



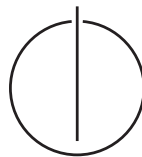
SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Effects of Linux VFIO for User Space I/O

Adrian Simon Würth





SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Effects of Linux VFIO for User Space I/O

Effekt von Linux VFIO auf User Space E/A

Author:	Adrian Simon Würth
Supervisor:	Prof. Dr. Thomas Neumann
Advisor:	Simon Ellmann, M.Sc.
Submission Date:	August 15, 2024



I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, August 10, 2024

Adrian Simon Würth

Abstract

Userspace storage drivers like SPDK rely on Direct Memory Access to read or write to memory. DMA bypasses the CPU and directly performs the I/O operation on the memory. This can be a huge security risk as malicious firmware or faulty operations could lead to detrimental access to memory, potentially extracting data or corrupting the system. The workaround to this is the IOMMU, which maps physical to I/O virtual addresses, similar to the CPU's MMU. As address translation can be a memory and performance intensive operation, it is necessary to look at how impactful the IOMMU is on the whole driver performance. In this thesis we implement IOMMU support for a userspace NVMe driver written in Rust and probe the performance, directly comparing using DMA with physical addresses and with IOMMU I/O virtual addresses. We show that it is possible that with the use of 2 MiB pages it is possible to achieve practically equivalent performance, with increased system security and being able to run the driver with no root permissions. Additionally, we implement support for IOMMUFD, a modern replacement for the VFIO IOMMU Interface.

Contents

Abstract	iii
1 Introduction	1
2 Background	3
2.1 vroom	3
2.2 I/O Memory Management Unit	3
2.3 I/O Translation Lookaside Buffer	4
2.4 Memory Mapped I/O	6
2.5 Direct Memory Access	6
2.6 Hugepages	6
2.7 Peripheral Component Interconnect Express	7
2.8 Character and Block Devices	7
2.9 Non-Uniform Memory Access	8
2.10 Rust	8
3 Related Work	9
3.1 Data Plane Development Kit	9
3.2 Storage Performance Development Kit	9
4 Implementation	10
4.1 syscalls	10
4.2 Virtual Function I/O	11
4.2.1 Regions	12
4.2.2 Groups and Containers	12
4.2.3 Binding NVMe to vfio-pci	12
4.2.4 IOMMU initialization	13
4.2.5 PCIe configuration	14
4.2.6 NVMe initialization	14
4.2.7 DMA (Un-)Mapping	14
4.2.8 I/O operations with VFIO	16
4.3 IOMMUFD	17

5	Evaluation	19
5.1	Setup	19
5.2	Overall Latency and Throughput	21
5.3	Determining IOTLB size	22
5.4	Multithreaded Random Writes	24
5.5	Increasing Queue Depth	26
5.6	Using multiple SSDs	26
5.7	IOMMU modes	26
6	Conclusion	28
	List of Figures	29
	List of Tables	30
	Listings	31
	Bibliography	32

1 Introduction

During his speech "Null Reference: The Billion Dollar Mistake" in 2009, Tony Hoare, a renowned computer scientist, well known for the invention of Quick-sort, proposed the idea of how null pointers are the reason for at least a billion dollars in damages [13]. This quote could not be more important than at this time. In July 2024, Microsoft devices faced what has been described as the "most spectacular IT meltdown the world has ever seen" [18]. This meltdown affected 8.5 million Microsoft Windows devices and severely impacted public institutions, including critical infrastructure like hospitals and airports [5]. In the root cause analysis paper by Crowdstrike, the cybersecurity company that deployed the faulty code, it was revealed that improper compile time validation and missing runtime array bounds checks were a big part of the error [6].

The damage that can be done by a single ring 0 driver like Crowdstrike's Falcon software shows how critical it is to ensure memory safety. By using Rust, a memory-safe yet highly performant programming language with a restrictive compiler, we could drastically improve security and memory safety. We can witness Rust's influence on the systems development community since even the Linux kernel, which has been using C for almost 30 years without accepting any other languages like C++, now allows Rust code in its codebase [17].

However, it's also important to consider Rust's safety limits. While using Rust for a driver improves the overall safety of the process while not compensating on performance, direct memory and I/O operations have to be implemented in an unsafe way. A userspace driver using physical DMA addresses enables a device to practically have full access to the memory and potentially do detrimental I/O operations. Malicious firmware attacks are a rising threat. To enforce safety at the device level, we need to make use of the IOMMU, a safe way of doing direct memory accesses. The IOMMU acts as a layer of isolation between devices and the CPU. By using virtual addresses, the IOMMU can provide a bigger virtual address space and enforce memory access rights [2].

The primary goal of this thesis is to examine how the IOMMU impacts performance in the context of userspace I/O. We demonstrate this by implementing IOMMU support on vroom, a NVMe driver written in Rust [16], and comparing it to using physical addresses. We use the Linux framework VFIO to implement the IOMMU functionality, which has the additional benefit of enabling the driver to run without root privileges.

Additionally, we will be looking at IOMMUFD, a modern replacement for VFIO's IOMMU API.

2 Background

2.1 vroom

Vroom is a userspace NVMe driver written in Rust. As of this writing, it offers good performance and the functionality required for general I/O operations, but it is not yet production ready. Unlike interrupt-driven drivers, vroom uses polling to determine the state of the I/O operations. Polling is often preferable in high-performance applications, as interrupts are relatively performance-intensive operations [21]. When using vroom without the IOMMU, the BAR of the NVMe is exposed in the pseudo-filesystem `sysfs` e.g. for the device with PCI address `0000:01:00.0` under the path `/sys/bus/pci/devices/0000:01:00.0/resource0`. The direct memory access is performed using physical addresses on hugepages.

A NVMe driver consists of submission and completion queues, implemented as ring buffers. The driver adds commands to the submission queue, which the NVMe controller reads and executes. The executed command gets placed on a corresponding completion queue. A deeper explanation of the steps will be provided in chapter 4. As vroom does not have a kernel driver part, we unbind the kernel driver and bind it to `pci-stub`. `Pci-stub` is a dummy driver to occupy the `pci` driver such that the kernel or another application cannot bind to the device.

2.2 I/O Memory Management Unit

Memory Management Units (MMU) for the CPU have been in use since the 1980s. After their first integrated application featuring on Intel's 80286 chip [10], they have since become the de facto standard for addressing memory on computers. By providing processes with a virtual address space instead of physical addresses, every process is isolated and prevented access to memory regions without permission. The MMU uses pages for the translation of addresses. Each address points to a region of memory called a page. These pages can have different sizes, with the default being 4 KiB pages on modern x86-64 architectures.

The translations of these pages are stored in a page table structure. A page table structure consists of multiple tables that store parts of the physical address. When

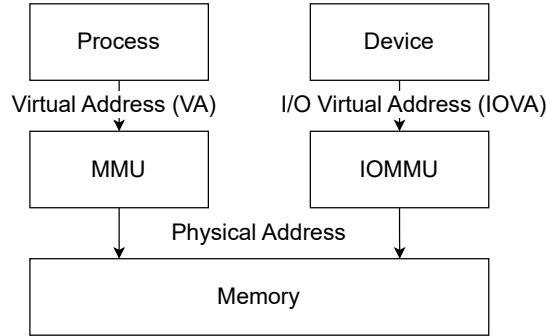


Figure 2.1: MMU and IOMMU relation to physical memory, adapted from [15]

an address is translated, a page table walk must be performed. On a 4 level page table structure as the IOMMU uses for 4 KiB pages, one address resolution results in 4 memory accesses. Thus, a page table walk is a performance costly operation. To circumvent this, a Translation Lookaside Buffer (TLB) is used to cache translation.

The TLB is very performant to access. Frequent access to the same address can be done at a fraction of the time needed to perform a page table walk. A TLB miss describes the scenario in which a physical address needs to be translated but it has no entry in the TLB, resulting in an expensive page walk.

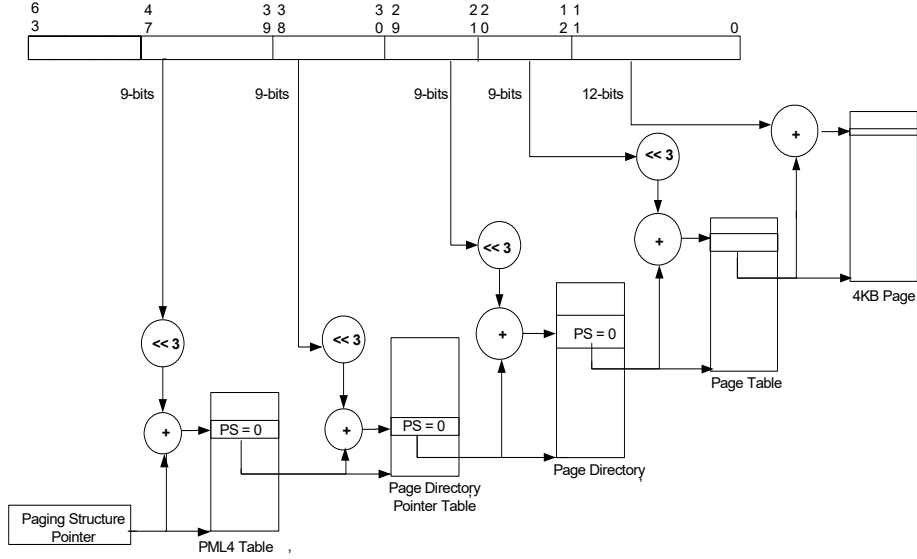
The advantages and success of the CPU's MMU as well as the introduction of the PCIe bus specification have incentivized hardware manufacturers to apply this concept on peripheral device buses. In 2006, Intel introduced their "Virtualization Technology for Directed I/O" (VT-d) and AMD their "AMD I/O Virtualization Technology" (AMD-Vi/IOMMU). In this thesis, the term IOMMU references both technologies.

The IOMMU paging structures consist of 4 KiB page tables storing 512 8-byte entries. The IOMMU uses the upper portions to determine the location of the stored page tables, and the lower portion of the address as page offset. In the case of 4 KiB pages its 12 bits, for 2 MiB its 21.

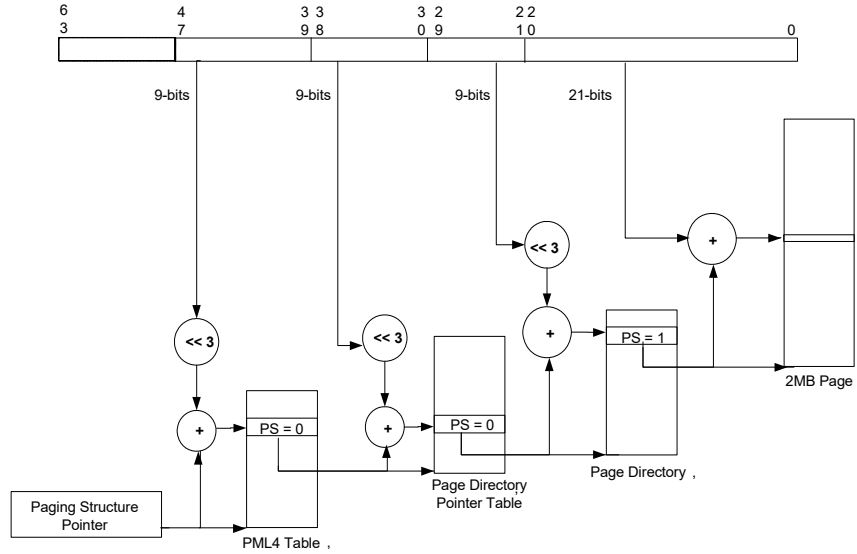
2.3 I/O Translation Lookaside Buffer

As page table walks are rather costly in performance, a cache on the IOMMU is used to store previously calculated addresses. This cache is called the Input/Output Translation Lookaside Buffer. The IOTLB possesses a limited capacity for entries which is not officially documented [8].

2 Background



(a) Translation of a 4 KiB page



(b) Translation of a 2 MiB page

Figure 2.2: VT-d Paging structure for translating a 48-bit address to a 4 KiB and a 2 MiB page, grabbed from [11]

2.4 Memory Mapped I/O

Memory Mapped I/O, in contrast to Port-Mapped I/O, places devices in the same address space as the main memory. Instead of using Assembly instructions like `in` or `out`, a device can be addressed just like the RAM can be. The addressing of this memory itself can currently be done using two ways. Either accessing it with the physical address or using a hardware translation device like the I/O Memory Management Unit (IOMMU). Before our implementation, vroom currently only supports the former.

Using the Linux syscall `mmap(2)` we can create a mapping in the virtual address space of our process. These mapping use so-called Pages. Pages are units of memory with a predefined size, the `pagesize`. The supported pagesizes can differ depending on the CPU architecture. As we use x86-64 for our tests, we have the default pagesize of 4 KiB and the bigger pagesizes 2 MiB and 1 GiB. Pages allocated with these bigger pagesizes are called Hugepages. In addition to mapping memory into our virtual address space, calling `mmap` with no address can be used as an allocation method for big data amounts [12].

2.5 Direct Memory Access

Using Direct Memory Access we can bypass the CPU for I/O operations. Previously this was handled by a separate DMA-controller hardware (third-party DMA) but using PCI, we can directly access it through bus mastering (first-party DMA). Using the IOMMU the request gets intercepted and translated to the physical address.

2.6 Hugepages

As the demand for bigger memory mappings e.g. for big files increased, the amount of TLB cache misses rose proportionally. With the CPUs TLB having space for only 4096 4 KiB pages, only an address space of 16 MiB could be stored and accessed quickly. To increase the amount of virtual memory space, hardware producers reacted by providing bigger page sizes on their architectures than the default 4 KiB. Linux currently provides two ways of using Hugepages. In the optimal case, using a 2 MiB or 1 GiB page size should result in a 512- or 262144-times reduction in cache misses compared to 4 KiB pages. Especially in high-performance computing this make a huge difference.

- **Persistent Hugepages:** Persistent Hugepages, are reserved in the kernel and cannot be swapped out or used for another purpose. To use these hugepages, they are mounted as a (pseudo) filesystem called `hugetlbfs`, which lays in

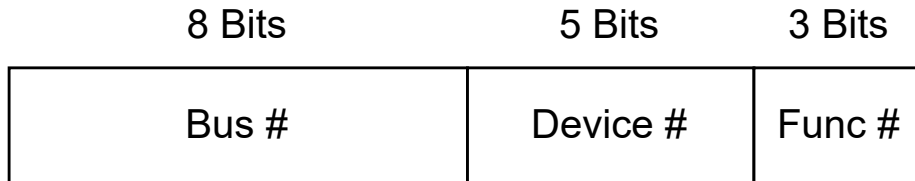


Figure 2.3: segmented PCI identifier

the directory `/mnt/huge`. The amount and size of the pages can be specified either during boot on the kernel commandline with e.g. `hugepagesz=1g hugepages=16` or dynamically using the Linux proc virtual filesystem e.g. `echo 16 > /proc/sys/vm/nr_hugepages`[9].

- **Transparent Hugepages:** Transparent Hugepages are a more recent addition to the kernel. Transparent hugepages are not fixed or reserved in the kernel and allow all unused memory to be used for other purposes. THPs provides a way of utilising the TLB effectively without reserving huge amounts of memory. The khugepaged daemon scans memory and collapses sequences of basic pages into bigger pages. THPs can either be enabled, disabled or only be used on `madvise(MADV_HUGEPAGE)` memory regions [22].

vroom currently uses hugepages for DMA and locks them with the Linux syscall `mlock` to prevent the Kernel from swapping them out. In order to ‘pin’ pages, preventing the Linux Kernel from switching them around, Hugepages have to be used. The kernel does not have the ability to move these pages like 4 KiB pages.

2.7 Peripheral Component Interconnect Express

PCIe is a standard for peripheral device busses. Each device on the PCI bus has a unique PCI address, which is segmented into for parts, as seen in Figure 2.3. The maximum data payload size of PCIe is 4 KiB.

2.8 Character and Block Devices

Unix/Linux use two types of devices: Character and Block devices. Character devices are used for devices with small amounts of data and no frequent seek queries, like keyboard and mouse. Block devices on the other hand have large data volumes, which

are organized in blocks and where search is common, like HDDs and SSDs. Read and Write operations on character devices are done sequentially byte-by-byte, while on block devices, read/write is done at the data block level. These constraints also impact how the drivers for these devices work. CDev drivers directly communicate with the device drivers, while block device drivers work in conjunction with the kernel file management and block device subsystem. This allows efficient asynchronous read/write operations for large data amounts, but small byte sized data transfer achieves lower latency on character devices.

2.9 Non-Uniform Memory Access

Non-Uniform Memory Access is a type of system architecture for multi-CPU systems. Hardware resources are bundled into "nodes". These nodes can contain multiple CPUs, memory or I/O buses [19]. This can relieve stress on the memory bus and improve performance if the memory is part of the same node as the CPU. On the other hand, performance may decrease if the memory is situated in another node. Using `numactl` it is possible to specify the locality of memory and cpu and run a process on one node [14].

2.10 Rust

While userspace drivers can theoretically and practically be written in any language, as proven by the network driver `Ixy` ??, Rust excels as it not only offers memory-safety but memory-safety without garbage collection. This is especially important, as garbage collected languages have overhead and latency spikes, which can lower performance. Another huge factor is that Rust, like C, does not use exceptions. By handling, or even better, being forced to handle errors assures that no rogue exception can take down critical code infrastructure. Additionally, Rust provides low-level access while also offering a high-level development experience through zero-cost abstractions.

3 Related Work

3.1 Data Plane Development Kit

The Data Plane Development Kit (DPDK) is a framework for developing user-space network card drivers. It allows for high performance network applications. It can either run using direct memory access with physical addresses or using VFIO [1]. DPDK offers polling drivers for a variety of network cards. It is one of the most successful projects in the world of userspace drivers and has influenced many advances in the IOMMU space.

3.2 Storage Performance Development Kit

The demand for high-speed userspace drivers in storage applications inspired the development of the Storage Performance Development Kit (SPDK). SPDK uses some shared libraries and architecture with DPDK. Especially through the wide adoption of the NVMe protocol and the standardization of said protocol, only one driver for all NVMe SSDs has to be developed. NVMe is a storage protocol which is widely used, modern and highly performant. Therefore it is a protocol for which many drivers have been written, including userspace drivers. The Storage Performance Development Kit (SPDK) provides “a collection of tools and libraries for writing high performance, scalable, user-mode storage applications” [20]. It includes an user-space NVMe driver which is fast and production-ready. While this driver supports the use of the driver without the IOMMU, the SPDK Documentation recommends using the IOMMU as using VFIO and the IOMMU is the “future proof...long-term foundation” for SPDK [4]. Even though SPDK is the established userspace NVMe driver option, the drawbacks include its high complexity even for simple applications, as well as it being written in C.

4 Implementation

4.1 syscalls

A variety of syscalls are used in the process of allocation and mapping. The syscalls that are used by vroom are `mmap`, `ioctl`, `pread`, `pwrite` (and `mlock` for the non IOMMU version). While there are crates that implement the syscall functionality, we only use the `libc` crate to avoid inflating the dependency list and executable size. As these require C-like syntax and an `unsafe` block in Rust, wrapper macros are used to provide locality of behaviour and secure error handling. In Listing 4.1 the macro for the `mmap` syscall can be seen. As part of our error handling, we introduce an error enum variant for each syscall. To not hide the inherit unsafety of these macros, we add the suffix `"_unsafe"`.

Listing 4.1: Syscall `mmap` macro, with own error variant

```
#[macro_export]
macro_rules! mmap_unsafe {
    ($addr:expr, $len:expr, $prot:expr, $flags:expr, $fd:expr, $offset:
    ↪ expr) => {{
        let ptr = unsafe { libc::mmap($addr, $len, $prot, $flags, $fd,
    ↪ $offset) };
        if ptr == libc::MAP_FAILED {
            Err(Error::Mmap {
                error: (format!("Mmap_with_len_{}_failed", $len)),
                io_error: (std::io::Error::last_os_error()),
            })
        } else {
            Ok(ptr)
        }
    }};
}
```


4.2 Virtual Function I/O

Virtual Function I/O (VFIO) is an IOMMU agnostic framework for exposing devices to userspace. VFIO consists of two parts, the `vfio-pci` driver and an IOMMU API. The IOMMU API can either be the type1 IOMMU API for x86 architectures or the SPAPR IOMMU API for ppc64 architectures. We will only use the type1 IOMMU API. The VFIO PCI driver can be bound to a PCI device. This allows using `mmap(2)` to map the PCI device registers into memory. The type1 VFIO IOMMU API is used for mapping and unmapping address translations in the IOMMU. Alternatively, the IOMMUFD API can be used instead of the container API, which currently is not feature complete, but will eventually replace the container-based solution [23]. The layers of VFIO can be seen on Figure 4.2a.

To use the IOMMU we need to initialize the IOMMU, VFIO, DMA and the NVMe device:

1. **Binding NVMe to `vfio-pci`:** Before initializing the devices, the NVMe device needs to be unbound from the kernel driver and bound to `vfio-pci`.
2. **IOMMU initialization:** As the first step of the actual driver, we need to initialize the IOMMU. This is done using the VFIO IOMMU API, which is an interface for the IOMMU driver. We can get the container file descriptor with VFIO. The group needs to be assigned to the container. In this step we also can attain the device file descriptor, which can be used to read/write/`mmap` the device and PCIe registers through the IOMMU.
3. **PCIe configuration:** Using the device fd, we can enable DMA by setting a bit in the PCIe command register.
4. **NVMe initialization:** We use `mmap` to map the NVMe base address register into memory for configuring the NVMe device.
5. **DMA (Un-)Mapping:** Using `mmap`, we can either map a file into the virtual address space or allocate space in memory. Using the container fd, we can create a mapping in the IOMMU, and therefore exposing it to the NVMe controller for DMA.

Communication with the VFIO interface works by using `ioctl` system calls. While there is Rust's extensive `libc` library providing the system calls `ioctl` and `mmap` and their flags, the Linux `vfio.h` constants and structs need to either be defined manually or with a crate like `bindgen`, which automates bindings for C and C++ libraries [3]. To keep the binary and dependency list as small as possible we chose the manual implementation.

4.2.1 Regions

Using regions, we can directly mmap device memory into host memory for easy access to the NVMe controller. VFIO provides structs for using mmap to directly map the NVMe device into memory. Using `VFIO_DEVICE_GET_REGION_INFO` we can attain the length and the offset needed for mmap.

4.2.2 Groups and Containers

VFIO uses group to distinguish between groups of devices which can be isolated from the host system. In the ideal case, every device would only be part of one group in order to increase security by providing single-device isolation. Groups are the smallest unit size on a system to ensure secure user access. To further reduce overhead from the IOMMU Containers are used in VFIO, which can hold multiple groups. These containers can be used to ease translation and reduce TLB page faults.

4.2.3 Binding NVMe to vfio-pci

To use the IOMMU for the driver, we first need to initialize the VFIO kernel module and bind the VFIO driver to the NVMe device. As this binds the driver to a device, it has to be run with root privileges. By changing the owner of the VFIO container to an unprivileged user, the driver can use the VFIO driver to interact with the device without root. In Listing 4.3 the initialization script for the Vfiio driver is shown.

Listing 4.2: vfio-pci binding function in bash

```
function vfio_bind {
    nvme_address=$1
    device_path="/sys/bus/pci/devices/$nvme_address"

    echo "vfio-pci" > "$device_path/driver_override"

    if [[ -d $device_path ]]; then
        curr_driver=$(readlink $device_path/driver)
        curr_driver=${curr_driver##*/}

        if [[ $curr_driver != "vfio-pci" ]]; then
            echo $nvme_address > "$device_path/driver/unbind"
        fi
    fi
}
```

```
    echo $nvme_address > /sys/bus/pci/drivers_probe  
}
```

Listing 4.3: Initializing vfio-pci using vfio_bind from Listing 4.2 and setting permissions

```
#!/bin/bash  
  
modprobe vfio-pci  
vfio_bind "0000:03:00.0"  
chown $user:$group /dev/vfio/*
```

1. Add VFIO kernel module using modprobe
2. Unbind the PCI device from other drivers and bind it to vfio-pci
3. Set VFIO group permissions to user/group using chown

4.2.4 IOMMU initialization

To initialize the IOMMU, we first need to get the container file descriptor. The container is accessible under the path `/dev/vfio/vfio`. Using the raw container file descriptor, we can use the following `ioctl` calls:

Listing 4.4: `ioctl` calls needed for IOMMU initialization

```
ioctl_unsafe!(container_fd, VFIO_GET_API_VERSION)  
ioctl_unsafe!(container_fd, VFIO_CHECK_EXTENSION, VFIO_TYPE1_IOMMU)  
ioctl_unsafe!(group_fd, VFIO_GROUP_GET_STATUS, &group_status)  
ioctl_unsafe!(group_fd, VFIO_GROUP_SET_CONTAINER, &container_fd)  
ioctl_unsafe!(container_fd, VFIO_SET_IOMMU, VFIO_TYPE1_IOMMU)  
ioctl_unsafe!(group_fd, VFIO_GROUP_GET_DEVICE_FD, pci_addr)  
ioctl_unsafe!(container_fd, VFIO_IOMMU_GET_INFO, &iommu_info)
```

Excluding the Status and Info calls, the functionality consists of initialising the IOMMU for the device groups by setting the container on the groups, enabling Type1 for the IOMMU and fetching the device file descriptor. With the device file descriptor, we gain access to the device regions through the VFIO device API, allowing us to `mmap` the NVMe BAR into memory. Using `VFIO_IOMMU_GET_INFO` we can see the supported pagesizes. As our IOMMU supports 4K, 2M and 1G pagesizes the field `iova_pgsizes` has the value `0x40201000`.

4.2.5 PCIe configuration

To enable DMA we need to set a bit in the PCIe device registers. This is done in the command register of the PCIe configuration space.

4.2.6 NVMe initialization

Using the Vfiio, the initialization process of the NVMe SSD works as followed.

1. Map the NVMe device memory into host memory using VFIO resource info.
2. Allocate Admin SQ, CQ and I/O SQ, CQ
3. Create a mapping on the IOMMU using VFIO
4. Configure the NVMe device
5. Pass I/O Queue addresses to NVMe device using admin queues

4.2.7 DMA (Un-)Mapping

In order to provide a section of memory on which the device can perform DMA operations, the user needs to either allocate some memory in the processes address space or map an existing file into the process virtual address space. This can be achieved by using the Linux syscall `mmap`. Using `mmap`'s flags we can also define the page size used. The `MAP_HUGETLB` flag is used in conjunction with the `MAP_HUGE_2MB` and `MAP_HUGE_1GB` flags for 2MiB and 1 GiB pages respectively. By default `mmap` uses the default page size of 4KiB. The main IOMMU work is done by then creating the map struct `vfio_iommu_type1_dma_map`. We set the DMA mapping to read and write, and provide the same IOVA as the Virtual address. By then passing it to an `ioctl` call with the according VFIO operation `VFIO_IOMMU_MAP_DMA` we can create a mapping in the page tables of the IOMMU. This way we can give the IOVA to the NVMe controller, which it will use to access the memory through the address translation of the IOMMU.

Listing 4.5: Mapping memory for DMA

```
let mut iommu_dma_map = vfio_iommu_type1_dma_map {
    argsz: mem::size_of::<vfio_iommu_type1_dma_map>() as u32,
    flags: IoctlFlag::VFIO_DMA_MAP_FLAG_READ
        | IoctlFlag::VFIO_DMA_MAP_FLAG_WRITE,
    vaddr: ptr as u64,
    iova: ptr as u64,
    size,
```

```
};

ioctl_unsafe!(
    *container_fd,
    IoctlOperation::VFIO_IOMMU_MAP_DMA,
    &mut iommu_dma_map
)?;

let iova = iommu_dma_map.iova as usize;
```

Unmapping DMA Unmapping DMA happens when the process exits, yet for performance and application reasons we implement the `unmap_dma` function which can be used to unmap a DMA. Using the `VFIO_IOMMU_UNMAP_DMA` `ioctl` operation we can unmap the memory, and finally free it by using `munmap`.

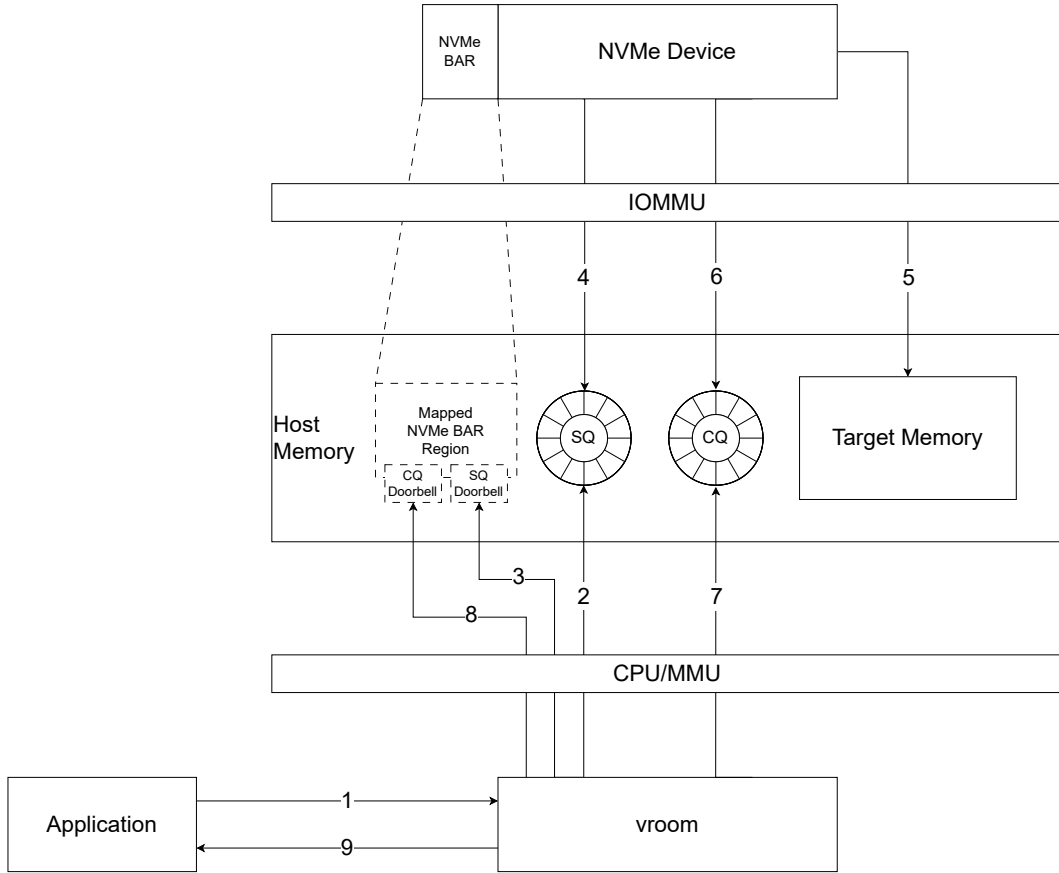


Figure 4.1: I/O operation using vroom with enabled IOMMU

4.2.8 I/O operations with VFIO

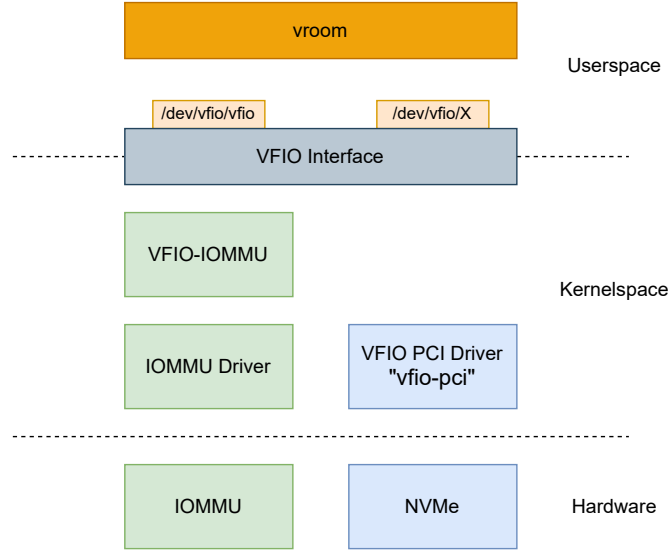
After initialization, the NVMe is ready to use. A sequential, single-threaded I/O operation is shown in Figure 4.1. The sequence of events are as followed:

1. **I/O function call**: The application calls a read/write method on vroom
2. **Command Submission**: Vroom creates a `NvmeCommand` struct and places it on the Submission Queue head.
3. **Ring SQ Doorbell**: Vroom places the submission queue head address in the doorbell register. The doorbell register is part of the NVMe BAR region, which is mapped to memory.

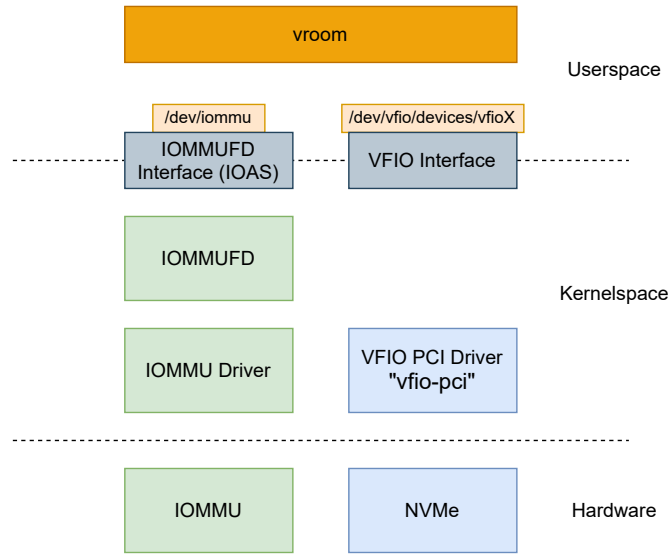
4. **Take Command:** The NVMe takes the command from the SQ.
5. **Perform I/O:** The NVMe uses the IOMMU to access the host memory via DMA and performs the read/write command.
6. **Complete I/O:** The NVMe places a `NvmeCompletion` struct instance on the head of the Completion Queue.
7. **Polled CQ:** By polling the CQ, vroom can process the CQ entry.
8. **Ring CQ Doorbell:** After processing the CQ entry, vroom rings the CQ Doorbell to notify the NVMe controller that the Completion Queue has been processed.
9. **Notify Application:** Vroom notifies the Application of the success of the I/O operation. The application can continue running.

4.3 IOMMUFD

The IOMMU File Descriptor user API (IOMMUFD) offers a way of controlling the IOMMU subsystem using file descriptors in user-space [7]. It replaces the VFIO IOMMU API using containers with the IOMMUFD API. It allows management of I/O address spaces (IOAS), enabling mapping user space memory on the IOMMU. IOMMUFD has only been recently added to the Linux Kernel in December 2022. E.g. Debian 12 does not include it, Fedora 40 does, but it is not enabled in the kernel configuration. Considering that it is not widely available or enabled on many distributions, our driver offers both options of using the IOMMU. The performance tests are done using the 'legacy' VFIO container way. The device file descriptor, which was previously attained with `VFIO_GROUP_GET_DEVICE_FD` can now be gotten through opening the character device `/dev/vfio/devices/vfioX`. By using this character device pointer we can claim the ownership over the VFIO device. That way VFIO does not rely on group/container/iommu drivers. I/O Address Spaces (IOAS) are the equivalent to a VFIO container in legacy VFIO.



(a) VFIO with Containers



(b) VFIO with IOMMUFD (IOAS)

Figure 4.2: Layer diagrams of VFIO with VFIO Container API and IOMMUFD, partly adopted from [24]

5 Evaluation

In this chapter, we analyse the performance impact of the IOMMU, directly comparing it to the physical address approach. To compare both approaches fairly, we don't include allocation and mapping times and perform them upfront. The focus lies on the IOMMU itself and how it performs with different page sizes. All performance tests use the Container IOMMU API as it currently remains the widely adopted VFIO variant. Due to the page size limitation of using physical addresses, we cannot compare the IOMMU to physical addresses using 4 KiB pages.

5.1 Setup

We use two systems to benchmark the driver's performance. Both systems run Ubuntu 23.10 with Linux kernel version 6.5.0-42. Despite the NVMe specifications maximum capability of 65536 I/O queues, our SSDs support a more reasonable amount of 128 I/O queues, which seems to be a typical amount. We use 1 thread per 1 I/O queue in our multithreaded tests. Turbowrite is a Samsung technology that drastically speeds up write latencies in the so-called "Turbowrite" buffer with the size of 42 GB of the NVMe, as shown in [16]. In order to use the Turbowrite NVMe to its maximum potential, most tests are conducted in said buffer to avoid the NVMe being the bottleneck instead of the IOMMU. As both systems are servers with NUMA architectures, we use `numactl -N 0 -m 0` for all tests to maintain consistency. We also disable performance scaling options. For the Intel CPU this is called `intel_pstate` and can be disabled using `echo "1" | sudo tee /sys/devices/system/cpu/intel_pstate/no_turbo`. The AMD equivalent can be disabled by using `echo "0" | sudo tee /sys/devices/system/cpu/cpufreq/boost`. The NVMe SSDs are formatted to 512-byte blocksize. Additionally, all standard tests are run with the `iommu=nomerge` kernel parameter to prevent the IOMMU from merging entries together, which can influence the results. Even though the driver can run without root, we use the root user to run the tests to avoid Linux' ulimits.

As Linux as well as our IOMMU supports 4 KiB, 2 MiB and 1 GiB page sizes we will test and analyse how it affects the overall performance.

CPU	Memory	NVMe	Capacity	Count
Intel Xeon E5-2660v2	251 GiB	Samsung Evo 970 Plus	1 TB	1
AMD EPYC 7713	1007 GiB	Samsung PM9A3	1.92 TB	8

Table 5.1: Specifications of systems used in performance testing

CPU	Clock	Cores	Virtualization	Year
Intel Xeon E5-2660v2	2.2 GHz	10	VT-d	2012
AMD EPYC 7713	2.0 GHz	64	AMD-V	2021

Table 5.2: CPUs of the systems

NVMe	Maximum Queue Count	Maximum Queue Size	Turbowrite	Usage
Samsung Evo 970 Plus	128	16384	Yes	Consumer
Samsung PM9A3	128	16384	No	Enterprise

Table 5.3: NVMe(s) of the systems

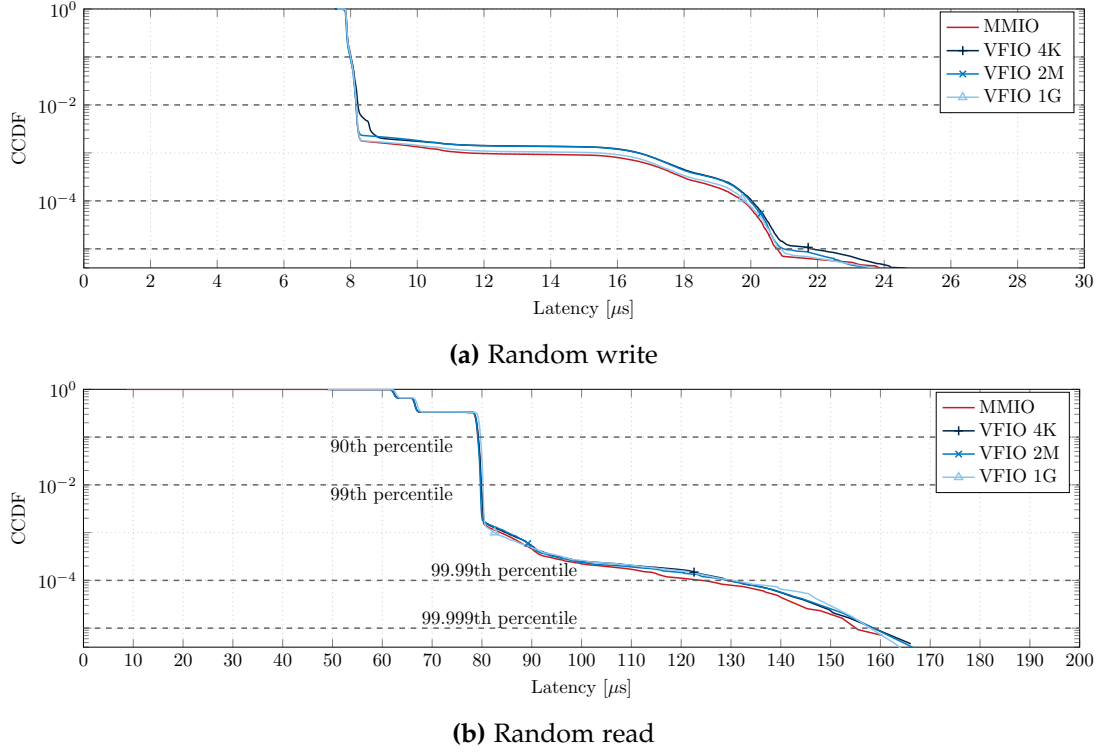


Figure 5.1: Tail latencies

5.2 Overall Latency and Throughput

First, we will compare the VFIO implementation to the MMIO implementation using latency and throughput tests. In these tests, we repeatedly write/read from a 4 KiB buffer in the memory. Each I/O operation uses a 4 KiB unit size.

In these tests, we can see that there is practically a negligible amount of overhead. As this test only uses a one page buffer at maximum, the buffer is constantly stored in the IOTLB.

This test uses one buffer from which the NVMe driver reads/writes to. This buffer and the Queues can fit on the IOTLB. Fetching addresses from the IOTLB is very efficient and thus, no significant performance impact occurs.

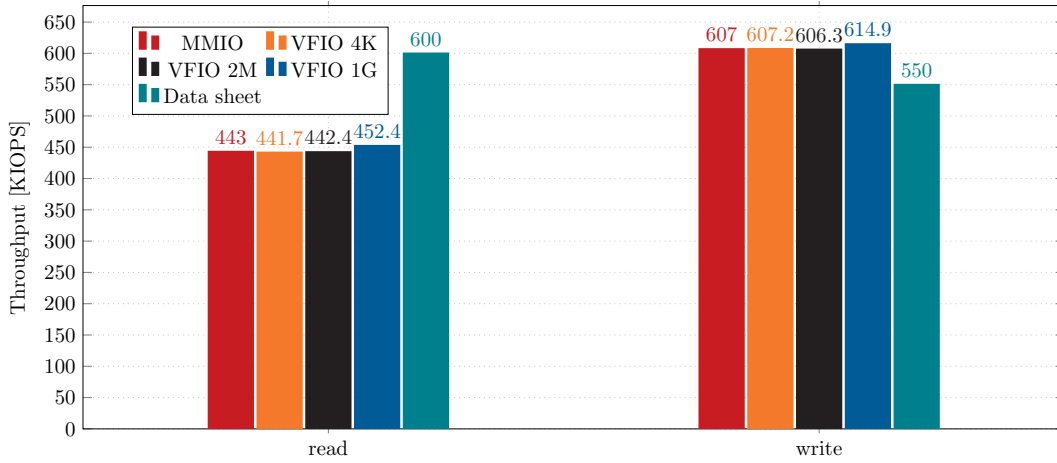


Figure 5.2: Write and Read Operations with queue depth 32 and 4 threads

5.3 Determining IOTLB size

As the size of the IOTLB is not stated in hardware and VT-d specifications, we use a latency test to analyse the behaviour of the IOMMU. We can assume that the IOTLB entry count must be a power of two. In order to isolate the effect of the IOMMU we aim to analyse the latencies of the fastest operation the NVMe can perform. The fastest operation is using random writes with the smallest blocksize of 512 B on an empty NVMe drive. By then repeating this random write 4096 times on page counts that are a power of two, taking the median and plotting it, we can figure out where a latency spike occurs. This power of two right before the latency spike should equal the IOTLB entry count. We configure the queues, buffer and prp-list to each take up one page, resulting in 6 allocated pages before the actual workload. This test is done using without the IOMMU with 2 MiB pages and the IOMMU with 4 KiB, 2 MiB and 1 GiB pages.

Results of Intel Xeon In the resulting graph Figure 5.3 we can observe a performance spike of around 300 nanoseconds for each write between 128 and 256 allocated pages. In the case of 4 KiB pages, this is a memory size of only 512 KiB. Using this information, we can assume that the IOTLB has the same size for each pagesize, as well as it being 128 entries of size.

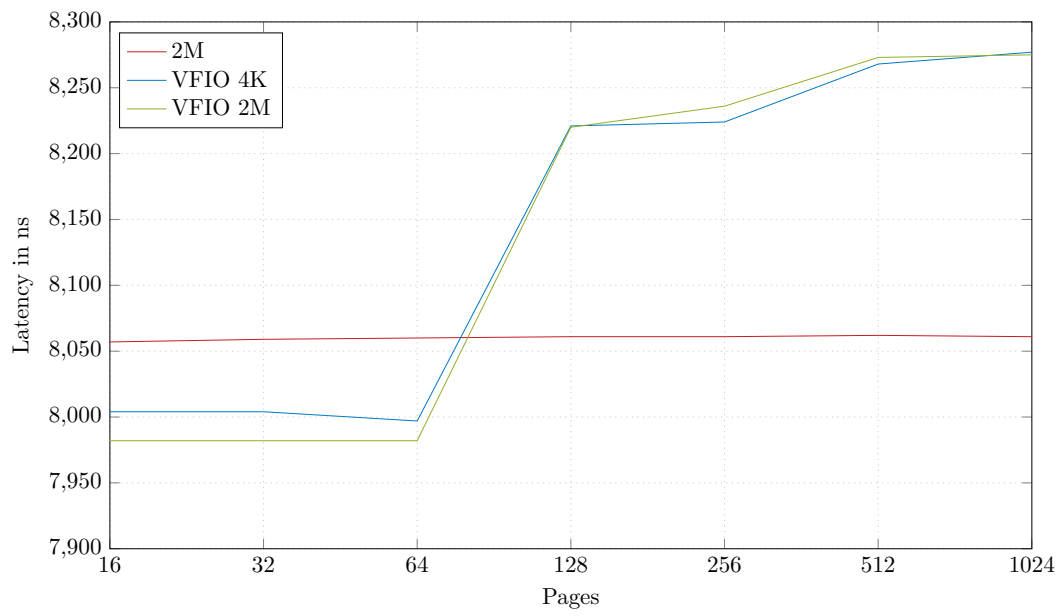


Figure 5.3: Latencies of random writes on an emptied SSD with increasing host memory pages on the Intel system

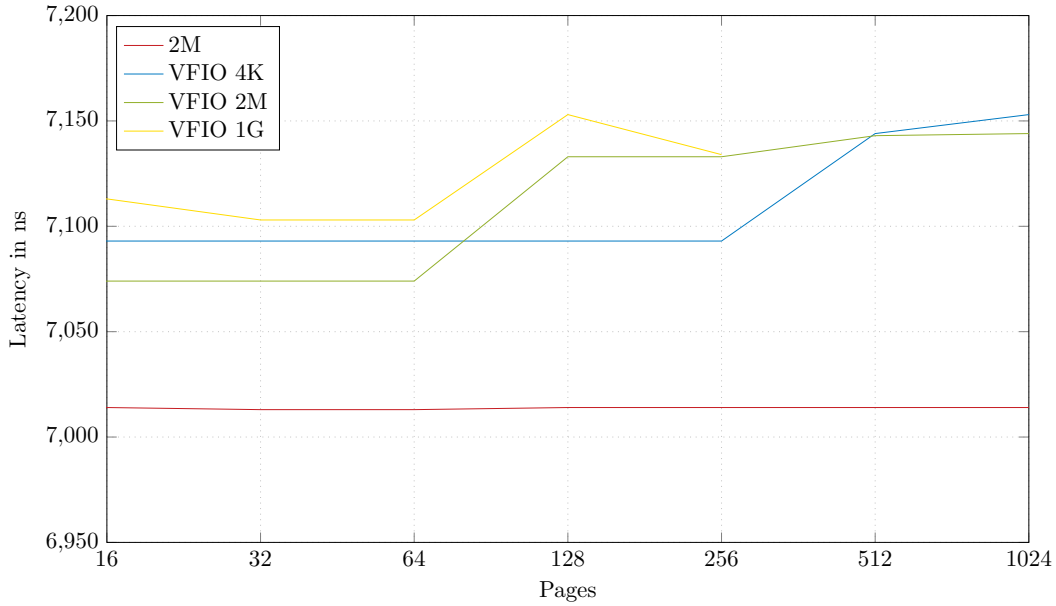
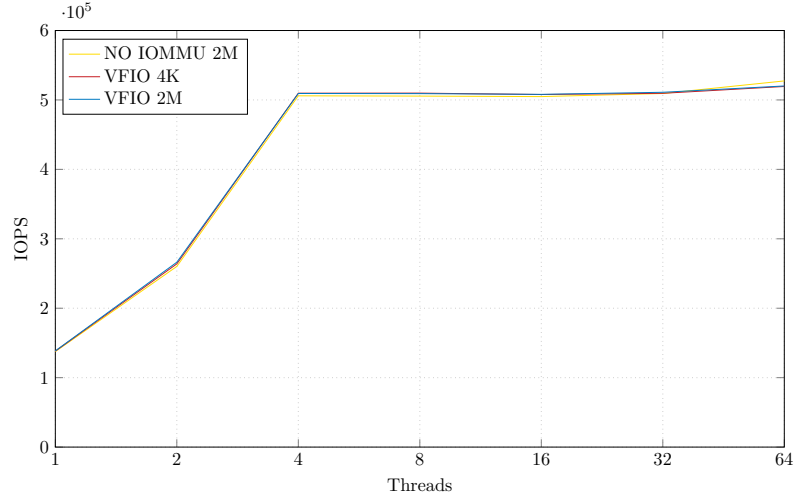


Figure 5.4: Latencies of random writes on an emptied SSD with increasing host memory pages on the AMD system

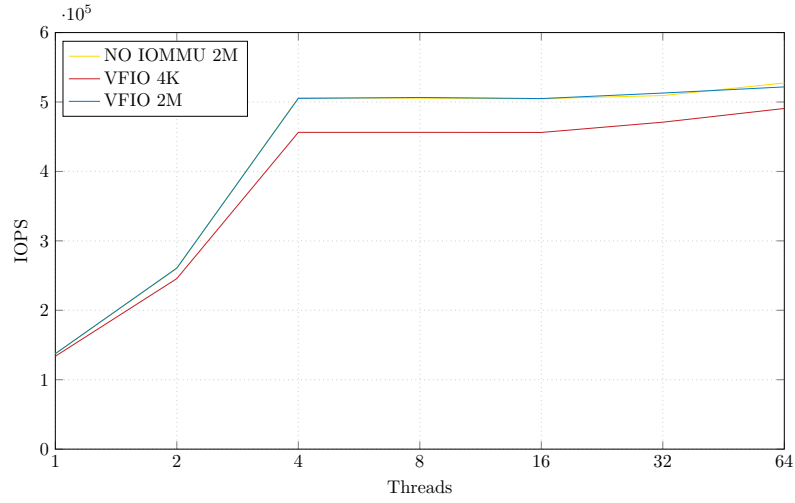
Results of AMD Epyc On the AMD IOMMU, we can see a performance spike that occurs at 64-128 pages for 2 MiB and 1 GiB page sizes and at 256-512 pages. We can therefore assume that the IOTLB size depends on the pagesize unlike on the Intel CPU. The performance itself only decreases by about 60 ns, which is a five fold performance increase of page walks compared to the intel cpu system.

5.4 Multithreaded Random Writes

When using multithreaded random writes with queue depth 1, in Figure 5.5 we can mostly see the same performance for 1 page, until 32 threads. This is because we allocate a buffer for each thread. For each thread, the submission queue, the completion queue and the buffer take up each one entry of the IOTLB. Therefore, with 32 threads we use 96 pages, exceeding the 64 entries we can store in the IOTLB. Using a 2 MiB buffer, we see around 10% less performance using 4 KiB pages, as even with one thread, the page count exceeds the possible entries in the IOTLB. The performance of the 2 MiB Vmio implementation stays the same as with 1 4 KiB buffer, as the 2 MiB buffer fits in one IOTLB entry.



(a) 1 4 KiB buffer per thread



(b) 1 2 MiB buffer per thread

Figure 5.5: QD1 Random Write Throughput with multithreading

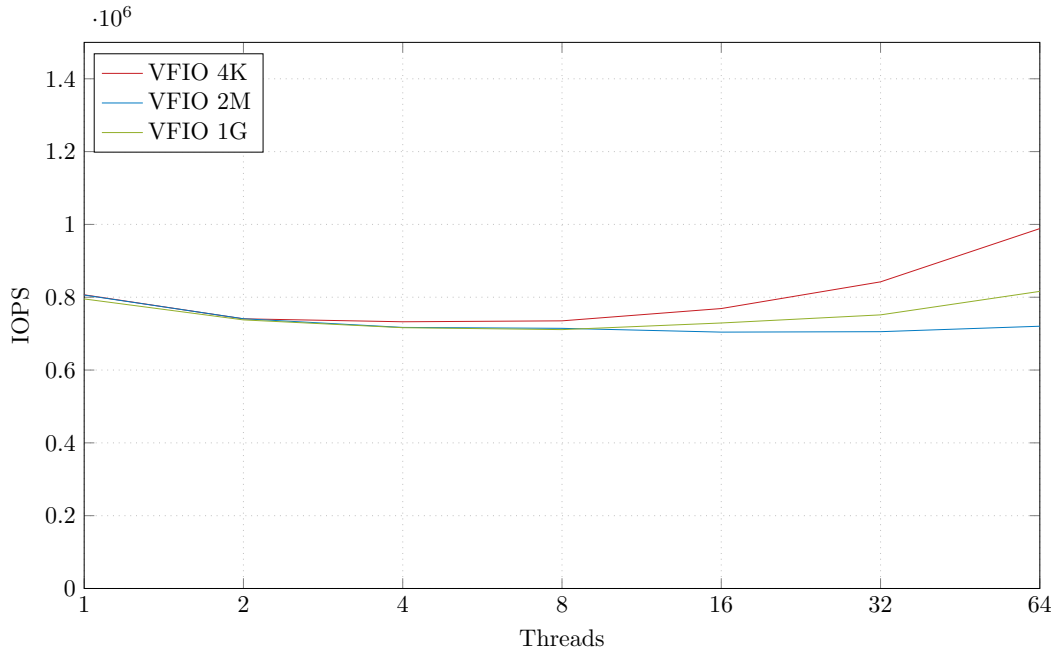


Figure 5.6: Latencies of random writes on an emptied SSD with queue depth 32 and increasing amount of threads

5.5 Increasing Queue Depth

To further test performance we can increase the queue depth. This metric describes how many outstanding commands can be placed on each queue.

5.6 Using multiple SSDs

All tests with multiple SSDs are run on the system with the AMD CPU.

5.7 IOMMU modes

There are a couple of kernel parameters that can be set at boot-time to influence the behaviour of the IOMMU. The availability is influenced by the iommu manufacturer, e.g. there's `amd_iommu` and `intel_iommu`, as well as the CPU architecture. Many of these manufacturer dependent options are either very specific, or shared behaviour is ported

to the general iommu parameter, e.g. `amd_iommu=fullflush` and `intel_iommu=strict`. We will be mainly looking at the general options `iommu`.

Strict To enable strict IOMMU mode, `iommu.strict=1` has to be set. Using strict mode, unmapping operation cause a complete IOTLB invalidation. Using relaxed mode, unmapping operations can be deferred and batched. This increases performance as an invalidation completely clears the IOTLB buffer, but reduces device isolation.

Passthrough Passthrough mode can be enabled using `iommu.passthrough=1`. Using passthrough DMA operations bypass the translation of the IOMMU, and instead directly access physical memory.

6 Conclusion

IOMMU In this thesis, we improved vroom’s safety by implementing IOMMU support and come to the same conclusion as SPDK. The advantages of using the VFIO such as access rights and bigger address spaces as well as the ability to run the driver without root privileges outweigh the small performance impact that can be registered in niche cases. Considering that IOMMU technology has seen a rise in popularity in the use of hardware passthrough for virtualization it is also likely that in the future the IOMMU performance and the IOTLB size will increase, further closing the gap. The ability to improve security drastically and increasing address space, while not compromising on performance are the reason the MMU succeeded, and it is likely that the IOMMU will as well.

Rust in driver development The viability of using Rust to develop drivers has been shown oftentimes, and it has proved that a modern, memory-safe language like Rust can compete with C in systems development. Using Rust does not only provide more safety, but also a modern ecosystem, a package manager and zero cost abstractions. Using Rust for drivers ensures in-process memory safety.

Future Work Future Work on the driver could include expanding the NVMe capabilities. Currently the driver is fixed to one namespace. Furthermore, the driver does not support a block device layer and a file system. Including sysfs support could also be a next step. To further push the throughput it could be investigated if and how many threads could operate on one I/O queue.

List of Figures

2.1	MMU and IOMMU relation to physical memory, adapted from [15] . .	4
2.2	VT-d Paging structure for translating a 48-bit address to a 4 KiB and a 2 MiB page, grabbed from [11]	5
2.3	segmented PCI identifier	7
4.1	I/O operation using vroom with enabled IOMMU	16
4.2	Layer diagrams of VFIO with VFIO Container API and IOMMUFD, partly adopted from [24]	18
5.1	Tail latencies	21
5.2	Write and Read Operations with queue depth 32 and 4 threads	22
5.3	Latencies of random writes on an emptied SSD with increasing host memory pages on the Intel system	23
5.4	Latencies of random writes on an emptied SSD with increasing host memory pages on the AMD system	24
5.5	QD1 Random Write Throughput with multithreading	25
5.6	Latencies of random writes on an emptied SSD with queue depth 32 and increasing amount of threads	26

List of Tables

5.1	Specifications of systems used in performance testing	20
5.2	CPUs of the systems	20
5.3	NVMe(s) of the systems	20

Listings

4.1	Syscall <code>mmap</code> macro, with own error variant	10
4.2	<code>vfio-pci</code> binding function in bash	12
4.3	Initializing <code>vfio-pci</code> using <code>vfio_bind</code> from Listing 4.2 and setting permissions	13
4.4	<code>ioctl</code> calls needed for IOMMU initialization	13
4.5	Mapping memory for DMA	14

Bibliography

- [1] *About DPDK*. DPDK. URL: <https://www.dpdk.org/about/> (visited on 08/08/2024).
- [2] M. Ben-Yehuda, J. Xenidis, M. Ostrowski, K. Rister, A. Bruemmer, and L. van Doorn. "The price of safety: Evaluating IOMMU performance." In: *Ottawa Linux Symposium (OLS)* (Jan. 2007), p. 13.
- [3] *Crate bindgen*. URL: <https://docs.rs/bindgen/0.69.4/bindgen/> (visited on 07/22/2024).
- [4] *Direct Memory Access (DMA) From User Space*. URL: <https://spdk.io/doc/memory.html> (visited on 07/22/2024).
- [5] L. Doan and M. Day. "CrowdStrike Crash Affected 8.5 Million Microsoft Windows Devices." In: *Bloomberg* (July 20, 2024). URL: <https://www.bloomberg.com/news/articles/2024-07-20/crowdstrike-crash-affected-8-5-million-microsoft-windows-devices> (visited on 07/23/2024).
- [6] *External Technical Root Cause Analysis — Channel File 291*. CrowdStrike, Aug. 6, 2024. URL: <https://www.crowdstrike.com/wp-content/uploads/2024/08/Channel-File-291-Incident-Root-Cause-Analysis-08.06.2024.pdf> (visited on 08/08/2024).
- [7] J. Gunthorpe and K. Tian. *IOMMUFD*. URL: <https://docs.kernel.org/userspace-api/iommufd.html> (visited on 07/08/2024).
- [8] S. Huber. "Using the IOMMU for Safe and Secure User Space Network Drivers." MA thesis. Technical University of Munich, 2019.
- [9] *HugeTLB Pages*. URL: <https://docs.kernel.org/admin-guide/mm/hugetlbpage.html> (visited on 07/25/2024).
- [10] Intel. *80286 Microprocessor with memory management and protection*. Sept. 1993. URL: <https://datasheets.chipdb.org/Intel/x86/286/datashts/210253-016.pdf> (visited on 07/23/2024).
- [11] Intel. *Intel Virtualization Technology for Directed I/O Architecture Specification Revision 4.1*. Mar. 22, 2023. URL: <https://www.intel.com/content/www/us/en/content-details/774206/intel-virtualization-technology-for-directed-i-o-architecture-specification.html> (visited on 08/02/2024).

- [12] *mmap(2) - Linux manual page*. URL: <https://man7.org/linux/man-pages/man2/mmap.2.html> (visited on 08/02/2024).
- [13] *Null References: The Billion Dollar Mistake*. URL: <https://www.infoq.com/presentations/Null-References-The-Billion-Dollar-Mistake-Tony-Hoare/> (visited on 07/23/2024).
- [14] *numactl(8) — Linux manual page*. URL: <https://man7.org/linux/man-pages/man8/numactl.8.html> (visited on 08/09/2024).
- [15] O. Peleg and A. Morrison. *Utilizing the IOMMU Scalably*. USENIX ATC '15. 2015. URL: https://www.youtube.com/watch?v=kL0Roes_cy0 (visited on 08/06/2024).
- [16] T. Pirhonen. "Writing an NVMe Driver in Rust." Bachelor's thesis. Technical University of Munich, 2024.
- [17] L. Proven. "Linux 6.1: Rust to hit mainline kernel." In: *The Register* (Oct. 5, 2022). URL: https://www.theregister.com/2022/10/05/rust_kernel_pull_request_pulled/ (visited on 08/02/2024).
- [18] D. Rovella. "Tech Meltdown Collapses Systems Worldwide." In: *Bloomberg* (July 20, 2024). URL: <https://www.bloomberg.com/news/newsletters/2024-07-19/bloomberg-evening-briefing-tech-meltdown-collapses-systems-worldwide> (visited on 07/23/2024).
- [19] K. Sarcar. *What is NUMA?* URL: <https://docs.kernel.org/mm/numa.html> (visited on 08/09/2024).
- [20] *Storage performance Development Kit*. URL: <https://spdk.io/> (visited on 07/22/2024).
- [21] *Submitting I/O to an NVMe Device*. SPDK. URL: https://spdk.io/doc/nvme_spec.html (visited on 08/05/2024).
- [22] *Transparent Hugepage Support*. URL: <https://docs.kernel.org/admin-guide/mm/transhuge.html> (visited on 07/23/2024).
- [23] *VFIO - "Virtual Function I/O"*. URL: <https://docs.kernel.org/driver-api/vfio.html> (visited on 07/08/2024).
- [24] C. Xia and Y. Cao. *Introducing New VFIO and IOMMU Framework to DPDK*. DPDK Summit 2023. Sept. 13, 2023. URL: <https://www.youtube.com/watch?v=ZhIOHEv50e0> (visited on 08/02/2024).