Contribution ID: **29** Type: **not specified**

VFIO/IOMMU/PCI MC

The PCI interconnect specification, the devices that implement it, and the system IOMMUs that provide memory and access control to them are nowadays a de-facto standard for connecting high-speed components, incorporating more and more features such as:

- Address Translation Service (ATS)/Page Request Interface (PRI)
- Single-root I/O Virtualization (SR-IOV)/Process Address Space ID (PASID)
- Shared Virtual Addressing (SVA)
- Remote Direct Memory Access (RDMA)
- Peer-to-Peer DMA (P2PDMA)
- Cache Coherent Interconnect for Accelerators (CCIX)
- Compute Express Link (CXL)/Data Object Exchange (DOE)
- Component Measurement and Authentication (CMA)
- Integrity and Data Encryption (IDE)
- Security Protocol and Data Model (SPDM)

These features are aimed at high-performance systems, server and desktop computing, embedded and SoC platforms, virtualisation, and ubiquitous IoT devices.

The kernel code that enables these new system features focuses on coordination between the PCI devices, the IOMMUs they are connected to, and the VFIO layer used to manage them (for userspace access and device passthrough) with related kernel interfaces and userspace APIs to be designed in-sync and in a clean way for all three sub-systems.

The VFIO/IOMMU/PCI MC focuses on the kernel code that enables these new system features, often requiring coordination between the VFIO, IOMMU and PCI sub-systems.

Following the success of LPC 2017, 2019, 2020, 2021, 2022, and 2023 VFIO/IOMMU/PCI MC, the Linux Plumbers Conference 2024 VFIO/IOMMU/PCI track will focus on promoting discussions on the PCI core and current kernel patches aimed at VFIO/IOMMU/PCI subsystems. Specific sessions will target discussions requiring coordination between the three subsystems.

See the following video recordings from 2023: LPC 2023 - VFIO/IOMMU/PCI MC.

Older recordings can be accessed through our official YouTube channel at @linux-pci and the archived LPC 2017 VFIO/IOMMU/PCI MC web page at Linux Plumbers Conference 2017, where the audio recordings from the MC track and links to presentation materials are available.

The tentative schedule will provide an update on the current state of VFIO/IOMMU/PCI kernel sub-systems, followed by a discussion of current issues in the proposed topics.

The following was a result of last year's successful Linux Plumbers MC:

- The first version of work on improving the IRQ throughput using coalesced interrupt delivery with MSI has been sent for review to be included in the mainline kernel
- The work surrounding support for /dev/iommufd continues with the baseline VFIO support replacing the "Type 1", has been merged into the mainline kernel, and discussions around introducing accelerated viommu to KVM are in progress. Both Intel and AMD are working on supporting iommufd in their drivers

- Changes focused on IOMMU observability and overhead are currently in review to be included in the mainline kernel
- The initial support for generating DT nodes for discovered PCI devices has been merged into the mainline kernel. Several patches followed with various fixes since then
- Following a discussion on cleaning up the PCI Endpoint sub-system, a series has been proposed to move to the genalloc framework, replacing a custom allocator code within the endpoint sub-system

Tentative topics that are under consideration for this year include (but are not limited to):

- PCI
  - Cache Coherent Interconnect for Accelerators (CCIX)/Compute Express Link (CXL) expansion memory and accelerators management
  - Data Object Exchange (DOE)
  - Integrity and Data Encryption (IDE)
  - Component Measurement and Authentication (CMA)
  - Security Protocol and Data Model (SPDM)
  - I/O Address Space ID Allocator (IOASID)
  - INTX/MSI IRQ domain consolidation
  - Gen-Z interconnect fabric
  - ARM64 architecture and hardware
  - PCI native host controllers/endpoints drivers' current challenges and improvements (e.g., state of PCI quirks, etc.)
  - PCI error handling and management, e.g., Advanced Error Reporting (AER), Downstream Port Containment (DPC), ACPI Platform Error Interface (APEI) and Error Disconnect Recovery (EDR)
  - Power management and devices supporting Active-state Power Management (ASPM)
  - Peer-to-Peer DMA (P2PDMA)
  - Resources claiming/assignment consolidation
  - Probing of native PCIe controllers and general reset implementation
  - Prefetchable vs non-prefetchable BAR address mappings
  - Untrusted/external devices management
  - DMA ownership models
  - Thunderbolt, DMA, RDMA and USB4 security
- VFIO
  - Write-combine on non-x86 architectures
  - I/O Page Fault (IOPF) for passthrough devices
  - Shared Virtual Addressing (SVA) interface
  - Single-root I/O Virtualization(SRIOV)/Process Address Space ID (PASID) integration
  - PASID in SRIOV virtual functions
  - Device assignment/sub-assignment
- IOMMU
  - /dev/iommufd development
  - IOMMU virtualisation
  - IOMMU drivers SVA interface
  - DMA-API layer interactions and the move towards generic dma-ops for IOMMU drivers
  - Possible IOMMU core changes (e.g., better integration with the device-driver core, etc.)

If you are interested in participating in this MC and have topics to propose, please use the Call for Proposals (CfP) process. More topics might be added based on CfP for this MC.

Otherwise, join us in discussing how to help Linux keep up with the new features added to the PCI interconnect specification. We hope to see you there!

Key Attendees:

- Alex Williamson
- Arnd Bergmann
- Ashok Raj
- Benjamin Herrenschmidt

- Bjorn Helgaas
- Dan Williams
- Eric Auger
- Jacob Pan
- Jason Gunthorpe
- Jean-Philippe Brucker
- Jonathan Cameron
- Jörg Rödel
- Kevin Tian
- Krzysztof Wilczyński
- Lorenzo Pieralisi
- Lu Baolu
- Marc Zyngier
- Peter Zijlstra
- Thomas Gleixner

Contacts:

- Alex Williamson (alex.williamson@redhat.com)
- Bjorn Helgaas (bhelgaas@google.com)
- Jörg Roedel (jroedel@suse.de)
- Lorenzo Pieralisi (lorenzo.pieralisi@linaro.org)
- Krzysztof Wilczyński (kw@linux.com)

Primary authors: WILLIAMSON, Alex; HELGAAS, Bjorn (Google); ROEDEL, Joerg (SUSE); WILCZYŃSKI, Krzysztof; PIERALISI, Lorenzo

Track Classification: LPC Microconference Proposals