Contribution ID: **5** Type: **not specified**

Containers and Checkpoint/Restore

The Containers and Checkpoint/Restore micro-conference brings together kernel developers, runtime maintainers, and developers working on container- and sandboxing related technologies in general to discuss current problems and agree on new features.

Both userspace and kernel related work is acceptable. The micro-conference targets the wider container ecosystem ideally with participants from all major container runtimes as well as init system developers.

Contributions to the micro-conference are expected to be problem statements, new use-cases, and feature proposals both in kernel- and userspace.

Last Year Summary

Last year's edition of the Containers and Checkpoint/Restore micro-conference inspired several major achievements:

- Discussion of outstanding problems even after the introduction of the openat2() syscall which did not guard against malicious procfs mounts and were considered a source of security issues led to multiple patches for hardening procfs and mount options with the most recent set targeting v5.13 merge window. More broadly, many discussions about proc-based hardening were the direct result of last year's session at the Containers and Checkpoint/Restore micro-conference.
- A session on the overlayfs filesystem and its use for containers led to a fruitful exchange between the developers of overlayfs and container developers. The inability to mount overlayfs inside an unprivileged container was identified as a major blocker. As a direct result of this session overlayfs has been reworked to enable it to be mountable inside unprivileged containers. This was a significant amount of work that is now complete and has been released as part of Linux v5.11.
- The ability to dynamically alter filesystem permissions for unprivileged containers (formerly known as "shiftfs") was an unsolved problem for over a decade. This topic was brought up several times at the mailing lists and at the Containers and Checkpoint/Restore micro-conference and at last the consensus was reached and the solution in the form of the idmapped mount patchset was merged during the v5.12 merge window. The patchset is a testament to the year-long coordination between filesystem and container kernel developers and stresses the strong tradition of this microconference to drive innovation across multiple years and editions.
- Non-privileged checkpoint restore has been a topic over the last three years with various features implemented both in the kernel and userspace to support it. Last year the focus was on the introduction of a new CAP_CHECKPOINT_RESTORE capability which has been merged in the meantime.
- Thanks to previous micro-conferences the connection between container technologies and checkpoint/restore support was always a part of the discussions which avoids breaking checkpoint/restore when introducing new container features.
- In addition there were fixes to systemd that enabled better integration of systemd-udevd with unprivileged containers, further evolution of pidfd APIs, further extensions of the system call interception feature aka seccomp notifiers, extension of clone3() with set_tid for checkpoint and restore, and many more.

Topics for this Year

This year's edition of the Containers and Checkpoint/Restore micro-conference will focus on a variety of topics that are in desperate need of discussion. The list of ideas is constantly evolving and we expected even more topics to pop up during the coming months as past experience has shown. Here is an excerpt:

- Extending the idmapped mount feature to unprivileged containers, i.e. agreeing on a sane and safe delegation mechanism with clean semantics.
- Porting more filesystems to support idmapped mounts.
- Making it possible for unprivileged containers and unprivileged users in general to install fanotify subtree watches.
- Discussing and agreeing on a concept of delegated mounts, i.e. the ability for a privileged process to create a mount context that can be handed of to a lesser privileged process which it can interact with safely.

- Fixing outstanding problems in the seccomp notifier to handle syscall preemption cleanly. A patchset for this is already out but we need a more thorough understanding of the problem and its proposed solution.
- Discussing an extension of the seccomp API to make it possible to ideally attach a seccomp filter to a task, i.e. the inverse of the current model instead of caller-based seccomp sandboxing enabling full supervisor-based sandboxing.
- Integration of the new Landlock LSM into container runtimes.
- Isolated user namespaces (each with full 32bit uid/gid range) and easier way for users to create and manage them.
- How to best use CAP_CHECKPOINT_RESTORE in CRIU to make it possible to run checkpoint/restore as non-root (with CAP_CHECKPOINT_RESTORE)
- With more container engines and orchestrators supporting checkpoint/restore there has come up the idea to provide an optional interface with which applications can be notified that they are about to be checkpointed. Possible example is a JVM that could do cleanups which do not need to be part of a checkpoint.
- Although checkpoint/restore can handle cgroupv1 correctly the cgroupv2 support is very limited and there is a need to figure out what is still missing to have v2 supported just as good as v1.
- Figure out what is missing on the checkpoint/restore level and maybe the container runtime level to support optimal checkpoint/restore integration on the orchestration level. Especially the pod concept of Kubernetes introduces new challenges which have not been part of checkpoint/restore before (containers sharing namespaces for example).

Key People

- Adrian Reber
- Christian Brauner
- Stéphane Graber
- Mike Rapoport
- Andrey Vagin
- James Bottomley
- Eric Biederman
- Seth Forshee
- David Howells
- Aleksa Sarai
- Tom Hromatka
- Sargun Dhillon
- Alban Crequy
- Amir Goldstein
- Matthew Bobrowski
- Mickael Salaün
- Kees Cook
- Andy Lutomirski
- Christoph Hellwig

I agree to abide by the anti-harassment policy

I agree

Primary authors: GRABER, Stéphane (Canonical Ltd.); RAPOPORT, Mike (IBM); REBER, Adrian (Red Hat); Mr BRAUNER, Christian

Session Classification: Containers and Checkpoint/Restore MC

Track Classification: Containers and Checkpoint/Restore MC