

# Object Storage mit Policy-Unterstützung für AFS

Felix Frank

DESY

GUUG Treffen Berlin, 05. März 2009

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies

## Deutsches Elektronensynchrotron

Ein Forschungszentrum der Helmholtz-Gemeinschaft

- ▶ Standorte Hamburg und Zeuthen
- ▶ Grundlagenforschung in der Teilchenphysik

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

### Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

#### AFS

Vergleich  
Konzepte  
Alternativen

#### OpenAFS+OSD

Idee  
Features  
Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung





## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



- ▶ verteiltes Dateisystem mit globalem Namensraum

```
diff /afs/ihf.de/user/f/ffrank/volume.c  
/afs/ipp-garching.mpg.de/.cs/volume.c
```

- ▶ Kerberos integriert, für WAN entwickelt

## Motivation

- ▶ viel flexibler als NFSv3, NFSv4 nicht bereit
- ▶ OpenSource

## Status

- ▶ OpenAFS 1.4.x stabil, 1.5.x testing und Windows-Client
- ▶ 194 Sites registriert, viele mehr unregistriert
- ▶ neue Features werden noch immer implementiert
- ▶ Pläne für Foundation

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



- ▶ verteiltes Dateisystem mit globalem Namensraum

```
diff /afs/ihf.de/user/f/ffrank/volume.c  
/afs/ipp-garching.mpg.de/.cs/volume.c
```

- ▶ Kerberos integriert, für WAN entwickelt

## Motivation

- ▶ viel flexibler als NFSv3, NFSv4 nicht bereit
- ▶ OpenSource

## Status

- ▶ OpenAFS 1.4.x stabil, 1.5.x testing und Windows-Client
- ▶ 194 Sites registriert, viele mehr unregistriert
- ▶ neue Features werden noch immer implementiert
- ▶ Pläne für Foundation

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



- ▶ verteiltes Dateisystem mit globalem Namensraum

```
diff /afs/ihf.de/user/f/ffrank/volume.c  
/afs/ipp-garching.mpg.de/.cs/volume.c
```

- ▶ Kerberos integriert, für WAN entwickelt

## Motivation

- ▶ viel flexibler als NFSv3, NFSv4 nicht bereit
- ▶ OpenSource

## Status

- ▶ OpenAFS 1.4.x stabil, 1.5.x testing und Windows-Client
- ▶ 194 Sites registriert, viele mehr unregistriert
- ▶ neue Features werden noch immer implementiert
- ▶ Pläne für Foundation

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Vergleich AFS / NFSv3

Setup	<b>NFSv3</b>	<b>AFS</b>
Administrierung	sehr einfach	komplex
Wartung	OK mit automount	einfach, transparent
max. # Clients	Downtime	Live
Caching (Client)	recht begrenzt	skaliert sehr gut
Einsatz im HPC	vorhanden	hochentwickelt
Durchsatz	schwierig	RO Daten repliziert
Authentisierung	schnell	Client langsam
ACLs	UNIX UIDs	Kerberos
Ausfallsicherheit	nur UNIX mode	nur für Directories
Einbindung	begrenzt	Live-Redundanz
Windows API	einfach	Kernel Modul notw.
Support	nicht trivial	1.5.x stabil
	gr. Userbase	aktive Community

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Volume

- ▶ Dateneinheit mit eigenem Quota
- ▶ RW + opt. 1...n RO + opt. Backup
- ▶ Mountpoints
  - ▶ Teil der Daten in einem Volume
    - ▶ Realisiert mittels Symlink
  - ▶ "User Level Automounter"
  - ▶ Client baut Verzeichnisbaum lokal auf
    - ▶ Flexibilität, Skalierbarkeit

```
$ fs mkm /afs/ihf.de/packages/RPMS p.rpm
```

## Access Control Lists

- ▶ gelten stets für gesamte Verzeichnisse
- ▶ read, lookup, insert, delete, write, lock, administer
- ▶ individuell zu vergeben für Nutzer und Gruppen
- ▶ nützlich: ACLs für IP-Adressen (Rechner, Netze)

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Volume

- ▶ Dateneinheit mit eigenem Quota
- ▶ RW + opt. 1...n RO + opt. Backup
- ▶ Mountpoints
  - ▶ Teil der Daten in einem Volume
    - ▶ Realisiert mittels Symlink
  - ▶ "User Level Automounter"
  - ▶ Client baut Verzeichnisbaum lokal auf
    - ▶ Flexibilität, Skalierbarkeit

```
$ fs mkm /afs/ihf.de/packages/RPMS p.rpm
```

## Access Control Lists

- ▶ gelten stets für gesamte Verzeichnisse
- ▶ read, lookup, insert, delete, write, lock, administer
- ▶ individuell zu vergeben für Nutzer und Gruppen
- ▶ nützlich: ACLs für IP-Adressen (Rechner, Netze)

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies





## Volume

- ▶ Dateneinheit mit eigenem Quota
- ▶ RW + opt. 1...n RO + opt. Backup
- ▶ Mountpoints
  - ▶ Teil der Daten in einem Volume
    - ▶ Realisiert mittels Symlink
  - ▶ "User Level Automounter"
  - ▶ Client baut Verzeichnisbaum lokal auf
    - ▶ Flexibilität, Skalierbarkeit

```
$ fs mkm /afs/ihf.de/packages/RPMS p.rpm
```

## Access Control Lists

- ▶ gelten stets für gesamte Verzeichnisse
- ▶ read, lookup, insert, delete, write, lock, administer
- ▶ individuell zu vergeben für Nutzer und Gruppen
- ▶ nützlich: ACLs für IP-Adressen (Rechner, Netze)



**Namensschema** für Volumes, z.B. Homes:

- ▶ user.ffrank
- ▶ user.hartmut
- ▶ ...

**Mountpoints** z.B.

/afs/afh.de/user/f/ffrank → user.ffrank

...

**Backups erzeugen**

```
$ vos backupsys -prefix `user.`  
/afs/afh.de/user/ffrank/OldFiles  
-r user.ffrank-backup
```

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



**Namensschema** für Volumes, z.B. Homes:

- ▶ user.ffrank
- ▶ user.hartmut
- ▶ ...

**Mountpoints** z.B.

/afs/afh.de/user/f/ffrank → user.ffrank

...

**Backups erzeugen**

```
$ vos backupsys -prefix ``user.``  
/afs/afh.de/user/f/ffrank.OldFiles  
→ user.ffrank.backup
```

...



**Namensschema** für Volumes, z.B. Homes:

- ▶ user.ffrank
- ▶ user.hartmut
- ▶ ...

**Mountpoints** z.B.

/afs/afh.de/user/f/ffrank → user.ffrank

...

**Backups erzeugen**

```
$ vos backupsys -prefix ``user.``  
/afs/afh.de/user/f/ffrank/.OldFiles  
→ user.ffrank.backup
```

...



**Namensschema** für Volumes, z.B. Homes:

- ▶ user.ffrank
- ▶ user.hartmut
- ▶ ...

**Mountpoints** z.B.

/afs/afh.de/user/f/ffrank → user.ffrank

...

**Backups erzeugen**

```
$ vos backupsys -prefix ``user.``
```

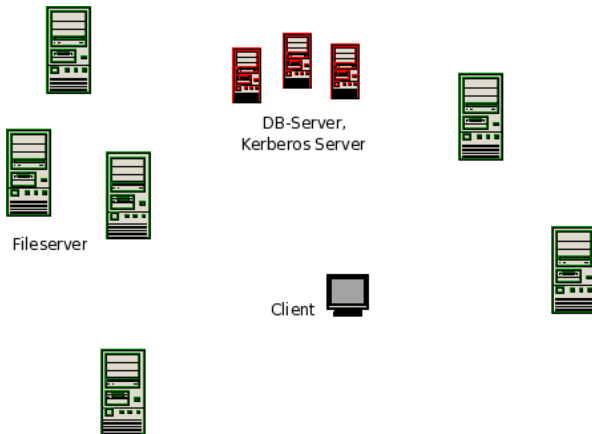
```
/afs/afh.de/user/f/ffrank/.OldFiles
```

```
→ user.ffrank.backup
```

...



# Übersicht: AFS Server



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies

# Übersicht: AFS Server

OpenAFS+OSD  
with Policies

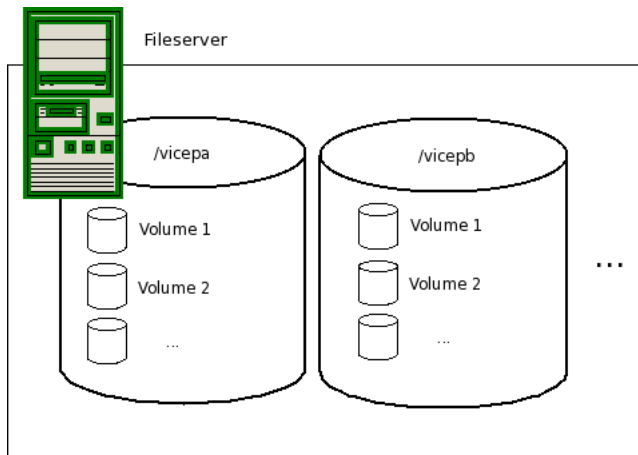
Felix Frank

AFS

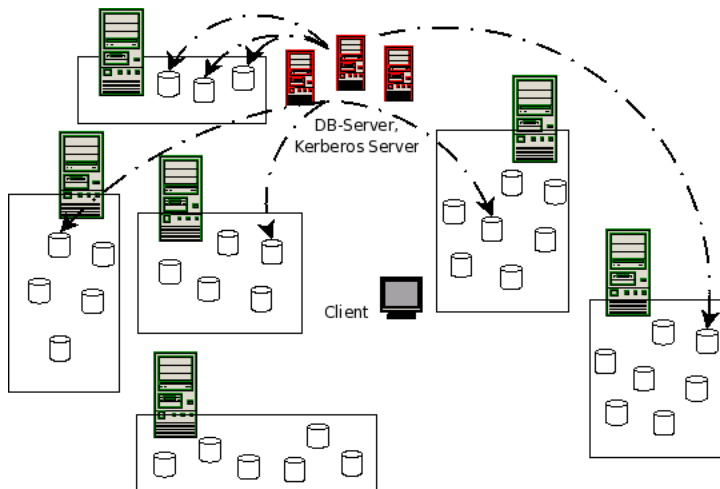
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Übersicht: AFS Server



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

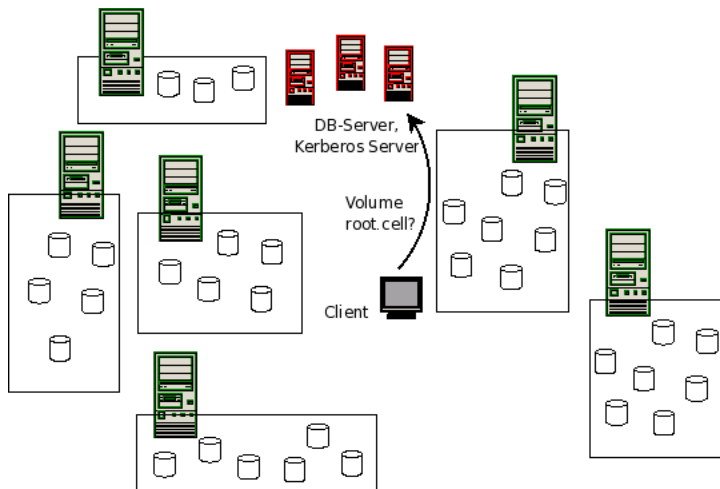
OpenAFS+OSD

Idee  
Features  
Policies





# Übersicht: AFS Server



OpenAFS+OSD  
with Policies

Felix Frank

AFS

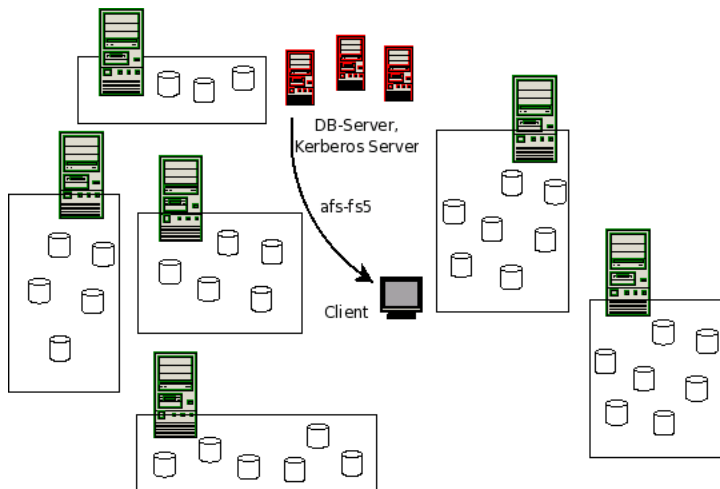
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Übersicht: AFS Server



OpenAFS+OSD  
with Policies

Felix Frank

AFS

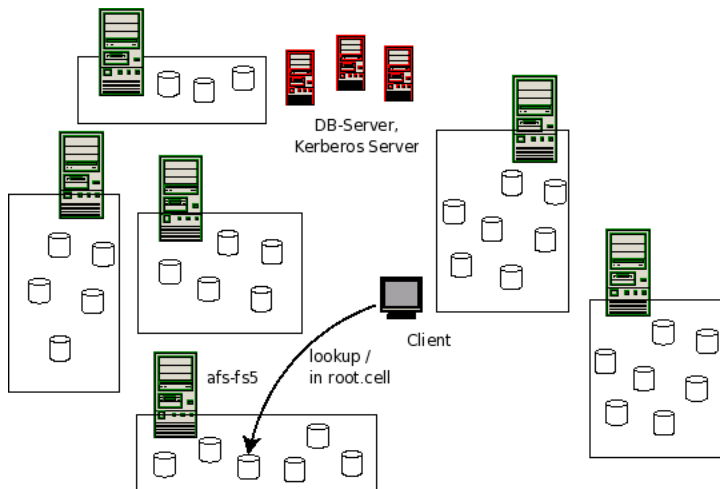
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Übersicht: AFS Server



OpenAFS+OSD  
with Policies

Felix Frank

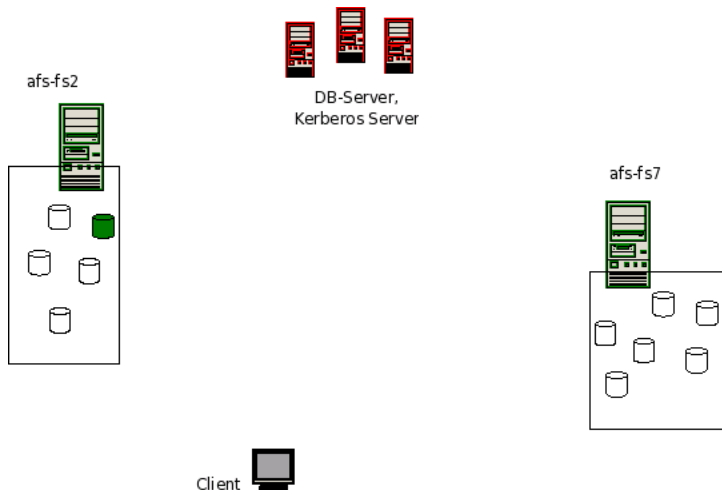
AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies

Volume: `software.linux` auf `afs-fs2` / `vicepa`



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

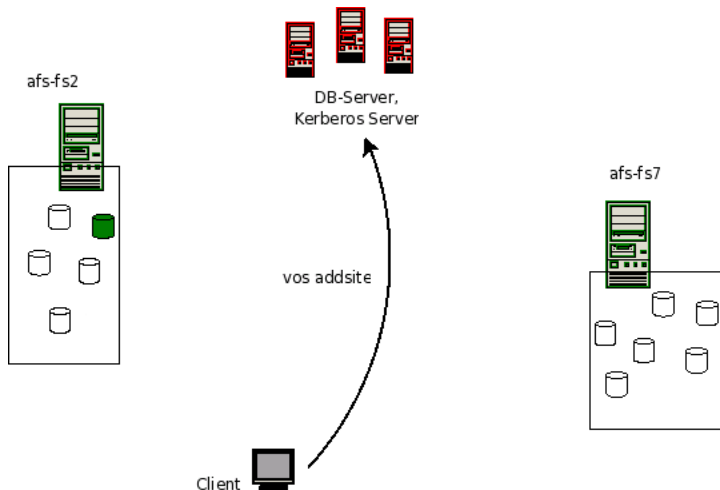
Features

Policies



# Datenreplikation

```
$ vos addsite software.linux afs-fs2 a
```



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

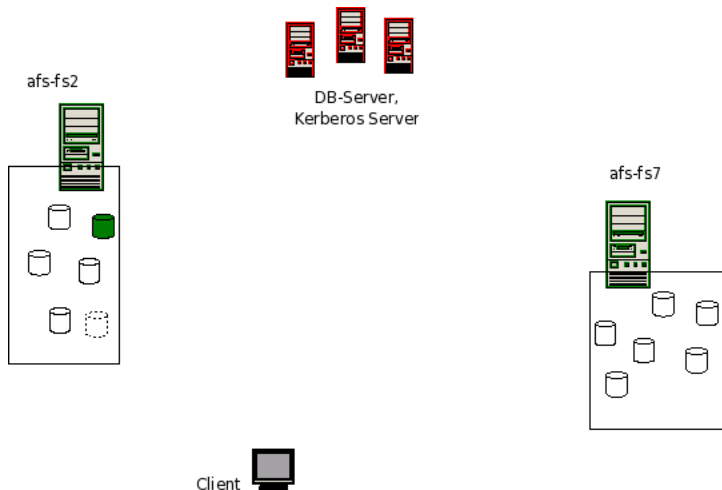
Idee

Features

Policies



## RO-Kopie nur in DB



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

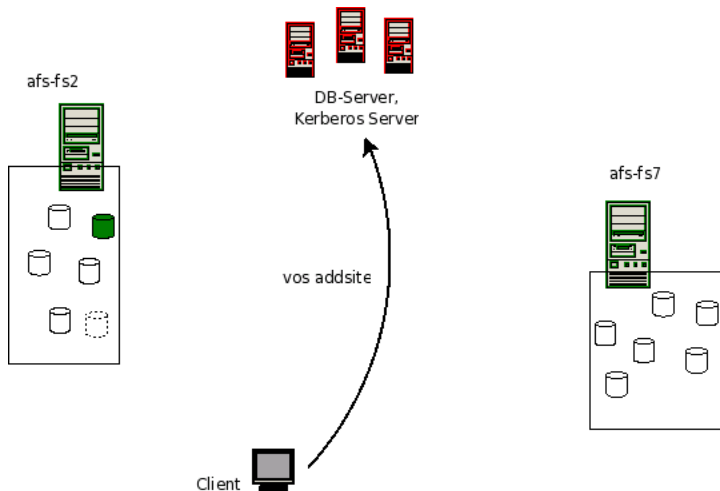
Features

Policies



# Datenreplikation

```
$ vos addsite software.linux afs-fs7 a
```



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

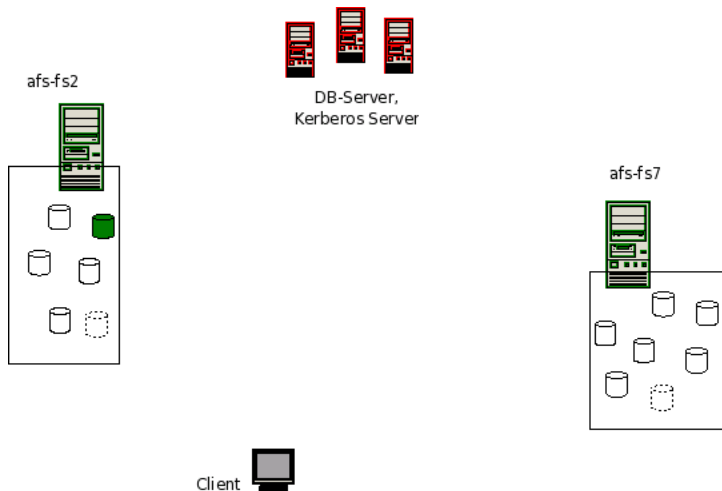
Idee

Features

Policies



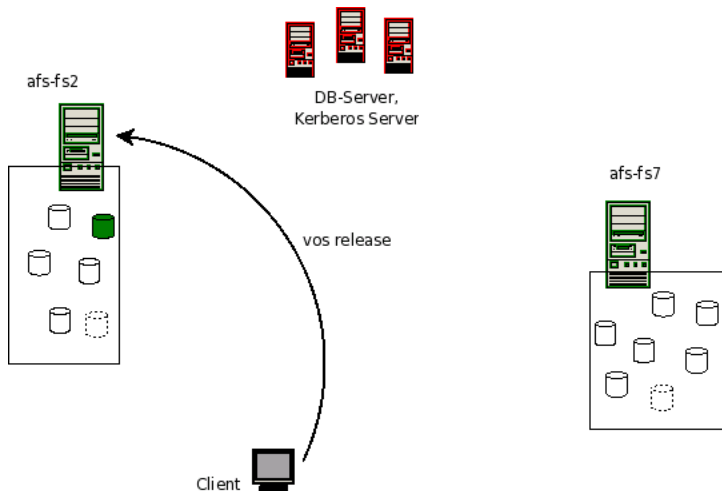
## Zweite RO-Kopie, nur in DB





# Datenreplikation

```
$ vos release software.linux
```



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

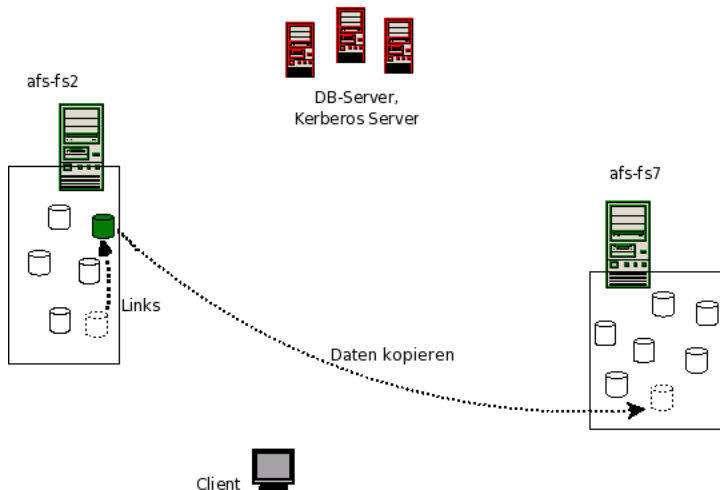
OpenAFS+OSD

Idee

Features

Policies

## Daten werden verteilt



OpenAFS+OSD  
with Policies

Felix Frank

AFS

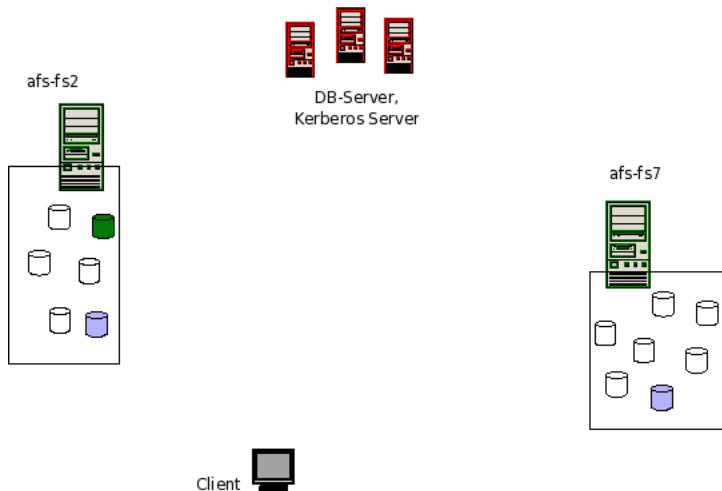
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## RO Volumes bereit



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

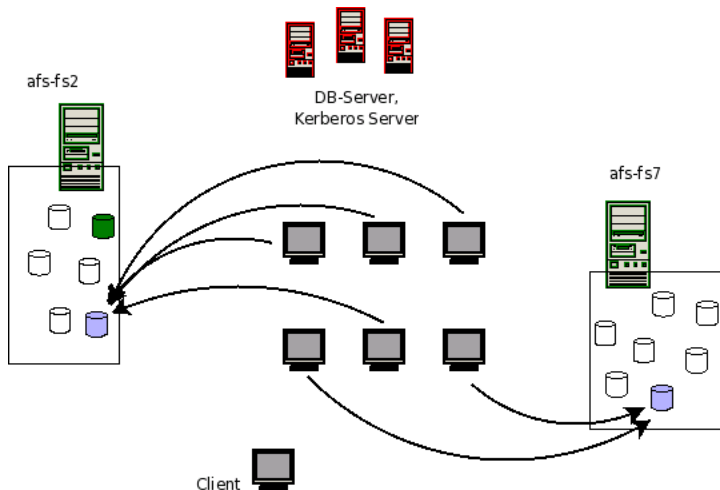
Idee

Features

Policies



## Verteilung der Lesezugriffe



OpenAFS+OSD  
with Policies

Felix Frank

AFS

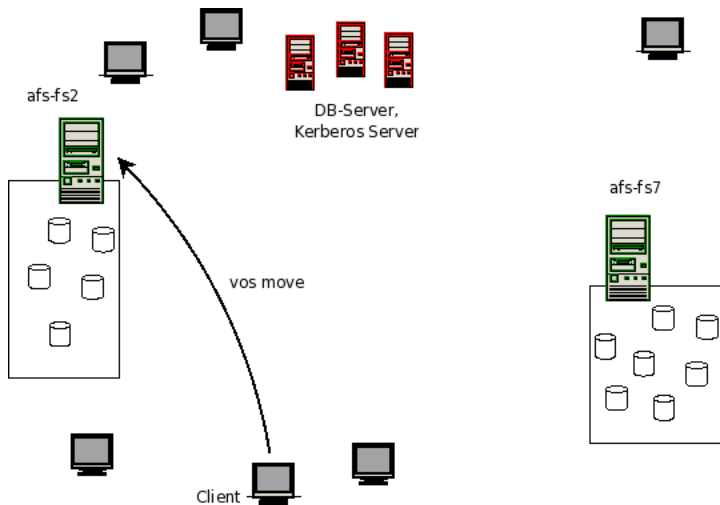
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies

# Datenmigration

```
$ vos move user.ffrank afs-fs2 a afs-fs7 a
```



OpenAFS+OSD  
with Policies

Felix Frank

AFS

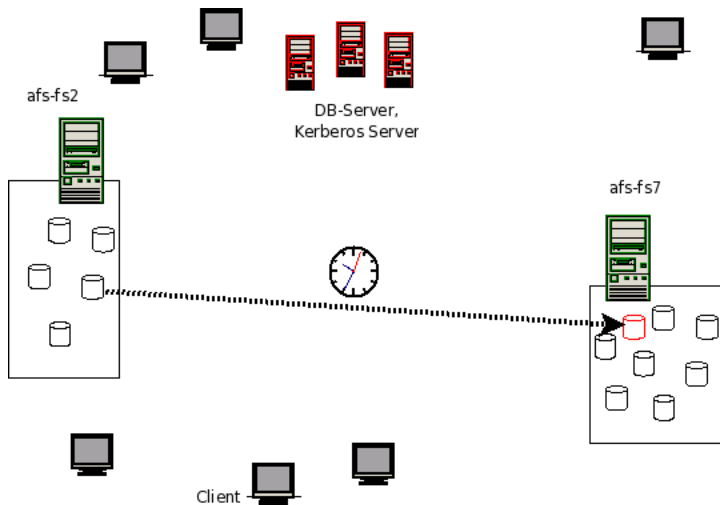
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Fileserver kopiert Klon



OpenAFS+OSD  
with Policies

Felix Frank

AFS

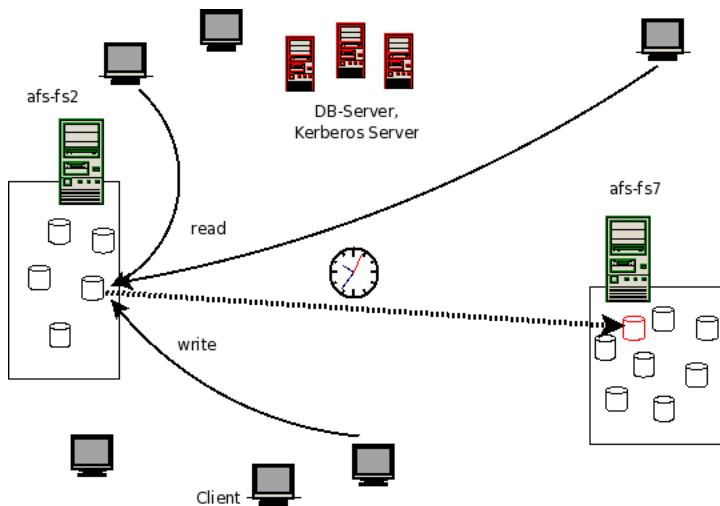
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Zugriffe funktionieren weiter



OpenAFS+OSD  
with Policies

Felix Frank

AFS

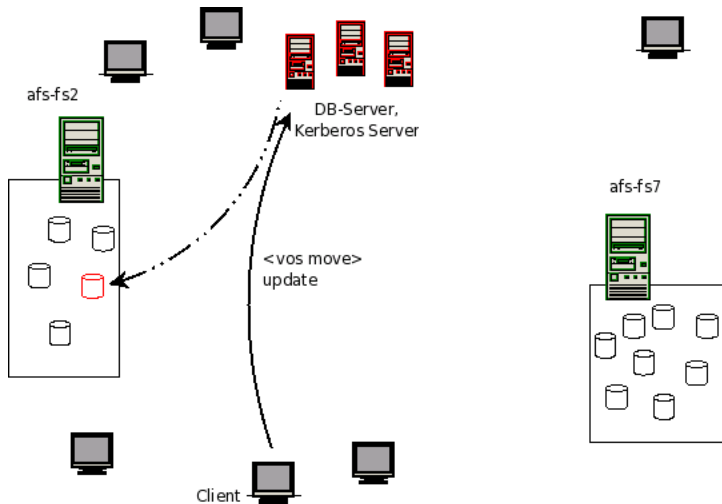
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Datenbank aktualisieren



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

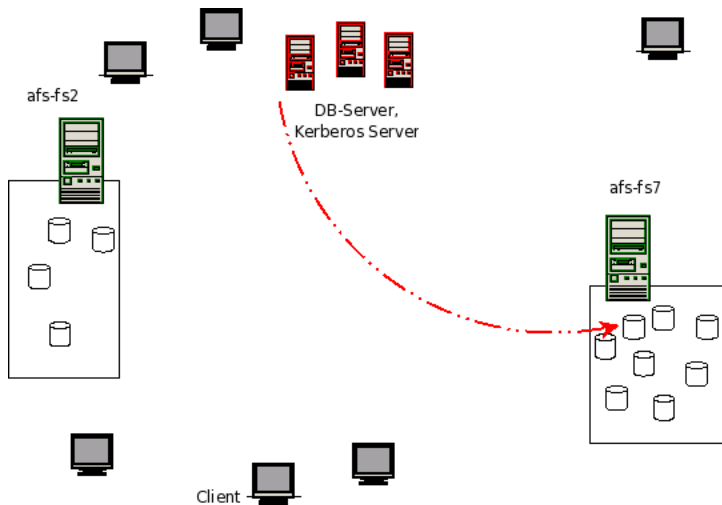
OpenAFS+OSD

Idee  
Features  
Policies





# Datenmigration



OpenAFS+OSD  
with Policies

Felix Frank

AFS

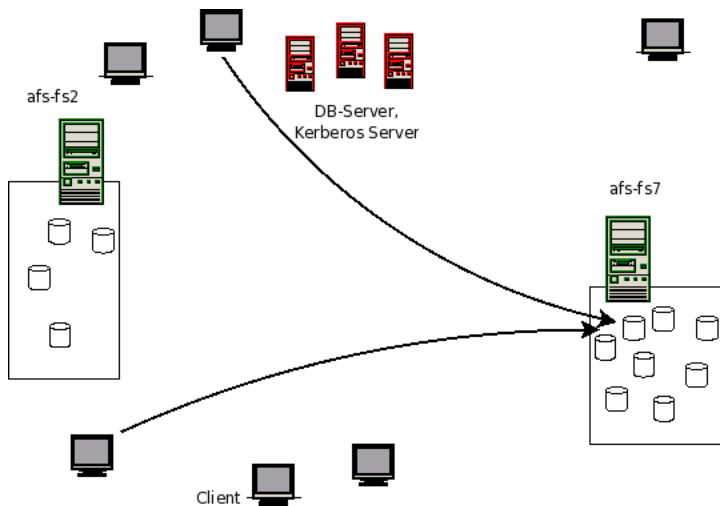
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Neue Position wird verwendet



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## NFS

- ▶ Version 3 verliert Bedeutung
- ▶ Version 4 kam zu spät, auch komplex, vieles noch TODO

## AFS

- ▶ Community arbeitet an Performance
- ▶ Hotspots machen Probleme mit Farm-Computing
- ▶ fest etabliert, NFSv4 (noch) keine Alternative
- ▶ Nutzung: 24.1 / 52 TB

## Lustre

- ▶ in Zeuthen Ersatz für kommerzielles **Panasas**
- ▶ schon jetzt wesentlich bessere Performance als AFS und NFS
- ▶ andere Nachteile

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## NFS

- ▶ Version 3 verliert Bedeutung
- ▶ Version 4 kam zu spät, auch komplex, vieles noch TODO

## AFS

- ▶ Community arbeitet an Performance
- ▶ Hotspots machen Probleme mit Farm-Computing
- ▶ fest etabliert, NFSv4 (noch) keine Alternative
- ▶ Nutzung: 24.1 / 52 TB

## Lustre

- ▶ in Zeuthen Ersatz für kommerzielles **Panasas**
- ▶ schon jetzt wesentlich bessere Performance als AFS und NFS
- ▶ andere Nachteile

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## NFS

- ▶ Version 3 verliert Bedeutung
- ▶ Version 4 kam zu spät, auch komplex, vieles noch TODO

## AFS

- ▶ Community arbeitet an Performance
- ▶ Hotspots machen Probleme mit Farm-Computing
- ▶ fest etabliert, NFSv4 (noch) keine Alternative
- ▶ Nutzung: 24.1 / 52 TB

## Lustre

- ▶ in Zeuthen Ersatz für kommerzielles **Panasas**
- ▶ schon jetzt wesentlich bessere Performance als AFS und NFS
- ▶ andere Nachteile

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Metadataserver**
- ▶ Daten stark verteilt auf **Objectstorage**servers
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformatoren erst bei MDS an
- ▶ Verbindet sich mit OSS(n) um Daten zu empfangen

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



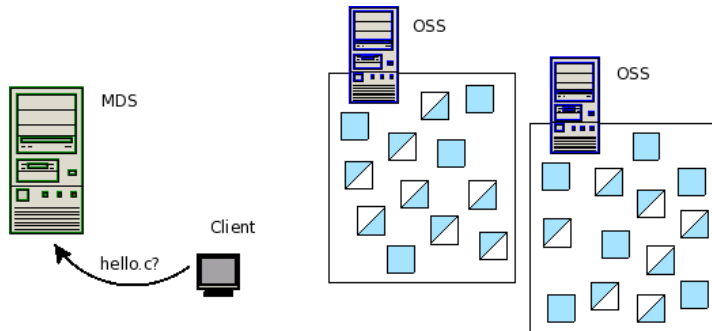
# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Metadataserver**
- ▶ Daten stark verteilt auf **Objectstorage**servers
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformationen erst bei **MDS** an

▶ Verbindet sich mit OSS(n) um Daten zu empfangen



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



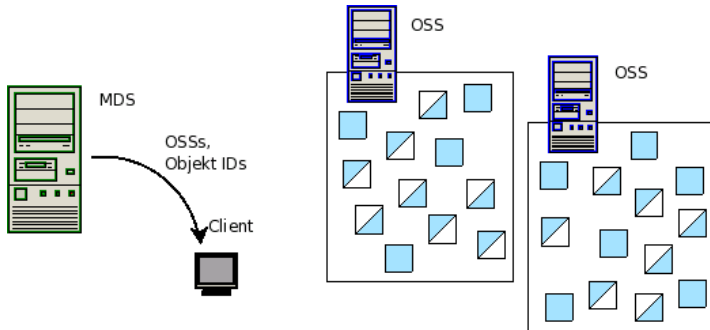


# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Metadataserver**
- ▶ Daten stark verteilt auf **Objectstorage**servers
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformationen erst bei MDS an
- ▶ Verbindet sich mit **OSS(n)** um Daten zu empfangen



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies

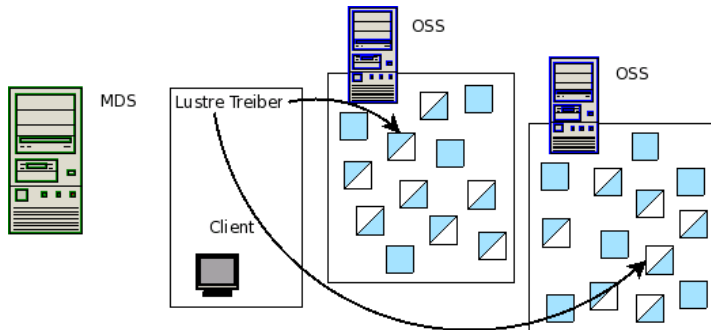


# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Meta**data**server**
- ▶ Daten stark verteilt auf **Object**storage**servern**
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformationen erst bei MDS an
- ▶ Verbindet sich mit OSS(n) um Daten zu empfangen



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies

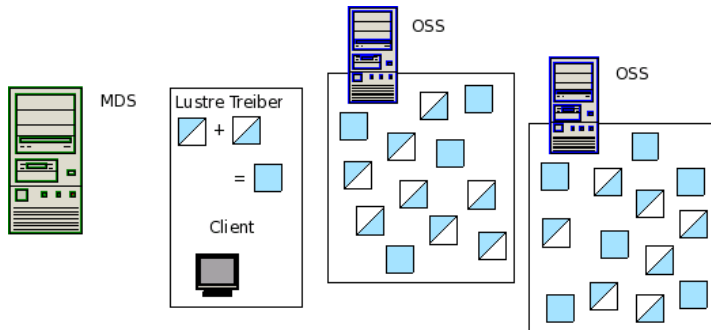


# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Meta**data**server**
- ▶ Daten stark verteilt auf **Object**storage**servern**
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformationen erst bei MDS an
- ▶ Verbindet sich mit OSS(n) um Daten zu empfangen



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies

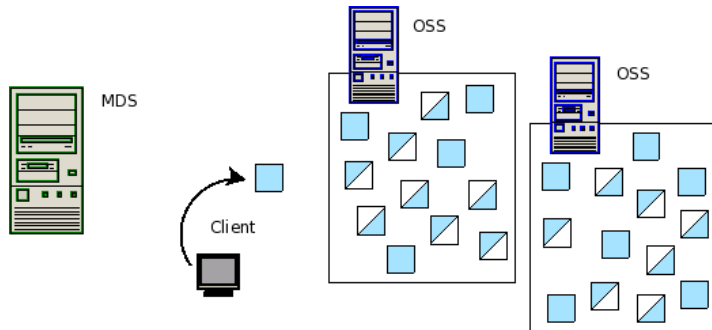


# Object Storage – Beispiel: Lustre

- ▶ zentrale Verwaltungsdatenbank: MGS
- ▶ Dateiverzeichnis: **Meta**data**server**
- ▶ Daten stark verteilt auf **Object**storage**servern**
  - ▶ halten nicht Dateien sondern *Objekte*
  - ▶ Datei kann aus mehreren Objekten bestehen

## Zugriff

- ▶ Client fragt Dateiinformationen erst bei MDS an
- ▶ Verbindet sich mit OSS(n) um Daten zu empfangen



OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



## Vorteile

- ▶ Datenverkehr stets verteilt
- ▶ extrem schneller Client und Server
- ▶ unterstützt RDMA über InfiniBand

## Nachteile

- ▶ bei Hochlast: Metadataserver wird Bottleneck
  - ▶ derzeit Live-Redundanz sehr beschränkt
  - ▶ Wartbarkeit bei AFS DB-Servern besser
- ▶ Storage Server nicht leicht ersetzbar
- ▶ Kernel Modul nicht portabel, gepatchter Kernel nötig (Server)
- ▶ Server nur auf Linux
- ▶ Erweiterbarkeit begrenzt
- ▶ Kerberos, ACLs, Quota etc. noch in Beta bzw. Roadmap

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



## Vorteile

- ▶ Datenverkehr stets verteilt
- ▶ extrem schneller Client und Server
- ▶ unterstützt RDMA über InfiniBand

## Nachteile

- ▶ bei Hochlast: Metadataserver wird Bottleneck
  - ▶ derzeit Live-Redundanz sehr beschränkt
  - ▶ Wartbarkeit bei AFS DB-Servern besser
- ▶ Storage Server nicht leicht ersetzbar
- ▶ Kernel Modul nicht portabel, gepatchter Kernel nötig (Server)
- ▶ Server nur auf Linux
- ▶ Erweiterbarkeit begrenzt
- ▶ Kerberos, ACLs, Quota etc. noch in Beta bzw. Roadmap

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



# Übersicht verteilte Dateisysteme

## NFS

sehr einfach

schnell nur für  
einzelne Clients

max.Geschw.OK

Wartung  
schwierig

keine Redundanz  
in v3

keine Verteilung

## AFS

Setup komplex

Lastverteilung  
(readonly)

Durchsatzlimits

transparente  
Datenmigration

redundante DB  
+ RO-Replikation

RW Hotspots

## Lustre

Setup komplex

schnelles Lesen  
und Schreiben

schnell!

keine  
Datenmigration

fehlende Redundanz  
der DB

gute Datenverteilung

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies

**Die Vereinigung: Das OpenAFS+OSD Projekt**

Idee und Funktion

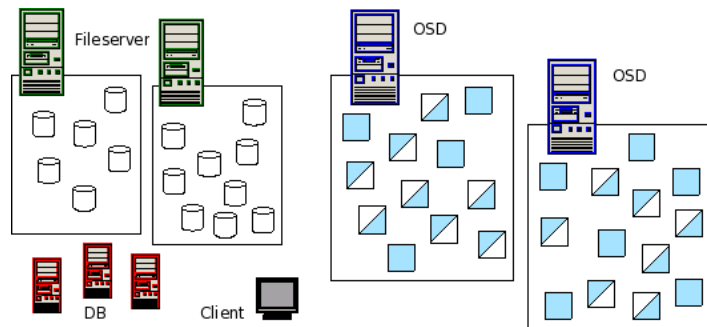
Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung

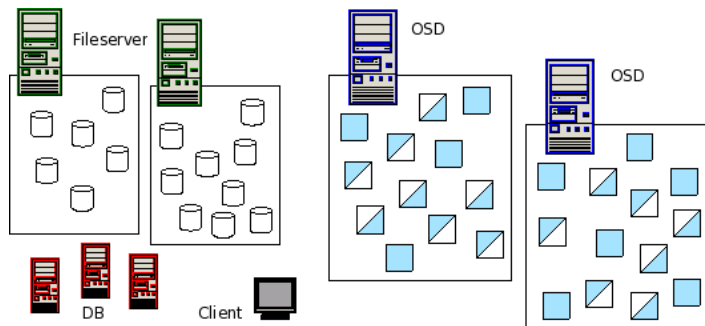




- ▶ (fast) vollständig kompatible Erweiterung von AFS
- ▶ herkömmlicher AFS Client wird unterstützt
- ▶ Object Storage Devices werden zusätzlich zu AFS Fileservern eingesetzt
- ▶ dadurch Vermeidung von Hotspots



- ▶ (fast) vollständig kompatible Erweiterung von AFS
- ▶ herkömmlicher AFS Client wird unterstützt
- ▶ Object Storage Devices werden zusätzlich zu AFS Fileservern eingesetzt
- ▶ dadurch Vermeidung von Hotspots



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als **MDS**
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

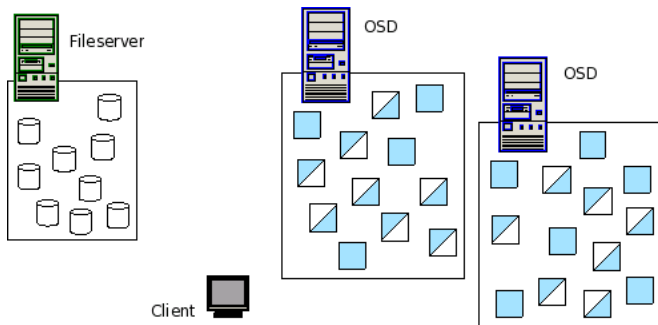
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur **Metadaten**
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

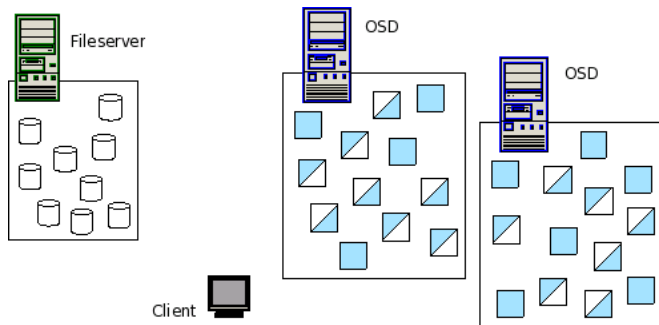
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

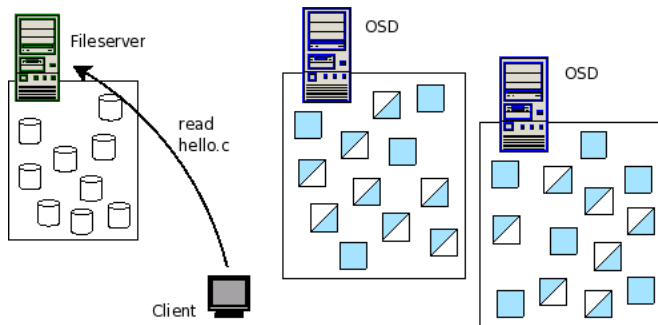
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

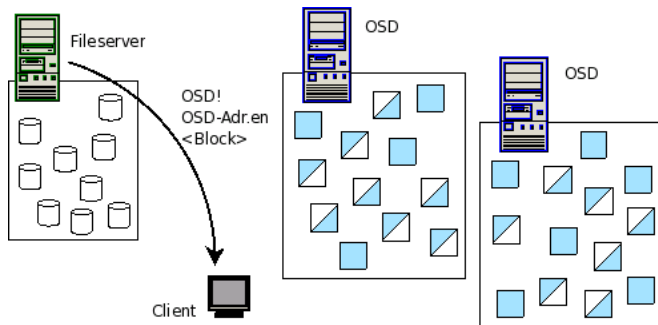
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

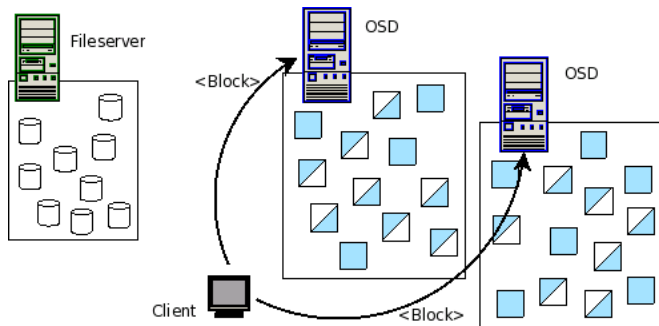
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

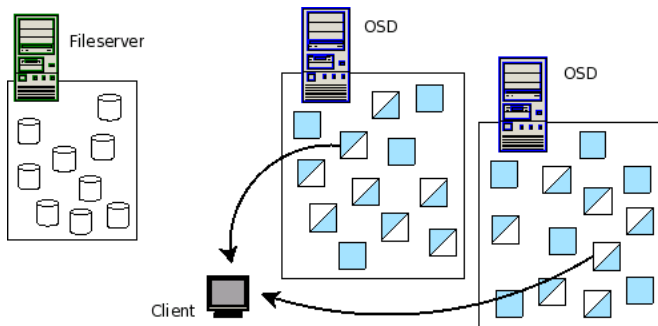
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies





# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

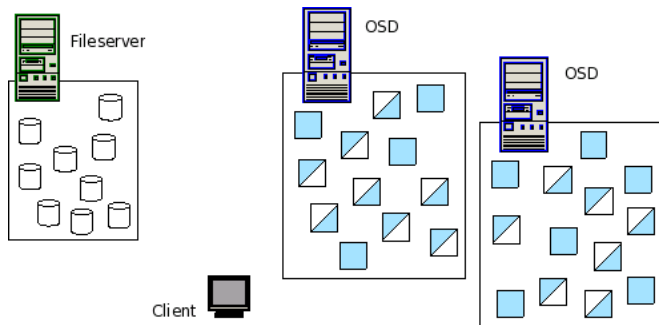
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Funktionsweise von AFS+OSD

- ▶ Fileserver agiert gleichzeitig als MDS
- ▶ Daten von File sind entweder direkt im Volume oder auf beliebigem OSD
- ▶ OSD wird verwendet wenn Datei  $> 1\text{MB}$  (z.B.)
- ▶ im Volume auf Fileserver dann nur Metadaten
- ▶ Client kommuniziert mit OSDs
- ▶ für "alte" Clients spricht Fileserver transparent mit OSDs

OpenAFS+OSD  
with Policies

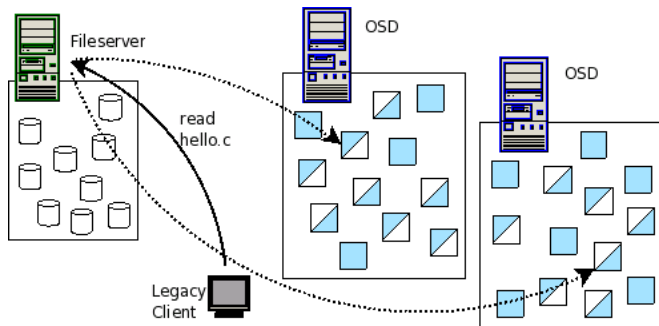
Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

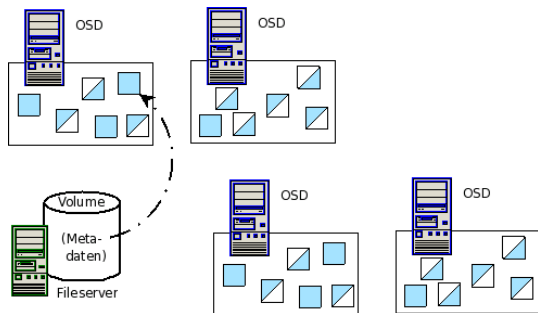
OpenAFS+OSD

Idee  
Features  
Policies



# Performance: Stripes und Copies

- ▶ OSD Dateien bestehen meist aus einem **Objekt**
- ▶ können aber auch gespiegelt oder verteilt sein
- ▶ z.B. zwei identische Objekte auf zwei OSDs (**copies**)
- ▶ z.B. zwei Hälften des Objekts (**stripes**)
- ▶ bis zu 8 Objekte pro Segment, z.B. 4 Stripes mit 2 Kopien



stripes=1, copies=1

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

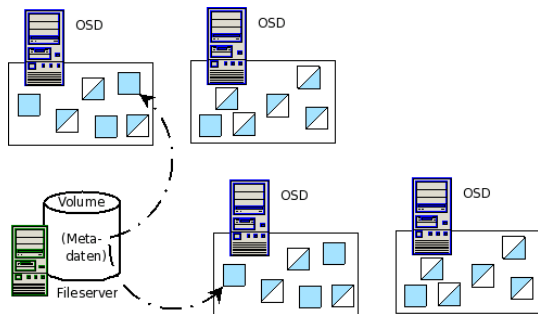
OpenAFS+OSD

Idee  
Features  
Policies



# Performance: Stripes und Copies

- ▶ OSD Dateien bestehen meist aus einem Objekt
- ▶ können aber auch gespiegelt oder verteilt sein
- ▶ z.B. zwei identische Objekte auf zwei OSDs (**copies**)
- ▶ z.B. zwei Hälften des Objekts (**stripes**)
- ▶ bis zu 8 Objekte pro Segment, z.B. 4 Stripes mit 2 Kopien



stripes=1, copies=2

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

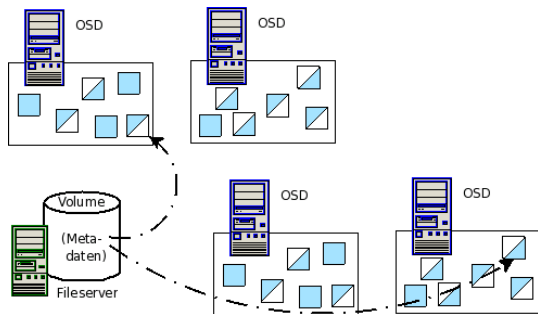
OpenAFS+OSD

Idee  
Features  
Policies



# Performance: Stripes und Copies

- ▶ OSD Dateien bestehen meist aus einem Objekt
- ▶ können aber auch gespiegelt oder verteilt sein
- ▶ z.B. zwei identische Objekte auf zwei OSDs (**copies**)
- ▶ z.B. zwei Hälften des Objekts (**stripes**)
- ▶ bis zu 8 Objekte pro Segment, z.B. 4 Stripes mit 2 Kopien



stripes=2,copies=1

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

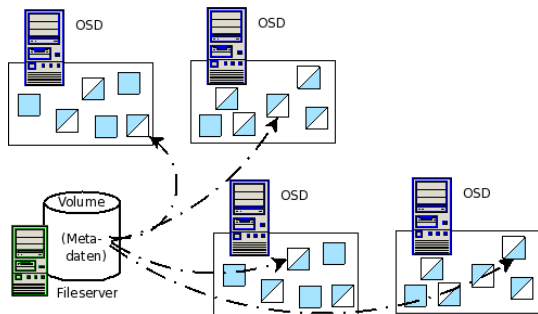
OpenAFS+OSD

Idee  
Features  
Policies



# Performance: Stripes und Copies

- ▶ OSD Dateien bestehen meist aus einem Objekt
- ▶ können aber auch gespiegelt oder verteilt sein
- ▶ z.B. zwei identische Objekte auf zwei OSDs (**copies**)
- ▶ z.B. zwei Hälften des Objekts (**stripes**)
- ▶ bis zu 8 Objekte pro Segment, z.B. 4 Stripes mit 2 Kopien



stripes=2,copies=2

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

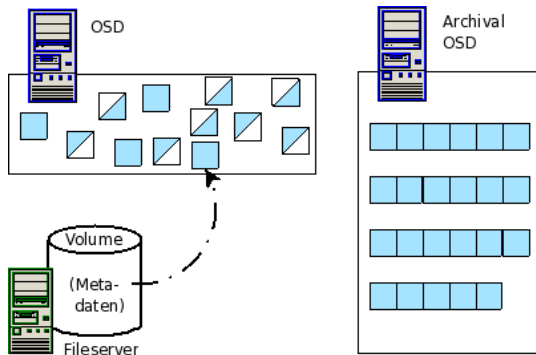
OpenAFS+OSD

Idee  
Features  
Policies



# Interface zu hierarchischem Storage

- ▶ spezieller Storage: “**Archival** OSD”
  - ▶ speichert Daten in z.B. TSM-HSM
- ▶ Files erhalten eine/mehrere Kopien in archival OSDs
- ▶ MD5-Summe ist in Metadaten auf AFS Fileserver
- ▶ Archivierung: `fs arch volume.c`
- ▶ aktive Kopie entfernen: `fs wipe volume.c`
  - ▶ bei Zugriff: transparentes Restore



OpenAFS+OSD  
with Policies

Felix Frank

AFS

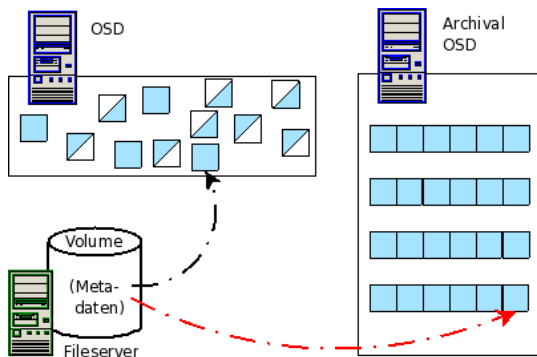
Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies

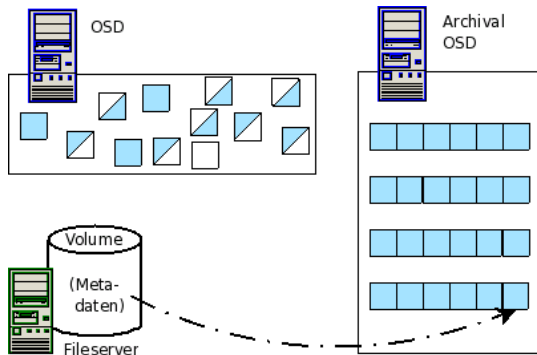


- ▶ spezieller Storage: “Archival OSD”
  - ▶ speichert Daten in z.B. TSM-HSM
- ▶ Files erhalten eine/mehrere Kopien in archival OSDs
- ▶ **MD5-Summe** ist in Metadaten auf AFS Fileserver
- ▶ Archivierung: `fs arch volume.c`
- ▶ aktive Kopie entfernen: `fs wipe volume.c`
  - ▶ bei Zugriff: transparentes Restore





- ▶ spezieller Storage: “Archival OSD”
  - ▶ speichert Daten in z.B. TSM-HSM
- ▶ Files erhalten eine/mehrere Kopien in archival OSDs
- ▶ MD5-Summe ist in Metadaten auf AFS Fileserver
- ▶ Archivierung: `fs arch volume.c`
- ▶ aktive Kopie entfernen: `fs wipe volume.c`
  - ▶ bei Zugriff: transparentes Restore



```
$ fs ls
```

```
d rwx ffrank 2048 13:49:08 .
m rwx root 2048 13:47:47 ..
o rwx ffrank 2521335 13:48:02 photos.tar
f rw- ffrank 35723 13:49:08 install.log
```

```
$ fs osd photos.tar
```

photos.tar has 132 bytes of osd metadata, v=3

On-line, 1 segm, flags=0x0

segment:

lng=0, ofs=0, stripes=1, strsize=0, cop=1, 1 objects

object:

obj=536870991.15108.1047363.0, osd=19, stripe=0

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



```
$ fs ls
```

```
d rwx ffrank 2048 13:49:08 .
m rwx root 2048 13:47:47 ..
o rwx ffrank 2521335 13:48:02 photos.tar
f rw- ffrank 35723 13:49:08 install.log
```

```
$ fs osd photos.tar
```

photos.tar has 132 bytes of osd metadata, v=3

On-line, 1 segm, flags=0x0

segment:

lng=0, offs=0, stripes=1, strsize=0, cop=1, 1 objects

object:

obj=536870991.15108.1047363.0, osd=19, stripe=0

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



# Archivierung und Wiping

```
$ fs arch photos.tar
photos.tar done
$ fs osd photos.tar
```

photos.tar has 284 bytes of osd metadata, v=3

**On-line**, 1 segm, flags=0x0

segment:

lmg=0, ofs=0, stripes=1, strsize=0, cop=1, 1 objects

object:

obj=536870991.15108.1047363.0, **osd=19**, stripe=0

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
**Features**  
Policies



# Archivierung und Wiping

```
$ fs arch photos.tar
```

```
photos.tar done
```

```
$ fs osd photos.tar
```

```
photos.tar has 284 bytes of osd metadata, v=3
```

```
On-line, 1 segm, flags=0x0
```

```
segment:
```

```
  lng=0, ofs=0, stripes=1, strsize=0, cop=1, 1 objects
```

```
object:
```

```
  obj=536870991.15108.1047363.0, osd=19, stripe=0
```

```
Archive, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x0
```

```
segment:
```

```
  lng=2521335, ofs=0, stripes=1, strsize=0, cop=1, 1 obj
```

```
object:
```

```
  obj=536870991.15108.1047363.1, osd=18, stripe=0
```

```
metadata:
```

```
md5=45e2377a073b7e0b3367a1bcd4c0823d
```

```
as from 2009-03-02 14:09:40
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



```
$ fs arch photos.tar
```

```
photos.tar done
```

```
$ fs osd photos.tar
```

```
photos.tar has 284 bytes of osd metadata, v=3
```

```
On-line, 1 segm, flags=0x0
```

```
segment:
```

```
  lng=0, ofs=0, stripes=1, strsize=0, cop=1, 1 objects
```

```
object:
```

```
  obj=536870991.15108.1047363.0, osd=19, stripe=0
```

```
Archive, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x0
```

```
segment:
```

```
  lng=2521335, ofs=0, stripes=1, strsize=0, cop=1, 1 obj
```

```
object:
```

```
  obj=536870991.15108.1047363.1, osd=18, stripe=0
```

```
metadata:
```

```
  md5=45e2377a073b7e0b3367a1bcd4c0823d
```

```
  as from 2009-03-02 14:09:40
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich

Konzepte

Alternativen

OpenAFS+OSD

Idee

Features

Policies



```
$ fs arch photos.tar
photos.tar done
$ fs osd photos.tar
```

photos.tar has 284 bytes of osd metadata, v=3

**On-line**, 1 segm, flags=0x0

segment:

lng=0, ofs=0, stripes=1, strsize=0, cop=1, 1 objects

object:

obj=536870991.15108.1047363.0, **osd=19**, stripe=0

**Archive**, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x0

segment:

lng=2521335, ofs=0, stripes=1, strsize=0, cop=1, 1 obj

object:

obj=536870991.15108.1047363.1, **osd=18**, stripe=0

metadata:

**md5**=45e2377a073b7e0b3367a1bcd4c0823d

as from 2009-03-02 14:09:40

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



# Archivierung und Wiping

```
$ fs wipe photos.tar  
done
```

```
$ fs ls ; fs osd photos.tar
```

```
d rwx ffrank 2048 13:49:08 .  
m rwx root 2048 13:47:47 ..  
w rwx ffrank 2521335 13:48:02 photos.tar  
f rw- ffrank 35723 13:49:08 install.log
```

photos.tar has 160 bytes of osd metadata, v=3

Archive, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x2

segment:

lng=2521335, offs=0, stripes=1, strsize=0, cop=1, 1 obj

object:

obj=536870991.15108.1047363.1, osd=18, stripe=0

metadata:

md5=45e2377a073b7e0b3367a1bcd4c0823d

as from 2009-03-02 14:09:40

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies





# Archivierung und Wiping

```
$ fs wipe photos.tar
```

```
done
```

```
$ fs ls ; fs osd photos.tar
```

```
d rwx ffrank 2048 13:49:08 .
m rwx root 2048 13:47:47 ..
w rwx ffrank 2521335 13:48:02 photos.tar
f rw- ffrank 35723 13:49:08 install.log
```

photos.tar has 160 bytes of osd metadata, v=3

Archive, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x2

segment:

lng=2521335, offs=0, stripes=1, strsize=0, cop=1, 1 obj

object:

obj=536870991.15108.1047363.1, osd=18, stripe=0

metadata:

md5=45e2377a073b7e0b3367a1bcd4c0823d

as from 2009-03-02 14:09:40

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



```
$ fs wipe photos.tar
```

```
done
```

```
$ fs ls ; fs osd photos.tar
```

```
d rwx ffrank 2048 13:49:08 .
m rwx root 2048 13:47:47 ..
w rwx ffrank 2521335 13:48:02 photos.tar
f rw- ffrank 35723 13:49:08 install.log
```

photos.tar has **160** bytes of osd metadata, v=3

Archive, dv=1, 2009-03-02 13:48:02, 1 segm, flags=0x2

segment:

lng=2521335, offs=0, stripes=1, strsize=0, cop=1, 1 obj

object:

obj=536870991.15108.1047363.1, osd=18, stripe=0

metadata:

md5=45e2377a073b7e0b3367a1bcd4c0823d

as from 2009-03-02 14:09:40

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
w rwx ffrank 2521335 13:48:02 photos.tar
```

```
$ fs flush photos.tar
```

```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
w rwx ffrank 2521335 13:48:02 photos.tar
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
**Features**  
Policies



```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
w rwx  ffrank  2521335  13:48:02  photos.tar
```

```
$ fs flush photos.tar
```

```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
o rwx  ffrank  2521335  13:48:02  photos.tar
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
w rwx ffrank 2521335 13:48:02 photos.tar
```

```
$ fs flush photos.tar
```

```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
o rwx ffrank 2521335 13:48:02 photos.tar
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
w rwx ffrank 2521335 13:48:02 photos.tar
```

```
$ fs flush photos.tar
```

```
$ tar tvf photos.tar
```

```
...
```

```
$ fs ls photos.tar
```

```
o rwx ffrank 2521335 13:48:02 photos.tar
```

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



**Einzig notwendig:** `fs archive` und `fs wipe`!

- ▶ Skripte zur Automatisierung sind vorhanden

## Stand der Entwicklung

- ▶ seit langem getestet
- ▶ in Garching seit über einem Jahr in Produktion
- ▶ bei DESY im Test, Produktion dieses Jahr

## Beispielzelle: [ipp-garching.mpg.de](http://ipp-garching.mpg.de)

- ▶ 46 AFS Fileserver, 30 OSDs
- ▶ ~280 TB in AFS Space
- ▶ ~550 TB in Object Storage
- ▶ Standorte in Garching und Greifswald
- ▶ Geographie wird berücksichtigt

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies





## Stand der Entwicklung

- ▶ seit langem getestet
- ▶ in Garching seit über einem Jahr in Produktion
- ▶ bei DESY im Test, Produktion dieses Jahr

## Beispielzelle: [ipp-garching.mpg.de](http://ipp-garching.mpg.de)

- ▶ 46 AFS Fileserver, 30 OSDs
- ▶ ~280 TB in AFS Space
- ▶ ~550 TB in Object Storage
- ▶ Standorte in Garching und Greifswald
- ▶ Geographie wird berücksichtigt

### AFS

Vergleich  
Konzepte  
Alternativen

### OpenAFS+OSD

Idee  
Features  
Policies



## Stand der Entwicklung

- ▶ seit langem getestet
- ▶ in Garching seit über einem Jahr in Produktion
- ▶ bei DESY im Test, Produktion dieses Jahr

## Beispielzelle: [ipp-garching.mpg.de](http://ipp-garching.mpg.de)

- ▶ 46 AFS Fileserver, 30 OSDs
- ▶ ~280 TB in AFS Space
- ▶ ~550 TB in Object Storage
- ▶ Standorte in Garching und Greifswald
- ▶ Geographie wird berücksichtigt



## Das AFS Dateisystem

Vergleich mit NFSv3

Grundlegende AFS-Konzepte

Alternativ: Object Storage Systeme, Beispiel Lustre

### AFS

Vergleich

Konzepte

Alternativen

### OpenAFS+OSD

Idee

Features

Policies

## Die Vereinigung: Das OpenAFS+OSD Projekt

Idee und Funktion

Nützliche Features von OpenAFS+OSD

Fileserver Policies für OSD-Verwendung



## Größenkriterium zu schwach

- ▶ bekanntes Problem: `root` Tool
  - ▶ erzeugt grosse Dateien, schreibt jedoch zuerst nur 64-byte Header
  - ▶ Dateien sind aber an suffix `.root` erkennbar
- ▶ es wäre wünschenswert, Dateien am Namen zu identifizieren
- ▶ Zusatzfeature: automatisches Striping/Mirroring
  - ▶ ist ursprünglich nur mit umständlichem Kommando zugänglich
  - ▶ `fs createstripedfile libhello.a -stripes 2 -copies 2`
  - ▶ danach nutzt Schreiboperation die Stripes und Kopien
- ▶ Policies erkennen
  - ▶ Dateigröße, Dateiname (Wildcard- oder Regex), Benutzer und (AFS-)Gruppe
- ▶ Policies gelten
  - ▶ für ganzes Volume oder einzelnes Directory

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Größenkriterium zu schwach

- ▶ bekanntes Problem: `root` Tool
  - ▶ erzeugt grosse Dateien, schreibt jedoch zuerst nur 64-byte Header
  - ▶ Dateien sind aber an suffix `.root` erkennbar
- ▶ es wäre wünschenswert, Dateien am Namen zu identifizieren
- ▶ Zusatzfeature: automatisches Striping/Mirroring
  - ▶ ist ursprünglich nur mit umständlichem Kommando zugänglich
  - ▶ `fs createstripedfile libhello.a -stripes 2 -copies 2`
  - ▶ danach nutzt Schreiboperation die Stripes und Kopien
- ▶ Policies erkennen
  - ▶ Dateigröße, Dateiname (Wildcard- oder Regex), Benutzer und (AFS-)Gruppe
- ▶ Policies gelten
  - ▶ für ganzes Volume oder einzelnes Directory

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Größenkriterium zu schwach

- ▶ bekanntes Problem: `root` Tool
  - ▶ erzeugt grosse Dateien, schreibt jedoch zuerst nur 64-byte Header
  - ▶ Dateien sind aber an suffix `.root` erkennbar
- ▶ es wäre wünschenswert, Dateien am Namen zu identifizieren
- ▶ Zusatzfeature: automatisches Striping/Mirroring
  - ▶ ist ursprünglich nur mit umständlichem Kommando zugänglich
  - ▶ `fs createstripedfile libhello.a -stripes 2 -copies 2`
  - ▶ danach nutzt Schreiboperation die Stripes und Kopien
- ▶ Policies erkennen
  - ▶ Dateigröße, Dateiname (Wildcard- oder Regex), Benutzer und (AFS-)Gruppe
- ▶ Policies gelten
  - ▶ für ganzes Volume oder einzelnes Directory

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## Größenkriterium zu schwach

- ▶ bekanntes Problem: `root` Tool
  - ▶ erzeugt grosse Dateien, schreibt jedoch zuerst nur 64-byte Header
  - ▶ Dateien sind aber an suffix `.root` erkennbar
- ▶ es wäre wünschenswert, Dateien am Namen zu identifizieren
- ▶ Zusatzfeature: automatisches Striping/Mirroring
  - ▶ ist ursprünglich nur mit umständlichem Kommando zugänglich
  - ▶ `fs createstripedfile libhello.a -stripes 2 -copies 2`
  - ▶ danach nutzt Schreiboperation die Stripes und Kopien
- ▶ Policies erkennen
  - ▶ Dateigröße, Dateiname (Wildcard- oder Regex), Benutzer und (AFS-)Gruppe
- ▶ Policies gelten
  - ▶ für ganzes Volume oder einzelnes Directory

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

Beispiele

- ▶ `~"*root"`  $\Rightarrow$  `location=osd stop`;
  - ▶ `true`  $\Rightarrow$  `stripes=2`;
  - ▶ `>10M`  $\Rightarrow$  `copies=2`;
  - ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2`;
- } 2 Regeln

OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies





Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

## Beispiele

- ▶ `~"*root"`  $\Rightarrow$  `location=osd stop;`
- ▶ `true`  $\Rightarrow$  `stripes=2;`
- ▶ `>10M`  $\Rightarrow$  `copies=2;`
- ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2;` } 2 Regeln

## OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

Beispiele

- ▶  $\sim$  `"*.root"`  $\Rightarrow$  `location=osd stop;`
- ▶ `true`  $\Rightarrow$  `stripes=2;`
- ▶ `>10M`  $\Rightarrow$  `copies=2;`
- ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2;` } 2 Regeln

OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

Beispiele

- ▶ `~"*root"`  $\Rightarrow$  `location=osd stop`;
  - ▶ `true`  $\Rightarrow$  `stripes=2`;
  - ▶ `>10M`  $\Rightarrow$  `copies=2`;
  - ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2`;
- } 2 Regeln

OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

Beispiele

- ▶ `~"*root"`  $\Rightarrow$  `location=osd stop`;
  - ▶ `true`  $\Rightarrow$  `stripes=2`;
  - ▶ `>10M`  $\Rightarrow$  `copies=2`;
  - ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2`;
- } 2 Regeln

OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



Regeln: *Bedingung*  $\Rightarrow$  *Ort, Stripes/Kopien*

- ▶ Ort: `osd`, `local`, `dynamic`
- ▶ Stripes/Kopien: 1,2,4,8 + Stripe-Size:12,13,...,19
- ▶ `stop` Option: wenn *Bedingung* dann beende Policy-Evaluierung

Beispiele

- ▶  $\sim$  `"*.root"`  $\Rightarrow$  `location=osd stop`;
  - ▶ `true`  $\Rightarrow$  `stripes=2`;
  - ▶ `>10M`  $\Rightarrow$  `copies=2`;
  - ▶ `>100M`  $\Rightarrow$  `copies=1, stripes=2`;
- } 2 Regeln

OSD Datenbank

- ▶ neuer Prozess auf Datenbankservern
- ▶ hält Policies und OSDs mit Prioritäten
- ▶ Policies können leicht ersetzt werden
- ▶ Fileserver hat immer Überblick über Platz auf OSDs

OpenAFS+OSD  
with Policies

Felix Frank

AFS

Vergleich  
Konzepte  
Alternativen

OpenAFS+OSD

Idee  
Features  
Policies



## OpenAFS Download:

- ▶ <http://www.openafs.org/>

## Mailing-Liste: openafs-info (at) openafs.org

- ▶ auch das DESY AFS-Team hilft gern per Mail

## OpenAFS+OSD:

- ▶ *svn co https://svnsrv.desy.de/public/openafs-osd/trunk/*
- ▶ Pakete leider derzeit nur für Scientific Linux 5

## Das OSD-Projekt wurde entwickelt von

- ▶ Hartmut Reuter (RZG)
- ▶ Rainer Többicke (CERN)
- ▶ Andrei Maslennikov (CASPUR)

## Veranstaltungen:

- ▶ European AFS meeting 2009 in Rome (September)
  - ▶ <http://www.dia.uniroma3.it/~afscon09/>
- ▶ ASF & Kerberos Best Practices Workshop (Juni)
  - ▶ <http://workshop.openafs.org/afsbpw09/cfp.html>



## OpenAFS Download:

- ▶ <http://www.openafs.org/>

## Mailing-Liste: openafs-info (at) openafs.org

- ▶ auch das DESY AFS-Team hilft gern per Mail

## OpenAFS+OSD:

- ▶ *svn co <https://svnsrv.desy.de/public/openafs-osd/trunk/>*
- ▶ Pakete leider derzeit nur für Scientific Linux 5

## Das OSD-Projekt wurde entwickelt von

- ▶ Hartmut Reuter (RZG)
- ▶ Rainer Többicke (CERN)
- ▶ Andrei Maslennikov (CASPUR)

## Veranstaltungen:

- ▶ European AFS meeting 2009 in Rome (September)
  - ▶ <http://www.dia.uniroma3.it/~afscon09/>
- ▶ ASF & Kerberos Best Practices Workshop (Juni)
  - ▶ <http://workshop.openafs.org/afsbpw09/cfp.html>



## OpenAFS Download:

- ▶ <http://www.openafs.org/>

## Mailing-Liste: openafs-info (at) openafs.org

- ▶ auch das DESY AFS-Team hilft gern per Mail

## OpenAFS+OSD:

- ▶ *svn co* <https://svnsrv.desy.de/public/openafs-osd/trunk/>
- ▶ Pakete leider derzeit nur für Scientific Linux 5

## Das OSD-Projekt wurde entwickelt von

- ▶ Hartmut Reuter (RZG)
- ▶ Rainer Többicke (CERN)
- ▶ Andrei Maslennikov (CASPUR)

## Veranstaltungen:

- ▶ European AFS meeting 2009 in Rome (September)
  - ▶ <http://www.dia.uniroma3.it/~afscon09/>
- ▶ ASF & Kerberos Best Practices Workshop (Juni)
  - ▶ <http://workshop.openafs.org/afsbpw09/cfp.html>





## OpenAFS Download:

- ▶ <http://www.openafs.org/>

## Mailing-Liste: openafs-info (at) openafs.org

- ▶ auch das DESY AFS-Team hilft gern per Mail

## OpenAFS+OSD:

- ▶ *svn co* <https://svnsrv.desy.de/public/openafs-osd/trunk/>
- ▶ Pakete leider derzeit nur für Scientific Linux 5

## Das OSD-Projekt wurde entwickelt von

- ▶ Hartmut Reuter (RZG)
- ▶ Rainer Többicke (CERN)
- ▶ Andrei Maslennikov (CASPUR)

## Veranstaltungen:

- ▶ European AFS meeting 2009 in Rome (September)
  - ▶ <http://www.dia.uniroma3.it/~afscon09/>
- ▶ ASF & Kerberos Best Practices Workshop (Juni)
  - ▶ <http://workshop.openafs.org/afsbpw09/cfp.html>



## OpenAFS Download:

- ▶ <http://www.openafs.org/>

## Mailing-Liste: [openafs-info \(at\) openafs.org](mailto:openafs-info@openafs.org)

- ▶ auch das DESY AFS-Team hilft gern per Mail

## OpenAFS+OSD:

- ▶ *svn co* <https://svnsrv.desy.de/public/openafs-osd/trunk/>
- ▶ Pakete leider derzeit nur für Scientific Linux 5

## Das OSD-Projekt wurde entwickelt von

- ▶ Hartmut Reuter (RZG)
- ▶ Rainer Többicke (CERN)
- ▶ Andrei Maslennikov (CASPUR)

## Veranstaltungen:

- ▶ European AFS meeting 2009 in Rome (September)
  - ▶ <http://www.dia.uniroma3.it/~afscon09/>
- ▶ ASF & Kerberos Best Practices Workshop (Juni)
  - ▶ <http://workshop.openafs.org/afsbpw09/cfp.html>



## Zusätzliche AFS Dokumentation

- ▶ <http://www.openafs.org/doc/index.htm>
- ▶ Grand Central Org <http://grand.central.org/>
- ▶ AFS Wiki  
<http://grand.central.org/twiki/bin/view/AFSLore/WebHome>
- ▶ AFS-Seiten des CERN in  
<http://consult.cern.ch/services/afs/>
- ▶ Universitäten
  - ▶ <http://www.tu-braunschweig.de/it/services/sys/afs/>
  - ▶ <https://rz-static.uni-hohenheim.de/netzwerkbetriebssysteme/afs/kurs-folien/sahyoun/afs-einfuehrungskurs.ppt>
  - ▶ <http://www.tu-chemnitz.de/urz/kurse/unterlagen/admin-afs/basics.html>

