Subject: Re: [RFC][PATCH] EXT3: problem with page fault inside a transaction
Posted by Andrew Morton on Thu, 12 Oct 2006 06:43:30 GMT

View Forum Message <> Reply to Message

On Thu, 12 Oct 2006 09:57:26 +0400
Dmitriy Monakhov <dmonakhov@openvz.org> wrote:

> While reading Andrew's generic_file_buffered_write patches i've remembered
> one more EXT3 issue.journal_start() in prepare_write() causes different ranking
> violations if copy_from_user() triggers a page fault. It could cause
> GFP_FS allocation, re-entering into ext3 code possibly with a different
> superblock and journal, ranking violation of journalling serialization
> and mmap_sem and page lock and all other kinds of funny consequences.

With the stuff Nick and I are looking at, we won't take pagefaults inside
prepare_write()/commit_write() any more.

> Our customers complain about this issue.

Really? How often?

What on earth are they doing to trigger this? writev() without the 2.6.18
writev() bugfix?

Subject: Re: [RFC][PATCH] EXT3: problem with page fault inside a transaction
Posted by Nick Piggin on Thu, 12 Oct 2006 07:01:03 GMT

View Forum Message <> Reply to Message

Andrew Morton wrote:
> On Thu, 12 Oct 2006 09:57:26 +0400
> Dmitriy Monakhov <dmonakhov@openvz.org> wrote:
>
>
>>While reading Andrew's generic_file_buffered_write patches i've remembered
>>one more EXT3 issue.journal_start() in prepare_write() causes different ranking
>>violations if copy_from_user() triggers a page fault. It could cause
>>GFP_FS allocation, re-entering into ext3 code possibly with a different
>>superblock and journal, ranking violation of journalling serialization
>>and mmap_sem and page lock and all other kinds of funny consequences.
>
>
> With the stuff Nick and I are looking at, we won't take pagefaults inside
> prepare_write()/commit_write() any more.

Yep. Because the page is locked, it is too much to deal with even
without a filesystem in the picture.

--
SUSE Labs, Novell Inc.

Subject: Re: [RFC][PATCH] EXT3: problem with page fault inside a transaction
Posted by Dmitriy Monakhov on Thu, 12 Oct 2006 07:53:56 GMT
View Forum Message <> Reply to Message

Andrew Morton <akpm@osdl.org> writes:

> On Thu, 12 Oct 2006 09:57:26 +0400
> Dmitriy Monakhov <dmonakhov@openvz.org> wrote:
>
>> While reading Andrew's generic_file_buffered_write patches i've remembered
>> one more EXT3 issue.journal_start() in prepare_write() causes different ranking
>> violations if copy_from_user() triggers a page fault. It could cause
>> GFP_FS allocation, re-entering into ext3 code possibly with a different
>> superblock and journal, ranking violation of journalling serialization
>> and mmap_sem and page lock and all other kinds of funny consequences.
>
> With the stuff Nick and I are looking at, we won't take pagefaults inside
> prepare_write()/commit_write() any more.

I'sorry may be i've missed something, but how cant you prevent this?

Let's look at generic_file_buffered_write:
#### force page fault
fault_in_pages_readable();

### find and lock page
 __grab_cache_page()

#### allocate blocks.This may result in low memory condition
#### try_to_free_pages->shrink_caches() and etc.
a_ops->prepare_write()

### can anyone guarantee that page fault hasn't happened by now ?
### user space buffer swapped out, or became invalid.
filemap_copy_from_user()

>
>> Our customers complain about this issue.
>
> Really? How often?

I have't concrete statistic

>
> What on earth are they doing to trigger this? writev() without the 2.6.18
> writev() bugfix?

Subject: Re: [RFC][PATCH] EXT3: problem with page fault inside a transaction
Posted by Nick Piggin on Thu, 12 Oct 2006 08:31:23 GMT
View Forum Message <> Reply to Message

Dmitriy Monakhov wrote:
> Andrew Morton <akpm@osdl.org> writes:
>
>>With the stuff Nick and I are looking at, we won't take pagefaults inside
>>prepare_write()/commit_write() any more.
>
> I'sorry may be i've missed something, but how cant you prevent this?
>
> Let's look at generic_file_buffered_write:
> #### force page fault
> fault_in_pages_readable();
>
> ### find and lock page
> __grab_cache_page()
>
> #### allocate blocks.This may result in low memory condition
> #### try_to_free_pages->shrink_caches() and etc.
> a_ops->prepare_write()
>
> ### can anyone guarantee that page fault hasn't happened by now ?

Yes. Do an atomic copy, which will early exit from the pagefault handler
and return a short copy. Then close up the write, drop the page lock,
and rerun the fault_in_pages_readable, which will do the full pagefaults
for us, then try again.

Regardless of what you do to ext3, the VM just can't handle a fault
here anyway.

--
SUSE Labs, Novell Inc.
Send instant messages to your online friends http://au.messenger.yahoo.com

Subject: Re: [RFC][PATCH] EXT3: problem with page fault inside a transaction
Posted by Andrew Morton on Thu, 12 Oct 2006 08:37:02 GMT
View Forum Message <> Reply to Message

On Thu, 12 Oct 2006 11:53:56 +0400
Dmitriy Monakhov <dmonakhov@sw.ru> wrote:

> > With the stuff Nick and I are looking at, we won't take pagefaults inside
> > prepare_write()/commit_write() any more.
> I'sorry may be i've missed something, but how cant you prevent this?

Start here: http://lkml.org/lkml/2006/10/11/12