PPC64 - Bridging the pSeries and PowerNV islands for VFIO and IOMMUFD

Shivaprasad G Bhat
Narayana Murty N

IBM Linux Technology Center

LINUX PLUMBERS CONFERENCE | Vienna, Austria
Sept. 18-20, 2024

pSeries VS PowerNV

Same Ocean two Islands

PowerNV – Baremetal Linux HyperVisor

- Based on the OPAL / OpenPower

Pseries – PowerVM Hypervisor

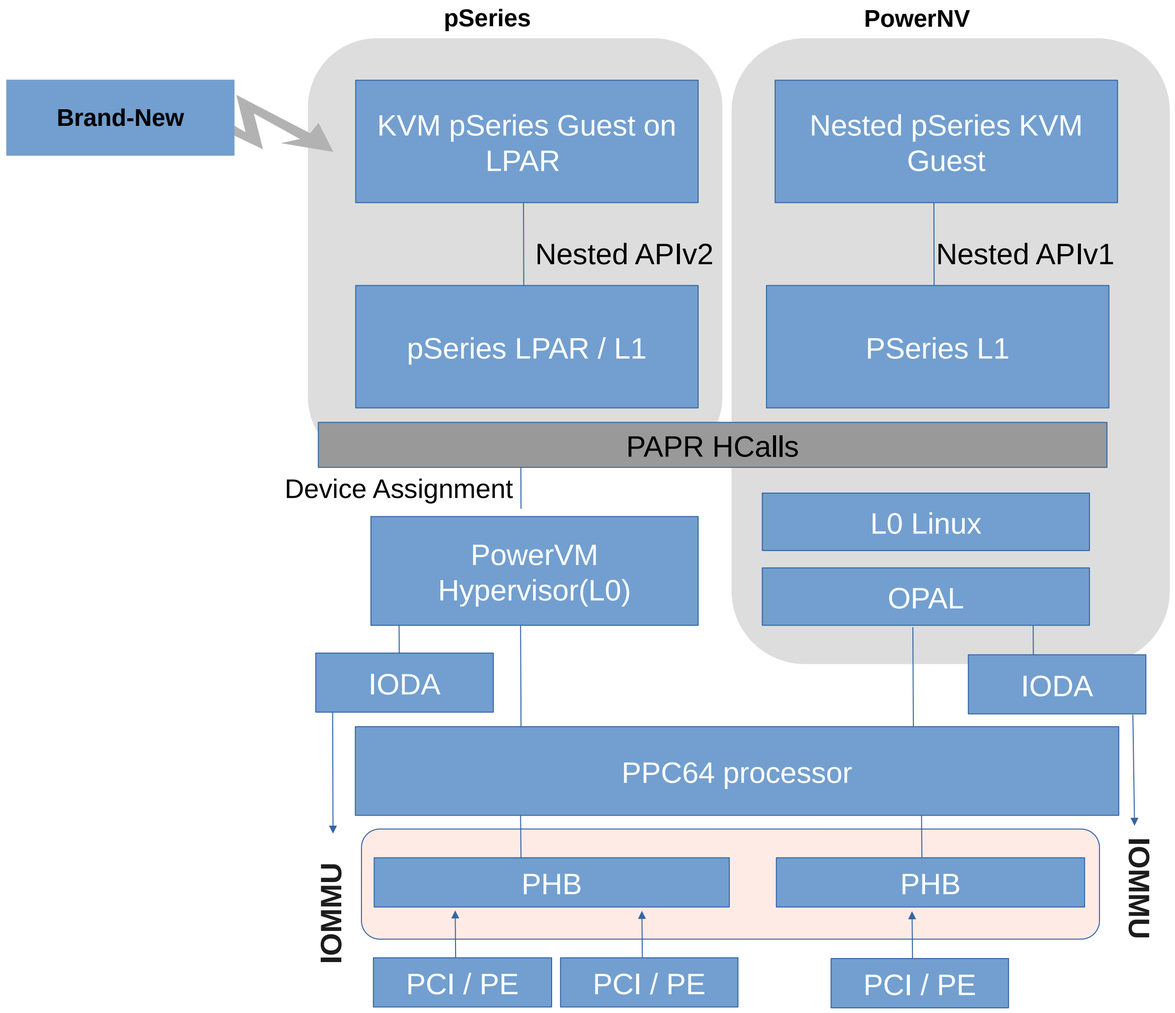
- Based on PAPR with PowerVM RTAS/Hcalls

Whats Common

- Both Hypervisors manage an IOMMU confirming to IODA
- Uniform PowerPC IOMMU APIs based on PAPR
- DMA window management using PAPR RTAS

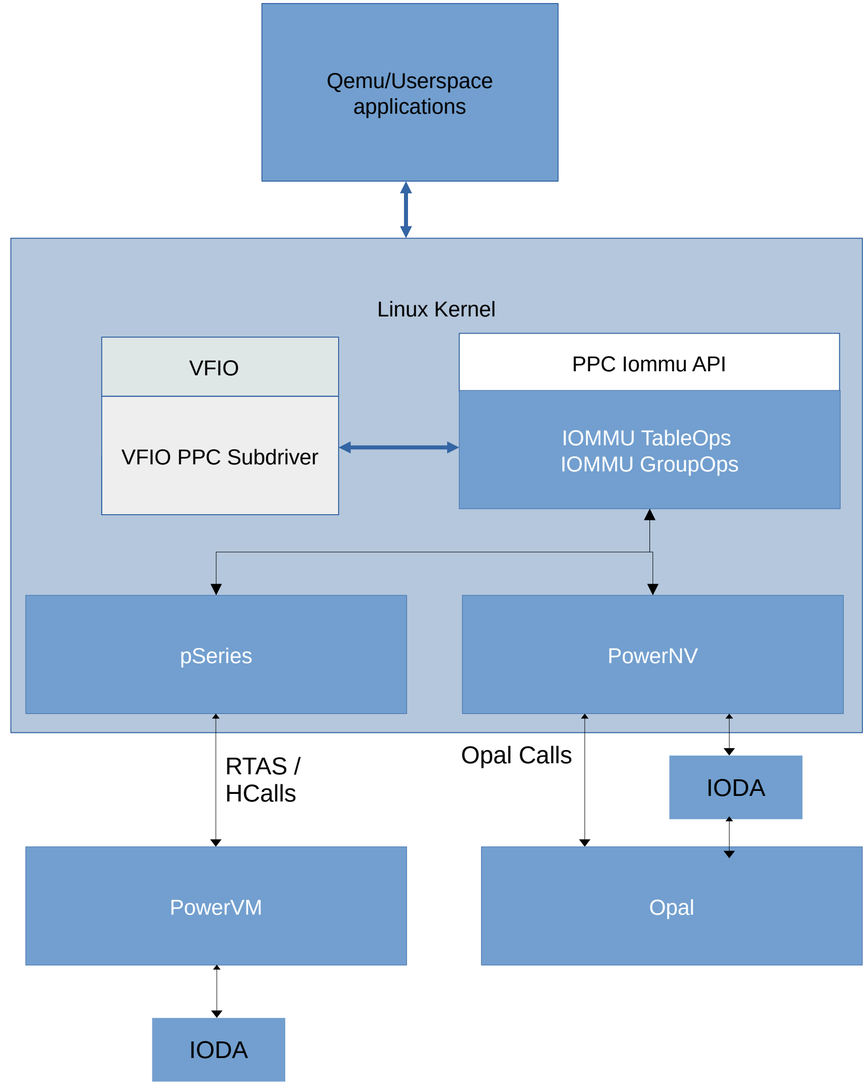
Key Differences

- Limited DMA Windows
- Multi-TCE vs Single TCE



Re-enabling VFIO on pSeries and PowerNV

- Skeleton iommu_ops with default_domain and blocking_domain was introduced[1] for both pSeries and PowerNV.
- Caused some breakages due to ongoing refactoring of platform_domain that were fixed in [2] [3] [4]
- Fixed the userspace view for Single Level TCE on pSeries that was broken due to Multi-tce support on PowerNV [5]
- For pSeries – Re-implementation of iommu_table_ops [5]
- pSeries – Better isolation by preventing DMA window "borrow" machenism [5]
- The VFIO is now back to fully functional state on both platforms!



- References:

1. powerpc/iommu: Add iommu_ops to report capabilities and allow blocking domains
2. iommu: Allow ops->default_domain to work when !CONFIG_IOMMU_DMA
3. powerpc/iommu: Fix the missing iommu_group_put() during platform domain attach
4. powerpc: iommu: Bring back table group release_ownership() call
5. powerpc: pSeries: vfio: iommu: Re-enable support for SPAPR TCE VFIO

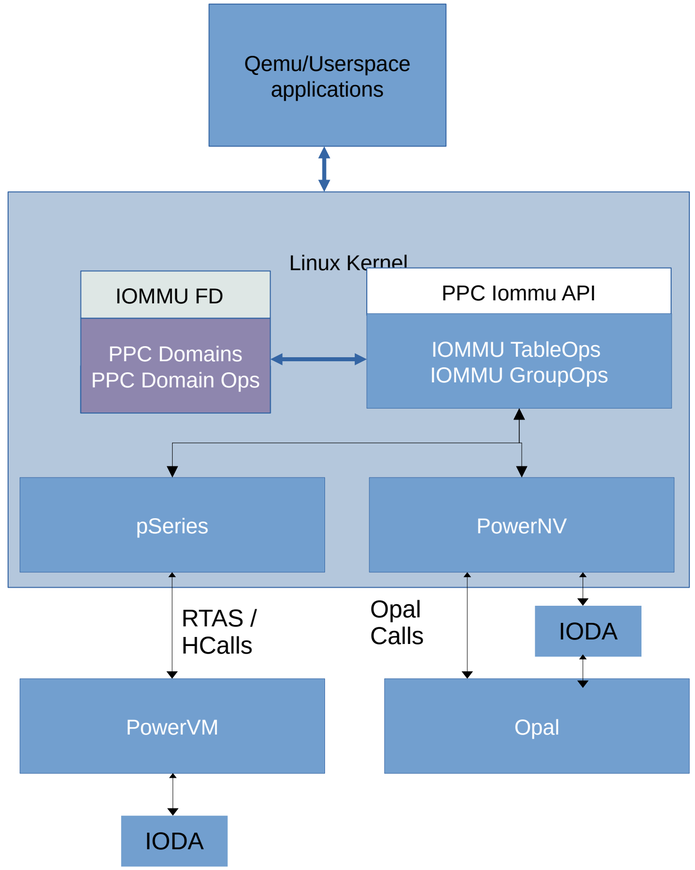
Way forward – IOMMUFD support

Plumbing needed

- Extend existing IOMMUFD driver instead of introducing a new PPC arch specific driver [1]
- Implement new IOMMU Alloc/domain ops for PPC64.
- Provide implementation of {blocked, release}_domain
- iommu_domain_ops.map/unmap use platform specific PPC Iommu {table, group}_ops
- Ensure one to one mapping of Host PE to guest Domain to simplify implementation
- DMA “windows” within a single large aperture of IOVA
- Implementation of viommu [2] for guests. This was discussed and answered well at [5]

Challenges :

- Do we really need support for vfio-compat
- Multi-window apperture within the same iommu_domain [1]
- Nested iommu_domains with SW IOPTE walker like s390 [1]
- Costly GUPs, and the need to pin before hand[1]
- IOVA_MAP with RESERVE flag, IOVA_MAP with FIXED alone can IOVA_COPY with (special/similar) flag
- Guest IOMMU emulation on host, the host can have different IOMMU context
  - x86 on PPC vice-versa(The allowed DMA ranges(IOVAs) are different/incompatible)



Open Questions

- Resolving conflicting ioctl number for vfio-compat thats common for both **vfio_iommu_type1** and **vfio_iommu_spapr_tce.**
  - include/uapi/linux/vfio.h:#define VFIO_DEVICE_BIND_IOMMUFD _IO(VFIO_TYPE, VFIO_BASE + 18)
  - include/uapi/linux/vfio.h:#define VFIO_IOMMU_SPAPR_UNREGISTER_MEMORY _IO(VFIO_TYPE, VFIO_BASE + 18)
- Do we need a new ioctl for creating the window inside nested domain. [1]
- Struct ‘require_iovas’[3] confirm if allow_iovas is achieving it by comparing with original RFC [4]
- Adding Error notifications (EEH support)

References

1 Re: [PATCH RFC 11/12] iommufd: vfio container FD ioctl compatibility

2 Re: [PATCH RFC 11/12] iommufd: vfio container FD ioctl compatibility

3 Re: [PATCH RFC 11/12] iommufd: vfio container FD ioctl compatibility

4 [PATCH RFC 08/12] iommufd: IOCTLs for the io_pagetable

5 Re: [PATCH RFC 11/12] iommufd: vfio container FD ioctl compatibility

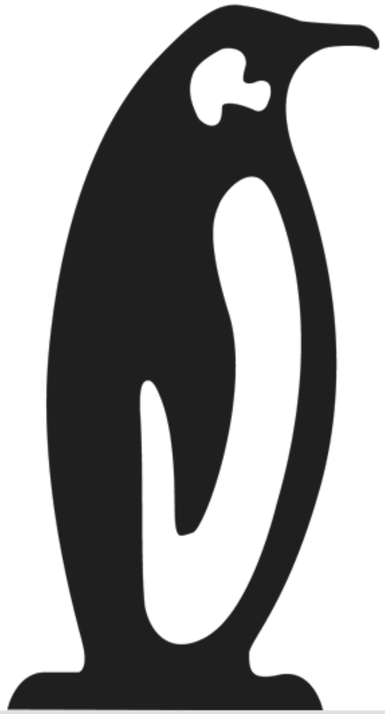
Legal Statement

- This work represents the view of the authors and does not necessarily represent the view of the employers (IBM Corporation).
- IBM and IBM (Logo) are trademarks or registered trademarks of International Business Machines in United States and/or other countries.
- Linux is a registered trademark of Linus Torvalds.
- Other company, product and service names may be trademarks or service marks of others.

Thank you!

Shivaprasad G Bhat
sbhat@linux.ibm.com

Narayana Murty N
nnmlinux@linux.ibm.com



Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024