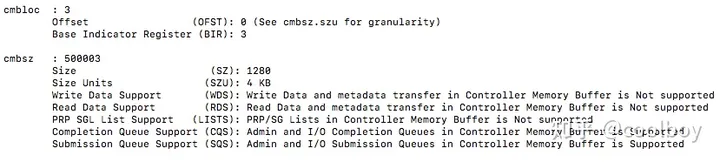
# NVME CMB原理和常规使用方案详解

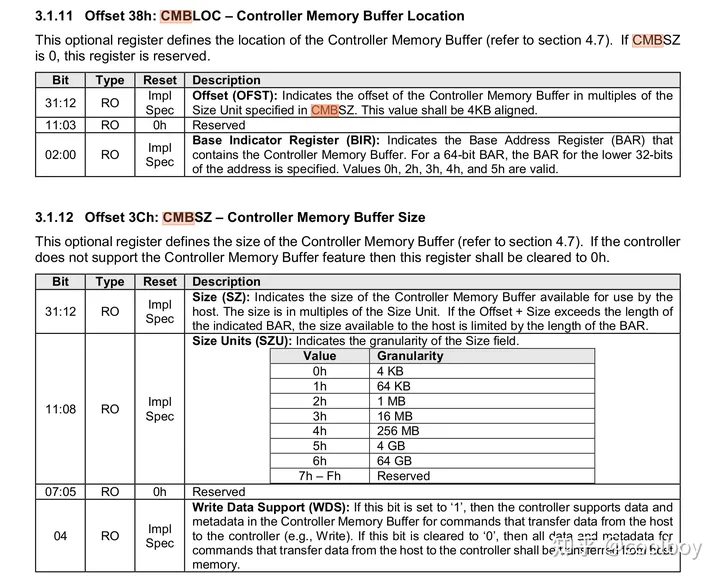
## 什么是CMB

在NVMe Express 1.2 Spec中开始支持一个特性，那就是CMB（Controller Memory Buffer），是指SSD控制器内部的读写存储缓冲区，与HMB（Host Memory Buffer）的不同处在于所使用的内存地址位于控制器自己的内存中，而不是位于主机内存中，但它们使用队列的方式都是一样的。

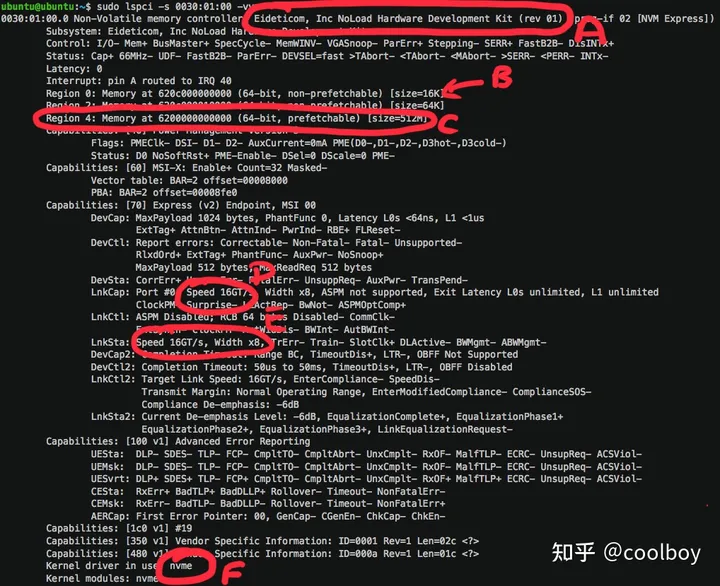
## 如何获取CMB的配置信息

在NVMe SSD Controller 中有两个寄存器CMBLOC和CMBSZ是描述CMB的基本信息。在主机中可以使用NVMe-cli工具查看寄存器信息（nvme show-regs /dev/nvme0n1 -H）。





* CMBLOC（Controller Memory Buffer Location),是指存储区的位置信息，其中OFST(Offset)表示存储区的偏移地址，单位是CMBSZ.SZ，注意要是4KB对齐。BIR（Base Indicator Register）则表示PCI BAR（基地址寄存器）的序号，。
* CMBSZ（Controller Memory Buffer Size),是指存储区的大小信息。其中Size是指CMB中可用空间的长度，注意单位也是CMBSZ.SZ。Size Units（SZU）表示CMB的单位是多少，从下表中我们可以看到，只要device有足够的空间，CMB的大小可以配置得非常大的。



A -该设备的制造商已经在PCIe数据库中注册了其供应商ID和设备ID。这意味着你得到了一个人类可读的描述。

B—该设备有3个PCIe BAR: BAR0为16KB，是任何合法NVMe设备都必须具有的标准NVMe BAR。

第三个BAR是控制器内存缓冲区(CMB)，它既可以用于NVMe队列，也可以用于NVMe数据。

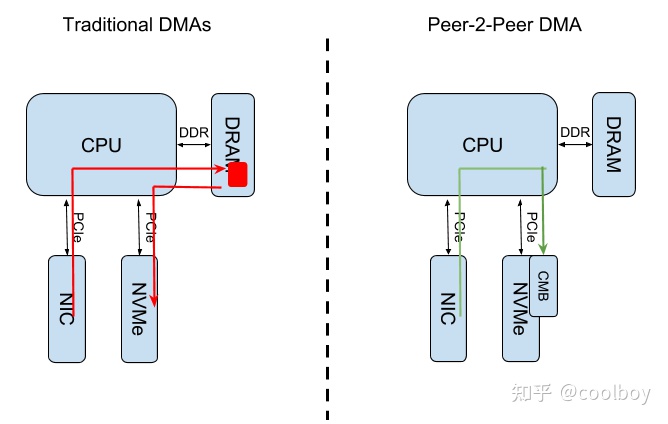
F -由于这个设备是一个NVMe设备，它被绑定到标准的Linux内核NVMe驱动程序。

## CMBs的一些应用

1. 将部分(或全部)NVMe队列放置在CMB而不是主机内存中,减少延迟。

2. 使用CMB作为DMA缓冲区允许卸载NVMe拷贝。可以提高主机性能，同时减轻主机CPU负担。

3.使用CMB作为DMA缓冲允许RDMA网卡直接将NVMe- of数据放入NVMe SSD中。减少延迟和CPU负载。



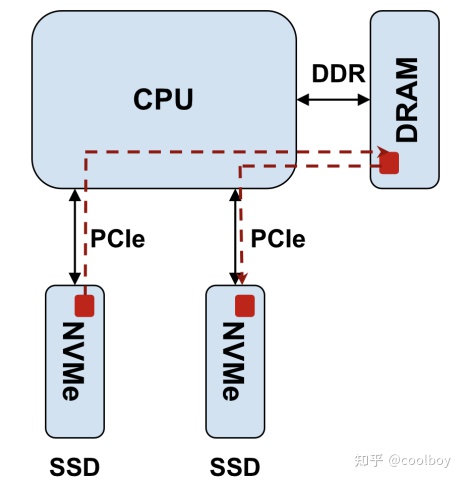
## CMB的软件- SPDK

存储性能开发工具包(Storage Performance Development Kit, SPDK)是一个免费和开源(FOSS)的高性能存储用户空间框架。

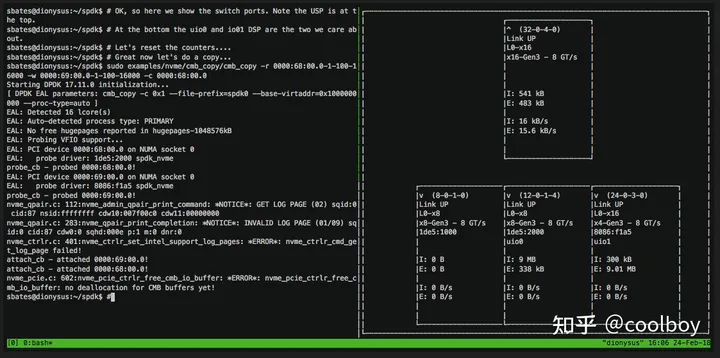
关注NVMe和NVMe- of。

2018年2月新增代码，在CMBs允许的情况下启用P2P NVMe copies。

一个简单的应用程序使用这个新的API的例子也在SPDK的例子(cmb\_copy)。



cmb\_copy是一个示例应用程序，使用SPDK的api通过P2P dma在NVMe ssd之间复制数据。这绕过了CPU的内存和PCIe子系统。



A -从SSD A拷贝9MB到SSD B。

B—PCIe交换机上行端口的数据小于1MB。

C - SPDK命令行

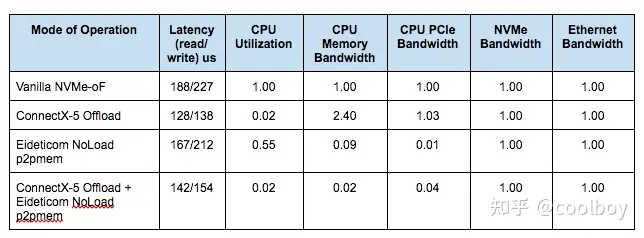
## CMB的软件 - The Linux Kernel

一个名为p2pdma的P2P框架被提议用于Linux内核。

比NVMe CMBs更通用。任何PCIe设备都可以使用它(网卡、gpgpu等)。

PCIe驱动可以注册内存(例如cmb)或请求访问内存的DMA。

初始补丁使用p2pdma优化NVMe-oF目标代码。



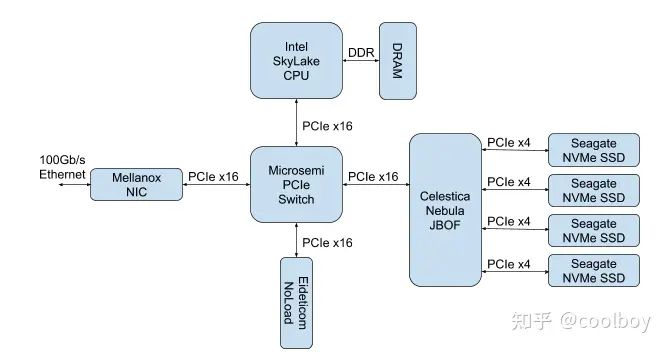
p2pdma框架可以用来改进NVMe-of目标。这里我们展示了一个通用NVMe-oF系统的结果。

p2pdma可以减少CPU内存负载x50, CPU PCIe负载x25。NVMe offload还可以将CPU核心负载降低50倍。

NVMe-oF p2pdma测试的硬件设置如下图所示。

软件安装由一个修改过的Linux内核和标准的NVMe-oF配置工具(主要是nvme-cli和nvmet)组成。

Linux内核使用了额外的NVMe offload和Peer-2-Peer DMAs支持，使用由Eideticom NVMe设备提供的NVMe CMB。



## CMB软件路标

NVMe cmb作为标准已经有一段时间了。然而，直到现在它们才开始变得可用，软件才开始使用它们。

SPDK和Linux内核是CMB软件的两个主要位置今天启用。

SPDK: NVMe P2P拷贝。NVMe-oF更新来了。SPDK Peer-2-Peer DMAs: https://spdk.io/doc/peer\_2\_peer.html

Linux内核。P2pdma框架即将上游。将扩展到支持其他NVMe/PCIe资源(例如门铃) https://www.kernel.org/doc/html/latest/driver-api/pci/p2pdma.html

2. 持久化内存区域增加了非易失性CMBs，也将需要(大量)软件支持。它们将在PCIe总线上启用一条通往持久内存存储的路径。