Subject: Re: [BUG]: Crash with CONFIG_FAIR_CGROUP_SCHED=y
Posted by Sukadev Bhattiprolu on Sat, 10 Nov 2007 23:13:47 GMT
View Forum Message <> Reply to Message

Serge E. Hallyn [serue@us.ibm.com] wrote:
| Quoting Srivatsa Vaddagiri (vatsa@linux.vnet.ibm.com):
| > On Fri, Nov 09, 2007 at 09:45:21AM +0100, Dmitry Adamushko wrote:
| > > Humm... the 'current' is not kept within the tree but
| > > current->se.on_rq is supposed to be '1' ,
| > > so the old code looks ok to me (at least for the 'leaf' elements).
| >
| > You are damned right! Sorry my mistake with the previous analysis and
| > (as I now find out) testing :(
| >
| > There are couple of problems discovered by Suka's test:
| >
| > - The test requires the cgroup filesystem to be mounted with
| >   atleast the cpu and ns options (i.e both namespace and cpu
| >   controllers are active in the same hierarchy).
| >
| > # mkdir /dev/cpuctl
| > # mount -t cgroup -ocpu,ns none cpuctl
| > (or simply)
| > # mount -t cgroup none cpuctl -> Will activate all controllers
| >   in same hierarchy.
| >
| > - The test invokes clone() with CLONE_NEWNS set. This causes a a new child
| >   to be created, also a new group (do_fork->copy_namespaces->ns_cgroup_clone->
| >   cgroup_clone) and the child is attached to the new group (cgroup_clone->
| >   attach_task->sched_move_task). At this point in time, the child's scheduler
| >   related fields are uninitialized (including its on_rq field, which it has
| >   inherited from parent). As a result sched_move_task thinks its on
| >   runqueue, when it isn't.
| >
| >   As a solution to this problem, I moved sched_fork() call, which
| >   initializes scheduler related fields on a new task, before
| >   copy_namespaces(). I am not sure though whether moving up will
| >   cause other side-effects. Do you see any issue?
| >
| > - The second problem exposed by this test is that task_new_fair()
| >   assumes that parent and child will be part of the same group (which
| >   needn't be as this test shows). As a result, cfs_rq->curr can be NULL
| >   for the child.
| >
| >   The solution is to test for curr pointer being NULL in
| >   task_new_fair().
| >
| >

| > With the patch below, I could run ns_exec() fine w/o a crash.
| >
| > Suka, can you verify whether this patch fixes your problem?
|
| Works on my machine. Thanks!

And mine too. Thanks,

|
| > --
| >
| > Fix copy_namespace() <-> sched_fork() dependency in do_fork, by moving
| > up sched_fork().
| >
| > Also introduce a NULL pointer check for 'curr' in task_new_fair().
| >
| > Signed-off-by : Srivatsa Vaddagiri <vatsa@linux.vnet.ibm.com>
|
| Tested-by: Serge Hallyn <serue@us.ibm.com>
Tested-by: Sukadev Bhattiprolu <sukadev@us.ibm.com>