MetaCentrum

CESNET

Souborové systémy a práce s daty

David Antoš

antos@ics.muni.cz

MetaCentrum

CESNET

Úvod

- obecný úvod do síťových souborových systémů
- souborové systémy v MetaCentru
  - jejich použití
  - práce s nimi
  - praktické poznámky
    - ⋆ kvóty
    - ⋆ efektivní použití
- do prezentace přispěl
  - Lukáš Hejtmánek (klasifikace souborových systému a graf)

Sdílené systémy souborů

- přístup přes síť
- předpokládáme mnoho klientů, kteří přistupují současně
- proč to
  - kapacita, výkon, potřeba data sdílet
  - většinou standardní rozhraní (POSIX) využitelné běžně aplikacemi
  - uživatelé jsou zvyklí používat souborový systém
- v prostředí MetaCentra jsou síťové systémy „transparentní“
  - nicméně to občas vede na nevhodné použití

Obecnější úvod do souborových systémů

- síťové souborové systémy dělíme
  - centralizované
    - ⋆ obvykle jeden server, obtížně škálovatelné
    - ⋆ single point of failure
    - ⋆ jednoduchý design
    - ⋆ NFSv3, v4, Samba/CIFS
  - distribuované
    - ⋆ realizované skupinou serverů
    - ⋆ komplikovaný design, potenciálně(!) výkonnější
    - ⋆ AFS, Lustre, PVFS2, GPFS, GlusterFS, ...

Obecnější úvod do souborových systémů II

- autentizace
  - slabá (UID/GID) vs. silná (např. Kerberos)
    - ⋆ určuje potřebu důvěryhodnosti klienta
- stavovost
  - nutná synchronizace při pádu klienta nebo rozpadu spojení
- nutnost řešit konzistenci dat
  - souběžný zápis několika klienty
    - ⋆ obvykle zamykání

MetaCentrum

CESNET

Souborové systémy v MetaCentru

- inventura:
  - rychlý lokální disk /scratch
  - síťový scratch přes PVFS
  - sdílený síťově připojený přes NFSv3 na lokálním clusteru
  - sdílený síťově připojený přes AFS na všech strojích
  - sdílený síťově připojený přes NFSv4 na téměř všech strojích
- každý se hodí na něco jiného, ale počet chceme omezit
- při výběru vhodného souborového systému používejte
  - pravidla MetaCentra
  - selský rozum ;) – na to je ale třeba vědět, jak to funguje

Svazky /scratch

- svazky /scratch jsou malé a rychlé
- slouží k ukládání dočasných dat výpočtu
  - tj. mezivýsledků běžící úlohy
  - nebo výsledků do doby ukončení úlohy
    - ⋆ z toho plyne: úloha má končit tím, že odsune svá data (stage-out) ze scratche
- současná praxe: zasahuje se při problému
- /scratch není zálohován
- experimentuje se se síťovým Lustre jako náhradou lokálních disků
  - přes IP nebo InfiniBand

Svazky /home

- vždy exportovány na jeden cluster
- postaveny na NFSv3
  - slabá autentizace (tj. nelze exportovat na pracovní stanice uživatelů)
  - výkon verze 3 není oslnivý
- slouží k uchování dat, se kterými uživatelé aktuálně pracují
  - kde „aktuálně“ znamená, že na ně v posledním roce přistoupili

MetaCentrum

CESNET

Svazky /afs

- dostupné a sdílené na všech strojích MetaCentra
- relativně pomalé, zato s velkou variabilitou nastavování ACL
- slouží k ukládání dat projektů, lze vytvářet svazky pro projekty
- kvóty
- nedoporučuje se provozovat svazky nad 2 GB
- je možno připojit k vlastní pracovní stanici
- přístup přes Kerberos

Svazky /storage (NFSv4)

- použití jako AFS, jen rychlejší
- „neomezená kapacita“
  - standardní kvóta 5 TB
  - fyzická kapacita 96 TB
- dostupné na strojích s vlastností `nfs4`, možno připojit k vlastní pracovní stanici
- přístup přes Kerberos
  - řešeny problémy s lístky (nemožnost revokace) a kvótami (špatně zobrazeny)
- server umístěn v Brně – výkonnostní důsledky

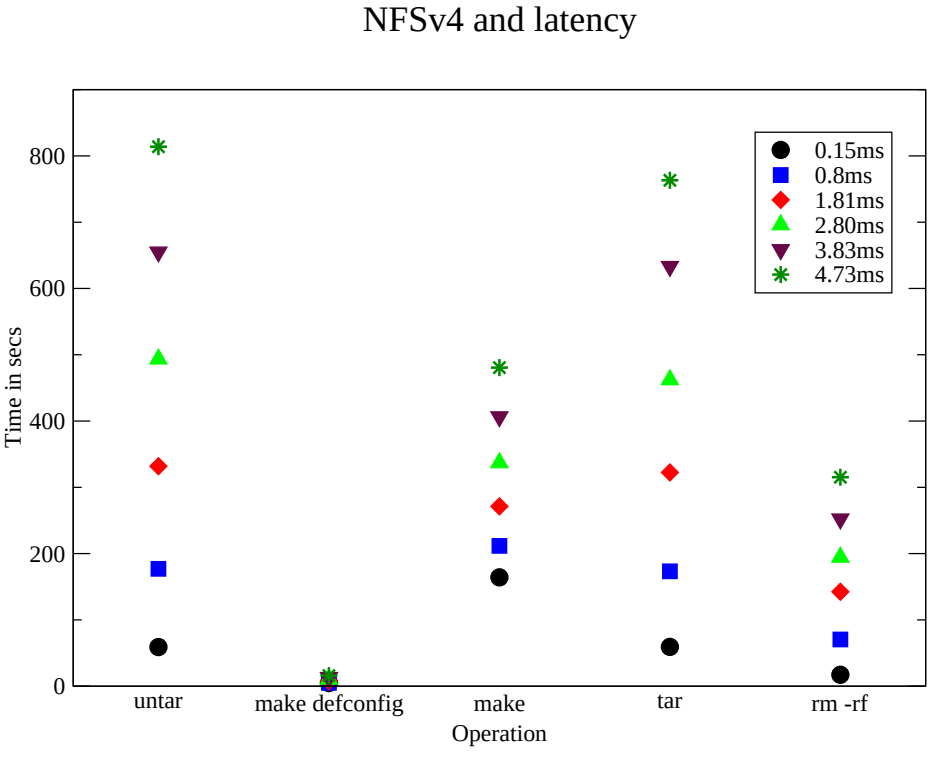
Kvóty

- slouží jako mechanismus hlídání dostupné kapacity pro správce
- a jako ochrana proti zaplnění celého prostoru chybou úlohy
- lze žádat kdykoli o změnu
  - `http://meta.cesnet.cz` – Můj účet – Souborové systémy – dole schované tlačítko „Požádat o změnu kvóty“

MetaCentrum

CESNET

Vliv geografické vzdálenosti na výkon



MetaCentrum

CESNET

Kterak sobě data ku clusteru nakopírovati

(a pak zpět domů)

- pro drobná data (desítky MB) je to jedno
  - přes frontend
- scp
  - pro větší data na interaktivní uzel
  - `qsub -I -l nodes=1:nfs4:brno`
- sftp na libovolný stroj
  - interaktivní uzel
- máte-li velká data, domluvte se předem

Kterak dopraviti data úloze do scratch

- do lokálního scratche
  - ze /storage: cp sem a tam
  - z /home: stage-in/stage-out skripty PBS

```
#PBS -W stagein=/scratch/pepa/vstup1@skirit:vstup1,
        /scratch/pepa/vstup2@skirit:vstup2
#PBS -W stageout=/scratch/pepa/vystup@skirit:vystup
cd /scratch/pepa
```

formát je následující

```
místní_soubor@jméno_stroje:vzdálený_soubor[,...]
```

MetaCentrum

CESNET

Diskuse

- kolik úrovní filesystémů
  - storage a scratch?
  - je zajímavé optimalizovat?
- které aplikace mají problém?
  - mám velký soubor a potřebuju zobat kousek?
  - masivní paralelismus nad velkým souborem?
- speciální požadavky aplikací?
- metadata? katalogy? databáze? OGSA-DAI? . . . ?