The Power of Simplicity

Herzlich Willkommen!

Shared Storage Clustering
mit
Solaris-Systemen

Autoren: B. Miemietz, C. Schmidt

Was heißt Shared Storage Clustering

Nutzung moderner RZ-Infrastrukturen für neuartige Management-Konzepte

The Power of Simplicity

OSL Storage Cluster:

- *Lösung zur Integration von Unix-Servern mit modernen, RAID-basierten Speicherinfrastrukturen*
- *erweitert OS um aufeinander abgestimmte Virtualisierungs-, Management- und Cluster-Funktionalitäten*
- *Speicher- bzw. Volume-Management, Virtualisierung, System- und Applikationsmanagement sowie Clustering werden als Einheit begriffen*
- *das administrative Konzept und die Software selbst zielen auf flexible, virtualisierte Administrations- und Ablaufumgebungen*
- *deutliche Vereinfachung der Abläufe und administrativen Aufgaben im RZ*

"Die Verbindung aus Softwaretechnologie und durchdachter, langfristig angelegter RZ-Organisation beim Anwender hilft, Ressourcen effektiv auszunutzen, Kosten zu senken und zusätzliche Freiheit bei der Auswahl der Systemplattformen zu gewinnen."

Kern des OSL Storage Clusters:

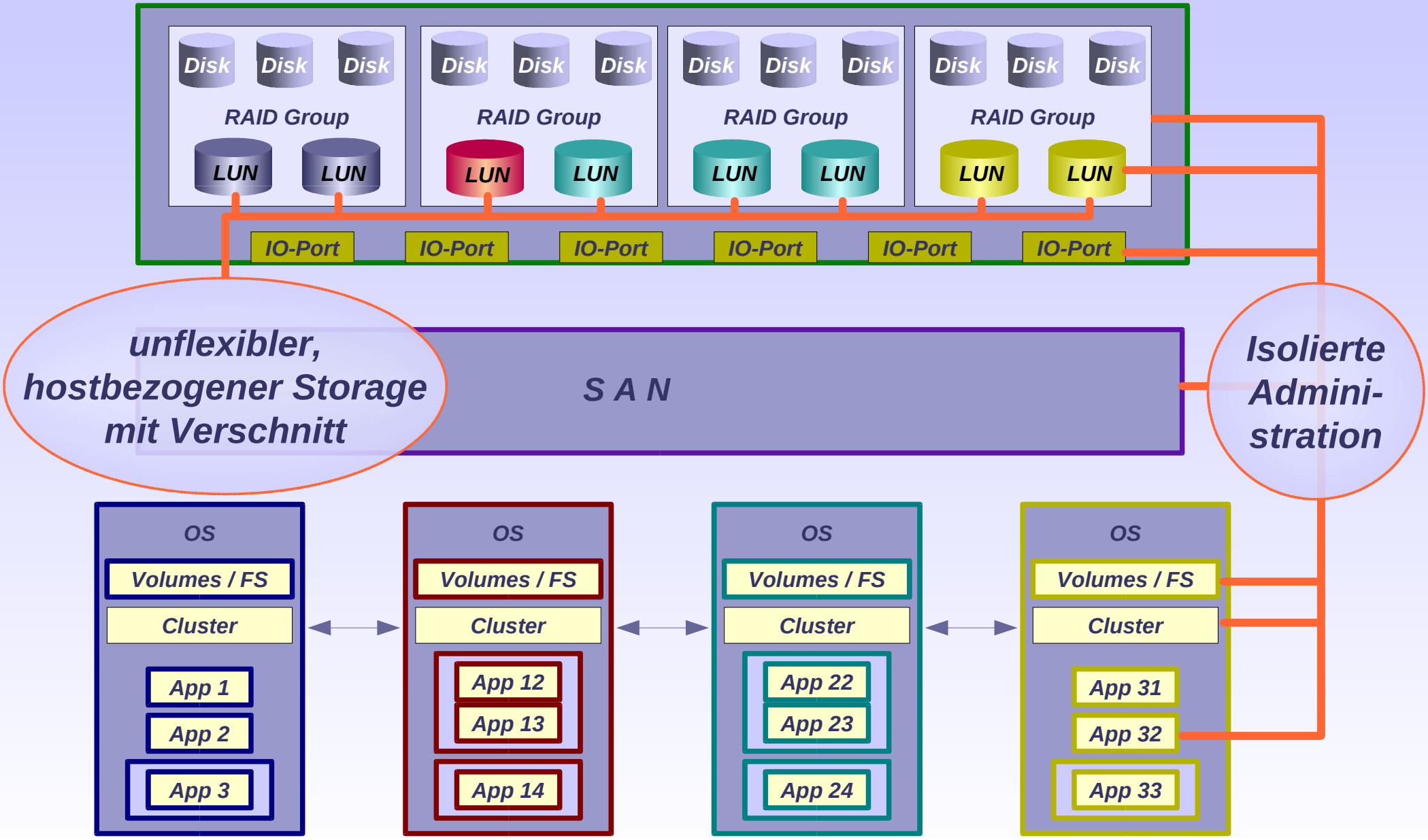
globale, hostbasierte Speichervirtualisierung

Autoren: B. Miemietz, C. Schmidt

Blick in das durchschnittliche Rechenzentrum

Moderne Infrastruktur, vielschichtige Administration

The Power of Simplicity

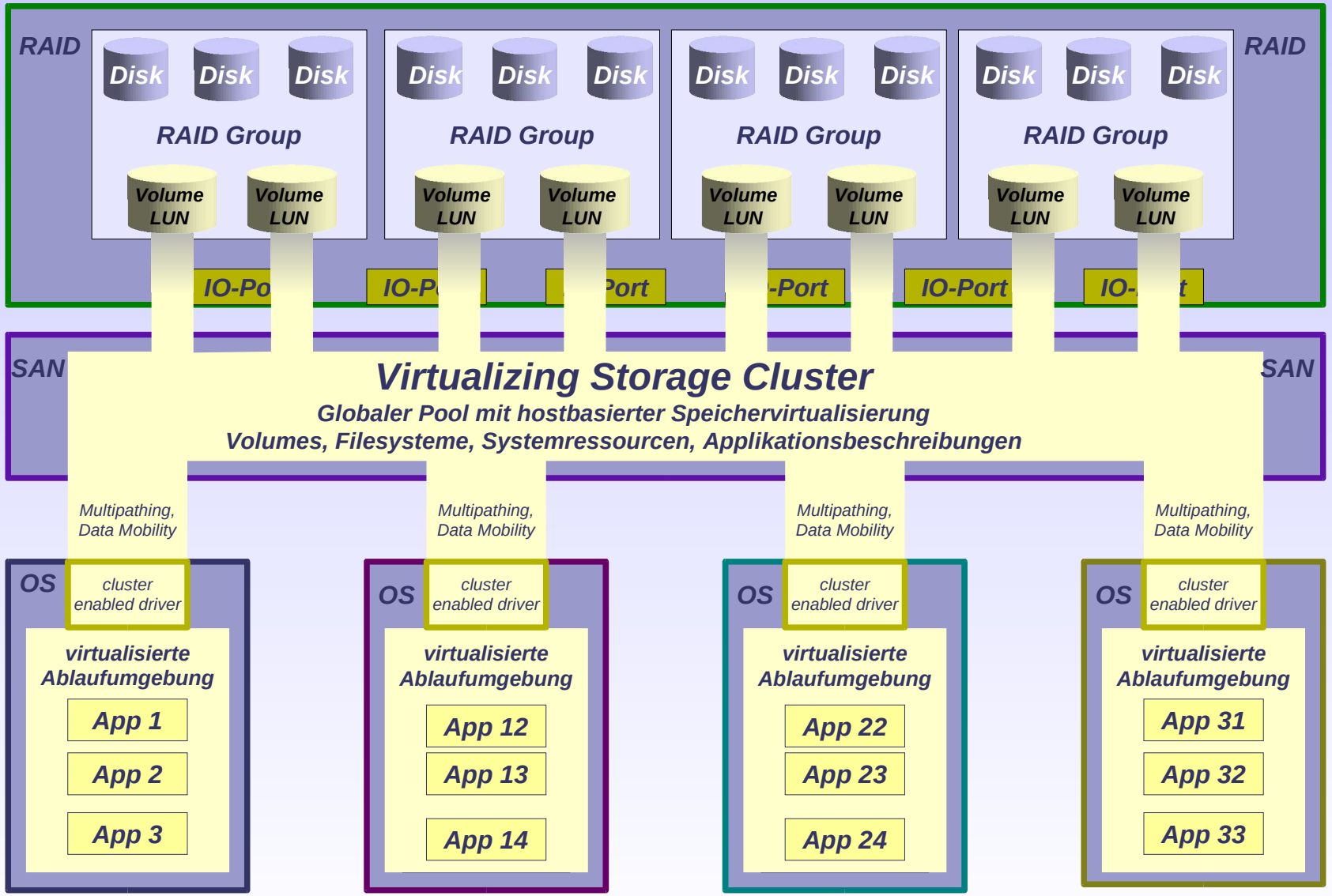


Autoren: B. Miemietz, C. Schmidt

Zentralisierung + Netztopologie = Cluster

OSL Storage Cluster: Vereinfachung durch Integration

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

Die großen Differenzierungsmerkmale

Was unseren Shared Storage Cluster von anderen Konzepten unterscheidet

The Power of Simplicity

- *Administration ausschließlich vom Host aus*
  (Allokieren, Volume erzeugen, Filesystem erzeugen ...)
- *Speichersystem einmal in Betrieb -> (fast) nichts mehr daran zu tun*
- *symmetrisches Konzept*
  - Administration von jeder Maschine aus
  - no single point of failure
- *keine zusätzliche Hardware für Heartbeat o.ä.*
- ***globaler** Storage-Pool*
  - enorme Flexibilität
  - kein Verschnitt
  - optimale Auslastung, auch unter Performance-Aspekten
  - erweitert damit auch Einsatzmöglichkeiten von ZFS
- *Clusterfähigkeit / global devices / namespace von Anfang an*
- *Integration mit Anwendungssteuerung - > Application Awareness*
- *beeindruckende Skalierbarkeit, keine Performance-Engpässe*

- ***enorme Vereinfachungen, enorme Stabilität***

Autoren: B. Miemietz, C. Schmidt

Leistungsumfang im Detail

Speicher-Virtualisierung, Anwendungssteuerung, Hochverfügbarkeit, Backup und DR

The Power of Simplicity

Application Awareness

Bandbreiten-steuerung

User-Management

Application Control Option

- clusterweite Steuerung von Anwendungen
- virtualisierte (hardwareabstrakte) Ablaufumgebungen
- Hochverfügbarkeit
- ressourcenbasiertes Selbstmanagement
- **Application Resource Description**

Application Mirrors

Application Clones

B2D / DASI / DR -Tools

Clusterfähige Speichervirtualisierung

- globale (hostübergreifende) Storage Pools
- Global Disk Inventory
- Global Devices / Global Namespace
- Cluster-Volumemanager mit automatischer Allokation
- Disk Access Management
- IO-Multipathing

Extended Data Management

- Integration RAID-basierter Datenkopien / Snapshots
- Hostbasierte Datenspiegelung
- Live Data Migration
- Data Clones

Autoren: B. Miemietz, C. Schmidt

Leistungsumfang im Detail

Speicher-Virtualisierung, Anwendungssteuerung, Hochverfügbarkeit, Backup und DR

The Power of Simplicity

Application Awareness

Bandbreiten-steuerung

User-Management

Application Control Option

- clusterweite Steuerung von Anwendungen
- virtualisierte (hardwareabstrakte) Ablaufumgebungen
- Hochverfügbarkeit
- ressourcenbasiertes Selbstmanagement

Application Resource Description

clusterweit verfügbare Beschreibung von:

- o genutzten Volumes, Filesystemen, IP-Adressen etc.
- o Start- und Stopmethoden
- o Methoden zum Abbruch einer Applikation
- o Methoden zum Recover einer Applikation
- o Methoden zum Monitoring / Auto-Restart einer Applikation

Dazu Informationen aus Anwendungsdefinition

- o Priorität / Verdrängungsmöglichkeiten
- o Ressourcensteuerung (IO-Bandbreite)
- o Verknüpfung von Usern mit Applikationen

Application Mirrors

Application Clones

B2D / DASI / DR -Tools

Clusterfähige Speichervirtualisierung

- globale (hostübergreifende) Storage Pools
- Global Disk Inventory
- Global Devices / Global Naming
- Cluster-Volumemanager mit automatischer Allokation
- Disk Access Management
- IO-Multipathing

Extended Data Management

- Integration RAID-basierter Datenkopien / Snapshots
- Hostbasierte Datenspiegelung
- Live Data Migration
- Data Clones

OSL Gesellschaft für offene Systemlösungen mbH

Autoren: B. Miemietz, C. Schmidt

Stand: 06/2009

The Power of Simplicity

Los geht's

clusterweite Speichervirtualisierung

Autoren: B. Miemietz, C. Schmidt

1. Aufgabe: Speicher an die Solaris-Systeme anbinden

Wie gehe ich mit OSL Storage Cluster vor?

The Power of Simplicity

Voraussetzungen

- Rechner mit fertig installiertem Betriebssystem
- vorhandene Infrastruktur zur gemeinsamen Speicheranbindung (SAN, (i)SCSI ...)
- Speichersystem, gemappte LUNs
- für SAN: geeignetes Zoning

Ablauf Inbetriebnahme

- OSL SC-Packages installieren
- Clusternamen setzen
- CCF (Cluster Configuration Facility) konfigurieren
- Disks inventarisieren
- Diskvirtualisierung starten

Autoren: B. Miemietz, C. Schmidt

1. Aufgabe: Speicher an die Solaris-Systeme anbinden

"Foreign" und "native" Disks

The Power of Simplicity

- *neue LUNs sind "foreign" und "unused"*
- *Inventarisierung nimmt Disks in den globalen Pool auf*
- *einheitliche, hardwareabstrakte Sicht durch alle Cluster-Nodes*

Vor Erst-Inventarisierung:

Nach Erst-Inventarisierung:

Nach RAID-Erweiterung:

Nach 2. Inventarisierung:

Vorteile

- *leichte Identifizierung neuer Disks*
- *Solaris-Geräteadressen für Alltagsadministration uninteressant*
- *sprechende und zugleich einfache Gerätenamen*
- *Bildung eines globalen Pools*

Autoren: B. Miemietz, C. Schmidt

1. Aufgabe: Speicher an die Solaris-Systeme anbinden

Ablauf der Inventarisierung

The Power of Simplicity

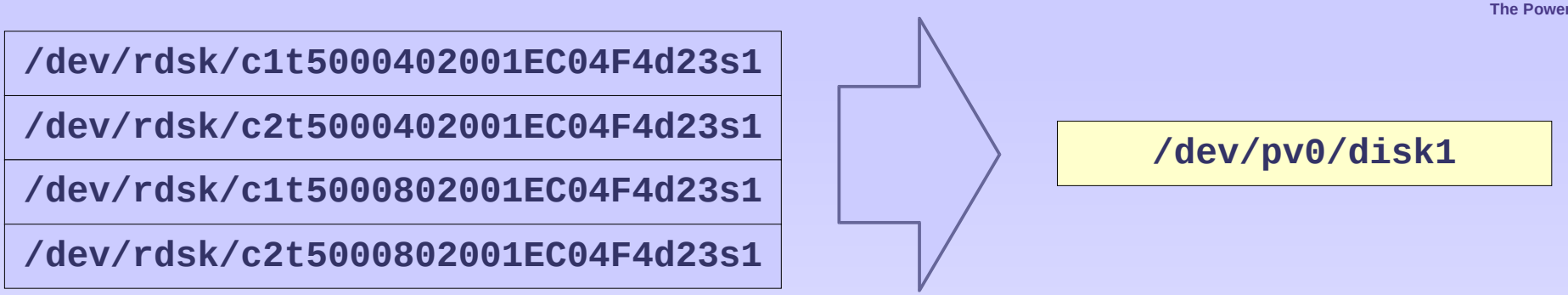
```
[root@erde] dkadmin -nie
Running with clustername:                sonne
Building device table:                   ok

Found foreign disk with following properties:
device path:             /dev/rdsk/c2t5000402001EC04F4d23s1
vendor / product:        NEXSAN / SATABl(C0A82C69)
serial number:           632D7033:"S;
capacity (MByte):        47683
disk format:             sunSPARC
dvsc product specific.: ????|                |                  |                |
former DVSC volume:      seems never used by DVSC with this device type
alternate disk info:     o.k.
current VTOC will be:    *** DESTROYED ***

new volume name or [RETURN] to skip device: disk1
new volume group or [RETURN] for "default":
```

dkadmin	***disk admin***
-n	***name***
-i	***interactively***
-e	***use EFI***

- *nicht inventarisierte (foreign) Disks werden automatisch angeboten*
- *passend zu Disks werden beim Laden der Virtualisierung (Systemstart)* ***Physical Volumes*** *erzeugt*
- *Physical Volumes sind auf allen Clusternodes identisch vorhanden*

Exkurs: Physical Volumes

Global wirksame Hardwareabstraktion



- *Gerätenamen werden vom Administrator gewählt -> es entfällt die Notwendigkeit, mit schwierigen Controller-Nummern oder SCSI-Adressen zu arbeiten*
- *SCSI-Adressen und Mappings des RAID-Systems spielen bei »native« Disks keine Rolle mehr und können (offline) geändert werden, ohne daß irgendwelche Konfigurationsänderungen in OSL SC erforderlich sind (wohl aber u. U. sd.conf etc).*
- *Slices (format) werden nicht mehr für die Aufteilung der Platten benutzt.*
- *Damit kann die Platte bei Neuaufteilungen online bleiben.*
- *Mehrere Datenpfade werden zu einem einzigen Geräteknoten zusammengefasst. Bei Ausfall eines Datenpfades ist weiter ein Zugriff auf die Platte möglich, sofern ein alternativer Pfad vorhanden ist (IO-Multipathing).*
- *Alle Platten werden optimal über die Kanäle verteilt (Load Balancing).*

1. Aufgabe: Speicher an die Solaris-Systeme anbinden

Diskvirtualisierung booten / aktivieren

The Power of Simplicity

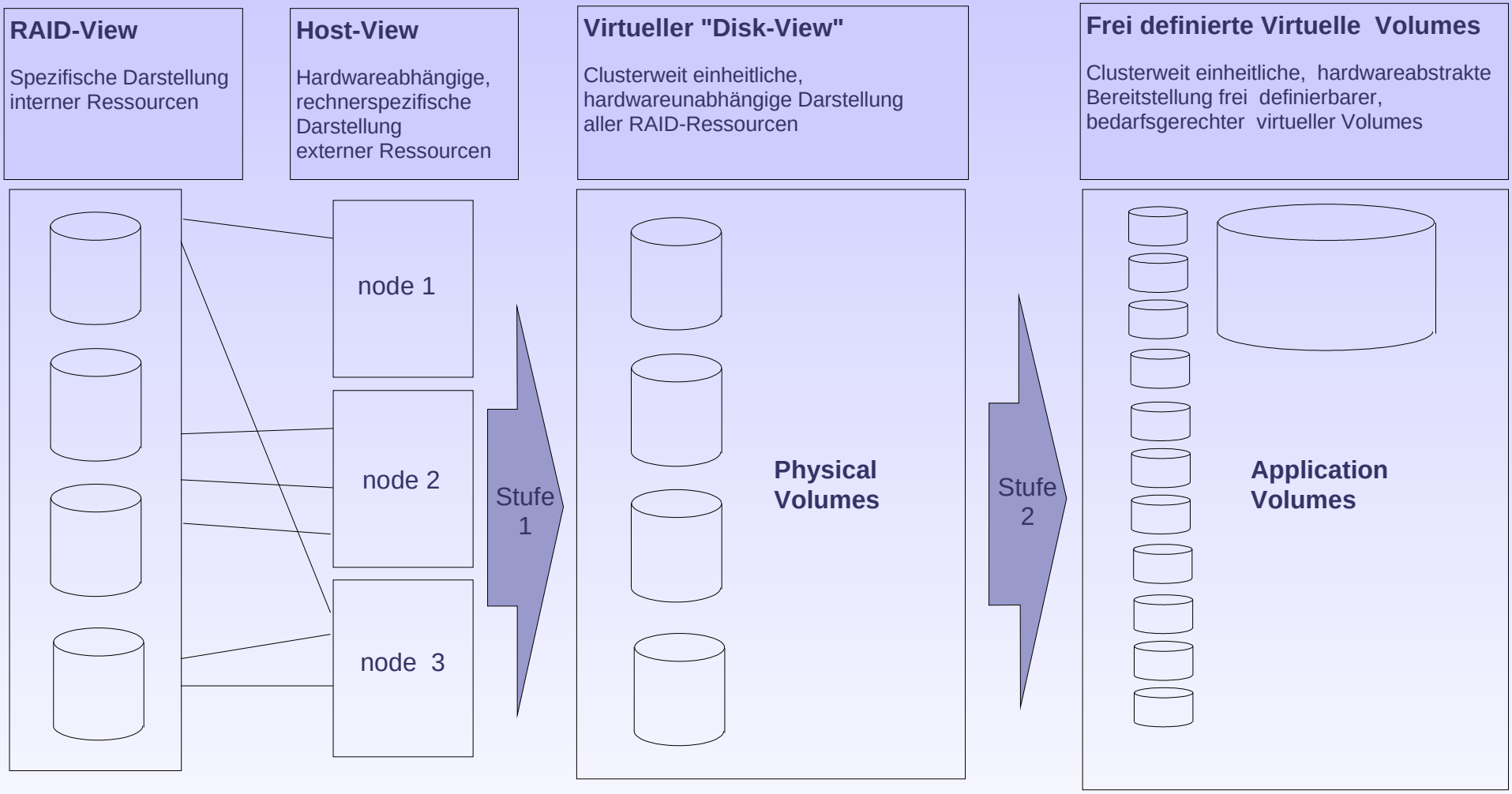
```
[root@erde] dvboot
dvboot
Booting with clustername:                sonne
*** DVBOOT PHASE 1 ***
Building device table:                   ok
Found following universes to boot:       0
PCCF already attached
my nodename is:      erde
my nodeid is:        1
my status is:        ONLINE
current clustersize: 1 of 1
*** DVBOOT PHASE 2 ***
*** DVBOOT PHASE 3 ***
[root@erde] smgr -q
0 pccf_erde                  1804288 blocks at                    2097152
0 disk1                     97638774 blocks at                          0
[root@erde] pvadmin -lvv disk1
0 disk1 (ok) 97638774 blocks over 2 path(s)
  >[ 1] (ok) /dev/rdsk/c2t5000402001EC04F4d23s1
   [ 2] (ok) /dev/rdsk/c1t5000402101EC04F4d23s1
[root@erde] ls -l /dev/pv0/disk1
lrwxrwxrwx   1 root     sys           31 Jun 24 15:09 /dev/pv0/disk1 -> ../../
devices/pseudo/vv@0:pv0_4
```


Autoren: B. Miemietz, C. Schmidt

Wie weiter?

Wie wir den globalen Pool nutzen

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

1. Aufgabe (Speicheranbindung) gelöst

Nach Inventarisierung steht ein globaler Speicherpool zur Verfügung

The Power of Simplicity

```
[root@erde] smgr -q summary
free on (0 dvsys         ):                875 MB            1 GB      0.001 TB
free on (0 default       ):            2098032 MB         2049 GB      2.001 TB
free on (1 default       ):            2193400 MB         2142 GB      2.092 TB
free on (2 default       ):            2004572 MB         1958 GB      1.912 TB
free on (3 default       ):            2860972 MB         2794 GB      2.728 TB

-------------------------------------------------------------------------------
TOTAL STORAGE POOL SUMMARY
free:          18755278423 bl          9157851 MB         8943 GB      8.734 TB
totl:          18757375575 bl          9158875 MB         8944 GB      8.735 TB
-------------------------------------------------------------------------------
```

- *jederzeit globale Übersicht über benutzte und freie Extents*
- *einfache Allokation von Storage für Filesysteme*
- *ist alles inventarisiert, brauche ich auf dem RAID-System oder im SAN nie mehr irgendetwas konfigurieren*

Autoren: B. Miemietz, C. Schmidt

2. Aufgabe: Filesysteme anlegen...

Wir bedienen uns einfach aus dem globalen Pool

The Power of Simplicity

```
[root@erde] smgr -c nfs1 -S 1g
[root@erde] newfs /dev/av0/rnfs1
newfs: construct a new file system /dev/av0/rnfs1: (y/n)? y
/dev/av0/rnfs1: 2097152 sectors in 256 cylinders of 64 tracks, 128 sectors
        1024.0MB in 26 cyl groups (10 c/g, 40.00MB/g, 19136 i/g)
super-block backups (for fsck -F ufs -o b=#) at:
 32, 82080, 164128, 246176, 328224, 410272, 492320, 574368, 656416, 738464,
 1312800, 1394848, 1476896, 1558944, 1640992, 1723040, 1805088, 1887136,
 1969184, 2051232
```

... gelöst

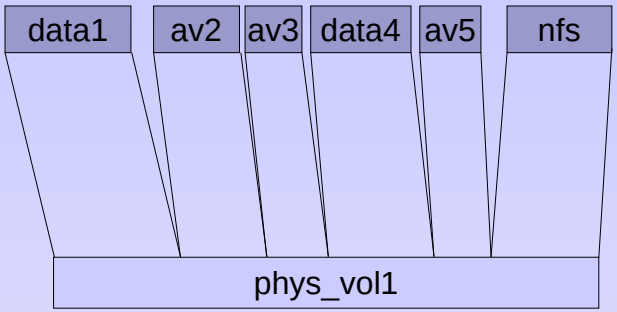
- *Virtuelle Volumes für Applikationsdaten ("**Application Volumes**") können einfach aus dem globalen Pool heraus erzeugt werden*
- *Application Volumes stehen nach dem Erzeugen clusterweit zur Verfügung*
- *Application Volumes stellen sich als normale Disk-Geräte dar, vgl. dkio(7I) und werden auch so genutzt: newfs ...*

Autoren: B. Miemietz, C. Schmidt

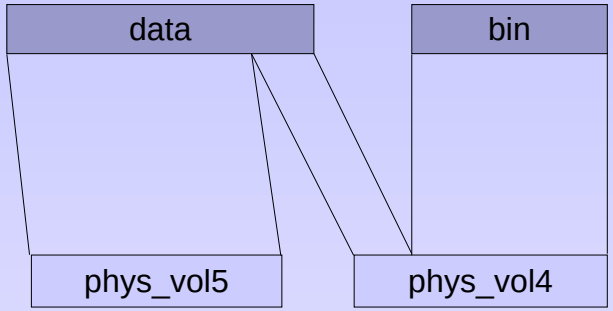
Exkurs: Application Volumes

Das Wichtigste im Schnelldurchlauf

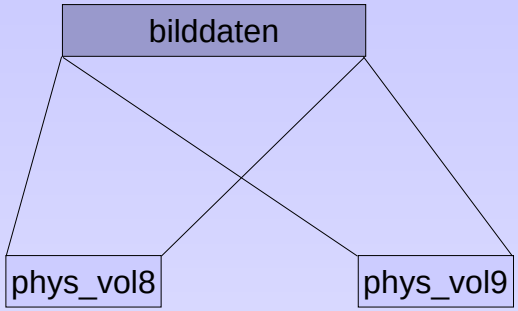
The Power of Simplicity



simple

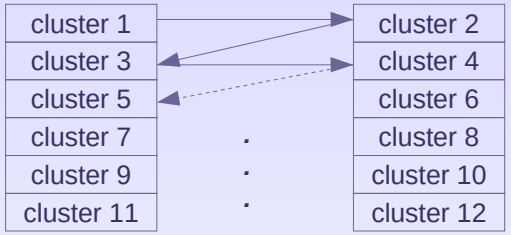


concat



stripe

- *es stehen unterschiedliche Typen zur Verfügung*
  - Bereitstellen benötigter Größen
  - mögliche Modifikation von Performance-Attributen
- *Erzeugung aus dem globalen Pool*
- *Application Volumes liegen direkt auf Physical Volumes (flache Hierarchie)*
- *frei wählbare Namen*



(Applikations-) Daten gehören per Definition auf Application Volumes!
Datenspeicherung direkt auf Physical Volumes oder Disks ist tabu.

Autoren: B. Miemietz, C. Schmidt

Was haben wir erreicht?

Vereinfachung in der Administration und mehr Flexibilität

The Power of Simplicity

RAID-View

Spezifische Darstellung interner Ressourcen

Host-View

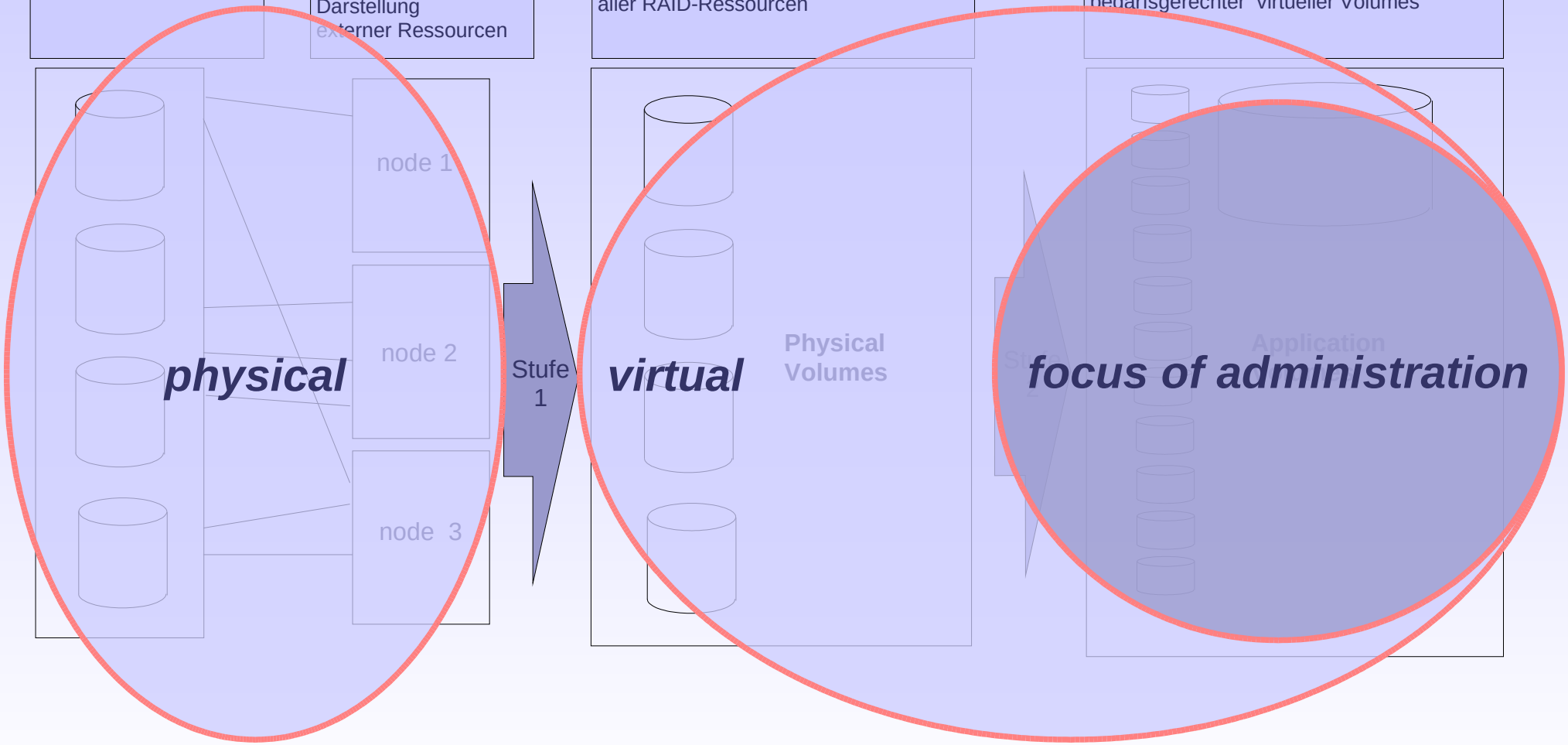
Hardwareabhängige, rechnerspezifische Darstellung externer Ressourcen

Virtueller "Disk-View"

Clusterweit einheitliche, hardwareunabhängige Darstellung aller RAID-Ressourcen

Frei definierte Virtuelle Volumes

Clusterweit einheitliche, hardwareabstrakte Bereitstellung frei definierbarer, bedarfsgerechter virtueller Volumes

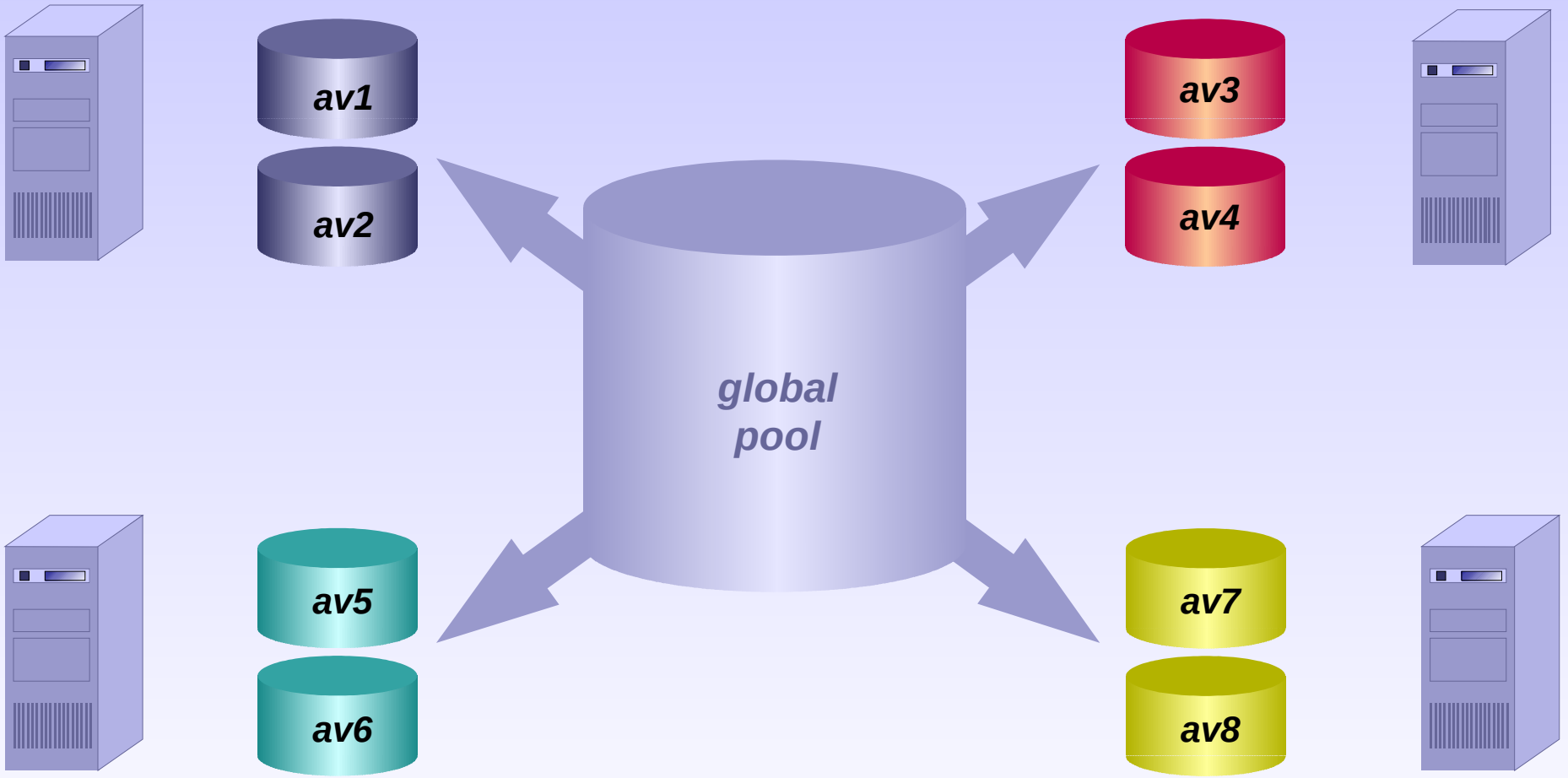


Autoren: B. Miemietz, C. Schmidt

Und was noch?

Verschnittfreie Nutzung der Ressourcen

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

Was heißt clusterweite Speichervirtualisierung?

Einfaches Management und einfacher Zugriff von jedem Knoten

The Power of Simplicity

- *vielleicht noch unbemerkt: Es ist ein Cluster entstanden...*

```
[root@erde] ndadmin -lvvv
nodename          id    state    os          cpu-isa          ncpu clock    memory

merkur             2   ONLINE   SunOS 5.10  amd64               2  3000      1964
venus              3   ONLINE   SunOS 5.10  amd64               2  3000      1964
erde               4   ONLINE   SunOS 5.10  amd64               2  3000      1964
```

- *und wir haben ganz stressfrei im gesamten Cluster Zugriff auf die Geräte*

```
[root@erde] mount /dev/av0/nfs1 /mnt
[root@erde] cd /mnt
[root@erde] echo hallo > hallo
[root@erde] cat hallo
hallo
[root@erde] cd /
[root@erde] umount /mnt
```

```
[root@venus] mount /dev/av0/nfs1 /mnt
[root@venus] cd /mnt
[root@venus] cat hallo
hallo
[root@venus] cd /
[root@venus] umount /mnt
```


Autoren: B. Miemietz, C. Schmidt

Was mir noch fehlt....

Es geht nicht nur um Speichervirtualisierung allein

The Power of Simplicity

- *ich habe:* - *virtuelle Geräte*
  - *clusterweit Zugriff auf diese Geräte*
  - *keinerlei Hardwarebezug in meiner Konfiguration*

- *aber:* - *am Ende muß ich den Anwendern Applikationen bereitstellen, flexibel, mit hoher Performance, gesicherten Daten und hoher Verfügbarkeit*

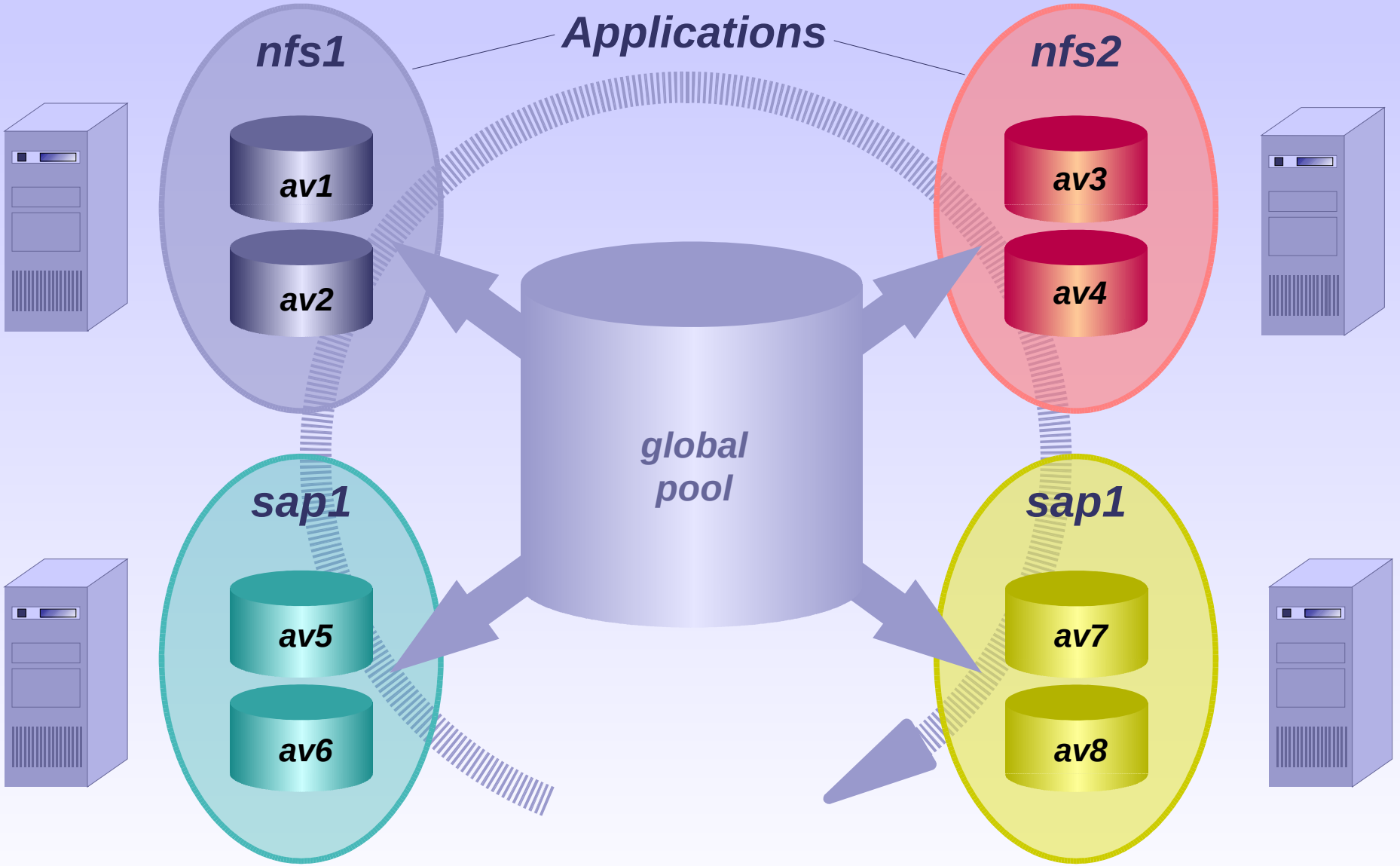
Was ich noch brauche, ...

Autoren: B. Miemietz, C. Schmidt

ist die Integration mit den Anwendungen

Speichervirtualisierung, Clustering und Applikationsmanagement gehören zusammen

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

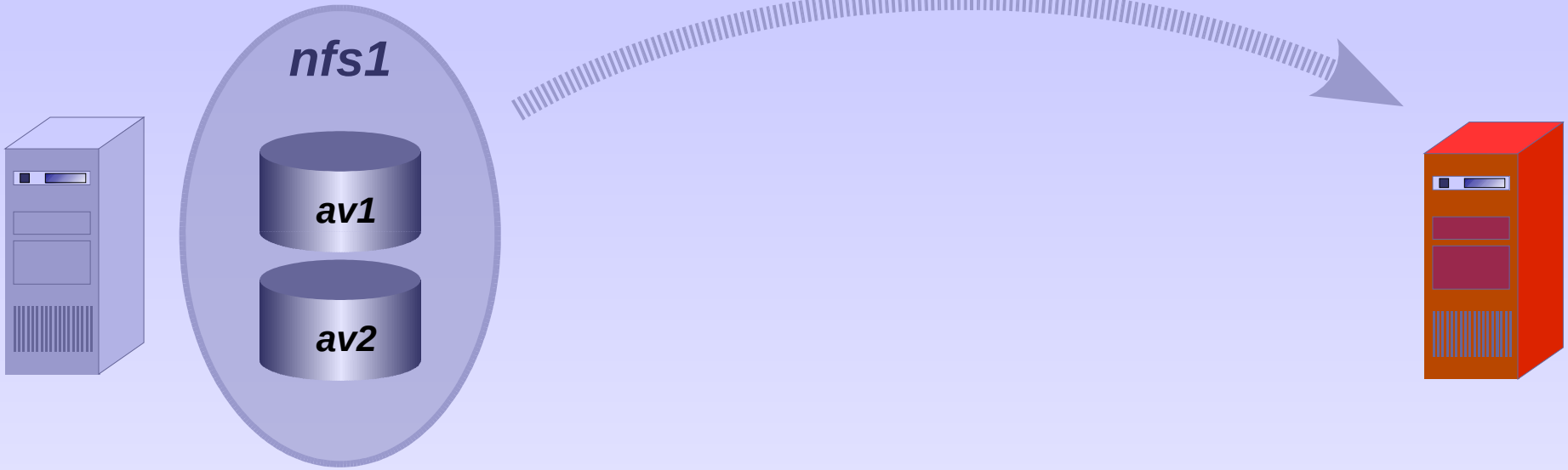
Anwendungen im Cluster organisieren + steuern

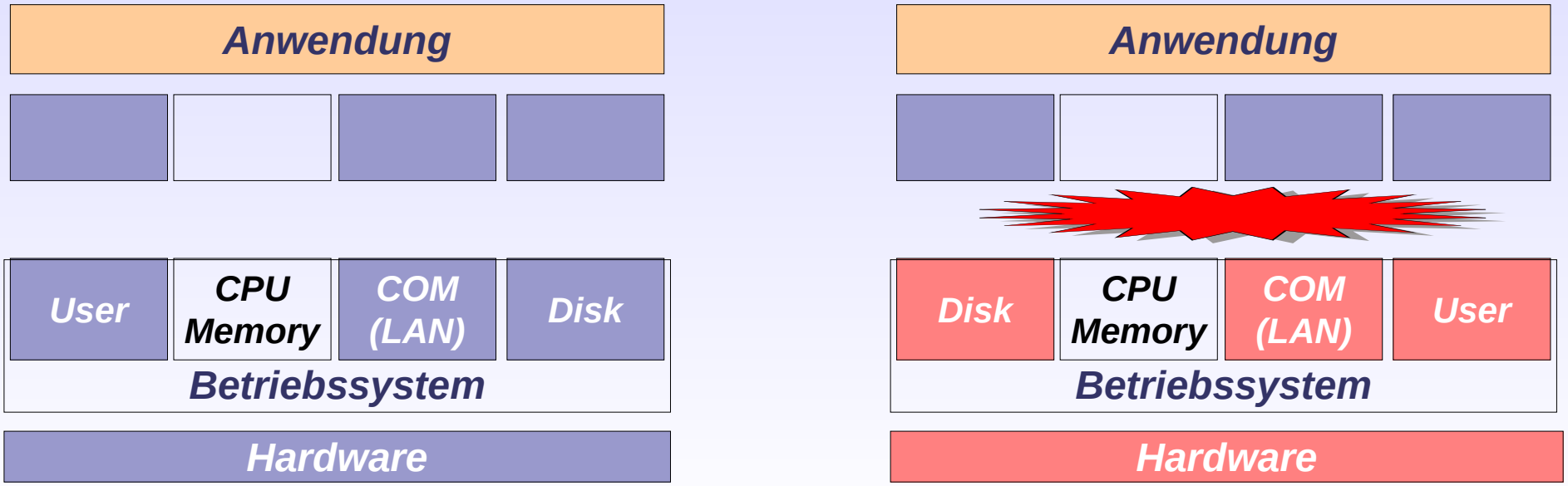
Autoren: B. Miemietz, C. Schmidt

Das Problem beim Verschieben von Anwendungen

Ohne Virtualisierung wenig Aussicht auf Erfolg

The Power of Simplicity

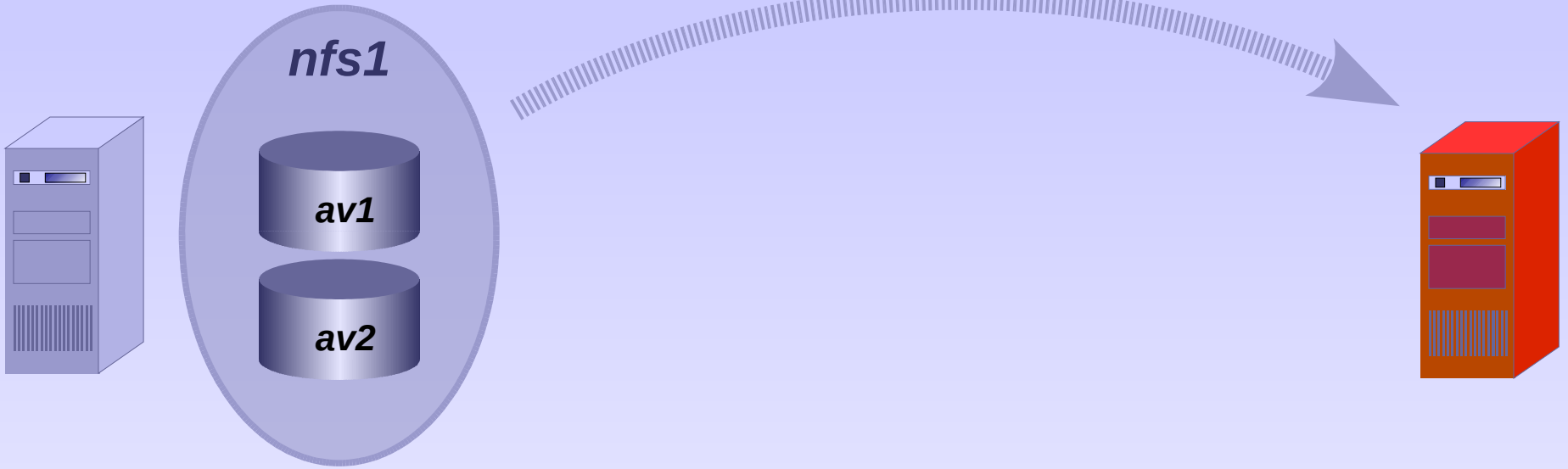


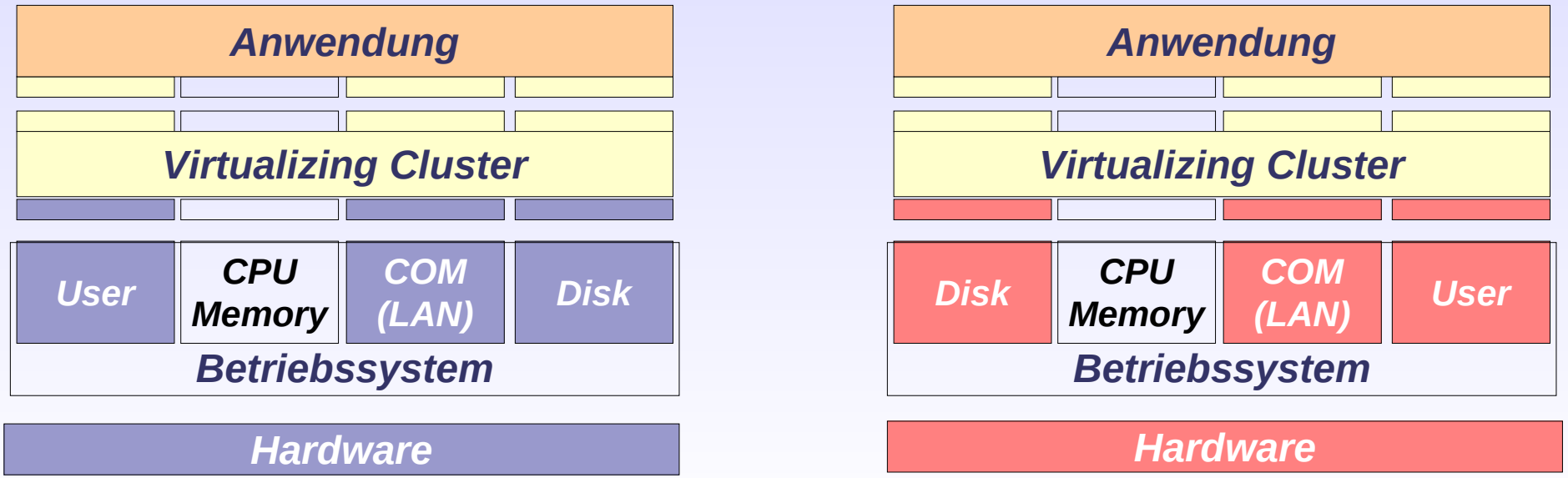


Kein Problem mit Virtualisierung

Hardwareabstraktion ist der Schlüssel zum Erfolg

The Power of Simplicity

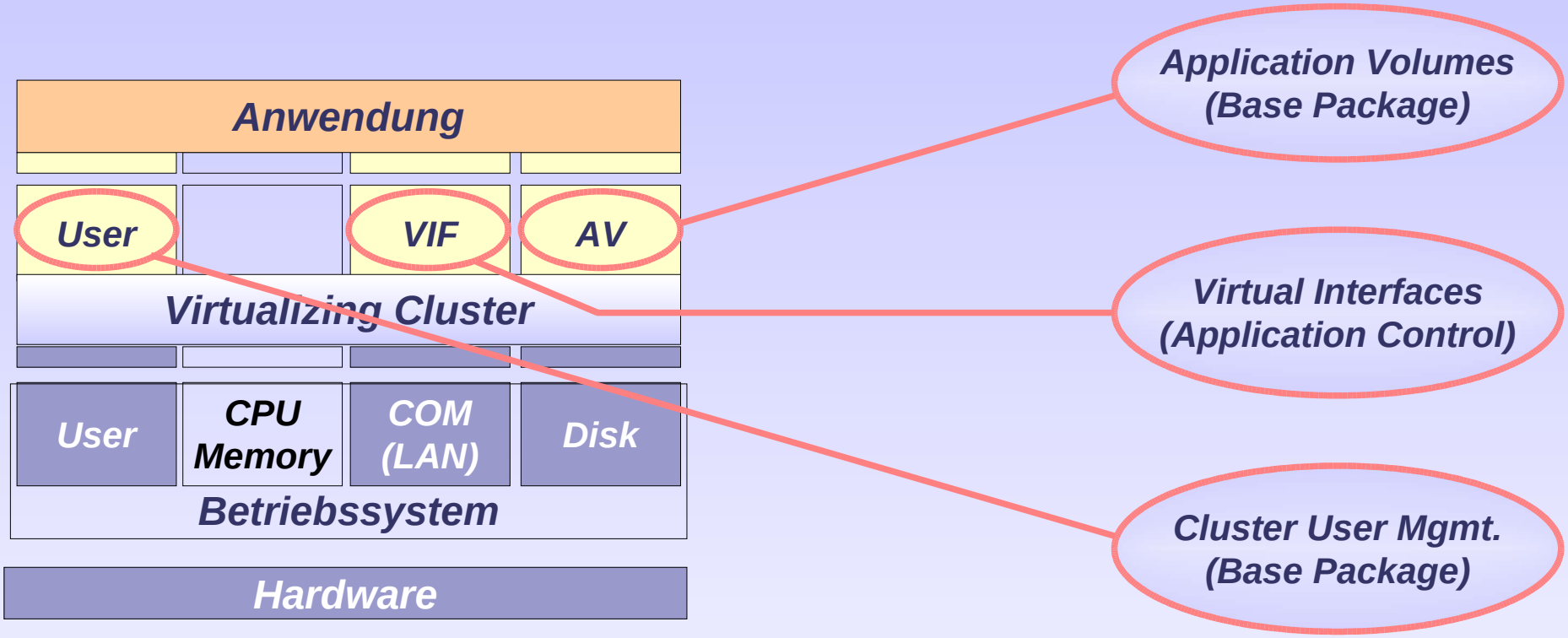




Virtualisierte Ablaufumgebungen für Applikationen

sind ganz ohne Virtuelle Maschinen oder Zonen möglich

The Power of Simplicity



Clusterfähige, virtualisierte Ablaufumgebungen bestehen aus:

- *anonymen und virtualisierten Ressourcen des Betriebssystems (RAM, CPU, VFS)*
- *aufsetzenden Virtualisierungs- und Clusterfunktionen (OSL Storage Cluster)*

Wem das zu einfach ist, dem bleiben zusätzlich Zonen und virtuelle Maschinen

Autoren: B. Miemietz, C. Schmidt

Bevor wir mit Applikationen loslegen...

OSL Storage Cluster -Applikationen im Überblick

The Power of Simplicity

Anwendungsdeklaration / -definition

o Priorität / Verdrängungsmöglichkeiten
o Migrationsstrategie und Execution Mode
o Ressourcensteuerung (IO-Bandbreite)
o Verknüpfung von Usern mit Applikationen

- *für jede Art von Applikationen*
- *applikationsabstrakt*
- *dienen der Steuerung durch die Cluster Engine*

```
# appadmin
```

Application Resource Description

clusterweit verfügbare Beschreibung von:
o genutzten Volumes, Filesystemen, IP-Adr. etc.
o Start- und Stopmethoden
o Methoden zum Abbruch einer Applikation
o Methoden zum Recover einer Applikation
o Methoden zum Monitoring / Auto-Restart

- *applikationsspezifisch*
- *einheitliches Schema*
- *freies Format*
- *dient der spezifischen Steuerung der einzelnen Applikation*

```
# ardadmin
```


Autoren: B. Miemietz, C. Schmidt

3. Aufgabe: Applikation anlegen

Anwendung anlegen und definieren

The Power of Simplicity

```
[root@erde] appadmin -c nfs1 -p 10 -d "1. NFS-Applikation"
NOTICE (appadmin): using local system platform "SunOS@amd64"
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control NOT_STARTED  -   -               -              2
```

Die neu angelegte (noch leere) Anwendung ist sofort im Cluster sicht-, startbar:

```
[root@erde] appstart nfs1
INFO (appstart): appstart for "nfs1" successful
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control STARTED      -   erde            [ONLINE]       1
[root@erde] appstop nfs1
INFO (appstop): appstop for "nfs1" successful
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control NOT_STARTED  -   -               -              2
```


Autoren: B. Miemietz, C. Schmidt

3. Aufgabe: Applikation anlegen

Die Anwendung mit Leben füllen: Application Resource Description

The Power of Simplicity

Nach Aufruf des Editiermodus findet sich die ARD unter /dvsc/ard/app_name.

Sie enthält folgende Verzeichnisse:

etc	*Eventual Technical Configuration* *Alle spezifischen Konfigurationsdaten zu Ihrer Anwendungen wie die dazugehörige vfstab, die Konfiguration virtueller IP-Adressen, der zugehörige Monitor oder anwendungsspezifische Konfigurationsparameter.*
start	*Application Start Scripts*
stop	*Application Stop Scripts*
break	*Application Break Scripts*

Applikationsbezogene Konfigurationsdaten gehören fortan nicht mehr in die Konfigurationsdateien des Betriebssystems, sondern in die der Applikation!

Die ARD hält dafür mindestens folgende Files bereit:

dfstab	*beim Start der Anwendung bereitzustellende NFS-Shares*
vfstab	*beim Start der Anwendung zu montierende Filesysteme auf Application Volumes bzw. NFS-Shares*
vif	*beim Start der Anwendung zu konfigurierende virtuelle IP-Adressen*

Autoren: B. Miemietz, C. Schmidt

3. Aufgabe: Applikation anlegen

Die Anwendung mit Leben füllen: Application Resource Description

The Power of Simplicity

Für NFS benötigen wir:

Serveradresse	*./etc/vif*
Dateisystem	*./etc/vfstab*
Shares	*./etc/dfstab*

Wir erstellen also Einträge in den entsprechenden Konfigurationsdateien der ARD:

```
[root@erde] ardadmin -e nfs1
NOTICE (ardadmin): application "nfs1" is now in edit mode.
[root@erde] cd /dvsc/ard/nfs1
[root@erde] ls
break  etc    start  stop
[root@erde] cd etc
[root@erde] ls
add_routes  del_routes  dfstab      vfstab      vif
[root@erde] tail -1 vif
VIF1=nfs1:192.168.44.201:0xffffff80
[root@erde] tail -1 vfstab
/dev/av0/nfs1  /dev/av0/rnfs1  /nfs1   ufs 6 yes rw
[root@erde] tail -1 dfstab
share -F nfs -d "1. NFS-Server" /nfs1
[root@erde] cd /
[root@erde] ardadmin -c nfs1
NOTICE (ardadmin): Changes to ARD of "nfs1" committed. Edit mode quit.
```


Autoren: B. Miemietz, C. Schmidt

3. Aufgabe (Applikation anlegen) gelöst

Wir können die Anwendung auf beliebigen Maschinen starten

The Power of Simplicity

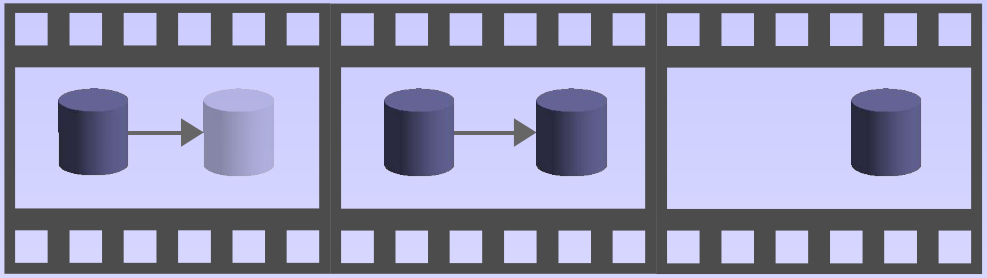
```
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control NOT_STARTED  -   -                -             3
[root@erde] appstart nfs1
INFO (appstart): appstart for "nfs1" successful
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control STARTED      -   erde             [ONLINE]      2
[root@erde] df -k
Filesystem            kbytes    used   avail capacity  Mounted on
/dev/dsk/c0d0s0      50431906 3825665 46101922     8%    /
...
/dev/av0/nfs1          985951    1042  925752     1%    /nfs1
[root@erde] vifadmin -l
The following virtual interfaces are configured:
e1000g0:1    192.168.44.201  netmask 0xffffff80        broadcast 192.168.44.255
[root@erde] share
-               /nfs1   rw   "1. NFS-Server"
[root@erde] appstop nfs1
INFO (appstop): appstop for "nfs1" successful
[root@erde] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)         node_state ready
  10 nfs1       e  no_control NOT_STARTED  -   -                -             2
```


Autoren: B. Miemietz, C. Schmidt

Wie war das mit Flexibilität im Storage-Management?

kurzer Exkurs: Online-Migration von Daten zwischen RAID-Systemen

The Power of Simplicity



online Daten verschieben / reorganisieren
automatische Priorisierung Anwendungs-IO

- *Zielkonfiguration der Migration als Shadow-Volume konfigurieren*
- *Move anstarten*
- *Fertig*

```
[root@erde] avadmin -lvv nfs1
0 nfs1 2097152 of 2097216 blocks "simple" in 1 pieces, 32 block clusters
  [ 1] old1 [0...2097215]
[root@erde] smgr -c ziel -S 1g new1
[root@erde] avmove nfs1 ziel
[root@erde] avadmin -lvv nfs1
0 nfs1 2097152 of 2097216 blocks "simple" in 1 pieces, 32 block clusters
  [ 1] new1 [0...2097215]
```


Autoren: B. Miemietz, C. Schmidt

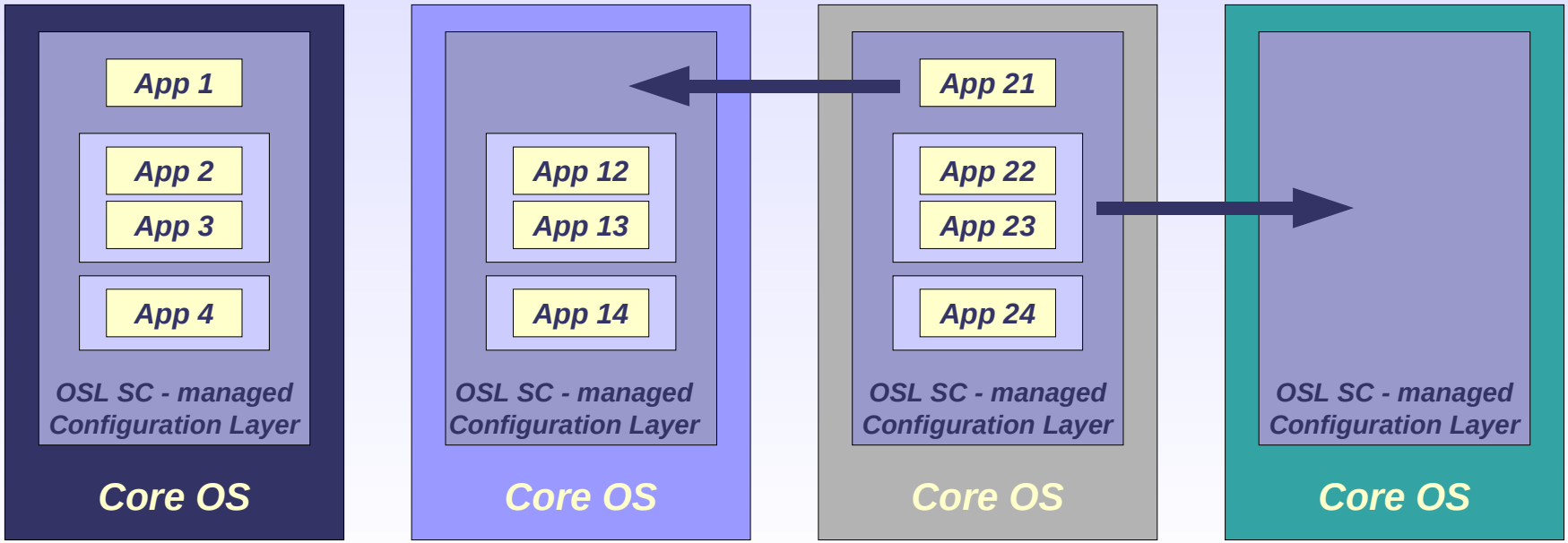
Was haben wir gewonnen?

Organisation in Applikationen ermöglicht Virtualisierte Ablaufumgebungen

The Power of Simplicity

- Global Devices und Application Resource Description
  - Raw- und Blockdevices + Dateisysteme
  - ZFS
  - IP-Adressen und NFS
- Globales Nutzer- und Gruppenmanagement
- Automatische Adaption ASCII-Konfigurationsdateien
- Globales Management und Migrationsdienste für Zonen

Umgebung lebt außerhalb der Maschine weiter



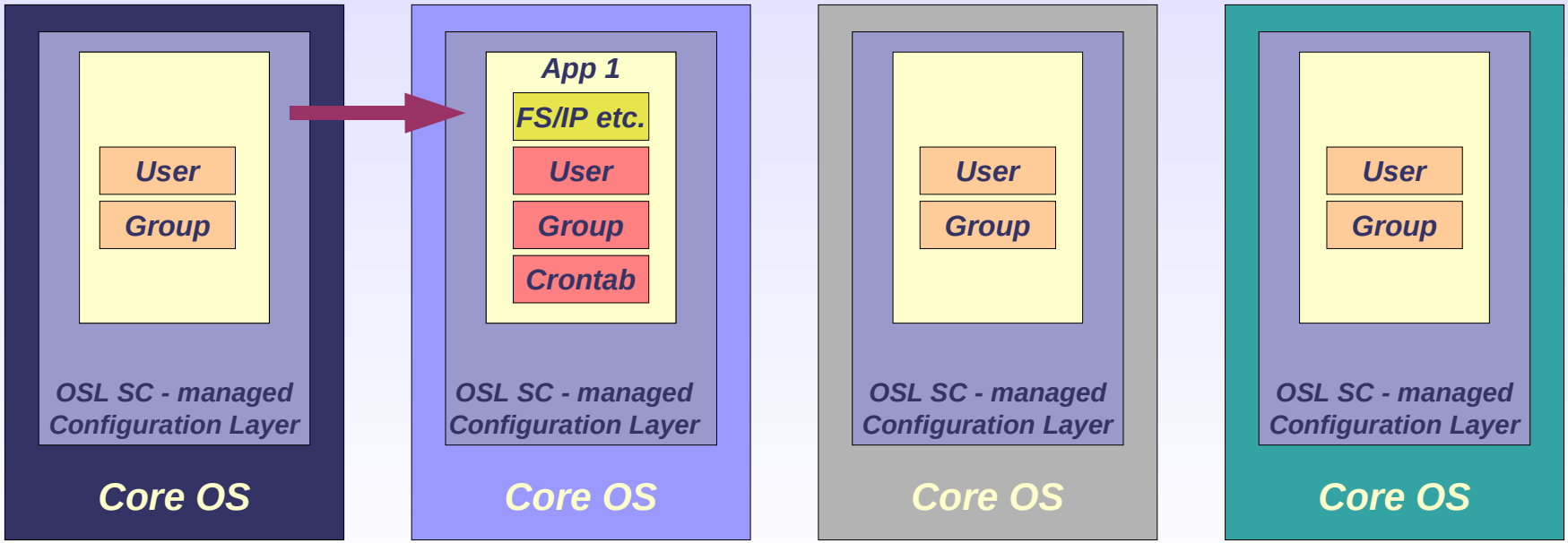
Autoren: B. Miemietz, C. Schmidt

Bisher noch unbeachtet: Global User Management

Globale Nutzerverwaltung komplettiert virtualisierte Ablaufumgebungen

The Power of Simplicity

- geeignet für Server / Application Service User
- Unabhängig von externen Services wie NIS/LDAP/ADS
- Vermeidung von Konflikten, Synchronisation, automatische Reparatur
- User kann einer Applikationen zugeordnet werden
- Crontab und Login-Möglichkeit wandern mit der Applikation
- auch nach Neuinstallation sofort wieder verfügbar

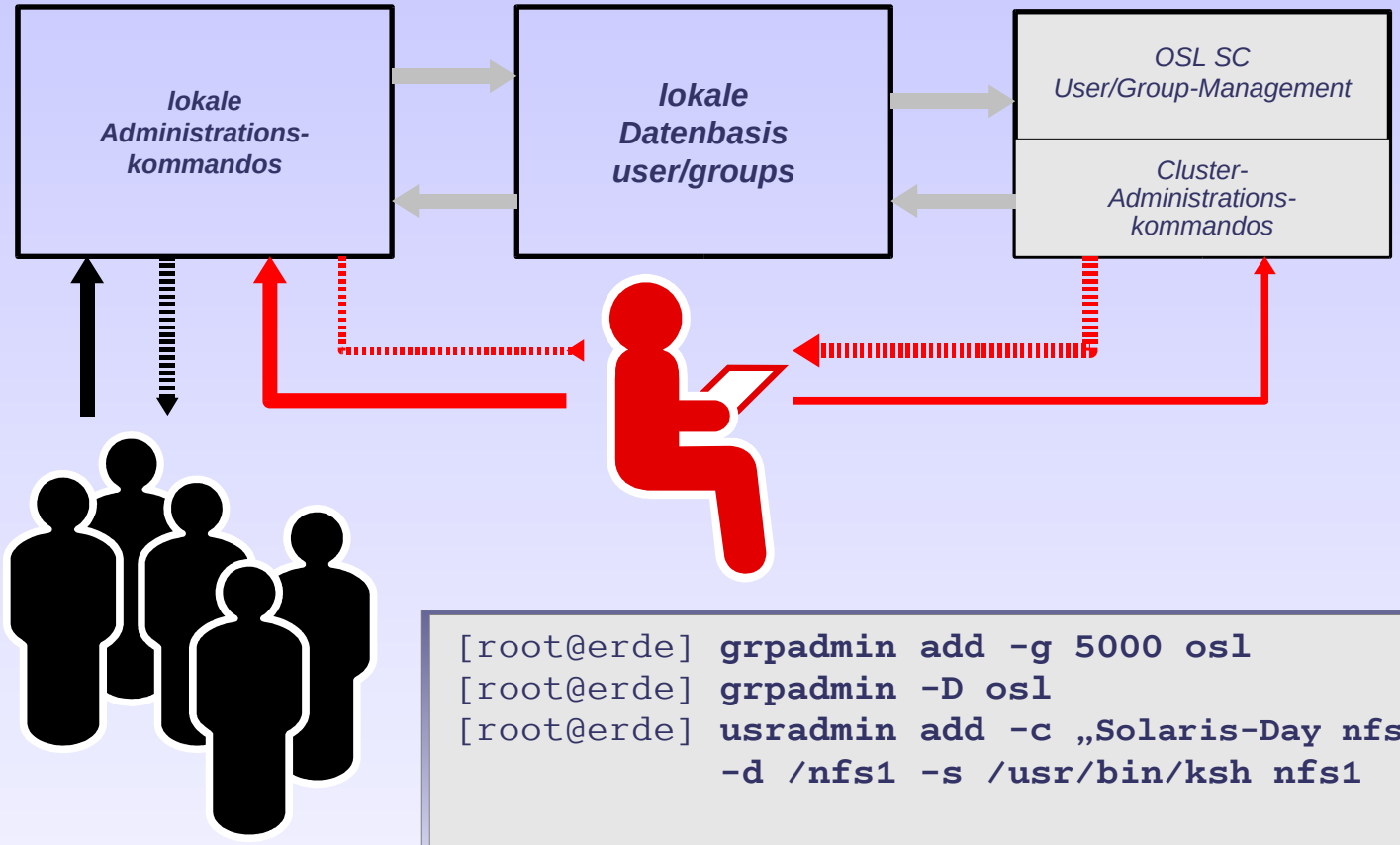


Autoren: B. Miemietz, C. Schmidt

Global User Management – so funktioniert es

Wir legen beispielhaft eine Gruppe und einen Nutzer global an

The Power of Simplicity




```
[root@erde] grpadmin add -g 5000 osl
[root@erde] grpadmin -D osl
[root@erde] usradmin add -c „Solaris-Day nfs1“ -u 5000 -g 5000\
            -d /nfs1 -s /usr/bin/ksh nfs1
```


Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Hochverfügbarkeit

Autoren: B. Miemietz, C. Schmidt

Nächster Schritt: Hochverfügbarkeit

Von manuellen zu automatischen Abläufen

The Power of Simplicity

*Anstelle der expliziten manuellen Steuerung können wir die Steuerung auch dem Cluster überlassen. Dazu gibt es drei Direktiven (= **target state**) für die Behandlung einer Applikation durch die Cluster Engine:*

no control	*keine Steuerung durch den Cluster, wohl aber Überwachung*
up	*Cluster versucht, Anwendung am Laufen zu halten*
down	*Cluster beendet Anwendung, falls nötig.*

Damit ist zugleich die Hochverfügbarkeit implementiert.

Was eine "richtige" Applikation noch braucht

Applikationsbezogene Start- und Stopp-Prozeduren

The Power of Simplicity

Über applikationsbezogene "User"-Start- und Stopp-Prozeduren können wir weitere Aktionen einbinden.

- *Die Scripts befinden sich in den ARD im Unterverzeichnis »start«, »stop«, »break«,*
- *Scripts werden beim Applikationsstart in folgender Reihenfolge ausgeführt:*

  1. Built-in

  *2. **Benutzerdefinierte Scripts** (S01-S99) **ähnlich** den **RC-Scripts** (/etc/rc2.d/S...)*

  3. Built-in

  Stop- und Break-Prozeduren analog mit K01-K99
- *Die Scripts werden mit 2 Argumenten aufgerufen:*

  $1 = "start" oder "stop"

  $2 = nickname
- *Die Scripts müssen einen definierten Return-Code liefern:*

  0 bei fehlerfreier Beendigung

  >0 bei aufgetretenen Fehlern, die weiteren App.-Start unmöglich machen
- *Gravierende Fehler (RC != 0) führen beim Start sofort zur Ausführung der Stop-Prozeduren.*

Autoren: B. Miemietz, C. Schmidt

Was man noch wissen könnte

Prioritäten und Knotenlisten

The Power of Simplicity

prio 1	node 1	node 2	node 3	node 4	node 5	node 6	node 7	node 8
prio 2	node 4	node 3	node 1	node 6				
prio 3								
prio 4								
prio 5								
prio 6								
prio 7								
prio 8								
prio 9								
prio 10								
prio 11								
prio 12								
prio 13								
prio 14								
prio 15								
prio 16								

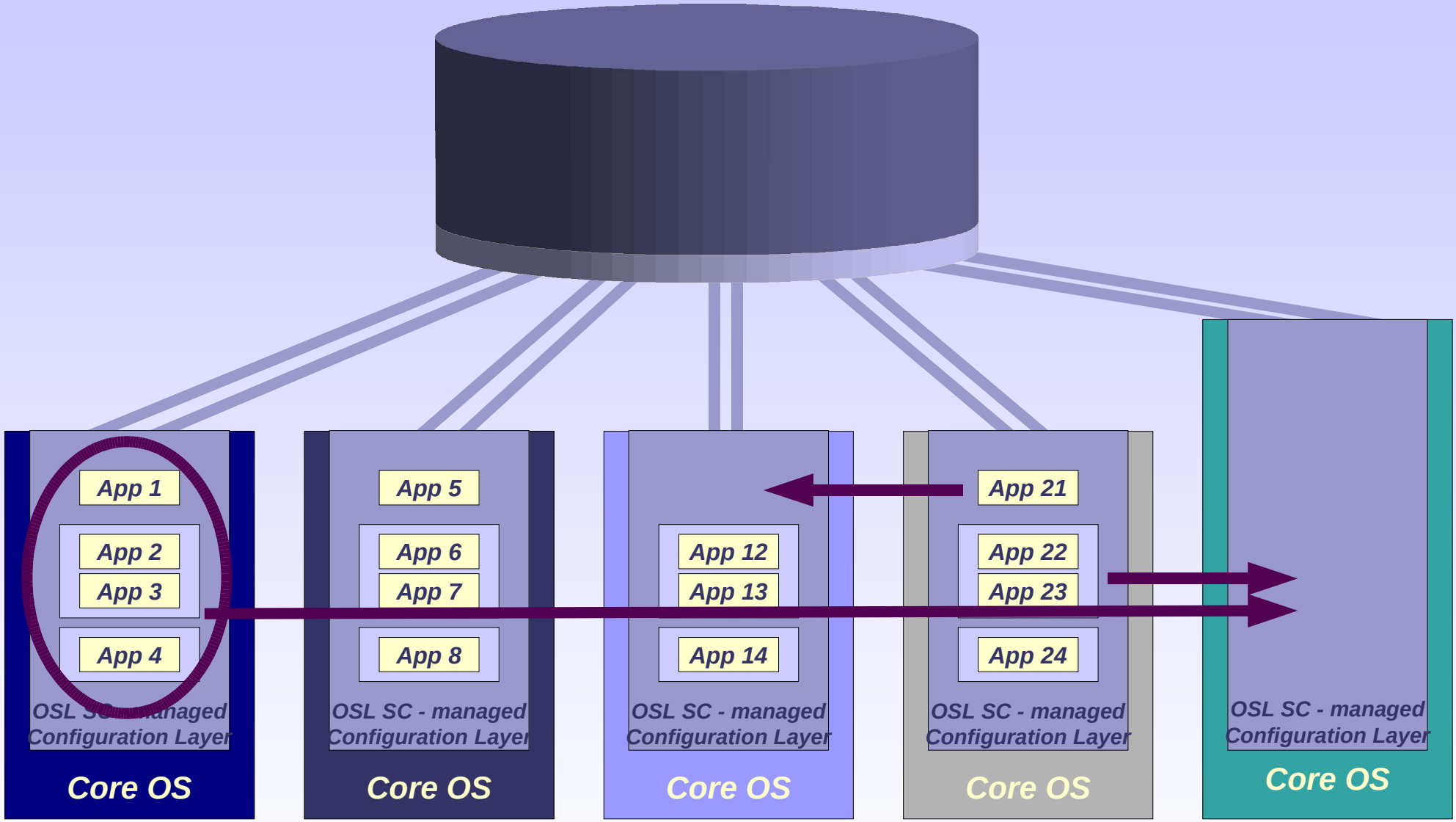
- *Anwendungen sind eineindeutig clusterweite Prioritäten zugeordnet*
  - *welche Anwendung wird zuerst gestartet?*
  - *Anwendungen höherer Priorität können bei Bedarf solche mit niederer Priorität verdrängen*
- *die Position eines Knotens in der Knotenliste einer Applikation bestimmt die Affinität der Applikation zum jeweiligen Knoten*
- *Es sind exklusive (default) und parallele Ausführungsmodi möglich*
- *Jeder Knoten kann mehreren Applikationen zugeordnet sein*
- *Bei der Auswahl des Zielknotens können Performanceaspekte berücksichtigt werden*
- *Es sind dynamische Knotengruppen möglich*

Autoren: B. Miemietz, C. Schmidt

Was sind dynamische Knotengruppen?

Hardware tauschen und Cluster verändern ohne Konfigurationen anzupassen

The Power of Simplicity

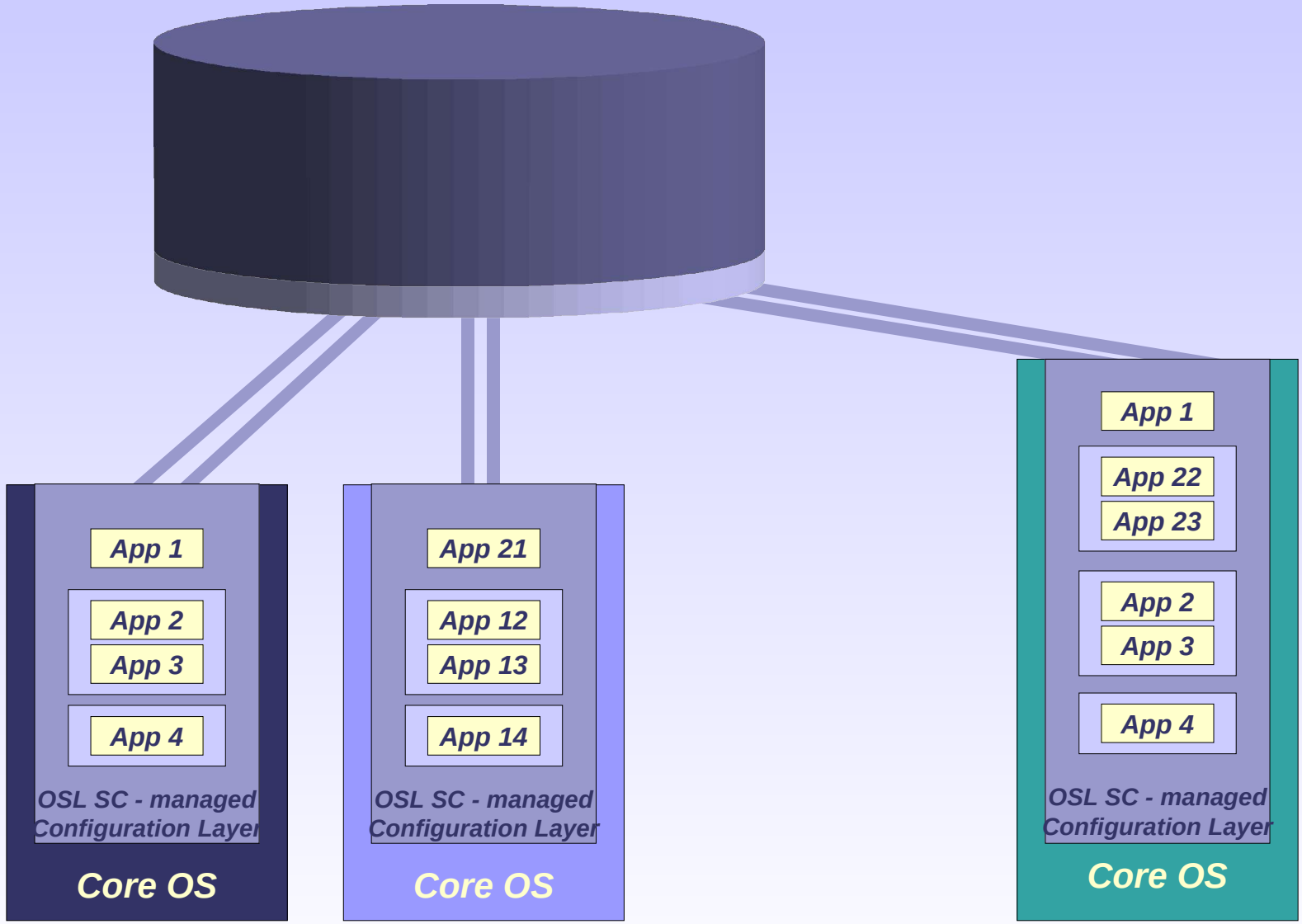


Autoren: B. Miemietz, C. Schmidt

Was sind dynamische Knotengruppen?

Hardware tauschen und Cluster verändern ohne Konfigurationen anzupassen

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Beispiel:
Hochverfügbarer Storage-Server mit Solaris + OSL Storage Cluster

Autoren: B. Miemietz, C. Schmidt

Anforderungen an einen Storage-Server

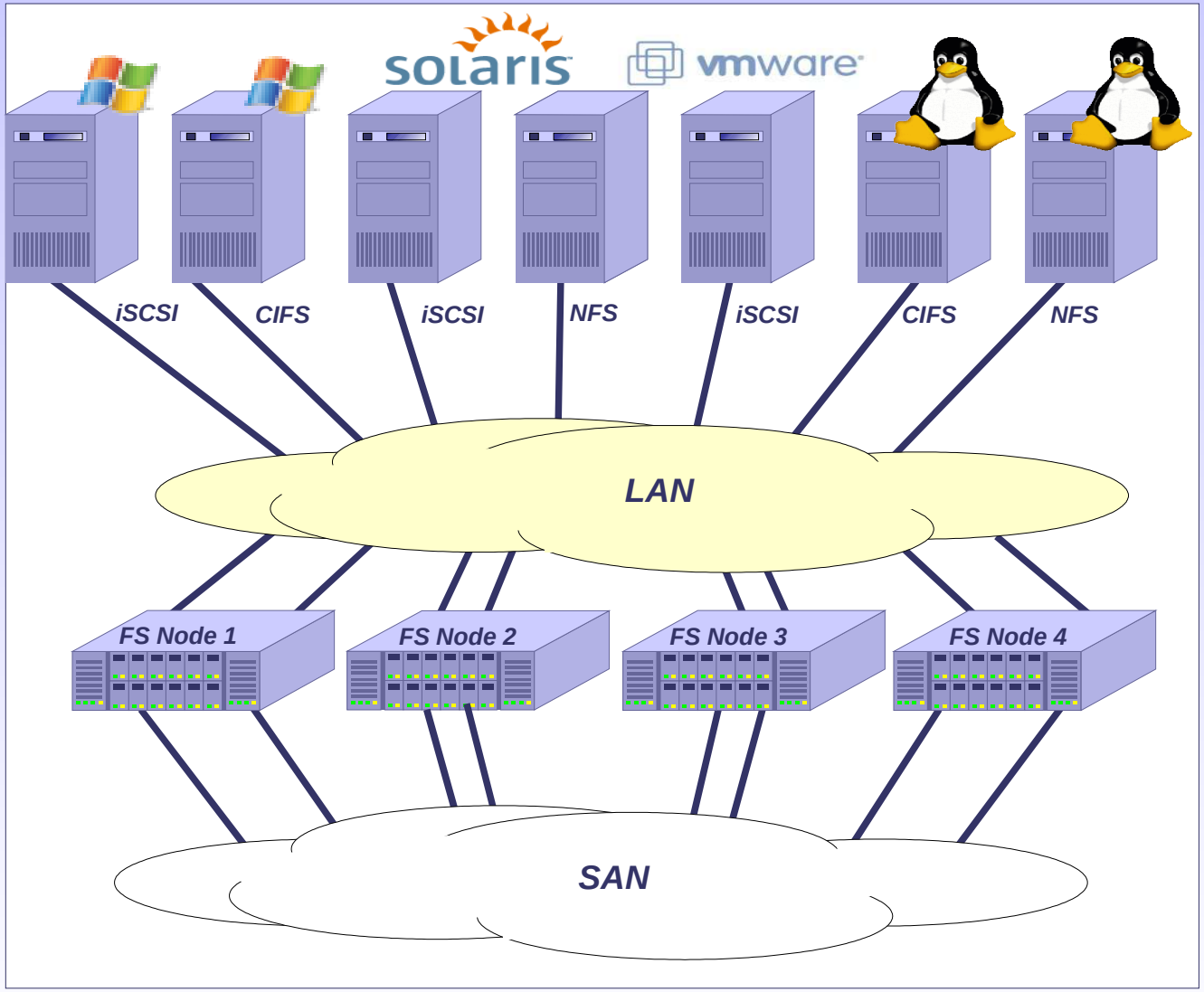
Welche Eigenschaften muss ein Storage-Server haben?

The Power of Simplicity

- ✔ *Der Client muss das Protokoll verstehen*
- ✔ *Einfaches Backup und Recovery*
- ✔ *Hohe Verfügbarkeit*
- ✔ *Einfach erweiterbar*
- ✔ *Leicht administrierbar*
- ✔ *Sicherheit auf Transport- und Zugriffsebene (Authentifzierung, Authorisierung)*

Autoren: B. Miemietz, C. Schmidt

Anforderungen an einen Storage-Server



Autoren: B. Miemietz, C. Schmidt

Anforderungen an einen Storage-Server
Mögliche Protokolle

The Power of Simplicity

SMB/CIFS

- ✔ Leichte Anbindung von Windows Clients
- ✔ Authentifizierung über die Windows Domain
- ✔ Server ist in Solaris 10 integriert
- ✗ zustandsbehaftetes Protokoll
- ✗ keine Solaris Client Software

NFS

- ✔ Leichte Anbindung von unixoiden Clients (Server und Client SW vorhanden)
- ✔ einfaches Erstellen von Freigaben
- ✔ zustandsloses Protokoll
- ✗ Schlechte Unterstützung von Windows (nur über SFU)

iSCSI

- ✔ kein Fileshare sondern ein Device Share – einfache Sicherheitsmechanismen
- ✔ zustandsloses Protokoll
- ✔ Initiator Software für alle gängigen Betriebssysteme

Autoren: B. Miemietz, C. Schmidt

Anforderungen an einen Storage-Server
Anforderungen an die Protokolle

The Power of Simplicity

	NFS (v3)	SMB/CIFS	iSCSI
Clientunterstützung	+	+	++
Serverunterstützung	++	++	++
Verhalten beim Failover	++	-	++
Implementierungsaufwand	++	o	+

Alle beschriebenen Fileserver Protokolle können als kombinierte oder eigenständige Applikation im Cluster abgebildet werden.

Somit sind für einen Fileserver alle Vorteile des OSL Storage Clusters nutzbar:

- *hochverfügbare Applikationen*
- *Volumes spiegeln, clonen, moven*
- *integriertes clusterweites Benutzermanagment*
- *integrierte Bandbreitensteuerung*
- *Übersicht des applikationsbezogenen Ressourcenverbrauchs*
- *Adaptive Applikation (Ressourcenbasiertes Selbstmanagement)*

Autoren: B. Miemietz, C. Schmidt

Storage-Server mit OSL Storage Cluster

Applikationstemplates erleichtern die Arbeit / das Rollout von Applikationen

The Power of Simplicity

- *Vorgefertigte Templates für verschiedene Applikationen (Datenbanken, Fileserver, Buisnessapplikationen)*
- *Einzige Aufgabe des Administrators: Konfiguration und Installation*
- *Getestete Start- Stop- und Breakprozeduren*
- *Integration von Backup- und Recovery*
  - *Backup-to-Disk / Backup-to-Tape mit Pre- und Postprocessing*
  - *Sicherung von Logfiles*
  - *Recoveryunterstützung (Backup finden, Rollforward der Datenbanken)*
- *Unterstützung beim Anlegen von Clonen*
  - *z.B. SAP Systemkopie*
  - *Clonen von Solaris Zonen im Cluster*
  - *Nachbereitung mit Modifikation der Umgebung*

Autoren: B. Miemietz, C. Schmidt

Storage-Server mit OSL Storage Cluster

Applikationstemplates erleichtern die Arbeit / das Rollout von Applikationen

The Power of Simplicity

… Zurück zum Fileserver

→ *NFS Shares können zu jeder Applikation hinzugefügt werden*

→ *was ist mit iSCSI und Samba?*

```
#> appadmin -lT
...
105 - Oracle 10 (single instance)
143 - Informix + ISM
405 - ISCSI-Target
410 - Samba fileserver
505 - EAS R3 4.6 Oracle 9 (single instance)
```

Templates erleichtern die Arbeit!

→ *Volumes anlegen*

→ *Konfigurieren*

→ *Starten*

Storage-Server mit OSL Storage Cluster

SAMBA-Template

The Power of Simplicity

Im OSL Storage Cluster gibt es eine Vielzahl von Templates. Sie erleichtern das Anlegen von hochverfügbaren Applikationen.

Beispiel:

Samba Server

- *benötigt clusterweit die gleichen Dienste*
- *benötigt clusterweit dieselben Volumes*
- *benötigt clusterweit dieselben User*
- *benötigt clusterweit dieselbe Konfiguration*

Anlegen eines Samba Fileservers Clusters vom Template

```
#> appadmin -c smbserver -p 201 -T 410
NOTICE (appadmin): using local system platform "SunOS@amd64"
```

Nächste Schritte:

→ Volumes anlegen / Konfigurieren / Starten

Autoren: B. Miemietz, C. Schmidt

Ist eine Applikation einmal erstellt worden, ist diese clusterweit bekannt und kann im ganzen Cluster genutzt werden.

Globaler Storage Pool:

- *Alle Clusterknoten sehen die selben Application Volumes*

Globale Userverwaltung:

- *Benutzer und Gruppen sind allen Knoten bekannt*
- *keine Probleme mit ACLs, Zugriffsrechten etc.*
- *selbst Cronjobs können an Applikationsuser gebunden werden und laufen nur auf dem Knoten mit der Applikation*

Knotenübergreifende Konfiguration

- *intelligente Start- und Stoppprozeduren bringen die Konfiguration auf jeden Server*
- *lokal laufende Applikationen werden nicht beeinflusst (Encapsulated Application Setup)*

Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Gleiches Vorgehen für ein iSCSI Target:

Anlegen

```
#> appadmin -c iscsi -p 56 -T 405
NOTICE (appadmin): using local system platform "SunOS@amd64"

#> smgr -c iscsitcfg -S 100m && newfs /dev/av0/riscsitcfg
...
#> smgr -c target -S 100g
```

- *Es wird eine neue Applikation angelegt*
- *Für diese Applikation gibt es vorgefertigte Start- und Stopproutinen*
- *Konfigurationsfiles müssen angepasst werden*

Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Anpassen der Applikation, Konfigurieren des virtuellen Interfaces, Anlegen von Volumes und Filesysteme

```
#>
#> ardadmin -e iscsi
NOTICE (ardadmin): application "iscsi" is now in edit mode.
#> cd /dvsc/ard/iscsi
#> ls
break  etc    start  stop
...
#> vi etc/vif
...
#> smgr -c iscsi_cfg -S 100m
#> newfs /dev/av0/riscsi_cfg
...
#> smgr -c target -S 100g
#> vi etc/vfstab
```


Autoren: B. Miemietz, C. Schmidt

Start der Applikation

```
#> appstart iscsi
```

Start beobachten

```
#> applogcat start iscsi
```

Stopp der Applikation

```
#> appstop iscsi
```

Stopp beobachten

```
#> applogcat stop iscsi
```


Autoren: B. Miemietz, C. Schmidt

Storage-Server mit OSL Storage Cluster

Übersichten

The Power of Simplicity

Im Storage-Cluster kann jede Applikation auf mehreren Knoten laufen und einen gemeinsamen Storage Pool benutzen.

Wie sehe ich, wie die Applikationen das SAN belegen?

```
[root@venus] smgr -qa
used by H05          :                  105472 MB          103 GB      0.101 TB
used by nfs1         :                       0 MB            0 GB      0.000 TB
used by C12          :                  614400 MB          600 GB      0.586 TB
used by X21          :                  333824 MB          326 GB      0.318 TB
-------------------------------------------------------------------------------
APPLICATION RELATED STORAGE POOL USAGE
used:           2157970592 bl          1053697 MB         1029 GB      1.005 TB
-------------------------------------------------------------------------------
TOTAL STORAGE POOL SUMMARY
free:          16009290592 bl          7817036 MB         7634 GB      7.455 TB
totl:          18171455552 bl          8872781 MB         8665 GB      8.462 TB
-------------------------------------------------------------------------------
```

Wo läut welche Applikation?

```
[root@venus] appadmin -qo
prio nickname   em tgt_state  availability mon node(s)          node_state ready
  10 nfs1       e  no_control NOT_STARTED  -   -                -              3
  51 HO5        e  no_control STARTED      -   venus            [ONLINE]       2
```


Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Was soll passieren beim Ausfall?

- Die Applikation soll schnell wieder laufen und es sollen keine Daten verloren gehen

Wie erreicht man dies?

- Die Applikation liegt im SAN
- Die Applikationsbeschreibung ist global verfügbar
- Es sind ausreichend freie Ressourcen vorhanden
- Ein anderer Knoten erkennt den Ausfall
- Der defekte Knoten ist mit Sicherheit beendet

Autoren: B. Miemietz, C. Schmidt

Gestaltung der Hochverfügbarkeit

Automatische Steuerung – Zustandsdirektiven (Target State)

The Power of Simplicity

Über den "Target State" einer Applikation kann festgelegt werden, dass diese durch den Cluster gesteuert werden soll.

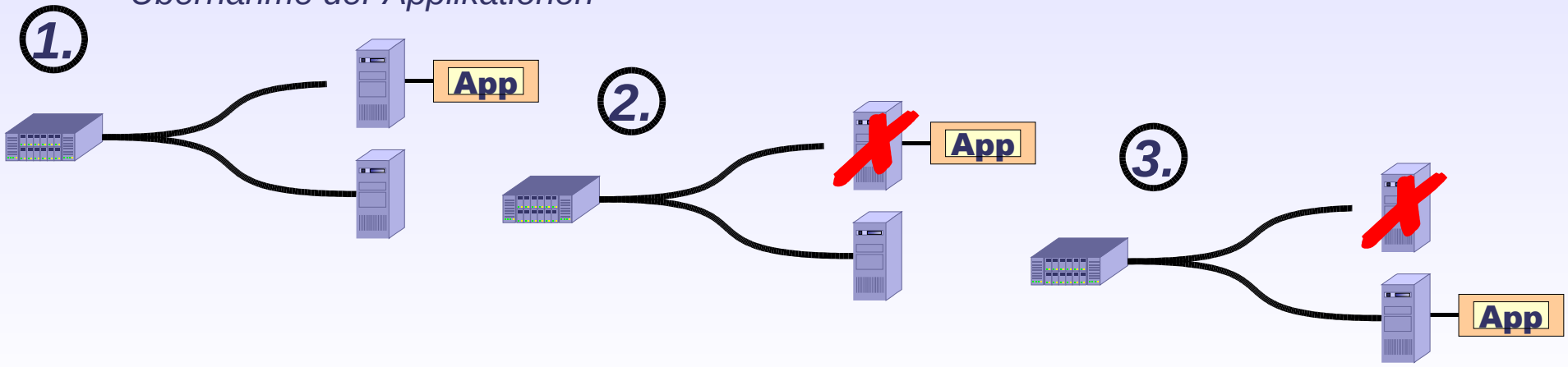
Mögliche Target States sind: ***no_control***
up *(Hochverfügbarkeit)*
down

Setzen des Target States

```
#> appadmin -s up iscsi
```

Applikationen mit Status "up" werden im Cluster verfügbar gehalten

- *Eliminierung fehlerhafter Knoten*
- *Übernahme der Applikationen*



Gestaltung der Hochverfügbarkeit

Festlegen von Node Power Control Parametern

The Power of Simplicity

- *Mit Hilfe von NPC Routinen können fehlerhafte Knoten von überlebenden Knoten beendet werden – STONITH Konzept.*
- *Die Node Power Control kann auch zum Ein- und Ausschalten der Rechner genutzt werden (einheitliche Bedienschnittstelle auch bei unterschiedlichen Hardware-/Firmware-Konzepten)*
- *Zustandsüberwachung erfolgt beim OSL SC über das SAN.*
- *Einstellungen zur "Node Power Control" sind unter /etc/dvsc/npccfg vorzunehmen. Es sind zahlreiche vorgefertigte Skripte vorhanden*
- *Nutzen und Überprüfen der Funktionalität mit "ndadmin"*

```
#> ndadmin -e power_off merkur
...
#> ndadmin -e power_on merkur
```


Autoren: B. Miemietz, C. Schmidt

Gestaltung der Hochverfügbarkeit

Beachtung von Applikationsressourcen

The Power of Simplicity

- *Applikationen können im OSL Storage Cluster grundsätzlich nebeneinander auf einem Clusterknoten laufen*
- *Sie beeinflussen sich nicht gegenseitig, wenn sie nach EAS Konzept erstellt wurden*
- *Manche Applikationen sind jedoch "ressourcenhungrig", sie können somit andere Applikationen auf dem selben Host beeinflussen*
- *Lösung im OSL SC: Prioritäten für Applikationen und Ressourcendefinitionen*

Anzeigen der Knotenperformance

```
[root@pluto] ndadmin -qp
nodename          id cpu-isa          ncpu  clock    memory      swap       rip
pluto              1 sparcv9+vis         1    650      2048     11265      3.49
jupiter            2 sparcv9+vis2        2   1100      2048     20480     16.84
```

Beim Erstellen einer Applikation muss eine Priorität angegeben werden (1 höchste, 512 niedrigste)

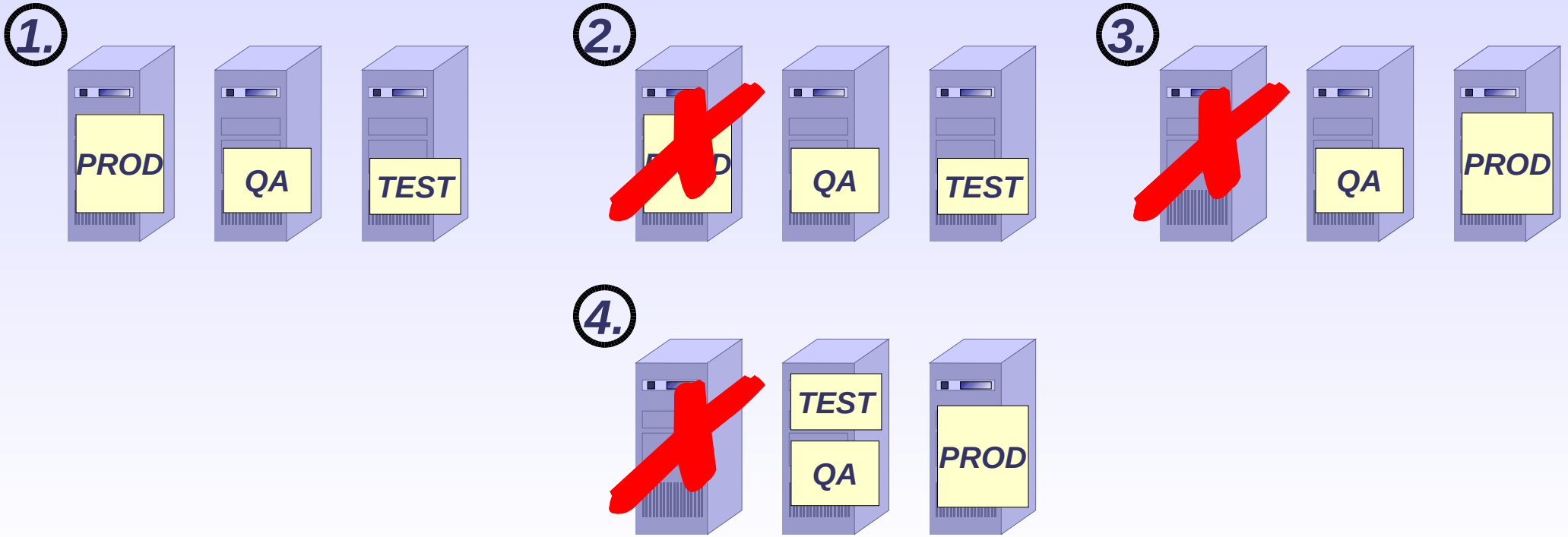
```
[root@pluto] appadmin -q
priority nickname    tgt_state    migration_strategy          rip          mem
  20     iscsi       up           node priority                10         1024
  78     samba       down         node priority                 2          100
```

Gestaltung der Hochverfügbarkeit

Definition und Nutzung von Applikationsressourcen

The Power of Simplicity

- *Ressourcenverbrauch wird anhand von RIP Werten und Hauptspeicherbedarf festgelegt*
- *Applikationen mit dem Status up werden im Cluster verfügbar gehalten*
- *Fällt ein Knoten mit laufenden Applikationen aus, werden die Applikationen auf anderen Knoten gestartet*
- *Hierbei können Applikationen mit niedriger Priorität verdrängt werden – z.B. ein produktives SAP System verdrängt das Testsystem, falls auf diesem Knoten keine ausreichenden Ressourcen vorhanden sind*



Autoren: B. Miemietz, C. Schmidt

Gestaltung der Hochverfügbarkeit

Mögliche Reaktion (Auswahl von Beispielen)

The Power of Simplicity

- **Pfadausfall**
  - Der OSL SC Multipathtreiber erkennt den Pfadausfall und es findet ein Pfadfailover statt
  - Sind keine Pfade mehr vorhanden wird der Knoten eliminiert (natürlich nur bei Nutzung der ACO und wenn es für den Neustart von Anwendungen erforderlich ist)
- **Stromausfall**
  - Failover in ein Backup RZ
  - Nutzen von permanent gespiegelten Volumes und Backupserver
- **RAID-Systemausfall**
  - Failover des RAID Systems ins Backup RZ
  - Nutzen von permanent gespiegelten Volumes
- **Hardwareausfall (Server)**
  - Eliminieren des Servers
  - Übernahme der Anwendungen durch andere Knoten des Clusters

Autoren: B. Miemietz, C. Schmidt

Storage-Server mit OSL Storage Cluster
Zusammenfassung

The Power of Simplicity

- *Fileserver und andere Applikationen können im OSL Storage Cluster mit Templates schnell erzeugt werden*
- *Durch virtuelle Ablaufumgebungen und einen globalen Storage Pool sind Applikationen auf allen Clusterknoten startfähig*
- *Mit NPC-Routinen kann ein HV Cluster mit automatischem Failover erstellt werden*
- *Applikationen können ungestört zusammen auf dem gleichen Knoten laufen*
- *Für Applikationen können Ressourcen und Prioritäten definiert werden*
- *Wenn eine hoch priorisierte Applikation Ressourcen benötigt, können andere Applikationen verdrängt werden*

Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

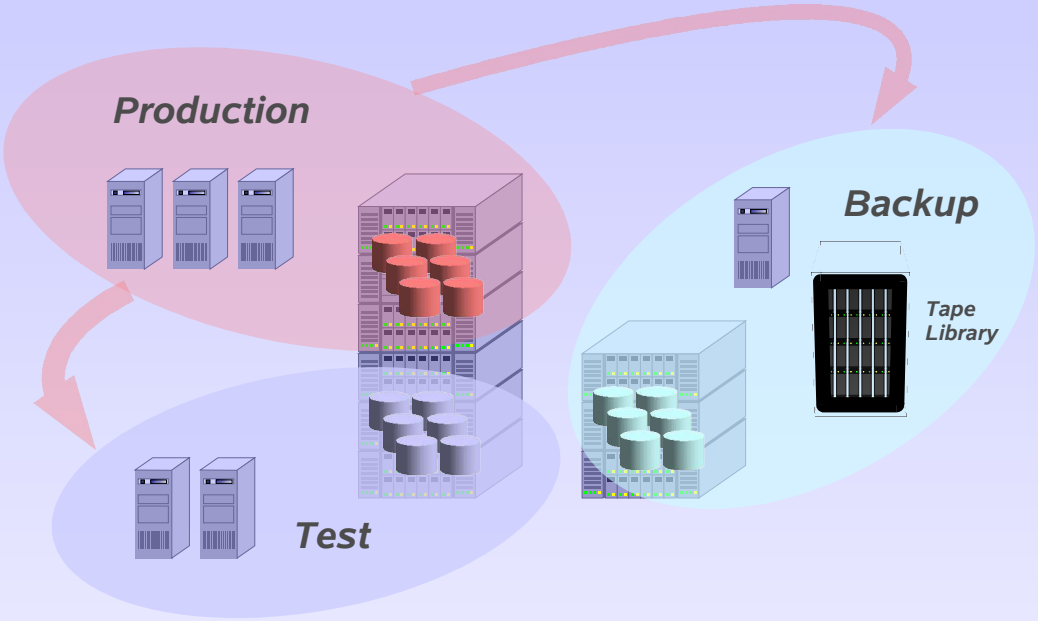
Extended Data Management & Application Aware Storage Management

Autoren: B. Miemietz, C. Schmidt

Storage-Universen im OSL-Storage-Cluster

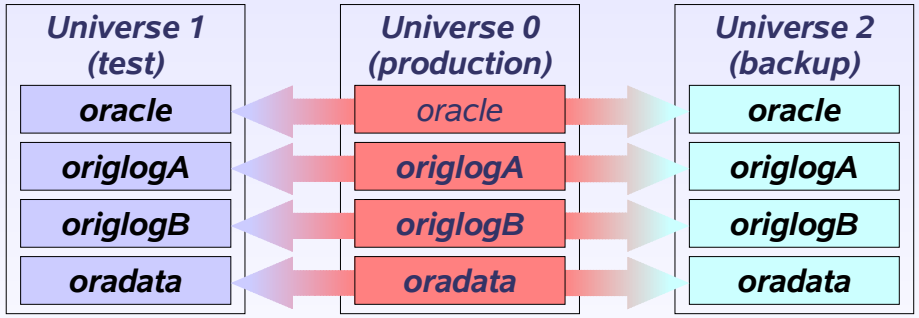
Abbildung logischer Beziehungen

The Power of Simplicity



- OSL Storage Universen bilden die Aufteilung der Ressourcen nach der Art der Nutzung ab
- Kopien eines Originals können jederzeit erstellt werden, auch unter Beibehaltung des Namens
- OSL Storage Cluster besitzt Informationen über die logischen Beziehungen zwischen den Universen

- Volle Integration der Universen in das Betriebssystem
- leichte Identifikation anhand des Namens
- Zugriff auf jede Instanz
  - jederzeit
  - von jedem Host aus

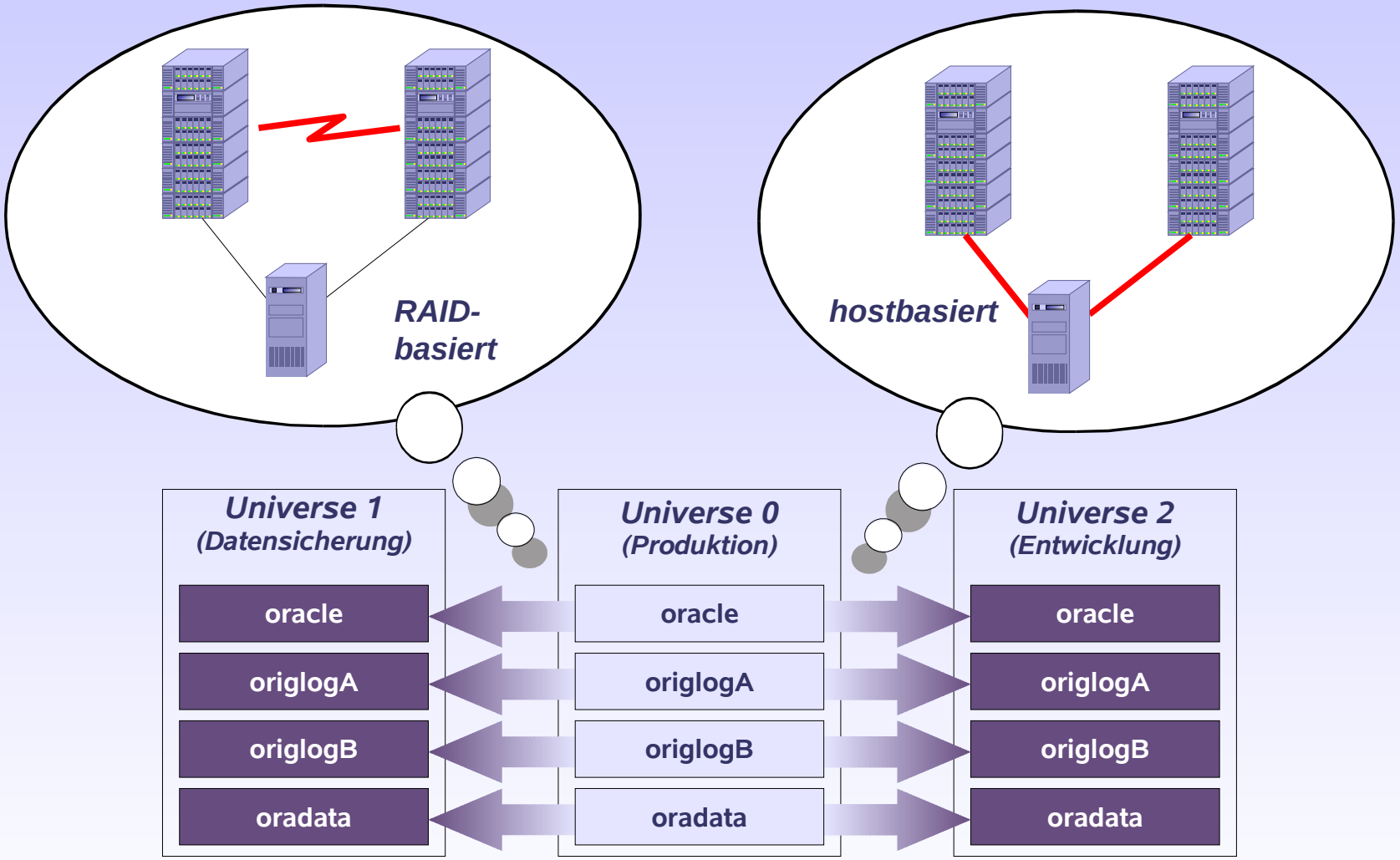


Autoren: B. Miemietz, C. Schmidt

Erzeugung von Storage-Universen

Zwei Möglichkeiten

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

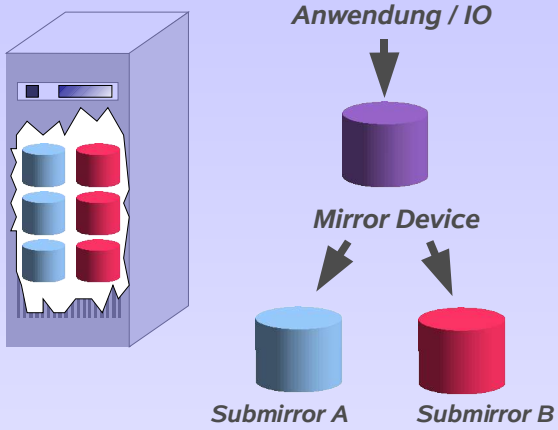
Warum noch eine Spiegel-Software?

Das gab es doch schon vor 30 Jahren

The Power of Simplicity

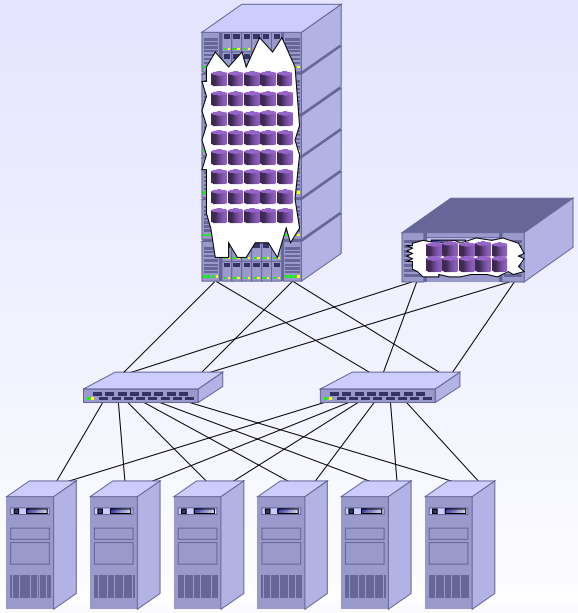
Stimmt! Und seitdem sah das so aus:

- Designschwerpunkt:
  Schutz vor Plattenausfällen
- statische Konfiguration
- Implementierung meist über hierarchisch organisierte Geräteknoten
- geringe Zahl von Geräten
- i. d. R. Administration für einen Rechner
- aufwendige Administration / OLR-Operationen



Aber die Anforderungen haben sich geändert:

- Schutz vor Plattenausfällen sollte keine Rolle mehr spielen
- Hostbasierte Spiegelung heute für:
  - Backup und Backup to Disk
  - schnellen Wiederanlauf nach logischen Fehlern
  - Clonen von Anwendungen (etwa Produktion auf Test)
  - ggf. „Mißbrauch“ für Disaster Precaution
- große Zahl von Geräten
- enge Verknüpfung mit der Anwendung
- Clustertauglichkeit

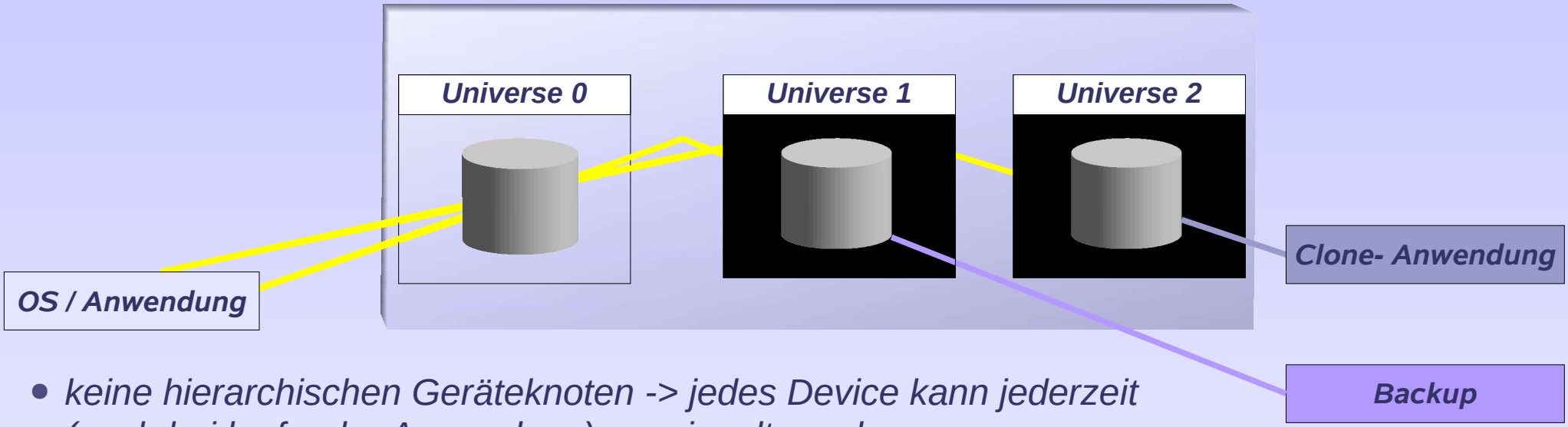


Autoren: B. Miemietz, C. Schmidt

Datenspiegelung mit XDM

ist auf RAID-Systeme, heutige RZ-Infrastrukturen und Anforderungen zugeschnitten

The Power of Simplicity



- *keine hierarchischen Geräteknoten -> jedes Device kann jederzeit (auch bei laufender Anwendung) gespiegelt werden*
- *Identische Gerätenamen für Master und Image dank OSL Storage Universen*
- *Überbrückung von Ausfällen des Masters (wenn Images im Status „connected“)*
- *Nach Disconnect der Images Zugriff auf diese vom selben oder von anderen Clusternodes*
- *Idle Synchronization, Idle Consistency Check, Incremental Synchronization*
- *Atomic Disconnect für beliebig zusammenstellbare Volumes und Volume-Gruppen*
- *synchrone, asymmetrische IO-Strategie mit Berücksichtigung wahrscheinlicher Anwendungsumgebungen: RAID-to-RAID Kopie, niedrigere Performance des Image-RAIDs*
- *Master und Images können unterschiedliche Volume-Typen und -Größen haben*

Autoren: B. Miemietz, C. Schmidt

Erzeugen eines Spiegels im OSL Storage Cluster

Ein kurzes Beispiel

The Power of Simplicity

```
[root@erde] smgr -c example -S 1g -R master
[root@erde] smgr -c example@1 -S 1g -R image
[root@erde] avmirror -q example
0       example ( simple, 1pc, 1024m) MASTER SOURCE -----  synchronized
1       example ( simple, 1pc, 1024m) image    -    -----  disconnected

[root@erde] avmirror -c example@1
INFO (avmirror): "1 example" connected as mirror instance, starting sync.
[root@erde] avmirror -q example
0       example ( simple, 1pc, 1024m) MASTER SOURCE s----  synchronized
1       example ( simple, 1pc, 1024m) image  target s---- synchronizing ( 10%)

[root@erde]  avmirror -q
0       example ( simple, 1pc, 1024m) MASTER SOURCE s----  synchronized
1       example ( simple, 1pc, 1024m) image  target s----  synchronized

total of 1 mirrors, 1 active, 0 need maintenance
```

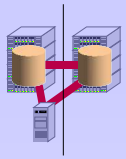
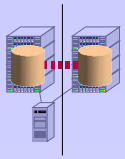
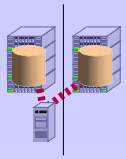
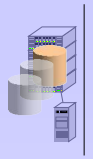
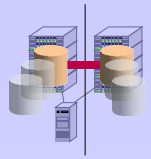
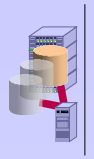
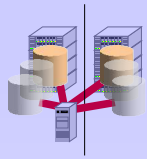

Autoren: B. Miemietz, C. Schmidt

Und warum überhaupt noch Spiegel?

Spiegel vs. Snapshots – Legenden und Fakten

The Power of Simplicity

Snapshots und Datenkopien im Vergleich

	 **synchron** (per RAID oder Host)	 **incr. per RAID** (periodischer Sync)	 **incr. per Host** (periodischer Sync)	 **HW-Snapshot** (lokal)	 **HW-Snapshot** (gespiegelt)	 **ZFS-Snapshot** (local)	 **ZFS-Snapshot** (gespiegelt)
Speicherbedarf	*200%*	*200-300%*	*200-300%*	*125-200%*	*250-400%*	*125-200%*	*250-400%*
mit o.g. Speicherbedarf mögliche Kopien bzw. Snapshots	*1*	*1-2*	*1-2*	*ca. 1-15*	*ca. 1-15*	*ca. 1-15*	*ca. 1-15*
OLTP-Performance Original	*o*	*++*	*++*	*++*	*o*	*++*	*+*
OLTP-Performance bei gleichzeitigem Zugriff auf Original und Kopie/Snapshot	*X*	*++*	*++*	*+*	*o*	*o*	*-*
simultaner Zugriff auf Original und Kopie/Snap vom gleichen Host	*nein*	*ja*	*ja*	*ja*	*ja*	*ja*	*ja*
simultaner Zugriff auf Original und Kopie/Snap von verschiedenen Hosts	*nein*	*ja*	*ja*	*ja*	*ja*	*nein*	*nein*
Backup-Performance Kopie/Snap (bei simultanem OLTP-Betrieb Original)	*X*	*++*	*++*	*o*	*o*	*o*	*-*
Integration mit Host-OS + Applikationen	*o*	*o*	*++*	*-*	*-*	*+*	*+*
Handhabung Komplettlösung	*o*	*o*	*+*	*-*	*-*	*o/+*	*o/+*
Performanceanforderung Original-Speichersystem	*hoch*	*hoch*	*hoch*	*sehr hoch*	*sehr hoch*	*sehr hoch*	*sehr hoch*
Performanceanforderungen an das Spiegel-Speichersystem (remote)	*hoch*	*mäßig*	*mäßig*	*X*	*sehr hoch*	*X*	*sehr hoch*
Verfügbarkeit Kopie/Snap nach User- oder SW-Fehler	*X*	*++*	*+/++*	*++*	*++*	*+*	*+*
Schutz gegen Ausfall Ausfall Orignal-Speichersystem	*++*	*+*	*+*	*X*	*++*	*X*	*+/++*
Brauchbarkeit Snap/Kopie als Sicherung	*X*	*++*	*++*	*--*	*++*	*--/-*	*-/o*
Belastung Host	*sehr gering*	*sehr gering*	*gering*	*sehr gering*	*sehr gering*	*mäßig*	*mäßig*

++ sehr gut + gut o mäßig - schwächer -- schlecht X entfällt/nicht vorhanden

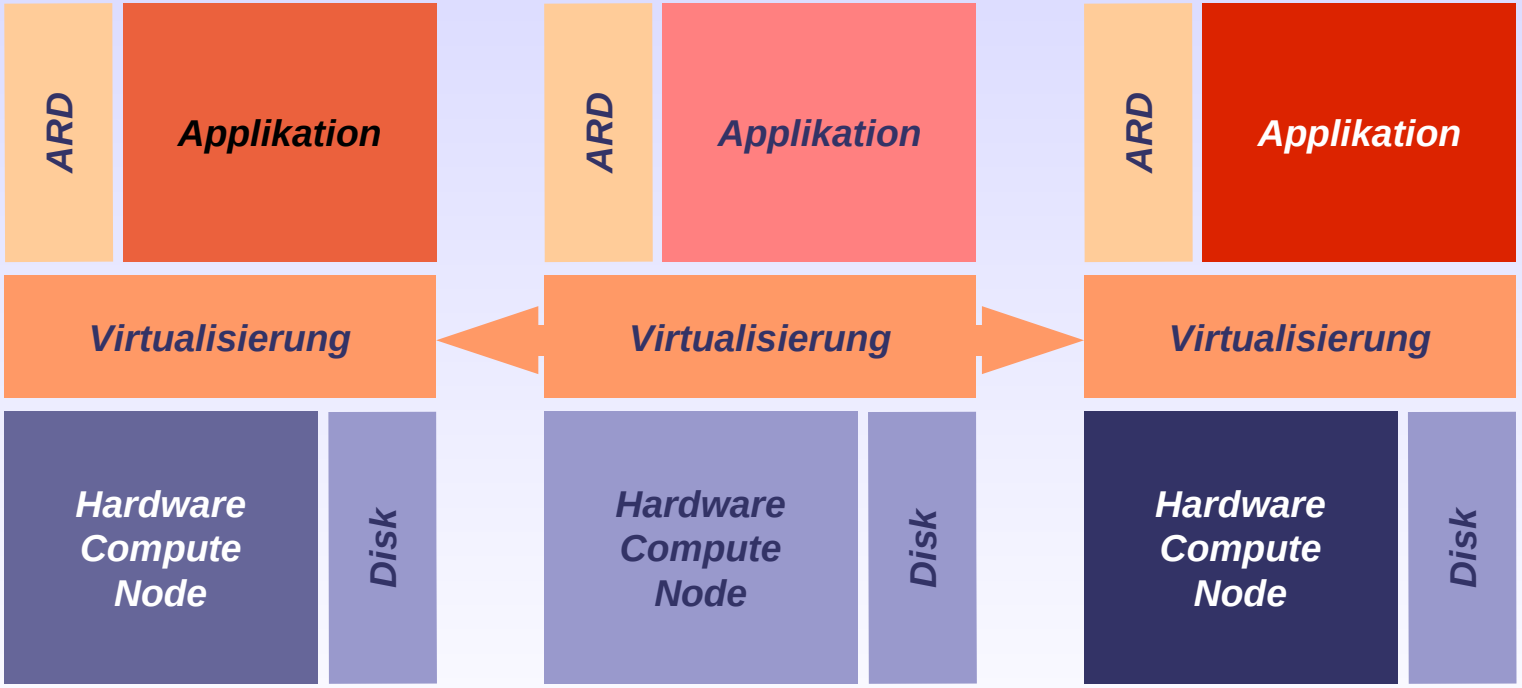
Autoren: B. Miemietz, C. Schmidt

EAS – die „technologiefreie“ Virtualisierung

Systemkopien mit XDM

The Power of Simplicity

- Gemeinsam mit Kunden erarbeitete EAS-Guides und Beispiele
- Tools für interaktive oder Batch-Bearbeitung
- Weitgehend automatisierte Abläufe
- Intelligente Nachsynchronisation



Autoren: B. Miemietz, C. Schmidt

Einige Demonstrationen
Application Aware Storage Management

The Power of Simplicity

- *Applikationsbezogene Übersichten mit smgr*
- *Spiegeln ganzer Applikationen mit appmirror*
- *Starten gespiegelter Applikationen*
- *inkrementelle Synchronisation*
- *schneller Neustart*
- *Applikationsbezogene Bandbreitensteuerung*

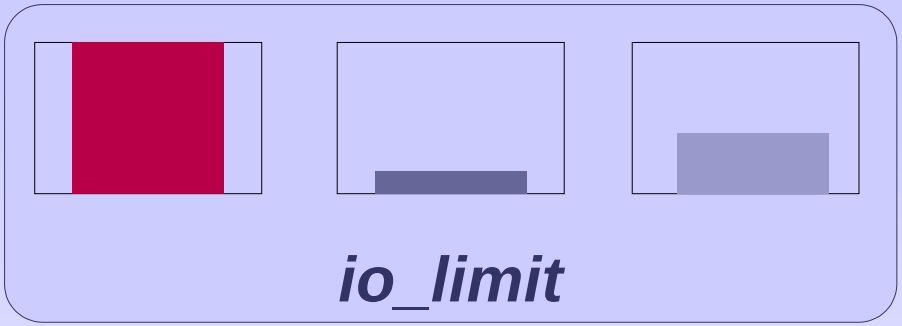
Autoren: B. Miemietz, C. Schmidt

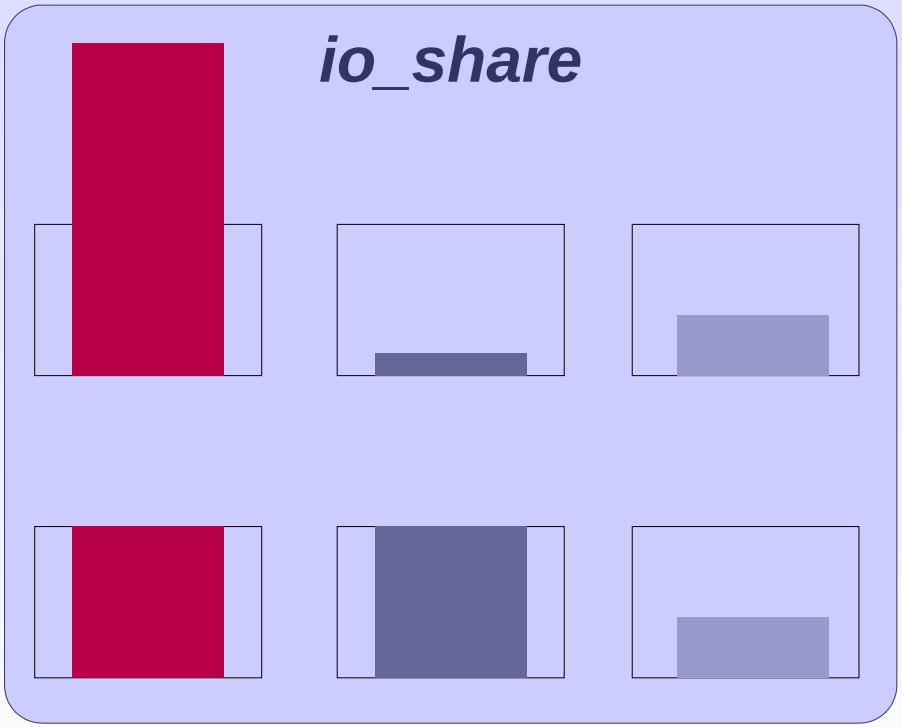
Bandbreitensteuerung

Vielfältige Möglichkeiten

The Power of Simplicity

- *Warum?*
  - *Sättigung IO-Kanäle*
  - *Sättigung Speichersystem(e)*
  - *Konkurrenz Applikationen*
- *Was?*
  - *einzelne Volumes*
  - *Gruppen von Volumes*
  - *Applikationen*
- *Wie?*
  - *absolute Bandbreite (io_limit)*
  - *adaptives Konzept (io_share)*
  - *Limit für Synchronisationsvorgänge (sync_limit)*
- *Mit welchem Resultat?*
  - *verbessertes Antwortzeitverhalten*
  - *faire Verteilung von IO und CPU-Bandbreite*
  - *reduzierte CPU-Belastung*
  - *gesteigerter Gesamtdurchsatz*



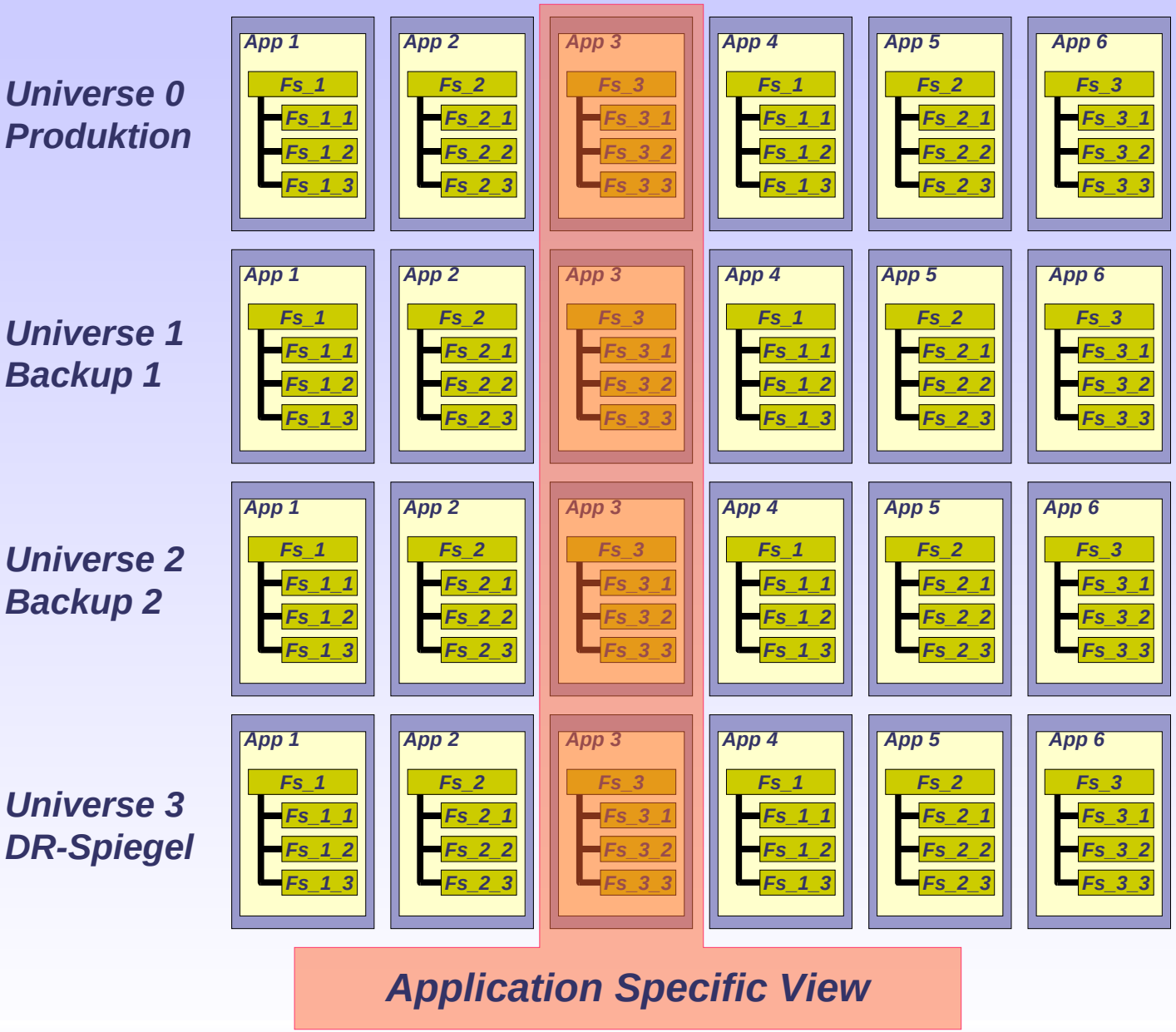


Autoren: B. Miemietz, C. Schmidt

Application Aware Storage Management

Anwendungsbeschreibungen und Volume Management integriert

The Power of Simplicity

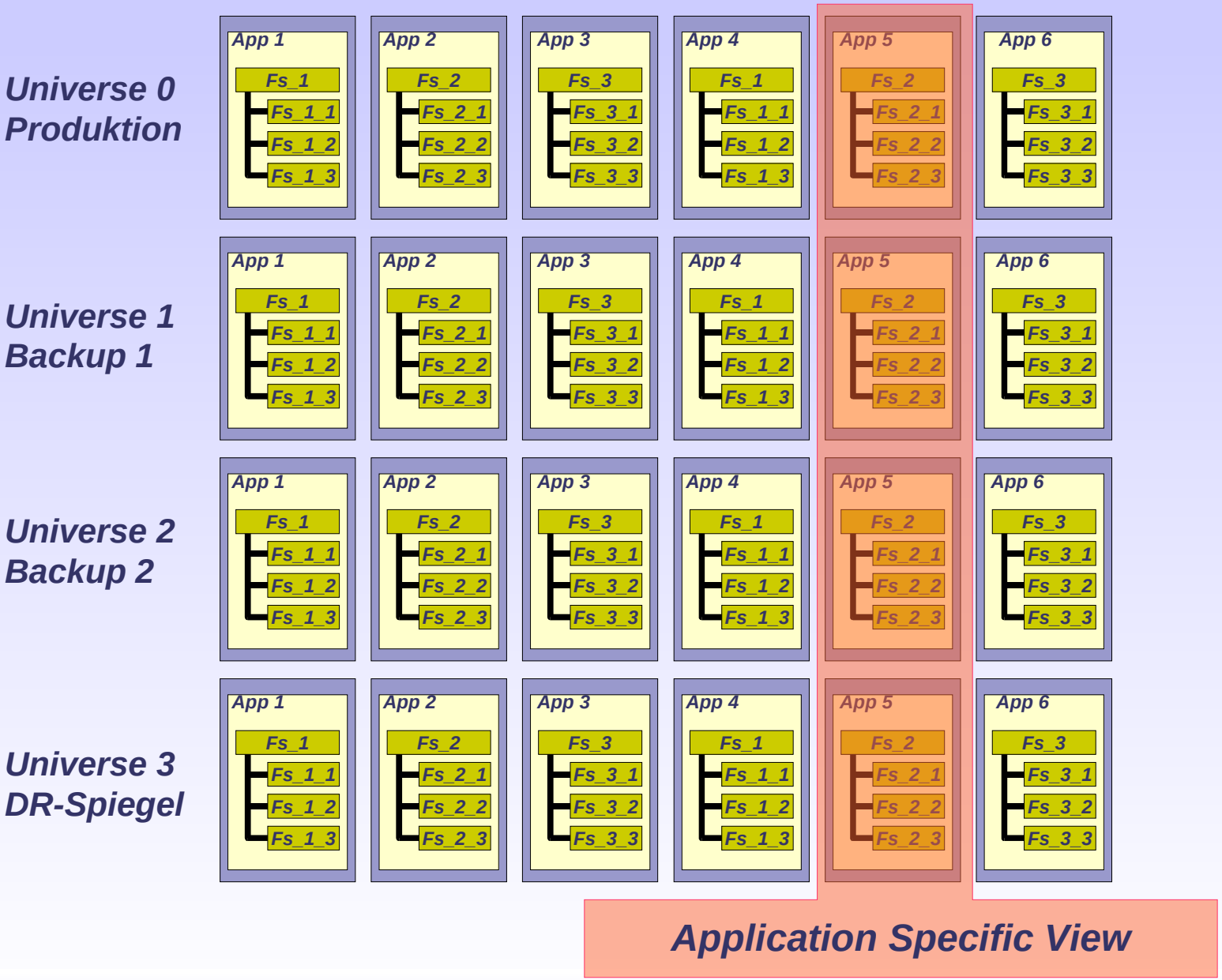


Autoren: B. Miemietz, C. Schmidt

Application Aware Storage Management

Anwendungsbeschreibungen und Volume Management integriert

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

Summary Application Aware Storage Virtualization

Anwendungsbeschreibungen und Volume Management integriert

The Power of Simplicity

- *Konfiguration der Applikation ordnet Geräte Applikationen zu*
- *Übersicht zu Ressourcenverbrauch einzelner Applikationen*
- *Basis für Applikations-Spiegel /-Clones*
- *Applikationsbezogene Spiegelzustände*
- *Applikationsbezogene Steuerung von Aktionen (z. B. set source)*
- *Applikationsbezogene Bandbreitensteuerung*

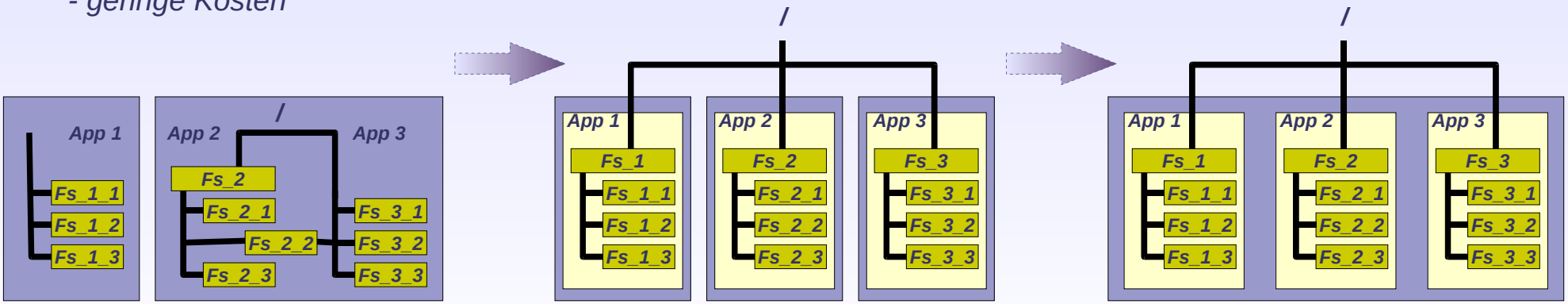
Autoren: B. Miemietz, C. Schmidt

Applikation Awareness braucht die passende Organisation

Technologie allein kann das Problem nicht lösen

The Power of Simplicity

- *organisatorische Maßnahme: Encapsulated Application Setup (EAS)*
  - *klare Strukturierung der Installation*
  - *kostet als organisatorische Maßnahme vergleichsweise wenig*
  - *Voraussetzung für flexible Zuordnung von Anwendungen zu Hosts*
  - *Datenaustausch und auch die Datensicherungswelt profitieren*
  - *Problem: katastrophales Installationsdesign vieler Anwendungen*
- *der Cluster*
  - *liefert das Framework für den EAS*
  - *stellt eine global einheitliche Ablaufumgebung bereit*
  - *sorgt für Datenkonsistenz (Konfiguration)*
  - *kann die Anwendungen und Hochverfügbarkeit steuern und überwachen*
- *beides zusammen:*
  - *Vereinfachung*
  - *Flexibilität und Verfügbarkeit*
  - *geringe Kosten*



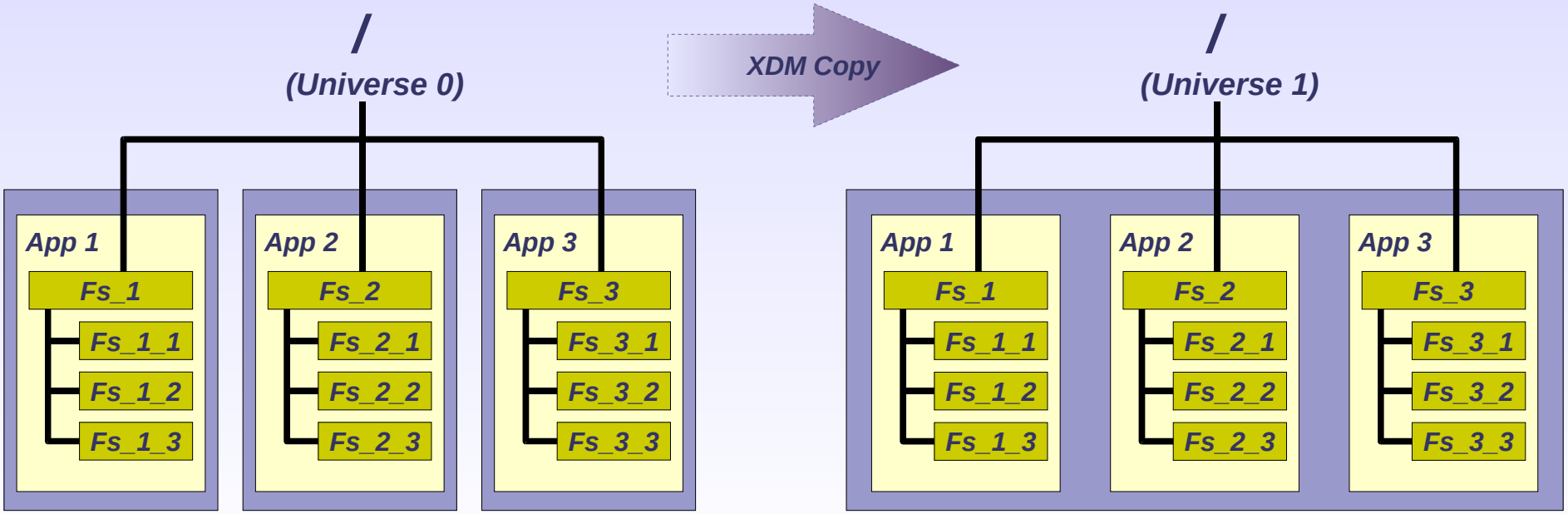
Autoren: B. Miemietz, C. Schmidt

EAS – die „technologiefreie“ Virtualisierung

Encapsulated Application Setup

The Power of Simplicity

- Trennung von Knoten, Betriebssystem und Applikationen
- Global eindeutige Pfade für Applikationen (Global Root)
- Freie Beweglicheit der Applikationen zwischen Rechnern
- Leichter Austausch von Knoten
- Vereinfachungen Backup
- Instant Restart in Verbindung mit XDM

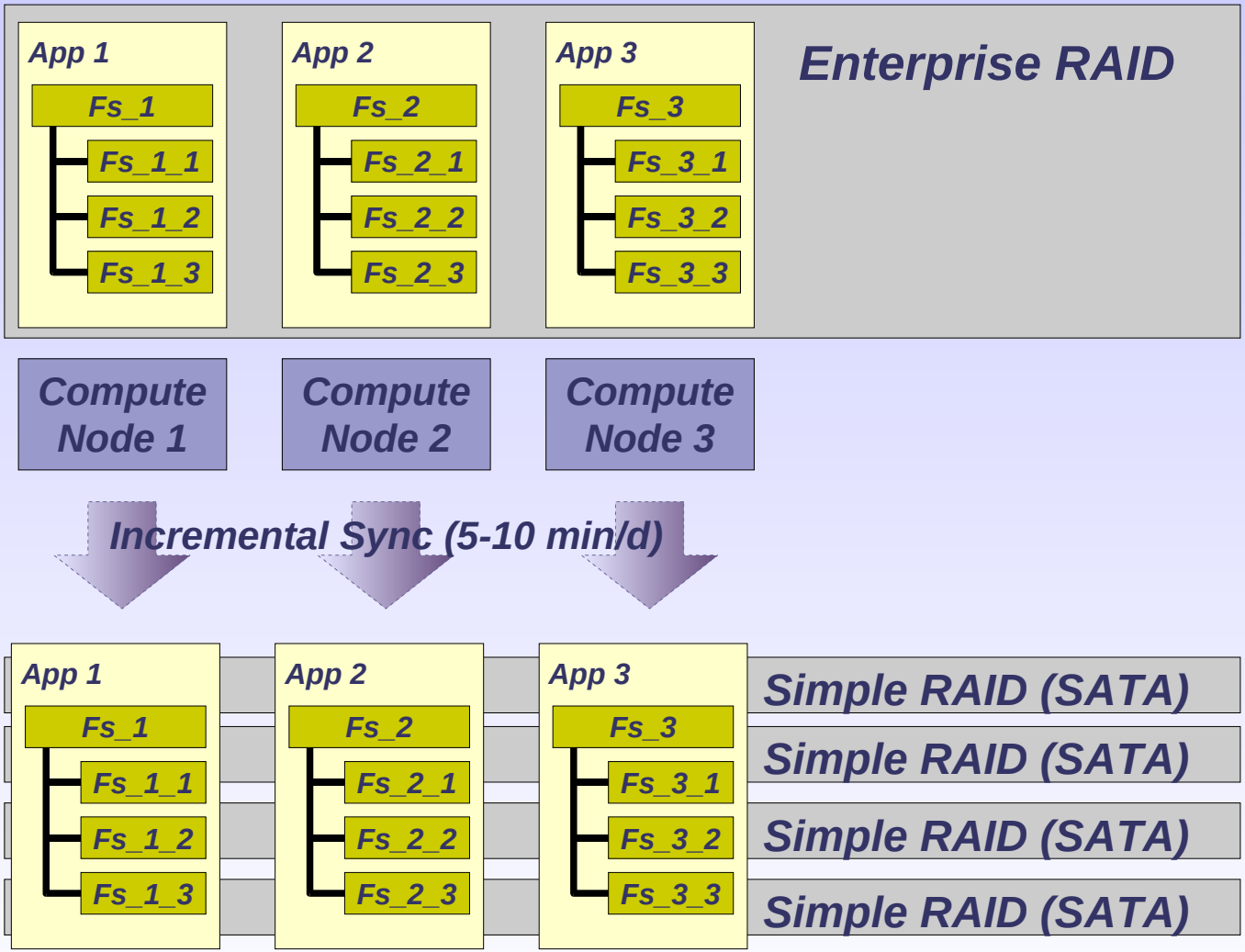


Autoren: B. Miemietz, C. Schmidt

High Speed Backup mit XDM

"Abfallprodukt" existierender Anwendungsbeschreibungen

The Power of Simplicity



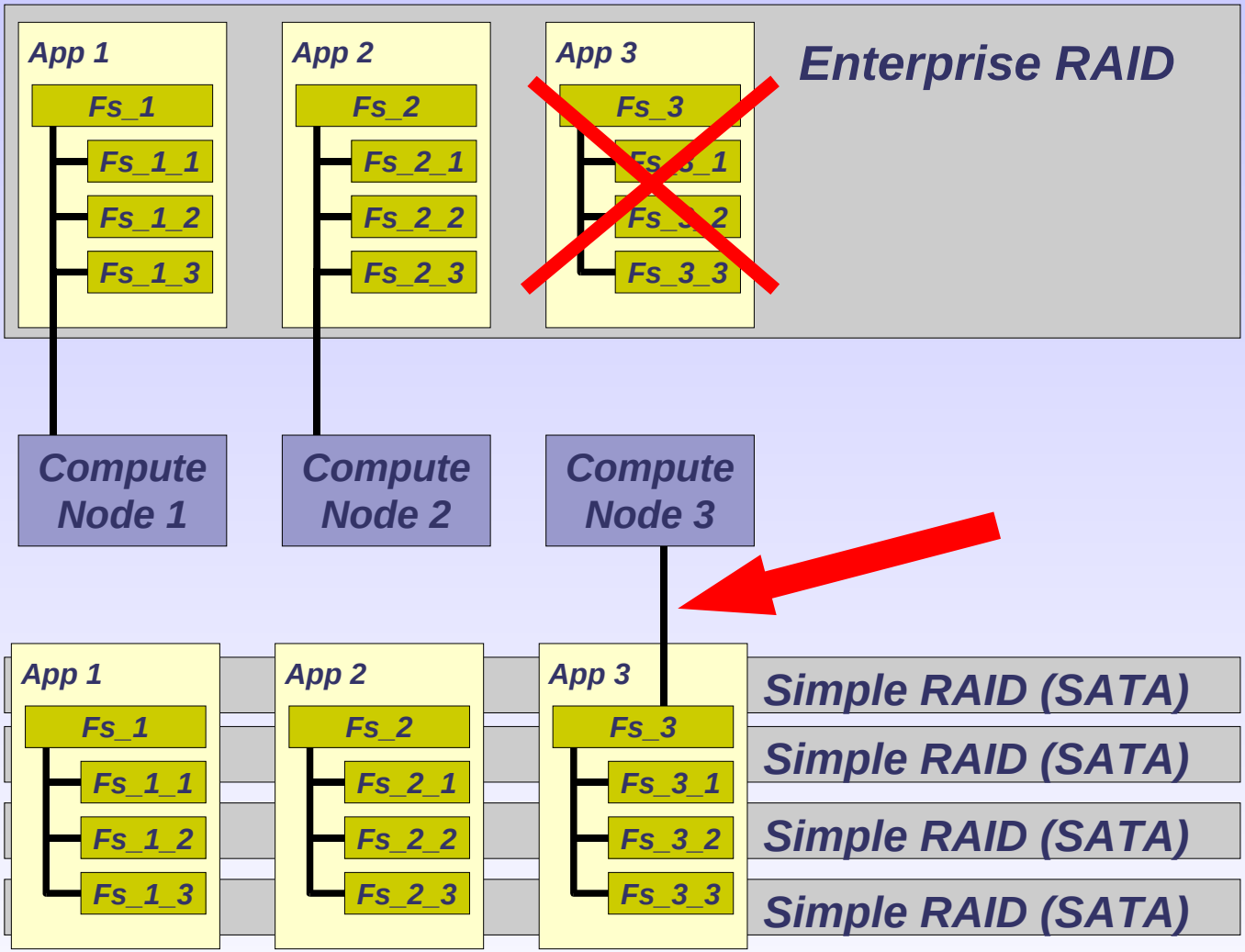
- *Extrem kurzes Backup für Compute-Nodes*
- *High-Speed Streaming to Tape*
- *Keine Belastung der Compute-Nodes während Backup to Tape*
- *Restart-fähige Images der Applikation im Backup-Universum*

Autoren: B. Miemietz, C. Schmidt

Restorefreies Recovery mit XDM

Tape-Backup wird nur im Ausnahmefall benötigt

The Power of Simplicity



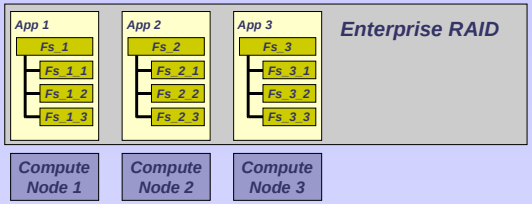
- *Sofortiger Wiederanlauf*
- *Kein Restore vom Tape*
- *Preview-Möglichkeit*
- *Bei Bedarf High-Speed Streaming from Tape*
- *Resync auf Enterprise-Storage bei bereits laufender Anwendung*

Autoren: B. Miemietz, C. Schmidt

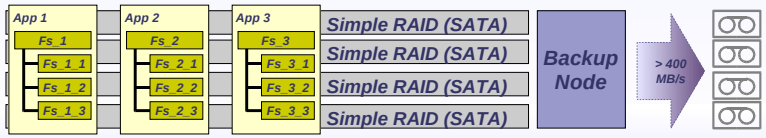
Differenzierung zu klassischen Backuplösungen

Möglichkeiten von Festplatten konsequent nutzen

The Power of Simplicity



Incremental Sync (5-10 min/d)



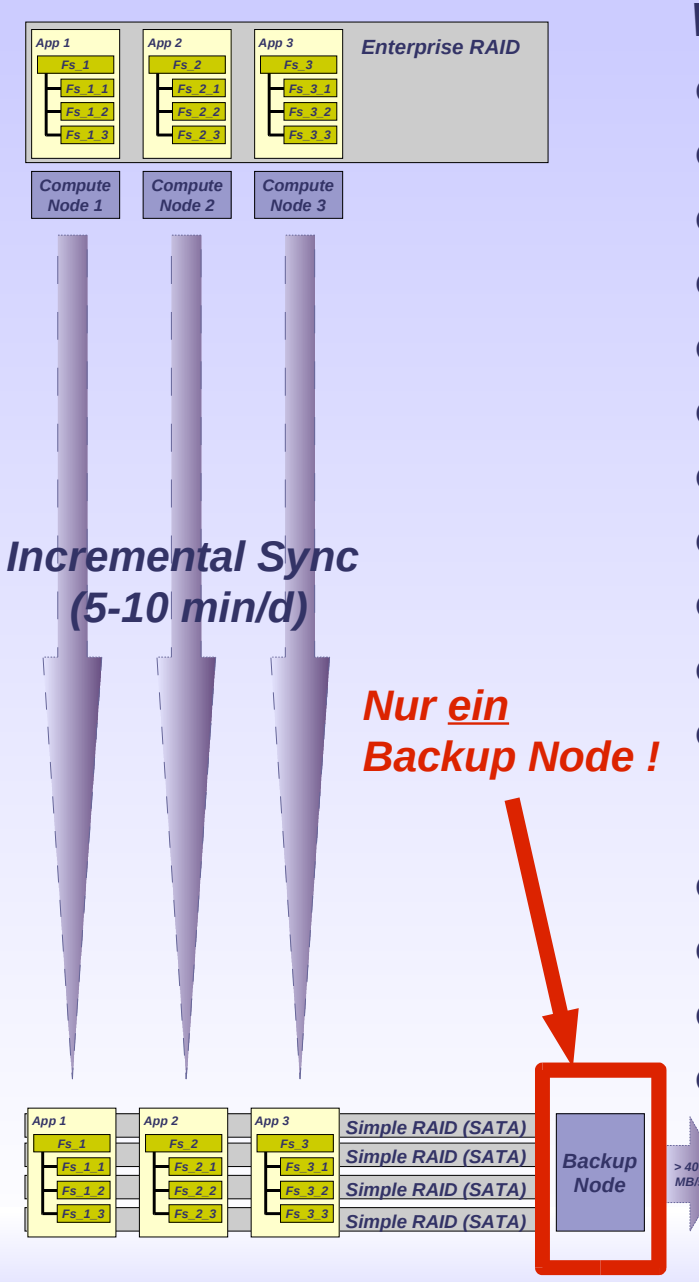
- *Extrem kurzes Backup für Application Nodes*
- *Minimale CPU-Belastung auf den Application Nodes (keine Verarbeitung der Daten)*
- *Nutzung SAN statt LAN*
- *kein Backup-Client auf Application Nodes (kein dezentrales Pflegen von Konfigurationen)*
- *Atomarer Backup – Dauer: NULL damit konsistenter Zustand*
- *Restartfähige Images der Applikation damit extrem schneller Wiederanlauf*
- *SW für Tape-Backup nur auf DASI-Server*
- *Zentrale Administration*
- *extreme Durchsätze bei Tape-Backup/-Restore möglich*
- *niedrige Anforderungen an Backup-RAID*
  - *ermöglicht SATA mit hoher Dichte*
  - *niedriger Platzbedarf*
  - *kürzere Backup-Zeiten*
  - *reduzierter Stromverbrauch / Wärmeabgabe*
- *adaptive Fahigkeiten bzw. "selbstlernend"*
- *Integration mit HV*
- *leicht zu DR-Umgebung ausbaubar*

Autoren: B. Miemietz, C. Schmidt

Integration mit Backup to Tape

Beispielimplementierung für Legato Networker

The Power of Simplicity



Was bietet die Integrationslösung von OSL?

- kombiniert B2D mit Bandsicherung
- applikationsorientiertes Verfahren
- sofortiger Neustart von Backup-Disk (kein Tape-Restore)
- Steuerung der Sicherungen über Networker-GUI oder CLI
- integriertes Pre- und Postprocessing
- mehrere Sicherungen pro Tag möglich
- differenzierte Erfolgskontrolle über die Networker-Indizes
- Aufzeichnung von Dateisystem-Informationen für Restore
- eindeutige Identifikation kompletter Sicherungen
- einfacher Restore kompletter Sicherungen
- weitere Funktionen für Oracle
  - Tool für Logrestore und Roll Forward
  - Archivierungslösung
- LAN-free Backup
- „cluster aware“, d.h. kein Eingriff bei Umschaltungen nötig
- automatische Gleichverteilung der Plattenlast
- hohe Durchsätze – VTL unter diesem Aspekt entbehrlich

Autoren: B. Miemietz, C. Schmidt

```
# dvamb2d -s nfs1@0 -t nfs1@1
```

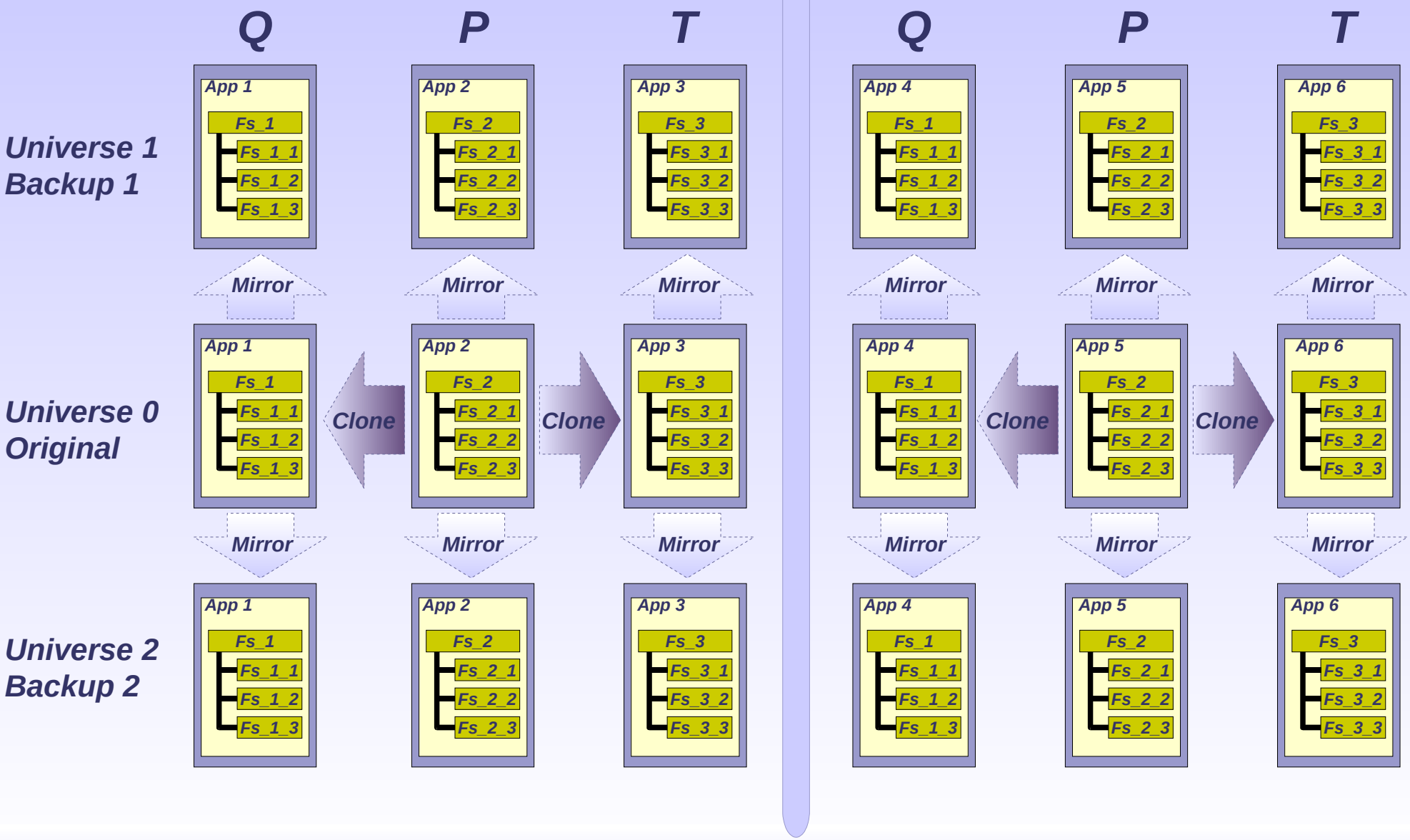
- *zentrale Steuerung*
- *zentrale Logfiles*
- *automatisiertes Auffinden der Applikation*
- *automatisierte Anpassung an notwendiges Pre- Postprocessing*
- *Übersichten und detaillierte Informationen*

Autoren: B. Miemietz, C. Schmidt

Spiegeln und Clonen

Für jeden Anwendungsfall den passenden Replikationstyp

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

Spiegeln und Clonen

Es geht auch anders ...

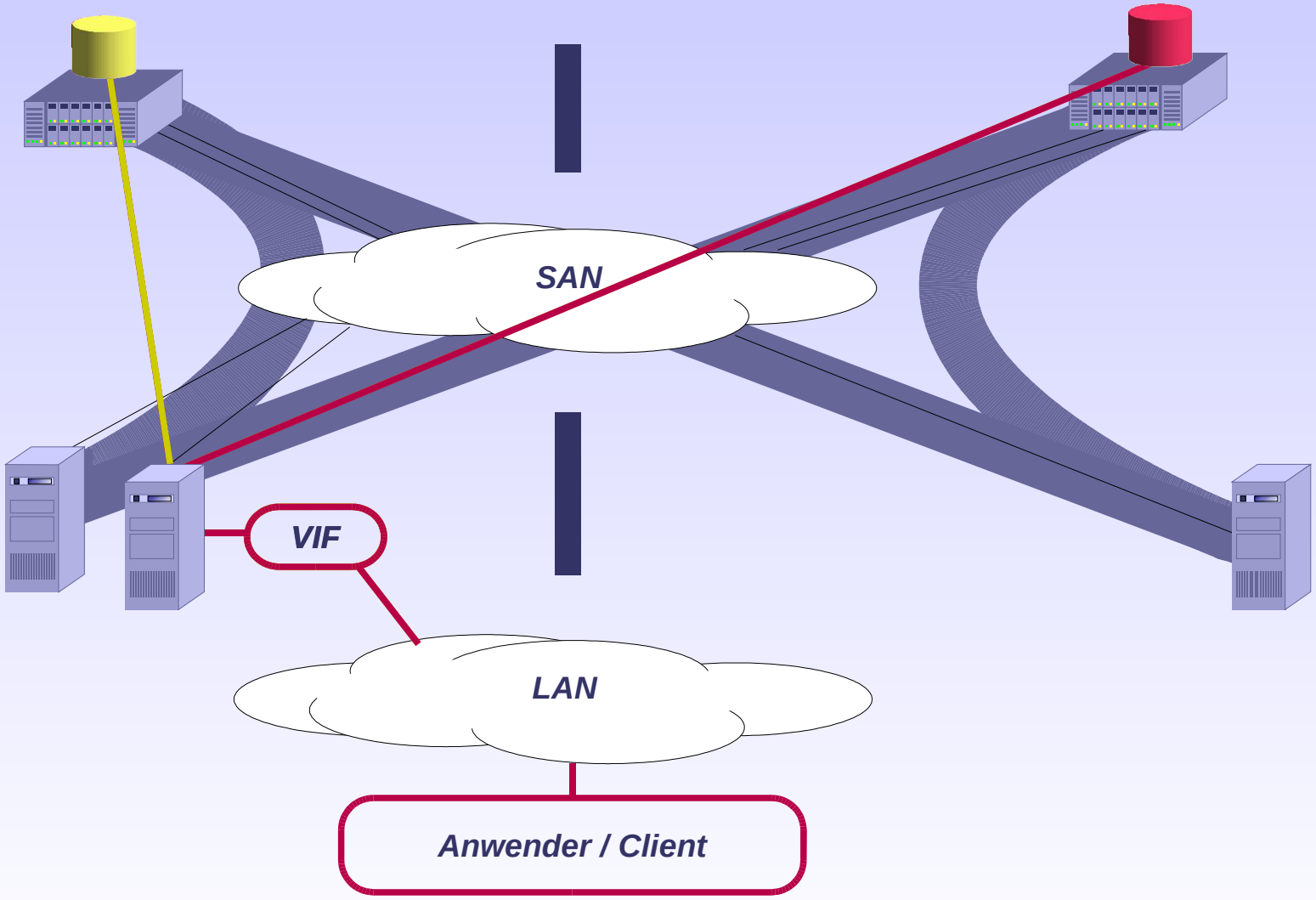
The Power of Simplicity

Autoren: B. Miemietz, C. Schmidt

Alles zusammen

Clusterfähige Storage Virtualisierung, Backup, HV, Disaster Recovery

The Power of Simplicity

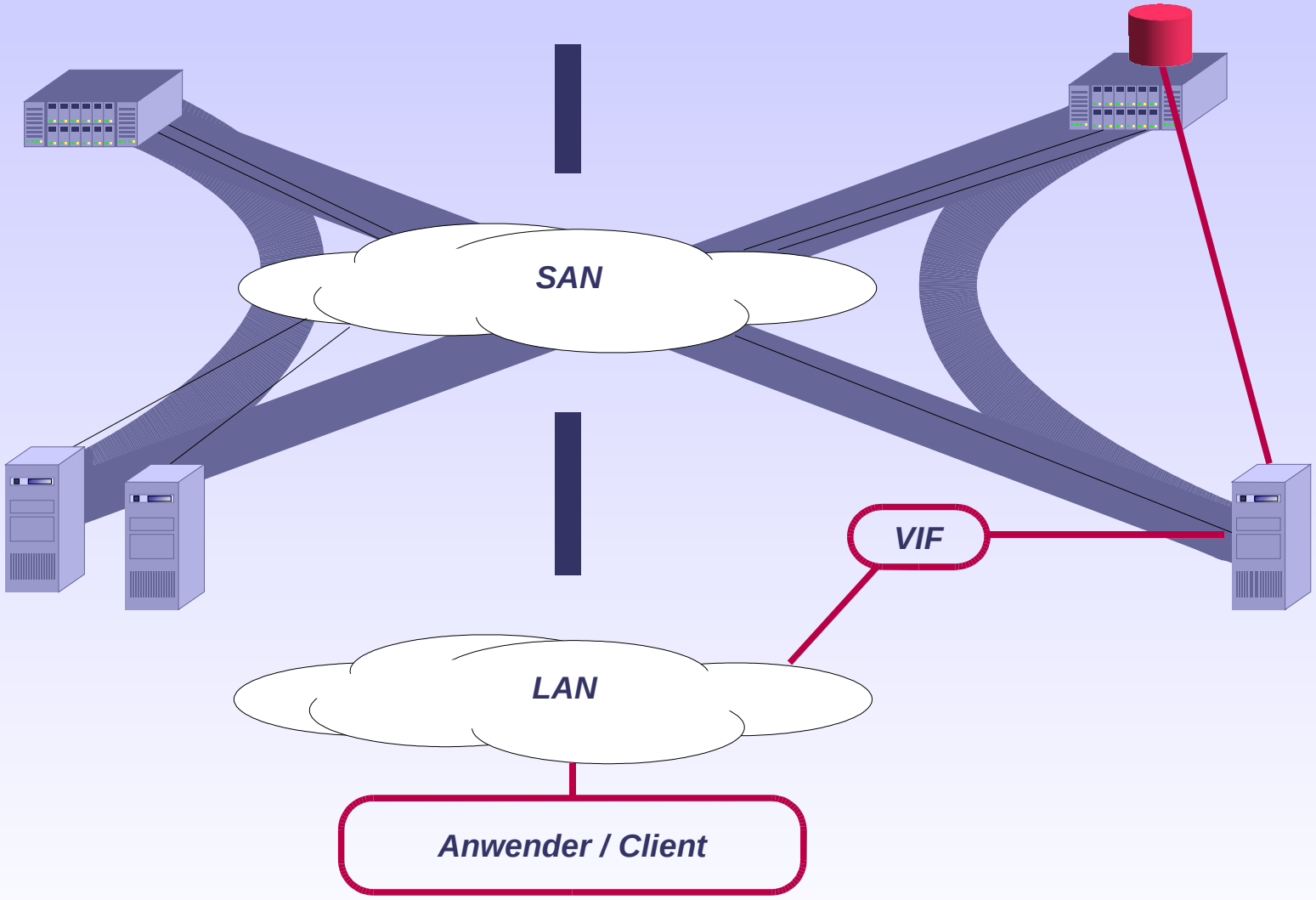


Autoren: B. Miemietz, C. Schmidt

Alles zusammen

Clusterfähige Storage Virtualisierung, Backup, HV, Disaster Recovery

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Zusammenfassung

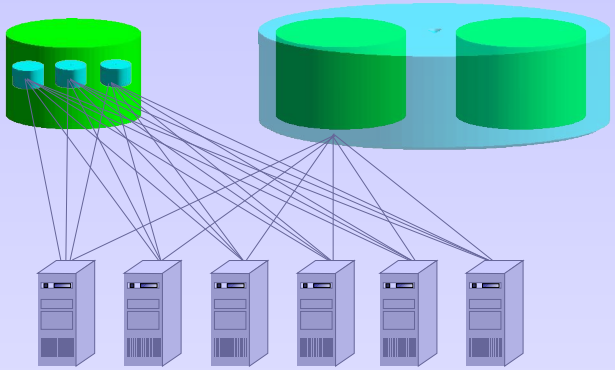
Autoren: B. Miemietz, C. Schmidt

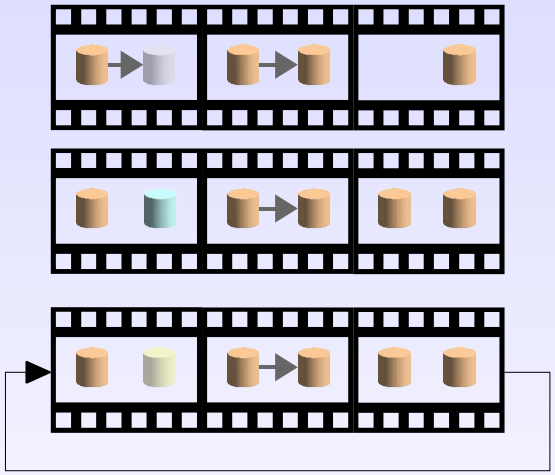
Was bietet OSL Storage Cluster im Speichermanagement?

Blockbasierte Virtualisierung – bedarfsgerecht und zuverlässig

The Power of Simplicity

Basis-Virtualisierung	*Physical Volumes + Application Volumes* *linear oder integriert (simple, concat, stripe)* *HW-Abstraktion und IO-Multipathing* *systemgestützte Allokation* *Online-Konfig./-Dekonfig./-Vergrößerung*
clusterweit	*global devices / global namespace* *integrated access management*
globale Pools	*rechnerübergreifend* *global inventory* *verschnittfreie Ausnutzung*
Daten verschieben	*online Daten verschieben / reorganisieren* *automatische Priorisierung Anwendungs-IO*
Daten clonen	*einmalig online auf beliebige Ziele kopieren* *atomare Operationen für mehrere Volumes*
Daten spiegeln	*Dauerhafte Beziehung Master -> Image* *bis zu 3 Images* *inkrementelle Resynchronisation* *atomare Operationen für mehrere Volumes* *Überbrückung Fehler auf Master*
Sonderfunktionen	*XVC (Extended Volume Controls)* *z. B. Pause, Stop, Trigger, Aktionen* *Bandbreitensteuerung* *detaillierte Statistik*





Autoren: B. Miemietz, C. Schmidt

Mehr als ein Volumemanager

Integration mit Clustertechnologie bringt weitere Vorteile

The Power of Simplicity

Speichervirtualisierung mit Anwendungsbezug

- Konfiguration der Applikation ordnet Geräte Applikationen zu
- Übersicht zu Ressourcenverbrauch einzelner Applikationen
- Basis für Applikations-Spiegel /-Clones
- Applikationsbezogene Spiegelzustände
- Applikationsbezogene Steuerung von Aktionen (z. B. set source)
- Applikationsbezogene Bandbreitensteuerung

Hochverfügbarkeit und Performance

- Zentrale und symmetrische Administration
- Einfache Migration von Applikationen zwischen Knoten
- Hochverfügbarkeit und Lastverteilung
- Verteilung von Funktionen im Cluster (Backup) -> Effektivitätssteigerung

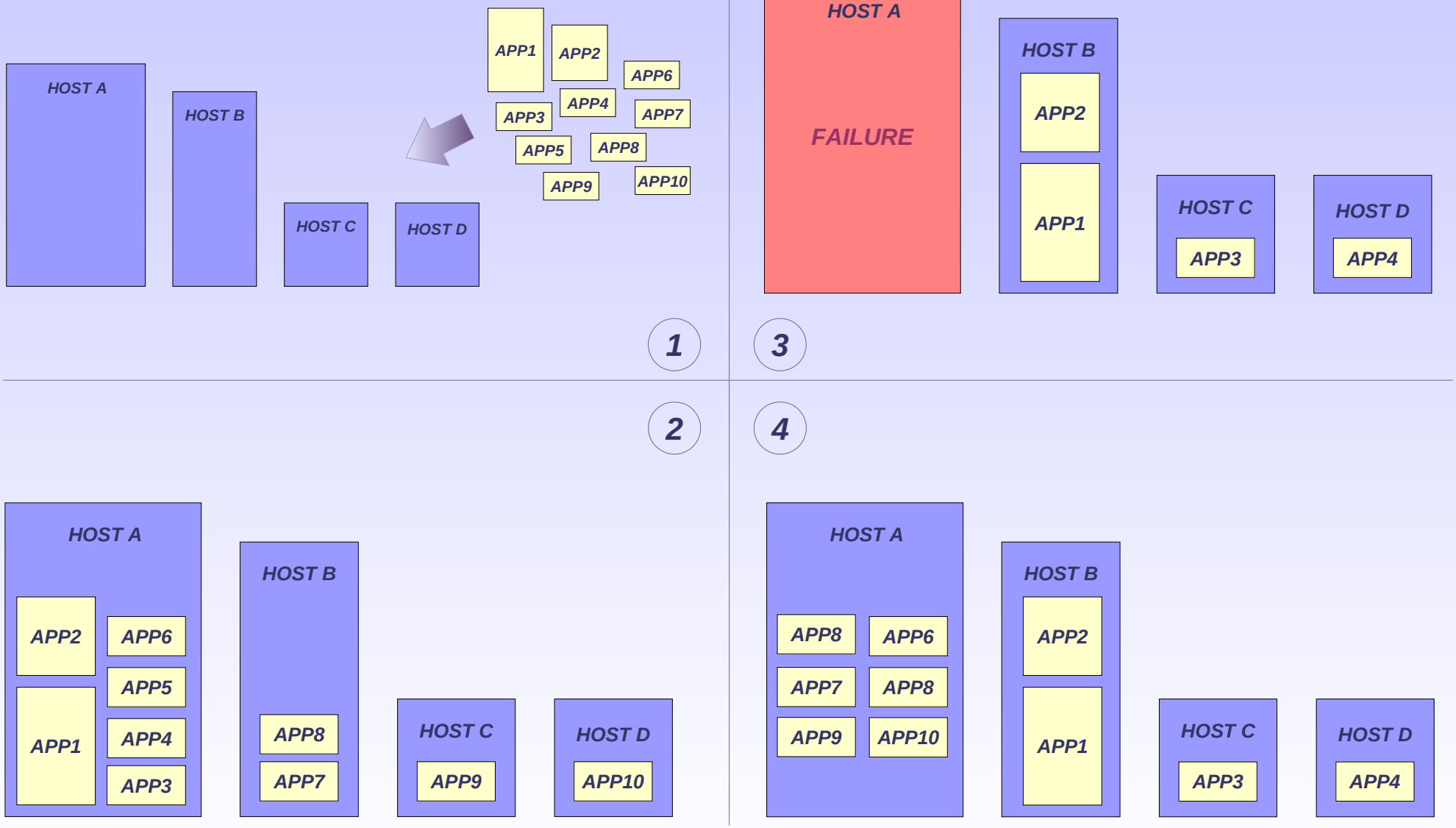
- Vereinfachung
- optimal für Konsolidierung (monolithisch oder parallel)

Autoren: B. Miemietz, C. Schmidt

Und das geht natürlich auch

High Availability und Adaptive Computing

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

The Power of Simplicity

Ausblick

RSIO – unser aktuelles Projekt

Autoren: B. Miemietz, C. Schmidt

Druck in Richtung NAS / Storage over IP

Vielfältige Angebote

The Power of Simplicity

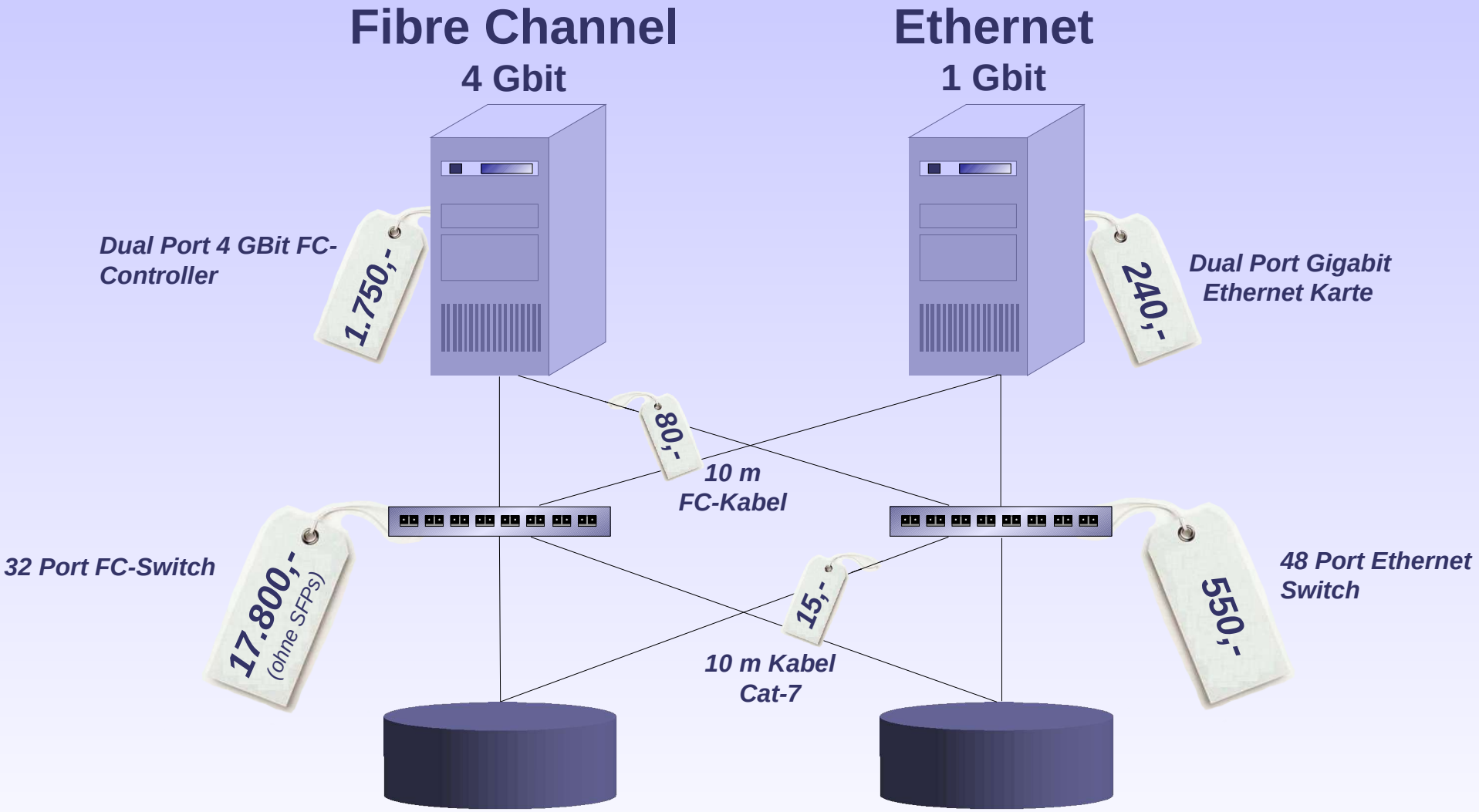
- *SNIA: - IP Storage Forum*
  - Ethernet Storage Forum (SIG iSCSI, SIG NFS, SIG CIFS?)
  (Compellent, Dell, EMC, HP, Intel, Microsoft, NEC, NetApp, Sun. Panasas)
- *lt. IDC 2009 ca. 30% des Marktes für Speichernetzwerke bei Ethernet*
- *reklamierte Vorteile:*
  - einheitliche Infrastruktur
  - Kostensenkung
  - Flexibilität
  - einfachere Handhabung, speziell auch mit virtuellen Maschinen
  - Data Sharing und weitere Zusatzfunktionen bei Filern
- *Ausprägung Fileserver:*
  - NFS
  - pNFS
  - CIFS
- *blockorientiert:*
  - iSCSI
  - FCIP
  - iFCP
  - FCoE
  - (e)NBD

Autoren: B. Miemietz, C. Schmidt

Kostenvorteile bei Storage over Ethernet

Was ist dran? Basis: Marktübliche A-Brand-Listenpreise 2008

The Power of Simplicity



Autoren: B. Miemietz, C. Schmidt

Storage over Ethernet
Was darf man noch erwarten?

The Power of Simplicity

- *sich anbahnendes 10GBase Ethernet ermöglicht Durchsätze in neuen Dimensionen*
- *Server bringen ab Werk bereits mehrere Ethernet Ports mit, weitere lassen sich preiswert nachrüsten*
- *Hardware kann damit effektives Multipathing ermöglichen*
- *Gigabit Ethernet liefert theoretisch Durchsätze bis 117 MB/s und wird in praxi heute kaum mehr als Shared Medium betrieben*
- *Durchsatz für viele Anwendungen also ausreichend*
- *SAN-Administration nicht immer einfach – vielleicht geht es hier besser ?*
- *ggf. Verbesserungen in der Bediensicherheit (Verfügbarkeit) möglich?*

Autoren: B. Miemietz, C. Schmidt

Brauchen wir ein neues Protokoll?

Anforderungen im Detail

The Power of Simplicity

- *Zuverlässigkeit*
- *Portabilität*
- *Skalierbarkeit*
- *einfache Handhabung auch in komplexeren Topologien (kein Zoning)*
- *Unterstützung heutiger wie zukünftiger Transport-Technologien*
- *Nutzbarkeit preiswerter Komponenten*
- *vollständige Abbildung aller relevanten IO-Aufrufe (read, write, ioctl ...)*
- *Multithreading-Support*
- *mit IP: Routingfähigkeit*
- *Erweiterbarkeit*
- *Einbindung in OSL-Clustertechnologie*

Autoren: B. Miemietz, C. Schmidt

Randbedingungen beim Design

Nützliche Eckdaten

The Power of Simplicity

Zeit zur Übertragung von 32 Byte an 1 GBit-Lanboard

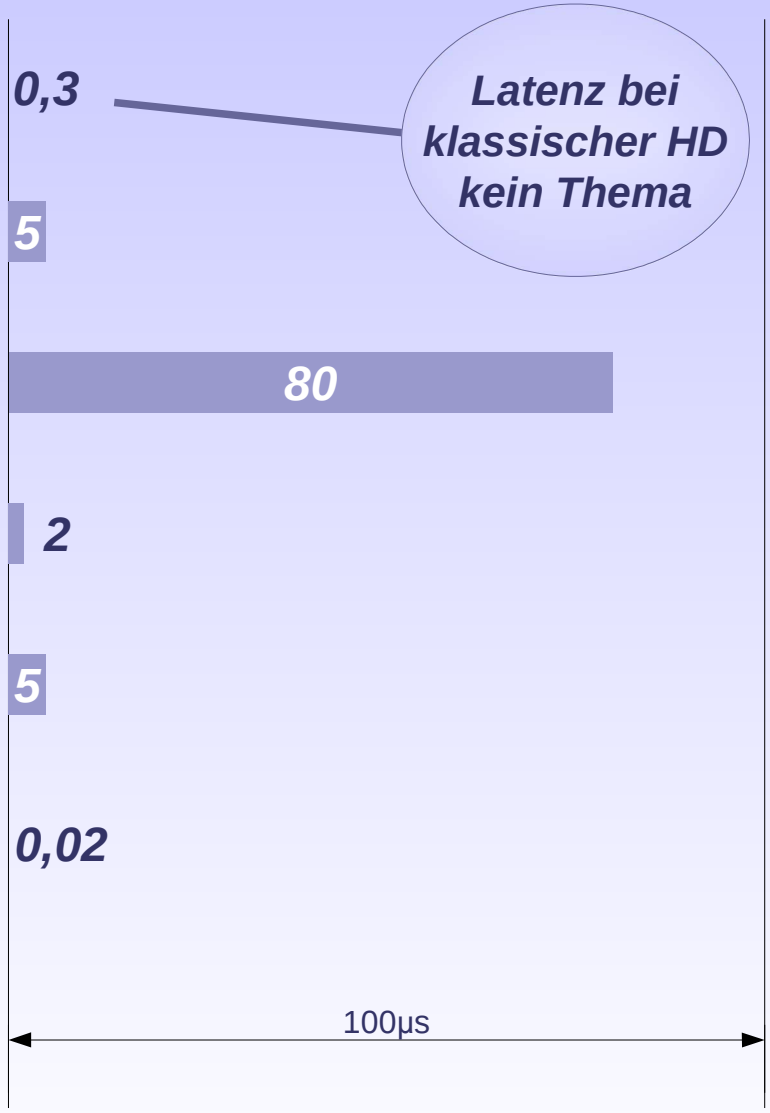
Zeit zur Übertragung von 512 Byte an 1GBit-Lanboard

Zeit zur Übertragung von 8kByte an 1GBit-Lanboard

memcpy 8k auf 600MHz-System

Threadwechsel per CV auf 600MHz-System

32Bit-Typkonvertierung auf 600MHz-System

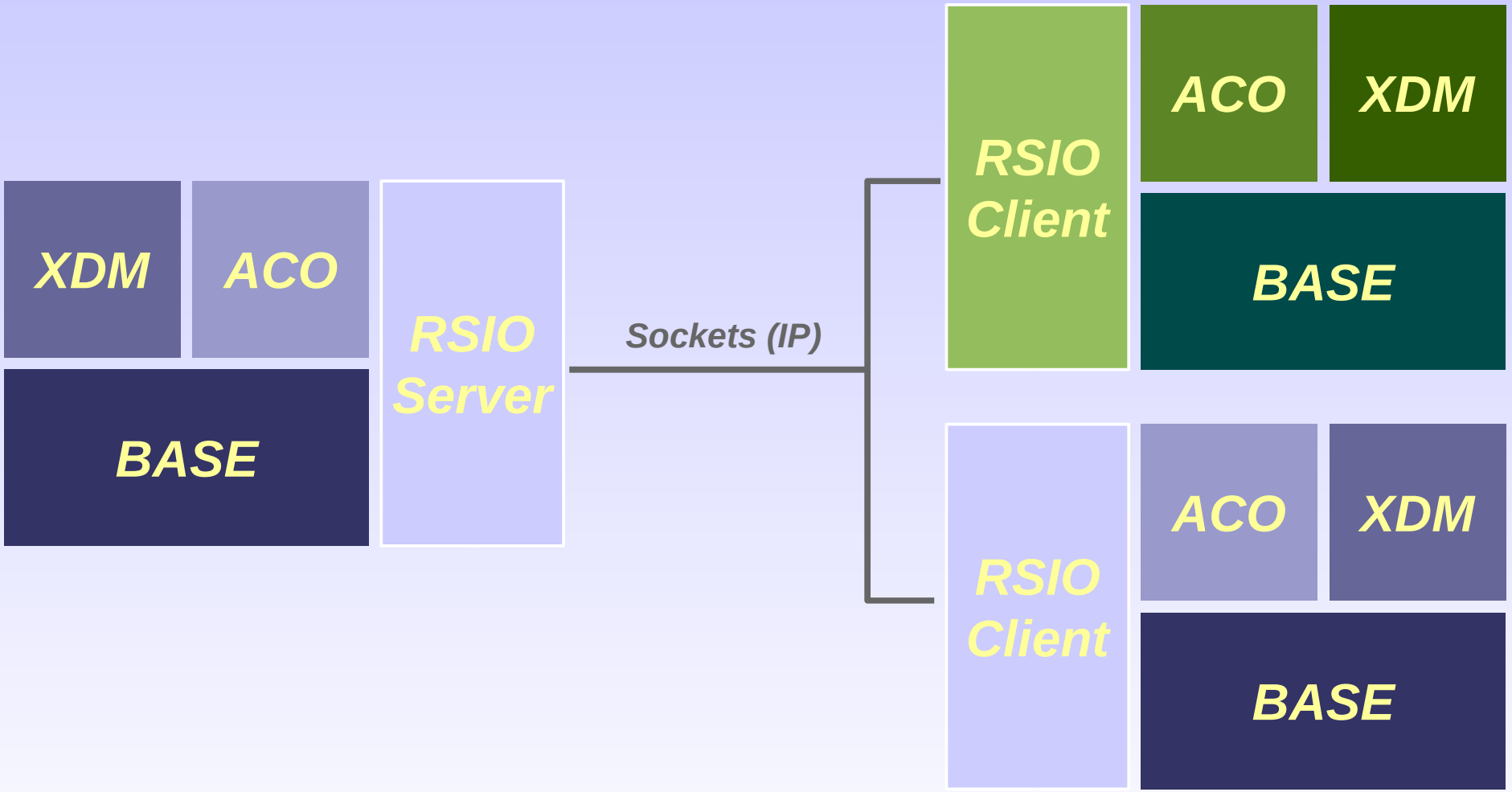


Autoren: B. Miemietz, C. Schmidt

Was kann ich damit anfangen?

Übersicht zur Ziellandschaft

The Power of Simplicity

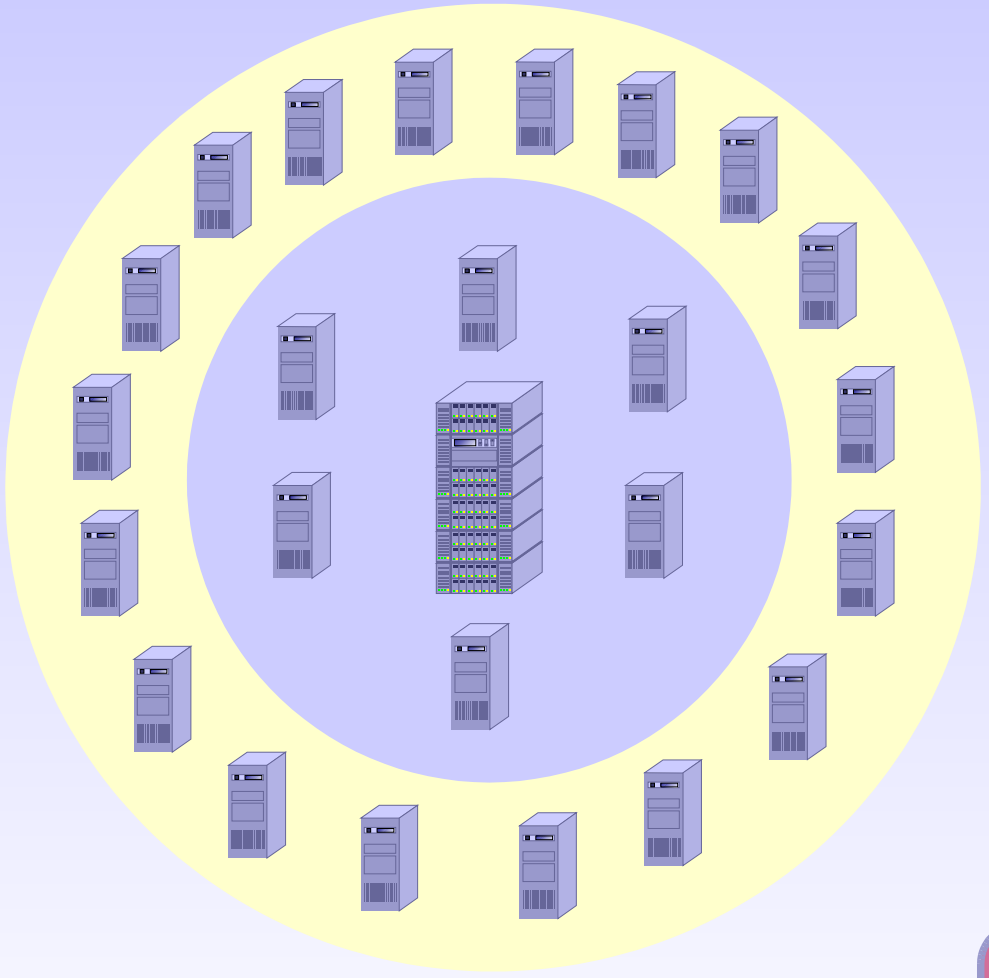


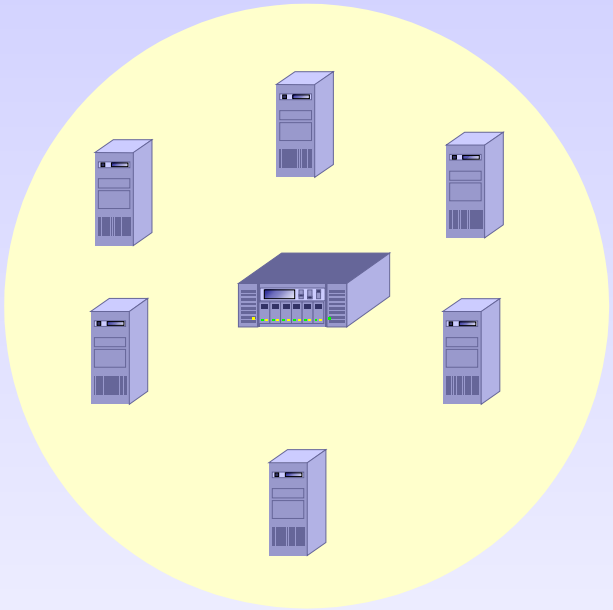
Autoren: B. Miemietz, C. Schmidt

Anwendungsszenarien mit OSL Storage Cluster

Kostengünstige Erweiterung der SANs oder autonom Storage over Ethernet

The Power of Simplicity





Integrierte Storage- und Clusterdienste über IP

OSL Gesellschaft für offene Systemlösungen mbH

Autoren: B. Miemietz, C. Schmidt

Stand: 06/2009

Roadmap

Die nächsten Termine (s. a. www.osl.eu)

The Power of Simplicity

29./30. Juni	**iX Solaris-Day**	**Stuttgart**
	Anwenderbericht OSL Storage Cluster *Vorstellung OSL Storage Cluster 3.1* *ausführliches Tutorial „Storage Server mit Solaris“*	
16./17. 9.	**OSL Technologietage**	**Berlin**
	Fachkonferenz und Anwendertreffen *Neues zu OSL Storage Cluster und RSIO* *Spannende Gastbeiträge*	
27./28. 10.	**SNW Europe**	**Frankfurt/M.**
	Fachbeitrag und Tutorial zu RSIO *Anwenderbericht* *im Ausstellungsteil OSL Storage Cluster*	

Autoren: B. Miemietz, C. Schmidt