Subject: Re: [PATCH v3 12/28] slab: pass memcg parameter to kmem_cache_create

Posted by Glauber Costa on Tue, 29 May 2012 16:59:54 GMT

View Forum Message <> Reply to Message

On 05/29/2012 08:52 PM, Christoph Lameter wrote:

> Well kmem_cache_alloc cache is the performance critical hotpath.
>
> If you are already there and doing all of that then would it not be better
> to simply count the objects allocated and freed per cgroup? Directly
> increment and decrement counters in a cgroup? You do not really need to
> duplicate the kmem_cache structure and do not need to modify allocators if
> you are willing to take that kind of a performance hit. Put a wrapper
> around kmem_cache_alloc/free and count things.

Well, I see it as the difference between being a big slower, and a lot slower.

Accounting in memcg is hard, specially because it is potentially hierarchical, (meaning you need to nest downwards until your parents).

I never discussed that this is, unfortunately, a hotpath. However, I did try to minimize the impact as much as I could.

Not to mention that the current scheme is bound to improvement as cgroups improve. One of the things being discussed is to having all cgroups always in the same hierarchy. If that ever happens, we can have the information about the current cgroup stored in a very accessible way, so to make this even faster.

This felt like the best way I could do with the current infrastructure, (and again, I did make it free for people not limiting kmem), and is way, way cheaper than doing accounting here.