Pedagogická fakulta
Faculty of Education

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích

Pedagogická fakulta

Katedra informatiky

Technologie zálohování dat se zaměřením na backup mezi vzdálenými pracovišti s využitím systému EMC Avamar deduplication backup

Data backup technology with a focus on backup between remote sites using the EMC Avamar deduplication backup

Bakalářská práce

Vypracoval: Kindelmann Ladislav

Vedoucí práce: PaedDr. Petr Pexa, Ph.D.

V Českých Budějovicích 2016

JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH
Fakulta pedagogická
Akademický rok: **2013/2014**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Ladislav KINDELMANN**
Osobní číslo:	**P120249**
Studijní program:	**B7507 Specializace v pedagogice**
Studijní obor:	**Informační technologie a e-learning**
Název tématu:	**Technologie zálohování dat se zaměřením na backup mezi vzdálenými pracovišti s využitím systému EMC Avamar Deduplication Backup**
Zadávající katedra:	**Katedra informatiky**

Zásady pro vypracování:

Cílem bakalářské práce je rozbor způsobu zálohování dat mezi vzdálenými pracovišti technologií deduplikace, využívané systémem EMC Avamar. Jedná se o speciální způsob komprese dat využitelný nejen při ukládání do diskových polí, ale i ve virtuálním prostředí. Deduplikace zajišťuje úsporu místa úložistě a spolehlivost při obnově dat. Systém EMC Avamar je jedním z technicky nejpropracovanějších a nejrozsáhlejších systémů, který výše uvedenou technologii využívá. V teoretické části student shrne dostupné technologie zálohování a rozebere způsoby vzdáleného zálohování včetně příkladů užití dostupného SW a nároků na HW. Poté se detailně zaměří na popis softwaru EMC Avamar, jeho implementaci, HW nároky a popis fungování celého systému. Student popíše technologii deduplikace dat, kterou systém využívá, vysvětlí princip přenosu dat a případné výhody či nevýhody oproti jiným systémům. V praktické části student vytvoří návrh vlastní realizace zálohování dat s využitím výše jmenovaného systému ve firemním prostředí mezi dvěma pobočkami společnosti, na kterém otestuje reálné vlastnosti této technologie, výhody i nedostatky. Závěr bakalářské práce bude doplněn přehledem využití zálohovacích systémů na současném světovém trhu.

Rozsah grafických prací: **CD ROM**

Rozsah pracovní zprávy: **40**

Forma zpracování bakalářské práce: **tištěná**

Seznam odborné literatury:

1. **EMC Avamar, Data protection suite [online]. 2014 [cit. 2014-21-03]. Dostupné z: https://www.emc.com/backup-and-recovery/avamar/avamar.htm**
2. **EMC Avamar, Documentations [online]. 2014 [cit. 2014-21-03]. Dostupné z: https://support.emc.com/products/759_Avamar-Server**
3. **VCE Company, VBLOCK SYSTEMS BACKUP AND RECOVERY-BEST PRACTISES [online]. 2012 [cit. 2014-21-03]. Dostupné z: http://www.vce.com/asset/documents/avamar-data-domain-whitepaper.pdf**
4. **Microsoft, Data deduplication overview [online]. 2013 [cit. 2014-21-03]. Dostupné z: http://technet.microsoft.com/en-us/library/hh831602.aspx**
5. **Roy Mikes, EMC Avamar: a deduplication way to go [online]. 2012 [cit. 2014-21-03]. Dostupné z: http://virtualization.blognotions.com/2012/01/05/emc-avamar-a-deduplication-way-to-go/**
6. **EMC Avamar: Deduplikace dat [online]. 2014 [cit. 2014-21-03]. Dostupné z: http://www.deduplikace.cz/emc-avamar.php**

Vedoucí bakalářské práce: **PaedDr. Petr Pexa, Ph.D.**
Katedra informatiky

Datum zadání bakalářské práce: **27. března 2014**
Termín odevzdání bakalářské práce: **30. dubna 2015**

Mgr. Michal Vančura, Ph.D.
děkan

L.S.

doc. PaedDr. Jiří Vaníček, Ph.D.
vedoucí katedry

V Českých Budějovicích dne 27. března 2014

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 8. prosince 2016.

Ladislav Kindelmann

Podpis

Abstrakt

Cílem bakalářské práce je rozbor způsobu zálohování dat vzdálených pracovišť technologií deduplikace využívané systémem EMC Avamar. Jedná se o speciální způsob komprese dat využitelný nejen při ukladání do diskových polí, ale i ve virtuálním prostředí. Deduplikace zajišťuje úsporu místa úložistě a spolehlivost při obnově dat. Systém EMC Avamar je jedním z technicky nejpropracovanějších a nejrozsáhlejších systémů, který výše uvedenou technologii využívá. Práce je rozdělena na dvě části - teoretickou a praktickou.

V teoretické části student stručně shrne dostupné technologie zálohování a rozebere způsoby vzdáleného zálohování včetně příkladů užití dostupného software a nároků na hardware. Poté se detailně zaměří na popis softwaru EMC Avamar, jeho implementaci, HW nároky a popis fungování celého systému. Student rozebere technologii deduplikace dat, kterou systém využívá, popis přenosu dat a případné výhody či nevýhody oproti jiným systémům.

V praktické části poté vytvoří vlastní návrh realizace zálohování dat s využitím výše jmenovaného systému ve firemním prostředí mezi dvěmi pobočkami společnosti. Cílem bude ukázat reálné vlastnosti technologie, výhody i nedostatky. Závěr bakalářské práce student navíc doplní přehledem využití zálohovacích systémů na současném světovém trhu.

Klíčová slova

Avamar, deduplikace, zálohování, LTO, DAS, NAS, SAN

Abstract

The aim of this dissertation is to analyze the way data backup of remote sites deduplication technology system used by EMC Avamar. This is a special type of data compression used not only when saving to disk arrays, but also in a virtual environment. Deduplication ensures saving storage space and reliability of data recovery. EMC Avamar system is one of the best-engineered and most extensive systems of which the above technology is used. The dissertation is divided into two parts - theoretical and practical.

In the theoretical part student briefly summarize the available backup technologies and describes the remote backup methods, including examples using available software and hardware requirements. Then it specifies details of EMC Avamar and its implementation, HW requirements and functioning of the entire system. Student will analyzes the data deduplication technology, which the system uses, description of data transferring and the possible advantages and disadvantages compared to other systems.

In the practical part it creates own proposal for implementation of data backup using the above mentioned system in a business environment between the two branches of the company. The aim will be to show the real technology features, advantages and shortcomings. The conclusion of the dissertation the student all accompanied overview of the use of backup systems in today's global market.

Keywords

Avamar, deduplication, data backup, LTO, DAS, NAS, SAN

Poděkování

Velmi si vážím a děkuji panu PaedDr. Petrovi Pexovi Ph.D. za ochotu, přátelský a profesionální přístup při vedení mé bakalářské práce, za rady i poznatky, které mi poskytl a hlavně za trpělivost a podporu, kterou mi byl ochoten obětovat při tvorbě této práce.

Obsah

1 Úvod **10**

1.1 Cíle práce 10

1.2 Východiska práce 11

1.3 Metody práce 12

2 Teoretická část **13**

2.1 Moderní technologie zálohování 13

2.2 Linear Tape Open (LTO) 14

2.3 Direct Attached Storage (DAS) 16

2.4 Network Attached Storage (NAS) 18

2.5 Storage Attached Network (SAN) 20

2.6 Co je deduplikace dat 21

2.7 Avamar 27

2.7.1 Redundant Array of Independent nodes (RAIN) 30

2.8 IBM Tivoli Storage Manager 34

2.9 Symantec NetBackup 38

3 Praktická část **43**

3.1 Představení projektu 43

3.2 Technické požadavky 43

3.3 Popis a ukázka praktického návrhu 44

3.3.1 Administrační rozhraní 44

3.3.2 Automatický proces zálohování 46

3.3.3 Proces deduplikace 53

4 Závěr **59**

Seznam použité literatury a zdrojů **61**

Seznam obrázků **66**

Seznam tabulek **67**

A Příloha **68**

1 Úvod

Informace, uložené v elektronické podobě na nejrůznějších platformách, jsou v současné době nezbytností. Na elektronických datech jsme již natolik závislí, že jejich ztráta nebo výpadek může být pro společnost likvidační. Prevence zálohováním byla často podceňovanou a obtěžující záležitostí a ve chvíli poruchy, ztráty dat, bylo samozřejmě pozdě.

Přicházely na řadu specializované firmy, pokoušející se o nápravu za nemalé finanční prostředky a s nejistým výsledkem. Situaci navíc komplikuje neustálý růst ukládaných dat nutící uživatele vynakládat větší a větší finanční prostředky na profesionální ochranu informací. Sílí tlak společnosti na pravidelné zálohování s maximálním využitím a úsporou datových úložišť a tím snížení vynakládaných prostředků. V moderní době plné nepravdivých mediálních kampaní je velmi komplikované vybrat spolehlivé řešení odpovídající potřebám konkrétní společnosti a zároveň schopné zálohovat a obnovovat data po celém světě.

Bakalářská práce se zabývá problematikou zálohování dat systémem EMC Avamar využívající deduplikace. Ta zabraňuje ukládání duplicitních informací, čímž snižuje objem zálohovaných dat a šetří potřebný prostor v datových centrech. Řešení EMC Avamar je jedním z velmi kvalitních možností uplatnění této technologie, zkrácení doby potřebné k vytvoření zálohy a ochrany dat lokálních sítí a vzdálených pracovišť.

1.1 Cíle práce

Cílem bakalářské práce je rozbor způsobu zálohování dat vzdálených pracovišť technologií deduplikace využívané systémem EMC Avamar. Jedná se o speciální způsob komprese dat využitelný nejen při ukládání do diskových

polí, ale i ve virtuálním prostředí. Deduplikace zajišťuje úsporu místa úložiště a spolehlivost při obnově dat. Systém EMC Avamar je jedním z technicky nejpropracovanějších a nejrozsáhlejších systémů, který výše uvedenou technologii využívá.

V teoretické části student stručně shrne dostupné technologie zálohování a rozebere způsoby vzdáleného zálohování včetně příkladů užití dostupného software a nároků na hardware. Poté se detailně zaměří na popis softwaru EMC Avamar, jeho implementaci, HW nároky a popis fungování celého systému. Student rozebere technologii deduplikace dat, kterou systém využívá, popis přenosu dat a případné výhody či nevýhody oproti jiným systémům.

V praktické části bude vytvořen vlastní návrh zálohování dat s využitím výše jmenovaného systému ve firemním prostředí mezi dvěmi pobočkami společnosti. Cílem bude ukázat reálné vlastnosti technologie, výhody i nedostatky.

Závěr bakalářské práce bude navíc doplněn přehledem využití zálohovacích systémů na současném světovém trhu.

1.2 Východiska práce

S rozvojem IT je stále důležitější zajistit ochranu dat a jejich zálohu. V moderní době je navíc vyžadováno zálohovat data nejen v místních počítačových sítích, ale také ze vzdálených poboček, například pomocí speciálního centrálního softwaru umožňujícího řídit celý proces. Neméně důležitým požadavkem je vysoká komprimace dat kvůli úspoře datového prostoru cílového zařízení.

Systém EMC Avamar je možností, jak využít jednu z nejmodernějších komprimačních metod, takzvanou deduplikaci. Propracovaný software nabízí ucelené řešení umožňující zálohu dat lokálních sítí i vzdálených poboček využívající privátní i veřejné internetové konektivity. Samozřejmostí je vysoká

komprimace dat, filtrování obsahu, plán úloh či obnova dat z datového centra do vzdálených poboček.

1.3 Metody práce

V úvodu budou stručně představené používané technologie zálohování dat vzdálených pracovišť dnešní doby, bude rozebrána technologie deduplikace dat, kterou systém využívá, přenos dat a případné výhody či nevýhody oproti jiným systémům. Poté je práce detailně zaměřena na produkt EMC Avamar, implementaci, systémové nároky a fungování celého systému.

Dostupnost aktuálních knih v českém jazyce na toto téma je bohužel omezená, v práci bylo čerpáno z cizojazyčných publikací, elektronických knih, internetových stránek a oficiálních zdrojů výrobců.

V praktické části bude vytvořen vlastní návrh zálohování dat s využitím výše jmenovaného systému ve firemním prostředí mezi dvěmi pobočkami společnosti. Budou popsány reálné vlastnosti systému EMC Avamar a upozorním na případné výhody či nedostatky. Praktický projekt bude opatřen obrázky dokumentujícími zásadní postupy i komplikace, které projekt provázely, navíc doplněné přehledem využití zálohovacích systémů na současném světovém trhu.

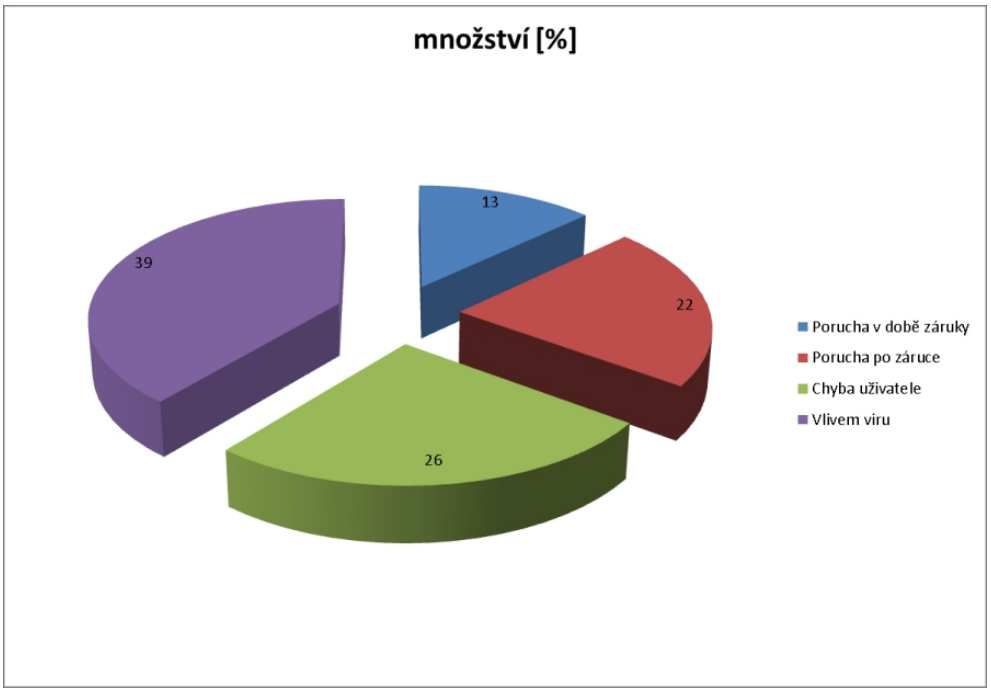
Testování softwaru byly omezeno reálnými podmínkami firemního prostředí a daným technickým vybavením. Z bezpečnostních důvodů byla skryta i některá označení testovaných prostředků v ilustračních screenech. To však nijak neovlivnilo záměr práce nebo její výsledek.

2 Teoretická část

2.1 Moderní technologie zálohování

Zálohování je procesem vytváření kopie a její uložení na jiné médium. Zálohovat lze celá úložiště, diskové oddíly nebo jednotlivé datové soubory. Zálohováno je z důvodu ochrany dat před poškozením, ať už úmyslným nebo neúmyslným, ztrátě nebo například k archivaci změn dokumentů.

Díky svým záznamům o servisních zásazích vám můžu sdělit následující data: Z celkového počtu 9742 klientů za 20 let se řešila ztráta nebo poškození dat v 6125 případech.[1]



Obrázek 1: Důvody poškození nebo ztráty dat [1]

Nejdůležitějším faktorem pro dosažení spolehlivé zálohy je správná volba techniky zálohování, odpovídající hardware[1] a v neposlední řadě volba zálo-

[1]Technické vybavení

hovacího softwaru[2], který vyhovuje požadavkům zákazníka. V dnešní široké nabídce produktů je možné nalézt nejen firemní profesionální řešení, ale i kvalitní a levné aplikace do domácností.

Každý uživatel by se o svá či sdílená data měl starat tak, aby nedošlo k jejich poškození, přesněji řečeno provádět zálohy na jiné dostupné médium v pravidelných intervalech. V případě operativních záloh není nutný žádný speciální software, postačí bežné kopírování souborů. Cestou je i využití nástroje *Synchronizace dat*, která je součástí operačního systému MS Windows.

Způsoby zálohování

- Záloha prostým kopírováním na jiné médium, například externí disk, flashdisk a podobně.
- Pomocí speciálního softwaru a páskových mechanik nebo technologiemi DAS[3], NAS[4], SAN[5].

Uživatelé počítačů se dělí na dvě základní skupiny: Na ty, kteří už někdy o svá data přišli a na ty, kteří o svá data teprve přijdou. [1]

Ustupujícím typem médií pro ukládání zálohovaných dat jsou páskové mechaniky, které postupně nahrazují NAS nebo SAN storage[6].

2.2 Linear Tape Open (LTO)

I přes zvýšení výkonu serverů a vysokokapacitních datových úložišt, stále mnoho administrátorů považuje zálohování na pásková média za jakousi „poslední linii obrany“. Na trhu se objevuje mnoho typů mechanik, například

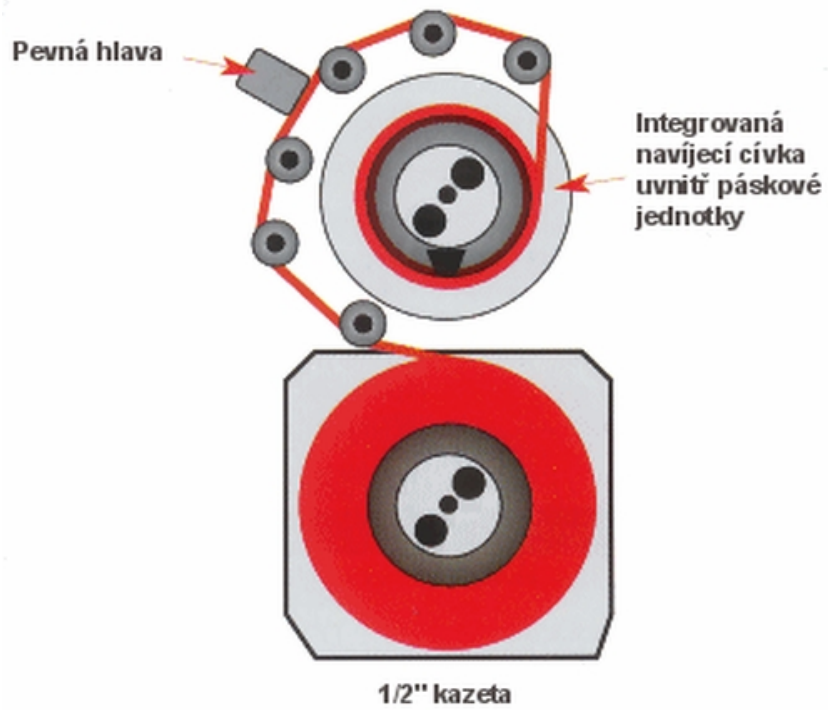
[2]Programové vybavení
[3]Direct Attached Storage
[4]Network Attached Storage
[5]Storage Area Network
[6]Úložiště

DAT, DLT, SDLT, LTO a další. Současným nejpoužívanějším standardem jsou LTO 6. generace vyvinuté společnostmi IBM, Quantum a HP garantující vzájemnou kompatibilitu.

Garance standardu zaručuje nejen kompatibilní kazetu, ale formát záznamu, komprese, samoopravných kódů i šifrování do takové míry, že kazeta zapsaná v mechanice jednoho výrobce musí být bez problémů čitelná v mechanice jiného výrobce.[2]

Na trhu jsou však k dispozici i LTO 7. a brzy LTO 8. generace. Média využívají formátů *Accelis* nebo *Ultrium*, lišící se přístupem k datům. Ultrium disponuje vysokou kapacitou a rychlým zápisem, naopak Accelis zrychluje čtení dat. Všechny generace LTO pásek mají stejné rozměry i vnitřní mechanismus, liší se pouze hustotou záznamu - záznamovou hlavou.

Životnost média, v zavislosti na používání, je 15 - 30 let.



Obrázek 2: Technologie LTO pásky [2]

Vlastnost	*LTO 1*	*LTO 2*	*LTO 3*	*LTO 4*	*LTO 5*	*LTO 6*	*LTO 7*	*LTO 8*
Kapacita bez komprese	100GB	200GB	400GB	800GB	1,5TB	2,5TB	6TB	12,8TB
Komprimovaná kapacita	200GB	400GB	800GB	1,6TB	3,0TB	6,25TB	15TB	32TB

Tabulka 1: Kapacita LTO všech generací

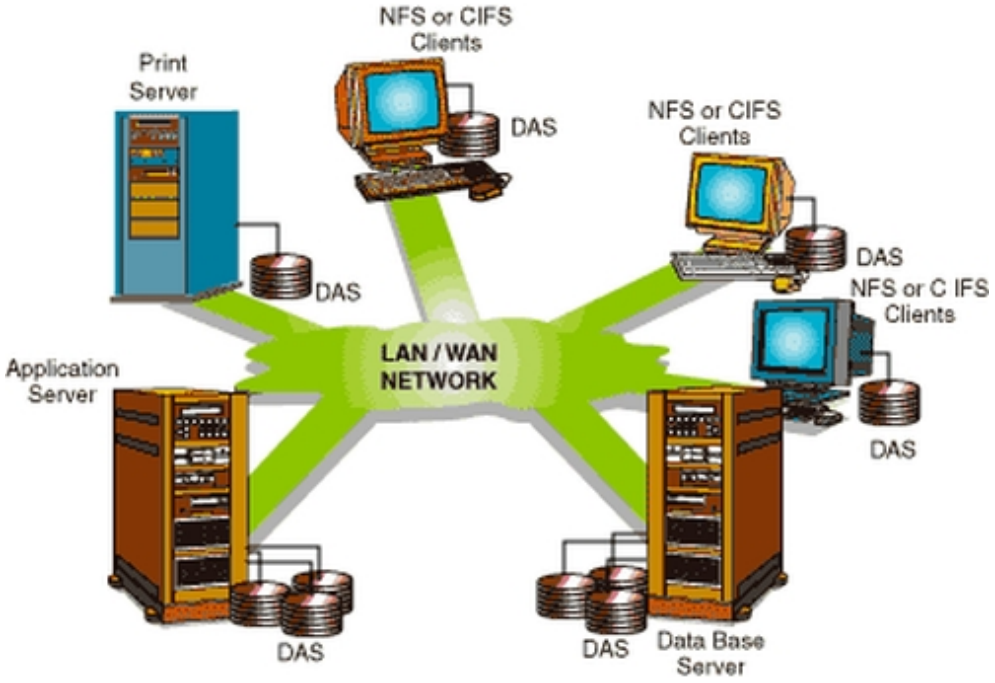
Standard byl navíc navržen tak, že mechanika každé generace umí číst dvě generace zpět a zapisovat na pásky jednu generaci zpět.[2]

Nová generace zálohovacích pásek implementuje Barium Ferritovou technologii zlešující magnetické vlastnosti záznamové vrtsvy dosahující kapacity až 35TB. Technologie vyvinutá společnostmi IBM a FUJIFILM Corporation se používá od LTO 6. generace.

2.3 Direct Attached Storage (DAS)

Datové úložiště je rozhraním ATA, SATA nebo SCSI[7] přímo spojeno se serverem přes který probíhá veškerá komunikace. Výhodou jsou nízké pořizovací náklady, jelikož úložištěm může být například i externí nebo interní disk počítače, záporem však potenciální přetěžování serveru, špatná rozšiřitelnost datové kapacity a problematické zálohování více strojů. Nejčastěji probíhá na každém serveru (stanici) samostatně.

[7]Počítačové sběrnice



Obrázek 3: Příklady implementace DAS [3]

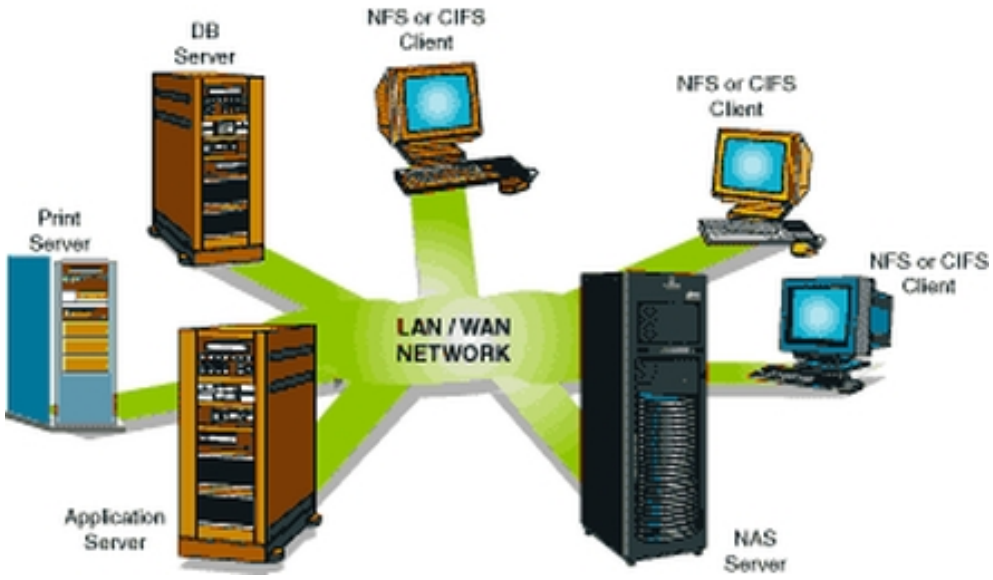
Není dosaženo centralizace, čímž narůstají náklady na odborný personál a zvyšuje se riziko chyby. DAS je zcela jistě aplikovatelný v malých společnostech a hlavně domácnostech připojením externího datového zařízení k vybranému počitači nebo serveru. Je však nutné si uvědomovat, že tato jednoduchá úložiště jsou jen minimálně, většinou vůbec, chráněna před fyzickým poškozením. Bezpečnějším způsobem aplikace DAS je připojení diskových polí chráněných některou z technologií RAID, což ale náklady na technické vybavení navyšuje.

V současné době je více než 95% datových úložišť, jako jsou pevné disky, disková pole a RAID systémy, přímo připojeno k počítači uživatele přes různé adaptéry standarizovanými protokoly jako SCSI, Fibre Channel a dalšími. Tento typ úložiště je také nazýván captive storage, server attached storage

nebo direct attached storage (DAS).[8][3]

2.4 Network Attached Storage (NAS)

NAS jsou specializované servery pro ukládání dat složené většinou z několika pevných disků (HDD) spravovaných metodou RAID, případně doplněné o páskové mechaniky. Připojují se do síťového prostředí (LAN), disponují vlastním mamagementem a v dnešní době i doplňujícími nástroji v podobě FTP serveru, Apache serveru, možnostmi sdílení, konfigurovatelnou bezpečnostní politikou a dalšími funkcemi, které se liší dle výrobce.



Obrázek 4: NAS topologie [3]

Skládají se z procesoru, řadiče RAID, síťové karty, uvedeného ovládacího softwaru a pevných disků SAS nebo SATA stejné kapacity. RAID techno-

[8]Today, greater than 95% of all computer storage devices such as disk drives, disk arrays and RAID systems are directly attached to a client computer through various adapters with standardized software protocols such as SCSI, Fibre Channel and others. This type of storage is alternatively called captive storage, server attached storage or direct attached storage (DAS).

logie zabezpečuje ochranu před fyzickým poškozením. Široká nabídka NAS storage dovoluje rackovou montáž (rackmount) stejně jako volně stojící variantu (standalone). Kapacita serverů, pohybující se v TB[9], možnosti škálovatelnosti a vybavení stroje jsou dnes vázané pouze na investici, kterou se zákazník rozhodne uvolnit. Server je umístěný v lokální sítí, kde provádí backup vybraných zařízení v nastavených intervalech, zálohování je tedy centralizované. Centralizace eliminuje rizika DAS technologie, doporučeno je též provádět replikaci dat na další NAS server, což umožňuje vhodný software.

Data mohou být na datový cíl bežně kopírovaná nebo řízena zálohovacím softwarem. V domácím prostředí zcela jistě postačí „ručně“ provedená záloha, naopak firemní prostředí vyžaduje profesionální přístup a konfiguraci automatických úloh. Správný výběr NAS může být limitován finančními prostředky, čím vyšší bezpečnost, příkladem RAID, tím je zapotřebí více pevných disků a vzrůstají pořizovací náklady.

Výhody síťového zálohování dat[4]

- Přechod k síťovým řešením zálohování dat přináší několik výhod najednou,
- zvyšuje se využívání kapacity zařízení na 75 % ve srovnání s 50% u DAS,
- snižuje se složitost a zvyšuje se produktivita IT pracovníků prostřednictvím centralizovaného managementu a rychlého a snadného zálohování,
- zvyšuje se dostupnost podnikových informací a zlepšuje se rychlost „zotavení“ po katastrofě.

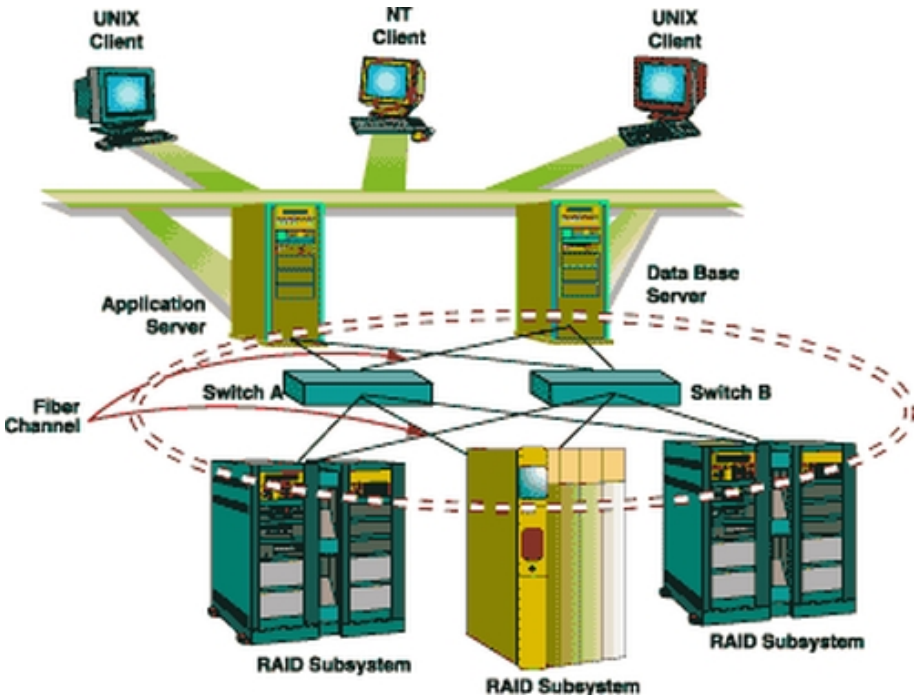
[9]Terabytes

2.5 Storage Attached Network (SAN)

SAN je zatím nejmodernějším způsobem uchovávání dat aplikovaným ve velkých počítačových sítích. Zálohovací zařízení jsou odděleně umístěná v datové síti, vzájemně propojená, komunikující se serverem nebo skupinou serverů.

V datové síti jsou umístěné vysokorychlostní switche kompatibilní s protokolem Fibre Channel nebo iSCSI, kontrolující hlavičky zasílaných datových paketů, které poté směřují do požadovaného SAN úložiště. Storage nedisponuje souborovým systémem, ovládají jej příkazy serverů. Každý server má vymezený určitý prostor (sektor) na storage, takzvaný LUN, který „vnímá“ jako svůj lokálně připojený pevný disk. Speciální clusterové systémy také umožňují sdílení konkrétního vyhrazeného LUNU mezi servery.

Právě použití sítě k vytváření sdíleného diskového úložiště (storage poolu) je to, co dělá SAN SANem.[5]



Obrázek 5: Storage Attached Network [3]

Fibre Channel protokol nejčastěji přenáší data optickým kabelem, alternativou je standardizované metalické médium. Je určený pro síťová úložiště, vyžaduje však dražší síťové switche. Naopak levnější alternativou je iSCSI, které používá stavající datovou síť a bežné ethernetové switche. Nižší pořizovací náklady jsou vykoupeny menší přenosovou rychlostí.

Výhodou Storage Attached Network architektury je rychlý přenos informací s využitím všech síťových tras k serverům, bezproblémová škálovatelnost, vysoká dostupnost a ochrana dat s podporovou bezpečnostní architektury *no single point of failure*[10]. Centralizovanost usnadňuje údržbu a snižuje provozní náklady. Vyšší pořizovací náklady jsou bohužel zápornou stránkou tohoto systému.

Gartner Group odhaduje, že na každý dolar vynaložený na hardware pro ukládání dat je třeba 3 dolarů na správu dat.[4]

2.6 Co je deduplikace dat

Důvodem implementace deduplikace byl obrovský narůst objemu dat a tím i zvýšené nároky na zálohovací kapacity. Společně s tím se samozřejmě navyšovaly i finanční prostředky k řešení těchto komplikací. Deduplikace je kompresní metoda eliminující duplicitu dat. Neukládá stejné souborové bloky dat, ale jen odkazy na ně.

Základní požadavky deduplikace

- Úspora kapacity úložište.
- Zkrácení zálohovacího času.
- Centralizace zálohování.

[10]Bezpečnostní část systému zajišťující vysokou dostupnost dat a spolehlivost přenosu

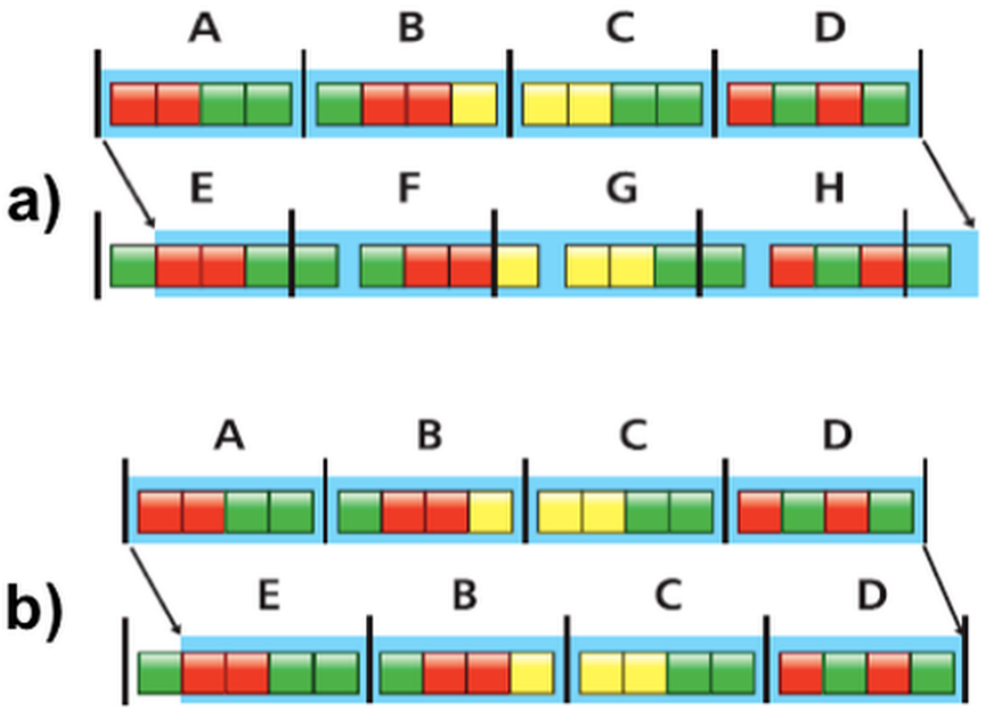
- Snížení požadavků na výkon datových linek.

Úspěšnost deduplikace samozřejmě závisí na mnoha parametrech, ale u vhodných datových „kolekcí“ – například mailových archivů, kde se tatáž příloha může vyskytovat znovu a znovu – může deduplikační poměr dosáhnout 100:1.[6]

Druhy deduplikace

- Na úrovni souboru
  - ▷ Málo efektivní metoda využití datového úložiště. Změnou jediného znaku v souboru je znovu celý ukládán.
- Datový blok[11] s konstatní velikostí
  - ▷ Mnohem efektivnější ukládání dat.
  - ▷ Velkou velikostí datového bloku vzniká nebezpečí neefiktivního ukládání dat stejně jako na úrovni souboru.
  - ▷ Při malé velikosti datového bloku se přiliš navyšují data nutná k obnově souborů.
- Datový blok s proměnnou velikostí
  - ▷ Nejúčinější alternativa efektivního uložení dat.
  - ▷ Soubory jsou dělené na co nejmenší části a tím snížena možnost duplicity informací.
  - ▷ Nevýhodou je vyšší výpočetní výkon.
  - ▷ EMC Avamar a EMC Data Domain.

[11] Část zálohovaného souboru



Obrázek 6: Fixní a variabilní délka datových bloků [7]

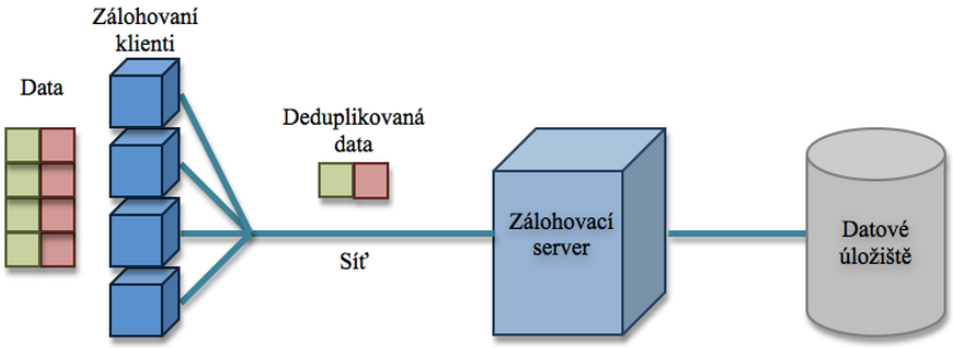
Účinnost deduplikace s proměnnou délkou bloků je ukázána na příkladu řešení Data Domain. Tyto systémy používá 3 500 zákazníků na celém světě. Jejich celková úložná kapacita je přibližně dva exabajty. Skutečně změřený deduplikační poměr u všech zákazníků je 1 : 13,4, což znamená 92,5% úsporu úložné kapacity. Na 100 TB ukládaných dat je tedy zapotřebí pouze 7,5 TB skutečné kapacity.[8]

Metody ukládání dat

- Na zdroji

  Systém přenese kompletní data jen jednou při první záloze. Poté probíhají plné zálohy dál, avšak technikou deduplikace se přenášejí pouze změny v souborech.

- ▷ Vyřadí duplicitní data ještě před uložením na cílový stroj.
- ▷ Rychlost datové linky může být 10-20krát menší oproti ukládání v cíli.
- ▷ Snižuje dobu zálohy.
- ▷ Více početních operací k nalezení duplicit zatěžuje hardware[12] zdroje.
- ▷ EMC Avamar.



Obrázek 7: Deduplikace na zdroji [7]

Deduplikace na zdroji je výhodná při potřebě centrálního zálohování vzdálených lokalit, ideální pro rozsáhlé virtuální prostředí nebo ochranu dat koncových zařízení.

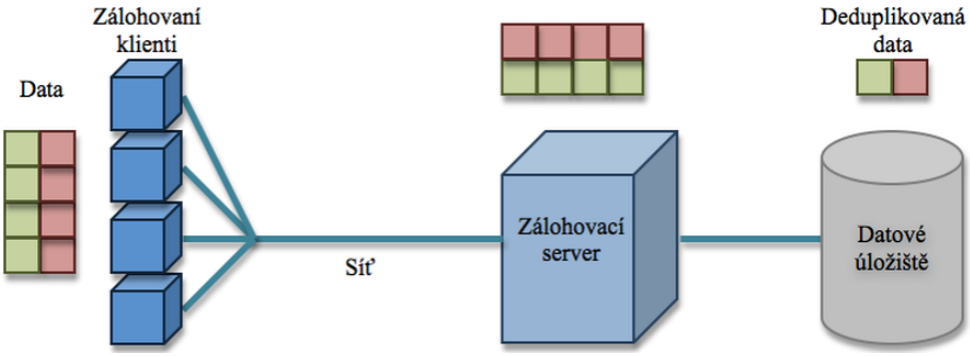
- Na cílovém úložišti

  Jsou přenášena kompletní data včetně duplicit. Teprve při jejich ukládání odstraňuje shodná data.

  - ▷ Prodlužuje čas zálohy.
  - ▷ Vyžaduje rychlejší datovou linku.

[12] Technické vybavení počítače

- Šetří diskovou či páskovou kapacitu úložiště.
- Lze využít na běžících systémech, například s Microsoft Windows Server 2012[13].
- EMC Data Domain.



Obrázek 8: Deduplikace na cíli [7]

Kvalita deduplikace závisí na množství totožných bloků v souborech a implementaci vhodně zvolené metody v daném prostředí. Deduplikují-li se pravidělně opakující se data, navíc s dlouhým retenčním obdobím[14], je možné dosáhnout kompresního poměru až 95%. Významným prvek snižujícím velikost komprese jsou samozřejmě také typy ukládaných souborů.

Efektivita řešení závisí na tom, kolik záloh je na úložišti udržováno. Samozřejmě, čím více záloh téhož systému je, tím je pravděpodobnější, že deduplikační systém uvidí více duplicitních dat. Provozovatelé často uvádějí poměr 20:1, ale to v případě, že je retence 16 týdnů, což je stáří nejstarší zálohy. V případě dvou týdnů to již může být jen 4:1.[6]

[13]Serverový operační systém

[14]Doba, po kterou je záloha uchovávána. Předpokladem je ukládání více záloh po určité období

Problematické zálohování souborů z pohledu deduplikace

- Databáze
  - ▷ Největším problémem nejen z hlediska deduplikace, ale samotné realizace zálohy, jsou běžící databáze provádějící neustálé změny.
- Multimediální soubory
  - ▷ Fotografie, obrázky.
- Komprimované soubory
  - ▷ Komprimační formáty ZIP, RAR a další.

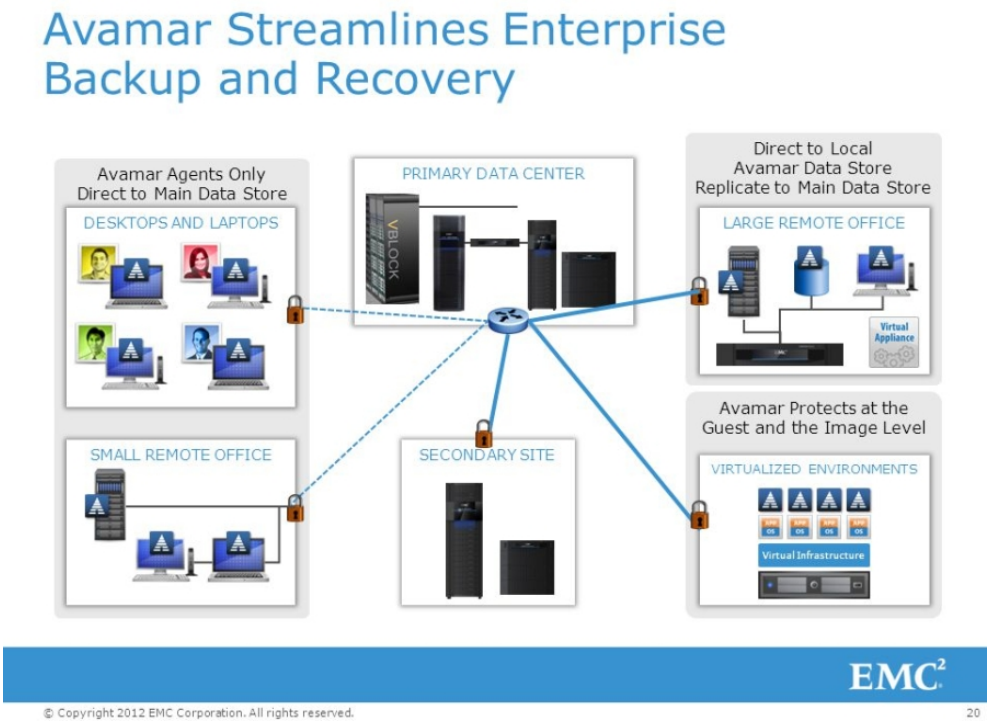
  Výše uvedené formáty jsou již komprimované a jejich rozdělení do bloků dle konceptu deduplikace je minimální.
- Kryptovaná data
  - ▷ Šifrovací nástroje[15].
- Virtualizované prostředí[16]
  - ▷ Virtuální servery, stanice.
  - ▷ WMvare, SUN.

Díky souborovému systému ZFS[17] využívající se u některých NAS[18] přichází s touto metodou do kontaktu i běžný koncový uživatel.

[15]Zpravidla aplikace sloužící pro ochranu dat uživatele.
[16]Speciální systém umožňující na jednom hardwaru provozovat více operačních systémů.
[17]Zettabyte File System
[18]Network Attached Storage

2.7 Avamar

Avamar je produktem společnosti EMC, která se v roce 2016 stala součástí koncernu Dell. Produkt nabízí hardwarové a softwarové řešení určené pro zálohování dat technologií deduplikace.



Obrázek 9: EMC Avamar [9]

EMC Avamar Data Store kombinuje certifikovaný EMC hardware se svým zálohovacím softwarem *Avamar deduplication backup and recovery*. Výhodou je eliminace problémů a chyb, které by mohli nastat implementací na nevhodný hardware nebo špatnou konfigurací softwarové části. Navíc, ve svém certfikovaném hardware, propaguje patentovou technologii RAIN[19], která ověřuje integritu diskových svazků a vysokou dostupnost mezi uzly (nodes) datového úložiště.

[19]Redundant Array of Independent nodes

Zálohovaná data dělí na datové bloky variabilní délky (segmenty), komprimuje a poté každému bloku přidělí unikátní jednoznačný identifikátor (hash). Pokračuje porovnáním jednotlivých bloků s těmi již zálohovanými, hledá nová data, která ukládá. Nikdy nezálohuje shodná data přes všechny zálohované stanice. Je jedno, zda byly uloženy například na nějakém serveru, potom na stanici. Avamar porovná všechna tato data, která kdy zálohoval a duplicity vynechá. Unikátní technologií deduplikace dosahuje velkých úspor místa datových úložišť, rychlého přenosu ukládaných dat i doby obnovy.

EMC Avamar je možné nasadit samostatně nebo jako rozšíření EMC NetWorker. V tomto případě slouží EMC Avamar jako tzv. deduplication node.[10]

Přednosti

- Redukce síťového provozu až o 99%.
    - ▷ Úspora datového úložiště až 30krát.
    - ▷ Doba zálohy zkrácena až o 50%.
    - ▷ V porovnání s tradičními systémy je doba obnovy až 30krát rychlejší.

Aktuální řada produktů EMC Avamar

- EMC Avamar
    - ▷ Součástí produktu Data Protection Suite Family.
    - ▷ Denní zálohování a jednoduchá obnova dat.
    - ▷ Dokáže zálohovat až 1000 klientů.
    - ▷ Ochrana dat a obnova po havárii.

- Zálohování
  - Virtuálního a fyzickéhé prostředí.
  - Klientské stanice a NAS úložiště.
  - Vzdálené lokality (remote offices).
  - Aplikace.
- Podpora zálohy databází IBM DB2, IBM Domino, MS Exchange Server, MS SQL Server, Oracle, SAP, Sybase.
- Kompatibilita s operačními systémy MS Windows, Apple OS X, Cent OS, Debian, FreeBSD, HP-UX, IBM AIX, Iomega, Linux, Novell Netware, Oracle Linux, Red Hat Linux, SUSE Linux, Ubuntu, SCO Unix, Sun Solaris.
- Systémy:
  - Avamar Deduplication Backup System.
    - Úložiště až 124 TB.
    - Deduplikace na zdroji.
  - Data Domain Deduplication Backup System.
    - Diskové pole až 570 TB.
    - Deduplikace na cíli.
- Hardwarová platforma:
  - EMC Data Store.
    - Diskové pole s technologií RAIN[20].
  - EMC Single Store
    - Zařízení s velkou diskovou kapacitou, ochrana diskového pole pomocí RAID[21].

[20]Redundant Array of Independent nodes
[21]Redundant Array of Inexpensive/Independent Disks

- EMC Avamar Virtual Edition.
  - ▷ Centrální zálohování virtuálních serverů i uživatelských stanic.
  - ▷ Ochrana dat a obnova po havárii.
  - ▷ Nabízí deduplikovaná úložiště od 0,5 TB do 4 TB.
  - ▷ Zálohování dat vnitřní sítě (LAN[22]), vzdálené lokality i cloudová úložiště.
  - ▷ Kompatibilní se systémy VMware vSphere, Microsoft Hyper-V, Azure[23].
- EMC Avamar Business Edition.
  - ▷ Denní zálohování včetně obnovy dat.
  - ▷ Ochrana dat a obnova po havárii.
  - ▷ Nabídka deduplikovaných úložišť 3,9 TB a 7,8 TB.
  - ▷ Podpora zálohy databází IBM DB2, IBM Domino, MS Exchange Server, MS SQL Server, Oracle, SAP, Sybase, Microsoft SharePoint.
  - ▷ Vhodné pro středně velké společnosti.
  - ▷ Hardwarová platforma
    - EMC Data Store.

2.7.1 Redundant Array of Independent nodes (RAIN)

Ztráta dat poškozeným HW[24] nebo SW[25] je skutečným problémem každé společnosti. RAIN technologie vznikla v California Institute of Technology

[22]Local Area Network
[23]Serverové systémy umožňující virtualizaci
[24]Hardware
[25]Software

(Caltech) ve spolupráci s NASA Jet Propulsion Laboratory a Defense Advanced Research Projects Agency (DARPA).

Pro ochranu dat v jednom diskovém poli (node) je využívána technologie RAID[26]. Ta v závislosti na její konfiguraci zajišťuje ochranu dat, například v případě havárie některého z datových disků pole. Data jsou rovnoměrně ukládána na všechna datová úložiště v nodu a v případě výpadku jednoho z nich jsou pak na nový zpět nakopírovaná z těch ostatních.

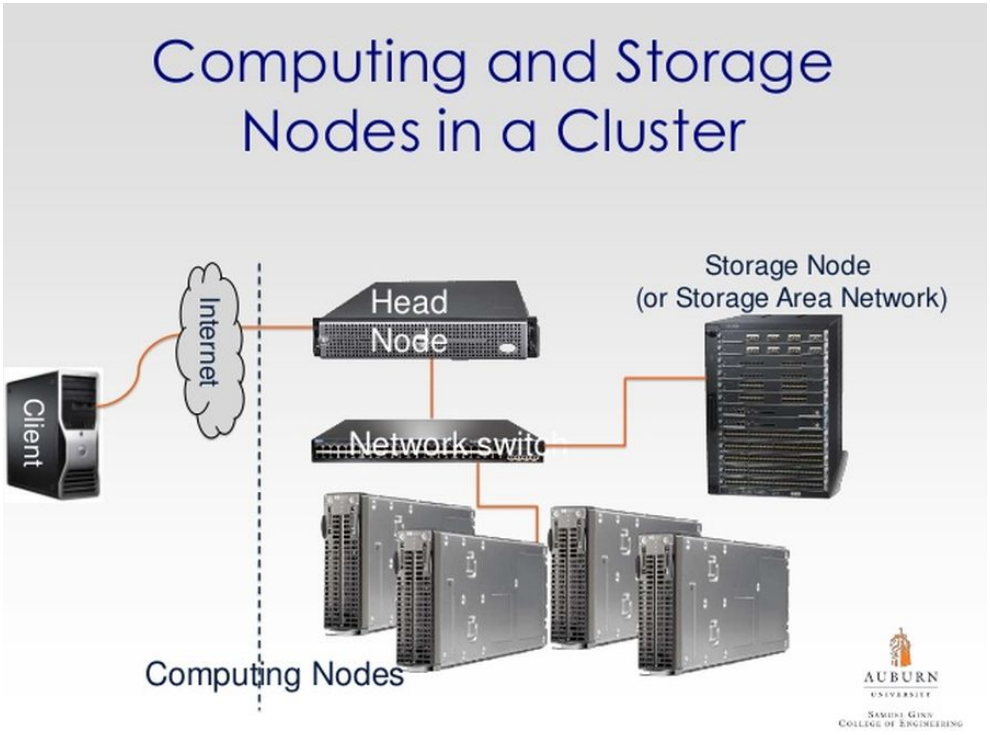
Nejběžnější typy RAID

- RAID 0
  - ▷ Ztráta jednoho disku znamená ztrátu všech dat.
  - ▷ Nejrychlejší z RAID, kapacita je rovna počtu disků v poli.
- RAID 1
  - ▷ Data se zrcadlí na všechny disky.
  - ▷ Rychlé čtení dat, avšak pomalý zápis.
  - ▷ Celková kapacita je kapacitou nejmenšího disku v poli.
  - ▷ Nesmí dojít ke ztrátě více než jednoho disku.
- RAID 5
  - ▷ Data a parita jsou rozmisťované na všechna datová úložiště v poli.
  - ▷ Minimálně 3 disky v poli.
  - ▷ Rychlé čtení, pomalejší zápis.
  - ▷ Poškození při selhání jednoho a více disků.

[26]Redundant Array of Inexpensive/Independent Disks

- RAID 6
  - ▷ Dva paritní bloky.
  - ▷ Ukládání informací totožné s RAID 5.
  - ▷ Minimálně 4 disky v poli.
  - ▷ Ztráta dat při poškození více jak dvou disků.
- RAID 10
  - ▷ Dvě RAID 1 pole implementované v RAID 0.
  - ▷ Dobrý výkon a efektivní ochrana dat.

Technologie RAIN je však ještě inovativnější. Zajišťuje spolehlivou a rychlou komunikaci mezi uzly (nodes). Zároveň disponuje systémem potvrzení, odpovídajícího za spolehlivé doručení paketů. K tomu využívá všechny dostupné síťové prostředky. Nestane-li se tak, upozorní horní vrstvu a funguje jako detektor chyby. Využívá virtuálních IP adres přiřazených clusteru, které jsou použité při potřebě sdílení zátěže nebo z poškozeného uzlu na jiném funkčním. Je kompatibilní s různými operačními systémy, počítačovými platformamy a síťovým prostředím.



Obrázek 10: Storage nodes [11]

Výhody

- Vysoká dostupnost a odolnost proti chybám.
- Různé operační systémy mezi uzly.
- Efektivní využití datových cest a odolnost proti chybám v síti.
  - ▷ Vyřazením síťového switche nedojde k výpadku.
- Optimální ukládání dat přes více datových disků.

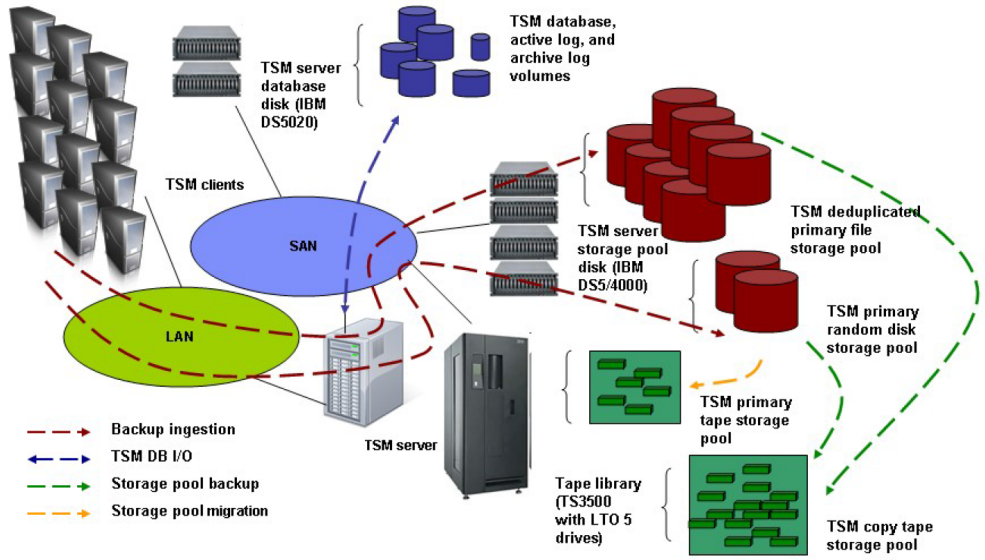
Kvalitní záloha dat je pro firmy jedním z klíčových bezpečnostních prvků. Zálohování je nutné pečlivě plánovat a prostředky pro jeho správu je nezbytné propojit s ostatními systémy. Vždycky jde o řešení, které kombinuje

software a hardware. Dobré řešení navíc umožní výměnu hardwaru za jiný bez toho, aby bylo nutné měnit softwarovou část – v takovém případě se přechod na modernější technologie (například takové, které využívají deduplikaci) obejde bez zbytečných nákladů, například na přeškolení obsluhy.[12]

2.8 IBM Tivoli Storage Manager

Společností IBM je pod značkou Tivoli nabízeno mnoho softwarových produktů. Tivoli Storage Manager (TSM) je modulární systém zaměřený na zálohování, správu a ochranu dat středních a menších organizací.

Většími společnostmi bývá využita robusnější aplikace Tivoli Storage Manager Extended Edition s rozšířenou správou páskových a diskových úložišť.



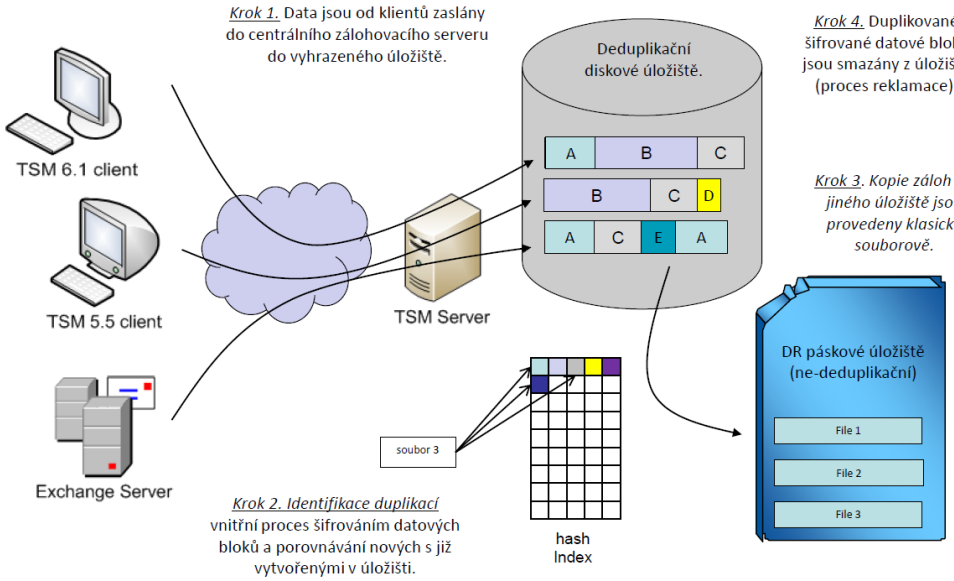
Obrázek 11: Architektura IBM Tivoli Storage Manager [13]

Aplikace je plně kompatibilní s IBM AIX provozovaným na platformě IBM Power Systems. Operační systém IBM AIX je UNIXOVÝM systémem[27] běžícím na kompatibilních serverech IBM.

[27]Operační systém

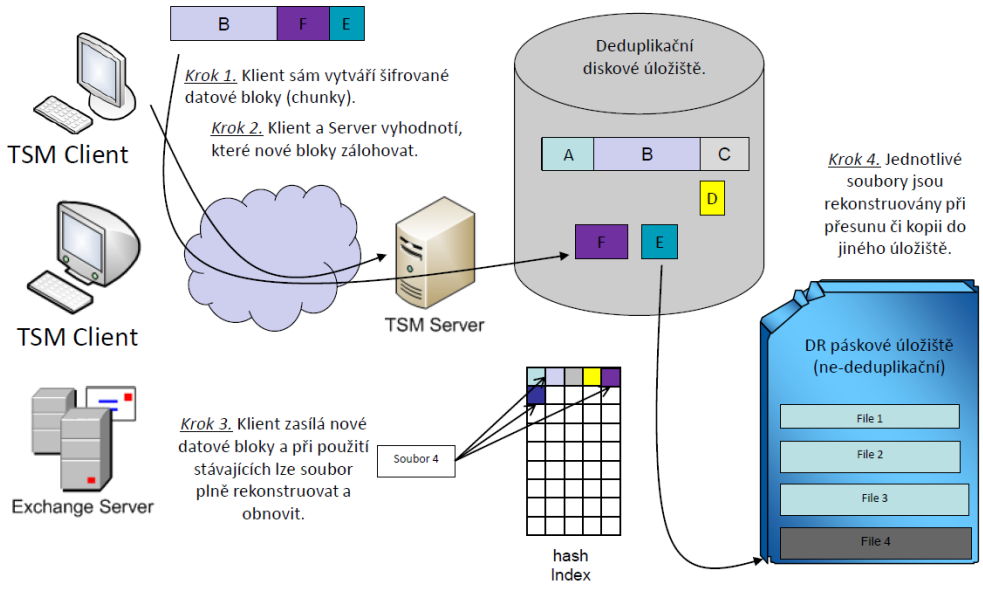
IBM Tivoli software je platforma pro správu podniku se specializovanými komponentami přizpůsobenými IT administrátorům spravujícím středně velká datová centra společnosti.[?][28]

Data jsou přenášena přes TSM server do deduplikačního úložiště. Kopie záloh ukládaných například na páskové mechaniky jsou prováděné souborově bez deduplikace. Lze využít deduplikační techniky na zdroji i cílovém úložišti.



Obrázek 12: Deduplikace na serveru [14]

[28]IBM Tivoli software is an enterprise system management platform with specialized components customized for IT administrators that manage midsize and enterprise data centers.



Obrázek 13: Deduplikace na zdroji [14]

Kompatibilita

- MS Windows verze 2000 a vyšší.
- Red Hat Linux, Suse a dalšími.
- IBM AIX, HP UX, Sun Solaris.
- Aplikace SAP, MS Exchange a MS SharePoint.
- Platformy Macintosh, Citrix nebo Novell NetWare.
- Databáze Oracle, IBM DB2, Informix i MS SQL.
- Podpora virtualizačních prostředí VMware a Hyper-V.

Licencování PVU[29]

- Dle procesorových jader u fyzických serverů.
- Podle množství jader přiřazených virtuálním zařizením.
  - ▷ Nutná manuální evidence nebo reporting nástrojem IBM ILMT.[30]

Příklady rozšíření (agenti pro TSM)

- TSM for Databases.
- TSM for Virtual Environments.
- TSM for Mail.
- TSM for SAN.
- TSM for Space Management.
- Tivoli Storage Fastback.
- Tivoli Storage Fastback for Exchange.

Přednosti

- Dodáváno s datovými úložišti IBM data storage.
- Zalohování otevřených souborů a databází.
- Integrace deduplikace na zdroji nebo cíli.
- Ochrana aplikací i databází a archivace dat.
- Podrobný reporting a monitorování procesů.

[29]Processor Value Unit
[30]License Metric Tool

- Jednoduchá administrace zálohovacích úloh.
- Vysoká kompatibilita s ostatními systémy.

Řešením IBM Tivoli Storage Manager lze rapidně snížit rizika ztráty nebo poškození. Umožňuje rychlou obnovu na úrovni souborů i operačních systémů. Škálovatelnost navíc zajišťuje vysokou kompatibilitu a rychlost požadovaných operací.

2.9 Symantec NetBackup

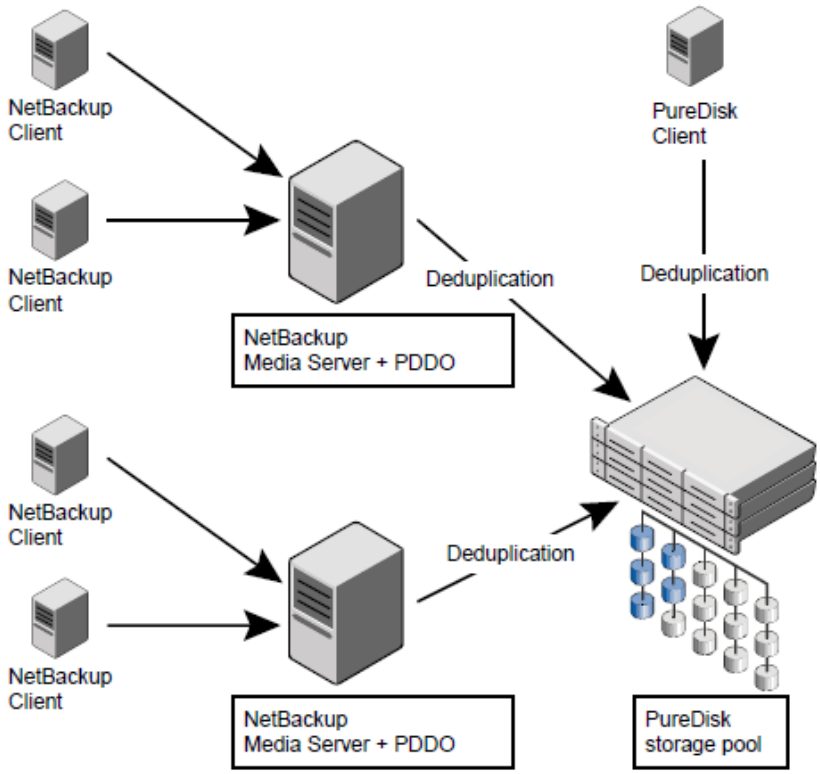
Kompletní systém ochrany dat s plně integrovaným deduplikačním softwarem PureDisk. Pomocí OpenStorage API rozhranní PDDO[31] probíhá zálohování do PureDisk datového úložiště, kde dochází k deduplikaci.

Obecně platí, že deduplikace uchovává pouze unikátní odkaz na zálohovaná data v datovém úložišti. [15][32]

K zálohování jsou využívány dva typy klientů. NetBackup server využívá svých vlastních, data se deduplikují a jsou ukládána do PureDisk serveru jako záložního úložiště. Další variantou je přímé využití PureDisk klientů a zálohování provádějící se rovnou do PureDisk serveru.

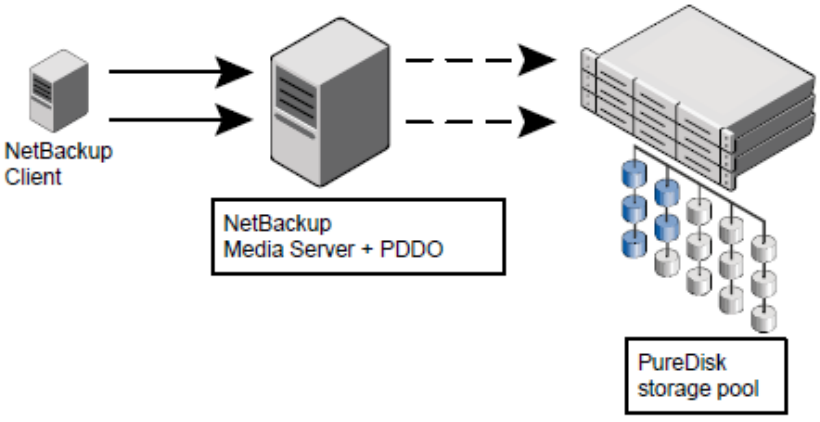
[31] PureDisk Deduplication Option

[32] In general, deduplication is a method of retaining only one unique instance of backup data on storage media.



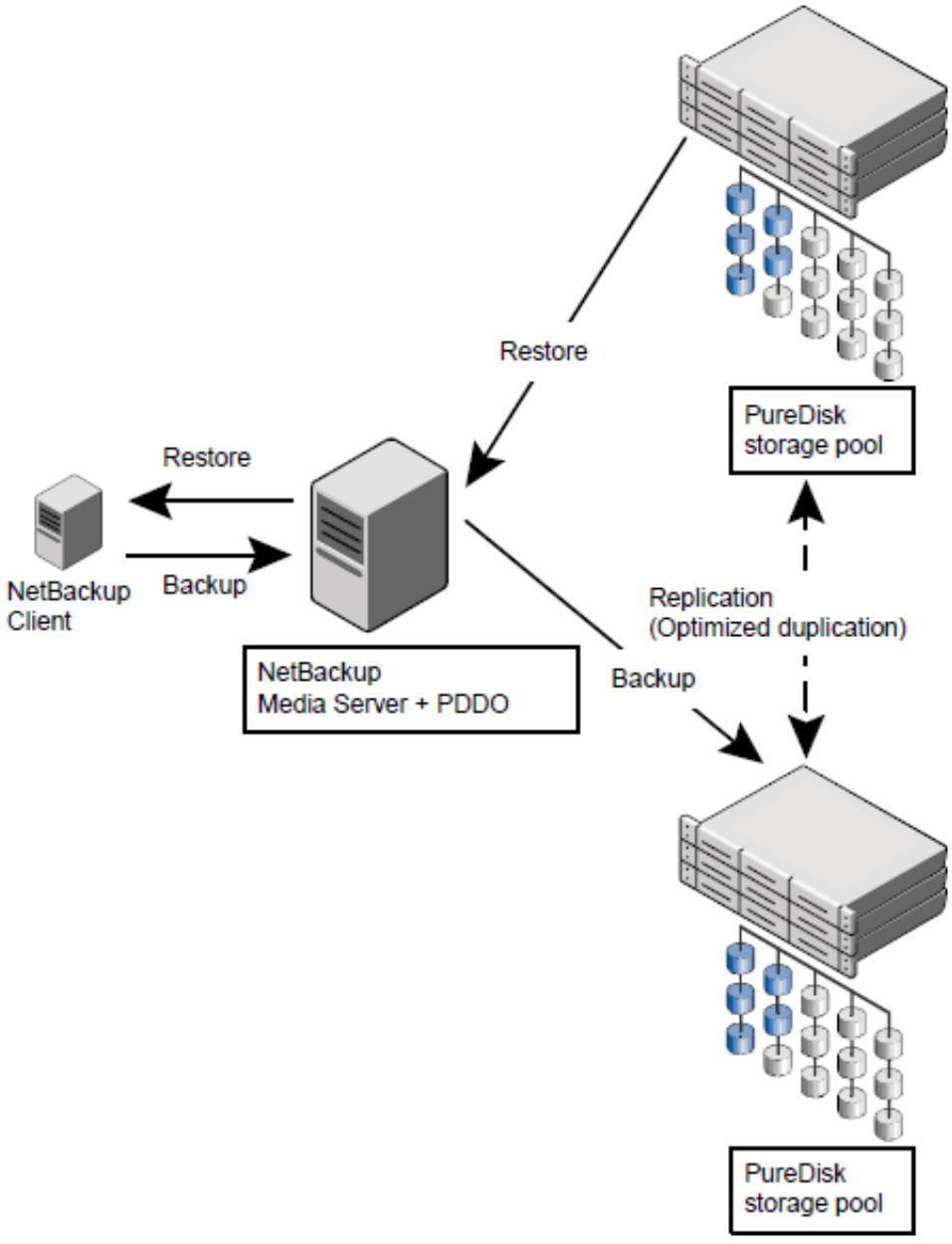
Obrázek 14: Deduplikace klientů [15]

NetBackup politikou je také možné klientské informace v serveru pouze zpracovávat, avšak deduplikovat a ukládat jen v PureDisk serveru.



Obrázek 15: Ukládání do PureDisk storage [15]

PDDO[33] umožňuje i vzájemné replikace mezi servery. Data mohou být současně uložena v několika PureDisk úložištích a z libovolného poté obnovena.



Obrázek 16: Replikace dat [15]

[33]PureDisk Deduplication Option

Kompatibilita

- MS Windows verze 7 a vyšší včetně Server edice.
- Virtuální prostředí VMware a MS Hyper-V.
- Operační systémy UBuntu, Linux, Cent OS, FreeBSD nebo Mac OS X.
- Databáze a aplikace IBM DB2, IBM Informix, IBM Lotus Domino, Microsoft Active Directory, Microsoft Exchange Server, Microsoft Office SharePoint Server, Microsoft SQL Server, Oracle, SAP HANA, SAP R/3, SAP Sybase a další.
- Kompatibilita s hardwarem EMC, HP, Hitachi, IBM, NetApp.

Způsoby licencování

- Tradiční
  - ▷ Licence podle počtu klientů a serverů.
- Dle kapacity [TB][34]
  - ▷ Za množství přenesených dat a počet klientů.

Velmi rozsáhlá licenční politika a s tím související množství vynaložených finančních prostředků je ovlivněno i konkrétní edicí NetBackup.

Přednosti

- Dle zvolené varianty deduplikuje data na zdroji nebo cíli.
- Dodáváno včetně kompatibilního hardware.

[34]Terabytes

- Replikace mezi oběma typy serverů.
- Ochrana dat technologií RAID6 a RAID1 pro systém serveru.
- Kompatibilita.
- Intuitivní uživatelské rozhraní.

Symantec NetBackup je systémem nabízejícím široké možnosti rozšíření, vysokou kompatibilitu a centrální ochranu dat.

3 Praktická část

Praktická část bakalářské práce demonstruje návrh zálohování mezi vzdálenými lokalitami systémem EMC Avamar. Bylo vytvořeno několik zálohovacích úloh s různými typy nejběžněji používaných souborů a zjištění rozdílu v množství přenesených dat a době potřebné pro jejich zálohu. Projekt zjednodušeně vysvětluje základní principy a postupy pro realizaci úlohy zálohování a její automatizaci v tomto software. Odborné veřejnosti může sloužit jako stručný manuál k základním operacím.

3.1 Představení projektu

V projektu bylo vytvořeno pět záloh rozdělených podle typu souborů následovně: první záloha soubory **.doc* a **.xls*, druhá záloha **.mdb* a **.dbf*, třetí záloha **.dwg*, čtvrtá záloha **.pdf*, poslední pátá záloha všech souborů dohromady. Simulace zálohování byla provedena mezi městy České Budějovice a Český Krumlov po datové lince 8 Mbit/s. Datový spoj mezi pobočkami využívá technologie MPLS[35]. Deduplikace dat probíhá na zdrojovém zařízení. Demonstrace zálohování se realizovala v prostředí stavební firmy s požadavkem na uchování velkého množství uvedených testovaných souborů.

3.2 Technické požadavky

Projekt využil systému Avamar 7.3 dodávaný včetně hardwaru Avamar Gen4. Jednalo se o kompletní řešení, takzvaný blackbox. Zdrojová data byla uložena na doménovém serveru Dell PowerEdge R520 s operačním systémem Microsoft Windows server 2008 R2. Webová administrace nástrojem EMC Avamar Web Restore se neobešla bez práce v internetovém prohlížeči, v našem pří-

[35]Multiprotocol Label Swapping

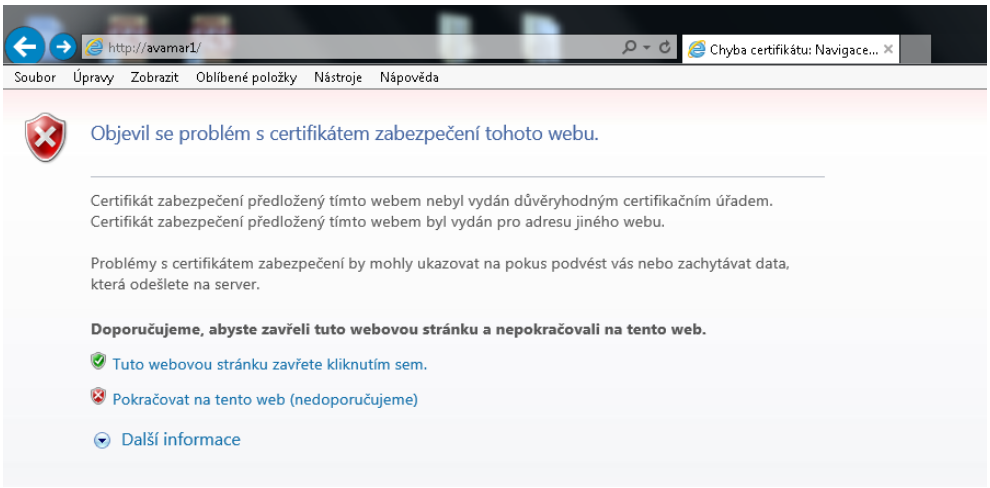
padě Internet Explorer 11 CZ, který běžel na operačním systému Microsoft Windows 7 Professional CZ 32bit a SW Java verze 7 nebo 8. Pod tímto operačním systémem bylo také spouštěno plnohodnotné administrační rozhraní Avamar Administrator 7.3.

3.3 Popis a ukázka praktického návrhu

3.3.1 Administrační rozhraní

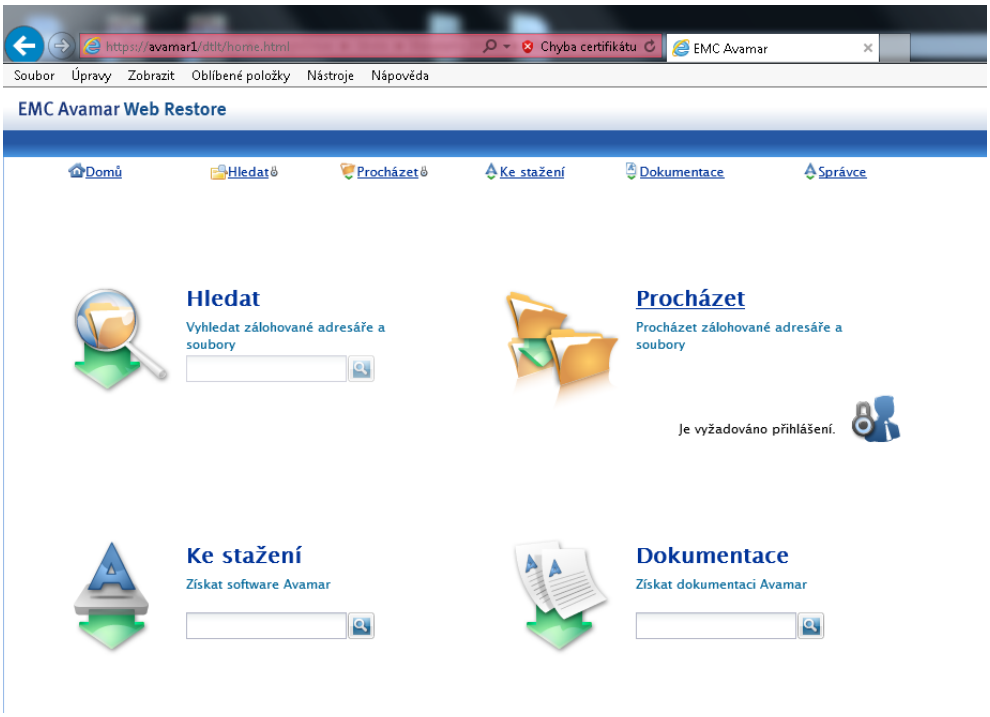
Aby bylo vůbec možné pracovat se zálohovacím systémem, muselo nejdříve dojít ke spuštění webového administračního prostředí EMC Avamar Web Restore.

Do webového prohlížeče se vepsal název nebo IP[36] adresa našeho Avamar serveru. Prohlížeč oznámí chybu certifikátu, kterou je nutno ignorovat a potvrdit volbu *Pokračovat na tento web*.



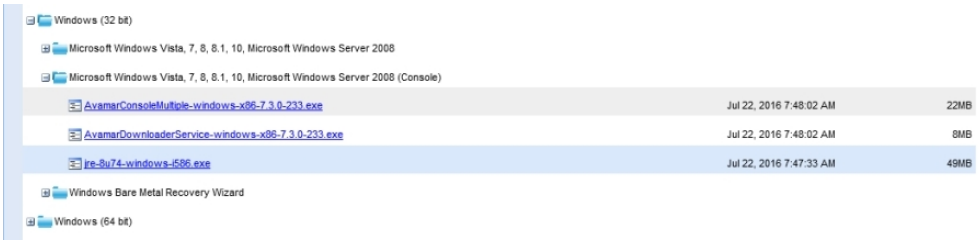
Obrázek 17: Chyba certifikátu přihlášení k webovému rozhraní

[36]Internet Protocol



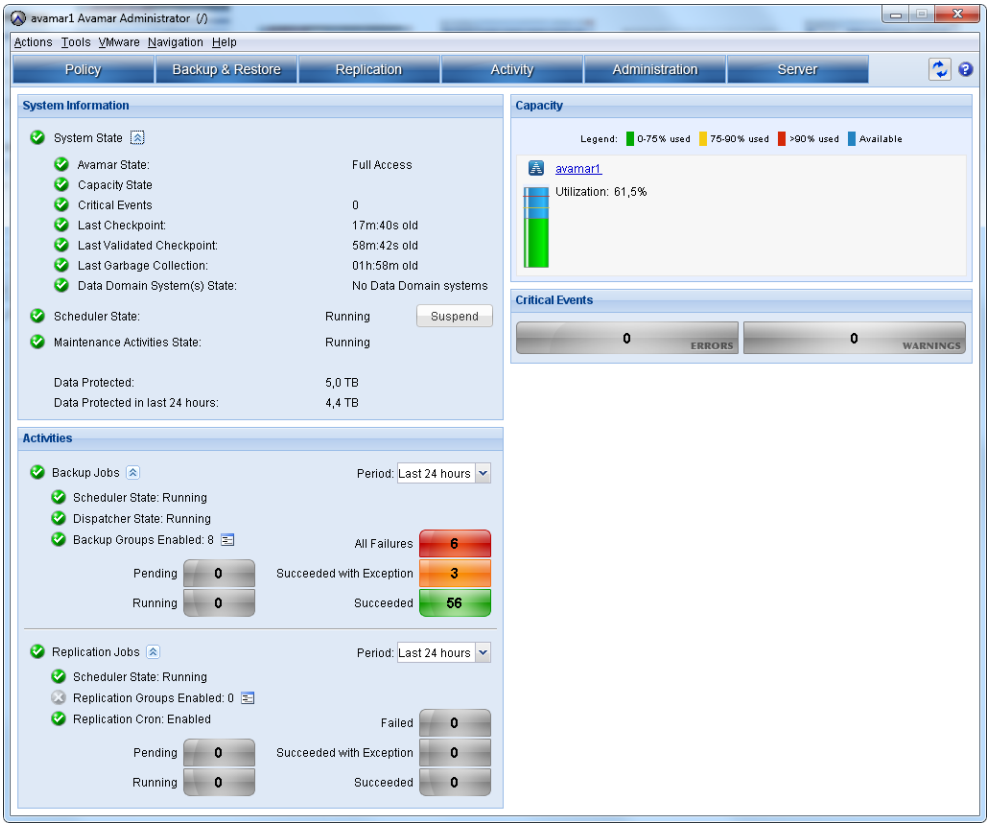
Obrázek 18: Webové rozhraní EMC Avamar Web Restore

V části *Ke stažení* je webovým rozhraním nabízen software Avamar administrator, který bude použitý pro konfiguraci systému.



Obrázek 19: Stažení aplikace Avamar Administrator

Software Avamar Administrator byl v projektu využit jako primarní rozhraní pro konfiguraci politiky zálohování, časových plánů, monitoringu, reportingu a dalších systémových funkcí.



Obrázek 20: Avamar Administrator

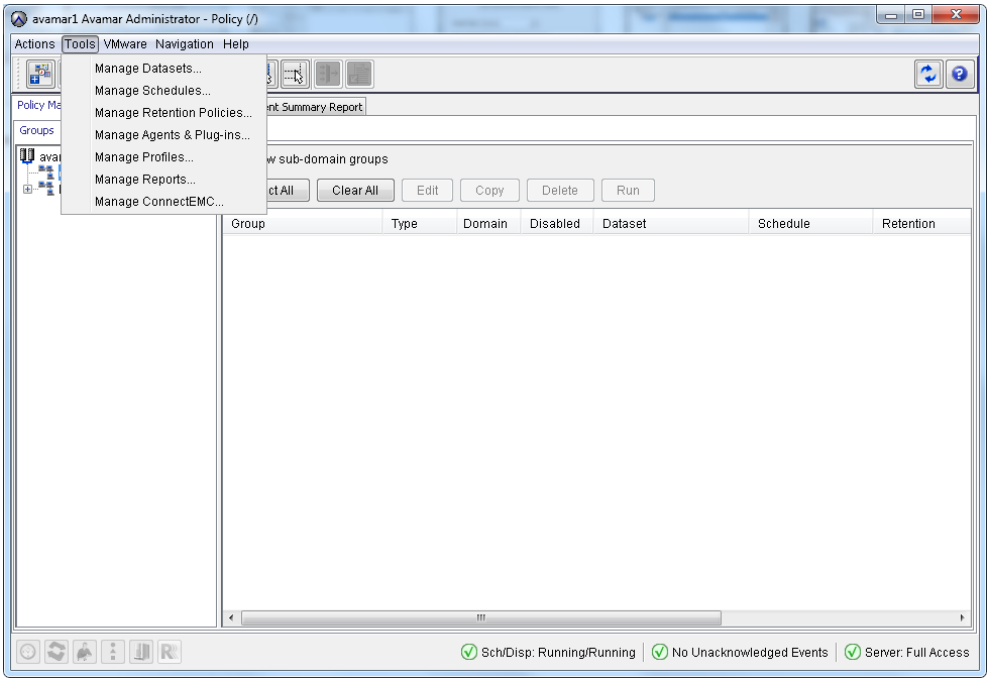
3.3.2 Automatický proces zálohování

Pro demostraci deduplikace různých dat postačuje provést jednorázovou zálohu (On Demand Touch), avšak zálohovací procesy by měly být automatizované, aby nezávisely na času a pracovní době IT personálu. V tomto projektu bude ukázáno základní nastavení automatického spouštění zálohovacích úloh.

Je vyžadována konfigurace následujících komponent

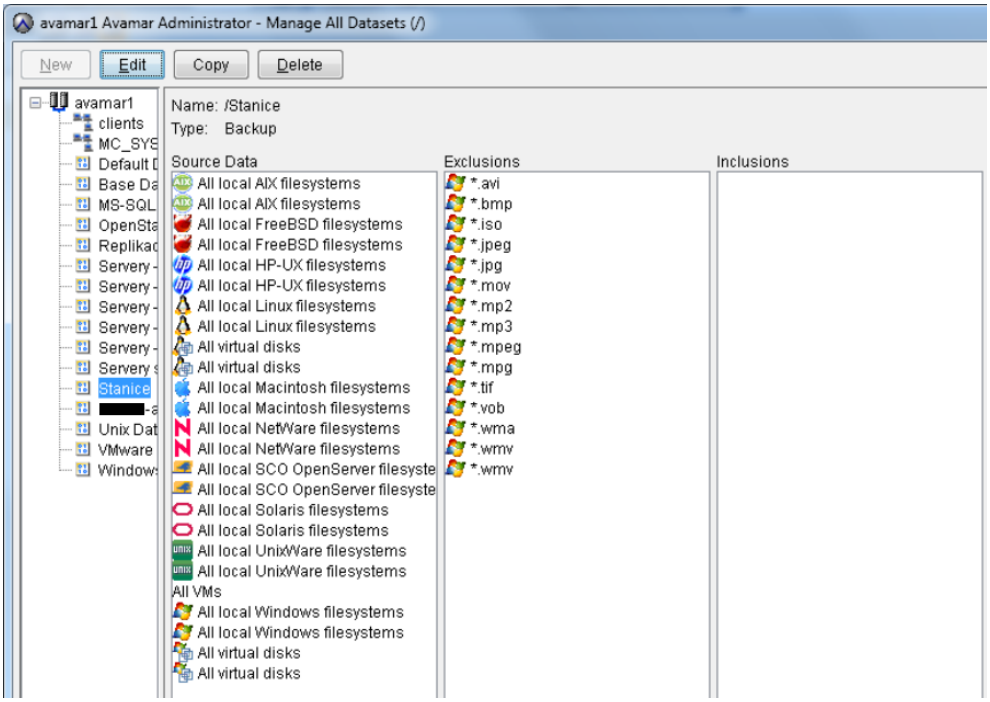
- *Datasets*
- *Schedules*
- *Groups*

V nástroji *Datasets* bylo vytvořeno několik skupin (*Groups*). Skupiny lze zjednodušeně považovat za filtry, definující co se má nebo naopak nemá zálohovat. Je velmi výhodné zvolit různé parametry pro zálohování serverů a uživatelských stanic nebo datových uložišť.



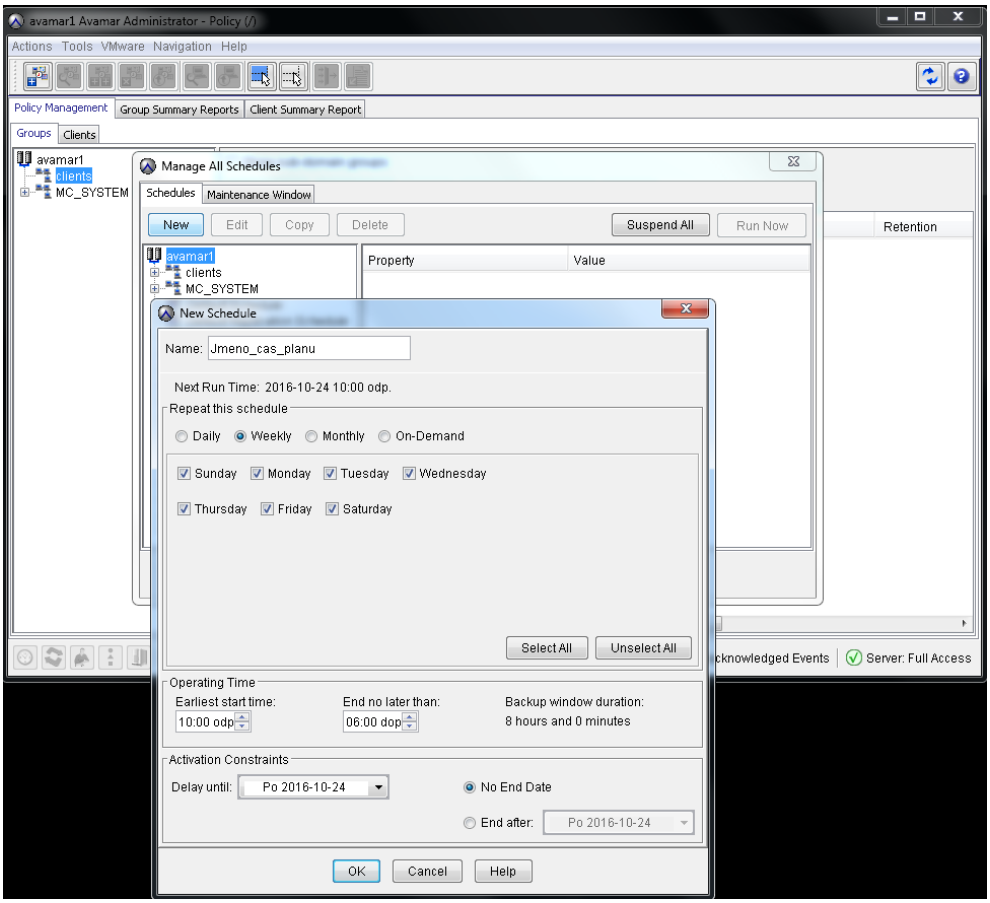
Obrázek 21: Datasets

V detailu Datasetu byla nastavena skupina *Stanice* pro zálohování uživatelských dat, filtrem se omezily typy nežádoucích souborů.



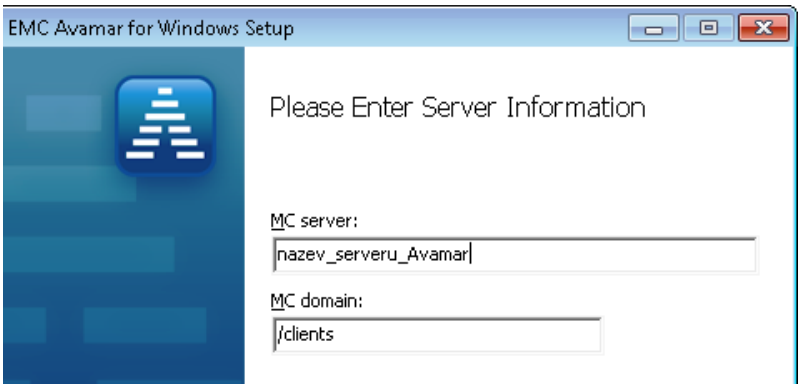
Obrázek 22: Datasets detail

Po vytvoření zálohovacích skupin se nastavil plánovaný čas spuštění - *Schedules*. V tomto projektu se jedná o každodenní dopolední zálohu. Nemělo by dojít ke kolizi s ostatními procesy zařízení, například údržbě databází nebo systému.



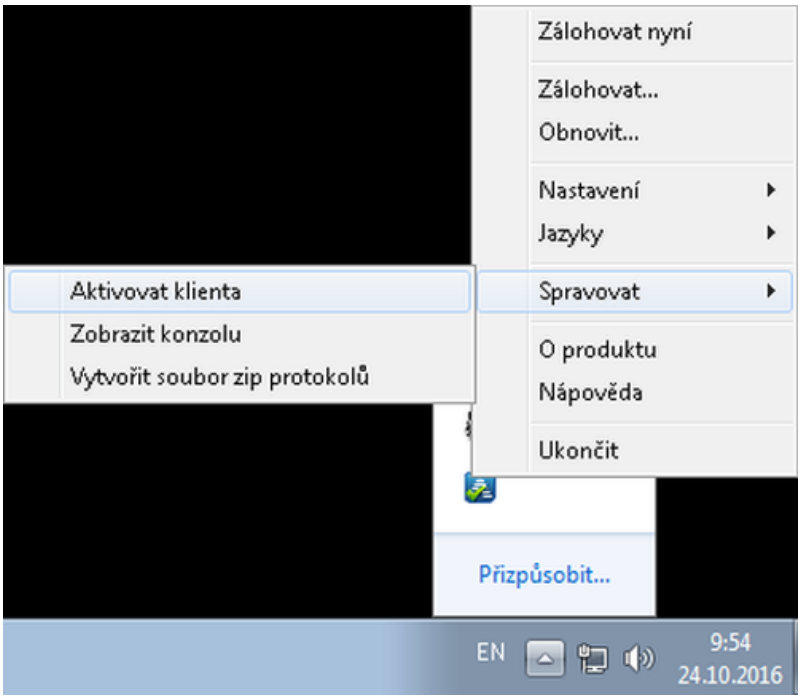
Obrázek 23: Schedules

V této fázi je možné nainstalovat a autorizovat klienta, to znamená zdrojové zařízení s daty k záloze. Instalační dávka je také uložena v části *Ke stažení* webové aplikace EMC Avamar Web Restore.

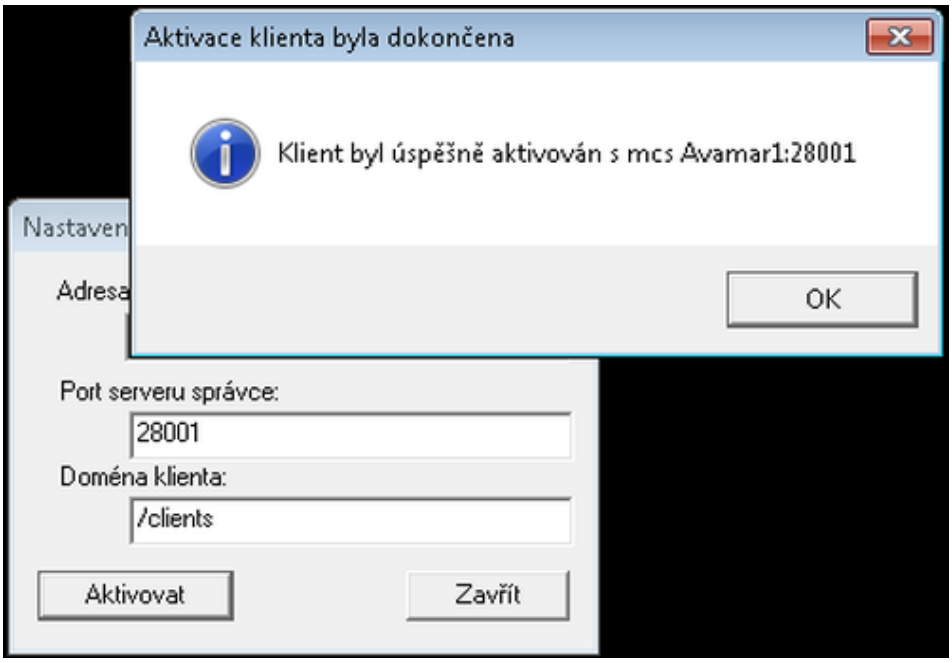


Obrázek 24: Instalace klienta Avamar

Posledním krokem klientské části je aktivace zařízení. Bez aktivace nebude zařízení Avamarem jednoznačně identifikované a zálohy nelze provádět.

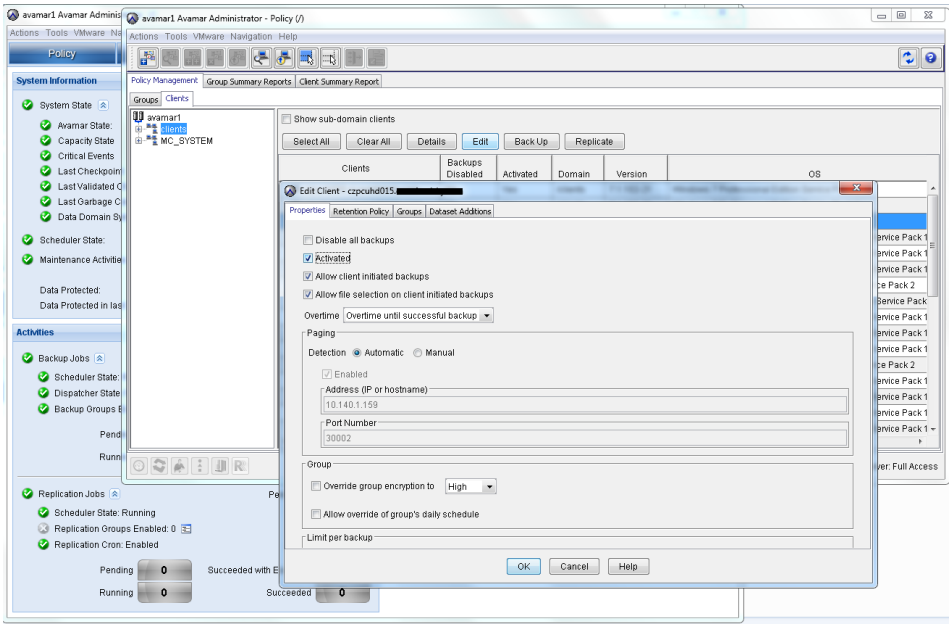


Obrázek 25: Aktivace klienta Avamar



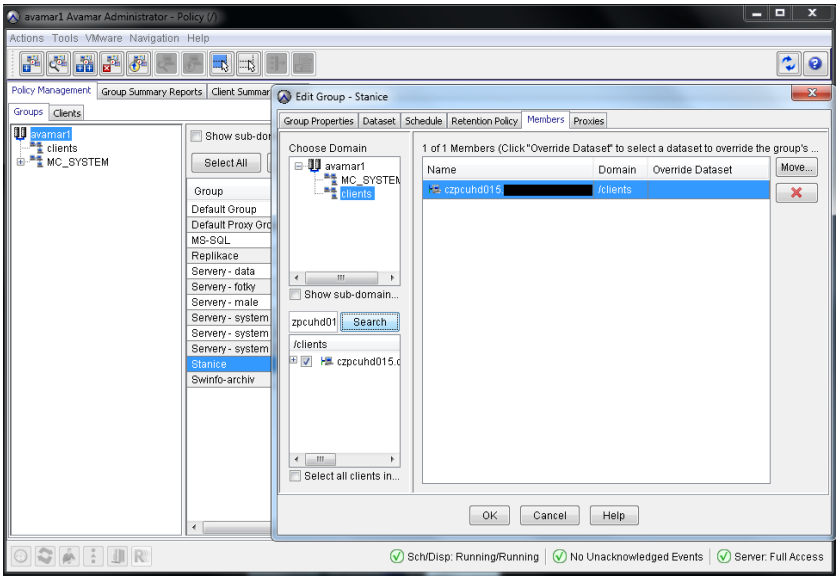
Obrázek 26: Úspěšná aktivace

V administraci Avamaru se ověří aktivaci a nastavení klienta.

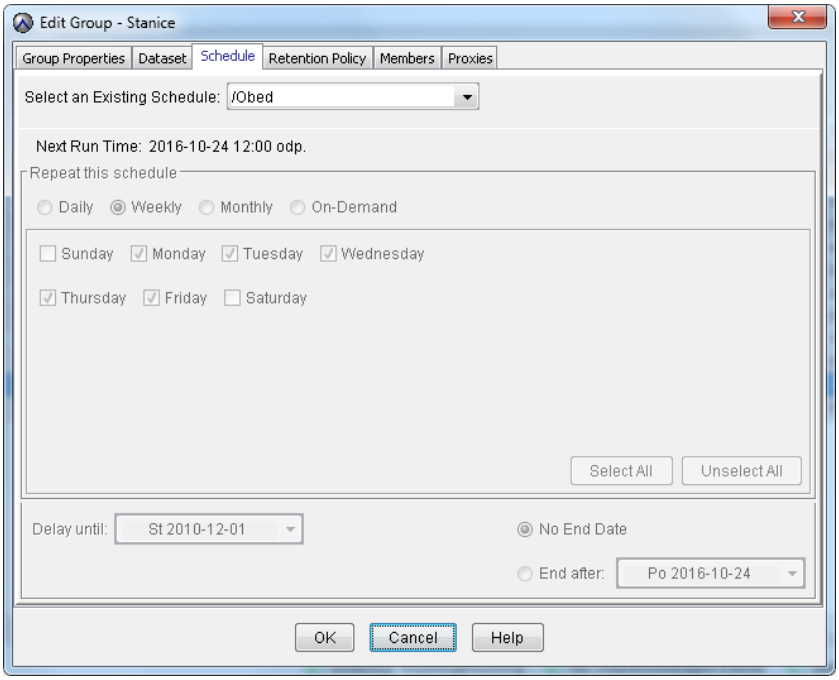


Obrázek 27: Správa klienta v Avamar Administrator

Jedním z posledních kroků je přiřazení klienta do předem vytvořené skupiny (*Groups*) v *Datasets* a volba odpovídajícího časového plánu v *Schedules*.



Obrázek 28: Přiřazení odpovídající zálohovací skupiny klientovi



Obrázek 29: Nastavení časového plánu zálohování

Těmito operacemi se provedla základní konfigurace důležitá pro automatizování zálohovacích úloh.

3.3.3 Proces deduplikace

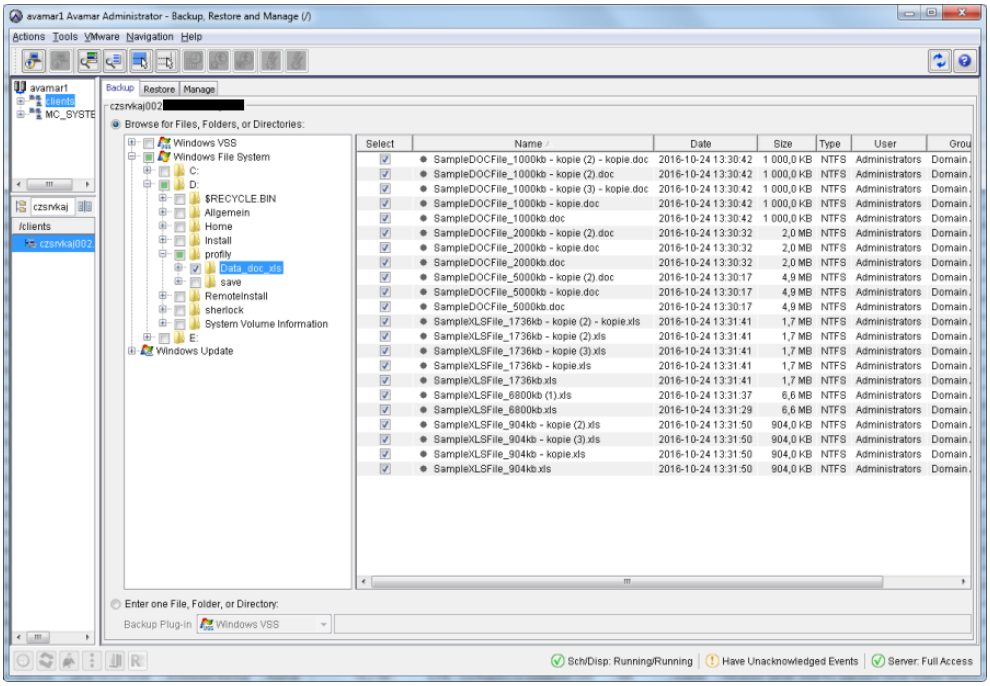
Jakmile se připravily všechny komponenty, bylo přistoupeno k demonstraci deduplikace systémem Avamar.

Projekt simuloval zálohování různých typů souborů:

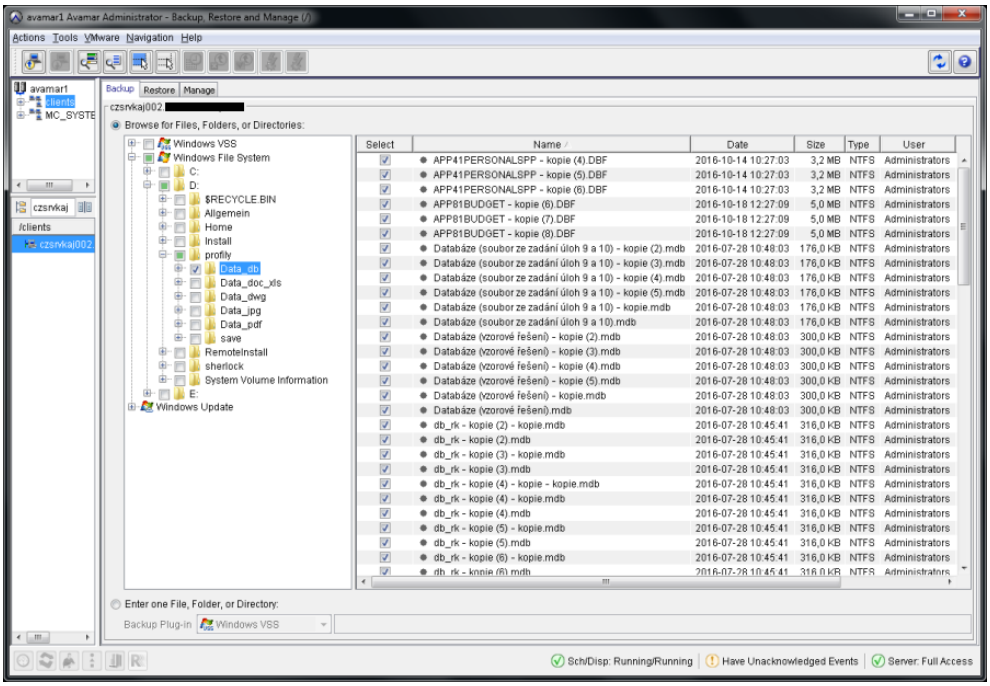
Typ souboru	*Počet souborů*	*Celková velikost [MB]*	*Název adresáře*
doc, xls	22	50,7	Data_doc_xls
dbf, mdb	94	50,6	Data_db
dwg	88	50,2	Data_dwg
jpg	105	50,2	Data_jpg
pdf	40	50	Data_pdf
všechny uvedené	349	251,7	Data_All

Tabulka 2: Soubory určené k zálohování

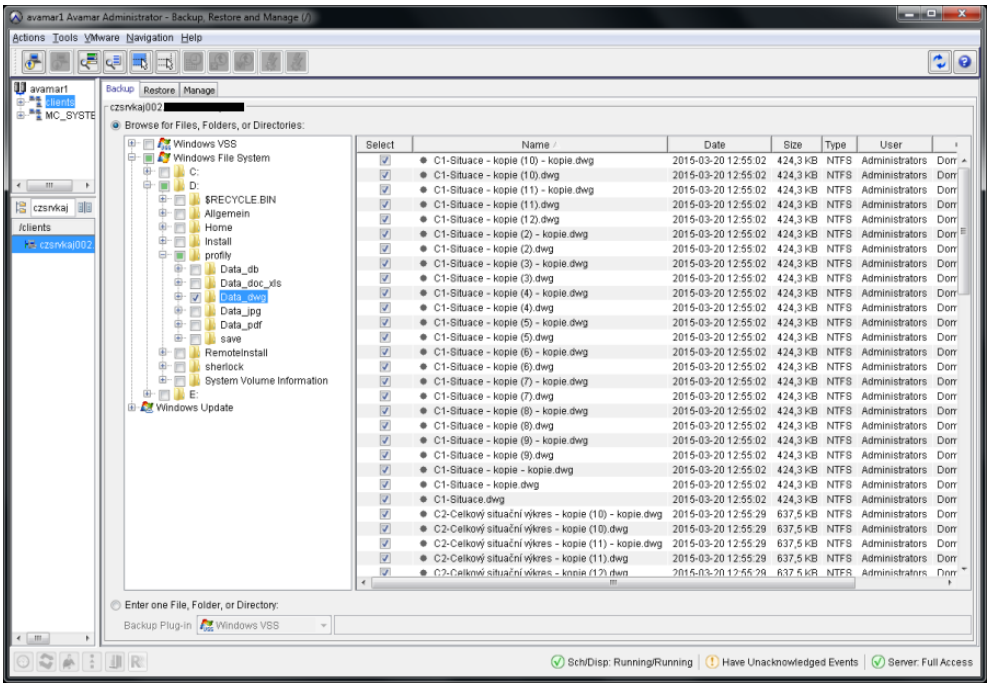
Soubory ve výše uvedené tabulce byly uloženy do adresářů a postupně zálohovány. Aby se více zviditelnil efekt deduplikace, nakonec se přesunuly do jednoho společného adresáře a opět zálohovány. Doba probíhající zálohy a přírůstek dat na serveru Avamar se monitoroval.



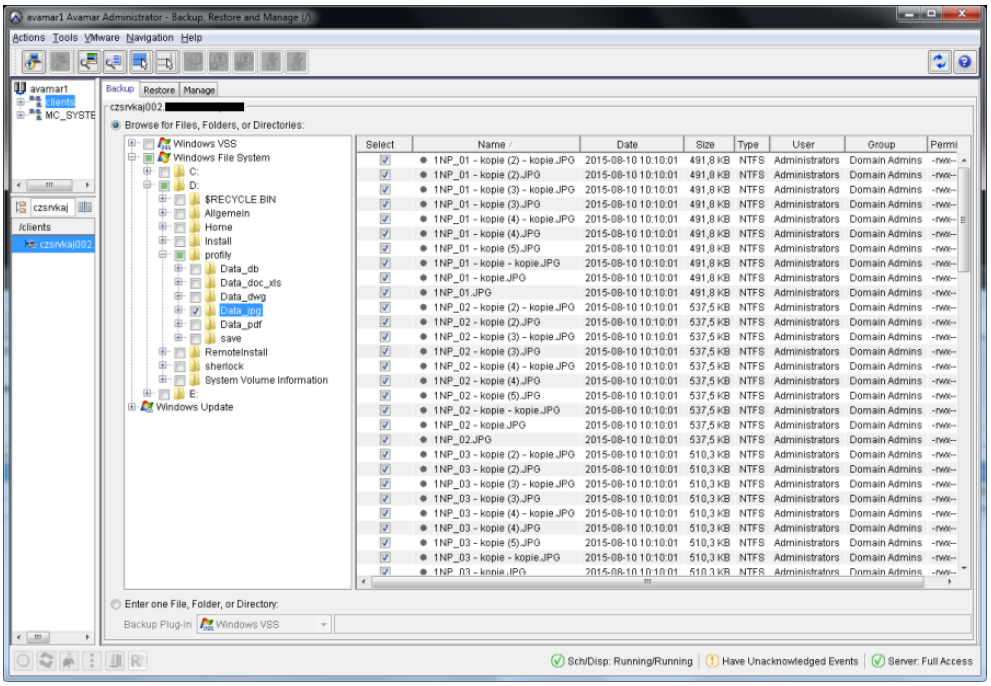
Obrázek 30: Záloha souborů MS Office



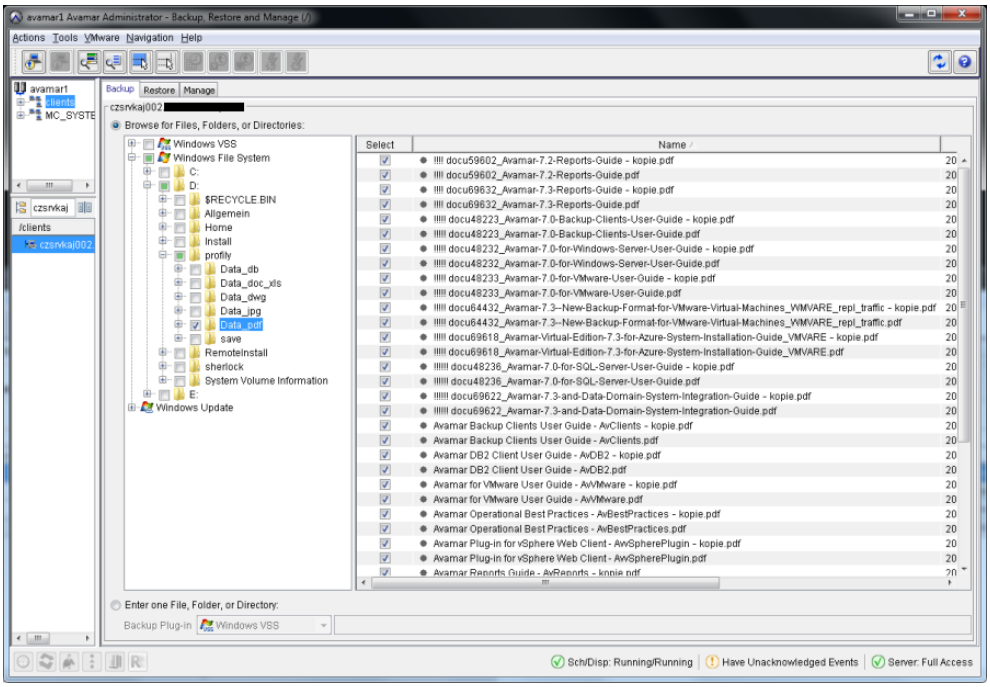
Obrázek 31: Záloha databázových souborů



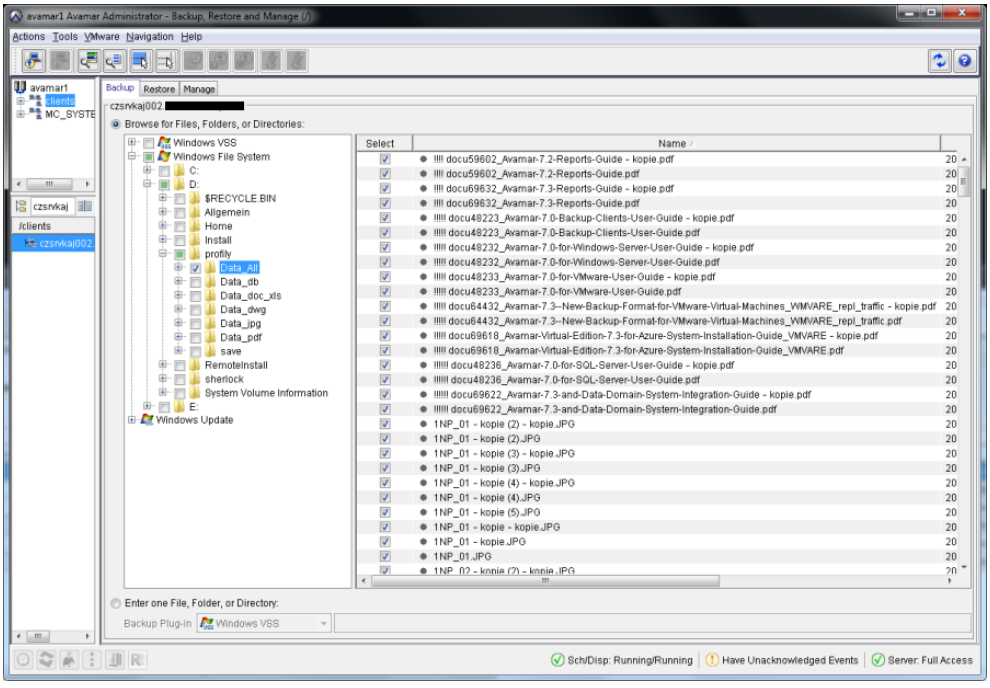
Obrázek 32: Záloha CAD souborů



Obrázek 33: Záloha fotografií



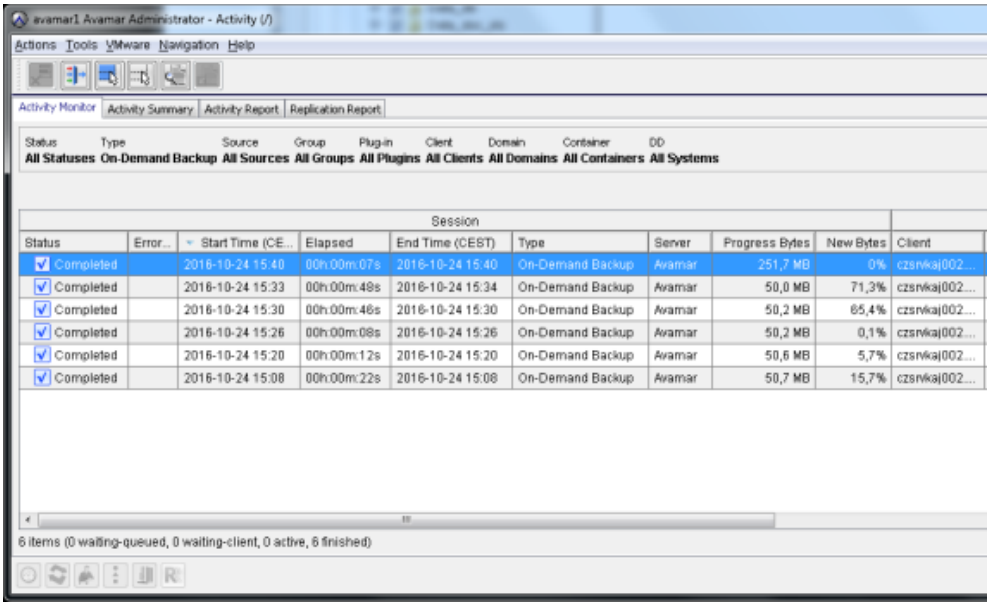
Obrázek 34: Záloha souborů pdf



Obrázek 35: Nastavení zálohování všech souborů

Využitím deduplikace nebyl přenesen celkový počet MB[37] zdrojových dat, ale pouze rozdílová data, která nebyla na zálohovací server Avamar nikdy předtím uložena. V poslední provedené záloze soubory sloučeny do jednoho společného adresáře. Jelikož již byla tyto data přenesená v předchozích zálohách, systém nezaznamenal změnu a nepřenesl žádná nová data.

Z výstupního reportu je patrný přírůstek dat uložených na serveru.



Obrázek 36: Výstupní report

V dlouhodobě běžícím systému Avamar byl sledován přírůstek rozdílových dat proti datům již zálohovaným. Rychlost provedení operace a množství přenesených dat je ovlivněno nejen datovými formáty určenými k záloze, HW[38] a SW[39] vybavením, ale také technologií deduplikace. Z naměřených hodnot vyplývá, že v případě tohoto projektu zálohování po datové síti mezi dvěmi pobočkami bylo za jednu sekundu přeneseno nejvíce MB[40] datových

[37] Megabytes
[38] Hardware
[39] Software
[40] Megabytes

souborů *pdf*, naopak nejméně formátů *dwg*. S ohledem na principy deduplikace se však výsledky jiných měření mohou lišit.

Příkladem může být projekční prostředí, kde je zálohováno velké množství *dwg* souborů a velmi malé množství souborů *doc*. V takové situaci může zálohování výkresové dokumentace proběhnout podstatně rychleji než v oblasti, kde se tyto dokumenty nepoužívají. To je velkou výhodou deduplikace. Je nezbytné analyzovat místní potřeby a požadavky na data určená k zálohování. Deduplikace byla detailně popsána v teoretické části této publikace.

Typ souboru	*Zdrojová data [MB]*	*Počet souborů [ks]*	*Přírůstek dat na serveru [%]*	*Přírůstek dat na serveru [MB]*	*Doba zálohování [s]*	*Přenos [MB/s]*
pdf	50,0	40	71,3	35,7	48	0,74
jpg	50,2	105	65,4	32,8	46	0,71
doc, xls	50,7	22	15,7	8,0	22	0,36
dbf, mdb	50,6	94	5,7	2,9	12	0,24
dwg	50,2	88	0,1	0,1	8	0,01
všechny uvedené	251,7	349	0,0	0,0	7	0,0

Tabulka 3: Přírůstek a rychlost přenosu dat ukládaných na server Avamar

4 Závěr

V bakalářské práci byla představena technologie deduplikace při zálohování vzdálených pracovišť demonstrovaná aplikací EMC Avamar deduplication backup. Deduplikací bylo umožněno uchování velkého množství dat a rychlý průběh zálohy. Bezproblémově lze provádět zálohu a obnovu vybraných souborů, diskových svazků i operačních systémů nejen v lokálních sítích, ale i prostřednictvím VPN[41] a PN[42] spojení.

Nepochybnou předností u velkého objemu zálohovaných dat, s ohledem na vynaložené finanční prostředky, je způsob licencování za množství uložených deduplikovaných dat, nikoliv za počet aktivních klientů nebo procesorových jader. Neméně důležitou vlastností je i hardwarová a softwarová podpora použité aplikace, on-line řešení problémů vzdáleným připojením odborného technika nebo přehledné administrátorské prostředí zvolené aplikace. V teoretické části byla také představena další softwarové řešení využívající této metody.

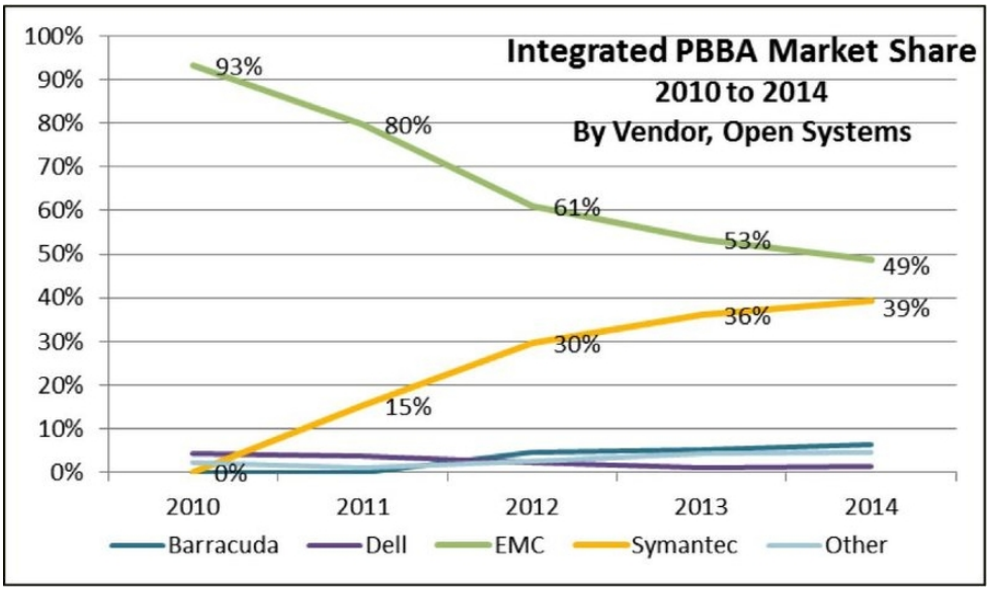
V případě realizace zálohy mezi lokalitami dominuje deduplikace na zdrojovém zařízení. To sice může mít negativní vliv na výkon stroje, přínosem je však kapacita datovým linek s podstatně menší rychlostí a tím i nižším cenovým nákladem než při duplikaci na cílovém serveru. První záloha stovek či tisíců zařízení současně může vytvářet problémy s propustnostní datové sítě, která je způsobena přenosem velkého množství dat. To je však jen otázkou první zálohy. Vhodným řešením je vytvoření menších skupin klientů a zálohovat postupně. V případě bezproblémového dokončení zálohy je možné počet zdvonásobit. Nedojde-li k problémům, lze dále navyšovat.

Slabinou systému je stabilita datového spoje. Jak je již výše uvedeno, není

[41] Virtual Private Network
[42] Private Network

důležitá jeho rychlost, ale v případě nestability či výpadku je zálohování přerušeno a po vyřešení problému nedojde k automatickému pokračování. Občasné problémy byly zaznamenány i u zálohování otevřených databází s neustálými změnami, například databáze Microsoft Exchange[43]. Je doporučeno tyto zálohovat v době nejmenší zátěže a mimo servisní plán, aby se eliminovalo riziko selhání. Bohužel se jedná o problém nemalé skupiny software pro zálohu dat.

V praktické části byl nasimulován přenos dat mezi firemními pobočkami a ověřena reálná úspora datového úložiště s použitím výše uvedeného řešení. Důkazem je i statistika využití techniky deduplikace různými zálohovacími prostředky a jejich postavení na světovém trhu.



Obrázek 37: Podíl na trhu zálohovacích zařízení dle výrobce [16]

[43]Poštovní server

Seznam použité literatury a zdrojů

[1] Miloslavlatal.cz: Zálohování dat [online]. Praha: Miloslav Látal, 2016 [cit. 2016-10-30]. Dostupné z: http://www.miloslavlatal.cz/zalohovani-dat-fakt-to-potrebuji/

[2] Vahal.cz: Proč se stále k zálohování používají pásky [online]. Prostějov: Vahal, 2015 [cit. 2016-11-01]. Dostupné z: http://www.vahal.cz/o-firme/clanky/proc-se-stale-k-zalohovani-pouzivaji-pasky.html

[3] Storagesearch.com: DAS, NAS and SAN storage architectures defined [online]. United States: Storagesearch.com, 2001 [cit. 2016-10-31]. Dostupné z: http://www.storagesearch.com/auspexart.html

[4] SystemOnLine.cz: DAS, NAS, SAN [online]. Brno: SystemOnLine.cz, 2004 [cit. 2016-07-13]. Dostupné z: https://www.systemonline.cz/clanky/das-san-nas.htm

[5] Abclinuxu.cz: Storage Area Network [online]. Praha: Nitemedia s.r.o., 2015 [cit. 2016-10-31]. Dostupné z: http://www.abclinuxu.cz/clanky/storage-area-network-1-uvod

[6] Computerworld.cz Deník pro IT profesionály: Deduplikace záloh pomůže snížit velikost úložiště [online]. Praha: IDG Czech Republic, a. s., 2016 [cit. 2016-10-27]. Dostupné z: http://computerworld.cz/technologie/deduplikace-zaloh-pomuze-snizit-velikost-uloziste-49836

[7] BERNAS, Karel. Deduplikace dat při zálohování [online]. Praha, 2012 [cit. 2016-11-01]. Dostupné z: https://www.vse.cz/vskp/47821_deduplikace_dat_pri_zalohovani.

Bakalářská práce. Vysoká škola ekonomická v Praze. Vedoucí práce Jan Pour.

[8] Deduplikace.cz: Druhy deduplikace [online]. Praha: GAPP System, 2016 [cit. 2016-10-20]. Dostupné z: http://www.deduplikace.cz/tri-druhy-deduplikace.php

[9] Dell EMC: EMC store products [online]. United States: Dell EMC, 2012 [cit. 2016-10-01]. Dostupné z: http://slideplayer.com/slide/1505685/

[10] EMC Avamar: Deduplikace dat [online]. 2016 [cit. 2016-21-09]. Dostupné z: http://www.deduplikace.cz/emc-avamar.php

[11] Slideshare.net: Energy Efficient Data Storage Systems [online]. Auburn: Xiao Qin Professor of Computer Science at Auburn University, 2012 [cit. 2016-10-13]. Dostupné z: http://www.slideshare.net/xqin74/energy-efficient-data-storage-systems

[12] Deduplikace.cz: Nástroje pro deduplikaci [online]. Praha: GAPP System, 2016 [cit. 2016-10-27]. Dostupné z: http://www.deduplikace.cz/nastroje-pro-deduplikaci.php

[13] Implementing IBM Tivoli Storage Manager [online]. United States: SearchDataBackup.com, 2012 [cit. 2016-11-28]. Dostupné z: http://searchdatabackup.techtarget.com/tip/Implementing-IBM-Tivoli-Storage-Manager

[14] Tivoli Storage Management [online]. Czech republic: IBM Corporation, Ondřej Bláha, 2010 [cit. 2016-11-29]. Dostupné z: http://www-05.ibm.com/cz/events/iif2010/resources/Efektivni_zalohovani_-_Novy_Tivoli_Storage_Manager__FastBack_-_part1_-_Ondrej_Blaha.pdf

[15] Symantec NetBackup PureDisk™ Deduplication Option Guide: Publication release 6.6.0.2, revision 1 [online]. United States: Symantec Corporation, 2016 [cit. 2016-11-28]. Dostupné z: https://symwisedownload.symantec.com

[16] IDC Worldwide Quarterly Purpose Built Backup Appliance Tracker – 2014 Q4 [online]. United States: Philip Wandrie, 2015 [cit. 2016-11-01]. Dostupné z: https://www.linkedin.com/pulse/why-do-i-think-emc-worried-losing-grounds-backup-abdul-rasheed

[17] Veritas NetBackup ™ Enterprise Server OS: Software Compatibility List [online]. United States: Veritas Technologies LLC / 2016-11-15, 2016 [cit. 2016-11-28]. Dostupné z: https://download.veritas.com

[18] IBM AIX [online]. United States: IBM, 2016 [cit. 2016-11-29]. Dostupné z: http://www-03.ibm.com/systems/power/software/aix/gettingstarted.html

[19] COUGIAS, Dorian J., E. L. HEIBERGER a Karsten KOOP. The BackupBook: Disaster Recovery from Desktop to Data Center. 1st ed. EN. Chicago: SV Books, 2003. ISBN 0-9729039-0-9.

[20] GÁLA, Libor, Jan POUR a Zuzana ŠEDIVÁ. Podniková informatika. 2., přeprac. a aktualiz. vyd. Praha: Grada, 2009. Expert (Grada). ISBN 978-80-247-2615-1.

[21] BARKER, Richard a Paul. MASSIGLIA. Storage area network essentials: a complete guide to understanding and implementing SANs [online]. 1st EN. New York: Wiley, c2002 [cit. 2016-11-01]. ISBN 04-710-3445-2.

[22] NELSON, Steven. Pro Data Backup and Recovery [online]. 1st EN. United States: Apress, 2011 [cit. 2016-08-02]. ISBN 978-1-4302-2663-5. Dostupné z: http://www.apress.com/gp/book/9781430226628

[23] EMC Avamar. EMC Account Login: Documentations [online]. United States, 2016 [cit. 2016-10-04]. Dostupné z: https://sso.emc.com/sso/login.htm?CTAuthMode=BASIC

[24] EMC® Avamar® Data Store Gen4T: Customer Installation Guide [online]. Massachusetts: EMC®, 2016REV2 [cit. 2016-10-28]. Dostupné z: https://support.emc.com/docu70649_Avamar-Data-Store-Gen4T-Customer-Installation-Guide.pdf?language=en_US

[25] Microsoft, Data deduplication overview [online]. 2013 [cit. 2016-10-27]. Dostupné z: http://technet.microsoft.com/en-us/library/hh831602.aspx

Seznam obrázků

1 Důvody poškození nebo ztráty dat [1] 13
2 Technologie LTO pásky [2] 15
3 Příklady implementace DAS [3] 17
4 NAS topologie [3] 18
5 Storage Attached Network [3] 20
6 Fixní a variabilní délka datových bloků [7] 23
7 Deduplikace na zdroji [7] 24
8 Deduplikace na cíli [7] 25
9 EMC Avamar [9] 27
10 Storage nodes [11] 33
11 Architektura IBM Tivoli Storage Manager [13] 34
12 Deduplikace na serveru [14] 35
13 Deduplikace na zdroji [14] 36
14 Deduplikace klientů [15] 39
15 Ukládání do PureDisk storage [15] 39
16 Replikace dat [15] 40
17 Chyba certifikátu přihlášení k webovému rozhraní 44
18 Webové rozhraní EMC Avamar Web Restore 45
19 Stažení aplikace Avamar Administrator 45
20 Avamar Administrator 46
21 Datasets 47
22 Datasets detail 48
23 Schedules 49
24 Instalace klienta Avamar 50
25 Aktivace klienta Avamar 50
26 Úspěšná aktivace 51

27 Správa klienta v Avamar Administrator 51
28 Přiřazení odpovídající zálohovací skupiny klientovi 52
29 Nastavení časového plánu zálohování 52
30 Záloha souborů MS Office 54
31 Záloha databázových souborů 54
32 Záloha CAD souborů 55
33 Záloha fotografií 55
34 Záloha souborů pdf 56
35 Nastavení zálohování všech souborů 56
36 Výstupní report 57
37 Podíl na trhu zálohovacích zařízení dle výrobce [16] 60

Seznam tabulek

1 Kapacita LTO všech generací 16
2 Soubory určené k zálohování 53
3 Přírůstek a rychlost přenosu dat ukládaných na server Avamar 58

A Příloha

CD/DVD