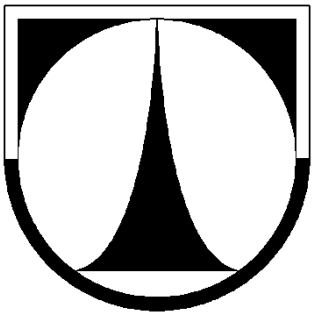
Technická univerzita v Liberci
Fakulta strojní



DIPLOMOVÁ PRÁCE

Implementace služeb vysoké dostupnosti ve virtualizovaném serverovém prostředí

Implementation of high availability services in a virtualized server environment

Liberec 2012 Aleš Pažout

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta strojní

Jméno a Příjmení:	Aleš Pažout
Studijní obor :	M2301 Strojní inženýrství - automatizované systémy řízení ve strojírenství
Zaměření :	Automatizace inženýrských prací

Katedra aplikované kybernetiky

Implementace služeb vysoké dostupnosti ve virtualizovaném serverovém prostředí

Implementation of high availability services in a virtualized server environment

Vedoucí diplomové práce : prof. Ing. Miroslav Olehla, CSc.

Technická univerzita v Liberci

Rozsah diplomové práce:

Počet stran……68

Počet příloh…….. 0

V Liberci dne 14.5.2012

ANOTACE

Technická univerzita v Liberci

Fakulta strojní

Katedra aplikované kybernetiky

Studijní obor :	M2301 Strojní inženýrství - automatizované systémy řízení ve strojírenství
Studijní zaměření :	Automatizace inženýrských prací
Diplomant :	Aleš Pažout
Téma práce :	Implementace služeb vysoké dostupnosti ve virtualizovaném serverovém prostředí
Theme of work :	Implementation of high availability services in a virtualized server environment
Rok obhajoby DP :	2012
Vedoucí DP :	prof. Ing. Miroslav Olehla, CSc.

ANOTACE:

Tato diplomová práce se zabývá implementací mechanismů pro zajištění vysoké dostupnosti vybraných služeb ve virtualizovaném serverovém prostředí. Úvodní kapitoly popisují infrastrukturu virtuálního prostředí a zaměřují se na výběr řešení pro doručení Bootstrap programu v režimu vysoké dostupnosti. Další kapitoly popisují clusterové mechanismy operačního systému Microsoft Windows Server 2008 a jejich praktickou implementaci pro služby souborového serveru a webového rozhraní.

ANOTATION:

This diploma thesis deals with the implementation of mechanisms for ensuring high availability of selected services in a virtualized server environment. Introductory chapters describe the infrastructure of virtual environment and focus on the choice of solutions for delivery of Bootstrap program in high availability mode. Other chapters describe the cluster mechanisms of operating system Microsoft Windows Server 2008 and their practical implementation for file server and web interface services.

Poděkování

Děkuji vedoucímu diplomové práce panu prof. Ing. Miroslavu Olehlovi, CSc. za pomoc a cenné připomínky při zpracování diplomové práce a Ing. Markovi Pažoutovi za odborné konzultace a praktické poznámky při realizaci.

Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména §60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

V Liberci 14. 5. 2012

……

Aleš Pažout

OBSAH

1. ÚVOD

Aktuální dynamické trendy a strategie v oblasti virtualizace dnes přetváří kompletní IT infrastrukturu v centralizovaný soubor služeb podnikového datacentra. Virtualizace se stává novým standardem infrastruktury s maximální provozní efektivitou umožňující flexibilní provedení změn a zajištění vysoké dostupnosti produkčních systémů.

Vývoj jednotlivých virtualizačních technologií a platforem umožnil, aby klíčové vlastnosti dříve téměř výhradně vlastní virtualizaci serverů, plynule pronikly i do oblasti uživatelských desktopů. Virtualizace desktopů přináší zcela nový pohled na poskytování desktopu a aplikací jako služby typu „on demand“ (na vyžádání). Centrální správa kompletní virtuální infrastruktury, vysoká dostupnost a nezávislost uživatelského prostředí od koncového zařízení výrazně snižuje náročnost individuální administrátorské správy a dramaticky také klesají pořizovací a energetické náklady na vybavení a provoz pracoviště.

Vysoká dostupnost služeb a aplikací je vyžadována z důvodů těsného propojení podnikových informačních systémů a jejich závislosti na jednotlivých prvcích IT infrastruktury. Termín vysoká dostupnost již není pouze definován pravděpodobností selhání fyzických komponent, ale obecně snahou značně minimalizovat výpadky a nedostupnost IT infrastruktury jako celku. Prvkem infrastruktury může tedy být hardware typu virtualizačního serveru, podnikový databázový systém nebo služby souborového a tiskového serveru. Prioritou je zde zajištění nepřetržité funkcionality všech nutných zdrojů a služeb pro spolehlivý chod podnikových procesů, kde v případě narušení kontinuity hrozí nemalé finanční ztráty.

Realizace a implementace vysoce dostupných systémů závisí na konkrétním scénáři a pochopení jednotlivých systémových vazeb. Tato diplomová práce se zabývá řešením vysoké dostupnosti kritických serverových služeb ve virtualizovaném prostředí platformy Citrix a operačního systému Microsoft Windows Server 2008.

1.1 Představení DENSO MANUFACTURING CZECH s.r.o.

DENSO MANUFACTURING CZECH s.r.o. se nachází v liberecké Průmyslové zóně – Jih, a je dceřinou firmou japonské nadnárodní společnosti DENSO CORPORATION. nTa vlastní 100% základního jmění společnosti. DENSO sídlí v japonském městě Kariya, v prefektuře Aichi. Je předním světovým dodavatelem moderních technologií, systémů a jejich součástí. V oborech klimatizace, řídících systémů motorů, elektroniky, kontroly řízení a bezpečnosti silničních vozidel, stejně jako v oborech informatiky a komunikace spolupracuje DENSO s hlavními výrobci automobilů po celém světě. Své patentově chráněné technologie a poznatky v odvětví průmyslových systémů a klimatizace využívá DENSO i mimo automobilový průmysl. V současnosti zaměstnává přibližně 123 tisíc zaměstnanců v 35 zemích světa včetně Japonska.

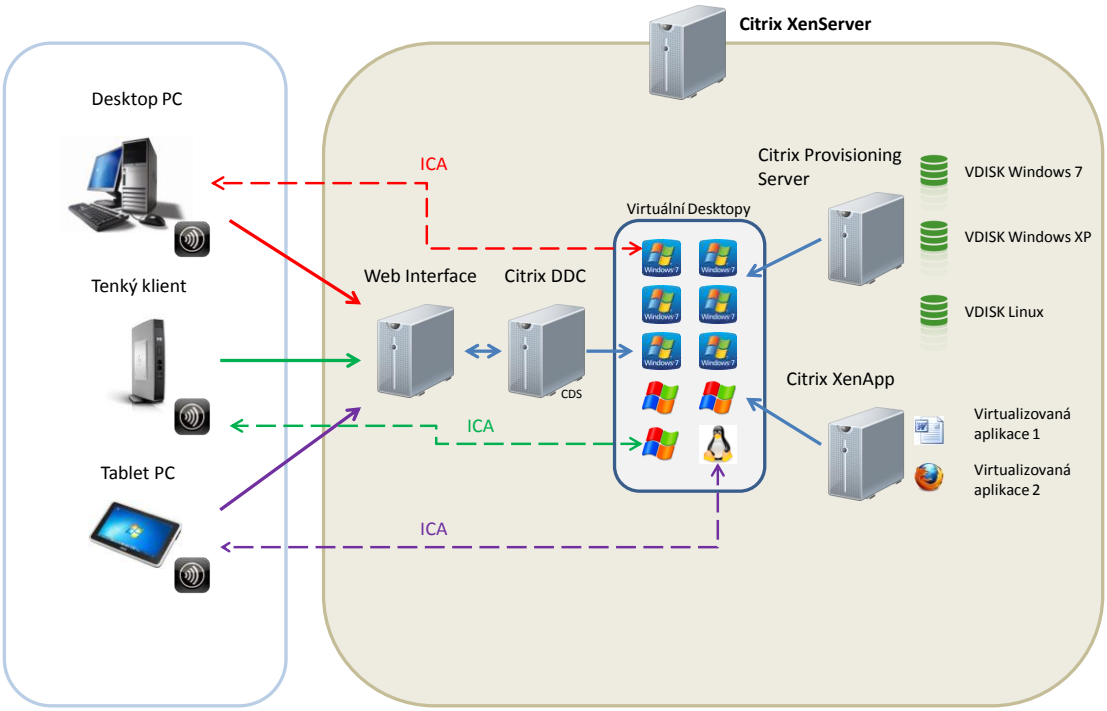
Hlavním výrobním programem společnosti DENSO MANUFACTURING CZECH s.r.o. je výroba klimatizačních jednotek pro osobní automobily a jejich příslušenství (topná tělesa, kondenzátory a chladiče). Zákazníky jsou přední evropské automobilky – VW, BMW, Toyota, Škoda Auto, Audi a Suzuki. V současnosti má firma 1400 zaměstnanců a její plánovaný obrat pro rok 2011 činil 7,6 mld.Kč.

1.2 Výchozí infrastruktura

Infrastruktura virtuálních klientských počítačů (desktopů) realizovaná v průběhu první poloviny roku 2011 ve společnosti Denso Manufacturing Czech (DMCZ) přinesla zcela nové možnosti pro poskytování a správu firemních počítačů. Klíčovými požadavky na aplikační části řešení bylo radikálně snížit časovou náročnost přípravy a správy operačního systému a aplikací pomocí centralizovaných nástrojů, zabezpečení uživatelských profilů a zajištění snadného přístupu pro mobilní zaměstnance, kteří potřebují vzdáleně pracovat s podnikovými aplikacemi.

Předchozí velmi dobré zkušenosti s virtualizačními platformami společnosti Citrix Systems, konkrétně produkty XenServer pro virtualizaci serverů a XenApp pro řešení terminálových služeb, byly jedním z logických předpokladů výběru komplexního systému virtuální desktopové infrastruktury (VDI). Společnost DMCZ zvolila pro řešení VDI další virtualizační platformu z portfolia Citrix Systems, produkt XenDesktop. Hlavní komponentou tohoto produktu je soubor doručovacích technologií FlexCast, který umožňuje sestavit a dynamicky měnit model „koncové zařízení-uživatel-virtuální desktop-aplikace“ spravovaný z jednoho centrálního místa.

Na obrázku 1 je znázorněno schéma VDI infrastruktury ve společnosti DMCZ. Pro připojení k virtuálním desktopům slouží jakákoliv HW platforma koncového zařízení (tencí klienti, stolní PC, chytré telefony, tablety…) s nainstalovanou podporou klientského SW Citrix Receiver. Uživatel se přihlašuje svým doménovým účtem prostřednictvím Citrix Web Interface. Po ověření doménovým řadičem a službami Citrix Desktop Delivery Controller je mu na základě členství ve skupinách generován příslušný datový soubor, který je lokálně interpretován pomocí Citrix Receiver na koncovém zařízení a obsahuje veškeré potřebné parametry pro navázání relace „koncové zařízení – virtuální desktop“ prostřednictvím přenosového protokolu ICA. Virtuální desktopy používají služeb farmy Citrix Provisioning Server pro streaming operačního systému z počítačové sítě a služeb farmy serverů Xenapp pro streaming virtualizovaných aplikací. Pro virtualizaci klientských desktopů a serverovových farem slouží farma hypervisorů XenServer. Komponenty a služby VDI infrastruktury jsou detailněji popsány v následující kapitole.

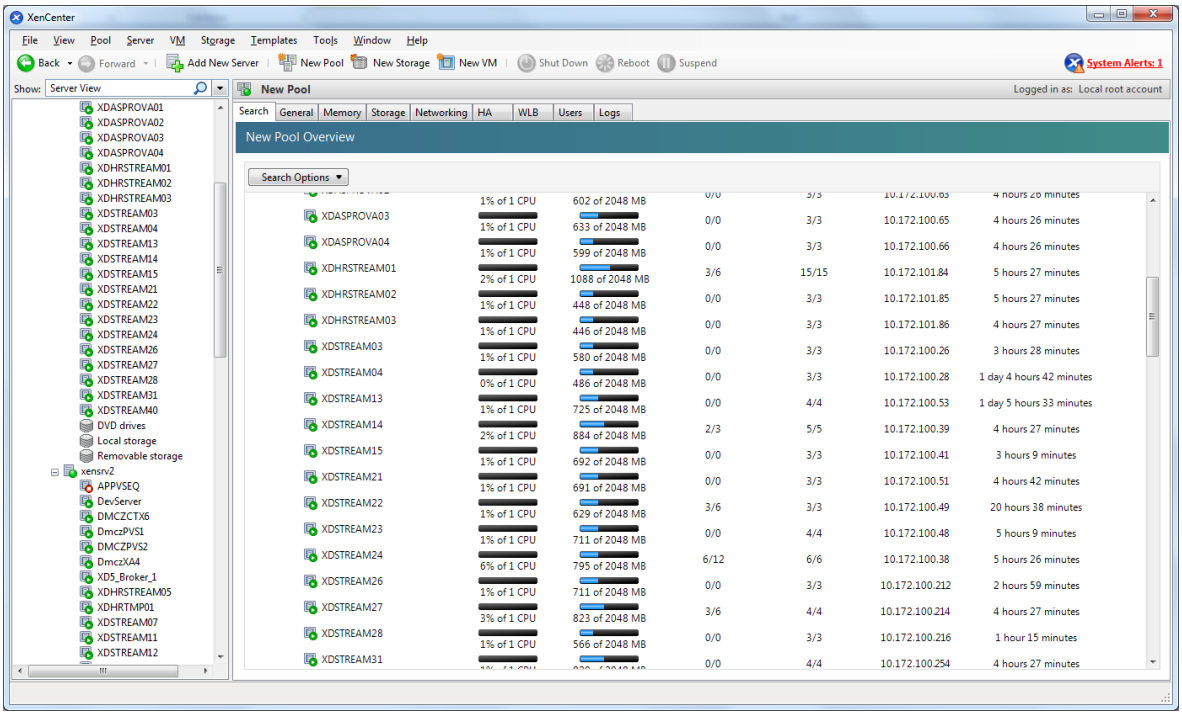


Obr. 1: VDI infrastruktura v DMCZ

1.2.1 Popis jednotlivých komponent infrastruktury

Citrix XenServer

Virtualizační hypervisor na bázi operačního systému Linux umožnující provozovat (hostovat) řadu virtuálních strojů (VM) s různými typy operačních systémů na jednom fyzickém serveru. Podporuje různé typy datových úložišť pro virtuální disky (NFS, HBA, ISCSI). Ve skupinovém režimu (tzv. Pool) podporuje funkce vysoké dostupnosti a umožnuje přesun VM za chodu mezi jednotlivými servery. Pro správu (převážně) slouží grafická konzole XenCenter, lze ovšem využít i příkazového interpreteru přímým připojením prostřednictvím protokolu SSH.



Obr. 2: Citrix XenCenter - administrátorská konzole

Citrix Desktop Delivery Controller (DDC)

DDC je klíčovou komponentou virtualizovaného prostředí XenDesktop. Soubor služeb controlleru zahrnuje:

- doménově autentifikovaný přístup uživatelů k jejich prostředkům s aplikací příslušných skupinových nebo individuálních politik.

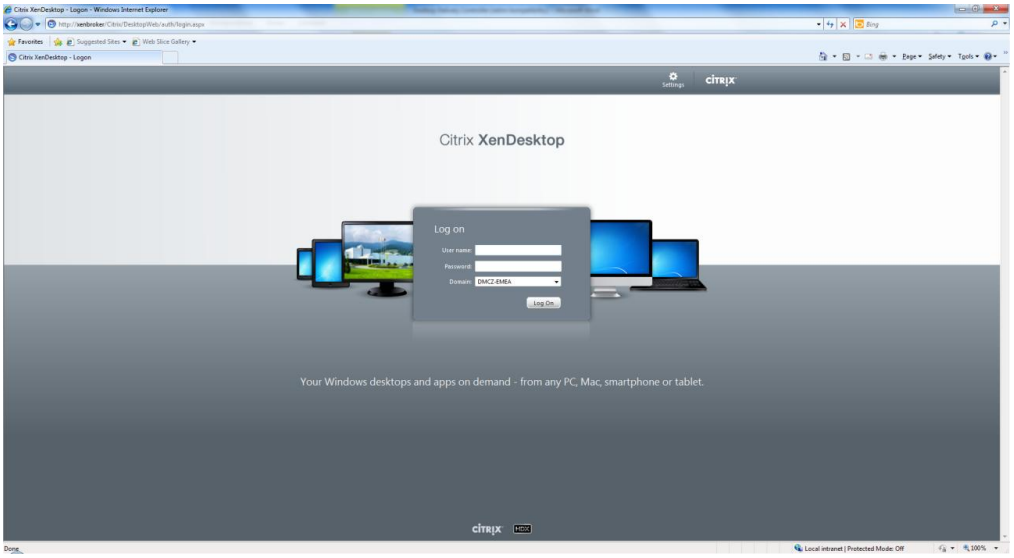
- mechanismus Desktop Broker pro vypublikovaní uživatelských virtuálních nebo fyzických desktopů umožňující takto vytvořené desktopy sdružovat do uživateli sdílených skupin (pool) nebo přiřazovat v relaci 1:1 (jeden uživatel – jeden konkrétní desktop)

- IMA služby pro vzájemnou komunikaci controllerů zařazených do tzv. farem a monitoring událostí (events) spravovaných desktopů (In use, Disconnected, Unregistered…)

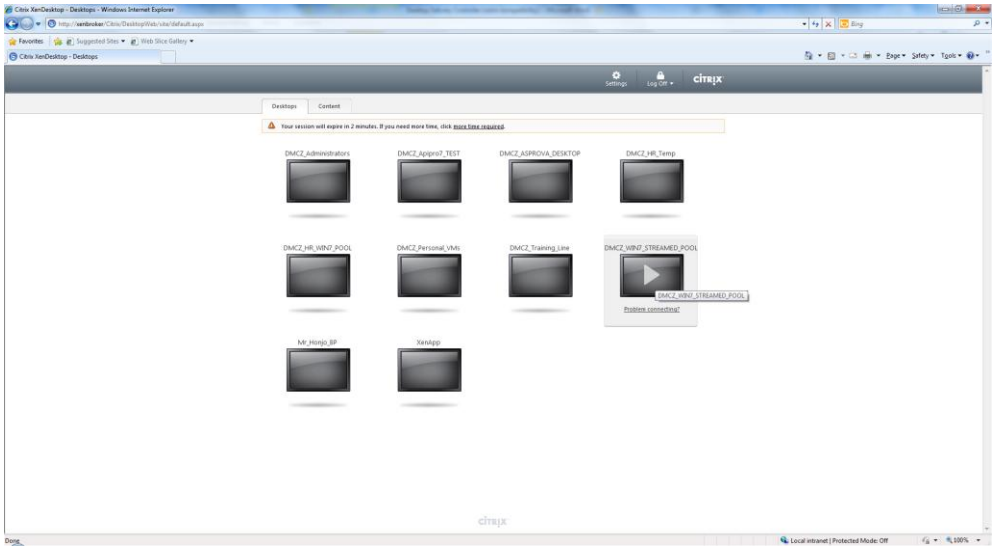
Skupiny DDC serverů umožnují vytvoření farmy pro zajištění vysoké dostupnosti svých služeb.

Citrix Web Interface

Webové rozhraní hostované Microsoft Internet Information Services (IIS) pro uživatelský přístup k virtuálním a desktopům a vypublikovaným aplikacím. Na základě autentifikace uživatele generuje grafické nebo textové aktivní linky pro připojení k požadovanému obsahu.



Obr. 3: Citrix Web Interface – úvodní přihlašovací stránka



Obr. 4: Citrix Web Interface – aktivní linky

ICA service

Služba využívající přenosový protokol ICA (Independent Computing Architecture) pro zprostředkování komunikace mezi virtuálním desktopem a koncovým zařízením uživatele (tenký klient, klasické stolní PC, tablet…). Zajišťuje přenos grafických dat a spravuje periferie (klávesnice, myš, tiskárny, USB disky) koncového zařízení.

Citrix Desktop Service (CDS)

Služba pro zprostředkování komunikace mezi virtuálním desktopem a DDC. Inicializuje uživatelskou relaci podle nastavených pravidel a slouží i k interakci relace se správcovskou konzolí.

Citrix Provisioning Server (PVS)

Služby PVS slouží pro zavedení a následného připojení (tzv. mount) obrazu systémového disku ze síťového úložiště prostřednictvím počítačové sítě. Tato vlastnost bývá označována jako provisioning operačního systému. Konkrétní obraz systémového disku (tzv. Vdisk) je zde reprezentovaný jedním datovým souborem uloženým v centrálním úložišti a jeho bloky jsou trvale přenášeny prostřednictvím síťě na základě požadavků klienta. Přiřazení příslušného Vdisku klientovi (virtuální nebo desktop PC) je realizované vazbou na příslušnou MAC adresu síťového adaptéru klienta.

Vdisk může být poskytován (streamován) z počítačové sítě jako:

- Privátní, kde jsou veškeré změny souborového systému trvale ukládány
- Standard (multi-device), kde jsou veškeré změny souborového systému ukládány do tzv. PVS cache a po restartu zahozeny

Klíčovou vlastností PVS jsou právě Vdisky typu Standard. Mechanismus PVS zde umožňuje současně připojit jeden konkrétní obraz systémového disku (tzv. gold image) několika

klientům, kde každý obdrží jinou IP adresu, jedinečné jméno v síti (Hostname) a bezpečnostní identifikátor (Security Identifier). Pro administrátory představuje tento centralizovaný způsob doručení operačního systému značnou časovou úsporu oproti individuální správě lokální instalace OS a umožňuje pružně reagovat na požadované změny konfigurace OS a sady programového vybavení. Centralizace zde také výrazně usnadňuje zálohovací proces. Skupiny PVS serverů umožňují vytvoření farmy pro zajištění vysoké dostupnosti svých služeb.

Name	MAC	Booting From	vDisk	IP Address	Server	Description
XDSTREAM01	8A-7B-88-71-CC-...	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.25	DMCZPVS2	
XDSTREAM02	8E-61-E3-81-9E-E7	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.27	DMCZPVS2	
XDSTREAM03	AE-41-E6-EC-A8-45	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.26	DMCZPVS2	
XDSTREAM04	AE-A3-00-88-57-93	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.28	DMCZPVS1	
XDSTREAM05	86-1B-8F-93-0A-45	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.30	DMCZPVS2	
XDSTREAM06	EA-6D-B3-29-80-41	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.34	DMCZPVS2	
XDSTREAM07	36-40-64-BB-17-D6	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.35	DMCZPVS2	
XDSTREAM08	AA-DE-AF-FF-F1-...	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.45	DMCZPVS1	
XDSTREAM09	FE-DE-F3-4C-33-07	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.40	DMCZPVS1	
XDSTREAM10	6A-23-5B-0B-4A-E6	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.36	DMCZPVS2	
XDSTREAM11	0E-96-52-14-36-A4	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.52	DMCZPVS1	
XDSTREAM12	B2-F6-2A-E5-71-BA	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.44	DMCZPVS2	
XDSTREAM13	0E-F4-9A-BA-00-35	vDisk	ipstor_150GB\XDSYSVOLt	10.172.100.53	DMCZPVS1	

Obr. 5: Citrix Provisioning Server – výpis síťových klientů

1.2.2 Klíčové požadavky společnosti DMCZ

Klíčovým požadavkem společnosti byla eliminace tzv. Single Point Of Failure (SPOF), tedy jednoho bodu reprezentujícího služby virtuální infrastruktury, při jehož selhání dojde k úplnému výpadku nebo značnému omezení funkčnosti celku. Konkrétně se zadání vymezilo na zajištění služeb vysoké dostupnosti (redundanci) pro následující služby a komponenty infrastruktury:

- Doručení bootstrap souboru prostřednictvím protokolu TFTP
- Souborový server hostující síťové profily uživatelů
- Webové rozhraní Citrix Web Interface pro přihlašování uživatelů

Dalšími obecnými požadavky na řešení bylo:

- Maximální využití dostupných zdrojů
- Možnost monitoringu kritických služeb
- Intuitivní správa

V následující kapitole jsou jednotlivě popsány body SPOF příslušných služeb infrastruktury s návrhem a realizací řešení s cílem jejich eliminace.

2. ZAJIŠTĚNÍ VYSOKÉ DOSTUPNOSTI SLUŽBY TFTP PRO DORUČENÍ BOOTSTRAP SOUBORU

Citrix Provisioning Server umožňuje doručit operační systém ze síťového úložiště prostřednictvím počítačové sítě typu LAN na běžné desktopové PC nebo virtuální stroje hostované v datovém centru. Pro zavedení operačního systému ze sítě je nutná prvotní konfigurace síťových parametrů prostřednictvím protokolu DHCP s pomocí Boot-rom zavaděče PXE a následné zajištění připojení prostřednictvím protokolu TFTP k vzdálenému serveru poskytujícímu provisioning služby.

2.1 Dynamic Host Control Protocol (DHCP)

DHCP protokol definovaný v RFC 2131 umožňuje prostřednictvím DHCP serveru nastavovat stanicím v počítačové síti sadu parametrů nutných pro komunikaci pomocí IP protokolu. Typicky se pomocí DHCP nastavují (přidělují) tyto parametry:

- IP adresa
- Maska sítě
- Implicitní brána (Default Gateway)
- DNS server (seznam jedné nebo více IP adres DNS serverů)

V případě požadavku na zavedení operačního systému z počítačové sítě je možné rozšířit konfigurační parametry o následující:

- IP adresu TFTP serveru v síti
- Jméno Bootstrap souboru (programu)

Princip klientské žádosti o přidělení IP adresy a dalších síťových konfiguračních parametrů je následující. Po připojení do sítě klient vyšle broadcastem (síťový rámec, který přijmou všechna připojená zařízení) DHCPDISCOVER paket. Na ten odpoví DHCP server paketem DHCPOFFER s nabídkou IP adresy. Klient si z (teoreticky několika) nabídek vybere jednu IP adresu a o tu požádá paketem DHCPREQUEST. DHCP server mu ji vzápětí potvrdí odpovědí DHCPACK. Jakmile klient obdrží DHCPACK, může už IP adresu a zbylá konfigurační nastavení používat po dobu zápůjčky (DHCP Lease Time).

2.2 Preboot Execution Environment (PXE)

PXE je označení technologie pro bootování (tj. start) počítačů z počítačové sítě. Typicky se využívá pro bezdiskové tenké klienty a pro automatické instalace operačních systémů. PXE standard přídává k firmware síťové karty část rozšiřujícího kódu, který umožní zavedení operačního systému z počítačové sítě.

Síťové bootování pomocí PXE využívá Internet Protocol (IP) z rodiny protokolů TCP/IP. Základní postup je následující:

1) Po inicializaci požádá síťový klient pomocí DHCP o základní konfigurační síťové parametry
2) Pokud byl v první odpovědi DHCP serveru určen TFTP server, PXE ho kontaktuje podruhé pomocí DHCPREQUEST se žádostí o sdělení jména souboru bootstrap programu (souboru). DHCP server odpoví pomocí DHCPACK.
3) Nyní dojde ke stažení bootstrap programu (NBP) do operační paměti RAM. Přenos NBP je z TFTP serveru proveden pomocí UDP protokolu.
4) Stažený NBP je spuštěn. Další síťovou komunikaci obstarává sám NBP. Bootstrap program může obsahovat libovolný kód (službu) pro inicializaci zavedení operačního systému z počítačové sítě.

2.3 Trivial File Transfer Protocol (TFTP)

TFTP (RFC 2347) je velice jednoduchý protokol pro přenos souborů, obsahující jen základní funkce protokolu FTP.

TFTP je určen pro přenos souborů v případech, kdy je běžný protokol FTP nevhodný pro svou komplikovanost. Typickým případem je právě bootování bezdiskových strojů ze sítě, kdy se celý přenosový protokol musí vejít do omezeného množství paměti, která je k dispozici na bezdiskovém stroji.

Jelikož TFTP funguje nad nespojovaným protokolem UDP, musí obsahovat vlastní řízení spojení. Koncepce relace je jednoduchá: v jednom spojení lze přenést jen jediný soubor, při komunikaci se na síti pohybuje vždy jen jediný paket (po odeslání jednoho paketu program

čeká na jeho potvrzení a teprve poté posílá další). Kvůli tomuto zjednodušení poskytuje protokol na linkách s velkou latencí jen malou přenosovou rychlost.

Oproti FTP má různá omezení a odlišnosti:

- Nelze procházet adresáře.
- Neumožňuje přihlášení uživatele ani zadání hesla.
- Neumožňuje navázání přerušeného přenosu
- Maximální velikost přenášeného souboru je 32 MB.

2.4 Výchozí stav pro doručení Bootstrap programu v DMCZ

Společnost Denso používá pro zajištění provisioning služeb dvou serverů (uzlů) v režimu clusteru:

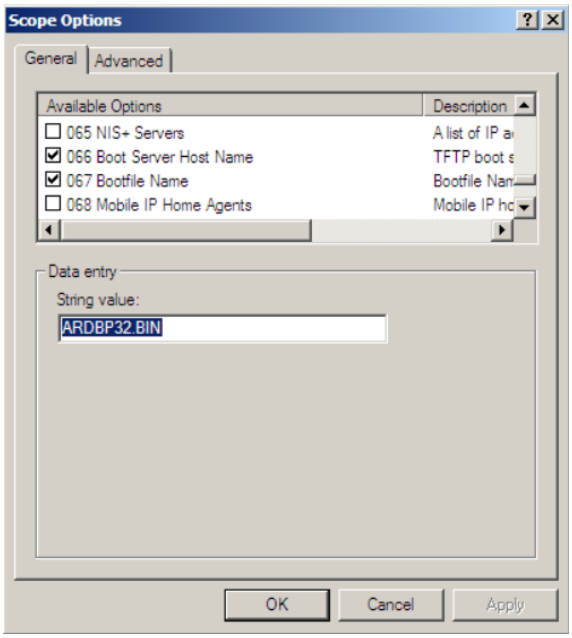
- DMCZPVS1 (IP adresa 10.172.103.1)
- DMCZPVS2 (IP adresa 10.172.103.2)

Režim clusteru typu active-active umožňuje využít mechanismů vysoké dostupnosti (high availability) pro vyvažování zátěže (load-balancing) a také pro automatické převzetí služeb při selhání (failover) jednoho uzlu druhým funkčním uzlem.

Oba servery poskytují službu TFTP serveru a hostují bootstrap program – soubor ARDBP32.BIN, který obsahuje obslužný kód a záznam IP adres Provisioning Serverů pro inicializaci a zajištění relace síťový klient (virtuální stroj) – služby Provisioning Serverů.

Pro doručení Bootstrap programu síťovému klientovi se využívá příslušných DHCP záznamů, konkrétně to jsou:

- DHCP Option 66 s IP adresou serveru DMCZPVS1 (server TFTP)
- DHCP Option 67 se jménem Bootstrap programu ARDP32.BIN

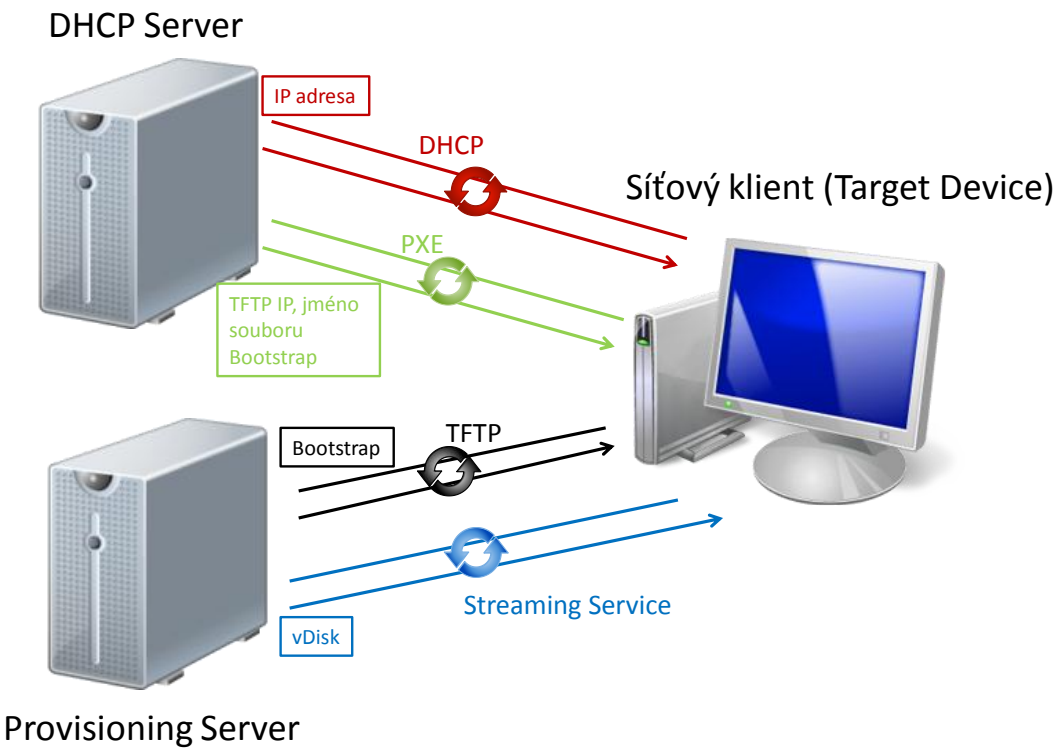


Obr. 6: DHCP volby 66 a 67

Proces startu (boot) virtuálního stroje z počítačové sítě zde probíhá následovně:

1) Nastavená volba pro bootování ze sítě pomocí PXE zapříčiní DHCP broadcast
2) DHCP server přijme klientský požadavek a odpoví zasláním síťových konfiguračních parametrů (IP adresa, síťová maska, výchozí brána, DNS servery)
3) Součástí odpovědi je i IP adresa Provisioning Serveru DMCZPVS1 (10.172.103.1) a jméno souboru bootstrap programu (ARDBP32.BIN)
4) Klient kontaktuje Provisioning Server se žádostí o zaslání bootstrap programu
5) Provisioning Server naváže spojení s klientem a odešle bootstrap program prostřednictvím protokolu TFTP
6) Klient stažený bootstrap program uloží do paměti, zpracuje a kontaktuje služby prvního dostupného Provisioning Serveru (tzv. login server)
7) Provisioning Server prohledá databázi definovaných klientských MAC adres adaptérů a v případě shody dále kontroluje přiřazení virtuálního disku (Vdisk)
8) V případě přiřazení Vdisku pro příslušnou MAC adresu klienta se pomocí load-balancing algoritmu vybere aktuálně nejméně vytížený Provisioning Server
9) Klient provede tzv. mount (připojení) Vdisku a vybraný server iniciuje síťový stream (přenos dat) pro zavedení operačního systému a jeho následný start

Klíčovým problémem tohoto scénáře je zde třetí bod - záznam IP adresy pouze jednoho Provisioning Serveru (DMCZPVS1), který zde představuje Single Point of Failure (SPOF). Při selhání jeho služby TFTP nelze virtuálnímu stroji doručit bootstrap program a je tedy kompletně přerušen jeho startovací proces. Následný opravný proces zahrnuje ať již manuální změnu administrátorem nebo pomocí monitoringu TFTP služby vyvolaným skriptem záznamu DHCP Option 66 na druhý Provisioning Server (DMCZPVS2) a dále manuální restart postižených virtuálních strojů prostřednictvím administrátorské konzole. Následující kapitola popisuje další možné alternativní scénáře doručení Bootstrap programu prostřednictvím služby TFTP s klíčovým požadavkem na zajištění její vysoké dostupnosti.

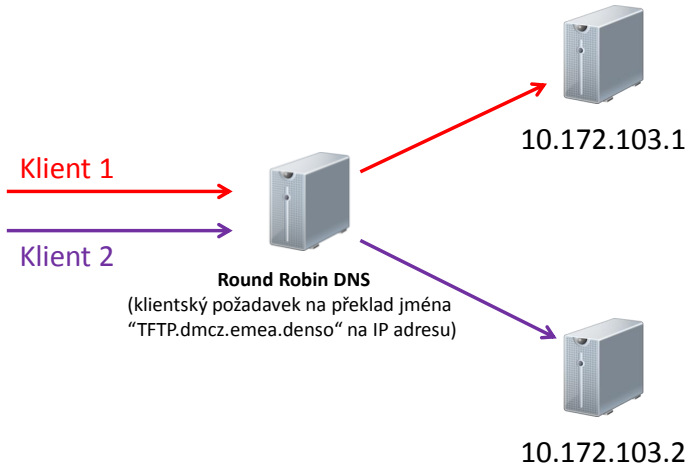


Obr. 7: Schéma doručení bootstrap programu pomocí DHCP voleb 66 a 67

2.5 Metody pro zajištění vysoké dostupnosti služby TFTP

2.5.1 DHCP Option 66 s použitím DNS Round Robin

Metoda DNS Round Robin (RFC 1794) poskytuje základní nízko úrovňovou implementaci vysoké dostupnosti a typicky se používá u webových farem pro rozložení zátěže. Princip vychází z jednoduchého algoritmu kruhového dotazování. Jmennému FQDN DNS záznamu typu A je přiřazeno několik cílových IP adres serverů, které v případě klientského dotazu na překlad jména na IP adresu rotují v pořadí a odpovědi tedy přichází střídavě. V našem případě by se jednalo o vytvoření jmenného záznamu, např. TFTP.dmcz.emea.denso, ke kterému by byly přiřazeny IP adresy obou Provisioning Serverů 10.172.103.1 a 10.172.103.2. Tento jmenný záznam by byl následně vepsán jako DHCP Option 66. První klient žádající systém DNS o překlad TFTP.dmcz.emea.denso na IP adresu by dostal odpověď v podobě IP adresy 10.172.103.1, druhý klient se stejným požadavkem by dostal odpověď v podobě IP adresy 10.172.103.2. Další požadavky klientů by byly opět vyřizovány v tomto pořadí. Toto řešení již ze svého principu není vhodné pro skutečné zajištění služeb vysoké dostupnosti, v případě selhání jednoho ze serverů zde totiž dojde k neúspěchu síťového startu u 50-ti procent virtuálních strojů.



Obr. 8: Mechanismus DNS Round Robin

2.5.2 DHCP Option 66 s více záznamy adres TFTP

Některá řešení DHCP serveru umožňují k zajištění vysoké dostupnosti služby TFTP záznam několika různých adres (typicky oddělených středníkem) při konfiguraci DHCP Option 66. V tomto případě je pro správnou interpretaci záznamu vyžadována příslušná podpora síťové části ve firmware klientského zařízení. Služba DHCP serveru od společnosti Microsoft používaná v DMCZ neumožňuje při použití DHCP Option 66 současný zápis více adres TFTP serverů.

2.5.3 Citrix Netscaler

Citrix Netscaler je HW zařízení typicky používané pro vyvažování zátěže síťového provozu (load-balancing) při poskytování webových aplikací.

Pro zajištění vysoké dostupnosti služby TFTP lze využít módu „Direct Server Return“ s přiřazením loopback síťových adaptérů PVS serverům s nastavenou IP adresou, která je stejná jako virtuální IP adresa Netscaleru (zapsaná jako DHCP Option 66) reprezentující TFTP služby jednotlivých serverů.

Druhým módem je „Global Server Load Balancing“ (nutná licence Platinum), kde se využívá schopnosti integrace služeb Netscaleru se systémem DNS pro překlad jmenného názvu na příslušnou IP adresu. V DHCP volbě 66 je zapsán jmenný alias a při požadavku klienta na přeložení tohoto jména na jeho IP adresu je pomocí služby Netscaleru nazvané „Authoritative DNS Listener“ vrácena IP adresa TFTP serveru, která byla aktuálně zvolená prostřednictvím vyvažovacích algoritmů v součinnosti s monitoringem samotné dostupnosti služeb TFTP.

Citrix Netscaler je poměrně nákladné zařízení (od 75 000 Kč). Jako samotné představuje SPOF a je tedy nutné vybavit infrastrukturu záložním.

Obr. 9: Citrix Netscaler

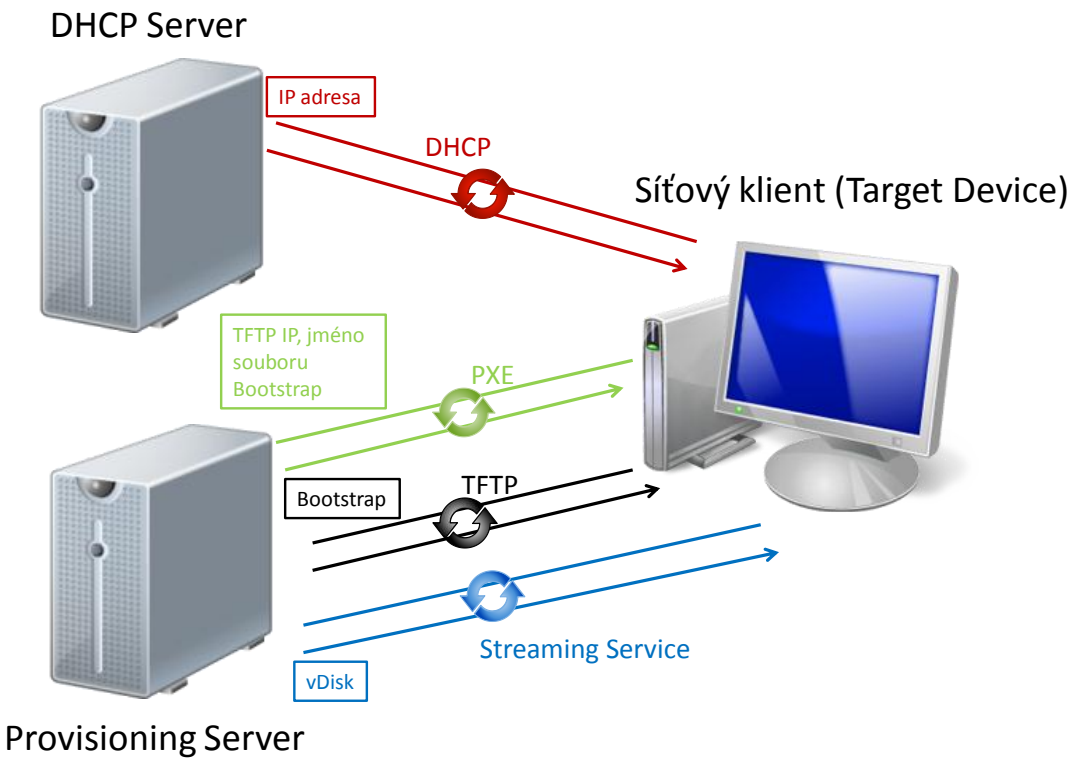
2.5.4 Proxy DHCP (Citrix PXE Services)

Velmi vhodným řešením vysoké dostupnosti služby TFTP je příslušná re-konfigurace Provisioning Serverů jako serverů typu Proxy DHCP a zavedení služby Citrix PXE Service. V tomto případě není nutný statický explicitní záznam TFTP serveru jako DHCP Option 66 a jména bootstrap programu jako DHCP Option 67.

Citrix PVS PXE Service	Citrix PVS ...	Started	Automatic	Local Service
Citrix PVS Ramdisk Server	Citrix PVS ...		Manual	DMCZ-EMEA\p..
Citrix PVS Soap Server	Citrix PVS ...	Started	Automatic	DMCZ-EMEA\p..
Citrix PVS Stream Service	Citrix PVS ...	Started	Automatic	DMCZ-EMEA\p..
Citrix PVS TFTP Service	Citrix PVS ...		Automatic	Local Service

Obr. 10: Služby PVS

Služba Citrix PXE Service reaguje na klientský požadavek typu DHCPDISCOVER, kde klient současně s odpovědí od DHCP serveru obsahující pouze základní konfigurační síťové parametry (IP adresa, maska podsítě, výchozí brána a servery DNS) obdrží prostřednictvím Citrix PXE služby potřebné informace o IP adrese TFTP serveru a jménu bootstrap programu. Odpověď PXE služby z Provisioning Serveru DMCZPVS1 tedy obsahuje jeho IP adresu (10.172.103.1) jako serveru TFTP, podobně odpověď PXE služby z Provisioning Serveru DMCZPVS2 obsahuje IP adresu 10.172.103.2 jako serveru TFTP.

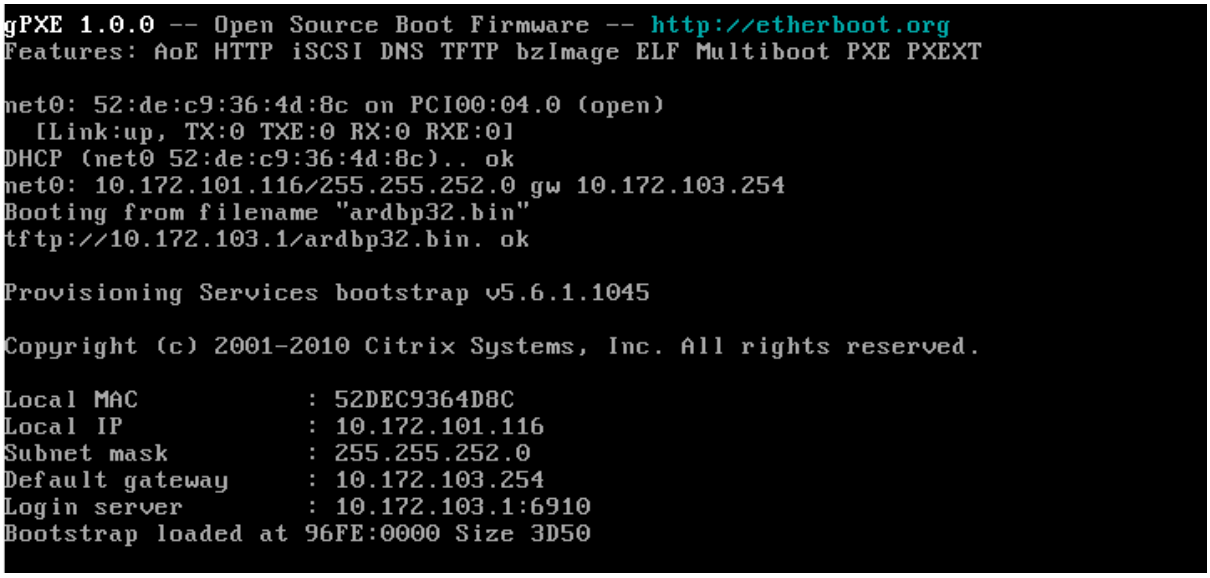


Obr. 11: Schéma doručení Bootstrap programu pomocí Citrix PXE služby

Paralelní běh služby Citrix PXE na obou Provisioning Serverech zde umožňuje zajistit vysokou dostupnost služby TFTP. Služba Citrix PXE pracuje podobně jako DHCP, kde klient zpracovává odpověď od serveru s nejrychlejší odezvou. V případě nedostupnosti prvního serveru tedy klient zpracuje odpověď od druhého serveru. Problematickými aspekty implementace tohoto řešení mohou být vzájemné kolidace s jinou PXE službou (např. Altiris) ve stejné broadcast doméně a také nutnost příslušnosti do stejné broadcast domény pro Provisioning Servery a jejich klienty.

Při implementaci služby PXE je také nutné myslet na fakt, že toto řešení nás neochrání před výpadkem služby TFTP na příslušném Provisioning Serveru. Pokud například klient obdrží odpověď od PXE služby z prvního serveru DMCZPVS1 a TFTP služba tohoto serveru je například z důvodu jejího dřívějšího selhání dočasně pozastavena operačním systémem, klient ztrácí možnost síťového startu. Vhodným doplňkem je tedy monitoring služby TFTP na Provisioning Serverech ať již prostřednictvím vlastních skriptů nebo komplexními

diagnostickými systémy typu NAGIOS a vytvoření závislostí stavu (start/stop) služby PXE na stavu služby TFTP.



Obr. 12: Síťový start (Boot) klienta s pomocí služby Citrix PXE

2.6 Boot Device Manager

Utilita Boot Device Manager (BDM) dostupná v instalaci Provisioning Serveru představuje na rozdíl od předchozích řešení zcela odlišný způsob doručení Bootstrap programu síťovému klientovi. V tomto případě není Bootstrap program doručen prostřednictvím sítě, ale je obsažen ve speciálním bootovatelném diskovém obrazu formátu ISO připojeného ke klientovi jako virtuální optická mechanika, pomocí kterého síťový klient startuje. Tato metoda doručení tedy nevyžaduje konfiguraci a přítomnost serverových služeb TFTP a PXE. Nutnou službou je zde pouze služba DHCP umožňující klientovi získat základní síťové parametry. Diskový obraz obsahující Bootstrap program (zde soubor TSBBDM.BIN) je sestaven dle požadovaných konfiguračních parametrů, klíčovým nastavením jsou zde IP adresy Provisioning serverů pro navázání klientské relace se streamovací službou.

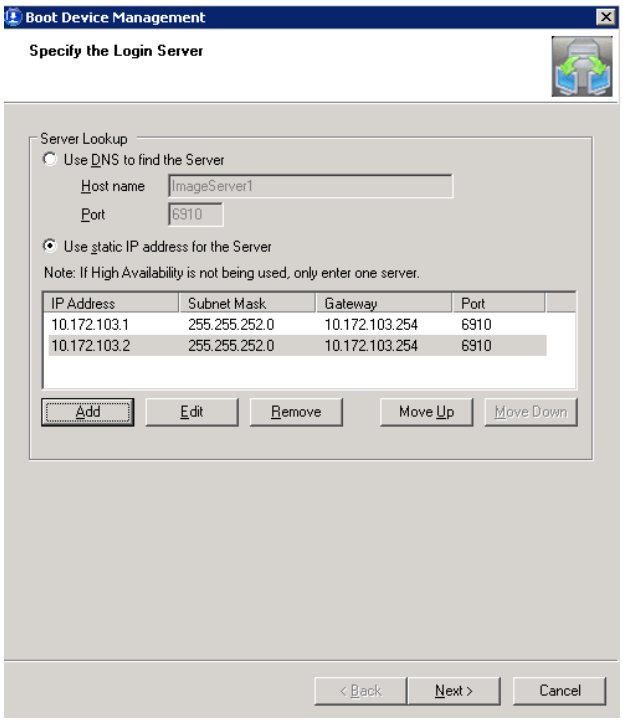
```
DEVICE MAC: 52 DE C9 36 4D 8C
DHCP Discover .
DHCP Request .
DEVICE IP: 10.172.101.116  NETMASK: 255.255.252.0  GATEWAY: 10.172.103.254
DHCP SERVER IP: 10.172.103.254  DNS IP: 172.20.93.60 172.20.93.62  DOMAIN: dmcz.
emea.denso


Bootstrap length: 33186 bytes.
Provisioning Services bootstrap v5.6.1.1045

Copyright (c) 2001-2010 Citrix Systems, Inc. All rights reserved.

Local MAC           : 52DEC9364D8C
Local IP            : 10.172.101.116
Subnet mask         : 255.255.252.0
Default gateway     : 10.172.103.254
Login server        : 10.172.103.1:6910
Bootstrap loaded at 9662:0000 Size 3DD0
```

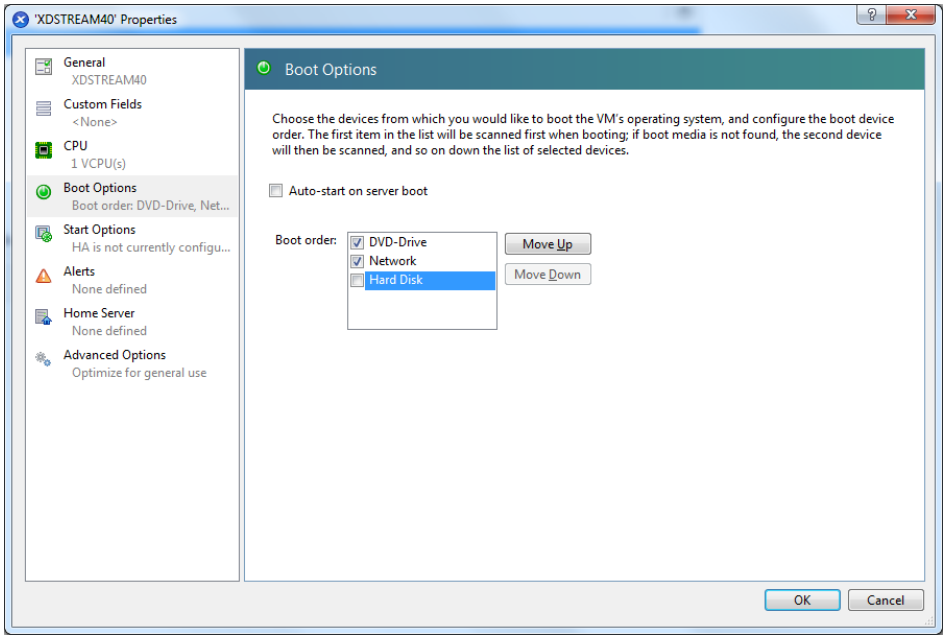
Obr. 13: Doručení Bootstrap programu z připojeného diskového obrazu ISO



Obr. 14: Nastavení PVS serverů v prostředí BDM

2.6.1 Zhodnocení

Testovací scénář probíhal použitím alternativních metod doručení prostřednictvím lokálního Bootstrap programu pomocí diskového obrazu a síťovou metodou pomocí Proxy DHCP služby. Obě metody poskytly zcela uspokojivé výsledky a lze je jednotlivě využít pro zajištění služeb vysoké dostupnosti. Jako ideální varianta byla zvolena kombinace obou služeb s příslušným nastavením pořadí bootování (1. ISO obraz, 2.síť), kde v případě nedostupnosti diskového obrazu obsahujícího Bootstrap program klient (virtuální stroj) automaticky přejde k režimu stažení Bootstrap programu s využitím potřebných informací od Proxy DHCP služby. Diskový obraz je vhodné umístit do sdílené složky poskytované souborovým serverem s vysokou dostupností, o kterém pojednává následující kapitola.



Obr. 15: Pořadí startu (Boot) virtuálního stroje (klienta)

3. ZAJIŠTĚNÍ VYSOKÉ DOSTUPNOSTI SLUŽBY SOUBOROVÉHO SERVERU FAILOVER CLUSTER MECHANISMEM OS WINDOWS SERVER 2008

Nezbytnou součástí virtuální desktopové infrastruktury je implementace vhodné strategie pro řešení síťových uživatelských profilů. Cílem je zajistit, aby uživatel pracující v relaci s libovolným virtuálním strojem pracoval vždy s konzistentní strukturou svého profilu obsahujícím jeho plochu, dokumenty, oblíbené položky, obecná nastavení atd. Společnost Denso používá pro řešení síťových uživatelských profilů a jejich správu komponentu Citrix Profile Manager (součást produktu Citrix XenDesktop) a mechanismus operačního systému Microsoft Windows pro přesměrování adresářů (Folder Redirection) pomocí doménových Group Policy Object (GPO) politik.

Citrix Profile Manager je aplikační služba běžící na úrovni operačního systému Windows 7 uvnitř virtuálních strojů vypublikovaných prostřednictvím broker mechanismu Citrix XenDesktop nebo na terminálových serverech Citrix XenApp s operačním systémem Windows Server 2008. Je obdobou tradičního řešení síťových profilů pomocí tzv. cestovních (Roaming) profilů operačního systému Windows. Nabízí však rozšířené funkce umožňující např. streaming síťového profilu, kde profil není vždy kompletně kopírován do uživatelské relace, funkci „Active Writeback“ pro periodický zápis změněných souborů profilu již během relace (nikoliv tedy až při odhlášení uživatele) zpět na síťové úložiště a další pokročilé funkce pro rozšířenou synchronizaci vybraných částí (změny registru, adresáře, soubory) uživatelského profilu.

Uživatelské síťové profily jsou uloženy na diskovém svazku obsahujícím jednotlivě sdílené složky pomocí příslušných omezení přístupových práv a diskových kvót souborového systému NTFS. Data svazku jsou fyzicky uložena v diskové storage v infrastruktuře SAN a připojeny prostřednictvím storage vrstvy hypervisoru XenServer jako lokální disková jednotka k virtuálnímu serveru s operačním systémem Windows Server 2008 v roli souborového serveru (File Server).

Tento server zde tedy představuje SPOF, při jeho výpadku je svazek s uživatelskými profily nedostupný a uživatelé se do operačního systému přihlásí tzv. dočasným profilem, který neobsahuje jejich data a osobní nastavení a neukládá provedené změny v profilu během

relace. V případě výpadku jsou nutné manuální akce administrátora pro jeho opětovné zprovoznění, které mohou být například v případě kompletní obnovy serveru ze zálohy časově náročné. Cílem následující kapitoly je realizace serverového clusteru na platformě operačního systému Windows Server 2008 poskytujícího vysokou dostupnost služby souborového serveru pro uložení uživatelských profilů.

3.1 Klíčové pojmy a terminologie

Pojem serverový cluster

Jako cluster označujeme skupinu vzájemně propojených nezávislých serverů, v clusterové terminologii označovaných jako uzly, které společně fungují jako jeden systém s cílem zajistit klientům vysokou dostupnost zvláště důležitých aplikací a prostředků. Clusterová řešení jsou obvykle nasazovány pro zvýšení dostupnosti systémů, rozložení zátěže a dále umožňují i lepší škálovatelnost.

Cluster s vysokou dostupností s podporou převzetí služeb při selhání

Cluster s vysokou dostupností (anglicky High-availability cluster nebo také failover cluster) zajišťuje pomocí několika serverů nepřetržité poskytování nějaké služby i při výpadku serveru z důvodu hardwarové poruchy nebo plánované údržby serveru. Službu poskytuje jeden server, který je v případě výpadku automaticky zastoupen jiným serverem clusteru. Tento typ clusteru je také nazýván jako „Active/Passive“.

Cluster s rozložením zátěže

Cluster s rozložením zátěže (anglicky load balancing) snižuje možnou míru zátěže tím, že službu poskytuje několik serverů, které mají stejný poskytovaný obsah (služba je tedy poskytována paralelně, jiný název je také cluster typu „Active/Active“). Stejný obsah je možné zajistit replikací poskytovaného obsahu mezi všechny propojené uzly clusteru nebo existencí speciálního centrálního úložiště. Takto spolupracující servery jsou také někdy označovány jako „serverové farmy“.

Podle společně současně sdílených nebo naopak nesdílených prostředků dále dělíme clusterová řešení na dvě kategorie:

Clustery typu Shared Everything

U tohoto typu může současně libovolný uzel clusteru přistupovat k prostředku (např. diskový svazek), který je připojen k jinému uzlu. Pokud dva či více uzlů musí současně pracovat se stejnými daty, musí být clusterové řešení opatřeno funkcionalitou pro zajištění jejich konzistence (podpora souborového systému, funkce zamykání dat a současně i tzv. koherence pamětí typu cache.)

Clustery typu Shared Nothing

Tento typ současně nesdíleného přístupu k prostředkům používají clusterové služby Microsoft Serveru 2008. Na rozdíl od předchozího typu clusteru, zde jednotlivé uzly nemohou přistupovat k prostředku, který již vlastní jiný uzel v clusteru. V případě výpadku uzlu, který je aktuálním vlastníkem prostředku, ovšem mohou další uzly v clusteru převzít vlastnictví prostředku. Například v případě diskového svazku se souborovým systémem typu NTFS nelze tento současně připojit k jednotlivým uzlům clusteru, vlastnictví tohoto svazku může být ovšem převedeno na zbývající uzly clusteru.

Operační systém Microsoft Windows Server 2008 podporuje v edicích Enterprise a Datacenter clusterovou službu pro vysokou dostupnost s podporou převzetí služeb při selhání (volitelná služba serveru nazvaná Failover Clustering) a ve všech jeho edicích clusterovou službu pro vyrovnávání zátěže sítě (volitelná funkce serveru nazvaná Network Load Balancing).

3.1.1 Služba Failover Clustering Microsoft Serveru 2008

Failover Clustering operačního systému MS Serveru 2008 v edicích Enterprise a Datacenter poskytuje řešení pro zajištění vysoké dostupnosti (HA) řady serverových služeb a aplikací. Implementace služby Failover Clustering zajistí v případě plánované údržby nebo neočekávaného selhání serveru automatické převzetí serverové služby nebo aplikace jiným serverem. Failover Clustering umožní zajistit vysokou dostupnost pro celou řadu kritických serverových služeb, jako jsou např. souborový server, DHCP server, služba zařazování tisku a další.

Windows Server 2008 podporuje následující základní typy Failover clusterů: clustery s jedním uzlem (Single-node Cluster), více uzlové clustery s jedním zařízením Quora (Single Quorum Device Multinode Cluster), více uzlové clustery s majoritní sadou uzlů (Majority Node Cluster) a hybridní více uzlové clustery.

Clustery s jedním uzlem

Clustery s jedním uzlem neumožňují převzetí služby jiným uzlem clusteru, ale dají se využít pro zjednodušenou správu a přístup ke sdíleným prostředkům a síťovým úložištím. Výhodou je zde monitorování spuštěné služby nebo aplikace clusterovou službou, která zajistí jejich restart v případě selhání. Clustery s jedním uzlem se také využívají í pro vývoj a testování clusterových aplikací.

Více uzlové clustery

Skutečné výhody Failover Clusteringu přináší řešení s využitím více uzlových clusterů. Každý uzel clusteru je zde připojen ke společně sdílenému úložišti, kde je uložena konfigurace clusteru (Disk Witness).

Jiným typem více uzlového clusteru jsou více uzlové clustery s majoritní sadou uzlů, kde je konfigurace clusteru uložena (replikována) lokálně.

Client Access Point (klientský přístupový bod)

Client Access Point je kombinací jména hostitele (Hostname) registrovanému ve službě DNS a příslušné IP adresy, které klienti clusteru použijí k připojení k dané službě, aplikaci nebo administrátoři k jeho správě.

3.1.2 Pojem Quorum

V terminologii Failover clusteru pojem Quorum definuje shodu, že dostatečný počet uzlů je schopný poskytovat clusterové služby. U MS Serveru 2008 je tato shoda založena na hlasování (Voting), kde v závislosti na příslušné konfiguraci quora může mít hlas (Vote) každý uzel clusteru s vlastní kopií konfigurace clusteru, File-Share Witness (sdílená složka s kopií konfigurace clusteru) a Disk Witness (disk s kopií konfigurace clusteru na sdíleném úložišti). Konfigurace quora určuje, kolik hlasů je potřeba k udržení konzistentní funkce clusteru. Hlasování je založeno na většinovém systému. Pokud počet hlasů klesne pod potřebnou většinu, cluster přestává pracovat.

Správná funkce Failover clusteru nezávisí pouze na dosažení quora, ale také na schopnosti (kapacitě) jednotlivých uzlů podporovat služby a aplikace, které v případě selhání převezmou. Cluster s pěti uzly by například po selhání dvou jeho uzlů mohl stále dosahovat quora, ale úroveň služeb poskytovaných jednotlivými zbývajícími uzly clusteru by závisela na kapacitě příslušného uzlu, které tento uzel převzal.

Proces dosažení Quora (shody)

1) Uzel, který je členem Failover clusteru, musí být schopný komunikovat s ostatními uzly clusteru.
2) Uzly clusteru souhlasí na konfiguraci quora a kontrolují, zda li je nutný počet hlasů pro správnou funkci clusteru.
3) Pokud není dostatečný počet hlasů pro získání quora, uzly přejdou do stavu nečinnosti až do doby, kdy je získán potřebný počet hlasů.
4) V případě získání dostatečného počtu hlasů, komponenta clusterové služby Resource Control Manager (RCM), která kontroluje konfiguraci a stav jednotlivých služeb a

zdrojů včetně vzájemných závislostí, zahájí činnost služeb a přechází do stavu monitoringu clusteru.

5) Jakmile je dosaženo quora, služby clusteru jsou plně funkční.

3.1.3 Konfigurace Quora

K dispozici jsou čtyři nastavení (modely) quora.

Node Majority (většina uzlů)

Nastavení doporučené pro clustery s lichým počtem uzlů. V této konfiguraci hlasují pouze jednotlivé uzly a quorum je dosaženo, když je online více než polovina uzlů. Cluster s celkem sedmi uzly tedy může například fungovat při selhání tří jeho uzlů.

Node and Disk Majority (většina uzlů a disků)

Nastavení doporučené pro clustery se sudým počtem uzlů. V této konfiguraci hlasují jednotlivé uzly a Disk Witness (disk s kopií clusteru). Quorum je dosaženo při nadpoloviční většině hlasů. Pokud Disk Witness zůstane v online režimu, může cluster vydržet selhání poloviny uzlů (zaokrouhleno nahoru). Cluster se šesti uzly, ve kterém je disk s kopií clusteru v online režimu, by například mohl vydržet selhání tří uzlů. Jestliže disk s kopií clusteru přejde do offline režimu nebo selže, může cluster vydržet selhání poloviny uzlů (zaokrouhleno nahoru) bez jednoho. Cluster se šesti uzly, jehož disk s kopií clusteru selhal, by například mohl vydržet selhání dvou uzlů.

Node and File Share Majority (většina uzlů a sdílená složka)

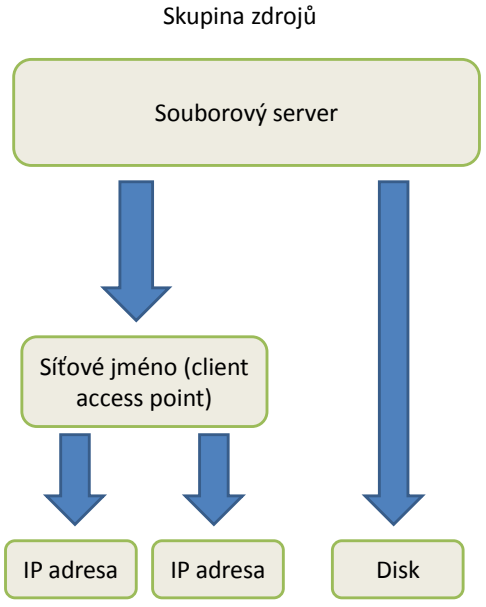
Toto nastavení je podobné jako Node and Disk Majority. V této konfiguraci hlasují jednotlivé uzly a File Share Witness (sdílená složka s kopií clusteru). Quorum je dosaženo při nadpoloviční většině hlasů.

No Majority : Disk Only (bez většiny: pouze disk)

Nedoporučované nastavení. V clusteru s touto konfigurací není způsob dosažení quora ovlivněn počtem uzlů, quorum zde představuje pouze sdílený disk s konfigurací clusteru. Pokud však dojde ke ztrátě komunikace s tímto diskem nebo k jeho poškození, nebude cluster k dispozici.

3.1.4 Komponenty Resource Hosting Subsystem (RHS) a Resource Control Manager (RCM)

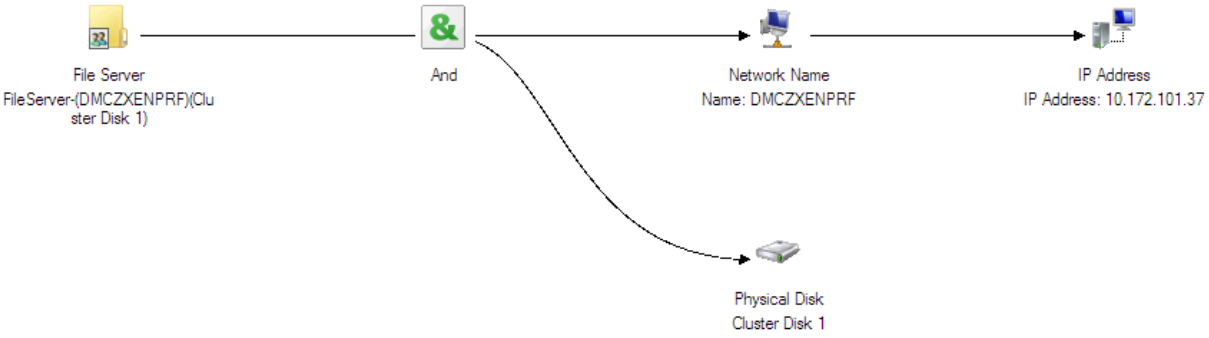
Operační systém MS Windows Server 2008 je schopný poskytovat řadu svých serverových služeb s vysokou dostupností pomocí různých zdrojů, z nichž některé jsou přímo součástí clusterové služby a některé serverové aplikace z konceptu jejich návrhu, jako jsou např. MS SQL server a mailserver Exchange, přímo podporují clusterová řešení. Jednotlivé zdroje (resources) jsou navrženy pro pro vzájemnou spolupráci a jsou typicky uspořádána do skupin zdrojů (resource groups). Např. skupina zdrojů pro vysoce dostupný souborový server obsahuje jeden či více následujících typů jednotlivých zdrojů: Client Access Point, diskové úložiště, souborový server. Vysoce dostupný SQL server se skládá z následujících zdrojů: Client Access Point, diskové úložiště, SQL server a SQL server agent. Clusterové zdroje obsahují podporu clusteringu použitím speciálních pluginů a knihoven (DLL), jejichž programový kód umožňuje správnou integraci a komunikaci pro účely clusterové služby.



Obr. 16: Skupina zdrojů pro souborový server v režimu HA

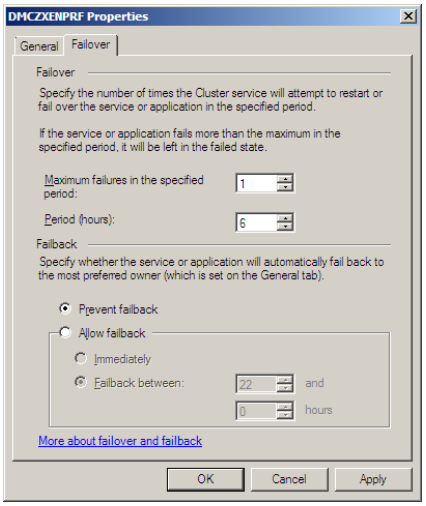
Správu těchto zdrojů obsluhuje komponenta Resource Control Manager (RCM) a Resource Hosting Subsystem (RHS), která poskytuje tuto funkcionalitu jako součást samotné clusterové služby.

Resource Control Manager je součástí celkové clusterové architektury. Implementuje mechanismy pro převzetí služeb (failover) pomocí nastavených pravidel a pro každý zdroj zakládá a udržuje tzv. strom závislostí (dependency tree). Např. zdroj souborový server má závislost na zdroji Client Access Point a zdroji typu disk.



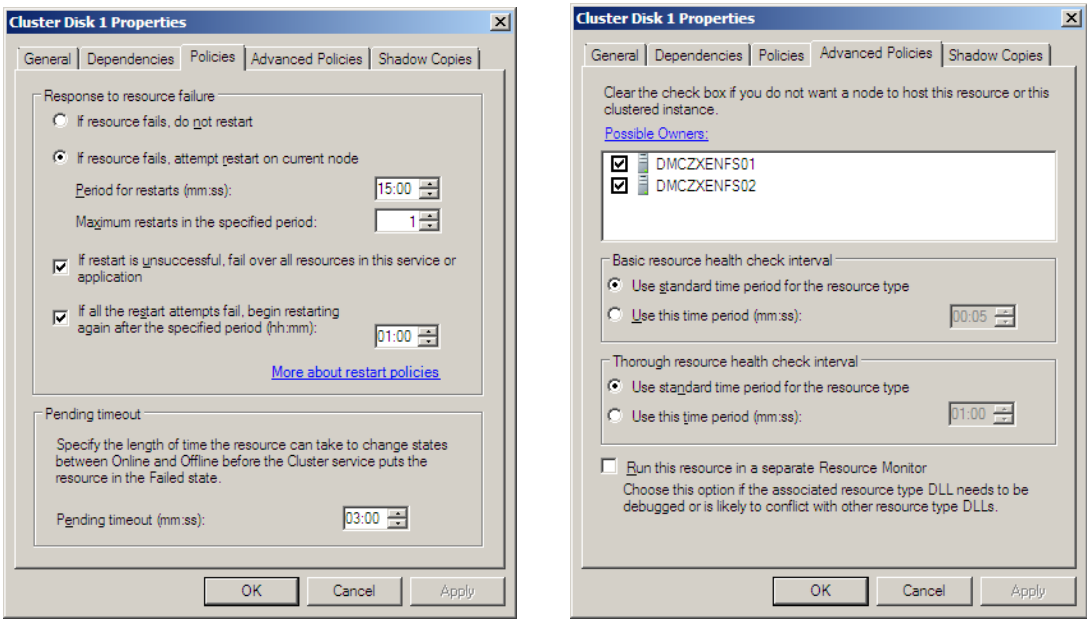
Obr. 17: Strom závislosti zdrojů

Resource Control Manager spravuje stavy jednotlivých zdrojů (Online, Offline, Failed, Online Pending, a Offline Pending) a skupin zdrojů (Online, Offline, Partial Online a Failed). Resource Control Manager může spustit následující akce nad skupinou zdrojů: Move, Failover a Failback. Provedení některé z těchto akcí závisí na různých faktorech jako je aktuální „zdraví“ (health) zdrojů ve skupině, administrátorská akce prováděná na skupině zdrojů (např. Move Group), nebo na nastavených politikách aplikovaných na skupinu zdrojů. Následující obrázky zobrazují nastavení akci typu Failover a skupinové politiky pro Failback.



Obr. 18: Nastavení Failover politiky pro zdroj typu Disk

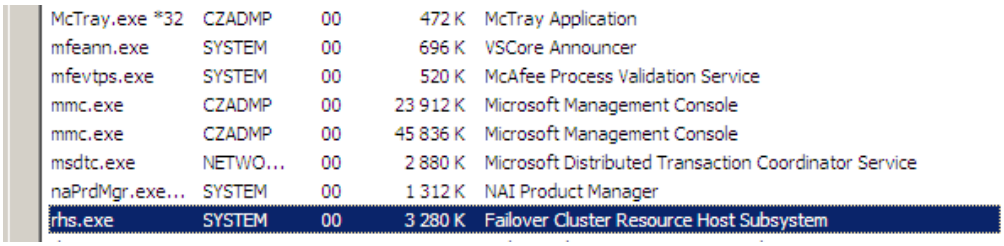
Politiky se aplikují též na jednotlivé zdroje, např. pro zdroj typu disk:



Obr. 19: Nastavení Failover politiky pro zdroj typu Disk

Resource Hosting Subsystem (RHS)

Resource Hosting Subsystem sdružuje jednotlivé zdroje clusteru v režimu online pod jedním procesem (rhs.exe).



Obr. 20: RHS ve výpisu procesů

Proces RHS provádí periodickou kontrolu „zdraví“ (Health Checking) těchto zdrojů pomocí diagnostických nízko úrovňových procesů LooksAlive a IsAlive, které jsou specifické pro jednotlivé typy clusterových zdrojů. Perioda kontroly je předurčena typem a příslušnou DLL knihovnou clusterového zdroje, nebo pomocí politik nastavených administrátorem. Pokud zdroj selže při kontrole procesem LooksAlive, je vykonána hloubková kontrola IsAlive.

V případě zjištěného selhání zdroje touto kontrolou jsou vykonávány další akce až do určité doby, kdy je vyhodnoceno, že zdroj, který selhal, nemůže být přiřazen ke konkrétnímu uzlu clusteru. V tomto případě proces RHS informuje komponentu Resource Control Manager, která zdroj v clusterové službě označí jako Failed a vykoná akci Failover k přesunutí zdroje (skupiny) na jiný uzel clusteru.

3.2 Přehled požadavků pro instalaci failover clusteru

Požadavky na hardware:

- serverové uzly clusteru mají vzájemně odpovídající konfiguraci a používají stejné nebo podobné komponenty
- všechny diagnostické testy Validation Tool musí být úspěšně splněny

Požadavky na software:

- na všech uzlech v clusteru musí běžet verze systému Windows Server 2008 R2 pro procesory x64, nebo verze založená na architektuře s procesorem Itanium (v uzlech v jednom clusteru běžet různé verze).
- na všech uzlech musí být nainstalovány stejné aktualizace softwaru (opravy) a aktualizace Service Pack.
- edice OS verze Enterprise nebo Datacenter

Požadavky na síťovou infrastrukturu :

- síťové adaptéry uzlů musí být stejné a používat shodné nastavení protokolu IP (statické/ DHCP přiřazení) a komunikaci (například rychlost, duplexní režim, řízení toku)
- je doporučována redundance pro eliminaci SPOF (selhání v jednom bodě)
- uzly clusteru musí k překladu názvů používat službu DNS (Domain Name System)
- uzly clusteru jsou členy stejné domény

- při prvním vytvoření clusteru nebo přidání uzlů do clusteru je nutné přihlášení pomocí účtu s právy a oprávněními správce pro všechny uzly v daném clusteru. Pokud se nejedná o účet ve skupině Domain Admins, je navíc nutné, aby bylo danému účtu přiděleno oprávnění „Vytvářet objekty počítačů“ v doméně.

Požadavky na disková úložiště

- technologie SAS (Serial Attached SCSI) nebo Fibre Channel: řadiče diskových úložišť, které jsou vyhrazeny pro úložiště clusteru, by měly být shodné. Měla by v nich být také použita stejná verze firmwaru.
- izolace zařízení úložiště - jeden cluster pro každé zařízení: servery z různých clusterů nesmějí být schopny získat přístup ke stejným zařízením úložiště.
- logická jednotka (LUN) použitá pro jednu skupinu serverů clusteru izolována od všech ostatních serverů prostřednictvím maskování logické jednotky (LUN masking) nebo jejího rozdělení na zóny (Zoning).
- disk s kopií konfigurace clusteru (Disk Witness) musí mít příslušný oddíl souborového systému typu NTFS

3.2.1 Cluster Validation Tool

Cluster Validation Tool je ověřovací nástroj, který pomocí testování určuje, zdali je konfigurace operačního systému, úložiště dat a sítě vhodná pro failover cluster. Toto ověření je nezbytně nutné pro vytvoření failover clusteru. V průběhu testu se kontrolují:

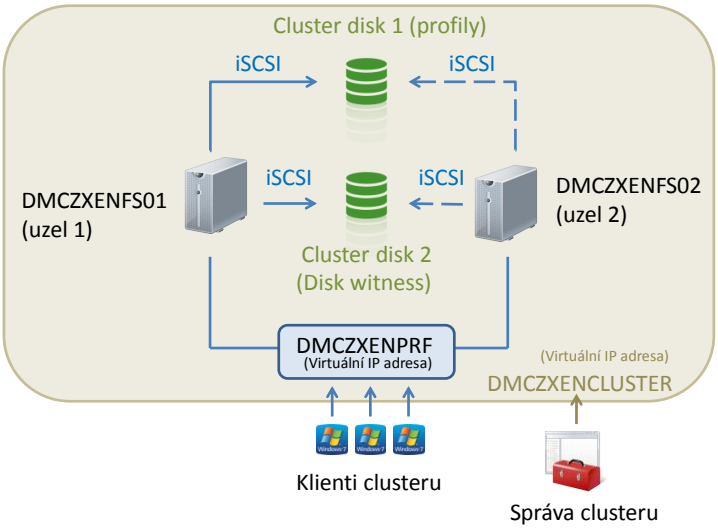
- Binární konzistence operačních systémů (kontrola, zdali uzly clusteru používají stejnou verzi operačního systému, opravné a servisní balíčky typu hotfix a service pack.
- Architektura procesorů a údaje o operační paměti
- Členství jednotlivých uzlu clusteru ve stejné doméně
- Digitálně nepodepsané ovladače
- Zařízení typu Plug and Play, síťové a HBA adaptéry
- Infrastruktura (vzájemná komunikace uzlů a kompatibilita SCSCI s Persistent Reservation)

- Možnost přístupu jednotlivých uzlů na sdílené clusterové diskové úložiště
- Přítomnost více síťových adaptérů s IP adresami jiné podsítě na jednotlivých uzlech clusteru
- Latence diskového úložiště a simulovaný test failoveru s převzetím služeb mezi uzly clusteru

Výsledkem ověření je formátovaný soubor typu HTML, který obsahuje podrobnou zprávu o validitě clusterového prostředí a prostředků. Tento ověřovací nástroj je možné spustit kdykoliv a může tedy opakovaně posloužit pro servisní účely jako komplexní diagnostika failover clusteru.

3.3 Realizace souborového serveru s vysokou dostupností

Realizovaný Failover Cluster hostující službu souborového server s vysokou dostupností je sestaven z dvou virtuálních parametrově identických serverů (uzlů) s operačním systémem MS Windows Server 2008 R2 64bit v edici DataCenter. Jako model clusterového Quora bylo zvoleno doporučené nastavení „Node and Disk Majority“ pro sudý počet uzlů, kde mají hlas jednotlivé uzly clusteru (hostitelská jména DMCZXENFS01 a DMCZXENFS02) a diskový svazek typu „Disk witness“ obsahující kopii konfigurace clusteru.



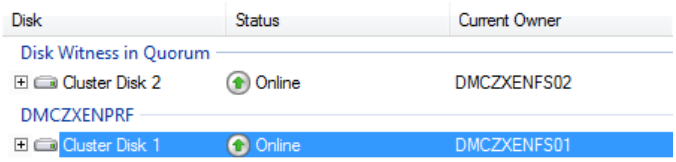
Obr. 21:Schéma realizovaného Failover clusteru

Virtualizační hypervisor XenServer neumožňuje prostřednictvím své storage vrstvy sdílet nebo dynamicky připojovat diskové svazky mezi jednotlivými virtuálními stroji, proto bylo nutné připojit virtuální diskové svazky z datové storage k serverovým uzlům prostřednictvím protokolu iSCSI.

Internet Small Computer System Interface (zkratka iSCSI) je síťový protokol, který umožňuje připojovat úložný prostor pomocí počítačové sítě. Koncepce iSCSI vychází ze dvou technologií, SCSI rozhraní pro připojování disků v serverech a protokolu TCP/IP. Z rozhraní SCSI se používá pouze protokol, kterým spolu zařízení komunikují, pro přenos paketů SCSI se použije jejich zapouzdření do protokolu TCP/IP. Disky připojené protokolem iSCSI se chovají úplně stejně, jako disky připojené na lokální diskový řadič, k jejich inicializaci a správě lze tedy využít nástroj „Disk Management“ (Správa disků) v operačním systému Windows.

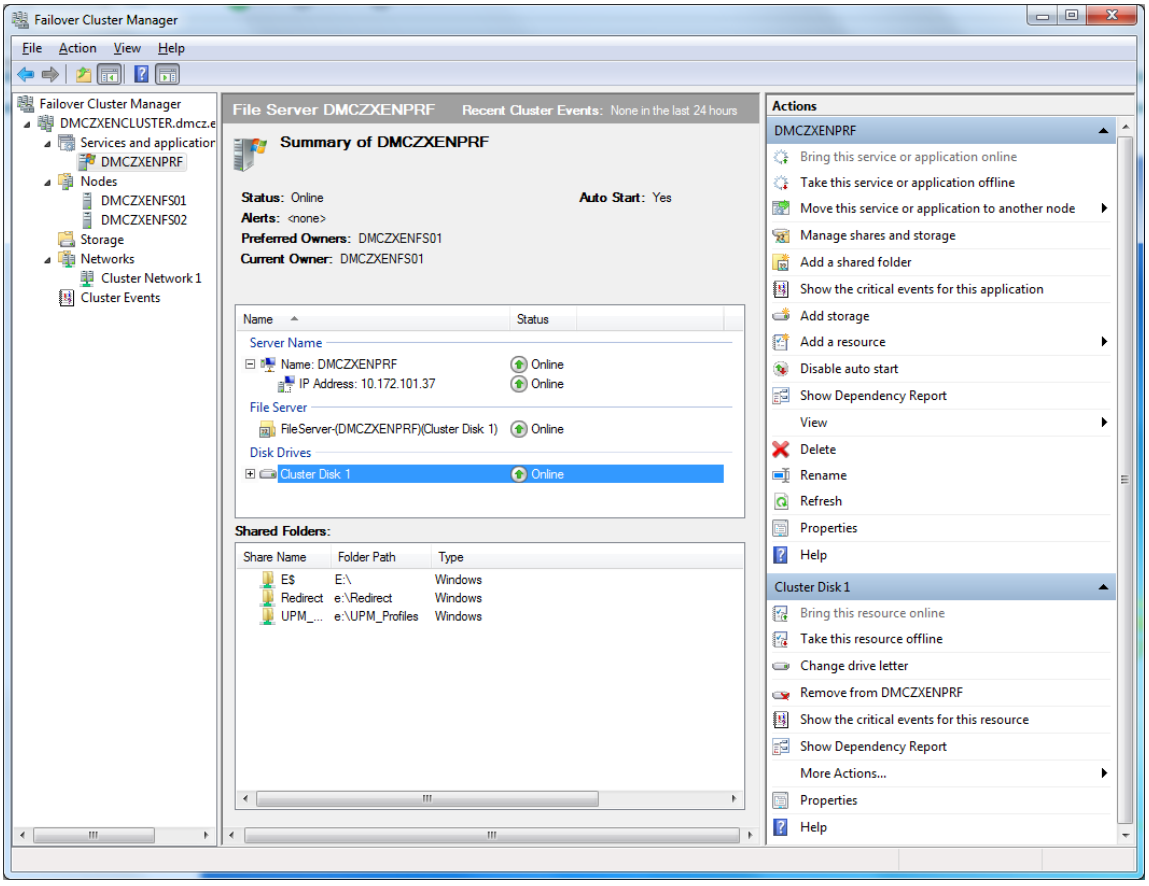
V terminologii iSCSI je klientské zařízení (server, desktop) označováno jako tzv. „iSCSI Initiator“ a jednotlivé (diskové) logické jednotky (LUN) definované na datové storage jako tzv. „iSCSI target.“ Softwarový iSCSI Initiator podporují všechny běžně používané serverové i klientské operační systémy. Značnou výhodou softwarového iSCSI Initiatoru je, že nevyžaduje žádný speciální hardware a je zdarma.

Oba uzly (DMCZXENFS01 a DMCZXENFS02) clusteru mají na úrovni služeb operačního systému nakonfigurováno připojení jako „iSCSI Initiators“ k datové storage v infrastruktuře SAN, tato storage zde tedy reprezentuje „iSCSI target“. Na úrovni datové storage byly vzhledem ke zvolenému modelu quora vytvořeny dva separátní LUNy, které z pohledu serverů (uzlů) představují lokální diskové svazky. Oba diskové svazky jsou naformátované souborovým systémem NTFS a představují clusterové zdroje (resources) typu disk. První svazek „Cluster Disk 1“ obsahuje data uživatelských profilů a druhý svazek „Cluster Disk 2“ slouží jako clusterový Disk Witness. Data svazků jsou chráněna proti fyzické chybě pomocí synchronního zrcadlení na jinou datovou storage a proti logické chybě pomocí tzv. snapshotů (snímků dat v čase).



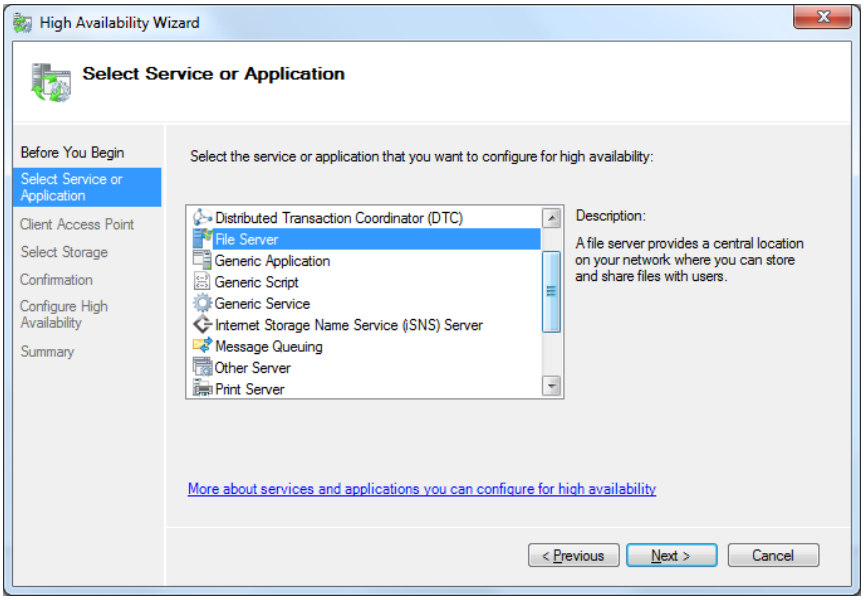
Obr. 22: Clusterové disky

Při instalaci byl nejprve vytvořen tzv. Přístupový bod pro správu clusteru (Access Point for Administering the Cluster) pojmenovaný DCMZXENCLUSTER s přiřazenou IP adresou. Tento název je virtuálním síťovým názvem (hostname) s doménovým členstvím registrovaným ve službě DNS a používá se k připojení ke clusteru a jeho následné správě. Tyto údaje se liší od názvu a IP adresy, kterou použijí klienti k připojení ke clusterové službě souborového serveru.



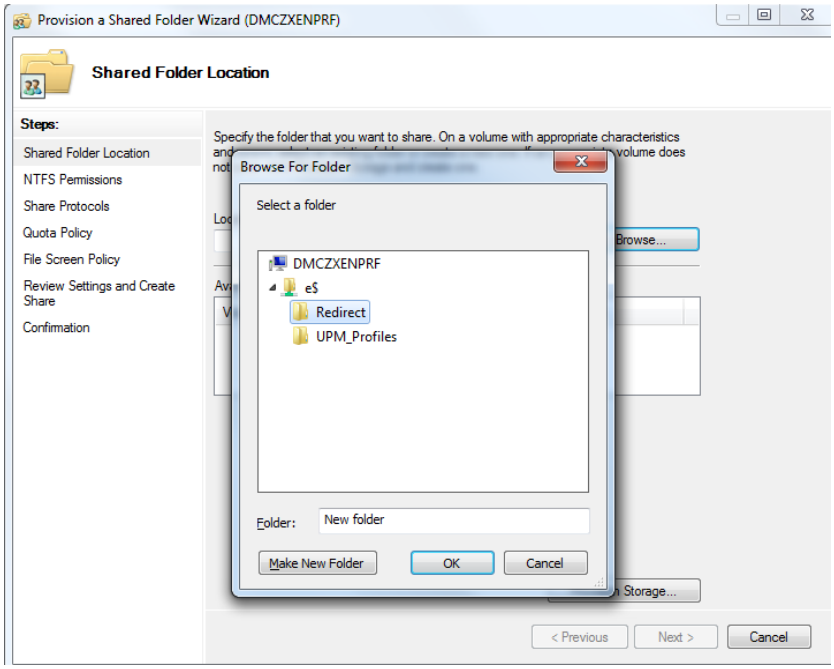
Obr. 23: Administrátorská konzole clusteru (připojení na Access Point DMCZXENCLUSTER)

Samotná clusterová služba souborového serveru se konfiguruje prostřednictvím průvodce vyvolaného z administrátorské konzole Failover Cluster Manager. Součástí konfigurace je vytvoření dalšího přístupového bodu (zde pojmenovaného DMCZXENPRF) s přiřazenou IP adresou, který opět reprezentuje virtuální síťový název s doménovým členstvím a je registrovaný ve službě DNS. Údaje o této IP adrese a síťovém jménu budou používat klienti pro připojení k souborovému serveru.



Obr. 24: Dialog pro přidání vybrané služby nebo aplikace v režimu HA

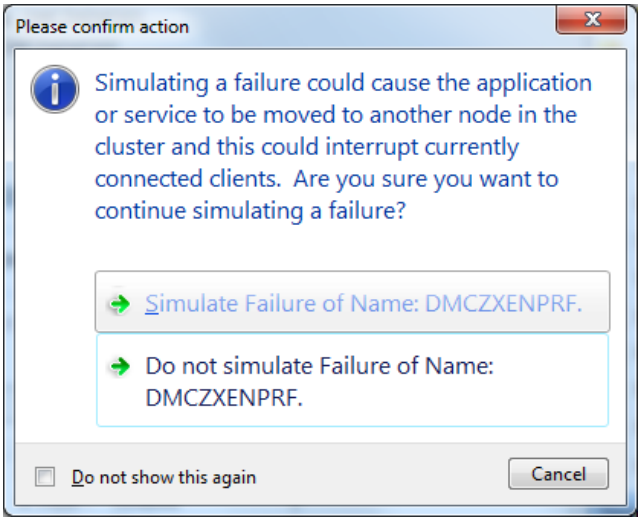
Data vybraných uživatelských profilů byla pro účely následného testu kompletně zkopírována včetně přístupových práv NTFS z původního umístění pomocí utility Robocopy do příslušných složek na datovém disku clusteru. Práva pro sdílení těchto složek, souborová práva NTFS a diskové kvóty jsou konfigurovatelné pomocí průvodce pro přidání sdílené složky (Shared Folder Location) z administrátorské konzole clusteru.



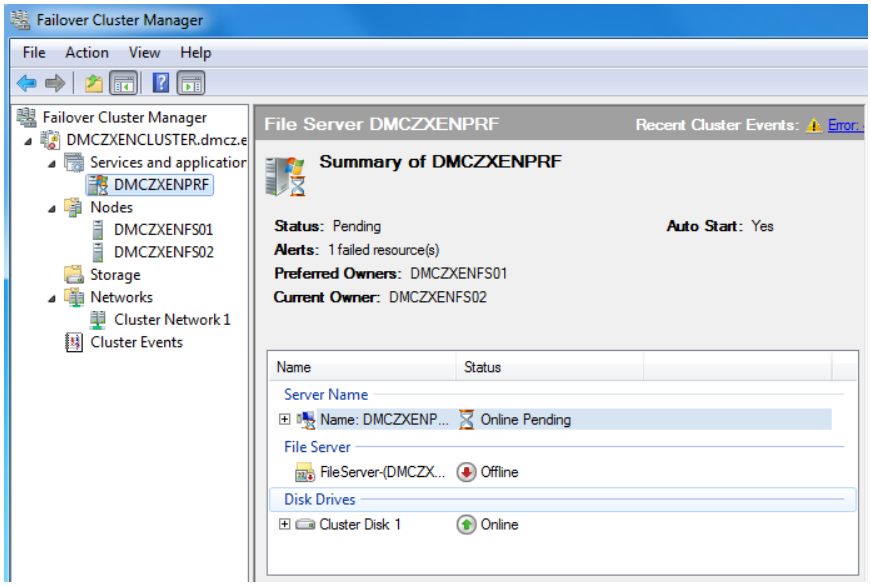
Obr. 25: Průvodce nastavením sdílených složek

Po nastavení sdílení a práv byly upraveny i údaje o sdílených složkách v konfiguračních parametrech služby Citrix Profile Manager a GPO politiky tak, aby reflektovaly provedenou změnu nového umístění síťových uživatelských profilů na clusterový datový disk (syntaxe pro přístup: \\DMCZXENPRF\název sdílené složky).

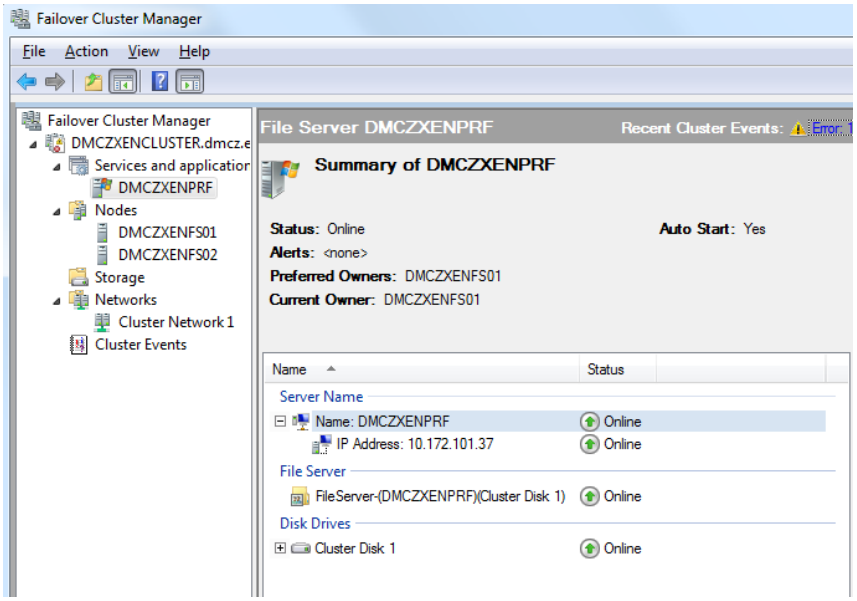
Test Failover mechanismu clusteru byl několikrát realizován prostřednictvím akce z administrátorské konzole simulující výpadek jednoho uzlu clusteru a převzetí služby souborového serveru DMCZXENPRF druhým uzlem. Doba nutná k převzetí služby a obnovení činnosti souborového serveru je přibližně čtyři sekundy. Následující obrázky postupně znázorňují proces simulovaného výpadku uzlu DMCZXENFS02 a převzetí služby souborového serveru (změna „Current Owner“) uzlem DMCZXENFS01 v prostředí administrátorské konzole clusteru.



Obr. 26: Akce pro simulaci výpadku



Obr. 27: Převzetí služby při simulovaném výpadku



Obr. 28: Dokončený Failover proces (služba je opět funkční)

Realizovaný cluster byl také podroben testováním pomocí utility Microsoft File Server Capacity Tool, která pomocí simulovaného zátěžového testu slouží pro odhad obslužné kapacity clusteru.

3.3.1 File Server Capacity Tool

Nástroj File Server Capacity Tool (FSCT) od společnosti Microsoft slouží pro odhad maximálního počtu uživatelů, které je souborový server v dané konfiguraci schopný obsloužit podle předem definovaných specifických scénářů operací simulujících reálné zatížení serveru. Typicky se používá scénář nazvaný „Home Folders“ (domácí adresáře), kde se simulují standarní uživatelské operace (vytvoření, čtení, zápis, mazání) se soubory z balíku Microsoft Office, operace se soubory prostřednictvím Windows Explorer (průzkumník) a příkazové řádky. Příprava testu a jeho realizace vyžaduje součinnost následujících komponent:

- doménového řadiče pro autentifikaci v doménovém prostředí a vytvoření testovacích uživatelských účtu
- testovaného serveru (serverového clusteru)

- tzv. Controlleru, který synchronizuje jednotlivé testovací aktivity a sbírá data pro závěrečné vyhodnocení testu
- klientských počítačů, které slouží pro generování zátěže (jeden klient může simulovat více uživatelů zároveň)

Souborový server byl ověřen scénářem „Home Folders“ se simulovaným limitem 300 (plánovaný počet je 150) uživatelů s následujícím uspokojivým výsledkem:

```
*** Results
Users  Overload  Throughput  Errors  Errors [%]  Duration [ms]
100          0%           7       0          0%         600027
200          0%          11       0          0%         600015
300          0%          16       0          0%         600049

*** Test's information
FSCT version: 1.2
Workload: HomeFolders
Time: 2012/09/04 10:37
```

Legenda:

Users (uživatelé) – počet simulovaných uživatelů, kteří přistupují k datům na souborovém serveru

Overload - nenulová hodnota indikující, že server není schopný obsloužit příslušný počet simulovaných uživatelů během jednoho testovacího časového intervalu

Throughput - počet operací se soubory v relaci klient/ souborový server za sekundu

Errors – nenulová hodnota indikuje, že klient obdržel chybová hlášení ze souborového serveru

Error Percentage – procentuálně vyjádřený poměr vykonaných testovacích operací nutných k dosažení chyby proti celkovému počtu vykonaných operací během testovacího časového intervalu

Duration - doba testovacího intervalu vyjádřená v [ms]

4. ZAJIŠTĚNÍ VYSOKÉ DOSTUPNOSTI SLUŽBY CITRIX WEB INTERFACE

Komponenta virtuální infrastruktury Citrix Web Interface slouží pro přihlašování uživatelů prostřednictvím webového rozhraní k virtuálním desktopům vypublikovaných prostřednictvím broker mechanismu XenDesktop a dále k terminálovému serveru XenApp. Služba Citrix Web Interface je spuštěna na virtuálních serverech se síťovými názvy XENBROKER a XENBROKER2, každý server má tedy spuštěnu vlastní službu IIS a používá vlastní kopii webového rozhraní.

Ve výchozím stavu se uživatelé přihlašovali prostřednictvím webového prohlížeče odkazem na http://xenapp , který měl nastavený záznam typu Alias (CNAME) v systému DNS na virtuální server se síťovým názvem XENBROKER.Tento server zde představoval SPOF, v případě jeho nedostupnosti byly nutné manuální akce administrátora pro obnovení jeho služeb nebo přesměrování záznamu v systému DNS na virtuální server XENBROKER2.

xdworkplan	Host (A)	10.172.100.197	4/24/2012 7:00:00 AM
xdxp	Host (A)	10.172.101.78	4/22/2012 8:00:00 PM
XenApp	Alias (CNAME)	xenbroker.dmcz.emea.denso.	static
XENBROKER	Host (A)	10.172.103.3	4/24/2012 2:00:00 PM
XENBROKER2	Host (A)	10.172.103.4	4/23/2012 6:00:00 PM
XPETERM01	Host (A)	172.20.95.212	4/24/2012 6:00:00 AM

Obr. 29: Alias záznam XenApp

Cílem následující kapitoly bylo realizovat službu Citrix Web Interface v režimu vysoké dostupnosti, kde při výpadku virtuálního hostitelského serveru dojde k automatickému přesměrování požadavku na funkční server.

4.1 Přehled řešení pro zajištění služeb vysoké dostupnosti Citrix Web Interface

Pro zajištění služby vysoké dostupnosti v režimu vysoké dostupnosti lze využít následující řešení a metody :

DNS Round Robin

Technika DNS Round Robin je popsána v kapitole 2.5.1. Značně limitujícím faktorem je zde nízká odolnost při selhání jednotlivých serverů vyžadující ruční odebrání příslušného záznamu nedostupného serveru v systému DNS.

Hardware Load Balancer

HW zařízení pro rozložení zátěže typicky používají překlad síťových adres NAT (Network Address Translation) pro přesměrování klientských požadavků na příslušné adresy cílových serverů. HW zařízení zde představuje SPOF a pro zajištění vysoké dostupnosti serverových služeb je nutné počítat se záložním.

Dispatch Server

SW řešení omezené propustností a výkonem serveru, který zde poskytuje služby pro směrování aplikačně modifikovaných klientských požadavků na cílové servery. Server se službou typu dispatch zde představuje SPOF. V případě selhání je nutné obnovit server se zálohy nebo přesunout službu na jiný, záložní server. Ve virtuálním prostředí je proto výhodné používat replikaci příslušné serverové VM.

Microsoft Network Load Balancing (NLB)

SW řešení používající distribuční architekturu pro vyvažování zátěže a síťové služby vysoké dostupnosti. Na rozdíl od Dispatch Serveru zde nedochází k centralizovanému přesměrování

klientských požadavků z jednoho bodu. Všechny servery (uzly) s nakonfigurovanou podporou NLB služby obdrží klientské požadavky okamžitě bez nutnosti jejich modifikace a opětovného přeposlání. Redundance je zde rovna počtu jednotlivých uzlů v clusteru.

4.1.1 Přehled vlastností jednotlivých řešení a metod

	DNS Round Robin	HW Loadalancer	Dispatch Server	MS Network Loadbalancing
Nutnost HW	-	Ano	-	-
SPOF	-	Ano	Ano	-
Snadná škálovatelnost	Ano	-	Omezená	Ano
Vysoký výkon	Ano	Ano	Omezený	Ano
Odolnost proti selhání	-	Omezená	Omezená	Ano

Tab. 1: Přehled vlastností jednotlivých metod

Porovnáním těchto řešení a s ohledem na požadované maximální využití dostupných zdrojů jsem pro zajištění služby Citrix Web Interface v režimu vysoké dostupnosti zvolil realizovat serverový cluster se službou Microsoft Network Loadbalancing, jehož realizace je popsána v následující kapitole.

4.2 Microsoft Network Load Balancing (NLB)

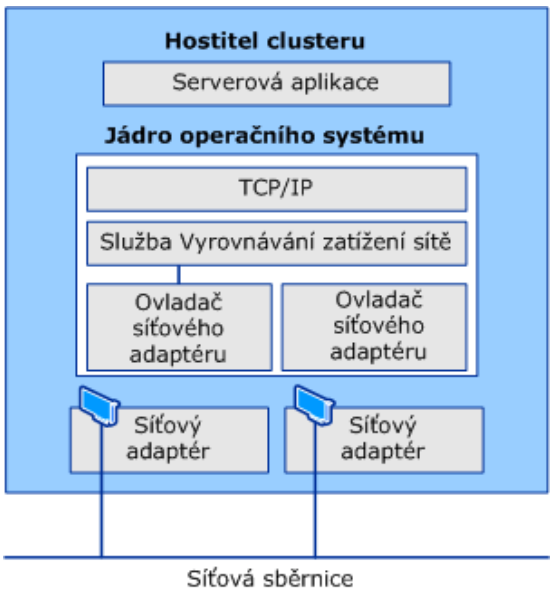
Clusterová služba Microsoft Serveru 2008 nazvaná **Network Load Balancing** (vyrovnávání zátěže sítě) představuje vysoce dostupné a škálovatelné řešení pro síťové aplikace a služby založené na komunikaci prostřednictvím sady protokolů TCP/IP jako jsou např. webové servery, FTP a proxy servery. Sloučením příslušných serverových služeb dvou

či více serverů do jediného clusteru poskytuje služba NLB redundanci dat včetně mechanismu pro vyvažování zátěže.

V NLB clusteru má každý jeho uzel spuštěnou separátní kopii požadované serverové aplikace. NLB služba podle nastavených pravidel distribuuje klientské požadavky přicházející na virtuální IP adresu clusteru mezi jednotlivé jeho uzly a v případě nedostupnosti uzlu (nedoručení monitorovacího Heartbeat signálu) automaticky iniciuje proces převzetí (Failover) příslušné síťové služby funkčními uzly NLB clusteru. Po obnovení činnosti uzlu, u kterého došlo k selhání, zpět do režimu online bude zatížení mezi ostatními uzly znovu rozloženo tak, aby byl zahrnut i obnovený uzel. Tento proces je pro klienty NLB clusteru zcela transparentní.

4.2.1 Architektura služby NLB

Služba NLB pracuje jako síťový ovladač v systému Windows. Pracuje nezávisle na síťovém zásobníku TCP/IP a její provoz vůči tomuto zásobníku je zcela transparentní. NLB cluster používá model distribuované softwarové architektury, kde na každém uzlu v clusteru současně běží identické kopie síťového ovladače NLB. Ovladač se nachází mezi zásobníkem TCP/IP a ovladači síťových adaptérů uzlu, služba NLB je v hierarchii spuštěna v úrovni nad nimi a to společně se serverovou aplikací.



Obr. 30: Služba NLB (vyrovnáváni zátěže sítě) jako síťový ovladač

Pro síťovou komunikaci používá služba NLB virtuální IP adresu (cluster IP) a této adrese podle módu nastavení NLB clusteru přiřazenou generovanou virtuální MAC adresu, které jsou nastaveny na každém uzlu NLB clusteru. Klientský dotaz doručený na tuto virtuální IP a MAC adresu je směrován na všechny uzly NLB clusteru a ty si podle určitých pravidel určí, který s uzlů bude na tento dotaz odpovídat. Rozdělení se provádí pomocí algoritmu randomizace hashovací funkce se vstupními hodnotami o klientské IP adrese, síťovém portu a počtu aktivních uzlů NLB clusteru. Jiné uzly NLB clusteru, než je ten, který posléze na klientský dotaz odpovídá, tento dotaz zahazují.

4.2.2 Proces konvergence NLB clusteru

Jednotlivé uzly v NLB clusteru si ve výchozím nastavení každou sekundu vyměňují monitorující Heartbeat signál o aktuálním stavu clusteru. Pokud dojde k události jako je přidání, odebrání, nebo selhání uzlu v NLB clusteru, dochází k procesu tzv. konvergence, kdy se NLB cluster automaticky rekonfiguruje podle typu události. Při přidání nebo odebrání uzlu dochází k přepočítání zátěže. V případě selhání uzlu, kdy se standardně jako selhání považuje nedoručení Heartbeat signálu po uplynutí doby pěti sekund, se funkční uzly snaží o udržení konzistentní funkce NLB clusteru změnou uzlu s definovanou vyšší prioritou na výchozí. Během procesu konvergence jednotlivé uzly stále obsluhují příchozí síťový provoz, služby nedostupného uzlu jsou automaticky rozloženy na zbylé funkční uzly. Proces konvergence NLB clusteru je pro síťové klienty zcela transparentní a nedochází k výpadku clusterové služby.

4.2.3 Konfigurační parametry služby Microsoft Network Load Balancing

Microsoft Network Load Balancing (NLB) cluster může pracovat v režimech Unicast (jednosměrné síťové volání), Multicast a rozšířený IGMP multicast (vícesměrové volání). Pro správnou funkci NLB clusteru musí ovšem být jednotlivé síťové adaptéry vždy nastaveny pro požadovaný režim síťového volání, kombinace Unicast/Multicast není možná. Následující kapitoly popisují jednotlivé režimy síťových volání.

Unicast

Režim Unicast je výchozím nastavením Microsoft NLB clusteru. V tomto režimu jsou MAC adresy síťových adaptéru jednotlivých uzlů nahrazeny jednou virtuální MAC adresou NLB clusteru a zároveň je přidána i druhá virtuální IP adresa NLB clusteru. V přepínaných síťových prostředích typu Ethernet, kde se předpokládá, že na jednom portu přepínače (switch) je vždy unikátní MAC adresa síťového adaptéru, by toto nefungovalo správně a na přepínači by neustále (= zvýšení zátěže přepínače) docházelo ke změnám záznamu tabulky přiřazení port na přepínači /MAC addresa adaptéru a vždy by fungoval pouze jeden (tzv. MAC adress flapping).

NLB řeší tento problém maskováním virtuální MAC adresy clusteru. NLB vytvoří falešné rozdílné MAC, které vychází z MAC adresy clusteru, ale na druhém oktetu MAC adresy je identifikační pořadové číslo uzlu clusteru. Komunikace potom probíhá tak, že v hlavičkách Ethernetového rámce se používají tyto falešné MAC adresy, ale na ARP dotazy odpovídají jednotlivé uzly virtuální MAC adresou. V ARP odpovědi je jiná MAC adresa v hlavičce a jiná MAC adresa v datech jako zdrojová. Maskování MAC adresy tedy zaručí, že přepínač nezapisuje do své tabulky přiřazení port/ MAC adresa virtuální MAC adresu NLB clusteru na různých portech, ale unikátní MAC adresu každého uzlu.

Multicast / IGMP Multicast

V režimu Multicast je každému uzlu NLB clusteru přidána speciální multicast MAC adresa s prefixem 03-bf a každý uzel si zachová i svou skutečnou MAC adresu síťového adaptéru. Multicast MAC adresa se používá pro síťovou komukaci klient-cluster a MAC adresa síťového adaptéru uzlu se používá pro běžnou síťovou komunikaci specifickou pro každý uzel v clusteru.

V režimu IGMP Multicast je každému uzlu NLB clusteru přidána speciální multicast MAC adresa s prefixem 01-00 a každý uzel si zachová i svou skutečnou MAC adresu síťového adaptéru. IGMP multicast MAC adresa se používá pro síťovou komukaci klient-cluster a MAC adresa síťového adaptéru uzlu se používá pro běžnou síťovou komunikaci specifickou pro každý uzel v clusteru.

Režimy typu Multicast přiřadí každému uzlu virtuální IP adresu a na ARP dotazy na IP adresu NLB clusteru odpovídají servery multicast MAC adresou. V závislosti na zvoleném režimu multicast přepínač odesílá ethernetový rámec na všechny porty kromě příchozího v příslušné podsíti (režim Multicast) nebo pouze na porty zařazené v IGMP multicast skupině (IGMP Multicast).

4.2.4 Možnosti implementace Microsoft Network Load Balancing

Microsoft NLB může být implementován ve čtyřech režimech, které se liší volbou volání Unicast nebo Multicast a počtem síťových adaptérů jednotlivých uzlů clusteru.

Unicast s jedním síťovým adaptérem

Tento režim clusteru je vhodný pro případy, kdy není nutná běžná síťová komunikace mezi jednotlivými uzly clusteru a vzhledem k intenzitě běžného síťového provozu na konkrétní uzel clusteru není nutné oddělit tento provoz od síťového provozu typu klient-cluster.

Multicast/ IGMP multicast s jedním síťovým adaptérem

Tento režim clusteru je vhodný pro případy, kdy je nutná běžná síťová komunikace mezi jednotlivými uzly clusteru a vzhledem k intenzitě běžného síťového provozu na konkrétní uzel clusteru není nutné oddělit tento provoz od síťového provozu typu klient-cluster.

Unicast s více síťovými adaptéry

Tento režim clusteru je vhodný pro případy, kdy je nutná běžná síťová komunikace mezi jednotlivými uzly clusteru a je požadováno oddělit síťový provoz pro správu clusteru od síťového provozu typu klient-cluster.

Multicast/ IGMP multicast s více síťovými adaptéry

Tento režim clusteru je vhodný pro případy, kdy je nutná běžná síťová komunikace mezi jednotlivými uzly clusteru a síťový provoz přicházející na konkrétní uzel clusteru z jiného než clusterového subnetu je značný.

4.2.5 Konfigurace síťových portů a client affinity

Síťový port je speciální číslo (1 až 65535), které slouží v počítačových sítích při komunikaci pomocí protokolů TCP a UDP k rozlišení aplikace v rámci počítače. Například HTTP protokol (implicitně) „naslouchá“ na portu 80, HTTPS protokol na portu 443, protokol FTP na portu 21 atd. Network load balancing může být nakonfigurován v režimu jednotlivých síťových portů nebo jejich rozsahu.

Pro každý definovaný port je možná následující konfigurace pro forwarding (přeposílaní) síťového provozu na cílový uzel v NLB clusteru:

Single host - síťový provozu na definovaném síťovém portu či rozsahu portů je směřován vždy na konkrétní uzel NLB clusteru

Multiple Hosts - síťový provoz na definovaném síťovém portu či rozsahu portů je rozložen na uzly NLB clusteru

Disabled - bez filtrování síťového provozu

Client affinity

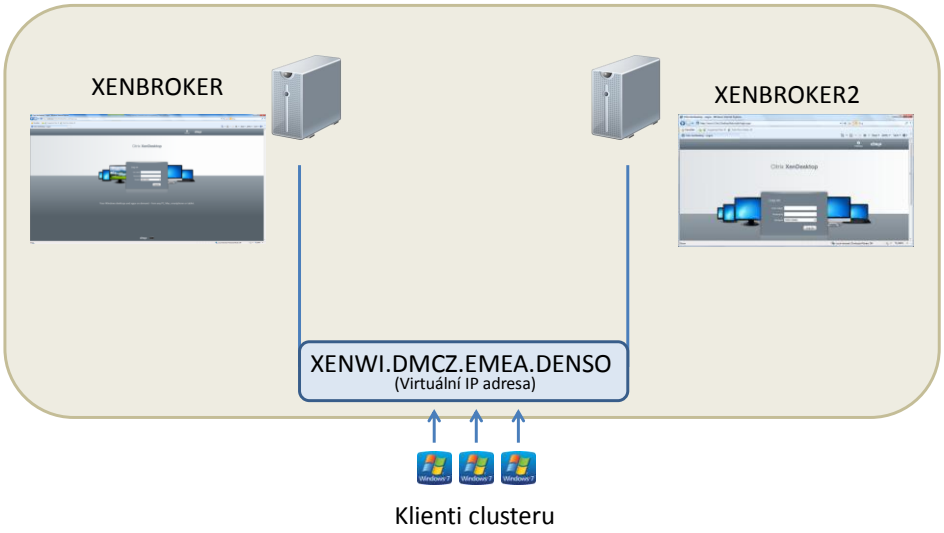
Single - požadavky z jedné klientské IP adresy jsou vždy směrovány na stejný uzel NLB clusteru po celou dobu trvání relace klient/ server. Pro šifrovanou SSL komunikaci (protokol HTTPS a další) je nutné zvolit toto nastavení z důvodu přenosu jedinečných stavových informací v zabezpečené relaci klient/ server, aby se předešlo cyklické SSL re-autentifikaci při odpovědi různých uzlů NLB clusteru.

Network – požadavky přicházející z třídy sítě typu C budou vždy směrovány na stejný uzel NLB clusteru po celou dobu trvání relace klient/ server. Toto nastavení je vhodné pro směrování požadavků klientů NLB clusteru přicházejícím prostřednictvím síťově vyvážených proxy serverů. Požadavek z každého proxy serveru bude mít jinou zdrojovou IP adresu, ale vždy přichází z adresního rozsahu jedné sítě typu C. Tato volba zajišťuje, že spojení bude stále udržováno i přes změnu zdrojové IP adresy.

None - požadavky z jedné klientské IP adresy jsou podle algoritmu vyvažování zátěže rozloženy na různé uzly NLB clusteru po celou dobu trvání relace klient/ server.

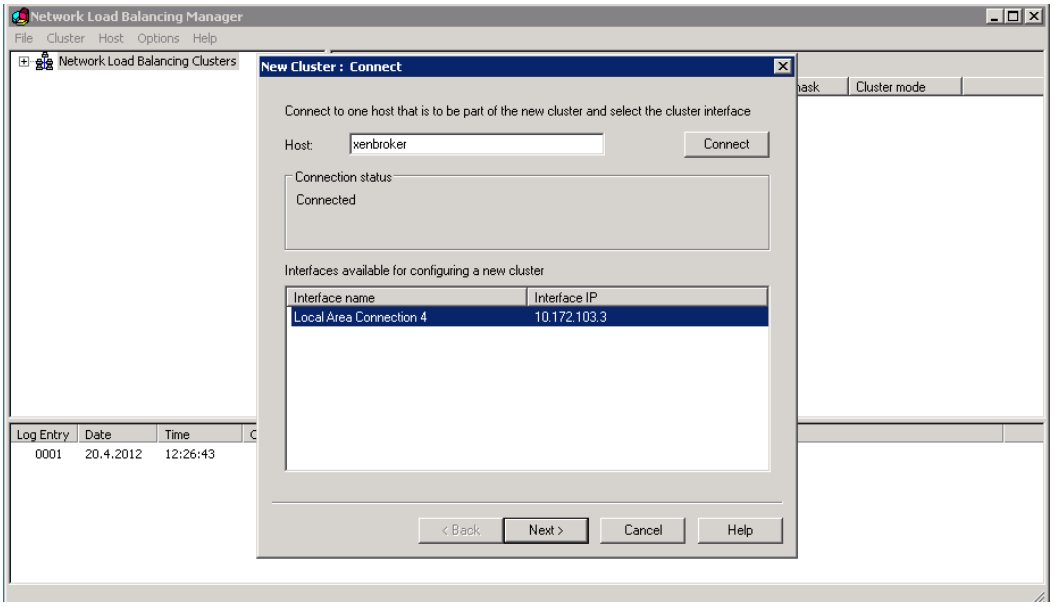
4.2.6 Implementace služby Network Load Balancing

Cluster byl sestaven z uzlů XENBROKER (IP adresa 10.172.103.3) a XENBROKER2 (IP 10.172.103.4). Serverové uzly hostující clusterovou službu NLB vyžadují podporu doplňkové funkce Network Load Balancing, kterou je možné nainstalovat prostřednictvím Server Manageru (správce serveru) a volby Add Feature (přidat funkci), nebo pomocí příkazové řádky spuštěním příkazu „*servermanagercmd -install nlb*“



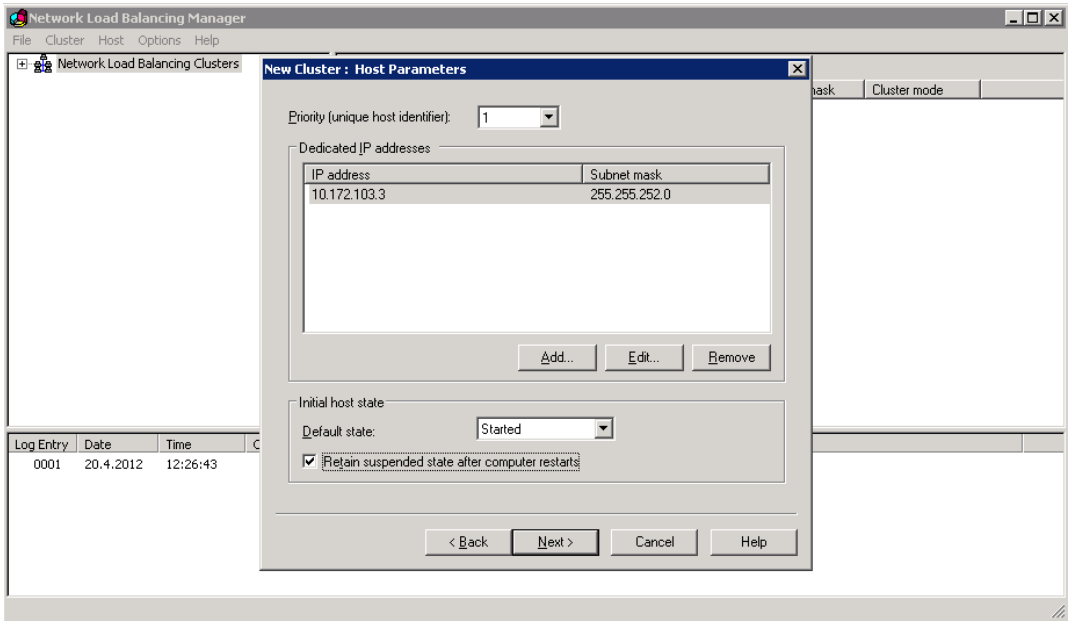
Obr. 31: Schéma realizovaného NLB clusteru

Při realizaci nejprve iniciujeme vytvoření NLB clusteru z menu administrátorské konzole pomocí položky „New Cluster a zadáme síťové jméno prvního uzlu XENBROKER.



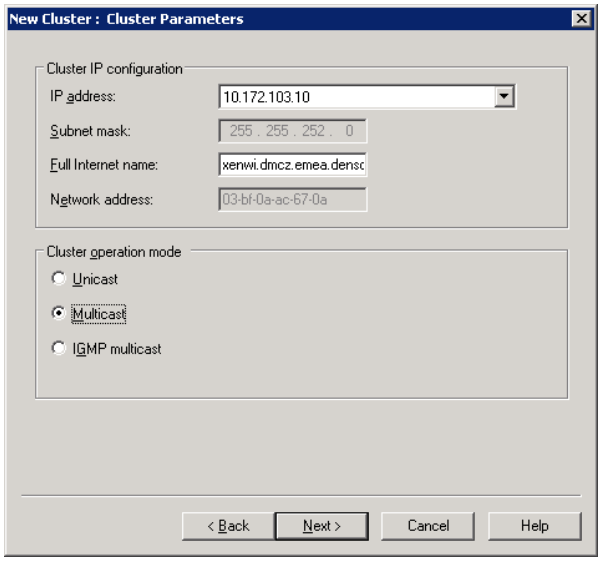
Obr. 32: Přidání prvního uzlu

V dalším kroku můžeme nastavit prioritu uzlu (zde automaticky vybraná nejvyšší) a kontrolovat výchozí stav pro spuštění NLB služby. Nastavíme výchozí stav na Started (spuštěno) s volbou pro vynucení zachování stavu Suspend (pozastavit) i po restartu uzlu. Tento stav se využívá pro účely administrace příslušného uzlu a může být žádoucí zachovat NLB službu pozastavenou až do doby, kdy jsou administrativní úkony zcela dokončeny.



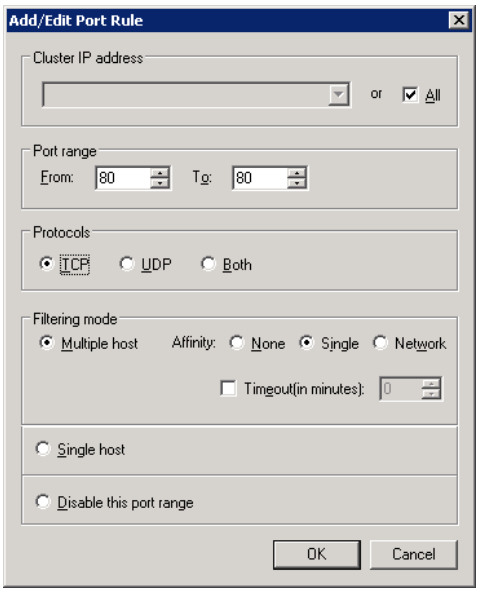
Obr. 33: Nastavení parametrů prvního uzlu

Prostřednictvím následujících dialogů nastavíme virtuální IP (zvoleno 10.172.103.10) adresu NLB clusteru, kterou použijí klienti pro připojení ke službě Citrix Web Interface a této IP adrese přiřadíme virtuální síťový název (zvoleno XENWI) NLB clusteru ve formátu FQDN, pro který později vytvoříme příslušný záznam v DNS. Zvolíme také příslušný mód režimu volání, z důvodů nutnosti vzájemné komunikace mezi servery XENBROKER1 a XENBROKER2 v roli DDC controllerů je nutná volba Multicast.



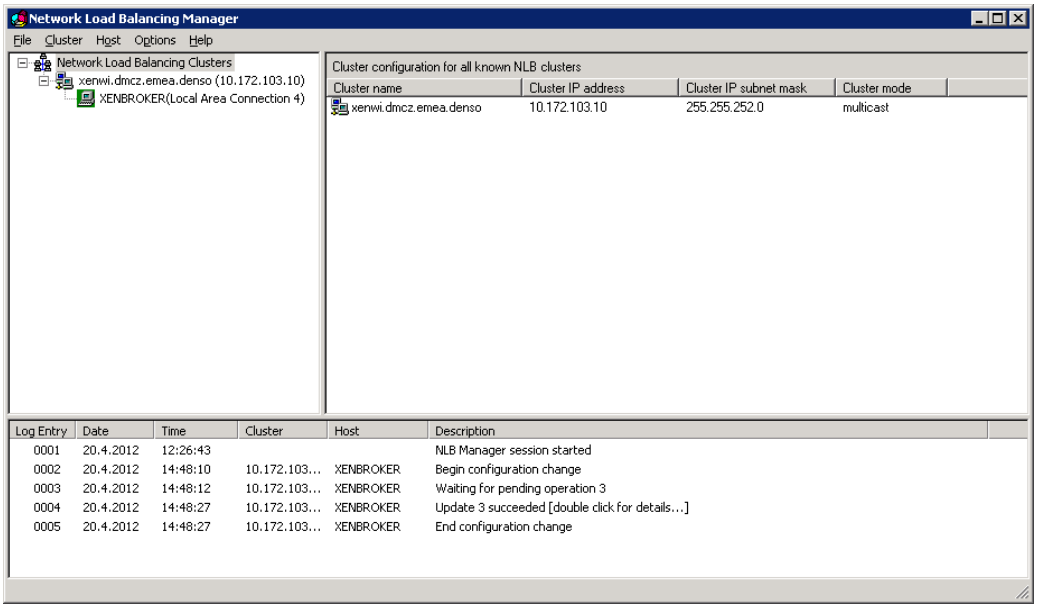
Obr. 34: Nastavení virtuální IP adresy NLB clusteru

Následující volbou změníme výchozí operační rozsah síťových portů clusteru pouze na požadovaný port 80 (služba HTTP) a protokol TCP.

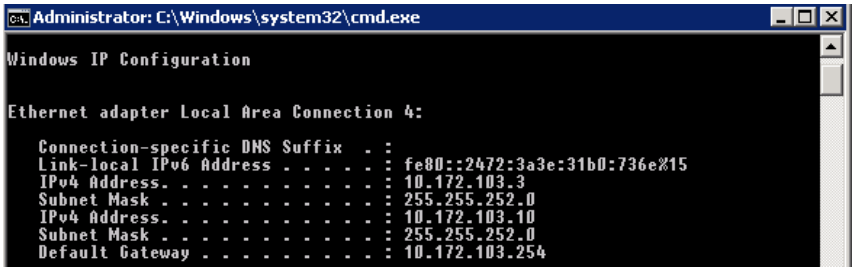


Obr. 35: Definice síťových portů

Po dokončení nastavení jednotlivých parametrů dojde k vytvoření NLB clusteru a automatické aktivaci clusterové služby na právě definovaném prvním uzlu XENBROKER. Na síťovém adaptéru tohoto uzlu dojde k přiřazení virtuální clusterové IP adresy (10.172.103.10).

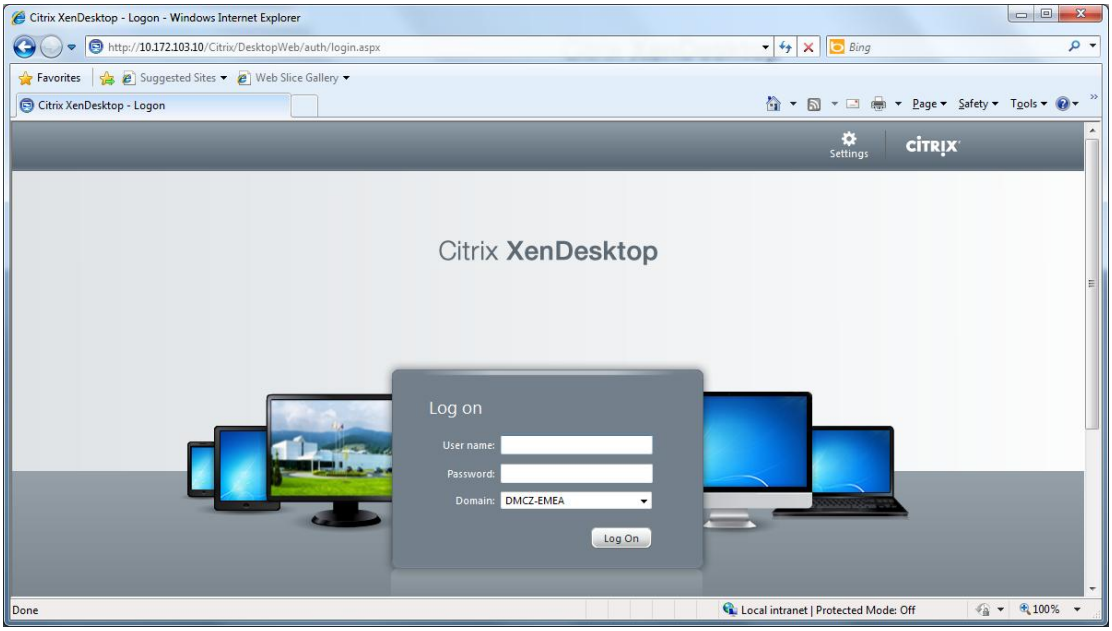


Obr. 36: Administrátorská konzole po konfiguraci prvního uzlu



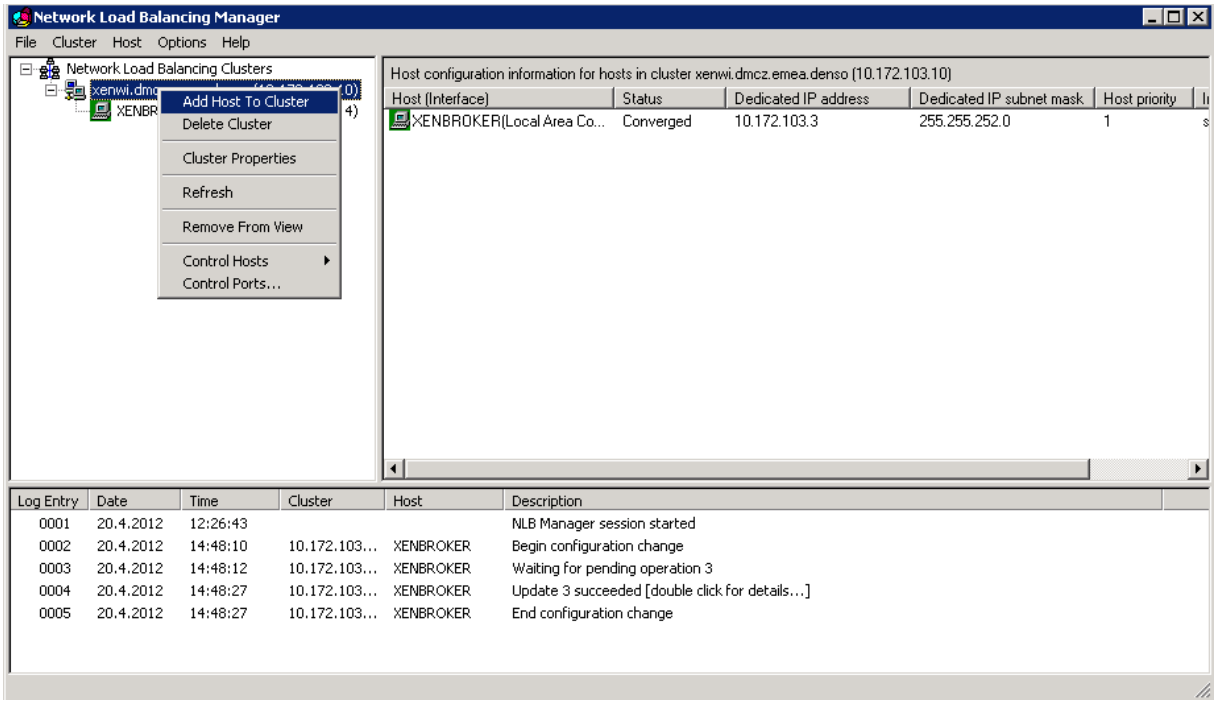
Obr. 37: Virtuální IP adresa na síťovém adaptéru prvního uzlu

Před přidáním druhého uzlu NLB clusteru ověříme funkčnost clusterové službu pomocí webového prohlížeče připojením na virtuální IP adresu clusteru (10.172.103.10).



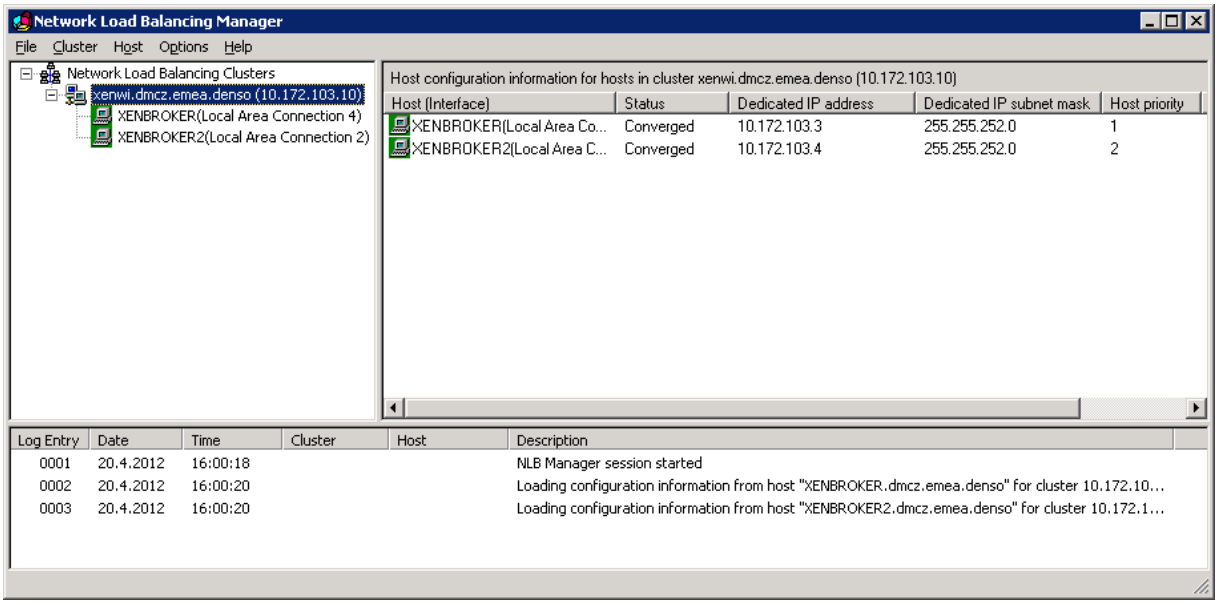
Obr. 38: Připojení na virtuální IP adresu clusteru 10.172.103.10

Druhý uzel clusteru (*XENBROKER2*) přidáme příkazem „Add Host To Cluster“ z administrátorské konzole.



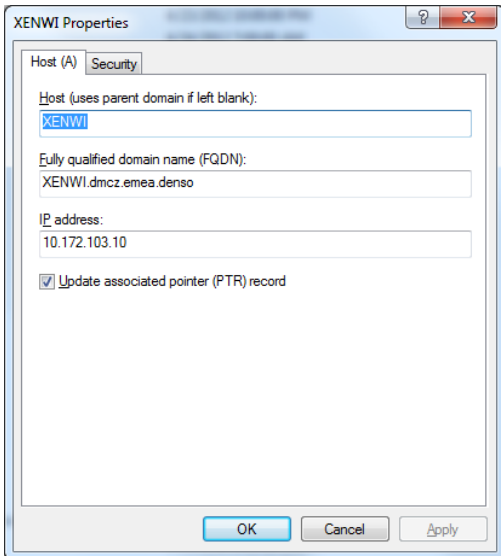
Obr. 39: Přidání druhého uzlu NLB clusteru

Podobně jako v případě konfigurace prvního uzlu vybereme síťový adaptér a nastavíme požadované identické parametry pro druhý uzel XENBROKER2. Po dokončení konfigurace a procesu konvergence dojde k přidání druhého uzlu do clusterové služby.

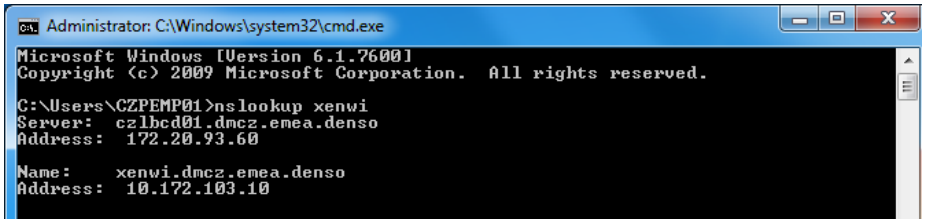


Obr. 40: Administrátorská konzole sestaveného NLB clusteru

Pro zvolenou virtuální IP adresu clusteru (10.172.103.10) vytvoříme záznam v systému DNS, správný překlad vytvořeného záznamu (*XENWI)* na IP adresu zkontrolujeme příkazem „nslookup“ z příkazového řádku.

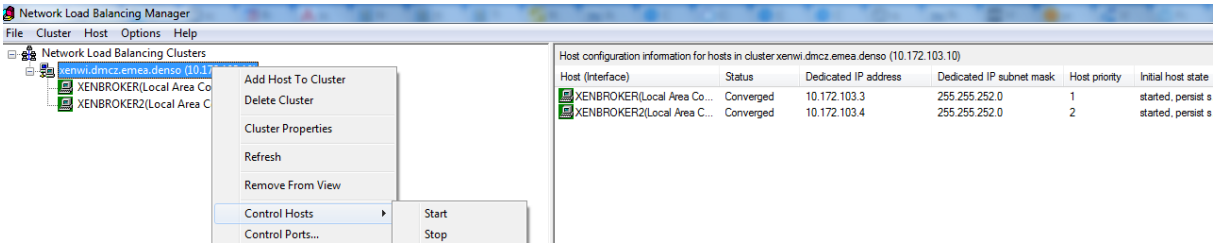


Obr. 41: Vytvoření jmenného DNS záznamu XENWI pro virtuální IP adresu NLB clusteru

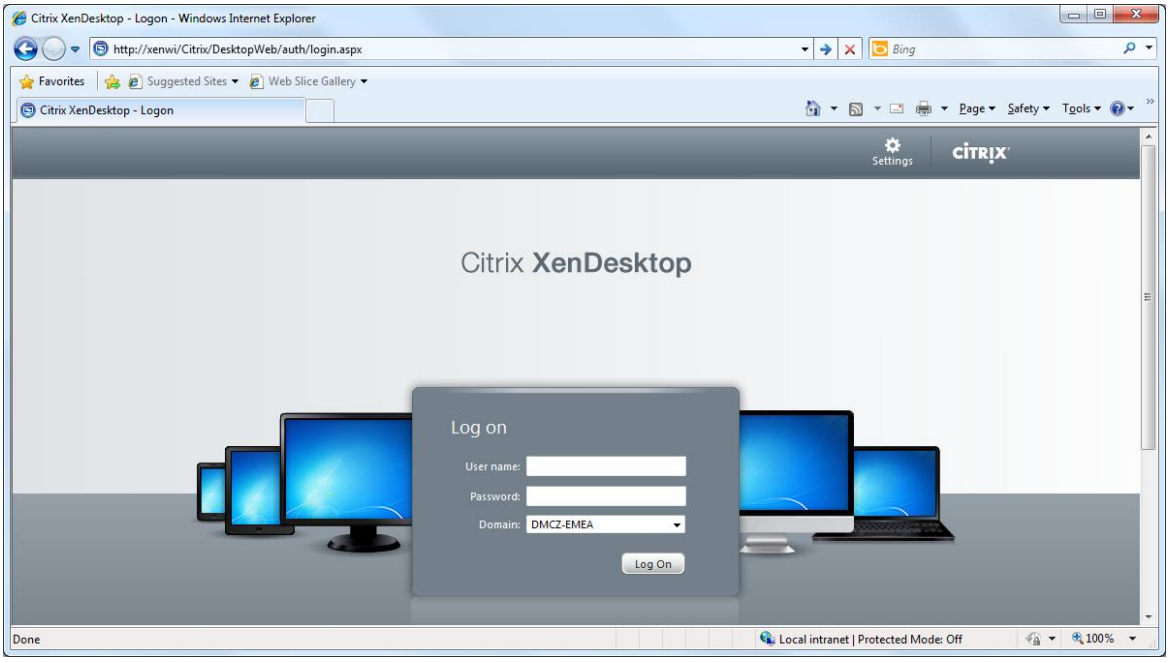


Obr. 42: Test překladu jmenného názvu XENWI na IP adresu

Správnou funkci clusterové NLB služby otestujeme z administrátorské konzole zastavením služby na uzlu XENBROKER (simulujeme tedy výpadek tohoto uzlu) a připojením na virtuální adresu clusteru http://xenwi, kde dojde k přesměrování požadavku na stále aktivní uzel XENBROKER2 (pro snažší identifikaci cílového serveru byla na tomto uzlu drobně pozměněna grafika úvodní přihlašovací stránky).

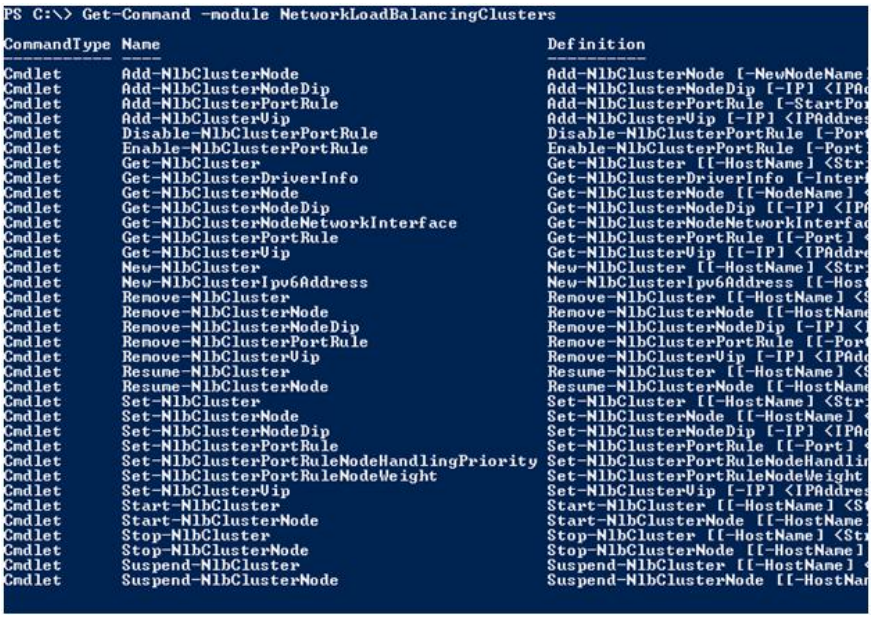


Obr. 43: Simulace výpadku prvního uzlu



Obr. 44: Přesměrování dotazu na http://xenwi na funkční druhý uzel

Správa služby NLB probíhá standardně přes GUI management konzoly se snap-in pluginem Network Load Balancing Manager připojením na virtuální IP adresu NLB clusteru. Všechny příkazy dostupné z konzole mají svůj ekvivalent v podobě příslušného command-let (cmdlet) obsaženého v příkazové sadě pro NLB Cluster interpreteru Powershell.



Obr. 45: Výpis Command-Let pro NLB Cluster

5. ZÁVĚR

Cílem diplomové práce byla implementace a konfigurace mechanismů pro zajištění vysoké dostupnosti u vybraných služeb virtuální infrastruktury. Předpokladem praktické realizace bylo detailní pochopení jednotlivých systémových a aplikačních vazeb včetně konkrétně definovaného návrhu řešení respektujícího přísný režim pro zajištění nepřetržitého provozu s ohledem na klíčové požadavky společnosti Denso Manufacturing Czech s.r.o.

Realizace s využitím clusterových mechanismů operačního systému Microsoft Windows představuje robustní a cenově značně výhodné řešení, které omezuje pravděpodobnost výpadku jednotlivých služeb na minimum a plně využívá výhody virtuální serverové infrastruktury. Clusterové služby umožní při výpadku serverového uzlu zcela automatický a transparentní proces migrace zdrojů a služeb na funkční serverový uzel, kde v původním pojetí by tento výpadek znamenal nedostupnost služby a vyžadoval následný ruční zásah administrátora.

Při návrhu řešení zajišťující služby v režimu vysoké dostupnosti je také nutné zvážit veškeré možné scénáře selhání jednotlivých komponent a následně kompletně eliminovat zjištěná kritická místa infrastruktury a to nejenom na aplikační úrovni. Před uvedením do ostrého provozu musí být důkladně otestována stabilita systémů a schopnost poskytování jejich služeb v režimu vysoké dostupnosti pomocí sady testů simulujících reálná selhání. Koncept musí také vycházet z předpokladu, že navrhnuté řešení musí být dlouhodobě funkční a umožnit intuitivní správu a dohled prostřednictvím monitorovacích systémů nebo administrátorských nástrojů. Realizované řešení splňuje kritéria zadání a clusterové služby mohou dále umožnit provoz dalších serverových služeb (např. tiskový server) v režimu vysoké dostupnosti.

Použitá literatura

[1] WILLIAM R. STANEK. Mistrovství v Microsoft Windows Server 2008. COMPUTER PRESS, Brno 2011

[2] SHARON CRAWFORD, CHARLIE RUSSEL. Microsoft Windows Server 2008. Velký průvodce administrátora. COMPUTER PRESS, Brno 2009

[3] Microsoft [online]. [2012-04-02] Failover Clustering in Windows Server 2008 R2 dostupné z WWW: <http://download.microsoft.com>

[4] Wikipedia [online]. 2012 [2012-03-28] Preboot Execution Environment Dostupné z WWW: < http://en.wikipedia.org/wiki/Preboot_Execution_Environment>

[5] Wikipedia [online]. 2012 [2012-03-28] Trivial File Transfer Protocol Dostupné z WWW: <http://en.wikipedia.org/wiki/Trivial_File _Transfer_Protocol)>

[6] Citrix [online]. 2012 [2012-04-10] Citrix Knowledge Center Dostupné z WWW: <http://support.citrix.com/>

[7] Citrix [online]. 2012 [2012-04-10] Citrix eDocs Dostupné z WWW: <http://support.citrix.com/>

[8] RINTALAN, NICK: The Citrix Blog [online]. 2011 [2012-04-10] Load Balancing TFTP Anything But Trivial

Dostupné z WWW: <http://blogs.citrix.com/2011/05/02/load-balancing-tftp-anything-but-trivial/>

[9] HEINZ, DOMINIK. Sign In Client Management Blog [online]. 18.3.2011 [2012-04-20] DHCP & PXE basics

Dostupné z WWW: < http://blogs.technet.com/b/dominikheinz/archive/2011/03/18/dhcp-amp-pxe-basics.aspx>

[10] MANSELL TREVOR. The Citrix Blog [online]: 25.2.2009 [2012-04-10] Using PVS Boot Device Manger with XenDesktop and Xenserver

Dostupné z WWW: < http://blogs.citrix.com/2009/02/25/using-pvs-boot-device-manger-with-xendesktop-and-xenserver/>

[11] BOUŠKA PETR. [online]. 9.8.2010 [2012-04-20] Microsoft Network Load Balancing (NLB) a Cisco switche

Dostupné z WWW: http://www.samuraj-cz.com/clanek/microsoft-network-load-balancing-nlb-a-cisco-switche/

[12] Techtopia [online].2011 [2012-04-20] Building a Windows Server 2008 Network Load Balancing Cluster

Dostupné z WWW: <http://www.techotopia.com/index.php/Building_a_Windows_Server_2008_Network_Load Balancing_Cluster>

[13] Denso [online]. 3.6.2011 [2012-03-05] Výroční zpráva 2010

Dostupné z WWW: <http://www.denso.cz/files/vyrocni%20zpravy/2440%20Denso%20-%20Annual%20Report%202010%20+%20(M)%20ticked_CZ.pdf>

[14] KUMAR NITISH. [online] 31.10.2011 [2012-04-02] Failover Clustering in Windows Server 2008 R2 Part 1

Dostupné z WWW: <http://winadmins.wordpress.com/2011/10/31/failover-clustering-in-windows-server-2008-r2-part-1/>

[15] HUGHES JEFF. [online] 23.11.2009 [2012-04-02] Resource Hosting Subsystem (RHS) In Windows Server 2008 Failover Clusters

Dostupné z WWW: <http://blogs.technet.com/b/askcore/archive/2009/11/23/resource-hosting-subsystem-rhs-in-windows-server-2008-failover-clusters.aspx>

Seznam zkratek:

IT	Information Technology
DMCZ	Denso Manufacturing Czech
VDI	Virtual Desktop Infrastructure
HW	Hardware
SW	Software
VM	Virtual Machine
NFS	Network File System
HBA	Host Bus Adapter
ISCSI	Internet Small Computer System Interface
SSH	Secure Shell
DDC	Desktop Delivery Controller
IMA	Independent Management Architecture
IIS	Internet Information Services
CDS	Citrix Desktop Service
PVS	Provisioning Server
SPOF	Single Point of Failure
TFTP	Trivial File Transfer Protocol
DHCP	Dynamic Host Control Protocol
DNS	Domain Name Systém
PXE	Preboot Execution Environment
BDM	Boot Device Manager
GPO	Group Policy Object

SAN	Storage Area Network
RCM	Resource Control Manager
HA	High Availability
RHS	Resource Hosting Subsystem
OS	Operating Systém
HTML	HyperText Markup Language
LUN	Logical Unit
NTFS	New Technology File Systém
NAT	Network Access Translation
NLB	Network Load Balancing
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

Seznam obrázků a tabulek: