

## Re: 2.6.12.2 dies after 24 hours

**Source:** <http://linux.derkeiler.com/Mailing-Lists/Kernel/2005-07/3324.html>

---

**From:** Rob Mueller ([robm\\_at\\_fastmail.fm](mailto:robm_at_fastmail.fm))

**Date:** 07/13/05

To: "Lars Roland" <[lroland@gmail.com](mailto:lroland@gmail.com)>, "Bron Gondwana" <[brong@fastmail.fm](mailto:brong@fastmail.fm)>  
Date: Wed, 13 Jul 2005 10:27:10 +1000

> > *We're also applying the attached patch. There's a bug in reiserfs that*  
> > *gets tickled by our huge MMAP usage (it's amazing what really busy*  
> > *Cyrus daemons can do to a server, ouch). It's fixed in generic\_write,*  
> > *so we take the few percent performance hit for something that doesn't*  
> > *break!*  
>  
> *Interesting – When I got the problem it was on mail servers under high*  
> *load (handling 60.000 emails pr. hour) with reiserfs as file system. I*  
> *have seen this problem on 5 different servers so I am confident that*  
> *it is not hardware failure.*  
>  
> *Sometimes the server load just rises and then the server dies other*  
> *times the load rises but the kernel manages to get it back alive*  
> *filling up syslog with messages like this*

Sounds like a different issue. The patch Bron included before fixes (or at least reduces to the point where it fixes it for us) a problem where processes get stuck in D state and are unkillable. A reboot is required to remove them. Apparently this is a known bug in ReiserFS (see messages below). As noted, the same bug exists in ext3. There appears to have been some patches to try and fix it for both reiserfs and ext3, but I'm not sure if they're in the mainline kernel yet.

<http://www.ussg.iu.edu/hypermail/linux/kernel/0409.0/2056.html>  
<http://hullug.principalhosting.net/archive/index.php/t-22774.html>

Rob

----- Original Message -----

From: "Vladimir Saveliev" <[vs@namesys.com](mailto:vs@namesys.com)>

To: "Jeremy Howard" <[jhoward@fastmail.fm](mailto:jhoward@fastmail.fm)>

Cc: "Hans Reiser" <[reiser@namesys.com](mailto:reiser@namesys.com)>; <[reiserfs-dev@namesys.com](mailto:reiserfs-dev@namesys.com)>

Sent: Friday, October 08, 2004 4:57 PM

Subject: Re: URGENT: Need fix for problem with copy\_from\_user inside a transaction

Linux-Kernel: Re: 2.6.12.2 dies after 24 hours

> Hello  
>  
> On Thu, 2004-10-07 at 21:35, Jeremy Howard wrote:  
>> On Fri, 01 Oct 2004 09:13:34 -0700, "Hans Reiser" <reiser@namesys.com>  
>> said:  
>>> Chris, can you comment please?  
>>>  
>>> Also... can you guys suggest any ways to minimise the problem, e.g.  
>>> external vs internal journal? metadata vs full journalling? changing the  
>>> elevator? ...  
>>>  
>>> Would you please check whether the attached patch for  
>>> ./fs/reiserfs/file.c fixes the problem?  
>>>  
>>> Quick benchmark shows that using generic\_file\_write does not hurt  
>>> reiserfs performance too much comparing to using original  
>>> reiserfs\_file\_write.  
>>>  
>>> ---Sequential Output (nosync)--- ---Sequential  
>>> Input-- --Rnd Seek--  
>>> -Per Char-- --Block---- -Rewrite-- -Per  
>>> Char-- --Block---- --04k (03)--  
>>> MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  
>>> K/sec %CPU /sec %CPU  
>>> generic\_file\_write 256 16690 99.7 28332 29.9 11528 16.8 12411 77.2  
>>> 28888 17.3 237.4 2.1  
>>> reiserfs\_file\_write 256 15647 86.8 30875 22.1 10822 14.1 11631 72.1  
>>> 29184 16.1 250.0 2.0  
>>>  
>>>  
>>>> Vladimir Saveliev wrote:  
>>>>  
>>>>> Hello  
>>>>>  
>>>>> On Fri, 2004-10-01 at 01:53, Hans Reiser wrote:  
>>>>>  
>>>>>  
>>>>>> vs, this is not enough of an answer to share your understanding of  
>>>>>> the  
>>>>>> problem. Please say much more.  
>>>>>>  
>>>>>>  
>>>>>>  
>>>>>>> Reiserfs\_write\_file locks set of pages, and then tries to copy data to  
>>>>>>> them. If it is to copy data from one of pages which are locked and if  
>>>>>>> that page is not uptodate, pagefault requires to lock that page, but  
>>>>>>> as  
>>>>>>> it is locked already - process deadlocks with itself.  
>>>>>>>  
>>>>>>>

Linux-Kernel: Re: 2.6.12.2 dies after 24 hours

```
>>> Is this when copying from one file in a formatted node to another file
>>> in that node?
>>>
>>> >As Chris said – fix is not trivial. Also, it is known that he did
>>> >already something about it, so, I thought that it would make sense to
>>> >find first what is his state at this problem.
>>>>
>>>>
>
> Content-Disposition: attachment; filename=file.c.diff2
>
> --- file.c~ 2004-10-02 12:29:33.223660850 +0400
> +++ file.c 2004-10-08 10:03:03.001561661 +0400
> @@ -1137,6 +1137,8 @@
> return result;
> }
>
> + return generic_file_write(file, buf, count, ppos);
> +
> if ( unlikely((ssize_t) count < 0 ) )
> return -EINVAL;
>
```

----- Original Message -----

From: "Chris Mason" <mason@suse.com>  
To: "Hans Reiser" <reiser@namesys.com>  
Cc: "Vladimir Saveliev" <vs@namesys.com>; "Oleg Drokin"  
<green@linuxhacker.ru>; "Jeremy Howard" <jhoward@fastmail.fm>;  
<reiserfs-dev@namesys.com>  
Sent: Saturday, October 09, 2004 1:05 AM  
Subject: Re: URGENT: Need fix for problem with copy\_from\_user inside  
atransaction

```
> On Fri, 2004-10-08 at 07:46 -0700, Hans Reiser wrote:
>> No, this is not the right answer.
>
>>>>--- file.c~ 2004-10-02 12:29:33.223660850 +0400
>>>>+++ file.c 2004-10-08 10:03:03.001561661 +0400
>>>>@@ -1137,6 +1137,8 @@
>>>> return result;
>>>> }
>>>>
>>>>+ return generic_file_write(file, buf, count, ppos);
>>>>+
>>>> if ( unlikely((ssize_t) count < 0 ) )
>>>> return -EINVAL;
>
> It's not right because ext3 has exactly the same bug. The only real
> solution is to change the order in which things happen during
> file_write.
>
```

Linux-Kernel: Re: 2.6.12.2 dies after 24 hours

> *Right now, we do this:*  
>  
> *prepare\_write() allocates space, starts a transaction*  
> *copy\_from\_user() can deadlock here because a transaction is running*  
> *commit\_write() end the transaction*  
>  
> *This is the same in generic\_file\_write and reiserfs\_file\_write, although*  
> *reiserfs\_file\_write doesn't call reiserfs\_prepare\_write, it has the same*  
> *basic structure in terms of when the transaction starts and ends.*  
>  
> *The solution is to move most of the work from reiserfs\_prepare\_write*  
> *into reiserfs\_commit\_write. I've made a number of different patches for*  
> *this, all have had problems, but I'm working through it and will have a*  
> *real tested fix.*  
>  
> *Jeremy the best thing you can do right now is to mount your filesystems*  
> *with -o nolargeio=1, and use the temporary workarounds I sent you*  
> *before. The -o nolargeio=1 will reduce the amount of work we try to do*  
> *with each pass in reiserfs\_file\_write (note that Vladimir's patch above*  
> *will have very similar effects).*  
>  
> *-chris*

-

To unsubscribe from this list: send the line "unsubscribe linux-kernel" in the body of a message to majordomo@vger.kernel.org

More majordomo info at <http://vger.kernel.org/majordomo-info.html>

Please read the FAQ at <http://www.tux.org/lkml/>