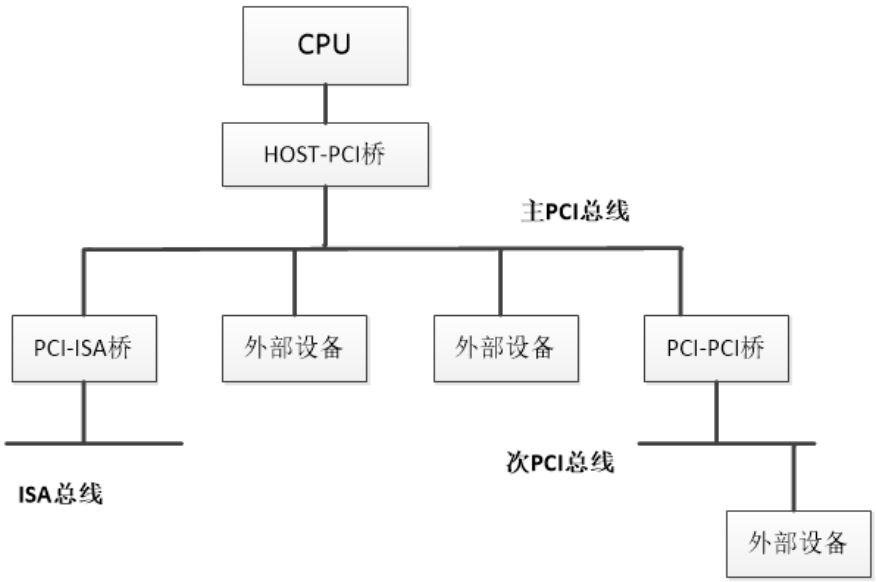
PCI 设备详解一

2016-10-09

其实之前是简单学习过PCI设备的相关知识，但是总感觉 自己的理解很函数，很多东西说不清楚，正好今天接着写这篇文章自己重新梳理一下，文章想要分为三部分，首先介绍PCI设备硬件相关的知识，然后介绍LINux内核中对PCI设备的支持。本节讲第一部分。

PCI总线在目前计算机总线系统中占据举足轻重的地位，其良好的扩展性，地址统一分配和总线竞争的处理相对于其他总线而言都具有绝对优势。

扩展性：
先说其扩展性，PCI总线上存在若干PCI设备插槽，当PCI插槽无法满足需求，就可以通过PCI桥扩展PCI设备，一个PCI桥把一个PCI总线连在一个PCI插槽上，作为PCI的一个设备。例如CPU通过"宿主-PCI桥与一条PCI总线相连，此总线成为"主PCI总线"，当通过PCI桥扩展PCI总线时，扩展的总线成为"从总线"，当然还可以通过其他的桥比如"PCI-ISA"桥扩展ISA总线，所以这样通过PCI-PCI桥可以构筑起一个层次的、树状的PCI系统结构，对于上层的总线而言，连接在这条总线上的PCI桥也是一个设备，但是这是一种特殊的设备。
其PCI树状结构如图所示：



一条PCI总线一般有32个接口，即可以连接32个PCI接口卡，而一个接口卡对应一个外部设备，注意这里的外部设备可以有多个功能（最多八个），每一个功能称为逻辑设备。每一个逻辑设备对应一个PCI配置空间。对于逻辑设备后面还会详细解释，这里先说配置空间的问题。PCI配置空间可以说是记录了关于此设备的详细信息。PCI配置空间最大256个字节，其中起先的64个字节的格式是预定义好的。当然并非所有的项都必须填充，位置是固定了，没有用到可以填充0。而前16个字节的格式是一定的。包含头部的类型、设备的总类、设备的性质以及制造商等。

PCi配置空间前64个字节格式如下：

邮箱：zhunxun@gmail.com

< 2020年4月 >

日	一	二	三	四	五	六
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

搜索

找找看

谷歌搜索

PostCategories
C语言(2)
IO Virtualization(3)
KVM虚拟化技术(26)
linux 内核源码分析(61)
Linux日常应用(3)
linux时间子系统(3)
qemu(10)
seLinux(1)
windows内核(5)
调试技巧(2)
内存管理(8)
日常技能(3)
容器技术(2)
生活杂谈(1)
网络(5)
文件系统(4)
硬件(4)

PostArchives
2018/4(1)
2018/2(1)
2018/1(3)
2017/12(2)
2017/11(4)
2017/9(3)
2017/8(1)
2017/7(8)
2017/6(6)
2017/5(9)
2017/4(15)
2017/3(5)
2017/2(1)
2016/12(1)
2016/11(11)
2016/10(8)
2016/9(13)

ArticleCategories
时态分析(1)

Recent Comments
1. Re:virtio前端驱动详解
我看了下，Linux-4.18.2中的vp_notify()函数。bool vp_notify(struct virtqueue *vq){ /* we write the queue's selec...
--Linux-inside
2. Re:virtIO之VHOST工作原理简析

再问一个问题，从设置ioeventfd那个流程来看的话是guest发起一个IO，首先会陷入到kvm中，然后由kvm向qemu发送一个IO到来的event，最后IO才被处理，是这样的吗？

--Linux-inside

3. Re:virtIO之VHOST工作原理简析
你好。设置ioeventfd这个部分和guest里面的virtio前端驱动有关系吗？
设置ioeventfd和virtio前端驱动是如何发生联系起来的？谢谢。

--Linux-inside

4. Re:QEMU IO事件处理框架
良心博主，怎么停跟了，太可惜了。

--黄铁牛

5. Re:linux 逆向映射机制浅析
小哥哥520脱单了么

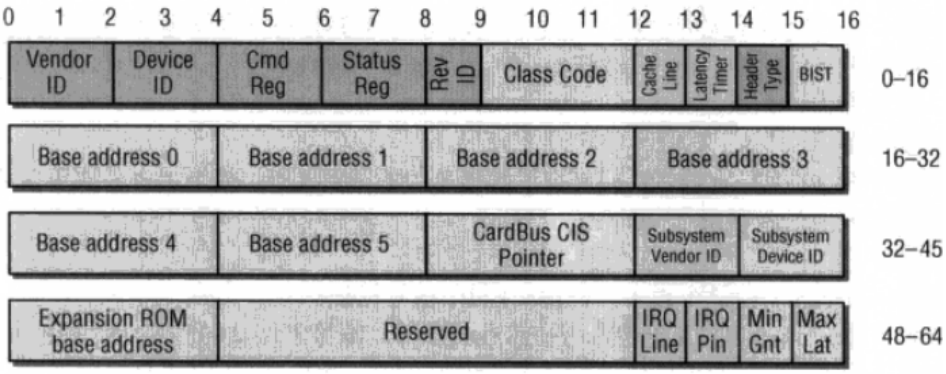
--黄铁牛

Top Posts

1. 详解操作系统中断(21009)
2. PCI 设备详解一(15698)
3. 进程的挂起、阻塞和睡眠(13575)
4. Linux下桥接模式详解一(13372)
5. virtio后端驱动详解(10478)

推荐排行榜

1. 进程的挂起、阻塞和睡眠(6)
2. 为何要写博客(2)
3. virtIO前后端notify机制详解(2)
4. 详解操作系统中断(2)
5. qemu-kvm内存虚拟化1(2)



需要注意的有一下几项：

ClassCode 用于将设备分到具体的功能组，该字段分为两部分，前8个bit表示基类即大类别，后8个比特表示基类的一个子类。比如PCI_BASE_CLASS_STORAGE表示大类大容量存储器，而PCI_CLASS_STORAGE_IDE表明这个IDE控制器。

HeaderType表明头部类型。一般区分为0型头部（PCI设备）1型头部（PCI桥），注意不同头部的配置空间格式有差异。这里我们主要先描述0型头部即普通PCi设备的配置空间。

PCI HeaderType为一个字节的大小，最高位为0表示单功能，最高位为1表示多功能（即前面描述的逻辑设备），单功能情况下一个PCI设备就是一个逻辑设备。低7位表示头部类型。

前16个字节都是一些基本的信息就不在多说，重点看下接下来的6个BAR空间。每个BAR记录了该设备映射的一段地址空间。为了区分IO空间和IO内存，这里我们分开描述：

当BAR最后一位为0表示这是映射的IO内存，为1是表示这是IO 端口，当是IO内存的时候1-2位表示内存的类型，bit 2为1表示采用64位地址，为0表示采用32位地址。bit1为1表示区间大小超过1M，为0表示不超过1M.bit3表示是否支持可预取。

而相对于IO内存，当最后一位为1时表示映射的IO地址空间。IO地址空间一般不支持预取，所以这里是29位的地址。

一般情况下，6个BAR是足够使用的，大部分情况都是3-4个BAR。而这些BAR映射的空间是连续的，即这些BAR共同描述设备的地址空间范围

除了6个基本的BAR空间们还有一个额外的配置ROM区间，区间最低位表示是否使用ROM区间，高21位表示地址。中间的是保留项。其余原理和上面类似。

接下来需要注意的就是中断，因为大部分外设和系统交互就是通过中断的方式。。

由配置空间中的IRQ Pin决定设备是否支持中断，1表示支持，0表示不支持。假如支持中断，IRQ Line表记录下中断号。

PCI桥的配置空间

PCI桥同样是连接在PCI总线接口卡上的一个设备，只不过是一个桥设备，连接一条PCI总线。既然同属于设备，那么它同样也就有设备的配置空间，只是它的配置空间和普通设备的有些差异。PCI桥的配置空间在系统软件遍历PCI总线树的时候配置，并不需要专门的PCI驱动，故称为透明桥。PCI桥连接两条总线，和HOst 桥近的称为上游总线（Primary Bus ）,远的一条称为下游总线（Secondry Bus）。PCI桥的配置空间在前16个字节的格式和普通PCI设备并无区别，另外，桥还保留了普通设备的前两个BAR空间。所以从配置空间的0x18开始有了桥设备自身的配置格式。如前所述，桥设备记录了上游总线和下游总线，以及桥下最大的总线号。这里还有一个比较重要的概念就是窗口。在PCI桥的配置空间有三个窗口：IO地址区间窗口、存储器区间窗口、可预取存储器地址窗口。实际上每个窗口都是一段地址区间，就像一个门，规定了桥下设备映射的区间。从北桥出来的地址，如果在该区间内，就可以穿过该桥到达次级总线，这样依次寻找设备。反过来，从南桥出来的地址及由设备发出的，只有地址不在该区间范围内才可以穿过该桥。因为同一条总线上的设备交互不需要外部空间。就像是内网传输和公网传输一样的道理。

内核中关于桥配置空间的定义如下：

```
/* Header type 1 (PCI-to-PCI bridges) */
#define PCI_PRIMARY_BUS         0x18    /* Primary bus number */
#define PCI_SECONDARY_BUS    0x19    /* Secondary bus number */
#define PCI_SUBORDINATE_BUS    0x1a    /* Highest bus number behind the bridge */
#define PCI_SEC_LATENCY_TIMER    0x1b    /* Latency timer for secondary interface */

#define PCI_IO_BASE         0x1c    /* I/O range behind the bridge */
#define PCI_IO_LIMIT         0x1d

#define  PCI_IO_RANGE_TYPE_MASK    0x0fUL    /* I/O bridging type */
#define  PCI_IO_RANGE_TYPE_16    0x00
#define  PCI_IO_RANGE_TYPE_32    0x01
#define  PCI_IO_RANGE_MASK    (~0x0fUL) /* Standard 4K I/O windows */
#define  PCI_IO_1K_RANGE_MASK    (~0x03UL) /* Intel 1K I/O windows */
#define PCI_SEC_STATUS         0x1e    /* Secondary status register, only bit 14 used */

#define PCI_MEMORY_BASE         0x20    /* Memory range behind */
#define PCI_MEMORY_LIMIT    0x22

#define  PCI_MEMORY_RANGE_TYPE_MASK 0x0fUL
#define  PCI_MEMORY_RANGE_MASK    (~0x0fUL)

#define PCI_PREF_MEMORY_BASE    0x24    /* Prefetchable memory range behind */
#define PCI_PREF_MEMORY_LIMIT    0x26

#define  PCI_PREF_RANGE_TYPE_MASK 0x0fUL
#define  PCI_PREF_RANGE_TYPE_32    0x00
#define  PCI_PREF_RANGE_TYPE_64    0x01
#define  PCI_PREF_RANGE_MASK    (~0x0fUL)
#define PCI_PREF_BASE_UPPER32    0x28    /* Upper half of prefetchable memory range */
#define PCI_PREF_LIMIT_UPPER32    0x2c
#define PCI_IO_BASE_UPPER16    0x30    /* Upper half of I/O addresses */
#define PCI_IO_LIMIT_UPPER16    0x32
/* 0x34 same as for htype 0 */
/* 0x35-0x3b is reserved */
#define PCI_ROM_ADDRESS1    0x38    /* Same as PCI_ROM_ADDRESS, but for htype 1 */
/* 0x3c-0x3d are same as for htype 0 */
#define PCI_BRIDGE_CONTROL    0x3e
#define  PCI_BRIDGE_CTL_PARITY    0x01    /* Enable parity detection on secondary
interface */
#define  PCI_BRIDGE_CTL_SERR    0x02    /* The same for SERR forwarding */
#define  PCI_BRIDGE_CTL_ISA    0x04    /* Enable ISA mode */
#define  PCI_BRIDGE_CTL_VGA    0x08    /* Forward VGA addresses */
#define  PCI_BRIDGE_CTL_MASTER_ABORT    0x20  /* Report master aborts */
#define  PCI_BRIDGE_CTL_BUS_RESET    0x40    /* Secondary bus reset */
#define  PCI_BRIDGE_CTL_FAST_BACK    0x80    /* Fast Back2Back enabled on secondary
interface */
```

下面说说PCI设备的地址空间：

前面也简单介绍了下PCI设备的地址空间支持PIO和MMIO，即IO端口和IO内存。下面详细分析下这两种方式：

PIO

IO端口的编址是独立于系统的地址空间，其实就是一段地址区域，所有外设的地址都映射到这段区域中。就像是一个进程内部的各个变量，公用进程地址空间一样。不同外设的IO端口不同。访问IO端口需要特殊的IO指令，OUT/IN，OUT用于write操作，in用于read操作。在此基础上，操作系统实现了读写不同大小端口的函数。为什么说是不同大小呢？？因为前面也说到，IO端口实际上是一段连续的区域，每个端口理论上是字节为单位即8bit,那么要想读写16位的端口只能把相邻的端口进行合并，32位的端口也是如此。

例如下面的汇编指令：
OUT 21h,al
0x21是8259A中断控制器的中断屏蔽寄存器，该指令将Intel X86处理器的al寄存器中的值写到中断寄存器的控制寄存器中，从而达到屏蔽某些中断信号的目的。类似的，在下面的汇编指令中：
IN al,20h
0x20是8259A中断控制器的正在服务寄存器，该指令将此寄存器中的状态值传递到处理器的al寄存器中。

无论是windows还是Linux都会上述指令做了封装以满足读写不同长度端口的需要。不过这种方式缺点也比较明显，一般情况CPU分配给IO端口的空间都比较小，在当前外设存储日益增大的情况下很难满足需要。另一方面，

MMIO

IO内存是直接把寄存器的地址空间直接映射到系统地址空间，系统地址空间往往会保留一段内存区用于这种MMIO的映射（当然肯定是位于系统内存区），这样系统可以直接使用普通的访存指令直接访问设备的寄存器，随着计算机内存容量的日益增大，这种方式更是显出独特的优势，在性能至上的理念下，使用MMIO可以最大限度满足日益增长的系统和外设存储的需要。所以当前其实大多数外设都是采用MMIO的方式。

还有一种方式是把IO端口空间映射到内存空间，这样依然可以通过正常的访存指令访问IO端口，但是这种方式下依然受到IO空间大小的制约，可以说并没有解决实际问题。

但是上述方案只适用于在外设和内存进行小数据量的传输时，假如进行大数据量的传输，那么IO端口这种以字节为单位的传输就不用说了，IO内存虽然进行了内存映射，但是其映射的范围大小相对于大量的数据，仍然不值一提，所以即使采用IO内存仍然是满足不了需要，会让CPU大部分时间处理繁琐的映射，极大的浪费了CPU资源。那么这种情况就引入了DMA，直接内存访问。这种方式的传输由DMA控制器控制，CPU给DMA控制器下达传输指令后就转而处理其他的事务，然后DMA控制器就开始进行数据的传输，在完成数据的传输后通过中断的方式通知CPU，这样就可以极大的解放CPU。当然本次讨论的重点不在DMA，所以对于DMA的讨论仅限于此。

那么CPU是怎么访问这些配置寄存器的呢？要知道配置寄存器指定了设备存储寄存器的映射方式以及地址区间，但是配置寄存器本身的访问就是一个问题。为每个设备预留IO端口或者IO内存都是不现实的。暂且不说x86架构下IO端口地址空间只有区区64K,就是内存，虽然现在随着科技的发展，内存空间越来越大，但是也不可能为每个设备预留空间。那么折中的方式就是为所有设备的配置寄存器使用同一个IO端口，系统在IO地址空间预留了一段地址就是0xCF8~0xCFF一共八个字节，前四个字节做地址端口，后四个字节做数据端口。CPU访问某个设备的配置寄存器时，先向地址端口写入地址，然后从数据端口读写数据。这里的地址是一个综合地址，结构如下：

最高位为1	保留不用（7位）	总线号（8位）	设备号（5位）	功能号（3位）	寄存器地址（8位）最低两位为0

这里就可以解释我们总线数量和逻辑设备数量的限制了。在寻找某一个逻辑设备时，先根据总线号找到总线，然后根据设备号找到总线上的某个接口，最后在根据功能号定位某一个逻辑设备。

到此，PCI总线架构以及设备的配置空间已经描述的差不多了，下面就是看系统如何探测和初始化设备了。

下篇文章会结合LInux源代码分析下Linux内核中对PCI设备的配置过程

分类: linux 内核源码分析, 硬件

jack.chen
关注 - 12
粉丝 - 44
+加关注

« 上一篇： SKBUFFER详解
» 下一篇： PCI 设备详解二

posted @ 2016-10-09 18:52 jack.chen Views(15699) Comments(0) Edit 收藏

相关博文：
· Linux驱动学习--初识PCI驱动(一)
· linux pci 协议一
· PCI 设备详解二
· PCI设备的地址空间
· PCI 设备详解三
» 更多推荐...

最新 IT 新闻:
· 李国庆发布的人事任命有效吗？律师是这样说的
· “负油价”都来了 国内油价为何不降？中国石油回应
· CarPlay新功能上线！宝马车主未来可免费使用
· 首次！美国防部公开3段UFO视频：官方证实真实性
· 中国网民分析：58.3%高中以下学历 6.5亿网民月收入不足5000元
» 更多新闻...

以马内利