Distribuované Datové Sklady

Lukáš Hejtmánek, FI MU,Luděk Matyska, ÚVT MU

1 Projekt DiDaS

Dostatečná disková kapacita se stále více stává jedním z hlavních hodnotících faktorů při pořizování nového počítače. Schopnost *počítat*, od níž mají i počítače své jméno, se s růstem výkonu procesorů stává stále podružnější a na první místo pozornosti se dostávají data a možnosti jejich zpracování. Analogický trend pozorujeme i v oblasti rozsáhlých distribuovaných systémů, kde se v posledních letech zkoumají možnosti nových způsobů využití diskové kapacity systémů propojených počítačovou sítí. Pracovníci ÚVT MU a studenti FI ve spolupráci s dalšími vysokými školami začali proto v roce 2003 řešit výzkumný projekt *Distribuované Datové Sklady (DiDaS)*, dotovaný z Fondu rozvoje sdružení CESNET. Cílem tohoto projektu byl výzkum možných přístupů k realizaci rozsáhlého distribuovaného datového úložiště a jeho následná realizace v prostředí akademické počítačové sítě CESNET. Výsledná datová kapacita je pak zpřístupněna akademickým uživatelům jako velkokapacitní dočasné úložiště. Projekt byl řešen od jara 2003 do června 2004 a jeho výsledky jsou v současné době postupně zpřístupňovány akademické veřejnosti České republiky.

2 Základní struktura

Distribuované datové úložiště vytvořené v rámci projektu DiDaS je tvořeno 10 diskovými poli, z nichž většina má neformátovanou kapacitu 1,5 TB. Každé diskové pole je řízeno osobním počítačem, vybaveným zpravidla jedním procesorem Intel Pentium IV s frekvencí 2,8 GHz, 1 GB paměti a gigabitovou síťovou přípojkou. Všechna disková pole jsou tvořena vysokokapacitními ATA disky (starší v provedení PATA, tedy parallel ATA, novější ve výkonnějším provedení SATA, serial ATA). Výjimku tvoří jedno experimentální diskové pole, tvořené SCSI disky s menší kapacitou, ale výrazně vyšší propustností. Disky každého diskového pole jsou zapojeny v uspořádání RAID 5, tj. pole jsou odolná proti výpadku jednoho z disků. Některá disková pole jsou externí – řadič disků je v samostatném boxu a je s řídícím počítačem propojen SCSI propojením, ostatní jsou interní – řadič disků je instalován přímo v počítači jako PCI karta.

Na řídícím počítači je instalován operační systém Linux (v současné době s jádrem 2.6) a počítače jsou zapojeny přímo do páteře akademické sítě CESNET2. Jednotlivá pole jsou umístěna na sedmi místech v ČR, konkrétně na ZČU v Plzni, JČU v Českých Budějovicích, TU v Liberci, UK a CESNETu v Praze, VŠB TUO v Ostravě a zbývající disková pole pak na MU v Brně. Tím je zaručena distribuce diskové kapacity a současně úplné pokrytí České republiky.

Datové sklady jsou zpřístupněné pomocí *protokolu IBP* (Internet Backplane Protocol) vyvinutého na univerzitě Tennessee Knoxwille v laboratoři LoCI. Podobně jako je IP protokol abstrakcí přenosu dat nad linkovou vrstvou, je IBP protokol abstrakcí, založenou na „datových blocích“, zpracovávaných jako pole bytů. IBP umožňuje vytvářet abstraktní síťovou vrstvu ukládání dat, která je nezávislá na konkrétní struktuře úložných systémů (disky, disková pole, ...). IBP používá *slabý* model konzistence a dostupnosti dat, tj. podobně jako IP reprezentuje pouze *best effort* službu. V rozsáhlé počítačové síti nelze garantovat dostupnost konkrétního místa (diskového pole) ani nelze garantovat, že někde nedojde ke ztrátě konkrétních dat. Protokol IBP proto data ukládá pouze dočasně, s každým uloženým objektem (souborem) je spojena doba expirace, po níž může systém data smazat. Uživatel je odpovědný za prodlužování této expirační doby, současně musí počítat s tím, že data na konkrétním místě nebudou dostupná. IBP protokol umožňuje několikanásobné ukládání stejných dat a touto redundancí fakticky zajišťuje dostatečně vysokou míru dostupnosti dat.

Podobně jako v prostředí protokolu IP, hovoříme o IBP stacku, který je tvořen několika vrstvami. Na nejnižší úrovni je transportní protokol IBP, který slouží ukládání dat do IBP skladů. Další je L-Bone vrstva, která registruje jednotlivé IBP sklady. L-Bone vrstva projektu DiDaS je k vidění na adrese `http://undomiel.ics.`

muni.cz/lors/lbone_list_view.cgi. Vedle L-Bone vrstvy stojí vrstva nazvaná Ex-node, jde o XML popis dat uložených v IBP skladech. Nad oběma vrstvami pak stojí LoRS vrstva, která poskytuje nástroje pro ukládání a stahování dat do/z distribuovaných skladů. LoRS vrstva současně „skrývá“ konkrétní sklady a poskytuje tak abstrakci datového úložného prostoru.

XML popis vrstvy Ex-node je nezbytný pro přístup k jednou uloženým datům. Obsahuje totiž „souřadnice“ jednotlivých bloků uloženého souboru (tedy na jakém datovém skladu či datových skladech je konkrétní blok uložen a jaká je jeho pozice na disku či diskovém poli). Bez znalosti Ex-node nelze soubor rekonstruovat. V datových skladech jsou přitom ukládána pouze uživatelská data, XML popis je v základní implementaci IBP ukládán u uživatele (na jeho lokálním disku) a uživatel je odpovědný za to, že Ex-node informaci neztratí. Pokud se tak stane, data budou nedostupná a systém je po uplynutí expirační doby sám smaže. V rámci projektu DiDaS byly vyvinuty nástroje, které umožňují Ex-node informaci ukládat do distribuovaného systému souborů AFS a tím uživatele zbavit odpovědnosti za uchování přístupových informací na lokálním disku.

3 Přístup na datové sklady

Pro přístup k datovým úložištím je k dispozici řada nástrojů, které se liší především v míře, s níž je konkrétní struktura datových skladů zpřístupněna uživateli.

3.1 Utility příkazové řádky

Utility příkazové řádky tvoří základní přístupové nástroje, určené především pro dávkové zpracování, případně pro pokročilé uživatele.

Z řádkových příkazů jsou nejdůležitější `lors_upload`, `lors_download`, `lors_trim` a `lors_ls`, které popíšeme dále.

- `lors_upload` slouží pro nahrání souboru do infrastruktury.
  Příklad použití je:
  ```
  lors_upload -f -H didas.ics.muni.cz
  -h -d 10d -c 1 soubor
  ```
  Soubor `soubor` bude nahrán na datového úložiště. Ex-node informace bude uložena na lokálním disku v souboru `soubor.xnd`. V případě ztráty tohoto souboru není žádná možnost, jak svá data získat, a to ani s pomocí administrátorů datových skladů.
  Volba `-H` uvádí konkrétní adresu L-Bone serveru. Pro zvýšení spolehlivosti bylo v rámci projektu DiDaS definováno generické jméno `didas.ics.muni.cz`, které je aliasem pro všechny sklady. Výběr konkrétního skladu provede jmenná služba Internetu (DNS) v okamžiku vyvolání tohoto příkazu. Konkrétní jméno je uloženo v souboru s Ex-node informací, takže se uživatel vůbec o fyzické umístění svých dat nemusí starat.
  Pokud je v okamžiku volání funkce konkrétní datový sklad nepřístupný, příkaz `lors_upload` skončí s chybou (L-Bone server nedostupný) a je nutno jej opakovat (při opakovaném zadání bude vybrán jiný sklad).
  Volba `-h` znamená, že se bude jednat o perzistentní uložení (server nesmí data smazat ani v případě nedostatku místa pro nová data).
  Volba `-d` udává, jak dlouho si přejeme data na skladech uchovat. Lze použít násobky dnů (d), hodin (h) a minut (m). Horní limit alokací je námi nastaven na 10000 dnů, ale pro běžný provoz počítáme s jeho významným snížením.
  Volba `-c` udává počet kopií. Bohužel v současné době LoRS vrstva neumí zaručit umístění kopií stejného bloku na různé sklady. LoCI laboratoř slibuje brzké vydání nové verze, která tuto vlastnost zaručuje. To znamená, že při dvou kopiích má uživatel jistotu dostupností dat i při úplném výpadku jednoho skladu.
  Příkaz `lors_upload` podporuje množství dalších voleb, které jsou popsány v nápovědě, dostupné přes volbu `--help`.
- `lors_download` slouží ke stažení dříve uloženého souboru.
  Příklad použití je:
  ```
  lors_download soubor.xnd -o soubor
  ```
  Volba `-o` specifikuje jméno výstupního souboru. V případě neuvedení je soubor vypisován na standardní výstup.

Příkaz `lors_download` podporuje množství dalších voleb, které jsou popsány v nápovědě (volba `--help`).

- `lors_trim` slouží ke smazání uložených dat. Příklad použití je:
  `lors_trim -d soubor.xnd`
  Samotný soubor `soubor.xnd` nebude z lokálního disku smazán, ale data, která popisuje, budou neplatná. V případě existence více kopií budou smazány všechny kopie současně. Tato operace je nevratná.
- `lors_ls` slouží k výpisu stavu uložených dat. Lze zjistit, jak dlouho budou data ještě přístupná. Pokud se u některé části dat objeví slovo *unknown*, znamená to, že příslušný server buď není v provozu nebo data již expirovala.

Utility jsou k dispozici na adrese `http://loci.cs.utk.edu/modules.php?name=Downloads&d_op=viewdownload&cid=5`.
V době psaní článku byla nejstabilnější verze 0.82. Verze pro Windows nepodporuje soubory větší než 2 GB.

3.2 GUI utility

Pro snazší práci bylo v rámci projektu vytvořeno grafické prostředí pro ukládání a stahování souborů. Aplikace je psaná v Javě a lze ji získat na adrese `http://undomiel.ics.muni.cz/presentation/download/JavaLors.jar`.
Pokud není *.jar* asociovaná přípona, lze aplikaci spustit `java -jar JavaLors.jar`.

Pokud chceme soubor nahrát na datová úložiště, vybereme jej jako vstupní soubor. Je možné vybrat jméno výstupního souboru, pokud není žádné vybráno, použije se vstupní soubor rozšířený o příponu *.xnd*. Zvolíme dobu uložení, počet kopií a můžeme soubor uložit, ostatní položky není nutné měnit. Jde o obdobu některých voleb utilit příkazového řádku.

Stažení souboru je jednodušší. Jako vstupní soubor zvolíme nějaký uložený *.xnd* soubor, vybereme jméno výstupního souboru a můžeme soubor stáhnout a uložit na lokální disk.

V obou případech se po skončení objeví okno oznamující úspěšnost přenosu.

3.3 Webový přístup

Další možností přístupu k úložišti je přes webový prohlížeč. Na adrese `http://didas.ics.muni.cz/cgi/` je rozhraní k souborovému systému úložiště. Toto rozhraní je stále ve vývoji, v současné době umožňuje pouze stažení souborů, které byly již dříve uloženy některým z výše popsaných způsobů. Hlavním omezením je velikost souboru, neboť většina prohlížečů neumí pracovat se soubory většími než 2 GB.

4 Nové aplikace

Součástí projektu DiDaS byla i tvorba resp. modifikace vhodných aplikací tak, aby byly schopny přímo pracovat se soubory uloženými v datových skladech, případně výsledky zpracování do datových skladů přímo ukládat.

Pro tento účel jsme vytvořili knihovnu `libxio`, která poskytuje standardní unixové rozhraní pro práci se soubory. Místo běžných systémových volání (`open`, `read`, `write`, `close`) je možné použít operace s prefixem `xio_` (tedy `xio_open`, `xio_read`, `xio_write`, `xio_close`) a aplikaci přeložit znovu s knihovnou `libxio` a knihovnami z LoRS balíčku.

Soubory uložené v datových skladech jsou zpřístupněny přes URI notaci, která umožňuje explicitně zadávat volby odpovídající jednotlivým parametrům příkazu `lors_upload`. Použitá syntaxe je následující: `lors://didas.ics.muni.cz/cesta/soubor?bs=cislo&duration=cislo&copies=cislo&threads=cislo&timeout=cislo&servers=cislo&size=cislo`

Přesný popis lze najít v technické zprávě `http://www.cesnet.cz/doc/techzpravy/2003/ibpdidas/` a v dokumentaci ke knihovně *libxio* na adrese `http://undomiel.ics.muni.cz/presentation/doc/libxio.html`. Samotnou knihovnu `libxio` je možné nalézt na adrese `http://undomiel.ics.muni.cz/presentation/projects.html`.

4.1 Zpracování videa

Hlavní testovací aplikací je překódování záznamů přednášek Masarykovy univerzity v Brně. Tato aplikace a související úprava programu

`transcode` byla podrobně popsána v jednom z předcházejících čísel Zpravodaje ÚVT[1].

Další modifikovanou aplikací je přehrávač záznamů pro Linux – Mplayer. Pomocí výše uvedené reprezentace souborů je Mplayer schopen přímo přehrávat (s podporou přetáčení) záznamy uložené v datových skladech.

Aplikace včetně dokumentace jsou k dispozici na adrese `http://undomiel.ics.muni.cz/presentation/projects.html`.

5 Další vývoj

Je zřejmé, že nutnost uchovávání metadat nedělá dosavadní způsob ukládání příliš pohodlným ani efektivním. Náš další vývoj proto směřuje k integraci distribuovaného indexu. To znamená, že uživatel bude schopen vidět adresářovou strukturu přímo v úložišti a bude moci se soubory přímo manipulovat.

V první fázi plánujeme integraci distribuovaného indexu do grafického nástroje v Javě. Následovat by měla integrace do knihovny `libxio` a jejím prostřednictvím do aplikací.

Paralelně pracujeme na vytvoření souborového systému pro operační systém Linux. Souborový systém zpřístupní data v datových úložištích formou souborů dostupných přímo standardními nástroji Linuxu. Bude podporovat koexistenci různých verzí téhož souboru, s možností práce pouze s nejaktuálnější nebo libovolnou verzí. Starší verze budou automaticky mazány v závislosti na době expirace.

6 Shrnutí

V rámci projektu DiDaS bylo vybudováno distribuované úložiště dat, integrované do vysokorychlostní sítě CESNET2. Pilotní projekty, spojené zejména s rozsáhlým zpracováním videa ze záznamů přednášek potvrdily stabilitu a robustnost celého řešení. V současné době jsou datové kapacity (celkem 15 TB) přístupné uživatelům projektu MetaCentrum a postupně budou zpřístupněny i široké akademické veřejnosti. Hlavní překážkou je v současné době plná anonymita uživatelů, která komplikuje kontrolu dodržování pravidel provozu akademické sítě CESNET2. Zejména nelze zábránit zneužití pro komerční či nelegální činnost. Tento problém bude vyřešen zavedením autentizace a autorizace přístupu k datovým skladům a rovněž plnou integrací obsahu, jak ji zmiňuje předchozí kapitolka.

Distribuované datové úložiště bude využitelné především jako dočasná datová kapacita – mezivýsledky rozsáhlých výpočtů, výsledky rozsáhlých experimentů před jejich dalším zpracováním, rozbalené archivy, atd. Datová úložiště bude ale možno využívat také pro rozložení zátěže při přístupu k často stahovaným datům. Na datová úložiště bude možné nahrát např. kompletní distribuci nové verze Linuxu a uživatelům prostřednictvím webového rozhraní zpřístupnit pouze Ex-node soubory. Vytvoření vícenásobných kopií přitom umožní rozložit zátěž při stahování přes několik datových skladů a tím snížit jak jejich zatížení, tak i zatížení páteřních linek.

Očekáváme rovněž postupný vznik dalších aplikací – např. služby pro zasílání velkých souborů mailem. Velký soubor (i několik GB) je možné uložit do datového úložiště a e-mailem poslat pouze soubor s Ex-node daty (zpravidla soubor s příponou `.xnd`), jehož velikost při přenosu e-mailem nedělá žádné problémy. Adresát si pak může soubor sám stáhnout (resp. může to přímo udělat vhodný plugin jeho e-mailového klienta). □

[1] `http://www.ics.muni.cz/bulletin/issues/vol14num05/holub/holub.html`