Ověření modelu Object-based Storage Device pro AFS

Michal Švamberg

Západočeská univerzita v Plzni
Centrum informatizace a výpočetní techniky
e-mail: svamberg@civ.zcu.cz

2. února 2009

Abstrakt

Cílem předkládaného projektu je seznámit se s architekturou *Object-based Storage Device (OSD)* a ověřit využitelnost deklarovaných vlastností systému ukládání dat pro distribuovaný souborový systém AFS. V pilotním prostředí bude provedeno testování základních provozních charakteristik (přenosové rychlosti, propustnost, stabilita apod.) a jejich porovnání vůči klasickému modelu uložení dat používaném v AFS.

Projekt počítá s aktivním zapojením několika studentů s cílem zprostředkovat jim přístup k moderní technologii a umožnit další vzdělávání v dané oblasti.

Současný stav řešeného problému

Distribuovaný souborový systém AFS[1] je v současnosti intenzivně využíván celou řadou akademických i komerčních institucí po celém

[1] Andrew File System byl vyvinutý na v rámci projektu *Andrew* na CMU v polovině osmdesátých let.

světě[2]. Standardní implementace je dostupná v rámci open source projektu OpenAFS[3] v produkční kvalitě a podporuje všechny běžně používané operační systémy.

V rámci aktivity OpenAFS je rozpracováno několik projektů[4], které si kladou za cíl doplnit novou funkcionalitu či zoptimalizovat existující koncept tak, aby vyhověl novým požadavků na moderní distribuované úložiště, zejména neustále se zvyšujícím nárokům na úložné kapacity a přenosové rychlosti.

Na Západočeské univerzitě v Plzni (ZČU) využíváme AFS více než 10 let jako hlavní sdílený souborový systém sloužící k uložení dat rozličného charakteru a určení (domovské adresáře uživatelů, webové stránky, konfigurační soubory, software, instalační obrazy, projektová data, výpočty apod.).

Postupný nárůst požadavků na objem ukládaných dat a rychlost přístupu k nim nutí i nás hledat cesty k optimalizaci stávající infrastruktury.

Potenciální problém s maximální přenosovou rychlostí se v podmínkách ZČU bohužel nepodařilo vyřešit ani postupným laděním parametrů datové komunikace a souborových serverů; současná rychlost se i na gigabitové infrastruktuře pohybuje okolo 6 MB/s. Naproti tomu, je pozitivní, že uvedená rychlost je dosahována stabilně a to i při vysoké zátěži serverů a paralelním přístupu stovek klientů (tato vlastnost vychází z návrhu AFS).

Domníváme se, že s ohledem na vznikající požadavky a dostupné hardwarové vybavení je dosahovaná přenosová rychlost relativně nízká a v blízké budoucnosti nedostačující. Taktéž využívání současného modelu tzv. AFS volumů[5] ve spojení s funkcemi lokálního souborového systému naráží při zpracovávaných objemech dat na řadu omezení a pohybuje se na hranici použitelnosti.

Přes uvedená omezení přináší využívání AFS vysokou přidanou hodnotu a považujeme za správné a perspektivní zabývat se podrobněji optimalizací a implementací navrhovaných rozšíření.

Jednou z efektivní metodou řešení výše uvedených omezení se zdá

[2] http://www.openafs.org/success.html

[3] http://www.openafs.org/

[4] http://openafs.org/projects.html

[5] Část adresářového stromu, který je uložen a spravován jako nedělitelný celek nástroji AFS.

být implementace rozšíření umožňující využívat *OSD modelu ukládání dat* v AFS namísto „klasického“ využití funkcí lokálního souborového systému. Implementace OSD navíc obohacuje (jako postranní efekt) AFS o absentující hierarchický model správy datového úložiště[6]

Cíle řešení

Projekt by měl z technického hlediska odpovědět, zda nasazení OSD modelu zvýší maximální přenosové rychlosti a propustnost AFS infrastruktury a zkvalitní možnosti správy uložených dat na souborových serverech včetně využití vlastností HSM. Podstatným výstupem bude také praktické ověření kompatibility s klasickou implementací.

Zavedení ukládání dat založeného na objektovém přístupu, řeší problém nedělitelnosti volumů. Distribuce volumů a jednotlivých souborů mezi více souborových serverů má efekt zvýšení propustnosti přenášených dat směrem ke klientům.

Pro následný převod produkční infrastruktury je zcela nezbytné ověřit kompatibilitu přístupu k datům uloženým s využitím OSD i HSM modelů.

Obdobně jako u několika předchozích grantů[7], bychom rádi zapojili do řešení tohoto projektu také studenty. Centrum informatizace a výpočetní techniky (CIV) již řadu let umožňuje studentům podílet se na návrhu koncepce a správě výpočetního prostředí Orion[8]. Zapojením vybraných studentů do projektu předpokládáme zvýšení jejich odborné kvalifikace a umožnění nahlédnutí do procesu správy a rozvoje rozsáhlé výpočetní infrastruktury. Získané znalosti mohou přímo zúročit nejen ve výuce, ale i v pozdější praxi.

V projektu předpokládáme účast na zahraniční konferenci orientované na správu a implementaci distribuovaných souborových systémů či přímo na správu AFS. Zde bychom rádi získali aktuální informace o stavu implementace rozšíření OSD či obdobných projektech souvisejících se zaměřením předkládaného grantu.

[6] Hierarchical storage management (HSM): http://en.wikipedia.org/wiki/Hierarchical_storage_management.

[7] Granty FR-CESNET 154R1/2005 a 192R1/2006)

[8] Komplexní distribuované výpočetní prostředí ZČU budované více než 15 let.

Střednědobým záměrem závislým přímo na výstupech grantu je optimalizace existující infrastruktury AFS na ZČU s cílem podstatně zvýšit její propustnost a zefektivnit správu dat v rámci existujících úložných kapacit tak, aby bylo možno poskytovat kvalitní, vysoce dostupné služby a flexibilně reagovat na požadavky koncových uživatelů.

Popis modelu OSD

Standard rozhraní OSD[9] byl vypracován v rámci odborné skupiny SNIA[10] a definuje novou sadu SCSI příkazů specifikovanou jako protokol ANSI T10 SCSI OSD V1[11].

V prvním přiblížení se můžeme na OSD dívat jako na obdobu *logické jednotky*. Na rozdíl od tradičních blokově orientovaných zařízení však zprostředkovává data na základě tzv. *úložných objektů*. Jedná se o virtuální entity spojující logicky související data definovaná uživatelem. Fyzický způsob a místo uložení těchto objektů jsou však plně v interní režii OSD a nezávislé na lokálním souborovém systému. OSD interně zabezpečuje všechny nízkoúrovňové funkce pro ukládání, správu a zabezpečení dat[12].

AFS implementace rozšíření OSD[13] vychází ze starší koncepce MR-AFS[14]. Přínos OSD je především v HSM funkcionalitě, vyšší škálovatelnosti a propustnosti na kapacitních linkách a poskytování podrobnějších informací využitelných pro optimalizaci a jednoduší správu AFS buňky. Podstatnou vlastností je i zachování zpětné kompatibility a s tím spojená možnost postupné migrace dat v produkčním prostředí. Podrobnější informace o současném stavu projektu byly prezentovány na OpenAFS konferenci pořádané v září 2008 v Grazu[15].

[9] V současnosti ve verzi OSD-2.

[10] http://www.snia.org/tech_activities/workgroups/

[11] http://www.t10.org/

[12] Podrobnější informace, přesahující rámec tohoto dokumentu, lze nalézt např. na http://developers.sun.com/solaris/articles/osd.html.

[13] Nemá nic společného s objektovým programováním.

[14] Multi-Resident AFS.

[15] http://www.openafs.at/drupal/files/slides/1Day_03/AFS-OSD.pdf

Způsob řešení

Pro potřeby projektu předpokládáme vytvoření testovacího prostředí pro nějž budou v rámci projetu zakoupeny dva serveru v konfiguraci, která umožní na každém provoz čtyř virtuálních strojů. Všechna data budou uložena na lokálních rychlých discích, tak aby se omezilo případné úzké hrdlo hardware. Přepínání virtuálních strojů umožní provoz klasického OpenAFS i rozšířeného OpenAFS s podporou OSD nad stejným fyzickým zařízením.

V první části bude potřeba instalovat testovací AFS buňku a seznámit se s konfigurací, ovládáním a novými vlastnostmi.

Dalším krokem bude příprava celé infrastruktury k testování, stanovení testů a vytvoření testovacích dat. Testovací prostředí se pokusíme částečně napojit na již existující infrastrukturu (PTS, VLDB databáze a Kerberos). Pro hromadné testy, kde čte více klientů shodná nebo „blízká“ data (shodný volume, různé volumy uložené na jednom serveru) bude použito klientů nainstalovaných ve veřejných učebnách. Testy se budou soustředit spíše na samostatného klienta, protože u něj očekáváme nejmarkantnější rozdíl.

Jako vhodný reálný test lze uskutečnit automatizovanou instalaci PC na veřejných učebnách, kde jsou veškerá data (konfigurace FAI[16], obraz MS Windows a mirror Debian GNU/Linux) uložena právě na AFS. V případě pozitivních výsledků budou tyto poznatky uplatněny a rozšířeny na další části distribuovaného souborového systému.

Ve výpočetním prostředí ZČU je provozováno řádově stovky AFS klientů v různých verzích na různých platformách (MS Windows, Linux, Solaris. . .). Na těchto klientech bude ověřena kompatibilita a dostupnost dat uložených na OSD a HSM.

Vlastnosti HSM budou vyzkoušeny na velkých řídce nepoužívaných souborech, které budou migrovány na pomalejší a levné velkokapacitní média (pásku, levné diskové pole). V projektu bude tento typ úložiště simulován na testovacím serveru s velkokapacitními levnými disky (2×1 TB).

Na projektu se bude podílet alespoň jeden student, jemuž pod vedením řešitelů bude umožněn přístup a možnost podílet se na přípravě a provozu nově zaváděné technologie do heterogenního prostředí ZČU.

[16] Fully Automatic Installation: http://www.informatik.uni-koeln.de/fai/

Tento přístup se nám velmi dobře osvědčil již v předchozích projektech a rádi bychom v něm pokračovali.

Cestovné a vložné bude využito pro účast dvou osob na mezinárodní zahraniční konferenci se zaměřením na problematiku implementace, poskytování, optimalizace a správy služeb distribuovaného datového úložiště. V případě konání a odpovídající odborné náplně předpokládáme účast na Evropské konferenci o AFS[17], která se koná pravidelně na podzim.

Očekávané přínosy

Jak již bylo v předchozím textu řečeno, mezi hlavní očekávané přínosy patří zvýšení kvalifikace pracovníků CIV a vybraných studentů a ověření deklarovaných technických přínosů, zejména:

- *zlepšení propustnosti* dat ke klientům,
- *optimalizované využití prostoru* na diskových polích,
- *vyžití HSM* pro minimalizaci nákladů a optimalizaci parametrů datového úložiště,
- *zlepšení správy* uložených dat na diskových subsystémech,
- *více informací* o volumech a souborech uložených na AFS umožňující optimalizaci AFS buňky,
- *zpětná kompatibilita* AFS klientů a možnost migrace produkčního prostředí,
- *zvýšení kvalifikace* zaměstnanců a vybraných studentů v oboru distribuovaných souborových systémů.

Podle dostupných informací od autorů, je vývojová fáze projektu dokončena a nyní bude následovat integrace kódu do standardní distribuce OpenAFS. Předpokládáme, že zkušenosti získané v předkládaném projektu nám podstatně ulehčí a urychlí následnou implementaci nového rozšíření do produkčního infrastruktury Orion.

[17] http://www.openafs.at/drupal/

Prezentace výsledků

Odborné materiály, získané poznatky, použité postupy a výsledky související s řešením projektu budou zájemcům dostupné v elektronické formě prostřednictvím WWW stránek CIV, ZČU v Plzni.

Klíčové výstupy budou prezentovány formou seminářů a přednášek. Ke stěžejní prezentaci výsledků bude patřit přednáška pro odbornou veřejnost na některé z národních konferencí zabývající se tématikou správy systémů. Předpokládat prezentaci na konferenci EurOpen.CZ[18] případně LinuxAlt[19].

Výsledky projektu budou prezentovány na odborném semináři pro studenty a zaměstnance ZČU, který bude zaměřen na konkrétní dopady a možnosti využití nových vlastností v prostředí Orion.

Charakteristika řešitelského týmu

Řešitelský tým je složen ze zkušených pracovníků Laboratoře počítačových systémů (LPS) Centra informatizace a výpočetní techniky (CIV) na Západočeské univerzitě v Plzni (ZČU). Řešitelský kolektiv má zkušenosti z oblasti distribuovaných výpočetních prostředí a implementace systémů do takové infrastruktury. Do řešitelského týmu bude také vybrán student[20], který bude mít zájem na projektu spolupracovat.

Ing. Michal Švamberg (hlavní řešitel) je absolventem Fakulty Aplikovaných věd Západočeské univerzity v Plzni v oboru Distribuované systémy. Od roku 2002 pracuje v Laboratoři počítačových systémů (LPS) Centra informatizace a výpočetní techniky (CIV), kde se účastnil návrhu a budování kolejních sítí Západočeské univerzity. Dále se zabývá správou operačního systému Linux a jeho integrací do distribuovaného výpočetního prostředí `Orion`. Spravuje také FibreChannel infrastrukturu, distribuovaný souborový systém AFS a XEN virtuální stroje. Na ZČU působí jako instruktor

[18] http://www.europen.cz/

[19] http://www.linuxalt.cz/

[20] Předpokládá se, že bude vybrán alespň jeden student ze skupiny HELPs. Tato skupina studentů pomáhá řešit běžné uživatelské problémy na stanicích jež provozuje CIV.

v certifikačních programech CCNA a CCNP[21]. Pro CESNET z.s.p.o. se podílí na správě výpočetních clusterů jako řešitel výzkumného záměru, projekt METACentrum národní gridové a superpočítačové infrastruktury.

Ing. Luboš Kejzlar (spoluřešitel) je absolventem Fakulty elektrotechnické Vysoké školy strojní a elektrotechnické v Plzni v oboru Automatizované systémy řízení. Od roku 1989 je v různých funkcích členem Laboratoře počítačových systémů (LPS) Centra informatizace a výpočetní techniky (CIV) ZČU v Plzni. V rámci své pracovní náplně se podílel na řešení řady rozsáhlých projektů z oblasti návrhu síťové infrastruktury a implementace distribuovaného výpočetního prostředí (WEBnet, Cassiopea, Orion). S několika přestávkami se od počátku podílí na realizaci projektů Superpočítačového centra VŠ a návazně *META Centrum*. Mezi jeho profesní zájmy patří problematika bezpečnostní infrastruktury a distribuovaných výpočetních prostředí. Od roku 2002 řídí ve funkci vedoucího LPS výjimečný kolektiv cca. dvaceti spolupracovníků, kteří se aktivně podílejí na tvorbě koncepce, implementaci a provozu komplexního výpočetního prostředí ZČU. Z povahy své funkce v současnosti řídí či se podílí na řešení řady střednědobých infrastrukturálních projektů v oblasti PKI, AAA, mobility, MOM či řízení přístupu a správy identit.

Navrhovaná doba trvání projektu

Doba trvání projektu je plánována na 12 měsíců.

Finanční rozvaha

Pro projekt jsou požadovány náklady ve výši 298 200,- Kč včetně DPH. Z prostředků Fondu Rozvoje budou čerpány prostředky ve výši 191 200,- Kč, zbylou částku 107 000,- Kč (tj. 36%) včetně dalších nákladů spojených s projektem hradí řešitelská organizace formou spoluúčasti.

[21] Jedná se o kurzy z Cisco Networking Academy Program, více viz http://www.netacad.cz/.

Celkové náklady na pokrytí projektu včetně DPH byly stanoveny následovně:

Položka	Cena v Kč	Hrazeno
2× server	171 200	FR-CESNET
2× HDD pro simulaci HSM	7 000	ZČU
tuzemské cestovné, vložné	30 000	ZČU
zahraniční cestovné, ubytování	40 000	ZČU
zahraniční vložné	20 000	FR-CESNET
odměny studentských řešitelů formou stipendia	30 000	ZČU
celkem	298 200	

Náklady na zahraniční cestovné, ubytování a vložné jsou odhadovány na základě informací z již konaných zahraničních konferencí a aktuálních cen za dopravu resp. ubytování.

Tuzemské cestovné a vložné pokrývá náklady odhadované na návštěvu a prezentaci výsledků projektu na některé z národních konferencí zabývající se administrací systémů.

Cena hardware, konfigurace a dodavatel (viz příloha) vychází z interního výběrového řízení ZČU pro rok 2008 na dodávku serverů. Předpokládaná životnost serveru je minimálně 5 let, čemuž odpovídá záruční podmínka a hardwarová konfigurace.