMWARE: VMware to Aquire Nicira, July 2012, http://www.vmware.com/company/news/releases/vmw-nicira-07-23-12.html .

[VXLAN] IETF Draft: VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, August 2012, http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-02 .

[Wang 11] WANGUHU, KAMAU: VXLAN-Primer Part 1, November 2011, http://www.borgcube.com/blogs/2011/11/vxlan-primer-part-1/ .

[ZJRR 12] ZHANG, YINQIAN, ARI JUELS, MICHAEL K. REITER und THOMAS RISTENPART: Cross-VM side channels and their use to extract private keys. In: Proceedings of the 2012 ACM conference on Computer and communications security, CCS '12, Seiten 305–316, New York, NY, USA, 2012. ACM, http://doi.acm.org/10.1145/2382196.2382230 .