Insights to Solaris

Vortrag bei den Linuxtagen 2009 in Graz

Jörg Möllenkamp
Senior Systems Engineer

Sun Microsystems
Geschäftsstelle Hamburg

Weblog
http://www.c0t0d0s0.org/

Moin!

Wer bin ich?

Jörg Möllenkamp
wie man an meinem Dialekt erkennt: nicht aus Österreich
Wie „sang“ Fettes Brot so schön: Nordisch by Nature

Senior Systems Engineer bei Sun Microsystems
und nebenaufgabendlich Solaris Evangelist
zu lange in der IT, um nicht hypefest und nicht zynisch zu sein.

Der Typ hinter http://www.c0t0d0s0.org
Reichweitenstärkstes Sun/Solaris-Blog im deutschsprachigen Bereich

c0t0d0s0.org? WTF?

c0t0d0s0 steht in einem Solarissystem für:

Controller 0
Target 0
Disk/Device 0
Slice 0

Hatte schon überlegt, hier im Anzug zu erscheinen.
So als Kontrapunkt.

Und das Linux-Tshirt, für das ich mich artig bedanke, wird man mir wohl über meinen kalten, toten Oberkörper ziehen müssen, um die Solaristätowierungen zu überdecken (ja, und jene von Apple). Sorry ...

;-)

Agenda

Ach ... lassen wir das mal mit der Agenda ...

Was will ich mit diesem Vortrag?

Mich in die Höhle des Löwens begeben,
den Kopf ins Maul stecken
und heil wieder rauskommen.

Gefahrensucher halt ...

Schliesslich heisst das hier Linuxtage ;)

Nein … ernsthaft:

Euch einige Gründe zu geben, mal in Solaris zu gucken!

Aber:

45 Minuten sind viel zu kurz ...
selbst für einen kurzen Überblick.

Achtung! Die folgende Präsenstation enthält dutzende Kommandozeilen. Kann Spuren von Marketing enthalten und ihre Sicht auf andere Betriebsysteme ändern. Wirkt unter Umständen pinguinozid ;)

Solaris ist anders!

Das hat Gründe. Auch diese möchte ich erläutern.

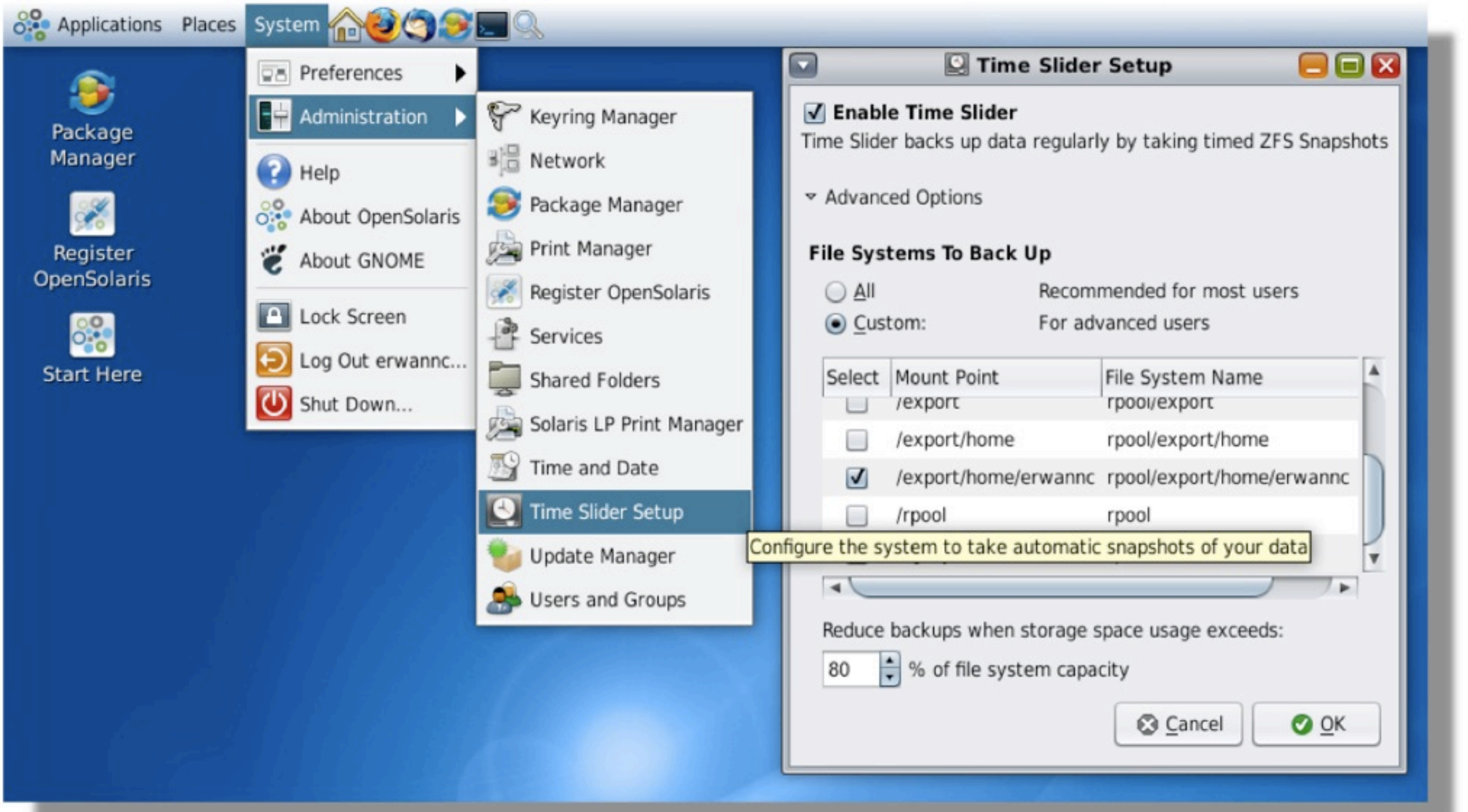
Insight 0

Ja ... Solaris kann auch gut aussehen und durchaus auch usable sein. ;)

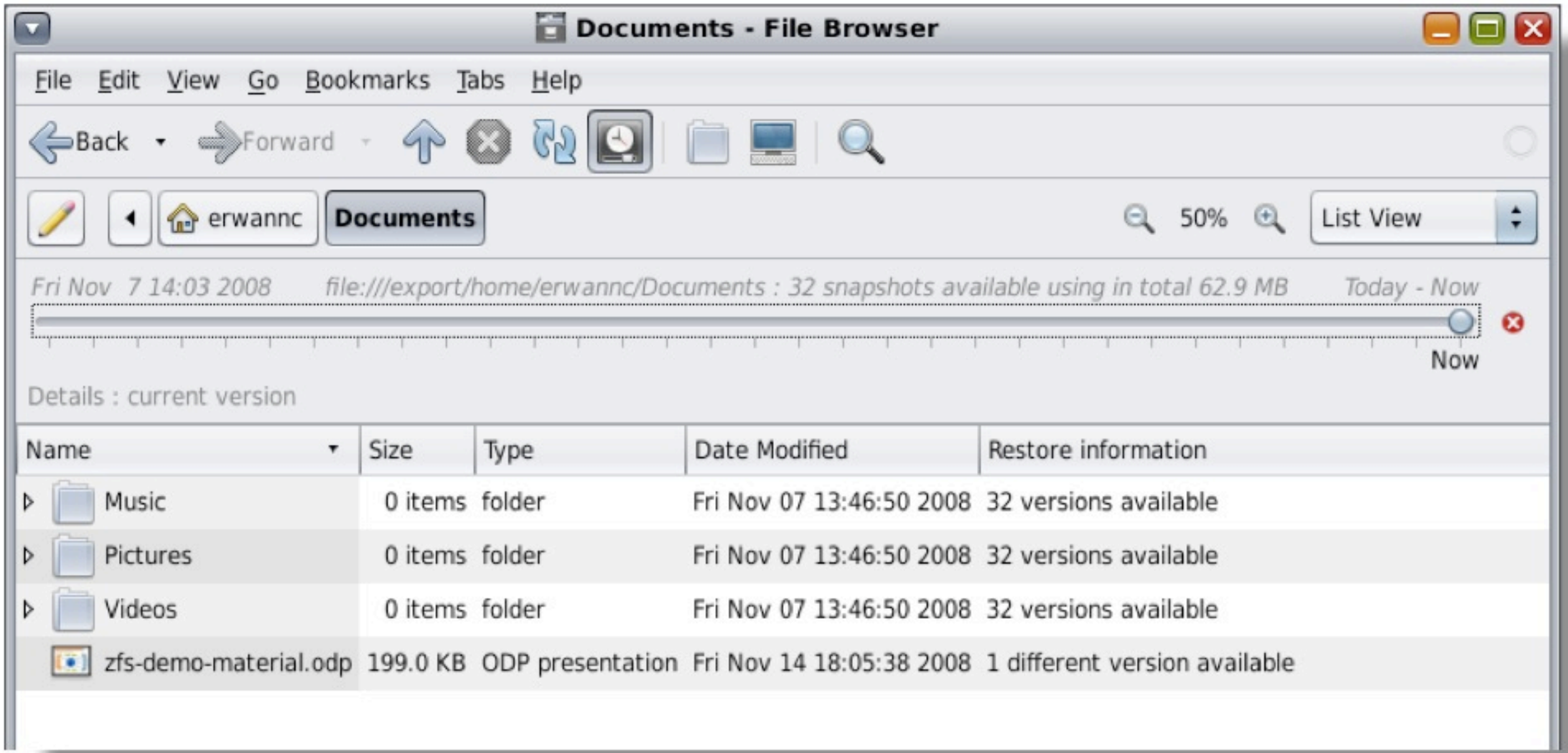
Mal ein Blick in 2008.11

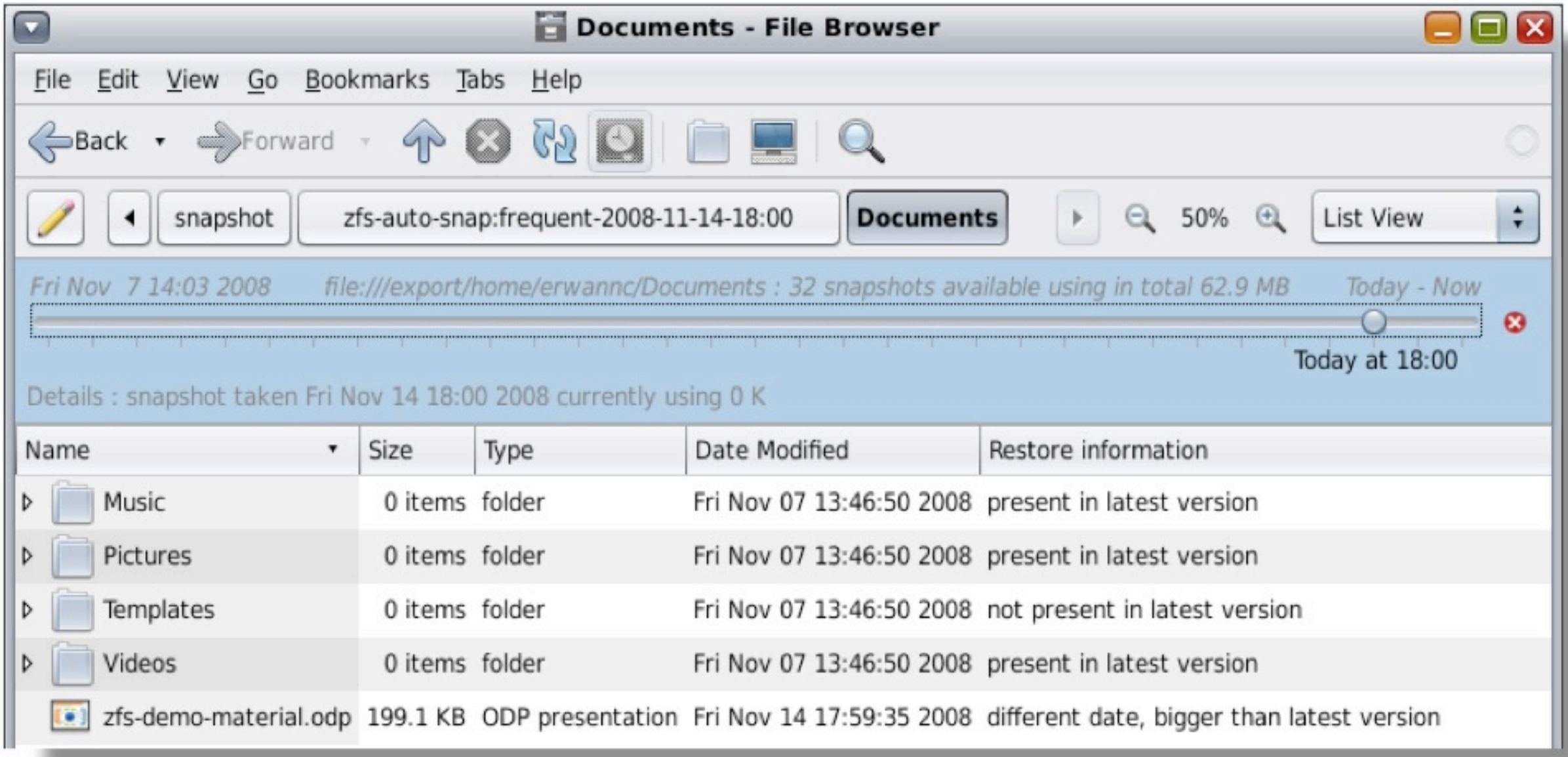
das aktuelle stabile Opensolaris Release
leider erst am 1.12.2008 veröffentlicht ... ;)

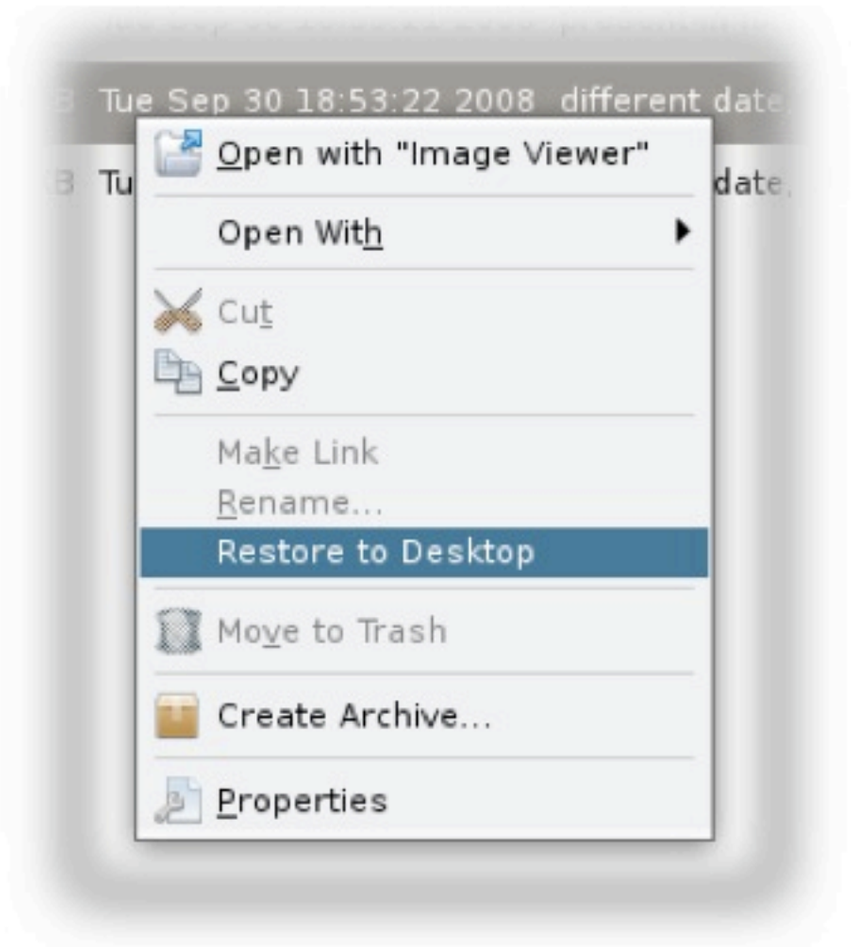
Wirklich sehr nettes Gimmick:
Der Timeslider


Applications
Places
System
Package Manager
Register OpenSolaris
Start Here
Preferences
Administration
Help
About OpenSolaris
About GNOME
Lock Screen
Log Out erwannc...
Shut Down...
Keyring Manager
Network
Package Manager
Print Manager
Register OpenSolaris
Services
Shared Folders
Solaris LP Print Manager
Time and Date
Time Slider Setup
Update Manager
Users and Groups
Time Slider Setup
Enable Time Slider
Time Slider backs up data regularly by taking timed ZFS Snapshots
Advanced Options
File Systems To Back Up
All
Recommended for most users
Custom:
For advanced users
Select
Mount Point
File System Name
/export
rpool/export
/export/home
rpool/export/home
/export/home/erwannc
rpool/export/home/erwannc
/rpool
rpool
Configure the system to take automatic snapshots of your data
Reduce backups when storage space usage exceeds:
80
% of file system capacity
Cancel
OK


erwannc - File Browser
File Edit View Go Bookmarks Tabs Help
Back
Forward
export
home
erwannc
50%
List View

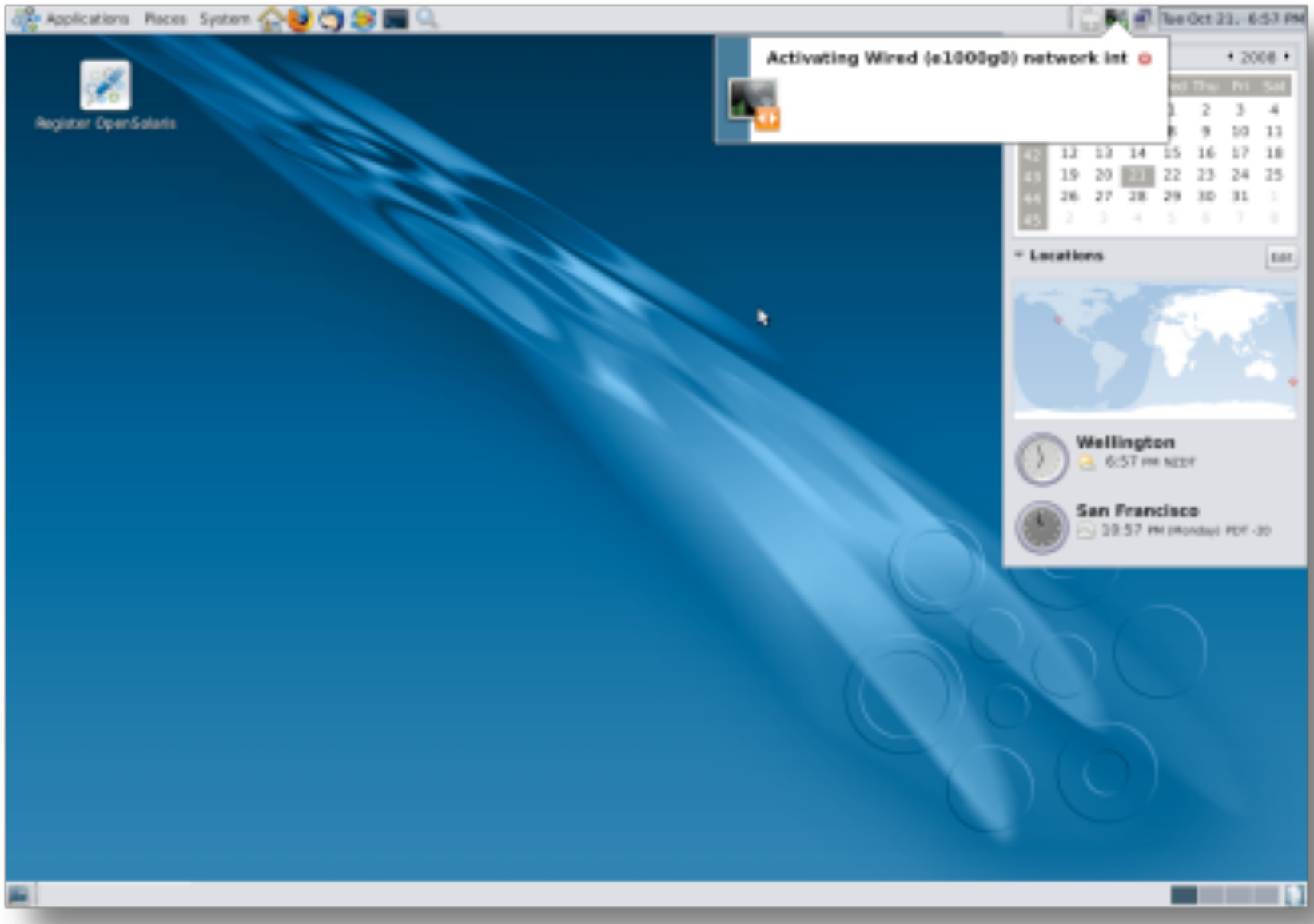

Documents - File Browser
File Edit View Go Bookmarks Tabs Help
Back
Forward
erwannc
Documents
50%
List View
Fri Nov 7 14:03 2008
file:///export/home/erwannc/Documents : 32 snapshots available using in total 62.9 MB
Today - Now
Now
Details : current version
Name
Size
Type
Date Modified
Restore information
Music
0 items
folder
Fri Nov 07 13:46:50 2008
32 versions available
Pictures
0 items
folder
Fri Nov 07 13:46:50 2008
32 versions available
Videos
0 items
folder
Fri Nov 07 13:46:50 2008
32 versions available
zfs-demo-material.odp
199.0 KB
ODP presentation
Fri Nov 14 18:05:38 2008
1 different version available


Documents - File Browser
File Edit View Go Bookmarks Tabs Help
Back
Forward
snapshot
zfs-auto-snap:frequent-2008-11-14-18:00
Documents
50%
List View
Fri Nov 7 14:03 2008
file:///export/home/erwannc/Documents : 32 snapshots available using in total 62.9 MB
Today - Now
Today at 18:00
Details : snapshot taken Fri Nov 14 18:00 2008 currently using 0 K
Name
Size
Type
Date Modified
Restore information
Music
0 items
folder
Fri Nov 07 13:46:50 2008
present in latest version
Pictures
0 items
folder
Fri Nov 07 13:46:50 2008
present in latest version
Templates
0 items
folder
Fri Nov 07 13:46:50 2008
not present in latest version
Videos
0 items
folder
Fri Nov 07 13:46:50 2008
present in latest version
zfs-demo-material.odp
199.1 KB
ODP presentation
Fri Nov 14 17:59:35 2008
different date, bigger than latest version


Tue Sep 30 18:53:22 2008 different date
Open with "Image Viewer"
Open With
Cut
Copy
Make Link
Rename...
Restore to Desktop
Move to Trash
Create Archive...
Properties

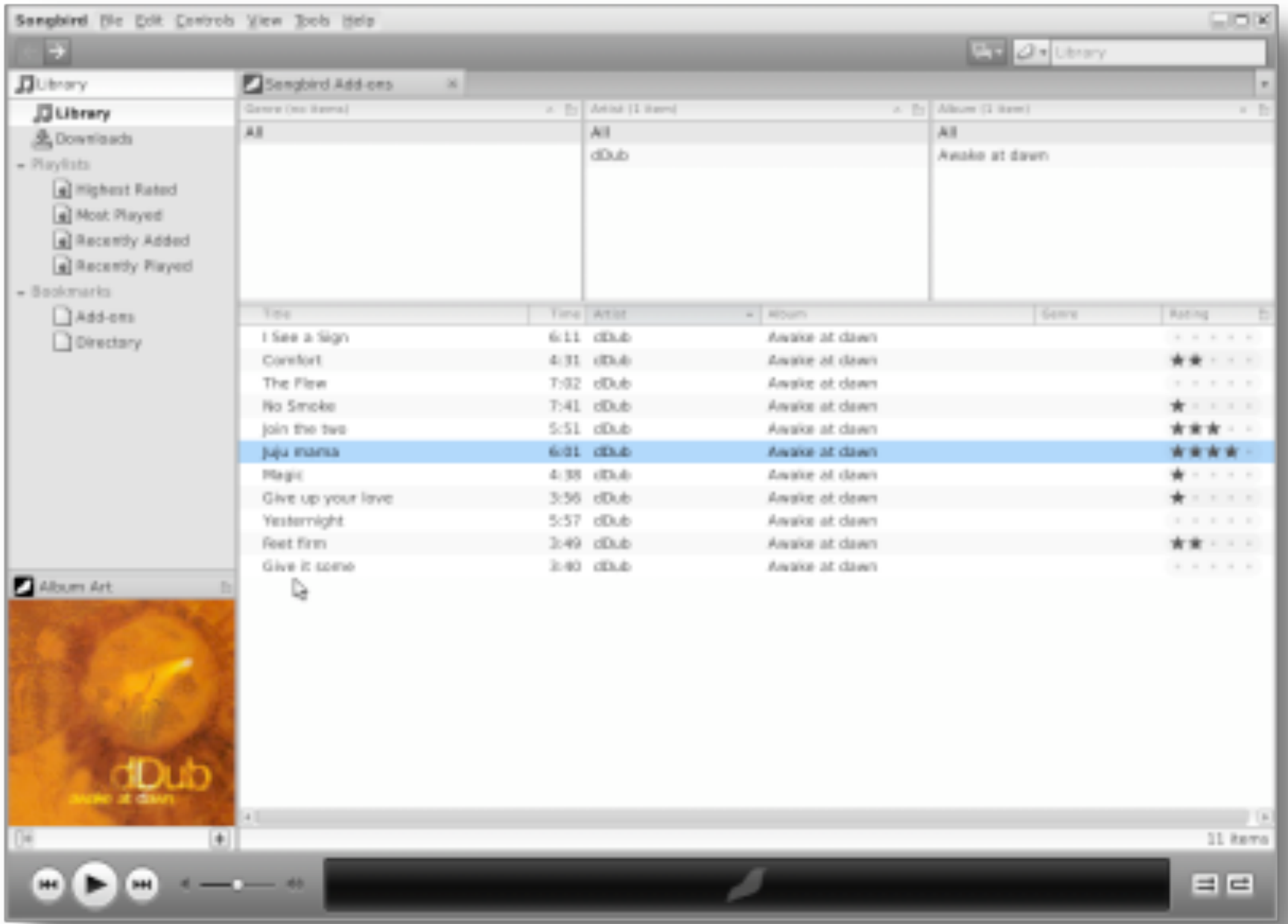
Aber auch an jeder Menge anderer Stellen hat sich eine Menge getan ...

Der Desktop basiert auf Gnome 2.23

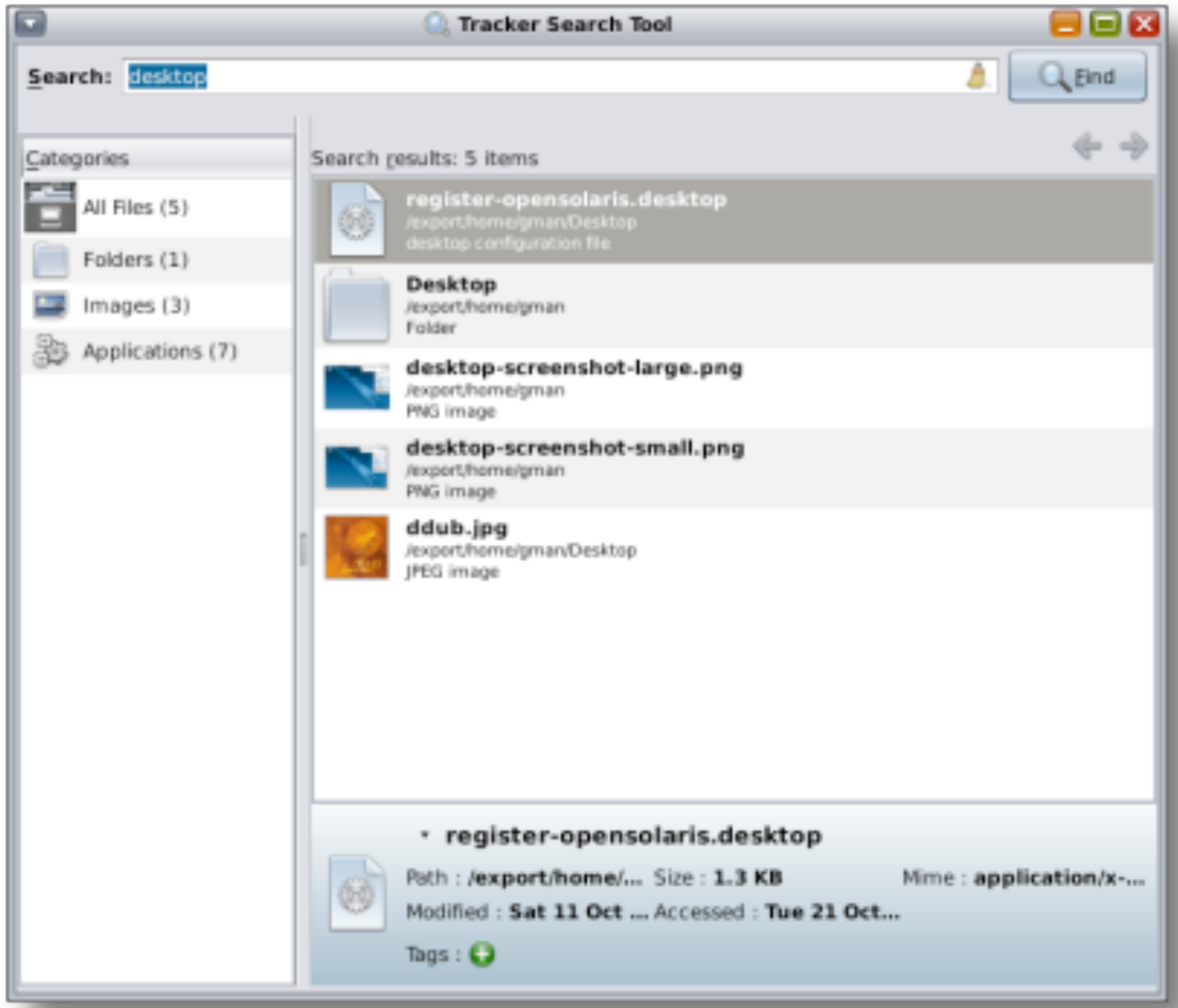


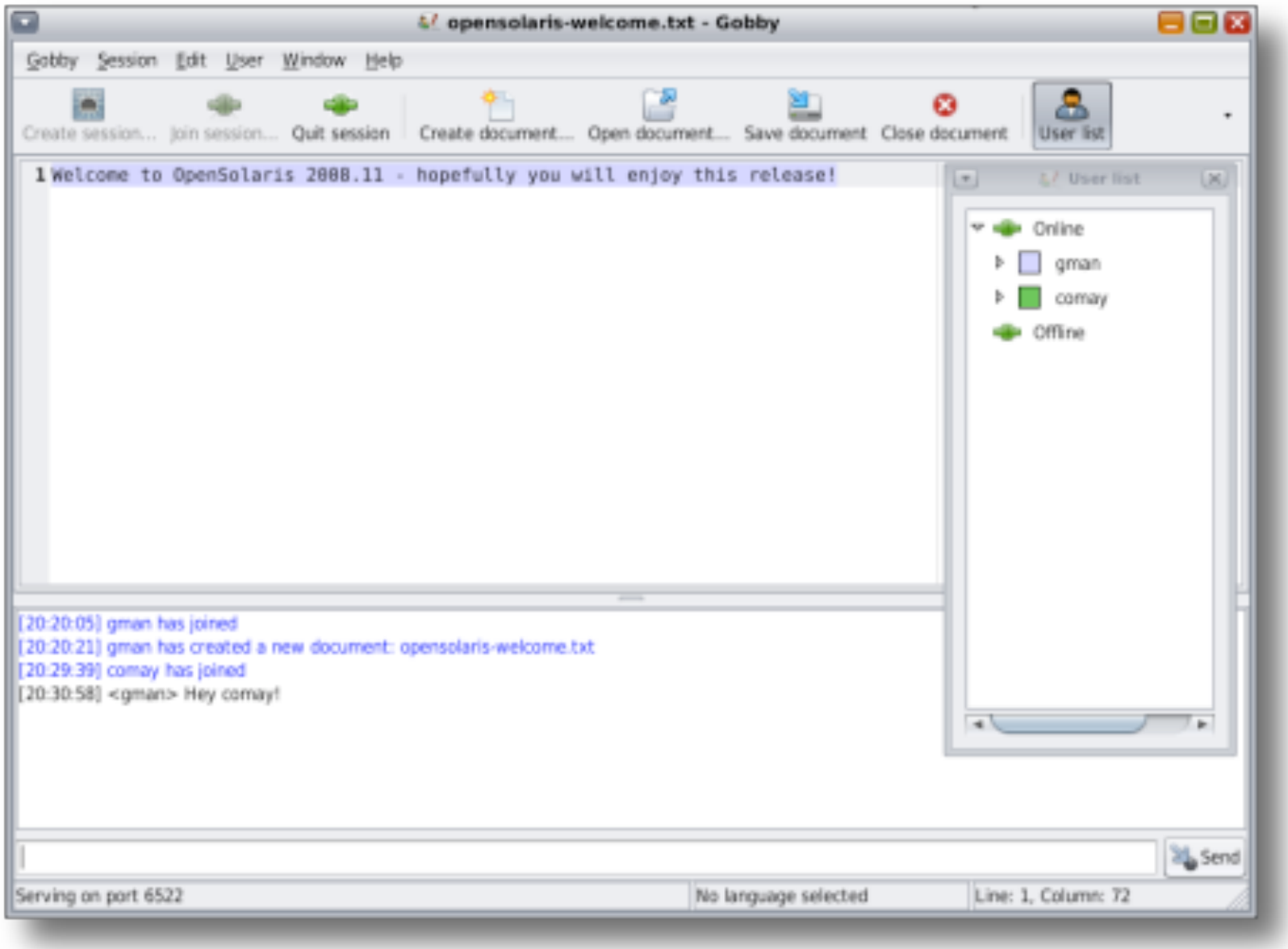
Firefox 3 natürlich...

Der Songbird MP3-Player



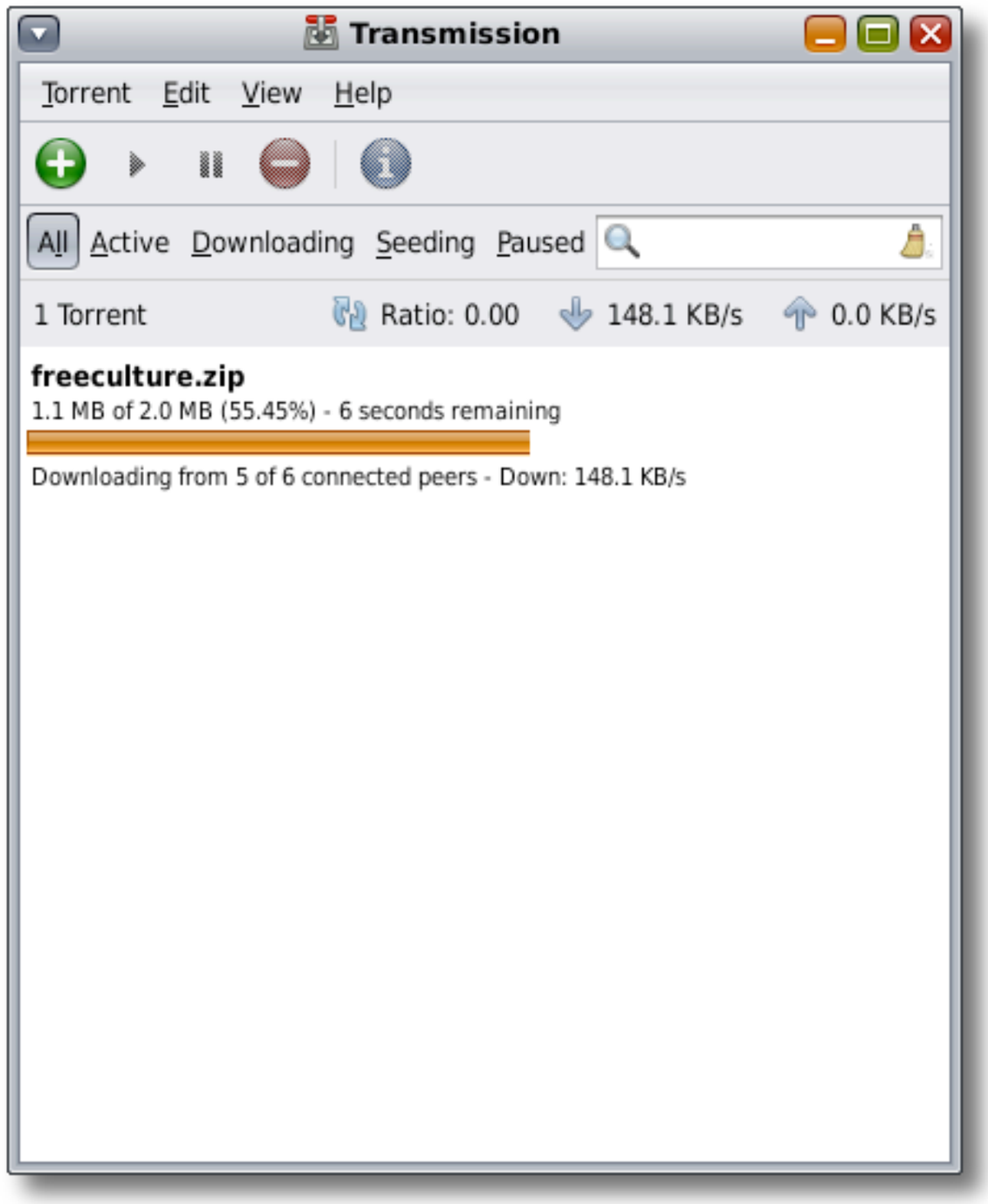
Tracker ist ein Tool zur Suche auf dem Desktop ... inclusive Suche in den eigenen Mails

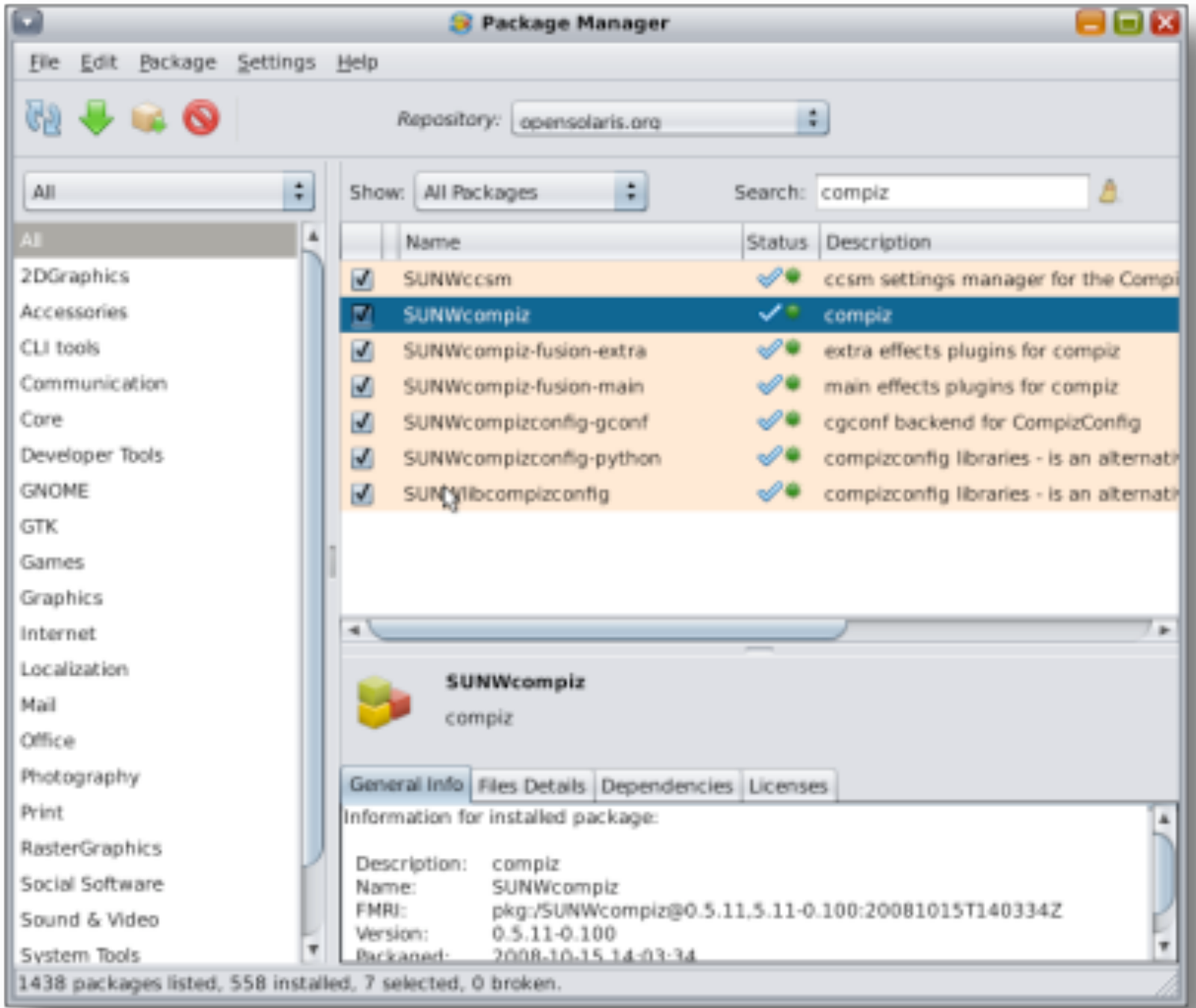


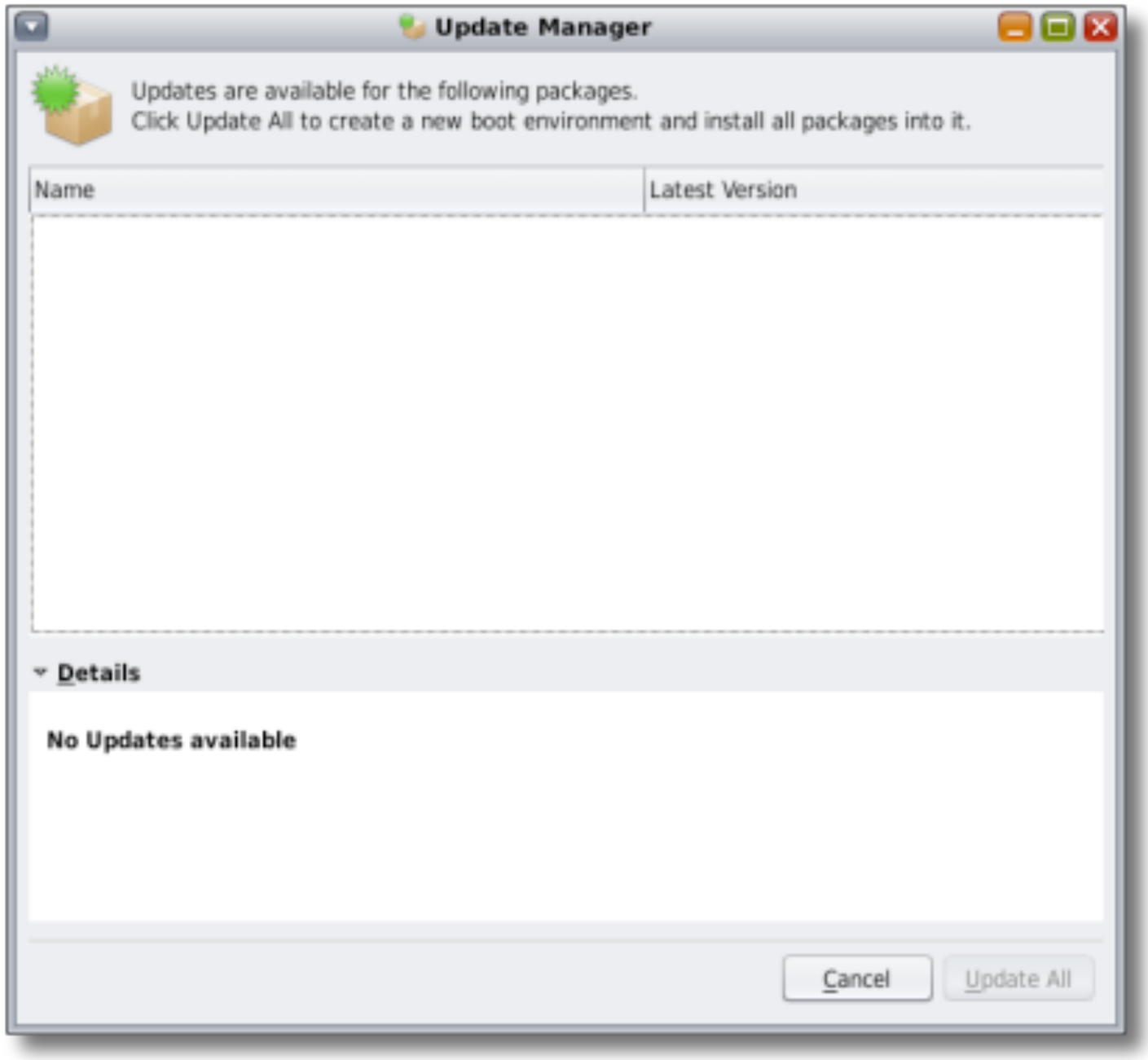


Gobby ist ein kollaborativer Editor mit Chat ... in Opensolaris verfügbar.

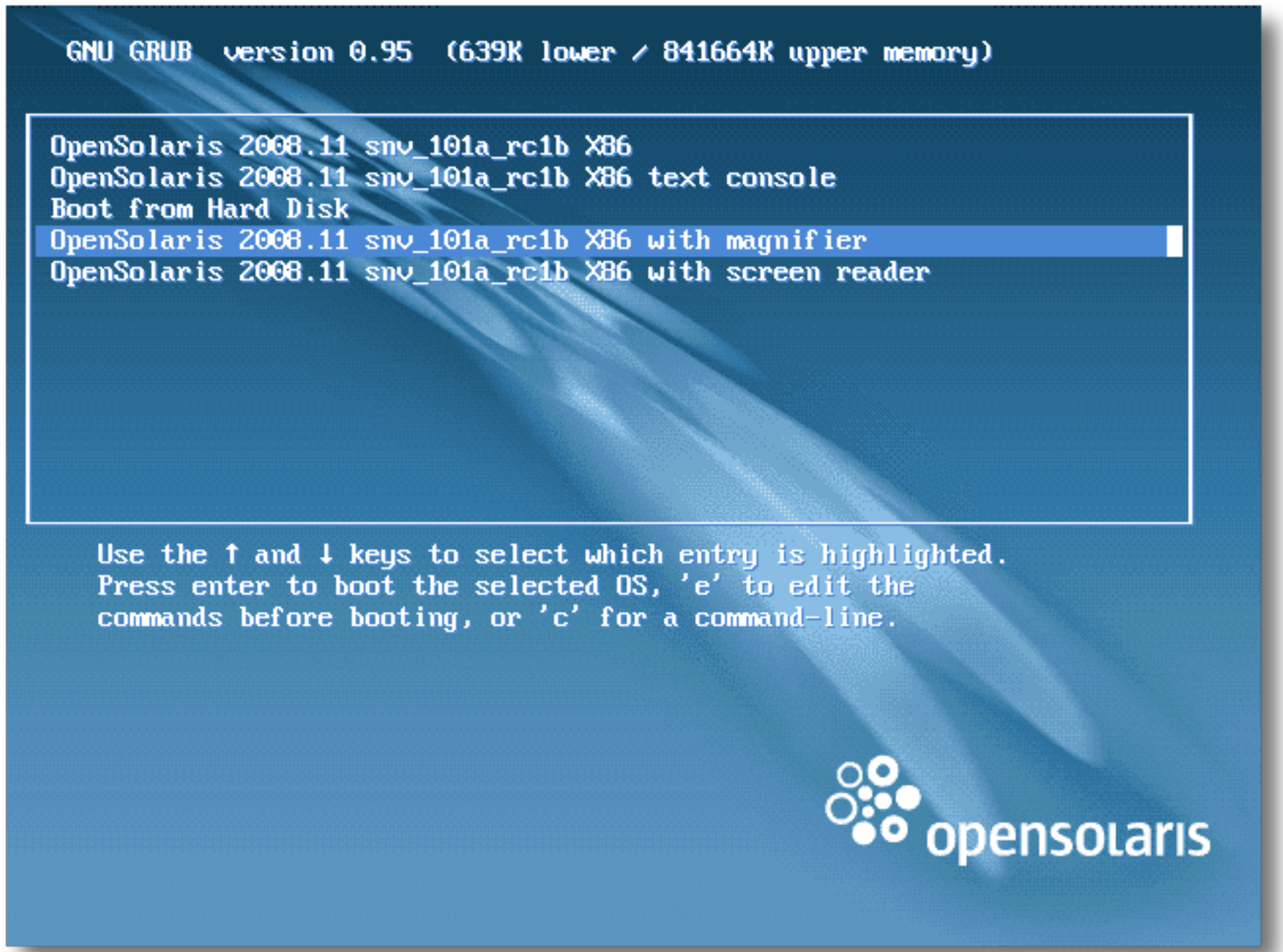
Wenn schon OpenSolaris via Bitorrent verfügbar ist, sollte vielleicht auch Bitorrent in Opensolaris verfügbar sein



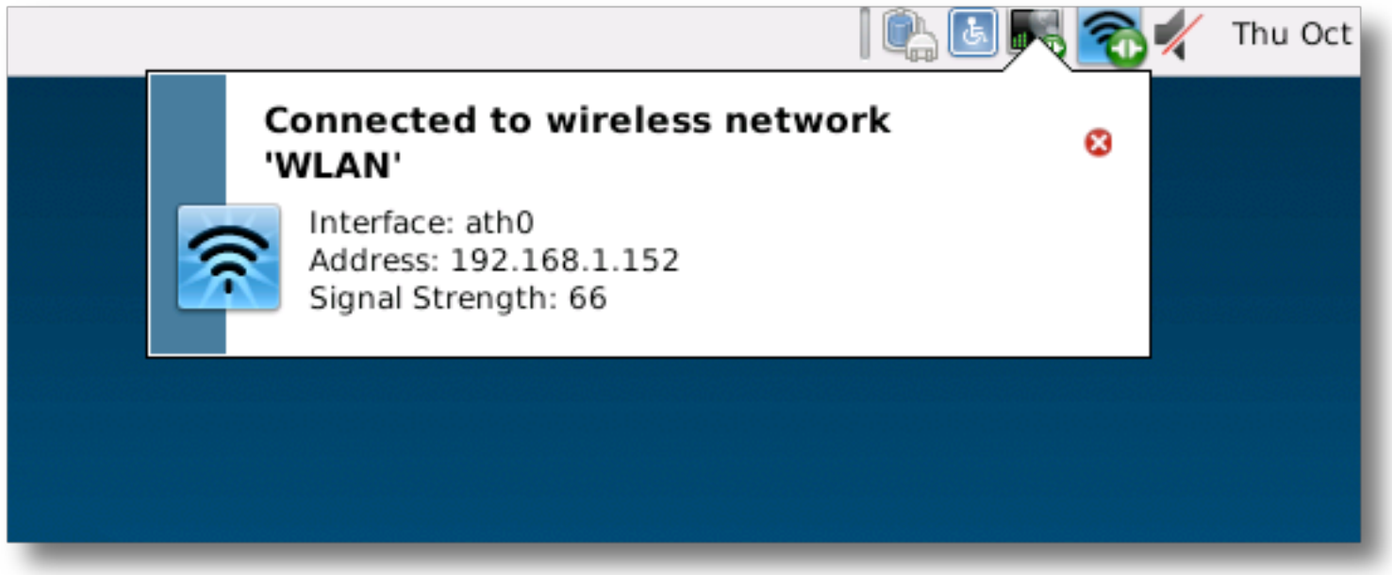

Package Manager
File Edit Package Settings Help
Repository: opensolaris.org
All
Show: All Packages
Search: compiz
All
2DGraphics
Accessories
CLI tools
Communication
Core
Developer Tools
GNOME
GTK
Games
Graphics
Internet
Localization
Mail
Office
Photography
Print
RasterGraphics
Social Software
Sound & Video
System Tools
Name
Status
Description
SUNWccsm
ccsm settings manager for the Compi
SUNWcompiz
compiz
SUNWcompiz-fusion-extra
extra effects plugins for compiz
SUNWcompiz-fusion-main
main effects plugins for compiz
SUNWcompizconfig-gconf
cgconf backend for CompizConfig
SUNWcompizconfig-python
compizconfig libraries - is an alternati
SUNWlibcompizconfig
compizconfig libraries - is an alternati
SUNWcompiz
compiz
General Info
Files Details
Dependencies
Licenses
Information for installed package:
Description: compiz
Name: SUNWcompiz
FMRI: pkg:/SUNWcompiz@0.5.11,5.11-0.100:20081015T140334Z
Version: 0.5.11-0.100
1438 packages listed, 558 installed, 7 selected, 0 broken.


Update Manager
Updates are available for the following packages.
Click Update All to create a new boot environment and install all packages into it.
Name
Latest Version
Details
No Updates available
Cancel
Update All

Der LiveCD Boot von Opensolaris unterstützt auch Hilfen für gehandicapte Personen



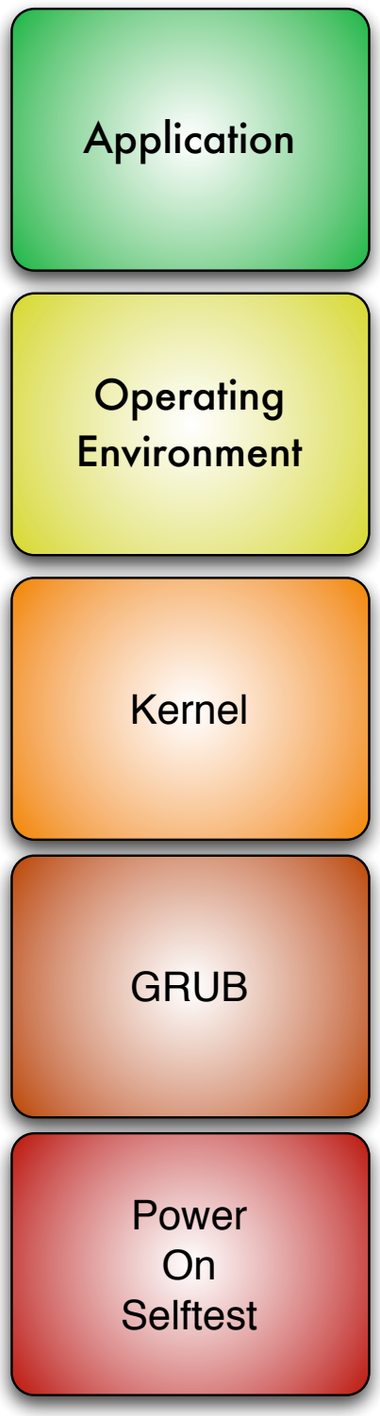
Network Automagic bringt Solaris ein ganzes Stück weiter in Richtung „It just works!“



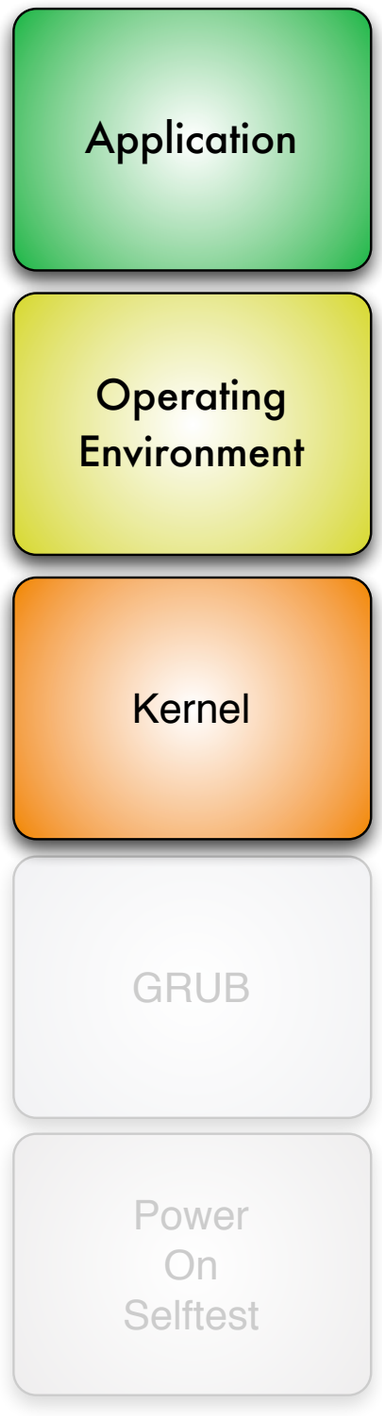
Fast Reboot für OpenSolaris

```
reboot -f
```

First Boot/Normal Reboot



Fast Reboot



Fast Reboot für OpenSolaris

```
reboot -f
```

Nein, wir haben das nicht gemacht,
weil Solaris so häufig rebootet
werden muss ....

Schon mal bei einem System mit 512 GB Hauptspeicher auf den Speichertest am Anfang gewartet ?

Insight 1

Es gibt durchaus die GNU-Tools in Solaris

Die meisten Tools findet Ihr unter `/usr/sfw/bin`

Warum sind sie dann bitte nicht in `$PATH`?

Und warum sind verhalten sich die Solariskommandos anders ?

Das sind ja gleich zwei Fragen auf einmal?

Insight

2

Kompatibilitätsgarantien

Warum nutzt Sun nicht einfach die GNU-Binaries direkt im Pfad?

Es gibt zwei heilige Kühe in Solaris:
Binärkompatibilitätsgarantie
Sourcecodekompatibilitätsgarantie

Sun garantiert:
Ein Programm das auf Solaris 7 lief, läuft auch unter Solaris 10

Ein Programm das auf Solaris x86 kompiliert, kompiliert ohne Änderungen auch auf Solaris SPARC.

(Es gibt ein paar Nebenbedingung, ein in C geschriebener Wrapper um Inline-Assembler hat naturgemäss mit der Sourcecodekompatibilität Probleme)

Das bedeutet aber auch:
Das Verhalten von Scripten darf sich nicht ändern.
Das Verhalten von Shell-Komandos darf sich nicht ändern.

auch Scripte sind irgendwie Binaries ;)

Die Kommandozeilenparameter dürfen sich nicht ändern.
Die Ausgabe darf sich nicht ändern.

Die Programmierinterfaces werden stabil gehalten:

Viele Treiber für Solaris 7 laufen auch unter Solaris 10

Wenn nicht, fragt der Treiber meistens ab, unter welcher Version er läuft ...

Versucht mal einen Treiber von einem Linux 2.0 Kernel unter Linux 2.6 zu laden ;)

Insight

3

Standards

Sun hält sich an Standards ...

Jahaa ... es gibt welche in Unix!

standards(5)

Unter anderem ist auch die Bedeutung von Kommandozeilenparametern teil dieser Standards.

Teilweise sind diese der BSD-Ahnenreihe entlehnt.

andere kommen aus der SysV-Historie

GNU Tools sind eine wilde Mischung aus beidem!

Die Syntax in GNU hält sich an keinen Standard.
Ausser seinen eigenen ....

Zynische Seitenbemerkung - Das ist wie mit OSS Lizenzen
Wenn die GPL nicht mit einer anderen Lizenz zusammenspielt, dann ist es die Schuld der anderen Lizenz, auch wenn es Dutzende andere gibt, die kompatibel sind.

Wenn die Syntax eines Kommandos anders ist, ist es die Schuld des nicht-GNU-Kommandos.

Leider kommen die GNU-Tools damit durch, weil viele Leute die Tools kennen und nutzen wollen.

Insight

4

ZFS

Solaris hat ein extrem hochentwickeltes Filesystem!

Ich sollte vielleicht mal sagen, das es einige Dinge nicht kann!

Es kann nicht:

- Volumemanagement
- Filesystemchecks
- Logging/Journaling
- RAID5/RAID6

Es kann dafür:

- Storage virtualisieren
- immer konsistent bleiben
- single/double Parity RAID ohne Write Hole
- die Datenvalidität garantieren.

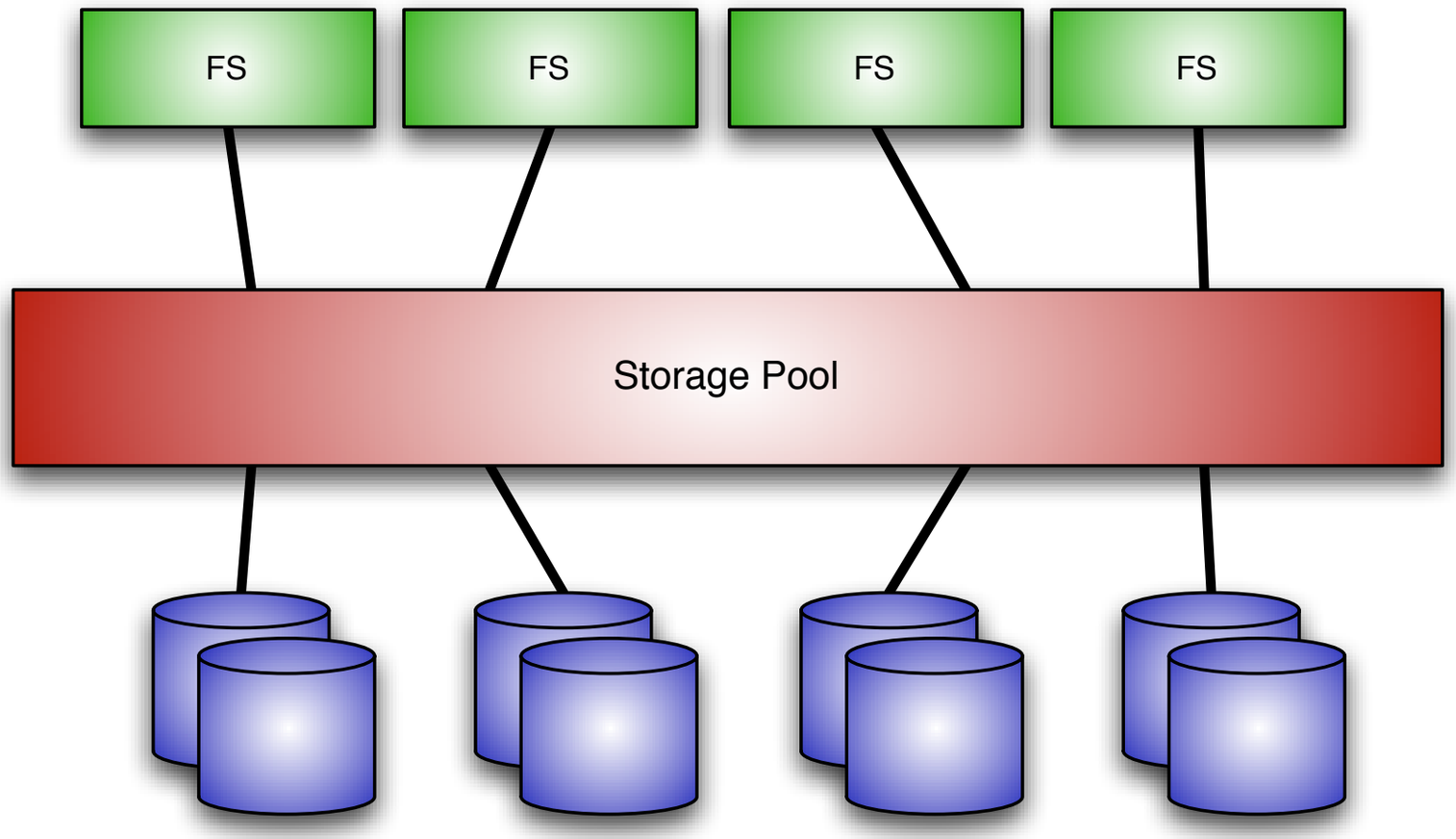
ZFS ist ...

seeehr gross.

2^128 Bit

- Das sind 256 Quadrillionen Zetabyte
- Das interessante Bit ist das 65ste Bit
- Ein Speicher dieser Grösse kann mit Erdmaterie nicht gebaut werden.

... anders

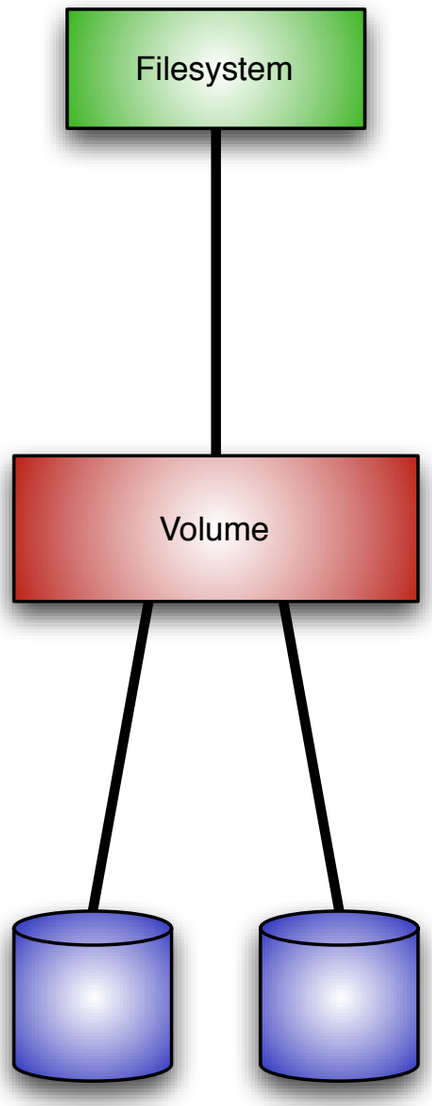


- Abstraktion: wie beim malloc für den Arbeitsspeicher
- Volumes gibt es nicht mehr
- automatisches Vergroessern/Verkleinern
- Jedem Filesystem steht die Bandbreite und die IOPS aller Platten zur Verfügung.
- Ein gemeinsamer Storagepool

Wo ist nun der Unterschied?

Block Device Interface

- „Schreib diesen Block, dann diesen, ...“
- Stromausfall = die Konsistenz auf der Disk geht verloren
- Workaround: Journaling, das ist aber komplex und langsam



Block Device Interface

- Jeden Block sofort auf beide Platten schreiben, um die Mirror synchron zu halten.
- Stromausfall = vollständiger Rsync notwendig
- ziemlich Langsam

Object Based Transactions

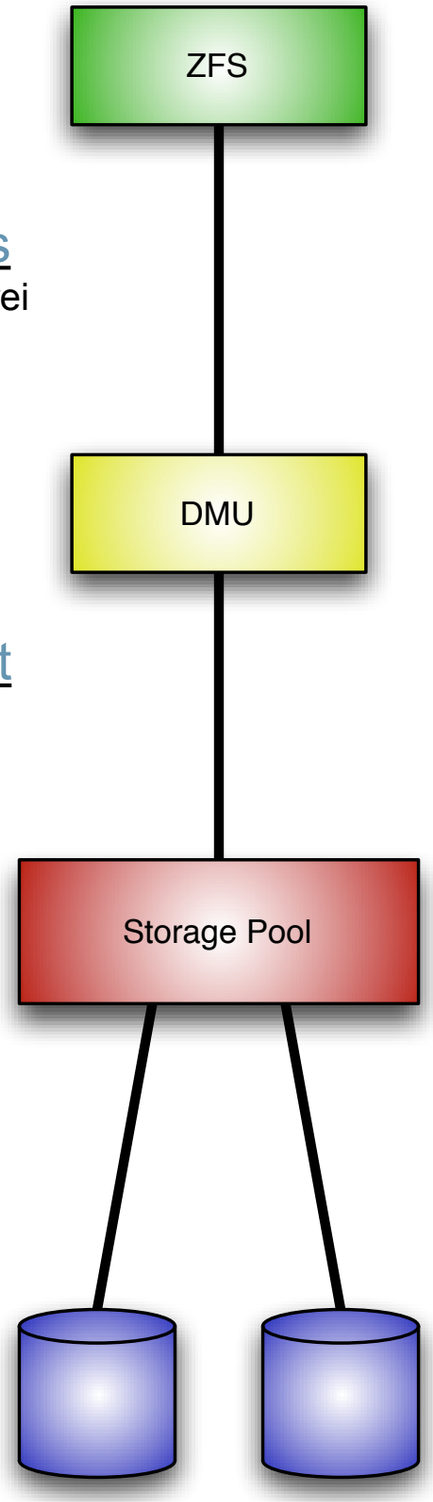
- „Ändere diese 7 Dinge an diesen drei Dateien“
- Alles oder nichts.

Transaction Group Commit

- Alles oder nichts.
- Zustand daher immer konsistent

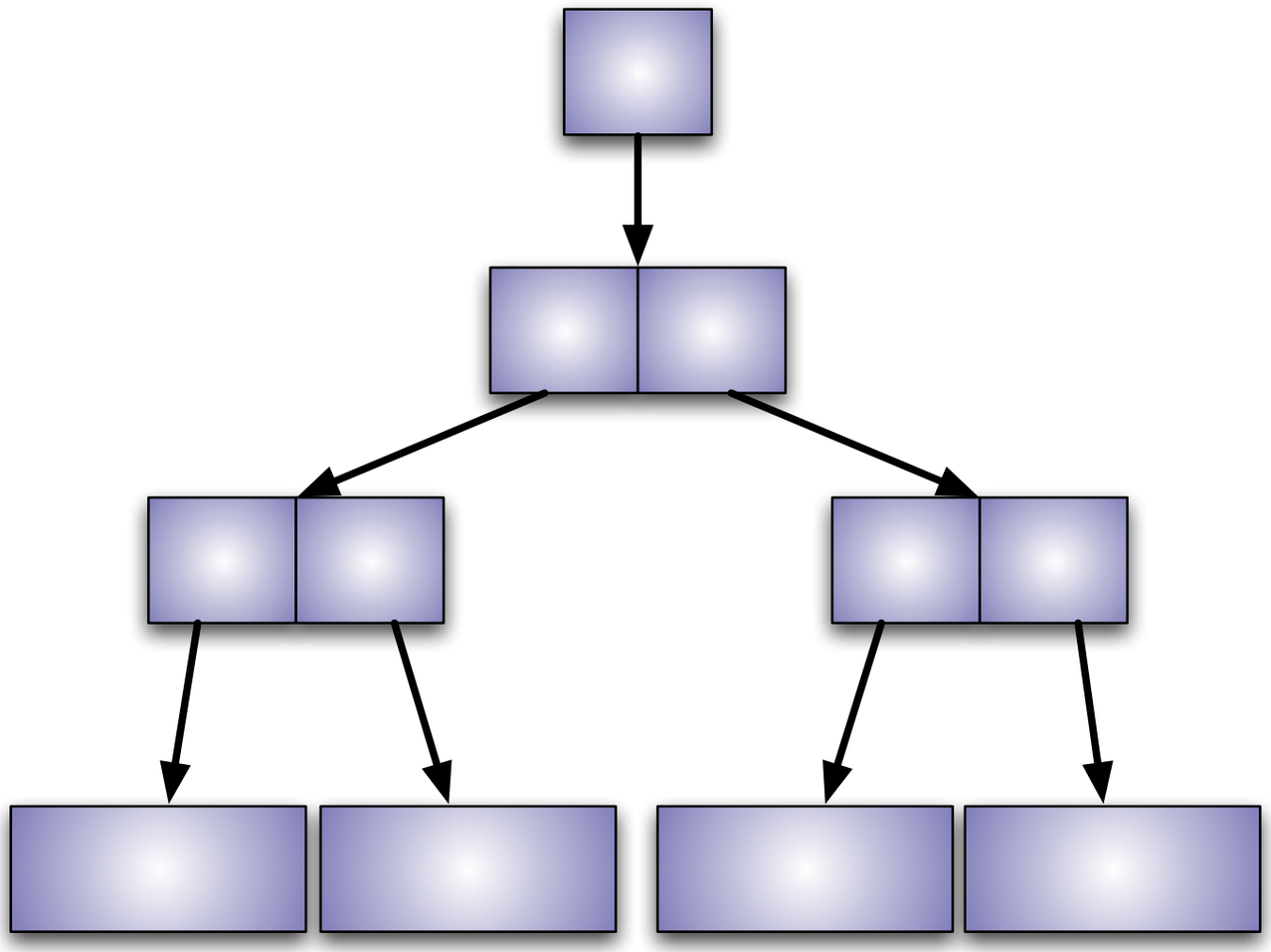
TX Group batch I/O

- Daten werden im Batch auf Platte geschrieben.
- Keine Repositionierung



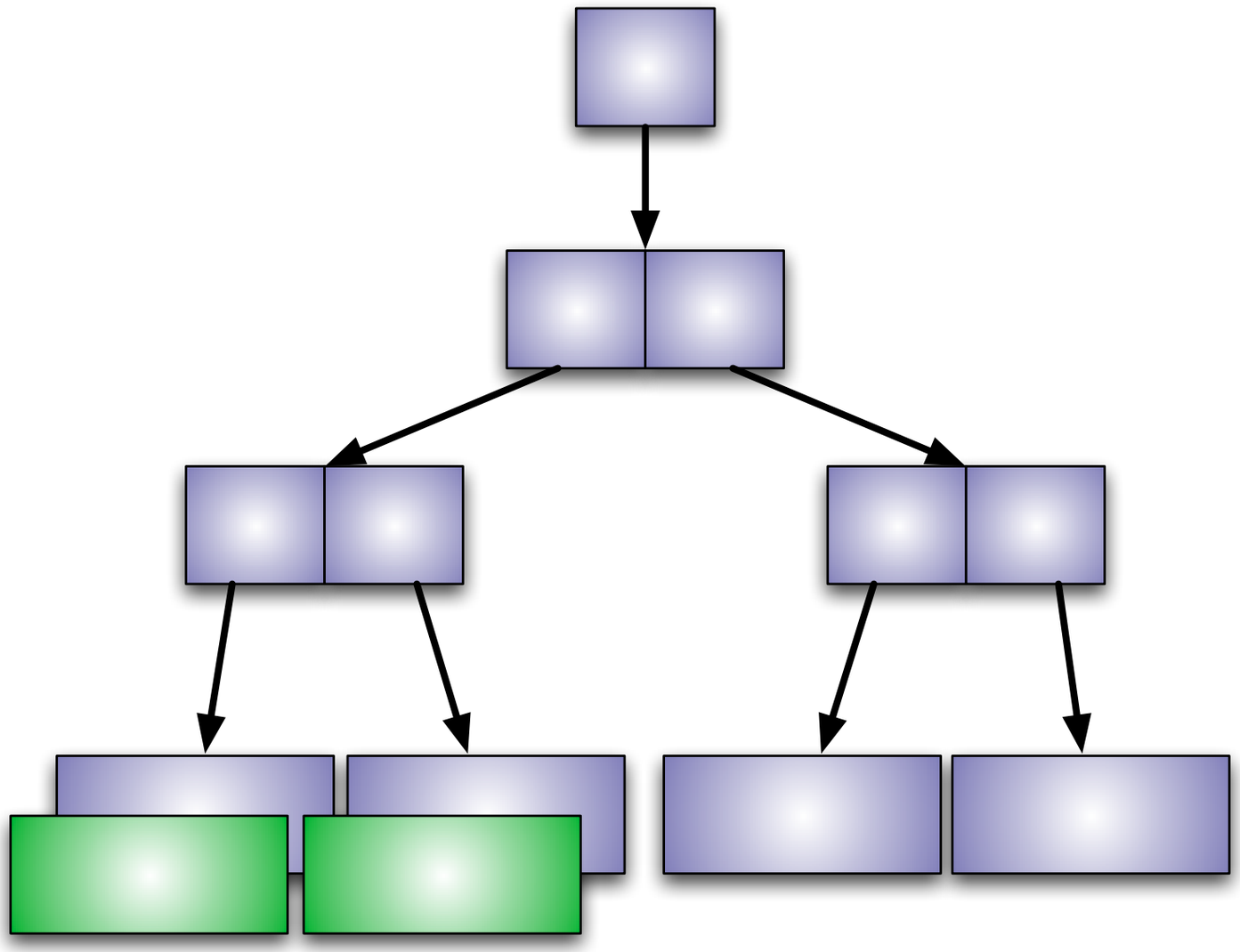
Copy on Write

So sieht das Filesystem vor der Änderung aus.



Copy on Write

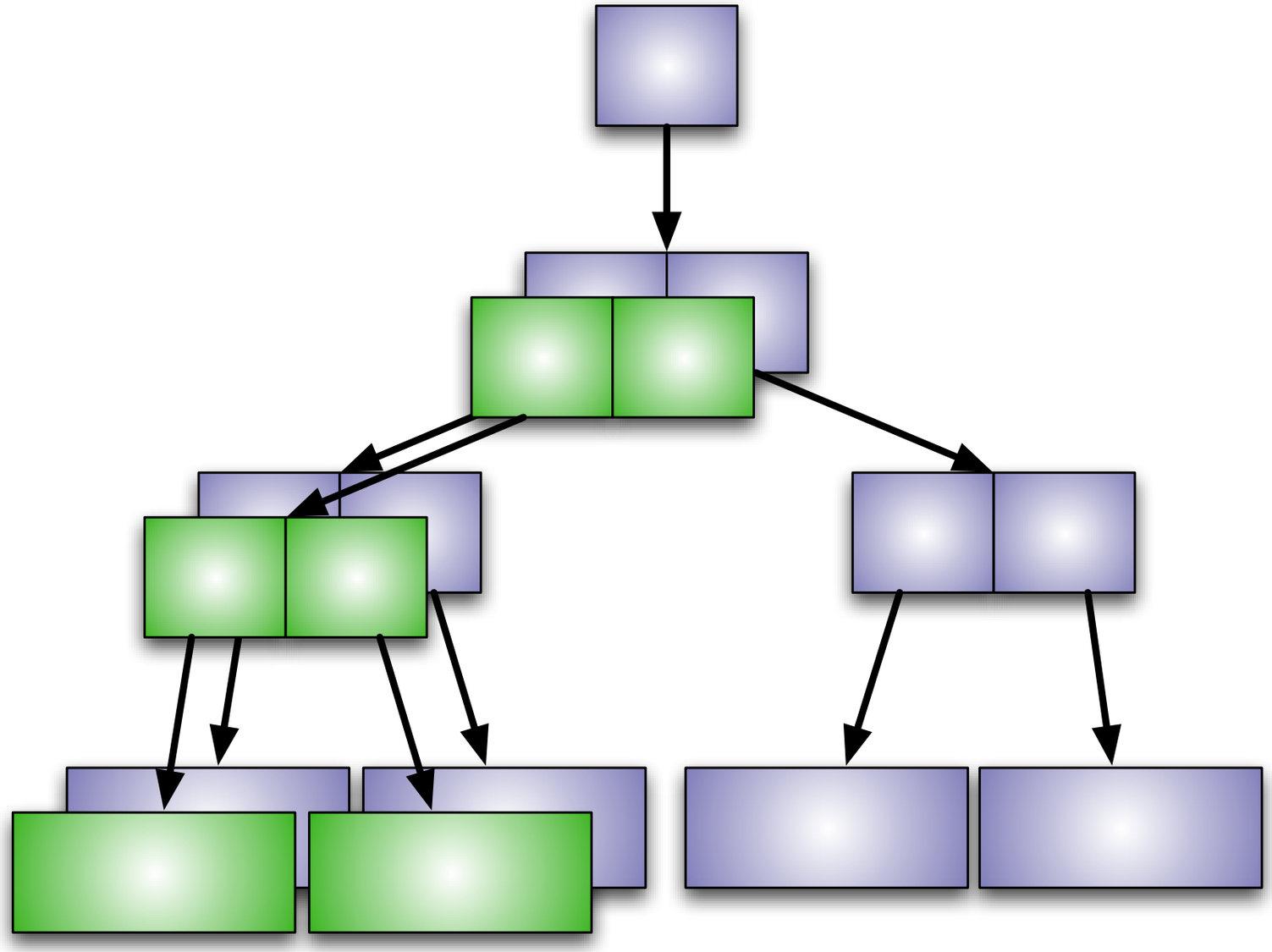
Die neuen Daten werden an eine neue Position geschrieben ...



User und Applikationen sehen den alten Zustand

Copy on Write

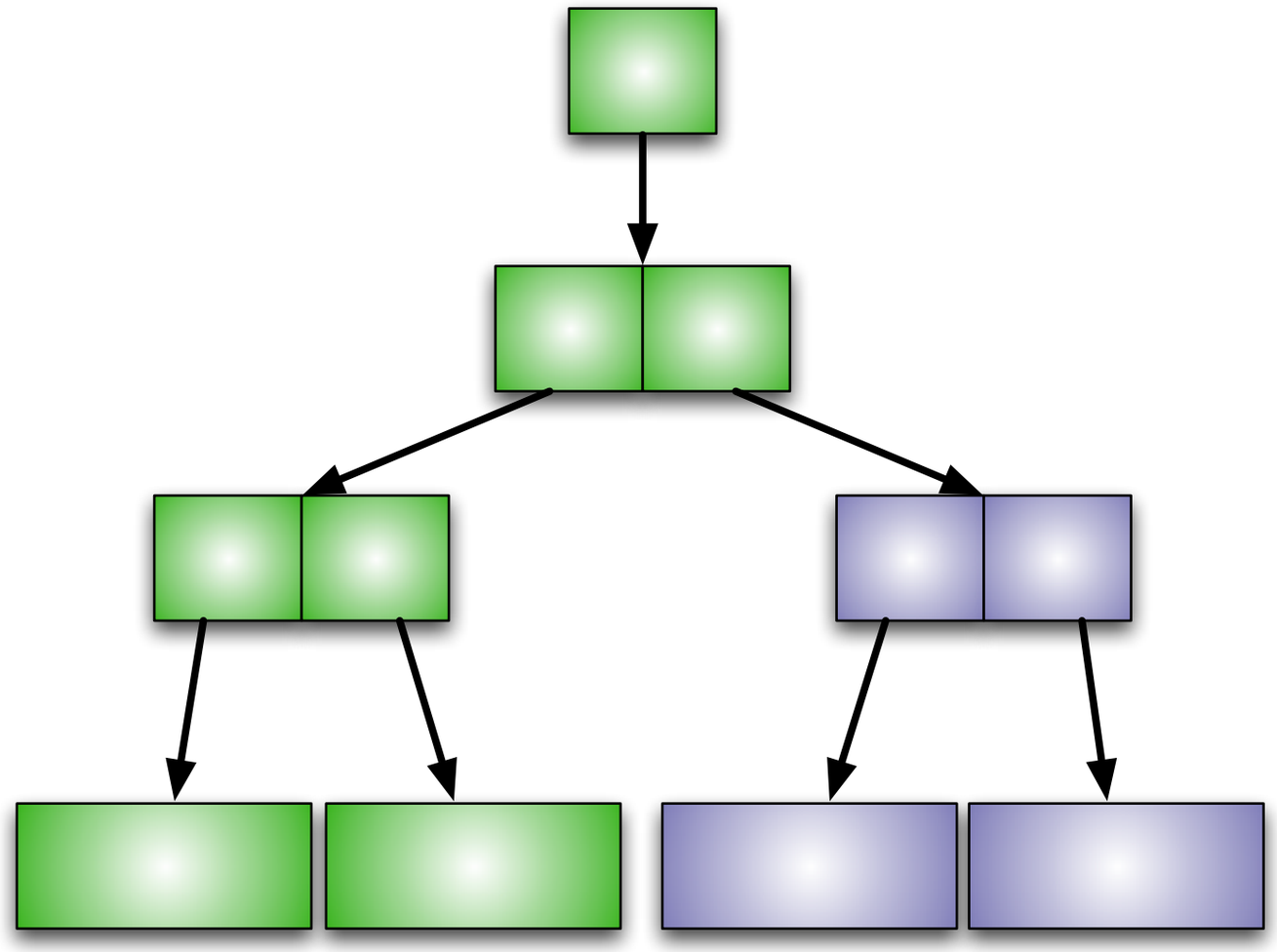
Copy on Write für die indirekten Blöcke. Auch diese werden an neue Positionen geschrieben



User und Applikationen sehen den alten Zustand

Copy on Write

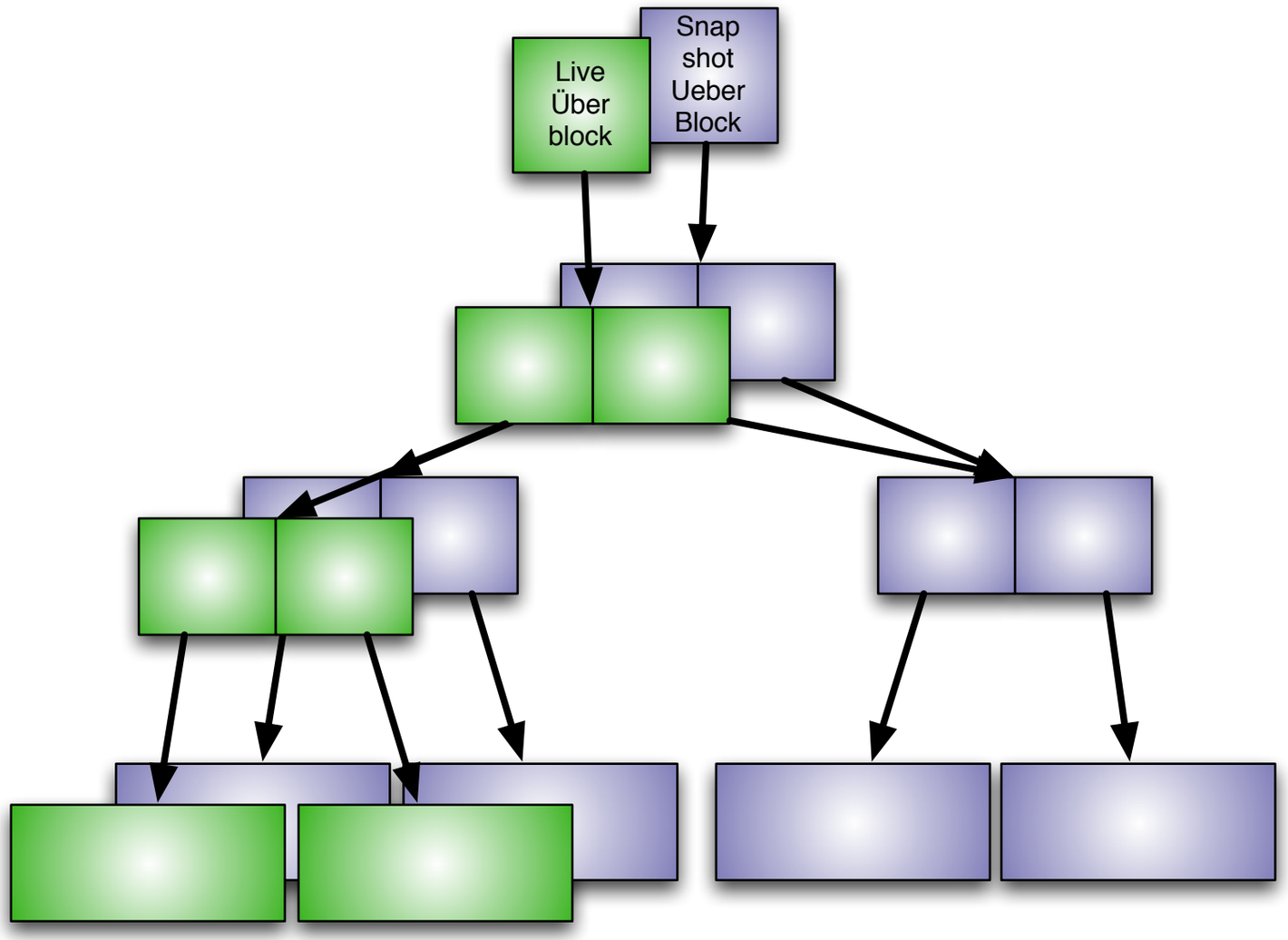
Schreiben des Überblocks in einer einzigen Schreiboperation



User und Applikationen sehen jetzt den neuen Zustand

Angenehmer Nebeneffekt: Snapshots for free

Nach Beenden der Transaktion werden die „cow-ten“ Blöcke nicht freigegeben.

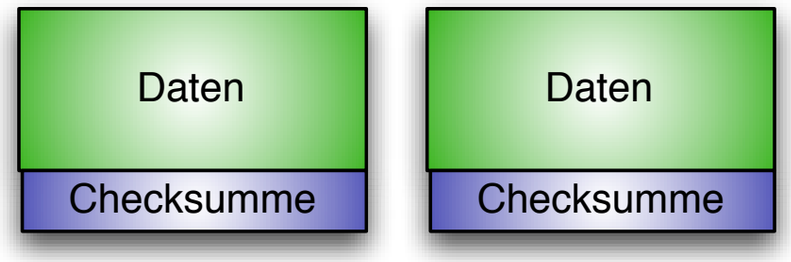


(genau genommen spart das Erstellen eines Snapshots sogar noch Schreiboperationen)

Was ist das besondere an den ZFS Checksummen?

Bei den Markbegleitern:

- Checksumme wird mit dem Block geschrieben
- Jeder in sich konsistente Block wird als korrekt festgestellt...
- ... auch wenn er an der falschen Stelle steht ...



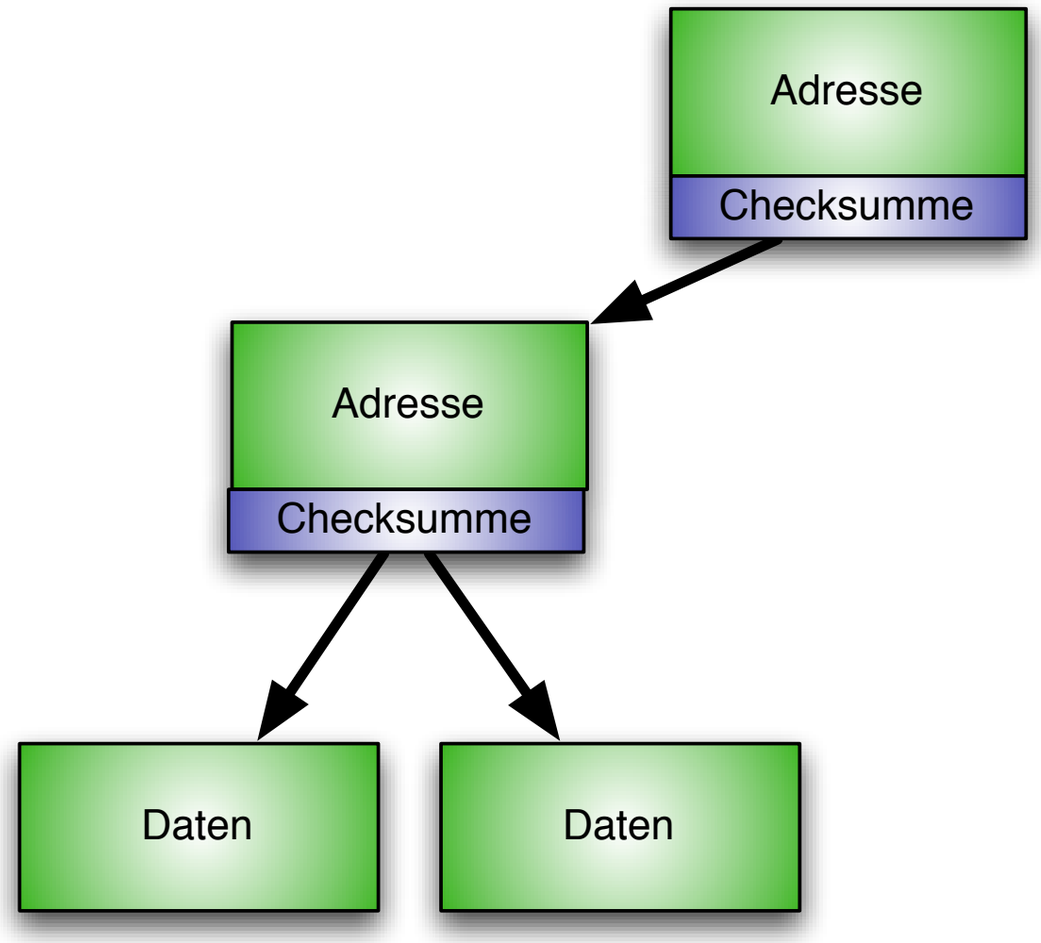
Was erkennt diese Lösung:

Bit Rot
Phantom Writes
Misdirected Reads/Writes
DMA Parity Errors
Driver Errors
Versehentliches Überschreiben

Was ist das besondere an den ZFS Checksummen?

ZFS Data Authentication

- Checksumme wird beim Parent Block gespeichert.
- Fehlerisolation zwischen Checksumme und Daten
- Die Validität des gesamten Baums kann überprüft werden



Was erkennt diese Lösung:

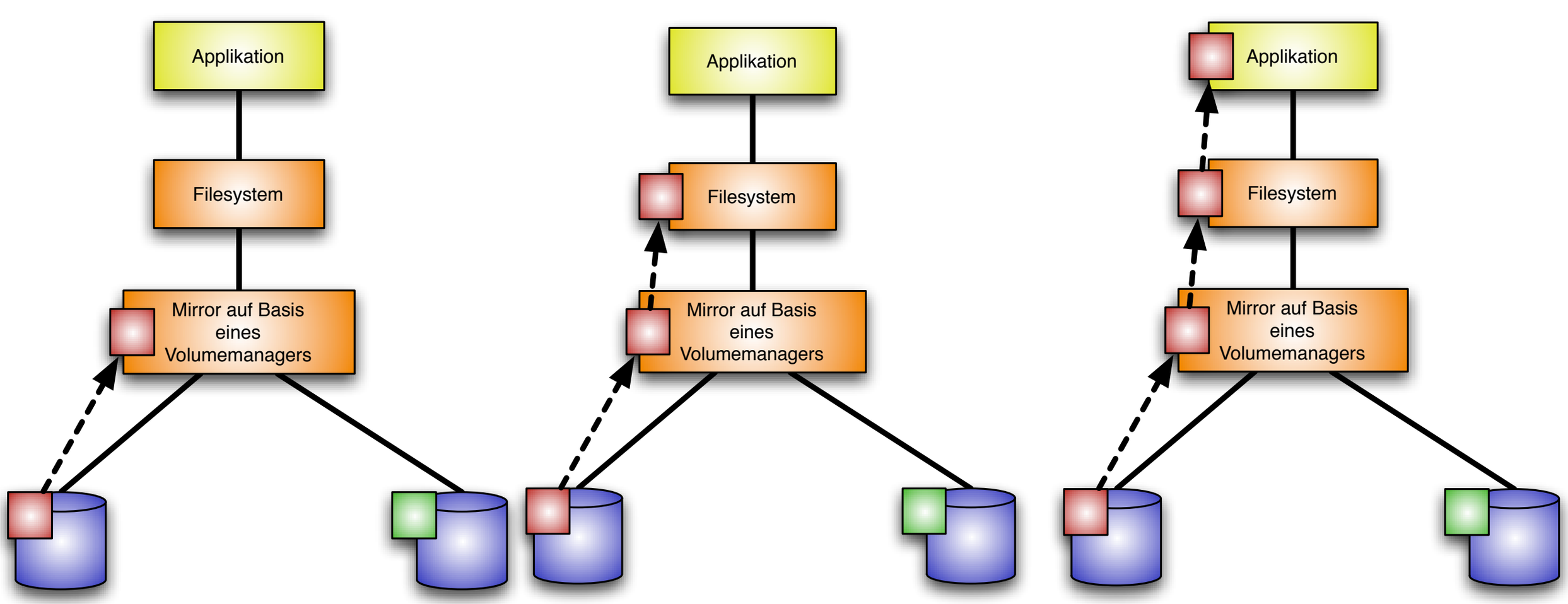
Bit Rot
Phantom Writes
Misdirected Reads/Writes
DMA Parity Errors
Driver Errors
Versehentliches Überschreiben

Chronologie einer Dateninkonsistenz

Applikation will Daten lesen.
Angenommen die Mirrorhälfte ist „schleichend“ defekt. Der VM liest falsche Daten, kann das aber nicht erkennen.

Der Volumemanager gibt die Daten ans Filesystem weiter.
Ist der Fehler in den Metadaten panict das System, ansonsten ...

... bekommt die Applikation falsche Daten

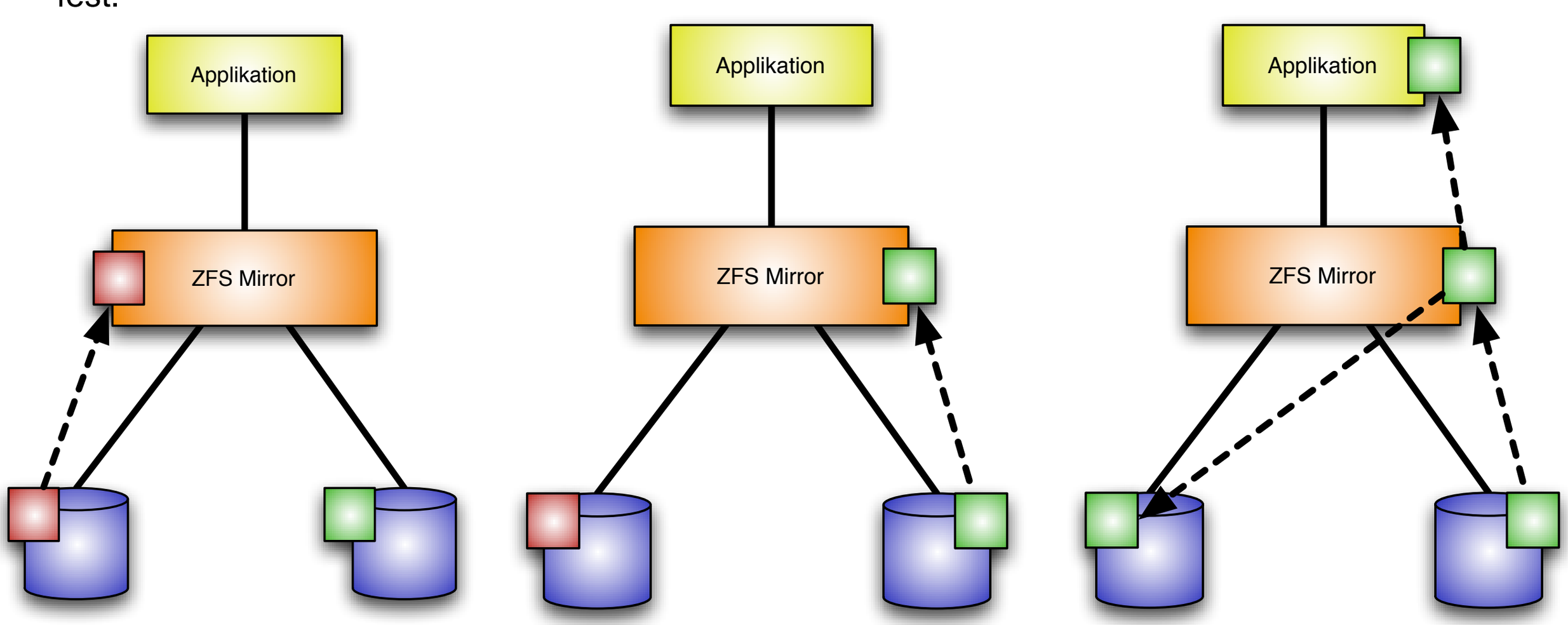


Was passiert nun bei ZFS ?

Applikation will Daten lesen. Angenommen die Mirrorhälfte ist „schleichend“ defekt. Der VM liest falsche Daten, überprüft die Checksume und stellt den Fehler fest.

ZFS versucht es auf dem zweiten Spiegel. Checksumme stimmt.

ZFS liefert die richtigen Daten aus und korrigiert die Daten auf der defekten Spiegelhälfte (oder schreibt sie wo anders hin)



Die eigentliche Herausforderung ?

Was passiert wenn der Strom ausfällt, wenn die Daten schon geschrieben sind, die Parity aber noch nicht?

Richtig ...


^
^
^
^
= garbage

Der Effekt nennt sich „write hole“

Workaround: batteriegebufferter NVRAM (€€€)

Was macht ZFS hier anders?

RAID-Z und RAID-Z2

- Dynamische Stripe-Size
  - Jeder logische Block ist sein eigener Stripe
- Dadurch ist jeder Write ein Full-Stripe-Write
  - Kein Lesen-Ändern-Schreiben
  - Kennt kein Write-Hole (... und braucht daher kein NVRAM)

ZFS ist trivial in der Administration

Beispiel:
Zwei Devices spiegeln, Filesystem drauf erstellen und mounten?

```
# zpool create tp mirror /testfile1 /testfil

# mount
[..]
/tp on tp read/write/setuid/devices/nonbmand/exec/xattr/atime/dev=2d90013 on Fri Mar 20
12:58:24 2009
```

Mehr nicht ...

Beispiel:
In diesem Pool weitere Filesystem anlegen ...

```
# zfs create tp/statler
# zfs create tp/gonzo
# zfs create tp/waldorf
# zfs create tp/kermit

# mount
[..]
/tp/statler on tp/statler read/write/setuid/devices/nonbmand/exec/xattr/atime/dev=2d90017 on
Fri Mar 20 12:59:34 2009
/tp/gonzo on tp/gonzo read/write/setuid/devices/nonbmand/exec/xattr/atime/dev=2d90018 on Fri
Mar 20 12:59:37 2009
/tp/waldorf on tp/waldorf read/write/setuid/devices/nonbmand/exec/xattr/atime/dev=2d90019 on
Fri Mar 20 12:59:40 2009
/tp/kermit on tp/kermit read/write/setuid/devices/nonbmand/exec/xattr/atime/dev=2d9001a on
Fri Mar 20 12:59:42 2009
```

Mehr nicht ...

BTW:
Durch das Poolkonzept ergibt sich ...

```
# zfs list | grep "tp/"
tp/gonzo                 18K  90.8M    18K  /tp/gonzo
tp/kermit                18K  90.8M    18K  /tp/kermit
tp/statler               18K  90.8M    18K  /tp/statler
tp/waldorf               18K  90.8M    18K  /tp/waldorf
```

Neat ...

Beispiel:
Snapshot erstellen ?

```
# zfs snapshot tp/gonzo@tuesdayevening
```

Beispiel:
Auf den Snapshot zugreifen?

```
# cd .zfs
# cd snapshot
# cd tuesdayevening/
# ls -l
total 40979
-rw-r--r--   1 root     root           0 Mar 20 13:10 monday
-rw------T   1 root     root     20971520 Mar 20 13:05 testfile
-rw-r--r--   1 root     root           0 Mar 20 13:10 tuesday
# cp monday /tp/gonzo/monday
```

Mehr nicht ...

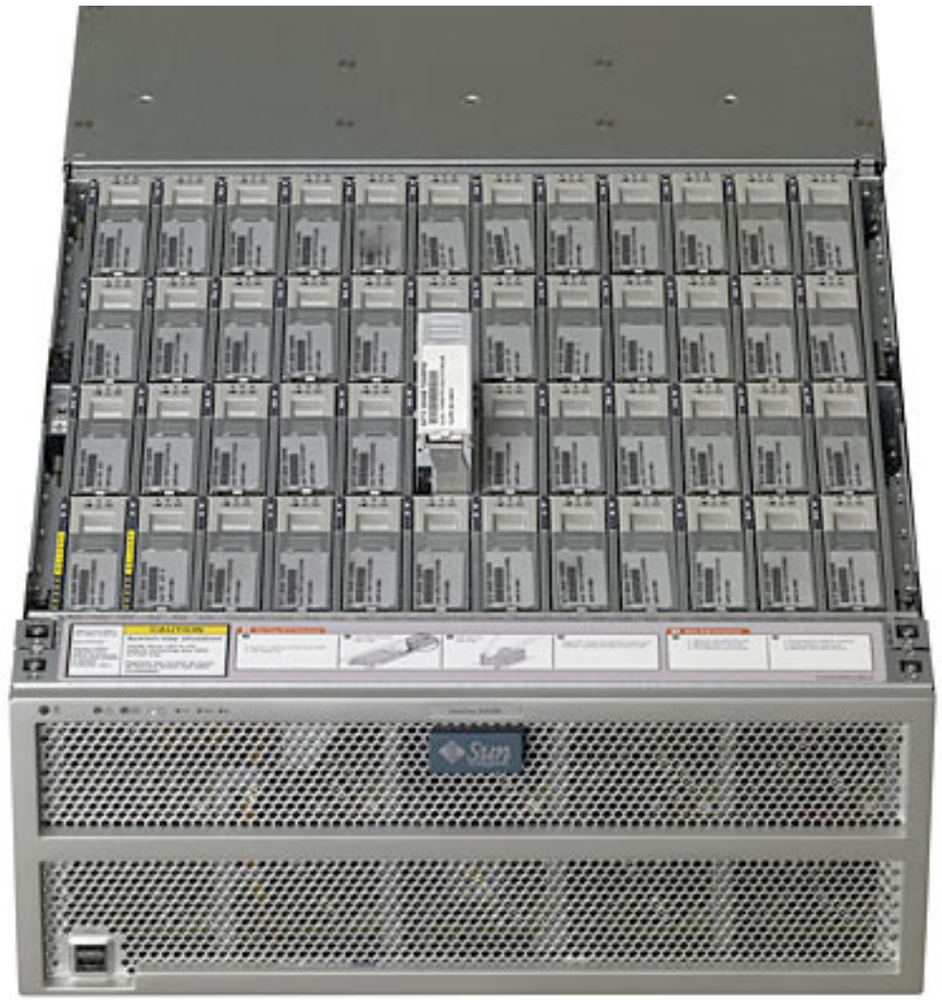
Richtig abgedreht ...
Sparse provisioniertes emulated ZFS volume anyone?

```
# zfs create -V 5g -s tp/ufsvolume
# zfs list | grep "tp/"
tp/gonzo                20.1M   194M    19K  /tp/gonzo
tp/kermit                 18K   194M    18K  /tp/kermit
tp/statler                18K   194M    18K  /tp/statler
tp/ufsvolume              16K   194M    16K  -
tp/waldorf                18K   194M    18K  /tp/waldorf

# ls -l /dev/zvol/dsk/tp/ufsvolume
lrwxrwxrwx   1 root     root          35 Mar 20 13:13 /dev/zvol/dsk/tp/
ufsvolume -> ../../../../devices/pseudo/zfs@0:6c

# newfs /dev/zvol/dsk/tp/ufsvolume
newfs: construct a new file system /dev/zvol/rdsk/tp/ufsvolume: (y/n)? y
Warning: 2082 sector(s) in last cylinder unallocated
/dev/zvol/rdsk/tp/ufsvolume:    10485726 sectors in 1707 cylinders of 48
tracks, 128 sectors
        5120.0MB in 107 cyl groups (16 c/g, 48.00MB/g, 5824 i/g)
super-block backups (for fsck -F ufs -o b=#) at:
 32, 98464, 196896, 295328, 393760, 492192, 590624, 689056, 787488,
885920,
 9539744, 9638176, 9736608, 9835040, 9933472, 10031904, 10130336,
10228768,
 10327200, 10425632
#
```

Dadurch werden solche Systeme erst möglich ...



Insight

4

dtrace

DTrace ist ein hochenwickeltes Tracingframework in Solaris. Es ist ständig verfügbar.

Solange man es nicht nutzt, braucht es keine Rechenzeit.

In Opensolaris Community Edition

Build 105 sind 75688 Probes verteilt.

Tendenz steigend.

```
# uname -a
SunOS master 5.11 snv_105 i86pc i386 i86pc
# dtrace -l | wc -l
   75688
```

DTrace enhält aber auch eine Scriptsprache.

Zum Aktivieren der Probes,
Deaktivieren der Probes,
Aggregierung,
Summierung,
etc.

Man kann damit richtig tolle Scripte zur Diagnostik von laufenden Systemen schreiben ...

IOtop

(zu finden unter: `http://brendangregg.com/dtrace.html#DTraceToolkit`)

```
# iotop -C
Sampling... Please wait.
2005 Jul 16 00:31:38,  load: 1.03,  disk_r:   5023 Kb,  disk_w:     22 Kb

  UID    PID   PPID CMD              DEVICE  MAJ MIN D            BYTES
    0  27740  20320 tar              cmdk0   102  16 W            23040
    0  27739  20320 find             cmdk0   102   0 R           668672
    0  27740  20320 tar              cmdk0   102  16 R          1512960
    0  27740  20320 tar              cmdk0   102   3 R          3108864

2005 Jul 16 00:31:43,  load: 1.06,  disk_r:   8234 Kb,  disk_w:      0 Kb

  UID    PID   PPID CMD              DEVICE  MAJ MIN D            BYTES
    0  27739  20320 find             cmdk0   102   0 R          1402880
    0  27740  20320 tar              cmdk0   102   3 R          7069696
[...]
```

dtruss - Ein truss ohne exorbitante Last

(zu finden unter: `http://brendangregg.com/dtrace.html#DTraceToolkit`)

```
# dtruss -eon bash
PID/LWP    ELAPSD OVERHD SYSCALL(args)            = return
 3911/1:        41     26 write(0x2, "l\0", 0x1)          = 1 0
 3911/1:   1001579     43 read(0x0, "s\0", 0x1)           = 1 0
 3911/1:        38     26 write(0x2, "s\0", 0x1)          = 1 0
 3911/1:   1019129     43 read(0x0, " \001\0", 0x1)                = 1 0
 3911/1:        38     26 write(0x2, " \0", 0x1)          = 1 0
 3911/1:    998533     43 read(0x0, "-\0", 0x1)           = 1 0
 3911/1:        38     26 write(0x2, "-\001\0", 0x1)               = 1 0
 3911/1:   1094323     42 read(0x0, "l\0", 0x1)           = 1 0
 3911/1:        39     27 write(0x2, "l\001\0", 0x1)               = 1 0
 3911/1:   1210496     44 read(0x0, "\r\0", 0x1)          = 1 0
[...]
```

Die Mächtigkeit zeigt sich aber
vielleicht besser an ein paar Onelinern

```
# dtrace -n 'proc:::exec-success { trace(curpsinfo->pr_psargs); }'
dtrace: description 'proc:::exec-success ' matched 1 probe
CPU     ID                    FUNCTION:NAME
  0   3297         exec_common:exec-success   man ls
  0   3297         exec_common:exec-success   sh -c cd /usr/share/man; tbl /usr/share/man/
man1/ls.1 |neqn /usr/share/lib/pub/
  0   3297         exec_common:exec-success   tbl /usr/share/man/man1/ls.1
  0   3297         exec_common:exec-success   neqn /usr/share/lib/pub/eqnchar -
  0   3297         exec_common:exec-success   nroff -u0 -Tlp -man -
  0   3297         exec_common:exec-success   col -x
  0   3297         exec_common:exec-success   sh -c trap '' 1 15; /usr/bin/mv -f /tmp/
mpzIaOZF /usr/share/man/cat1/ls.1 2> /d
  0   3297         exec_common:exec-success   /usr/bin/mv -f /tmp/mpzIaOZF /usr/share/man/
cat1/ls.1
  0   3297         exec_common:exec-success   sh -c more -s /tmp/mpzIaOZF
  0   3297         exec_common:exec-success   more -s /tmp/mpzIaOZF
```

```
# dtrace -n 'sysinfo:::readch { @dist[execname] = quantize(arg0); }'
dtrace: description 'sysinfo:::readch ' matched 4 probes
^C
[...]
  gnome-terminal
           value  ------------- Distribution ------------- count
              16 |                                         0
              32 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@   15
              64 |@@@                                      1
             128 |                                         0

  Xorg
           value  ------------- Distribution ------------- count
              -1 |                                         0
               0 |@@@@@@@@@@@@@@@@@@@@                     26
               1 |                                         0
               2 |                                         0
               4 |                                         0
               8 |@@@@                                     6
              16 |@                                        2
              32 |@                                        2
              64 |                                         0
             128 |@@@@@@@@                                 11
             256 |@@@                                      4
             512 |                                         0
```

Aber dtrace kann auch für Applikationen genutzt werden

Seit 5.1 sind DTrace-Probes in Mysql enthalten
Seit 5.4 sind diese per Default aktiviert.

Stellt euch mal vor:
Mysql Server ist irgendwie langsam, aber in Produktion!
Ausser Betrieb nehmen und Debugging einschalten geht nicht.

Wahrscheinlich hat wieder ein Entwickler nicht aufgepasst ...
... irgendwo fehlt wahrscheinlich mal wieder ein Index.

```
#cat query_load.d
#!/usr/sbin/dtrace -s
#pragma D option quiet
dtrace:::BEGIN
{
 printf("Tracing... Hit Ctrl-C to end.\n");
}
mysql*:::query-start
{
   this->query = copyinstr(arg0);
   this->who   = strjoin(copyinstr(arg3),strjoin("@",copyinstr(arg4)));
   this->connid = arg1;
}
mysql*:::query-done
{
    @queries[this->who,this->connid,this->query] = count();
}
dtrace:::END
{
    printf(" %-18s %5s %s %s\n", "USER@HOST","CONNECT_ID", "COUNT",
"QUERY");
    printa(" %-18s %6d %@5d %s\n", @queries);
}
```

```
# ./query_load.d
Tracing... Hit Ctrl-C to end.
USER@HOST CONNECT_ID COUNT QUERY
root@localhost4 247 SELECT * FROM city_huge ORDER BY population DESC LIMIT 3
root@localhost4 249 INSERT INTO city_huge (Name, CountryCode, District, Population)
SELECT Name, CountryCode, District, Population FROM City LIMIT 1
root@localhost4 249 UPDATE city_huge SET population=population+1 WHERE id =
last_insert_id()
root@localhost4 500 SET @var = floor(rand() * (SELECT max(id) FROM city_huge))
root@localhost4 500 UPDATE city_huge SET population = population +  1 WHERE id = @var
root@localhost4 502 SELECT * FROM city_huge WHERE id = @var
root@localhost4 749 COMMIT
root@localhost4 750 BEGIN
```

```
#cat query_stime.d
#!/usr/sbin/dtrace -s
#pragma D option quiet
dtrace:::BEGIN
{
   printf("Tracing... Hit Ctrl-C to end.\n");
}
mysql*:::query-start
{
   this->query = copyinstr(arg0);
   this->who   = strjoin(copyinstr(arg3),strjoin("@",copyinstr(arg4)));
   this->connid = arg1;
   this->querystart = timestamp;
}
mysql*:::query-done
{
    this->elapsed = (timestamp - this->querystart) /1000000;
    @time[this->who,this->connid,this->query] = sum(this->elapsed);
}
dtrace:::END
{
    printf(" %-18s %5s %10s %10s\n", "USER@HOST","CONNECT_ID", "QUERY
ms",
"QUERY");
    printa(" %-18s %6d %@5d %s\n", @time);
}
```

```
# ./query_stime.d
Tracing... Hit Ctrl-C to end.
USER@HOST CONNECT_ID ms QUERY
root@localhost4 0 BEGIN
root@localhost4 0 COMMIT
root@localhost4 1 INSERT INTO city_huge (Name, CountryCode, District, Population)
SELECT Name, CountryCode, District, Population FROM City LIMIT 1
root@localhost4 1247 UPDATE city_huge SET population=population+1 WHERE id =
last_insert_id()
root@localhost4 1500 SELECT * FROM city_huge WHERE id = @var
root@localhost4 1501 SET @var = floor(rand() * (SELECT max(id) FROM city_huge))
root@localhost4 1502 UPDATE city_huge SET population = population + 1 WHERE id =
@var
root@localhost4 2748 SELECT * FROM city_huge ORDER BY population DESC LIMIT 3
```

Wohlgemerkt immer nur kurz während des Betriebes das Script starten und damit Daten sammeln.

Insight

6

Crossbow

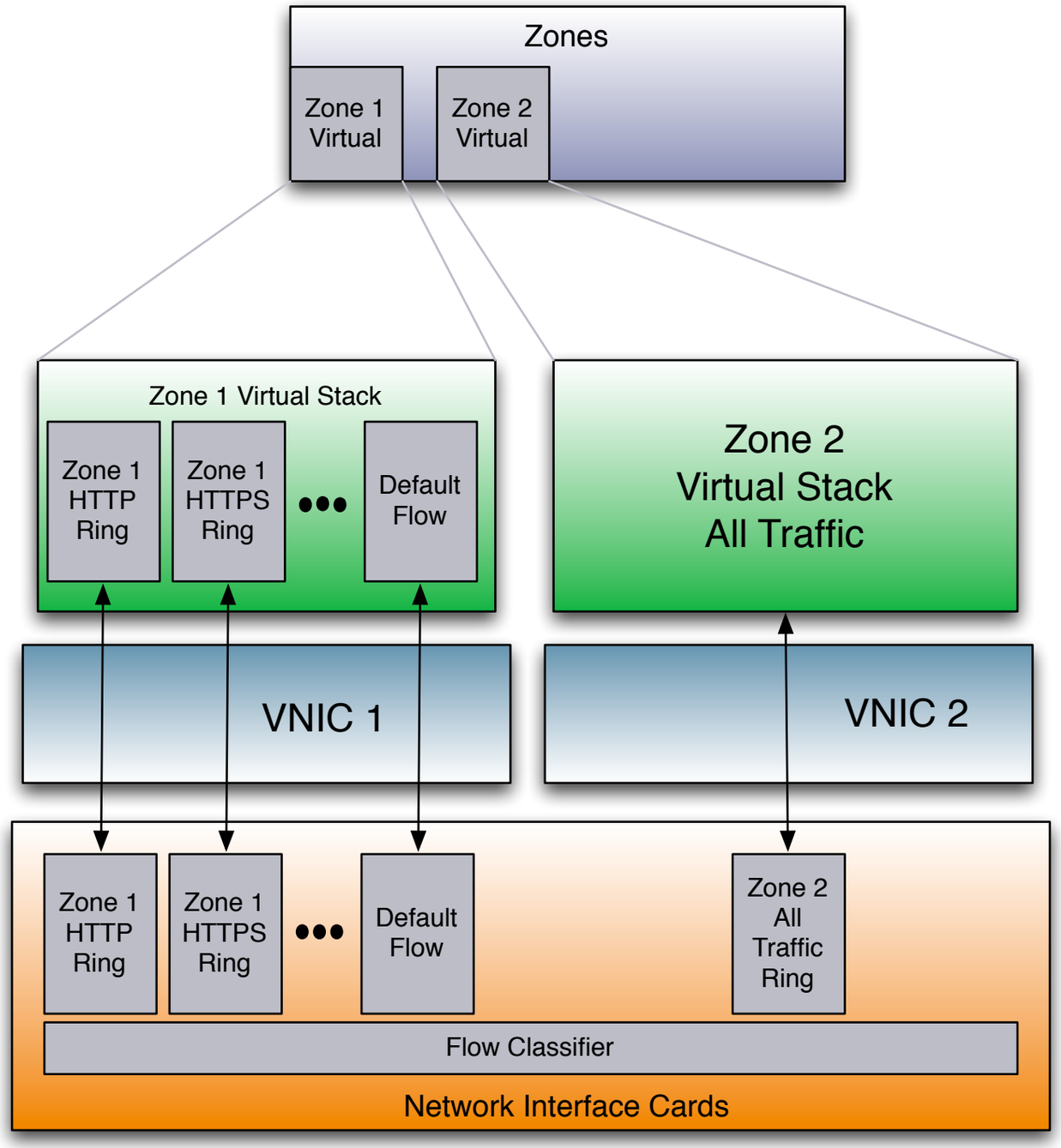
Bei Crossbow geht es vordergründig um Virtualisierung ...

Dafür wird es auch schon in Solaris 10 eingesetzt:
Die IP-Instanzen der Zones kommen aus Crossbow

Das Gesamtkonzept ist sehr viel weitergreifender ...

Fragestellungen:

Wie skaliert man Netzwerktraffic über viele Cores ?
Wie stellt man in konsolidierten Umgebung eine Qualität sicher ?
Wie konsolidiert mit Zonen, wenn das verbindende Netzwerk komplexer ist ?


Zones
Zone 1
Virtual
Zone 2
Virtual
Zone 1 Virtual Stack
Zone 1
HTTP
Ring
Zone 1
HTTPS
Ring
Default
Flow
Zone 2
Virtual Stack
All Traffic
VNIC 1
VNIC 2
Zone 1
HTTP
Ring
Zone 1
HTTPS
Ring
Default
Flow
Zone 2
All
Traffic
Ring
Flow Classifier
Network Interface Cards

Mal ein wenig zur Praxis mit Crossbow ...

Einrichten von VNICs und virtuellen Switchen

```
# dladm create-etherstub etherstub0
# dladm create-etherstub etherstub1
```

```
# dladm create-vnic -l etherstub0 vnic1
# dladm create-vnic -l etherstub1 vnic2
# dladm create-vnic -l etherstub1 vnic3
```

Resourcenmanagement auf Basis von Flows und VNICs

```
#dladm set-linkprop -p maxbw=10mb vnic2
```

```
# flowadm add-flow -l vnic0 -a transport=tcp,local_port=80 httpflow
# flowadm add-flow -l vnic0 -a transport=tcp,local_port=443 httpsflow
# flowadm show-flow
FLOW         LINK         IP_ADDR                                PROTO   PORT
DSFLD
httpflow     vnic0        --                                     tcp     80      --
httpsflow    vnic0        --                                     tcp     443     --
```

```
# flowadm set-flowprop -p maxbw=500M,priority=high httpsflow
# flowadm show-flow httpsflow
FLOW         LINK         IP ADDR                PROTO   PORT     DSFLD
httpsflow     bge0         --                     tcp     443      --
```

```
# flowadm show-flowprop https-1
FLOW         PROPERTY    VALUE     DEFAULT       POSSIBLE
https-1      maxbw       500       --            --
https-1      priority    HIGH      --            LOW,NORMAL,HIGH
```

```
# flowadm add-flow -l bge0 -a transport=UDP -p maxbw=100M, priority=low limit-udp-1
```

Flows zur Verarbeitung an CPU binden

```
# flowadm set-flowprop -p cpus=4,5 httpsflow
# flowadm set-flowprop -p cpus=6,7 httpflow
```

Netzwerkaccounting auf Basis von Flows

```
# acctadm -e extended -f /var/log/net.log net
# acctadm net
Network accounting: active
Network accounting file: /var/log/net.log
Tracked Network resources: extended
Untracked Network resources: none
```

```
# flowadm show-usage -f /var/log/net.log
        FLOW      DURATION  IPACKETS RBYTES       OPACKETS OBYTES      BANDWIDTH
        flowtcp   100       1031     546908       0        0           43.76 Kbps
        flowudp   0         0        0            0        0           0.00 Mbps
```

```
# flowadm show-usage -d -f /var/log/net.log
02/19/2008
```

```
# flowadm show-usage -s 02/19/2008,10:39:06 -e 02/19/2008,10:40:06 -f /var/log/
net.log flowtcp
FLOW      TIME        IPACKETS RBYTES      OPACKETS OBYTES     BANDWIDTH
flowtcp   10:39:06    1        1546          4       6539       3.23 Kbps
flowtcp   10:39:26    2        3586          5       9922       5.40 Kbps
flowtcp   10:39:46    1        240           1       216      182.40 bps
flowtcp   10:40:06    0        0             0       0          0.00 bps
```

Insight 7

Systemic Features

Prelude:
Wie entstehen eigentlich Features in Solaris ?

Bevor überhaupt eine Zeile Code geschrieben wird, geht ein neues Feature erst einmal ein durch ein Kommitee ...

das

Architecture

Review Commitee

Ziele des ARC ist es, …

… sicherzustellen, das für jedes Projekt eine Dokumentation verfügbar ist, damit andere verstehen, was Inhalt des Projektes ist.

Ziele des ARC ist es, ...

... ein Rahmen der Beratung zu schaffen mit anderen Entwicklern die sich schon lange in diesem oder ähnlichen Themengebieten betwegen.

Ziele des ARC ist es, ...

... Doppelentwicklungen, Overengineering, Quality Problems oder unerwünschte Effekte auf die Gesamtarchitektur zu entdecken.

Ziele des ARC ist es, ...

... einen formalen Rahmen zur Diskussion und Lösung von architekturellen Problemen zur Verfügung zu stellen.

Ziele des ARC ist es, …

… diese architekturellen Probleme überhaupt erst zu identifizieren

Ziele des ARC ist es, …

- Die Kommunikation zwischen den Projekten zu verbessern …

Erst wenn das ARC zugestimmt hat, kann der Code später in die Codebase integriert werden ..

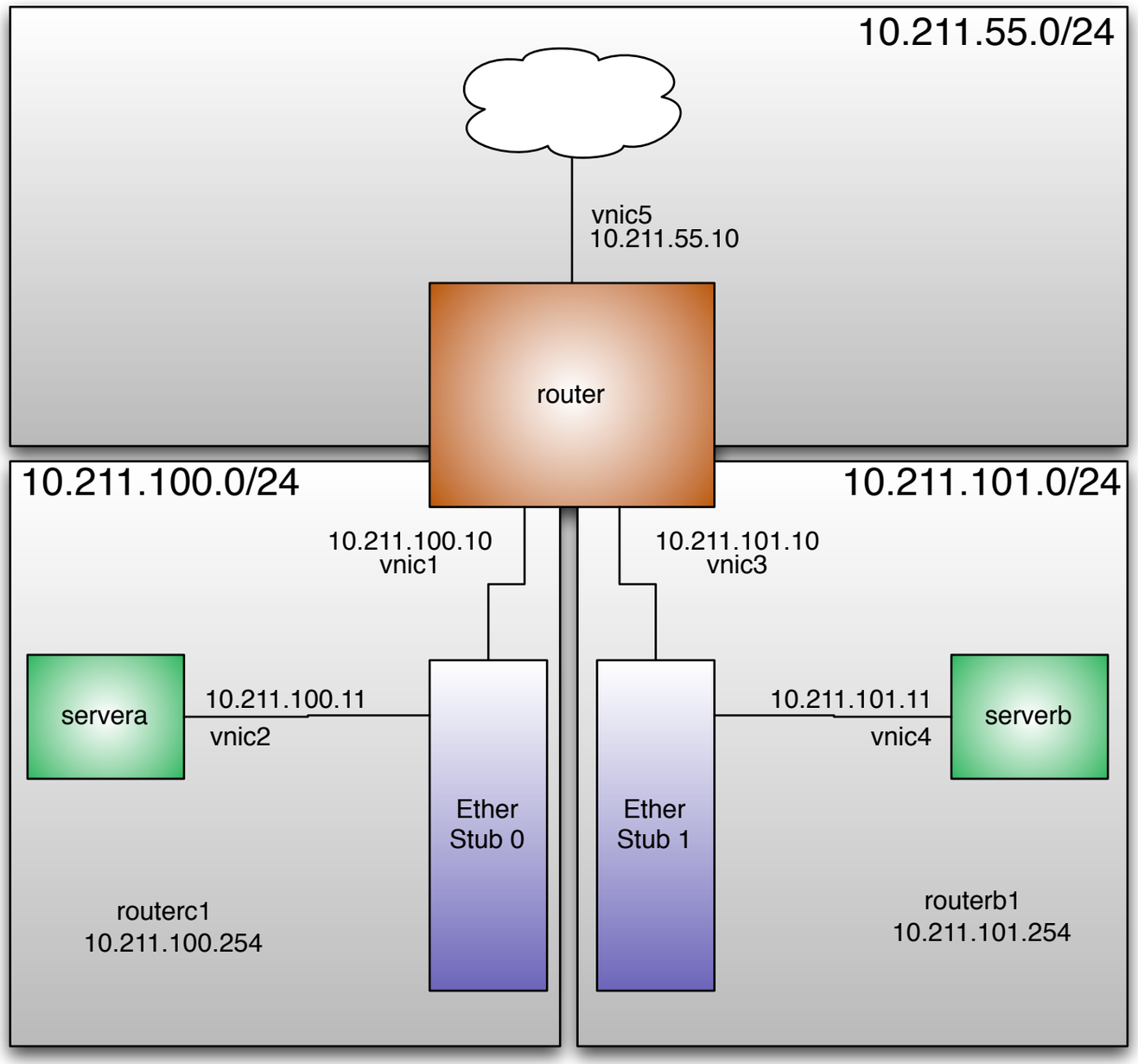
Das mag zunächst sehr formal erscheinen …

Hat aber einige Vorteile:

- eine sehr zielgerichtete Entwicklung
  - wenns fertig ist, kommt es auch in die Codebase (entsprechende Qualität vorausgesetzt)
  - ich habe schon häufiger eine Raffinierung der Architektur in diesem Prozess gesehen.
  - Es wird immer auf die Gesamtarchtektur geachtet.

Insbesondere letzteres führt zu etwas, das ich Systemische Features nenne:

Einzelfeatures, die durch ihre enge Kombination erst ihre ganze Fähigkeit aufweisen und so zu Gesamtkonstrukten führt, die mehr sind als die Summe ihrer Einzelfunktionen ...


10.211.55.0/24
vnic5
10.211.55.10
router
10.211.100.0/24
10.211.101.0/24
10.211.100.10
vnic1
10.211.101.10
vnic3
servera
10.211.100.11
vnic2
10.211.101.11
vnic4
serverb
Ether
Stub 0
Ether
Stub 1
routerc1
10.211.100.254
routerb1
10.211.101.254

```
# dladm create-etherstub etherstub0
# dladm create-etherstub etherstub1
```

```
# dladm create-vnic -l etherstub0 vnic1
# dladm create-vnic -l etherstub1 vnic2
# dladm create-vnic -l etherstub1 vnic3
```

```
# dladm showlink
LINK         CLASS     MTU    STATE    OVER
ni0          phys      1500   unknown  --
etherstub0   etherstub 9000   unknown  --
etherstub1   etherstub 9000   unknown  --
vnic1        vnic        9000   up       etherstub0
vnic2        vnic        9000   up       etherstub0
vnic3        vnic        9000   up       etherstub1
```

```
# zfs create rpool/zones
# zfs set compression=on rpool/zones
# zfs set mountpoint=/zones rpool/zones
```

```
create -b
set zonepath=/zones/template
set ip-type=exclusive
set autoboot=false
add inherit-pkg-dir
set dir=/lib
end
add inherit-pkg-dir
set dir=/platform
end
add inherit-pkg-dir
set dir=/sbin
end
add inherit-pkg-dir
set dir=/usr
end
add inherit-pkg-dir
set dir=/opt
end
commit
```

```
# zonecfg -z template -f template
# zoneadm -z template install
A ZFS file system has been created for this zone.
Preparing to install zone <template>.
Creating list of files to copy from the global zone.
Copying <3488> files to the zone.
Initializing zone product registry.
Determining zone package initialization order.
Preparing to initialize <1507> packages on the zone.
Initialized <1507> packages on zone.
Zone <template> is initialized.
The file </zones/template/root/var/sadm/system/logs/
install_log> contains a log of the zone
installation.
#
```

```
create -b
set zonepath=/zones/serverA
set ip-type=exclusive
set autoboot=false
add inherit-pkg-dir
set dir=/lib
end
add inherit-pkg-dir
set dir=/platform
end
add inherit-pkg-dir
set dir=/sbin
end
add inherit-pkg-dir
set dir=/usr
end
add inherit-pkg-dir
set dir=/opt
end
add net
set physical=vnic2
end
commit
```

```
# zonecfg -z servera -f serverA
# zoneadm -z servera clone template
Cloning snapshot rpool/zones/template@SUNWzone3
Instead of copying, a ZFS clone has been created for this
zone.
# cp serverA_sysidcfg /zones/serverA/root/etc/sysidcfg
# cp site.xml /zones/serverA/root/var/svc/profile
# zoneadm -z servera boot
```

```
# zonecfg -z servera -f serverA
# zoneadm -z servera clone template
Cloning snapshot rpool/zones/template@SUNWzone3
Instead of copying, a ZFS clone has been created for this
zone.
# cp serverA_sysidcfg /zones/serverA/root/etc/sysidcfg
# cp site.xml /zones/serverA/root/var/svc/profile
# zoneadm -z servera boot
```

```
# zonecfg -z servera -f serverA
# zoneadm -z servera clone template
Cloning snapshot rpool/zones/template@SUNWzone3
Instead of copying, a ZFS clone has been created for this
zone.
# cp serverA_sysidcfg /zones/serverA/root/etc/sysidcfg
# cp site.xml /zones/serverA/root/var/svc/profile
# zoneadm -z servera boot
```

```
system_locale=C
terminal=vt100
name_service=none
network_interface=vnic2 {primary hostname=server1
ip_address=10.211.100.11 netmask=255.255.255.0 protocol_ipv6=no
default_route=NONE}
nfs4_domain=dynamic
root_password=cmuL.HSJtwJ.I
security_policy=none
timeserver=localhost
timezone=US/Central
```

```
# zoneadm list -v
  ID NAME             STATUS     PATH                           BRAND    IP
   0 global           running    /                              native   shared
  13 router           running    /zones/router                  native   excl
  15 servera          running    /zones/serverA                 native   excl
  19 serverb          running    /zones/serverB                 native   excl
```

```
# ifconfig vnic2 plumb
vnic2 is used by non-globalzone: servera
```

```
# routeadm -e ipv4-forwarding
# routeadm -e ipv4-routing
# routeadm -u
# routeadm
             Configuration   Current             Current
                    Option   Configuration       System State
---------------------------------------------------------------
               IPv4 routing   enabled             enabled
               IPv6 routing   disabled            disabled
             IPv4 forwarding  enabled             enabled
             IPv6 forwarding  disabled            disabled

           Routing services   "route:default ripng:default"
Routing daemons:

                   STATE    FMRI
                disabled    svc:/network/routing/zebra:quagga
                disabled    svc:/network/routing/rip:quagga
                disabled    svc:/network/routing/ripng:default
                disabled    svc:/network/routing/ripng:quagga
                disabled    svc:/network/routing/ospf:quagga
                disabled    svc:/network/routing/ospf6:quagga
                disabled    svc:/network/routing/bgp:quagga
                disabled    svc:/network/routing/isis:quagga
                disabled    svc:/network/routing/rdisc:default
                  online    svc:/network/routing/route:default
                disabled    svc:/network/routing/legacy-routing:ipv4
                disabled    svc:/network/routing/legacy-routing:ipv6
                  online    svc:/network/routing/ndp:default
```

```
# netstat -nr

Routing Table: IPv4
  Destination           Gateway           Flags  Ref     Use     Interface
-------------------- -------------------- ----- ------ --------- ---------
default              10.211.100.10        UG        1          0 vnic2
10.211.100.0         10.211.100.11        U         1          0 vnic2
127.0.0.1            127.0.0.1            UH        1         49 lo0
```

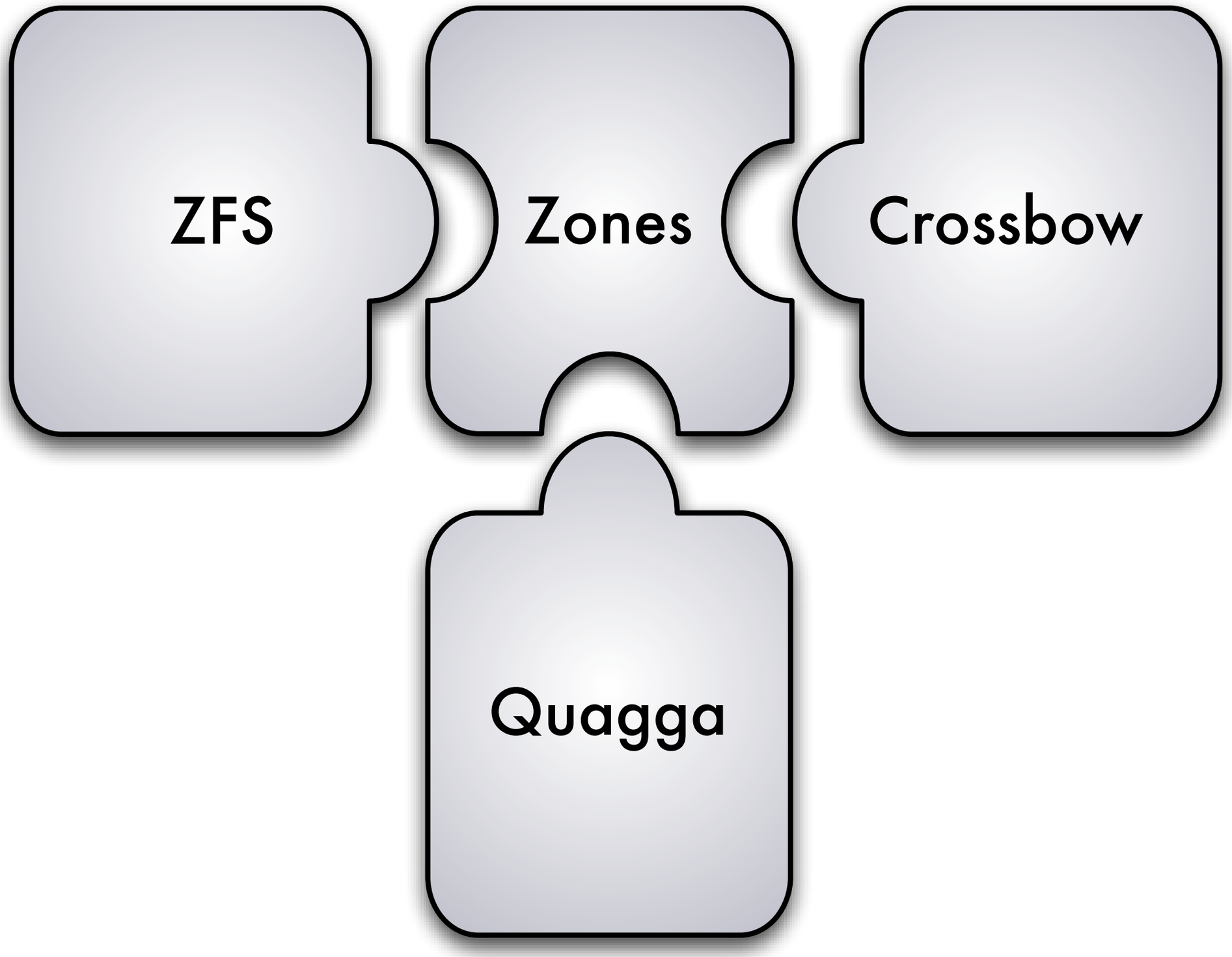
```
# ping 10.211.100.11
10.211.100.11 is alive
# traceroute 10.211.100.11
traceroute to 10.211.100.11 (10.211.100.11), 30 hops max, 40 byte packets
 1  10.211.101.10 (10.211.101.10)  0.285 ms  0.266 ms  0.204 ms
 2  10.211.100.11 (10.211.100.11)  0.307 ms  0.303 ms  0.294 ms
#
```

Solaris Zones liefern die virtuellen Systeme
ZFS beschleunigt die Generierung der Zones
Crossbow liefert die das Netzwerk

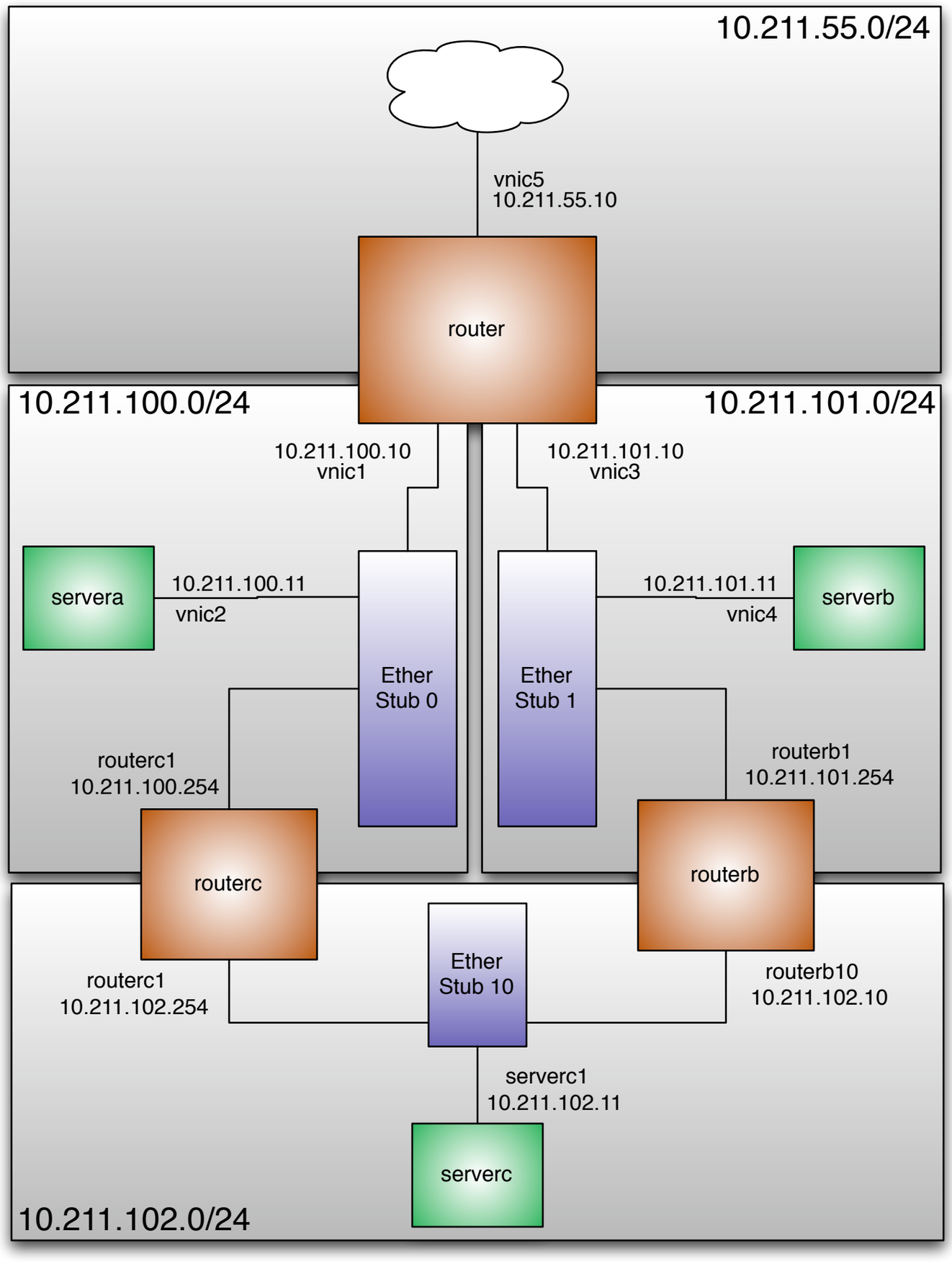
```
system_locale=C
terminal=vt100
name_service=none
network_interface=vnic2 {primary hostname=server1
ip_address=10.211.100.11 netmask=255.255.255.0 protocol_ipv6=no
default_route=NONE}
nfs4_domain=dynamic
root_password=cmuL.HSJtwJ.I
security_policy=none
timeserver=localhost
timezone=US/Central
```

Und genau genommen aktiviert dieses
`default_route=none`
ein viertes Feature.

Ist keine Defaultroute angegeben,
startet das System automatisch
bei einer Netzwerkkarte den RDISC-Prozess
bei zwei oder mehr Karten einen RIPv2-Daemon


ZFS
Zones
Crossbow
Quagga

Das wird bei solchen Konfigurationen sehr interessant ....


10.211.55.0/24
vnic5
10.211.55.10
router
10.211.100.0/24
10.211.101.0/24
10.211.100.10
vnic1
10.211.101.10
vnic3
servera
10.211.100.11
vnic2
10.211.101.11
vnic4
serverb
Ether
Stub 0
Ether
Stub 1
routerc1
10.211.100.254
routerb1
10.211.101.254
routerc
routerb
Ether
Stub 10
routerc1
10.211.102.254
routerb10
10.211.102.10
serverc1
10.211.102.11
serverc
10.211.102.0/24

Auf einem der Server:

```
Routing Table: IPv4
  Destination           Gateway           Flags  Ref    Use     Interface
-------------------- -------------------- ----- ----- ---------- ---------
default              10.211.100.10        UG        1          0 vnic2
default              10.211.100.254       UG        1          0 vnic2
10.211.100.0         10.211.100.11        U         1          0 vnic2
127.0.0.1            127.0.0.1            UH        1         49 lo0
```

Auf einem der Router:

```
Routing Table: IPv4
  Destination           Gateway           Flags  Ref    Use     Interface
-------------------- -------------------- ----- ----- ---------- ---------
default              10.211.101.10        UG        1          0 routerb1
10.211.100.0         10.211.102.254       UG        1          0 routerb10
10.211.101.0         10.211.101.254       U         1          0 routerb1
10.211.102.0         10.211.102.10        U         1          0 routerb10
127.0.0.1            127.0.0.1            UH        1         23 lo0
```

ZFS

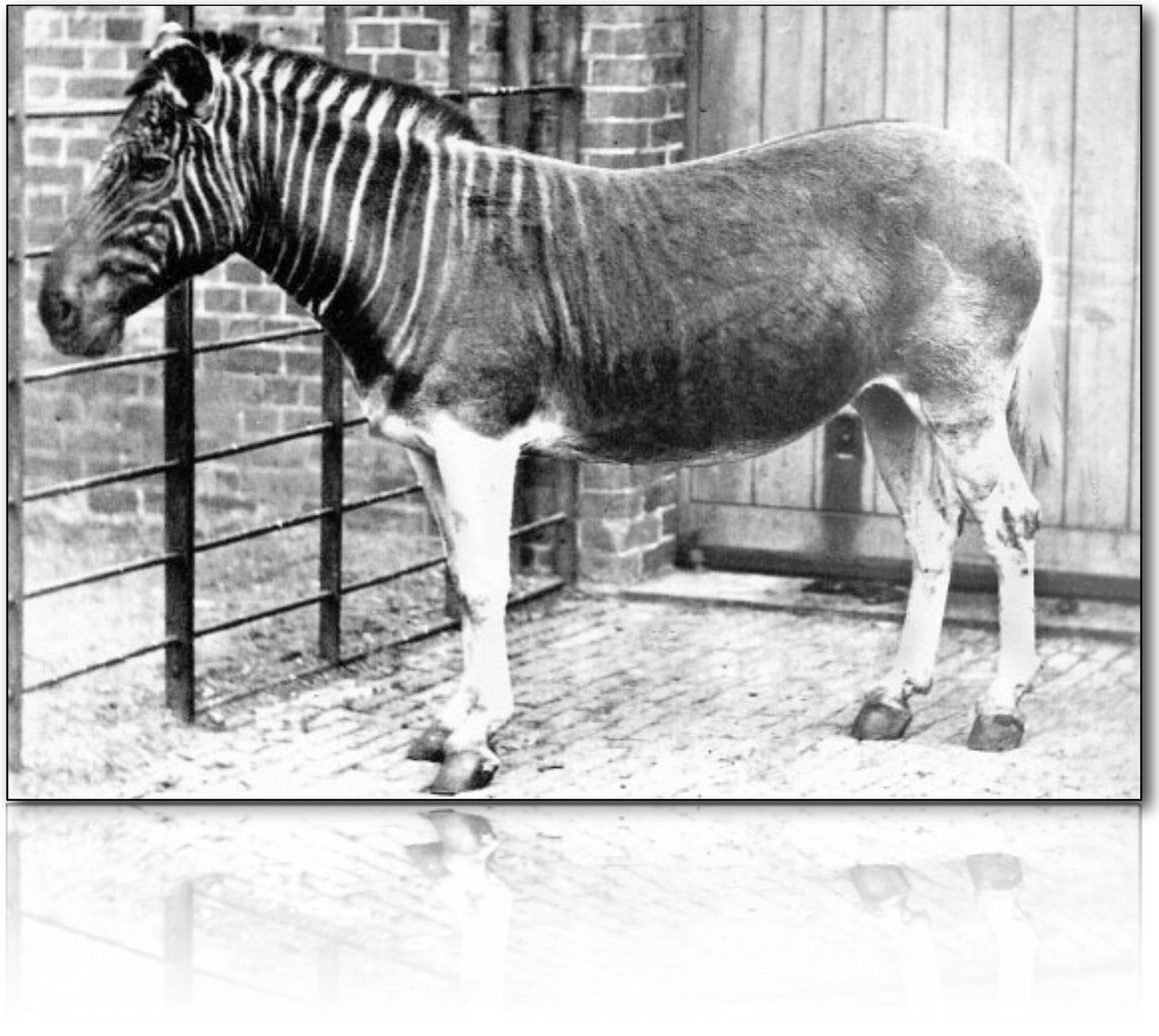
Zones

Crossbow

Quagga ?????

Quagga

Aus der Wikipedia:



Ein Quagga ist ein Zebraart ...

GNU Zebra war eine Suite von Routingdaemons für verschiende Protokolle (inaktiv seit 2006)

GNU Quagga ist eine aktive Weiterentwicklung von GNU Zebra ...

ZFS

Zones

Crossbow

Quagga

Ironie der Geschichte:

In der Natur ist das Quagga ausgestorben ...
... in der GNUsphäre das Zebra.

Insight

8

Was habe ich nicht genannt?

Solaris Resource Manager

Unter Solaris kann man den CPU- und Speicher mit Resource Manager und mit Crossbow den Bandbreitenverbrauch steuern.

Availability Suite

In OpenSolaris ist ein System enthalten zur synchronen und asynchronen Replikation von Blockdevices auch über IP-Leitungen.

Least Privileges

There is no root. Okay ... es gibt ihn noch, aber nur aus Kompatibilitäts-gründen. Problemlösung für: Wie wird man den http-prozess mit root-Rechten los

Role Based Access Control

Aufteilung der Rechte des Admins auf verschiedene Rollen. Erweitert durch Authorisations zur Weiterführung dieser Aufteilung innerhalb einer Applikation

Aussagekräftige Crash/Coredumps inclusive mdb

Wie holt man sinnvolle Informationen aus dem Crashdump

Zonen

Bis zu 8191 virtuelle Betriebsysteminstanzen mit extrem wenig Virtualisierungoverhead (es läuft ein Kernel, der sich so benimmt wie viele)

Service Management Framework

Überwachen von Services auf dem System. Automatisches Restarten. Starten in der richtigen Reihenfolge. Ein init.d startet nicht neu, wenn der Prozess terminiert.

Fault Management Architecture

Überwachung des Systems auf Hinweise, das sich ein Problem anbahnt und automatische, präventive Reaktion darauf (CPU offlining, Memory Page Retirement)

Open HA Cluster

Zwar nicht auf dem Datenträger dabei, jedoch ein professionelles, quelloffenes Clusterframework für Solaris.

SCSI Target Framework

iSCSI-Target, iSCSI-over-RDMA-over-IB-Target, iSER-Target, FC-Target, SAS-Target,

NFS

NFSv4 in einer standardkonformen Implementation. NFSv4.1-Features wie parallelNFS kurz vor Release.

Hierarchical Storage Management (upcoming)

ZFS erhält bald ein hierarchisches Storage Management. Festplatten sind nur noch Cache. Daten liegen auf Tape, langsamen Festplatten, einer Storage Cloud.

Momentan verwendet man unter Solaris dafür SamFS.

Und noch viel mehr ... !

Interesse geweckt?

Download unter:

http://www.opensolaris.org/os/downloads/

Ein guter Einstieg:

http://www.solaristutorials.org

jede Menge Feature-Walkthroughs

Danke!

Jörg Möllenkamp
Senior Systems Engineer

Sun Microsystems
Geschäftsstelle Hamburg

Weblog
http://www.c0t0d0s0.org/