Gefährdungen und Gegenmaßnahmen beim Einsatz von VCE Vblock

Version: 2.5

Bundesamt
für Sicherheit in der
Informationstechnik

Eine Studie in Zusammenarbeit mit dem
Bundesamt für Sicherheit in der Informationstechnik

Autoren

Stephan Bohnengel, VMware

Klaus Böttcher, EMC

Dr. Clemens Doubrava, BSI

Alex Didier Essoh, BSI

Yves Fauser, Cisco

Isabel Münch, BSI

Norbert Olbrich, RSA

Michael Otto, EMC

Gerald Pernack, RSA

Wolfgang Reh †, EMC

Hinweis

Inhaltsverzeichnis

Verzeichnis der Gefährdungen

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Aktuellen Umfragen zufolge wird sich die Nachfrage nach Cloud-Dienstleistungen in den nächsten Jahren weltweit stark erhöhen. Die Gründe für das zunehmende Interesse an Cloud Computing und die steigende Nutzung von Cloud-Diensten sind vielfältig. Cloud Computing verspricht hohe Flexibilität bei der Buchung und Nutzung sowie Stilllegung von Ressourcen, je nach aktuellem Bedarf. Erwartet wird auch ein hohes Einsparpotential im Bereich der ansonsten lokal vorzuhaltenden, zu wartenden und zu erneuernden IT-Systeme. Weitere Vorteile sind die Standardisierung und die ubiquitäre Verfügbarkeit von Geschäftsanwendungen.

Diesen potentiellen Vorteilen steht eine Reihe von Risiken gegenüber, die mit einer Auslagerung der Daten bzw. Anwendungen in eine Public Cloud verbunden sind. Daten bzw. Anwendungen werden außer Haus verlagert und sind somit dem direkten Zugriff durch die eigene IT entzogen. Darüber hinaus könnten geltende Richtlinien und Vorgaben wie z. B. Datenschutzanforderungen verletzt werden, wenn sensitive Daten in eine Public Cloud ausgelagert werden.

Um die potentiellen Vorteile von Cloud Computing nutzen zu können und dennoch die Kontrolle über die IT-Infrastruktur zu behalten, greifen viele Anwender zur Bereitstellung der Dienste auf eigene virtualisierte Rechenzentren (Private Clouds) zurück. Nichtsdestotrotz müssen Vertraulichkeit, Integrität und Verfügbarkeit auch in einer Private Cloud gewährleistet werden, indem infrastrukturelle, organisatorische und technische Maßnahmen zum sicheren Betrieb und zur sicheren Nutzung umgesetzt werden (siehe BSI-Eckpunktepapier Cloud Computing; [18]).

Viele Kunden wollen heute intern die Flexibilität von Virtualisierungstechniken ausnutzen, um interne Clouds im Unternehmen zu realisieren. Cloud Computing ist keine Technologie, sondern ein Paradigma, wie IT-Dienste innerhalb eines Unternehmens möglichst schnell bereitgestellt werden und wie diese aus Anwendersicht konsumiert werden. Ein vordefinierter IT-Warenkorb legt fest, welche Dienste und Leistungen die IT-Abteilung erbringt, und wie diese verbrauchsabhängig verrechnet werden. Idealerweise wäre eine Fachabteilung in der Lage, den benötigten IT-Dienst im Rechenzentrum mittels eines Self-Service Portals eigenhändig zu provisionieren.

Die Virtual Computing Environment Coalition (kurz "VCE-Koalition") wurde von Cisco, EMC und VMware als Joint Venture gegründet, um die IT-Virtualisierung zu vereinfachen und technische Risiken auf dem Weg zur Infrastrukturvirtualisierung bis hin zur Private Cloud zu verringern. Ein Vblock ist ein Infrastruktur-Paket der VCE-Koalition, in dem Blade-Server, Virtualisierung, Netze, Speicher, Sicherheitskomponenten und Management in einer Komplettlösung vereint sind.

Mit den Vblock-Produkten werden den Kunden Pakete angeboten, welche vorkonfigurierte und aufeinander abgestimmte Hard- und Softwarekomponenten enthalten. Diese hochintegrierten Komponenten stellen Standardeinheiten für die Infrastruktur dar, mit denen man leicht und schnell das Rechenzentrum erweitern kann.

Die Einzelkomponenten für die Vblock-Architektur sind vorgegeben. Architekturvorgaben und getestetes Design der drei Hersteller bieten den Kunden Vorteile gegenüber selbst zusammengestellten Komponenten, die ähnliche Funktionalität bieten, in Bezug auf Performance, Skalierbarkeit und Verfügbarkeit. Alle Hardware-Komponenten des Vblocks sind redundant ausgelegt, um höchste Verfügbarkeit zu ermöglichen. Durch die unveränderbaren Komponenten können Kunden auf ein validiertes Design mit all seinen Performance-Aspekten zurückgreifen.

1.2 Zielsetzung

Das vorliegende Dokument betrachtet die Sicherheitsaspekte, die sich aus Betrieb und Nutzung von VCE Vblock 1/1u, Vblock 300 und Vblock 700 ergeben (Vblock 0 wird nicht betrachtet). Diese Studie beschreibt die damit verbundenen Gefährdungen und zeigt, in Anlehnung an die IT-Grundschutz-Kataloge des Bundesamts für Sicherheit in der Informationstechnik (BSI), Maßnahmen zum sicheren Betrieb eines Vblocks auf. Der Fokus wird auf Cloud-spezifische Gefährdungen gelegt oder auf solche, die aufgrund der Einsatzumgebung besondere Relevanz erhalten. Es werden jedoch ausschließlich

technische und organisatorische Aspekte behandelt. Gefährdungen und Maßnahmen, die bereits heute in den IT-Grundschutz-Katalogen aufgeführt sind, werden in der vorliegenden Studie nicht betrachtet.

1.3 Adressatenkreis

Die Studie wendet sich primär an IT-Planer, Architekten für Virtualisierung, Architekten für Informationssicherheit, und Administratoren. Sie soll als Anregung dienen für alle Personen, die sich in irgendeiner Weise mit der Sicherheit der Vblock Infrastructure Packages beschäftigen.

Es wird angenommen, dass der Leser / die Leserin Vorkenntnisse über Informationssicherheit besitzt, besonders in Bezug auf Storage, Netze und Virtualisierung.

2 Vblock-Überblick

2.1 Vblock Infrastructure Packages

In diesem Kapitel werden die verschiedenen Vblock Infrastructure Packages kurz vorgestellt, die zum Zeitpunkt der Erstellung der Studie verfügbar waren.

Unter dem Ausdruck "Vblock Infrastructure Package" werden folgende Lösungsangebote verstanden:

- **Vblock Series 700** (früher als "Vblock 2" bezeichnet)
  "Highend" System: Entwickelt für den Einsatz mit einer sehr hohen Anzahl von virtuellen Maschinen und Benutzern; erfüllt höchste Performance- und Verfügbarkeitsansprüche von geschäftskritischen Anwendungen.
- **Vblock Series 300**
  "Midrange" System: Entwickelt für den Einsatz mit einer hohen Anzahl von virtuellen Maschinen und Benutzern; erfüllt hohe Performance- und Verfügbarkeitsansprüche von Anwendungen.
- **Vblock1, Vblock 1u**
  "Midrange" System: Entwickelt für den Einsatz mit einer mittleren bis hohen Anzahl von virtuellen Maschinen und Benutzern; geeignet für Anwendungen wie z.B. E-Mail-, File- und Print-Services, virtuelle Desktops, etc.
- **Vblock 0**
  "Einsteiger-Variante": Entwickelt für den Einsatz mit einer kleinen bis mittleren Anzahl von virtuellen Maschinen und Benutzern; geeignet für Entwicklungs- und Testumgebungen oder kleinere Produktionsumgebungen.

Die nachfolgende Tabelle 1 zeigt im Überblick die Einsatzbereiche der Vblock-Plattformen.

	Vblock 0	Vblock 1 / 1U	Vblock 300	Vblock 700
Einsatzbereich	Small und Medium Business	Commercial und Enterprise	Commercial und Enterprise und Service Provider	Enterprise und Service Provider
Private / Public Cloud	Private Cloud	Private Cloud	Public und Private Cloud	Public und Private Cloud

Tabelle 1: Einsatzbereiche der verschiedenen Vblock-Plattformen

Die nachfolgende Tabelle 2 zeigt pro Vblock drei verschiedene Werte für die Anzahl der VMs pro physischem Server sowie für die VM P2V Rate (Best Practices). Dabei bezeichnet die "VM P2V Rate" das Verhältnis von physischen zu virtuellen Maschinen.

Der mittlere Wert (fett markiert) zeigt die wahrscheinliche Anzahl von VMs in einer üblichen IT-Umgebung. Der Wert 1:15 wird üblicherweise zur überschlägigen Berechnung der Kapazität verwendet. Zu einer exakten Bestimmung wird ein Vorab-Assessment benötigt.

Ist der Workload der einzelnen VMs entsprechend hoch, geht die Anzahl der VMs in Richung des linken (kleineren) Wertes. Ist der Workload der einzelnen VMs entsprechend niedrig, geht die Anzahl der VMs in Richtung des rechten (größeren) Wertes.

	Vblock 0	Vblock 1 / 1U	Vblock 300	Vblock 700
Anzahl VMs	320 ; **960** ; 2560	320 ; **960** ; 2560	320 ; **960** ; 2560	1.280 ; **3.840** ; 10.240
VM P2V Rate	1:5 ; **1:15** ; 1:40	1:5 ; **1:15** ; 1:40	1:5 ; **1:15** ; 1:40	1:5 ; **1:15** ; 1:40

Tabelle 2: Anzahl VMs und VM P2V Rate bei den verschiedenen Vblock-Plattformen

Die verschiedenen Vblocks können mit unterschiedlichen UCS Blade Servern ausgestattet werden. Die Anzahl der maximalen CPU Cores hängt dann direkt von der Art und Konfiguration der eingesetzten Blades (Typ 200, 230, 250, 440) ab.

Die nachfolgende Tabelle 3 zeigt einige grundsätzliche Unterschiede bei verschiedenen Vblock-Plattformen.

	Vblock 0	Vblock 1 / 1U	Vblock 300	Vblock 700
Anzahl Chassis (ist erweiterbar, siehe Kapitel 2.5.3)	1	2	4	8
Anzahl Blades	4-16	16-64 / 8-64	8-64	max. 160 full-width max. 320 half-width
Storage Protokoll	NFS (optional iSCSI)	FC (optional NFS)	FC (optional NFS, iSCSI)	FC

Tabelle 3: Grundsätzliche Unterschiede bei verschiedenen Vblock-Plattformen

Hinweise:

(1) Vblock 0 wird in dieser Studie nicht betrachtet.

(2) Vblock 1 enthält optional iSCSI; Vblock 1U enthält optional NFS; Vblock 300 enthält optional iSCSI oder NFS. Die Gefährdungen, die sich durch den Einsatz der Protokolle iSCSI oder NFS ergeben können, werden im vorliegenden Dokument nicht betrachtet.
Begründung: Das Dokument ist fokussiert auf Enterprise-Kunden und Internet Service Provider. Bei diesen Zielgruppen kommen die genannten Protokolle im produktiven Umfeld selten vor.

Für weiterführende Detailinformationen zu den Vblock Infrastructure Packages wird auf die Dokumente von VCE verwiesen:

Vblock Infrastructure Platforms; [01]

Vblock Infrastructure Platform Architecture Overview; [02]

Vblock Infrastructure Platforms Technical Overview; [03]

2.2 Vblock-Architektur

Grundsätzlich besteht ein Vblock aus folgenden Schichten:

- Storage
- SAN
- Network (optional)
- Compute
- Management

Die verschiedenen Schichten und ihre Ausprägung mit Produkten sind in der nachfolgenden Abbildung 1 dargestellt.

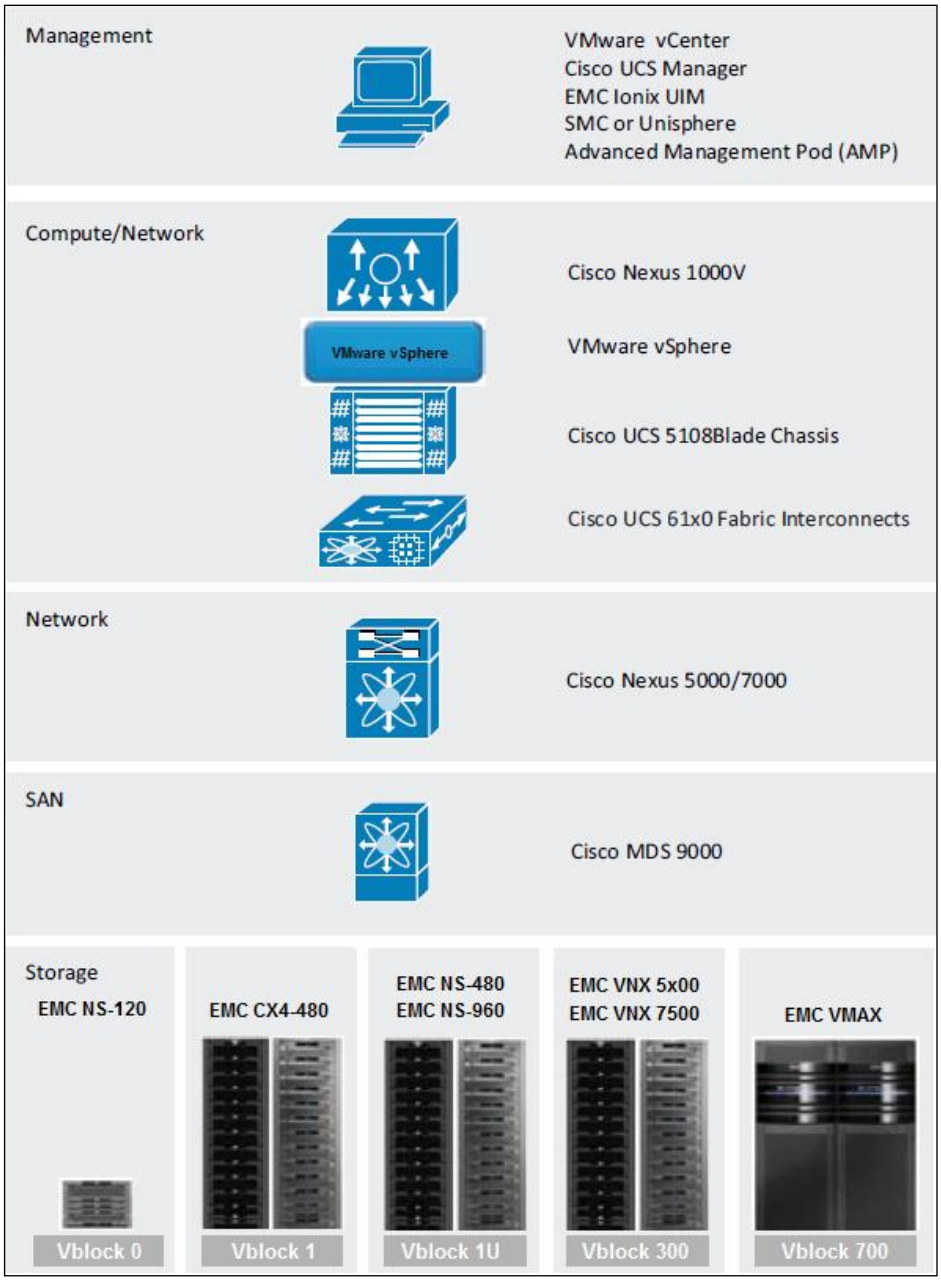


Abbildung 1: Schichten eines Vblocks

2.2.1 Vblock Storage

In dieser Schicht befindet sich der Plattenspeicher des gesamten Systems. Eingesetzt werden hier je nach Typ des Vblocks verschiedene Speichersysteme von EMC (siehe Tabelle 4).

Eine Anbindung an andere bestehende Plattenspeichersysteme ist nur zu Migrationszwecken temporär möglich.

2.2.2 Vblock SAN

In dieser Schicht befinden sich Fibre Channel (FC-)Switche zur Anbindung der Plattenspeichersysteme über Fibre Channel an die Compute Schicht. Eingesetzt werden Systeme der Cisco MDS-Serie.

2.2.3 Vblock Network

In dieser Schicht befinden sich optionale Netzaggregationskomponenten, um mehrere Vblock-Systeme miteinander verbinden zu können. Eingesetzt werden Systeme der Cisco Nexus 5000/7000-Serie.

Normalerweise wird der Vblock allerdings über den Fabric Interconnect der Compute Schicht direkt an bestehende RZ-Netze des Kunden angeschlossen.

2.2.4 Vblock Compute

In dieser Schicht befinden sich die Computing Ressourcen (Blade Server), der Fabric Interconnect und der Hypervisor der Lösung. Eingesetzt werden Systeme der Cisco UCS B-Serie für die Computing Ressourcen sowie ESX/ESXi-Systeme von VMware für den Hypervisor.

2.2.5 Vblock Management

In dieser Schicht befinden sich zwei Systemkategorien: Element Manager und übergeordnete Provisionierungs- und Überwachungssysteme.

Element Manager werden verwendet zur detaillierten Verwaltung einzelner Komponenten. Ein Element Manager kann sowohl Bestandteil der Komponente als auch eine externe Applikation sein.

Die Element Manager bekommen von den Provisionierungssystemen Konfigurationsinformationen, um komplexe Gesamtaufgaben (z.B. die Provisionierung von Speicherplatz, Computing Ressourcen, Fibre Channel- und Netzkonfigurationen) umzusetzen.

2.3 Vblock-Hauptkomponenten

Die nachfolgende Tabelle 4 zeigt die Hauptkomponenten der verschiedenen Vblock-Pakete.

	VBLOCK 0	VBLOCK 1U / VBLOCK 1	VBLOCK SERIES 300	VBLOCK SERIES 700
STORAGE	EMC Celerra Unified Storage NS-120 **Drive Types** • EFD • Fibre Channel • SATA	**Vblock 1U-** EMC Celerra Unified Storage NS-480 or NS960 **Vblock1-** EMC CLARiiON CX4 Model 480 **Drive Types** • EFD • Fibre Channel • SATA	**300HX** EMC VNX 7500 **300GX** EMC VNX 5700 **300FX** EMC VNX 5500 **300EX** EMC VNX 5300 **Drive Types** • EFD • SAS • NL-SAS	**700MX** EMC Symmetrix VMAX **Drive Types** • EFD • Fibre Channel • SATA
COMPUTE	Cisco UCS • B200M2 • B250M2 • B230M1	Cisco UCS • B200M2 • B250M2 • B440M1 • B230M1	Cisco UCS • B200M2 • B250M2 • B440M1 • B230M1	Cisco UCS • B200M2 • B250M2 • B440M1 • B230M1
NETWORKING	Cisco Nexus 1000V Cisco Nexus 5000 series switches Cisco MDS 9000 Series SAN Switch	Cisco Nexus 1000V Cisco Nexus 5000 series switches Cisco MDS 9000 Series SAN Switch	Cisco Nexus 1000V Cisco Nexus 5000 series switches Cisco MDS 9000 Series SAN Switch	Cisco Nexus 1000V Cisco MDS 9000 Series SAN Switch
VIRTUALIZATION	VMware vSphere 4 Enterprise Plus Suite			
SECURITY	Individual component security tools and protocols RSA enVision, RSA SecurID® (both optional)			
ORCHESTRATION	Ionix Unified Infrastructure Manager 2.1 Advanced Management POD (AMP) (optional)			
VIRTUALIZATION	Virtualization vSphere ESX 4, 4.1, or ESXi 4.1i, vCenter, Nexus 1000V with per CPU license			
COMPUTE/ NETWORKING MANAGEMENT	Cisco UCS Manager / Cisco Fabric Manager			
STORAGE	EMC Unisphere®	EMC Unisphere®	EMC Unisphere®	EMC Symmetrix Management Console

Tabelle 4: Hauptkomponenten der verschiedenen Vblock-Plattformen

2.4 Vblock Management

Element Manager sind:

- **VMware vCenter Server**
  Dieser Element Manager dient zur Verwaltung der Hypervisor Server.
- **Cisco UCS Manager**
  Dieser Element Manager dient zur Verwaltung der Computing Ressourcen (Blade Server) sowie deren Peripherie (z.B. FC-Uplinks, Netz-Uplinks, etc.).
- **EMC Symmetrix Management Console oder EMC Unisphere Manager**
  Diese Element Manager dienen zur Verwaltung der verschiedenen Plattenspeichersysteme.
- **Cisco Datacenter Manager (DCNM)**
  Dieser Element Manager dient als optionale Komponente zur Verwaltung der FC-Switche über ein graphisches Interface. Er wird auch zur Verwaltung der optionalen Nexus 5000/7000 Komponenten verwendet.

Die nachfolgende Abbildung 2 zeigt die Schnittstellen zwischen den Element Managern, welche ein erweiterbares offenes Management Framework darstellen.

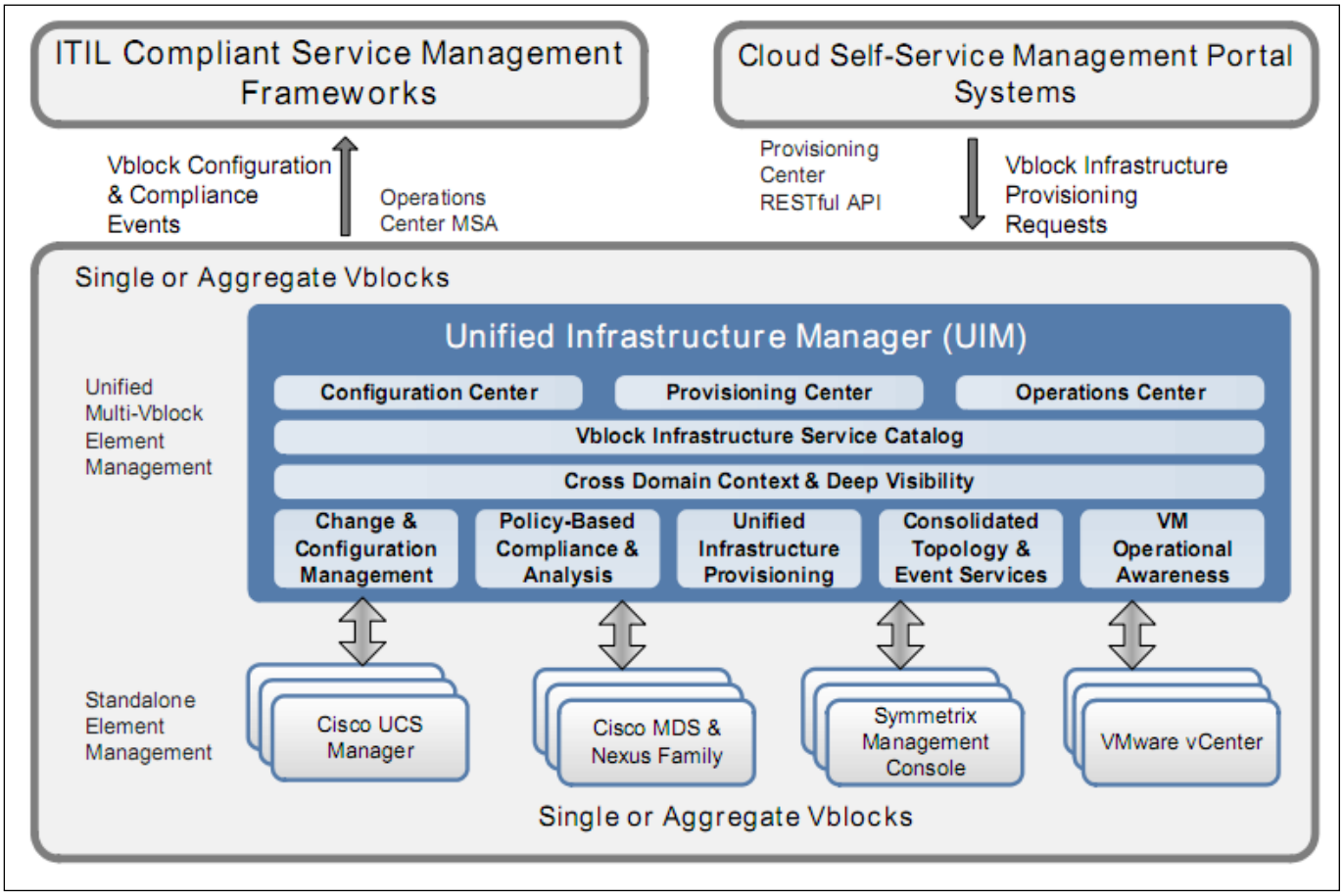


Abbildung 2: Verwaltung der Vblock-Plattform

Der Unified Infrastructure Manager (UIM) dient als zentrale Schnittstelle, um die Systemkonfiguration, das Ausrollen und die Konformität der Konfiguration untereinander zu verwalten. Durch den hohen Abstraktionsgrad erleichtert der UIM mittels seiner Serverprofile sowie seiner programmierbaren Schnittstellen das Einbinden der Vblock-Plattform in Servicekataloge und Workflow Orchestrierung.

Er kommuniziert zur Realisierung der Aufgaben, wie in der obigen Abbildung 2 dargestellt, mit den jeweiligen Element Managern und dient als Meta Werkzeug.

VMware vCenter hat unter den Element Managern einen Sonderstatus, da es als zentrale Verwaltungseinheit der Virtualisierung über dem UIM ansetzt. D.h. der UIM provisioniert vor allem die Infrastruktur (Storage, Netz, Computing und Hypervisor). Das vCenter nutzt diese Infrastruktur sobald sie bereitgestellt ist, um VMs auszurollen.

2.5 Einsatzbereiche

2.5.1 Vblock im Enterprise-Umfeld

Applikationen werden virtualisiert oder nicht-virtualisiert (auch mit dem Begriff "*Bare Metal*" beschrieben) auf Vblock-Komponenten ausgerollt; primärer Anwendungsfall ist aber das Ausrollen auf virtuellen Maschinen.

Für die Vblock-Varianten gibt es bereits mehrere getestete und validierte Designs für bestimmte Tier1-Applikationen, z.B. SAP, VMware View.

Um die Performance von Applikationen zu gewährleisten, sind das Management, die Überwachung und die Skalierbarkeit folgender vier Kernkomponenten essentiell:

- CPU
- Memory
- Network
- Disk

Im Enterprise-Umfeld muss zudem gewährleistet sein, dass die administrative Kontrolle der Komponenten an die jeweiligen Administrationsabteilungen delegiert werden kann.

Um in der Administration Gewaltenteilung zu erreichen, müssen die Management Tools (z.B. VMware vCenter, Cisco UCS Manager und EMC Ionix UIM) die Fähigkeit besitzen, Rechte zu delegieren.

Teile des Vblocks können auch von unterschiedlichen IT-Administrationsteams verwaltet werden (Storage, Netz, Virtualisierung, Sicherheit), womit auch eine Gewaltenteilung innerhalb der Lösung realisierbar ist.

2.5.2 Vblock bei Cloud Service Provider (CSP)

Im CSP- Bereich gelten die gleichen Vorteile wie im Enterprise-Umfeld: Hoher Automatisierungsgrad bei der Erweiterung der Infrastrukturkomponenten unter Einhaltung definierter Sicherheitsbereiche.

Aus rechtlicher oder Compliance Sicht muss man hier zwischen interner und externer Cloud unterscheiden.

Wenn man in einer internen Cloud ein "Infrastructure as a Service"-Modell (IaaS) unter dem Aspekt eines Self Service Portals (z.B. VMware vCloud Director) betreiben möchte, können Mandanten im Unternehmen unterschiedliche Abteilungen sein. Diese internen Mandanten wären z.B. eine Schulungsabteilung und eine Entwicklungsabteilung, die man voneinander isolieren möchte und dafür im Servicekatalog bestimmte Service Level Agreements definiert hat.

Als externer Cloud-Anbieter betreibt man für eigenständige Organisationen deren Anwendungen auf einer gemeinsamen geteilten Plattform unter Einhaltung kompletter Isolationsmechanismen. Neben der Mehrmandantenfähigkeit sind hierbei die Ressourcenplanung und die Skalierbarkeit der Umgebung von entscheidender Bedeutung.

2.5.3 Skalierung des Vblock

Die nachfolgende Abbildung zeigt beispielhaft die einfachen Skalierungsmöglichkeiten durch definierte Infrastrukturkomponenten. Die anfängliche Konfiguration *Vblock 1 Base* kann zunächst um die Komponenten von *Vblock 1 Storage Expansion* und *Vblock 1 Compute Expansion* erweitert werden. In einem weiteren Schritt können dann z.B. die Komponenten von *Vblock 700 Base* hinzugefügt werden.

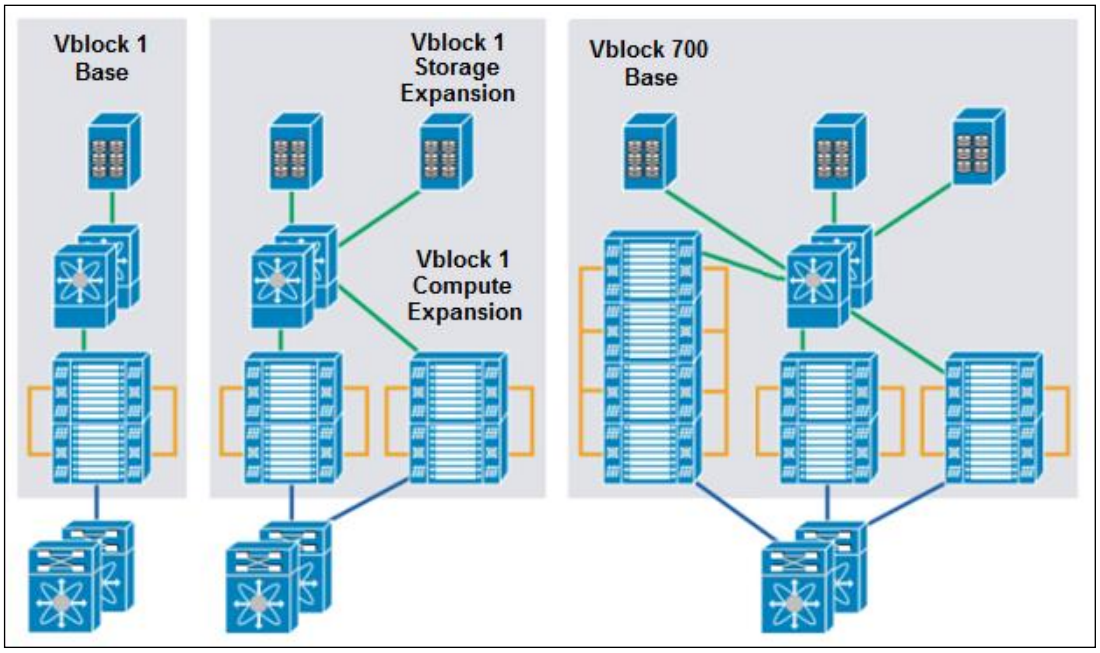


Abbildung 3: Einfache Skalierung durch definierte Infrastrukturpakete

3 Eingesetzte Techniken

Das nachfolgende Kapitel gibt einen Überblick über die technologischen Neuerungen, die im Vblock eingesetzt werden. Dabei wird das Hauptaugenmerk auf die Kernunterschiede zwischen klassischen und virtualisierten Umgebungen gelegt.

3.1 VMware Hypervisor anstatt "Bare Metal"

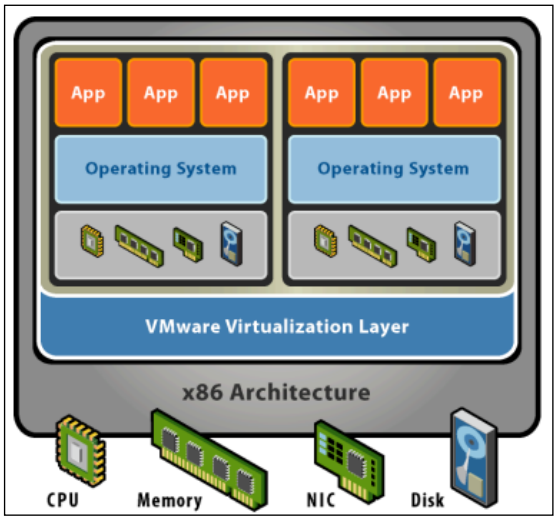


Abbildung 4: VMware Hypervisor Architektur

Die Abbildung 4 zeigt die Architektur des VMware Hypervisors.

Bei der Virtualisierung von IT-Systemen werden ein oder mehrere virtuelle IT-Systeme auf einem physischen Computer (Host) betrieben (beim Vblock der sogenannte ESX-Server).
Eine wesentliche Komponente bei der Servervirtualisierung ist der Hypervisor (in der nebenstehenden Abbildung mit "VMware Virtualization Layer" bezeichnet). Der Hypervisor teilt dem Gast dynamisch physische Ressourcen wie CPU, Memory, Storage und Netz zu.

Um den besonderen Herausforderungen gerecht zu werden, die sich beim Einsatz virtueller Infrastrukturen ergeben, hat VMware eine Reihe von Techniken eingeführt, wie z.B.

- vMotion
- Dynamic Resource Scheduling (DRS)
- High Availability (HA)
- Fault Tolerance (FT)
- Distributed Power Management (DPM)

Diese werden im Folgenden beschrieben.

Enkapsulierung

Eine virtuelle Maschine ist eigentlich ein Software-Container, der Hardware Ressourcen, Betriebssystem und Anwendungen innerhalb eines Software-Pakets bündelt.

Eine VM läuft im nicht-privilegierten Modus auf dem Hypervisor und besteht aus wenigen Dateien, die leicht migriert werden können. Die Hardware wird von der VM abstrahiert – dadurch entstehen keine zusätzlichen Migrationskosten bei einem Hardwaretausch.

Durch die Abstrahierung des Betriebssystems und dessen Applikationen ergeben sich neue Möglichkeiten, die auf reiner Hardware nicht möglich gewesen wären. Zum Beispiel können nun Server hochverfügbar auf virtueller Infrastruktur betrieben werden. Darüber hinaus erleichtert die Hardwareunabhängigkeit sowohl die Wartung der Systeme als auch die Lastverteilung.

Höhere Verfügbarkeit

Eine VM wird auf einem Cluster von physischen **ESX-Servern** betrieben, welcher von VMware vCenter zentral verwaltet wird. Mittels VMware **vMotion** (Live Migration) lassen sich VMs im laufenden Betrieb unterbrechungsfrei zwischen physischen ESX-Servern verschieben. Dies hilft z.B. geplante Wartungsfenster für physische Systeme während normaler Geschäftszeiten zu realisieren. Dazu setzt man setzt einen ESX-Server in den sogenannten "Maintenance"-Modus, dann wird der ESX-Server von

den virtuellen Diensten befreit. Somit lässt sich auch z.B. der Hypervisor selbst unterbrechungsfrei patchen oder man kann Erweiterungen an der Hardware durchführen.

Eine weitere Technik ist **VMware HA**, die den Kunden vor ungeplanten Ausfällen schützt. Dabei beobachten sich die ESX-Server gegenseitig im Cluster. Gibt es nun einen Defekt auf einem ESX-Server, werden die darauf betriebenen VMs automatisch auf den verbleibenden ESX-Server "crash-konsistent" neu gestartet. Um eine ähnliche Verfügbarkeit in einem klassischen Rechenzentrum zu erreichen, müsste für jeden Server ein Cold-Standby-Cluster einrichtet werden.

Lastenausgleich

Die ESX-Server im Cluster teilen untereinander auch Lastinformationen. Mit dem **VMware Dynamic Resource Scheduler (DRS)** werden VMs mittels vMotion immer dorthin migriert, wo sie vorhandene freie Compute Ressourcen finden. Zur Migration der VMs werden Mindestanforderungen an die CPU-Leistung und den Arbeitsspeicher herangezogen. Damit findet ein Loadbalancing von VMs statt, und die IT-Dienste laufen immer auf dem ESX-Server, der eine optimale Performance im Cluster bietet. Durch DRS-Regeln kann erreicht werden, dass redundante VMs immer auf unterschiedlichen physischen ESX-Servern laufen (z.B. Domaincontroller, Microsoft Cluster Dienste, Load Balancer, etc.). Darüber hinaus kann mit Hilfe der DRS-Host-Affinität eine Beschänkung einer VM innerhalb eines Hosts erreicht werden.

Fault Tolerance (FT)

In der Administrationsschicht einer VM kann definiert werden, dass beim Ausfall der primären VM eine automatische Erstellung einer sekundären VM erfolgt. Damit ist ein unterbrechungsfreier ("nahtloser") Failover mit Statuserhalt der VMs gewährleistet, d.h. es gibt keinen Datenverlust und es entstehen keine Ausfallzeiten.

Distributed Power Management (DPM)

DPM sorgt für die kontinuierliche Optimierung des Energieverbrauchs im Rechenzentrum. Dabei werden VMs, die in einem DRS-Cluster konfiguriert sind, z.B. am Wochenende oder nachts auf eine geringere Anzahl von Servern konsolidiert, indem die durch weniger Last nicht mehr benötigten ESX-Server automatisch ausgeschaltet werden. Bei steigender Last werden die ausgeschalteten ESX-Server dann automatisch wieder zugeschaltet.

Green-IT

Es ist keine Seltenheit, dass auf einem Host 40 VMs und mehr betrieben werden. Dies führt zu erheblichen Energieeinsparungen und ist dadurch sehr kosteneffizient.

Isolierung und Sicherheit

Aufgrund der Tatsache, dass sich bei der Servervirtualisierung viele VMs einen gemeinsamen Host teilen, ist eine sichere Trennung der VMs durch den Hypervisor notwendig.

Der Hypervisor von VMware weist starke Isolierungsmechanismen auf, um zu verhindern, dass VMs über den Hypervisor untereinander kommunizieren können. Für weitere Details zur Isolierungsmechanismen siehe Kapitel 4.1.1.

Die ESX-Plattform wurde einer Common Criteria Evaluation mit dem Assurance Level (EAL) 4+ unterzogen und entsprechend zertifiziert. Certification Report und Security Target siehe: http://www.cse-cst.gc.ca/its-sti/services/cc/vmware-esx-v40-dec-eng.html

3.2 Verstärkte Nutzung von SAN und NAS Storage anstatt DAS

Die klassische Versorgung isolierter Server mit Datenspeicherkapazität erfolgt über DAS (Direct Attached Storage). DAS ist für die Virtualisierung nicht geeignet. Als dynamische Speicherform kommen für die Virtualisierung nur SAN-(Storage Area Network)- oder NAS-(Network Attached Storage)-Speicher in Frage, die folgende Eigenschaften aufweisen:

- Zentralisiert
- Gemeinsam nutzbar (shared)
- Partitionierbar (Multi-Tenant)
- Abstrahierbar bzw. virtualisierbar
- Hochverfügbar
- Zentral verwaltbar und wartbar

3.3 Nutzung von Netz-Virtualisierung

Der Großteil der Komponenten der Netz-Virtualisierung, die im Vblock eingesetzt werden, sind bereits seit Jahren Bestandteil von üblichen Netzdesigns. Die im Vblock eingesetzten Techniken sind:

- VLANs
- VSANs
- VN-Link

Diese werden nachfolgend detaillierter beschrieben.

3.3.1 VLAN

Ein Virtual Local Area Network (VLAN) ist ein logisches Teilnetz innerhalb eines Switches oder eines gesamten physischen Netzes. Ein VLAN kann sich hierbei über ein ganzes geswitchtes Netz hinziehen und braucht nicht nur auf einen einzelnen Switch beschränkt zu bleiben. Ein VLAN trennt physische Netze in Teilnetze auf, indem es dafür sorgt, dass VLAN-fähige Switche Frames (Datenpakete) eines VLANs nicht in ein anderes VLAN weiterleiten. Dies gilt auch, wenn die Teilnetze an gemeinsame Switches angeschlossen sind. Ein VLAN bildet gleichzeitig eine separate Broadcast-Domäne. Das bedeutet, dass Broadcasts nur innerhalb des VLANs verteilt werden.

Um VLANs über mehrere Switches auszudehnen, werden sogenannte Trunking-Protokolle eingesetzt. Hierbei werden pro Switch ein oder mehrere physische Ports für die Inter-Switch-Kommunikation reserviert; die logische Verbindung zwischen den Switches wird als *Trunk* bezeichnet. Ein Ethernet-Frame wird beim Informationsaustausch zwischen den Switches in das Trunking-Protokoll gekapselt. Dadurch ist der Ziel-Switch in der Lage, die Information dem entsprechenden VLAN zuzuordnen. Als Standard hat sich das Protokoll IEEE 802.1Q durchgesetzt. Frühere proprietäre Protokolle, wie z.B. ISL (Inter Switch Link) des Herstellers Cisco, werden kaum noch verwendet und werden von neueren Systemen auch nicht mehr unterstützt.

VLANs gehören seit Ende der 1990er Jahre zum normalen Handwerkszeug von Netzdesignern; deswegen wird hier auf die grundsätzliche Funktion eines VLANs und auf den IEEE 802.1Q Standard "Virtual Bridged Local Area Networks" nicht weiter eingegangen.

Im Gegensatz zu traditionellen Computing Umgebungen, bei denen häufig immer noch physisch getrennt wird, wird beim Vblock ausschließlich eine logische Trennung angewendet.

Die physische Trennung ist in der nachfolgenden Abbildung 5 beispielhaft gezeigt.

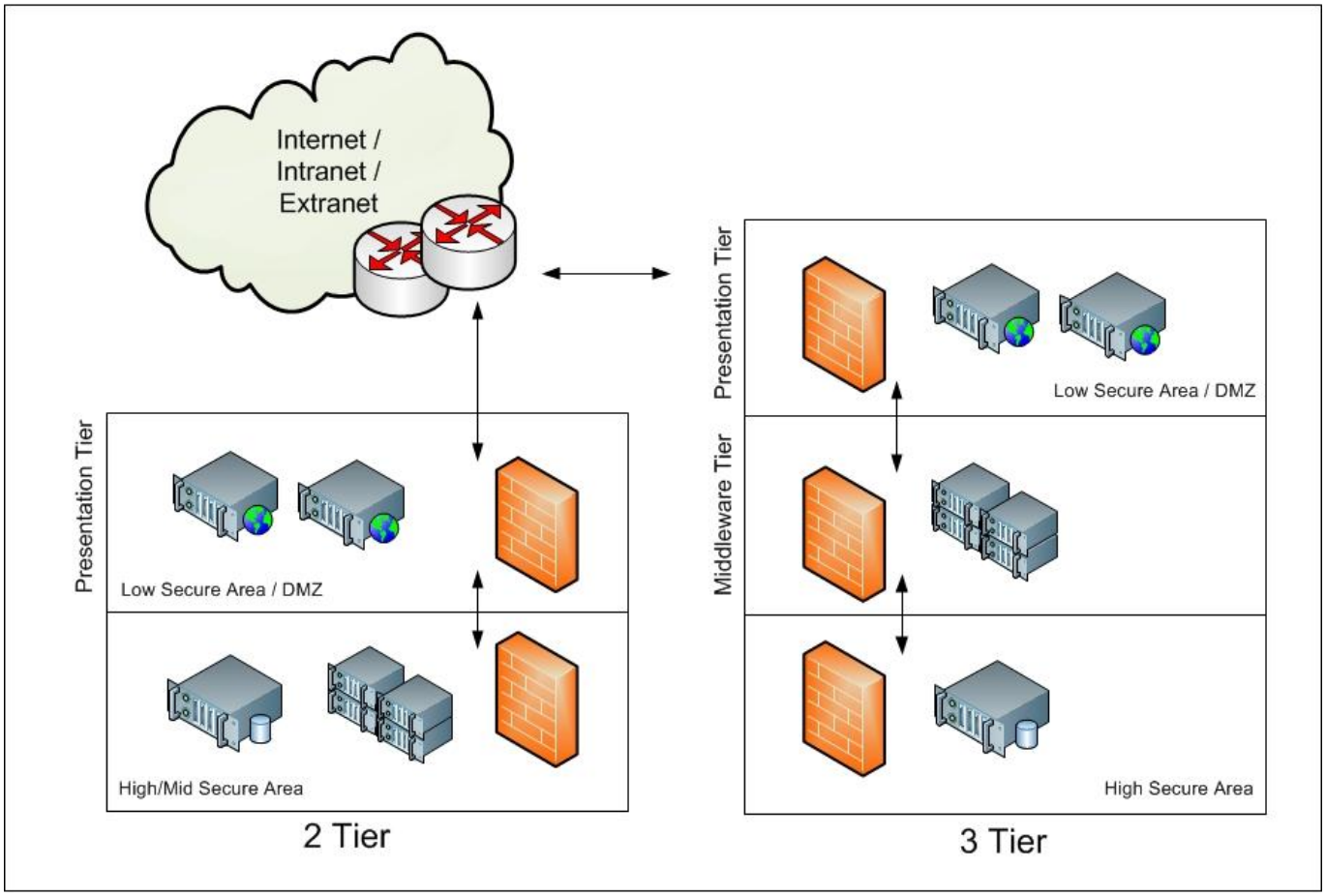


Abbildung 5: Trennung von Applikationen über mehrere Tiers

Wie in Abbildung 5 dargestellt, werden Applikationen im Normalfall in zwei oder drei Layers ("Tiers" oder Zonen) aufgebaut.

Der Presentation Layer besteht aus Servern, die dem Nutzer z.B. ein Web-Frontend bieten. Da diese Server von außen (z.B. dem Internet/Extranet/Intranet) erreichbar sein müssen, werden diese üblicherweise in einer DMZ-Zone aufgebaut. Der nächste Layer ist der Middleware Layer, der die Anwendungslogik darstellt. Hier werden die Daten verarbeitet, umgewandelt, usw. Der dritte Layer ist ein Datenbank Layer, der für die Speicherung der Daten sowie für das Liefern von Ergebnissen von Datenbankabfragen zuständig ist.

Bei 2-Tier Architekturen werden, wie oben dargestellt, der Middleware und Database Layer zusammengefasst. Zwischen den Layers werden Firewallsysteme eingesetzt, die eine Kommunikation zwischen ihnen nur über bestimmte Protokolle und Ports zulassen.

Von Interesse ist hier die Frage, ob und wann für die einzelnen Zonen komplett getrennte Hardware-Komponenten verwendet werden müssen.

Im Vblock werden Ressourcen gebündelt und den verschiedenen Tiers zur Verfügung gestellt. So können ein Presentation Server und ein Middleware Server gemeinsam als Gast auf einem VMware ESX-Hypervisor laufen.

Die Netzkomponenten im Vblock sorgen über VLANs dafür, dass der Netzverkehr zwischen Zonen getrennt zum nächsten System (z.B. Firewall) geführt wird.

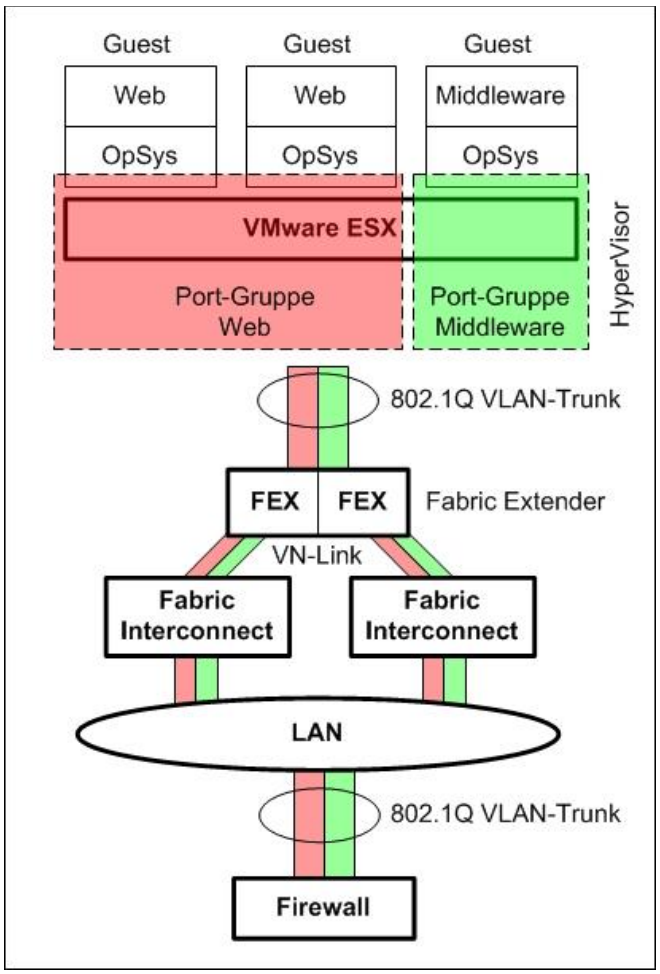


Abbildung 6: Anwendung von VLANs

Die in Abbildung 6 dargestellte Trennung von Application Tier Zonen, die in eine zentrale Firewall im Datacenter geführt wird, ist eines der möglichen Szenarien.

In Hosting-Umgebungen wird der Traffic üblicherweise für jeden Mandanten in seinen eigenen WAN Transport geführt. Dort werden Mittel wie Virtuelle Routing und Forwarding Tabellen (VRFs), MPLS Transport, IPSec-Tunnel Overlays, usw. verwendet, um eine Layer3-Trennung auch beim WAN-Transport bis zum Endnutzer aufrecht zu halten.

Oftmals führt die Trennung von Mandanten über VLANs in Hosting-Umgebungen aber zu Skalierungsproblemen. Wenn jedem Mandanten ein VLAN zugeordnet wird, steigt nicht nur die Anzahl von VLANs, sondern auch der "Verschnitt" der IP-Subnetze, dadurch, dass jedes Subnetz seine eigene Netz- und Broadcast Adresse bekommt. Wenn jeder Mandant ein eigenes VLAN- und IP-Subnetz erhält, gehen dem Hosting Provider dadurch schnell die ihm zugeordneten öffentlichen IP-Adressen aus. Aus diesem Grund werden "private VLANs" eingesetzt. Hierbei handelt es sich um spezielle VLANs, bei denen alle Access-Interfaces in einem VLAN- und IP-Subnetz gesetzt werden, diese Access-Interfaces aber untereinander keine Verbindung haben. Da sich alle Access-Interfaces im gleichen IP-Subnetz befinden, benötigen sie also keine eigene Subnetz-Adressierung.

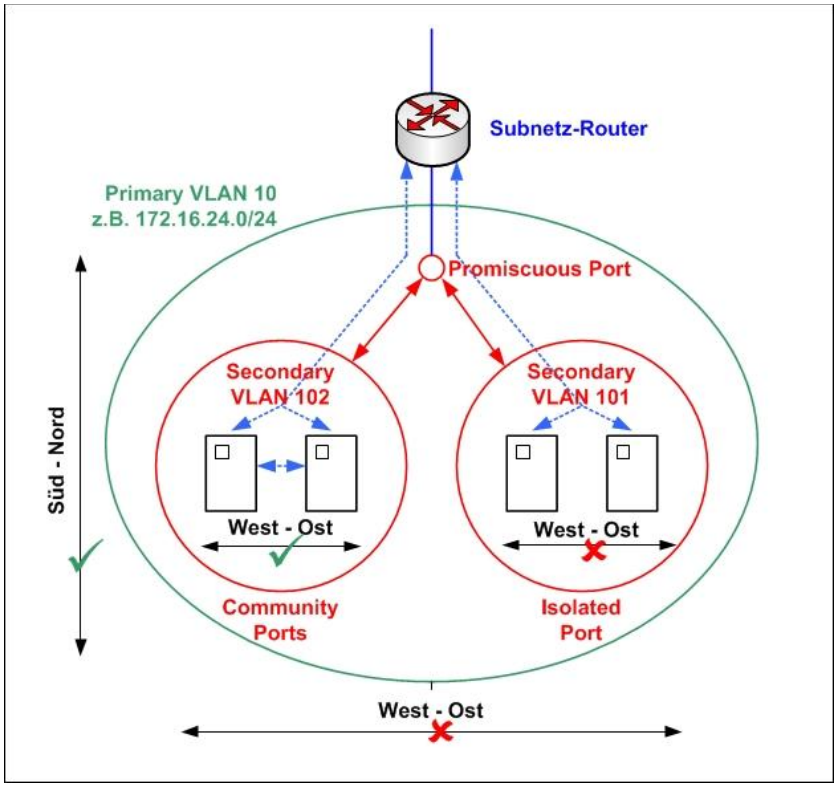


Abbildung 7: Private VLANs

Die Abbildung 7 zeigt die drei verschiedenen Port Typen:

- **Promiscuous Ports**
  Ports, die als Promiscuous Port konfiguriert sind, können jeden anderen Port im Private VLAN erreichen. Z.B. wird der Port des Subnetz-Routers (Default Gateway) als Promiscuous Port konfiguriert.
- **Isolated Ports**
  Ports, die einem isolated VLAN angehören, sind nur in der Lage Netzverkehr zu Promiscuous Ports zu senden. Jeglicher Netzverkehr zwischen Hosts an Isolated Ports wird unterbunden.
- **Community Ports**
  Dieser spezielle Typ ist in der Lage, Netzverkehr zu Promiscuous Ports zu senden, sowie Netzverkehr zu anderen Ports im selben Community VLAN zu senden.

Community Ports und Isolated Ports werden hierbei einem Secondary VLAN zugeordnet, welches aber kein eigenes IP-Subnetz bildet.

Auf diese Art wird "Nord-Süd"-Netzverkehr, d.h. z.B. zum WAN-Ausstieg, erlaubt, "West-Ost"-Verkehr zwischen Endsystemen im gleichen Subnetz wird untersagt. Dabei bilden Community Ports die Ausnahme von der Regel, und werden für Hosts desselben Mandanten verwendet.

Private VLANs können im Vblock z.B. auf dem optional einsetzbaren Nexus 1000v Virtual Switch oder direkt auf dem vSwitch des ESX Hypervisors konfiguriert werden.

Der Einsatz von privaten VLANs löst also das Skalierungsproblem in Hosting Lösungen, indem eine Flut von VLANs und IP-Subnetzen verhindert wird.

Der Einsatz von privaten VLANs ist sehr gut geeignet für Hosting-Lösungen, bei denen zwischen den Hosts entweder jeglicher Verkehr erlaubt (Community Ports), oder verboten (Isolated Ports) ist. Beim Multi-Tier Applikation Model ist aber eine granulare Traffic-Filterung zwischen den Zonen gefordert.

Um hierfür eine Lösung zu bieten, kann im Vblock optional eine virtuelle Firewall-Lösung verwendet werden, die die VMsafe APIs verwendet; Produktbeispiele wären die VMware VShield Produktfamilie oder Cisco VSG (Virtual Security Gateway).

Die nachfolgende Abbildung 8 beschreibt exemplarisch eine Hypervisor-basierte Firewall-Lösung auf der Basis von VMsafe. Diese kann auf jedem VMware ESX-Server installiert werden, und dient zur Kontrolle und Analyse des gesamten Datenverkehrs zwischen virtuellen Maschinen, auch wenn diese logisch im selben VLAN und damit IP-Subnetz sind.

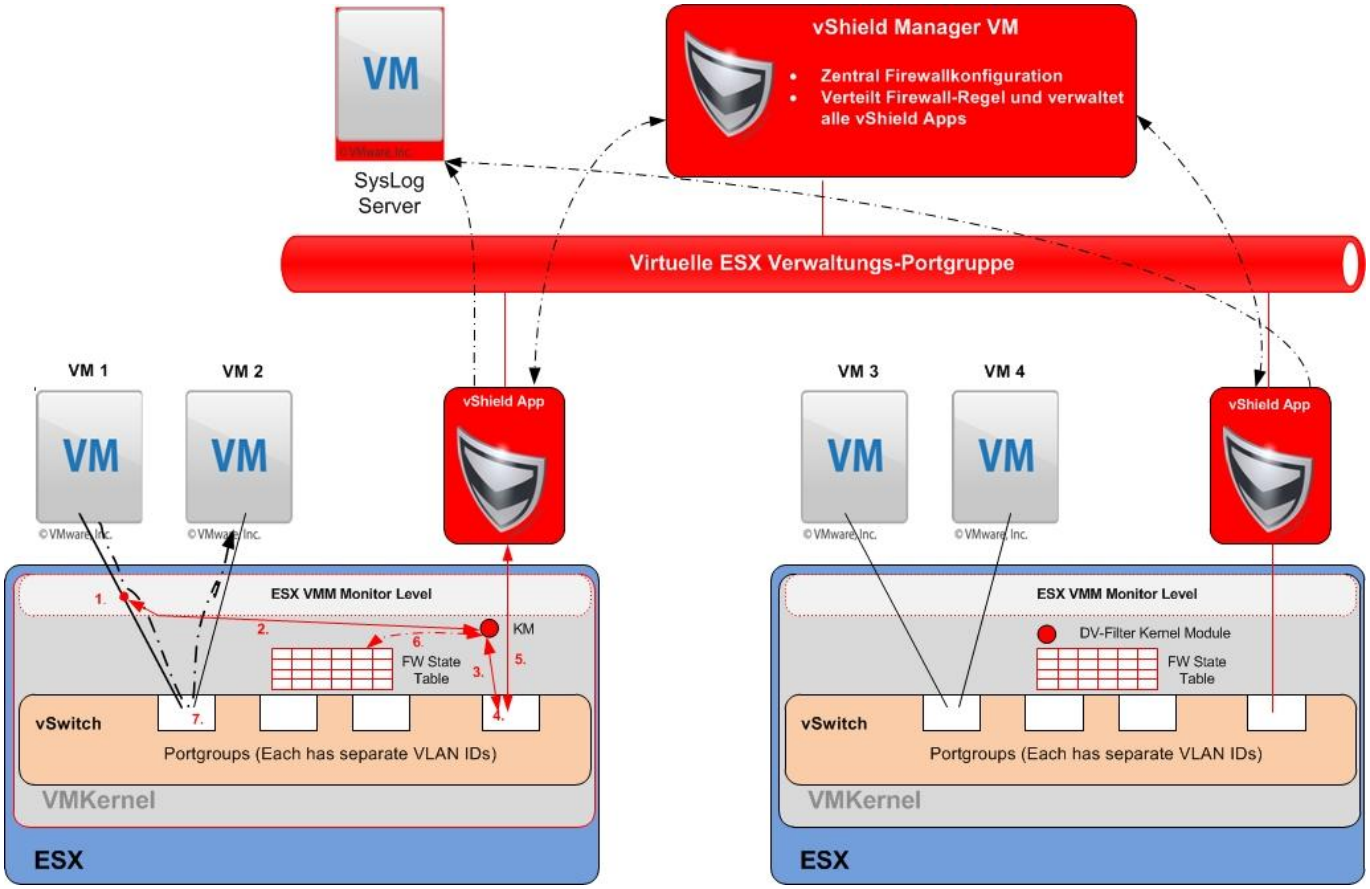


Abbildung 8: Einsatz von vShield App

Erklärungen zur obenstehenden Abbildung:

Schritt 1: VM1 möchte eine TCP/IP-Verbindung zu VM2 in derselben Portgruppe aufbauen, jedoch wird der Verbindungsaufbau auf dem VMM-Level abgefangen.

Schritt 2: Die Verbindung wird zur Überprüfung an das DV-Filter Module gesandt. Hier wird auch geprüft, ob es schon einen Eintrag für die TCP/IP-Verbindung in der State Table gibt.

Schritte 3-5: Existiert kein Eintrag, dann wird die Verbindung über die lokale Management Portgruppe zur vShield Appliance gesendet.

Schritt 6: Ist der Verkehr erlaubt, wird der Eintrag in die State Table geschrieben.

Schritt 7: Nun darf der Netzverkehr in den virtuellen Switch zur normalen Weiterverarbeitung gegeben werden; somit kann VM1 mit VM2 kommunizieren.
Weitere Pakete, die zu dieser TCP/IP-Session gehören, können über die lokale State Table geprüft werden.

Firewalls, die VMsafe verwenden, haben gegenüber privaten VLANs den Vorteil, dass granular entschieden werden kann, welcher Traffic zwischen Applikations-Zonen im gleichen VLAN / IP-Subnetz möglich sein soll, und welcher nicht.

3.3.2 VSAN

Virtual Storage Area Networks (VSANs) sind virtuelle Speichernetze, mit denen unabhängige logische Speichernetze in Fibre Channel Switches (FC-Switches) definiert werden können. Die Switch-Ports sind jeweils nur einem VSAN zugeordnet. Die einzelnen VSANs arbeiten vollkommen getrennt voneinander, sind skalierbar, verfügen über eigene Sicherheitskriterien und unterstützen eigene SAN-Services wie den Fibre-Channel Namensservice, Fabric Login, Zoning, etc.

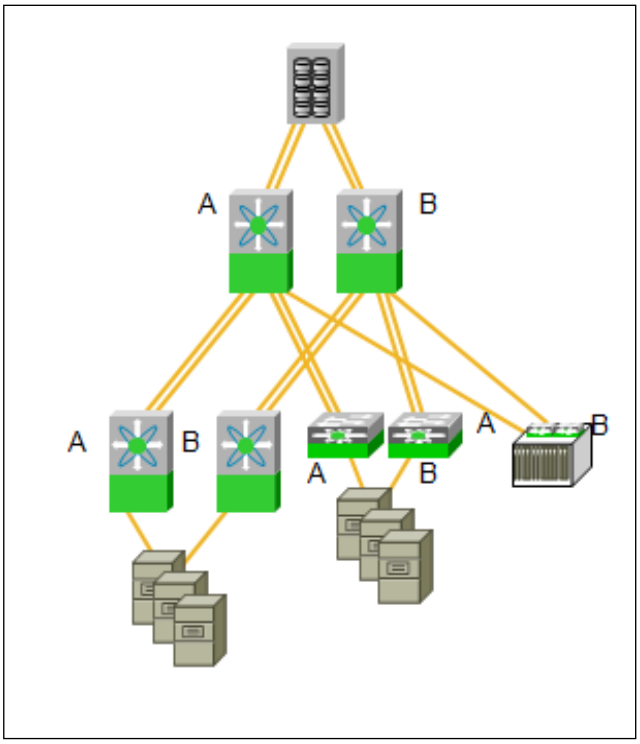


Abbildung 9: Fabrics

Im Vblock werden VSANs innerhalb des Fabric Interconnects verwendet, um zu entscheiden, in welche "Fabric" ein bestimmter virtueller Host Bus Adapter (vHBA) aufgenommen wird.

In Fibre Channel Umgebungen, wie in der Abbildung 9 dargestellt, werden üblicherweise zwei komplett getrennte Netze (Fabrics) aufgebaut (genannt Fabric A und Fabric B). Die Host Bus Adapter (HBAs) der Serversysteme werden so aufgeteilt, dass es von jedem Host zu jedem Target (Speicher-Subsystem) mindestens zwei Wege über zwei getrennte Netze (Fabrics) gibt.

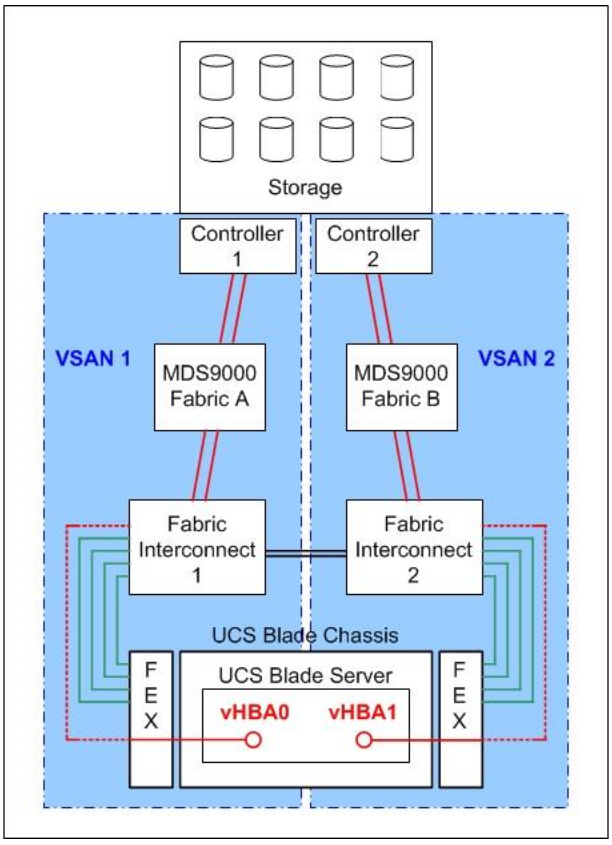


Abbildung 10: Fabrics im Vblock

Wie in der Abbildung 10 gezeigt, werden im Vblock ebenfalls zwei Fabrics A und B aufgebaut. Innerhalb des Vblocks werden zwei Cisco MDS 9000 SAN Switche eingesetzt, die die Fabrics A und B darstellen. Über diese MDS 9000 SAN Switche werden die Fabric Interconnects mit den EMC Storage Controllern verbunden.

VSANs werden im Vblock verwendet, um die vHBAs in den Blade Servern einer Fabric zuzuordnen. Üblicherweise werden verschiedene VSAN-IDs für Fabric A und B verwendet. Das VSAN ist dann einem Fibre Channel over Ethernet (FCoE)-VLAN zugeordnet, das zwischen den Adaptern in den Blades und den Fabric Interconnect verwendet wird. Auf diese Art hat jeder Blade Server jeweils einen Zugang zur Fabric A und B. Die Zuordnung kann dabei pro Adapter dynamisch erfolgen.

3.3.3 IEEE 802.1BR – Bridge Port Extension, VN-Link, vNICs und vHBAs

Im Gegensatz zu VLANs und VSANs, die bereits zum normalen Arbeitswerkzeug in Netzumgebungen gehören, ist 802.1BR eine Neuentwicklung. 802.1BR ist ein Projekt des IEEE unter dem Titel "Bridge Port Extensions" innerhalb der 802.1 Datacenter Bridging Task Group.

802.1BR führt einen neuen Ethernet Header ein. Dieser wird verwendet, um einen Port auf einem "*Port Extender*" von einer "*Controlling Bridge*" anzusprechen. Vereinfacht gesagt, handelt es sich bei dem Port Extender um eine "*Remote Line Card (entfernte Schnittstellenkarte)*" eines Switches.

802.1BR addressiert mehrere Probleme:

- In modernen Datacenter-Umgebungen wird versucht die Menge an Kupferkabeln zwischen Schränken über Deckenträger zu reduzieren. Die Doppelböden werden meist freigehalten, und die Kabelwege werden über Deckenträger geführt. Um die Brandlast und Deckenlast durch Kupferkabel zu verringern, werden meist in jedem Rack 24- und 48-Port-Switches eingebaut, die wiederum über LWL-Verkabelung an den Datacenter Distributionsbereich angebunden werden. Dieses führt zu einer Vielzahl von einzeln zu verwaltenden Switches. 802.1BR wird auch dazu verwendet "*remote linecards*", bei Cisco "Fabric Extender (FEX)" genannt, durch die zentrale Instanz der Controlling Bridge zu verwalten. Diese Fabric Extender werden anstatt der 24/48 Port Switche in die Racks eingebaut. Damit wird die Anzahl von Verwaltungspunkten im Datacenter sehr stark verringert.
- Durch Blade Chassis und Server "explodiert" ebenfalls die Anzahl von einzeln zu verwaltenden Switches. Jedes Blade Chassis hat ein bis zwei eigene Switches; bei ca. sechs Blade Chassis pro Rack kann hochgerechnet werden, wie viele Switches verwaltet werden müssen. 802.1BR wird dazu verwendet jedem internen Ethernet Port zu den Blade Servern einen virtuellen Ethernet Port zu geben, der durch die Controlling Bridge verwaltet wird.
- Oftmals ist es notwendig einem Server / ESX-Hypervisor mehrere Ethernet Adapter zu geben, die wiederum z.B. zur Trennung von Mandanten verwendet werden, sowie zur Trennung von Management und Mandanten Traffic. Über spezielle Adaptertechnologien können den Servern / ESX Hypervisor eine Vielzahl von virtuellen Adaptern gegeben werden. Jeder virtuelle Adapter ist wiederum auf der Controlling Bridge einem virtuellen Ethernet Ports (vEth), bzw. virtuellen HBA (vHBA) zugeordnet.
- VM-Gäste sind mobil; sie können von einem ESX-Hypervisor zu einem anderen verschoben werden. In der "realen Welt" kann ein Endsystem anhand seines Ethernet Access Ports festgemacht werden. Es können Regeln (ACLs, QoS, etc.) an dem Ethernet Access Ports definiert werden. 802.1BR ermöglicht es, dass ein oder mehrere virtuelle Ethernet Ports an der VM anhand ihrer UUID (Universal Unique Indentifier) festgemacht werden und von ESX-Host zu ESX-Host mitwandern.

In der nachfolgenden Abbildung 11 werden diese Optionen dargestellt.

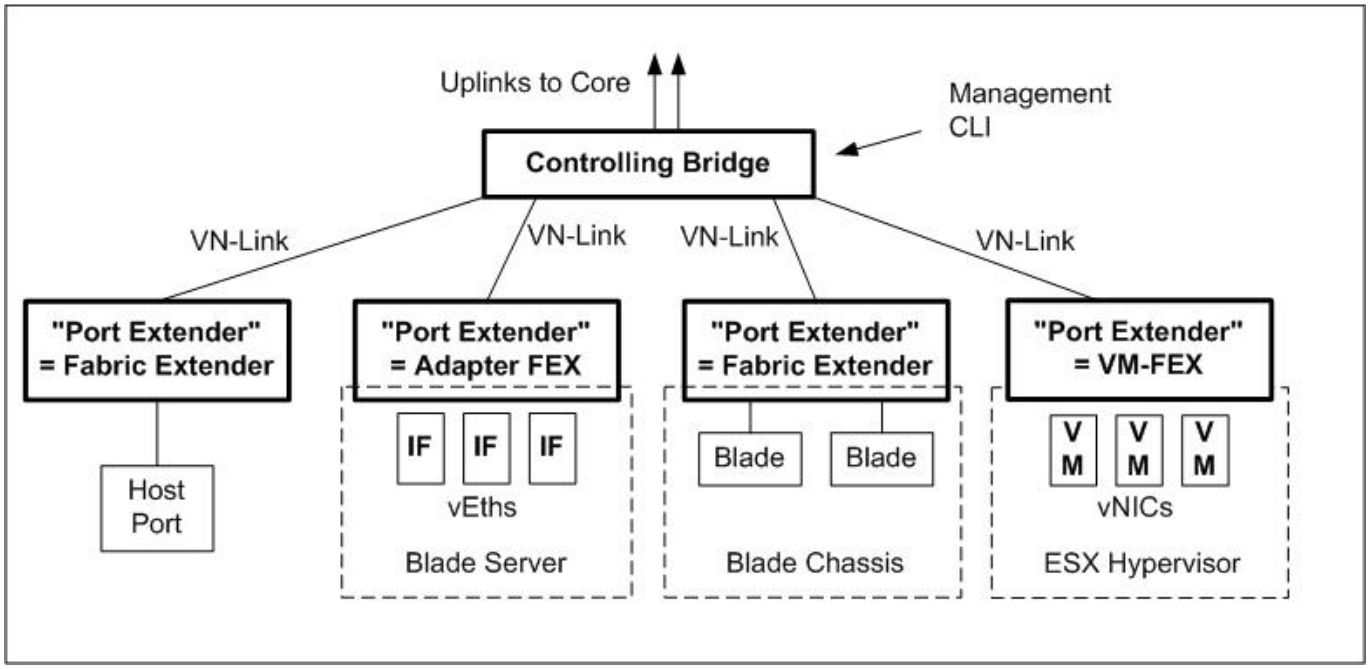


Abbildung 11: VN-Link Technologie

Im Vblock wird die VN-Link Technologie, eine Cisco proprietäre Vorab-Version des Header-Formates von 802.1BR, eingesetzt. Die grundsätzlichen Management-Modelle von 802.1BR und VN-Link sind aber gleich.

An zwei Stellen ist Cisco VN-Link im Einsatz:

Erstens wird VN-Link auf der Verbindung zwischen dem "Fabric Extender" im UCS Blade Server Chassis, und dem UCS Fabric Interconnect eingesetzt (siehe nachfolgende Abbildung 12).

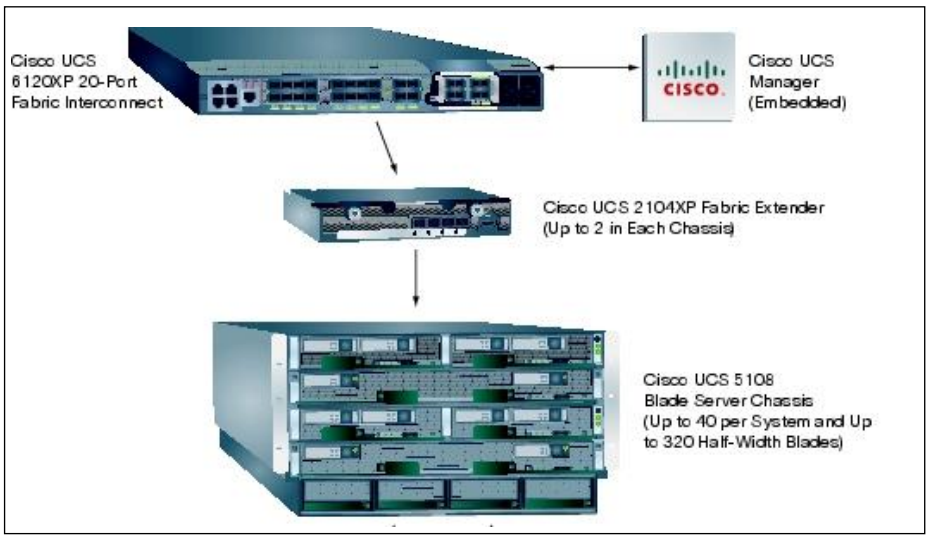


Abbildung 12: VN-Link zwischen Fabric Extender und UCS Fabric Interconnect

Durch VN-Link wird jeder 10GE Port auf der Backplane des Blade-Chassis als virtueller 10GE Port auf dem Fabric Interconnect geführt. Jeder der derzeit zwei Fabric Extender im Blade Chassis ist eine "Remote Line Card" des Fabric Interconnect. Der Vorteil ist, dass die Fabric-Extender im Chassis nicht als separater Switch verwaltet werden müssen, und dass die gesamte Kontrolle des Netzverkehrs im Fabric Interconnect geschieht.

Zweitens wird VN-Link auf der Verbindung zwischen der Cisco UCS M81KR Virtual Interface Card und dem Fabric Interconnect eingesetzt (siehe Abbildung 13). Die UCS M81KR Vitual Interface Card wird als

Aufsteck-Karte (mezzanine card) im Blade Server installiert, und bietet dem Server derzeit bis zu 56 virtuelle PCIe Ethernet Netzkarten und FC-HBAs (theoretisch sind 128 virtuelle Interfaces möglich; die genaue Anzahl ist vom Betriebssystem abhängig). Jeder virtuelle PCIe-Adapter kann entweder als Ethernet Netzkarten oder FC-HBA konfiguriert werden. Dabei handelt es sich aber nicht um eine SR-IOV-Implementierung (Single-Root I/O Virtualization), wie sie von der PCI-SIG-Working Group standardisiert wird, sondern die Karte stellt sich dem Host als Standard PCIe-Riser mit multiplen Karten dar. Diese Option wird bei Cisco als "Adapter Fabric Extender (Adapter-FEX)" bezeichnet.

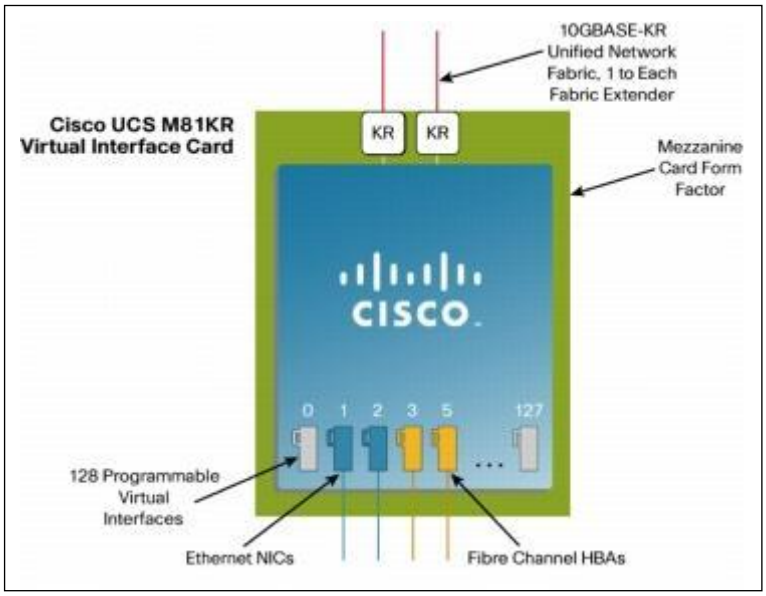


Abbildung 13: VN-Link zwischen UCS M81KR Virtual Interface Card und UCS Fabric Interconnect

Die Karte verbindet ihre Uplink Ports über die Backplane des Chassis mit dem Fabric-Extender, und damit mit dem Fabric-Interconnect. Durch die Verwendung von VN-Tag wird jede virtuelle Netzkarte/HBA und der Port auf dem Fabric-Extender als virtueller Port auf dem Fabric Interconnect geführt.

Wie bereits in Kapitel 3.3.1 dargestellt, werden häufig VLANs eingesetzt, um den Netzverkehr zwischen Mandanten vom VMware ESX Hypervisor bis zum Ausstieg aus dem Datacenter-Netz getrennt zu transportieren. Die UCS M81KR Virtual Interface Card bietet eine Möglichkeit, dem ESX Hypervisor pro Mandant/Guest eigene Netzkarten zu liefern, und den Netzverkehr dann auf dem Fabric Interconnect in VLANs zu führen.

Im Vblock wird zum Zeitpunkt der Erstellung dieses Dokumentes VN-Link zu den VMs (bei Cisco als "VM-FEX" bezeichnet) noch nicht eingesetzt, wird jedoch in den nächsten Releasezyklen auch im Vblock eingesetzt werden.

Zum Verwalten der Ethernet Ports der VMs steht aber die optionale Komponente Cisco Nexus 1000v zur Verfügung, die ebenfalls virtuelle Ethernet Ports zu VM-Gästen, die an die UUID gebunden sind, anbietet. Nexus 1000v wird im Kapitel 4.1.3.3 vorgestellt.

3.4 Unified Fabric

Im Vblock wird, anders als bei herkömmlichen LAN- und SAN-Verkabelungen, das Konzept der "Unified Fabric" verfolgt. Damit ist die Konvergenz von LAN (Ethernet) und SAN (Fibre Channel) auf eine gemeinsame Anschlussmethode gemeint.

Dadurch werden Kosten eingespart sowohl durch Reduzierung der Verkabelung wie auch durch Reduzierung von Ports sowohl im LAN- wie auch im SAN-Bereich.

Hierbei werden die Entwicklungen im Bereich Fibre Channel over Ethernet (FCoE) und Data Center Bridging (DCB) verwendet, um LAN- und SAN-Traffic auf einen 10GE Ethernet Link zu vereinen.

3.4.1 Fibre Channel over Ethernet (FCoE)

FCoE wurde im Fibre Channel Backbone 5 (FC-BB-5) Projekt des T11 Komitees des INCITS (InterNational Committee for Information Technology Standards) spezifiziert. Darin wurden (vereinfacht gesagt) zwei Hauptthemen für FCoE definiert:

1. Der Transport von Fibre Channel (FC) Frames innerhalb eines Ethernet Frames, und die damit verbundenen "Mappings" von FC-Bestandteilen wie Adressen und Control-Informationen in Ethernet Header Informationen.
2. Die Initialisierung eines virtuellen FC-Links durch eine Ethernet-"Wolke", über das FCoE Initialization Protocol (FIP).

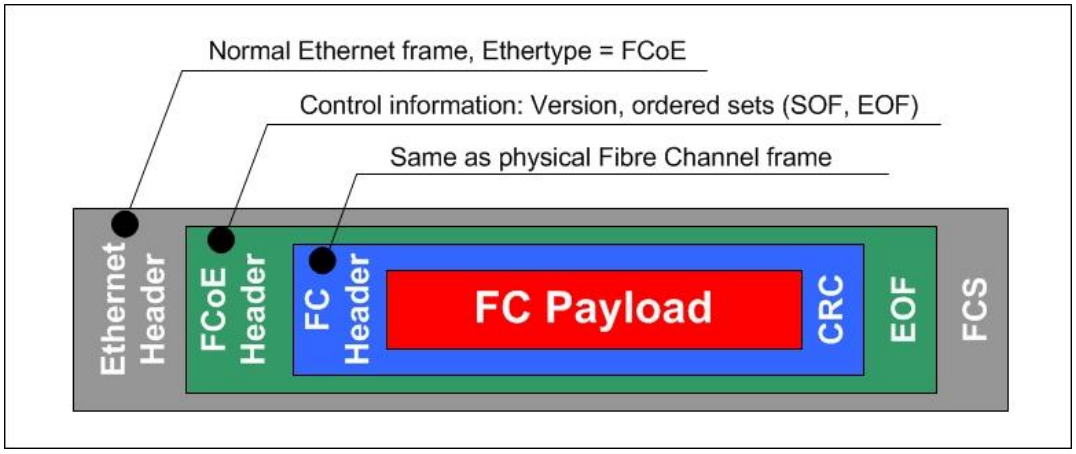


Abbildung 14: Fibre Channel Frames innerhalb eines Ethernet Frames

FC-BB-5 spezifiziert, dass der Transport von FC innerhalb von Ethernet verlustfrei (lossless) erfolgen muss. "Lossless" bedeutet in diesem Fall nicht völlig fehlerfrei, denn FC erlaubt eine Bit-Fehlerrate (Bit Error Rate) von 10^{-12}.

Lossless bedeutet aber, dass ein Frame z.B. nicht durch einen volllaufenden Puffer in einem Netzgerät verloren gehen darf. FC-BB-5 macht aber keine konkreten Vorgaben, wie dieses zu erfolgen hat. Diese Problemstellung wurde von der Data Center Bridging (DCB) Task Group des IEEE aufgegriffen.

3.4.2 Data Center Bridging (DCB)

Innerhalb des IEEE 802.1 wurde die DCB Task Group gegründet, um Ethernet so zu verbessern, dass es den Ansprüchen und Anforderungen im Rechenzentrum Rechnung trägt. Als eine der Kernanforderung wird die Verlustfreiheit in Ethernet angestrebt ("Lossless Ethernet"). Im Kern des Projektes stehen zwei neue Standards, die für FCoE entscheidend sind: Priority-based flow control (IEEE 802.1Qbb) und Enhanced Transmission Selection (IEEE 802.1Qau). Diese werden im Folgenden näher beschrieben.

<u>802.1Qbb: Priority based Flow Control (PFC)</u>

Auf normalen 1 Gigabit und 10 Gigabit Ethernet Verbindungen gibt es schon seit langer Zeit den PAUSE Mechanismus. Dabei teilt der empfangende Switch dem sendenden Switch mit, dass seine Empfangspuffer (buffer) volllaufen, und instruiert den sendenden Switch für eine gewisse Zeit das Senden einzustellen (siehe Abbildung 15). Dieser PAUSE Mechanismus gilt unabhängig von der geförderten Quality of Service (QoS) aber für alle Verkehrsklassen auf der Verbindung (z.B. RIP, OSPF Updates, Video, Audio, etc.).

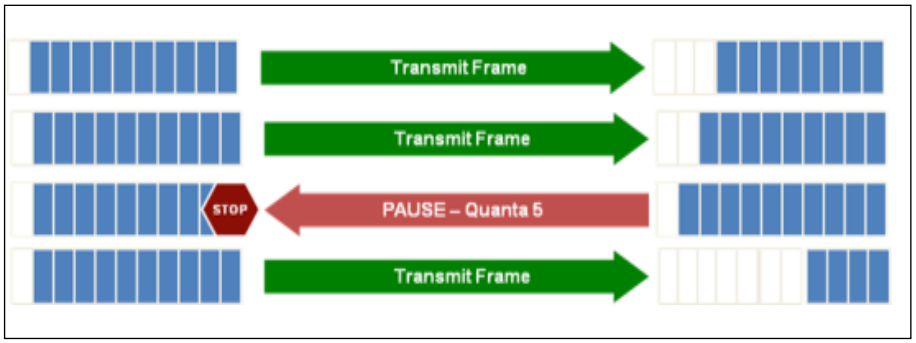


Abbildung 15: PAUSE Mechanismus für alle Verkehrsklassen

802.1Qbb – PFC erweitert den PAUSE Mechanismus derart, dass eine individuelle PAUSE pro QoS-Klasse gesendet wird. Dabei werden die acht QoS-Klassen (CoS) verwendet, die im 802.1p Standard für die Verwendung auf 802.1Q VLAN-Trunks verwendet werden (dieser Mechanismus ist für eine Verkehrklasse in der nachfolgenden Abbildung 16 dargestellt).

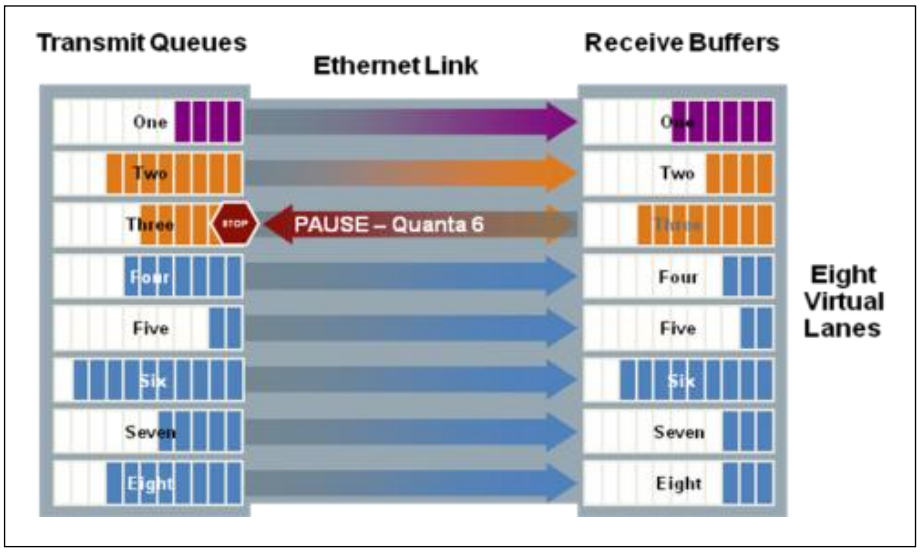


Abbildung 16: PAUSE Mechanismus für eine Verkehrsklasse

Auf diese Art können für bestimmte Verkehrsklassen, z. B. der Klasse 3, in der FCoE transportiert wird, ein "lossless" Verhalten eingeschaltet werden. Andere Verkehrsklassen, in denen hauptsächlich normaler IP-Traffic transportiert wird, sind auf Frame-/Paketverluste optimiert (z.B. TCP-Flusskontrolle), und arbeiten besser, wenn kein PAUSE Mechanismus für diese Klassen eingeschaltet ist.

802.1Qaz: Enhanced Transmission Selection (ETS), inkl. Data Center Bridging Exchange (DCBX)

Enhanced Transmission Selection (ETS) führt ein festes QoS-Queuing ein, das auf acht Deficit Weighted Round Robin (DWRR) plus einer Strict Priority Queue basiert. Die Strict Priority Queue wird hierbei immer als erste geleert, die anderen Queues werden nach ihrer Gewichtung geleert. Dieses Queuing setzt aber erst ein, wenn die Verbindung ausgelastet ist.

Die nachfolgende Abbildung 17 zeigt beispielhaft dieses Verhalten. Insgesamt steht eine Bandbreite von 10 Gbit/s zur Verfügung.

- Zum Zeitpunkt t1 können die garantierten Bandbreiten genutzt werden (keine volle Ausnutzung).
- Zum Zeitpunkt t2 wird die verbleibende Bandbreite durch den LAN Traffic ausgenutzt.
- Zum Zeitpunkt t3, wird die vom HPC Traffic frei werdende Bandbreite durch den LAN Traffic ausgenutzt.

- Zum Zeitpunkt t4 hingegen wird die Bandbreite für den LAN Traffic wieder reduziert, um die garantierte Bandbreite für den HPC Traffic wieder zu erreichen.

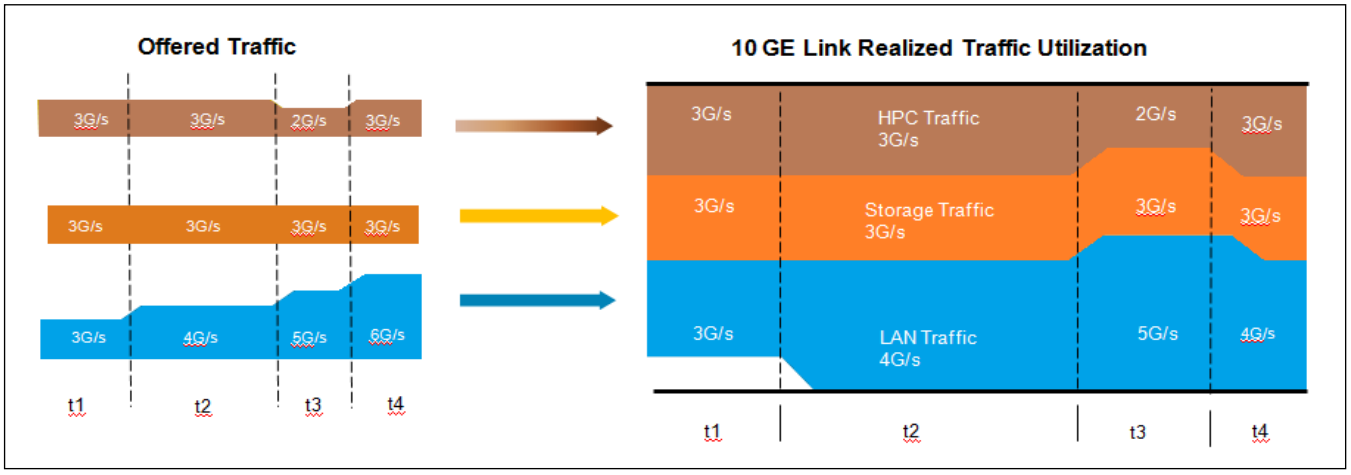


Abbildung 17: QoS-Queuing mit Enhanced Transmission Selection (ETS)

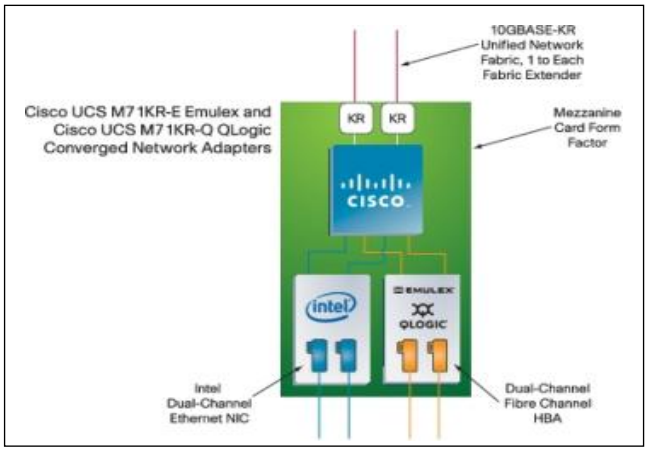
Dieses Queuing Schema ist dem Class Based Weighted Fair Queuing auf Cisco Platformen sehr ähnlich. ETS harmonisiert die bestehenden Queuing Schemata und erweitert diese bis in den Netzadapter im Server hinein.

Datacenter Bridging Exchange (DCBX) nutzt das Link Layer Dicovery Protocol (LLDP), um PFC- und ETS-Parameter (wie z.B. Queuing Einstellungen, oder Einstellungen, welche Priorität "lossless" sein soll, etc.) zwischen den Teilnehmern in dem Netz abzustimmen. Z.B. werden die Parameter zwischen den Converged Network Adaptern (CNA) in den Servern und den Switchen ausgetauscht.

DCBX wird auch in Multihop FCoE-Topologien verwendet, wenn FCoE-Switche (sogenannte Fibre Channel Forwarder), über Ethernet Verbindungen miteinander verbunden werden. Über DCBX werden unterstützte Merkmale sowie Queuing Strukturen, PFC-Parameter, etc. ausgetauscht. Erst wenn sich die Switche auf ein gemeinsames "Vorgehen" einigen können, wird der Link als "up" deklariert.

3.4.3 Converged Network Adapter (CNA)

Unter einen Converged Network Adapter (CNA) verstehen wir eine PCIe-Karte für Blade-Serversysteme, die sich auf der PCI-Bus Seite sowohl als Dual Fibre Channel Host Bus Adapter (HBA), sowie als Dual Ethernet Karte darstellt. Innerhalb des Adapters wird FC und Ethernet Verkehr in die 10GE-Verbindung "gemultiplexed".



Im Vblock hat die Karte, wie in Abbildung 18 dargestellt, nach außen zwei Verbindungen zur Chassis Backplane, die sie wiederum zu zwei Fabric Extender (Port Extender) führt. Die Fabric Extender sind dann über jeweils 4x 10GE mit dem Fabric-Interconnect verbunden.

Abbildung 18: Converged Network Adapter

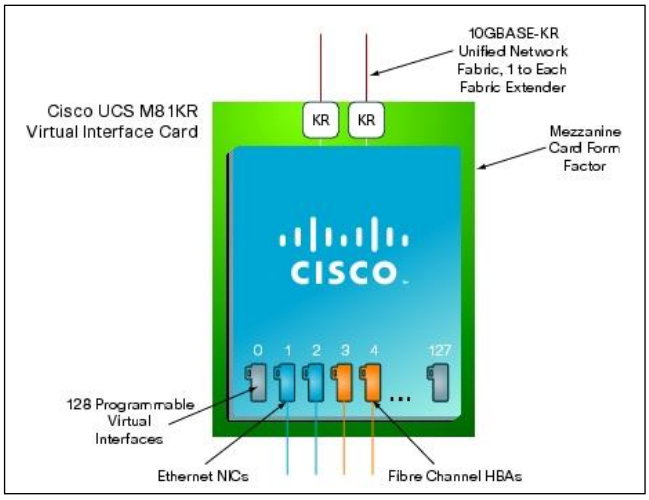


Abbildung 19: Cisco UCS M81KR Virtual Interface Card

Der bereits erwähnte Cisco UCS M81KR geht sogar weiter, und bietet dem Hypervisor über den PCI-Bus bis zu 128 (derzeit 56) dynamisch konfigurierbare virtuelle Interfaces (siehe Abbildung 19). Jedes dieser Interfaces kann sowohl als Virtual HBA (vHBA) oder Virtual Ethernet (vEth) konfiguriert werden.

Der CNA hat nicht nur die Aufgabe die Encapsulation von FC in Ethernet vorzunehmen, sondern die Karte muss auch die virtuellen FC-Links durch das **Ethernet Netz** über das FIP-Protokoll initialisieren. Hierzu hat sie einen (logischen) FCoE-Controller, der für das Verarbeiten der DCBX- und FIP-Nachrichten zuständig ist.

Ein mit einem CNA ausgestattetes System wird im FCoE auch als End-Node (Enode) bezeichnet.

3.4.4 Fibre Channel Forwarder (FCF)

Als *Fibre Channel Forwarder* (FCF) verstehen wir ein Fibre Channel Switching Element (Fibre Channel Switch) der mit einer "lossless" Ethernet Bridge (Switch) kombiniert ist (siehe Abbildung 20).

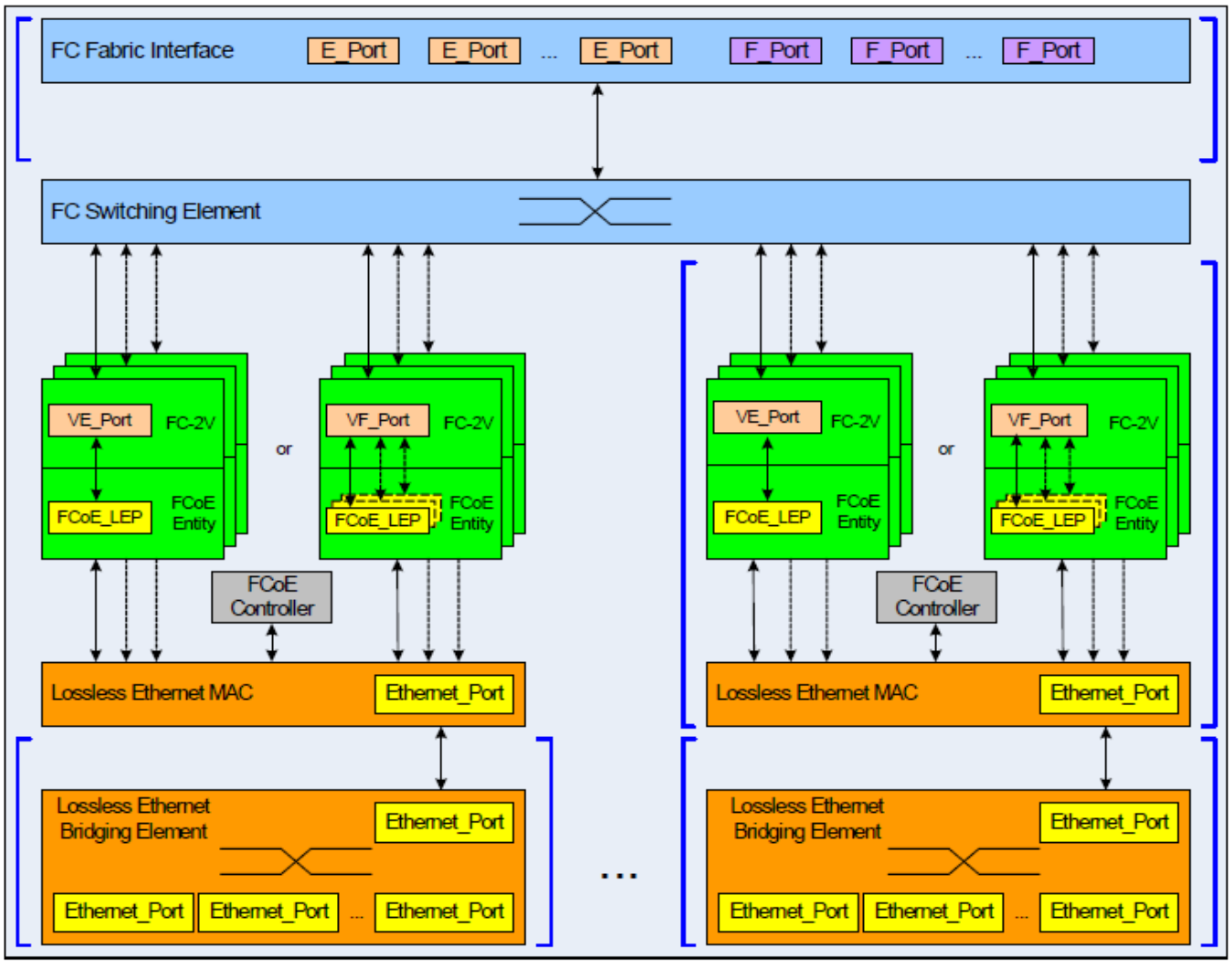


Abbildung 20: Fibre Channel Forwarder (FCF)

Der FCF kann dabei sowohl als Bridge zwischen FCoE- und klassischen FC-Netzen verwendet werden, sowie als FCoE-zu-FCoE-Switch. Dabei ist entscheidend, zu verstehen, dass aus FC-Sicht innerhalb des FCF kein Unterschied zwischen einem FC-Port und einem FCoE-Port besteht.

In klassischen FC Umgebungen werden Ports typisiert anhand der Funktion, die sie erfüllen (siehe die nachfolgende Abbildung 21):

- ***N-Port***:
  Der *Node Port* ist der Port an einem End-Node, wie z.B. einem HBA in einem Host oder Storage Controller.
  Der *NP-Port* ist ein spezieller N-Port, der sich beim FC Switch als multiple Nodes anmeldet.
- ***F-Port***:
  Der *Fabric Port* ist der Port an einem FC-Switch, an dem ein N-Port angeschlossen wird.
- ***E-Port***:
  Der *Expansion Port* ist ein Port an einem FC-Switch an dem ein anderer FC-Switch angeschlossen wird. Die Switche bilden dann eine "Fabric".
- ***TE-Port***:
  Der *Trunking E-Port* ist ein spezieller E-Port, auf dem der Verkehr von multiplen VSANs zwischen FC Switches transportiert wird.

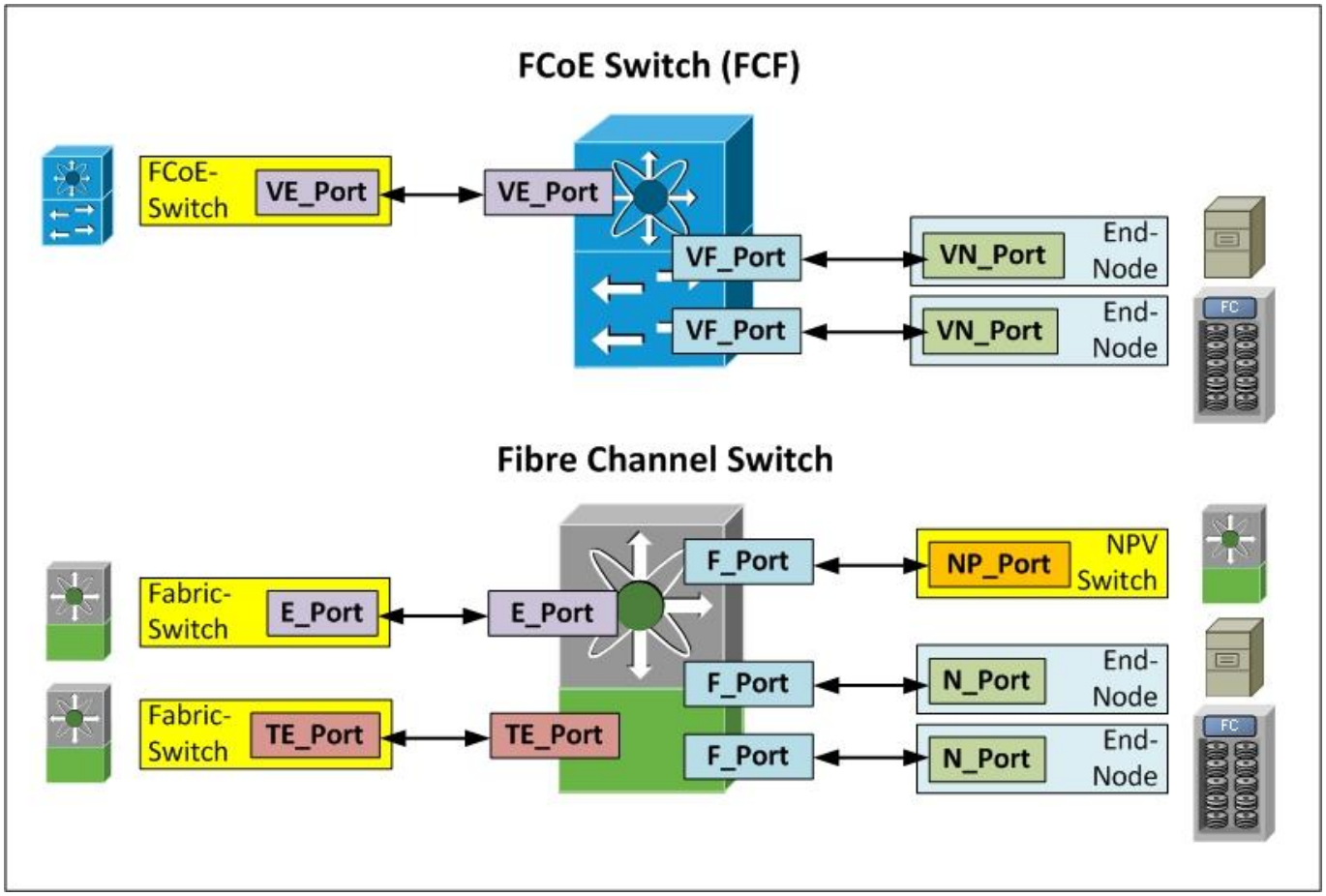


Abbildung 21: Porttypen bei FC- und FCoE-Switchen

Analog zu den klassischen FC Port-Typen werden bei FCoE die Ports wie folgt bezeichnet:

- ***VN-Port***:
  Der *Virtual N-Port* ist der Port eines CNAs (ENode), oder genauer gesagt des FC Anteils der CNAs.
- ***VF-Port***:
  Der *Virtual F-Port* ist die virtuelle Abbildung des F-Ports innerhalb eines FCF. D.h. er ist der F-Port

am Fibre-Channel Switching Element des FCF, der die virtuelle FC-Verbindung zwischen dem CNA und dem FCF durch das Ethernet Netz darstellt.

- ***VE-Port***:
  Wenn zwei FCF miteinander über eine Ethernet-Verbindung gekoppelt werden, verwendet die FCF *Virtual E-Ports*. Bei einem ***Virtual TE Port*** werden dann auch mehrere VSANs transportiert.

Der FCF verfügt wie der CNA über einen (logischen) FCoE-Controller, auf dem das FIP-Protocol verarbeitet wird, und der mit den CNAs FIP-Nachrichten austauscht, um den virtuellen VN zu VF Port Link durch ein "lossless" Ethernet Netz aufzubauen.

3.4.5 FCoE Initialization Protocol (FIP)

In Fibre Channel ist die "State Maschine" stark abhängig vom Zustand des physischen Links. Wenn der physische Link zwischen dem N- und F-Port (HBA zu FC-Switch) aktiv wird, sendet der HBA zum FC-Switch als erstes einen Fabric Login (**`FLOGI`**), bei dem u.a. auch dem HBA die FC-ID vom FC-Switch vergeben wird. Anschließend wird eine Registrierung beim Name Server auf dem FC-Switch durchgeführt. Die Details hierzu sind nicht Teil dieses Dokumentes, da dieses Dokument sich hauptsächlich mit den Unterschieden zu klassischen Umgebungen befasst.

Wenn der physische Link inaktiv wird, wird auch der Name Server Eintrag gelöscht, und sogenannte State Change Notifications zu den anderen Teilnehmern im FC-Netz geschickt, etc.

In FCoE-Netzen ist es möglich, dass die Verbindung zwischen dem CNA (ENode) und dem FCF über mehrere "lossless" Ethernet Bridges (Switches) aufgebaut wird. Daher kann die "State Machine" nicht am Status des physischen Links angehängt werden. Dies ist einer der Gründe warum FIP eingeführt wurde.

Um nun einen virtuellen FC-Link aufzubauen, tauschen die FCoE-Controller im CNA (ENode) und FCF FIP-Nachrichten über einen eigenen Ethertype 0x8914h aus (siehe Abbildung 22).

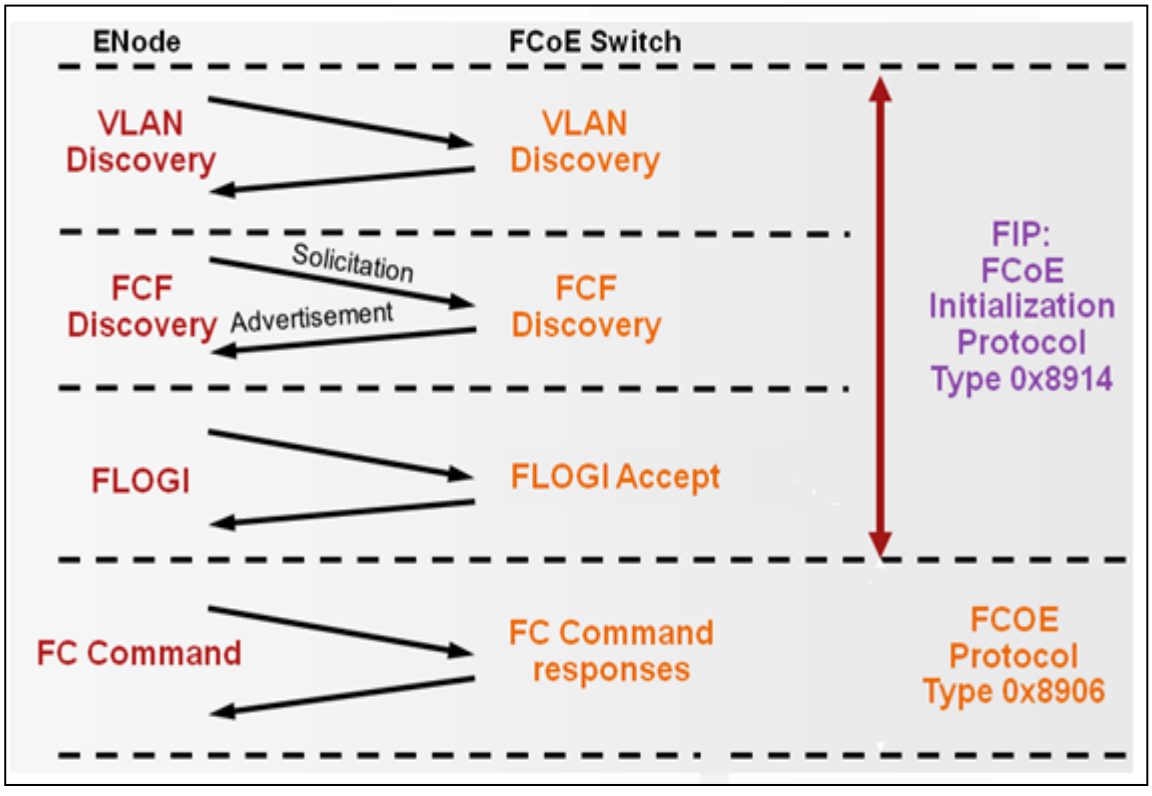


Abbildung 22: FIP-Nachrichten

1. VLAN Discovery:
   Als erste FIP-Nachricht wird ein **`VLAN-Request`** vom CNA (ENode) an die Ethernet Multicast Adresse ALL_FCF_MACs (01-10-18-01-00-02) gesendet. Dieser Request wird im "native" VLAN

des CNAs (ENode) geschickt, d.h. auf dem Ethernet Interface ohne einen 802.1Q VLAN-Tag zu verwenden. Der/die FCFs senden dann eine Liste von VLANs zurück, auf denen sie FCoE unterstützen. Der Administrator vom CNA (ENode) kann dann auswählen oder automatisch auswählen lassen, welches dieser VLANs verwendet wird. Üblicherweise wird hier sichergestellt, dass nur ein VLAN zurückgegeben wird, so dass die CNAs (ENodes) nur das gültige FCoE VLAN auswählen.

2. FIP Discovery:
   Über FIP Advertisement und Solicitation Nachrichten "finden" sich CNAs (ENodes) und FCFs. FCFs senden in regelmäßigen Abständen FIP-Advertisements an die Ethernet Multicast Adresse ALL_ENode_MACs. Die CNAs (ENodes), bzw. deren FCoE-Controller die diese FIP-Advertisements empfangen, bilden sich dann eine Liste von FCFs, die für sie erreichbar sind.
   Um nicht auf dieses Advertisement warten zu müssen, werden CNAs (ENodes) auch eine **`FIP-Solicitation Message`** an die Ethernet Multicast Adresse ALL_FCF_MACs schicken, wenn ihr Interface aktiv wird. Alle FCFs, die diese Nachricht empfangen, antworten dann mit einem Unicast Advertisement direkt zum CNA (ENode). In diesen Advertisement werden auch Prioritäten für die FCFs angegeben. Innerhalb des Advertisements werden auch Parameter abgestimmt, wie häufig FIP-Multicast Advertisements als "Keep-Alives" zu erwarten sind. Wenn zwei Advertisements in der abgestimmten Periode nicht empfangen werden, wird der virtuelle Link als "Down" deklariert. Es kann über einen Flag aber auch angegeben werden, dass es sich um eine Punkt-zu-Punkt Ethernet Verbindung handelt, und daher keine Keep-Alives gesendet und empfangen werden müssen.
3. FIP Virtual Link Instantiation
   Nachdem sich CNAs (ENodes) und FCFs "gefunden" haben, wird der virtuelle Link aufgebaut (instantiated). Hierbei handelt es sich um in FIP transportierte **`FC-FLOGI`** Nachrichten. Hierbei werden den CNAs (ENodes) auch MAC-Adressen vergeben, die fortan als Quell-MAC-Adressen beim versenden von FCoE-Daten verwendet werden.
4. Nachdem der Link aufgebaut ist, werden "normale" **`FC-Command/Response`** Nachrichten, in Ethernet verpackt, vom CNA (ENode) über den FCF mit anderen FC Zielen ausgetauscht. Hierbei wird fortan aber nicht mehr der Ethertype von FIP verwendet, sondern der von FCoE (0x8906h).
5. Der virtuelle Link kann entweder über eine in FIP eingepackte Fabric Logout (**`FLOGO)`** Nachricht abgebaut werden, oder wenn zwei FIP-Advertisements vom FCF nicht mehr beim CNA (ENode) empfangen werden.

Im Vblock gibt es, wie weiter unten erläutert, nur Punkt-zu-Punkt FCoE-Links; daher ist die Behandlung von FIP, z.B. was Keep-Alives angeht, sehr viel einfacher als bei FCoE über multiple "lossless" Ethernet Switche.

An dieser Stelle ist auch FIP-Snooping zu erwähnen. Um mehr Sicherheit für FCoE über multiple "lossless" Ethernet Switche zu bieten, wurde die Funktion von FIP-Snooping in "lossless" Ethernet Switche eingeführt. FIP-Sooping funktioniert wie folgt:

1. Für alle FCoE-"Access-Ports" gilt eine Access-Control-Liste, die Frames mit dem Ethernet Type von FIP 0x8914h nur zu speziell konfigurierten Ports erlaubt, die zu FCFs führen. Dadurch wird sichergestellt, dass kein CNA (ENode) von einem anderen CNA (ENode) FIP-VLAN-Requests und FIP-Solicitation Nachrichten empfängt.
2. Über die FIP-Snooping Funktion wird die FIP-Link-Instantiation abgehört. Bei einer erfolgreichen FIP-Instantiation zwischen einem CNA (ENode) und einem FCF wird dann eine neue Access-Control-Liste eingefügt, die FCoE-Verkehr mit dem Ethertype 0x8906h zwischen dem CNA (ENode), und dem FCF erlaubt.

Diese Maßnahmen sollen die Sicherheit von FCoE, wenn es über multiple Ethernet Switches geführt wird, erhöhen. Da im Vblock aber nur Punkt-zu-Punkt FCoE-Links zwischen den CNAs (ENodes) und dem Fabric-Interconnect bestehen, wird FIP-Snooping nicht benötigt und auch nicht eingesetzt. Der Fabric Interconnect filtert die FIP-Nachrichten zwischen seinen Ports und leitet diese nicht weiter.

3.4.6 NPV und N-Port ID Virtualization (NPIV)

Bevor wir zu einer zusammenfassenden Übersicht über "Unified Fabric" im Vblock kommen, muss noch die N-Port ID Virtualisierung betrachtet werden (siehe Abbildung 23).

Wie schon erwähnt, melden sich in FC-Netzen HBAs über den Fabric Login (**FLOGI**) am FC-Switch an, tauschen Parameter aus, und erhalten ihre 24 Bit FC-ID. Danach melden sich die Hosts beim Fibre Channel Names Service auf dem FC-Switch an, registrieren sich für State Changes, etc.

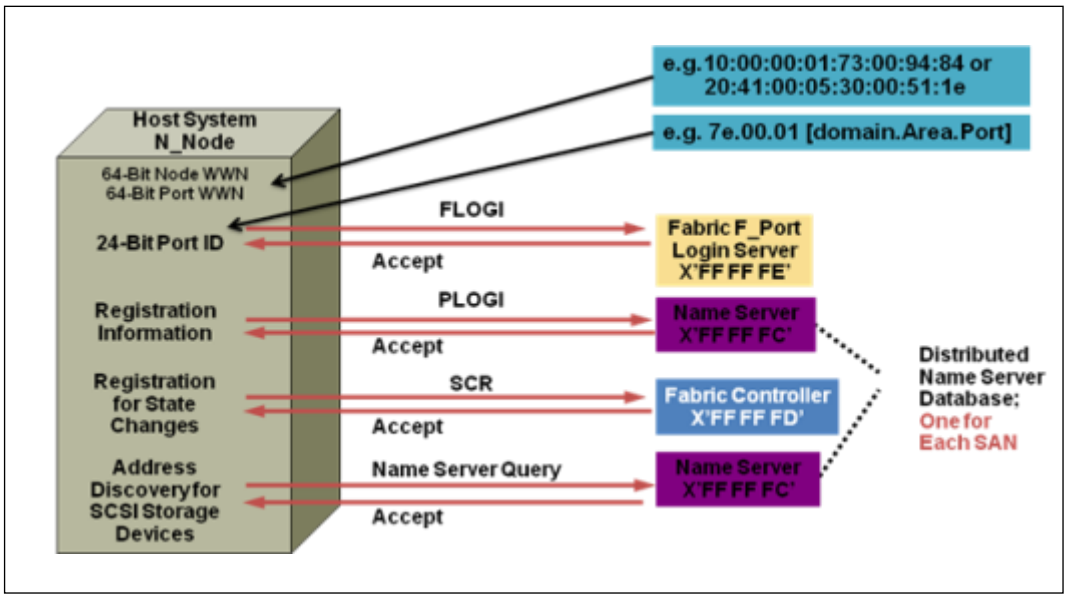


Abbildung 23: N-Port ID Virtualization (1)

NPIV, das ebenfalls ein T11-Standard ist, ermöglicht es einem N-Port multiple **FCIDs** zu bekommen. Der erste **FLOGI** initialisiert den Link, anschließend werden sogenannte Fabric Discovery (**FDISC**) Nachrichten gesendet, um weitere **FCID** auf demselben Port zu erhalten. Dies ist in Abbildung 24 dargestellt.

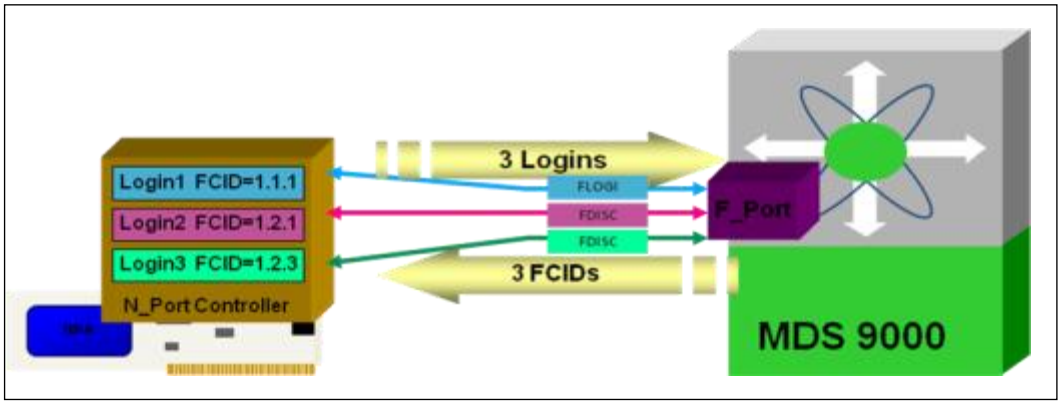


Abbildung 24: N-Port ID Virtualization (2)

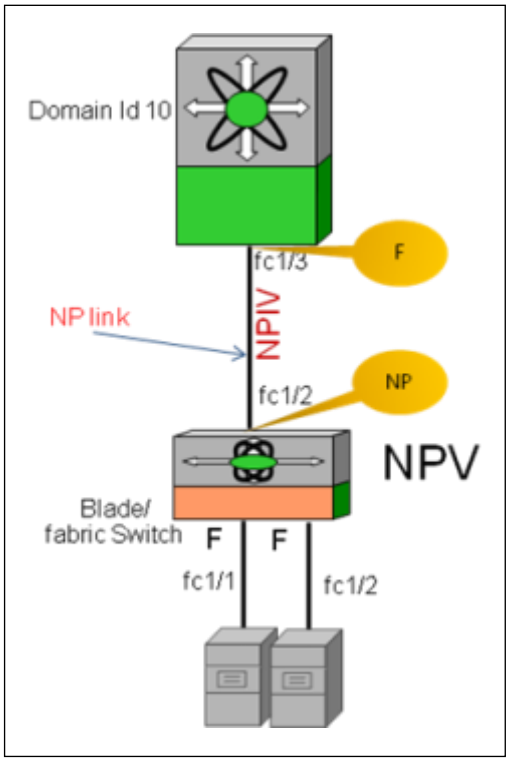


Abbildung 25: NPV-Switch

NPIV kann auch von Fibre Channel Switches verwendet werden. In diesem Fall kann der Fibre Channel Switch aus dem Switching Mode in den NPV-Mode (N-Port Virtualization) umgeschaltet werden.

Der NPV-Mode hat den Vorteil, dass der NPV-Switch nicht Teil der FC-Fabric wird. In einer FC-Fabric erhält jeder Switch eine Domain-ID. Diese ID kann im Bereich 1 - 239 liegen. Allerdings wird empfohlen, in einer Fabric nicht mehr als 40 Domain-IDs zu haben. Dies limitiert die Skalierungsmöglichkeiten in einer FC-Fabric recht stark, gerade wenn alle Switche in einem Blade Server Chassis und "Top of Rack" Switche in größerer Zahl eingesetzt werden.

Mit NPV verhält sich der Switch wie ein NPIV-Host. D.h. wenn der Link initialisiert wird, sendet der Switch einen **`FLOGI`** und erhält eine **`FCID`** (siehe Abbildung 25). Die an dem NPV-Switch angebundenen Hosts senden dann bei der Initialisierung ihres Links selbst einen **`FLOGI`** zum NPV-Switch. Dieser **`FLOGI`** wird vom NPV-Switch zu einem **`FDISC`** gewandelt, und auf dem NP Port zum upstream FC-Switch gesendet. Dieser vergibt dann eine **`FCID`**, die vom NPV-Switch an den Host weitergeleitet wird.

3.4.7 Unified Fabric im Vblock

Im Vblock werden in jedem Blade Server CNAs eingebaut. Jeder CNA (1x in den Halb-Breiten Blades (B200, B210), 2x in den Voll-Breiten Blades (B250, B440)) ist über die Chassis Backplane per 10GBASE-KR mit jedem der zwei Fabric Extender verbunden (siehe nachfolgende Abbildung 26). Der Fabric Extender ist dann mit jeweils 4x 10G Verbindungen mit dem Fabric Interconnect verbunden. Dabei handelt es sich um eine logische Punkt-zu-Punkt-Verbindung vom CNA bis zum Fabric-Interconnect.

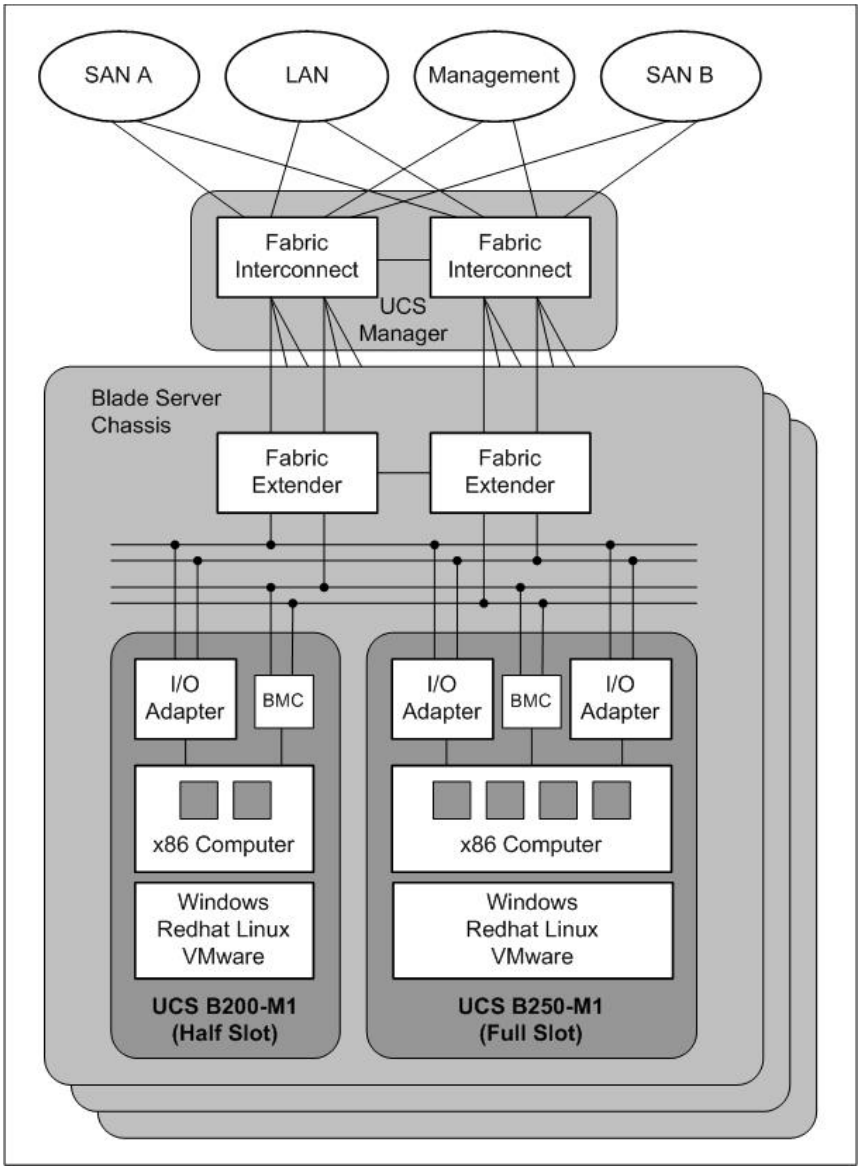


Abbildung 26: Unified Fabric im Vblock

Der Fabric Interconnect ist aus FCoE-Sicht ein FCF, und aus FC-Sicht ein NPV-Switch. D.h. die CNAs bauen zum Fabric Interconnect einen VN-zu-VF-Link über ein Punkt-zu-Punkt Ethernet Interface auf.

Zum FC-Netz, bestehend aus den zwei im Vblock eingebauten MDS 9000 FC-Switchen, werden vom Fabric Interconnect mehrere FC-Verbindungen geschaltet. Dabei sind die Ports vom Fabric Interconnect zum MDS 9000 FC-Switch als NP-Ports konfiguriert, d.h. das komplette Unified Computing System (UCS) stellt sich den FC-Switchen als ein NPIV-Host dar.

Zum Ethernet-LAN bestehen 802.1Q Trunk Verbindungen, so dass die VLANs vom Hypervisor bis zum LAN, an dem der Vblock angeschlossen ist, durchgezogen werden. Zusätzlich wird der Vblock auch an ein Out-Of-Band Management-Netz angeschlossen, über das die Management Interfaces aller Komponenten erreichbar gemacht werden.

3.5 Zentrales Management und Betriebsführung

3.5.1 EMC Ionix Unified Infrastructure Manager (UIM)

Der Unified Infrastructure Manager (UIM) stellt die zentrale Management Komponente im Vblock dar. Im Unterschied zu klassischen Computing-Umgebungen ist der Vblock zustandslos ("stateless"), d.h. die darunterliegenden Ressourcen (Compute, Netz, Storage) können dynamisch definiert und verschoben werden. Durch Hardware-Servicekataloge werden Ressourcen aus dem Bereich Server, Netz und Storage definiert und sind zentral provisionierbar (siehe Abbildung 27).

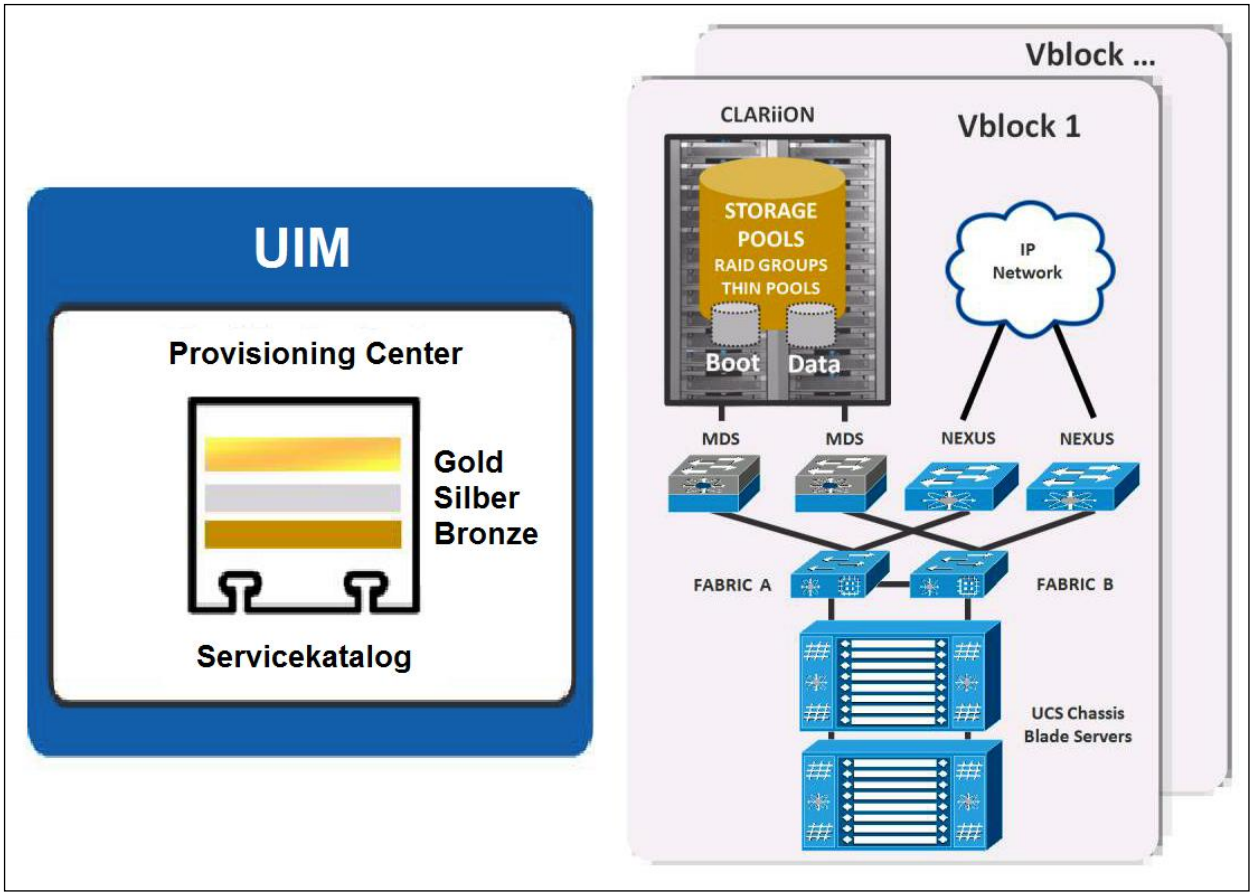


Abbildung 27: Servicekataloge im UIM

Der UIM erkennt und verwaltet die im Vblock verfügbaren physischen Ressourcen; er erlaubt außerdem eine Klassifizierung der Güte-Unterschiede innerhalb gleicher Ressourcenarten und ermöglicht damit die Bildung von Servicekatalogen für die physische Infrastruktur. Ebenso führt der UIM die Provisionierung als zentrale Instanz unter Zuhilfenahme der einzelnen Element-Manager (wie z.B. vCenter, UCS Manager, etc.) aus. In den jeweiligen Element-Managern und Systemkonfigurationen werden die Ressourcenvorgaben (Templates) definiert, die vom UIM dann über die APIs erkannt und ausgerollt werden.

In der Abbildung 27 sind im Servicekatalog beispielsweise drei Serviceklassen "Gold", "Silber" und "Bronze" definiert (jeweils für Storage, Compute und Netz). Diese unterscheiden sich durch unterschiedliche "Tiers", um entsprechende SLAs erfüllen zu können. Aus diesem Angebot kann der Anwender sich dann eine Kombination für seinen Bedarf zusammenstellen (z.B. Storage = "Gold", Compute = "Bronze" und Netz = "Bronze").

Die nachfolgende Abbildung 28 zeigt beispielhaft ein Service Offering Template "Training Example". In dem Fenster rechts unten sind die Kapazitäten für Compute, Storage und Netz vorbestimmt und können bei Bedarf individuell geändert werden.

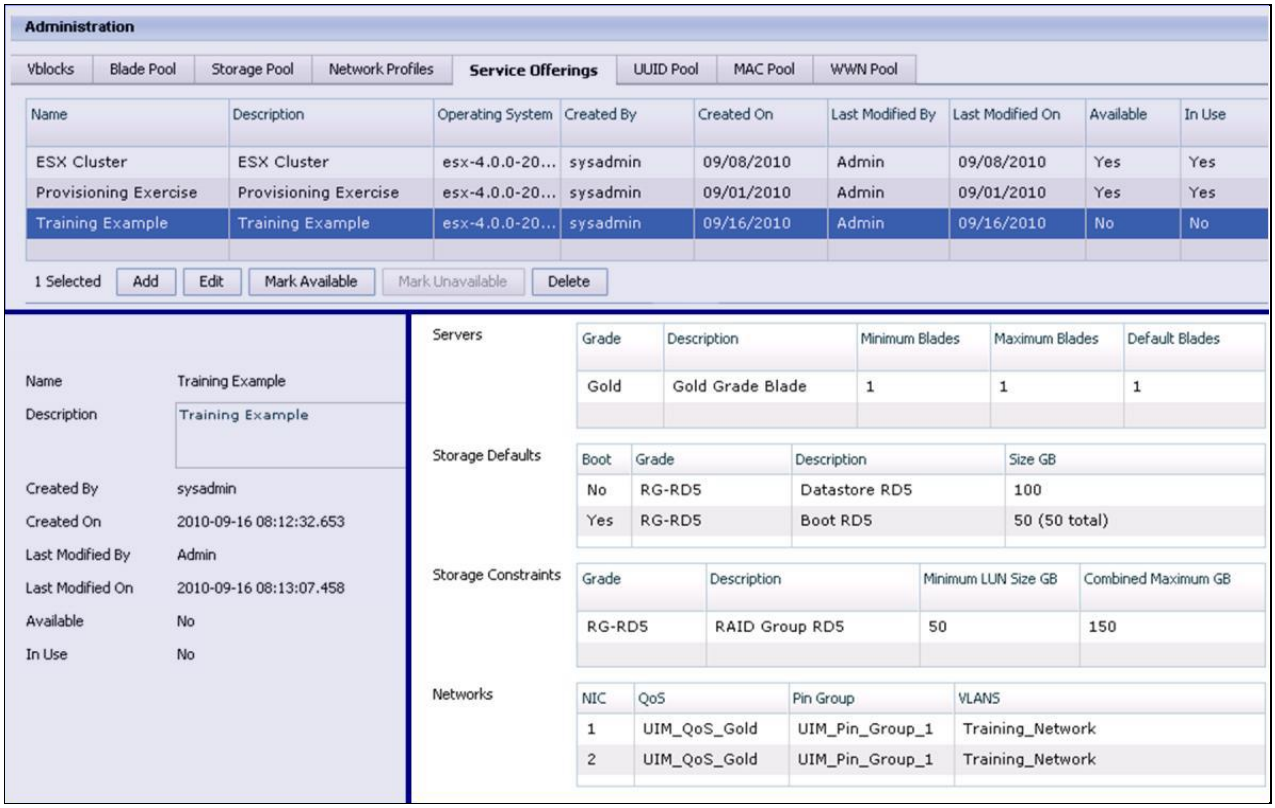


Abbildung 28: UIM - Service Offering Template (Beispiel)

3.5.2 Cisco Unified Computing System Manager (UCSM)

Das Cisco UCS System im Vblock besteht aus dem Blade Chassis, den Blade Servern, den Fabric Extendern und dem Fabric Interconnect (siehe Abbildung 29).

Der UCS Manager (UCSM) ist ein Bestandteil des Fabric Interconnect und ist der Element Manager für alle Konfigurationen und das Monitoring der Computing Komponenten, inkl. ihrer Netzbestandteile (LAN/SAN).

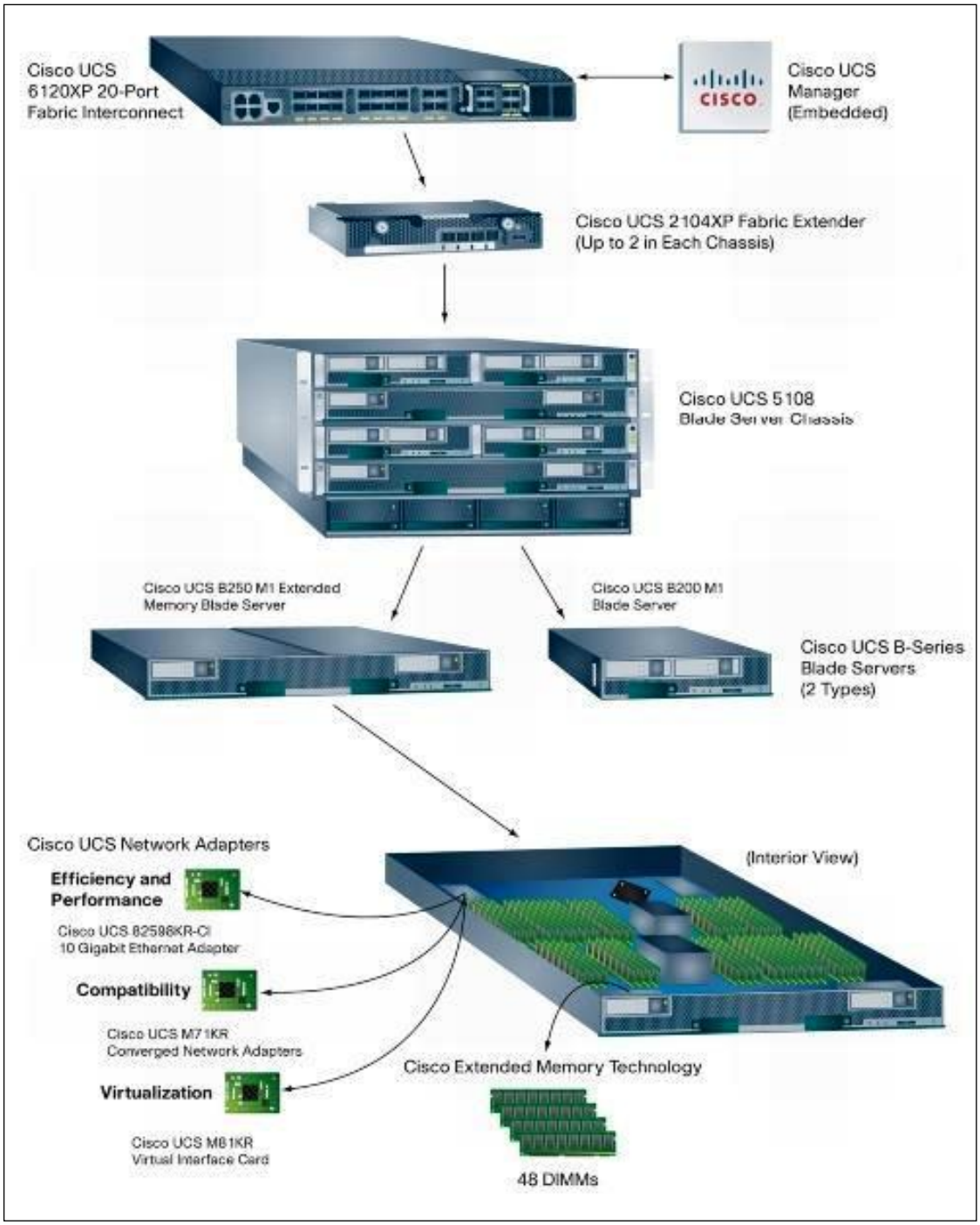


Abbildung 29: Cisco Unified Computing System Manager (UCSM)

Der "native" Zugriff auf den UCSM erfolgt entweder durch ein Command Line Interface (CLI) über SSH, oder über eine Web-GUI über HTTPS.

Neben den "nativen" Zugriffsmöglichkeiten kann der UCSM aber auch von übergeordneten Automations- und Management-Systemen über eine Reihe von Schnittstellen angesprochen werden. Dazu gehören:

- **SNMP** (Simple Network Management Protocol); read-only.
- **SMASH-CLP** (Systems Management Architecture for Server Hardware - Command Line Protocol); read-only.
- **CIM-XML** (Common Information Model - Extensible Markup Language); read-only.
- eine eigene **XML API** (Application Programing Interface)

Neben diesen Schnittstellen in den UCSM hinein gibt es noch ein "cut-through" Interface um einen direkten Zugriff zu einem individuellen Blade-Server zu ermöglichen. Funktionen dieses cut-through Interface sind:

- **KVM (Keyboard-Video-Mouse)**: Zugriff auf eine remote KVM-Session über eine Java Client Console, die vom UCSM aus gestartet werden kann.
- **SOL (Serial-Over-LAN)**: Zugriff auf ein serielles Interface in den Blade-Server hinein, z.B. für das Management von UNIX basierten Systemen.
- **IPMI (Intelligent Platform Management Interface)**: IPMI ist ein Protokoll für das Monitoring und Management eines einzelnen Blade-Servers. Über IPMI können z.B. Statuswerte von Temperatur, Spannungs- und Stromverbrauch, etc. abgerufen werden. Außerdem können über dieses Interface auch out-of-band Befehle für reboot oder power-off/power-on abgesetzt werden.

Der UCSM besteht aus mehreren Schichten, in deren Mitte die Data Management Engine (DME) liegt (siehe Abbildung 30). Die DME ist die einzige Komponente im UCS-System, in der Status- und Konfigurationsdaten für die Einzelkomponenten gehalten werden. Die DME ist objektbasiert, so dass für jede Einzelkomponente ein **"Managed Object" (MO)** existiert. Administratoren machen Änderungen an MOs über die zur Verfügung stehenden schreibberechtigten Schnittstellen (CLI, GUI und XML API).

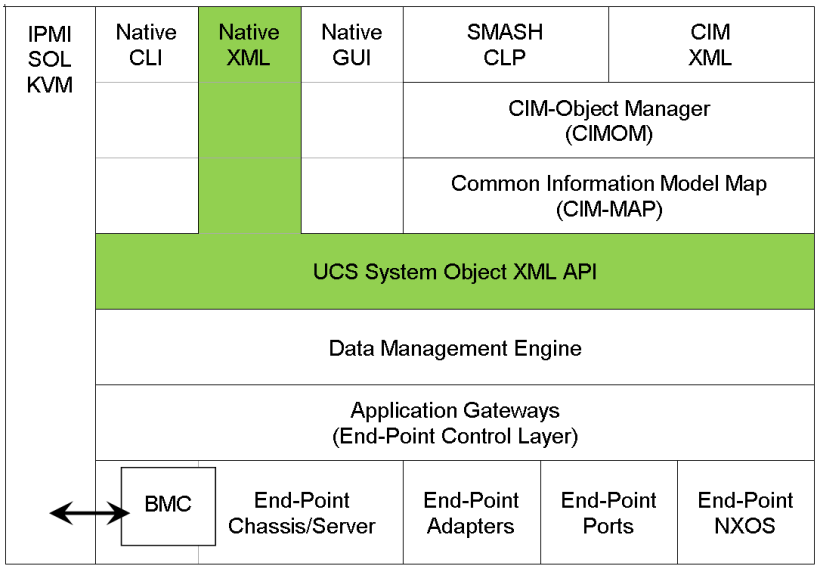


Abbildung 30: UCSM Schichten

Jede Änderung an einem MO wird von der DME validiert, und dann über die "Application Gateway / End-Point Control Layer" Schicht in die Hard-/Soft-/Firmware der Einzelkomponente im System einprogrammiert.

Für übergeordnete Management- und Automationssysteme, wie den UIM, besteht eine XML API, die über sämtliche Funktionen verfügt, die einem Administrator über das CLI und die GUI zur Verfügung stehen. Dabei wird über die XML API das gleiche Role Based Access Model (RBAC) wie beim CLI und GUI verwendet, indem der Zugriff zu Gruppen von Objekten über die Rollen-Zugehörigkeit geregelt ist.

Als weiteres Interface ist noch das Syslog Interface zu nennen. Über Syslog werden aus dem UCSM-Log Meldungen zu übergeordneten Management Systemen geschickt. Details hierzu werden im Kapitel 4.4 betrachtet.

Regeln und Policies

Der UCSM erlaubt es Policies für Teilbereiche zu definieren, z.B. für Netze, SAN-Storage, Serverkonfigurationen, die dann von Betriebsmitarbeitern beim Konfigurieren einer Computing Ressource ausgewählt werden können. So definiert ein LAN-Experte beispielsweise LAN Adapter QoS Policies, der Server Experte definiert BIOS Policies, etc. Das hierzu gehörige Role Based Access (RBAC) Modell wird in diesem Dokument im Kapitel 4.3.3.2 dargestellt.

Pools

Pools sind "Container" für Ressourcen und Identitätsdefinitionen. Dabei gibt es zwei Haupttypen:

- **Blade Pools**: In einem Blade Pool sind Server Blades zusammengefasst. Ein Betriebsmitarbeiter kann einen Blade-Server sowohl aus einem Pool automatisch auswählen lassen, oder einen bestimmten Blade Server auswählen. Über die Rollenzugehörigkeit kann bestimmten Betriebsmitarbeitern Zugriff auf einen beschränkten Pool von Blade Server Ressourcen gewährt werden.
- **Identity Pools**: Ein Identity Pool enthält drei Typen von Identitäten: WWN Adressen für die Node/Port FC-Namen, MAC-Adressen für die vEth-Netzkarten und UUIDs für die Server. Diese Identity Pools werden von Service Profiles "verbraucht", die im Weiteren erläutert werden.

Service Profiles und Templates

Ein Service Profile ist eine logische Abbildung (Objekt) eines physischen Servers (siehe Abbildung 31). Es enthält Schnittstellen nach außen wie virtuelle HBAs (vHBAs), virtuelle Netzkarten (vEth) und ihre Identitäten. Es inkludiert aber auch gewünschte Firmware/BIOS-Versionen, Boot Informationen, gewünschtes Hardware Layout (z.B. gewünschte Memory, CPU-Typ, CNA-Typ).

Ein Service Profile nutzt dabei entweder Policies und Pools als Informationsquelle, oder das Service Profile wird komplett manuell erstellt.

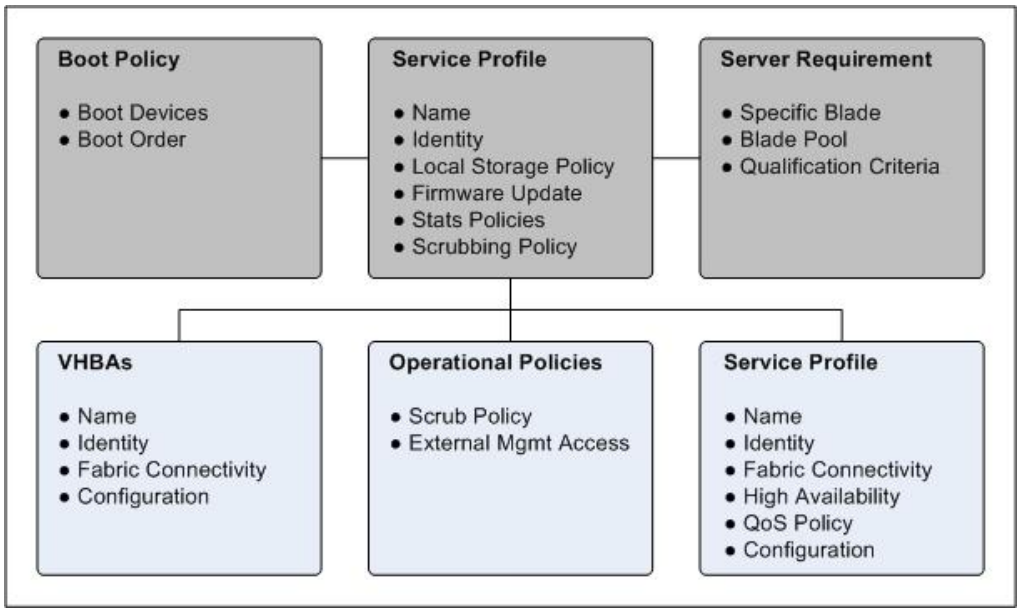


Abbildung 31: Service Profile

Service Profiles werden den Blade Servern zugeordnet. Sobald ein Service Profile einem Blade Server zugeordnet wird, wird erst einmal die Eignung des physischen Servers für das Service Profile überprüft. Wenn z.B. eine Policy im Service Profile einen bestimmten CPU-Typ oder verfügbares Memory vorgibt, das nicht zu dem verfügbaren Blade Server passt, schlägt die Zuordnung des Service Profiles zum Blade Server fehl.

Wenn die Policies alle eingehalten werden können, speichert der UCSM die Identitäten in dem Blade System. So wird z.B. den vEth-Adaptern eine MAC-Adresse vergeben. Über Policies werden dann auch z.B. QoS-Parameter dem System vergeben. Die Identitäten können beim Anlegen des Service Profiles entweder manuell vergeben oder aus einem ID-Pool bezogen werden.

Wird die Zuordnung des Service Profile vom Blade Server wieder entfernt, so werden auch die Identitäten wieder vom Blade Server entfernt.

Ein *Service Profile Template* hat den gleichen Aufbau wie ein Service Profile, nur kann es nicht einem Blade Server zugeordnet werden. D.h. ein Service Profile Template ist eine Vordefinition eines

gewünschten Service Profiles mit bestimmten Eigenschaften. Dabei können Identitäten aus Pools herangezogen werden. Wenn beispielsweise aus einem Service Profile Template ein Service Profile "geklont" wird, wählt der UCSM aus einem dem Service Profile Template zugeordneten Pool z.B. die WWNs für die vHBAs aus.

Zusätzlich zu Service Profile Templates gibt es noch z.B. Adapter Templates, die Vordefinitionen für Adapter bieten, wie z.B. konfigurierte VLANs, Anzahl und Typ von virtuellen Ethernet Karten und virtuellen HBAs, etc.

Die hier dargestellten Mittel von Policies, Pools, Service Profiles und Templates bilden einen wichtigen Bestandteil der Automation im Vblock.

Übergeordnete Management- und Automationssysteme wie der Ionix Unified Infrastructure Manager (UIM) greifen auf diese von Administratoren vordefinierten Objekte über die XML-Schnittstelle zu, um dynamisch Computing Ressourcen im Vblock anzufordern und zu provisionieren.

3.5.2.1 Cisco MDS & Nexus Family CLI / Cisco Datacenter Network Manager (DCNM)

Die Cisco MDS9000 SAN-Switche und der Nexus 1000v, die im Vblock ihren Einsatz finden, verwenden beide als Betriebssystem das Nexus OS (NX-OS).

NX-OS kann über ein CLI (Telnet und SSH), über SNMP und über CIM (Common Interface Model / http und https) konfiguriert werden. Dabei wird der Zugriff über eine Role Based Access Model (RBAC) gesteuert (siehe Kapitel 4.3.3.2).

Der UIM greift über das CLI über SSH auf den SAN Switch MDS9000, sowie auf den Nexus 1000v zu, und provisioniert die Systeme.

Als zusätzliche Komponente kann der Cisco Datacenter Network Manager (DCNM) eingesetzt werden. Diese Managementoberfläche bietet eine GUI, über welche die SAN- und LAN-Umgebungen grafisch verwaltet werden können. So kann über den DCNM z.B. das Fibre-Channel Zoning konfiguriert werden. In diesem Dokument werden wir uns aber für das Management der Cisco MDS und Nexus Komponenten auf das CLI fokussieren.

3.5.2.2 EMC Symmetrix Management Console

Die Symmetrix Management Console (SMC) stellt den Element Manager für Storage im Vblock Series 700 dar. Die SMC ist eine Java-Anwendung mit eigenem TOMCAT Application Server.

Die SMC basiert ebenso wie das CLI auf dem Symmetrix Application Program Interface (SYMAPI). Die SYMAPI ist die Programmier-Schnittstelle für Symmetrix Management Software. Diese Schnittstelle wird auch vom UIM genutzt.

Ein Zugriff auf die SMC erfolgt über HTTPS und gegebenenfalls über HTTP.

Zugriff auf die SMC hat grundsätzlich jeder. Die Authentisierung und Autorisierung erfolgt ausschließlich über Mechanismen der SYMAPI-Schnittstelle und sind damit durch SYMAUTH (siehe Kapitel 4.3.3.1) und SYMACL definiert.

Der Zugriff auf die zu verwaltenden Symmetrix-Systeme erfolgt entweder "inband" über FC- oder iSCSI-Verbindung zum Storage Array, entweder lokal vom Server, auf dem die SMC installiert ist (siehe Abbildung 32), oder remote über eine weitere HTTPS-Verbindung zu einem SYMAPI-Server, der seinerseits direkten "outband"-Zugriff auf die zu verwaltenden Storage Arrays hat (siehe Abbildung 33).

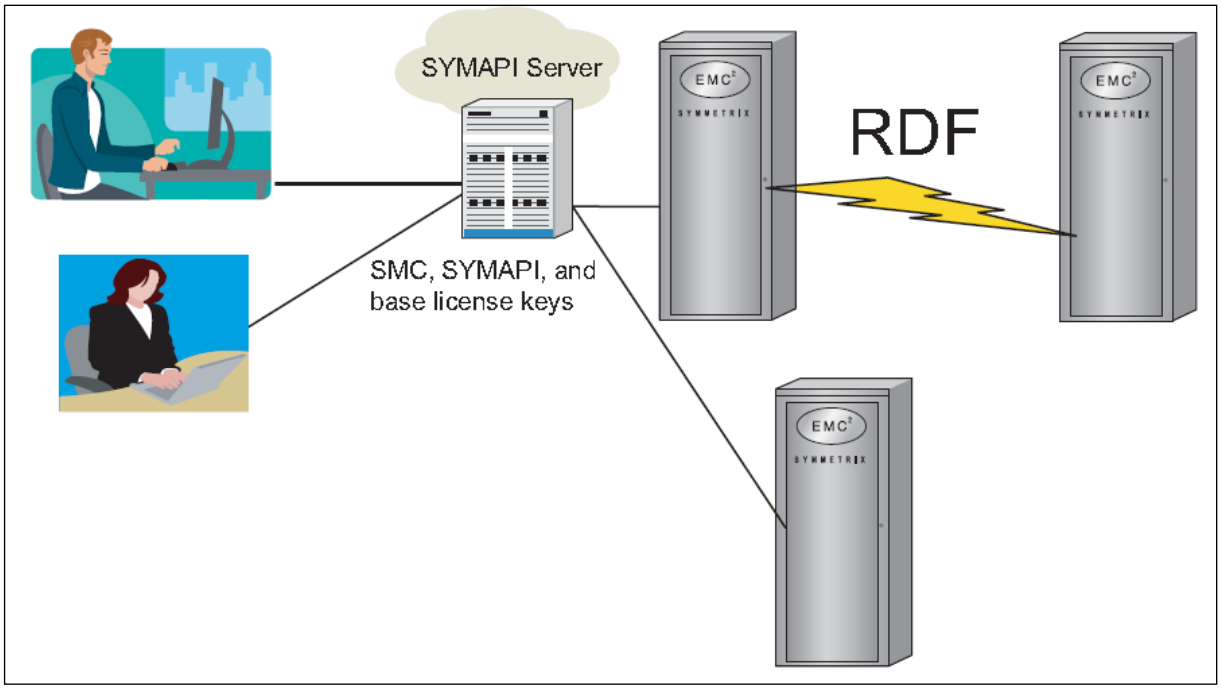


Abbildung 32: Lokale SMC SYMAPI Konfiguration

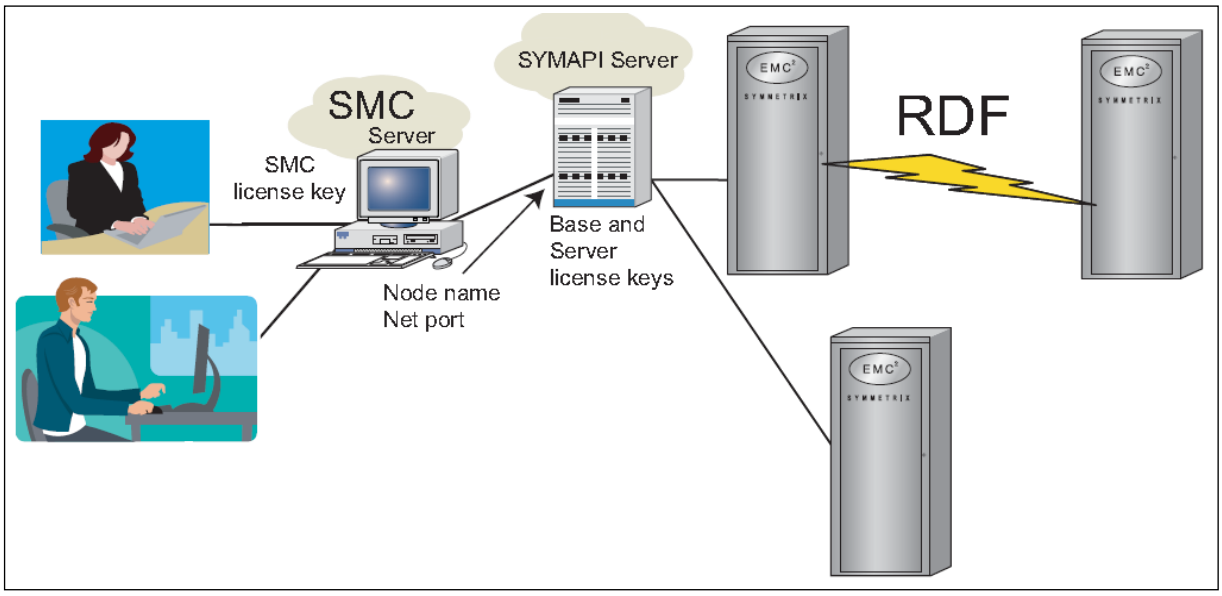


Abbildung 33: Remote SMC SYMAPI Konfiguration

Zusätzlich dazu verwendet der UIM unabhängig von dem Vblock zugrundeliegenden Storage Array den EMC SMI-S Provider.

3.5.2.3 EMC UniSphere

EMC UniSphere stellt den Storage Element Manager für die Vblocks 0, 1/1U und Series 300 dar. Hier ist statt der EMC Symmetrix ein auf CLARiiON-Technologie basierendes Storage Array vorhanden.

Ebenso wie die SMC ist der Zugriff auf UniSphere über einen Web-Browser notwendig. Im Unterschied zur SMC befindet sich der Web-Server jedoch nicht auf einem Server außerhalb des Arrays sondern läuft auf den Storage-Prozessoren der CLARiiON selbst. Zusätzlich gibt es ein Kommandozeilen-Interface (NaviSecCLI), das ebenfalls vom UIM genutzt wird.

Die Zugriffsbeschränkung basiert ähnlich wie bei der Symmetrix auf den im Array angegebenen und konfigurierten Rollen und Benutzern.

3.5.2.4 VMware vCenter

Die VMware vSphere Produktfamilie beinhaltet das zentrale Management einer ESX-Serverfarm, das sogenannte vCenter. Dort wird an zentraler Stelle Zugriff auf Ressourcen gewährt, die durch den UIM oder den UCSM an den Hypervisor herangeführt werden. Der vSphere Client dient als zentrales Administrationstool und kann sich sowohl mit der zentralen Management Konsole verbinden als auch mit jedem einzelnen ESX-Server.

Im Normalfall erfolgt 90% der Administration über das vCenter. Deshalb ist es sehr wichtig, diese zentrale Management Komponenten zu schützen und dort entsprechende Hardening Guidelines von VMware strikt zu implementieren (siehe dazu auch den VMware vSphere 4.0 Security Hardening Guide; [04]. Eine rollenbasierende Benutzerverwaltung ist dort sehr granular beschrieben).

Weiterführende Informationen zur Delegierung von administrativen Rechten sind auch im Kapitel 4.3.3.5 zu finden.

3.5.3 VMware vCloud Director (VCD)

Der vCloud Director ist die Cloud Portal-Lösung von VMware, und kann von einem Service Provider auch auf einem Vblock betrieben werden. Er fügt über die vSphere Basisinfrastruktur eine weitere Abstraktionsschicht hinzu, in der Mandanten angelegt werden. Der vCloud Director ist ein Self Service Portal zum Provisionieren von Diensten aus einem definierbaren Service Katalog und ermöglicht ein effizienteres und flexibleres Modell zur Bereitstellung von IT-Services z.B. in Form von *Infrastructure as a Service (IaaS)*.

Die nachfolgende Abbildung 34 zeigt die Architektur des vCloud Directors im Überblick.

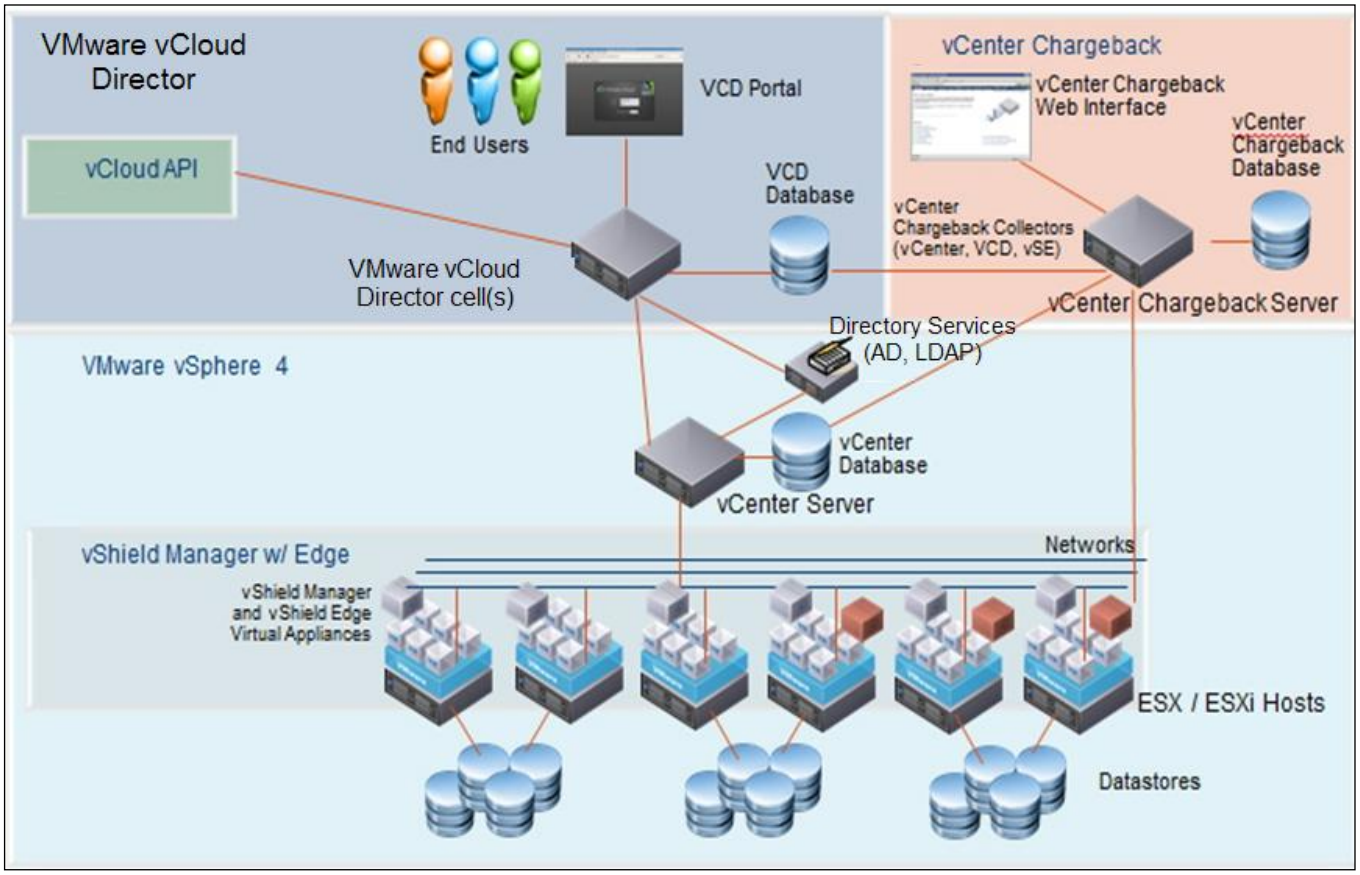


Abbildung 34: Architektur des vCloud Directors

Das Portal verbessert den Service und die Provisionierungszeiten unter Einhaltung aller Isolations- und Sicherheitsvorgaben. Das Infrastrukturteam entscheidet, welche Ressourcen (Netze, Storage, CPU, Memory) einen bestimmten Mandanten zugewiesen werden.

Damit kann in derselben ESX-Farm eine interne Cloud betrieben werden, die z.B. die internen Mandanten Training, Entwicklung, Vorproduktion und Produktion bedient. Dedizierte Ressourcen werden für jede Abteilung allokiert und können mit einem Pay-Per-Use-Modell mittels vCenter Chargeback rückverrechnet werden. Die Netzsegmentierung wird mittels VLANs und den bereits vorgestellten virtuellen Firewall-Systemen erreicht. Im Falle eines Public Cloud Providers sind die Abteilungen im oben genannten Beispiel einfach verschiedene Mandanten.

Der vCloud Director verwendet ein *RBAC (Role Based Access Control)* Modell, das komplett von den Administrationswerkzeugen des Infrastrukturteams losgelöst ist.

Der Organisationsadministrator ist, wie in Abbildung 34 dargestellt, ein Enduser. Seine Rolle ist die Benutzerrolle mit den meisten Rechten. Er kann jedoch Rechte und Ressourcen nur im Rahmen der ihm zugewiesenen Organisationsressourcen im vCD Portal verteilen. Er hat keinen Zugriff auf die darunterliegenden vSphere Administrationswerkzeuge.

Wie in der obigen Abbildung dargestellt, besteht der vCloud Director aus einem oder mehreren *Director Cells*, welche Verbindungspunkte nach aussen für Benutzer darstellen und das Portal bereitstellen. In Richtung Infrastruktur kommuniziert der vCloud Director mit dem vShield Manager und dem vCenter, um beispielsweise virtuelle Firewalls auszurollen oder Aufgaben zur Provisionierung an das oder die zugrundeliegenden vCenter zu übertragen. Alle Prozesse und Funktionen und das Aussehen des Cloud Directors lassen sich verändern und über die vCloud Director API ansprechen.

3.5.4 VMware Configuration Manager

Der VMware Configuration Manager (vCM) ist kein Standardbestandteil eines Vblocks. Dennoch soll er hier erwähnt werden, da er in der Lage ist Best Practices durch Richtlinien (ISO, NIST, DISA (Data Interchange Standards Association), CIS, PCI-DSS, u.v.m.) für virtuelle und physische Maschinen anzuwenden. Auch herstellerspezifische Härtungsempfehlungen, wie z.B. im VMware vSphere 4.0 Security Hardening Guide [04] gegeben, können überprüft werden. Damit können Betriebsysteme nach Best Practices und bestimmten Regulatorien gehärtet werden.

Das Sicherheitsteam innerhalb eines Unternehmens kann sich einen Überblick verschaffen, ob bestimmte Richtlinien eingehalten werden. Damit kann die Vorbereitungszeit für einen Audit reduziert werden, da Compliance und Änderungen als fortlaufender Prozess regelmäßig durchgeführt und korrigiert werden können. So werden bei Windows bis zu 80.000 Prüfpunkte abgefragt, dazu zählen Registry Active Directory, Benutzer Accounts, Share Permissions und vieles mehr. Deshalb lassen sich schnell Schwachstellen aufdecken, die der vCM durch Konfigurationsänderung innerhalb des verwalteteten Knotens (z.B. Registry Härtung in Windows) oder durch Softwareinstallation (z.B. zu alter AV Client) beheben kann. Richtlinientemplates können jederzeit an die jeweiligen Bedürfnisse des Unternehmens angepasst werden. Auch Linux oder Unix Systeme lassen sich damit härten.

Darüberhinaus existiert pro verwalteten Knoten ein Audit Logging, um gewünschte oder ungewünschte Änderungen zu beobachten, und um sie eventuell wieder zurückgängig zu machen. Bei Poblemen kann man damit auch eine *Root Cause Analyse* vornehmen, ob vielleicht letzte Änderungen am System das Problem verursacht haben. Eine Integration in eine CMDB wie z.B. BMC Remedy oder VMware Service Manager ist möglich.

VMware Configuration Manager ist darüber hinaus in der Lage sowohl physische als auch virtuelle Betriebssysteme zu installieren (PXE-Installation) und deren Applikationen zu patchen. Zu den unterstützten Betriebssystemen gehören alle Windows Betriebssysteme, aber auch viele Unix/Linux Derivate wie z.B. Red Hat, Suse, Debian, HP UX, IBM AIX, Solaris und Mac OS Server.

4 Mögliche Gefährdungen und Gegenmaßnahmen

Die allgemeinen Gefährdungen bei der Nutzung von Vblock entsprechen den relevanten Gefährdungen in den IT-Grundschutz-Katalogen G1 bis G5 (siehe [17]). Diese allgemeinen Gefährdungen werden in dieser Studie nicht mit ausgeführt. In diesem Kapitel werden einige Gefährdungen, die beim Betrieb einer Private Cloud-Infrastruktur entstehen, beschrieben und im Anschluss tabellarisch aufgelistet. Der Fokus wird auf Vblock-spezifische Gefährdungen gelegt. Dies orientiert sich an einer Gefährdungsübersicht der ergänzenden Sicherheitsanalyse, wie sie der BSI-Standard 100-2 (siehe [19]) vorsieht.

4.1 Unzureichende Mandanten Isolation

Das Ziel der Virtualisierung in einem Vblock ist das Zusammenlegen von Computing Ressourcen auf die Vblock-Infrastruktur (Blade-Server, Hypervisor, SAN Storage, etc). Dabei werden vorher getrennt aufgebaute Systeme von Mandanten und Applikations-Tiers auf der virtualisierten Plattform zusammengefasst. Das kann z.B. bedeuten, dass Ressourcen verschiedener Abteilungen innerhalb eines Unternehmens auf den Vblock zusammengelegt werden, es kann z.B. aber auch bedeuten, dass völlig getrennte Endkunden, wie beim Einsatz in Providerinfrastrukturen, voneinander abgeschottet im Vblock zusammengelegt werden. Neben Fehlern bei der Konfiguration von Komponenten ergeben sich die größten Gefährdungen, wenn eine Möglichkeit gefunden wird, die Isolation zu umgehen.

Bei der Zusammenlegung entstehen verschiedenste Gefährdungen, die in diesem Kapitel betrachtet werden. Die Gefährdungen können grob unterteilt werden in:

- Gefährdungen innerhalb des Hypervisor Systems, wie z.B. die nach jetzigem Kenntnisstand eher theoretische Möglichkeit, dass ein Mandant von einem Gastsystem, auf das er vollen Zugriff hat, unerlaubt Zugriff auf ein fremdes Gastsystem bekommt. Diese Angriffe werden unter dem Begriff "*VM Escape*" zusammengefasst. Mögliche Angriffsszenarien sind z.B. das Auslesen fremder physischer Memory Bereiche im Hypervisor.
- Gefährdungen durch DoS-Attacken innerhalb des Hypervisors, wie z.B. der Einsatz von CPU Last Tools mit dem Ziel, alle CPU Ressourcen des Hypervisors an sich zu ziehen und damit andere Gast-Systeme "lahmzulegen". Andere Szenarien sind z.B. die Inanspruchnahme von zu viel physischem Memory, um den Hypervisor zum exzessiven *Swappen* zu veranlassen.
- Gefährdungen durch Remote Attacken über das Netz, wie z.B. *Man-in-the-Middle* Attacken mit dem Ziel, den Netzverkehr von einem VM-Gast durch einen anderen VM Gast abzufangen und aufzuzeichnen.
- Gefährdungen durch DoS-Attacken über das Netz, wie z.B. *Syn Floods*, *Ping of Death*, etc. von einem Gast auf einen anderen Gast.
- Gefährdungen durch Zugriff auf fremde Storage Ressourcen, wie z.B. das Vortäuschen von falschen Identitäten mit dem Ziel Zugriff auf die Daten eines anderen Mandanten zu erlangen.

4.1.1 Hypervisor Isolation

Gefährdung 1:	Virtual Machine Escape – Erlangen von privilegiertem Zugriff auf ein fremdes Gastsystem oder den Hypervisor selbst durch einen kompromittierten Gast.

Bei der Virtual Machine Escape handelt es sich um einen Angriff, bei dem ein externer Angreifer zuerst die Kontrolle über das VM-Gast-System erlangen muss, welches auch schon durch geeignete Schutzmechanismen wie Antivirus, Antispyware, Backdoor-Detection, etc. gehärtet worden ist.
Ein Innentäter hat natürlich bereits den vollen Zugriff auf den Gast. Dies gilt natürlich auch für Angebote der Form IaaS-Compute. Auch hier hat ein Nutzer de facto volle Zugriffsrechte auf die Gast-VM. Hat man die Kontrolle über den Gast, könnte man potentiell eine aktuell bestehende Verwundbarkeit im Hypervisor ausnutzen, um privilegierten Zugriff auf ein fremdes Gastsystem und den Hypervisor selbst zu erlangen.

Mögliche Angriffsszenarien sind z.B. das Auslesen oder Beschreiben fremder Memory Bereiche anderer Gäste oder des Hypervisors selbst. Mögliche Folgen wären das Ausspähen fremder Daten, die im Memory gehalten werden, oder das Auslösen der Ausführung von eingeschleustem *shellcode* durch das Beschreiben fremder Memorybereiche.

Im Nachfolgenden wird beschrieben, wie der beim Vblock eingesetzte VMware ESX/ESXi Hypervisor bereits im Produktdesign solche Angriffsszenarien verhindert.

Der Hypervisor im VMware ESX/ESXi Produkt präsentiert dem Gastbetriebssystem eine generische x86-Hardwareplattform durch Virtualisierung von vier Hauptkomponenten: CPU, Memory, Netz und Disk. Ein Betriebssystem wird dann in die virtualisierte Hardwareplattform installiert.

Der Hypervisor umfasst neben dem VMkernel weitere Softwarekomponenten, welche für zentrales Monitoring und Betriebsführung zuständig sind.

Die zentrale Komponente im ESX/ESXi Hypervisor ist der *VMkernel* (siehe Abbildung 35). Der VMkernel ist ein speziell von VMware entworfener Kernel, um virtuelle Maschinen zu betreiben. Er kontrolliert die benutzte Hardware des ESX/ESXi Hosts, und er plant die Verteilung der Hardwareressourcen zwischen virtuellen Maschinen (*Scheduler*).

Wie in der Abbildung dargestellt, umfasst der VMkernel Netzkarten, Eingabe-/Ausgabe-Treiber, die Filesystem-Treiber, sowie den Softwarestack für den virtuellen Switch. Desweiteren ist seine Aufgabe, für die sichere Verteilung von physischen Hardwareressourcen zu sorgen und deren Verteilung an die VMs vorzunehmen (Scheduler, Memory Allocator). Der Virtual Machine Monitor (VMM) wird einmal pro Gastbetriebssystem gestartet und fängt privilegierte Systemcalls der VMs ab. Darüberhinaus stellt er der VM virtuelle x86-Standardhardware zur Verfügung.

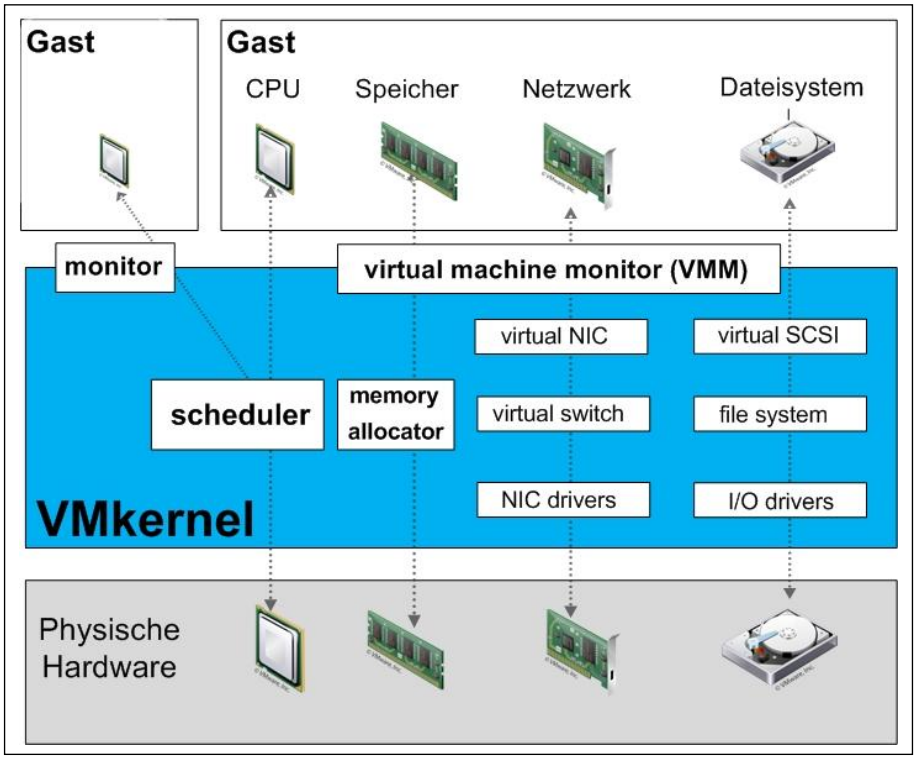


Abbildung 35: VMware VMkernel Stack

Weil der VMkernel ausschließlich dafür ausgelegt ist virtuelle Maschinen zu betreiben, ist der Zugriff auf den VMkernel strikt auf die Schnittstelle limitiert, die man für diese virtuellen Maschinen benötigt. Es gibt keine öffentlichen Schnittstellen zum VMkernel, und er kann keinen "*Arbitrary Code*" ausführen.
Durch den VMM benutzt die VM (d.h. der Gast) die Kerntechnologien des VMkernels wie Memory Management oder die Netz- oder Storage Stacks des ESX/ESXi Servers. Pro VM wird ein weiterer VMM-Prozess gestartet, wobei jeder dieser Prozesse CPU, Memory, I/O Devices partitioniert und teilt, damit die VM erfolgreich virtualisiert betrieben werden kann.

Dabei kann der VMM die Virtualisierung im Hardware-, Software- oder Paravirtualisierungsmodus betreiben. Bei Paravirtualisierung betreibt das Betriebssystem einen für die Virtualisierung angepassten Kernel; es "weiß" also, dass es virtualisiert wurde. Unmodifizierte Betriebs-Systeme wie z.B. Windows 2000 oder Windows XP, können nicht paravirtualisiert werden. Paravirtualisierung hat zum Ziel, die Kommunikation zwischen Gastbetriebssystem und Hypervisor zu verbessern. Der Hauptvorteil liegt in einem geringeren Performance Overhead. Der Leistungsvorteil kann je nach Gastbetriebssystem sehr unterschiedlich ausfallen. Die Kompatibilität und Portierbarkeit der VMs ist ein wesentlicher Nachteil der Paravirtualisierung.

Die virtuelle CPU einer VM besteht aus dem virtuellen *Instruction Set* und der virtuellen *Memory Unit* (MMU). Die MMU ist die virtuelle Hardware, die das Mapping zwischen virtuellen RAM und dem physischen Memory des ESX-Servers verwaltet. Die Kombination der Techniken, die verwendet werden, um das Instruction Set der CPU und den Hauptspeicher zu virtualisieren, bestimmt den verwendeten "*Monitor Execution Mode*". Der VMM identifiziert die ESX/ESXi Hardware Plattform und deren verfügbare CPU-Funktionen und wählt dann den Monitor Execution Mode für ein bestimmtes Gastbetriebssystem aus. Der VMM kann als Monitor Execution Modus die Softwarevirtualisierung (siehe Abbildung 37), Hardwarevirtualisierung (siehe Abbildung 38) oder eine Kombination aus beiden auswählen.

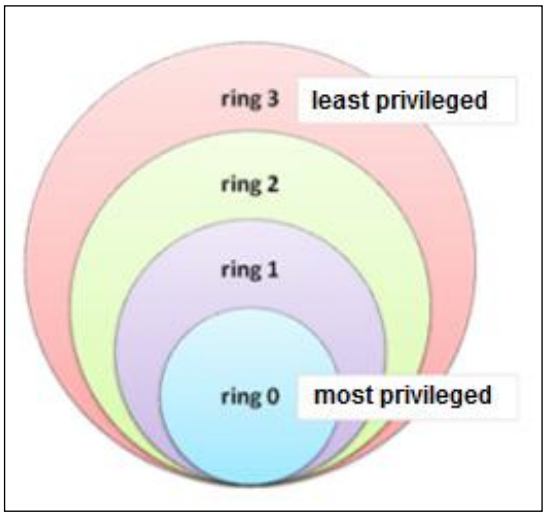


Abbildung 36: x86 Ringmodell

Das x86-Ringmodell beinhaltet vier Schutzstufen zur Ausführung von x86-CPU-Befehlssätzen (siehe Abbildung 36).

Wenn man CPUs konventionell ohne Virtualisierung betreibt, hat das Betriebssystem volle Kontrolle über die Hardware, während Applikationen auf Ring 3 betrieben werden.

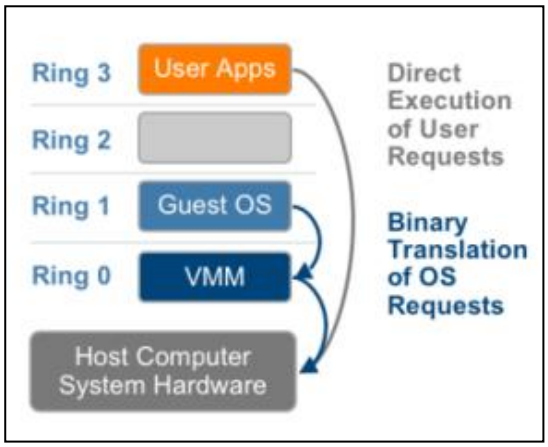


Abbildung 37: VMware Binary Translation ohne Hardwareunterstützung

Binary Translation ohne Hardwareunterstützung

Wie in der Abbildung 37 dargestellt, läuft der VMM (Virtual Machine Monitor) im Hypervisor auf Ring 0 im privilegierten Modus, während das Gastbetriebssystem auf Ring 1, also im nicht privilegierten Modus läuft. Bei dieser Vollvirtualisierung weiß der Gast nicht, dass er virtualisiert wurde, und braucht deshalb auch keine Anpassungen. Der VMM stattet die VM mit allen Diensten eines physischen Systems aus, welchem z.B. virtuelles BIOS, virtuelle Geräte und virtueller Hauptspeicher angehören.

Applikationen (User Apps) laufen auf Ring 3 und können mittels „Direct Execution“ direkt auf die Hardware zugreifen, während privilegierte Instruktionen des Gastbetriebssystems die Binary Translation durchlaufen müssen, bevor sie auf die Hardware zugreifen können.

VMware kann jedes x86-Betriebssystem durch Benutzung von *Binary Translation* und *Direct Execution* Technologien virtualisieren.

Der Binary Translator ersetzt privilegierte Instruktionen mit einer Sequenz von Instruktionen, welche die privilegierten Operationen in der VM ausführen, nicht aber auf der physischen Maschine. Diese Übersetzung erzwingt die Enkapsulierung und Isolierung der VM unter Einhaltung der x86-Befehlssemantik aus Sicht der VM.

User Level Code wird hingegen direkt auf der CPU ausgeführt, um eine hochperformante Virtualisierung zu ermöglichen. Dies ist sicher, da User Level Code keine priviligierten Instruktionen ausführen kann.

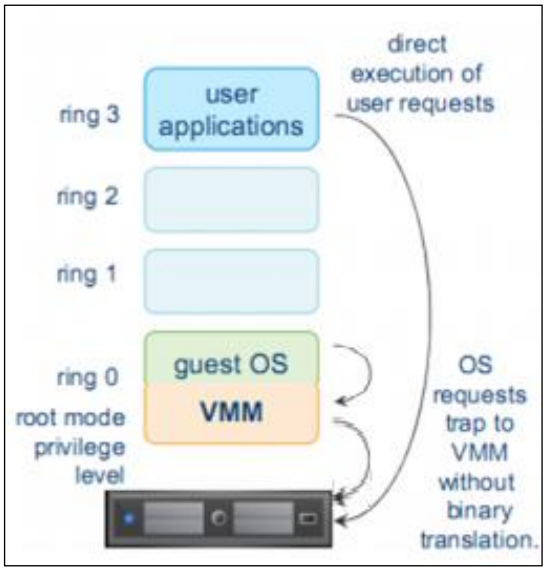


Abbildung 38: VMware Binary Translation mit Hardwareunterstützung (Intel VT und AMD-V)

Binary Translation mit Hardwareunterstützung

In diesem Modus läuft der VMM auf Ring 0 im "root mode priviledged level", während das virtualisierte Gastbetriebssystem ebenfalls auf Ring 0 im "non-priviledged level" läuft (siehe Abbildung 38).

Applikationen laufen weiterhin auf Ring 3. Wie dargestellt werden nicht-privilegierte User Requests weiterhin direkt an die Hardware CPU gesendet ohne abgefangen werden zu müssen. Privilegierte Instruktionen des Gastbetriebssystems werden von Hardwareerweiterungen neuerer CPU-Generationen direkt in der Hardware sicher abgearbeitet, wobei der VMM die volle Kontrolle behält.

Die CPU-Hersteller Intel und AMD haben diese Funktion unter dem Namen Intel VT-x und AMD-V implementiert. Beide Designs haben das gleiche Ziel: Virtualisierung einfacher und effektiver zu machen. Beide Technologien erlauben es dem VMM Binary Translations unter Beibehaltung der vollen Kontrolle der privilegierten Instruction Sets auf die CPU auszulagern.

Weitergehende Informationen hierzu finden Sie im Whitepaper Software und Hardware Techniques for x86 Virtualization; [20].

Ein gegenseitiges Auslesen von Speicher der VMs ist aufgrund des Designs der CPU Instruction Set Virtualisierung nicht möglich. Es können grundsätzlich auch keine Buffer Overflow Attacken gegen den VMkernel gefahren werden, weil Kernelinhalte immer durch die ASLR-Technik (Address Space Layout Randomization) in unterschiedlichen Adressbereichen des Hauptspeichers geladen werden. Desweiteren besteht beim richtigen Sicherheitsdesign der virtuellen Infrastruktur keine Verbindung zwischen virtuellen Maschinen, Netzen und Verwaltungsnetzen. Eine Buffer Overflow Attacke kann nur erfolgen, wenn der Angreifer die Sicherheitsmaßnahmen im Managementnetz bereits überwunden hat (Firewall, Intrusion Detection, etc.).

Abgesehen von den Enkapsulierungsmechanismen, die in den Hypervisor eingebaut sind, sind virtuelle Maschinen generell so zu behandeln wie physische Maschinen. Regelmäßiges Patchen des Hypervisors und der virtuellen Maschinen, sowie Gastschutzmechanismen wie Antivirus, Intrusion Detection und Härtung des Gast-Betriebssystems sind durchzuführen bzw. zu implementieren.

Auch wenn VM Escape Angriffe eher unwahrscheinlich sind, ist es dennoch sehr wichtig, beim Hypervisor auf Verwundbarkeit zu achten. Diesbezügliche Informationen werden von VMware auf deren Internet-Seite publiziert, und der ESX Hypervisor ist entsprechend privilegiert zu patchen.
Dabei sei erwähnt, dass es sich bei ESX(i) und seinem VM-Kernel nicht um ein "General Purpose" Betriebssystem handelt, und damit die Patch-Häufigkeit vergleichsweise gering ist.

4.1.2 Ressourcenverteilung zwischen Mandanten

Gefährdung 2:	DoS-Attacken auf Hypervisor-Ressourcen (z.B. durch CPU Testtools, Memory Überallokierung, etc.).

Eine potentielle Attacke könnte darauf abzielen, sämtliche Ressourcen im Hypervisor zu beanspruchen. Dadurch würden andere VMs in ihrer Leistungsfähigkeit stark beeinträchtigt werden.

Der VM-Gast kann jedoch nur Ressourcen verbrauchen, die ihm der Hypervisor von außen zugewiesen hat. Im Beispiel einer DoS-Attacke auf die CPU-Ressourcen kann eine VM nur die Anzahl der physischen und logischen CPUs im Host verwenden, die ihr im Rahmen der VM-Konfiguration zugewiesen worden sind. Hat eine Maschine zwei vCPUs, so können maximal zwei logische oder physische Cores des Hosts "verbraucht" werden. Dies lässt sich weiter einschränken, indem man die maximalen CPU-Zyklen in MHz nach oben hin reglementiert. Gleiches gilt auch für den Hauptspeicher. Ein Ressourcenmaximum für RAM kann dort auch administrativ festgelegt werden.

Beim Einsatz eines Vblocks können physische Ressourcen wie z.B. CPU, Arbeitsspeicher oder Storage überbucht werden, d. h. den virtuellen Maschinen werden in Summe mehr Ressourcen zugesagt, als tatsächlich vorhanden ist. Da in der Regel nicht alle VMs die zugesicherten Ressourcen gleichzeitig voll ausnutzen, ist dies ein praktikabler Ansatz. Zum Beispiel ist die Größe des Arbeitsspeichers pro Host durch den physisch zur Verfügung stehenden Arbeitsspeicher begrenzt. Hat ein Angreifer eine VM unter seiner Kontrolle, dann könnte er über ein Schadprogramm so viel Hauptspeicher anfordern, dass dieser für andere virtuelle Maschinen knapp wird.

Dies kann jedoch mit *Ressource Pools* im VMware vCenter verhindert werden, indem definiert wird wie viel RAM und CPU ein Mandant oder eine Abteilung insgesamt für die Gesamtheit aller provisionierten VM-Gäste verbrauchen kann.

Weiterführende Informationen findet man im vSphere Resource Management Guide; [05].

4.1.3 Netz Isolation

Eine ganze Gruppe von Gefährdungen ergibt sich durch Abhören von Netzverkehr mit dem Ziel unerwünschten Zugriff auf die Daten anderer VMs zu erlangen (z.B. durch "Man-in-the-Middle-Attacken").

Angriffsszenarien auf dem Hypervisor selbst (VM Escape) erfordern hohe Programmierkenntnisse sowie Kenntnisse über noch nicht behobene Schwachstellen. Daher ist es für einen Angreifer meist attraktiver einen remote Angriff von einem VM-Gast, der sich in seinem Zugriff befindet, durch das Netz auf andere VM-Gäste oder die Infrastruktur zu versuchen.

Dieser Abschnitt beschäftigt sich mit den Gefährdungen, die aufgrund des gemeinsam genutzten Netzes entstehen und beschreibt Maßnahmen, die zur sicheren Trennung des Netzverkehrs zwischen Mandanten eingesetzt werden können.

Wie bereits im Kapitel 4.1 beschrieben, gibt es verschiedene Möglichkeiten der Trennung zwischen Mandanten. Im Vblock beginnt die Trennung der Mandanten immer bei Port-Gruppen im ESX/ESXi Hypervisor. Von dort aus gibt es verschiedene Optionen, wie die Mandantentrennung durchgeführt wird.

Daher betrachten wir nachfolgend detailliert die Optionen:

- Trennung der Mandanten durch VLANs
- Trennung der Mandanten durch Private-VLANs
- Trennung der Mandanten durch virtuelle Firewalls

Alle diese Optionen, sowie die im VMware ESX/ESXi Hypervisor schon im Produktdesign erhaltenen Maßnahmen, adressieren die Gefährdung des unerwünschten Zugriffs und Abhören von fremdem Netzverkehr durch einen VM-Gast.

4.1.3.1 vSwitch im ESX Hypervisor

Gefährdung 3:	Netzangriffe wie MAC Flooding, Rogue Root Bridge, Spanning Tree Angriffe.

Ein Angreifer, der Zugriff zu einem VM-Gast im Vblock hat, könnte versuchen diverse Netzangriffe zu starten. Dazu gehören unter anderem:

- ***MAC Flooding***:
  Beim MAC-Flooding versucht ein Angreifer die Netz-Switche durch das Senden von vielen Ethernet Frames mit wechselnden Quell-MAC-Addressen zu "überfluten". Die Netz-Switche würden, sobald die Limits ihrer MAC-Adressen erreicht werden, anfangen Netzverkehr zu fluten und der Angreifer kann dadurch den Netzverkehr fremder Teilnehmer sehen.
- ***Spanning Tree* Angriffe**:
  Bei diesem Angriff versendet der Angreifer sogenannte BPDUs (Bridge Protocol Data Units) mit dem Ziel, die Netzswitche dazu zu bringen den VM-Gast des Angreifers als Root Bridge (*rogue root bridge*) anzusehen. Im besten Fall für den Angreifer könnte der Netzverkehr dann über seine Maschine übertragen werden, so dass er als Man-in-the-Middle (MITM) den Netzverkehr mitschneiden könnte. Gelingt es einem Angreifer sich als MITM zu etablieren, dann könnte er dieses aber auch dazu nutzen DoS-Attacken zu fahren und durch falsche BPDUs das Netz dazu zwingen die Spanning-Tree Topologie neu aufzubauen, was zu Netzausfällen führt.

Nachfolgend werden der VMware vSwitch und Distributed vSwitch betrachtet und analysiert inwiefern eine Verwundbarkeit bei diesen beiden o.g. Angriffen gegeben ist.

In der nachfolgenden Abbildung 39 ist links ein ESXi-Server und rechts daneben ein ESX-Server (mit Service Konsole) zu sehen. Auf jedem ESX-Server können ein oder mehrere virtuelle Switche angelegt werden, die wiederum über mehrere Portgruppen verfügen können. Jeder Portgruppe können unterschiedliche physische Uplink Adapter zugeordnet werden. Am oberen Rand kann gesehen werden, dass virtuelle Maschinen auch über ein oder mehrere virtuelle Netzkarten (vNICs) verfügen, die in einen virtuellen Port einer Portgruppe gepatcht sind.

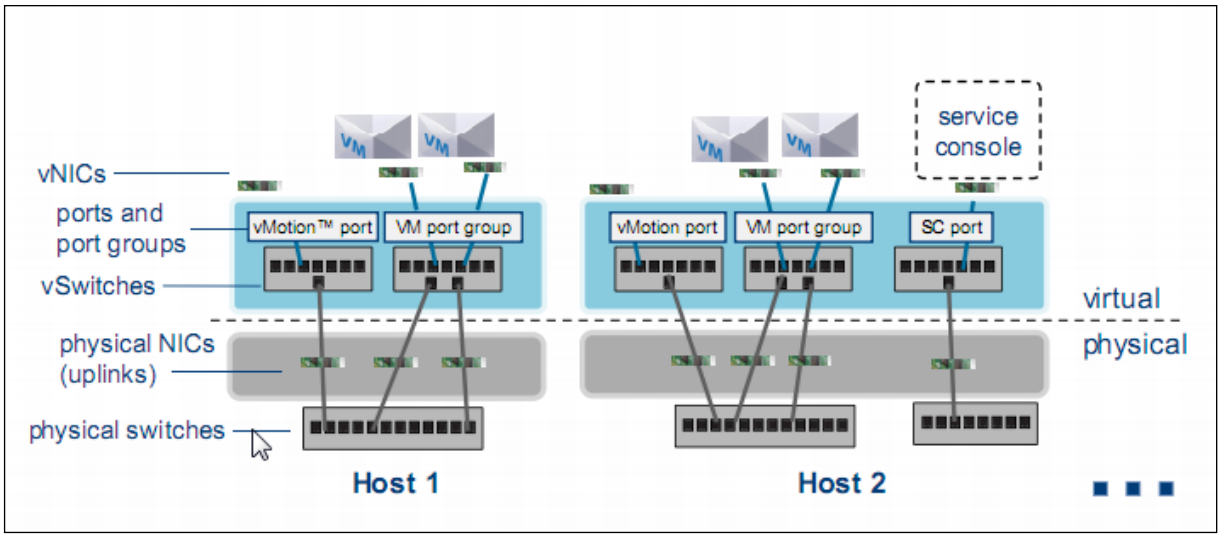


Abbildung 39: VMware Netz Stack - Beispiel bei einem klassischen VMware vSwitch

Der virtuelle Switch arbeitet auf Layer2; er routet also nicht. Meist werden die virtuellen Switche mit der physischen Switch Infrastruktur mit einem 802.1Q-VLAN-Trunk versehen, der also nur *tagged* Pakete zwischen dem virtuellen Switch und seinem physischen Uplink Pendant erlaubt. Das Taggen in das jeweilige VLAN erfolgt dann mittels *Virtual Switch Tagging* (VST). Damit lassen sich die VLAN-Tags der jeweiligen Portgruppe direkt in den VMware Administrationswerkzeugen setzen.

Eine Ausnahme bildet hier der Nexus 1000v Softwareswitch; hier werden VLAN-IDs innerhalb des NX-OS von Cisco gesetzt und nicht mittels VMware Administrationswerkzeugen. Dies hat den Vorteil einer besseren Gewaltenteilung zwischen VMware Administratoren und Netz Administratoren.

Generell kann zwischen drei virtuellen Switcharten unterschieden werden:

- Standard vSwitch
- Verteilter Switch / distributed Switch (vDS)
- 3rd Party Switch, welcher die VMware vNetwork API benutzt (z.B. Cisco Nexus 1000v)

Der Vorteil bei einem verteilten virtuellen Switch (vDS) liegt in der zentralen Konfiguration über das vCenter. Die Konfiguration wird von dort aus auf die einzelnen ESX-Server verteilt (siehe die nachfolgende Abbildung 40). Damit kann innerhalb eines ESX-Clusters eine Fehlkonfiguration durch Konfigurationsdrift zwischen den ESX/ESXi-Servern vermieden werden. Zudem verfügt der vDS über weitergehende Funktionen wie z.B. Private VLANs aber auch QoS-Mechanismen, die der Standard VMware vSwitch nicht beherrscht.

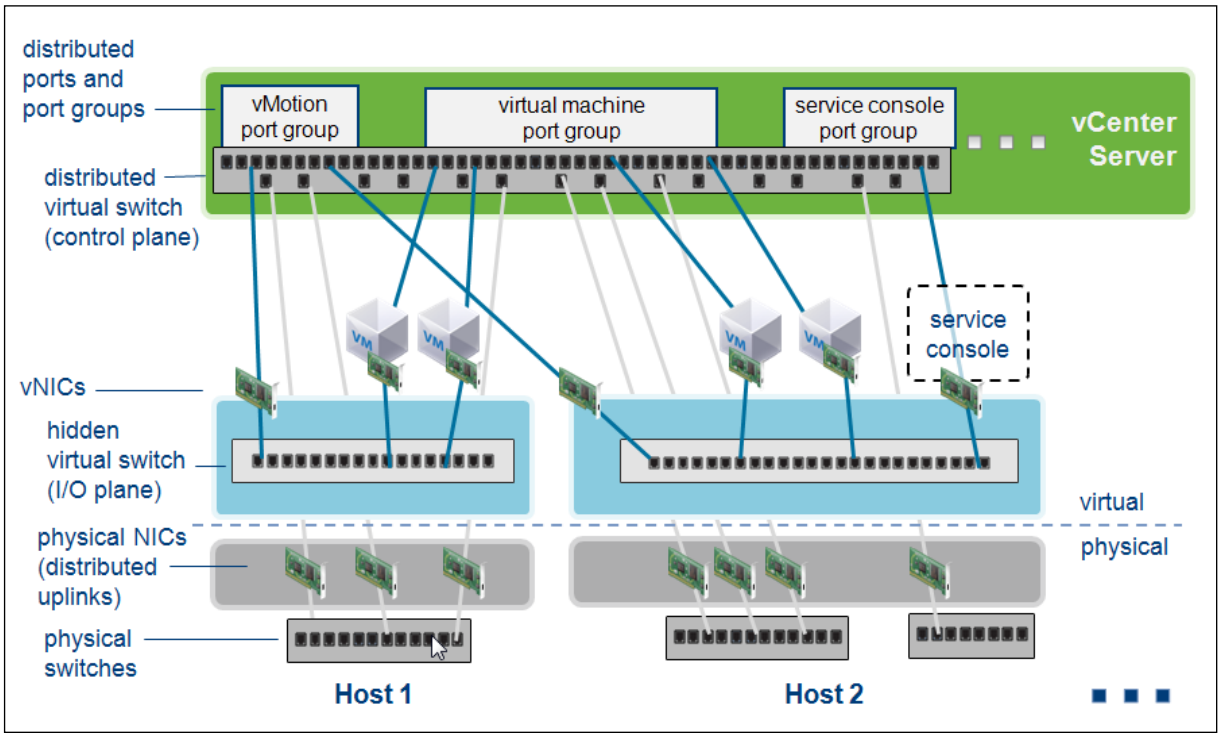


Abbildung 40: VMware Netz Stack - Beispiel mit Virtual Distributed Switch

Die ESX/ESXi-Server verfügen zusätzlich zu den vNICs der VM-Gäste über einige besondere Netzverbindungen, die ausschließlich vom VMkernel verwaltet werden. Hierzu zählen:

- VMkernel Interface für vMotion (Live Migration)
- VMkernel Interface für FT (Fault Tolerance Logging)
- VMkernel Interface für Management (nur ESXi)
- VMkernel Interface für Software iSCSI oder NFS

Die Absicherungen dieser internen Schnittstellen wird in Kapitel 4.2.1.3 behandelt.

Der VMware vSwitch und Distributed vSwitch unterscheiden sich von physischen Netz-Switchen an entscheidenden Punkten:

- VMware vSwitche lernen keine MAC Adressen. Jeglicher Netzverkehr der VM-Gäste wird direkt an einen Uplink Port, oder an einen VM-Gast auf demselben vSwitch und selben VLAN weitergegeben.
- VMware vSwitche nehmen nicht am Spanning-Tree teil.

D.h. der vSwitch selbst ist schon im Produkt Design immun gegen MAC-Flooding und Spanning-Tree Angriffe. Zusätzlich hierzu sollte der vSwitch so eingestellt werden (siehe nachfolgende Abbildung 41), dass er das Versenden von Ethernet Frames mit falscher Quell-MAC-Adresse unterbindet.

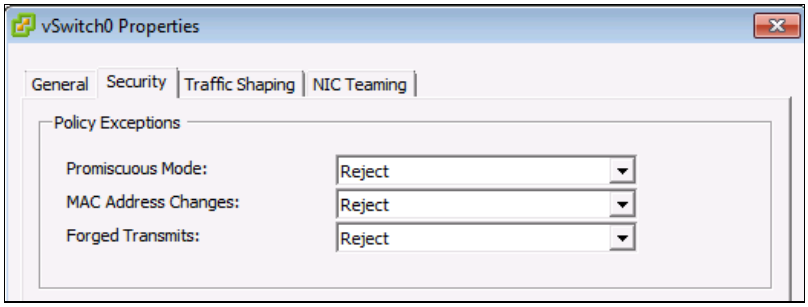


Abbildung 41: Einstellungen der Policy Exceptions bei einem vSwitch

Wie in der Abbildung dargestellt, sollte eingestellt werden:

- **Promiscuous Mode = "Reject"**
  Die Grundeinstellung "Reject" sollte beibehalten werden. Wird "Accept" eingestellt, so kann der VM-Gast den Netzverkehr im vSwitch mitlesen. Dieses wird nur durch Administratoren verwendet, um im Fehlerfall eine Fehleranalyse mit *packet sniffern* vorzunehmen.
- **MAC Address Changes = "Reject"**
  Ist "Accept" eingestellt (Grundeinstellung), so kann ein VM-Gast die MAC-Adresse seines virtuellen Adapters (vNIC) ändern.
- **Forged Transmits = "Reject"**
  Ist "Accept" eingestellt (Grundeinstellung), so kann ein VM-Gast Ethernet Frames mit beliebiger Quelladresse senden. Bei "Reject" werden Ethernet Frames mit falschen Quell-MAC-Adressen verworfen.

Die genannten Einstellungen zu "MAC Address Changes" und "Forged Transmits" verhindern effektiv MAC-Flooding und MAC-Spoofing Attacken.

Auch wenn der VMware vSwitch Spanning-Tree selbst nicht verwendet, werden BPDUs weitergeleitet. Um gegen die Gefährdungen vorzugehen, die von Spanning-Tree Attacken ausgehen, kann optional der Cisco Nexus 1000v eingesetzt werden. Über Features wie Root Guard und BPDU Guard, die im Kapitel 4.1.3.3 beschrieben werden, werden diese Angriffe unterbunden.
Wird der optionale Cisco Nexus 1000v nicht verwendet, so müssen auf der Switching Infrastruktur außerhalb des Vblocks entsprechend geeignete Massnahmen durchgeführt werden.

4.1.3.2 VLANs

Gefährdung 4:	Durch Fehlkonfiguration von Trunks in den Upstream-Netzkomponenten besteht die Gefahr des Überspringens von einem VLAN in ein anderes.

Gefährdung 5:	Durch eine unzureichende Konfiguration des "*Next Hop Routers*" (Default Gateways) können private VLANs vom Mandanten umgangen werden.

Die größten Risiken bestehen bei VLANs durch Fehlkonfigurationen. Diese Fehlkonfigurationen können unwissentlich oder wissentlich erfolgen. Die notwendigen Maßnahmen, um Fehlkonfigurationen zu verhindern oder nachweisen zu können, wenn sie doch erfolgt sind, werden in den Kapiteln 4.4.6 und 4.6 dieses Dokumentes beschrieben.

Zur Sicherheit von VLANs im Allgemeinen ist bereits viel in den letzten Jahren geschrieben worden. Eine gute Übersicht zu VLAN Sicherheitsaspekten findet sich in VLAN Security White Paper; [06].

Folgende Angriffsmöglichkeiten seien an dieser Stelle erwähnt:

Ausnutzen von "offenen" 802.1Q Trunks

Bei diesem Angriff gibt es zwei Ausprägungen: entweder ist bei der Konfiguration des Access-Ports an einem Netz-Switch ein 802.1Q VLAN Trunk fest konfiguriert worden, ohne die erlaubten VLANs einzuschränken, oder es wurde ein dynamisches Trunk Aushandlungs-Protokoll (DTP) verwendet. Im ersten Fall ist es klar, dass ein Angreifer leicht in ein fremdes VLAN kommt, indem er einfach den richtigen 802.1Q VLAN Header verwendet.

Im zweiten Fall muss der Angreifer mit Netz-Switchen das dynamische Trunk Protokoll sprechen und einen 802.1Q VLAN Trunk mit dem Switch aushandeln. Anschließend kann er wie im ersten Fall den Netzverkehr mit den richtigen VLAN Header versehen, um in das fremde VLAN zu gelangen.

Double-Encapsulated 802.1Q / Nested VLAN Attacke

Ein 802.1Q VLAN Trunk hat sogenannte "tagged" VLANs und ein "natives" VLAN. Das native VLAN verwendet keinen 802.1Q Header, d.h. der Netzverkehr im nativen VLAN des Trunk Ports verhält sich wie regulärer Ethernet Netzverkehr. Im nativen VLAN werden auch die Managementprotokolle wie VTP (VLAN Trunking Protocol), CDP (Cisco Discovery Protocol), DTP (Dynamic Trunking Protocol) transportiert.

Der Angreifer kann dieses native VLAN verwenden, um ein entferntes System anzugreifen.

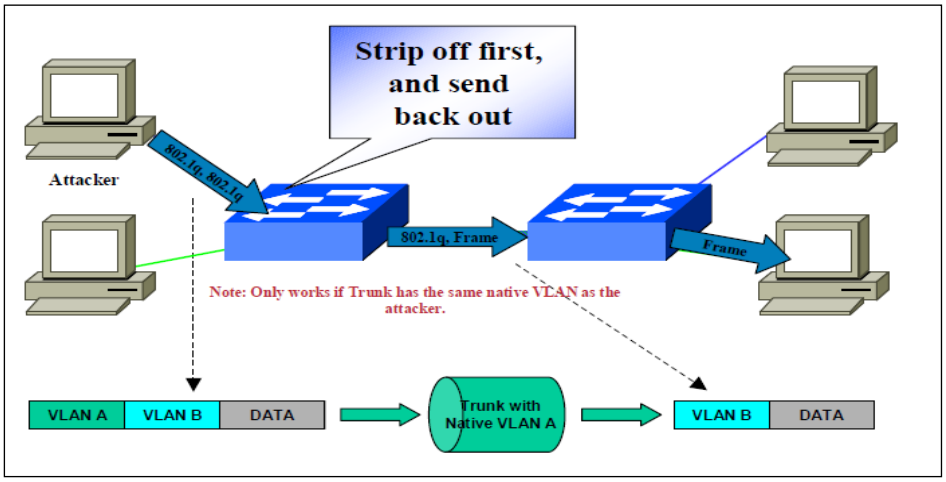


Abbildung 42: Nested VLAN Attacke

Wie die Abbildung 42 zeigt, sendet der Angreifer Frames mit zwei 802.1Q VLAN Headern (daher der Name "Double-Encapsulated VLAN Attacke"). Der erste Switch entfernt den ersten VLAN Header und leitet den Frame inklusive dem zweiten noch vorhandenen VLAN Header normal weiter. Ist nun das native VLAN auf dem Trunk zwischen zwei Switches dasselbe VLAN, das der Angreifer im ersten Header verwendet hat, so wird der zweite Switch einen Frame mit dem noch vorhandenen zweiten Header bekommen. Der zweite Switch wird diesen VLAN Header akzeptieren, entfernen, und den Frame ohne Header an das Ziel mit der richtigen MAC Adresse senden.

Eine Rückantwort kann es bei diesem Angriff nicht geben, trotzdem ist das potentielle Risiko einer DoS-Attacke gegeben.

Damit dieser Angriff möglich ist, muss auf dem Netz-Port, der dem Angreifer zur Verfügung steht, ein 802.1Q VLAN Trunk zur Verfügung stehen. Üblicherweise würde bei einem Access-Port der Switch den 802.1Q Header bzw. den ganzen Frame verwerfen. Es gibt hier allerdings Ausnahmen. So gibt es Switche, die 802.1Q Header an Access-Ports erlauben, wenn entweder VLAN ID = 0 oder VLAN ID = VLAN ID des Access-Ports verwendet wird. Zusätzlich muss auf dem 802.1Q VLAN Trunk zu dem Angreifer das gleiche VLAN verfügbar sein, welches als natives VLAN auf den Trunks im Backbone zwischen den Switchen konfiguriert ist.

Umgehen von Private VLANs über das Default Gateway – Next Hop

Private VLANs erzeugen eine Isolation auf Layer 2 Ebene, d.h. Systeme, die sich in Isolated VLANs oder in verschiedenen Community VLANs befinden, haben keine Möglichkeit direkt über Layer 2

miteinander zu kommunizieren. Durch mangelnde Sorgfalt bei der Konfiguration im Next Hop Router, der sich außerhalb des Vblocks befindet, könnte ein Angreifer einen fremden VM-Gast von seinem Isolated Port aus erreichen.

Der Angriff funktioniert wie in Abbildung 43 dargestellt. Der Angreifer sendet ein IP-Paket mit der Ziel-MAC-Adresse vom Default Gateway im Ethernet Frame Header sowie der Ziel-IP-Adresse vom Opfer im IP-Header. Das Default-Gateway fällt eine Entscheidung anhand seiner IP-Routing Tabelle und sendet das IP-Packet zurück zum Opfer, über dasselbe Interface auf dem es das IP-Paket empfangen hat. Auch dieser Angriff funktioniert nur in eine Richtung, reicht aber für eine DoS-Attacke aus.

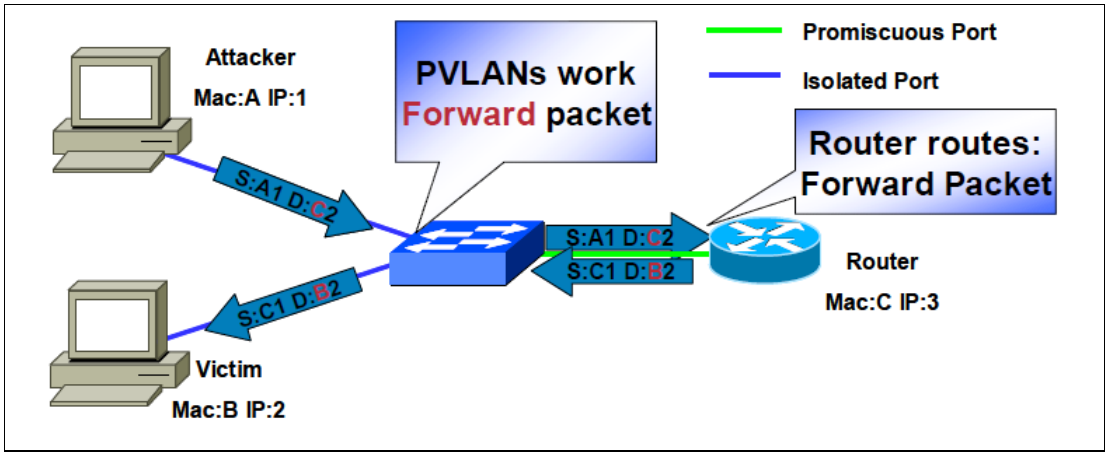


Abbildung 43: IP Next Hop (Default Gateway)

Nachfolgend wird die Verwendung und Konfiguration von VLANs im Vblock erklärt und anschließend die Gegenmaßnahmen gegen diese Angriffsmöglichkeiten anhand von Best Practices aufgezeigt.

Um dem ESX Virtual Switch (vSwitch) die notwendigen VLANs für die VM-Gäste zuzuführen, wird im Normalfall *802.1Q VLAN Tagging* verwendet. Eine andere Methode ist das Verwenden von multiplen realen oder virtuellen Ethernet Adaptern, d.h. ein Adapter pro zugeführtem VLAN. Werden VLAN Trunks zum vSwitch geführt, werden auf den virtuellen Ethernet Karten die "erlaubten" VLANs konfiguriert. Außerdem wird normalerweise ein VLAN konfiguriert, welches als "natives" VLAN ohne Tagging auf der virtuellen Ethernet Karte anliegt.

Die im UCS-System zur Verfügung stehenden VLANs werden vom Administrator oder Teilbereichsexperten für Netze im UCS Manager (UCSM) definiert, und es wird dem VLAN ein Name innerhalb des Systems vergeben. Durch das Anlegen des VLANs im UCSM wird das VLAN auch auf den Uplinks zum übergeordneten LAN ausserhalb des Vblock erlaubt, so dass vom/zum LAN-Netzverkehr in diesem VLAN über 802.1Q Trunking übertragen wird. Dieses wird in der nachfolgenden Abbildung 44 dargestellt.

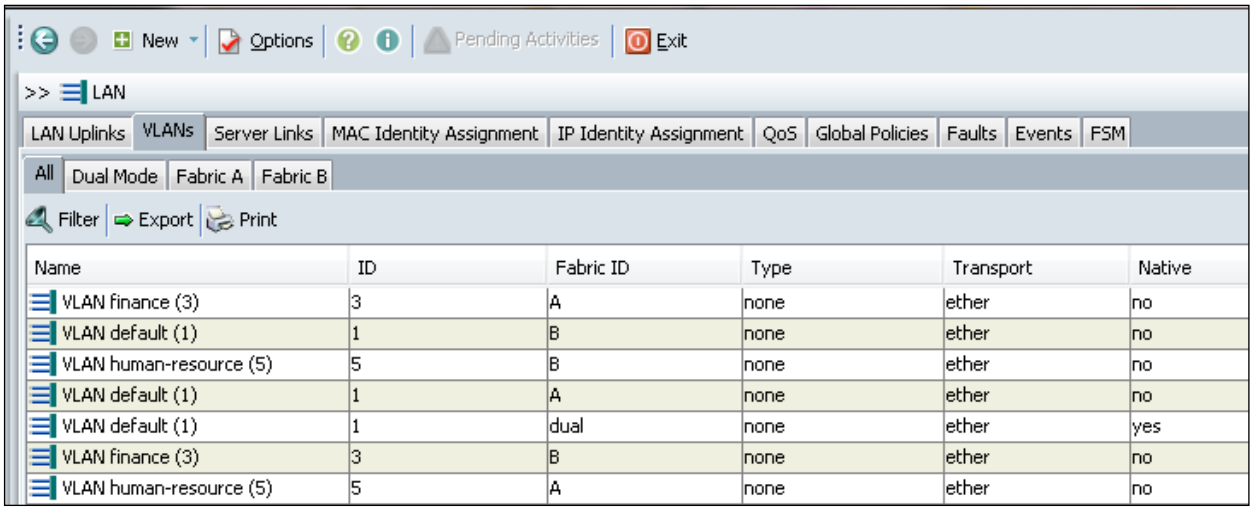


Abbildung 44: UCSM: VLANs

Im Service-Profile wird dann in der Konfiguration der virtuellen Ethernet Karten angegeben, welche VLANs zum Hypervisor transportiert werden sollen, und welches VLAN das "native" (un-tagged) VLAN ist (siehe nachfolgende Abbildung 45).

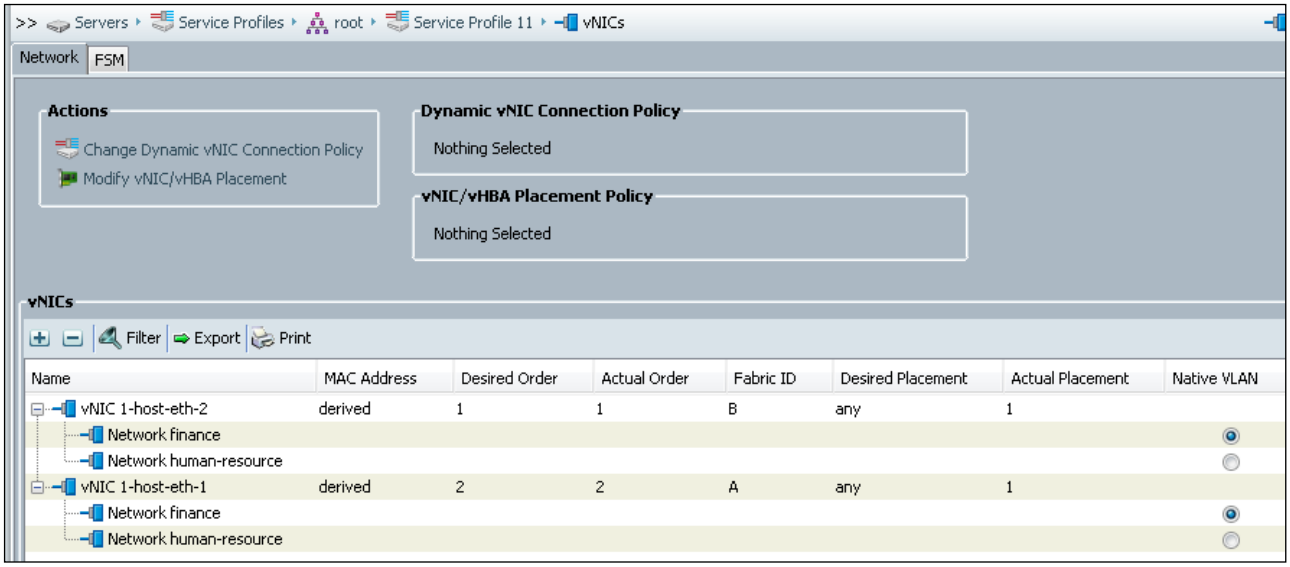


Abbildung 45: UCSM: vNICs (M81KR Virtual NIC)

Beim Einsatz der optionalen Netzkarte UCS M81KR Virtual NIC, die VN-Link verwendet, ist es aufgrund der hohen Anzahl von derzeit 56 virtuellen Netzkarten möglich, für jedes VLAN eine eigene virtuelle Netzkarte in den Hypervisor zu führen. Dabei wird üblicherweise jedes dieser VLANs als "natives" VLAN (un-tagged) definiert (siehe nachfolgende Abbildung 46).

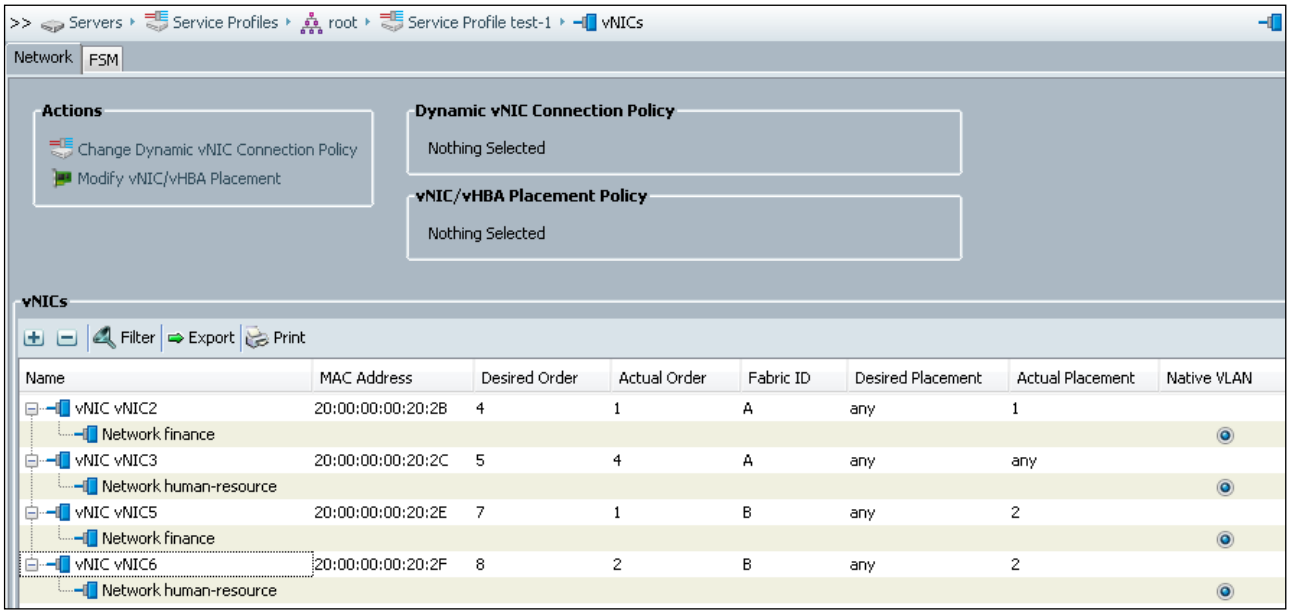


Abbildung 46: UCSM: vNICs (M81KR)

Es ist gleichwohl aber auch bei der M81KR Netzkarte möglich, pro virtueller Ethernet Karte mehrere VLANs zuzuführen.

Folgende "Best Practices" verhindern die Angriffsmöglichkeiten Ausnutzen von "offenen" 802.1Q Trunks und Double-Encapsulated 802.1Q / Nested VLAN Attacke:

- **VLAN ID 1**
  Das VLAN mit der ID 1 hat dadurch eine besondere Bedeutung, da Netz-Switche ab Werk üblicherweise mit einer Konfiguration ausgeliefert werden, die nur das VLAN mit der ID 1 beinhaltet und alle Access-Ports mit dem "nativen" VLAN in das VLAN 1 legt. Weiterhin gibt es eine Reihe von

Management Protokollen wie CDP, LLDP (Link Layer Discovery Protocol), LACP (Link Aggregation Control Protocol), VTP, etc., die ihren Netzverkehr im VLAN 1 übertragen.
Aus diesem Grund wird empfohlen, das VLAN 1 nicht für regulären Netzverkehr zu verwenden, und folgende Maßnahmen vorzunehmen:

- VLAN1 sollte nicht für "In-Band" Management Netzverkehr der Komponenten verwendet werden. Es sollte hierfür ein eigenes Management VLAN verwendet werden.
- VLAN1 sollte nicht zu Access-Ports erlaubt werden; in unserem Fall im Vblock sollte VLAN1 also nicht auf den virtuellen Ethernet Karten der Blade-Servern anliegen.

- **Entfernen von nicht benötigten VLANs zu Access-Ports (vEth Ports der Blade-Server)**
  Jegliche nicht für einen VM-Gast-, oder für Management- und VMotion-Verkehr des ESX-Hypervisors benötigten VLANs sollten nicht an den virtuellen Ethernet Karten der Blade Server anliegen.
- **Unterschiedliches "natives" VLAN für 802.1Q Trunks und Access-VLANs**
  Das native VLAN eines Trunks zwischen Switchen, also z.B. vom Fabric Interconnect zum übergeordneten LAN, sollte niemals ein VLAN sein, das auch als VLAN an einer der virtuellen Ethernet Karten der Blade-Server verfügbar ist. Im besten Fall ist das native VLAN von Trunk Verbindungen zu entfernen (d.h. alle VLANs verwenden ein VLAN Tag); alternativ kann als natives VLAN ein VLAN definiert werden, das nirgendwo anders genutzt wird. Dieses verhindert die beschriebene Double-Encapsulated 802.1Q / Nested VLAN Attacke.

Folgende "Best Practice" verhindert die Angriffsmöglichkeit Umgehen von Private VLANs über das Default Gateway:

So einfach der Angriff ist, so einfach ist auch die Gegenmaßnahme. Eine IP-Access-Liste auf dem Router-Interface kann eingesetzt werden, die es verbietet, Netzverkehr, der aus dem lokalem Subnetz kommt, wieder in das gleiche Subnetz zu senden.

Zusätzlich dazu sollte "Unicast Reverse Path Forwarding Check" (URPF) im "strict mode" eingesetzt werden, um zu verhindern, dass diese IP-Access-Liste umgangen wird, indem die Quell-IP-Adresse gefälscht wird. URPF wird Netzverkehr, dessen Quell-IP-Adresse nicht zum Subnetz des empfangenden Interface passt, verwerfen. Weitere Details hierzu siehe Understanding Unicast Reverse Path Forwarding; [07].

4.1.3.3 Nexus 1000v - zusätzliche Sicherheit

Gefährdung 6:	Im Übergang zwischen Server- und Netzadministration kann es zu Fehlern bei der Zuordnung von VLAN zu Portgruppen kommen.

Im Vblock gibt es einen Übergang der Aufgaben zwischen dem Teilbereichsexperten für den Server/Hypervisor und dem Teilbereichsexperten für das Netz. An diesem Übergang müssen die richtigen VLAN-IDs ausgehandelt werden. So kann z.B. VLAN 1034 eine VLAN-ID von Mandant 1 sein, und VLAN-ID 1234 die VLAN-ID von Mandant 2. Verwendet der Server/Hypervisor Teilbereichsexperte nun die falsche VLAN-ID, so kann das dazu führen, dass ein potentieller Angreifer aufgrund dieser Fehlkonfiguration den Zugang in ein VLAN eines fremden Mandanten bekommt.

Gefährdung 7:	Layer-2-Attacken wie Rogue DHCP-Server, GARP Attacken, IP und MAC Address Spoofing, MAC Flooding und Spanning Tree-Angriffe.

Zusätzlich zu den im vSwitch enthaltenen Sicherheitsmaßnahmen gegen MAC Flooding und seine Immunität gegen Spanning Tree-Angriffen gibt es eine Reihe von bekannten Layer-2-Attacken, die nicht vom vSwitch verhindert werden. Diese können durch den Einsatz des optionalen Nexus 1000v adressiert werden. Vor allem in virtuellen Desktop Umgebungen kann damit die Sicherheit stark erhöht werden.

Auch wenn diese Art von Angriffen gut bekannt und vielfach dokumentiert ist, werden sie im Kontext des Vblocks im Folgenden kurz angerissen:

- ***Rogue DHCP Server***
  Startet ein auf DHCP eingestelltes System, in unserem Fall ein VM-Gast, neu oder läuft seine *DHCP lease time* ab, so sendet es einen Broadcast aus seinem Ethernet Adapter und fragt nach einem neuen DHCP lease. Dabei erhält er als Antwort seine IP-Adresse, aber auch Parameter, wie das Default-Gateway, DNS Server, usw.
  Ein Angreifer, der sich im selben L2 VLAN / Subnetz wie sein Opfer befindet, kann diesen DHCP Request oder Discover nun abfangen und beantworten bevor der eigentliche DHCP Server antwortet. Da der DHCP Server meist mehrere IP-Hops vom Opfer entfernt ist, antwortet der Angreifer mit hoher Wahrscheinlichkeit auch schneller als der echte DHCP-Server.
  Bei der DHCP-Antwort trägt sich der Angreifer mit seiner IP-Adresse als Default Gateway ein. Fortan sendet das Opfer nun jeglichen Netzverkehr, der zu Zielen außerhalb seines lokalen Subnetzes gesendet wird, zum Angreifer. Der Angreifer sendet den Netzverkehr dann weiter zum Default Gateway. Damit ist der Angreifer jetzt in der Lage, sämtliche Kommunikation zwischen dem Opfer und dem echten Default Gateway mitzuschneiden.
  Im Vblock ist eine solche Attacke vor allem bei Virtual Desktop Anwendung innerhalb eines Mandanten denkbar, da diese Attacke nur innerhalb eines VLANs / Subnetzes funktioniert. Im Virtual Desktop Bereich kann ein Innentäter versuchen mit diesem Angriff den Netzverkehr im gleichen VLAN / Subnetz mitzuschneiden.

- **GARP Attacken**
  Bei *Gratuitous ARP* (GARP) Attacken sendet der Angreifer unaufgeforderte ("gratuitous") ARP-Antworten an bestimmte Opfer, oder an alle Systeme im gleichen VLAN / Subnetz. In dieser gefälschten ARP-Antwort trägt der Angreifer seine MAC-Adresse als Zuordnung zu einer fremden IP-Adresse ein und bringt das Opfer dazu seine ARP-Tabelle so zu verändern, dass der Netzverkehr nun zum Angreifer geht anstatt zum validen Ziel gesendet wird.
  Möchte ein Angreifer z. B. den Netzverkehr zwischen zwei Opfern abhören, so sendet er GARPs zu beiden Opfern, und trägt seine MAC-Adresse als Zuordnung zu den IP-Adressen der Opfer ein. Fortan empfängt er den Netzverkehr zwischen den Opfern und leitet diesen zwischen den Opfern weiter.
  Auch bei diesem Angriff kann eine solche Attacke vor allem bei Virtual Desktop Anwendung innerhalb eines Mandanten für den Angreifer interessant sein, da diese Attacke nur innerhalb eines VLANs / Subnetzes funktioniert.

- ***IP Address Spoofing***
  Bei diesem Angriff versucht der Angreifer IP-Access-Listen bzw. Firewall Regeln zu umgehen, indem er sich eine falsche IP-Identität, d.h. eine falsche Quell-IP-Adresse gibt. Damit kann er dann eine DoS-Attacke gegen ein System zu fahren, das er anhand von IP-Access-Listen bzw. Firewall-Regeln eigentlich nicht erreichen sollte.

- ***MAC Address Spoofing***
  Bei diesem Angriff verwendet ein Angreifer die MAC-Adresse seines Opfers, um den Netzverkehr des Opfers an sich zu ziehen. Netzswitche lernen die Zuordnung der MAC-Adresse zum Port über die Quell-MAC-Adressen des empfangenen Verkehrs. Wenn ein Angreifer eine gefälschte MAC-Adresse verwendet, so ändert der Switch diese Zuordnung und sendet den Verkehr anstatt zum Opfer zum Angreifer. Auch in diesem Fall handelt es sich um einen lokalen Angriff im gleichen VLAN / Subnetz, der vor allem bei Virtual Desktop Anwendung innerhalb eines Mandanten für den Angreifer interessant sein kann.

- ***MAC Flooding***
  (siehe Kapitel 4.1.3.1).

- ***Spanning-Tree-Attacken***
  (siehe Kapitel 4.1.3.1).

Eine Lösungsmöglichkeit sowohl gegen die Gefahr durch Fehler bei der Übergabe von VLAN-IDs zwischen dem Server/Hypervisor Teilbereichsexperten und dem Netz-Teilbereichsexperten als auch gegen die Gefahren durch die beschriebenen Layer-2-Netzattacken bietet der Nexus 1000v.

Der **Nexus 1000v** ist eine optionale Komponente des Vblock. Der Nexus 1000v ersetzt in den ESX-Servern den vSwitch mit virtuellen Ethernet Modulen (VEM) und verwaltet diese VEMs über Virtuelle Supervisor Module (VSM) (Active/Standby).

Vereinfacht gesagt werden die VEMs zu Schnittstellenkarten des VSMs. Das VSM kann auf einer eigenen Appliance installiert werden oder als virtuelle Maschine auf einem ESX-Server.

Das VSM steht mit dem VMware VCenter Server in Verbindung, um Port-Profile im VSM und Port-Gruppen Namen im VMware vCenter Server zu synchronisieren (siehe Abbildung 47).

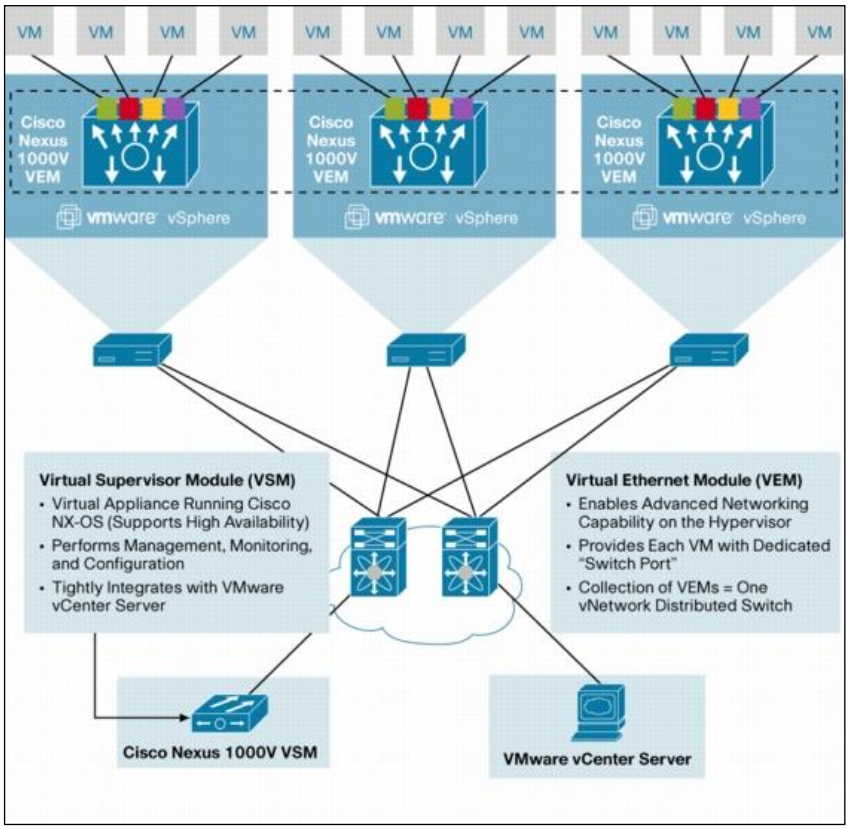


Abbildung 47: Cisco Nexus 1000v

Wird im vCenter für einen VM-Gast eine Port-Gruppe ausgewählt, die an das VSM gebunden ist, so wird auf dem VSM beim Start des VM-Gasts ein neuer virtueller Ethernet Port angelegt, dessen Einstellungen vom Port-Profile vererbt werden. Nachfolgend ein beispielhafter Auszug aus den virtuellen Ethernet Ports eines Nexus 1000v:

```
dcv-n1kv# sh interface virtual
--------------------------------------------------------------------------------
Port        Adapter        Owner                     Mod Host
--------------------------------------------------------------------------------
Veth1       vmk0           VMware VMkernel           3   ucs-esx1.fra-lab.net
Veth2       vmk1           VMware VMkernel           3   ucs-esx1.fra-lab.net
Veth3       Net Adapter 1  UCS-WS-VC                 6   ucs-esx4.fra-lab.net
Veth4       Net Adapter 1  Win2003-File-Web-SQL      4   ucs-esx3.fra-lab.net
Veth5       Net Adapter 2  Win2003-DC,DNS            4   ucs-esx3.fra-lab.net
Veth6       Net Adapter 1  Win2003-DC,DNS            4   ucs-esx3.fra-lab.net
Veth7       Net Adapter 1  DCV-N1KV-PROD-VSM-HA      3   ucs-esx1.fra-lab.net
```

Hier z.B. ist das Veth4 Interface mit dem VM-Gast `Win2003-File-Web-SQL` auf dem ESX-Server `ucs-esx3.fra-lab.net` verbunden. Ganz gleich, wohin die Maschine durch vMotion bewegt wird, die Zuordnung zu dem VM-Gast wandert innerhalb des ESX-Clusters mit.

Der Teilbereichsexperte für das Netz definiert in den Port-Profilen auch das VLAN, bzw. bei 802.1Q VLAN Trunks die erlaubten VLANs. Die Port-Profile werden einem Port-Group Namen im vCenter zugeordnet. D.h. der Server/Virtualisierungs Teilbereichsexperte wählt für den virtuellen Netzadapter eine Port-Gruppe anhand eines sprechenden Namens wie z.B. "DMZ Netz Mandant 1" aus. Damit sind Fehlkonfigurationen zwar nicht mehr ausgeschlossen, aber die Wahrscheinlichkeit von Fehlkonfigurationen wird vermindert.

Durch die Vererbung von Regeln aus den Port-Profilen im Nexus 1000v heraus werden auch Netz-Policies auf VMware vCenter Port-Gruppen automatisch zugeordnet. Für Port-Profile können die Standard Leistungsmerkmale, wie sie auf physischen Cisco Switche vorhanden sind, an die virtuellen Ethernet Ports eines VM-Gasts gehängt werden. Dazu gehören unter anderem die Leistungsmerkmale der Switche der Catalyst Serien zum Schutz gegen die bei den Gefährdungen beschriebenen Layer-2-Attacken:

- ***DHCP Snooping***:
  Über DHCP-Snooping können rogue DHCP Server Attacken unterbunden werden. Dabei wird auf dem Nexus 1000v definiert, ob der Port ein Uplink Port oder ein Host Port ist. Auf Host Ports unterbindet DHCP-Snooping DHCP-Offers und verhindert damit, dass ein VM-Gast innerhalb eines VLANs vorgibt ein DHCP-Server zu sein. Damit werden Man-in-the-Middle Attacken durch rogue DHCP-Server effektiv verhindert.
- ***Dynamic ARP Inspection (DAI)***:
  Über DAI werden ARP-Requests und -Responses validiert. DAI überprüft ob eine ARP-Response auf einem Access-Port eine valide Kombination von IP-zu-MAC Binding aufweist. Dabei wird überprüft, ob die Information im ARP-Response zu der IP-Adresse des sendenden Endgerätes passt. Dazu wird die DHCP-Snooping-Binding-Tabelle verwendet. DAI verhindert die beschriebenen GARP Attacken.
- ***IP Source Guard***:
  Mit IP Source Guard überprüft der Nexus 1000v, ob die Quell-IP-Adresse eines Paketes zum Endgerät passt. Hierzu wird wieder die DHCP-Snooping-Binding-Tabelle verwendet. Sollte ein Endgerät eine falsche Quell-IP verwenden, wird das Paket verworfen.
- ***Port Security***:
  Über Port-Security kann z.B. die maximale Anzahl von MAC-Adressen an einem Port überprüft und Massnahmen ergriffen werden, wenn diese Anzahl überschritten wird. Port Security wird gegen MAC Flooding Attacken und MAC Spoofing eingesetzt.
- ***BPDU Guard und Root Guard***
  BPDU Guard wird auf dem virtuellen Ethernet Port zu den VM-Gästen konfiguriert. Sendet ein VM-Gast unberechtigterweise eine BPDU, so wird der virtuelle Ethernet Port abgeschaltet und eine Security Violation Warnung wird protokolliert.
  Zusätzlich wird Root Guard auf den Uplink Interfaces des Nexus 1000v zu den übergeordneten Netzkomponenten konfiguriert. Sollte der Nexus 1000v auf irgendeinem anderen Port als dem konfigurierten Port eine BPDU empfangen, die eine andere Root-Bridge mit höherer Priorität angibt, so wird dieser Port abgeschaltet und eine Security Violation Warnung wird protokolliert.

Neben diesen auf Switching Attacken ausgelegten Eigenschaften, bietet der Nexus 1000v auch viele weitere wichtige Tools wie Private-Vlans mit Isolated und Community Ports, Switch Port Analyser (SPAN), RMON und Netflow, etc.

Weitere Details zum Nexus 1000v in VMware ESX-Umgebungen auf UCS siehe Best Practices in Deploying Cisco Nexus 1000v Series Switches on Cisco UCS B Series Blade Servers; [08] und Cisco Nexus 1000v Security Configuration Guide; [09].

4.1.4 Storage Isolation und Fibrechannel Security

Unter *Storage Isolation* versteht man die Abgrenzung von SAN-Speicherressourcen gegenüber den zahlreichen Konsumenten dieser Speicherbereiche. Der Zugriff kann in unterschiedlichen Bereichen gesteuert werden.

Gefährdung 8:	Zugriff auf Datenträger anderer Mandanten durch WWN-Spoofing bei unzureichender Konfiguration.

4.1.4.1 Zugriffssteuerung im SAN

Der Zugriff auf das Speichernetz kann einerseits mittels VSANs geregelt werden (siehe Kapitel 3.3.2). Andererseits gibt es den Mechanismus des *Zonings* als traditionellen Weg, den Zugriff auf im Netz verfügbare Ressourcen einzuschränken. Hier werden Zugriffe von Initiatoren (Hosts) auf sogenannte Targets (Speichersysteme, Tape Libraries, etc.) beschränkt. Die Sichtbarkeit der Ressourcen für die Netzteilnehmer wird auf die freigegebenen Ressourcen beschränkt. Alle nicht gezonten Kommunikationswege werden ausgeblendet.

In diesem Bereich unterscheidet man zwei unterschiedliche Ansätze: das Port-Zoning und das WWN-Zoning (WWN = World Wide Name)

- Beim **Port-Zoning** erfolgt die Definition der erlaubten Kommunikationswege über die Gruppierung von Ports im SAN. Der erzielte Effekt ist, dass diese miteinander kommunizieren dürfen, unabhängig davon welche Geräte an den Ports angeschlossen sind.
- Das **WWN-Zoning** bildet die Kommunikationsgruppen anhand der weltweit eindeutigen Identifikatoren der Kommunikationspartner (vgl. MAC-Addresse im Ethernet Umfeld).
  Der administrative Vorteil dieses Verfahrens besteht in der Flexibilität, bei einem defektem Anschluss-Port am SAN einfach auf einen funktionsfähigen freien Port ausweichen zu können.

Die Angriffsmöglichkeiten bei Verwendung der einzelnen Verfahren unterscheiden sich demnach wie folgt:

- Bei der Nutzung des WWN-Zonings kann ein Angreifer mittels WWN-Spoofing (das Imitieren einer fremden WWN) Zugang zu der Kommunikationgruppe und somit Zugriff auf die für diese WWN freigegebenen Netzressourcen erlangen.
- Dem WWN-Spoofing beugt das Port-Zoning vor. Da hier die Kommunikation mithilfe der Ports definiert ist, können Angriffe mittels WWN-Spoofing hier nicht gelingen. Das Risiko dieser Methode besteht in der Tatsache, dass sofern sich der Angreifer physischen Zugang zum Switch verschaffen und er sich am entsprechenden Port in das Netz einhängen kann, er automatisch Zugang zu allen für diesen Port freigegebenen Ressourcen erhält.

Die Kombination beider Verfahren – WWN-Zoning mit **Port-Lockdown** – verhindert beide Angriffsmöglichkeiten, bedingt aber gleichzeitig einen wesentlich höheren administrativen Aufwand.

4.1.4.2 Zugriffsbeschränkung auf Ebene der Speichersysteme

Nach erfolgter Einschränkung der Kommunikationswege durch das Netz kann der Speicherzugriff auf Ebene des Speichersystems weiter eingeschränkt bzw. überhaupt erst freigegeben werden.

Im Fibre Channel Netz werden am Speichersystem nur die Datenträger oder LUNs für angemeldete WWNs (Initiatoren, Hosts) freigegeben. Alle nicht freigegebenen Ressourcen werden ausgeblendet. Dieses Verfahren wird "*Masking*" genannt.

Auch hier droht die Gefahr des WWN-Spoofings. Sofern sich ein Angreifer mit der WWN eines berechtigten Systems Zugang zu dem Speichersystem verschafft, hat er Zugriff auf die für diese WWN freigegebenen Ressourcen. Einige Speichersysteme können diesem Angriff mittels *FCID-Lockdown* begegnen. Die FCID ist ein beim Fabric-Login jeder WWN-/Portkombination zugewiesener, innerhalb des Netzes (Fabric) eindeutiger Identifikator. Damit ist es unmöglich, dass zwei Initiatoren – selbst mit der gleichen WWN – die gleiche FCID erhalten. Das Speichersystem legt neben der

Berechtigungsinformation der WWN auch die berechtigte FCID in der Zugriffsdatenbank ab. Somit wird sichergestellt, dass Angreifer, die eine erfolgreiche WWN-Spoofing-Attacke im Netz geschafft haben, dennoch keinen Zugriff auf Datenträger anderer Systeme erhalten.

Die Möglichkeiten einer Zugriffsbeschränkung für die einzelnen im Vblock verwendeten Speichersysteme stellt folgende Auflistung dar:

Vblock 1/1U und Series 300 (CLARiiON, VNX)

- iSCSI CHAP, IPSec
- LUN Masking (Storage Groups, WWN oder IQN (iSCSI Qualified Name) basierend)

Vblock Series 700 (Symmetrix)

- iSCSI CHAP, DH-CHAP (Diffie Hellman – Challenge Handshake Authentication Protocol), IPSec
- LUN Masking (WWN oder IQN basierend; FCID locking)
- Data-at-rest Encryption

4.1.4.3 Fibre Channel Best Practices

Gefährdung 9:	Erlangen eines physischen Zugriffs auf den SAN-Switch mit der Möglichkeit Spoofing- und DoS-Attacken durchzuführen.

Im Vblock erfüllen die zwei MDS 9000 Switche die Funktion der redundanten Fibre Channel Fabrics. Im Normalfall sind die Switche nur an die Systeme im Vblock angeschlossen. Nur in Ausnahmefällen können die Switche, z.B. zur Migration von Daten von einem bestehenden SAN-Storage, an ein übergeordnetes SAN angeschlossen werden. Trotzdem gelten auch für die im Vblock verbauten MDS 9000 Switche die gleichen Best Practices zum Absichern der Fibre Channel Switche, um u.a. folgende Angriffsmöglichkeiten auszuschließen:

- Ein Angreifer schließt fremde SAN Switche an den MDS 9000 an, mit der Absicht z.B.
  - auf die verteilte Zoning Datenbank zuzugreifen und diese so zu verändern, dass er Zugriff auf die Speichersysteme bekommt
  - sich in den Datenstrom einzuschleifen, indem er z.B. das FC-Routing verändert, mit der Absicht Daten mitzuschneiden
  - die *Name Server Database* so zu verändert, dass er sich als Target darstellt und damit die Daten der Initiatoren bekommt und an den echten Target weiterleitet. Dieses wieder in der Absicht Daten mitzuschneiden
- Ein Angreifer versucht eine DoS-Attacke auf die Fabric, indem er z.B. *Reconfigure Fabric (RCF) Nachrichten* an den MDS 9000 Switch sendet, mit der Absicht den Netzverkehr zu stören und damit z.B. "*Blue Screens*" und/oder *Kernel Panics* bei den VM-Gästen und Hypervisor Systemen auszulösen.

Folgende Best Practices sollten zur Erhöhung der Sicherheit und Integrität des SANs angewendet werden. Der MDS 9000 Switch ist entsprechend zu konfigurieren:

- Unbenutzte Ports auf den Fibre Channel Switches deaktivieren und Default-Einstellung auf "deaktiviert" setzen. Dies verhindert, dass ein Angreifer unbemerkt "offene“" Ports am FC-Switch ausnutzt, um in die Fabric zu kommen.
- Default Port Typ von "Auto" auf "F-Port" und default Trunk Einstellung auf "off" umstellen. Damit können keine Interswitch-Links aufgebaut werden. Dies verhindert dass der Angreifer über offene E-Ports schreibenden Zugriff auf die Zoning-, Name-, Routing-Database bekommt.
- Port Security verwenden, und alle WWNs aus den WWN-Pools dediziert auf den Eingangsport vom Fabric-Interconnect erlauben. Dies ist eine erste Maßnahme, um zu verhindern, dass ein Angreifer an einem Port ein Kabel absteckt und einfach einen fremden HBA anschließt. Zumindest muss der

Angreifer dann über das Wissen verfügen, welche WWNs an dem Port erwartet werden, und eine der WWNs spoofen.

- Fabric Binding mit Static Domain IDs verwenden. Dies verhindert das Anschließen von fremden FC-Switchen, zumindest solange der Angreifer nicht die in der Fabric vorhandenen WWNs der Switche und die verwendeten Domain-IDs kennt.
- Principle Switch vorgeben über Priorities.

Details zur Konfiguration sind zu finden im
Cisco MDS 9000 Family NX-OS Security Configuration Guide; [10].

Abschließend sollte noch erwähnt werden, dass alle in diesem Kapitel beschriebenen Angriffe einen **physischen** Zugriff zum Vblock voraussetzen.

Die Hauptgefährdung durch remote Angriffe besteht durch ein WWN Spoofing bei den Blade Servern. In Kapitel 4.1.4 ist beschrieben, wie die WWN Pools der Blade Server konfiguriert werden, und wie verhindert wird, dass ein WWN Spoofing erfolgt.

4.2 Unzureichender Schutz interner Managementnetze

Beim unzureichenden Schutz der Vblock internen Managementnetze (Administrationszugriffe und interne Interfaces zwischen den Komponenten) können Angreifern diverse Angriffsmöglichkeiten ermöglicht werden, die in diesem Kapitel erläutert werden.

Die im Vblock eingesetzten Hard-/Software Komponenten haben untereinander diverse Kommunikationswege (Heartbeats, State Synchronisation, usw.), sowie Interfaces zur Administration der Komponenten. Werden diese internen Managementnetze und Administrationsinterfaces unzureichend geschützt, so werden eine Reihe von Angriffen möglich, wie z.B.

- Mitschneiden der State Synchronisation zwischen den Hypervisoren über eine Man-in-the-Middle-Attacke. Dabei könnte ein Angreifer z.B. Speicherinhalte mitschneiden in denen sich Datenbankdaten befinden.
- Ausnutzen offener Dienste auf den Management-Interfaces, um Daten abzugreifen oder Konfigurationsänderungen vorzunehmen.
- Ausnutzen von Diensten auf den Management-Interfaces, um evtl. Vulnerabilities der Softwarekomponenten des vBlocks für Exploits auszunutzen oder um einfache DoS Attacken zu starten ("exploit of vulnerabilities" = Ausnutzen von Schwachstellen im Code).
- Mitschneiden von unverschlüsselten Managementverbindungen z.B. über Telnet, http mit dem Ziel Passwörter mitzuschneiden.
- Man-in-the-Middle-Attacke auf die Fibre Channel over Ethernet (FCoE) Storage Kommunikation zwischen Hypervisor und Storage Subsystem, mit dem Ziel Daten mitzuschneiden oder DoS-Attacken zu starten.

4.2.1 Unzureichende Trennung zwischen internen Management- und Mandantennetzen

Gefährdung 10:	Durch eine unzureichende Trennung von internen Managementnetzen und Mandantennetzen können diverse Angriffe ausgeführt werden, wie z.B. das Mitschneiden von Passwörtern durch Man-in-the-Middle-Attacken oder der Versuch von Exploits auf die Management-Interfaces der vBlock-Komponenten.

Wird die Trennung von internen Managementnetzen und Mandantennetzen nicht korrekt durchgeführt, können die in Kapitel 4.2 genannten Gefahren auftreten.

Grundsätzlich sollten alle Managementnetze (VLANs) des vBlocks und Out-of-Band Interfaces der vBlock-Komponenten strikt von den Mandantennetzen (VLANs) der VM Gäste getrennt werden.

Empfohlen wird die Verwendung einer durch Firewalls abgetrennten Admin Zone. Zugriffe zu den Managementnetzen und Out-of-Band Interfaces des vBlocks sollten nur aus dieser Admin Zone heraus möglich sein.

4.2.1.1 EMC Symmetrix, VNX und CLARiiON

Das EMC CLARiiON Speichersystem und das EMC VNX Speichersystem verfügen auf beiden Storage Prozessoren über Out-of-Band Management Ethernet Management Interfaces. Diese sollten in einem Administrations-Netzsegment getrennt von öffentlichen Netzen eingerichtet werden. Hierüber erfolgt die gesamte Konfiguration.

Das EMC Symmetrix Speichersystem verfügt über einen sogenannten Service Prozessor, über den sich der EMC Support zur Entstörung oder zur Wartung des Enginuity Betriebsystems mit dem System verbinden kann (entweder per Modem oder per IP-Verbindung mittels des EMC Remote Support Systems; siehe dazu auch Kapitel 4.5). Auf diesem Service Prozessor ist auch eine Symmetrix Management Console installiert, mit der die jeweilige Symmetrix konfiguriert werden kann.

Der Management Zugriff per SMC außerhalb des Service Prozessors erfolgt inband über eine Fibre Channel oder iSCSI-Verbindung zu den zu verwaltenden Storage Arrays.

4.2.1.2 Cisco UCS, MDS und Nexus

Sowohl der UCS Fabric Interconnect, wie auch die MDS 9000 Fibre Channel Switch Familie verfügen über Out-of-Band Ethernet Management Interfaces. Diese Interfaces sind komplett getrennt von den Mandantennetzen. Über die IP-Adressen dieser Management Interfaces werden alle Verbindungen zu den Systemen aufgebaut.

Neben der Verbindung zum UCSM und NX-OS (MDS 9000) gibt es im UCS System auch die pass-through Verbindung zu den Server-Blades selbst. Auch diese pass-through Verbindung erfolgt durch die Out-of-Band Ethernet Management Interfaces der Fabric Interconnect. Dieses betrifft z.B. die KVM-Verbindung, etc.

D.h. im Vblock wird per Design verhindert, dass Mandanten Zugriff auf die Managementinterfaces bekommen, unter der Voraussetzung, dass keine Fehlkonfigurationen auf den übergeordneten Switches, Router und Firewalls bestehen, die einen Zugriff von den Mandantennetzen auf die Out-of-Band Interfaces ermöglichen.

4.2.1.3 VMware

Applikationsverkehr auf dem Hypervisor ist im Wesentlichen die Netzkommunikation der VMs, also der Gäste. Dieser ist durch VLAN-Separierung strikt vom Managementverkehr zu trennen. Das Management Segment der vSphere Server Farm ist besonders schützenswert und sollte durch strikte Zugriffsmöglichkeiten (Firewalls) und Protokollierung überwacht werden.

Client Sitzungen zum vCenter Server können über vSphere API Clients wie z.B. den vSphere Client oder die PowerCLI aufgebaut werden. Standardmäßig werden diese Verbindungen via SSL AES 256 Bit verschlüsselt, aber die Standardzertifikate werden nicht durch eine vertrauenswürdige Zertifizierungsstelle ausgestellt und liefern deshalb nicht den Schutz, den man sich in einer Produktionsumgebung wünscht. Diese eigensignierten Zertifikate sind durch "Man-in-the-Middle-Attacken" verwundbar. Deswegen empfiehlt VMware die selbstsignierten Zertifikate durch eigengenerierte oder externe Zertifikate einer vertrauenswürdigen Zertifizierungsstelle zu ersetzen.

Weitere Informationen sind zu finden im VMware vSphere 4.0 Security Hardening Guide; [04].

4.2.2 Deaktivieren optionaler Services

Gefährdung 11:	Nutzung von überflüssig geöffneten Kommunikations-Ports, z.B. für das Ausnutzen von Schwachstellen im Code ("exploit of vulnerabilities").

Sollte sich ein Angreifer in die Lage versetzt haben Zugriff auf die internen Managementnetze zu erlangen, so kann er versuchen Angriffe auf die Management-Interfaces der vBlock-Komponenten zu fahren. Alle offene Ports – d.h. offene Dienste auf den Komponenten des vBlocks, ermöglichen Zugriffe auf die Software der Systeme von aussen. Sollten bei diesen Diensten Programmierfehler bestehen, die z.B. *buffer overflows* oder fehlerhafte *memory pointer* auslösen, so können diese Schwachstellen potentiell einen *exploit* ermöglichen, der dem Angreifer einen Zugriff auf das System oder fremde Mandantenressourcen ermöglicht.

In den nachfolgenden Unterkapiteln wird beschrieben, welche Massnahmen getroffen werden müssen, um unbenötigte Dienste abzuschalten.

4.2.2.1 EMC Symmetrix, VNX und CLARiiON

Non-secure CLI Access muss explizit abgeschaltet sein. Sobald dieser Zugang abgeschaltet ist, wird automatisch der Secure CLI verwendet.

Die Symmetrix bietet selbst keine Dienste im Netz an.

4.2.2.2 Cisco UCS

Um das Cisco UCS System zu verwalten, sind eine Reihe von Management-Protokollen im Einsatz. Folgende Dienste sind zu beachten:

- **HTTP**
  Es wird empfohlen, den HTTP-Zugriff abzuschalten.
- **HTTPS und XML über HTPS**
  Um auf die UCSM GUI und das XML Interface zugreifen zu können, sollte mindestens HTTPS eingeschaltet bleiben. Es kann für das HTTPS- und SSH-Management ein von einer CA signiertes Zertifikat verwendet werden. Im Auslieferungszustand wird ein "*self signed certificate*" verwendet.
- **Telnet**
  Es wird empfohlen, den Telnet Zugriff abzuschalten. Im Auslieferungszustand ist dieser bereits ausgeschaltet.
- **CIM XML**
  Es wird empfohlen, diesen "read-only" Zugriff abzuschalten. Im Auslieferungszustand ist dieser bereits ausgeschaltet.
- **SNMP**
  Es sind SNMP v1, 2c und 3 möglich. Es wird empfohlen, wenn SNMP eingeschaltet wird, ausschließlich die Version 3 zu verwenden sowie die Authentisierung und Verschlüsselung von SNMPv3 zu verwenden. SNMP v1 und 2c sollten ausgeschaltet bleiben.

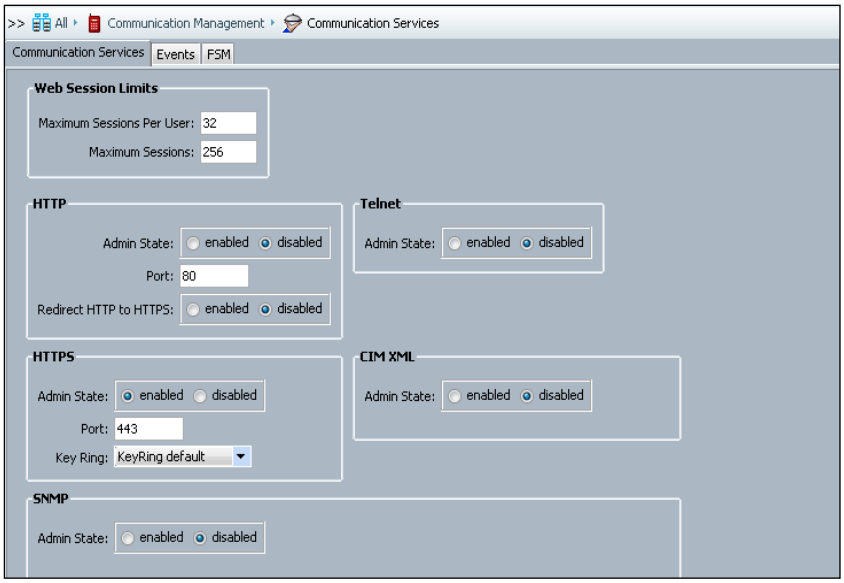


Abbildung 48: UCSM: Empfohlene Settings der Communication Services

4.2.2.3 MDS und Nexus (NX-OS)

Der Zugriff auf die Konfiguration kann bei NX-OS mittels CLI über Telnet und/oder SSH erfolgen.

Der Zugriff über das unsichere Telnet Protokoll auf das CLI kann und sollte abgeschaltet werden. Dieses kann im Initialen Setup geschehen oder nachträglich über das Kommando **`no telnet server enable`** erfolgen. Der UIM greift über das CLI-Interface über SSH auf die Konfigurationen in NX-OS zu. Weitere zu betrachtende Interfaces sind:

- **SNMP read/write Zugriff**
  Bei Zugriff über SNMP sollte nur SNMPv3 mit AES-128 Verschlüsselung eingesetzt werden. Die Userverwaltung von SNMPv3 erfolgt dann ähnlich wie für den SSH/CLI-Zugriff, und wird im Kapitel 4.3.2.5 erklärt.
  Details zu SNMP Konfiguration für NX-OS sind in Cisco MDS 9000 Family NX-OS System Management Configuration Guide; [14], Kapitel "Configuring SNMP" zu finden.
- **CIM (Common Interface Model) / SMI-S über http oder https**
  Der Zugang über CIM (Common Interface Model) / SMI-S erfolgt per wbem http/https über die Ports TCP 5988 (wbem-http) oder TCP 5989 (wbem-https). Im default Zustand ist die CIM-Schnittstelle auf NX-OS abgeschaltet. Wenn diese Schnittstelle für Management Systeme verwendet werden soll, so wird empfohlen nur HTTPS mit zertifikatsbasierter Authentisierung zuzulassen.
  Details zu der CIM Konfiguration sind zu finden im Cisco MDS 9000 Family NX-OS Fabric Configuration Guide; [15].
- **HTTP/HTTPS-Zugriff zum Download des Device Managers**
  Der Zugriff über http/https zum Download des Device Manager kann und sollte abgeschaltet werden. Dieses erfolgt über das Kommando **`no feature http-server`**.

4.2.2.4 VMware

Per default ist die Netzkommunikation zwischen einem administrativen Host und ESX/ESXi- und vCenter-Servern verschlüsselt. Es wird empfohlen, nicht benötigte Dienste wie z.B. den Web Access Zugriff der ESX/ESXi-Server zu deaktivieren.

Viele andere Härtungsmaßnahmen sind insbesondere auf den vCenter-Server anzuwenden. Das installierte Betriebssystem Windows Server 2003 oder 2008 64 Bit ist zu härten und es ist genauestens darauf zu achten, welche Konten administrativen Zugriff auf den Host selbst haben dürfen.

Weitere Informationen sind zu finden im VMware vSphere 4.0 Security Hardening Guide; [04].

4.2.3 Unsicheres Management Interface

Gefährdung 12:	Mitschneiden von Administrationspasswörtern durch Man-in-the-Middle-Attacken auf die Managementnetze.

Sollte sich ein Angreifer Zugriff auf die Managementnetze verschafft haben, so kann er versuchen sich durch Man-in-the-Middle-Attacken in die Kommunikation zwischen den Administrationstools und den vBlock-Komponenten zu schleifen. Damit könnte er bei der Verwendung von unverschlüsselten Managementprotokollen Passwörter der Administrator mitschneiden, und diese später verwenden um das System umzukonfigurieren um z.B. einen Zugriff auf fremde Daten zu erlangen.

4.2.3.1 EMC Symmetrix, VNX und CLARiiON

Der Zugriff auf die Verwaltungsoberfläche UniSphere der VNX- und CLARiiON basierenden Storage Arrays erfolgt per HTTPS und ist damit verschlüsselt.

Gleichermaßen verhält es sich bei der EMC Symmetrix Management-Konsole. Hier erfolgt der Zugriff ebenfalls über HTTPS. Der optionale HTTP-Zugriff sollte abgeschaltet werden. Zugriffe mittels des CLI erfolgen inband über Fibrechannel auf eigens dafür freigegebene Symmetrix Logical Volumes.

4.2.3.2 Cisco UCS

Der Zugriff auf den UCS Manager, sowie auf die Blade Systeme über die pass-through Interfaces sind bis auf wenige Ausnahmen über verschlüsselte Protokolle ausgeführt:

- **SSHv2**
  CLI Management zum UCSM. Dies kann nicht ausgeschaltet werden.
  Die SSH-Kommunikation ist verschlüsselt.
- **HTTPS**
  GUI-Management; ist ab Werk eingeschaltet. Auch wenn es ausgeschaltet werden kann, ist dieses nicht zu empfehlen, da das System dann nur noch über das CLI konfiguriert werden kann. Weiterhin verwendet die XML-Schnittstelle ebenfalls HTTPS als Basis, und der UIM greift darüber auf den UCSM zu.
  Die HTTPS-Kommunikation ist verschlüsselt.
- **KVM-Zugriff (TCP Port 2068)**
  Der Zugriff auf die KVM-Konsole der Blade Server ist mit RC4 verschlüsselt.
- **SOL-(Serial over LAN)-Zugriff**
  Der Zugriff auf die serielle Konsole der Blade-Server wird über SSH verschlüsselt.
- **IPMI-Zugriff**
  Der IPMI-Zugriff (Intelligent Platform Management Interface) zu den Blade-Servern wird innerhalb des Service-Profiles definiert. Dort wird angegeben, ob der Zugriff "read-only" oder "read-write" ist. Für den Zugriff wird ein lokaler User auf dem Server (Profile) angegeben, der verwendet werden kann, um über IPMI Statistiken auszulesen oder Befehle zum Server zu schicken.
  Wird IPMI für ein übergeordnetes Management System benötigt, so kann man über eine IPMI-Policy diesen einschalten. Im anderen Fall empfehlen wir den Zugriff ausgeschaltet zu lassen.
  Die Kommunikation erfolgt unverschlüsselt.
- **SMASH CLP**
  Das "read-only" SMACH CLP (Systems Management Architecture for Server Hardware Command-Line Protocol) Interface zum Fabric Interconnect ist ab Werk eingeschaltet und kann nicht abgeschaltet werden. Das SMASH CLP Interface ermöglicht allerdings nur einen geringen Zugriff auf "show" Kommandos.
  Die Kommunikation erfolgt unverschlüsselt.

4.2.3.3 MDS und Nexus (NX-OS)

Der Zugriff auf NX-OS kann und sollte über SSH und SNMPv3 mit AES-Verschlüsselung erfolgen. Telnet ist abzuschalten und SNMPv1/2c sollten nicht verwendet werden.

Wenn CIM/SMI-S verwendet werden soll, dann sollte dieses über HTTPS mit zertifikatsbasierter Authentisierung erfolgen.

4.2.3.4 VMware

Der Zugriff auf ESX/ESXi Server und das vCenter erfolgt über unterschiedliche Protokolle

- **HTTPS/SSL** wird benutzt, um sich vom vSphere Client mit dem ESX-Server oder dem vCenter zu verbinden. Das gleiche gilt, wenn sich Third-Party Plug-Ins mit dem Web-API des vCenters verbinden (z.B. das Nexus1000v VSM, der UIM, etc.).
- **SSHv2** wird benutzt, um sich mit der Servicekonsole der ESX-Systeme zu verbinden.

Nicht benötigte Dienste, wie z.B. den Web Access Zugriff der ESX/ESXi-Server sind zu deaktivieren.

Weitere Informationen sind zu finden im VMware vSphere 4.0 Security Hardening Guide; [04].

4.2.4 Sicherheitsaspekte von FCoE und DCB

Gefährdung 13:	Man-in-the-Middle-Angriffe und DoS-Angriffe auf die FCoE-Storageanbindung innerhalb des vBlocks.

Im Vblock haben die ESX Hypervisor einen Fibre Channel Zugriff über FCoE auf die Storage Ressourcen. Diese Kommunikation ist besonders schützenswert. Bei unzureichendem Schutz könnte ein Angreifer:

- DoS Attacken starten, mit dem Ziel die Hypervisor und VM-Gäste zu einem Systemcrash zu bringen. Dieser Systemcrash kann ausgelöst werden, wenn die Betriebssysteme und ESX Hypervisor nicht in der Lage sind auf den Storage zuzugreifen.
- Man-in-the-Middle-Attacken ausführen, mit dem Ziel die Kommunikation zwischen Storage und Hypervisor mitzuschneiden und damit fremde Daten zu erlangen.

Zu betrachten sind bei FCoE und DCB (Data Center Bridging) vor allem die Sicherheit der Protokolle FCoE Initialization Protocol (FIP) und Data Center Bridging Exchange (DCBX).

Bei FIP besteht z.B. die Gefahr, dass den CNAs ein falsches FCoE-VLAN als Antwort auf einen VLAN Request gesendet wird (beim Booten der Hypervisor, z.B. nach einem Software Upgrade), mit der Absicht alle ESX Hypervisor in ein falsches FCoE VLAN zu setzen, und damit das Booten der Hypervisor zu verhindern (DoS Attacke).

Außerdem bestehen weitere Gefahren durch DoS-Attacken auf den FCF (Fibre Channel Forwarder), sowie das Abhören des FCoE-Verkehrs zwischen dem CNA (ENode) und dem FCF, und damit zwischen Host und Target (Storage).

Nachfolgend wird dargestellt wieso diese Angriffe im vBlock per Produktdesign nicht möglich sind.

Im Vblock UCS System wird FIPund FCoE nur auf den Punkt-zu-Punkt Ethernet-Verbindungen zwischen CNA und Fabric Interconnect eingesetzt. Der Fabric Interconnect verarbeitet dabei die FIP-Nachrichten, welche vom CNA im Blade Server empfangen werden, lokal ohne diese Nachrichten weiterzuleiten. D.h. der Fabric Interconnect sendet empfangene FIP-VLAN-Requests, die zur Multicast MAC-Adresse All-FCF-MACs gesendet werden, weder an seine LAN Uplinks, noch an andere CNAs. Daher können keine FIP Nachrichten von Blade Server zu Blade Server gesendet werden.

Diese Eigenschaft verhindert, dass ein Angreifer in die Lage versetzt wird FIP Nachrichten zu spoofen und zu DoS oder Man-in-the-Middle-Attacken zu verwenden.

Trotzdem sollten einige Best-Practices eingehalten werden, auch wenn FCoE Angriffe durch das Produktdesign des UCS Systems bereits unterbunden werden.

FIP wird vom CNA als erstes im "default" FCoE-VLAN gesprochen, anschließend bekommt der CNA die FCoE-VLAN-Response und stellt sich zu seinem eigentlichen FCoE-VLAN um, das wiederum in das richtige VSAN zwischen dem Fabric Interconnect und dem MDS 9000 Switch mündet.

Um zu verhindern, dass ein VM-Gast falsche FIP-Informationen zum Fabric Interconnect (FCF) schickt, sollte das "default" FCoE-VLAN des CNAs von keinem der VM-Gäste aus erreichbar sein. Das gleiche gilt für die produktiven FCoE-VLANs. Am sichersten ist es, für alle FCoE-VLANs eigene VLAN-IDs zu verwenden, sowie FCoE und "normales" Ethernet niemals im gleichen VLAN zu transportieren. Auch ist ab Werk das VLAN mit der ID 1 das "default" FCoE-VLAN. Die VLAN-ID für das "default" FCoE-VLAN muss geändert werden.

DCBX stellt einem potentiellen Angreifer sehr viele Informationen zur Verfügung. Diese Informationen über die Details des Netzes könnten von einem Angreifer genutzt werden, um gezieltere Angriffe zu gestalten. Daher sollte DCBX, wie LLDP (Link Layer Discovery Protocol) und CDP (Cisco Discovery Protocol), auf Access-Ports abgeschaltet sein, die es nicht benötigen. Im Vblock wird DCBX aber wiederum nur zwischen dem CNA und dem Fabric Interconnect gesprochen, und kommt damit nicht mit dem VMware Hypervisor oder den VM-Gasts in Verbindung.

Zusammenfassend gesagt werden, dass die Sicherheitsbetrachtung von FCoE und DCBX sehr stark vom eingesetzten Design abhängig ist. Vor allem bei möglichen zukünftigen Topologien über "lossless" Ethernet Clouds über mehrere Ethernet "hops" gibt es viele Aspekte der Sicherheit zu betrachten. Bei der im Vblock UCS System eingesetzten Punkt-zu-Punkt-Topologie sind die Gefährdungen, außerhalb von Fehlkonfigurationen, als gering anzusehen.

4.3 Unzureichende Admin-Rechtevergabe, Beschränkung und Protokollierung

Gefährdung 14:	Unzureichende Admin-Rechtevergabe und Beschränkung begünstigt Fehlkonfigurationen und Missbrauch der Privilegien.

Die Vblock-Infrastrukturplattform ermöglicht eine Konsolidierung und gemeinsame Nutzung von Computing, Netz und Storage Ressourcen. Obwohl diese Konsolidierung einen guten Weg darstellt auch mehr Sicherheit durch zentralisierte und standardisierte Sicherheitsfunktionen einzuführen, birgt sie auch gleichzeitig ein erhöhtes Risiko durch die Möglichkeiten, die sich den Administratoren eröffnen. Missbrauch von Administratorrechten sowie versehentliche oder absichtliche Fehlkonfiguration bilden das Hauptsicherheitsrisiko im Vblock.

Als Beispiel ist hier die Verwaltung der Service Profile im Unified Computing System Manager (UCSM) zu nennen. In den Service Profilen werden Identitäten verwaltet, an denen wiederum Zugriffsrechte festgemacht werden. Insbesondere bei den WWN-Adressen, die im FC-Switch und Storage Controller verwendet werden, um Schreib- und Leseberechtigungen zu Storage Ressourcen zu kontrollieren, besteht die Gefahr, dass bewusste und unbewusste Fehlkonfigurationen zu Security Incidents führen können.

4.3.1 Administrator Authentisierung und Autorisierung

Die nachfolgenden Kapitel beschäftigen sich im Detail mit den "Best Practices" der Konfiguration der individuellen Komponenten des Vblocks im Bereich Authentisierung und Autorisierung.

Kurz zusammengefasst wird empfohlen:

- Die Authentisierung und Rechteverwaltung sollten zentral über einen Directory Service mittels LDAP umzusetzt werden.
- Wo immer möglich, sollten die Rechte der individuellen Administratoren und Teilbereichsexperten auf das für sie Nötigste eingeschränkt werden.

- Event Protokolle sollten zu einem remote Syslog Server oder Security Incident und Event Management (SIEM) gesendet werden, und dort ausgewertet werden.
- Ein SIEM sollte eingesetzt werden, um Ereignisse zu korrelieren, Security Incidents zu erkennen und zu melden. Dieses ermöglicht auf solche Security Incidents schnell zu reagieren sowie forensische Analysen auf vergangene Events durchzuführen.

4.3.2 Authentisierung

4.3.2.1 Unified Infrastructure Manager

Zur Authentisierung von Benutzern unterstützt der Unified Infrastructure Manager folgende Methoden:

- TACACS+
- RADIUS
- LDAP (Lightweight Directory Access Protocol) / ADS (Active Directory Service)
- Internal registry
- SAML (Security Assertion Markup Language)

In Abbildung 49 sind sämtliche Kommunikationswege des UIM dargestellt:

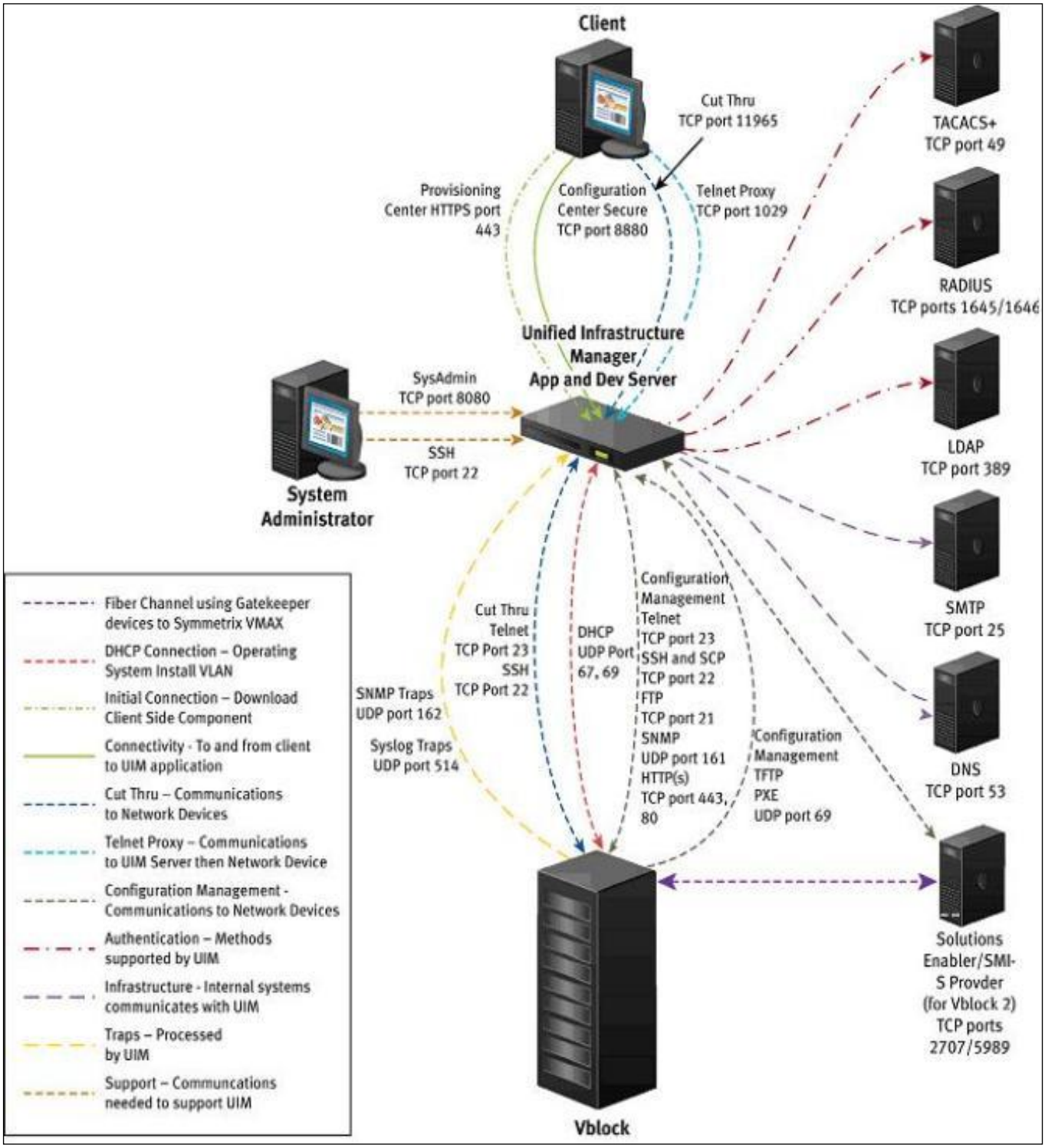


Abbildung 49: EMC UIM: Authentisierung

4.3.2.2 vCenter

Das vCenter unterstützt folgende Methoden der Authentisierung und die Autorisierung für den rollenbasierten Zugriff:

- Lokale Benutzer und Gruppen des vCenter Servers (NTLMv2)
- Kerberos für die ADS Benutzer und Gruppen Authentisierung.

Der vSphere Client kann sich als Front End Applikation sowohl mit dem vCenter als auch mit dem einzelnen ESX(i) Server verbinden. Jegliche Kommunikation wird dabei mittels SSL sicher übertragen.

Der Benutzer wird nach erfolgreicher Authentisierung nur Entitäten im vCenter sehen (Need to Know), für welche er auch laut Role Based Access Control Zugriffsrechte besitzt.

4.3.2.3 Storage

Vblock 1/1U und Series 300 (CLARiiON, VNX)

Die VNX und CLARiiON unterstützt zur Authentisierung neben einer internen Benutzerdatenbank auch die Standard-Systeme LDAP und Active Directory als Benutzerdatenbank.

Vblock Series 700 (Symmetrix)

Die Symmetrix untertützt einen administrativen Zugriff ausschließlich inband über Fibre Channel. Es findet hier keine Authentisierung statt. Die Voraussetzung für einen Zugriff ist eine erfolgreiche Authentisierung (und Autorisierung als Administrator) am Management-Host, über den die Administration der Symmetrix erfolgt. Die Benutzer- und Host-Informationen werden aber zum Zweck der Autorisierung und Protokollierung an das Symmetrix-System weitergegeben.

Die eigentliche Kommunikation zwischen Management Server und Storage Array erfolgt dabei über die SYMAPI-Schnittstelle. Sie wandelt alle Informations- und Änderungsanfragen zur Konfiguration in spezielle SCSI Kommando-Pakete um, die dann eingebettet in FC Frames an spezielle Konfigurations-LUNs der Symmetrix (sogenannte *Gatekeeper*) gesendet werden. Diese LUNs sind gerade so groß, dass sie von allen gängigen Betriebssystemen als Datenträger erkannt werden (ca. 3 MB). ESX-Server blenden diese Datenträger in der Standardkonfiguration aus.

Die Symmetrix erkennt die speziellen Kommando-Pakete und legt sie nicht als Daten auf den LUNs ab, sondern wertet deren Informationen zur Beantwortung der Anfrage beziehungsweise zur Änderung der Konfiguration aus.

Die zur Konfiguration benötigten LUNs werden ausschließlich innerhalb des Managementbereichs des Vblocks bereitgestellt. Ein Zugriff über die Mandantenumgebung ist daher ausgeschlossen.

Bei Verwendung der Element-Manager findet die Authentisierung mithilfe der dort unterstützten Methoden statt. Im Falle der Symmetrix Management-Konsole können systeminterne und Active Directory Benutzer authentisiert werden.

4.3.2.4 Cisco UCS

Der Unified Computing System Manager (UCSM) verfügt über zwei Möglichkeiten zur Authentisierung von Administratoren und Teilbereichsexperten:

- Lokale Benutzerdatenbank auf dem UCSM-System selbst. Diese lokale Datenbank sollte auf sehr wenige Notfall-Accounts beschränkt bleiben.
- Remote Authentisierung über folgende Protokolle:
  - LDAP
  - RADIUS
  - TACACS+

Die Methoden und Protokolle können gemischt verwendet werden. Beim Anmeldefenster kann die Methode (Authentisierungsgruppe) ausgewählt werden. Für die Konsolen Authentication (SSH), kann nur eine Authentication Methode ausgewählt und verwendet werden.

Für alle Methoden können zwei Attribute mitgegeben werden, die Rolle des Nutzers, sowie das "Locale", d.h. die Zugehörigkeit in der Hierarchie. Beides ist detailliert im Kapitel 4.3.3.2 beschrieben.

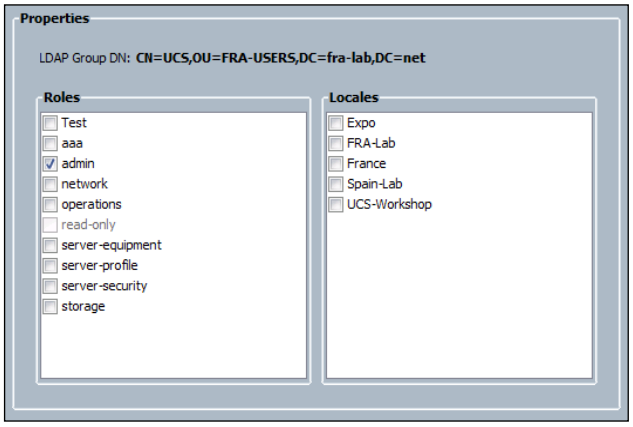


Abbildung 50: UCSM: Zuordnung von Rolle und Locales zu einem LDAP-Attribut

Für RADIUS und TACACS+ wird die Rolle und das "Locale" in einem Cisco Attribute-Value Paar (AV-Pair) mitgegeben, z.B.
`shell:roles="admin,aaa"`
`shell:locales="L1,abc"`

Für LDAP kann angegeben werden, zu welchem LDAP-Attribut die Rolle und das "Locale" zugeordnet wird. Zum Beispiel wird in der Abbildung die Rolle "admin" mit dem LDAP-Attribut "CN=UCS,OU=FRA-USERS,DC=fra-lab,DC=net" verbunden. Wenn dieses Attribut für einen Nutzer übergeben wird, wird ihm im UCSM die Rolle "admin" vergeben.

Mehr Details hierzu siehe Cisco UCS Manager GUI Configuration Guide; [11]
Kapitel "System Configuration" → "Configuring Authentication".

4.3.2.5 MDS und Nexus (NX-OS)

NX-OS verfügt über zwei Authentisierungsmöglichkeiten für administrative Benutzer:

- Lokale Benutzerdatenbank: Die lokale Datenbank sollte auf sehr wenige Notfall-Accounts beschränkt bleiben.
- Remote Authentisierung über folgende Protokolle:
  - RADIUS
  - TACACS+

Das CLI und SNMPv3 verwenden gemeinsame User und Rollendefinitionen (RBAC). Die Passwörter von SNMPv3 und CLI Usern werden synchronisiert.

Für die lokalen User gibt es Passwortrichtlinien. Auch kann den Passwörtern ein Ablaufdatum vergeben werden. Dieses ist nur für reguläre User zu empfehlen. Mindestens ein Admin Account sollte als "Notfalluser" ohne Ablaufdatum bestehen.

Generell ist die Nutzung eines zentralen Authenisierungsdienstes auf RADIUS oder TACACS+ Basis zu empfehlen, und der lokalen Authentisierung vorzuziehen. Sollte kein RADIUS oder TACACS+ Server mehr erreichbar sein, gibt es immer ein Fallback zu der lokalen Nutzerdatenbank.

Details zur Konfiguration von Nutzern und Nutzerrechten bzw. RADIUS und TACACS+ sind zu finden im Cisco MDS 9000 Family NX-OS Security Configuration Guide; [10]
Kapitel "Configuring Users and Common Roles" → "Configuring RADIUS and TACACS+".

Für den SSH-Zugriff kann auch eine Zertifikatsauthentisierung erfolgen. Diese wird z.B. für die Verbindung des UIM zum NX-OS verwendet.

4.3.3 Autorisierung

Autorisierung bedeutet in der IT die Zuweisung und Überprüfung von Zugriffsrechten auf Daten und Dienste an Systemnutzer. Die Autorisierung erfolgt meist nach einer erfolgreichen Authentisierung.

4.3.3.1 Storage

Vblock 1/1U und Series 300 (CLARiiON, VNX)

Zum Zweck der Autorisierung sind bei der VNX und der CLARiiON (UniSphere / NaviSphere) folgende Rollen vorgehanden:

Rolle	Rechte
Monitor	Dieser Account hat nur lesenden Zugriff.
Manager	Dieser Account ist in der Lage neue Storage Ressourcen (Pools, LUNs, RAID Gruppen) zu erstellen.
Security Administrator	Dieser Account ist in der Lage User- und Domain-Sicherheitsaufgaben auszuführen.
Administrator	Dieser Account hat die volle Kontrolle über alle Storage und Security Ressourcen.
Replication Roles	
Local	Nur für Basis SnapView Aufgaben.
Replication	Basis MirrorView, SAN Copy und SnapView Kommandos.
Replication and Recovery	Alle MirrorView, SAN Copy, SnapView Aufgaben inklusive Wiederherstellung.

Eine Begrenzung der Administration auf Teilbereiche oder -komponenten ist nicht möglich. So kann zum Beispiel ein Manager alle Speicherbereiche konfigurieren.

Vblock Series 700 (Symmetrix)

Die Autorisierung ist mit verschiedenen Rollen über SYMAUTH konfigurierbar (Host und User Roles):

Rolle	Rechte
None	Kein Zugriff erlaubt.
Monitor	Diese Rolle verfügt über Rechte für Read-only / View Aufgaben.
Auditor	Diese Rolle verfügt über Rechte für Security Aufgaben und über den Zugriff zum Symmetrix audit log.
StorageAdmin	Diese Rolle verfügt über die Rechte alle Storage Management Aufgaben durchzuführen.
SecurityAdmin	Diese Rolle verfügt über Rechte zur Durchführung von Security Aufgaben.
Admin	Diese Rolle verfügt über Rechte zur Durchführung von allen Storage Management und Security Aufgaben.

Hier werden Benutzern oder Gruppen entsprechende Rollen zugewiesen und erhalten so die Administrationsfreigaben auf Systemebene. Die Autorisierung kann hier entweder nur überwacht und bei Verstoß entsprechend protokolliert (advisory) oder erzwungen werden (enforce).

Eine weitere Autorisierungsmöglichkeit bietet der Einsatz von Symmetrix Access Control Lists. Diese Methode erlaubt die Beschränkung der Administration auf definierbare Teilbereiche des bereitgestellten Speicherplatzes. Dazu werden in der Symmetrix-Datenbank User-IDs und Host-IDs erzeugt und hinterlegt. Diese werden dann sogenannten Access Groups zugewiesen. Weiterhin können Access Pools definiert werden, denen dann die Speicherressourcen zugewiesen werden, für die der eingeschränkte Zugriff gelten soll.

Benutzern, Hosts oder Gruppen an Pools können verschiedene Rollen zugewiesen werden. Für Details siehe EMC Solutions Enabler Symmetrix Array Management CLI Product Guide; [16], Kapitel "Host-based Access Control" → "Creating and managing access control entries".

Damit kann die Autorisierung auch auf Hosts und Teilbereiche der Symmetrix für unterschiedliche Aufgaben erfolgen. Dieser Weg ist immer strikt (enforced).

4.3.3.2 Cisco UCS Role-Based Access Control (RBAC)

Wie bereits erwähnt, ist das RBAC-Modell im UCSM ein entscheidender Bestandteil der Sicherheit im Vblock. Über Service Profile und Identity Pools werden die Identitäten verwaltet, die in den SAN-Komponenten zur Zugriffsteuerung verwendet werden. In den Service Profilen werden VLANs, QoS Parameter, etc. vergeben.

Es ist also entscheidend das Rollenmodell so umzusetzen, dass Administratoren und Teilbereichsexperten nur Schreibzugriff auf den Teil der UCS-Konfiguration erhalten, den sie wirklich benötigen.

Der UCSM arbeitet mit Rollen (Roles) und Hierarchie-Zugehörigkeit (Locales).

Eine Rolle enthält eine oder mehrere Privilegien mit denen auf Objekte (MOs – Managed Objects), d.h. auf Komponenten, Pools, Profile, etc. zugegriffen werden kann. Dabei gibt es eine Reihe von vorkonfigurierten Rollen:

- **AAA Administrator**
  Lese- und Schreibzugriff auf User, Rollen und Locales, sowie auf Authentication, Authorisation und Accounting (logging) (AAA) Konfigurationen. Lesezugriff auf den Rest des Systems.
- **Administrator**
  Kompletter Lese- und Schreibzugriff auf das gesamte System. Dem "default" Admin-Account ist diese Rolle zugeordnet; seine Rollenzugehörigkeit kann nicht geändert werden.
- **Facility Manager**
  Lese- und Schreibzugriff auf das "power management" durch das "power-mgmt" Privileg. Lesezugriff auf den Rest des Systems.
- **Network Administrator**
  Lese- und Schreibzugriff auf die Fabric Interconnect Infrastruktur und Netz-Sicherheitsfunktionen, z.B MAC-ID-Pools, Netz-Uplinks, etc. Lesezugriff auf den Rest des Systems.
- **Operations**
  Lese- und Schreibzugriff auf System Logs und Syslog Server Konfigurationen. Lesezugriff auf den Rest des Systems.
- **Read-Only**
  Ausschließlicher Lesezugriff, keine Schreibberechtigungen.
- **Server Equipment Administrator**
  Lese- und Schreibzugriff auf physische Blade-Server Operationen, z.B. ein Blade-Server zu de-kommissionieren und aus dem System zu entfernen. Lesezugriff auf den Rest des Systems.
- **Server Profile Administrator**
  Lese- und Schreibzugriff auf Service Profile Bestandteile, die mit dem logischen Server zu tun haben, z.B. Anzahl und Art von virtuellen Netzkarten, Identitäten, etc. Lesezugriff auf den Rest des Systems.

- **Server Security Administrator**
  Lese- und Schreibzugriff auf die Bestandteile der Service Profile, die mit Sicherheitsfunktionen im Service Profile zu tun haben, wie z.B. der Definition von Profilen für den Zugriff über IPMI. Lesezugriff auf den Rest des Systems.
- **Storage Administrator**
  Lese- und Schreibzugriff auf Storage Operationen, z.B. WWN-Id-Pools, FC-Uplinks, etc. Lesezugriff auf den Rest des Systems.

Zusätzlich zu den vorkonfigurierten Rollen können auch individuelle Rollen definiert werden. Den individuellen Rollen können dann Privilegien zugeordnet werden.

Details zu den Rollen und Privilegien siehe Cisco UCS Manager GUI Configuration Guide; [11] Kapitel "System Configuration" → "Configuring Role-Based Access Control".

Hierarchie von Pools und Policies

Objekte wie Pools, Policies für Service Profiles werden im UCSM in einer Organisationshierarchie verwaltet. Die Spitze dieser Baumstruktur "root" ist die erste, vorkonfigurierte und nicht löschbare Hierarchiestufe. Aus ihr verzweigen sich die untergeordneten Organisationen.

Pools und Policies werden Organisationen zugeordnet, wobei die Pools innerhalb eines "Astes" zur "root" gemeinsam genutzt werden können, wenn sie den gleichen Namen haben.

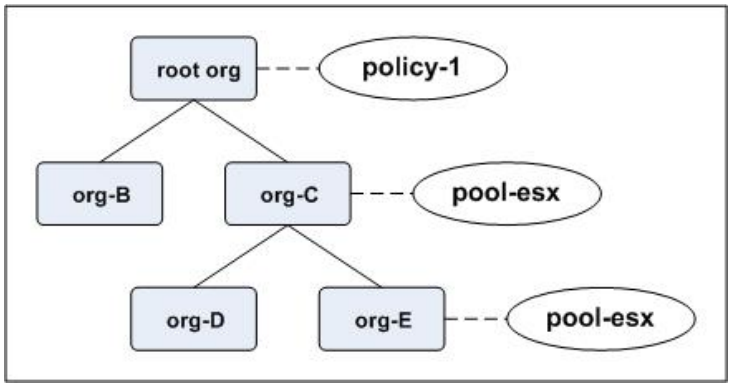


Abbildung 51: UCSM: Organisationshierarchie für Pools und Policies

Z.B. kann in dem dargestellten Beispiel (siehe Abbildung 51) "org-E" auf ihren "pool-esx" zugreifen. Wenn in diesem keine Ressourcen mehr verfügbar sind, kann "org-E" auf Ressourcen im "pool-esx" von "org-C" zugreifen.

"org-D" und "org-C" können aber nicht auf Ressourcen und Policy Definitionen in "org-E" zugreifen. Wird innerhalb des Baums keine Ressource oder Policy mit dem richtigen Namen und verfügbaren Ressourcen gefunden, so werden Default Pools und Default Policies im "root" verwendet.

Über die "Locales" kann die Zugehörigkeit zur Organisation, und damit der Einstieg in die Hierarchie für den Administrator geregelt werden. In dem in der Abbildung 52 dargestellten Beispiel wird für den Administrator (User) "Bob" die Rolle "Server" übergeben.

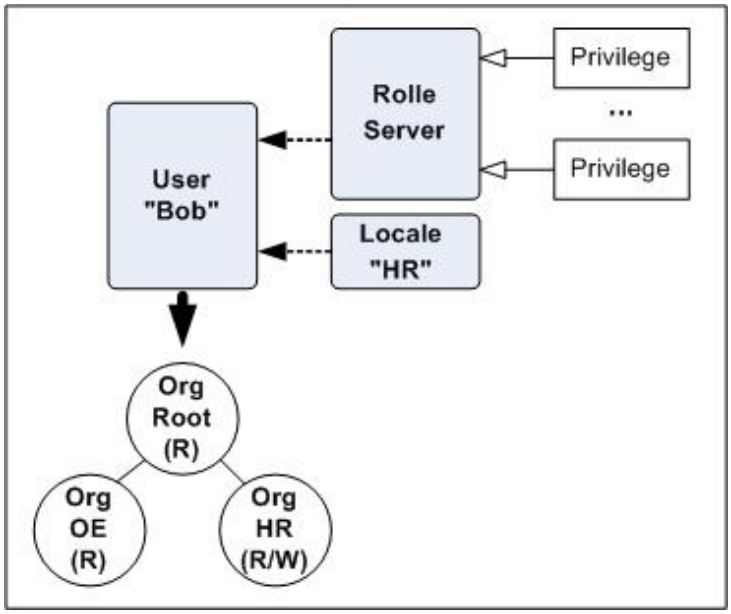


Abbildung 52: UCSM: Organisationshierarchie für User, Rollen und Locales

Diese Rolle hat dann Privilegien wie z.B. die Berechtigung einen Server zu de-kommissionieren. Der Administrator (User) "Bob" bekommt auch den Locale "HR" übergeben, über den er in die Organisation "HR" innerhalb des Baumes eingefügt wird. Der Administrator (User) "Bob" hat dadurch die Berechtigung bestimmte Server Operationen innerhalb der Ressourcen in seiner Organisation vorzunehmen. Dabei kann er auf Pools und Policies in der "root" Organisation zugreifen, diese allerdings nicht selbst verwalten oder ändern.

Weitere Details siehe Cisco UCS Manager GUI Configuration Guide; [11]
Kapitel "System Configuration" → "Configuring Organisations".

4.3.3.3 MDS und Nexus (NX-OS)

NX-OS verfügt über ein Rollenmodel zur Autorisierung von Konfigurations-Kommandos. NX-OS hat dabei drei vordefinierte Rollen:

- **Network-Admin**
  Nutzer, die diese Rolle bekommen, dürfen Änderungen an sämtlichen Konfigurationselementen vornehmen.
- **Network-Operator**
  Nutzer, die diese Rolle bekommen, dürfen die Konfiguration einsehen (show Befehle).
- **Default-Rolle**
  Nutzer, die diese Rolle bekommen, dürfen nur über die GUI Tools (Fabric Manager, Device Manager) zugreifen, und nur die Konfiguration einsehen.

Bis zu 64 weitere Rollen können individuell definiert werden, d.h. es werden einzelne Konfigurations- und "show"-Befehle für eine Rolle erlaubt.

Wird RADIUS oder TACACS+ verwendet, so wird die Rolle über ein Attribute (cisco-av-pair, shell-attribute) mitgegeben. Wird keine Rolle bei der RADIUS oder TACACS+ Authentisierung angegeben, wird als Rolle die „Network-Operator“ Rolle verwendet.

Es wird empfohlen, die Network-Admin Rolle auf wenige "Super-User" zu beschränken und sonst mit Rollen zu arbeiten, die nur den Zugriff auf die unbedingt benötigten Konfigurationselemente erlaubt.

4.3.3.4 Unified Infrastructure Manager

Die Autorisierung ist in verschiedenen Rollen konfigurierbar:

Rolle	Rechte
User	Diese Rolle verfügt über Rechte zur Sichtung des Provisioning Center, und zur Durchführung von Basisaufgaben. Diese Rolle kann keine Blade Server Aufgaben durchführen, und hat keinen Zugriff auf den Configuration Manager.
Service Admin	Diese Rolle verfügt über Rechte für alle Lifecycle Operationen, hat aber keinen Zugriff auf den Configuration Manager.
System Admin	Voller Zugriff.

4.3.3.5 vCenter

VMware vCenter gestattet sehr granular Rechte zu delegieren. Zusätzlich lässt sich der direkte Zugriff auf die Konsolen der ESXi-Server im "*Lock Down Modus*" noch weiter einschränken. Dadurch werden lokale ESXi-Server Accounts deaktiviert, und es ist nur noch der Zugriff über das vCenter möglich.

Das vCenter Rollenkonzept verwendet lokale Benutzer und Gruppen oder/und Active Directory Benutzer oder Active Directory Gruppen. Diesen Objekten werden vCenter Rollen zugeordnet. Es gibt vordefinierte Rollen im vCenter (sample Roles), nicht veränderbare System Rollen, und benutzerdefinierte Rollen.

Rollen werden Systemrechten zugeordnet. Die vCenter Objekte sind hierarchisch angeordnet. Berechtigungen werden in der vCenter Hierarchie wie bei einem Dateisystem von oben nach unten vererbt.

Die Kombination aus Rolle und Benutzer wird als effektive Berechtigung auf die einzelnen Objekte im vCenter gesetzt.

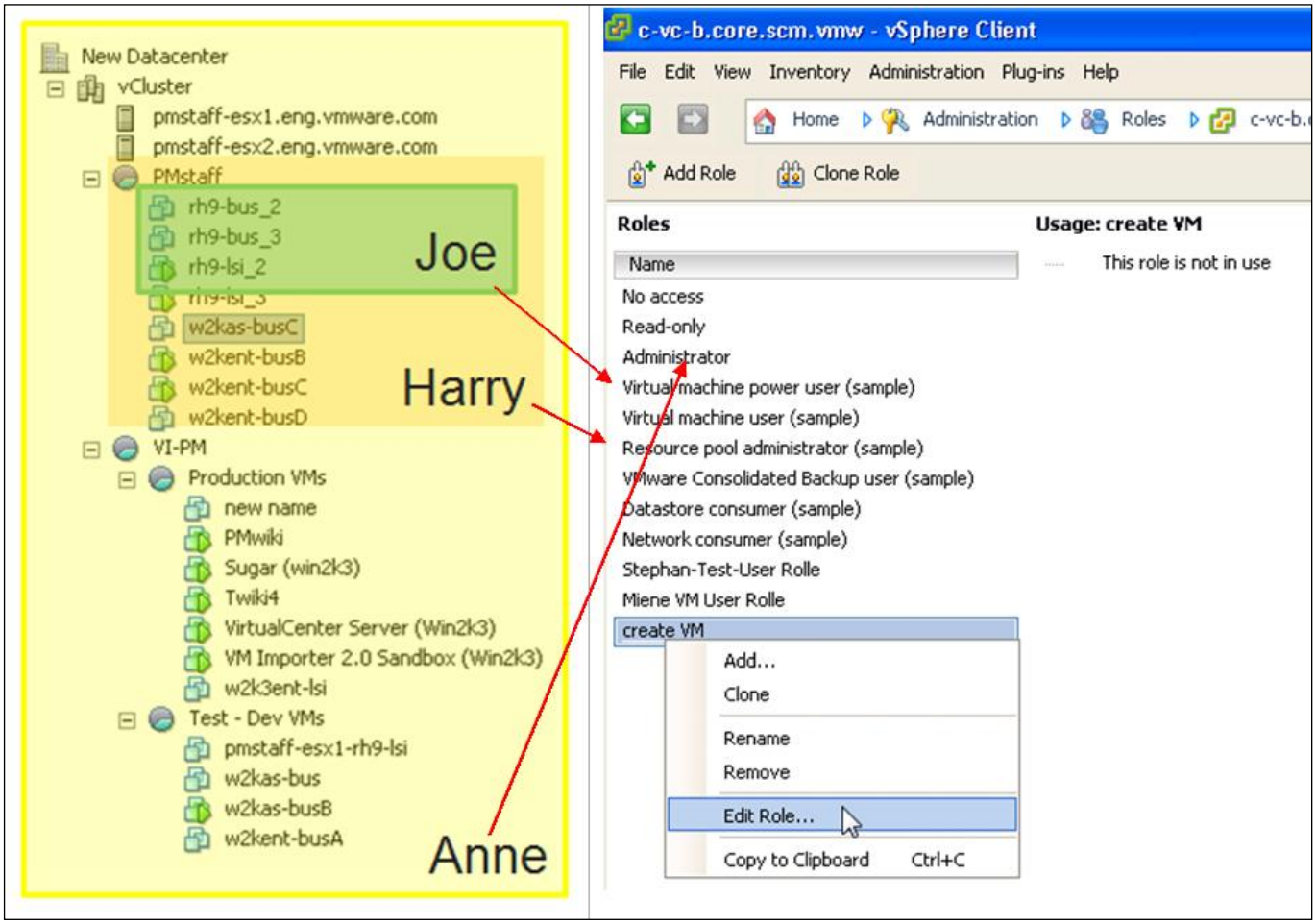


Abbildung 53: Beispiel der Objekt-, Benutzer- und Rollenzuordnung

In der obigen Abbildung 53 hat Joe z.B. nur das Recht die drei RedHat Maschinen als "Virtual Machine Power User" zu bedienen. Er kann im Gegensatz zu Harry keine Ressource-Pool Änderungen am Ressource-Pool PMstaff vornehmen.

In der nachfolgenden Abbildung 54 wird beispielhaft gezeigt, aus welchen Privilegien sich die Rolle "Virtual Machine Power User" zusammensetzt.

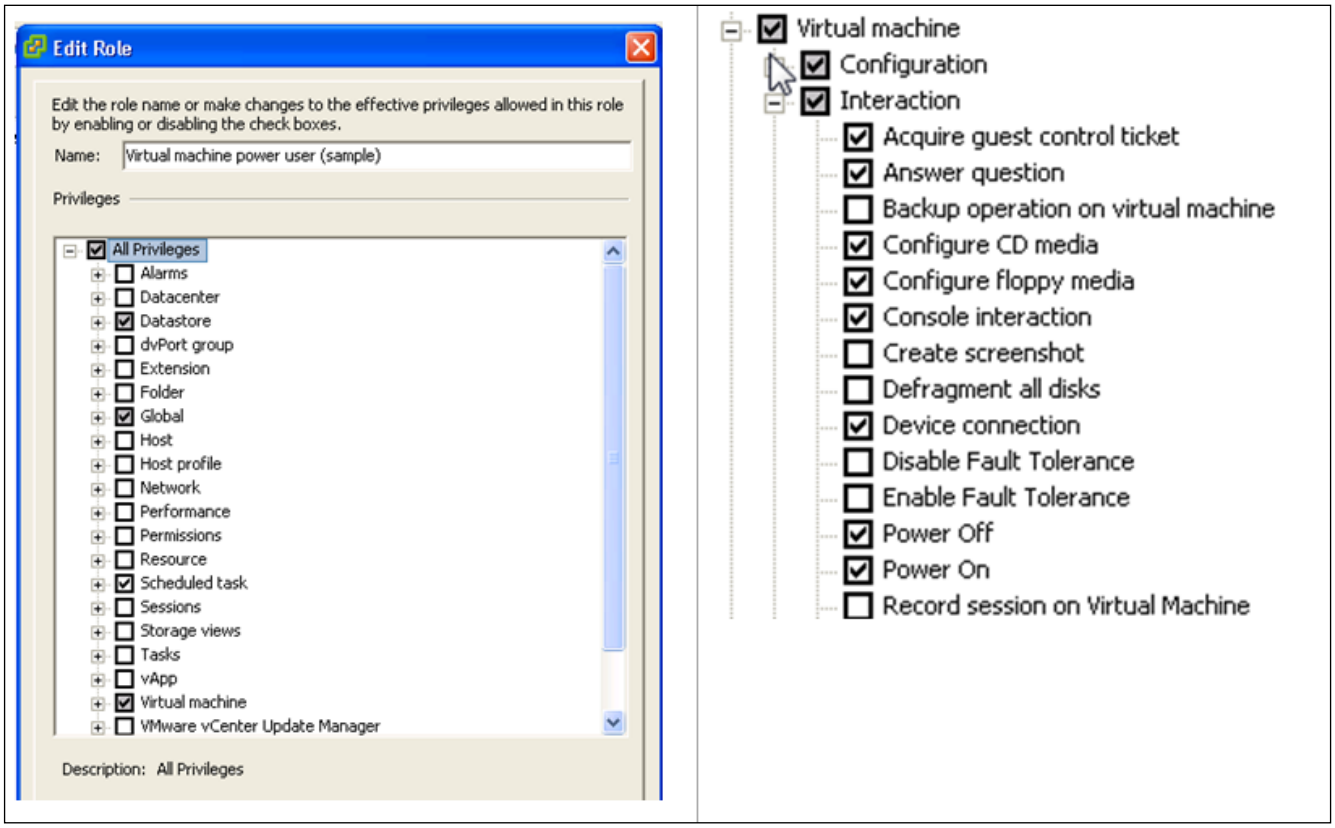


Abbildung 54: Die Privilegien der Rolle "Virtual Machine Power User"

Rechte lassen sich nicht nur auf VMs anwenden, sondern auch auf vCenter Objekte wie Storage oder Netze. So könnte man z.B. eine Rolle erstellen, welche zwar das Recht hat, virtuelle Maschinen zu erstellen, aber nicht das Recht besitzt, diese virtuellen Maschinen auf einen vSwitch zu patchen, da dies in der Kontrolle des Netz Administrationsteams liegt.

Für weiterführende Information zum Rechte Management siehe das Whitepaper Managing VMware VirtualCenter Roles and Permissions; [12].

4.4 Prüfung und Rechenschaftspflicht (Protokollierung)

Gefährdung 15:	Angriffe, unbeabsichtigte Zugriffsverletzungen und Fehlkonfigurationen bleiben unbemerkt. Dadurch entsteht wirtschaftlicher Schaden durch unbemerkte eventuell dauerhaft auftretende Verletzungen von SLAs, Policies und Compliance Vorgaben.

Gefährdung 16:	Ein Administrator kann durch entsprechende Berechtigungen lokale Logging-Daten löschen oder verändern und damit seine Spuren verwischen.

Logging Daten sind entscheidend für die Aufklärung von Security Incidents, sei es durch gewollte oder ungewollte Fehlkonfigurationen oder Angriffe von Aussen. Dabei ist es entscheidend, dass zuverlässig überprüft werden kann, wer eine Änderung vorgenommen hat, und wann diese erfolgt ist. Fehlen diese Informationen, so können Security Incidents nicht aufgeklärt werden, und somit keine Gegenmaßnahmen zum Verhindern von zukünftigen Incidents ergriffen werden.

Auch muss zwischen lokalem und remote Logging unterschieden werden. Bei ausschließlich lokalem logging besteht die Gefahr, dass Angreifer ihre Spuren verwischen indem sie Logging Daten löschen.

4.4.1 Storage

Vblock 1/1U und Series 300 (CLARiiON, VNX)

Alle Aktivitäten werden permanent überwacht und protokolliert (Source IP, Username, Aktion).

Die Audit Informationen sind Teil des Storage Processor's Event Log und umfassen:

- Event code
- Description of event
- Name of storage system
- Name of corresponding Storage Processor
- Hostname of SP (Storage Processor)
- Requestor (Unisphere user name)
- Type of request
- Target of request
- Success or Failure of request

Ein Event Log kann nur von einem Administrator oder Security Administrator gelöscht werden. Auch diese Aktion wird (im dann leeren Eventlog als erster Eintrag) protokolliert.

Ein Event Log kann außerhalb des Storage Arrays archiviert werden, z.B. um Compliance Anforderungen zu genügen.

Da auch der Anmeldevorgang Teil der Überwachung ist, enthält das Protokoll auch Informationen darüber von welchem Host aus die Anmeldung erfolgt ist.

Da die Protokolle bei Platzbedarf neue Einträge zu Lasten der ältesten überschreiben, muss bei bestehender Compliance-Anforderung ein entsprechender Mechanismus das Protokoll dauerhaft sichern. Dazu eignet sich besonders die Integration der CLARiiON Protokollierung in eine Lösung wie RSA enVision. Dabei sollte für das periodische Übertragen der Protokolldaten ein möglichst kleines Zeitintervall gewählt werden. Besser ist die Nutzung von SNMP-Traps, die einen Verlust von Protokollinformationen ausschließen. Jedes Event wird sofort per Trap versandt.

Vblock Series 700 (Symmetrix)

Alle Aktivitäten werden permanent überwacht und protokolliert (Source IP, Username, Aktion).

Die Audit Informationen sind Teil des Event Log und umfassen:

- Record number
- Records in Sequence
- Offset in Sequence
- Time
- Vendor-ID
- Application-ID
- Application Version
- API-Library
- API-Version
- Host Name
- OS-Name
- OS-Revision

- Client Host
- Process-ID
- Task-ID
- Function Class
- Action Code
- Description / Text
- Username
- Activity-ID

Das Protokoll wird intern in der Symmetrix in einem besonderen Bereich des Symmetrix File Systems gespeichert. Darauf kann von außen nur lesend zugegriffen werden. Auch ein Löschen von außerhalb ist nicht möglich. Die Einträge werden bei Bedarf nach größtem Alter zugunsten neuer Einträge verworfen. Das Extrahieren und Sichern des Protokolls ist möglich.

Neben der Interaktion der Administratoren finden sich hier alle Aktionen und Zugriffe samt Zugriffsverletzungen aller Hosts.

4.4.2 VMware

4.4.2.1 Hypervisor Logging

ESXi unterstützt sowohl lokales als auch remote Logging. Relevante Logfiles sind das vCenter Server Agent Logfile, das hostd Logfile, das VMkernel Logfile und das VMkernel warnings Logfile.

Folgende Informationen finden sich im hostd Log:

- Erfolgreiche und fehlgeschlagene Anmeldungen.
- Änderungen an Rollen und Berechtigungen.
- Änderungen an virtuellen Switches und Maschinen.
- Das Ein- und Ausschalten des Technischen Support Modus beim ESXi (der Lockdown Modus kann das direkte Anmelden verhindern).

4.4.2.2 Virtual Machine Logging

Gefährdung 17: DoS-Attacke durch Überfüllen des Datastores durch VM Gast Logging Daten.

Ein besonderes Augenmerk ist auf die Logging Daten der VM Gäste zu legen, da sich hier auch ein potentielles Angriffsszenario bietet.

Bei jedem Neustart eines VM Gast auf dem Hypervisor werden logging Daten zu diesem Ereignis erzeugt, und auf dem Datastore abgelegt auf dem die VM Gäste liegen. Ein Angreifer könnte sich dieses für eine DoS Attacke zu Nutze machen, indem er einen VM Gast regelmäßig neu startet und/oder zum Absturz bringt. Bleibt dieses über eine lange Zeit unbemerkt, so besteht die Gefahr, dass die Logging Daten den Datastore vollschreiben. Die Konsequenz wäre, dass andere VM Gäste nicht mehr in der Lage wären auf den Storage zu schreiben, was dann sehr wahrscheinlich einen Systemabsturz zu Folge hätte.

Durch Begrenzung der Protokolldateigröße und die Reduktion der aufbewahrten Anzahl der Protokolldateien kann diese Gefährdung aber ausgeschlossen werden.

Weitere Details zur Konfiguration siehe VMware vSphere 4.0 Security Hardening Guide; [04].

4.4.2.3 vCenter Logging

Das vCenter empfängt als zentrale Managementinstanz alle wesentlichen Protokollinformationen, Ereignisse und Alarme aller verwalteten ESX-Server. Das vCenter selbst kann für CIM-Provider (Hardware-Events) task-basierende Events (z.B. Patchen einer bestimmten VM in eine bestimmte Portgruppe) oder performance-basierende Events, Alarmbenachrichtigungen mit bestimmten Aktionen verknüpfen, d.h. diese Statusmeldungen via SMTP oder SNMP an Administrationsteams oder zentrale Protokoll-Kollektoren wie RSA enVision versenden, und/oder benutzerdefinierte Skripte und vordefinierte Aktionen ausführen. Zum Beispiel stellt der ESX über CIM-Provider fest, dass der Lüfter einer CPU ausgefallen ist, verschickt dann einen Trap an das Admin-Team und setzt den ESX-Server in den Wartungsmodus, sodass virtuelle Maschinen vollautomatisch vor dem Fehlschlagen des physischen Hosts wegmigriert werden.

Außerden ist das Objekt Framework des vCenters öffentlich, sodass andere Hersteller wie RSA entsprechende Events in eigenen Produkten empfangen können.

4.4.2.4 vShield Logging

vShield kann neben lokalen Logfiles innerhalb des vShield Managers und der vShield App und Edge Firewalls Ereignisse an einen Syslog Server senden.

4.4.3 Cisco UCS

4.4.3.1 Allgemeines Logging

Der UCSM unterstützt zwei Arten von Logging:

- **Lokales Logging**
  Fault, Events und Audit Daten werden lokal gespeichert und können im UCSM lokal durchsucht werden. Um die Logging Daten zu sichten und Teile davon zu löschen, benötigt ein Benutzer die "AAA Role" oder eine Rolle mit vergleichbaren Privilegien.
- **Remote Logging**
  Fault, Events und Audit Daten werden per Syslog an einen Syslog Server gesendet. Remote und Lokales Logging können gleichzeitig verwendet werden.

Beim lokalen Logging gibt es Einstellungen für "Console, Monitor und File" Logging:

- **Console Logging** betrifft die Meldungen, die auf der seriellen Konsolenverbindung der Fabric Interconnect ausgegeben werden. Hier kann eingestellt werden bis zu welchem Level Meldungen ausgegeben werden (emergencies, alerts oder critical).
- **Monitor Logging** betrifft die Meldungen, die über das CLI bei einer SSH-Verbindung ausgegeben werden, wenn der Befehl "terminal monitor" eingegeben wird. Auch hier kann eingestellt werden bis zu welchem Level Meldungen ausgegeben werden (emergencies, alerts, critical, errors, warnings, notifications, informations oder debugging).
- **File Logging** betrifft die Meldungen, die lokal auf dem Fabric Interconnect gespeichert werden. Auch hier kann eingestellt werden bis zu welchem Level Meldungen ausgegeben werden (emergencies, alerts, critical, errors, warnings, notifications, informations oder debugging). Über die maximale Größe (Size) kann angegeben werden wie viele Logging Einträge maximal gespeichert werden. Überschreitet das Logfile die maximale Größe, werden die ältesten Einträge gelöscht. Dieses File Logging kann nur über das CLI über SSH abgerufen werden. Diese logging location ist getrennt von den im UCSM gespeicherten Events zu verstehen.

4.4.3.2 Events, Faults und Audit Logging

Der UCSM generiert und speichert Events, Faults und Audit Daten.

- Ein **Event** ist eine Meldung über ein Ereignis, das gerade auf dem System vorgegangen ist. So wird z.B. ein Event generiert, wenn ein Blade Server abgeschaltet wird, ein Uplink zum LAN oder SAN "down" geht, etc.
  Events sind Meldungen, die persistent sind und sich nicht ändern.
- Ein **Fault** wird ausgelöst, wenn etwas in dem System ausgefallen ist oder ein Schwellwert überschritten wurde.
  Anders als bei Events kann ein Fault wieder aufgehoben werden, wenn er bestätigt ("cleared") wird. Diese Bestätigung kann automatisch oder manuell erfolgen (einstellbar).
  So kann z.B. ein Fault generiert werden, wenn ein Uplink zum LAN oder SAN "down" geht, und sich automatisch bestätigen ("cleared"), wenn der Uplink wieder zur Verfügung steht. Nachdem der Fehler bestätigt ist, kann er entweder automatisch gelöscht werden, oder persistent im Fault Log stehen bleiben.
  Für eine genaue Beschreibung der generierten Faults siehe Cisco UCS Faults Reference; [13].
- Ein **Audit** Eintrag im Audit Log wird generiert, wenn ein Nutzer eine Änderung an einem Objekt im USCM vornimmt. Dabei wird das genaue Objekt, das geändert wurde, gespeichert, das Datum und die Uhrzeit der Änderung sowie der User (Administrator), der diese Änderung vorgenommen hat.

Der Event, Fault und Audit Log kann lokal auf dem UCSM über die GUI oder das CLI ausgewertet werden, sowie als CSV-Datei heruntergeladen und gespeichert werden. Ein Abruf der Logging Daten über die XML-API ist ebenfalls möglich. Eine manuelle Löschung von Einträgen ist nicht möglich.

Der UCSM speichert 10.000 Einträge aus diesen drei Kategorien. Wenn das Limit erreicht wird, werden automatisch 250 der ältesten Einträge gelöscht. Aus diesem Grund können selektiv Faults, Events und Audit Log Einträge an zu bis zu drei remote Syslog Server gesendet werden.

Für die Syslog Server kann ebenfalls eingestellt werden, bis zu welchem Level Meldungen ausgegeben werden (emergencies, alerts, critical, errors, warnings, notifications, informations oder debugging).

Die Verwendung eines remote Syslog Servers zur zentralen Speicherung und Auswertung von Events, Faults und Audit Daten wird empfohlen.

4.4.3.3 Core Files

Sollte ein katastrophaler Software- oder Hardwaredefekt die UCSM oder Fabric Interconnect Software zum Absturz bringen, wird ein "Core File" geschrieben (auch "Core Dump" genannt; enthält Debug-Daten, die ein Entwickler lesen kann). Das Core File wird auf das Filesystem des Fabric Interconnect geschrieben. Zusätzlich kann ein "Core File Exporter" konfiguriert werden, der die Datei automatisch auf einen remote TFTP-Server überträgt. Die Verwendung dieses "Core File Export" wird empfohlen.

4.4.3.4 System Event Log (SEL) – Blade Server Log

Zusätzlich zu dem UCSM-Logging besteht noch pro Blade Server ein Log im NVRAM (nicht volatiler Speicher) auf dem Blade Management Controller (BMC) des Servers. Dieser SEL speichert Server bezogene Events wie Über- und Unterspannung, thermale Ereignisse sowie eine Reihe anderer hardwarenaher Ereignisse aus dem BIOS. Die SEL-Datei ist 40KB groß. Wenn die maximale Größe überschritten wird, können keine weiteren Ereignisse protokolliert werden. Die SEL-Datei muss geleert werden, um neue Ereignisse zu speichern.

Auf dem UCSM kann und sollte eine SEL-Policy konfiguriert werden, um die SELs der Server zu einem remote Server zu übertragen und lokal zu löschen. Diese Backup Operation kann manuell oder in regelmäßigen Abständen angestoßen werden. Als Protokolle zur Übertragung des Logs stehen FTP, TFTP, SCP oder SFTP zur Verfügung. Es wird empfohlen SCP oder SFTP zu verwenden, weil diese Protokolle verschlüsselt arbeiten.

4.4.3.5 NTP

Für alle Logging Operationen ist es natürlich entscheidend, dass die Uhrzeit und Datum, sowie die Zeitzone auf dem UCSM richtig gesetzt ist. Daher kann und sollte ein NTP-Server für die Synchronisation der Zeit auf dem System verwendet werden.

Erfolgt keine Zeitsynchronisation, dann besteht die Gefahr, dass entscheidende Logging Informationen nicht miteinander in Verbindung gebracht werden können, was die Aufklärung von Incidents erschwert oder sogar unmöglich macht.

4.4.4 MDS und Nexus (NX-OS)

NX-OS verfügt sowohl über ein lokales Logging als auch über die Möglichkeit Logging Daten über Syslog an ein zentrales System wie z.B. enVision zu senden.

Wir empfehlen das Senden sämtlicher Logging Daten über Syslog Nachrichten zum Archivieren und Auswerten an ein zentrales System.

Für weitere Details zur Logging Konfiguration siehe Cisco MDS 9000 Family NX-OS System Management Configuration Guide; [14], Kapitel "Configuring System Message Logging".

Weiterhin wird empfohlen, wenn core files vorliegen, diese automatisch per TFTP vom NX-OS zu einem zentralen Server zu übertragen.

Details hierzu siehe Cisco MDS 9000 Family NX-OS System Management Configuration Guide; [14], Kapitel "Monitoring System Processes and Log" → "Core and Log Files".

Um Logging Daten mit einem korrekten Zeitstempel zu versehen, ist es zwingend notwendig eine Zeitsynchronisierung per NTP zu konfigurieren.

4.4.5 Unified Infrastructure Manager

Unified Infrastructure Manager nutzt oder basiert auf folgenden Diensten:

- vcmaster — Unified Infrastructure Manager Wrapper Script
- controldb — PostgreSQL Database Server
- sonicmq — SonicMQ JMS Server
- httpd — Apache HTTP Web Server
- controldaemon — Unified Infrastructure Manager Application Communication Service
- jboss — JBoss Application Server
- slm-tomcat — Apache Tomcat Server
- voyence — Unified Infrastructure Manager Device Communication Server
- sysadmin — Unified Infrastructure Manager Watchdog Service
- sshdaemon — Unified Infrastructure Manager SSH Proxy Service
- emc_storapid — EMC SMI Provider
- emc-homebase-serverdb — EMC HomeBase Server Database
- emc-homebase-server — EMC HomeBase Server
- healthcheck — Unified Infrastructure Manager Health Check Service

Diese Dienste protokollieren Ereignisse und Informationen u.a. in folgenden Protokoll-Dateien:

Lokation:	Inhalt:
<Product Directory>/slm/logs/slm.log	Service Lifecycle Manager Log
<Product Directory>/jboss/server/vc-server/log/server.log	JBoss Server Log
<Product Directory>/db/controldb/logs/server.postmaster	PostgreSQL Database Log
<Product Directory>/logs/autodisc.log	Autodiscovery Log
<Product Directory>/logs/commmgr.log	Device Communication Log
<Product Directory>/logs/daemon.log	Controldaemon Log
<Product Directory>/logs/backup.log	Data Backup Log
<Product Directory>/logs/restore.log	Data Restore Log
<Product Directory>/logs/smieventreceiver.log	SMI Provider Events Receiver Log
<Product Directory>/logs/ssxfrcgi.log	SysSync Transfer CGI Log
<Product Directory>/logs/syssyncm.log	SysSync Master Log
<Product Directory>/logs/syssyncs.log	SysSync Slave Log
<Product Directory>/logs/sysmonm.log	System Monitor Master Log
<Product Directory>/logs/sysmons.log	System Monitor Slave Log
<Product Directory>/logs/ucsmeventreceiver.log	UCSM Events Receiver Log
<Product Directory>/logs/session.log	Device Session Log
<Product Directory>/logs/healthcheck-0.log	Healthcheck Service Log
/opt/emc/ECIM/ECOM/log/EMCProvider-<date>.log	EMC SMI Provider Log

Mit diesen Protokoll-Dateien ist es z.B. möglich nachzuvollziehen, wer welche Änderungen durchgeführt hat.

4.4.6 Erweiterte Maßnahmen zur Erfüllung von Compliance Anforderungen

Um die o.g. Protokoll-Dateien aus einer Sicherheitsperspektive sinnvoll und mit vertretbarem Aufwand auswerten zu können, ist es empfehlenswert hierfür Werkzeuge einzusetzen, die sicherheitsrelevante Ereignisse miteinander korrelieren, und die daraus resultierende Gefährdung zu interpretieren. Durch diese Heuristiken und Vorfilterungsmechanismen soll der Sicherheitsadministrator auf wesentliche Ereignisse hingewiesen werden.

So können die o.g. Protokoll-Dateien für Compliance-Anforderungen an z.B. RSA enVision übergeben und dort ausgewertet werden.

RSA enVision kann optional zu den Vblock-Varianten erworben werden.

Die RSA enVision Plattform für das Log-Management dient zur Erfassung, Benachrichtigung und Analyse von Protokolldaten, mit der Unternehmen die Compliance vereinfachen und schnell auf gefährliche Sicherheitsprobleme reagieren können. Die RSA enVision 3-in-1-Plattform ist ein SIEM- (Security Information and Event Management) und Log-Managementsystem, mit dem sich große Datenmengen in Echtzeit aus jeder Ereignisquelle und beliebig großen Rechnerumgebungen zusammenstellen und analysieren lassen. Die RSA enVision-Plattform lässt sich einfach skalieren, sodass Filterung und Bereitstellung von Agenten nicht mehr notwendig sind.

4.5 Sichere Remote Systemwartung

Im Rahmen der vereinbarten Wartungsverträge eines oder mehrerer Vblocks kann eine Systemwartung vom Systemhaus "remote" erfolgen.

Gefährdung 18: Benutzung des Remote Administrationszugangs zu Angriffen.

Wird der remote Zugang nicht ausreichend geschützt, dann befindet sich ein Angreifer im Managementnetz und ihm stehen alle Möglichkeiten zu Angriffen offen, wie bereits im Kapitel 4.2 beschrieben.

EMC stellt zur Absicherung des Remote Administrationszugangs eine IP basierende sichere Fernwartungslösung **EMC Secure Remote Support IP (ESRS IP)** zur Verfügung.

ESRS IP ist eine konsolidierte sichere Alternative zu den bisher üblichen Modem-Zugängen der einzelnen Storage Arrays. Die Abbildung 55 zeigt die Architektur der Lösung im Überblick.

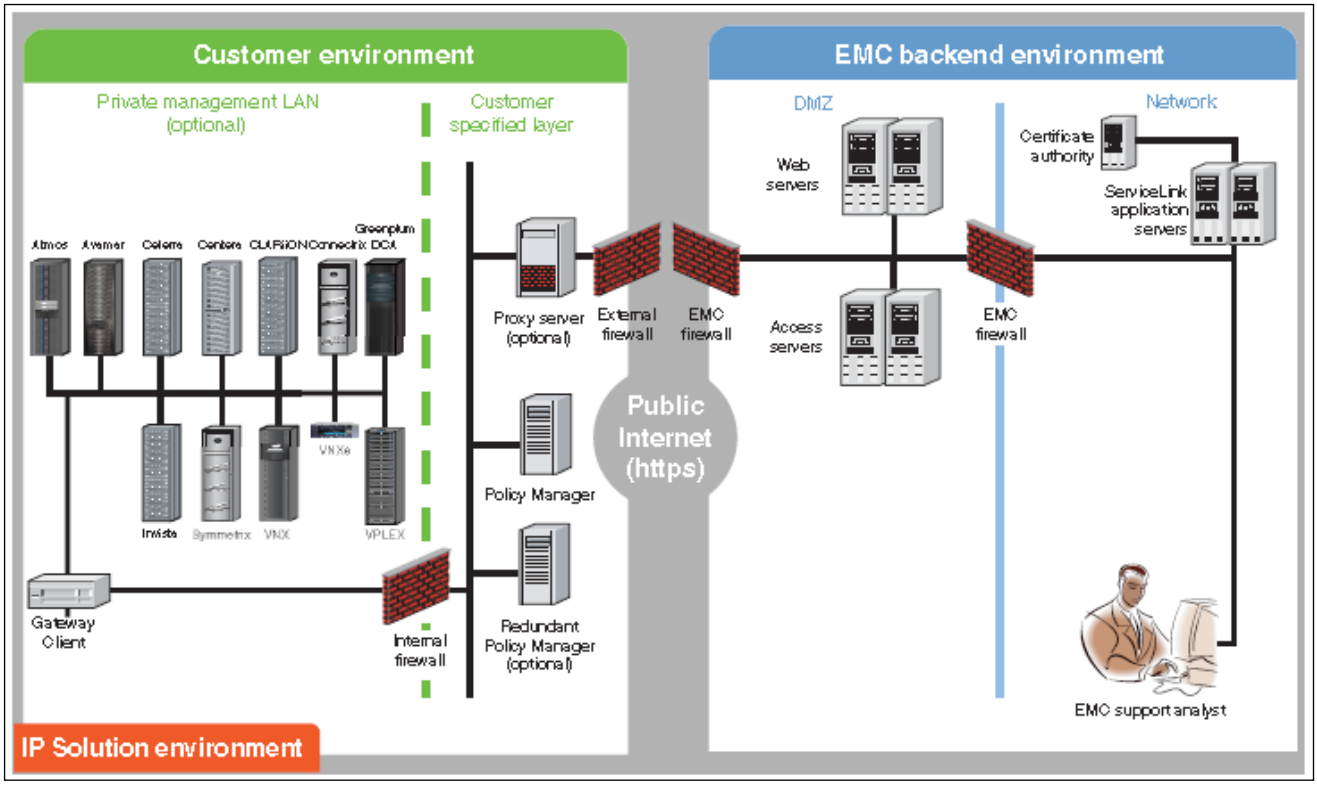


Abbildung 55: EMC Secure Remote Support IP (Übersicht)

Die ESRP IP Lösung baut auf folgenden Funktionalitäten auf:

- **Digitale Sicherheit**
  SSL Datenverschlüsselung durch TLS 1.0 Tunnel mit erweitertem Verschlüsselungsstandard (AES 256 bit), Entitäten Authentisierung mittels privaten digitalen Zertifikaten und Authentisierung via EMC Network Security.
- **Autorisierungskontrolle**
  Zugriffsregeln ermöglichen angepasste Autorisierung: Zulassen, Verweigern oder Erfordern dynamischer Zustimmung zu Systemen auf Ebene der Support-Anwendung und Geräte.
- **Auditing**
  Der ESRS IP Policy Manager protokolliert alle Fernwartungszugriffsversuche, die Ausführung von Skripten und Dateiübertragungen. Die Verwaltung der Protokoll-Dateien erfolgt durch den Kunden.**Sichere Session Tunnel** für Fernwartungszugriff.

4.6 Restrisiken und mögliche Eindämmungen

Die nachfolgende Abbildung 56 zeigt allgemein die Typen von Gefährdungen / Angriffsszenarien, die nicht mit rein technischen Gegenmaßnahmen abgefangen werden können. Diese wurden in den vorhergehenden Kapiteln 4.1 bis 4.5 behandelt.

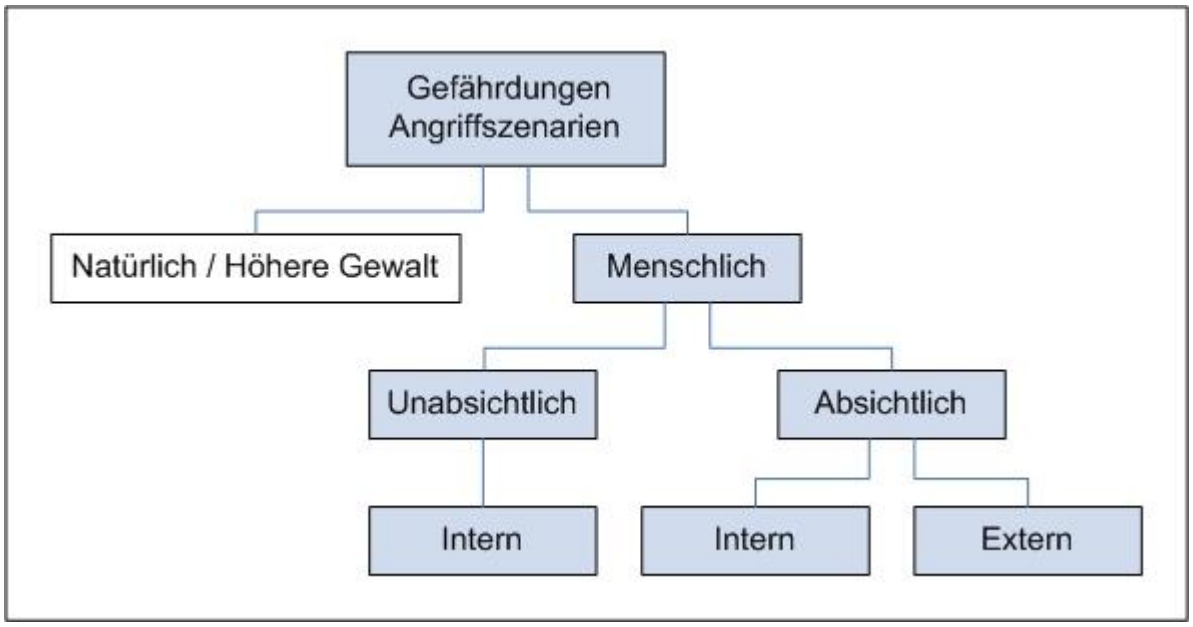


Abbildung 56: Typen von Gefährdungen / Angriffszenarien

Gefährdungen / Angriffsszenarien lassen sich wie folgt gruppieren:

- **Natürliche Gefährdungen / Höhere Gewalt**
  Z.B. Stromausfall, Überschwemmung, Erdbeben, etc.
  Bemerkung: Natürliche Gefährdungen werden in dieser Studie nicht betrachtet.
- **Unabsichtliche Gefährdungen durch Menschen**
  Z.B. fehlerhafte Konfigurationen, fehlerhafte Administration, etc.
- **Absichtliche Gefährdungen durch Menschen (intern)**
  Z.B. Angriffe durch neugierige oder verärgerte Angestellte, mutwillige Fehlkonfiguration, etc.
- **Absichtliche Gefährdungen durch Menschen (extern)**
  Z.B. Angriffe durch Hacker, etc.

Die durch Menschen verursachten möglichen Gefährdungen / Angriffsszenarien für den VCE Vblock sind nachfolgend aufgelistet. Die Reihenfolge stellt dabei keine Priorisierung oder Gewichtung dar.

Unabhängig vom Angriffsszenario müssen wohldefinierte Zugriffsregeln zur Administration definiert und eingerichtet sein. Ebenso muss eine Überwachung und Protokollierung der administrativen und anwenderbedingten Tätigkeiten erfolgen, sowie eine periodische Überprüfung der Konfiguration auf Änderungen stattfinden. Diese Vorgänge können nur mithilfe entsprechender Governance, Risk und Compliance Werkzeuge erreicht werden.

4.6.1 Unabsichtliche Gefährdungen durch Menschen (intern)

Gefährdung 19:	Gefahren durch achtlosen Einsatz von "Virtual Appliances", z.B. könnten "Virtual Appliances" mit Malware befallen sein.

Gegenmaßnahmen:

Testen von "Virtual Appliances" in isolierter Umgebung.

Verweis:

VLANs (Kapitel 3.3.1 und 4.1.3.2), vShield App (Kapitel 3.3.1), Zoning (Kapitel 4.1.4.1), VSANs (Kapitel 3.3.2).

Weitere Gefährdungen sind durch **Fehlkonfigurationen** gegeben. Beispiele hierfür sind:

Gefährdung 20:	Versehentliches Zuweisen einer VM in die falsche Sicherheitszone / Mandant.

Gefährdung 21:	Versehentliches Zuweisen eines oder mehrerer Datastores zu einer VM.

Gefährdung 22:	Falsches Zuweisen / Entfernen von ESX-Servern zu VSANs / Zonen (z.B. falsches Service-Profil gewählt, falsche WWN-Pools verwendet).

Gefährdung 23:	Falsches Zuweisen / Entfernen von Datastores oder Raw-Devices zu / von einem ESX-Cluster.

Gegenmaßnahmen:

- Organisatorische Gegenmaßnahmen wie das Vier-Augen-Prinzip bei kritischen Administrationstätigkeiten.
- Einsatz von geschultem und vertrauenswürdigem Personal.
- Einhaltung der definierten Prozessworkflows zur Administration, Governance, Risk, Compliance.

Verweis:

enVision (Kapitel 4.4.6).

4.6.2 Absichtliche Gefährdungen durch Menschen (intern)

Gefährdung 24:	VM-zu-VM-Kommunikation über "Side Channels"

Gegenmaßnahmen:

- VM Communication Interface (VMCI) standardmäßig deaktivieren und nur im Ausnahmefall nutzen.
- Auditing

Verweis:

enVision (Kapitel 4.4.6).

Gefährdung 25:	Vorsätzliche Fehlkonfiguration durch einen Administrator, die den Zugriff zwischen Mandanten und Sicherheitszonen ermöglicht.

Gegenmaßnahmen:

- Organisatorische Gegenmaßnahmen wie das Vier-Augen-Prinzip bei kritischen Administrationstätigkeiten.
- Einsatz von geschultem und vertrauenswürdigem Personal.

- Einhaltung der definierten Prozessworkflows zur Administration, Governance, Risk, Compliance.

Verweis:

enVision (Kapitel 4.4.6).

Gefährdung 26:	Vorsätzliches falsches Zuweisen / Entfernen von ESX Servern zu VSANs.

Gefährdung 27:	Vorsätzliches falsches Zuweisen / Entfernen von Datastores oder Raw-Devices zu / von einem ESX-Cluster.

Gefährdung 28:	Vorsätzliches falsches Zuweisen / Entfernen von VMs zu / von Netzfreigaben (NAS).

Gefährdung 29:	Vorsätzliches falsches Zuweisen / Entfernen von ESX Servern zu / von Netzfreigaben (NAS).

Gegenmaßnahmen:

- Einsatz von geschultem und vertrauenswürdigem Personal.
- Einhaltung des definierten Prozessworkflows zur Administration.
- Nutzen und Überwachen von Konfigurationsprofilen.
- Protokollierung von bestimmten Administrationsvorgängen wie Zuweisen und Entfernen von Diensten oder Systemen.

Verweis:

enVision (Kapitel 4.4.6).

Gefährdung 30:	Unberechtigtes Löschen von LUNs (Datastores oder RAW Devices).

Gefährdung 31:	Stehlen von vertraulichen Informationen durch Storage-basierte Replikationsmethoden.

Gefährdung 32:	Manipulation von Mandantendaten durch Wiederherstellen älterer oder gefälschter Replikate.

Gegenmaßnahmen:

- Definition von Storagesystem basierten ACLs gemäß der Rollen und Rechte.
- Definition der Zugriffsregeln zur Administration.
- Monitoring, Auditing und Reporting aller administrativen Vorgänge.
- Einsatz von geschultem und vertrauenswürdigem Personal.
- Sicherstellen einer aktuellen und konsistenten Datensicherung.

5 Anhang: Referenzierte Dokumente

Ref. Nr.	Titel	Autor	Quelle
[01]	Vblock Infrastructure Platforms	VCE	http://vce.com/pdf/solutions/vce-vblock-infrastructure-brochure.pdf
[02]	Vblock Infrastructure Platform Architecture Overview	VCE	http://vce.com/pdf/solutions/vce-vblock-infrastructure-reference-architecture.pdf
[03]	Vblock Infrastructure Platforms Technical Overview	VCE	http://vce.com/pdf/solutions/vce-vblock-infrastructure-technical-overview.pdf
[04]	VMware vSphere 4.0 Security Hardening Guide	VMware	http://www.vmware.com/resources/techresources/10109
[05]	vSphere Resource Management Guide	VMware	http://www.vmware.com/pdf/vsphere4/r41/vsp_41_resource_mgmt.pdf
[06]	VLAN Security White Paper	Cisco	http://www.cisco.com/en/US/products/hw/switches/ps708/products_white_paper09186a008013159f.shtml
[07]	Understanding Unicast Reverse Path Forwarding	Cisco	http://www.cisco.com/web/about/security/intelligence/unicast-rpf.html
[08]	Best Practices in Deploying Cisco Nexus 1000v Series Switches on Cisco UCS B Series Blade Servers	Cisco	http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/white_paper_c11-558242.html
[09]	Cisco Nexus 1000v Security Configuration Guide	Cisco	http://www.cisco.com/en/US/docs/switches/datacenter/nexus1000/sw/4_0_4_s_v_1_2/security/configuration/guide/n1000v_security.html
[10]	Cisco MDS 9000 Family NX-OS Security Configuration Guide	Cisco	http://www.cisco.com/en/US/docs/switches/datacenter/mds9000/sw/5_0/configuration/guides/sec/nxos/sec.html
[11]	Cisco UCS Manager GUI Configuration Guide	Cisco	http://www.cisco.com/en/US/docs/unified_computing/ucs/sw/gui/config/guide/141/b_UCSM_GUI_Configuration_Guide_141.pdf
[12]	Managing VMware VirtualCenter Roles and Permissions	VMware	http://www.vmware.com/resources/techresources/826
[13]	Cisco UCS Faults Reference	Cisco	http://www.cisco.com/en/US/docs/unified_computing/ucs/ts/faults/reference/UCSFaultsRef.pdf
[14]	Cisco MDS 9000 Family NX-OS System Management Configuration Guide	Cisco	http://www.cisco.com/en/US/docs/switches/datacenter/mds9000/sw/5_0/configuration/guides/sysmgnt/nxos/snmp.html
[15]	Cisco MDS 9000 Family NX-OS Fabric Configuration Guide	Cisco	http://www.cisco.com/en/US/docs/switches/datacenter/mds9000/sw/5_0/configuration/guides/fabric/nxos/adv.html
[16]	EMC Solutions Enabler Symmetrix Array Management CLI Product Guide	EMC	http://powerlink.emc.com/km/live1/en_US/Offering_Technical/Technical_Documentation/453-004-913.pdf

Ref. Nr.	Titel	Autor	Quelle
[17]	BSI IT-Grundschutzkataloge	BSI	https://www.bsi.bund.de/DE/Themen/weitereThemen/ITGrundschutzKataloge/itgrundschutzkataloge_node.html
[18]	BSI-Eckpunktepapier Cloud Computing	BSI	https://www.bsi.bund.de/ContentBSI/Themen/CloudComputing/Eckpunktepapier/CloudComputing-Eckpunktepapier.html
[19]	BSI-Standard 100-2	BSI	https://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzStandards/ITGrundschutzStandards_node.html
[20]	Software und Hardware Techniques for x86 Virtualization	VMware	http://www.vmware.com/files/pdf/software_hardware_tech_x86_virt.pdf