Subject: Re: [PATCH] BC: resource beancounters (v2)
Posted by Balbir Singh on Tue, 29 Aug 2006 17:08:45 GMT
View Forum Message <> Reply to Message

Kirill Korotaev wrote:
>>> ------------- cut ----------------
>>> The task of limiting a container to 4.5GB of memory bottles down to the
>>> question: what to do when the container starts to use more than assigned
>>> 4.5GB of memory?
>>>
>>> At this moment there are only 3 viable alternatives.
>>>
>>> A) Have separate memory management for each container,
>>>   with separate buddy allocator, lru lists, page replacement mechanism.
>>>   That implies a considerable overhead, and the main challenge there
>>>   is sharing of pages between these separate memory managers.
>>>
>>> B) Return errors on extension of mappings, but not on page faults, where
>>>   memory is actually consumed.
>>>   In this case it makes sense to take into account not only the size
>>> of used
>>>   memory, but the size of created mappings as well.
>>>   This is approximately what "privvmpages" accounting/limiting
>>> provides in
>>>   UBC.
>>>
>>> C) Rely on OOM killer.
>>>   This is a fall-back method in UBC, for the case "privvmpages" limits
>>>   still leave the possibility to overload the system.
>>>
>>
>>
>> D) Virtual scan of mm's in the over-limit container
>>
>> E) Modify existing physical scanner to be able to skip pages which
>>   belong to not-over-limit containers.
>>
>> F) Something else ;)
> We fully agree that other possible algorithms can and should exist.
> My idea only is that any of them would need accounting anyway
> (which is the most part of beancounters).
> Throtling, modified scanners etc. can be implemented as a separate
> BC parameters. Thus, an administrator will be able to select
> which policy should be applied to the container which is near its limit.
>
> So the patches I'm trying to send are a step-by-step accounting of all
> the resources and their simple limitations. More comprehensive limitation
> policy will be built on top of it later.

>

One of the issues I see is that bean counters are not very flexible. Tasks
cannot change bean counters dynamically after fork()/exec() that is - can they?

> BTW, UBC page beancounters allow to distinguish pages used by only one
> container and pages which are shared. So scanner can try to reclaim
> container private pages first, thus not influencing other containers.
>

But can you select the specific container for which we intend to scan pages?

> Thanks,
> Kirill
>

--
Thanks,
Balbir Singh,
Linux Technology Center,
IBM Software Labs