Subject: [RFC][PATCH 0/4] Containers: Pagecache accounting and control subsystem (v4)
Posted by Vaidyanathan Srinivas on Wed, 20 Jun 2007 11:33:44 GMT
View Forum Message <> Reply to Message

Containers: Pagecache accounting and control subsystem (v4)

This patch extends the RSS controller to account and reclaim pagecache
and swapcache pages. This is a prototype to demonstrate that the existing
container infrastructure is useful to build different VM controller.

This patch is based on RSS Controller V3.1 by Pavel and Balbir. This patch
depends on

1. Paul Menage's Containers(V10): Generic Process Containers
http://lwn.net/Articles/236032/
2. Pavel Emelianov's RSS controller based on process containers (v3.1)
http://lwn.net/Articles/236817/
3. Balbir's fixes for RSS controller as mentioned in
http://lkml.org/lkml/2007/6/04/185

This is very much work-in-progress and it have been posted for comments
after some basic testing.

Comments, suggestions and criticisms are welcome.

Thanks,
Vaidy

Features:
---
* No new subsystem is added. The RSS controller subsystem is extended
  since most of the code can be shared between pagecache control and
  RSS control.
* The accounting number include pages in swap cache and filesystem
  buffer pages apart from pagecache, basically everything under
  NR_FILE_PAGES is counted as pagecache.
* Limits on pagecache can be set by echo -n 100000 > pagecache_limit on
  the /container file system. The unit is in pages or 4 kilobytes
* If the pagecache utilisation limit is exceeded, the container reclaim
  code is invoked to recover pages from the container.

Advantages:
---
* Minimal code changes to RSS controller to include pagecache pages

Limitations:

---
* All limitation of RSS controller applies to this code as well
* Page reclaim needs to be reworked to select correct pages when the
  respective limits are exceeded
* Concurrent and recursive triggering of reclaimer code is a mess leading
  to deadlocks. Reclaimer needs to be serialised and reworked to
  do the right job and also improve performance

Usage:
---
* Add all dependent patches before including this patch
* No new config settings apart from enabling CONFIG_RSS_CONTAINER
* Boot new kernel
* Mount container filesystem
 mount -t container none /container
 cd /container
* Create new container
 mkdir mybox
 cd /container/mybox
* Add current shell to container
 echo $$ > tasks
* There are two files pagecache_usage and pagecache_limit
* In order to set limit, echo value in pages (4KB) to pagecache_limit
 echo -n 100000 > pagecache_limit
 #This would set 409MB limit on pagecache usage
* Trash the system from current shell using scp/cp/dd/tar etc
* Watch pagecache_usage and /proc/meminfo to verify behavior

Tests:
---
* Simple dd/cat/cp test on pagecache limit
* rss_limit was tested with simple test application that would malloc
  predefined size of memory and touch them to allocate pages.

ToDo:
---
* Optimise the reclaim. Currently isolate_container_pages does not distinguish
  between whether pagecache limit is hit or rss limit is hit
* Prevent concurrent reclaim and recursive reclaim when both limits are set.

Patch Series:
---
pagecache-controller-v4-setup.patch
pagecache-controller-v4-acct.patch
pagecache-controller-v4-acct-hooks.patch
pagecache-controller-v4-reclaim.patch

ChangeLog:

---

v4:
* Patch remerged to Container v10 and RSS v3.1
* Bug fixes
* Tested with kernbench

v3:
* Patch merged with Containers v8 and RSS v2

Subject: [RFC][PATCH 1/4] Pagecache controller setup
Posted by Vaidyanathan Srinivas on Wed, 20 Jun 2007 11:35:40 GMT
View Forum Message <> Reply to Message

Pagecache controller setup

This patch basically adds user interface files in container fs
similar to the rss control files.

pagecache_usage, pagecache_limit and pagecache_failcnt are added
to each container. All units are 'pages' as in rss controller.

pagecache usage is all file backed pages used by the container
which includes swapcache as well.

Separate res_counter for pagecache has been added.

Signed-off-by: Vaidyanathan Srinivasan <svaidy@linux.vnet.ibm.com>
---
 mm/rss_container.c | 43 ++
 1 file changed, 43 insertions(+)

--- linux-2.6.22-rc2-mm1.orig/mm/rss_container.c
+++ linux-2.6.22-rc2-mm1/mm/rss_container.c
@@ -24,6 +24,8 @@ struct rss_container {
   * the counter to account for RSS
   */
  struct res_counter res;
+ /* the counter to account for pagecache pages */
+ struct res_counter pagecache_res;
  /*
   * the lists of pages within the container.
   * actually these lists store the page containers (see below), not
@@ -347,6 +349,7 @@ static int rss_create(struct container_s
   return -ENOMEM;

```
  res_counter_init(&rss->res);
+ res_counter_init(&rss->pagecache_res);
  INIT_LIST_HEAD(&rss->inactive_list);
  INIT_LIST_HEAD(&rss->active_list);
  rss_container_attach(rss, cont);
@@ -359,6 +362,21 @@ static void rss_destroy(struct container
  kfree(rss_from_cont(cont));
 }

+static ssize_t pagecache_read(struct container *cont, struct cftype *cft,
+  struct file *file, char __user *userbuf,
+  size_t nbytes, loff_t *ppos)
+{
+ return res_counter_read(&rss_from_cont(cont)->pagecache_res,
+    cft->private, userbuf, nbytes, ppos);
+}
+
+static ssize_t pagecache_write(struct container *cont, struct cftype *cft,
+  struct file *file, const char __user *userbuf,
+  size_t nbytes, loff_t *ppos)
+{
+ return res_counter_write(&rss_from_cont(cont)->pagecache_res,
+    cft->private, userbuf, nbytes, ppos);
+}

 static ssize_t rss_read(struct container *cont, struct cftype *cft,
   struct file *file, char __user *userbuf,
@@ -418,6 +436,25 @@ static struct cftype rss_reclaimed = {
  .read = rss_read_reclaimed,
 };

+static struct cftype pagecache_usage = {
+ .name = "pagecache_usage",
+ .private = RES_USAGE,
+ .read = pagecache_read,
+};
+
+static struct cftype pagecache_limit = {
+ .name = "pagecache_limit",
+ .private = RES_LIMIT,
+ .read = pagecache_read,
+ .write = pagecache_write,
+};
+
+static struct cftype pagecache_failcnt = {
+ .name = "pagecache_failcnt",
+ .private = RES_FAILCNT,
+ .read = pagecache_read,
```

```
+};
+
 static int rss_populate(struct container_subsys *ss,
   struct container *cont)
 {
@@ -431,6 +468,12 @@ static int rss_populate(struct container
   return rc;
  if ((rc = container_add_file(cont, &rss_reclaimed)) < 0)
   return rc;
+ if ((rc = container_add_file(cont, &pagecache_usage)) < 0)
+  return rc;
+ if ((rc = container_add_file(cont, &pagecache_failcnt)) < 0)
+  return rc;
+ if ((rc = container_add_file(cont, &pagecache_limit)) < 0)
+  return rc;

  return 0;
 }
```

Subject: [RFC][PATCH 2/4] Pagecache accounting
Posted by Vaidyanathan Srinivas on Wed, 20 Jun 2007 11:37:11 GMT
View Forum Message <> Reply to Message

Pagecache Accounting

The rss accounting hooks have been generalised to handle both anon pages
and file backed pages and charge the respective resource counters.

New flags and ref count has been added to page_container structure.
The ref count is used to ensure a page is added or removed from page_container
list only once independent of repeated calls from pagecache, swapcache and mmap
to RSS.

Flags in page_container is used to keep the accounting correct between RSS and
unmapped pagecache (includes swapcache) pages.

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

```
---
 include/linux/rss_container.h |   25 ++++--
 mm/rss_container.c            |  171 +++++++++++++++++++++++++++++------------
 2 files changed, 140 insertions(+), 56 deletions(-)

--- linux-2.6.22-rc2-mm1.orig/include/linux/rss_container.h
+++ linux-2.6.22-rc2-mm1/include/linux/rss_container.h
@@ -12,6 +12,14 @@
 struct page_container;
```

```
 struct rss_container;

+/* Flags for container_page_xxx calls */
+#define PAGE_TYPE_RSS  0x00000001
+#define PAGE_TYPE_PAGECACHE 0x00000002
+#define PAGE_TYPE_SWAPCACHE 0x00000004
+
+#define PAGE_SUBTYPE_ANON 0x00010000
+#define PAGE_SUBTYPE_FILE 0x00020000
+
 #ifdef CONFIG_RSS_CONTAINER
 /*
  * This is how the RSS accounting works.
@@ -68,11 +76,12 @@ struct rss_container;
  *
  */

-int container_rss_prepare(struct page *, struct vm_area_struct *vma,
-  struct page_container **);
-void container_rss_add(struct page_container *);
-void container_rss_del(struct page_container *);
-void container_rss_release(struct page_container *);
+int container_page_prepare(struct page *page, struct mm_struct *mm,
+  struct page_container **ppc, unsigned int flags);
+
+void container_page_add(struct page_container *, unsigned int flags);
+void container_page_del(struct page_container *, unsigned int flags);
+void container_page_release(struct page_container *, unsigned int flags);

 void container_out_of_memory(struct rss_container *);

@@ -92,15 +101,15 @@ static inline int container_rss_prepare(
  return 0;
 }

-static inline void container_rss_add(struct page_container *pc)
+static inline void container_page_add(struct page_container *pc, unsigned int flags)
 {
 }

-static inline void container_rss_del(struct page_container *pc)
+static inline void container_page_del(struct page_container *pc, unsigned int flags)
 {
 }

-static inline void container_rss_release(struct page_container *pc)
+static inline void container_page_release(struct page_container *pc, unsigned int flags)
 {
```

```
 }

--- linux-2.6.22-rc2-mm1.orig/mm/rss_container.c
+++ linux-2.6.22-rc2-mm1/mm/rss_container.c
@@ -58,6 +58,8 @@ struct rss_container {
  */

 struct page_container {
+  unsigned long flags;
+  unsigned long ref_cnt;
   struct page *page;
   struct rss_container *cnt;
   struct list_head list; /* this is the element of (int)active_list of
@@ -65,6 +67,11 @@ struct page_container {
     */
 };

+/* Flags for Page Container */
+#define PAGE_IN_PAGECACHE  0x0001
+#define PAGE_IN_RSS   0x0002
+#define PAGE_IN_SWAPCACHE  0x0004
+
 static inline struct rss_container *rss_from_cont(struct container *cnt)
 {
   return container_of(container_subsys_state(cnt, rss_subsys_id),
@@ -95,27 +102,39 @@ void mm_free_container(struct mm_struct
   * I bet you have already read the comment in include/linux/rss_container.h :)
   */

-int container_rss_prepare(struct page *page, struct vm_area_struct *vma,
-  struct page_container **ppc)
+int container_page_prepare(struct page *page, struct mm_struct *mm,
+   struct page_container **ppc, unsigned int flags)
 {
- struct rss_container *rss;
- struct page_container *pc;
-
- rcu_read_lock();
- rss = rcu_dereference(vma->vm_mm->rss_container);
- css_get(&rss->css);
- rcu_read_unlock();
-
- pc = kmalloc(sizeof(struct page_container), GFP_KERNEL);
- if (pc == NULL)
-  goto out_nomem;
-
- while (res_counter_charge(&rss->res, 1)) {
-  if (try_to_free_pages_in_container(rss)) {
```

```
-   atomic_inc(&rss->rss_reclaimed);
-   continue;
-  }
+   struct rss_container *rss;
+   struct page_container *pc;
+  int rc = 1;
+
+   /* Page may have been added to container earlier */
+   pc = page_container(page);
+ /* Check if this is fist time addition or not */
+ if (!pc) {
+  rcu_read_lock();
+  rss = rcu_dereference(mm->rss_container);
+  css_get(&rss->css);
+  rcu_read_unlock();
+ } else {
+  rss = pc->cnt;
+ }

+  /* Charge the respective resource count */
+  while (rc) {
+   if (flags & PAGE_TYPE_RSS)
+    rc = res_counter_charge(&rss->res, 1);
+   else
+    rc = res_counter_charge(&rss->pagecache_res, 1);
+
+   if (!rc)
+    break; /* All well */
+
+    if (try_to_free_pages_in_container(rss)) {
+     atomic_inc(&rss->rss_reclaimed);
+    continue; /* Try agin to charge container */
+    }
   /*
     * try_to_free_pages() might not give us a full picture
     * of reclaim. Some pages are reclaimed and might be moved
@@ -126,61 +145,117 @@ int container_rss_prepare(struct page *p
     */
   if (res_counter_check_under_limit(&rss->res))
    continue;
+  if (res_counter_check_under_limit(&rss->pagecache_res))
+   continue;

-  container_out_of_memory(rss);
-  if (test_thread_flag(TIF_MEMDIE))
-   goto out_charge;
- }
-
```

```
- pc->page = page;
- pc->cnt = rss;
- *ppc = pc;
- return 0;
-
-out_charge:
- kfree(pc);
-out_nomem:
- css_put(&rss->css);
- return -ENOMEM;
+    /* Unable to free memory?? */
+     container_out_of_memory(rss);
+     if (test_thread_flag(TIF_MEMDIE))
+      goto out_charge;
+    }
+
+    /* First time addition to container: Create new page_container */
+    if (!pc) {
+     pc = kzalloc(sizeof(struct page_container), GFP_KERNEL);
+     if (pc == NULL)
+      goto out_nomem;
+
+    pc->page = page;
+    pc->cnt = rss;
+    }
+
+    *ppc = pc;
+    return 0;
+
+  out_charge:
+  /* Need to zero page_container?? */
+  out_nomem:
+    css_put(&rss->css);
+    return -ENOMEM;
 }

-void container_rss_release(struct page_container *pc)
+void container_page_release(struct page_container *pc, unsigned int flags)
 {
  struct rss_container *rss;

  rss = pc->cnt;
- res_counter_uncharge(&rss->res, 1);
+ /* Setting the accounts right */
+ if (flags & PAGE_TYPE_RSS) {
+  res_counter_uncharge(&rss->res, 1);
+ } else {
+  res_counter_uncharge(&rss->pagecache_res, 1);
```

```
+ }
+
+ if (!(pc->flags)) {
+  BUG_ON(pc->ref_cnt);
+  set_page_container(pc->page, NULL);
+  kfree(pc);
+ }
  css_put(&rss->css);
- kfree(pc);
 }

-void container_rss_add(struct page_container *pc)
+void container_page_add(struct page_container *pc, unsigned int flags)
 {
  struct page *pg;
  struct rss_container *rss;
+ unsigned long irqflags;

  pg = pc->page;
- rss = pc->cnt;

- spin_lock_irq(&rss->res.lock);
- list_add(&pc->list, &rss->active_list);
- spin_unlock_irq(&rss->res.lock);
+ if (pg == ZERO_PAGE(0))
+  return;
+
+ /* Update flage in page container */
+ if (flags & PAGE_TYPE_RSS) {
+  pc->flags |= PAGE_IN_RSS;
+ } else {
+  pc->flags |= PAGE_IN_PAGECACHE;
+ }
+
+ rss = pc->cnt;
+ spin_lock_irqsave(&rss->res.lock, irqflags);
+ if (!pc->ref_cnt)
+  list_add(&pc->list, &rss->inactive_list);
+ pc->ref_cnt++;
+ spin_unlock_irqrestore(&rss->res.lock, irqflags);

  set_page_container(pg, pc);
 }

-void container_rss_del(struct page_container *pc)
+void container_page_del(struct page_container *pc, unsigned int flags)
 {
+ struct page *page;
```

```
  struct rss_container *rss;
+ unsigned long irqflags;

+ page = pc->page;
  rss = pc->cnt;
- spin_lock_irq(&rss->res.lock);
- list_del(&pc->list);
- res_counter_uncharge_locked(&rss->res, 1);
- spin_unlock_irq(&rss->res.lock);

+ if (page == ZERO_PAGE(0))
+  return;
+ BUG_ON(pc->flags & ~3);
+ /* Setting the accounts right */
+ if (flags & PAGE_TYPE_RSS) {
+  res_counter_uncharge(&rss->res, 1);
+  pc->flags &= ~PAGE_IN_RSS;
+  /* If it is a pagecache page the move acct to pagecache */
+ } else {
+  res_counter_uncharge(&rss->pagecache_res, 1);
+  pc->flags &= ~PAGE_IN_PAGECACHE;
+ }
+ spin_lock_irqsave(&rss->res.lock, irqflags);
+ pc->ref_cnt--;
+ if (!pc->ref_cnt) {
+  list_del(&pc->list);
+  set_page_container(page, NULL);
+ }
+ spin_unlock_irqrestore(&rss->res.lock, irqflags);
+
+ if (!pc->ref_cnt) {
+  BUG_ON(pc->flags);
+  kfree(pc);
+ }
  css_put(&rss->css);
- kfree(pc);
 }

 /*
```

Subject: [RFC][PATCH 3/4] Pagecache and RSS accounting hooks
Posted by Vaidyanathan Srinivas on Wed, 20 Jun 2007 11:39:13 GMT
View Forum Message <> Reply to Message

Pagecache and RSS accounting Hooks

New calls have been added from swap_state.c and filemap.c to track pagecache
and swapcache pages.

All existing RSS hooks have been generalised for pagecache accounting as well.

Most of these are function prototype changes.

Signed-off-by: Vaidyanathan Srinivasan <svaidy@linux.vnet.ibm.com>
---
 fs/exec.c | 4 ++--
 mm/filemap.c | 18 ++++++++++++++++++
 mm/memory.c | 18 +++++++++---------
 mm/migrate.c | 4 ++--
 mm/rmap.c | 12 ++++++------
 mm/swap_state.c | 16 ++++++++++++++++
 mm/swapfile.c | 4 ++--
 7 files changed, 55 insertions(+), 21 deletions(-)

```
--- linux-2.6.22-rc2-mm1.orig/fs/exec.c
+++ linux-2.6.22-rc2-mm1/fs/exec.c
@@ -321,7 +321,7 @@ void install_arg_page(struct vm_area_str
  if (unlikely(anon_vma_prepare(vma)))
   goto out;

- if (container_rss_prepare(page, vma, &pcont))
+ if (container_page_prepare(page, vma->vm_mm, &pcont, PAGE_TYPE_RSS))
   goto out;

  flush_dcache_page(page);
@@ -343,7 +343,7 @@ void install_arg_page(struct vm_area_str
  return;

 out_release:
- container_rss_release(pcont);
+ container_page_release(pcont, PAGE_TYPE_RSS);
 out:
  __free_page(page);
  force_sig(SIGKILL, current);
--- linux-2.6.22-rc2-mm1.orig/mm/filemap.c
+++ linux-2.6.22-rc2-mm1/mm/filemap.c
@@ -30,6 +30,7 @@
 #include <linux/security.h>
 #include <linux/syscalls.h>
 #include <linux/cpuset.h>
+#include <linux/rss_container.h>
 #include "filemap.h"
 #include "internal.h"
```

```
@@ -117,6 +118,9 @@ void __remove_from_page_cache(struct pag
  struct address_space *mapping = page->mapping;

  radix_tree_delete(&mapping->page_tree, page->index);
+ /* Uncharge before the mapping is gone */
+ if (page_container(page))
+  container_page_del(page_container(page), PAGE_TYPE_PAGECACHE);
  page->mapping = NULL;
  mapping->nrpages--;
  __dec_zone_page_state(page, NR_FILE_PAGES);
@@ -440,6 +444,8 @@ int add_to_page_cache(struct page *page,
   pgoff_t offset, gfp_t gfp_mask)
 {
  int error = radix_tree_preload(gfp_mask & ~__GFP_HIGHMEM);
+ struct page_container *pc;
+ struct mm_struct *mm;

  if (error == 0) {
   write_lock_irq(&mapping->tree_lock);
@@ -453,6 +459,18 @@ int add_to_page_cache(struct page *page,
    __inc_zone_page_state(page, NR_FILE_PAGES);
   }
   write_unlock_irq(&mapping->tree_lock);
+  /* Unlock before charge, because we may reclaim this inline */
+  if(!error) {
+   if (current->mm)
+    mm = current->mm;
+   else
+    mm = &init_mm;
+   if (!container_page_prepare(page, mm, &pc, PAGE_TYPE_PAGECACHE))
+    container_page_add(pc, PAGE_TYPE_PAGECACHE);
+   else
+    BUG();
+  }
+
   radix_tree_preload_end();
  }
  return error;
--- linux-2.6.22-rc2-mm1.orig/mm/memory.c
+++ linux-2.6.22-rc2-mm1/mm/memory.c
@@ -1755,7 +1755,7 @@ gotten:
   cow_user_page(new_page, old_page, address, vma);
  }

- if (container_rss_prepare(new_page, vma, &pcont))
+ if (container_page_prepare(new_page, vma->vm_mm, &pcont, PAGE_TYPE_RSS))
   goto oom;
```

```
  /*
@@ -1791,7 +1791,7 @@ gotten:
    new_page = old_page;
    ret |= VM_FAULT_WRITE;
   } else
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS);

   if (new_page)
    page_cache_release(new_page);
@@ -2217,7 +2217,7 @@ static int do_swap_page(struct mm_struct
    count_vm_event(PGMAJFAULT);
   }

- if (container_rss_prepare(page, vma, &pcont)) {
+ if (container_page_prepare(page, vma->vm_mm, &pcont, PAGE_TYPE_RSS)) {
    ret = VM_FAULT_OOM;
    goto out;
   }
@@ -2235,7 +2235,7 @@ static int do_swap_page(struct mm_struct

   if (unlikely(!PageUptodate(page))) {
    ret = VM_FAULT_SIGBUS;
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS);
    goto out_nomap;
   }

@@ -2271,7 +2271,7 @@ unlock:
 out:
  return ret;
 out_nomap:
- container_rss_release(pcont);
+ container_page_release(pcont, PAGE_TYPE_RSS);
  pte_unmap_unlock(page_table, ptl);
  unlock_page(page);
  page_cache_release(page);
@@ -2302,7 +2302,7 @@ static int do_anonymous_page(struct mm_s
    if (!page)
     goto oom;

-  if (container_rss_prepare(page, vma, &pcont))
+  if (container_page_prepare(page, vma->vm_mm, &pcont, PAGE_TYPE_RSS))
     goto oom_release;

    entry = mk_pte(page, vma->vm_page_prot);
@@ -2338,7 +2338,7 @@ unlock:
  pte_unmap_unlock(page_table, ptl);
```

```
  return VM_FAULT_MINOR;
 release_container:
- container_rss_release(pcont);
+ container_page_release(pcont, PAGE_TYPE_RSS);
 release:
  page_cache_release(page);
  goto unlock;
@@ -2442,7 +2442,7 @@ static int __do_fault(struct mm_struct *

  }

- if (container_rss_prepare(page, vma, &pcont)) {
+ if (container_page_prepare(page, vma->vm_mm, &pcont, PAGE_TYPE_RSS)) {
   fdata.type = VM_FAULT_OOM;
   goto out;
  }
@@ -2483,7 +2483,7 @@ static int __do_fault(struct mm_struct *
   update_mmu_cache(vma, address, entry);
   lazy_mmu_prot_update(entry);
  } else {
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS);
   if (anon)
    page_cache_release(page);
   else
--- linux-2.6.22-rc2-mm1.orig/mm/migrate.c
+++ linux-2.6.22-rc2-mm1/mm/migrate.c
@@ -159,7 +159,7 @@ static void remove_migration_pte(struct
    return;
   }

- if (container_rss_prepare(new, vma, &pcont)) {
+ if (container_page_prepare(new, vma->vm_mm, &pcont, PAGE_TYPE_RSS)) {
   pte_unmap(ptep);
   return;
  }
@@ -194,7 +194,7 @@ static void remove_migration_pte(struct

 out:
  pte_unmap_unlock(ptep, ptl);
- container_rss_release(pcont);
+ container_page_release(pcont, PAGE_TYPE_RSS);
 }

 /*
--- linux-2.6.22-rc2-mm1.orig/mm/rmap.c
+++ linux-2.6.22-rc2-mm1/mm/rmap.c
@@ -578,14 +578,14 @@ void page_add_anon_rmap(struct page *pag
```

```
  if (atomic_inc_and_test(&page->_mapcount)) {
   __page_set_anon_rmap(page, vma, address);
   /* 0 -> 1 state change */
-  container_rss_add(pcont);
+  container_page_add(pcont, PAGE_TYPE_RSS|PAGE_SUBTYPE_ANON);
  } else {
   __page_check_anon_rmap(page, vma, address);
   /*
    * we raced with another touch or just mapped the page
    * for the N-th time
    */
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS|PAGE_SUBTYPE_ANON);
  }
 }

@@ -606,7 +606,7 @@ void page_add_new_anon_rmap(struct page
 {
  BUG_ON(address < vma->vm_start || address >= vma->vm_end);
  atomic_set(&page->_mapcount, 0); /* elevate count by 1 (starts at -1) */
- container_rss_add(pcont);
+ container_page_add(pcont, PAGE_TYPE_RSS|PAGE_SUBTYPE_ANON);
  __page_set_anon_rmap(page, vma, address);
 }

@@ -628,10 +628,10 @@ void page_add_file_rmap(struct page *pag
     * are not added to the container as they do not imply
     * RSS consumption  --xemul
     */
-   container_rss_add(pcont);
+   container_page_add(pcont, PAGE_TYPE_RSS|PAGE_SUBTYPE_FILE);
   __inc_zone_page_state(page, NR_FILE_MAPPED);
  } else if (pcont)
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS|PAGE_SUBTYPE_FILE);
 }

 #ifdef CONFIG_DEBUG_VM
@@ -689,7 +689,7 @@ void page_remove_rmap(struct page *page,
   }

   if (pcont)
-   container_rss_del(pcont);
+   container_page_del(pcont, PAGE_TYPE_RSS);
   /*
    * It would be tidy to reset the PageAnon mapping here,
    * but that might overwrite a racing page_add_anon_rmap
--- linux-2.6.22-rc2-mm1.orig/mm/swap_state.c
```

```
+++ linux-2.6.22-rc2-mm1/mm/swap_state.c
@@ -17,6 +17,7 @@
 #include <linux/backing-dev.h>
 #include <linux/pagevec.h>
 #include <linux/migrate.h>
+#include <linux/rss_container.h>

 #include <asm/pgtable.h>

@@ -74,6 +75,8 @@ static int __add_to_swap_cache(struct pa
         gfp_t gfp_mask)
 {
  int error;
+ struct page_container *pc;
+ struct mm_struct *mm;

  BUG_ON(PageSwapCache(page));
  BUG_ON(PagePrivate(page));
@@ -92,6 +95,17 @@ static int __add_to_swap_cache(struct pa
   }
   write_unlock_irq(&swapper_space.tree_lock);
   radix_tree_preload_end();
+  /* Unlock before charge, because we may reclaim this inline */
+  if(!error) {
+   if (current->mm)
+    mm = current->mm;
+   else
+    mm = &init_mm;
+   if (!container_page_prepare(page, mm, &pc, PAGE_TYPE_PAGECACHE))
+    container_page_add(pc, PAGE_TYPE_PAGECACHE);
+   else
+    BUG();
+  }
  }
  return error;
 }
@@ -130,6 +144,8 @@ void __delete_from_swap_cache(struct pag
  BUG_ON(PagePrivate(page));

  radix_tree_delete(&swapper_space.page_tree, page_private(page));
+ if (page_container(page))
+  container_page_del(page_container(page), PAGE_TYPE_PAGECACHE);
  set_page_private(page, 0);
  ClearPageSwapCache(page);
  total_swapcache_pages--;
--- linux-2.6.22-rc2-mm1.orig/mm/swapfile.c
+++ linux-2.6.22-rc2-mm1/mm/swapfile.c
@@ -535,7 +535,7 @@ static int unuse_pte_range(struct vm_are
```

```
  int found = 0;
  struct page_container *pcont;

- if (container_rss_prepare(page, vma, &pcont))
+ if (container_page_prepare(page, vma->vm_mm, &pcont, PAGE_TYPE_RSS))
   return 0;

  pte = pte_offset_map_lock(vma->vm_mm, pmd, addr, &ptl);
@@ -552,7 +552,7 @@ static int unuse_pte_range(struct vm_are
  } while (pte++, addr += PAGE_SIZE, addr != end);
  pte_unmap_unlock(pte - 1, ptl);
  if (!found)
-  container_rss_release(pcont);
+  container_page_release(pcont, PAGE_TYPE_RSS);
  return found;
 }
```

Subject: [RFC][PATCH 4/4] Pagecache reclaim
Posted by Vaidyanathan Srinivas on Wed, 20 Jun 2007 11:40:28 GMT
View Forum Message <> Reply to Message

Pagecache controller reclaim changes

Reclaim path needs performance improvement.
For now it is minor changes to include unmapped
pages in our list of page_container.

Signed-off-by: Vaidyanathan Srinivasan <svaidy@linux.vnet.ibm.com>
---
 mm/rss_container.c | 3 ---
 1 file changed, 3 deletions(-)

```
--- linux-2.6.22-rc2-mm1.orig/mm/rss_container.c
+++ linux-2.6.22-rc2-mm1/mm/rss_container.c
@@ -274,9 +274,6 @@ void container_rss_move_lists(struct pag
  struct rss_container *rss;
  struct page_container *pc;

- if (!page_mapped(pg))
-  return;
-
  pc = page_container(pg);
  if (pc == NULL)
   return;
```
