Contribution ID: **316** Type: **not specified**

Unsolved CRIU problems

Thursday, 19 September 2024 15:00 (20 minutes)

Unsolved CRIU problems.

1) Restoring complex process trees.

Processes can not enter into pre-existing process-session (sid), sessions can only be inherited. (Same for process-groups (pgid) in nested pid namespaces.)

Probable solution 1 - CABA:
The idea was to save as much of the original historical tree topology as possible in an auxiliary in-kernel tree, but it didn't go well. I also have the same thing in eBPF but obviously it is unreliable.
See my previous talk on this matter with a deeper dive.

Probable solution 2 - Allow setsid to pre-existing session + Allow setsid/setpgid to "sid 0":
Is it safe? - We can prohibit entering into a session with controlling ttys, so that there is no way someone can use this change to steal your passwords.

2) The clone3 syscall's set_tid feature is unusable in nested pid and user namespaces (nested containers).

Because, for pid namespace init creation, we need at the same time:
a) be checkpoint_restore_ns_capable at all levels of pid namespace's owner user namespaces;
b) be inside user namespace which is an owner of the pid namespace to be created;

Probable solution 1 - Hack clone3 syscall to receive second user namespace (b) somewhere in arguments.

Probable solution 2 - Make it possible to create a pid namespace separately from creating its init, create init through setns:
Here we need to carefully handle races of two processes created at the same time in a not yet fully setup pid namespace.

3) CPU mismatch.

If we have different cpu features returned by cpuid (or different xsave features/sizes) between nodes, we can not migrate a process between those nodes as glibc of the process may have detected cpu features on the first node and will try to use these features after migration even if the feature is not available on the destination.

Probable solution - Do it like in OpenVZ:
Using cpuid faulting cpu feature to return restricted cpuid mask for containers (need to patch kernel).

Primary author: TIKHOMIROV, Pavel (Virtuozzo)

Co-author: VAGIN, Andrei

Presenter: TIKHOMIROV, Pavel (Virtuozzo)

Session Classification: Containers and checkpoint/restore MC

Track Classification: Containers and checkpoint/restore MC