Subject: Re: Linux-VServer example results for sharing vs. separate mappings ...
Posted by Balbir Singh on Mon, 26 Mar 2007 06:05:54 GMT
View Forum Message <> Reply to Message

Andrew Morton wrote:
> On Mon, 26 Mar 2007 08:06:07 +0530 Balbir Singh <balbir@in.ibm.com> wrote:
>
>> Andrew Morton wrote:
>>>> Don't we break the global LRU with this scheme?
>>> Sure, but that's deliberate!
>>>
>>> (And we don't have a global LRU - the LRUs are per-zone).
>>>
>> Yes, true. But if we use zones for containers and say we have 400
>> of them, with all of them under limit. When the system wants
>> to reclaim memory, we might not end up reclaiming the best pages.
>> Am I missing something?
>
> If a zone is under its min_pages limit, it needs reclaim. Who/when/why
> that reclaim is run doesn't really matter.
>
> Yeah, we might run into some scaling problems with that many zones.
> They're unlikely to be unfixable.
>

ok.

>
>>>>> b) Create a new memory abstraction, call it the "software zone", which
>>>>> is mostly decoupled from the present "hardware zones". Most of the MM
>>>>> is reworked to use "software zones". The "software zones" are
>>>>> runtime-resizeable, and obtain their pages via some means from the
>>>>> hardware zones. A container uses a software zone.
>>>>>
>>>> I think the problem would be figuring out where to allocate memory from?
>>>> What happens if a software zone spans across many hardware zones?
>>> Yes, that would be the tricky part. But we generally don't care what
>>> physical zone user pages come from, apart from NUMA optimisation.
>>>
>>>> The reclaim mechanism proposed *does not impact the non-container users*.
>>> Yup. Let's keep plugging away with Pavel's approach, see where it gets us.
>>>
>> Yes, we have some changes that we've made to the reclaim logic, we hope
>> to integrate a page cache controller soon. We are also testing the
>> patches. Hopefully soon enough, they'll be in a good state and we can
>> request you to merge the containers and the rss limit (plus page cache)
>> controller soon.
>

> Now I'm worried again. This separation between "rss controller" and
> "pagecache" is largely alien to memory reclaim. With physical containers
> these new concepts (and their implementations) don't need to exist - it is
> already all implemented.
>
> Designing brand-new memory reclaim machinery in mid-2007 sounds like a very
> bad idea. But let us see what it looks like.
>

I did not mean to worry you again :-) We do not plan to implement brand new memory reclaim, we intend to modify some bits and pieces for per container reclaim. We believe at this point that all the necessary infrastructure is largely present in container_isolate_pages(). Adding a page cache controller should not require core-mm surgery, just the accounting bits.

We basically agree that designing a brand new reclaim machinery is a bad idea, non-container users will not be impacted. Only container driver reclaim (caused by a container being at it's limit), will see some change in reclaim behaviour and we shall try and restrict the changes to as small as possible.

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

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers