Distribuované filesystémy a Linux

Miroslav Ruda
http://www.fi.muni.cz/~ruda

Superpočítačové centrum
Ústav Výpočetní techniky
Masarykova Univerzita, Brno

Pokročilé distribuované filesystémy

Jednotný jmenný prostor – znalost o umístění souborů obsahují servery, není na klientech, databáze umístění dat (volumů) je distribuovaná

Replikace dat – některá data mohou být zreplikována

Cache – klient používá cache, zrychlení opakovaných přístupů

Zvýšená bezpečnost – pro autentizaci se často používá systém Kerberos

Autorizace – možnost definovat přístupová práva v ACL

Heterogenní prostředí – různé operační systémy, spolupráce s NFS

Mobilní počítání – disconnected operace

Správa systému – musí podporovat správu rozsáhlých systémů

Dostupné implementace

NFS

- mapování jména souboru na umístění na serveru je statické, neexistuje možnost dynamické změny, musí se požívat automounter, jeho mapy jsou na filesystému nezávislé
- nepodporuje replikace – jedině automounter
- stateless server
- close-to-open konzistence
- pro autentizaci se používá obyčejná unixová autentizace – UID

NFS verze 3

- definuje možnost asynchroního zápisu
- podporuje soubory větší než 4 GB
- několik dalších úprav pro zrychlení
- stále pouze weak-cache konzistence
- URL:`http://www.netsys.com/unix/NFSv3_Rev_3.eps`

Secure NFS

- rozšíření normy NFS navržené firmou Sun
- definuje autentizaci typu UNIX, DES nebo Kerberos
- používá Secure RPC
- DES verze používá public kryptografii na výměnu DES klíčů, používá NIS
- Kerberos verze používá ticket a session key z Kerbera
- dostupné minimálně pro Solaris a AIX, existuje i implementace pro Linux
- URL:`ftp://ftp.nada.kth.se/pub/krb/doc/nfs.ps`
- URL:`ftp://ftp.muni.cz/pub/drafts/draft-eisler-nfssec-01.txt`

Kerberised NFS

- NFS rozšířené o autentizaci pomocí Kerbera
- jiný protokol než v Secure NFS, v podstatě existuje stroj, kterému se dokáže, že mám ticket a tento stroj za mě přistupuje k diskům
- používá KTH Kerberos
- NFS filter machine, nfsuid démon, kinit, nfsuid
- možnost použít i Hesiod
- URL:`ftp://ftp.monash.edu.au/pub/lknfs/1.0/README`

TCFS (Transparent Cryptographic FileSystem)

- linuxová implementace kryptovaného filesystému
- rozšíření NFS o možnost kryptovat jednotlivé soubory
- používá attributy z ext2fs
- URL:`http://www.itplanet.it/~ermmau/tcfs/`

AFS

- původně Andrew File System vyvinutý na Carnegie Mellon University, nyní vyvíjený a prodávaný firmou Transarc pod jménem AFS
- Irix, SunOS, Solaris, Digital Unix, Ultrix, AIX, HP-UX, NCR
- klienti i na Windows NT, NetBSD
- pro Linux prozatím klient, v nově ohlášené verzi i server
- jednotný jmenný prostror `/afs`
- podporuje replikaci read-only dat, klient si vybírá nejbližší dostupnou kopii
- používá velkou (většinou diskovou) cache
- callback – server se zavazuje informovat o změně dat
- pro autentizaci se používá Kerberos
- ACL – seznam přístupových práv pro adresáře
- práva read, write, lookup, lock, insert, delete, administrate
- systémové skupiny, každý uživatel si také může vytvářet svoje skupiny a volně je používat v AFS
- @sys se při přístupu do AFS překládá na jméno platformy
- AFS/NFS translator

- jednotná databáze umístění souborů, ACL a autentizačních informací
- backup, rekonfigurace a rutinní správa bez výpadků
- administrace možná z libovolného stroje
- nástroje na monitorování serverů i klientů
- podpora geograficky rozsáhlých systémů – decentralizace
- nemá příliš velké nároky na velikost paměti
- URL:`http://www.dementia.org/~shadow/afs.html`
- URL:`http://www.transarc.com/`

Rozšíření

- disconected AFS
- MR-AFS

Problémy

- data jsou dostupná jen pomocí AFS, lokálně nejsou čitelná
- pomalejší přístup do lokálních filesystémů
- nutnost synchronizovat čas na všech strojích
- složitější systém než NFS
- komerční řešení
- na PC zatím pouze klient, ne server (neplatí u AFS 3.5 !)

Arla

- volně dostupná implementace AFS klienta pro *BSD, Linux, Solaris, SunOS
- částečně i Irix, Digital Unix, AIX
- zatím není příliš stabilní, neumí memory cache, cachuje pouze celé soubory do velikosti cache
- vyvíjeno s Kerberem z KTH, lze použít i MIT Kerberos
- URL:`http://www.stacken.kth.se/projekt/arla/`

Coda

- následník AFS vyvíjený opět na CMU, hlavní motivací je podpora „Mobile Computing"
- autentizace podobná systému Kerberos, RPC2 – bezpečná autentizace a bezpečné komunikační kanály
- nové vlastnosti
  - „disconnect operace" pro mobilní klienty – hoaring
  - reintegrace dat po následném připojení, řešení konfliktů vzniklých v době odpojení
  - replikace i read-write dat, read once - write all
- součást Linuxového jádra 2.1, dostupné i pro *BSD
- větší nároky na paměť, (prozatím) pomalejší a méně stabilní
- URL:`http://www.cs.cmu.edu/afs/cs/project/coda/Web/coda.html`

DFS

- následník AFS, vyvíjený firmou Transarc jako součást DCE/DFS
- DCE (Distributed Computing Enviroment) definováno OSF (Open Software Foundation)
- DCE Secure Core

  Threads – založeno na POSIX 1003.1c normě, draft 4
  Remote Procedure Call (RPC)
  Directory Services (CDS) – založeno na X.500
  Security Service – Kerberos 5
  Time Service – NTP
- lepší spolupráce s lokálním filesystémem – do DFS světa lze nabízet i lokální filesystém, rychlejší přístup do lokálních filesystémů
- ACL pro každý soubor
- AFS callback je zobecněn na token-passing – lepší emulace UNIXové sémantiky
- dostupné na Solaris, Digital Unix, HP-UX, Irix, AIX; bez DFS existuje pro řadu dalších systémů
- existuje linuxový (spolu s *BSD) port (DCE 1.2.2 je dostupné ve zdrojácích) klientské části DCE, DFS neexistuje
- URL:`http://www.opengroup.org/tech/dce/`
- URL:`http://www.bu.edu/~jrd/FreeDCE/`
- URL:`http://www.transarc.com/`

Kerberos

Kerberos 4
- MIT – původní implementace, dnes už neudržovaná
- KTH Kerberos – verze vyvinutá v PDC KTH ve Švédsku, dnes asi nejpoužívanější Kerberos 4
- URL:`http://www.pdc.kth.se/kth-krb/`

AFS Kerberos
- součást AFS
- původně Kerberos 4 implementovaný na CMU, má doplněny některé vlastnosti Kerbera 5
- v AFS ho lze nahradit KTH Kerberem, MIT V5, Heimdalem

MIT Kerberos 5

- další kryptovací algoritmy (triple-DES), modularita
- podpora hierarchických realmů
- forwadable, renewable, proxiable tickety
- reply-cache
- ASN.1 kódování zpráv
- GSS-API
- URL:`http://web.mit.edu/kerberos/www/`

Heimdal
- implementace Kerbera 5 z PDC KTH
- hodně nová implementace, ale základ už je funkční
- URL:`http://www.pdc.kth.se/heimdal/`

DCE Secure daemon

- základ je Kerberos 5
- je to nadmnožina Kerbera 5, navíc podporuje komunikace přes RPC a poskytuje Privilege Service

Sesame
- projekt podporovaný EU
- původně založený na Kerberovi 5, dnes vedle Kerbera podporuje i autentizaci pomocí metod asymetrické kryptografie
- navíc obsahuje Privilege Service a nástroje pro distribuci klíčů
- podpora pro Solaris, HP-UX, AIX, Linux
- URL:`http://www.esat.kuleuven.ac.be/cosic/sesame/`

Příklady použití

- učebna s linuxovými stroji
- robusní webserver
- *METACentrum*

Robusní webserver

- více počítačů – duplikovaných webserverů
- data typu read-only – replikována mezi několika fileservery, zrychlení přístupu k datům, zajištění konzistence dat
- data typu read-write – zrychlení pomocí cache, zajištění konzistence dat

Učebna

- sdílené home adresáře – AFS, nejsou nutné AMD mapy na klientech a exportování na serverech
- bezpečná autentizace, po síti nejdou hesla, kryptovaná spojení, časově omezená platnost ticketů
- autorizace pro přístup do adresářů, lokální root nemá přístup k cizím home adresářům
- ACL
- šíření `/etc/shadow` – nemusí být, řeší Kerberos
- šíření `/etc/passwd (+shadow)`
  - kopírování pomocí rcp/ssh
  - přes AFS – nic tajného tam není !
  - Hesiod - rozšíření DNS
- `/usr` montovaný po síti, AFS cache
- další vlastnosti – RO kopie, backupování
- `/` může být generovaný

META Centrum

Využití principu projektu `Athena` na MIT
(též projekt `Orion` na ZČU Plzeň)

Uzly *META Počítače*:

- ÚVT UK Praha:
  SGI Power Challenge XL (12 CPU R10000)
  SGI Challenge S (1 CPU R4400)
  2× Digital AlphaServer 2100 (à 2 CPU každý)
- VC ČVUT Praha:
  IBM SP2 (19 CPU)
- ÚVT MU Brno:
  SGI Origin 2000 (32 CPU R10000)
  SGI POWER Challenge XL (12 CPU R10000)
- CVIS VUT Brno:
  SGI POWER Challenge L (8 CPU R10000)
- ZČU Plzeň:
  Digital AlphaServer 8400 (8 CPU)

Celkový výkon	≈90 procesorů
	≈ 38 GFLOPS
Celková paměť	≈ 25 GB
Disková kapacita	≈650 GB

META Centrum technické řešení

Základem distribuovaný systém souborů AFS:

- Dostupný na heterogenní základně
- Určený pro geograficky rozsáhlé systémy
- Zajišťuje zvýšenou bezpečnost dat
- Relativně snadná správa – důsledné oddělení centrálních a lokálních informací
- Akceptovatelná režie
- Každý uzel je samostatnou buňkou AFS

Autentizaci uživatelů zajišťuje systém Kerberos

- Umožňuje lokální autentizaci přenášet na vzdálené počítače
- Nabízí zabezpečení síťového spojení šifrováním
- Integrován do AFS
- Každý uzel je samostatnou říší (realm)
- Jednotlivé uzly si důvěřují