Subject: [RFC][PATCH 0/3] Containers: Integrated RSS and pagecache control v5
Posted by Vaidyanathan Srinivas on Fri, 29 Jun 2007 06:17:23 GMT
View Forum Message <> Reply to Message

Containers: Integrated RSS and pagecache accounting and control v5
--- ---

Based on the discussions at OLS yesterday, the consensus was to try an integrated pagecache controller along with RSS controller under the same usage limit.

This patch extends the RSS controller to account and reclaim pagecache and swapcache pages. The same 'rss_limit' now applies to both RSS pages and pagecache pages. When the limit is reached, both pagecache and RSS pages are reclaimed in LRU order as per the normal system wide reclaim policy.

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 with kernbench.

Comments, suggestions and criticisms are welcome.

--Vaidy

Features:
---
* Single limit for both RSS and pagecache/swapcache pages
* 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 under rss_usage.
* The usage limit set in rss_limit applies to sum of both RSS and pagecache pages
* Limits on pagecache can be set by echo -n 100000 > rss_limit on the /container file system. The unit is in pages or 4 kilobytes
* If the pagecache+RSS utilisation exceed the limit, 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
* Per-container reclaim knobs like dirty ratio, vm_swappiness may
  provide better control

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
* In order to set limit, echo value in pages (4KB) to rss_limit
 echo -n 100000 > rss_limit
 #This would set 409MB limit on pagecache+rss usage
* Trash the system from current shell using scp/cp/dd/tar etc
* Watch rss_usage and /proc/meminfo to verify behavior

Tests:
---
* Simple dd/cat/cp test on pagecache limit/reclaim
* rss_limit was tested with simple test application that would malloc
  predefined size of memory and touch them to allocate pages.
* kernbench was run under container with 400MB memory limit

ToDo:
---
* Optimise the reclaim.
* Per-container VM stats and knobs

Patch Series:
---
pagecache-controller-v5-acct.patch
pagecache-controller-v5-acct-hooks.patch
pagecache-controller-v5-reclaim.patch

ChangeLog:
---

v5: Integrated pagecache + rss controller

* No separate pagecache_limit
* pagecache and rss pages accounted in rss_usage and governed by rss_limit
* Each page counted only once in rss_usage. Mapped or unmapped
  pagecache pages are counted alike in rss_usage

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

v3:
* Patch merged with Containers v8 and RSS v2