Saturday, May 3. 2008

New features of Solaris: Alternate boot environments based on snapshots

One of the limitations of Opensolaris 2008.05 will be the missing LiveUpgrade. But ... well ... you have something better. The whole concept of LifeUpgrade was transformed into the future by using the capabilities of ZFS.
Using snapshots for boot environmentsOne of the nice features of ZFS is the fact, that you get snapshots for free. The reason lies in the copy-on-write nature of ZFS. You can freeze the filesystem by not freeing the old blocks. as new data is written is written to new blocks, you don't even have to copy the blocks (in this sense the COW of ZFS is more like a ROW ... redirect on write).

ZFS boot enables the system to work with such snapshots, as you can use one of these to boot from. You can establish multiple boot environments just by snapshoting the bootfilesystems, clonimg them and promoting them to real filesystems. This are features inherent to ZFS.

A practical exampleA warning at first: Don´t try this example without a backup of your system. Or use a test system or test VM. We will fsck up the system during this example. Okay....

I've updated my system, so i have alread two boot environments on my system:jmoekamp@glamdring:~# beadm list

```
BE           Active Active on Mountpoint Space
Name          reboot          Used
----         ------ --------- ---------- -----
opensolaris-1 yes   yes     legacy   2.31G
opensolaris  no    no       -       62.72M
```
This mirrors the actual state in your ZFS pools. You will find filesystems with accordings names.
```
NAME                                 USED  AVAIL  REFER  MOUNTPOINT
rpool                                2.39G  142G  56.5K  /rpool
rpool@install                        18.5K    -    55K  -
rpool/ROOT                           2.37G  142G   18K  /rpool/ROOT
rpool/ROOT@install                       0    -    18K  -
rpool/ROOT/opensolaris               62.7M  142G  2.23G  legacy
rpool/ROOT/opensolaris-1             2.31G  142G  2.24G  legacy
rpool/ROOT/opensolaris-1@install     4.66M    -  2.22G  -
rpool/ROOT/opensolaris-1@static:-:2008-04-29-17:59:13    5.49M    -  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                         18.9M  142G   19K  /export
rpool/export@install                  15K    -    19K  -
rpool/export/home                    18.9M  142G  18.9M  /export/home
rpool/export/home@install             18K    -    21K  -
```
After doing some configuration, you can create an boot environment called opensolaris-baseline:
It's really easy. You just have to create a new boot environment:# beadm create -e opensolaris-1 opensolaris-baselineBut we will not work with this environment. We use it as a baseline, as a last resort when we destroy our running environment. To run the system we will create another snapshot: # beadm create -e opensolaris-1 opensolaris-workNow let´s look into the list of our boot environments.jmoekamp@glamdring:~# beadm list

```
BE               Active Active on Mountpoint Space
Name             reboot          Used
----             ------ --------- ---------- -----
opensolaris-baseline no    no      -       53.5K
opensolaris-1     yes   yes     legacy   2.31G
opensolaris      no    no       -       62.72M
opensolaris-work  no    no       -       53.5K
```
Okay, now we activate the opensolaris-work boot environment:jmoekamp@glamdring:~# beadm activate opensolaris-workOkay, let´s look at the list of boot environments again.jmoekamp@glamdring:~# beadm list

```
BE              Active Active on Mountpoint Space
Name            reboot           Used
----            ------ --------- ---------- -----
opensolaris-baseline no    no       -        53.5K
opensolaris-1    yes   no      legacy  24.5K
opensolaris      no    no      -       62.72M
opensolaris-work no    yes     -        2.31G
```
jmoekamp@glamdring:~#You will see that the opensolaris-1 snapshot is still active, but that the opensolaris-work will be active at the next reboot. Okay, now reboot:jmoekamp@glamdring:~# beadm list

```
BE              Active Active on Mountpoint Space
Name            reboot           Used
----            ------ --------- ---------- -----
opensolaris-baseline no    no       -        53.5K
opensolaris-1    no    no      -       54.39M
opensolaris      no    no      -       62.72M
opensolaris-work yes   yes     legacy   2.36G
```
Okay, you see ... the boot environment opensolaris-work is now active and it´s activated for the next reboot (until you activate another boot environment).

Now we can reboot the system. The GRUB comes up and it will default to the opensolaris-work environment. Please remember on whicht position you find opensolaris-baseline in the boot menu. You need this position in a few moments. After a few seconds, you can log into the system and work with it.

Okay ... now let's drop the atomic bomb of administrative mishaps to your system. Log into your system, assume the root role and do the following stuff:# cd /
rm -rf *You known what happens. Depending from how fast you are able to interrupt this run to get an slightly damaged system up to a system fscked up beyond any recognition. Normaly the system would send you to the tapes now. But remember. You have some alternate boot environments.

Reboot the system, wait for the grub. You may have an garbeled output, so it's hard to read the output from the grub. Choose opensolaris-baseline. The system will boot up quite normaly.

You need a terminal window now. How you get such a terminal window depends from incurred damage. The boot environment snapshots doesn't cover the home directories. So you may have no home directory any longer. I will assume this for this example: You can get a terminal window by clicking on "Options", then "Change Session" and choose "Failsafe Terminal" there.

Okay, login via the graphical login manager, a xterm will appear. At first we delete the defunct boot environment:# beadm destroy opensolaris-work1
Are you sure you want to destroy opensolaris-work1? This action cannot be undone (y/[n]):
yOkay, now we clone the opensolaris-baseline environment to form a new opensolaris-work environment.# beadm create -e opensolaris-baseline opensolaris-workWe reactivate the opensolaris-work boot environment:# beadm activate opensolaris-workNow check, if you still have a homedirectory for your user:# ls -l /export/home/jmoekamp
/export/home/jmoekamp: No such file or directoryIf your home directory doesn't exist any longer, create a new one:# mkdir -p /export/home/jmoekamp
chown jmoekamp:staff /export/home/jmoekampNow reboot the system:# rebootWait a few moments. The system starts up. The GRUB defaults to opensolaris-work and the system starts up normaly without any problem in that condition the system had, when you create the opensolaris-baseline boot environment.# beadm list

```
BE              Active Active on Mountpoint Space
Name            reboot           Used
----            ------ --------- ---------- -----
opensolaris-baseline no    no       -        3.18M
opensolaris-1    no    no      -       54.42M
opensolaris      no    no      -       62.72M
opensolaris-work yes   yes     legacy   2.36G
```
Obviously you may have to recover your directory with data. It's a best practice to make snapshots of this directories on a regular schedule. So you can simply promote a snapshot to your actual version of the directory.

ConclusionYou see, this is a really neat feature. Recovering from a disaster in a minute or two. Snapshotting opens a completely new way to recover from errors. Unlike with Liveupgrade you don't need extra disks or extra partitions, and

as ZFS snapshots are really fast, creating alternate boot environments on zfs are extremly fast as well.

At the moment this feature is available on Opensolaris 2008.05 only. But with future updates it will find it´s way into Solaris as well.

Posted by Joerg Moellenkamp in English, Solaris at 13:40