
Subject: Re: [PATCH 04/10] memcg: Introduce __GFP_NOACCOUNT.
Posted by [Glauber Costa](#) on Thu, 01 Mar 2012 00:24:11 GMT
[View Forum Message](#) <> [Reply to Message](#)

On 02/29/2012 09:10 PM, KAMEZAWA Hiroyuki wrote:
> On Wed, 29 Feb 2012 11:09:50 -0800
> Suleiman Souhlal<suleiman@google.com> wrote:
>
>> On Tue, Feb 28, 2012 at 10:00 PM, KAMEZAWA Hiroyuki
>> <kamezawa.hiroyu@jp.fujitsu.com> wrote:
>>> On Mon, 27 Feb 2012 14:58:47 -0800
>>> Suleiman Souhlal<ssouhlal@FreeBSD.org> wrote:
>>>
>>>> This is used to indicate that we don't want an allocation to be accounted
>>>> to the current cgroup.
>>>>
>>>> Signed-off-by: Suleiman Souhlal<suleiman@google.com>
>>>
>>> I don't like this.
>>>
>>> Please add
>>>
>>> __GFP_ACCOUNT "account this allocation to memcg"
>>>
>>> Or make this as slab's flag if this work is for slab allocation.
>>
>> We would like to account for all the slab allocations that happen in
>> process context.
>>
>> Manually marking every single allocation or kmem_cache with a GFP flag
>> really doesn't seem like the right thing to do..
>>
>> Can you explain why you don't like this flag?
>>
>
> For example, tcp buffer limiting has another logic for buffer size controlling.
> _AND_, most of kernel pages are not reclaimable at all.
> I think you should start from reclaimable caches as dcache, icache etc.
>
> If you want to use this wider, you can discuss
>
> + #define GFP_KERNEL (.....| __GFP_ACCOUNT)
>
> in future. I'd like to see small start because memory allocation failure
> is always terrible and make the system unstable. Even if you notify
> "Ah, kernel memory allocation failed because of memory.limit? and
> many unreclaimable memory usage. Please tweak the limitation or kill tasks!!"
>

> The user can't do anything because he can't create any new task because of OOM.
>
> The system will be being unstable until an admin, who is not under any limit,
> tweaks something or reboot the system.
>
> Please do small start until you provide Eco-System to avoid a case that
> the admin cannot login and what he can do was only reboot.
>
Having the root cgroup to be always unlimited should already take care
of the most extreme cases, right?
