

Re: Filesystem snapshots dog slow

Source: <http://unix.derkeiler.com/Mailing-Lists/FreeBSD/hackers/2007-10/msg00136.html>

- *From:* Eric Anderson <anderson@xxxxxxxxxxx>
 - *Date:* Tue, 16 Oct 2007 06:54:11 -0500
-

Jeremy Chadwick wrote:

Since the snapshot code (e.g. `mksnap_ffs(8)` and friends) was introduced, `dump(8)` was modified to nag you if you didn't use the `-L` argument. "Um, okay, I'd better use `-L`" is what came out of my mouth, and I'm sure a lot of other administrators' when they saw this message.

But it seems the making a snapshot is an incredibly slow/intensive task. The documentation I've read indicates that making a snapshot "is incredibly fast" — based on my experiences, it isn't. At least it's no where near as fast as, say, a Netapp filer.

The problem is the way the snapshots work in UFS2. It has to do a lot of work to create that snapshot, and the amount of work it does goes up with the amount space you have available (because it relates to the number of cylinder groups you have). The UFS2 snapshot and the WAFL (NetApp's file system) snapshot are **completely** different, and should not be compared in this way. The functionality is (in the end) the same, but otherwise, they are different.

I've found 3 threads (dating 2003, 2005, and 2007) about this problem:

<http://lists.freebsd.org/pipermail/freebsd-current/2003-August/009135.html>

<http://lists.freebsd.org/pipermail/freebsd-fs/2005-July/001216.html>

<http://lists.freebsd.org/pipermail/freebsd-stable/2007-January/031882.html>

Only three threads? :) There's probably hundreds like them..

This issue is still present on RELENG_7, and I can confirm it on multiple machines (some running **completely** different hardware than others).

It's a UFS2 problem, and the docs that say 'incredibly fast' are actually referring to small filesystems, that are

Re: Filesystem snapshots dog slow

not busy (with writes). Maybe the docs should be clarified for now. You can submit patches to the docs you found that say that if you'd like to help out.

```
osiris# df -ki /disk2
Filesystem 1024-blocks Used Avail Capacity iused ifree %iused Mounted on
/dev/ad6s1d 236511738 4 217590796 0% 2 30570492 0% /disk2
```

```
osiris# time mksnap_ffs /disk2 /disk2/mysnapshot
0.000u 1.012s 5:12.23 0.3% 5+1149k 7803+18819io 0pf+0w
```

While `mksnap_ffs` runs, the process remains in `wdrain` state. `gstat(8)` shows immense disk I/O. `ms/r` occasionally jumps up to 1100 or higher, but usually hovers around 40–60.

[..snip..]

The time doubled. This isn't good.

Disks are getting larger, filesystems growing, people storing more data. Hitachi, for example, has guaranteed 4TB disks by the end of 2011. If this problem has sat idle for at least 4 years already, we'll be in a lot of trouble come 2011. And let's not forget that every piece of FreeBSD documentation tells admins to "use `dump`, it's the best!". This issue is a good reason to consider using tools like `rsync` or `tar` instead. :-(

I recommend reading up a little bit on how the snapshots for UFS2 work. It will give you a good understanding of what the issue is. Essentially, your disk is hammered making copies of all the cylinder groups, skipping those that are 'busy', and coming back to them later. On a 200Gb disk, you could have 1000 cylinder groups, each having to be locked, copied, unlocked, and then checked again for any subsequent changes. The stalls you see are when there are lock contentions, or disk IO issues. On a single disk (like your setup above), your snapshots will take forever since there is very little random IO performance available to you.

I will gladly work with anyone who wishes to tackle this, either by providing hardware (MB/disks/etc.) for free, or by giving the individual access to a box that has serial console + a serial debugger available.

FreeBSD 7 includes ZFS. Have you thought about using it? The problem isn't that developers don't know the problem exists, or that they don't have hardware, or a serial console access to a system. The problem is that there are only so many developers, and so much time, and this is a big mountain to climb. It's hard to find an experienced person to do the work (for free), when they could be doing anything else they wish. I think, that

Re: Filesystem snapshots dog slow

in the end, for some of these aging issues to get resolved, there needs to be another bounty put out on it. I think rsync.net might even have one started for this issue already – you might think about adding to the bounty, or officially offering hardware through there.

Eric

freebsd-hackers@xxxxxxxxxxx mailing list

<http://lists.freebsd.org/mailman/listinfo/freebsd-hackers>

To unsubscribe, send any mail to "freebsd-hackers-unsubscribe@xxxxxxxxxxx"