Subject: Re: [PATCH, v3 2/2] cgroups: introduce timer slack subsystem
Posted by jacob.jun.pan on Thu, 03 Feb 2011 17:51:17 GMT
View Forum Message <> Reply to Message

On Thu, 3 Feb 2011 11:22:29 +0200
"Kirill A. Shutemov" <kirill@shutemov.name> wrote:

> On Wed, Feb 02, 2011 at 02:56:05PM -0800, jacob pan wrote:
> > On Wed, 2 Feb 2011 22:47:36 +0200
> > "Kirill A. Shutsemov" <kirill@shutemov.name> wrote:
> >
> > > From: Kirill A. Shutemov <kirill@shutemov.name>
> > >
> > > Provides a way of tasks grouping by timer slack value. Introduces
> > > per cgroup max and min timer slack value. When a task attaches to
> > > a cgroup, its timer slack value adjusts (if needed) to fit min-max
> > > range.
> > >
> > > It also provides a way to set timer slack value for all tasks in
> > > the cgroup at once.
> > >
> > > This functionality is useful in mobile devices where certain
> > > background apps are attached to a cgroup and minimum wakeups are
> > > desired.
> > >
> > > Signed-off-by: Kirill A. Shutemov <kirill@shutemov.name>
> > > Idea-by: Jacob Pan <jacob.jun.pan@linux.intel.com>
> > > ---
> > > include/linux/cgroup_subsys.h | 6 +
> > > include/linux/init_task.h | 4 +-
> > > init/Kconfig | 10 ++
> > > kernel/Makefile | 1 +
> > > kernel/cgroup_timer_slack.c | 242
> > > ++
> > >
> >
> >
> >
> > >
> > > +
> > > +static struct cftype files[] = {
> > > + {
> > > + .name = "set_slack_ns",
> > > + .write_u64 = tslack_write_set_slack_ns,
> > > + },
> > should we also allow reading of the current slack_ns?
> >
> There is no 'current slack_ns' for a cgroup since any process free to
> change it with prctl().

>

I think there is still a need for current slack_ns. e.g. if i created a cgroup_1 then attach task_A and task_B to it such that their individual timer_slack got adjusted based on the limit in the cgroup. Now I set cgroup_1 timer_slack to be ts_1, then timer_slack for both task_A and task_B are set to ts_1.

If i attach another task_C to cgroup_1, timer_slack for task_C will be adjusted based on min/max setting of cgroup_1, which can be different than ts_1. User has to manually set cgroup time_slack again to make them identical.

I think this logic defeats the purpose of having timer_slack subsystem in the first place. IMHO, the original intention was to have grouping effect of tasks in the cgroup.

So my suggestion is to keep a per cgroup current timer_slack value, which can be default to the system default at 50us. Like Arjan suggested, we can enforce the timer_slack value in the timer code when it is used. This way we can solve another problem where when a task is detached from the cgroup, it would be desirable to restore its original slack value.