best SYSTEME GMBH

10Gbit und Network FS

best Open Systems Day
Fall 2007

Unterföhring

Marco Kühn
best Systeme GmbH
marco.kuehn@best.de

Agenda

- **10Gbit – Wie bringe ich soviel PS auf die Strasse ?**
- pNFS
- 10Gbit Anwendungen und Lösungen von AT

10Gbit Ethernet Performance:

- 10Gbit Ethernet schneller als Netzwerk (File-) Protokolle (NFS, FTP)
- Single Stream Performance für
  - NFS: 200-250 MB/s
  - FTP: 300-350 MB/s
  - Netperf: 750-850 MB/s
- Multi Stream Performance wird genutzt
  - Mehrere Clients und ein Server

- Wie kann man die Single Stream Performance erhöhen ?
- Wie kann man die Single Server Performance erhöhen ?
  - Horizontale Skalierung ?

NFS-over-RDMA

- Remote DMA erlaubt Client / Server Kommunikation ohne Overhead von XDR/TCP/IP Protokoll
  - Client und Server teilen sich den Addressraum
  - Data Transfer folgt direkt via DMA Operationen
  - Implementiert auf Infiniband
  - Angewandt für NFS oder pNFS

- Schnell und Effizient
  - 980 MB/s (IB) vs. 235 MB/s (10Gbit)
    - 980 MB/s limitiert durch HW x2100 M1

- Gibt es eine Lösungen ohne Infiniband ?
- Kann man horizontal skalieren ?

Agenda

- 10Gbit – Wie bringe ich soviel PS auf die Strasse ?
- **pNFS Architektur**
- 10Gbit Ethernet: Anwendungen und Lösungen von AT

pNFS Architektur

Was ist pNFS ?

- pNFS = parallel NFS
- Erweiterung auf NFS
  - Standardisiert durch die NFSv4 Arbeitsgruppe IETF
  - Teil der NFS Version 4.1 Spezifikation

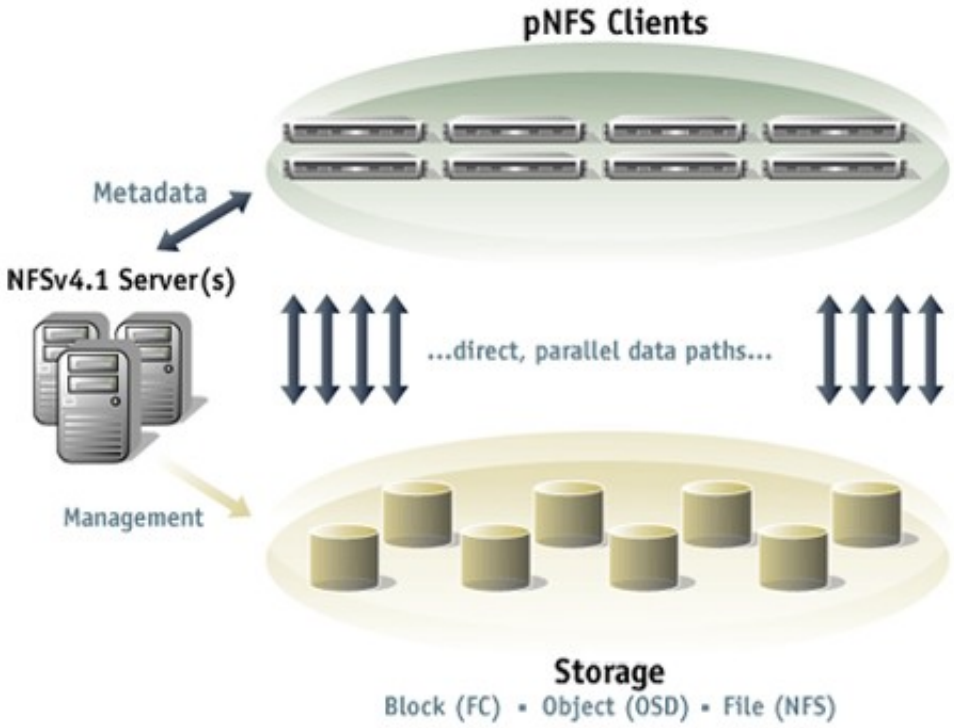
- Wichtige Features:
  - Parallele NFS-Übertragung von mehreren Servern zu einem Client
  - Ein einheitlicher Namensraum
  - Horizontale Skalierung
  - Trennung von Daten und Metadaten

pNFS Architektur

- Klassische Cluster Filesytem Architektur
  - Metadaten werden von Metadaten Server (MDS) verwaltet
  - Benuzterdaten werden auf mehrere Daten Server (DS) verteilt

- pNFS unterstützt 3 Storage Protokolle für den Datenpfad
  - Block (FC)
  - Object (OSD)
  - File (NFS)

Vorteile:
- Parallel I/O
- Standard NFS



pNFS Architektur

Client Daten Transfer:

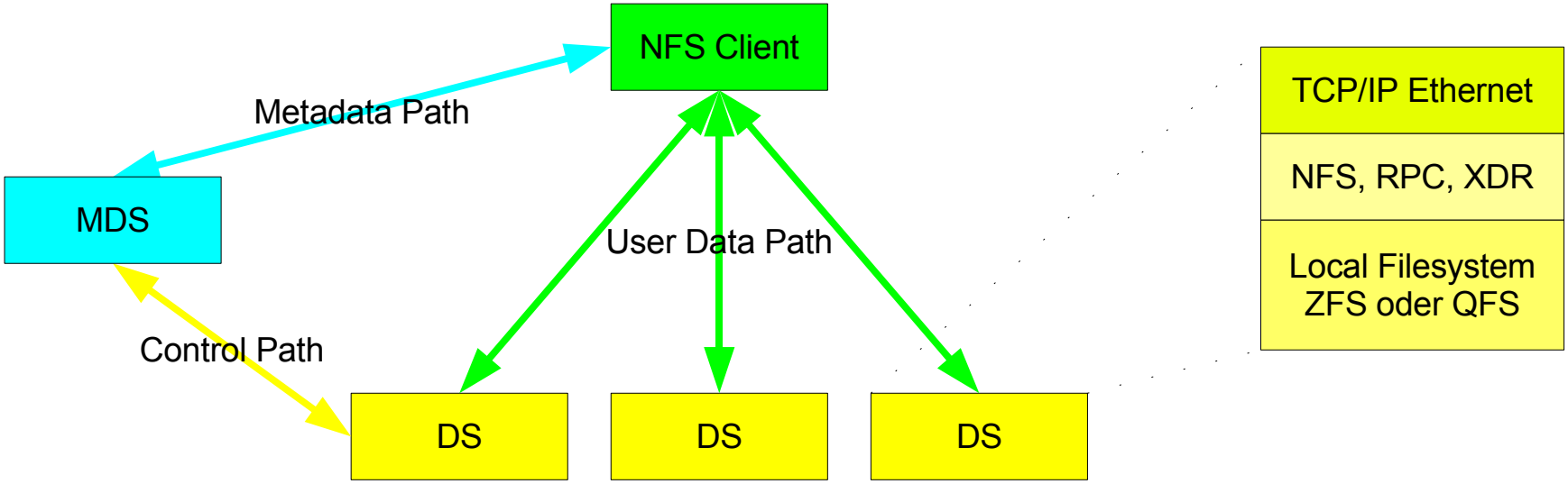
- Client öffnet Datei, indem der MDS gefragt wird
- MDS erlaubt den Client Zugriff und sendet Layout zurück
- Layout beschreibt wie Datei Daten auf den DS verteilt sind

Daten Organisation:

- Standard Support für Raid0 und replizierte Daten
  - Striping auf Datei Ebene
- Dateien werden kundenspezifisch verteilt im globalen Namensraum
  - /pnfs/file1 kann ein 5-wege Stripe sein: ds1, ds2, ds3, ds4, ds5
  - /pnfs/file2 kann ein einelementige Datei sein: ds3

pNFS Architektur

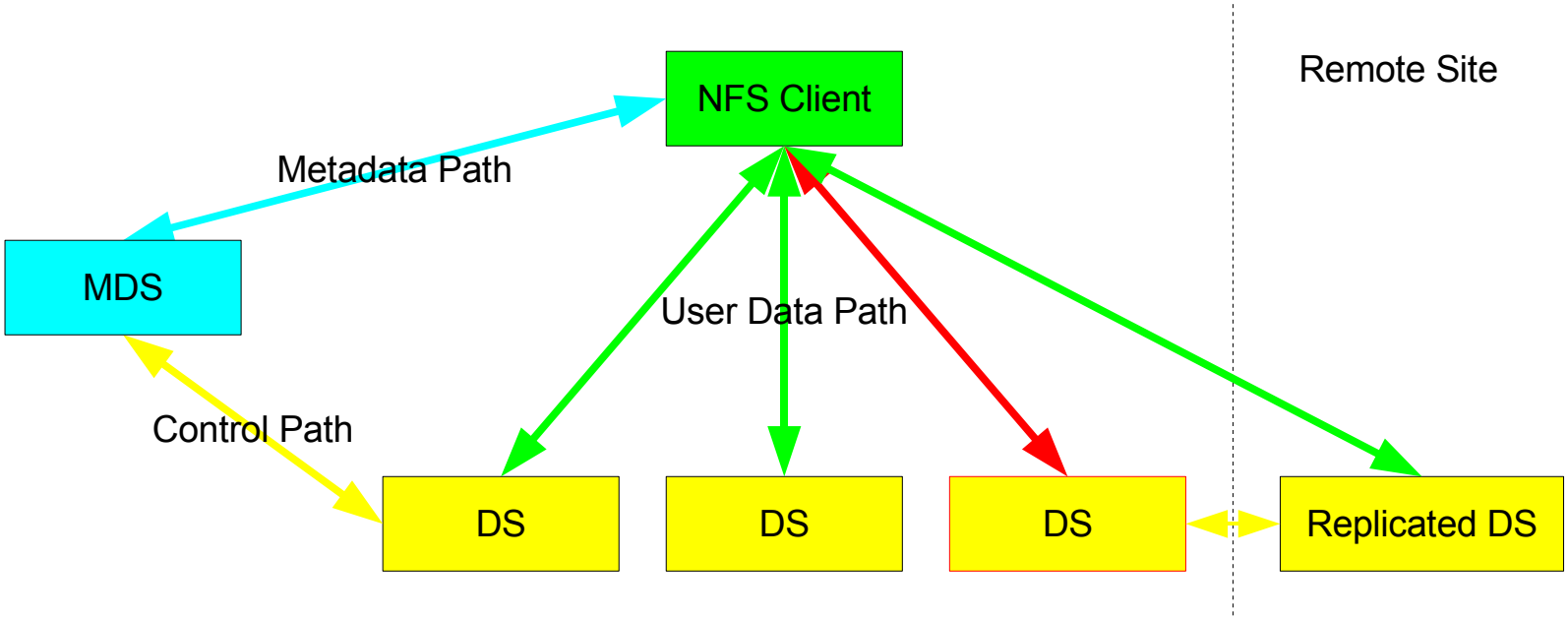
- Solaris DS legen Sub-Files in ZFS ab
- Solaris pNFS Community (Single Mount Point)
  - 50 Milliarden Dateien verwalten
  - aus mehreren Tausenden DS bestehen
  - Dateien mit tatsächlich unlimitierter Grösse beinhalten
  - Praktisch kein oberes Limit auf Community-weite Bandbreite



pNFS Architektur

Repliziertes Layout:

- DS Remote Copy (SNDR)
- Transport Failover
  - MDS erkennt DS Verlust und fragt nach neuen Layout
  - Client meldet Layout Fehler und fragt nach neuem Layout
  - MDS antwortet dem Client neues Layout mit Replicated DS
- Kann während Client NFS Zugriff passieren bei synchroner Replikation



pNFS Architektur

Praktische Folgerung:

- pNFS mit NFS-over-RDMA
  - Schneller und großer Datentransfer
  - Große Clients (X4600) kann fast 2.5 GB/s von bzw. zu den DS's
- Datenbank via NFS (Oracle RAC)
  - pNFS + RDMA machen DSS und OLTP praktikabel

Danke !

Agenda

- 10Gbit – Wie bringe ich soviel PS auf die Strasse ?
- pNFS Architektur
- **10Gbit Ethernet: Anwendungen und Lösungen von AT**