Friday, October 10. 2008

Separated ZIL on ramdisk.

I used ramdisks today in a live presentation of ZFS, L2ARC and sZIL (you can´t use files for sZIL and L2ARC testing, you need a device). One of the customers asked me if there is a real use case for this configuration. My answer was: Not really. But this question haunted me the whole day while sitting in the train.
I had some weird ideas like using an UPS backed server to act as a sZIL device by provding access to a ramdisk via Fibre Channel (by using the FC target in COMSTAR) , but i had no idea if this would be really clever idea. Ramdisk based sZIL? Okay, i know there are now several admins with a cardiac arrest out there. The problem of a ramdisk based sZIL is the volatile nature of the ramdisk. When you loose the power on your systems, you have still a consistent pool but you loose the transactions stored in the sZIL.This is a less than desirable effect. Between you and loosing transactions is just the power switch There was a missing piece...

The good thing about a large community: Sometimes someone had already similar ideas long ago and had some time to do some experiments. So i found a really neat idea on the zfs-discuss mailing list with the important extra thought compared to my "dozing in the train"-idea.

The idea of Chris Greer has this nice additional twist making the concept of using a ramdisk sZIL more sensible. He writes in "An slog experiment (my NAS can beat up your NAS)": So I tried this experiment this week...
On each host (OpenSolaris 2008.05), I created an 8GB ramdisk with ramdiskadm. I shared this ramdisk on each host via the iscsi target and initiator over a 1GB crossconnect cable (jumbo frames enabled). I added these as mirrored slog devices in a zpool.
This is really a neat trick. You can mirror a sZIL device so you could distribute the ramdisks over several systems. When one of the system fails, you have still the other device with the same data. This configuration had an impressive effect on the performanceThe big thing here is I ended up getting a MASSIVE boost in performance even with the overhead of the 1GB link, and iSCSI. The iorate test I was using went from 3073 IOPS on 90% sequential writes to 23953 IOPS with the RAM slog added. The service time was also significantly better than the physical disk.I think you can drive this concept even farther by substituting the iSCSI over TCP over Gigabit with iSER (iSCSI over RDMA over Infiniband) and using separate small x86 systems for this task, each backed by a small UPS keeping the system up until an dd from the ramdisk to a hard disk has completed.

Chris, sounds like a really clever idea ...

Posted by Joerg Moellenkamp in English, Solaris at 20:28

COMSTAR with SAS Target sounds very cool for this setup...But how long is data kept in the sZIL device before it is flushed to disk? Isnt it enough to keep the data in the write cache of the sZIL computer in case your zpool node crashes? So this way you could use a disk-backed sZIL device that survives everything except rebooting both/all sZIL nodes in parallel.
Anonymous on Oct 11 2008, 14:06

LOL. Mirrored Ramdisk ZILs don't work in production environment because if the servers are on the same power grid, you are screwed if the entire power grid went offline.

Now mirrored ramdisks backed by battery and flash(Acard) might work. Although the Acard ram drive cannot be hot swapped so in production, if one of the ram drives failed, you have to offline the system anyways, which defeats the purpose.

Some one please make a battery backed ramdrive that's backed by integrated flash and make it hotswappable 2.5inch. That device would sell like hotcakes. For now mirroring SLC drives would do. (Look out for Samsung's enterprise drives)
Anonymous on Dec 9 2008, 20:21

That´s easy ... small rackable USV at the server ... just for long enough to dump the ramdisk to a real disk ...
Anonymous on Dec 9 2008, 20:30

ACADR or any other RAM-SSD are limited to 1.2-1.3 GB/s RAID cards throughput limit, Ramdisk on the server with 20GB Infiniband (40GB might be used later with Nehalem Xeon platform) will work at 1.6-1.8GB/s. Same goes for IOPS, as Infiniband connection is less latent than RAID one.
Anonymous on Feb 16 2009, 20:18