

Wednesday, April 30, 2008

Spring 4

Posted by Joerg Moellenkamp in Photographie at 20:06

Pagerank

This made my day ...

Homepage of Sun Germany

Pagerank 6

Homepage of Jörg Möllenkamp

Pagerank 6

Posted by Joerg Moellenkamp in Blogosphere at 12:44

Open Storage

A few months ago i've speculated about the death of high end storage. At the end, all storage boxes from NetApp, EMC, HDS, HP, Sun et. al. are just custom build-computers with a custom build operating system (and to a part not even that) with a bulkload of harddisks connected to it. There is not really a technological barrier, that keeps other companies out of the market.

The biggest problem to build commodity storage boxes out of commodity components was the problem of having an operating system for it. Stable and with all needed features (beginning with a good filesystem and ending with the target modes for various interconnect technologies). And more important: This software has to be non-proprietary. Albeit some vendors use a Unix derivative under the hood, the storage part is mostly closed down.

Okay, yesterday Sun announced the Open Storage Platform. When you look at Opensolaris you already find several components of a storage os. So it's quite logical to use Opensolaris as the basis for an open storage system. The basic idea of Open Storage is simple: OpenStorage = commodity industry standard hardware + OpenSolaris. When you can combine standard hardware with a standard operating system, you can build even high end storage system with standard components thus building them to a price point much lower than today. As i've speculated in my older article: The problems of technology commodization will hit the vendors of high storage soon. When it's not Opensolaris, it will be one of the other operating systems. But it's inevitable. Well, i strongly believe, that Sun has a headstart because of all components in Opensolaris like ZFS, COMSTAR, FMA and so on. Time will tell.

Posted by Joerg Moellenkamp in Solaris at 10:33

Tuesday, April 29. 2008

Playing around with Opensolaris 2008.05 Release Candidates

Today, i upgraded my workstations at home to Opensolaris 2008.05 as the primary operating system. Went smoothly. When you don't have to keep the Nvidia drivers out of your code because of political reasons, even installing and configuring the graphic card is a no-brainer. You have to do nothing. Correct settings out of the box.

ZFS snapshot based boot environments I've started with the Release Candidate 2

```
jmoekamp@glamdring:~# cat /etc/release
```

```
Open Solaris 2008.05 snv_86_rc2 X86
```

```
Copyright 2008 Sun Microsystems, Inc. All Rights Reserved.
```

```
Use is subject to license terms.
```

```
Assembled 21 April 2008 Okay, now i've updated the package list and started the update of my
```

```
installation.jmoekamp@glamdring:~# pkg refresh
```

```
jmoekamp@glamdring:~# pkg image-update
```

```
DOWNLOAD          PKGS    FILES  XFER (MB)
```

```
Completed          4/4     3/3   1.24/1.24
```

```
PHASE              ACTIONS
```

```
Update Phase      13/13
```

A clone of opensolaris exists and has been updated and activated.

On next boot the Boot Environment opensolaris-1 will be mounted

on '/. Reboot when ready to switch to this updated BE. A really neat feature of the package manager is the automatic generation of an zfs-snapshot based boot environment:jmoekamp@glamdring:~# zfs list

```
NAME              USED AVAIL REFER MOUNTPOINT
rpool             2.37G 142G 56.5K /rpool
rpool@install     18.5K - 55K -
rpool/ROOT        2.28G 142G 18K /rpool/ROOT
rpool/ROOT@install 0 - 18K -
rpool/ROOT/opensolaris 122K 142G 2.23G legacy
rpool/ROOT/opensolaris-1 2.28G 142G 2.23G legacy
rpool/ROOT/opensolaris-1@install 4.66M - 2.22G -
rpool/ROOT/opensolaris-1@static:-:2008-04-29-17:59:13 562K - 2.23G -
rpool/ROOT/opensolaris-1/opt 3.60M 142G 3.60M /opt
rpool/ROOT/opensolaris-1/opt@install 0 - 3.60M -
rpool/ROOT/opensolaris-1/opt@static:-:2008-04-29-17:59:13 0 - 3.60M -
rpool/ROOT/opensolaris/opt 0 142G 3.60M /opt
rpool/export      85.8M 142G 19K /export
rpool/export@install 0 - 19K -
rpool/export/home 85.8M 142G 85.8M /export/home
rpool/export/home@install 18K - 21K -
```

The update automatically triggers the creation of a new boot-environment and the integration of this boot-environment to GRUB. Next time when you start your system, it will start with this boot environment.jmoekamp@glamdring:~# beadm list

```
BE      Active Active on Mountpoint Space
```

```
Name      reboot      Used
```

```
----      -
opensolaris-1 yes yes legacy 2.29G
```

```
opensolaris no no - 57.37MI rebooted here, and after the reboot the operating system came up as a
```

Release Candidate 2a system:

```
jmoekamp@glamdring:~$ cat /etc/release
```

```
Open Solaris 2008.05 snv_86_rc2a X86
```

```
Copyright 2008 Sun Microsystems, Inc. All Rights Reserved.
```

```
Use is subject to license terms.
```

```
Assembled 23 April 2008 Nice ... switching to the old environment is really easy
```

```
jmoekamp@glamdring:~# beadm activate opensolaris Now we can reboot again. Et voila, after the reboot you are back in your Release Candidate 2 operating system. jmoekamp@glamdring:~# cat /etc/release
```

```
Open Solaris 2008.05 snv_86_rc2 X86
```

```
Copyright 2008 Sun Microsystems, Inc. All Rights Reserved.
```

Use is subject to license terms.

Assembled 21 April 2008
When you look into the list of boot environments, you will see that the both environment swapped their roles.
jmoekamp@glamdring:~# beadm list
BE Name Active reboot Active on Mountpoint Space Used

opensolaris-1 no no - 63.57M
opensolaris yes yes legacy 2.30G
Of course, you can jump back to the newer installation again.
jmoekamp@glamdring:~# beadm activate opensolaris-1
Reboot the system and afterwards you have your Release Candidate 2a operating system online again.
jmoekamp@glamdring:~\$ cat /etc/release
Open Solaris 2008.05 snv_86_rc2a X86
Copyright 2008 Sun Microsystems, Inc. All Rights Reserved.
Use is subject to license terms.
Assembled 23 April 2008

A nice side effect of snapshot based boot environments
Every now and then even a experienced admin tends to do really dumb errors. Like me i've accidentally deleted the /etc/hosts:
jmoekamp@glamdring:~# echo "192.168.1.xxx fileserver.internal" > /etc/hosts
Fsck! But then i thought: "The system makes snapshots when upgrading packages".
Okay, snapshots are accessible by using the .zfs directory in the root of the filesystem. So i just had to go into a snapshot directory and gather an older /etc/hosts version.
jmoekamp@glamdring:~# cd .zfs
jmoekamp@glamdring:~/.zfs# ls
snapshot
jmoekamp@glamdring:~/.zfs# cd snapshot/
jmoekamp@glamdring:~/.zfs/snapshot# ls
install static-:2008-04-29-17:59:13
jmoekamp@glamdring:~/.zfs/snapshot# cd static-:2008-04-29-17:59:13/
jmoekamp@glamdring:~/.zfs/snapshot/static-:2008-04-29-17:59:13# cd etc/
jmoekamp@glamdring:~/.zfs/snapshot/static-:2008-04-29-17:59:13/etc# cat hosts
CDDL HEADER START

[..]

::1 localhost
127.0.0.1 glamdring glamdring.local localhost loghost
jmoekamp@glamdring:~/.zfs/snapshot/static-:2008-04-29-17:59:13/etc# cp hosts /etc/hosts
jmoekamp@glamdring:~/.zfs/snapshot/static-:2008-04-29-17:59:13/etc# cd /
And now i was able to add the hostname in a correct way without deleting it:
jmoekamp@glamdring:~# echo "192.168.1.xxx fileserver.internal" >> /etc/hosts

Posted by Joerg Moellenkamp in Solaris at 22:13

Spring 3

Posted by Joerg Moellenkamp at 22:00

Comparative advertising

The problem with the X300 ... it runs this mediocre operation system from Redmond. Dear Levono ... you could make it smaller, you could make it faster, you could even make it silver and cool-looking ... your problem is the OS.

(via: fakesteve)

Posted by Joerg Moellenkamp in Apple at 18:18

links for 2008-04-29

John Resig - Ruby VM in JavaScript

Ruby Interpreter in Javascript 5 on Firefox 3.0b times faster than original Ruby ? WTF?

(tags: Ruby)

Posted by del.icio.us in del.icio.us at 13:33

easySMF

You need a manifest for Service Management Facility. It's simple service, so no complex property configurations or multiple instances. Then easySMF may a helpful tool for you. EasySMF is a web based tool to create simple manifests for direct import to the repository or as a starting point.

(found via cuddletech)

Posted by Joerg Moellenkamp in Solaris at 09:53

Spring 2

Posted by Joerg Moellenkamp in Photographie at 08:34

Monday, April 28. 2008

Aha ...

53%

Posted by Joerg Moellenkamp in Fundsache at 19:05

ZFS at digitar

In "Democratizing Storage" Jason writes about his opinion of ZFS and the usage of ZFS at his employer digitar. It's really positive. You can summarize the article at best with the following quote: When it comes to storing data, you'll pry OpenSolaris (and ZFS) out of our cold dead hands. We won't deploy databases on anything else.

Posted by Joerg Moellenkamp in Solaris at 18:01

Less Known Solaris features: lockfs

Okay, quite often the configuration of a feature or application mandates that the data on the disk doesn't change while you activate it. Okay, an easy way would be simply unmounting the disk. Okay, that's possible but then you can't access the data on the disk at all. You can't even read from the filesystem, albeit this doesn't change anything on the disk (okay, as long you've mounted the disk with noatime).

So: How do you ensure, that the content of a filesystem doesn't change while you work with the disk. ufs has an interesting feature. It's called lockfs you can lock the filesystem. You can lock it to an extent that you can only unmount and remount it to gather access to the data, but you can lock only some ways to access it.

Types of Locks

The lockfs command can establish a number of locks on an UFS filesystem:

- Delete Lock (-d): suspends access that could remove directory entries
- Hard Lock (-h): suspends all access to the filesystem. It can't be unlocked. You have to unmount and remount it
- Name Lock (-n): suspends all access to the filesystem, that could remove or change directory entries
- Error Lock (-n): This lock is normally established, when UFS detects an internal inconsistency. It's released by the usage of fsck. It suspends all access to the filesystem.
- write lock (-w): This lock suspends all accesses that could modify the filesystem
- flush log (-f): This isn't really a lock. This option forces a synchronous flush of the named filesystem. It returns wenn all data has been written to disk and the log has been rolled

Write Lock Okay, let's assume we've mounted an UFS filesystem at /mnt. # echo "test" > testfile1

ls

lost+found testfile1 No problem. Our testfile found it's way into file system. Okay, now we establish a write lock on our file system. # lockfs -w /mnt You set the locks with the lockfs command, the switch -w tells command to set an write lock. With a write lock, you can read a filesystem, but you can't write to it. Okay, let's check the existing locks. You use the lockfs command without any further options. # lockfs

```
Filesystem      Locktype  Comment/mnt      write
When we try to add an additional file, the write system call simply blocks.
```

echo "test" > testfile2

^C bash: testfile2: Interrupted system call We have to break the echo command with CTRL-C. Okay, now let's release the lock. # lockfs -u /mnt The -u commands lockfs to release the lock. When you list the existing locks, the lock on /mnt is gone. # lockfs And just to check it, we try to write "test" to testfile2 again. # echo "test" > testfile2 Not problem. The command returns an an instance. When you check the filesystem, you will see both files. # ls -l

total 20

```
drwx----- 2 root  root    8192 Apr 25 18:10 lost+found
-rw-r--r--  1 root  root     5 Apr 27 03:49 testfile1
-rw-r--r--  1 root  root     5 Apr 27 03:50 testfile2
```

Delete lock A rather strange kind of lock is the delete lock. It blocks all delete operations. It's of less practical use, as you would think at first. You can't delete a file, but you can zero it or fill it with other content. Okay, let's use the testfile from our last example. At first we try to delete the first file: # rm testfile1 No problem. Now we establish the delete lock. This time we add a comment. You can use this command to tell other admins why you have established the lock.

```
# lockfs -c "no deletes today" -d /mnt
When you check for existing locks, you will see the delete lock on /mnt and the comment:
# lockfs
Filesystem      Locktype Comment
/mnt            delete  no deletes today
When you try to delete the file, the rm just blocks and you have to break it with CTRL-C again:
# rm testfile2
^C
When you've delete-locked an filesystem, you can create new files, you can append data to it, you can overwrite it:
# echo "test" > testfile3
# echo "test" >> testfile3
# echo "test" > testfile3
There is only one thing you can't do with this new file: You simply can't delete it.
# rm testfile3
^C
Now we release the lock
# lockfs -u /mnt
Now you can clean up your testdirectory /mnt again.
# rm testfile2
# rm testfile3
```

Conclusion The lockfs is a really neat feature to deny certain accesses to your filesystem without unmounting it completely. Some locks are more useful for general use than other other. For example the write lock is really useful, when you want to freeze the content of the filesystem while working with tools like AVS. Delete lock or name lock are more of use, when you need a stable directory. The usecase is more an internal one for other tools.

Do you want to learn more?

Documentation

docs.sun.com: man page lockfs

Posted by Joerg Moellenkamp in Solaris at 13:22

Spring

Posted by Joerg Moellenkamp in Photographie at 08:07

Sunday, April 27, 2008

links for 2008-04-27

Quercus

Quercus is Caucho Technology's 100% Java implementation of PHP 5 released under the Open Source GPL license.
(tags: php java opensource programming web webdev scripting)

Resin backed PHP drives 4x performance improvements for Drupal | WorkHabit.org

(tags: php java performance drupal programming optimization)

Posted by del.icio.us in del.icio.us at 13:32

Less known Solaris Features: Point-in-time copy with AVS

Okay, the next installment of the "Less known Solaris Features" series is online. This time i will discuss the feature Point in time copies with AVS. As i have to go through some theory at first, it's a quite long tutorial. It's the longest so far. Point-in-time copies are a rather unusual feature for small installations, but they are absolutely essential for many enterprise customers. And when you really think about it, there are many usecases even for small installations.

Part 1: Introduction

Part 2: Basics

Part 3: Independent copy

Part 4: Dependent copy

Part 5: Compact dependent copy

Part 6: Preparation of the test environment

Part 7: Starting a point-in-time copy

Part 8: Working with point-in-time copies

Part 9: Disaster Recovery

Part 10: Administration

Part 11: Conclusion

Have fun while trying out the feature!

Posted by Joerg Moellenkamp in Solaris at 11:04

Less known Solaris Features: Point-in-time copy with AVS - Part 11: Conclusion

I hope i gave you some insight into this really interesting feature of Solaris and Opensolaris. There are vast possibilities to use it in your daily use. It's not limited to disaster recovery or backups. One of my customers uses this tool to create independent copies of their database. They take a snapshot at midnight and export it on a different database server. The rationale for this process: They run some long running analytics with a huge load on the I/O system on this independent copy. By using the copy the analysis doesn't interfere with the production use of the database. Another customer uses this feature for generating test copies of their production data for testing new software versions. You see, the possibilities are vast and virtually endless.

Do you want to learn more?

Documentation

docs.sun.com: Sun StorageTek AVS 4.0 Point-in-Time Copy Software Administration Guide

docs.sun.com: Manpage of iadm

Misc.

blogs.sun.com/AVS: The Blog of the Availability Suite

Posted by Joerg Moellenkamp in Solaris at 10:12

Less known Solaris Features: Point-in-time copy with AVS - Part 10: Administration

Okay, there are several administrative procedures with the point-in-time copy functionality. I will describe only the most important ones, as I don't want to substitute the manual with this tutorial.

Deleting a point-in-time copy configuration Okay, let's assume you used the following configuration so far:

```
# iiadm -l
```

```
dep /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s6 /dev/rdisk/c1d1s4
```

It's really easy to delete this config. As I mentioned before, the name of the shadow volume clearly indicates a point-in-time copy configuration, as there can be only one configuration for any given shadow volume. So you use the name of the shadow volume to designate a configuration. Thus the command to delete the configuration is fairly simple:

```
# iiadm -d /dev/rdisk/c1d1s6
```

The -d tells iiadm to delete the config.

When we recheck the current AVS configuration, the config for /dev/rdisk/c1d1s6 is gone:

```
#
```

Forcing a full copy resync of a point-in-time copy Whenever you are in doubt of the consistency of your point-in-time copy (flaky disks, you've swapped a disk) it may be sensible to force a full copy resync instead of copying only the changed parts. Let's assume the following config of an independent copy:

```
# iiadm -l
```

```
ind /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s3 /dev/rdisk/c1d1s4
```

Again you use the name of the shadow volume to designate the configuration. You force the full copy resync with a single command:

```
# iiadm -c s /dev/rdisk/c1d1s3
```

When we check the status of the dependent copy, you will see that a full copy is in progress:

```
# iiadm -i
```

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s3: (shadow volume)
```

```
/dev/rdisk/c1d1s4: (bitmap volume)
```

```
Independent copy, copy in progress, copying master to shadow
```

```
Latest modified time: Sun Apr 27 01:49:21 2008
```

```
Volume size: 273105
```

```
Shadow chunks total: 4267 Shadow chunks used: 0
```

```
Percent of bitmap set: 69
```

```
(bitmap dirty)
```

Let's wait for a few moments and check the status again:

```
# iiadm -i
```

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s3: (shadow volume)
```

```
/dev/rdisk/c1d1s4: (bitmap volume)
```

```
Independent copy
```

```
Latest modified time: Sun Apr 27 01:49:21 2008
```

```
Volume size: 273105
```

```
Shadow chunks total: 4267 Shadow chunks used: 0
```

```
Percent of bitmap set: 0
```

```
(bitmap clean)
```

The full copy resync has completed.

Grouping point-in-time copies Sometimes the data of an application is distributed over several disks. For example because your application is rather old can use only volumes sized at 2 Gigabytes each. When you want to make a consistent point-in-time copy of all volumes, you have to do it at the same time. To enable the admin to do so, you can group point-in-time copies. When you use the groupname, all members of the group get the commands at the same time.

Okay, let's assume we have an independent copy so far. # iiadm -l

```
ind /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s3 /dev/rdisk/c1d1s4
```

Now we want to configure another one for the volume

```
/dev/rdisk/c1d0s5 with /dev/rdisk/c1d1s5 as the shadow volume and /dev/rdisk/c1d1s6 as the bitmap volume.
```

At first we move the existing configuration into a group. I will name it database in my example but you could choose any other name for it.

```
# iiadm -g database -m /dev/rdisk/c1d1s3
```

With -g we designate the groupname and with -m we move the volume into the group. As usual we use the name of the shadow volume to designate the configuration.

Now we create the point-in-time copy of the second volume. But we will create it directly in the group. To do so, we need the -g switch.

```
# iiadm -g database -e dep /dev/rdisk/c1d0s5 /dev/rdisk/c1d1s5 /dev/rdisk/c1d1s6
```

Please notice, that we used a different copy mechanism for the point-in-time copy. The don't have to be identical in the group.

Let's check the state of our copies:

```
# iiadm -i
```

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s3: (shadow volume)
```

```
/dev/rdisk/c1d1s4: (bitmap volume)
```

```
Group name: database
```

Independent copy

Latest modified time: Sun Apr 27 01:49:21 2008

Volume size: 273105

Shadow chunks total: 4267 Shadow chunks used: 0

Percent of bitmap set: 0
(bitmap clean)

/dev/rdisk/c1d0s5: (master volume)

/dev/rdisk/c1d1s5: (shadow volume)

/dev/rdisk/c1d1s6: (bitmap volume)

Group name: database

Dependent copy

Latest modified time: Sun Apr 27 02:05:09 2008

Volume size: 273105

Shadow chunks total: 4267 Shadow chunks used: 0

Percent of bitmap set: 0

(bitmap clean)Now let's initiate a full copy resync on the group database: # iiadm -c s -g databaseWhen you check the state of your copies again, you will recognize that you initiated a full resync on both copies at the same time:# iiadm -i

/dev/rdisk/c1d0s3: (master volume)

/dev/rdisk/c1d1s3: (shadow volume)

/dev/rdisk/c1d1s4: (bitmap volume)

Group name: database

Independent copy, copy in progress, copying master to shadow

Latest modified time: Sun Apr 27 02:08:09 2008

Volume size: 273105

Shadow chunks total: 4267 Shadow chunks used: 0

Percent of bitmap set: 42
(bitmap dirty)

/dev/rdisk/c1d0s5: (master volume)

/dev/rdisk/c1d1s5: (shadow volume)

/dev/rdisk/c1d1s6: (bitmap volume)

Group name: database

Dependent copy, copy in progress, copying master to shadow

Latest modified time: Sun Apr 27 02:08:09 2008

Volume size: 273105

Shadow chunks total: 4267 Shadow chunks used: 0

Percent of bitmap set: 40
(bitmap dirty)

Posted by Joerg Moellenkamp in Solaris at 09:12

Less known Solaris Features: Point-in-time copy with AVS - Part 9: Disaster recovery

The process of syncing master and shadow is bidirectional. You can't not only update the shadow from the master, you can update the master from the shadow as well. This is a really neat feature for disaster recovery.

Let's assume, you tried a new version of your software. At first all is well, but a minute later the system is toast. Later you will find out, that there was a race condition in the new code, that only manifested on your production system. But you don't know this now. And to add insult to injury, your face go white after looking into the directory.# ls -l

```
/mnt/testindex*
```

```
-rw-r--r-- 1 root root 0 Apr 25 19:40 /mnt/testindex1
```

```
-rw-r--r-- 1 root root 0 Apr 25 19:41 /mnt/testindex2
```

The new code killed your testindex-files. Zero bytes. And you hear the angry guy or lady from customer support shouting your name. But you were cautious, you've created a point-in-time copy before updating the system.

So, calm down and recover before a customer support lynch mob reach your office with forks and torches. Leave the filesystem and unmount it.

```
# cd /
```

```
# umount /mntNow sync the master with the slave. Yes, the other way round.# iiadm -u m /dev/rdisk/c1d1s3
```

Overwrite master with shadow volume? yes/no yesOkay ... after a few moments the shell prompt appears again. Now you can mount it again.
mount /dev/dsk/c1d0s3 /mnt
cd /mntLet's check our work and check the testindex files.
ls -l /mnt/testindex*
-rw-----T 1 root root 1024 Apr 25 18:11 /mnt/testindex1
-rw-----T 1 root root 3072 Apr 25 19:33 /mnt/testindex2Phew ... rescued ... and the lynch mob in front of your office throws the torches out of the window, directly on the car of the CEO (of course by accident)

Posted by Joerg Moellenkamp in Solaris at 08:45

Less known Solaris Features: Point-in-time copy with AVS - Part 8: Working with point-in-time copies

We've created the point-in-time-copy in the last part of the tutorial, but this is only one half of the story. In this part, we will use the this feature for backup purposes. The procedures are independent from the chosen point-in-time copy mechanism.

Okay, let's play around with our point-in time copy. At first we check the filesystem on our master volume mounted at /mnt. Nothing has changed. The AVS doesn't touch the master volume. But now let's create some files.

```
# ls
lost+found test1 test2 test3 test4 testindex1
# touch test5
# touch test6
# mkfile 2k testindex2We check our dependent copy again:# iadm -i
/dev/rdisk/c1d0s3: (master volume)
/dev/rdisk/c1d1s3: (shadow volume)
/dev/rdisk/c1d1s4: (bitmap volume)
Independent copy
Latest modified time: Fri Apr 25 18:16:59 2008
Volume size: 273105
Shadow chunks total: 4267 Shadow chunks used: 0
Percent of bitmap set: 0
```

(bitmap dirty)Please look at the highlighted part. The system detected the changes to the master volume and marked the changed block on the bitmap volumes. The bitmap is "dirty" now.

Okay, now let's use our copy. We create a mountpoint and mount our shadow volume at this mountpoint.
mkdir /backup
mount /dev/rdisk/c1d1s3 /backupJust for comparision, we have a short look at our master volume again:
cd /mnt
ls
lost+found test2 test4 test6 testindex2
test1 test3 test5 testindex1Now we check our copy:
cd /backup
ls
lost+found test1 test2 test3 test4 testindex1We see the state of the filesystem at the moment we've created the point-in-time copy. Please notice the difference. The files created after initiating the copy are not present in the shadow.

You can make everything you want with the filesystem on the shadow volume. You can even write to it. But for this tutorial, we will make a backup from it. Whatever happens with the master volume during this backup, the data on the shadow won't change. Okay, that's isn't so interesting for a few bytes, but important for multiterabyte databases or filesystems. # tar cfv /backup20080424.tar /backup

```
a /backup/ OK
a /backup/lost+found/ OK
a /backup/test1 1K
a /backup/test2 1K
a /backup/test3 1K
a /backup/test4 1K
a /backup/testindex1 1KAs you see, no test5, test6 or testindex2. Okay, we have made our backup, now let's sync our copy.# iadm -u s /dev/rdisk/c1d1s3That's all. What have we done. We told the AVS to update the shadow copy on /dev/c1d1s3. Whenever you specify a disk or volume directly, you use the name of the shadow volume. A master volume can have several shadow volumes, but there can be only one shadow on a volume. So the copy configuration can be specified with the shadow volume. The -u s tells AVS to do an update (not a full copy) to the slave (from the master). Okay, now let's check the copy again.# iadm -i
/dev/rdisk/c1d0s3: (master volume)
/dev/rdisk/c1d1s3: (shadow volume)
```

/dev/rdisk/c1d1s4: (bitmap volume)

Independent copy

Latest modified time: Fri Apr 25 19:30:19 2008

Volume size: 273105

Shadow chunks total: 4267 Shadow chunks used: 0

Percent of bitmap set: 0

(bitmap clean)Please look at the highlighted part again. The bitmap is clean again. The master and the shadow are in sync.

Okay, let's check it by mounting the filesystem.

```
# mount /dev/dsk/c1d1s3 /backup
```

```
# cd /backup
```

```
# ls -l
```

```
total 30
```

```
drwx----- 2 root root 8192 Apr 25 18:10 lost+found
```

```
-rw-----T 1 root root 1024 Apr 25 18:10 test1
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test2
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test3
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test4
```

```
-rw-r--r-- 1 root root 0 Apr 25 18:20 test5
```

```
-rw-r--r-- 1 root root 0 Apr 25 18:20 test6
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 testindex1
```

```
-rw-----T 1 root root 2048 Apr 25 18:20 testindex2It's the exact copy of the filesystem in the moment when you've initiated the copy.
```

Okay, now let's play again with our point-in-time copy. Let's create some files in our master volume:

```
# cd /mnt
```

```
# touch test7
```

```
# touch test8
```

```
# mkfile 3k testindex2Please note, that i've overwritten the 2k sized version of testindex2 with 3k sized version.
```

A quick check of the directories:

```
# ls /mnt
```

```
lost+found test2 test4 test6 test8 testindex2
```

```
test1 test3 test5 test7 testindex1
```

```
# ls /backup
```

```
lost+found test2 test4 test6 testindex2
```

```
test1 test3 test5 testindex1Okay, the directory are different. Now let's start the backup again.
```

```
# tar cfv
```

```
backup20080425.tar /backup
```

```
a /backup/ OK
```

```
a /backup/lost+found/ OK
```

```
a /backup/test1 1K
```

```
a /backup/test2 1K
```

```
a /backup/test3 1K
```

```
a /backup/test4 1K
```

```
a /backup/testindex1 1K
```

```
a /backup/test5 0K
```

```
a /backup/test6 0K
```

```
a /backup/testindex2 2KOkay, test7 and test8 didn't made it into the tarball, as they were created after updating the point-in-time copy. Futhermore we've tared the 2k version of testindex2 not the 3k version. So you can backup a stable version of your filesystem, even when you modify your master volume during the backup.
```

Okay, now we can unmount the filesystem again.

```
# cd /
```

```
# umount /backupAfter this we sync the slave volume with the master volume.
```

```
# iiadm -u s /dev/rdisk/c1d1s3And when we compare the filesystems, they are identical again.
```

```
# mount /dev/dsk/c1d1s3 /backup
```

```
# ls /mnt
```

```
lost+found test2 test4 test6 test8 testindex2
```

```
test1 test3 test5 test7 testindex1
```

```
# ls /backup
```

```
lost+found test2 test4 test6 test8 testindex2
```

```
test1 test3 test5 test7 testindex1You can play this game forever, but i will stop now, before it gets boring.
```

Blog Export: c0t0d0s0.org, <http://www.c0t0d0s0.org/>

Posted by Joerg Moellenkamp in Solaris at 06:30

Saturday, April 26, 2008

Less known Solaris Features: Point-in-time copy with AVS - Part 7: Starting a point-in-time copy

Okay, before using the merits of point-in-time copies, we have to configure such copies. The configuration of this copies is done with the with the `iiadm` command. In this part of the tutorial i will show you how to configure the different kinds of point-in-time copies.

Common prerequisite At first we have to enable the Availability suite. This is independent from the method of doing the point-in-time copy. When you've used the AVS before, you don't need this step

```
# dscfgadm
```

```
Could not find a valid local configuration database.
```

```
Initializing local configuration database...
```

```
Successfully initialized local configuration database
```

If you would like to start using the Availability Suite immediately, you may start the SMF services now. You may also choose to start the services later using the `dscfgadm -e` command.

Would you like to start the services now? [y,n,?] y Please answer the last question with y . By doing so, all the services of the AVS we need in the following tutorial are started (besides the remote replication service)

Create an independent copy Now we can configure the point in time copy. This is really simple. `# iiadm -e ind /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s3 /dev/rdisk/c1d1s4` That's all. What does this command mean: Create an independent copy of the data on the slice `/dev/rdisk/c1d0s3` on `/dev/rdisk/c1d1s3` and use `/dev/rdisk/c1d1s3` for the bitmap. As soon as you execute this command, the copy process starts. We decided to use an independent copy, thus we start a full copy of the master volume to the shadow volume. As long this fully copy hasn't completed, the point-in-time copy behaves like an dependent copy. Now we check the configuration. `# iiadm -i`

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s3: (shadow volume)
```

```
/dev/rdisk/c1d1s4: (bitmap volume)
```

```
Independent copy
```

```
Latest modified time: Fri Apr 25 18:16:59 2008
```

```
Volume size: 273105
```

```
Shadow chunks total: 4267 Shadow chunks used: 0
```

```
Percent of bitmap set: 0
```

```
(bitmap clean)
```

The highlighted part is interesting. The bitmap is clean. This means, that there are no changes between the master and the shadow volume.

Create an independent copy Creating an dependent copy is quite easy. You have just alter the command a little bit, you've used to create independent one. `# iiadm -e dep /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s3 /dev/rdisk/c1d1s4` Just substitue the `ind` with the `dep` and you get a dependent copy. `# iiadm -i`

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s3: (shadow volume)
```

```
/dev/rdisk/c1d1s4: (bitmap volume)
```

```
Dependent copy
```

```
Latest modified time: Sat Apr 26 23:50:19 2008
```

```
Volume size: 273105
```

```
Shadow chunks total: 4267 Shadow chunks used: 0
```

```
Percent of bitmap set: 0
```

```
(bitmap clean)
```

Create an compact independent copy How do you get a compact depedent copy? Well, there is no command to force the creation of such a copy. But it's quite easy to get one. When the shadow volume is smaller than the master volume, the system chooses the compact independent copy automatically. We've created two small slices, when we formatted the harddrive. One of the small slices is `/dev/rdisk/c1d1s6`. Let's use it as the shadow volume. This volume has only a size 32 MB while the master volume is 256 IMB large. At first we create an dependent copy again, but with different volumes: `# iiadm -e dep /dev/rdisk/c1d0s3 /dev/rdisk/c1d1s6 /dev/rdisk/c1d1s4` Now let's check the status of the point-in-time copy configuration: `# iiadm -i`

```
/dev/rdisk/c1d0s3: (master volume)
```

```
/dev/rdisk/c1d1s6: (shadow volume)
/dev/rdisk/c1d1s4: (bitmap volume)
Dependent copy, compacted shadow space
Latest modified time: Sat Apr 26 23:55:05 2008
Volume size: 273105
Shadow chunks total: 1255 Shadow chunks used: 0
Percent of bitmap set: 0
(bitmap clean)Et volia, you´ve configured a compact dependent copy.
```

Posted by Joerg Moellenkamp at 23:27

Less known Solaris Features: Point-in-time copy with AVS - Part 6: Preparation of the test environment

After all this theory, i will go into more practical stuff. In the following parts of this tutorial i will give you an introduction to point-in-time copies with AVS. But at first we have to prepare some things.

At first: We need only one system for this example, so we don´t need any networking configuration. Furthermore you need to assume the root role to do the configuration in this example.

Okay, i will use two harddisks in my example: /dev/dsk/c1d0 and /dev/dsk/c1d1. I´ve chosen the following layout for the disk..

. Partition	Tag	Flags	First Sector	Count	Last Sector	Mount	Directory
2	5	01	0	65480940	65480939		
3	0	00	48195	273105	321299		
4	0	00	321300	80325	401624		
5	0	00	401625	273105	674729		
6	0	00	674730	80325	755054		
8	1	01	0	16065	16064		
9	9	00	16065	32130	48194		

With this configuration i have two 128 mb sized slices. I will use them for data in my example. Additionally i´ve create two 32 mb small slices for the bitmaps. 32 mb for the bitmaps is too large, but i didn´t wanted to calculate the exact size. You will find the exact math behind the size of the bitmap in the manuals.

It´s important to have exactly the same layout on the second disk, at least, when you use independent or non-compact dependent copies. Okay, to be more precise, the slices under the control of the point-in-time copy functionality has to have the same size. To simplify the fulfillment of this requirement, i copy the layout from my master disk to the shadow disk.

```
# prtvtoc /dev/dsk/c1d0s2 | fmthard -s - /dev/rdisk/c1d1s2
```

```
fmthard: New volume table of contents now in place.
Okay, now let´s create a file system for testing purposes on the master disk.
# newfs /dev/dsk/c1d0s3
```

```
newfs: construct a new file system /dev/rdisk/c1d0s3: (y/n)? y
```

```
Warning: 3376 sector(s) in last cylinder unallocated
```

```
/dev/rdisk/c1d0s3: 273104 sectors in 45 cylinders of 48 tracks, 128 sectors
```

```
133.4MB in 4 cyl groups (13 c/g, 39.00MB/g, 18624 i/g)
```

```
super-block backups (for fsck -F ufs -o b=#) at:
```

```
32, 80032, 160032, 240032
Okay, as an empty filesystem is a boring target for point-in-time copies, we play around a little bit and create some files in our new filesystem.
# mount /dev/dsk/c1d0s3 /mnt
```

```
# cd /mnt
```

```
# mkfile 1k test1
```

```
# mkfile 1k test2
```

```
# mkfile 1k test3
```

```
# mkfile 1k test4
```

```
# mkfile 1k testindex1
```

```
# ls -l
```

```
total 26
```

```
drwx----- 2 root root 8192 Apr 25 18:10 lost+found
```

```
-rw-----T 1 root root 1024 Apr 25 18:10 test1
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test2
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test3
```

```
-rw-----T 1 root root 1024 Apr 25 18:11 test4
```

-rw-----T 1 root root 1024 Apr 25 18:11 testindex1 Okay, that's all ... now let's try point-in-time copies.

Posted by Joerg Moellenkamp in Solaris at 23:09

Less known Solaris Features: Point-in-time copy with AVS - Part 5: Compact dependent copy

The compact dependent copy is similar to the normal dependent copy. But this dog knows an additional trick: The shadow and the master doesn't have to be at the same size.

Like the dependent copy this methods uses the concept of the virtual shadow volume. So the bitmap is really important. The bitmap of the compact dependent copy tracks the changes. This exactly as with the normal dependent snapshots. But the bitmap for compact dependent snapshots is enabled to store an important additional information. It's enabled to track, where the system has written the changed data blocks on the shadow volume. So you don't have to write it at the same logical position, you can write it at any position on your shadow volume and it can retrieve the old data with this information.

Deeper dive Okay, at start this picture looks pretty much the same like it's counterpart at the normal dependent snapshots. Please note, the additional information on the bitmap volume. To make the compact dependent copy we just initialize the bitmap again to it's "clean" state.

Let's assume that we change the fourth block on our disk. As with the normal copy, the block is declared as dirty. But now starts to work differently. The original data of the master volume is stored to the first free block on the physical shadow volume. In addition to that the position is stored at the bitmap.

The way to read from the shadow volume changes accordingly. When the bitmap signals, that a block is clean, it just passed the data from the master volume to the user or application. When the bitmap signals a dirty, thus changed block, it reads the position of the block on the physical shadow volume from the bitmap, reads the block from there and delivers it to the application or user.

When we change the next block, for example the third one, the same procedure starts. The original data is stored to the next free block, now the second one, on the physical shadow volume and this position is stored in the bitmap together with the dirty state of the block.

Okay, resyncing the shadow with the master is easy again. Just initializing the bitmap.

Advantages and Disadvantages The trick of storing the position with the dirty state has an big advantage. You don't need master and shadow volumes with the same size. When you know, that only a small percentage of blocks change between two point in time copies, you can size the shadow volume much smaller, thus saving space on your disk. In my opinion compact dependent copies are the only reasonable way to go when you want more than one copy of your master volume. The disadvantages are pretty much the same of the normal dependent copies.

Posted by Joerg Moellenkamp in Sun at 22:19

Less known Solaris Features: Point-in-time copy with AVS - Part 4: Dependent copy

The mechanism of dependent copies was introduced to get rid of this initial sync, as there are circumstances where this initial copy would pose to much load to the system.

The dependent copy uses the bitmap a little bit different. The shadow disk doesn't contain all the data, it just contains the changed data. The bitmap is used to keep track of the block on the masters which have a copy on the shadow.

Deeper dive The dependent copy is one of the two mechanisms in AVS using the concept of the virtual shadow. Thus the model is a little more complex. Let's assume you create an dependent copy. The initialisation is simple. You move no data. You just initialize the data. When an user or application access the virtual shadow volume, it checks in the bitmap if the blocks has changed. If the bitmap signals no change, it just delivers the original block of the master volume.

When you change some data on the master volume, AVS starts to copy data. It copies the original content of the block onto the physical shadow volume at the same logical position in the volume. This is the reason, why master and shadow

volumes have to have the same size when using dependent copies. Furthermore AVS logs in the bitmap that there is data on the shadow volumes, the block is dirty in the bitmap. When you access the virtual shadow volume now, the bitmap is checked again. But for blocks declared dirty in the bitmap, the data is delivered from the copy on the physical shadow volume, for all other "clean" blocks the data comes from the master volume.

Resyncing the shadow to the master is easy. Just reinitializing the bitmap. Now all data comes from the master volume again, until you change some data on it.

Advantages and DisadvantagesThe advantage of this method is it's short time of initialisation and syncing from master to the shadow. It's virtually zero. But you lose the advantages of the independent copy. The dependent copy can only be used in conjunction with the original data. This has two main effects. You can't export the copied volume to a different server and you can't use it as an copy for disaster recovery when your master volume has failed.

Furthermore the master and the slave volume still have to be the same size. But the next method for making a point-in-time copy was derived from the dependent copy exactly to circumvent this problem.

Posted by Joerg Moellenkamp in Solaris at 21:34

Less known Solaris Features: Point-in-time copy with AVS - Part 3: Independent copy

The most important point about independent copies are in their name. The point in time copy is an independent copy of your original data. You can use it on its own and the copy doesn't need the original data to work.

This method is quite similar to doing a copy with dd or cp. At first a full copy is done by the PiT functionality. Now you've created the first point in time copy. But now the advantages of tracking the changes in a database come in to the game. Whenever data on the master is changed, the system tracks this in the bitmap volume. At the next resync of the shadow with the master, the system only copies the changed blocks to the shadow. This vastly reduces the time for the update.

Deeper diveAt first you have your master volume. It's the volume with your data. Before configuration of point-in-time copy both shadow and bitmap copy are uninitialized. We use some special manufactured disks for tutorial purposes. They have only five blocks.

When you've configured the independent copy, a full sync takes place. Each block is copied from the master volume to the shadow volume, and the bitmap volume is initialized. At the end of this process master and shadow are identical and the bitmap is in the "clean" state. No differences between the both.

There is one important fact: After the initial copy, the bitmap doesn't have to be clean. During the full sync the independent copy behaves like a dependent copy. This is done to enable you to use the master volume directly after initiating the independent copy. So, when you change the master volume during the full sync, you will have a "dirty" bitmap (I will explain this condition in a few moments).

Okay, now we change the fourth block of the disk. As the old data is already on the shadow volume, we don't have to move any data. But we log in the bitmap volume, that a block has changed, the block is "dirty". From now the bitmap is in the "dirty" state. The dirty state tells us, that there are differences between the master and the shadow volume.

Okay, we don't need to move data, why do we need the bitmap volume. The bitmap volume makes the synchronisation of master and shadow much more efficient. With the bitmap volume you know the position of changed blocks. So when you resync your shadow with the shadow you just have to copy these blocks, and not the whole disk. After copying the block, the adjacent bit in the bitmap is set to zero, the system knows that the synced block on master and shadow are identical again.

Advantages and DisadvantagesThis has some interesting advantages: You can export these disks in your SAN and use it from a different server. For example you can mount it on a backup server, thus you don't need to transport all the traffic across your local area network.

But there are some disadvantages. You have a high initial amount of data to sync and the size of the master and the shadow have to be equal. This isn't much of a problem of the time needed for the sync (because of the fact, that it behaves as a dependent copy) but it poses more load to CPU and I/O system to copy all the data.

Posted by Joerg Moellenkamp in Sun at 20:51

Less known Solaris Features: Point-in-time copy with AVS - Part 2: Basics

One of this methods is the usage of the point in time copy functionality of the Availability Suite. I've wrote about another function of AVS not long ago when i wrote the tutorial about remote replication. Point-in-time-copy and remote replication are somewhat similar (you detect and record changes and transmit those to a different disk, albeit the procedures are different). Thus it was quite logical to implement both in the AVS.

Availability Suite (AVS)The AVS is a Suite consisting out of two important parts: The "Remote Replication" functionality and the "Point-in-time Copy" functionality. Regular readers of this blog will remember the remote replication as i've already written a tutorial about it. The Availability Suite is integrated to Solaris Express Community and Developer Edition. You can use it for free. It's available for Solaris as well, but when you want support for it, you have to buy the product, as it's a add-on product for Solaris 10

The jargon of Point in Time Copies with AVSOkay, as every technology the mechanisms of Point-in-time copies have their own jargon and i will use it quite regulary in this tutorials.

Master volumeThe master volume is the source of the point in time copy. This is the original dataset

Shadow volumeThe shadow volume is the volume, which contains the point in time copy

Virtual shadow volume There are certain methods to establish a point in time copy, that copies only the original data in the case the data is changed on the master volume. But such a shadow volume is incomplete, as the unchanged parts are missing on the shadow volume . For this the virtual shadow volume was introduced.The idea is simple, but effective. Whenever a block wasn't changed since the last sync of your point-in-time copy, the data is delivered by the master volume. When the block has changed, the data is delivered by the shadow volume. This is transparent to the user or the application, as this virtual shadow volume is created by the AVS point-in-time copy drivers. You access this virtual shadow volume simply by using the name of the shadow volume, even albeit the volume doesn't contain all the needed data.

Bitmap volume All this mechanisms need a logbook about all the changes made to the master volume. This job is done by bitmap volume. Whenever a block on the disk is changed, AVS marks this in the bitmap volume. The bitmap volume is used at several occasions. By using the bitmap volume it can efficienctly sync the master and the shadow volume, you can construct the virtual shadow volume with it in an effcient way.

All types of volumes can be placed on real disk or volumes represented by Veritas Volume Manager or Solaris Volume Manager.

Types of copiesPoint-in-time Copy in AVS supports three types of copies:

independent copies

dependent copies

compact independent copies

All three have a basic idea. Using a bitmap to track changes and using it generate a point-in-time copy. But the methods behind it are quite different. In the next three parts of this tutorial i will dig deeper into this methods.

Posted by Joerg Moellenkamp at 20:43

Less known Solaris Features: Point-in-time copy with AVS - Part 1: Introduction

The basic idea of Point-in-Time copies is the idea to freeze the contents of a disk or a volume at a certain time, thus other processes can work on data of a certain point in time, while your application works on the original dataset and changes it.

Why is this important? Let's assume you want to make a backup. The problem is quite simple. When a backup takes longer than a few moments, the files backup first my represent an different state of the system than the files backup last. You backup is inconsistent, as you've done a backup of a moving target. Okay, you could simply freeze your application, copy it's data to another disk (via cp or dd) and backup it from there or backup it directly and restart your application, but most of the time, this isn't feasible. Let's assume you have a multiterabyte database on your system. A simple copy can take quite a time, in this time your application doesn't work (at least when you database has no backup mode).

Okay, simple copy is to ineffective. We have to do it with other methods. This tutorial will show you the usage of a method integrated into Opensolaris and Solaris.

Posted by Joerg Moellenkamp in Solaris at 20:37

Friday, April 25. 2008

Cringely about Apples acquisition of PA Semiconductor

Cringely has a really interesting opinion about the recent acquisition of PA Semi by Apple. He thinks, that this move will lead us to an Apple system, where even the CPU is a custom-build circuit from Apple optimized for MacOS X. The arguments for his ideas sound quite reasonable.

Posted by Joerg Moellenkamp in Apple at 22:23

Sun buys Montalvo Systems?

The Inquirer reports, that Sun will buy Montalvo System. I don't think that this will lead to Suns own x86 procs. At first Montalvo was founded by people from Transmeta, thus such a move would add some really bright talents to ours and they've worked (and patented this work) on asymmetrical multicore architectures. I assume, this will give us some nice additional capabilities for future multicore CPUs.

PS: I wonder, if anybody thought about asymmetrical cores in the way of having cores with different command sets in one processor ... for example an x86 and a SPARC CPU on one die ...

Posted by Joerg Moellenkamp in Sun at 17:43

Urban decay

Posted by Joerg Moellenkamp in Photographie at 17:38

Walkthrough to Opensolaris 2008.05 RC2

You can download a test image of Opensolaris 2008.05 Release Candidate 2 (the distro formerly and still better known as Indiana). It starts to look really good.

Just to give you a short impression of OpenSolaris 2008.05 i've prepared a short walkthrough to the installation of the distribution, albeit it's so simple you won't need one. The URL of the torrent is available in this announcement mail in the indiana-discuss forum. Please read this announcement completely. But now to the walkthrough....

1. At first, Opensolaris 2008.05 starts with a standard GRUB.
2. Before starting up the graphical desktop, the system asks for the keyboard layout and the language of the live system boot.
3. The graphical Desktop based on Gnome starts up.
4. Now the Licence Agreement appears, read it and then close the window.
5. Now you can play around with the desktop, but i will proceed with the Harddisk install. Please click on the "Install Opensolaris" icon.

6. Okay, the installer starts with some explanations. Press "Next".

7. Choose your installation disk. It defaults to using the the full disk for solaris. At the moment, the installer installs Opensolaris on only one disk.

8. Now have to choose the timezone. We´ve added a world map for easier selection (okay, i find it harder, but your milage may vary. The people like it)

9. For this installlation i´ve choosed English as the language for the installation.

10. Now the system ask for some basic data: root password, an initial user account and a name for the server.

11. The system summarizes your settings. Please press "Install".

12. The system starts to install. You can leave your machine alone for a while and drink a coffee. On my virtual machine on Parallels this took 20 minutes.

13. Now press reboot.

14. Okay, the GRUB again, but now it´s the grub installed on the harddisk.

15. The system starts up in text mode and parses the SMF manifests.

16. The login screen appears. Type in your login ...

17. ... and your password.

18. Now you have your Opensolaris Desktop installed on your system. Play around with it Test it and send bug reports to us. If you'd like submit issues, please feel free to use defect.opensolaris.org, under the distribution/opensolaris product

19. But i want to point you to an important tool: Please click on System -> Administration -> Package Manager. This is

the graphical tool to install software on your system. But wait until the update announcement mentioned in the mail

20. With the Package Manager you can install additional software on your system. It's really similar to tools like synaptic, the graphical apt front end.

Posted by Joerg Moellenkamp in Solaris at 12:58

Phoronix reviews Opensolaris 2008.05

A really positive review of the upcoming Opensolaris 2008.05 release at phoronix - OpenSolaris 2008.05 Gives A New Face To Solaris: Sun Microsystems still has some work to do in improving the OpenSolaris experience, but they are now making some excellent headway. The theme for OpenSolaris 2008.05 is great by our tastes and goes to providing a feel for a richer experience. The biggest feature we feel that has been added to OpenSolaris is the graphical package manager for IPS. Sun Microsystems is clearly trying to use OpenSolaris to increase their desktop OS market share and expecting some users to use the command-line to install packages is not an option

Posted by Joerg Moellenkamp in Solaris at 09:34

solaristutorials.org

As the old link to the list of my tutorials was a little bit long and hard to remember, i had to change this. All tutorials are available at solaristutorials.org (and .de). At the moment it's just a redirection to the old, but restructured page at c0t0d0s0.org.

Futhermore Winnie had the idea of an own feed for the tutorials. So ... here the freshly burned feed.

Posted by Joerg Moellenkamp at 07:44

Thursday, April 24. 2008

Probleme des anderen Geschlechts

Als Mann hat man es ja relativ einfach, was die Ausstattung mit Bekleidung angeht. Farbliche Bemalung der Haut findet nicht statt, jedenfalls bei einem Grossteil der Männerwelt. Ausserdem ist das Auge bei den XY-Chromosomenträgern sowieso nur darauf ausgelegt, drei Rottöne unterscheiden zu koennen. So sind einem doch viele Problem der Frauenwelt ziemlich fremd. Um so göttlicher ist dann dieser Text bei der Katze mit Wut: Hoeschenroulette

Posted by Joerg Moellenkamp in Fundsache at 21:34

Ksplice

Interesting concept: Ksplice: Rebootless Linux kernel security updates. As long a patch doesn't change data structures in the kernel, ksplice can update a running kernel. Ksplice allows system administrators to apply security patches to the Linux kernel without having to reboot. Ksplice takes as input a source code change in unified diff format and the kernel source code to be patched, and it applies the patch to the corresponding running kernel. The running kernel does not need to have been prepared in advance in any way.

Posted by Joerg Moellenkamp in The IT Business at 19:26

Automated exploit generation

A really interesting article about Automatic Patch-Based Exploit Generation: Attackers can simply wait for a patch to be released, use these techniques, and with reasonable chance, produce a working exploit within seconds. Coupled with a worm, all vulnerable hosts could be compromised before most are even aware a patch is available, let alone download it. Thus, Microsoft should redesign Windows Update. We propose solutions which prevent several possible schemes, some of which could be done with existing technology.
(found via Bruce Schneiers blog)

Posted by Joerg Moellenkamp at 11:22

Milax 0.3

The Opensolaris distribution Milax 0.3 was released yesterday: MilaX is a small size Live CD distribution which runs completely off a CD or a USB pendrive. It is based on Solaris Nevada and includes its basic features. What originally started as an experiment to see how much Solaris software could fit in miniCD eventually became a full-fledged OpenSolaris distribution. MilaX also it is possible to use as Rescue-CD. It can be installed on storage media with small capacities, like bootable business cards, USB flash drives, various memory cards, and Zip drives. MilaX is free to use, modify and distribute.

Posted by Joerg Moellenkamp in Solaris at 09:56

Practical Dtrace

Ben Rockwood shows in the blog of his company, how dtrace can help you to find bottlenecks in a fast an efficient manner: DTrace, MySQL, Ganglia, and Digging for Solutions.

Posted by Joerg Moellenkamp in Solaris at 09:25

Less Known Solaris features: fssnap

An ever reoccurring problem while doing backups is the problem, that you have to keep the state of the backup consistent. For example: You use the cyrus imapd for storing mails on a system. As it's a system used for years, the message store is still on an UFS file system. Okay, now there is a little problem: You want to make make backups, but it's a mail server.

With the amount of mail junkies in modern times, you can't take the system offline for an hour or so, just to make a backup. But you have to take it offline. Especially a mail server is a moving target, as cyrus imapd has indexes for its

mailboxes, and index files for the mailboxes itself. Let's assume a backup takes 1 hour, and users delete mails in this time, create mailboxes and so on. It's possible that you backup your mailserver in an inconsistent state, as the mail directory of the user may represent the state one hour ago, but the mailboxes file represent the state of one minute ago.

fssnapUFS has a little known feature, that comes to help in such a situation. You can do a filesystem snapshot of the system. This is a non-changing point-in-time view to the filesystem, while you can still change the the original filesystem. fssnap is a rather old tool. We've introduced in the 1/01 update of Solaris 8.

There is a restriction with this snapshots: This snapshots are solely for the purpose of backups, thus they are not boot persistent. For boot persistent snapshots of a filesystem you will need the Sun Availability Suite.

A practical example. Let's assume that we have a directory called /mailserver. This directory is the mountpoint for a UFS filesystem on /dev/dsk/c1d1s1. At the moment it contains some files: # ls -ls

```
total 10
 2 -rw-----T 1 root  root    1024 Apr 23 01:41 testfile1
 2 -rw-----T 1 root  root    1024 Apr 23 01:41 testfile2
 2 -rw-----T 1 root  root    1024 Apr 23 01:41 testfile3
 2 -rw-----T 1 root  root    1024 Apr 23 01:41 testfile4
 2 -rw-----T 1 root  root    1024 Apr 23 01:41 testindex1
```

Now we want to make a backup. It's sensible to take the mail server offline for a few seconds to keep the files consistent. In this moment we make the filesystem snapshot. This is really easy: # fssnap -o bs=/tmp /mailserver /dev/fssnap/0 With this command we told fssnap to take an snapshot of the filesystem mounted at /mailserver. Furthermore we configured that the snapshot uses the /tmp for it's backing store. In the backing store the changes since the snapshot will be recorded. When fssnap is able to create the snapshot, it will return the name for the pseudo device containing the filesystem snapshot. In our case it's /dev/fssnap/0. Please remember it, we need it later.

When you look at the /tmp directory you will find an backing store file for this snapshot. It's called snapshot0 for the first snapshot on the system: # ls -l /tmp

```
total 910120
-rw-r--r-- 1 root  root    30 Apr 22 14:50 ogl_select305
```

-rw----- 1 root root 465976320 Apr 23 01:47 snapshot0 Now we bring the mailserver online again, and after a few seconds we see changes to the filesystem again (okay, in my example i will do this manually): # mkfile 1k testfile5

```
# mkfile 1k testfile6
# mkfile 2k testindex1
```

```
# ls -l
total 16
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile1
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile2
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile3
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile4
-rw-----T 1 root  root    1024 Apr 23 01:53 testfile5
-rw-----T 1 root  root    1024 Apr 23 01:53 testfile6
-rw-----T 1 root  root   2048 Apr 23 01:53 testindex1
```

Now we want to make the backup itself. At first we have to mount the filesystem. Thus we create a mountpoint # mkdir /mailserver_forbackup Now we mount the snapshot. You will recognize the pseudo device here again. The snapshot is read only thus you have to specify it at mount time: # mount -o ro /dev/fssnap/0 /mailserver_forbackup

Okay, when we look into the filesystem, we will see the state of the filesystem at the moment of the snapshot. testfile5, testfile6 and the bigger testindex1 are missing. # ls -l

```
total 10
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile1
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile2
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile3
-rw-----T 1 root  root    1024 Apr 23 01:41 testfile4
-rw-----T 1 root  root    1024 Apr 23 01:41 testindex1
```

Now we can to a backup without any problems and in a consistent manner: # tar cfv /backup_mailserver_20080424.tar *

```
a testfile1 1K
a testfile2 1K
a testfile3 1K
a testfile4 1K
```

a testindex1 1K After this step we should clean-up. We unmount the snapshot and delete the snapshot with it's backing store file: # umount /mailserver_forbackup

```
# fssnap -d /mailserver  
Deleted snapshot 0.  
# rm /tmp/snapshot0
```

Conclusion With the fssnap command you have an easy way to do consistent backups on UFS filesystems. While it's not as powerful as the functions of the point-in-time copy functionality in the Availability Suite, it's a perfect match for it's job.

Do you want to learn more [fssnap\(1M\)](#)
[fssnap_ufs\(1M\)](#)

Posted by Joerg Moellenkamp at 07:04

Wednesday, April 23. 2008

Signal

Posted by Joerg Moellenkamp in Photographie at 09:13

Tuesday, April 22. 2008

OSnews about Solaris Filesystems

OSNews published a good and positive article about the filesystem choices at Opensolaris - Solaris Filesystem Choices: True, some of these features can be found in any enterprise-ready UNIX OS. But Solaris 10 integrates all of them into one well-tested package
(Thanks to Christopher for the link)

Posted by Joerg Moellenkamp in Solaris at 14:14

links for 2008-04-22

Modifying and Respinning a Bootable Solaris ISO Image
(tags: Solaris ISO)

Gary Voth Photography: The Forgotten Lens
About 50mm lenses
(tags: photography)

BigAdmin Feature Article: Patch Management Best Practices
(tags: Solaris Patching)

Posted by del.icio.us in del.icio.us at 13:34

Solaris as a storage OS

Otmar Meier of otmanix.de wrote a nice summary of all the features in Solaris making it a good choice for storage systems: Solaris als Storageserver. Sorry, it's in german, but the links in the text are very useful even without the rest of the article.

Posted by Joerg Moellenkamp in Solaris at 12:36

Monday, April 21. 2008

What people think about necrophiliacs?

(found via graphjam. Thanks to Chris for the tip)

Posted by Joerg Moellenkamp in Fundsache at 19:24

An Google Ad too far on Facebook

I can't remember, that i allowed Facebook to advertise on Google Search pages with my name. When i search for "Jörg Möllenkamp", this Google Ad appeared:

I've tried the same with some names of my colleagues:

or

Dear Facebook, please stop this!

PS: Interestingly my colleague Constantin Gonzalez has no ads on his google page.

Posted by Joerg Moellenkamp at 18:27

Jupiter

Official informations about UltraSPARC VII aka Jupiter is a little bit sparse at the moment, but a look at Opensolaris and opensolaris.org yields some interesting informations. Bob Hueston collected the existing fragments in his blog

Posted by Joerg Moellenkamp in Sun at 15:48

links for 2008-04-21

Gerhards' Weblog: Amazing easy - Solaris 10 as iSCSI target
(tags: iSCSI Solaris)

Posted by del.icio.us in del.icio.us at 13:32

Mysql and Sun

I planed to write an article about this "Sun will close the source of Mysql" nonsense. Alec Muffet summarizes the real story of this Enterprise/Community Edition discussion on his blog in Whats the SQL? by linking to an authoritative source: A Slashdot article written by Marten Mickos, the former CEO of Mysql AB wait ... a CEO writing at Slashdot????

Blog Export: c0t0d0s0.org, <http://www.c0t0d0s0.org/>

Posted by Joerg Moellenkamp in Sun at 11:27

Sunday, April 20. 2008

links for 2008-04-20

BigAdmin Feature Article: Sun Java System Directory Server 6.0 as an LDAP Naming Service: Part 1 -- Installation and Configuration
(tags: ldap solaris sysadmin tutorial sun howto directory)

Bringing identity home : Media Influencer
Really interesting article about identity
(tags: Identity)

Ten typographic mistakes everyone makes | Life, Tutorials | Receding Hairline
(tags: typography)

Passat CC > Modelle > Volkswagen Deutschland
(tags: newcar)

Posted by del.icio.us at 13:32

Kampnagel

Nicht nur eine Theater- und Zappelbude:

Posted by Joerg Moellenkamp in Photographie at 08:39

The future of mysql (The Project)

Michael Widenius gave an interesting presentation on the MySQL Users Conference: The future of MySQL. He talks about the shortcomings of mysql and their solution. A really interesting read.

PS: I wondered a little bit about the last slide ... many observers thought about it the other way round ...

(found via Kris)

Posted by Joerg Moellenkamp in Sun at 08:22

Saturday, April 19. 2008

About frogs, storks and Linux - or: The nature of pseudo correlations

This is really interesting. I've talked about this effect at my presentation on Friday. There is an effect in statistics called pseudo correlation. You have an pseudo correlation, when you can proof a correlation between two attributes , albeit the real relation lies somewhere else.

A good example: In years with many storks, more babies are born. And well, we were able to proof this relation in the past. But i hope all my reader are aware of the fact, that babies are not delivered by storks. The real corelation is somewhere else. Animals tend to reproduce them more effectively in years with good conditions. And albeit we try to hide it, humans are nothing than mamals. Okay, in former years wet years were good years for growing food. But wet years are good for frogs as well. And storks like frogs ... for lunch. More frogs, more storks. More water, more frogs. More water, more food. More food more babies. Albeit the amount of storks and the amount of babies seems to have a correlation, it's an indirect one.

Okay, getting from mamals, storks, babies and frogs back to the topic of this blog entry. I found an article in SearchEnterpriseLinux.com: "Move from Solaris to RHEL boosts performance for the Chicago Mercantile Exchange. In the first moment i thought: What the fsck ...? But then i've read the article and it's pseudo coreleation time again.

The CME had Sun Server in the past. These Sun systems operated with Solaris. They purchased new servers. They use RedHat on them. Okay, that's bad for us, but hey ... that's life. The strange part are the conclusions in this articles.

What would you think, when you read the headline. "Solaris is so slow ... and Linux is much better". I want Linux. I want this boost, too.

When you read the article, you come to a different conclusion: Their baseline were Sun servers from 2000-2003. That's 8 years ago. In IT this 32 hype cycles (a hype a quarter ... good rule of thumb) in the market and and least 3 real technology cycles (UltraSPARC III, UltraSPARC IV and IV+, Sparc64 VI and CMT) at Sun. They've redesigned their business logic. And obviously their application got faster by the help of Mr. Moore.

And now i come to the big pseudo correlation: The migration to Linux has paid off in improved performance, lower cost and greater stability, Kutty said.Mr. Kutty ... i don't want to insult you, but this is nonsense. Your performace boost didn't came from Linux, it came from using faster servers and redesigning your business logic. At least this is the real conclusion out of this article. I'm pretty sure, that with Solaris under the hood you would have yielded the same effect with new hardware and the rewritten application.

But wait ... when you really read the quote, he is pretty aware of the fact, that it wasn't Linux that gave him the boost and that it was an effect of the migration instead. But SearchEnterprise linux got this totally wrong. So, what's the sense behind this story? There is none ... i assume, at least in regard of the headline.Well ... pseudo correlations are still a standard tactic of marketing and to answer the "Let us know what you think about the story". Well, Mrs. Derringer ... i have to admit .. nothing good.

Posted by Joerg Moellenkamp in The IT Business at 22:05

links for 2008-04-19

YouTube - ZFS-Man
(tags: ZFS fun)

Mac and Solaris: Fix the "xterm-color"• issue
(tags: Mac Solaris)

IBM Ponders Future of System x
(tags: IBM xSeries)

Murphys Law
(tags: article)

Posted by del.icio.us in del.icio.us at 13:34

Successful advertisement

I have to keep this "New Soul" from Yael Naim from my ears. When i hear this music, i have the strong urge to search my credit card and to buy a notebook

Posted by Joerg Moellenkamp in Apple at 08:59

Friday, April 18. 2008

c0t0d0s0@Heldenfunk

Ich habe ja eigentlich schon so jede Schandtut in Bezug auf Bloggen, Twittern (und was es da noch so an neumodischen Diensten gibt) ausprobiert. Nur gepodcastet habe ich noch nicht, das hat sich nun geaendert: Constantin hat mich dazu ueberredet, das ich mich fuer den Heldenfunk interviewen lasse. Also wer schon immer mal meine Stimme hoeren wollte, kann das jetzt mit der vierzehnten Folge des Heldenfunks tun. Zudem wird in diesem Heldenfunk auch ueber die UltraSPARC T2plus basierten Systeme gesprochen.

Die Zusammenfassung von dem was ich so bei Sun tue, ist uebrigens so summarisch fuer die letzten sieben Jahre gesehen. Momentan habe ich den Eindruck, das ich nur noch Vortraege halte Frueher wollte ich nie Verkaeuer werden, heute bin ich Senior Systems Engineer (ein vertreibender Techniker eben). Dann wollte ich nicht marketoide Tuetigkeiten uebernehmen mal sehen, wann das faellt ... bei dieser Praesentationsfrequenz

PS: Constantin,Rolf,Marc ... das am Ende nehm ich Euch uebel ... so stehe ich nicht beim Kunden

Posted by Joerg Moellenkamp in General at 18:35

Hardware review: Ultimate Ears super.fi 4vi

I wrote some days ago in Twitter, that i purchased some Ultimate Ears super.fi 4vi as my old original iPhone phones. This was a little bit tricky, as i wanted headphones with an integrated mic for telephone calls.

Now i used them for a few day and it's time for a recommendation. These headphones are excellent. Good sound, good bass. I suspect, that the users complaining about a weak bass weren't able to seal the ear with the ear pieces. In-ear headphones need to seal the ear to get good bass. Once this seal is closed you get a good and precise bass response from the headphones. My tip: Throw away the original headphones, they are an insult to the sound of an iPhone or iPod, there are much better alternatives.

The build quality seems to be good from my view. The metal casing of the drivers looks excellent. There is a big disadvantage with this ear phones. They close out the ambient sounds around you very effectivly. Nice for hearing music, but somewhat disturbing when you stand on the platform for your train, and you wonder from where all the people come because you didn't heard the incoming train on the other plattform.

They are not effective as active noise cancellation headphones, at least in a turboprop Dornier 328. My QC2 are better in this category, but the sound of super.fi's is better.

The Ultimate Ears burn quite a big hole in your pocket. I've paid 129 Euros at Amazon. That's an huge price tag for some in-ears. But at the moment i think the headphones are worth each Euro i've spend.

Posted by Joerg Moellenkamp in Music at 14:58

No data was harmed in this movie

... but two hard disks had a really bad day. A demonstration of the capabilities of ZFS:

Posted by Joerg Moellenkamp in Solaris at 14:17

Two interesting presentations.

I found two really interesting interesting presentation. At first Brent Paulson wrote an really interesting presentation about practical Opensolaris Security. It's a good overview about the possibilities of Solaris to increase the level of security at your installation. Ben Rockwood of cuddletech.com wrote an good preso about the usage of dtrace in conjunction with mysql.

Blog Export: c0t0d0s0.org, <http://www.c0t0d0s0.org/>

Posted by Joerg Moellenkamp in Solaris at 13:34

Wednesday, April 16, 2008

Less known Solaris Feature: Solaris Security Toolkit

When you want to place a system into a network, it's a good practice to harden the system. Hardening is the configuration of a system to minimize the attack vectors for an intruder by closing down services, configuring stricter security policies and activating a more verbose logging or auditing.

But hardening is not a really simple task: You have to switch off as much services as possible and modify the configuration of many daemons. Furthermore you have to know, what your application needs to run, you can't close down a service that another service needs to execute. Those dependencies may be simple for a server with an apache daemon, but to harden a Sun Cluster needs a little bit more knowledge. Furthermore you have to keep the configuration in a way, that's supported by Sun.

What is the Solaris Security Toolkit? People at Sun do this at their everyday job, thus we have experience to do such hardings. It would be nice to have this knowlege in a framework to automate all this steps. The Sun Security Toolkit was designed with this objective. As the SST website states: The Solaris Security Toolkit, formerly known as the JumpStart Architecture and Security Scripts (JASS) toolkit, provides a flexible and extensible mechanism to harden and audit Solaris Operating Systems (OSs). The Solaris Security Toolkit simplifies and automates the process of securing Solaris Operating Systems and is based on proven security best practices and practical customer site experience gathered over many years. This toolkit can be used to secure SPARC-based and x86/x64-based systems. The SST is an old tool. I use it for years to harden my own systems and in the past any Jumpstart installation made at customer sites contained this toolkit to give them automatic hardening. Futhermore it's a good practice to use this toolkit on freshly iinstalled systems as a first step in the deployment process of new server hardware before you start to install your application.

How to install the Solaris Security Toolkit? Installation of the Toolkit is really easy. At first you have to gather it from the Sun Download Center. Sorry, you need a account for it, but you can register for free. You will find it here. Before login in as root, i've copied the file SUNWjass-4.2.0.pkg.tar.Z via scp to my freshly installed system with Solaris 10 Update 5#

```
cd /tmp
# ls
SUNWjass-4.2.0.pkg.tar.Z  hsperrdata_root      typescript
hsperrdata_noaccess    ogl_select216
# bash
# uncompress SUNWjass-4.2.0.pkg.tar.Z
# tar xfv SUNWjass-4.2.0.pkg.tar
x SUNWjass, 0 bytes, 0 tape blocks
x SUNWjass/pkgmap, 33111 bytes, 65 tape blocks
[...]
x SUNWjass/install/preremove, 1090 bytes, 3 tape blocks
x SUNWjass/install/tsolinfo, 52 bytes, 1 tape blocks
Now let's install the package:
# pkgadd -d . SUNWjass
```

Processing package instance from

```
Solaris Security Toolkit 4.2.0(Solaris) 4.2.0
Copyright 2005 Sun Microsystems, Inc. All rights reserved.
Use is subject to license terms.
Using / as the package base directory.
## Processing package information.
## Processing system information.
## Verifying package dependencies.
## Verifying disk space requirements.
## Checking for conflicts with packages already installed.
## Checking for setuid/setgid programs.
```

Installing Solaris Security Toolkit 4.2.0 as

```
## Installing part 1 of 1.
```

```
/opt/SUNWjass/Audit/disable-llim.aud
[...]
/opt/SUNWjass/sysidcfg
[ verifying class ]
```

Installation of was successful. Now i can use the Sun Security Toolkit for system hardening. It's installed at /opt/SUNWjass/

A look into the frameworkAt the end the SST is a collection of scripts and files and some code to distribute those changes throughout the system.

Each script was developed to execute a single job in the process of system hardening. When you look into the directory /opt/SUNWjass/Finish you will find a vast amount of them. For example enable-bart.fin for the automatic execution of BART to generate a system baseline or disable-remote-root-login.fin to automatically disable root logins, when an admin had activated those logins.

On the other side you will some configuration files in the Sun Security Toolkit as well. Sometimes a service needs some additional configuration for hardining, for example an really paranoid hosts.deny. Those configuration templates are contained in the directory /opt/SUNWjass/Files.

But you don't use both types of directly ... you use them in collections called drivers. You find those drivers in /opt/SUNWjass/Drivers. These drivers execute all such scripts in a certain sequence. Some drivers are really simple ... they just call other drivers. A good example is the secure.driver which just calls the hardening.driver and the configuration.driver. The hardening.driver is a better example. This driver calls many of the scripts mentioned before.

At this moment, you should take some time and examine some files in /opt/SUNWjass/Drivers, /opt/SUNWjass/Files and /opt/SUNWjass/Finish to get more insight into the inner workings of SST. Don't be shy, it's just clever shell scripting.

Use the Solaris Security Toolkit for hardeningA tip from the field at first: Open a second ssh connection to the system, login as root, and leave this window untouched and minimized. This connection is quite handy to have a second limb, when you sawed away the one you sit on.

Okay, how do you use such a driver? This is really simple. But you shouldn't execute the drivers blindly. They can lock you out of your system in the case you use them without caution. So change in to the directory for the Drivers:# cd /opt/SUNWjass/DriversNow you should look into the desired drivers. An example: The hardening.driver contains a like to disable the nscd. disable-nscd-caching.fin

But you want another behaviour for some reason. You just have to add an # in front of the line:# disable-nscd-caching.finWell, there is another behaviour i don't want. The default locks sshd via tcpwrapper to accesses from the local host. But there is a better template at /opt/SUNWjass/Files/etc/hosts.allow allowing ssh access from all hosts. You can force SST to use it by adding another line to the hardening.driver. I've added the bold line to do so:JASS_FILES="

```
    /etc/hosts.allow
    /etc/dt/config/Xaccess
    /etc/init.d/set-tmp-permissions
    /etc/issue
    /etc/motd
    /etc/rc2.d/S00set-tmp-permissions
    /etc/rc2.d/S07set-tmp-permissions
    /etc/syslog.conf
```

" Now the Toolkits copies the file /opt/SUNWjass/Files/etc/hosts.allow to /etc/hosts.allow. As you may have noticed, the template as to be in the same directory as the file you want to substitue with the difference, that the directory of the template has to be relative to /opt/SUNWjass/Files/ and not to /

Okay, now we have modified our driver, now we can execute it:# cd /opt/SUNWjass/bin
./jass-execute secure.driver

[NOTE] The following prompt can be disabled by setting JASS_NOVICE_USER to 0.

[WARN] Depending on how the Solaris Security Toolkit is configured, it is both possible and likely that by default all remote shell and file transfer access to this system will be disabled upon reboot effectively locking out any user without console access to the system.

Are you sure that you want to continue? (yes/no): [no]

yesThis warning is not a joke. Know what you do, when you use this toolkit. Hardening means "real hardening" and this process may leave you with a paranoid hosts.allow locking you out from accessing the sshd on your system. Without console access you would be toast now. But as we use the more sensible template for hosts.allow, we can proceed by answering with yes:

Executing driver, secure.driver

```
=====
secure.driver: Driver started.
=====
```

```
=====
Toolkit Version: 4.2.0
Node name:   gondor
Zone name:   global
Host ID:     1911578e
Host address: 10.211.55.200
MAC address: 0:1c:42:24:51:b9
OS version:  5.10
Date:       Wed Apr 16 16:15:19 CEST 2008
=====
```

After a first status report, a long row of scripts will print log messages to the terminal. For example the finish script enable-coreadm.fin:=====

```
==
secure.driver: Finish script: enable-coreadm.fin
=====
```

Configuring coreadm to use pattern matching and logging.

[NOTE] Creating a new directory, /var/core.

[NOTE] Copying /etc/coreadm.conf to /etc/coreadm.conf.JASS.20080416161705

coreadm change was successful.Many reports of this kind will scroll along and at the end jass.execute prints out some diagnostic:=====

```
secure.driver: Driver finished.
=====
```

```
=====
[SUMMARY] Results Summary for APPLY run of secure.driver
[SUMMARY] The run completed with a total of 97 scripts run.
[SUMMARY] There were Failures in 0 Scripts
[SUMMARY] There were Errors in 0 Scripts
[SUMMARY] There were Warnings in 3 Scripts
[SUMMARY] There were Notes in 81 Scripts
=====
```

[SUMMARY] Warning Scripts listed in:
/var/opt/SUNWjass/run/20080416161504/jass-script-warnings.txt

[SUMMARY] Notes Scripts listed in:
/var/opt/SUNWjass/run/20080416161504/jass-script-notes.txt

=====When you look around at you system, you will notice some new files. Every file in the system changed by the SST will be backedup before the change is done. For example you will find a file named vfstab.JASS.20080416161812 in /etc. JASS contains a finish script to limit the size of the /tmp. As the /tmp filesystem resides in the main memory, this is a sensible thing to do.

Let's check for the differences:# diff vfstab vfstab.JASS.20080416161812
11c11

Blog Export: c0t0d0s0.org, http://www.c0t0d0s0.org/

```
< swap - /tmp tmpfs - yes size=512m
```

```
---
```

```
> swap - /tmp tmpfs - yes -The script has done it's job and added the size=512m option to the mount.
```

Effects of the hardening

Many effects are not directly obvious like changes to security or password policies. You will recognize them, when the system forces you to change your password and when it's getting harder to change a new one because of higher requirements for new passwords enforces by solaris.

A more obvious change is the new message of the day. It doesn't print out the version. The new version is a little bit ...
uhmmm ... more unfriendly: aragorn:~ joergmoellenkamp\$ ssh jmoekamp@10.211.55.200

Password:

Last login: Wed Apr 16 19:12:19 2008 from 10.211.55.2

```
|-----|
| This system is for the use of authorized users only.          |
| Individuals using this computer system without authority, or in |
| excess of their authority, are subject to having all of their   |
| activities on this system monitored and recorded by system    |
| personnel.                                                    |
|                                                                |
| In the course of monitoring individuals improperly using this  |
| system, or in the course of system maintenance, the activities |
| of authorized users may also be monitored.                    |
|                                                                |
| Anyone using this system expressly consents to such monitoring |
| and is advised that if such monitoring reveals possible       |
| evidence of criminal activity, system personnel may provide the |
| evidence of such monitoring to law enforcement officials.     |
|-----|
```

Undo the hardening All these file ending with .JASS. are not just for you to lookup the changes of SST This files enables the SST to undo the changes and to fall back to a different configuration.

```
# ./jass-execute -u
```

```
Executing driver, undo.driver
```

Please select a Solaris Security Toolkit run to restore through:

1. April 16, 2008 at 16:15:04 (/var/opt/SUNWjass/run/20080416161504)

Choice ('q' to exit)? 1

The choice is easy, we have just one old version, so we choose 1

[NOTE] Restoring to previous run from /var/opt/SUNWjass/run/20080416161504

```
=====
undo.driver: Driver started.
=====
```

You may have changed some files since using the toolkit, thus the SST will ask you if you what it should do with those files. For example, i've changed the password of my account, thus the /etc/shadow has changed:

```
=====
undo.driver: Undoing Finish Script: set-flexible-crypt.fin
=====
```

[NOTE] Undoing operation COPY.

[WARN] Checksum of current file does not match the saved value.

[WARN] filename = /etc/shadow

[WARN] current = 8e27a3919334de7c1c5f690999c35be8

[WARN] saved = 86401b26a3cf38d001fdf6311496a48c

Select your course of action:

1. Backup - Save the current file, BEFORE restoring original.

2. Keep - Keep the current file, making NO changes.

3. Force - Ignore manual changes, and OVERWRITE current file.

NOTE: The following additional options are applied to this and ALL subsequent files:

4. ALWAYS Backup.
5. ALWAYS Keep.
6. ALWAYS Force.

Enter 1, 2, 3, 4, 5, or 6:

2

After this command the Solaris Security Toolkit has reverted all changes.

Conclusion With the Solaris Security Toolkit you can deploy an certain baseline of security configurations to all your systems in an automatic manner. But it isn't limited to run once after the installation, you can run it as often as you want to ensure that you get to an known secured state of your system after patching or reconfigurations of your system. By doing this automatically, you get a big advantage. Once you've developed your own driver for your site, nobody forgets to set a certain configuration leaving an attack vector open, you assumed as being closed down.

This tutorial demonstrated only a small subset of the capabilities of the toolkit. For example you can integrate it into Jumpstart to automatically harden systems at their installation, you can use it to install a minimal patch cluster on each system where you execute the toolkit. So you should really dig down into the documentation of this toolkit to explore all the capabilities.

Do you want to learn more?

Documentation

docs.sun.com: Solaris Security Toolkit 4.2 Administration Guide

Posted by Joerg Moellenkamp in Solaris at 19:52

About the UltraSPARC T2plus memory controller discussion

This is really interesting in regard of the discussion about the reduced amount of memory controller in the UltraSPARC T2: Stream Benchmark Results on Sun SPARC Enterprise T5240 Server. On a two socket UltraSPARC T2plus systems the STREAMS TRIAD benchmark yields a result of 30 GB per second. On a single socket UltraSPARC T2 the same benchmark (executed by a customer) yields 16 GB/s. Obviously counting the amount of controller circuits isn't enough to estimate the memory bandwidth a system is capable to use or to draw some more or less sensible conclusions out of it (as TPM tried to do).

Posted by Joerg Moellenkamp at 12:00

Tuesday, April 15. 2008

Opensolaris 2008.05

Another important Solaris release isn't that far in the future: Sun plans to release the first version of the Opensolaris distribution named Opensolaris 2008.05 in May. Jyothi Srinath linked in her blog to a preliminary version of the "Getting Started" documentation. It's a first public version, so feel free to help by sending suggestions to indiana-discuss@sun.com.

Posted by Joerg Moellenkamp at 20:19

Out now: Solaris 10 5/08 aka Update 5

Solaris 10 Update 5 aka 5/08 went live: You can download at the usual place. You will find a short overview of the new features in earlier blog entry: [What's New in the Solaris 10 5/08 Release?](#)

Posted by Joerg Moellenkamp in Solaris at 19:51

Pac-Man

The only way to wake up your audience after 20 slides with statistics and pie charts

(inspired ... eehm ... stolen from presentation zen)

Posted by Joerg Moellenkamp at 18:59

Ich werde es nicht mehr erleben ...

Ich werde den Tag nicht mehr erleben, an dem der gemeine S-Bahnfahrer endlich lernt, das man sich nicht genau vor die Tür der S-Bahn stellt um auszusteigen, während die im Zug befindlichen Personen noch dabei sind, aus dem Einstiegsbereich zu quellen ...

Posted by Joerg Moellenkamp in Bahn at 18:41

At last: Sun Fire X4140 and X4440

Sun introduced the Sun Fire X4140 and X4440 today. The X4140 is dual socket 1 RU server with up to 8 internal disks:

But the X4440 is even more interesting. We hadn't a quad socket Opteron system in the portfolio for quite a time, but this gap is closed now. As far as i remember, this is the first 2U quad socket Opteron server from a Tier-1 vendor.

The mechanical construction is somewhat similar to the Xeon based system. They use a mezzanine board too, but unlike the Xeon mezzanine board, which houses only memory, the board of the X4440 houses two sockets and the adjacent memory of both procs.

At the moment, both system are available with dualcores only, but quadcores will be announced really soon.

Posted by Joerg Moellenkamp in Sun at 15:14

links for 2008-04-15

Sun, Solaris, and a new chance to shine | The Open Road - The Business and Politics of Open Source by Matt Asay - CNET Blogs
(tags: Sun Solaris)

Hunderttausende Chinesen klicken tagesschau.de-Umfrage | tagesschau.de
(tags: politik propaganda china)

Linux.com :: Commentary: the Linux Foundation and the future of Linux
Is Linux still controlled by the users and developers or is it an shared asset of some of the big companies in this field?
(tags: business server desktop Linux)

Home - cmt - wikis.sun.com
The CMT wiki at wikis.sun.com
(tags: CMT Sun)

Posted by del.icio.us in del.icio.us at 13:42

Paul Murphy about UltraSPARC T2plus

As usual Paul brings an interesting perspective into discussion: On the other hand.. the way the processors are coupled - done by replacing the the T2's on board 10Gbyte facility - demonstrated that Sun can now produce highly customized versions of the core CPU set and suggests what I believe may be a unique performance opportunity for this product line. Taking into consideration the additional transistor budget by the usage of upcoming process technologies, it should be feasible to integrate other interconnect technologies as well, for example Infiniband on die (just think about the latency advantages of an Infiniband port directly connected to the crossbar) or the integration of additional support circuits for special tasks. I assume, there is only one limit ... the pin count of the proc ... at a certain point you can't get all the interfaces out of the chip in an economical feasible manner.

Posted by Joerg Moellenkamp in Sun at 12:24

Location-aware Twitter for the iPhone: Twinkle

I revived my twitter account a few days ago, when I discovered, that my preferred feedreader (Netnewswire) is capable of sending URLs to twitter with a single command. So I use twitter as a public clipboard again for links I do not want to put on del.icio.us. Yesterday I've found yet another really neat tool for my iPhone: Twinkle. Twinkle is a Twitter-client for the iPhone, but with a spin. You can add location informations to your tweets (the messages in twitter). Resulting from this information you can query for tweets that were published by people near of you:

In the message entry dialog you can upload images and it let you choose if you want to add the location. As the iPhone has no GPS, I assume that the location is derived from the WLAN and GSM footprint at your location.

Okay, you can argue about the sense of Twitter, but given enough computing power and a good user interface at a mobile, services start to converge in an interesting manner, and Twinkle is an interesting example. You can download on your phone via Installer.app

PS: And think about that: 1995 - 13 years ago - nobody forecasted that sending 140 chars long messages in the GSM control channel would be such an roaring success.

Posted by Joerg Moellenkamp in iPhone at 06:48

Monday, April 14. 2008

Is hypervisor based virtualisation the correct way?

I thought a little bit about virtualisation this weekend after some interesting comments of a customer about his experiences with virtualisation.

Are hypervisor based virtualisation mechanisms really the way to go? Think about the timing problems of VMware. The problems to virtualize dozens of timer ticks, eating away a good amount of the capacity of your system. Or the problems with I/O intensive tasks. You can throw more CPUs to the problem. But: I'm tempted to think, that solutions like branded zones are a more viable solution to virtualize for unix systems. You stay with a single kernel infrastructure mapped to different target operating systems by special mechanisms ... like the branded zones, perhaps in conjunction with resource management. Looks more efficient to me than simulating the complete system.

Posted by Joerg Moellenkamp in The IT Business at 18:08

Wisdom of the day: About Opensource

This quote found at Dave Edstrom's blog made my day ... "There are two types of users - those who are ready to spend a lot of time in order to save money, and those who are ready to spend a lot of money in order to save time."
- Mårten Mickos, CEO MySQL

Posted by Joerg Moellenkamp in The IT Business at 16:39

"Real" Grid Computing

My brother is consultant for increasing the energy efficiency of buildings. Not long ago i had an interesting discussion with him: In IT we use the example of the electricity grid and a central power generating plants as the future of a utility based computing. The problem: My brother told me, that this is a model of the past. There is a massive trend in the generation of electricity away from the big central plant, the future is the small decentralized power plant to generate the electric power near the consumers of power to get rid of the transport losses (those losses accumulate to an primary energy factor of 1:3. For each consumed kilwatt, you have to put 3 into the power grid).

There are even thought games about virtual power plants by generating electrical power directly at the households. All this micro plants are operated and controlled by a central control. When you need more energy in a region, the microplants in the region start to produce more energy to feed the energy in the regional grid.

Perhaps this is the way to go in utility computing. The systems are at the consumer premises, operated and managed by a central . When other customers has a surge in it's need for more compute power , the compute power is generated by the micro-datacenters at the others sites, for example at sites, where the customer can't load it's compute system (a good argument for cryptography everywhere). In this case you have a fast and low-latency access to your compute power at normal operation but access to a vast amount additional capacity. Smaller customers in a region could even start without own machines and when they demand more than a certain amount of capacity, and own micro compute plant will be installed at this site.

Posted by Joerg Moellenkamp in The IT Business at 10:30

Sunday, April 13. 2008

A system in crisis

A really interesting presentation about the economy of materials - The Story of Stuff:

Posted by Joerg Moellenkamp at 19:42

links for 2008-04-13

Memo to Petraeus
About the usage of language
(tags: awesome)

Posted by del.icio.us in del.icio.us at 13:41

On market trends

I've thought a little bit about prediction of market trends. There is a large plethora of companies making a living out of predicting the next big thing. I find this forecasts of the future a little bit suspicious. Or to ask the question in a different way: Are this predictions good forecasts or just self fulfilling prophecies.

Market trend: Virtualisation

At the moment, i opt for the second one. My favorite example for this is the rise of virtualisation. Okay, i hear the "Oh, no, not a virtualisation rant again" in front of a dozens monitors, but this article is not about the "virtues" of virtualisation.

From my view, the hype surrounding virtualisation is a combination of consultancy companies searching for topics, sales peoples searching for hardware business, CIOs seaching for solutions for the sins of the past and the media searching for coverstories.

Believe it or not, many customers are in a group we call "blueshifters" at Sun. The don't want to talk about bigger new systems, they want to talk about opportunities to make their IT much cheaper, as they think about IT as a cost factor, not a competitive weapon.

The problem: Servers do not rot away. A decent server will run for years. So when you already have reaction times in a dialog software in the subsecond range on a slow and old server, customers think twice about buying a new system. Or better: They won't think about an update.

So you need a reason to sell new hardware, new consultancy services and so on. Thus a new hype topic is created. A trend is born: Virtualisation. And now you have a new opportunity to sell new systems. You calculate an OPEX reduction with the customer and maybe the customer will buy a new server. The natural path for such customers will be the reduction to just two systems. The singularity of OPEX driven system architecture. But thats a problem for the account manager in a few years ...

The interesting question: Did the need for virtualisation really existed before the industry stepped on this idea or did we create the trend? Is the next big thing hyped by analysts really choosen on business needs. Okay, it's choosen by business needs, but are they choosen on the needs of the end customer or on the business needs of the consultancy companies?

And such a predicted market trend can a self fulfilling prophecy very fast: Let's assume a technical race that's absolutly undecided. A well known analyst company comes to the conclusion, that the technology A will win this race. The usual suspicious IT media sites report about this. Customers read this articles and talk about this with their vendors. No insult

here, there are so many trends out there and not much time . Most salesman won't try to bet against a market trend. This effect is amplified by regular survey about hot topics for customers. Often you can correlate them to the hot topics in media. The salesman will sell the Technology A to them, technology B doesn't get sold, thus research on it will be decreased or it will be discontinued. Technology A won't because of a real advantage, just because of an avalanche effect.

The role of the gut at making decisions

The human being is a herd animal by nature. There are several examples that people walked into a disaster, just because some alpha "animals" chose to do so. Most decisions are not rational ones, they are gut decisions at a large part. They are made from the stomach. You may say: Those people are professionals. Decisions are made on the basis of comprehensible and adequate criteria, not on the basis on market trends. But this is simply not true. The more the differences of products are in the depth of the technology only comprehensible to experts and the more difficult and complex a technology is, the more the gut decides.

To get back to the "walking to disaster" example ... i should tell a story from the field: We've tried to sell a product to a customer. Our proposal was more thought-out and the product superior. The problem: We all had a bad day, thus our presentation wasn't really a good one, i felt from one dead faint to another at that day. We left with a customer unsure of our capability to execute, although we implemented the solution before and after the proposal of this customer. Obviously we lost the deal. A year later the implementation of our competitor ended in a disaster. Project stopped after really a vast amount of money spend, not a single objective fully reached. We knew it before, but our doubts were dismissed as trying to win back the deal.

This was a really important experience for me. When you lose gut of the alpha animals at a customer, even an otherwise extremely professional will decide against you. Even when you decide about a multi-million project, it's more about the gut feeling than about a rational decision.. When a customer is unsure about a technology, the customer will follow the herd, aka a perceived or real market trend or the salesmen and experts giving him a good feeling about a decision.

So what?

This combination of customers searching for guidance at betting their career on a technology, analysts searching for the next big thing, salesmen searching for the easier sell, media searching for something to write about to fill the pages between the ads leads to a situation, where i really believe that market trends in IT are just self fulfilling prophecies at a large portion.

After all, what should you think about market trends? In my opinion you should ignore them. Completely. Make rational decision make on hard data and proof-of concepts matching on your own needs. And my personal opinion: Vendors should monitor market trends closely, but not obey them slavishly. At the end you stop to act, and start to react only. And this is a safe way to disaster.

Posted by Joerg Moellenkamp at 12:45

Saturday, April 12, 2008

Flag-day for ZFS boot support

The support for booting from ZFS had it's flag day yesterday. It looks like this will appear in build 88 for x86 and SPARC. I hope this find it's way in Update 6 of Solaris, stopping this ever reoccurring question at presentation about this feature.

Posted by Joerg Moellenkamp at 21:04

What's New in the Solaris 10 5/08 Release?

The next release of Solaris 10 is imminent. As the Sun website states: Solaris 10 5/08 will be available for download on 4/16/08. So it may be interesting to you, what new features found their way in 5/08. Well, you can already look into the What's New in the Solaris 10 5/08 Release document.

System Resource Enhancements

- CPU Caps
- projmod(1M) Option

Device Management Enhancements

- Tape Self-Identification
- x86: Enhanced Speedstep CPU Power Management
- iSNS Support in the Solaris iSCSI Target

Security Enhancements

- Solaris Trusted Extensions Supports Mounting Labeled File Systems With the NFSv3 Protocol
- SPARC: Hardware -Accelerated Elliptical Curve Cryptography (ECC) Support

Networking Enhancements

- Sockets Direct Protocol
- Tunable inetd Backlog Queue Size

X11 Windowing Enhancements

- Xvnc Server and Vncviewer Client

System Administration Enhancements

- Solaris Trusted Extensions Administrator Procedures
- Flash Update Tool
- PPD File Management Utility
- Internet Printing Protocol Client-Side Support
- Selectable Use of localhost for Solaris Print Server Database Hostname
- Fault Management for T5140/T5240 Platforms
- SunVTS 7.0

Desktop Tools Enhancements

- new version StarOffice 8
- Flash Player 9
- Pidgin 2.0
- PAPI Print Commands

System Performance Enhancements

- 64-bit SPARC: Memory Placement Optimization Support for sun4v Platforms
- SPARC: Shared Contexts Support
- x86: CPUID-Based Cache Hierarchy Awareness

Language Support Enhancements

Locale Creator
libchewing 0.3.0
File Encoding Examiner

Kernel Functions Enhancements
x86: MONITOR and MWAIT CPU Idle Loop

Driver Enhancements
x86: Support Sun Fire X4540 Disk Status Indicators
MPxIO Extension for Serial Attached SCSI Devices on mpt(7D)
x86: SATA ATAPI Support in AHCI Driver
x86: GLDv3 Version bnx II Driver
x86: AMD-8111
SATA NCQ Support in AHCI Driver
x86: bnx II Ethernet Driver
USB-to-Serial Driver for Keyspan Adapters

Freeware Enhancements
32-bit: pgAdmin III
p7zip

Posted by Joerg Moellenkamp in Solaris at 18:58

Do not walk over the apron

Processes on the apron of an airport are really tight ... yesterday i've got aware of this fact again. Yesterday evening i've took the plane to Hamburg, Gate 22 in MUC. A "finger" position. The plane was already there ... but the finger was out of order, it collided with a ground vehicle. Albeit the plane was visible, we had to board the bus, drive along of half of the terminal, then to the apron and the complete distance back to the plane just to board it from the rear exits. The gag: we stood directly in front of the plane ... maybe 5-10 meters. But you aren't allowed to simply walk to the stairs, you have to drive by bus.

BTW: This was better than on the flight to MUC. We were on time, standing on our outside parking position. This was seemingly so unusual that we had to wait quite long for some ground personal to move a stair to the plane.

Posted by Joerg Moellenkamp at 14:59

UltraSPARC T2+ Scaling

Stefan Hinker looks to an interesting fact in regard of our new UltraSPARC T2+ systems: They are scaling really well.

BenchmarkT5x20 T5x40 SkalierungSPECint_rate2006 78.5 157x2 SPECfp_rate2006 62.3 119x1.91 SAP SD
2-tier 2,175 4,170x1.91 Lotus R6 iNotes 43,000 65,000x1.51 SPECjbb2005 192,055 373,405x1.94

Another interesting fact: From view of the operating system this was a jump from 64 CPUs to 128 CPUs, thus the results are even more impressive. Additionally this should stop those comments about the fact, that the UltraSPARC T2 had two memory controllers more than the T2+. You don't reach a scaling factor near of two in SAP with a proc starving for memory bandwidth.

BTW: I really wait for public results of our quad-socket system

Posted by Joerg Moellenkamp in Sun at 14:31

Friday, April 11. 2008

Last preso of the week

Well, let's call this a somewhat tough week. My virtualisation preso today was really good. Many good questions and good feedback. I'm sitting in the FTL lounge in MUC now ... batteries empty. Need a weekend. Thank god it's friday. Normal blog operations will resume on monday ...

Posted by Joerg Moellenkamp in Business Travel at 16:46

Thursday, April 10, 2008

@home @last

I'm at home. At last. This was really a day to make two of it and still having some spare hours. It began with sleeping really bad last night. Looked at the clock at 3, at 4 and at quarter past 5. At six, i stood up, but was still tired. 9 o'clock meeting, the first presentation of the day at quarter past 10. Presented with Constantin for round about an hour. Left the hotel in Schwaebisch-Hall much to late. Standard road back to Stuttgart. Well, this way stopped right after the city border of Schwaebisch-Hall traffic jam. Forced me to drive the scenic but slower road to Stuttgart. Returning my rental car, no time for refueling it.

And it was still to late ... but i had luck ... as the loading didn't started, they took my luggage. I think all my charm (yes, i have something like that) , a really honest and nice smile (yes, i can do something like that) and my FTL card got me in that plane. I think, the real difference. Was a good flight to TXL ... besides of spilling the rest of my apple juice on my trousers. Well, i made it to the customer (an education partner ... i've mad a presentation in front of their customers. It was a "Solaris at a glance" presentation) At 17 o'clock the next presentation. It started with the usual beamer problems ... as soon ultra portable beamers for the notebook bag come available, i will buy one ... at the end we changed rooms because the beamer are fixed in the rooms.

I've finally made the train to hamburg at quarter past eight and now i'm finalizing the last slides for my presentation tomorrow. Have a good night

Posted by Joerg Moellenkamp at 22:45

links for 2008-04-10

RAS in the T5140 and T5240
Interesting informations about RAS and our new systems
(tags: UltraSPARCT2+ Sun RAS)

Posted by del.icio.us in del.icio.us at 13:38

Being partisan

Sometimes it's really easy to see, if a writer is partisan. Okay, i'm pro-sun ... that's obvious and i'm admitting that. Hey, i'm working for Sun. But i saw a really nice example for an partisan "independent" writer. He criticize some design decisions at UltraSPARC T2+over a length of two paragraphs. That's okay, but it throws an interesting light, when you celebreate a watercooled IBM p575 in a nonstandard formfactor of 23" width as the way to go in computing at the same time.

But, hey ... it was pretty obvious that TPM is heavily IBM biased before as i'm biased for Sun. But at least I don't pretend to be independent

Posted by Joerg Moellenkamp in The IT Business at 08:06

Wednesday, April 9, 2008

Only fly-fishing?

A really interesting reflection in the sunglasses of Dick Cheney

(found via fefe)

Posted by Joerg Moellenkamp in Fundsache at 22:19

Podcasting presentation

This was the presentation i´ve finalized on this weekend. It was intended for an internal audience in Germany to introduce the idea of podcasting to them in the context of a project i´m working on together with Constantin Gonzalez. I think i will add an audio track to this preso soon:

PS: The diagram on page 58 is from the wikipedia entry for podcasting.

Posted by Joerg Moellenkamp in Blogosphere at 21:57

Solaris 9 Containers for Solaris 10

There was an additional announcement today. It´s at the end of the T5140 release: Today, we´ve announced the availability of Solaris 9 containers in Solaris 10 in addition to the Solaris 8 containers, thus it get really easy to migrate you existing Solaris 9 servers to CMT servers (which need Solaris 10). There is even an P2V tool to make this even more easier.

Posted by Joerg Moellenkamp in Sun at 18:46

UltraSPARC T2+ and the Sparc Enterprise T5240

Today Sun and Fujitsu announced the the first new multiproc Victoria Falls systems: Sun Microsystems And Fujitsu Expand SPARC Enterprise Server Line With New UltraSPARC T2 Plus Processor-Based Systems. The benchmark results for this new system are really good: The SPARC Enterprise T5240 server supported 4,170 SAP SD Benchmark users - a result more than double the 2,035 SAP SD Benchmark users achieved by the IBM p570 system with two POWER6 processors. or The dual-socket SPARC Enterprise T5240 server running the Solaris 10 OS excels on the Lotus iNotes email benchmark with 65,000 users -- the highest ever number of supported users..

The announcement already resulted in a huge amount of blog articles. Some really interesting entries so far:

Scaling Solaris on Large CMT Systems
Fast! Java on UltraSPARC T2 Plus
T5240 PCI-E I/O Performance
10 Gigabit Ethernet on UltraSPARC T2 Plus

Posted by Joerg Moellenkamp in Sun at 18:31

Hotels and WLAN

I hate hotels promising WLAN in rooms when this is only true for rooms adjacent to the lobby.

Posted by Joerg Moellenkamp in Business Travel at 06:16

Monday, April 7, 2008

Presentation finalized

Pheewwww I think i've finalized my presentation for Tuesday afternoon. 110 slides Lessig style.

PS: I will post my part of the preso tomorrow evening after the "show"

Posted by Joerg Moellenkamp in Sun at 21:39

links for 2008-04-07

On Choosing Type | i love typography, the typography blog
(tags: typography)

Super.fi 4vi Description - Products - Ultimate Ears Earphones Headphones Personal Monitors
(tags: gadgets hardware headphones phones iphone earphones)

Sun sues NetApp, take three | The Register
(tags: ZFS lawsuit NTAP NetApp)

OpenSolaris: A Sneak Peak with Ian Murdock - Barton's Blog
(tags: IPS Opensolaris Murdock)

Posted by del.icio.us at 13:39

Backups irrelevant?

It's nice that he references to ZFS in his article, but i think his opinion stated in "#PostgreSQL to Scale to 1 Billion Users, Dr Evil would be proud" is dangerous:Backup is irrelevant for those of you who care about this discussion. LVM/ZFS snapshots are the rule of the land.Well, when the first disaster hits your datacenter (and such disaster begin with a failed disk too much) , you will recognize that you are toast without tapes or at least with a disk-based backup on a different storage as a last line of defence. Neverever rely on a single pool of disk blocks on a single machine for disaster recovery, even when it gives you features like snapshot to freeze the disk state at one or multiple points in time.

Posted by Joerg Moellenkamp in The IT Business at 10:49

Sunday, April 6, 2008

Less known Solaris Features: Long support cycles

I'm extremely busy this weekend to prepare some presentations, thus i decided to write a rather short article about a topic nevertheless important to many customers. This time, i don't talk about a function of Solaris or a neat trick with the tools of the operating Environment. I just want to talk about a underrated feature of Solaris: The long support cycle. The support cycleEnterprise customers want to protect their investments. They don't want to use an operating system only to be forced in two years to use a different one, because they won't get any support. Most companies have a really long process of validating new configurations and they don't want to go through it without a good reason (bigger hardware updates or when they use a new software). Thus you have to support a commercial unix quite long.

Out of this reason, Sun has an well defined and long life cycle for the releases of our operating environment.

Event

Name

Description

E1

General Availability (GA)

GA is a day of joy, celebrations and marketing. A new major release of Solaris is born, for example the first release of Solaris 10 In at least the next 4 and a half years your will see several updates to this version of the perating system. Until Solaris 9 those updates were in a quarterly schedule, but since we use the update for new features, it isn't that regular.

E2

End of Life (EOL) Pre-Notification

Okay, one year before we announce the formal End of Life of the product, Sun sends the first notification/warning to our customers.

E3

End of Life (EOL) Announcement

Okay, we announce the end of the development of a major Solaris release. As i told before, this is at least 54 month after the GA, sometimes longer than that. When we announce the EOL, we trigger the start of the next phase in the lifecycle of a Solaris releaseo, the last order phase.

E4

Last Order Date (LOD)

90 days after the EOL announcement is the Last-Order-Date. This is the last day, you can order a Solaris version as an preinstalled image and it's the last a purchased system includes the license for this specific operating system. This Last Order Date isn't effective for Support Contracts. You can order a support contract for a Solaris version until it's End

of Service Life. With the Last Order day the next phase ist started: The last ship phase

E5

Last Ship Date (LSD)

In the next 90 days all orders for a new version has to be fullfilled. Yeah, you can´t order an EOled operating system for deliverya year after it´s end-of-life (besides of special agreements). With the Last-Order-Date the retirement of the Solaris Version starts.

E6

End of Retirement Support Phase 1

For the next two years you have essentially the same service than before EOL with some exceptions. No fixes for cosmetic bugs, no feature enhancements, no quarterly updates.

E7

End of Retirement Support Phase 2 / End of Service Life (EOSL)

In the last phase of the lifecycle, you still get telephone support for a version and you can still download patches for the system, but there will be no new patches.

after E7

After EOSL you can´t get further support or patches with the exception of special aggreements between the customer and Sun.

This is the policy of Sun for lifecycles. We won´t shorten the time, but often the effective lifetime ist much longer, as you will see in the next paragraph.

An example: Solaris 8Okay, we started to ship Solaris 8 in February 2000.Last order date (E4) was at Novemer 16, 2006. After that we shipped Solaris 8 Media Kits until February 16, 2007. Solaris 8 entered Retirement support mode Phase I on Mach 31, 2007. Thus it will reach Retirement Support Mode Phase II on March 31, 2009. End of Service Life is on March 31, 2012. Thus Solaris 8 has an service life of 12 years.

SidenoteFor a customer this long release cycles are optimal, but there is a problem for Sun in it. We don´t force our customer to use new versions early. Some customers still use old Solaris 8 versions and they use Solaris 10 like Solaris 8 to keep the processes in sync. There are some technological leaps between 8 and 10, but they don´t use the new features. They think they know Solaris, but they know just 8, not 10. The reputation of beeing somewhat outdated has it´s root partly in this habit. This is the bad side of the medal, but long support cycles are too important to change this policy...

Do you want to learn moreDisclaimer: The following documents are the authoritative source. If i made a mistake at summarizing them, the informations in these both documents are the valid ones.

Solaris Operating System Life Cycle

Solaris Operating System Retirement End Of Life Matrix

Posted by Joerg Moellenkamp in Solaris at 14:56

Fremdschämen im Zug

Ich finde ja auch, das persoenliche Durchsagen mehr Charme haben als jene vom Band und ich weiss, das auch mein Englisch nicht sehr berauschend ist ... aber "Thö wiekent tikket is no valid in trein" geht auch mir dann zu weit ..

Posted by Joerg Moellenkamp in Bahn at 11:45

Dieses Wochenende ...

Nicht wundern ... dieses Wochenende wird es keinen Teil in der Reihe "Less known Solaris Features" geben. Hoechstens einen relativ untechnischen Beitrag, den ich in der letzten Woche angefangen habe, aber da muss ich mal gucken, wann ich da letzte Hand anlegen kann, um den online zu stellen. Über ein "Feature", auf das mich ein Kunde hingewiesen hat. Apropos Kunden: Ich bekomme ganz gutes Feedback fuer die Serie von Kunden, aber habe schon wieder eine Halde von Themenvorschlaegen. Viele der Vorschlaege wuerden aber den Namen "Less known" auesserst weitgehend sprengen, beispielsweise Zones. Zumal es da ein excellentes Werk bereits gibt, naemlich den Container Leitfaden von Detlef Drewanz, Uli Graef et al. Auf meiner Todo-Liste stehen aber mittlerweile Themen wie GridEngine (ein sehr unterschaeetzes Feature, auch wenn es nicht so richtig zu Solaris dazugehoert,), eine Neuauflage des JET-Tutorials, das Solaris Security Toolkit.

Warum aber wird es dieses Wochenende keinen neuen Teil geben? Aus dem gleichen Grund, warum es etwas ruhig die letzten Tage hier ist. Die nächste Woche wirft ihre Schatten voraus. Dienstag morgen Dienstreise irgendwo die Pampa, Dienstag Vortrag, Mittwoch Moderation, Donnerstag Morgen Vortrag, Rückflug nach Berlin, Donnerstag nachmittag Vortrag in Berlin und Freitag Vortrag in München. Der Vortrag fuer Dienstag ist fast fertig, fuer die Moderation muss ich nichts tun ausser ein wenig polarisieren, um die Diskussion anzuheizen. Und naja ... polarisieren kann ich gut ... sogar ohne Vorbereitung. Der Vortrag am Freitag macht mir noch ein wenig Sorgen, mal sehen, was ich da mache. 2 Stunden über Virtualisierung. Ich weiss allerdings schon was ich nächsten Samstag machen werde .. Komapennen ...

Posted by Joerg Moellenkamp in General at 11:38

Friday, April 4. 2008

links for 2008-04-04

Scott Brown on Why 'Battlestar Galactica' Must Self-Destruct
Have i mentioned in the past, that i'm a regular viewer of the new BSG series ?
(tags: BSG)

Posted by del.icio.us in del.icio.us at 13:38

Gleise

Also irgendwie scheinen Gleise momentan eine ziemliche Anziehung auszuueben. Vorgestern S-Bahnzuege zwischen Hammerbrook und Hamburg gestoppt wegen Polizeieinsatz im Gleis. Heute morgen RTW-Einsatz im Zug wegen wasweissich ... auf jeden Fall blieb der Zug erstmal im Bahnhof Hasselbrook stehen. Letzten Samstag stuerzt jemand auf die Gleise im Bremer Hauptbahnhof. Er hat Glueck, das auf dem Gleis zu diesem Zeitpunkt kein Zug faehrt ... ein Güterzug kommt spaeter rechtzeitig zum Stehen waehrend die Person aus dem Gleisbett gezogen wird. Die Person duerfte ihr Leben der Geistesgegenwart des auf dem Gleis anwesenden Personals verdanken ...

Der Tote auf auf den Gleisen bei Halstenbek/Krupunder letzte Woche hat sich laut Hamburger Abendblatt ueberigens als ein 19 jähriger Mann herausgestellt.

Posted by Joerg Moellenkamp in Bahn at 10:37

Froward old man

This picture describes the situation of President Bush quite well. At the end, he is just an froward, old man waiting for the end of his presidency. He just kept his face at the NATO summit because the rest made a nebulous promise to Georgia and the Ukraine for a NATO membership in the remote future. I think, this is diplomatic lingo for "Never. At least not in this century". Nobody wants to follow his warmongering in the Iran. And his legacy: Irak is just a big mess with over 4000 dead soldiers now. An recession is imminent and he just send checks instead of using the money for infrastructure programs. A really luckless presidency.

Posted by Joerg Moellenkamp in policy of ... at 07:21

Thursday, April 3, 2008

Leakage through Opensource

One "nice" side effect of Opensourcing: You don't need leaks at Sun, you have just to wait until platform support code gets integrated into Opensolaris to learn more about upcoming hardware, for example OPL Ikkaku platform support in Solaris Nevada: Official product name for Ikkaku is not yet finalized. Ikkaku is a 2U, single CPU version of SPARC Enterprise M-series (sun4u) server utilizing the SPARC64-VII (Jupiter) processor.

Posted by Joerg Moellenkamp in Sun at 19:30

Who are you?

Inspired by this article in Kris Koehntopps blog, i want to ask you the same questions: 1. Who are you?
2. What tools to you use to read this blog?
3. Why do you read my blog?

Posted by Joerg Moellenkamp at 19:18

Infoworld about the T5120

Paul Venezia of Infoworld wrote a really positive review of the UltraSPARC T2 based Sun SPARC Enterprise T5120 - "Lab test: Sun's octo-core SPARC is made to multitask". He even concludes at the end, that reports of SPARCs death are greatly exaggerated in the light of this system: Overall, the UltraSPARC T2 and the T5120 build upon the hallmarks of the first-generation UltraSPARC T1-based servers, and remind us that although the SPARC CPU may have been marginalized in recent years, it hasn't surrendered, and may in fact be making a comeback.

Posted by Joerg Moellenkamp in Sun at 09:27

Wednesday, April 2. 2008

Ich bin ein Nasenduscher

Um mal die Dunkelziffer ins Helle zu zerren:

Zur freundlichen Übernahme und Verlinkung

Posted by Joerg Moellenkamp at 20:45

links for 2008-04-02

smcFanControl 2.1
(tags: mac tools utilities freeware hardware software osx)

Posted by del.icio.us in del.icio.us at 13:38

Dedupalicious

Deduplication is one of the big hype topics in the storage world today. Deduplication is about finding blocks or files on a storage, that exist multiple times with the same content. For example one of this unfunny joke presentation mailed to a big mail alias, which got saved on 50% of the desktops. Deduplication find this blocks or files and leaves only one to the disk, replacing all others with a pointer to the remaining copy. The basic idea: Saving capacity by reducing redundancy.

ZFS may have such an functionality in the future. Eric Kustarz writes in [how dedupalicious is your pool?](#) about some basics of such a functionality. The checksums integrated in ZFS may be of great use for Dedup, albeit you would use a cryptographically really strong hash algorithm.

There is an BugID for this RFE at [bugs.opensolaris.org](#). Hmm i hope this RFE will find it's way to Opensolaris or Solaris soon. Think about a Thumper as a Backup2Disk device with automatically dedups all data coming to it's disks. Or about an fileserver: Think about a combination of the in-kernel CIFS and the dedup functionality to save the storage needed for all this joke presentations.

(found via Robert Milkowski)

Posted by Joerg Moellenkamp at 07:35

Tuesday, April 1. 2008

Update 1 of SunCluster 3.2 released

Sun released the first update for Sun Cluster 3.2 yesterday. You will find the wikified Release Notes at wikis.sun.com. This update offers some really interesting features. For example you can put branded zones under the control of the cluster framework, even the lx brand. Yeah, you can cluster Linux applications with Sun Cluster

Posted by Joerg Moellenkamp in Sun at 21:48

Less known Solaris Features: Remote Mirror with AVS

As i told you before, Remote Mirror of the Availability Suite is a rather unusual feature. Nevertheless it's a really useful feature. The possibilities of such tools are endless, but it isn't heavy wizardry as you may have noticed already. So it's definitely a feature you should try.

Part 1: Introduction

Part 2: The fundamentals of Remote mirroring

Part 3: Setting up a synchronous mirror

Part 4: Testing the replication

Part 5: Replication Groups

Part 6: Deactivating a replication

Part 7: Truck based synchronisation

Part 8: Conclusion

As usual: Have fun !

Posted by Joerg Moellenkamp in Solaris at 20:51

Less known Solaris Features: Remote Mirror with AVS - Part 8: Conclusion

How can you use this feature? Some use cases are really obvious. It's a natural match for disaster recovery. The Sun Cluster Geographic Edition even supports this kind of remote mirror out of the box to do cluster failover with wider distances than just a campus. But it's usable for other jobs as well, for example for migrations to a new datacenter, when you have to transport a large amount data over long distances without a time window for a longer service interruption.

Do you want to learn more?

Documentation

Sun StorageTek Availability Suite 4.0 Software Installation and Configuration Guide

Sun StorageTek Availability Suite 4.0 Remote Mirror Software Administration Guide

misc. Links

OpenSolaris Project: Sun StorageTek Availability Suite

Posted by Joerg Moellenkamp in Solaris at 20:10

Less known Solaris Features: Remote Mirror with AVS - Part 7: Truck based synchronisation

Andrew S. Tannenbaum said "Never underestimate the bandwidth of a truck full of tapes hurling down the highway". This sounds counterintuitive at the first moment, but when you start to think about it, it's really obvious. The math behind the phrase Let's assume that you have two datacenters, thousand kilometers apart from each other. You have to transport 48 Terabytes of storage. We will calculate with the harddisk marketing system, 48.000.000 Megabytes. Okay ... now we assume, that we have a 155Mps leased ATM line between the locations. Let's assume that we can transfer 15,5 Megabytes per second of this line under perfect circumstances. Under perfect circumstances we can tranfer the amount of data in 3096774 seconds. Thus you would need 35 days to transmit the 48 Terabytes. Now assume a wagon car with two thumpers (real admins don't use USB sticks, they use the X4500 for their data transportation needs)in the trunk driving at 100 kilometers per hour. The data would reach the datacenter within 10 hours. Enough time to copy the data to the transport-thumpers and after the arrival from the thumper to the final storage array.

Blog Export: c0t0d0s0.org, http://www.c0t0d0s0.org/

Truck based replication with AVSAVS Remote mirror supports a procedure to exactly support such a method: Okay, let's assume you want to migrate a server to a new one. But this new server is 1000km away. You have multiple terabytes of storage, and albeit your line to the new datacenter is good enough for the updates, an full sync would take longer than the universe will exist because of the proton decay.

AVS Remote Mirror can be configured in a way, that relies on a special condition of primary and secondary volumes: The disks are already synchronized, before starting the replication. For example by doing a copy by dd to the new storage directly or with the redirection of an transport media like tapes. When you configure AVS Remote Mirror in this way, you don't need the initial full sync.

```
On our old serverTo play around, we create at first a new filesystem:[root@theoden:~]$ newfs /dev/rdisk/c1d1s1
newfs: construct a new file system /dev/rdisk/c1d1s1: (y/n)? y
/dev/rdisk/c1d1s1: 968704 sectors in 473 cylinders of 64 tracks, 32 sectors
473.0MB in 30 cyl groups (16 c/g, 16.00MB/g, 7680 i/g)
super-block backups (for fsck -F ufs -o b=#) at:
32, 32832, 65632, 98432, 131232, 164032, 196832, 229632, 262432, 295232,
656032, 688832, 721632, 754432, 787232, 820032, 852832, 885632, 918432, 951232Now mount it , play around with
it and put a timestamp in a file.[root@theoden:~]$ mount /dev/dsk/c1d1s1 /mnt
[root@theoden:~]$ mount /dev/dsk/c1d1s1 /mnt
[root@theoden:~]$ touch /mnt/test1
[root@theoden:~]$ mkfile 1k /mnt/test2
[root@theoden:~]$ mkfile 1k /mnt/test3
[root@theoden:~]$ mkfile 1k /mnt/test4
[root@theoden:~]$ date >> /mnt/test5Okay, now unmount it again.[root@theoden:~]$ umount /mntNow we can generate
a backup of this filesystem. You have to make a image of the volume, making a tar or cpio file backup isn't
sufficient.[root@theoden:~]$ dd if=/dev/rdisk/c1d1s1 | gzip > 2migrate.gz
968704+0 records in
968704+0 records out
Okay, now activate the replication on the primary volume. Don't activate it on the secondary one! The important
difference to a normal replication is the -E. When you use this switch, the system assumes that the primary and
secondary volume are identical already.
[root@theoden:~]$ sndradm -E theoden /dev/rdisk/c1d1s1 /dev/rdisk/c1d1s0 gandalf /dev/rdisk/c1d1s1 /dev/rdisk/c1d1s0
ip sync
Enable Remote Mirror? (Y/N) [N]: yOkay, we've used the -E switch again to circumvent the need for a full
synchronisation. When you look at the status of volume, you will see the volume in the "logging" state:[root@theoden:~]$
dsstat
name      t s  pct role ckps dkps tps svt
dev/rdisk/c1d1s1 P L  0.00 net  -  0  0  0
dev/rdisk/c1d1s0      bmp  0  0  0  0This means, that you can do changes on the
volume.[root@theoden:~]$ mount /dev/dsk/c1d1s1 /mnt
[root@theoden:~]$ cat /mnt/test5
Mon Mar 31 14:57:04 CEST 2008
[root@theoden:~]$ date >> /mnt/test6
[root@theoden:~]$ cat /mnt/test6
Mon Mar 31 15:46:03 CEST 2008
Now we transmit our image of the primary volume to our new system. In my case it's scp, but for huge amount of data
sending the truck with tapes would be more sensible.
[root@theoden:~]$ scp 2migrate.gz jmoekamp@gandalf:/export/home/jmoekamp/2migrate.gz
Password:
2migrate.gz      100% |*****| 1792 KB  00:00
```

```
On our new serverOkay, when the transmission is completed, we write the image to the raw device of the secondary
volume:[root@gandalf:~]$ cat 2migrate.gz | gunzip | dd of=/dev/rdisk/c1d1s1
968704+0 records in
968704+0 records out
```

```
Okay, now we configure the replication on the secondary host:# sndradm -E theoden /dev/rdisk/c1d1s1 /dev/rdisk/c1d1s0
gandalf /dev/rdisk/c1d1s1 /dev/rdisk/c1d1s0 ip syncA short look into the status of replication:[root@gandalf:~]$ dsstat
name      t s  pct role ckps dkps tps svt
dev/rdisk/c1d1s1 S L  0.00 net  -  0  0  0
dev/rdisk/c1d1s0      bmp  0  0  0  0
Okay, our primary and secondary volumes are still in logging mode. How do we get them out of this? In our first example
```

we did an full synchronisation, this time we need only an update synchronisation. So login as root to our primary host and initiate such an update sync. This is the moment, where you have to stop working on the primary volume.

```
[root@theoden:~]$ sndradm -u
```

Refresh secondary with primary? (Y/N) [N]: y After this step all changes we did after creating the image from our primary volume will be synced to the secondary volume.

Testing the migrationWell ... let's test this: Do you remember, that we created /mnt/test6 after the dd for the image?

Okay, at first, we put the replication in logging mode again. So login as root on our secondary host.

```
[root@gandalf:~]$ sndradm -l
```

Put Remote Mirror into logging mode? (Y/N) [N]: y

Now we mount the secondary volume:# mount /dev/dsk/c1d1s1 /mnt

```
[root@gandalf:~]$ cd /mnt
```

```
[root@gandalf:~]$ ls
```

```
lost+found test2 test4 test6
```

test1 test3 test5By the virtues of update synchronisation, the test6 appeared on the secondary volume.Let's have a look in /mnt/test6:

```
[root@gandalf:~]$ cat test6
```

```
Mon Mar 31 15:46:03 CEST 2008Cool, isn't it ?
```

Posted by Joerg Moellenkamp in Solaris at 20:00

Less known Solaris Features: Remote Mirror with AVS - Part 6: Deactivating a replication

In the last parts of this tutorial i've explained to you, how you set up a replication relation. But it's important to know how you deactivate and delete the replication as well.

Deleting the replication configurationIt's quite easy to delete the replication.

At first we look up the existing replication configuration.[root@gandalf:~]\$ sndradm -P

```
/dev/rdisk/c1d1s1 gandalf:/dev/rdisk/c1d1s1
```

```
autosync: off, max q writes: 4096, max q fbas: 16384, async threads: 2, mode: sync, state: logging
```

```
[root@theoden:~]$ sndradm -d gandalf:/dev/rdisk/c1d1s1
```

Disable Remote Mirror? (Y/N) [N]: y

```
[root@theoden:~]$ sndradm -P
```

```
[root@theoden:~]$
```

Posted by Joerg Moellenkamp in Solaris at 19:59