探究虚拟化环境下 SPDK 存储设备的性能

实验目的

在这个实验中，我们将学习并实践多种在虚拟机中访问宿主机的存储设备的方法。通过亲身实践，你将会：

- 对常见存储介质的种类、性能指标和各项参数有基本的了解。
- 了解在 SSD 上进行性能基准测试的方法，理解各项性能指标的含义。
- 自行设计实验，理解并亲身实践在虚拟机中访问宿主机存储介质的各种方法，特别是基于 SPDK 的方案。
- 理解 SPDK、QEMU、设备驱动、Linux 内核在用户访问存储介质过程中所起到的作用，以及这些组件之间的相互关系。

整个实验如果顺利进行，大概只需要 3-6 个小时即可完成。其中 1-2 小时用于安装相关的环境，3-4 个小时用于在不同的配置下跑 benchmark。同时，我们提供了大量的参考资料帮助你更好地理解实验过程。当然不读这些参考资料也没什么问题，只要体验到实验的乐趣，**提供真实可信的数据和报告**即可。

Good luck and have fun!

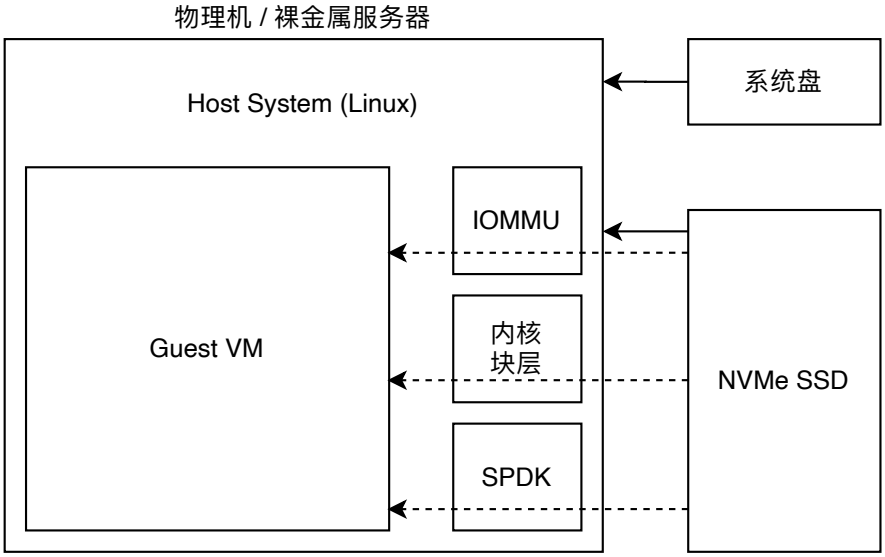
实验所需环境

推荐环境

我们推荐使用物理机或裸金属服务器进行实验。物理机/裸金属服务器应该：

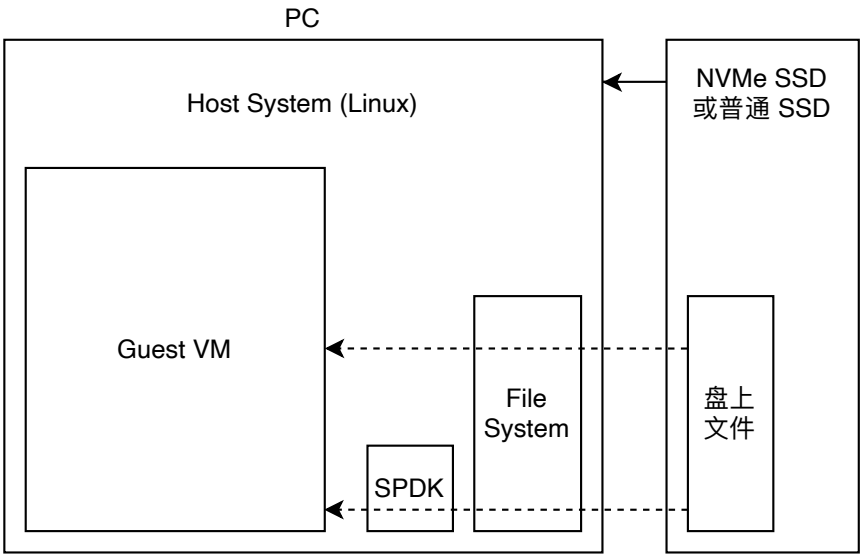
- 支持硬件虚拟化。（物理机、裸金属服务器当然应该支持吧）
  - 运行较新的 Linux 发行版。（内核版本 >= 5.10）
- 除系统盘外，有一块不含用户数据的 NVMe SSD 盘。（在实验过程中，数据盘上的信息将会丢失）
  - 可以直接插在主板的 PCIe M.2 接口上。
  - 也可以通过 Thunderbolt 外置 NVMe SSD 接入。
  - 也可以直接使用裸金属服务器提供的 NVMe SSD。
- 使用推荐环境测试的实验结果更有参考价值。
- 我们鼓励同学间互相借用实验环境、互相借用盘，但禁止抄袭或互相参考实验过程。如果你要向其他同学提供实验环境，请提前清理实验相关的文件、bash_history 等信息。

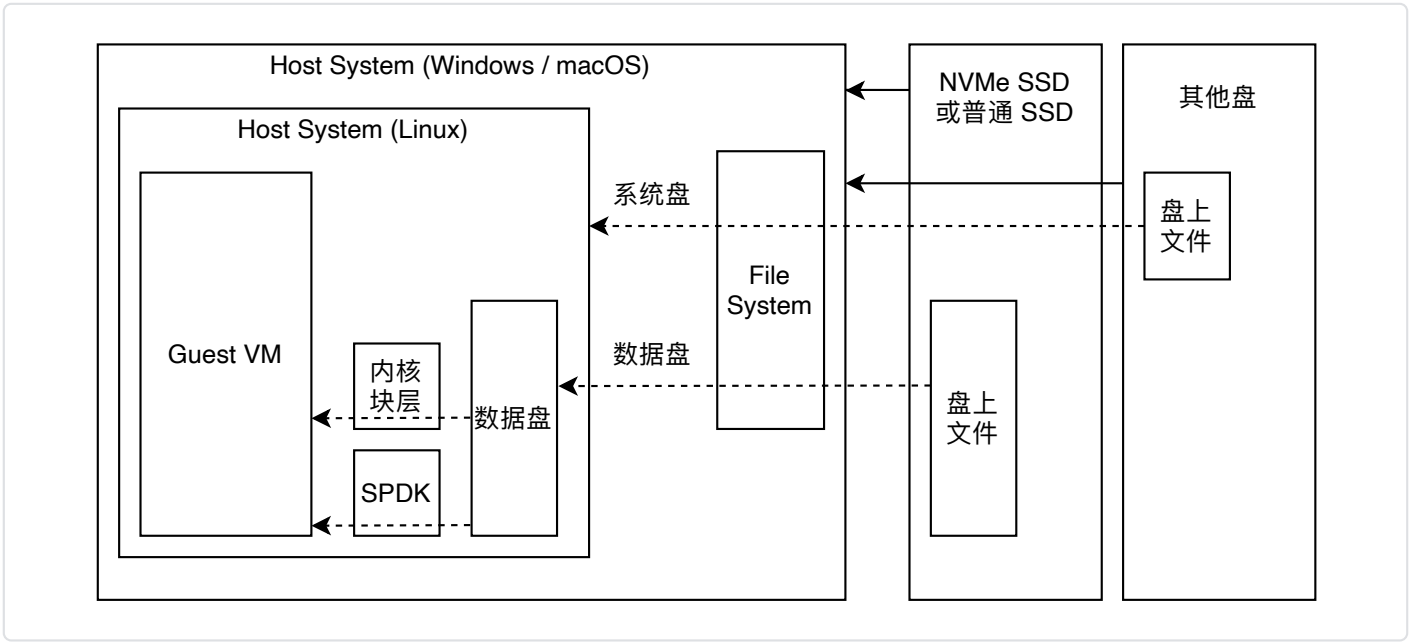
在实验报告的致谢部分，请明确说明你使用了哪位同学的环境。



最小要求环境

- 需要支持在 Linux 中开虚拟机。
  - 对于安装了双系统/直接运行 Linux 发行版的机器，需要支持硬件虚拟化。
  - 对于在虚拟机中开始实验的机器，需要支持嵌套虚拟化。
    - 运行 Linux 的第一层嵌套虚拟化机器的系统盘可以放在慢的盘上，比如机械盘。但之后运行基准测试的盘，最好使用 NVMe SSD 或普通 SSD。
- 使用最小要求环境测试的实验结果可能不太准确，有各种因素会影响测得的数据。当然，我们的实验以体验感受为主，对实验结果的精确性并没有强制要求。





其他建议软件硬件环境

- Ubuntu 20.04 (w/ Kernel >= 5.10)
- QEMU >= 5.2.0
- SPDK >= 20.07
- 内存 >= 16GB
- CPU >= 4 线程
- fio >= 3.27

没有验证过的实验环境

- 各种不支持硬件虚拟化的云主机
- 普通 SSD
- HDD 盘
- SPDK 内存设备

背景知识

工具介绍

fio

fio 是实验推荐的对盘做基准测试的工具。

默认配置下，fio 的输出如下图所示。

```
4K-RR-QD1J1: (groupid=0, jobs=1): err= 0: pid=1033: Sun Aug 15 03:35:55 2021
  read: IOPS=9899, BW=38.7MiB/s (40.5MB/s)(4640MiB/120001msec)
    clat (usec): min=11, max=405, avg=100.63, stdev=24.79
     lat (usec): min=11, max=405, avg=100.70, stdev=24.79
    clat percentiles (usec):
     |  1.00th=[   69],  5.00th=[   71], 10.00th=[   73], 20.00th=[   76],
     | 30.00th=[   86], 40.00th=[   90], 50.00th=[   93], 60.00th=[  108],
     | 70.00th=[  121], 80.00th=[  124], 90.00th=[  129], 95.00th=[  147],
     | 99.00th=[  165], 99.50th=[  167], 99.90th=[  176], 99.95th=[  186],
     | 99.99th=[  253]
   bw (  KiB/s): min=39300, max=39887, per=100.00%, avg=39604.39, stdev=123.24, samples=119
   iops        : min= 9825, max= 9971, avg=9900.97, stdev=30.81, samples=119
  lat (usec)   : 20=0.15%, 50=0.03%, 100=57.58%, 250=42.23%, 500=0.01%
  cpu          : usr=99.98%, sys=0.00%, ctx=771, majf=0, minf=3
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=1187955,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1
```

我们需要关注的就是其中的 latency 分布、bandwidth 和 IOPS。

一些注意事项

- 为了实验准确性，我们建议先把把盘先覆写两遍，保证 SSD 主控达到一个稳定的状态。你可以用 dd 或者 fio 完成这一步。
- fio 测试的 workload：randread 4K（推荐）或 randwrite 4K。**如果使用 randwrite 进行测试，建议在每次测试前都将盘写两遍，不然前后 n 次实验之间会互相影响，产生难以解释的数据。**设置 direct=1 不走内核 page cache。如果是在普通文件系统上运行，需要开 fsync。
- 关注指标：bandwidth，IOPS，latency 分布。
- 实验中需要测试不同 QD 下的性能。
  - QD=1 可以用来测试端到端延迟。
  - QD=32 可以用来测试设备最大能支撑的 throughput。
- 在对块设备测试时，我们推荐使用 io_uring 或 libaio engine 做测试。
- 盘的温度可能影响性能。可以用 smartctl 关注盘传感器的温度。在测试时，保持高温测试长期性能，或保持低温测试短时性能。
- fio 测试时看到明显不对劲的值应当舍弃
  - 消费级 NVMe 盘的顺序写入速度大多在 2GB/s 左右，随机 4K 写可以到 300MB/s 左右。6GB/s 的速度大多是打到了写缓存，**请尝试关闭缓存或仅测试随机读性能。**
  - 随机读写 QD1 IOPS 可以打到 50k 左右。

注：在这个实验中，你也可以使用其他可以反映存储介质性能的工具，比如：RocksDB 的

db_bench、在关系型数据库上运行 sysbench、或使用 filebench 等。

fio 可以用配置文件或命令行传参配置。在配置实验环境一章中，我们提供了多种参考的配置。

SPDK

全称 Storage Performance Development Kit。在这个实验中，我们使用 SPDK 的用户态 NVMe 驱动绕过内核块层，从而达到更高的 I/O 效率。

SPDK 接管 NVMe 设备后，可以向上层提供多种接口。在这个实验中，我们会将 SPDK 暴露的 vhost target 交给 QEMU 中的 VM 使用。

如果实验中没有 NVMe 盘可以用，也可以使用 SPDK 的 aio bdev 或 uring bdev 进行实验。使用这两个 bdev 就无法绕过内核块层了。

QEMU

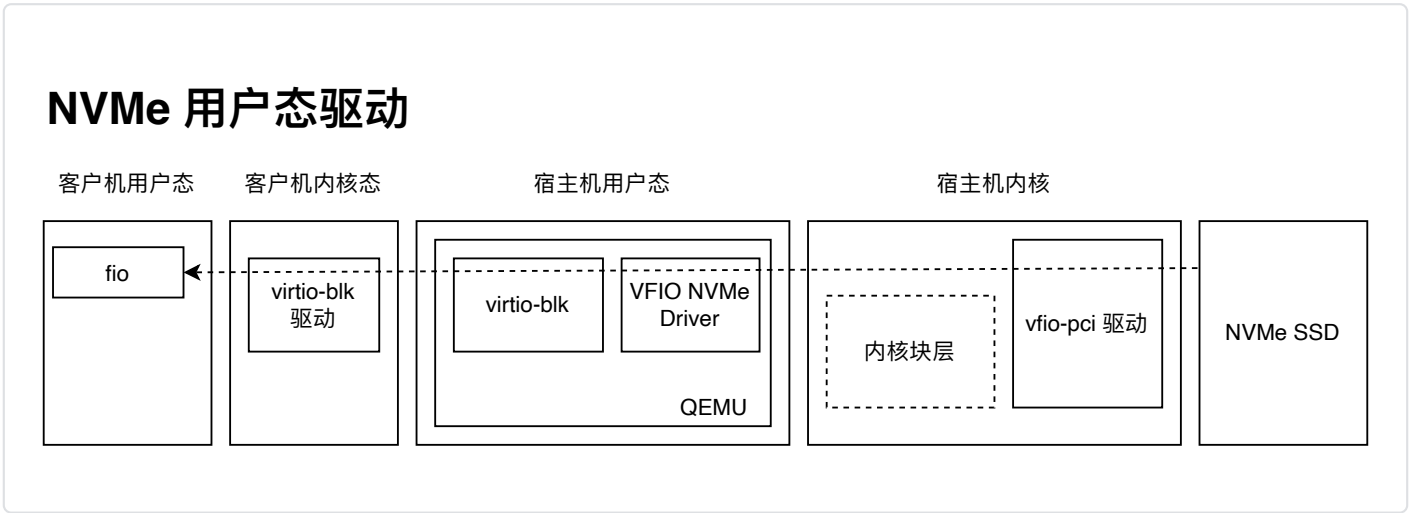
QEMU 是一个 VMM，可以用来运行虚拟机。在这个实验中，QEMU 会通过多种不同的接口访问宿主机的存储设备。

在这个实验中，也可以使用 libvirt 等工具管理虚拟机。

下面我们将简要介绍实验中可能用到的模式，和 QEMU 对应的启动参数。

NVMe 用户态驱动

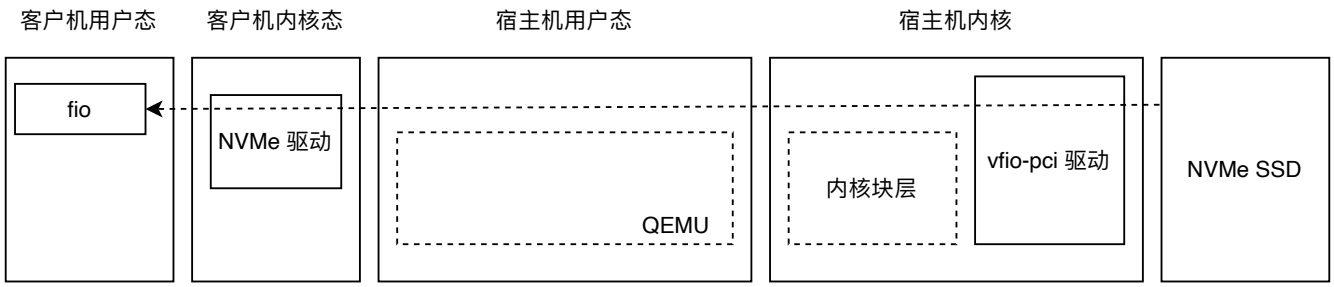
`--drive file=nvme://0000:03:00.0/1`



IOMMU

`--device vfio-pci,host=03:00.0`

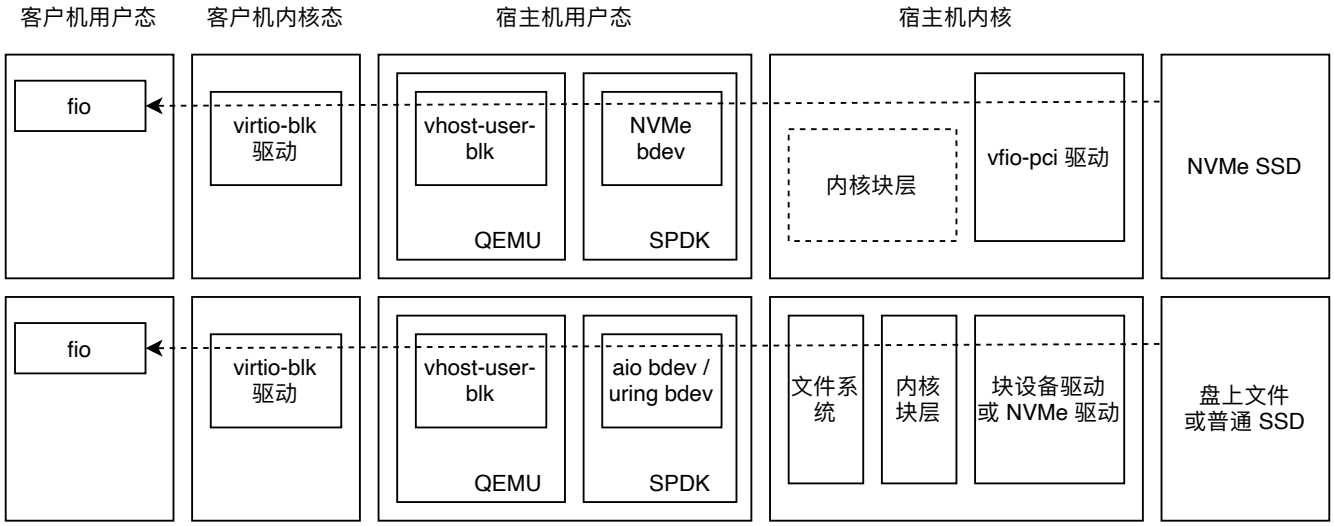
IOMMU



SPDK 方案

```
-device vhost-user-blk-pci,chardev=vhost0,num-queues=2
```

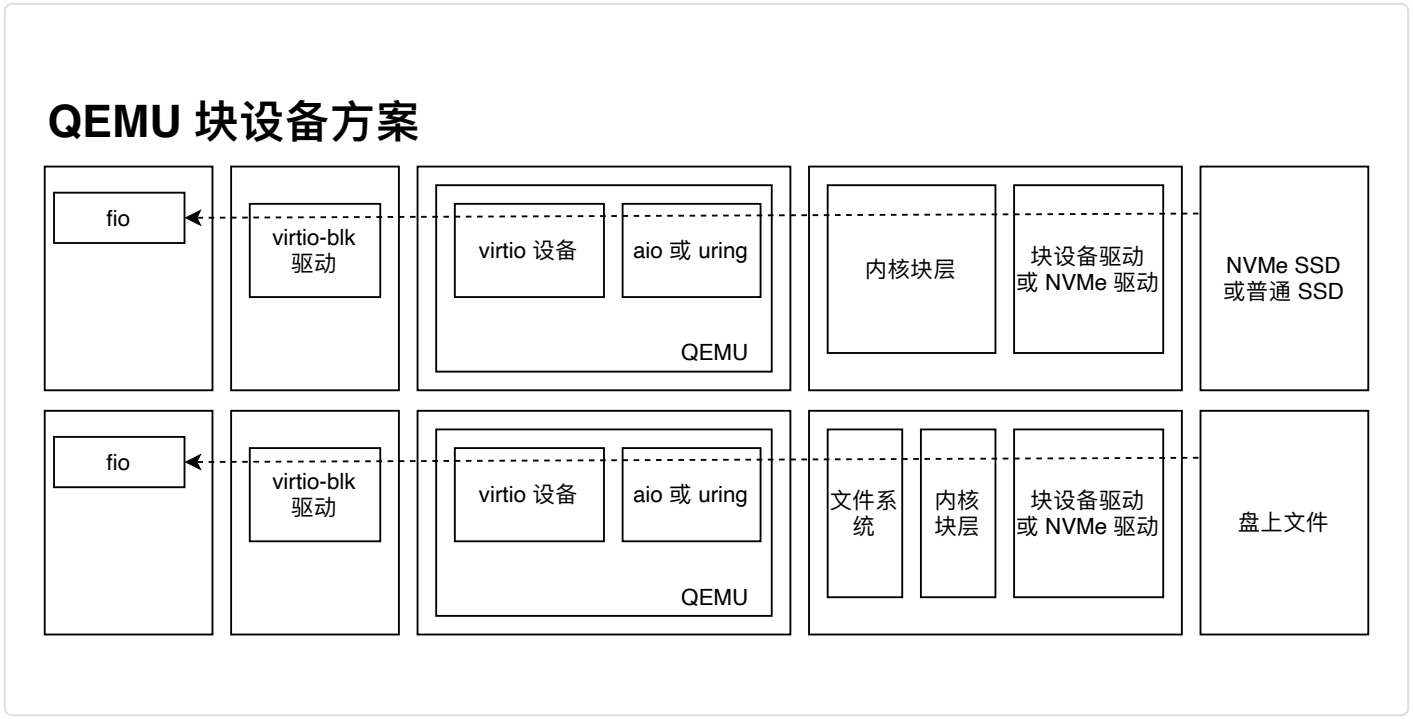
SPDK



* vhost-user-blk 对应 virtio-blk 驱动
* vhost-user-scsi 对应 virtio-scsi 驱动（实验设计时未测试）
* vhost-user-nvme 对应 NVMe 驱动（需要 SPDK 自己 fork 的 QEMU）

QEMU 原生方案

```
--drive file=/dev/nvme1n1,if=virtio,format=raw
```



配置实验环境

每个实验环境中的配置方法都有所不同。在这里我们提供两种不同的配置方式：

- 裸金属服务器 (NVMe)
- 物理机 (NVMe)

这个文档中，我们只粗略描述大概的配置步骤。详细每一步的操作请到各个配置文档中阅读。

- 安装软件：包括 QEMU、NVMe 管理工具、编译工具链
- 调整内核参数，开启 IOMMU 和内核大页
- 如果 SSD 的 IOMMU Group 有其他设备，打 ACS Override 补丁。
- 确认系统配置：确认机器支持硬件虚拟化、查看 CPU 拓扑结构（用于绑定核心）、查看 NVMe 盘的 PCIe 地址。
- 配置 NVMe 盘：确认没有挂载文件系统，关闭内核调度器，关闭 SSD 内部写缓存（如果有），把盘写两遍。
- 安装实验软件：fio 和 SPDK。
- 下载镜像并启动虚拟机。

实验设计指导与要求

- 记录所使用的环境。
- 在宿主机上使用 fio 测试性能，用于和客户机性能对比。

    - 可以直接访问块设备 `/dev/nvmexnx`
    - 也可以使用 SPDK plugin
    - 运行任何实验前，请仔细检查写入位置是否正确，千万不要把自己的数据（系统盘）写坏。
- 在客户机上自行选择 2 种方法测试，必须包括一种 SPDK 的方案。
    - 非 SPDK 方案包括
        - QEMU 块设备
        - passthrough
        - NVMe userspace driver
    - SPDK 方案包括
        - vhost-user-scsi
        - vhost-user-blk

	IOMMU + vfio-pci	块设备 + virtio-blk / virtio-scsi	NVMe + NVMe userspace driver	SPDK + vhost-user-scsi / vhost-user-blk-pci
QEMU 支持?	✅	✅	✅	✅
技术成熟度	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐
预估实验难度	⭐⭐	⭐	⭐⭐⭐⭐	⭐⭐⭐⭐

- 运行 benchmark 记录实验结果，分析产生性能差异的原因。如果使用 fio 做 benchmark。建议分别测试 QD=1 **和** QD=32 情况下 randomread 4K **或** randomwrite 4K 的 P50, P95, P99, P999, MAX latency，IOPS 和 bandwidth。使用其他 benchmark 工具也请从 latency 分布、throughput 两个维度评估效果。

思考与练习

- 关于实验过程的分析 **（三选一）**
    - 简述你设计的实验中可能存在的问题，可能产生数据误差的地方，并给出可能的解决方案和效果。
    - 如果有不合常理的数据点，请叙述你期望的结果，并分析为什么结果不合常理。
    - 简述实验中碰到的困难，和解决的方案。
- 关于 SPDK 和 Linux 内核 I/O 栈的理解 **（三选一）**

    - 简述不同挂载方法的区别与优劣。
    - 任选三组（或更多）实验，探究从用户态到盘上经过了哪些软件/硬件队列，并猜测（或阅读资料确认）每个队列的参数（如支持的队列深度、队列个数、队列中 I/O 请求的存在形式）。
        - 举例：使用 fio io_uring engine 时，fio 和 VM kernel 之间有一个 io_uring 队列。队列深度由 fio 的用户配置决定，队列个数和 fio numjobs 相关。队列中的 I/O 请求以系统调用的方式存放，如 pread + 位置 + 长度。
    - 尽可能地列出设备从盘到 fio 的可能硬件软件组合。
        - 举例：NVMe 盘 -> SPDK NVMe BDev -> vhost-user-blk -> virtio-blk 是一种组合，SATA SSD -> SCSI 驱动 -> 文件系统 -> QEMU -> virtio-blk 是另一种组合
- 关于基准测试的性能理解
    - 在基准测试中，latency 分布、throughput 的大小是否有数值上的关联？如果有，定性分析它们之间可能的关系。如果没有，简述理由。

拓展部分

这一部分的内容只是实验设计者想的 idea，**没有对下面 idea 的可行性做验证。**
如果你感兴趣，可以尝试下面的内容，并在实验报告中附上相关的内容。**拓展部分不会作为评分的依据。**

- 尝试热迁移虚拟机的同时热迁移盘。
- 盘的温度可能对实验结果产生影响。试着跑一个长时间的测试 (>= 10min)，用 smartctl 观测盘的温度，并使用工具画出盘的性能随时间变化的情况。消费级的盘容易达到比较高的温度。https://github.com/louwrentius/fio-plot
- 根据 fio 结果的 Histogram 绘制 latency 热力图，观察对比不同场景下的 latency 分布。如果绘制了热力图，则无须用表格记录 P95, P99 等 latency 信息。下面是一个热力图的例子：

- 尝试搭建 SPDK 的 NVMe-oF 环境，并进行测试。
- 尝试使用 vhost-nvme 做测试。
- 尝试 NVMe MDev https://www.usenix.org/conference/atc18/presentation/peng
- 尝试关掉 Specter / Meltdown 做对比测试。
- 比较不同方式访问存储设备的 CPU 占用情况。内核/用户/irq 分别占了多少比例?
- 尝试使用多个虚拟机访问同一块盘，验证哪些挂载方法支持多租户。
- 在实验手册中，我们不建议做 randwrite 的实验；即使要做，也需要在每次实验前填充两次盘。试做实验探究写入次数和读性能之间的关系，分析对 SSD pre-condition 的必要

性。比如每随机写 100G 后测试随机读性能。

提交要求

- 成功提交，包含需要的提交材料 5 分
  - 提交内容需要包括一份实验报告“学号_姓名_report.{docx/doc/pdf}”，和一个（可选）包含截图和 log 的文件夹“学号_姓名_data”。
  - 如果使用在线文档提交，文档标题命名为“学号_姓名_report”，文档附录内需要有所有要求的截图。
- 原始实验数据记录 5 分
  - 所有实验相关运行的指令（包括 qemu 启动指令、fio 启动指令、如果有用到 libvirt 则提供配置过程等）**都要带时间截图**，截图需另外打包提交或出现在实验报告的附录部分。
    - 如果使用配置文件或脚本启动 QEMU / fio，请在附录中附上**每一次运行的**配置文件和脚本，**无需提供截图。**
    - 如果运行次数较多，且脚本/配置文件内容几乎不变，可以简要叙述每次配置的不同之处。
  - 所有出现的数据应当有对应的**带时间截图**（包括 fio 运行结果等），截图可以另外打包提交，也可以出现在实验报告的附录部分。
    - 如果实验结果是直接输出到文件而非控制台，请在附录中附上所有的 log 文件，**无需提供截图。**
  - 编译、安装软件过程无需截图。
  - 原始数据会被用于评估实验的真实性。请不要篡改任何实验的运行结果。
  - 拓展部分不受此限制，不需要对结果截图，不需要原始 log 文件，也不会作为评分依据。但我们仍会根据提交的材料做一些评估，**请不要在实验的任何部分弄虚作假，不论是否计分**。
- 实验报告 5 分
  - 主体部分不超过 5 页。附录、拓展部分不受此限制，也不会（除原始数据真实性以外）作为评分依据。
  - **禁止抄袭他人的实验数据、实验内容。禁止不加标注地引用外部文字。禁止重复本实验文档和实验附加材料已经详细介绍的内容。**
  - 主体部分需要包含标题、实验设计、实验数据、实验分析、思考与练习、致谢（可选）、参考文献/资料这几个部分。
    - 标题和个人信息：《探究虚拟化环境下存储设备的性能》或其他标题，并附上姓名学号邮箱。
    - 实验设计：所使用的环境 (物理机/裸金属服务器，盘的型号，内存大小，CPU 线

程数等信息)+测试的方法指标参数+数据从盘到 fio 的路径。
    - 实验数据：表格形式记录不同场景下的各项指标。（可选）能绘制成图进行比较最好。
    - 实验分析：准确总结数据的规律与趋势。根据数据流过的路径，分析不同方法的性能为什么有差异。
    - 思考与练习：简单回答几个问题。
    - 致谢：可选。如果借用了其他同学的环境或和其他同学进行了答疑/交流，可以在这里说明。
    - 参考文献/资料：列出参考过的文献、网页即可。
  - 无需出现具体运行的指令。可以简要介绍运行的指令、参数等信息。其他内容在截图中体现即可。
  - （非必需）可以参考我们提供的模版。
- 实验分析+思考与练习 10 分
  - 有理有据，言简意赅。
  - 实验分析准确表述了实验结果的规律与趋势，并分析了不同实验之间结果区别背后的原因。
  - 思考与练习没有标准答案，有理有据、基本理解正确即可。
- 加分项
  - 解决同学遇到的问题。（若能总结成文档或加入本实验手册最好）
  - 发现实验手册的问题，提出自己的见解。（可以直接在这个文档上加评论）
- 总分 25，计入总成绩 20 分，超过 20 分按 20 分记。