Tracer Namespaces

Mathieu Desnoyers
Michael Jeanson
EfficiOS Inc.

Problem and Goals

- Allow tracing of kernel and user-space instrumentation to be available (consumed) from containers:
    - System calls,
    - Uprobes,
    - User events.
- Tenants should have the ability to observe their own user-space instrumentation and the kernel system calls instrumentation related to their own activity.
    - Also eventually delegated NIC and block I/O devices.
- Should be hierarchical, each layer able to trace itself and its children.
- Would like to create a “fast strace” using kernel ring buffers which can be used within containers.

Per-Namespace Filtering

- In-kernel tracer filtering by tracer namespace including children namespaces
  - Instrumentation sources
    - Syscall events
    - Uprobes
    - User events
  - Filter by namespace comparison needs to be hierarchical
    - When an event is hit, filter comparison needs to compare with all namespace IDs going upwards in the namespace hierarchy.

Namespace Filtering vs Dispatch

- Dispatch events efficiently upwards in the tracer namespace hierarchy.
  - Serialize a copy of the event into each ring buffer which subscribed to this event.
  - Only ring buffers associated with current and parent namespaces may be interested in the event.
  - Take advantage of this to eliminate useless iteration on all system's ring buffers and useless filtering comparisons.
- Only possible if we attach ring buffers to the tracer namespace they belong to.

Tracing Buffers Resource Control

- Resource control to allocate tracer ring buffers within the tracer namespace hierarchy.
  - How can we combine cgroup to control resources with namespace ?

User events "namespace" RFC in tracefs

- [RFC PATCH v2 0/7] tracing: Add tracing namespace API for user
  - https://lore.kernel.org/lkml/20220728235241.2249-4-beaub@linux.microsoft.com/t/
- "To create a tracing namespace, run mkdir under the new tracefs directory named "namespaces" (/sys/kernel/tracing/namespaces typically)”
- I am not convinced that tracefs is the natural place for a namespace
  - How about creating a new namespace instead ?
    - CLONE_NEWTRACER,
    - Use with clone, setns, unshare.
  - Use cgroup to delegate tracing buffer resources to containers ?