Linux Plumbers Conference 2025

Contribution ID: **233** Type: **not specified**

Supporting Hypervisor Kexec with Modern Devices

Restarting a node running a stateful workload for an infrastructure software upgrade can be an extremely costly operation. Modern infrastructure software upgrades must also account for applications which are using accelerators such as GPUs, RDMA NICs and NIC stateful flow accelerators. While these workloads may typically run in isolated VMs, a hypervisor reboot for a kernel update can lead to several minutes of downtime and large-scale application restarts. To achieve minimal (tens of milliseconds) to zero downtime, it is critical to enable live updates of a hypervisor kernel using kexec while stateful applications continue to run within their VMs.

In this talk, we discuss changes made to the mlx5 driver stack which bifurcates the physical function's privileges into multiple distinct functions: for the hypervisor and the switchdev offloads. A minimal mlx5 driver is moved to a user-space application using a generic vfio-pci driver and enables delegation of management of vports to another function. The switchdev offload driver is then able to manage these delegated vports. This presentation discusses this design and shows how this driver bifurcation and switchdev extension achieves zero downtime for stateful applications with hypervisor live updates.

https://lore.kernel.org/netdev/20250829223722.900629-1-saeed@kernel.org/

Primary authors: JAYACHANDRAN, Adithya (NVIDIA); MAHAMEED, Saeed (Nvidia)

Presenters: JAYACHANDRAN, Adithya (NVIDIA); MAHAMEED, Saeed (Nvidia)

Session Classification: Live Update MC

Track Classification: Live Update MC