Subject: Re: [RFC] mm-controller
Posted by Vaidyanathan Srinivas on Mon, 25 Jun 2007 18:02:37 GMT
View Forum Message <> Reply to Message

Peter Zijlstra wrote:
> On Fri, 2007-06-22 at 22:05 +0530, Vaidyanathan Srinivasan wrote:
>
>> Merging both limits will eliminate the issue, however we would need
>> individual limits for pagecache and RSS for better control. There are
>> use cases for pagecache_limit alone without RSS_limit like the case of
>> database application using direct IO, backup applications and
>> streaming applications that does not make good use of pagecache.
>
> I'm aware that some people want this. However we rejected adding a
> pagecache limit to the kernel proper on grounds that reclaim should do a
> better job.
>
> And now we're sneaking it in the backdoor.
>
> If we're going to do this, get it in the kernel proper first.
>

Good point. We should probably revisit this in the context of containers, virtualization and server consolidation. Kernel takes the best decision in the context of overall system performance, but when we want the kernel to favor certain group of application relative to others then we hit corner cases. Streaming multimedia applications are one of the corner case where the kernel's effort to manage pagecache does not help overall system performance.

There have been several patches suggested to provide system wide pagecache limit. There are some user mode fadvice() based techniques as well. However solving the problem in the context of containers provide certain advantages

* Containers provide task grouping
* Relative priority or importance can be assigned to each group using resource limits.
* Memory controller under container framework provide infrastructure for detailed accounting of memory usage
* Containers and controllers form generalised infrastructure to create localised VM behavior for a group of tasks

I would see introduction of pagecache limit in containers as a safe place to add the new feature rather than a backdoor. Since this feature has a relatively small user base, it be best left as a container plugin rather than a system wide tunable.

I am not suggesting against system wide pagecache control. We should definitely try to find solutions for pagecache control outside of containers as well.

--Vaidy