Subject: Re: [RFC PATCH ext3/ext4] orphan list corruption due bad inode
Posted by Eric Sandeen on Tue, 05 Jun 2007 03:15:47 GMT
View Forum Message <> Reply to Message

Vasily Averin wrote:
> After ext3 orphan list check has been added into ext3_destroy_inode() (please see my previous patch) the following situation has been detected:
> EXT3-fs warning (device sda6): ext3_unlink: Deleting nonexistent file (37901290), 0
> Inode 00000101a15b7840: orphan list check failed!
> 00000773 6f665f00 74616d72 00000573 65725f00 06737270 66000000 616d726f
> ...
> Call Trace: [<ffffffff80211ea9>] ext3_destroy_inode+0x79/0x90
> [<ffffffff801a2b16>] sys_unlink+0x126/0x1a0
> [<ffffffff80111479>] error_exit+0x0/0x81
> [<ffffffff80110aba>] system_call+0x7e/0x83
>
> First messages said that unlinked inode has i_nlink=0, then ext3_unlink() adds this inode into orphan list.
>
> Second message means that this inode has not been removed from orphan list. Inode dump has showed that i_fop = &bad_file_ops and it can be set in make_bad_inode() only. Then I've found that ext3_read_inode() can call make_bad_inode() without any error/warning messages, for example in the following case:
> ...
>
```
> if (inode->i_nlink == 0) {
>         if (inode->i_mode == 0 ||
>             !(EXT3_SB(inode->i_sb)->s_mount_state & EXT3_ORPHAN_FS)) {
>                 /* this inode is deleted */
>                 brelse (bh);
>                 goto bad_inode;
```
> ...
>
> Bad inode can live some time, ext3_unlink can add it to orphan list, but
> ext3_delete_inode() do not deleted this inode from orphan list. As
> result we can have orphan list corruption detected in ext3_destroy_inode().

Ah, I see - so you have confirmed that this inode does have bad ops...?
I did notice on another orphan inode bug investigation that
ext3_inode_delete won't clear orphan from a bad_inode...

> However it is not clear for me how to fix this issue correctly.
>
> As far as i see is_bad_inode() is called after iget() in all places excluding ext3_lookup() and ext3_get_parent(). I believe it makes sense to add bad inode check to these functions too and call iput if bad inode detected.

That seems reasonable to me in any case, though as Andrew said, do you
know for sure how the bad inodes were getting on the orphan list...?

Is it possible that they were recycled after being freed while still on the orphan list, as in my previous reply to your previous message?

-Eric