

Správa dat v gridových systémech

Systemy souborů

Lukáš Hejtmánek
Cesnet, z.s.p.o
Zikova 4, 160 00 Praha
xhejtman@ics.muni.cz

Sdílené systémy souborů

- Co je sdílený systém souborů?
 - Přístup přes síť
 - Předpokládáme větší množství klientů
 - Mohou přistupovat současně k uloženým datům
- Proč uvažujeme sdílené systémy souborů?
 - Lokální úložiště obvykle nestačí kapacitou i výkonem, neumožňují data sdílet
 - Rozhraní standardního (POSIX) souborového systému využije každá aplikace
 - Uživatelé jsou „zvyklí“ používat souborový systém

Sdílené systémy souborů

- Sdílené (síťové) souborové systémy dělíme
 - Centralizované
 - Realizované obvykle jedním serverem
 - Problémy s výkonem
 - Single-point-of-failure
 - Jednoduchý design
 - NFS (verze 3, 4), Samba/CIFS
 - Distribuované
 - Realizovaná skupinou serveru
 - Potenciálně výkonnější
 - Komplikovaný design
 - AFS, Lustre, PVFS2, GPFS, GlusterFS, ...

Sdílené systémy souborů

- Autentizace
 - Ověření identity uživatele
 - Slabá
 - Identity pomocí číselných UID/GID
 - Silná
 - Identity jsou ověřovány např. systémem Kerberos
- Stavovost/nestavovost
 - Stavový systém souborů přechází ze stavu do stavu
 - Nutná synchronizace stavu při
 - Pádu klienta nebo serveru
 - Rozpojení a znovu navázání spojení
 - Nestavový
 - Každá operace je izolovaná bez návaznosti

Sdílené systémy souborů

- Sdílené systémy souborů musí řešit konzistenci dat
 - Souběžné aktualizace dat několika klienty
 - Obvykle řešeno zamykáním
 - U distribuovaných systémů souborů – protokoly pro distribuované zamykání
- Komunikace mezi klienty a servery
 - Obvykle lze použít TCP/IP
 - Některé systémy podporují pokročilejší možnosti komunikace
 - InfiniBand (RDMA) – vysoká rychlost, nízká latence

Sdílené systémy souborů a Gridy

- Co požadujeme po sdíleném systému souborů v Gridech?
 - Závisí na oblasti použití
 - Obecně však preferujeme distribuované systémy
 - Obvykle rozlišujeme
 - Obecné úložiště
 - Trvanlivé (zálohované), se silnou autentizací
 - Domovské adresáře uživatelů
 - Oblast pro data uživatelů (a-la storage element)
 - Přístupné odkudkoliv (ne jen uvnitř Gridu)
 - Heterogenní prostředí může být problém
 - Úložiště pro výpočty
 - Data na něm uložená mají dočasný charakter
 - Obvykle po dobu běhu úlohy
 - Nezálohované, důraz na rychlost
 - Přístupné jen uvnitř Gridu / uvnitř clusteru

Souborový systém NFS

- Centralizovaný síťový souborový systém
- Hlavní verze 2, 3, 4
- Verze 2
 - První masivně rozšířená verze
 - Bezstavová
 - Restart klienta nebo serveru nepřinášel žádné problémy
 - Nepodporovala zamykání
 - Autentizace založena na důvěryhodných klientech
 - UID/GID
 - Nabízí lokální jmenný prostor
 - Různí klienti mohou vidět sdílený svazek různě

Souborový systém NFS v. 3

- Přidává podporu zámků
 - Doplnkový protokol, podpůrné aplikace (rpc.lockd)
 - Zavádí stav – zámek
- Některé implementace podporují silnější autentizaci
 - K dispozici Kerberos a SPKM
 - Klienti nemusí být důvěryhodní
- Experimentální implementace nabízí podporu ACL

Souborový systém NFS v. 4

- Poslední hlavní verze
- Stavový protokol
- Standardní součástí je autentizace pomocí GSS API
 - Nabízený Kerberos, SPKM
- Zavádí globální jmenný prostor
- Kromě standardních protokolů TCP/IP podporuje InfiniBand (RDMA)

Souborový systém NFS

- Ačkoliv nejde o distribuovaný systém
 - Na řadě míst se používá pro domovské adresáře uživatelů
- Díky podpoře GSS API je možné systém otevřít i mimo Grid
- Výkon nemusí být dostačující pro velké množství uživatelů
 - Některé firmy nabízí clusterované NFS
 - Velké diskové pole
 - Několik NFS serverů
 - Clusterovaný souborový systém (GPFS, CXFS)
- Je podporován řadou platforem a architektur

Souborový systém NFS

- Výhled do budoucna
 - Stále centralizovaný síťový systém
 - Verze 4.1
 - Podpora pNFS
 - Částečně distribuovaná verze NFS
 - Rozdělení na
 - Metadata (informace o uložených souborech)
 - Data (samotný obsah souborů)
 - Metadatový server zůstává jeden
 - Datových serverů může být více

Souborový systém Lustre

- Distribuovaný souborový systém
- Dělí ukládání dat na
 - Metadata – informace o souborech – ukládané centrálně
 - Data – vlastní data souborů – ukládána distribuovaně
- Data i metadata jsou ukládána do diskových oddílů (ve formátu blízkému ext3/ext4)

Souborový systém Lustre

- Autentizace
 - Do verze 1.6 pouze slabá (UID/GID)
 - Od verze 1.6 i silná (Kerberos)
(současná stabilní verze je 1.8)
- Vedle TCP/IP protokolu podporuje Infiniband (RDMA)

Souborový systém Lustre

- Podpora pouze pro systémy Linux
 - Navíc pouze konkrétní jádra
- Obvyklé využití je pro dočasná data výpočtů
- Existují velké instalace Lustre
 - 50 serverů pro data souborů, více jak 1000 klientů
 - TACS University USA
 - Úložná kapacita v řádu PB
- Budoucí nová verze 2.0 bude obsahovat spíše interní změny, nemá významné nové vlastnosti

Souborový systém PVFS2

- Distribuovaný souborový systém
- Dělí ukládání dat na
 - Metadata a Data
 - Obě části lze ukládat distribuovaně
- Data a metadata jsou ukládána do existujícího systému souborů
 - Pomocí speciální databáze Trove
- Autentizace
 - Pouze slabá pomocí UID/GID

Souborový systém PVFS2

- Komunikace mezi klienty a servery
 - TCP/IP (přidali jsme podporu IPv6)
 - InfiniBand
 - GM, a další
- Systém je dostupný pro operační systém Linux
 - Podporuje většinu jader
- Klienti nemají zápisovou cache
 - Degradace výkonu při použití malých datových bloků
- Využití pro dočasná data výpočtů

Souborový systém GlusterFS

- Experimentální distribuovaný souborový systém
- „Překryvový“ souborový systém
 - Data jsou ukládána jako soubory
 - GlusterFS z nich vytváří jednotný síťový souborový systém
 - Lze nastavovat pravidla – replikace, spojování úložišť, preference úložiště
- Autentizace řízena pomocí UID/GID

Souborový systém GlusterFS

- Komunikace mezi klienty a servery
 - TCP/IP
 - InfiniBand
- Systém je dostupný pro operační systém Linux
 - Podporuje většinu jader
 - Využívá FUSE modul jádra
 - Součástí je patch na zvýšení výkonu
- Klient umožňuje využití cache
 - Vyšší výkon proti PVFS2
- Různorodé využití

Praktické zkušenosti v národním Gridu MetaCentrum

- Historie
 - AFS svazek pro distribuované domovské adresáře
 - Poskytuje silnou autentizaci
 - Nevyhovovalo výkonově
 - NFSv3 domovský svazek, pouze pro daný cluster
 - Použita slabá autentizace, verze 3 není příliš výkoná
 - Lokální oblasti pro dočasná data výpočtů

Praktické zkušenosti v národním Gridu MetaCentrum

- Nasazení NFS verze 4
 - Náhrada AFS
 - Velkokapacitní úložiště sdílené přes celý Grid
- Nasazení Lustre
 - Pokusné nasazení síťových oddílů pro dočasná data

Nasazení NFS v. 4

- Podporuje silnou autentizaci (Kerberos)
- Díky rozšířením protokolu poskytuje vyšší výkon oproti verzi 3
- Pro transport umožňuje využít technologii InfiniBand
- Nasazeno jako alternativa domovských adresářů uživatelů přes celý Grid

Nasazení NFS v. 4

- Několik problémů s nasazením
 - Špatné vyhledávání lístků uživatelů
 - NFS občas vybíralo špatné lístky
 - Podařilo se opravit
 - Nemožnost revokace lístku
 - Autentizace uživatele je zrušena až expirací lístku
 - Vyřešeno násilným omezením platnosti lístku
 - Vhodné řešení pomocí Keyringu/pagsh je ve vývoji
 - Příliš velké kvóty
 - Protokol kvót nepočítá s kvótami o velikosti 5 TB
 - Problémy s latencí
 - Některé programy (gaussian) neběží nad NFS

Nasazení Lustre

- Náhrada lokálních disků
 - Výpočty mohou používat rychlé sdílené úložiště
 - Větší kapacita
 - Podpora pro virtualizaci – snadná preempce
- Paralelní souborový systém
 - Navržen pro obsluhu velkého množství klientů

Nasazení Lustre

- Problémy s nasazením
 - Podpora jádra Linuxu max. 2.6.22
 - Obsahuje řadu chyb
 - Klient ve verzi 1.8 způsoboval pády aplikaci
 - Verze 1.6 byla bezproblémová
 - Nepříliš vhodný pro velké množství malých souborů
 - Vždy se přenáší 1 MB dat

Experimenty

- Pro český vyhledávač jsme testovali GlusterFS ve spojení s InfiniBand
 - Problémy se stabilitou a pamětí
- Plány na nasazení GPFS
 - Jediný souborový systém použitelný jako úložiště pro Sambu – nemá problémy se zámky

Vliv latencí na souborový systém

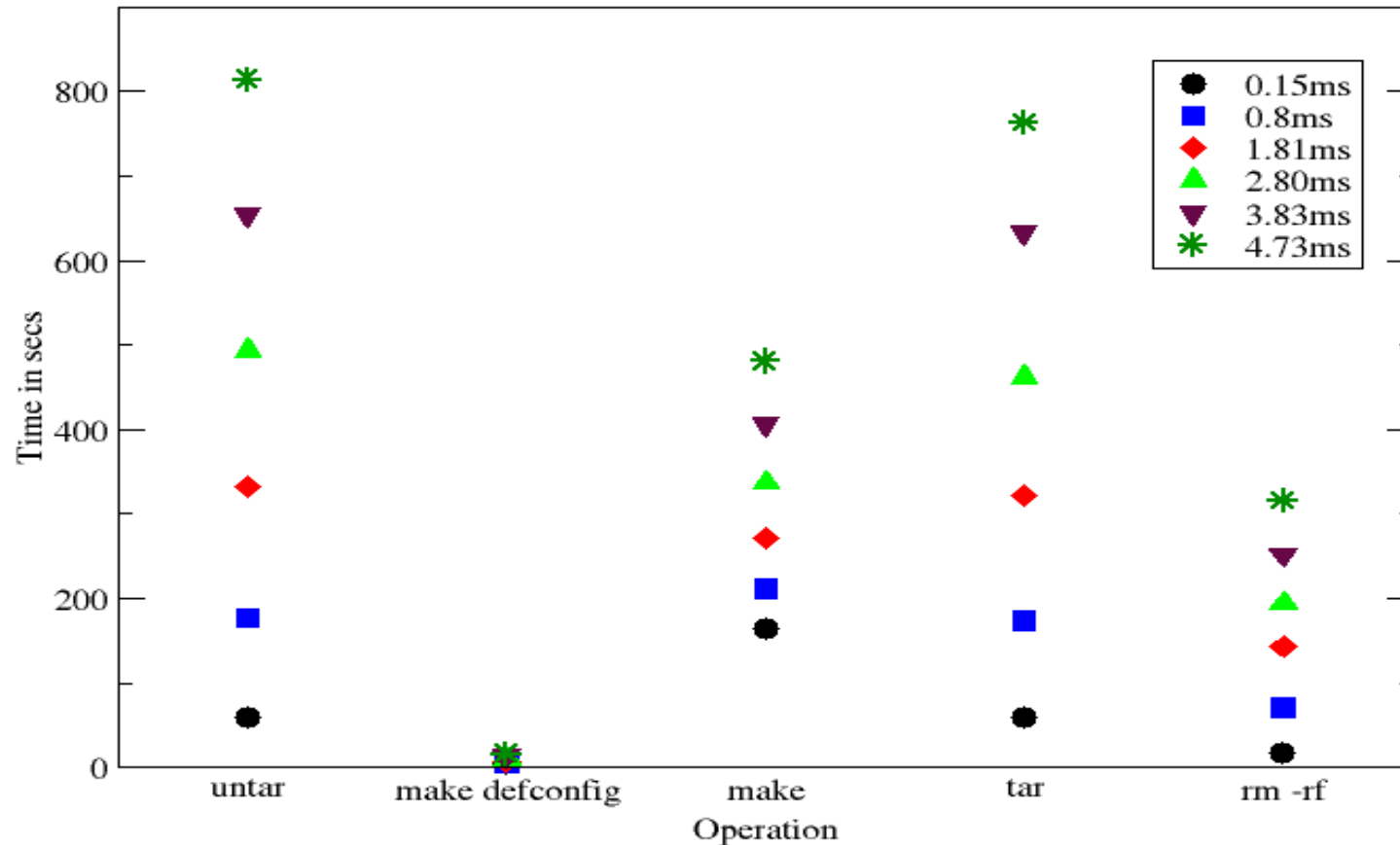
- Rozsáhlé Gridy jsou charakteristické vyšší latencí
- V rámci jediné organizace (v jednom městě) latence do 1.5 ms
- Latence na území ČR do 5 – 7 ms
- Latence má zásadní vliv na chování souborového systému
 - Problematické jsou synchronní operace
 - Doba trvání synchronní operace odpovídá latenci sítě
 - Např. založení souboru

Vliv latencí na souborový systém

- Série měření na souborovém systému NFS v. 4
 - Nasazujeme v MetaCentru jako globální síťový systém
 - Server v Brně, klienti v Brně, Praze, Plzni
 - Latence od 0.15 ms do 4.73 ms
 - V měření jsme se zaměřili na operace, které latence ovlivňuje nejvíce
 - Manipulace s malými soubory

Vliv latencí na souborový systém

NFSv4 and latency



Děkuji za pozornost