Subject: Re: [RFC][-mm] [2/2] Simple stats for memory resource controller
Posted by Balbir Singh on Mon, 28 Apr 2008 18:03:05 GMT
View Forum Message <> Reply to Message

```
Andrew Morton wrote:
> On Mon, 28 Apr 2008 21:30:29 +0530 Balaji Rao <balajirrao@gmail.com> wrote:
>
>> On Monday 14 April 2008 08:09:48 pm Balbir Singh wrote:
>>> Balaji Rao wrote:
>>>> This patch implements trivial statistics for the memory resource controller.
>>>>
>>>> Signed-off-by: Balaji Rao <balajirrao@gmail.com>
>>>> CC: Balbir Singh <balbir@linux.vnet.ibm.com>
>>>> CC: Dhaval Giani <dhaval@linux.vnet.ibm.com>
>>>>
>>>> diff --git a/mm/memcontrol.c b/mm/memcontrol.c
>>>> index a860765..ca98b21 100644
>>>> --- a/mm/memcontrol.c
>>>> +++ b/mm/memcontrol.c
>>>> @@ -47,6 +47,8 @@ enum mem_cgroup_stat_index {
>>>>    */
>>>>   MEM_CGROUP_STAT_CACHE,     /* # of pages charged as cache */
>>>>   MEM_CGROUP_STAT_RSS,    /* # of pages charged as rss */
>>>> + MEM_CGROUP_STAT_PGPGIN_COUNT, /* # of pages paged in */
>>>> + MEM_CGROUP_STAT_PGPGOUT_COUNT, /* # of pages paged out */
>>>>
>>>>   MEM_CGROUP_STAT_NSTATS,
>>>>  };
>>>> @@ -198,6 +200,13 @@ static void mem_cgroup_charge_statistics(struct mem_cgroup
*mem, int flags,
>>>>    __mem_cgroup_stat_add_safe(stat, MEM_CGROUP_STAT_CACHE, val);
>>>>   else
>>>>    __mem_cgroup_stat_add_safe(stat, MEM_CGROUP_STAT_RSS, val);
>>>> +
>>>> + if (charge)
>>>> +  __mem_cgroup_stat_add_safe(stat,
>>>> +    MEM_CGROUP_STAT_PGPGIN_COUNT, 1);
>>>> + else
>>>> +  __mem_cgroup_stat_add_safe(stat,
>>>> +    MEM_CGROUP_STAT_PGPGOUT_COUNT, 1);
>>>>  }
>>>>
>>>>  static struct mem_cgroup_per_zone *
>>>> @@ -897,6 +906,8 @@ static const struct mem_cgroup_stat_desc {
>>>>  } mem_cgroup_stat_desc[] = {
>>>>   [MEM_CGROUP_STAT_CACHE] = { "cache", PAGE_SIZE, },
>>>>   [MEM_CGROUP_STAT_RSS] = { "rss", PAGE_SIZE, },
>>>> + [MEM_CGROUP_STAT_PGPGIN_COUNT] = {"pgpgin", 1, },
```

>>>> + [MEM_CGROUP_STAT_PGPGOUT_COUNT] = {"pgpgout", 1, },
>>>> };
>>>>
>>>> static int mem_control_stat_show(struct cgroup *cont, struct cftype *cft,
>>>>
>>> Acked-by: Balbir Singh <balbir@linux.vnet.ibm.com>
>>>
>>> Hi, Andrew,
>>>
>>> Could you please include these statistics in -mm.
>>>
>>> Balbir
>>>
>>>
>> Hi Andrew,
>>
>> Now that Balbir Singh has ACKed it, could you please include it in -mm ?
>
> <looks>
>
> I guess we can add this one, sure. But [patch 1/2] needs work.
>
> - The local_irq_save()-around-for_each_possible_cpu() locking doesn't
>   make sense.
>

Yes, that needs re-work. Peter Zijlstra had detailed review comments for the patch

> - indenting is busted in account_user_time() and account_system_time()
>
> - The use of for_each_possible_cpu() can be grossly inefficient. It
>   would be preferred to use for_each_possible_cpu() and add a cpu-hotplug
>   notifier.
>
> - The proposed newly-added userspace interfaces are undocumented
>

Yes, we need more documentation

> - The changelogs don't explain why we might want this feature in Linux.
>

We need more accurate utime/stime per cgroup. Summing them in user space is insufficient, since tasks can move across groups and what we have is accumulated time per task.

> - Generally: there are a heck of a lot of different ways of accounting
>   for things in core kernel and it's really sad to see yet another one

> being added.
>

We thought of summing up stuff in user space, we've look harder. The plan is to finally send all the data using cgroupstats.

>
> Actually, [patch 2/2] adds new kerenl->user interfaces and doesn't document
> them. But afaict the existing memcgroup stats are secret too.
>

The statistics was added as a part of git commit d52aa412d43827033a8e2ce4415ef6e8f8d53635. I'll go ahead and try to document them. These patches piggy back on the statistics patches and add pagein/pageout counts, which is a useful statistic for the memory controller.

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