Kata-Containers on openSUSE

Ralf Haferkamp, Container Software Engineer, SUSE
Dario Faggioli, Virtualization Software Engineer, SUSE
What is Kata Containers

A container runtime providing stronger isolation by using hardware virtualization technologies.
Traditional Containers

Process A
Filter:
- Seccomp
- MAC
- CAPS
Namespaces

Process B
Filter:
- Seccomp
- MAC
- CAPS
Namespaces

Process C
Filter:
- Seccomp
- MAC
- CAPS
Namespaces

Linux Kernel

CPU  Memory  Network  Storage

Isolation by namespaces, cgroups with shared kernel
Kata Containers

Additional isolation with a lightweight VM and individual kernels
Why Virtualization

• Threat Model: untrusted code in a (Kata) Container attacks the host

• Attack surface--
  – Containers: the shared host kernel: all syscalls (files, directories, MMIO, AIO, different kinds of sockets, different IPC mechanisms, futexes, shared memory, ioctl, TTY, …)
  – Virtualization/Kata: the hypervisor + the VMM: hypercalls + devices.

• Defense in Depth
  – Containers: escape the container ==> Host!
  – Virtualization/Kata: escape the container ==> escape the hypervisor ==> Host

• Isolation++
  – Containers: crash the kernel ==> crash the host ==> DoS for everyone
  – Virtualization/Kata: crash the kernel ==> crash your VM only
Lightweight Virtualization

Low CPU and Memory Overhead
• Small and Fast VMs == More VMs == More Kata Containers

Small & Fast kernel
• Little, tailored, optimized kernel image
  • On openSUSE, currently, kvmsmall as temporary solution
  • Ship Kata upstream kernel?
  • Make one ourselves?

Small & Fast VMM
• QEMU, rust-vmm, FireCracker, CloudHypervisor
  • In openSUSE, currently QEMU
  • Firecracker (available, not fully functional)
  • QEMU MicroVM (when supported in Kata)
OCI compatible runtimes
What Kata Containers is NOT

It’s NOT meant as a mechanism to run „normal“ VM workloads inside Kubernetes.
The above architecture is looking slightly different when container-shim-kata-v2 (shimv2) is used e.g. with containerd
Kata Architecture

- kata-runtime creates a VM per pod (using a pretty minimal kernel and initrd)
- Inside the VM the kata-agent responsible for launching containers and multiplex I/O streams to the outside (either via vsock or virtio-serial)
- If a pod has multiple containers all of the containers are launched within the same VM
- On the host kata-shib communicates with the kata-agent inside the VM. Providing a seamless interface for the upper layer services (cri-o, docker, podman)
Kata Details

- Storage (i.e. the container rootfs and volumes) is shared with the VMs via 9pfs. (when using QEMU/KVM)

- 9pfs has some known performance issues. Work is on the way to move to virtio-fs in the future.

- For networking, kata transparently connects the veth pair from the host to the TAP interface of the VM
Kata-containers on openSUSE

- Tumbleweed is tracking the latest release
- Leap Packages available via the devel:kubic Project in OBS
  https://download.opensuse.org/repositories/devel:/kubic/openSUSE_Leap_15.1
- Packages:
  - katacontainers.rpm
  - katacontainers-image-initrd.rpm for a prebuilt kernel and initrd
Demo
Podman
Kubernetes/Kubic