Subject: Re: [PATCH] Cpu statistics accounting based on Paul Menage patches
Posted by xemul on Thu, 12 Apr 2007 16:19:50 GMT
View Forum Message <> Reply to Message

Andrew Morton wrote:
> On Wed, 11 Apr 2007 19:02:27 +0400
> Pavel Emelianov <xemul@sw.ru> wrote:
>
>> Provides a per-container statistics concerning the numbers of tasks
>> in various states, system and user times, etc. Patch is inspired
>> by Paul's example of the used CPU time accounting. Although this
>> patch is independent from Paul's example to make it possible playing
>> with them separately.
>
> Why is this actually needed? If userspace has a list of the tasks which
> are in a particular container, it can run around and add up the stats for
> those tasks without kernel changes?

Well, the per-conainer nr_running and nr_uninterruptibe
accounting is the only way to calculate loadavg and idle
time for *container*.

Look, the idle time for container (for one CPU) is the time
when no tasks within this container were ready to run on
this CPU. That's the definition implicitly used in global
idle time accounting. Current per-rq counters can't solve
this problem and neither can the update_process_times()
method.

The same with loadavg. To calculate it per-container we need
to have some nr_active_in_container() function, but it must
work faster than walking the tasks within the container.

> It's a bit irksome that we have so much accounting of this form in core
> kernel, yet we have to go and add a completely new implementation to create
> something which is similar to what we already have. But I don't

Creating something similar would be a bit problematic.

Some stats are stored on task_struct some are pointed
from a signal_struct, some are reported via /proc files,
some via netlink sockets, some statistics can be per-task
some cannot.

On the other hand containers provide a generic way to group
the tasks and report the statistics for it. So we can keep
the stats in one place and report in a similar way.

> immediately see a fix for that. Apart from paragraph #1 ;)
>
> Should there be linkage between per-container stats and
> delivery-via-taskstats? I can't think of one, really.

Since this patch uses Paul's containers to define the term
of a group it uses the provided facilities for reporting the
results :)

> You have cpu stats. Later, presumably, we'll need IO stats, MM stats,
> context-switch stats, number-of-syscall stats, etc, etc. Are we going to
> reimplement all of those things as well? See paragraph #1!

Not reimplement, but collect it in two or three stages: per-task (if
needed), per-container and overall.

There are two ways of doing so:
1. collect it on-demand by walking tasks in container;
2. collect it on-the-fly reporting the values accumulated.

The first way is less intrusive, while the second one is
probably faster. Moreover - the first way is inapplicable
to some statistics, e.g. loadavg. The same stays true for
overall statistics - we may report nr_running by walking
tasklist each time, but it is not used. We suggest to make
it the same way in per-container accounting as well.

> Bottom line: I think we seriously need to find some way of consolidating
> per-container stats with our present per-task stats. Perhaps we should
> instead be looking at ways in which we can speed up paragraph #1.
> -
> To unsubscribe from this list: send the line "unsubscribe linux-kernel" in
> the body of a message to majordomo@vger.kernel.org
> More majordomo info at http://vger.kernel.org/majordomo-info.html
> Please read the FAQ at http://www.tux.org/lkml/
>