Subject: [-mm PATCH 4/10] Memory controller memory accounting (v7)
Posted by Balbir Singh on Fri, 24 Aug 2007 15:20:30 GMT
View Forum Message <> Reply to Message

Changelog for v6
1. Do a css_put() in the case of a race in allocating page containers
   (YAMAMOTO Takashi)

Changelog for v5
1. Rename meta_page to page_container
2. Remove PG_metapage and use the lower bit of the page_container pointer
   for locking

Changelog for v3

1. Fix a probable leak with meta_page's (pointed out by Paul Menage)
2. Introduce a wrapper around mem_container_uncharge for uncharging pages
   mem_container_uncharge_page()

Changelog

1. Improved error handling, uncharge on errors and check to see if we are
   leaking pages (review by YAMAMOTO Takashi)

Add the accounting hooks. The accounting is carried out for RSS and Page
Cache (unmapped) pages. There is now a common limit and accounting for both.
The RSS accounting is accounted at page_add_*_rmap() and page_remove_rmap()
time. Page cache is accounted at add_to_page_cache(),
__delete_from_page_cache(). Swap cache is also accounted for.

Each page's page_container is protected with the last bit of the page_container
pointer, this makes handling of race conditions involving simultaneous
mappings of a page easier. A reference count is kept in the page_container to
deal with cases where a page might be unmapped from the RSS of all tasks, but
still lives in the page cache.

Credits go to Vaidyanathan Srinivasan for helping with reference counting work
of the page container. Almost all of the page cache accounting code has help
from Vaidyanathan Srinivasan.

Signed-off-by: Vaidyanathan Srinivasan <svaidy@linux.vnet.ibm.com>
Signed-off-by: Balbir Singh <balbir@linux.vnet.ibm.com>
---

 include/linux/memcontrol.h |   20 +++++
 mm/filemap.c               |   12 ++-
 mm/memcontrol.c            |  166 ++++++++++++++++++++++++++++++++++++++++++++-
 mm/memory.c                |   43 ++++++++++-
 

```
 mm/migrate.c              |    6 +
 mm/page_alloc.c           |    3
 mm/rmap.c                 |   17 ++++
 mm/swap_state.c           |   12 ++-
 mm/swapfile.c             |   41 ++++++-----
 9 files changed, 293 insertions(+), 27 deletions(-)

diff -puN include/linux/memcontrol.h~mem-control-accounting include/linux/memcontrol.h
--- linux-2.6.23-rc2-mm2/include/linux/memcontrol.h~mem-control-accounting 2007-08-24
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/include/linux/memcontrol.h 2007-08-24 20:46:07.000000000
+0530
@@ -30,6 +30,13 @@ extern void mm_free_container(struct mm_
 extern void page_assign_page_container(struct page *page,
      struct page_container *pc);
 extern struct page_container *page_get_page_container(struct page *page);
+extern int mem_container_charge(struct page *page, struct mm_struct *mm);
+extern void mem_container_uncharge(struct page_container *pc);
+
+static inline void mem_container_uncharge_page(struct page *page)
+{
+ mem_container_uncharge(page_get_page_container(page));
+}

 #else /* CONFIG_CONTAINER_MEM_CONT */
 static inline void mm_init_container(struct mm_struct *mm,
@@ -51,6 +58,19 @@ static inline struct page_container *pag
  return NULL;
 }

+static inline int mem_container_charge(struct page *page, struct mm_struct *mm)
+{
+ return 0;
+}
+
+static inline void mem_container_uncharge(struct page_container *pc)
+{
+}
+
+static inline void mem_container_uncharge_page(struct page *page)
+{
+}
+
 #endif /* CONFIG_CONTAINER_MEM_CONT */

 #endif /* _LINUX_MEMCONTROL_H */
diff -puN include/linux/page-flags.h~mem-control-accounting include/linux/page-flags.h
diff -puN mm/filemap.c~mem-control-accounting mm/filemap.c
```

```
--- linux-2.6.23-rc2-mm2/mm/filemap.c~mem-control-accounting 2007-08-24 20:46:07.000000000
+0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/filemap.c 2007-08-24 20:46:07.000000000 +0530
@@ -31,6 +31,7 @@
 #include <linux/syscalls.h>
 #include <linux/cpuset.h>
 #include <linux/hardirq.h> /* for BUG_ON(!in_atomic()) only */
+#include <linux/memcontrol.h>
 #include "internal.h"

 /*
@@ -116,6 +117,7 @@ void __remove_from_page_cache(struct pag
 {
  struct address_space *mapping = page->mapping;

+ mem_container_uncharge_page(page);
  radix_tree_delete(&mapping->page_tree, page->index);
  page->mapping = NULL;
  mapping->nrpages--;
@@ -442,6 +444,11 @@ int add_to_page_cache(struct page *page,
  int error = radix_tree_preload(gfp_mask & ~__GFP_HIGHMEM);

  if (error == 0) {
+
+  error = mem_container_charge(page, current->mm);
+  if (error)
+   goto out;
+
   write_lock_irq(&mapping->tree_lock);
   error = radix_tree_insert(&mapping->page_tree, offset, page);
   if (!error) {
@@ -451,10 +458,13 @@ int add_to_page_cache(struct page *page,
    page->index = offset;
    mapping->nrpages++;
    __inc_zone_page_state(page, NR_FILE_PAGES);
-  }
+  } else
+   mem_container_uncharge_page(page);
+
   write_unlock_irq(&mapping->tree_lock);
   radix_tree_preload_end();
  }
+out:
  return error;
 }
 EXPORT_SYMBOL(add_to_page_cache);
diff -puN mm/memcontrol.c~mem-control-accounting mm/memcontrol.c
--- linux-2.6.23-rc2-mm2/mm/memcontrol.c~mem-control-accounting 2007-08-24
```

```
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/memcontrol.c 2007-08-24 20:46:07.000000000 +0530
@@ -21,6 +21,9 @@
 #include <linux/memcontrol.h>
 #include <linux/container.h>
 #include <linux/mm.h>
+#include <linux/page-flags.h>
+#include <linux/bit_spinlock.h>
+#include <linux/rcupdate.h>

 struct container_subsys mem_container_subsys;

@@ -31,7 +34,9 @@ struct container_subsys mem_container_su
  * to help the administrator determine what knobs to tune.
  *
  * TODO: Add a water mark for the memory controller. Reclaim will begin when
- * we hit the water mark.
+ * we hit the water mark. May be even add a low water mark, such that
+ * no reclaim occurs from a container at it's low water mark, this is
+ * a feature that will be implemented much later in the future.
  */
 struct mem_container {
  struct container_subsys_state css;
@@ -49,6 +54,14 @@ struct mem_container {
 };

 /*
+ * We use the lower bit of the page->page_container pointer as a bit spin
+ * lock. We need to ensure that page->page_container is atleast two
+ * byte aligned (based on comments from Nick Piggin)
+ */
+#define PAGE_CONTAINER_LOCK_BIT  0x0
+#define PAGE_CONTAINER_LOCK   (1 << PAGE_CONTAINER_LOCK_BIT)
+
+/*
  * A page_container page is associated with every page descriptor. The
  * page_container helps us identify information about the container
  */
@@ -56,6 +69,8 @@ struct page_container {
  struct list_head lru;  /* per container LRU list */
  struct page *page;
  struct mem_container *mem_container;
+ atomic_t ref_cnt;  /* Helpful when pages move b/w  */
+     /* mapped and cached states     */
 };


@@ -88,14 +103,157 @@ void mm_free_container(struct mm_struct
```

```
 css_put(&mm->mem_container->css);
}

+static inline int page_container_locked(struct page *page)
+{
+ return bit_spin_is_locked(PAGE_CONTAINER_LOCK_BIT,
+     &page->page_container);
+}
+
 void page_assign_page_container(struct page *page, struct page_container *pc)
 {
- page->page_container = (unsigned long)pc;
+ int locked;
+
+ /*
+  * While resetting the page_container we might not hold the
+  * page_container lock. free_hot_cold_page() is an example
+  * of such a scenario
+  */
+ if (pc)
+  VM_BUG_ON(!page_container_locked(page));
+ locked = (page->page_container & PAGE_CONTAINER_LOCK);
+ page->page_container = ((unsigned long)pc | locked);
 }

 struct page_container *page_get_page_container(struct page *page)
 {
- return page->page_container;
+ return (struct page_container *)
+  (page->page_container & ~PAGE_CONTAINER_LOCK);
+}
+
+void __always_inline lock_page_container(struct page *page)
+{
+ bit_spin_lock(PAGE_CONTAINER_LOCK_BIT, &page->page_container);
+ VM_BUG_ON(!page_container_locked(page));
+}
+
+void __always_inline unlock_page_container(struct page *page)
+{
+ bit_spin_unlock(PAGE_CONTAINER_LOCK_BIT, &page->page_container);
+}
+
+/*
+ * Charge the memory controller for page usage.
+ * Return
+ * 0 if the charge was successful
+ * < 0 if the container is over its limit
```

```
+ */
+int mem_container_charge(struct page *page, struct mm_struct *mm)
+{
+ struct mem_container *mem;
+ struct page_container *pc, *race_pc;
+
+ /*
+  * Should page_container's go to their own slab?
+  * One could optimize the performance of the charging routine
+  * by saving a bit in the page_flags and using it as a lock
+  * to see if the container page already has a page_container associated
+  * with it
+  */
+ lock_page_container(page);
+ pc = page_get_page_container(page);
+ /*
+  * The page_container exists and the page has already been accounted
+  */
+ if (pc) {
+  atomic_inc(&pc->ref_cnt);
+  goto done;
+ }
+
+ unlock_page_container(page);
+
+ pc = kzalloc(sizeof(struct page_container), GFP_KERNEL);
+ if (pc == NULL)
+  goto err;
+
+ rcu_read_lock();
+ /*
+  * We always charge the container the mm_struct belongs to
+  * the mm_struct's mem_container changes on task migration if the
+  * thread group leader migrates. It's possible that mm is not
+  * set, if so charge the init_mm (happens for pagecache usage).
+  */
+ if (!mm)
+  mm = &init_mm;
+
+ mem = rcu_dereference(mm->mem_container);
+ /*
+  * For every charge from the container, increment reference
+  * count
+  */
+ css_get(&mem->css);
+ rcu_read_unlock();
+
+ /*
```

```
+  * If we created the page_container, we should free it on exceeding
+  * the container limit.
+  */
+ if (res_counter_charge(&mem->res, 1)) {
+  css_put(&mem->css);
+  goto free_pc;
+ }
+
+ lock_page_container(page);
+ /*
+  * Check if somebody else beat us to allocating the page_container
+  */
+ race_pc = page_get_page_container(page);
+ if (race_pc) {
+  kfree(pc);
+  pc = race_pc;
+  atomic_inc(&pc->ref_cnt);
+  res_counter_uncharge(&mem->res, 1);
+  css_put(&mem->css);
+  goto done;
+ }
+
+ atomic_set(&pc->ref_cnt, 1);
+ pc->mem_container = mem;
+ pc->page = page;
+ page_assign_page_container(page, pc);
+
+done:
+ unlock_page_container(page);
+ return 0;
+free_pc:
+ kfree(pc);
+ return -ENOMEM;
+err:
+ unlock_page_container(page);
+ return -ENOMEM;
+}
+
+/*
+ * Uncharging is always a welcome operation, we never complain, simply
+ * uncharge.
+ */
+void mem_container_uncharge(struct page_container *pc)
+{
+ struct mem_container *mem;
+ struct page *page;
+
+ if (!pc)
```

```
+  return;
+
+ if (atomic_dec_and_test(&pc->ref_cnt)) {
+  page = pc->page;
+  lock_page_container(page);
+  mem = pc->mem_container;
+  css_put(&mem->css);
+  page_assign_page_container(page, NULL);
+  unlock_page_container(page);
+  res_counter_uncharge(&mem->res, 1);
+  kfree(pc);
+ }
  }

  static ssize_t mem_container_read(struct container *cont, struct cftype *cft,
@@ -150,6 +308,8 @@ mem_container_create(struct container_su
    return NULL;

   res_counter_init(&mem->res);
+ INIT_LIST_HEAD(&mem->active_list);
+ INIT_LIST_HEAD(&mem->inactive_list);
   return &mem->css;
  }

diff -puN mm/memory.c~mem-control-accounting mm/memory.c
--- linux-2.6.23-rc2-mm2/mm/memory.c~mem-control-accounting 2007-08-24
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/memory.c 2007-08-24 20:46:07.000000000 +0530
@@ -50,6 +50,7 @@
  #include <linux/delayacct.h>
  #include <linux/init.h>
  #include <linux/writeback.h>
+#include <linux/memcontrol.h>

  #include <asm/pgalloc.h>
  #include <asm/uaccess.h>
@@ -1136,14 +1137,18 @@ static int insert_page(struct mm_struct
   pte_t *pte;
   spinlock_t *ptl;

+ retval = mem_container_charge(page, mm);
+ if (retval)
+  goto out;
+
   retval = -EINVAL;
   if (PageAnon(page))
-  goto out;
+  goto out_uncharge;
```

```
  retval = -ENOMEM;
  flush_dcache_page(page);
  pte = get_locked_pte(mm, addr, &ptl);
  if (!pte)
-  goto out;
+  goto out_uncharge;
  retval = -EBUSY;
  if (!pte_none(*pte))
   goto out_unlock;
@@ -1155,8 +1160,11 @@ static int insert_page(struct mm_struct
  set_pte_at(mm, addr, pte, mk_pte(page, prot));

  retval = 0;
+ return retval;
 out_unlock:
  pte_unmap_unlock(pte, ptl);
+out_uncharge:
+ mem_container_uncharge_page(page);
 out:
  return retval;
 }
@@ -1629,6 +1637,9 @@ gotten:
   goto oom;
  cow_user_page(new_page, old_page, address, vma);

+ if (mem_container_charge(new_page, mm))
+  goto oom_free_new;
+
  /*
   * Re-check the pte - we dropped the lock
   */
@@ -1660,7 +1671,9 @@ gotten:
   /* Free the old page.. */
   new_page = old_page;
   ret |= VM_FAULT_WRITE;
- }
+ } else
+  mem_container_uncharge_page(new_page);
+
  if (new_page)
   page_cache_release(new_page);
  if (old_page)
@@ -1681,6 +1694,8 @@ unlock:
   put_page(dirty_page);
  }
  return ret;
+oom_free_new:
+ __free_page(new_page);
```

```
 oom:
  if (old_page)
   page_cache_release(old_page);
@@ -2085,6 +2100,11 @@ static int do_swap_page(struct mm_struct
  }

  delayacct_clear_flag(DELAYACCT_PF_SWAPIN);
+ if (mem_container_charge(page, mm)) {
+  ret = VM_FAULT_OOM;
+  goto out;
+ }
+
  mark_page_accessed(page);
  lock_page(page);

@@ -2121,8 +2141,10 @@ static int do_swap_page(struct mm_struct
  if (write_access) {
   /* XXX: We could OR the do_wp_page code with this one? */
   if (do_wp_page(mm, vma, address,
-    page_table, pmd, ptl, pte) & VM_FAULT_OOM)
+    page_table, pmd, ptl, pte) & VM_FAULT_OOM) {
+   mem_container_uncharge_page(page);
    ret = VM_FAULT_OOM;
+  }
   goto out;
  }

@@ -2133,6 +2155,7 @@ unlock:
 out:
  return ret;
 out_nomap:
+ mem_container_uncharge_page(page);
  pte_unmap_unlock(page_table, ptl);
  unlock_page(page);
  page_cache_release(page);
@@ -2161,6 +2184,9 @@ static int do_anonymous_page(struct mm_s
  if (!page)
   goto oom;

+  if (mem_container_charge(page, mm))
+   goto oom_free_page;
+
  entry = mk_pte(page, vma->vm_page_prot);
  entry = maybe_mkwrite(pte_mkdirty(entry), vma);

@@ -2178,8 +2204,11 @@ unlock:
  pte_unmap_unlock(page_table, ptl);
  return 0;
```

```
 release:
+ mem_container_uncharge_page(page);
  page_cache_release(page);
  goto unlock;
+oom_free_page:
+ __free_page(page);
 oom:
  return VM_FAULT_OOM;
 }
@@ -2290,6 +2319,11 @@ static int __do_fault(struct mm_struct *

  }

+ if (mem_container_charge(page, mm)) {
+  ret = VM_FAULT_OOM;
+  goto out;
+ }
+
  page_table = pte_offset_map_lock(mm, pmd, address, &ptl);

  /*
@@ -2325,6 +2359,7 @@ static int __do_fault(struct mm_struct *
   /* no need to invalidate: a not-present page won't be cached */
   update_mmu_cache(vma, address, entry);
  } else {
+  mem_container_uncharge_page(page);
   if (anon)
    page_cache_release(page);
   else
diff -puN mm/migrate.c~mem-control-accounting mm/migrate.c
--- linux-2.6.23-rc2-mm2/mm/migrate.c~mem-control-accounting 2007-08-24 20:46:07.000000000
+0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/migrate.c 2007-08-24 20:46:07.000000000 +0530
@@ -29,6 +29,7 @@
 #include <linux/mempolicy.h>
 #include <linux/vmalloc.h>
 #include <linux/security.h>
+#include <linux/memcontrol.h>

 #include "internal.h"

@@ -157,6 +158,11 @@ static void remove_migration_pte(struct
    return;
   }

+ if (mem_container_charge(new, mm)) {
+  pte_unmap(ptep);
+  return;
```

```
+ }
+
   ptl = pte_lockptr(mm, pmd);
   spin_lock(ptl);
  pte = *ptep;
diff -puN mm/page_alloc.c~mem-control-accounting mm/page_alloc.c
--- linux-2.6.23-rc2-mm2/mm/page_alloc.c~mem-control-accounting 2007-08-24
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/page_alloc.c 2007-08-24 20:46:07.000000000 +0530
@@ -42,6 +42,7 @@
 #include <linux/backing-dev.h>
 #include <linux/fault-inject.h>
 #include <linux/page-isolation.h>
+#include <linux/memcontrol.h>

 #include <asm/tlbflush.h>
 #include <asm/div64.h>
@@ -1025,6 +1026,7 @@ static void fastcall free_hot_cold_page(

  if (!PageHighMem(page))
   debug_check_no_locks_freed(page_address(page), PAGE_SIZE);
+ page_assign_page_container(page, NULL);
  arch_free_page(page, 0);
  kernel_map_pages(page, 1, 0);

@@ -2654,6 +2656,7 @@ void __meminit memmap_init_zone(unsigned
   set_page_links(page, zone, nid, pfn);
   init_page_count(page);
   reset_page_mapcount(page);
+  page_assign_page_container(page, NULL);
   SetPageReserved(page);

   /*
diff -puN mm/rmap.c~mem-control-accounting mm/rmap.c
--- linux-2.6.23-rc2-mm2/mm/rmap.c~mem-control-accounting 2007-08-24 20:46:07.000000000
+0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/rmap.c 2007-08-24 20:46:07.000000000 +0530
@@ -48,6 +48,7 @@
 #include <linux/rcupdate.h>
 #include <linux/module.h>
 #include <linux/kallsyms.h>
+#include <linux/memcontrol.h>

 #include <asm/tlbflush.h>

@@ -550,8 +551,14 @@ void page_add_anon_rmap(struct page *pag
  VM_BUG_ON(address < vma->vm_start || address >= vma->vm_end);
  if (atomic_inc_and_test(&page->_mapcount))
```

```
   __page_set_anon_rmap(page, vma, address);
- else
+ else {
   __page_check_anon_rmap(page, vma, address);
+  /*
+   * We unconditionally charged during prepare, we uncharge here
+   * This takes care of balancing the reference counts
+   */
+  mem_container_uncharge_page(page);
+ }
 }

 /*
@@ -582,6 +589,12 @@ void page_add_file_rmap(struct page *pag
 {
  if (atomic_inc_and_test(&page->_mapcount))
   __inc_zone_page_state(page, NR_FILE_MAPPED);
+ else
+  /*
+   * We unconditionally charged during prepare, we uncharge here
+   * This takes care of balancing the reference counts
+   */
+  mem_container_uncharge_page(page);
 }

 #ifdef CONFIG_DEBUG_VM
@@ -642,6 +655,8 @@ void page_remove_rmap(struct page *page,
    page_clear_dirty(page);
    set_page_dirty(page);
   }
+  mem_container_uncharge_page(page);
+
   __dec_zone_page_state(page,
    PageAnon(page) ? NR_ANON_PAGES : NR_FILE_MAPPED);
  }
diff -puN mm/swapfile.c~mem-control-accounting mm/swapfile.c
--- linux-2.6.23-rc2-mm2/mm/swapfile.c~mem-control-accounting 2007-08-24
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/swapfile.c 2007-08-24 20:46:07.000000000 +0530
@@ -27,6 +27,7 @@
 #include <linux/mutex.h>
 #include <linux/capability.h>
 #include <linux/syscalls.h>
+#include <linux/memcontrol.h>

 #include <asm/pgtable.h>
 #include <asm/tlbflush.h>
@@ -506,9 +507,12 @@ unsigned int count_swap_pages(int type,
```

```
  * just let do_wp_page work it out if a write is requested later - to
  * force COW, vm_page_prot omits write permission from any private vma.
  */
-static void unuse_pte(struct vm_area_struct *vma, pte_t *pte,
+static int unuse_pte(struct vm_area_struct *vma, pte_t *pte,
   unsigned long addr, swp_entry_t entry, struct page *page)
 {
+ if (mem_container_charge(page, vma->vm_mm))
+  return -ENOMEM;
+
  inc_mm_counter(vma->vm_mm, anon_rss);
  get_page(page);
  set_pte_at(vma->vm_mm, addr, pte,
@@ -520,6 +524,7 @@ static void unuse_pte(struct vm_area_str
   * immediately swapped out again after swapon.
   */
  activate_page(page);
+ return 1;
 }

 static int unuse_pte_range(struct vm_area_struct *vma, pmd_t *pmd,
@@ -529,7 +534,7 @@ static int unuse_pte_range(struct vm_are
  pte_t swp_pte = swp_entry_to_pte(entry);
  pte_t *pte;
  spinlock_t *ptl;
- int found = 0;
+ int ret = 0;

  pte = pte_offset_map_lock(vma->vm_mm, pmd, addr, &ptl);
  do {
@@ -538,13 +543,12 @@ static int unuse_pte_range(struct vm_are
    * Test inline before going to call unuse_pte.
    */
   if (unlikely(pte_same(*pte, swp_pte))) {
-   unuse_pte(vma, pte++, addr, entry, page);
-   found = 1;
+   ret = unuse_pte(vma, pte++, addr, entry, page);
    break;
   }
  } while (pte++, addr += PAGE_SIZE, addr != end);
  pte_unmap_unlock(pte - 1, ptl);
- return found;
+ return ret;
 }

 static inline int unuse_pmd_range(struct vm_area_struct *vma, pud_t *pud,
@@ -553,14 +557,16 @@ static inline int unuse_pmd_range(struct
 {
```

```
  pmd_t *pmd;
  unsigned long next;
+ int ret;

  pmd = pmd_offset(pud, addr);
  do {
   next = pmd_addr_end(addr, end);
   if (pmd_none_or_clear_bad(pmd))
    continue;
-  if (unuse_pte_range(vma, pmd, addr, next, entry, page))
-   return 1;
+  ret = unuse_pte_range(vma, pmd, addr, next, entry, page);
+  if (ret)
+   return ret;
  } while (pmd++, addr = next, addr != end);
  return 0;
 }
@@ -571,14 +577,16 @@ static inline int unuse_pud_range(struct
 {
  pud_t *pud;
  unsigned long next;
+ int ret;

  pud = pud_offset(pgd, addr);
  do {
   next = pud_addr_end(addr, end);
   if (pud_none_or_clear_bad(pud))
    continue;
-  if (unuse_pmd_range(vma, pud, addr, next, entry, page))
-   return 1;
+  ret = unuse_pmd_range(vma, pud, addr, next, entry, page);
+  if (ret)
+   return ret;
  } while (pud++, addr = next, addr != end);
  return 0;
 }
@@ -588,6 +596,7 @@ static int unuse_vma(struct vm_area_stru
 {
  pgd_t *pgd;
  unsigned long addr, end, next;
+ int ret;

  if (page->mapping) {
   addr = page_address_in_vma(page, vma);
@@ -605,8 +614,9 @@ static int unuse_vma(struct vm_area_stru
   next = pgd_addr_end(addr, end);
   if (pgd_none_or_clear_bad(pgd))
    continue;
```

```
-  if (unuse_pud_range(vma, pgd, addr, next, entry, page))
-   return 1;
+  ret = unuse_pud_range(vma, pgd, addr, next, entry, page);
+  if (ret)
+   return ret;
  } while (pgd++, addr = next, addr != end);
  return 0;
 }
@@ -615,6 +625,7 @@ static int unuse_mm(struct mm_struct *mm
     swp_entry_t entry, struct page *page)
 {
  struct vm_area_struct *vma;
+ int ret = 0;

  if (!down_read_trylock(&mm->mmap_sem)) {
   /*
@@ -627,15 +638,11 @@ static int unuse_mm(struct mm_struct *mm
   lock_page(page);
  }
  for (vma = mm->mmap; vma; vma = vma->vm_next) {
-  if (vma->anon_vma && unuse_vma(vma, entry, page))
+  if (vma->anon_vma && (ret = unuse_vma(vma, entry, page)))
    break;
  }
  up_read(&mm->mmap_sem);
- /*
-  * Currently unuse_mm cannot fail, but leave error handling
-  * at call sites for now, since we change it from time to time.
-  */
- return 0;
+ return ret;
 }

 /*
diff -puN mm/swap_state.c~mem-control-accounting mm/swap_state.c
--- linux-2.6.23-rc2-mm2/mm/swap_state.c~mem-control-accounting 2007-08-24
20:46:07.000000000 +0530
+++ linux-2.6.23-rc2-mm2-balbir/mm/swap_state.c 2007-08-24 20:46:07.000000000 +0530
@@ -17,6 +17,7 @@
 #include <linux/backing-dev.h>
 #include <linux/pagevec.h>
 #include <linux/migrate.h>
+#include <linux/memcontrol.h>

 #include <asm/pgtable.h>

@@ -80,6 +81,11 @@ static int __add_to_swap_cache(struct pa
  BUG_ON(PagePrivate(page));
```

```
  error = radix_tree_preload(gfp_mask);
  if (!error) {
+
+  error = mem_container_charge(page, current->mm);
+  if (error)
+   goto out;
+
   write_lock_irq(&swapper_space.tree_lock);
   error = radix_tree_insert(&swapper_space.page_tree,
       entry.val, page);
@@ -89,10 +95,13 @@ static int __add_to_swap_cache(struct pa
    set_page_private(page, entry.val);
    total_swapcache_pages++;
    __inc_zone_page_state(page, NR_FILE_PAGES);
-  }
+  } else
+   mem_container_uncharge_page(page);
+
   write_unlock_irq(&swapper_space.tree_lock);
   radix_tree_preload_end();
  }
+out:
  return error;
 }

@@ -132,6 +141,7 @@ void __delete_from_swap_cache(struct pag
  BUG_ON(PageWriteback(page));
  BUG_ON(PagePrivate(page));

+ mem_container_uncharge_page(page);
  radix_tree_delete(&swapper_space.page_tree, page_private(page));
  set_page_private(page, 0);
  ClearPageSwapCache(page);
```

_

--
Warm Regards,
Balbir Singh
Linux Technology Center
IBM, ISTL