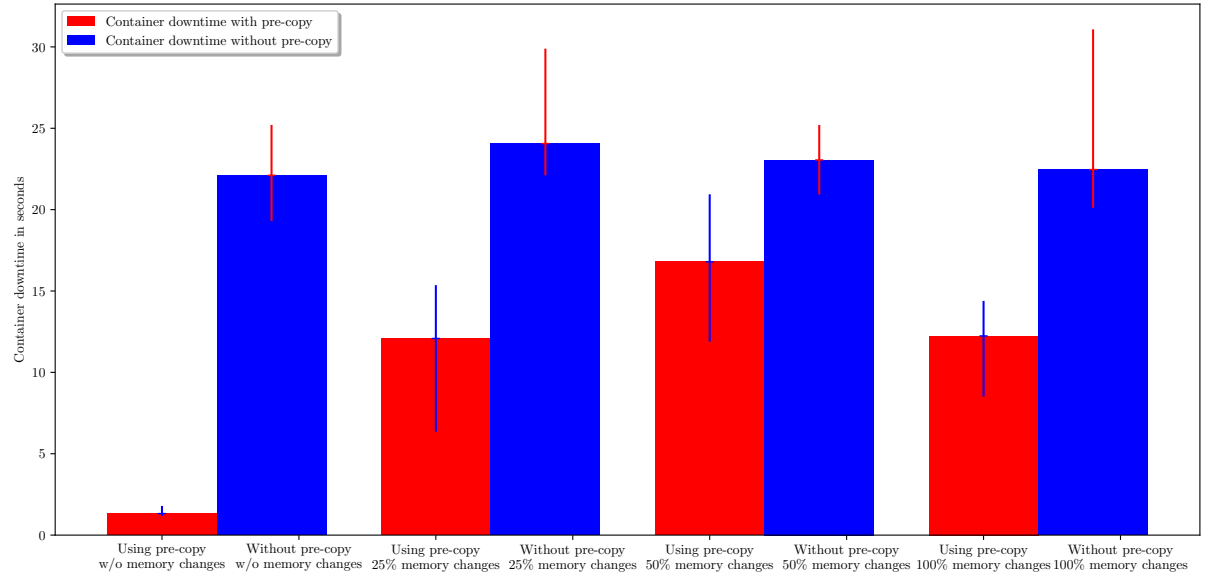
Optimized Container Live-Migration

Incremental Live-Migration of LXD System Containers with CRIU

Adrian Reber <areber@redhat.com>
Christian Brauner <christian.brauner@ubuntu.com>

Tübix 2018
June 09, Tübingen


Container downtime with pre-copy
Container downtime without pre-copy
Container downtime in seconds
0
5
10
15
20
25
30
Using pre-copy
w/o memory changes
Without pre-copy
w/o memory changes
Using pre-copy
25% memory changes
Without pre-copy
25% memory changes
Using pre-copy
50% memory changes
Without pre-copy
50% memory changes
Using pre-copy
100% memory changes
Without pre-copy
100% memory changes

From the Beginning

LXD

- System container manager
- Based on LXC
- Containers can be migrated from one instance to another
  - `lxc move <container> <remote>:<container>`

lxc move

- `lxc move <container> <remote>:<container>`
- Based on lxc-checkpoint
- Based on CRIU - Checkpoint/Restore In Userspace
- Migration steps:
  - Sync file-system (container running)
  - Dump all processes using CRIU (container stopped)
  - Transfer CRIU dump (container stopped)
  - Final file-system sync (container stopped)
  - Restart container on destination
- Migration time depends on
  - Memory size of processes
  - File system change rate

Available CRIU optimizations - 1

- Pre-copy migration
  - Uses the soft-dirty bit from the page table entry
  - Freeze the process
  - Dump the memory
  - Process continues running
  - Memory is transferred to the destination
  - Dump only the memory changes which changed
  - Process continues running
  - Memory is transferred to the destination
  - Dump only the memory changes which changed
  - ...Final dump
- Post-copy migration (lazy migration)
- Combination of pre-copy and post-copy migration

Available CRIU optimizations - 2

- Pre-copy migration
- Post-copy migration (lazy migration)
  - Based on userfaultfd
  - Freeze the process
  - Transfer everything besides memory pages
  - Restart process
  - Transfer memory pages on page fault
- Combination of pre-copy and post-copy migration

Available CRIU optimizations - 3

- Pre-copy migration
- Post-copy migration (lazy migration)
- Combination of pre-copy and post-copy migration
  - First (multiple) pre-copy iterations
  - Migrate to destination
  - Restore missing pages lazily

$\Rightarrow$ To optimize LXC container live migration we started with pre-copy migration.

Adding Pre-Copy-Migration to LXC

- It already existed
- Minimal fixes required to re-enable it
- Added Pre-Copy-Migration Support Check
  - Soft-dirty bit in the PTE is not implemented for all architectures
  - Depends on:
    - Architectures
    - Kernel Version
    - Soft-dirty bit support enabled
    - CRIU Version

Adding Pre-Copy-Migration to LXD

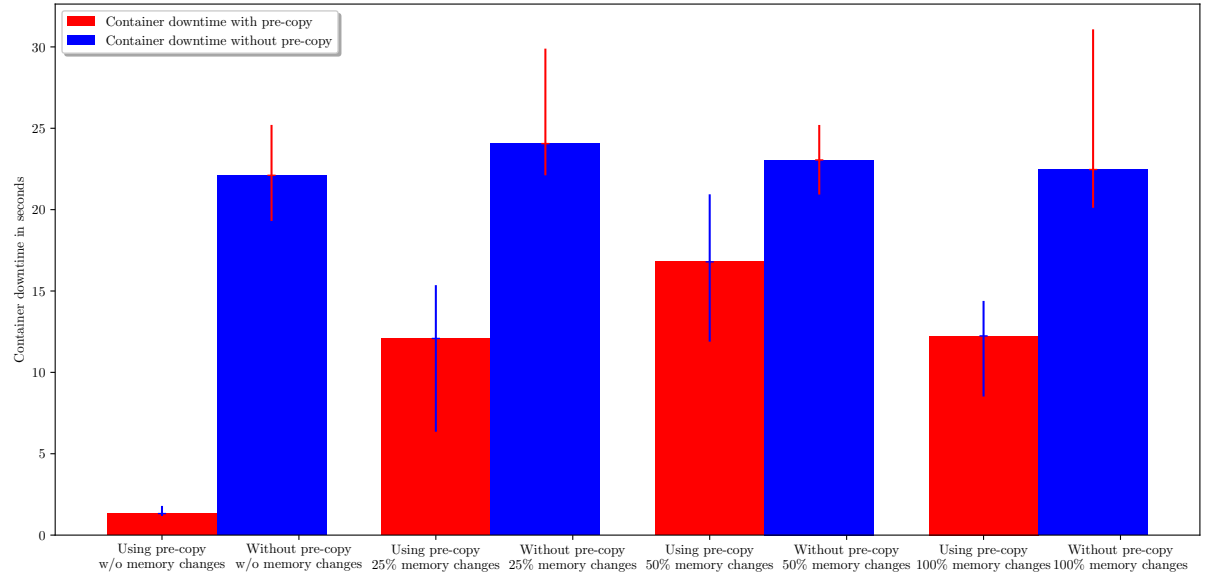
- Check if both sides support pre-copy migration
- Check if user did not disable pre-copy migration
- Do first pre-copy dump using CRIU through LXC
- Check if the percentage of unchanged memory pages is over threshold
- Check if the maximum number of pre-copy cycles has been reached
- ...
- Final dump
- Restore on destination

Pre-Copy-Migration Configuration Options in LXD

- `migration.incremental.memory`
  Incremental memory transfer of the container's memory to reduce downtime.
- `migration.incremental.memory.goal`
  Percentage of memory to have in sync before stopping the container.
- `migration.incremental.memory.iterations`
  Maximum number of transfer operations to go through before stopping the container.

Test Case

- Allocate 1GB of memory
- Read out clock
- Sleep 100µs
- Read out clock and compare
- If difference greater than 400µs $\Rightarrow$ we probably missed a sleep
- Write every/every second/every fourth/no memory page
- Two VMs: RHEL 7.4, latest LXC/LXD from git


Container downtime with pre-copy
Container downtime without pre-copy
Container downtime in seconds
0
5
10
15
20
25
30
Using pre-copy
w/o memory changes
Without pre-copy
w/o memory changes
Using pre-copy
25% memory changes
Without pre-copy
25% memory changes
Using pre-copy
50% memory changes
Without pre-copy
50% memory changes
Using pre-copy
100% memory changes
Without pre-copy
100% memory changes

Current container migration support status

- LXC/LXD
  - CRIU based migration with pre-copy
- runc
  - CRIU based migration with pre-copy and post-copy
- docker
  - CRIU based migration

The end.

Thanks for listening.