

Addressing Challenges for Container Migration in Heterogeneous Clusters



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

Introduction

- **Operational Reality:**
 - User fleets mix older and newer CPU generations.
 - Snapshotting and live migration require hiding CPU differences.
- **Challenge 1: Feature Discovery**
 - Workloads must only use features supported across the whole cluster.
- **Challenge 2: Signal Frame Compatibility**
 - Signal delivery saves CPU registers on the user stack.
 - `rt_sigreturn` can reject a signal frame saved on another CPU.



LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



Advertise Common Feature Set via AUXV

Hardware capability vectors can be used as the source of truth for available CPU features:

- Container runtime sets hardware capability vectors (AT_HWCAP .. AT_HWCAP4) via `prctl(PR_SET_MM, PR_SET_MM_AUXV)`.
- Enables inheritance via `prctl(PR_SET_USER_HWCAP)`.
- On `execve()`, capability vectors are propagated to the new process.
- Privilege Protection:
 - `PR_SET_USER_HWCAP` requires `CAP_SYS_ADMIN` in the user namespace.
 - Any subsequent `saved_auxv` update clears `MMF_USER_HWCAP`.



LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



XSAVE - Hardware-Level State Compatibility

- Fixed Component Layout:
 - Each CPU extension has a CPUID-defined offset and size in the XSAVE area.
- Tracking Only Used Features (XINUSE → XSTATE_BV):
 - Hardware tracks which register states are modified from their initial configuration.
 - XSAVE header's XSTATE_BV bitmap only sets bits for features actually used by the task.
- Cross-CPU Compatibility by Design:
 - XRSTOR only restores state components marked active in XSTATE_BV and resets all other enabled features to their initial state.
 - #GP exception is triggered if XSTATE_BV contains unsupported bits on the target CPU.



Software Checks Blocked Hardware Portability

- Signal frames wrap the XSAVE buffer with software metadata (struct `_fpx_sw_bytes`: `xstate_size`, `xfeatures`, `FP_XSTATE_MAGIC2`).
- Older → Newer CPU (`xstate_size < user_size`):
 - On page fault during `XRSTOR`, the kernel pre-faulted the destination's full `fpstate->user_size`, reading past the end of a smaller migrated signal frame.
 - Fix: Only fault in `xstate_size` bytes from the user signal frame.
- Newer → Older CPU (`xstate_size > user_size`):
 - `check_xstate_in_sigframe()` rejected any frame with `xstate_size > fpstate->user_size` upfront, before `XRSTOR` could see that `XSTATE_BV` only contained supported features.
 - Fix: Allow `xstate_size > user_size` when `XSTATE_BV` only uses supported features.



ARM64: Why Compatibility Is Harder

- Instructions in the HINT space (like PAC, BTI, etc.):
 - Apps execute these instructions without checking anything.
 - On older CPUs, hardware treats them as NOPs.
 - On newer CPUs, they do real work and change process state.
- Signal Frames Include Unused Features:
 - A list of records ({magic, size}) for each feature (FPSIMD, SVE, TPIDR2, etc.).
 - Build the frame based on what the CPU supports; even for features the app never used.
 - `rt_sigreturn` triggers a fault if a frame contains any unknown feature records.

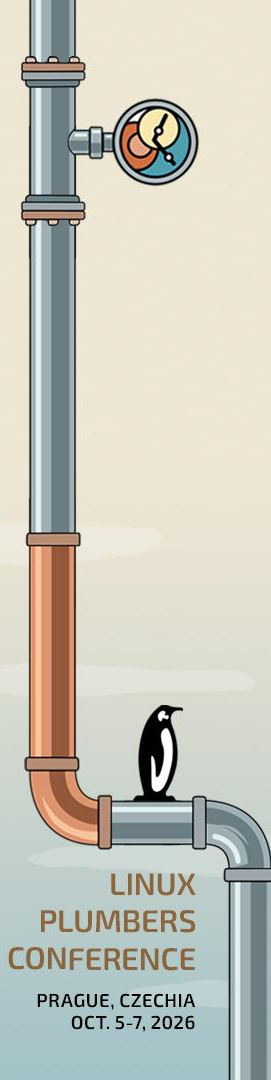


LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



Questions



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026