Contribution ID: **188** Type: **not specified**

CRIU and the PID dance

Tuesday 10 September 2019 15:10 (20 minutes)

CRIU only restores processes with the same PID the processes used to have during checkpointing. As there is no interface to create a process with a certain PID like fork_with_pid() CRIU does the **PID dance** to restore the process with the same PID as before checkpointing.

The PID dance consists of open()ing /proc/sys/kernel/ns_last_pid, write()ing PID-1 to /proc/sys/kernel/ns_last_pid and close()ing it. Then CRIU does a clone() and a getpid() to see if the clone() resulted in the desired PID. If the PID does not match, CRIU aborts the restore.

This PID dance is slow, racy and requires CAP_SYS_ADMIN.

Fortunately the newly introduced clone3() offers the possibility to be extended to support clone3() with a certain/desired PID. There are currently (July 2019) discussions how to extend clone3() to be able to use it with a certain PID. By the time LPC has started these patches will probably be already posted. With these patches it should be possible to solve the problems that the PID dance is slow and racy.

Which leaves the problem of CAP_SYS_ADMIN. This is a problem for CRIU because it is the major reason why CRIU needs to be run as root during restore. If the root and CAP_SYS_ADMIN requirement could be somehow relaxed it would solve the problems for people running CRIU as non-root for container migration as reported during last year's LPC and it would also open up easy CRIU usage in areas like HPC with MPI based checkpointing and restoring running as non-root.

In this talk we want to give some background how and why CRIU does the PID dance, we want to present our changes based on clone3() to be able to create a process with a certain PID. Then we would like to get feedback from the community if a rootless restore is important and how to relax the CAP_SYS_ADMIN requirement and how this relaxation could be implemented.

I agree to abide by the anti-harassment policy

Yes

Primary author: REBER, Adrian (Red Hat)

Presenter: REBER, Adrian (Red Hat)

Session Classification: Containers and Checkpoint/Restore MC