Gespiegelte Solaris Server mit AVS

Ein Erfahrungsbericht

Jochen Schulz

Georg-August Universität Göttingen

29ter September 2009

Aufbau

1. **Einleitung**
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. **AVS**
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

Aufbau

1. **Einleitung**
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. **AVS**
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

Aufbau

1. **Einleitung**
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. **AVS**
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

Praxis-Beispiel I

Anforderungen

- Zentraler Fileserver, inklusive Auslieferung Homedirectories.
- 20-70 parallele Clients. Aufgrund der Homedirectories einige Performance nötig.
- Backups und/oder Snapshots, die vom Nutzer bedient werden können.
- Redundanz (Hardware und Daten).
- Kostengünstig.

Praxis-Beispiel II

Folgerungen

- Snapshots: ZFS ! Weitere externe Backups können bequem zentral durchgeführt werden.
  $\Rightarrow$ OpenSolaris
- Redundanz Hardware: zwei identische Server mit redundanten Teilen (Netzteil).
- Redundanz Daten: Ziel: Raid 1 über Netzwerk
  Linux: LINBIT® DRBD® .
  OpenSolaris: Sun™ StorageTek Availability Suite (AVS)
- Leistung/Kosten: Rack-Server mit 6 1TB-Platten.

Aufbau

1. **Einleitung**
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. **AVS**
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

Am Anfang... Das Setup

- 2 identische Server mit
  - (SATA)-RAID-Controller
  - 6 SATA Platten mit 1 TB
  - redundanten Netzteilen
- 2 Netzwerkkabel extern.
- 1 Netzwerkkabel (crossover) für die direkte Verbindung beider Server.
- 50 wartende clients ..

Die ersten Schritte

1ter Versuch mit OpenSolaris 2008.11

- AVS-Pakete waren fehlerhaft $\Rightarrow$ auch mit Hilfe war kein funktionierender Betrieb zu erreichen
- Ein Hinweis brachte mich auf...
  $\Rightarrow$ Nexenta

Aufbau

1. **Einleitung**
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. **AVS**
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

Nexenta

Nexenta (`nexenta.org`) erklärt anhand von einem Vergleich:

OpenSolaris vs. Nexenta (aus der Sicht von einem Linux-Admin)

Solaris-„Feel “	Ubuntu-„Feel“
rel. geringe Software-Auswahl	∼Ubuntu Hardy-Repository
Desktop-orientiert	Server-orientiert
-AVS	+AVS
Autom. Snapshots	Manuelle Nachinst.

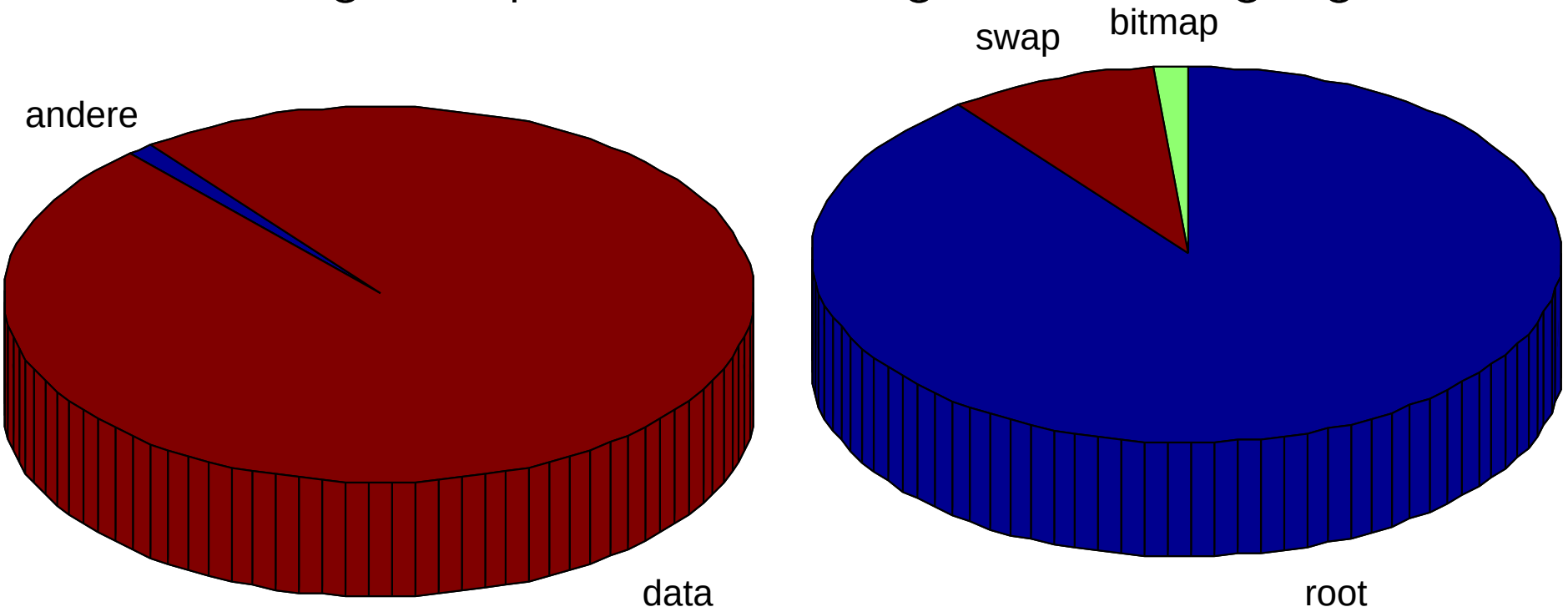
Aufbau

1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. **Basis-Installation**
3. AVS
   - Einführung
   - Installation
   - Tuning
4. Die (kurze) Kür: LDAP, NFSv4 und Samba

Festplatten-Aufteilung

- Es sind keine separaten System-Platten vorhanden $\Rightarrow$ Aufteilung in mehrere Slices.

Den Anforderungen entsprechend wurde folgende Aufteilung angestrebt:



Installation

- RAID-Controller auf JBOD-Modus stellen.
- Nexenta-Installer: installiert syspool auf eine oder mehrere Platten (dann Raid 1).
- Verkleinerung root-slice (und syspool) nach der Standard-Installation *gescheitert*. (durch snapshot, zfs send, zfs receive..) (Boot-Fähigkeit!).

Installation

- RAID-Controller auf JBOD-Modus stellen.
- Nexenta-Installer: installiert syspool auf eine oder mehrere Platten (dann Raid 1).
- Verkleinerung root-slice (und syspool) nach der Standard-Installation *gescheitert*. (durch snapshot, zfs send, zfs receive..) (Boot-Fähigkeit!).
- $\Rightarrow$ Anpassung des Installers: Der Installer ist ein Bash-Skript:
  - Installations-CD einlegen
  - Bevor die richtige Installation beginnt F2 drücken
  - in der Shell: `vim `which nexenta-install.sh``
  - in Zeile 1023 von `nexenta-install.sh` stand die Berechnung des root-slices; diese den Bedürfnissen anpassen.
  - Installation starten.

Abschluss Partitionierung

- Erzeugen von Daten- und bitmap-slice: `format`.
- Partitionierungstabelle auf alle anderen 5 Festplatten kopieren

```
prtvtoc /dev/rdsk/c0t0d0s2 | fmthard -s - /dev/rdsk/c0t1d0s2
prtvtoc /dev/rdsk/c0t0d0s2 | fmthard -s - /dev/rdsk/c0t2d0s2
...
```

- Syspool zum Raid 1 machen:

```
zpool attach -f rpool c0t0d0s0 c0t1d0s0
```

- Grub für die andere Platte installieren

```
installgrub -m /boot/grub/stage1 /boot/grub/stage2 \
  /dev/rdsk/c0t1d0s0
```

Aufbau

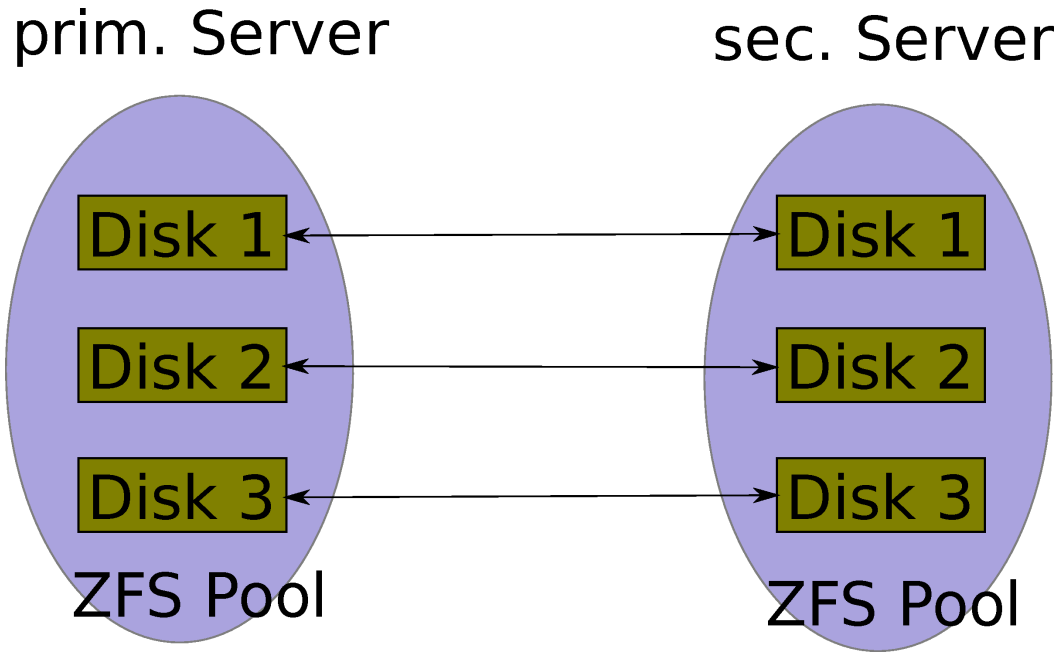
1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. Basis-Installation
3. AVS
   - Einführung
   - Installation
   - Tuning
4. Die (kurze) Kür: LDAP, NFSv4 und Samba

Aufbau

1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. Basis-Installation
3. AVS
   - Einführung
   - Installation
   - Tuning
4. Die (kurze) Kür: LDAP, NFSv4 und Samba

Was ist AVS ?

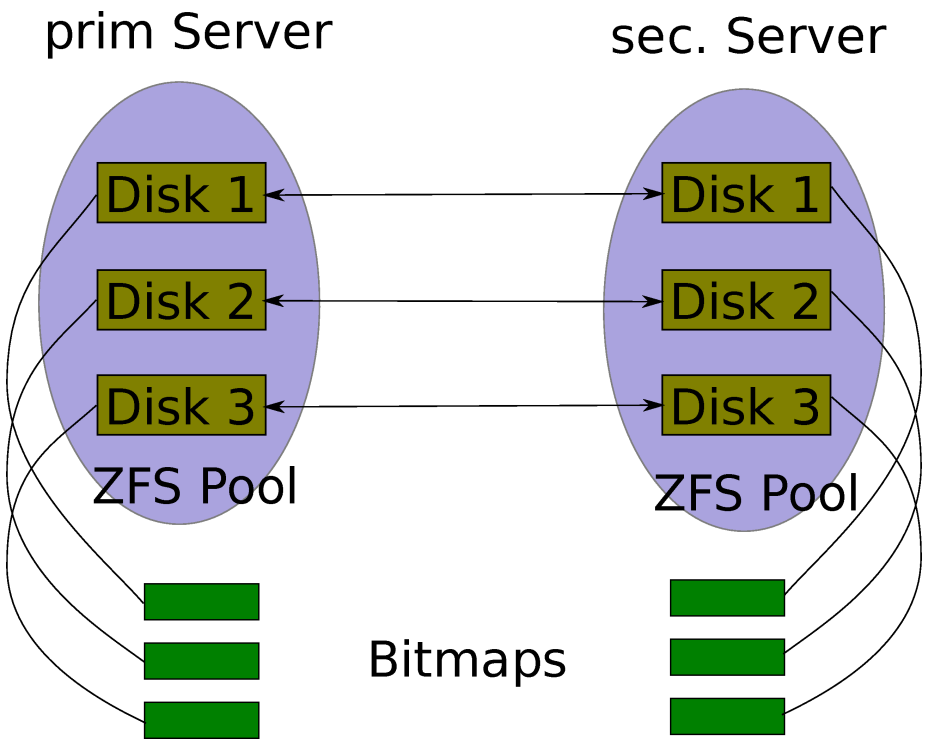
Sun™ StorageTek Availability Suite (AVS): Daten einer Partition/Slice können über das Netzwerk an identische Partitionen gesendet werden (Daten werden Synchron gehalten).



Bitmaps

AVS benötigt für jede gespiegelte Partition eine sogenannte bitmap-volume:

enthaltene Informationen: veränderte Blocks .



Sync, Async und Logging

sync

- Daten $\Rightarrow$ Platten vom primary und secondary Server $\Rightarrow$ Schreibvorgang fertig.

Sync, Async und Logging

sync

- Daten $\Rightarrow$ Platten vom primary und secondary Server $\Rightarrow$ Schreibvorgang fertig.

async

- Daten, Bitmap-Information $\Rightarrow$ Queue und primary Server $\Rightarrow$ Schreibvorgang fertig.
- Benötigt eine memory- oder disk-Queue (evtl. Überlauf).
- Queue $\Rightarrow$ Netz $\Rightarrow$ sec. Server.

Sync, Async und Logging

sync

- Daten $\Rightarrow$ Platten vom primary und secondary Server $\Rightarrow$ Schreibvorgang fertig.

async

- Daten, Bitmap-Information $\Rightarrow$ Queue und primary Server $\Rightarrow$ Schreibvorgang fertig.
- Benötigt eine memory- oder disk-Queue (evtl. Überlauf).
- Queue $\Rightarrow$ Netz $\Rightarrow$ sec. Server.

logging

- Daten, Bitmap-Information $\Rightarrow$ primary Server $\Rightarrow$ Schreibvorgang fertig.

Sync vs. Async

sync

- grosser negativer Einfluss auf die Performance.
- sehr hohe Datensicherheit.

async

- leichter negativer Einfluss auf die Performance.
- relativ hohe Datensicherheit (evtl. Datenverlust von Sekunden).
- etwas schwieriger zu konfigurieren; mehr Hardware-Anforderungen.

$\Rightarrow$ `async`-Modus (geringfügig schlechtere Datensicherheit und signifikant bessere Perfomance)

Queue - memory oder disk ?

Blocking Modus:
Queue voll $\Rightarrow$ blockt weitere Schreibvorgänge bis Queue abgearbeitet wird
Nonblocking Modus:
Queue voll $\Rightarrow$ AVS geht in den logging-Modus

Disk-Queue

- Kann prinzipiell mehr Daten aufnehmen
- Blocking Modus / nonblocking Modus
- Benötigt weitere slices

Memory-Queue

- Nur blocking Modus

Aufbau

1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. Basis-Installation
3. AVS
   - Einführung
   - Installation
   - Tuning
4. Die (kurze) Kür: LDAP, NFSv4 und Samba

Vorbereitung

- installation der Pakete:

```
apt-get install sunwspsvu sunwrdcr sunwscmr sunwrdcu \
    sunwscmu sunwspsvr
```

- Bestimmung der Grösse der Bitmap-Volumes:

```
> dsbitmap -r /dev/rdsk/c0t0d0s3
Data volume (/dev/rdsk/c0t0d0s3) size: 1931013000 blocks
Required bitmap volume size:
  Sync replication: 7368 blocks
  Async replication with memory queue: 7368 blocks
  Async replication with disk queue: 66304 blocks
```

Vorbereitung II

- Partitionierung anpassen (hier bereits gemacht!).
- Eintrag in `/etc/hosts` (wichtig!)

```
192.168.0.2    secondary
192.168.0.1    primary
```

- AVS-Dienste initialisieren (und starten)

```
dscfgadm
```

Konfiguration

- Setup der Replikation pro Slice:

```
sndradm -E primary /dev/rdsk/c0t0d0s3 /dev/rdsk/c0t0d0s4 \
   secondary /dev/rdsk/c0t0d0s3 /dev/rdsk/c0t0d0s4 \
   ip async g datapool
sndradm -E primary /dev/rdsk/c0t1d0s3 /dev/rdsk/c0t1d0s4 \
   secondary /dev/rdsk/c0t1d0s3 /dev/rdsk/c0t1d0s4 \
   ip async g datapool
...
```

Konfiguration

- Setup der Replikation pro Slice:

```
sndradm -E primary /dev/rdsk/c0t0d0s3 /dev/rdsk/c0t0d0s4 \
   secondary /dev/rdsk/c0t0d0s3 /dev/rdsk/c0t0d0s4 \
   ip async g datapool
sndradm -E primary /dev/rdsk/c0t1d0s3 /dev/rdsk/c0t1d0s4 \
   secondary /dev/rdsk/c0t1d0s3 /dev/rdsk/c0t1d0s4 \
   ip async g datapool
...
```

- Alle Schritte müssen auf beiden Servern durchgeführt werden.

Start!

AVS befindet sich im logging-Modus.

- Synchronisation starten:

```
sndradm -g datapool -u
```

- Datapool erzeugen (erst jetzt!)

```
zpool create -f datapool raidz2 c0t0d0s3 c0t1d0s3 \
  c0t2d0s3 c0t3d0s3 c0t4d0s3 c0t5d0s3
```

Start!

AVS befindet sich im logging-Modus.

- Synchronisation starten:

```
sndradm -g datapool -u
```

- Datapool erzeugen (erst jetzt!)

```
zpool create -f datapool raidz2 c0t0d0s3 c0t1d0s3 \
  c0t2d0s3 c0t3d0s3 c0t4d0s3 c0t5d0s3
```

Up and running!

Monitoring und Alltags-Aufgaben

- Status der Replikation

```
sndradm -g datapool -P
```

- Falls AVS im logging-Modus (Überläufe):

```
sndradm -g datapool -u
```

  $\Rightarrow$ AVS im syncing-Modus.

- Beobachten des Sync durch

```
dsstat -m sndr
```

Aufbau

1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. Basis-Installation
3. AVS
   - Einführung
   - Installation
   - Tuning
4. Die (kurze) Kür: LDAP, NFSv4 und Samba

Tuning - Parameter

Die Queue hat 3 Hauptparameter:

- Maximum Schreibvorgänge
- Maximum 512-Byte Blocks
- Anzahl async flusher threads

Bemerkungen:

- je grösser die Werte umso mehr Performance, aber weniger Sicherheit.
- Grösse der Queue $>$ gepufferte zu verarbeitende Daten (da sonst Stagnierung!).
- async threads: parallele Threads die die Queue abarbeiten (Daten übers Netz senden).
- je mehr Threads, umso schnellere Datenverarbeitung.
- Begrenzung: Speicher-Verbrauch.

Queue-Tuning

- Konfigurationswerte des asynchronen Modus ansehen

```
> sndradm -g datapool -P
autosync: off, max q writes: 81920, max q fbas: 3276800, \
  async threads: 5, mode: async, group: datapool, \
  state: replicating
```

- Monitoring der Queue-Parameter

```
kstat sndr:0:setinfo | grep async
```

Queue-Tuning II

- Ändern des Maximums der Schreibvorgänge

```
sndradm -W 81953 -g datapool
```

- Ändern des Maximums der Blocks

```
sndradm -F 586780 -g datapool
```

- Ändern der Anzahl der Flusher Threads.

```
sndradm -A 5 -g datapool
```

Aufbau

1. Einleitung
   - Motivation
   - Am Anfang...
   - Nexenta
2. Basis-Installation
3. AVS
   - Einführung
   - Installation
   - Tuning
4. **Die (kurze) Kür: LDAP, NFSv4 und Samba**

LDAP client

- Installation

```
apt-get install sunwlldap
```

- Konfiguration

```
ldapclient manual  -a defaultServerList=134.76.80.xxx \
-a serviceSearchDescriptor=password:ou=people,..,dc=de \
-a serviceSearchDescriptor=group:ou=groups,dc=..,dc=de \
-a authenticationMethod=simple  \
-a defaultSearchScope=sub \
-a defaultSearchBase=dc=math,dc=uni-goettingen,dc=de \
-a proxyDN=cn=auth,dc=zimbra,dc=num,dc=math,..,dc=de \
-a proxyPassword=xxxxxxx \
-a credentialLevel=proxy
```

LDAP client II

- Konfiguration ausgeben

```
ldap_cachemgr -g
```

- Starten

```
svcadm start svc:/network/ldap/client:default
```

- `nsswitch.conf.ldap` benutzen
- pam anpassen

```
/etc/pam.conf

login   auth sufficient          pam_unix_auth.so.1
login   auth required            pam_ldap.so.1
```

NFSv4 und Samba

NFSv4

- NFSv4 ist standardmässig aktiviert.
- Auf client und server müssen die user bekannt sein (daher LDAP).
- `idmapd.conf` anpassen (User-mapping).
- In lstinline!/etc/default/nfs-common! idmapd auf `yes` stellen

Samba

- samba von `samba.org`.
- LDAP-backend konfigurieren.
- Ansonsten *Standard*-Konfiguration in `/etc/samba/smb.conf`

Ende

Vielen Dank für ihre Aufmerksamkeit!