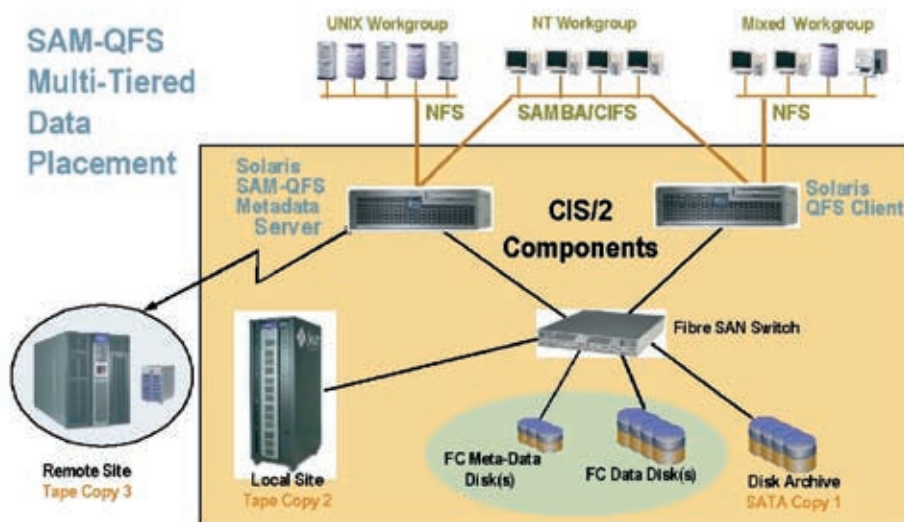


Problémy s archivací a jejich řešení

Sun Content Infrastructure System – jasná volba pro archivaci

Žijeme ve světě, ve kterém často informace představují to nejcennější, čím může člověk disponovat. Získat včas potřebnou informaci může znamenat, že vyděláme peníze, získáme vliv nebo naopak zabráníme katastrofě, či zachráníme lidské životy.



Správná informace ve správnou chvíli může mít nevyčíslitelnou hodnotu. Informace však stárnou a jejich hodnota klesá, přesto se v mnoha případech vyplatí takovou informaci uložit a později se k ní vrátit, tak abychom např. byli schopni analyzovat vývoj problému a na základě této analýzy najít snáze jeho adekvátní řešení. Existují ale také informace, které se vyplatí uchovávat prostě proto, že jejich hodnota je trvalá a že představují určité intelektuální dědictví, které je potřeba zachovat pro příští generace. Mohou to být umělecká díla stejně tak dobře jako třeba vědecké publikace. Zkrátka a dobře existuje celá řada informací, které je potřeba uchovávat dlouho nebo dokonce napořád. V některých případech nám takovou povinnost dokonce ukládá zákon.

Na první pohled na tom není nic složitého, nakonec lidstvo uchovává informace v různé podobě od nepaměti. V tomto ohledu se jako médium skvěle osvědčil papír, který nám některé informace uchoval po celá staletí. Nicméně i tato osvědčená metoda má svá úskalí. Patří mezi ně především složité vyhledávání informace, její problematický přenos, kopírování, ale také zabezpečení a v neposlední řadě nároky na fyzický prostor. Proč tedy nepřevést papírovou informaci do digitální podoby? Většina problémů se jistě vyřeší sama. Ale i uchování informace

ve digitální podobě má svá úskalí. Jak zaručit autentičnost informací? Jak zaručit čitelnost informace za 100 let? A kolik mě to všechno bude stát? To jsou pouze některé z otázek, které se člověku vybaví při vyslovení pojmu „digitální archiv“.

Co to je archiv?

Obecně lze digitální archiv chápat jako datové úložiště, s možností snadného vyhledání požadované informace. Důležitými charakteristikami digitálního archivu jsou mechanismy zajišťující bezpečnost uchovávaných dat, způsob jejich zpřístupnění a způsob, jakým je zabezpečena integrita dat. Protože objem informací většinou rychle roste, dostává se do popředí také požadavek na škálovatelnost. Klíčovou vlastností digitálního archivu je jeho životnost. Ta je úzce svázána s životností technologické infrastruktury, která předurčuje, jak dlouho lze informaci uchovat a zajistit její čitelnost. Posledním neméně důležitým aspektem archivace informací v digitální podobě jsou náklady spojené s uchováním informace. Ty jsou úzce spojeny nejen s pořizovací cenou technologické infrastruktury, ale také s náklady na její provoz, které stoupají s časem, po něž jsou data uchovávána. Ve struktuře nákladů na provoz hraje klíčovou úlohu spotřeba elektrické energie. Ta ve výsledku ovlivňuje nejen náklady, ale také

dopad na životní prostředí, který je v současnosti akcentován stále častěji a vzhledem k dramatickému růstu objemu uchovávaných informací jej nelze určitě do budoucna bagatelizovat.

Náklady na provoz a dopady na životní prostředí

Bezkonkurenčně nejrychleji rostoucím odvětvím ukládání dat je archivace nestrukturovaného obsahu. Data jsou ukládána zpravidla ve formě souborů. K tomu, aby bylo možné vyhledávat potřebné informace podle různých a dynamicky se měnících kritérií, je společně s těmito soubory potřeba uchovávat také jejich popis ve formě metadat. Metadata mohou být ukládána a spravována přímo na úrovni datového úložiště, případně na úrovni aplikace jako např. ECM. S velkým objemem dat v nestrukturované podobě se ale potkáváme nejen v rámci archivace dokumentů, ale také v souvislosti s archivací jiného obsahu. Může se jednat např. o archiv Roentgenových snímků na bázi PACS ve zdravotnictví, o archiv satelitních snímků zemského povrchu, nebo prostě jen o archiv starších e-mailů. V mnoha oborech lidské činnosti navíc dobu, po kterou musí být data bezpečně uložena a přitom kdykoliv dostupná, ukládá zákon. A to je právě hlavní kámen úrazu, protože nezřídka zákon ukládá archivovat data po dobu několika desítek let. V takovém časovém horizontu už se vyplatí zvažovat ekonomické a také ekologické aspekty spojené s provozem archivu.

Datová úložiště sice v celkových nákladech za spotřebovanou energii v datovém centru nedosahují úrovně serverů, problém je ale v tom, že náklady s nimi spojené rostou rychleji, než je tomu v případě serverů. Důvod je jednoduchý – objem dat prostě dramaticky roste. Jak by měla tedy vypadat technologická infrastruktura? Jedním z nejrozšířenějších mýtů spojených s digitální archivací je představa, že nejjednodušším způsobem je data uchovávat na levných SATA discích. Vždyť cena těchto disků vytrvale klesá a jejich kapacita naopak roste. Dokonce se v diskových systémech objevují technologie umožňující snižovat odběr energie na bázi snižování otáček disků. Disk jako médium pro ukládání dat nabízí sice výhody spojené s rychlým vyhledáním informace, má však několik významných nevýhod. Jednou z nich je právě již jmenovaná spotřeba energie spojená s mechanickými otáčkami, kterou se při

veškeré snaze nedaří radikálně snížit. Mechanický pohyb navíc významným způsobem snižuje životnost disků, a proto je nutné vzdorovat tomuto faktu různými metodami zvyšování redundance ukládané informace. Jen těžko si lze představit, že důležitou informaci uchováme na disku více než 10 let. Se spotřebovanou energií souvisí také zvýšené nároky na klimatizaci, omezená hustota disků v racku a ve výčtu by se jistě dalo pokračovat. Disky se tedy prodlouhodobou archivací nehodí. Čím je ale nahradit?

Jisté řešení nabízejí magnetopáskové technologie, které jsou v porovnání s disky o něco starším médiem. Také tato technologie prošla překotným vývojem. Vždyť rychlost čtení i zápisu dnes u pásek překračuje rychlosti dosahované v případě disků a také kapacitou se dnešní magnetické pásky diskům vyrovnají. Magnetická páska je navíc mnohem levnější a náklady spojené s odběrem energie jsou v případě magnetopáskové knihovny minimální. Některé studie uvádějí až 26x nižší energetické náklady v porovnání s disky a až osmkrát nižší náklady na klimatizaci. Když k tomu připočteme nižší pořizovací cenu, nezbyvá než zajásat. V reálném světě to nicméně tak jednoduché není. Páska má totiž jako médium jednu zásadní nevýhodu. Informace se na ní ukládá sekvenčně. To znamená, že je požadovanou informaci na pásce potřeba nejprve najít, a to znamená její převinutí. Vyhledání informace tak trvá podstatně déle než u disků. Optimální je tedy výhody obou technologií při budování archivu kombinovat. Disková cache spolu s promyšlenými pravidly pro přesun dat na pásky zajistí rychlý přístup k nejčastěji využívaným datům, pásky jako hlavní médium umožní snížit cenu řešení.

Životnost

Použití pásek pro archivaci má ještě jednu výhodu. Životnost dat lze totiž garantovat po delší dobu. Moderní páskové technologie umožňují garanci po dobu až 30 let. To je hodnota, které u disků dosáhnout nelze. Pokud však disky plní pouze roli cache, lze je periodicky obměňovat v návaznosti na překotný technologický vývoj v této oblasti. Kopie dat jsou přitom bezpečně uloženy na páskách třeba i v několika kopiích a po technologické obměně diskové cache je možné tato data do cache znovu přesunout. Kombinovaný archiv tak poskytuje rozumnou

úroveň flexibility. Klíčem přitom zůstávají inteligentně definovaná pravidla pro přesun dat z disků na pásky a naopak. Jakýsi hierarchický způsob ukládání dat.

Sun v takových případech s úspěchem využívá virtuálního hierarchického souborového systému SAM-QFS. Z hlediska přístupu k datům se jedná o transparentní souborový systém, kde složitost hierarchického uspořádání dat zůstává běžnému uživateli skryta. Data uložena v souborovém systému se nachází buď na discích nebo na páskách, přístup k informacím je však v obou případech stejný. Jsou-li načítána data, která nejsou přítomna v diskové cache paměti, dojde nejprve k jejich stažení do cache a pak teprve jejich načtení. Uživatel to sice pozná na době odezvy, způsob přístupu nicméně zůstává zachován. Takto strukturovaný souborový systém lze navíc sdílet přímo prostředky SAN pro čtení i zápis, a to díky technologii QFS. QFS klient přistupuje k datům přímo po SAN, nebo dále zprostředkovává přístup dalším klientům v síti LAN prostřednictvím protokolů NFS nebo CIFS. Největší výhodou SAM-QFS je však fakt, že přestože se jedná o komerční produkt, jeho zdrojový kód byl v rámci strategie Open Storage uvolněn a je vývojářům k dispozici. Management software, jehož zdrojový kód je k dispozici, přitom snižuje rizika vyplývající z otázky: „Bude tento software podporován ještě za 30 let?“

Sun Content Infrastructure System

Závěrem lze říci, že popsané principy archivace lze najít v komerčním produktu firmy SUN skrývajícím se pod zkratkou CIS. Tento produkt je kompletním zařízením, skládajícím se z dvouúrovňové hierarchie diskových polí s FC a SATA disky a třetí storage úrovně na bázi páskové knihovny SL500. Vše je úhledně namontováno do 19 palcového racku a pospojováno do malé SAN prostřednictvím vestavěných FC switchů. Součástí balíku je také Sun server se software SAM-QFS řídicí přesun dat mezi zmíněnými úrovněmi storage. CIS tak pro mnoho zákazníků potýkajících se s problémy spojenými s dlouhodobým uchováním dat může představovat rychlé a snadné řešení.

Více informací naleznete na www.sun.com