
Subject: [-mm PATCH 3/9] Memory controller accounting setup (v5)

Posted by [Balbir Singh](#) on Mon, 13 Aug 2007 17:43:46 GMT

[View Forum Message](#) <> [Reply to Message](#)

From: Pavel Emelianov <xemul@openvz.org>

Changelog for v5

1. Remove inclusion of memcontrol.h from mm_types.h

Changelog

As per Paul's review comments

1. Drop css_get() for the root memory container
2. Use mem_container_from_task() as an optimization instead of using mem_container_from_cont() along with task_container.

Basic setup routines, the mm_struct has a pointer to the container that it belongs to and the the page has a meta_page associated with it.

Signed-off-by: Pavel Emelianov <xemul@openvz.org>

Signed-off-by: <balbir@linux.vnet.ibm.com>

```
include/linux/memcontrol.h | 35 ++++++
include/linux/mm_types.h   | 3 ++
include/linux/sched.h      | 4 +++
kernel/fork.c              | 11 ++++++
mm/memcontrol.c            | 57 ++++++
5 files changed, 103 insertions(+), 7 deletions(-)
```

```
diff -puN include/linux/memcontrol.h~mem-control-accounting-setup include/linux/memcontrol.h
--- linux-2.6.23-rc1-mm1/include/linux/memcontrol.h~mem-control-accounting-setup 2007-08-13
23:06:11.000000000 +0530
```

```
+++ linux-2.6.23-rc1-mm1-balbir/include/linux/memcontrol.h 2007-08-13 23:06:11.000000000
+0530
```

```
@@ -3,6 +3,9 @@
```

```
* Copyright IBM Corporation, 2007
```

```
* Author Balbir Singh <balbir@linux.vnet.ibm.com>
```

```
*
```

```
+ * Copyright 2007 OpenVZ SWsoft Inc
```

```
+ * Author: Pavel Emelianov <xemul@openvz.org>
```

```
+ *
```

```
* This program is free software; you can redistribute it and/or modify
```

```
* it under the terms of the GNU General Public License as published by
```

```
* the Free Software Foundation; either version 2 of the License, or
```

```

@@ -17,5 +20,37 @@
#ifndef _LINUX_MEMCONTROL_H
#define _LINUX_MEMCONTROL_H

+struct mem_container;
+struct meta_page;
+
+#ifdef CONFIG_CONTAINER_MEM_CONT
+
+extern void mm_init_container(struct mm_struct *mm, struct task_struct *p);
+extern void mm_free_container(struct mm_struct *mm);
+extern void page_assign_meta_page(struct page *page, struct meta_page *mp);
+extern struct meta_page *page_get_meta_page(struct page *page);
+
+#else /* CONFIG_CONTAINER_MEM_CONT */
+static inline void mm_init_container(struct mm_struct *mm,
+    struct task_struct *p)
+{
+}
+
+static inline void mm_free_container(struct mm_struct *mm)
+{
+}
+
+static inline void page_assign_meta_page(struct page *page,
+    struct meta_page *mp)
+{
+}
+
+static inline struct meta_page *page_get_meta_page(struct page *page)
+{
+    return NULL;
+}
+
+#endif /* CONFIG_CONTAINER_MEM_CONT */
+
#endif /* _LINUX_MEMCONTROL_H */

```

```

diff -puN include/linux/mm_types.h~mem-control-accounting-setup include/linux/mm_types.h
--- linux-2.6.23-rc1-mm1/include/linux/mm_types.h~mem-control-accounting-setup 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/include/linux/mm_types.h 2007-08-13 23:06:11.000000000
+0530
@@ -83,6 +83,9 @@ struct page {
    unsigned int gfp_mask;
    unsigned long trace[8];
#endif
#ifdef CONFIG_CONTAINER_MEM_CONT

```

```

+ unsigned long page_container;
+ #endif
};

# endif /* _LINUX_MM_TYPES_H */
diff -puN include/linux/sched.h~mem-control-accounting-setup include/linux/sched.h
--- linux-2.6.23-rc1-mm1/include/linux/sched.h~mem-control-accounting-setup 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/include/linux/sched.h 2007-08-13 23:06:11.000000000 +0530
@@ -89,6 +89,7 @@ struct sched_param {

#include <asm/processor.h>

+ struct mem_container;
+ struct exec_domain;
+ struct futex_pi_state;
+ struct bio;
@@ -433,6 +434,9 @@ struct mm_struct {
/* aio bits */
+ rwlock_t ioctx_list_lock;
+ struct kiocx *ioctx_list;
+ #ifdef CONFIG_CONTAINER_MEM_CONT
+ struct mem_container *mem_container;
+ #endif
};

+ struct sighand_struct {
diff -puN kernel/fork.c~mem-control-accounting-setup kernel/fork.c
--- linux-2.6.23-rc1-mm1/kernel/fork.c~mem-control-accounting-setup 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/kernel/fork.c 2007-08-13 23:06:11.000000000 +0530
@@ -50,6 +50,7 @@
#include <linux/taskstats_kern.h>
#include <linux/random.h>
#include <linux/tty.h>
+ #include <linux/memcontrol.h>

#include <asm/pgtable.h>
#include <asm/pgalloc.h>
@@ -328,7 +329,7 @@ __cacheline_aligned_in_smp DEFINE_SPINLO

#include <linux/init_task.h>

- static struct mm_struct * mm_init(struct mm_struct * mm)
+ static struct mm_struct * mm_init(struct mm_struct * mm, struct task_struct * p)
{
+ atomic_set(&mm->mm_users, 1);
+ atomic_set(&mm->mm_count, 1);

```

```

@@ -345,11 +346,14 @@ static struct mm_struct * mm_init(struct
    mm->iocx_list = NULL;
    mm->free_area_cache = TASK_UNMAPPED_BASE;
    mm->cached_hole_size = ~0UL;
+ mm_init_container(mm, p);

    if (likely(!mm_alloc_pgd(mm))) {
        mm->def_flags = 0;
        return mm;
    }
+
+ mm_free_container(mm);
    free_mm(mm);
    return NULL;
}
@@ -364,7 +368,7 @@ struct mm_struct * mm_alloc(void)
    mm = allocate_mm();
    if (mm) {
        memset(mm, 0, sizeof(*mm));
- mm = mm_init(mm);
+ mm = mm_init(mm, current);
    }
    return mm;
}
@@ -378,6 +382,7 @@ void fastcall __mmdrop(struct mm_struct
{
    BUG_ON(mm == &init_mm);
    mm_free_pgd(mm);
+ mm_free_container(mm);
    destroy_context(mm);
    free_mm(mm);
}
@@ -498,7 +503,7 @@ static struct mm_struct *dup_mm(struct t
    mm->token_priority = 0;
    mm->last_interval = 0;

- if (!mm_init(mm))
+ if (!mm_init(mm, tsk))
    goto fail_nomem;

    if (init_new_context(tsk, mm))
diff -puN mm/memcontrol.c~mem-control-accounting-setup mm/memcontrol.c
--- linux-2.6.23-rc1-mm1/mm/memcontrol.c~mem-control-accounting-setup 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/memcontrol.c 2007-08-13 23:06:11.000000000 +0530
@@ -3,6 +3,9 @@
 * Copyright IBM Corporation, 2007
 * Author Balbir Singh <balbir@linux.vnet.ibm.com>

```

```

*
+ * Copyright 2007 OpenVZ SWsoft Inc
+ * Author: Pavel Emelianov <xemul@openvz.org>
+ *
* This program is free software; you can redistribute it and/or modify
* it under the terms of the GNU General Public License as published by
* the Free Software Foundation; either version 2 of the License, or
@@ -17,6 +20,7 @@
#include <linux/res_counter.h>
#include <linux/memcontrol.h>
#include <linux/container.h>
+#include <linux/mm.h>

struct container_subsys mem_container_subsys;

@@ -35,6 +39,13 @@ struct mem_container {
    * the counter to account for memory usage
    */
    struct res_counter res;
+ /*
+  * Per container active and inactive list, similar to the
+  * per zone LRU lists.
+  * TODO: Consider making these lists per zone
+  */
+ struct list_head active_list;
+ struct list_head inactive_list;
};

/*
@@ -56,6 +67,37 @@ struct mem_container *mem_container_from
    css);
}

+static inline
+struct mem_container *mem_container_from_task(struct task_struct *p)
+{
+ return container_of(task_subsys_state(p, mem_container_subsys_id),
+  struct mem_container, css);
+}
+
+void mm_init_container(struct mm_struct *mm, struct task_struct *p)
+{
+ struct mem_container *mem;
+
+ mem = mem_container_from_task(p);
+ css_get(&mem->css);
+ mm->mem_container = mem;
+}

```

```

+
+void mm_free_container(struct mm_struct *mm)
+{
+ css_put(&mm->mem_container->css);
+}
+
+void page_assign_meta_page(struct page *page, struct meta_page *mp)
+{
+ page->meta_page = mp;
+}
+
+struct meta_page *page_get_meta_page(struct page *page)
+{
+ return page->meta_page;
+}
+
static ssize_t mem_container_read(struct container *cont, struct cftype *cft,
    struct file *file, char __user *userbuf, size_t nbytes,
    loff_t *ppos)
@@ -91,14 +133,21 @@ static struct cftype mem_container_files
    },
};

+static struct mem_container init_mem_container;
+
static struct container_subsys_state *
mem_container_create(struct container_subsys *ss, struct container *cont)
{
    struct mem_container *mem;

- mem = kzalloc(sizeof(struct mem_container), GFP_KERNEL);
- if (!mem)
- return -ENOMEM;
+ if (unlikely((cont->parent) == NULL)) {
+ mem = &init_mem_container;
+ init_mm.mem_container = mem;
+ } else
+ mem = kzalloc(sizeof(struct mem_container), GFP_KERNEL);
+
+ if (mem == NULL)
+ return NULL;

    res_counter_init(&mem->res);
    return &mem->css;
}
@@ -123,5 +172,5 @@ struct container_subsys mem_container_su
    .create = mem_container_create,
    .destroy = mem_container_destroy,
    .populate = mem_container_populate,

```

```
- .early_init = 0,  
+ .early_init = 1,  
};
```

—

--

Warm Regards,
Balbir Singh
Linux Technology Center
IBM, ISTL

Containers mailing list
Containers@lists.linux-foundation.org
<https://lists.linux-foundation.org/mailman/listinfo/containers>

Subject: [-mm PATCH 4/9] Memory controller memory accounting (v5)
Posted by [Balbir Singh](#) on Mon, 13 Aug 2007 17:43:56 GMT
[View Forum Message](#) <> [Reply to Message](#)

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 meta_page is protected with a bit in page flags, this makes handling of race conditions involving simultaneous mappings of a page easier. A reference count is kept in the meta_page 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 meta page. 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@linux.vnet.ibm.com>

```
include/linux/memcontrol.h | 33 ++++++--
mm/filemap.c               | 12 +-
mm/memcontrol.c            | 171 ++++++-----
mm/memory.c               | 44 ++++++--
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, 303 insertions(+), 36 deletions(-)
```

```
diff -puN include/linux/memcontrol.h~mem-control-accounting include/linux/memcontrol.h
--- linux-2.6.23-rc1-mm1/include/linux/memcontrol.h~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/include/linux/memcontrol.h 2007-08-13 23:06:11.000000000
+0530
```

```
@@ -21,14 +21,22 @@
```

```
#define _LINUX_MEMCONTROL_H
```

```
struct mem_container;
-struct meta_page;
+struct page_container;
```

```
#ifdef CONFIG_CONTAINER_MEM_CONT
```

```
extern void mm_init_container(struct mm_struct *mm, struct task_struct *p);
extern void mm_free_container(struct mm_struct *mm);
-extern void page_assign_meta_page(struct page *page, struct meta_page *mp);
-extern struct meta_page *page_get_meta_page(struct page *page);
+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,
@@ -40,16 +48,29 @@ static inline void mm_free_container(str
{
}

-static inline void page_assign_meta_page(struct page *page,
- struct meta_page *mp)
+static inline void page_assign_page_container(struct page *page,
+ struct page_container *pc)
{
}

-static inline struct meta_page *page_get_meta_page(struct page *page)
+static inline struct page_container *page_get_page_container(struct page *page)
{
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-rc1-mm1/mm/filemap.c~mem-control-accounting 2007-08-13 23:06:11.000000000
+0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/filemap.c 2007-08-13 23:06:11.000000000 +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->numpages--;
@@ -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->numpages++;
        __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-rc1-mm1/mm/memcontrol.c~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/memcontrol.c 2007-08-13 23:06:11.000000000 +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;
@@ -49,13 +54,23 @@ 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 meta page is associated with every page descriptor. The meta page
 * helps us identify information about the container
 */
-struct meta_page {
+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 +103,156 @@ void mm_free_container(struct mm_struct
    css_put(&mm->mem_container->css);
}

-void page_assign_meta_page(struct page *page, struct meta_page *mp)
+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)
+{
+ 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)
+{
+ page->meta_page = mp;
+ 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);
+     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;
+ }

-struct meta_page *page_get_meta_page(struct page *page)
+/*
+ * Uncharging is always a welcome operation, we never complain, simply
+ * uncharge.
+ */
+void mem_container_uncharge(struct page_container *pc)
+ {
+     - return page->meta_page;
+     + 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 +307,8 @@ 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-rc1-mm1/mm/memory.c~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/memory.c 2007-08-13 23:06:11.000000000 +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>
@@ -1225,14 +1226,18 @@ 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;
```

```

@@ -1244,8 +1249,11 @@ 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;
}
@@ -1725,6 +1733,9 @@ gotten:
    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
 */
@@ -1757,7 +1768,9 @@ 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)
@@ -1778,6 +1791,8 @@ unlock:
    put_page(dirty_page);
}
return ret;
+oom_free_new:
+ __free_page(new_page);
oom:
if (old_page)
    page_cache_release(old_page);
@@ -2182,6 +2197,11 @@ 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);

@@ -2218,8 +2238,10 @@ 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;
}

@@ -2230,6 +2252,7 @@ unlock:
out:
return ret;
out_nomap:
+ mem_container_uncharge_page(page);
pte_unmap_unlock(page_table, ptl);
unlock_page(page);
page_cache_release(page);
@@ -2250,6 +2273,7 @@ static int do_anonymous_page(struct mm_s
pte_t entry;

if (write_access) {
+
/* Allocate our own private page. */
pte_unmap(page_table);

@@ -2259,6 +2283,9 @@ 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_mkdirty(entry, vma);

@@ -2291,8 +2318,11 @@ 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;
}
@@ -2403,6 +2433,11 @@ 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);

/*
@@ -2439,6 +2474,7 @@ static int __do_fault(struct mm_struct *
  update_mmu_cache(vma, address, entry);
  lazy_mmu_prot_update(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-rc1-mm1/mm/migrate.c~mem-control-accounting 2007-08-13 23:06:11.000000000
+0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/migrate.c 2007-08-13 23:06:11.000000000 +0530
@@ -28,6 +28,7 @@
#include <linux/mempolicy.h>
#include <linux/vmalloc.h>
#include <linux/security.h>
+#include <linux/memcontrol.h>

#include "internal.h"

@@ -156,6 +157,11 @@ 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-rc1-mm1/mm/page_alloc.c~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/page_alloc.c 2007-08-13 23:06:11.000000000 +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>
@@ -1017,6 +1018,7 @@ 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);

@@ -2580,6 +2582,7 @@ 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-rc1-mm1/mm/rmap.c~mem-control-accounting 2007-08-13 23:06:11.000000000
+0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/rmap.c 2007-08-13 23:06:11.000000000 +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>

@@ -551,8 +552,14 @@ 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);
+ }
}

/*
@@ -583,6 +590,12 @@ 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
@@ -643,6 +656,8 @@ 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-rc1-mm1/mm/swapfile.c~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/swapfile.c 2007-08-13 23:06:11.000000000 +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-rc1-mm1/mm/swap_state.c~mem-control-accounting 2007-08-13
23:06:11.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/swap_state.c 2007-08-13 23:06:11.000000000 +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>

@@ -79,6 +80,11 @@ 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;
}

@@ -129,6 +138,7 @@ 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

Containers mailing list
 Containers@lists.linux-foundation.org
<https://lists.linux-foundation.org/mailman/listinfo/containers>

Subject: [-mm PATCH 5/9] Memory controller task migration (v5)
 Posted by [Balbir Singh](#) on Mon, 13 Aug 2007 17:44:08 GMT

Allow tasks to migrate from one container to the other. We migrate mm_struct's mem_container only when the thread group id migrates.

Signed-off-by: <balbir@linux.vnet.ibm.com>

mm/memcontrol.c | 35 +++++
1 file changed, 35 insertions(+)

```
diff -puN mm/memcontrol.c~mem-control-task-migration mm/memcontrol.c
--- linux-2.6.23-rc1-mm1/mm/memcontrol.c~mem-control-task-migration 2007-08-13
23:06:12.000000000 +0530
+++ linux-2.6.23-rc1-mm1-balbir/mm/memcontrol.c 2007-08-13 23:06:12.000000000 +0530
@@ -325,11 +325,46 @@ static int mem_container_populate(struct
     ARRAY_SIZE(mem_container_files));
 }

+static void mem_container_move_task(struct container_subsys *ss,
+ struct container *cont,
+ struct container *old_cont,
+ struct task_struct *p)
+{
+ struct mm_struct *mm;
+ struct mem_container *mem, *old_mem;
+
+ mm = get_task_mm(p);
+ if (mm == NULL)
+ return;
+
+ mem = mem_container_from_cont(cont);
+ old_mem = mem_container_from_cont(old_cont);
+
+ if (mem == old_mem)
+ goto out;
+
+ /*
+ * Only thread group leaders are allowed to migrate, the mm_struct is
+ * in effect owned by the leader
+ */
+ if (p->tgid != p->pid)
+ goto out;
+
+ css_get(&mem->css);
+ rcu_assign_pointer(mm->mem_container, mem);
+ css_put(&old_mem->css);
+
+ }
```

```
+out:
+ mmpout(mm);
+ return;
+}
+
+struct container_subsys mem_container_subsys = {
+ .name = "memory",
+ .subsys_id = mem_container_subsys_id,
+ .create = mem_container_create,
+ .destroy = mem_container_destroy,
+ .populate = mem_container_populate,
+ .attach = mem_container_move_task,
+ .early_init = 1,
+};
```

—

--

Warm Regards,
Balbir Singh
Linux Technology Center
IBM, ISTL

Containers mailing list
Containers@lists.linux-foundation.org
<https://lists.linux-foundation.org/mailman/listinfo/containers>
