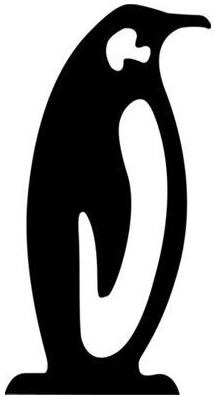
Google



東京 2025 LINUX PLUMBERS CONFERENCE

TOKYO, JAPAN / DECEMBER 11-13, 2025

Google

Live Update IOMMU Preservation

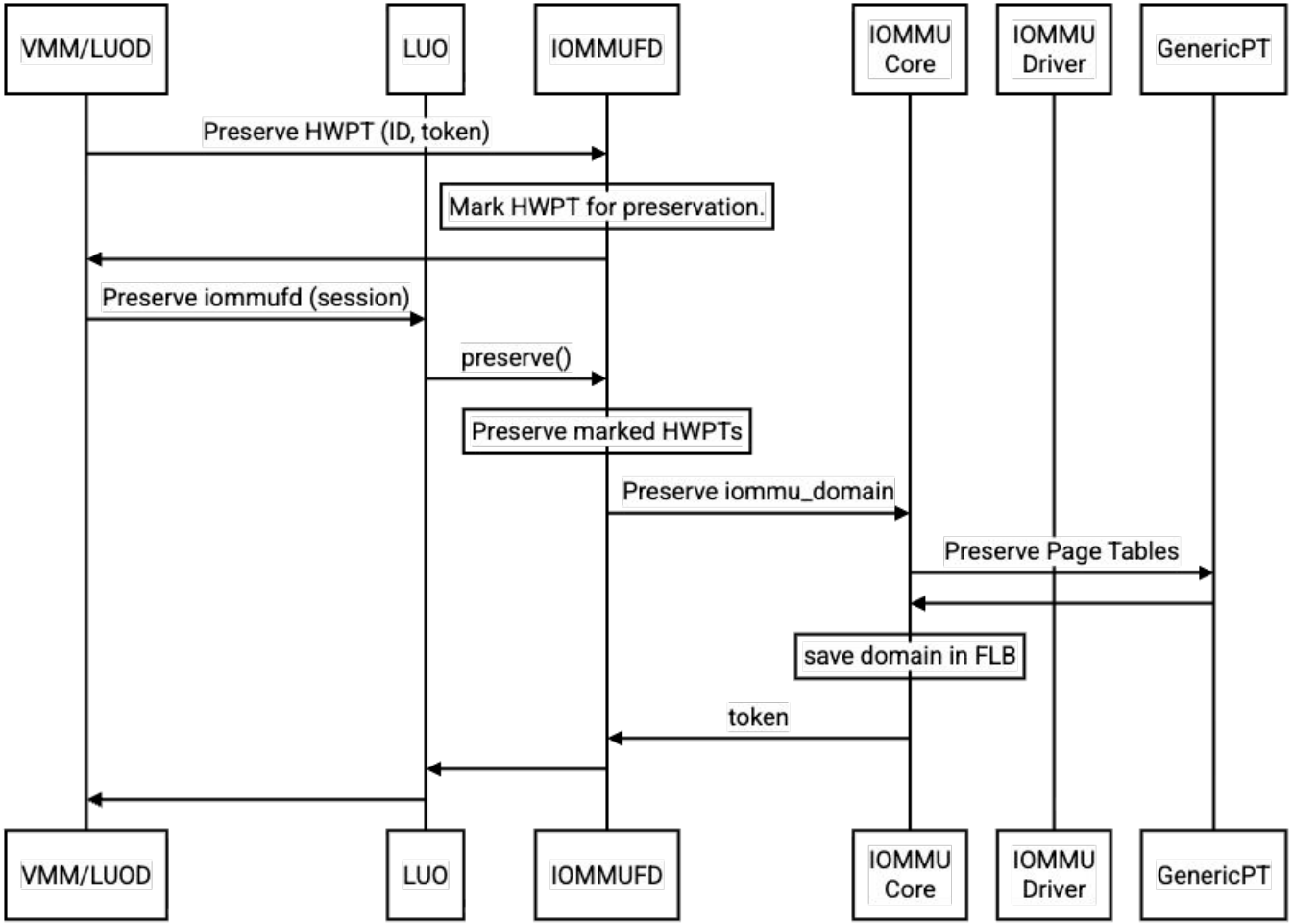
Live Update MC

Samiullah Khawaja <skhawaja@google.com>

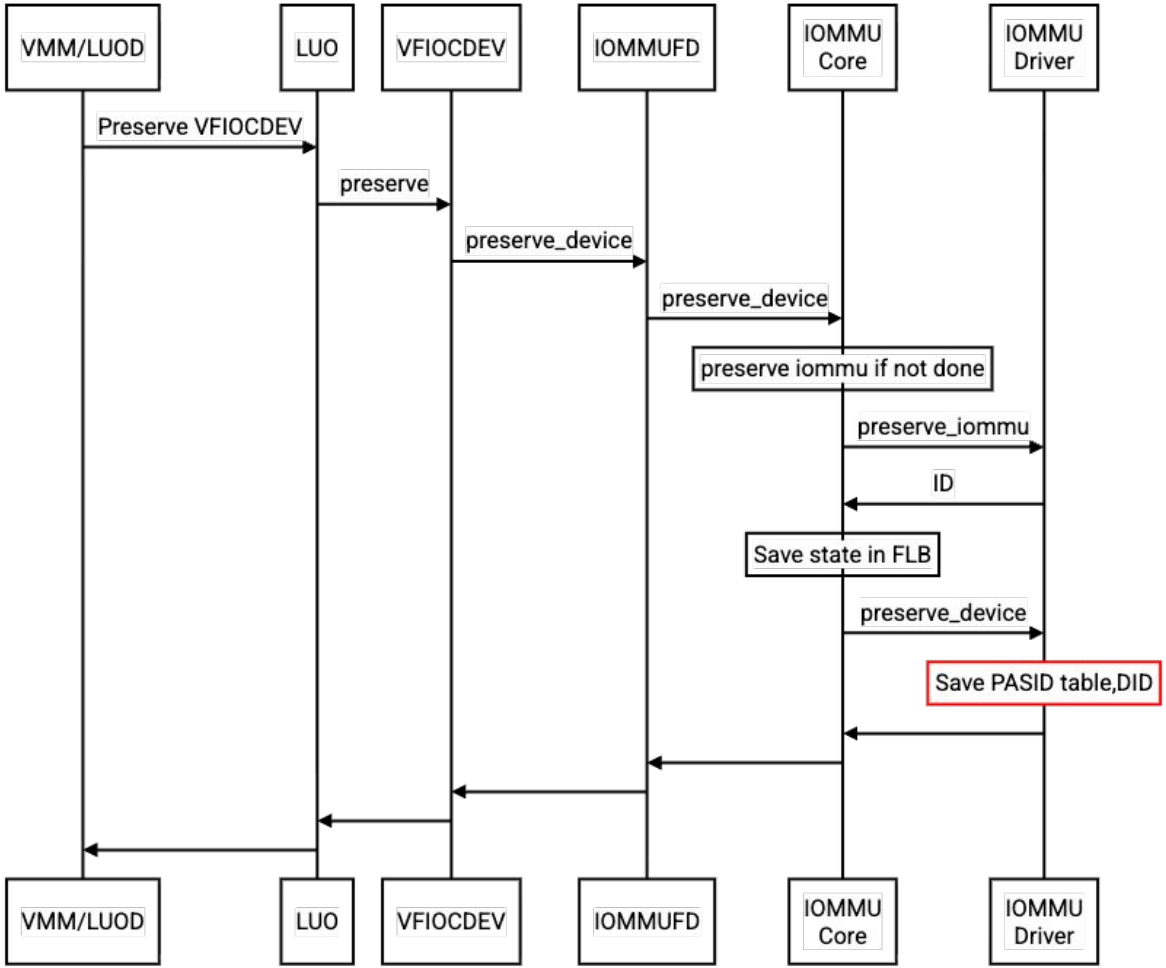
東京 2025
LINUX PLUMBERS CONFERENCE
TOKYO, JAPAN / DEC. 11-13, 2025

Live Update IOMMU Preservation

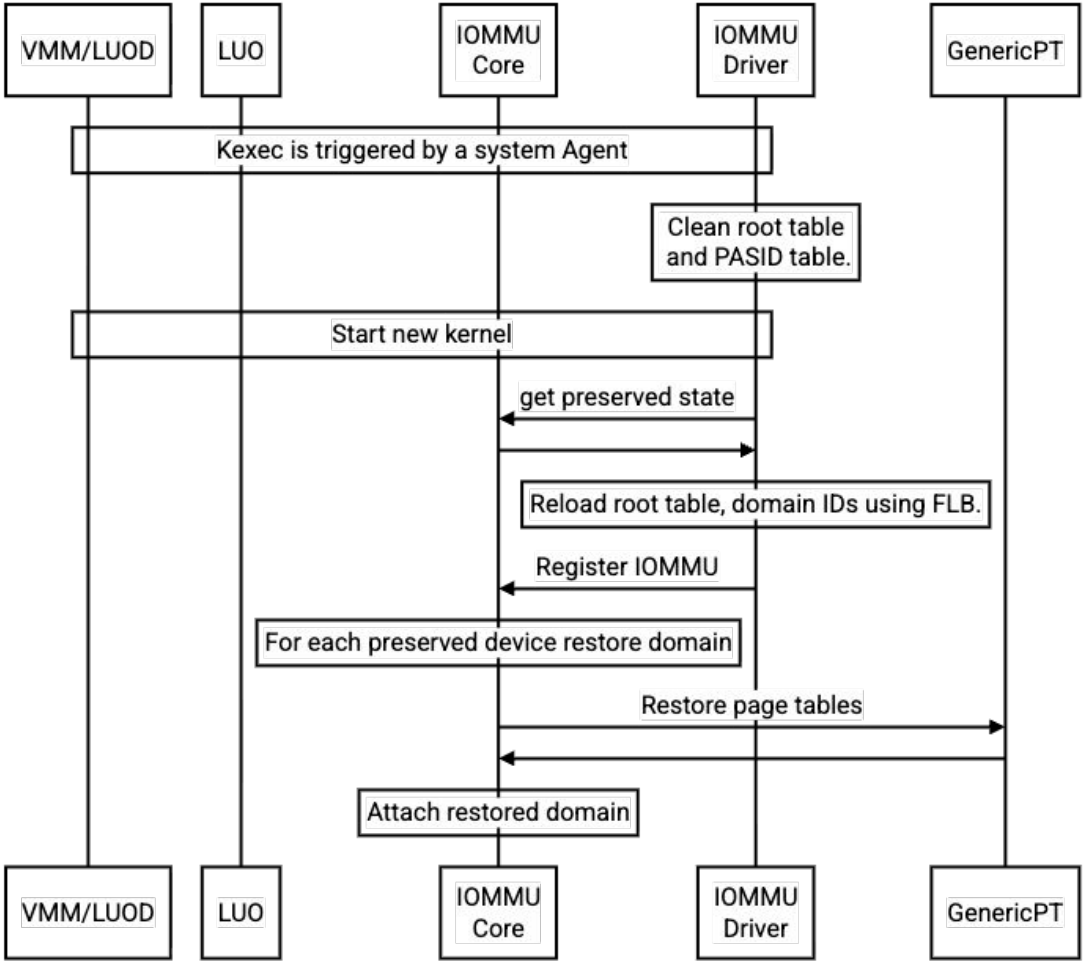
- Preserve the IOMMU domain,
  - Preserve DMA mappings setup by the VMM.
  - Preserve page tables
- Preserve IOMMU device state,
  - Root Tables, DIDs.
- Preserve IOMMU specific state of preserved devices
  - Pasid Tables.
- RFC Patch Series:
  - [RFC PATCH v2 00/32] Add live update state preservation
- Integrated with VFIO cdev preservation:
  - [PATCH 00/21] vfio/pci: Base support to preserve a VFIO device file across Live Update



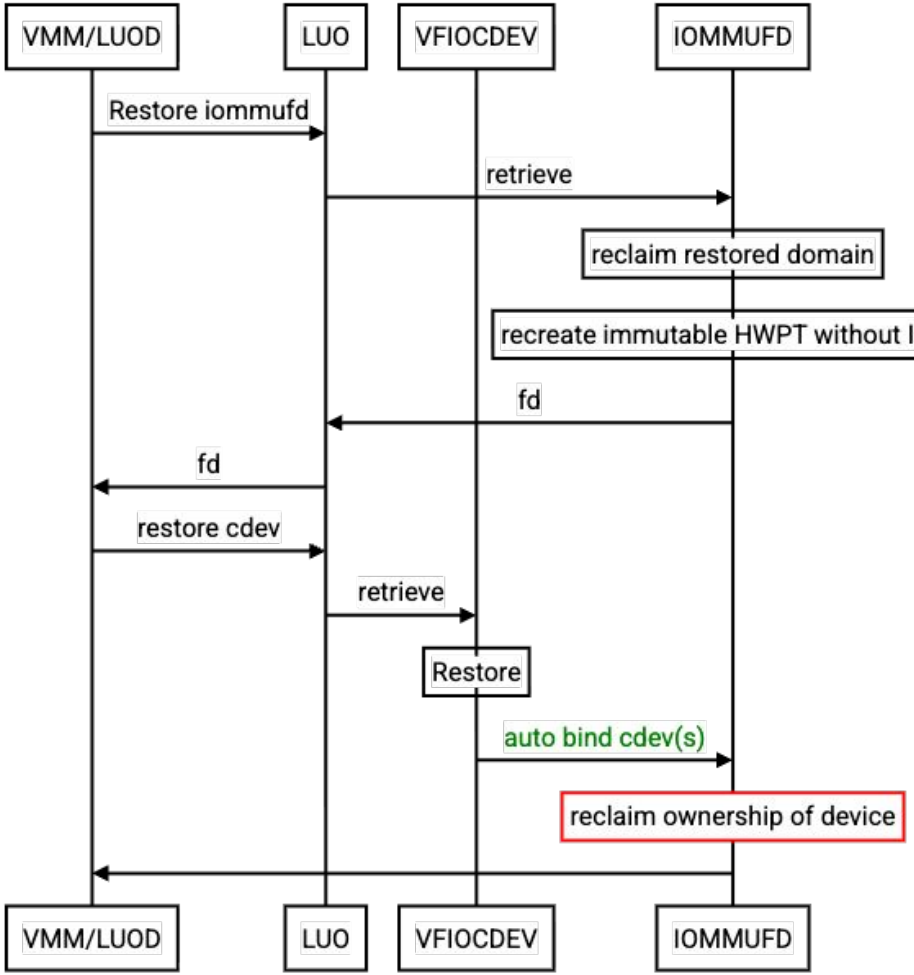
Preservation before Live Update.



Preserve device context using vfio cdev preservation

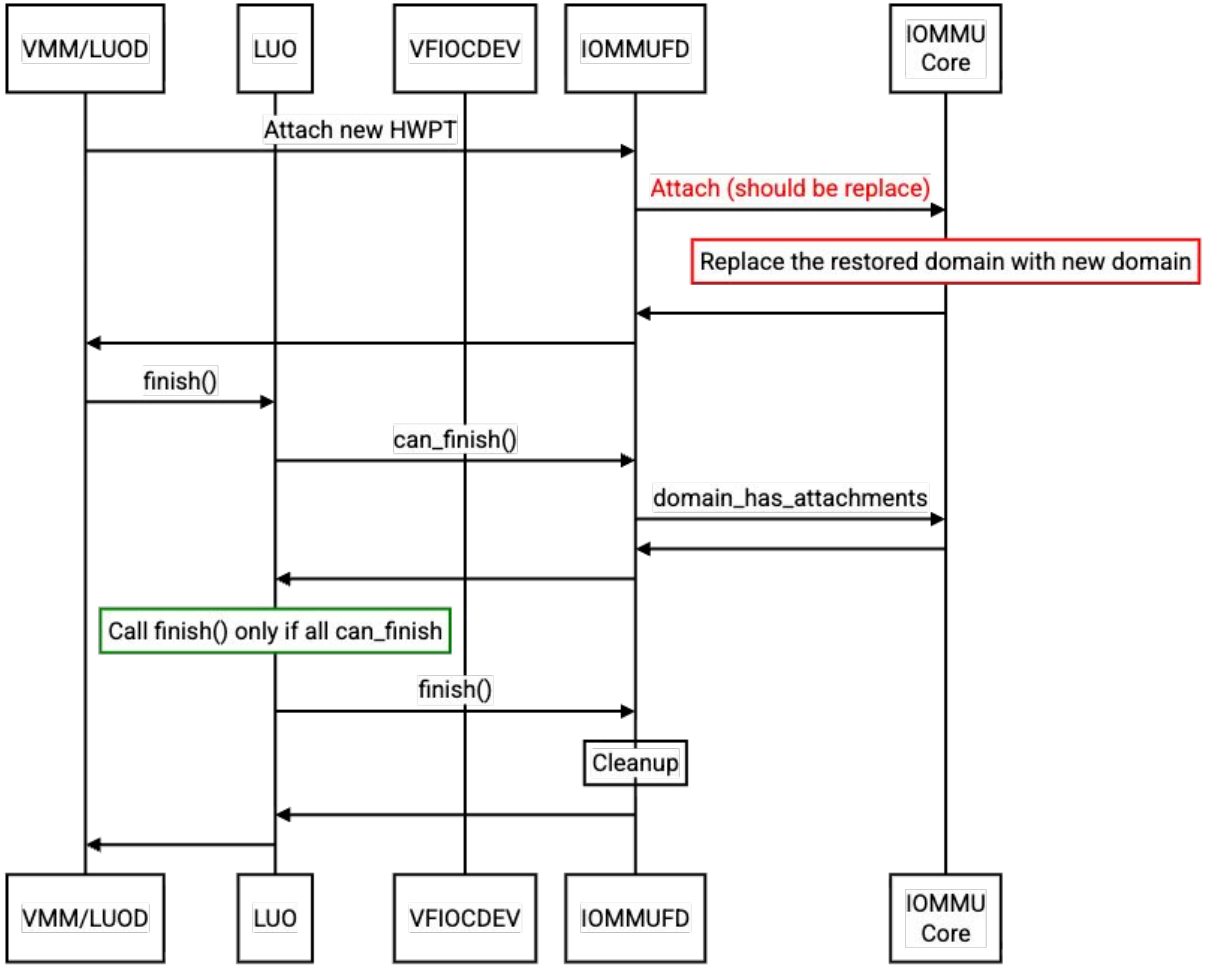


After kexec during boot in next kernel



Autobind: Auto bind needs token of preserved iommufd during preservation. Use LUO API to get it.

Reclaim: iommu group already has a restored domain attached, make Live Update aware.



Reassociate with preserved HWPT during bind so replace is triggered?

Discussion: Replace HWPT with a new one

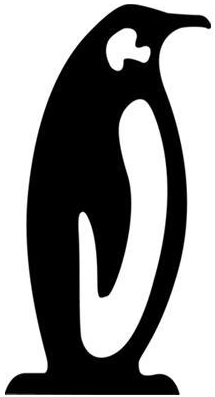
- The restored iommu domain needs to be replaced with a new one.
  - Basically trigger replace from IOMMUFD when attaching the preserved device to a new HWPT after kexec.

- Possible Solutions:
  - Do HWPT replace instead of normal attach when attaching to a device (+PASID) that was previously preserved. (Currently implemented).
  - Reassociate the HWPT with the VFIO Cdev (including PASIDs on restore).
    - During bind iommufd reattaches/associates the restored HWPTs with the preserved device.
    - User does PT attach and it triggers a replace as restored HWPT is attached to it.
    - Allows clean replace using existing logic.

Ongoing/Future work

- Intel IOMMU Driver hitless domain swap.
  - Replace IOMMU domains without blocking translations.
- Arm sMMUv3 support
- Pasid preservation
  - Intel IOMMU Pasid Tables
  - IOMMUFD
  - VFIO CDEV
    - Selectively preserve domains of some PASIDs

Google

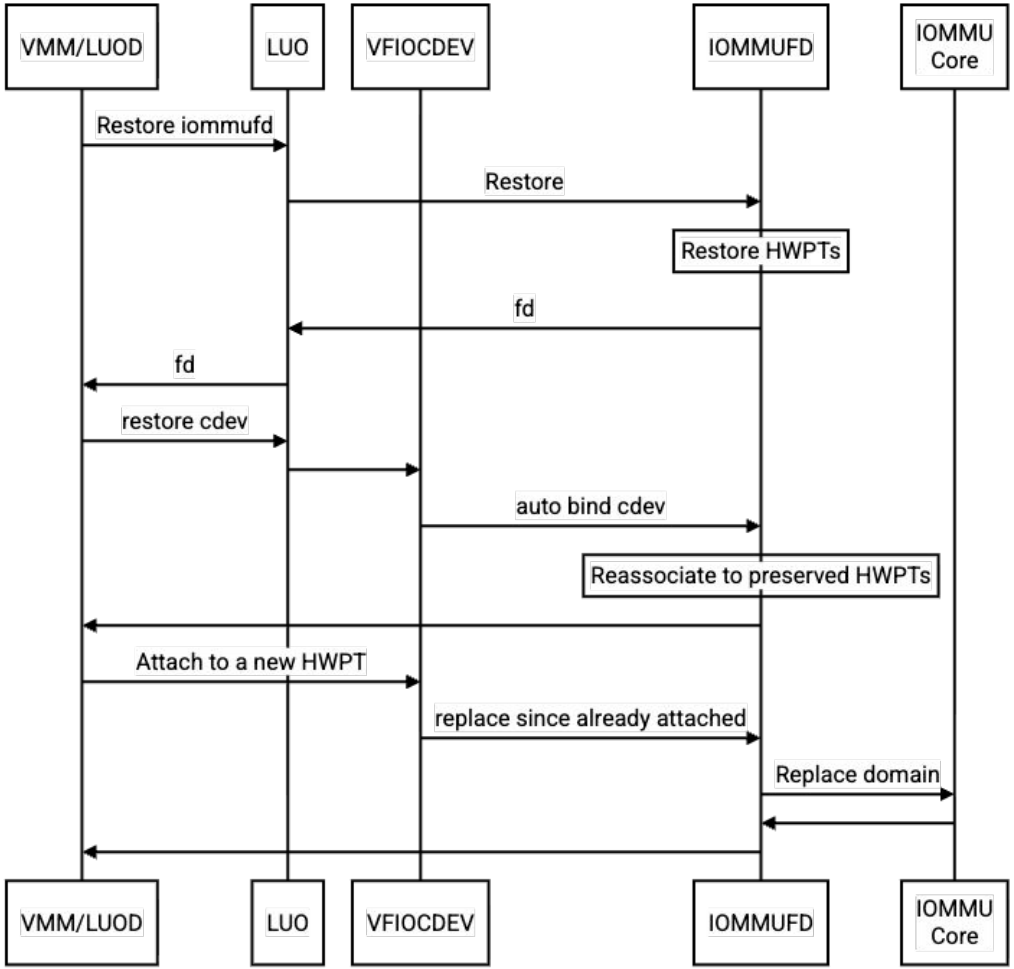


東京 2025 LINUX PLUMBERS CONFERENCE

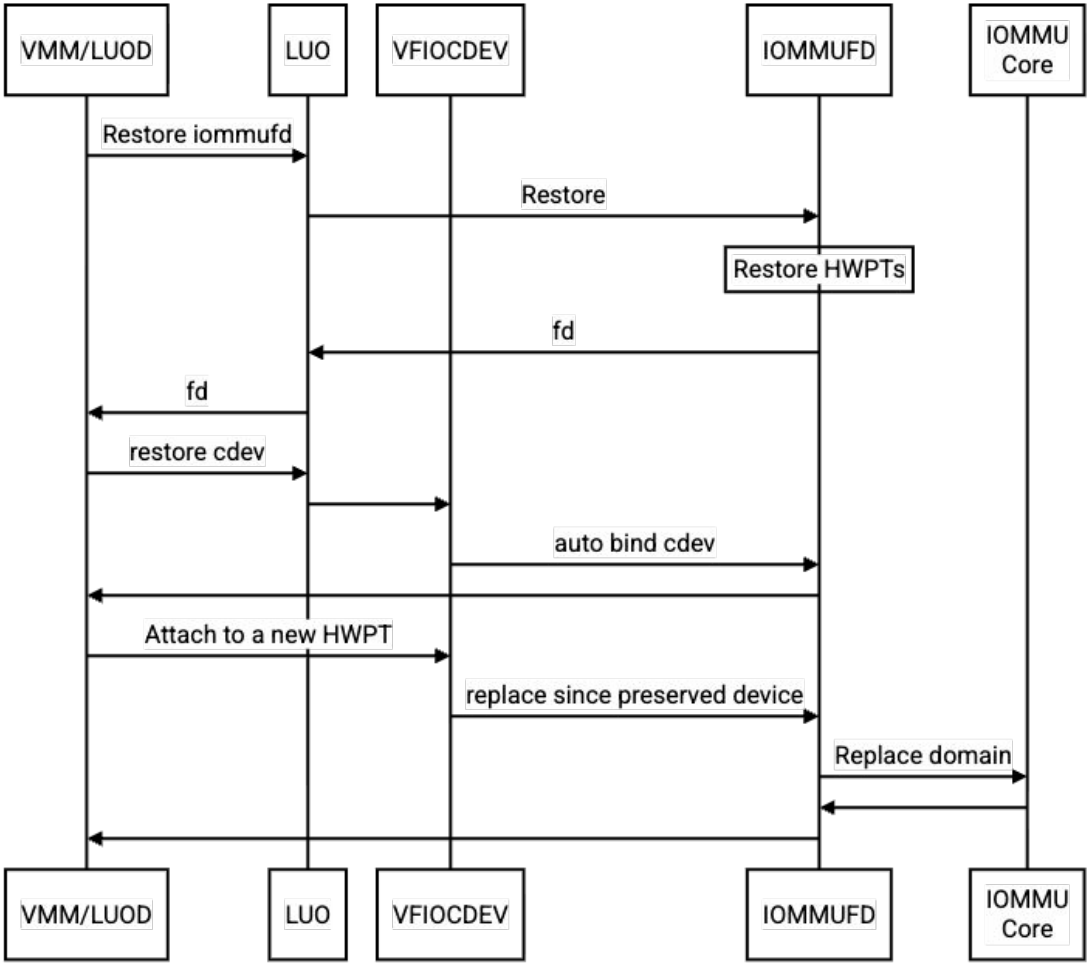
TOKYO, JAPAN / DECEMBER 11-13, 2025

How it works

- User marks HWPTs for preservation.
- Preserve iommufd FD using LUO
  - iommufd preserves iommu domains backed by HWPTs that were marked for preservation.
- During vfio-cdev FD preservation using LUO
  - Call iommufd_preserve_device() to preserve device attachments.
  - Call iommu_preserve_device(dev, domain) to preserve device associated iommu context entry and root table.
- Restore preserved IOMMUs during boot.
- Restore preserved iommu domains during boot.
  - Reattach them with the preserved devices.
- User retrieves iommufd
  - re-associate HWPTs with preserved iommu domains.
- Rebind restored vfio cdev with iommufd
  - User creates a new HWPT (with all the required mappings)
  - Replace restored iommu domain with the new one backed by new HWPT.
  - Destroy old HWPT



Restore HWPT<->cdev association



Do replace HWPT if vfio cdev was preserved

Discussion: Claim DMA ownership

- On iommufd bind claim_dma_ownership returns -EBUSY.
  - This is because the iommu domains are restored and reattached to preserved devices

- Possible Solutions:
  - Allow claim_dma_device_ownership if the device being claimed was preserved
    - Claim can only be done by VFIO driver and VFIO Cdev FD can only be retrieved through LUO.
  - Use preserved device token from previous kernel to verify and transfer ownership.