

## Re: Sysinstall automatic filesystem size generation.

*Source:* <http://unix.derkeiler.com/Mailing-Lists/FreeBSD/stable/2005-08/0567.html>

---

*From:* Jon Dama ([jd\\_at\\_ugcs.caltech.edu](mailto:jd_at_ugcs.caltech.edu))

*Date:* 08/30/05

Date: Mon, 29 Aug 2005 19:59:10 -0700 (PDT)

To: Matthias Buelow <[mkb@incubus.de](mailto:mkb@incubus.de)>

Well, I think one issue is that it destroys one of the fundamental advantages of softupdates which was that you could interleave streams of strongly ordered metadata writes without demanding a sequence for the streams collectively. By using request barriers, you are effectively forcing an additional synchronization requirement, the secret will be not forcing us all the way back to having effectively synchronous metadata writes (see below).

As I understand, metadata operations are only added to the WORKLIST when their dependents have already been "completed" i.e., at the lowest level have had biodone called to mark the write operation completed. I am not sure how ffs\_softdeps checks this property.

It seems you need to add a layer of indirection. (owing to biodone being called merely when the drive has cached the request). What you know is that those operations marked completed by biodone are in fact done only after a (costly) flush cache operation is executed.

Therefore you want to delay this operation for as long as possible, in fact until you actually depend on biodone being honest. I.e., at the time another operation is inserted into the WORKLIST.

The secret I think is to keep track of which bp's marked B\_DONE by biodone that have been certified by a flush cache. Thus permitting you to avoid some cache flushes. Furthermore, the softdep code has to be responsible for envoking the flush cache operation when it notices that the B\_DONE flag that it cares about does not have a matching B\_REALLY\_DONE flag, which every block should have that had B\_DONE set before the flush cache operation happened.

I do not really know how GEOM has changed this situation. biodone seems to have been stripped of much of its old responsibilities?

-Jon

freebsd-stable: Re: Sysinstall automatic filesystem size generation.

I'd guess that it belongs

On Tue, 30 Aug 2005, Matthias Buelow wrote:

> *Jon Dama wrote:*

>

> > *Ironically, phk backed out the underlying support for this safety fix*

> > *from the FreeBSD kernel b.c. it wasn't integrated into the softupdates*

> > *code*

> > *whereas in reality the proper course of action would have been to hook it*

> > *in. :-/*

>

> *Can it be put into softupdates at all? From what I understand (which*

> *is probably a rather sketchy idea of the matter), write barriers*

> *work because they are only used here to separate journal writes*

> *from data writes (i.e., to make sure the log is written, by flushing*

> *the cache, before any filesystem data hits the platters). I've read*

> *the softupdates paper some time ago and haven't found similar*

> *sequence points where one could insert such flushing. One would*

> *have to "flush" all the time, either continuously or in very short*

> *intervals, in order to keep the ordering, which then would amount*

> *to the same effects as if one simply disabled the cache. But probably*

> *I'm wrong here (I hope).*

>

> *mkb.*

>

---

freebsd-stable@freebsd.org mailing list

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

To unsubscribe, send any mail to "freebsd-stable-unsubscribe@freebsd.org"