Skalierbarer Fileservice mit pNFS

Dipl.-Chem. Rainer Orth
Technische Fakultät
Universität Bielefeld
ro@TechFak.Uni-Bielefeld.DE

Übersicht

- Das Problem: skalierbarer Fileservice jenseits einzelner Server
- Bisherige proprietäre Lösungen
- Jetzt mit Standards: Parallel NFS (pNFS)
- Storage Server: Block/Volume, Files und Object Storage

Skalierbarer Fileservice: das Problem

- Fileserver halten mit Wachstum von Clientanforderungen (z.B. Compute-Cluster) und Storage-Kapazität (SANs) nicht mit
- einzelner Server als Bottleneck
- aber: manuelle Verteilung auf mehrere Server unbefriedigend oder unwirksam: erfordert Anpassungen der Clientsoftware/-konfiguration, nicht handhabbar für Einzelfiles

Bisherige Lösungen

- Clients ausnutzen
  - Clientseitiges Caching (AFS, CacheFS)
  - NFSv4-Delegationen
  - aber: Workloads teilweise zu groß, zu wenig Reuse
- Server-Forwarding
  - aber: Verkehr durch Front- und Backend-Server
  - Bandbreiten-Limit durch Backend-Server

Bisherige Lösungen (ii)

- Out of band (OOB)-Zugriff auf Daten
- Symmetrisch, jeder Client ist Server: Cluster-Filesysteme, Locks und Metadaten werden von anderen Servern geholt (z.B. Sun QFS)
- Asymmetrisch: Trennung in Metadaten-Server und mehrere Storage Devices (z.B. SGI CXFS)
- Vorteile: Bandbreite und Kapazität skalieren
- aber: proprietär, eingeschränkter Client-Support, nicht interoperabel, keine Zugangskontrolle zu Daten, kein kompletter Ersatz für NFS

Jetzt mit Standards: Parallel NFS

- pNFS: NFS mit OOB-Zugriffen auf Daten
- Erweiterung von NFSv4: V4.1 (minor version)
- NFSv4-Server als Metadatenserver
- Clients bekommen Layouts mit Zugriffsinformationen für Daten
- Zugriffsprotokolle für Daten: NFSv4 (Files), SCSI SBC (Blocks/Volumes), SCSI OSD (Objekte)
- erweiterbar um weitere Zugriffsprotokolle
- Standardisierung in NFSv4-WG der IETF: EMC, IBM, NetApp, Panasas, Sun, ...
- Implementierungen: Prototypen von Calsoft (FreeBSD), NetApp, Sun (Solaris), U. Michigan/CITI (Linux), ...

Storage Server: Files, Blocks . . .

- Files:
  - Files ggfs. über mehrere NFSv4-Server gestriped
- Blocks:
  - Nutzung von SAN-Storage (FC oder iSCSI), NFSv4-Server alloziert Blocks
  - diverse Volume-Konfigurationen: Simple, Slice, Concat, Stripe
  - aber: keine Zugriffskontrolle auf Storage (max. auf LUN-Ebene)

Storage Server: . . . und Objekte

- SCSI Object Storage Devices (OSD)
- Objekte statt Blocks als adressierbare Einheiten
- ANSI T10-Spezifikation, erste Implementierungen: z.B. Seagate 2004
- Operationen: Create/Delete/Read/Write Object, Adressierung: Objekt-Id, Byte-Bereich
- Security auf Objekt-Ebene: kryptographische Capabilities entscheiden über Zugriff, werden von Security Manager (pNFS-Server) an Clients gegeben, HMAC mit Shared Secret zwischen OSD und Security Manager
- diverse pNFS-Konfigurationen möglich: Simple, Stripe, Mirror

Neue NFSv4.1-Operationen für pNFS

- Layout: Mapping von File auf (mehrere) Storage-Server, abhängig vom Storage-Server-Typ: `<clientid, filehandle, offset, length, iomode, type, type-specific>`, Möglichkeit zum direkten Zugriff auf Storage Server
- Layout-Verwaltung: `LAYOUTGET, LAYOUTCOMMIT, LAYOUTRETURN`
- Device-Identifikation: Umsetzung von 32-Bit-Device-Ids in Zugriffsinformationen: `GETDEVICEINFO, GETDEVICELIST`
- Neue Callbacks: `CB_LAYOUTRECALL, CB_SIZECHANGED`

Literatur: pNFS

- NFSv4-WG der IETF:
  `http://www.ietf.org/html.charters/nfsv4-charter.html`
  - G. Gibson u.a., pNFS Problem Statement, expired I-D, Juli 2004
  - G. Gibson u.a., Parallel NFS Requirements and Design Considerations, expired I-D, Oktober 2004
  - S. Shepler ed., NFSv4 Minor Version 1, I-D, Dezember 2005
  - D. L. Black u.a., pNFS Block/Volume Layout, I-D, Dezember 2005
  - J. Zelenka u.a., Object-based pNFS Operations, I-D, Oktober 2005
- G. Gibson, Parallel NFS (pNFS), SNIA Developers Solutions Conference, August 2005

Literatur: Object Storage Devices (OSD)

- ANSI T10 Technical Commitee on SCSI Storage Interfaces: `http://www.t10.org/`
- Information technology—SCSI Object-Based Storage Device Commands-2 (OSD-2), ANSI T10 Working Draft T10/1731-D, Oktober 2004
- A. Krimkevich, Object-Based Storage, NAS Industry Conference, Oktober 2005