Subject: Re: [PATCH v3 04/13] kmem accounting basic infrastructure
Posted by Tejun Heo on Wed, 26 Sep 2012 22:10:46 GMT
View Forum Message <> Reply to Message

Hello, Glauber.

On Thu, Sep 27, 2012 at 01:24:40AM +0400, Glauber Costa wrote:
> "kmem_accounted" is not a switch. It is an internal representation only.
> The semantics, that we discussed exhaustively in San Diego, is that a
> group that is not limited is not accounted. This is simple and consistent.
>
> Since the limits are still per-cgroup, you are actually proposing more
> user-visible complexity than me, since you are adding yet another file,
> with its own semantics.

I was confused. I thought it was exposed as a switch to userland (it being right below .use_hierarchy tripped red alert). This is internal flag dependent upon kernel limit being set. My apologies.

So, the proposed behavior is to allow enabling kmemcg anytime but ignore what happened inbetween? Where the knob is changes but the weirdity seems all the same. What prevents us from having a single switch at root which can only be flipped when there's no children?

Backward compatibility is covered with single switch and I really don't think "you can enable limits for kernel memory anytime but we don't keep track of whatever happened before it was flipped the first time because the first time is always special" is a sane thing to expose to userland. Or am I misunderstanding the proposed behavior again?

Thanks.

--
tejun