Subject: Re: [PATCH 20/33] memory controller add documentation
Posted by Randy Dunlap on Tue, 18 Sep 2007 16:53:13 GMT
View Forum Message <> Reply to Message

On Mon, 17 Sep 2007 14:03:27 -0700 Paul Menage wrote:

> From: Balbir Singh <balbir@linux.vnet.ibm.com>
> (container->cgroup renaming by Paul Menage <menage@google.com>)
>
> Signed-off-by: Balbir Singh <balbir@linux.vnet.ibm.com>
> Signed-off-by: Paul Menage <menage@google.com>
>
> ---
>
> Documentation/controllers/memory.txt | 259 ++++++++++++++++++++++++++++++++++
> 1 files changed, 259 insertions(+)
>
> diff -puN /dev/null Documentation/controllers/memory.txt
> --- /dev/null
> +++ a/Documentation/controllers/memory.txt
> @@ -0,0 +1,259 @@

> +Benefits and Purpose of the memory controller
> +
> +The memory controller isolates the memory behaviour of a group of tasks
> +from the rest of the system. The article on LWN [12] mentions some probable
> +uses of the memory controller. The memory controller can be used to
> +
> +a. Isolate an application or a group of applications
> +   Memory hungry applications can be isolated and limited to a smaller
> +   amount of memory.
> +b. Create a cgroup with limited amount of memory, this can be used

"," makes run-on sentence. Please use ";" or 2 sentences.

> +   as a good alternative to booting with mem=XXXX.
> +c. Virtualization solutions can control the amount of memory they want
> +   to assign to a virtual machine instance.
> +d. A CD/DVD burner could control the amount of memory used by the
> +   rest of the system to ensure that burning does not fail due to lack
> +   of available memory.
> +e. There are several other use cases, find one or use the controller just

Ditto.

> +   for fun (to learn and hack on the VM subsystem).

[snip]

> +
> +2.1. Design
> +
> +The core of the design is a counter called the res_counter. The res_counter
> +tracks the current memory usage and limit of the group of processes associated
> +with the controller. Each cgroup has a memory controller specific data
> +structure (mem_cgroup) associated with it.
> +
> +2.2. Accounting
> +
[snip]
> +
> +The accounting is done as follows: mem_cgroup_charge() is invoked to setup
> +the necessary data structures and check if the cgroup that is being charged
> +is over its limit. If it is then reclaim is invoked on the cgroup.
> +More details can be found in the reclaim section of this document.
> +If everything goes well, a page meta-data-structure called page_cgroup is

^drop second hyphen (use space)

> +allocated and associated with the page. This routine also adds the page to
> +the per cgroup LRU.
> +
> +2.2.1 Accounting details
> +
> +All mapped pages (RSS) and unmapped user pages (Page Cache) are accounted.
> +RSS pages are accounted at the time of page_add_*_rmap() unless they've already
> +been accounted for earlier. A file page will be accounted for as Page Cache;
> +it's mapped into the page tables of a process, duplicate accounting is carefully

"," makes run-on sentence...

> +avoided. Page Cache pages are accounted at the time of add_to_page_cache().
> +The corresponding routines that remove a page from the page tables or removes

s/removes/remove/

> +a page from Page Cache is used to decrement the accounting counters of the
> +cgroup.
> +
> +2.3 Shared Page Accounting
> +
> +Shared pages are accounted on the basis of the first touch approach. The
> +cgroup that first touches a page is accounted for the page. The principle
> +behind this approach is that a cgroup that aggressively uses a shared
> +page will eventually get charged for it (once it is uncharged from
> +the cgroup that brought it in -- this will happen on memory pressure).

> +
> +2.4 Reclaim
> +
> +Each cgroup maintains a per cgroup LRU that consists of an active
> +and inactive list. When a cgroup goes over its limit, we first try
> +to reclaim memory from the cgroup so as to make space for the new
> +pages that the cgroup has touched. If the reclaim is unsuccessful,
> +an OOM routine is invoked to select and kill the bulkiest task in the
> +cgroup.
> +
> +The reclaim algorithm has not been modified for cgroups, except that
> +pages that are selected for reclaiming come from the per cgroup LRU
> +list.
> +
> +2. Locking

2. ?? Section numbering is "off."

> +The memory controller uses the following hierarchy
> +
> +1. zone->lru_lock is used for selecting pages to be isolated
> +2. mem->lru_lock protects the per cgroup LRU
> +3. lock_page_cgroup() is used to protect page->page_cgroup
> +
> +3. User Interface
> +
> +0. Configuration

  0. Kernel build configuration
> +
> +a. Enable CONFIG_CGROUPS
> +b. Enable CONFIG_RESOURCE_COUNTERS
> +c. Enable CONFIG_CGROUP_MEM_CONT
[snip]

~Randy

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