

---

Subject: Re: [RFC][PATCH 4/7] RSS accounting hooks over the code  
Posted by [Nick Piggin](#) on Tue, 13 Mar 2007 10:25:49 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:

> Herbert Poetzl <herbert@13thfloor.at> writes:

>

>

>>On Mon, Mar 12, 2007 at 09:50:08AM -0700, Dave Hansen wrote:

>>

>>>On Mon, 2007-03-12 at 19:23 +0300, Kirill Korotaev wrote:

>>>

>>>>For these you essentially need per-container page->\_mapcount counter,

>>>>otherwise you can't detect whether rss group still has the page

>>>>in question being mapped in its processes' address spaces or not.

>>

>>>What do you mean by this? You can always tell whether a process has a

>>>particular page mapped. Could you explain the issue a bit more. I'm

>>>not sure I get it.

>>

>>OpenVZ wants to account \_shared\_ pages in a guest

>>different than separate pages, so that the RSS

>>accounted values reflect the actual used RAM instead

>>of the sum of all processes RSS' pages, which for

>>sure is more relevant to the administrator, but IMHO

>>not so terribly important to justify memory consuming

>>structures and sacrifice performance to get it right

>>

>>YMMV, but maybe we can find a smart solution to the

>>issue too :)

>

>

> I will tell you what I want.

>

> I want a shared page cache that has nothing to do with RSS limits.

>

> I want an RSS limit that once I know I can run a deterministic

> application with a fixed set of inputs in I want to know it will

> always run.

>

> First touch page ownership does not guarantee give me anything useful

> for knowing if I can run my application or not. Because of page

> sharing my application might run inside the rss limit only because

> I got lucky and happened to share a lot of pages with another running

> application. If the next I run and it isn't running my application

> will fail. That is ridiculous.

Let's be practical here, what you're asking is basically impossible.

Unless by deterministic you mean that it never enters the a non trivial syscall, in which case, you just want to know about maximum RSS of the process, which we already account).

> I don't want sharing between vservers/VE/containers to affect how many  
> pages I can have mapped into my processes at once.

You seem to want total isolation. You could use virtualization?

> Now sharing is sufficiently rare that I'm pretty certain that problems  
> come up rarely. So maybe these problems have not shown up in testing  
> yet. But until I see the proof that actually doing the accounting for  
> sharing properly has intolerable overhead. I want proper accounting  
> not this hand waving that is only accurate on the third Tuesday of the  
> month.

It is basically handwaving anyway. The only approach I've seen with a sane (not perfect, but good) way of accounting memory use is this one. If you care to define "proper", then we could discuss that.

--

SUSE Labs, Novell Inc.

Send instant messages to your online friends <http://au.messenger.yahoo.com>

---

Containers mailing list  
Containers@lists.osdl.org  
<https://lists.osdl.org/mailman/listinfo/containers>

---