

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

## Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

---

*Source:* <http://linux.derkeiler.com/Mailing-Lists/Kernel/2006-12/msg06122.html>

---

- *From:* Peter Zijlstra <[a.p.zijlstra@xxxxxxxxx](mailto:a.p.zijlstra@xxxxxxxxx)>
  - *Date:* Wed, 20 Dec 2006 14:56:18 +0100
- 

On Wed, 2006-12-20 at 13:00 +0000, Hugh Dickins wrote:

On Wed, 20 Dec 2006, Peter Zijlstra wrote:

fix page\_mkclean\_one()

Congratulations on getting to the bottom of it, Peter (if you have: I haven't digested enough of the thread to tell).

Well, I thought I understood, you just shattered that.

I'm mostly offline at present, no time for dialogue, I'll throw out a few remarks and run...

I wondered where you were ;-) Enjoy your time away from the computer.

it had several issues:  
- it failed to flush the cache

It's unclear to me why it should need to flush the cache, but I don't know much about that, and mprotect does flush the cache in advance - I think others will tell you that if it does need to be flushed,

I was still thinking about why exactly, but indeed since mprotect does I thought it prudent to also do it.

it must  
be flushed while there's still a valid pte (on some arches at least).

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

Ah, good point, makes sense I guess.

– it failed to flush the tlb

Eh? It flushed the TLB inside ptep\_establish, didn't it?  
I guess you mean you've found a race before it flushed the TLB.

Hmm, quite right indeed. I missed that. So moving the flush inside the pte cleared section closed a race. It seems I must have a long hard look at these architecture manuals...

– it failed to do s390 (s390 guys, please verify this is now correct)

Hmm, I thought we cleared it with them back at the time.

/me queries mail folder...  
can't seem to find it.

Also, clear in a loop to ensure SMP safeness as suggested by Arjan.

Yikes. Well, please compare with mprotect's change\_pte\_range. I think I took that as the relevant standard when checking your implementation, and back then satisfied myself that what you were doing was equivalent. If page\_mkclean\_one is now agreed to be significantly defective, then I suspect change\_pte\_range is also; perhaps others too.

Arjan argued that mprotect and msync would mostly race with themselves in userspace.

(But I haven't found time to do more than skim through the thread, I've not thought through the issues at all: I am surprised that it's now found defective, we looked at it long and hard back then.)

---

page\_mkclean\_one() fix

it had several issues:

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

- it failed to flush the cache
- a race wrt tlb flushing
- it failed to do s390 (s390 guys, please verify this is now correct)

Also, clear in a loop to ensure SMP safeness as suggested by Arjan.

Signed-off-by: Peter Zijlstra <a.p.zijlstra@xxxxxxxxxx>

---

mm/rmap.c | 23 ++++++-----  
1 file changed, 13 insertions(+), 10 deletions(-)

Index: linux-2.6/mm/rmap.c

=====

--- linux-2.6.orig/mm/rmap.c

+++ linux-2.6/mm/rmap.c

@@ -432,7 +432,7 @@ static int page\_mkclean\_one(struct page

```
{  
    struct mm_struct *mm = vma->vm_mm;  
    unsigned long address;  
    - pte_t *pte, entry;  
    + pte_t *pte;  
    spinlock_t *ptl;  
    int ret = 0;
```

```
@@ -444,17 +444,20 @@ static int page_mkclean_one(struct page  
if (!pte)  
goto out;
```

```
- if (!pte_dirty(*pte) && !pte_write(*pte))  
- goto unlock;  
+ while (pte_dirty(*pte) || pte_write(*pte)) {  
+ pte_t entry;
```

```
- entry = ptep_get_and_clear(mm, address, pte);  
- entry = pte_mkclean(entry);  
- entry = pte_wrprotect(entry);  
- ptep_establish(vma, address, pte, entry);  
- lazy_mmu_prot_update(entry);  
- ret = 1;  
+ flush_cache_page(vma, address, pte_pfn(*pte));  
+ entry = ptep_get_and_clear(mm, address, pte);  
+ flush_tlb_page(vma, address);  
+ (void)page_test_and_clear_dirty(page); /* do the s390 thing */  
+ entry = pte_wrprotect(entry);  
+ entry = pte_mkclean(entry);  
+ set_pte_at(vma, address, pte, entry);  
+ lazy_mmu_prot_update(entry);  
+ ret = 1;  
+ }
```

```
-unlock:
```

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

Re: [PATCH] mm: fix page\_mkclean\_one (was: 2.6.19 file content corruption on ext3)

```
pte_unmap_unlock(pte, ptl);  
out:  
return ret;
```

—

To unsubscribe from this list: send the line "unsubscribe linux-kernel" in  
the body of a message to majordomo@xxxxxxxxxxxxxxxxxxx  
More majordomo info at <http://vger.kernel.org/majordomo-info.html>  
Please read the FAQ at <http://www.tux.org/lkml/>