Chapter 3 Page tables

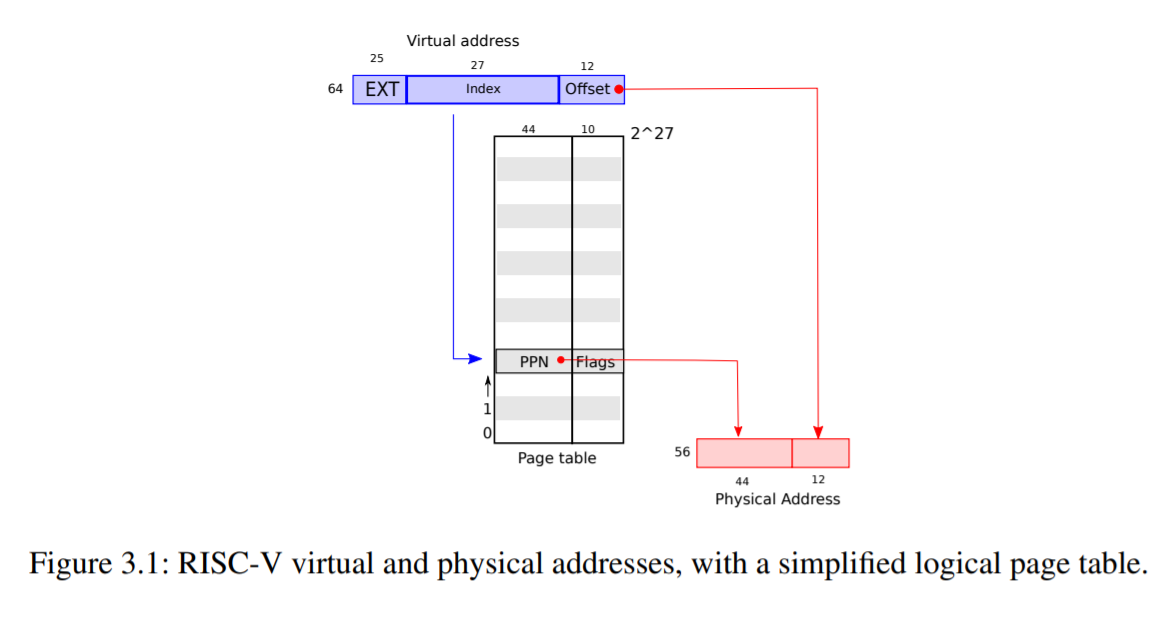
page table是操作系统为每个进程提供自己私有地址空间和内存的机制。page table决定了内存地址的含义，以及物理内存的哪些部分可以被访问。它们允许xv6隔离不同进程的地址空间，并将它们映射到物理内存上。page table还提供了一个间接层次，允许xv6执行一些技巧：在几个地址空间中映射同一内存(一个trampoline页)，以及用一个未映射页来保护内核和用户stack。本章其余部分将解释RISC-V硬件提供的page table以及xv6如何使用它们。

### 3.1 Paging hardware

提醒一下，RISC-V指令(包括用户和内核)操作的是虚拟地址。机器的RAM，或者说物理内存，是用物理地址来做索引的，RISC-V page table 硬件将这两种地址连接起来，通过将每个虚拟地址映射到物理地址上。

xv6 运行在 Sv39 RISC-V 上，这意味着只使用 64 位虚拟地址的底部 39 位，顶部 25 位未被使用。在这种Sv39配置中，一个RISC-V page table 在逻辑上是一个2^27（134,217,728）page table entries（PTE）的数组。每个PTE包含一个44位的physical page number（PPN）和一些标志。分页硬件通过利用39位中的前27位索引到页表中找到一个PTE来转换一个虚拟地址，并计算出一个56位的物理地址，它的前44位来自于PTE中的PPN，而它的后12位则是从原来的虚拟地址复制过来的。图3.1显示了这个过程，逻辑上把页表看成是一个简单的PTE数组（更完整的故事见图3.2）。页表让操作系统控制虚拟地址到物理地址的转换，其粒度为4096（2^12）字节的对齐块。这样的分块称为页。

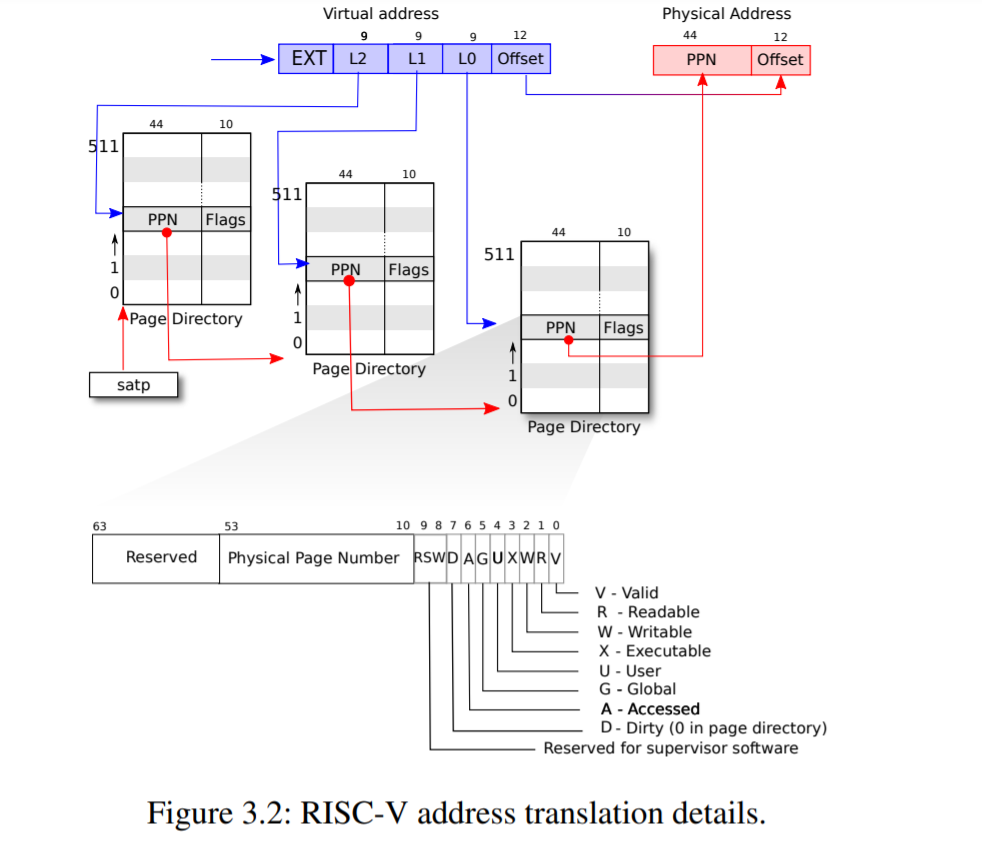
在Sv39 RISC-V中，虚拟地址的前25位不用于转换地址；将来，RISC-V可能会使用这些位来定义更多的转换层。物理地址也有增长的空间：在PTE格式中，物理页数还有10位的增长空间。



如图3.2所示，实际转换分三步进行。一个page table以三层树的形式存储在物理内存中。树的根部是一个4096字节的page table页，它包含512个PTE，这些PTE包含树的下一级page table页的物理地址。这些page中的每一个都包含512个PTE，用于树的最后一级。分页硬件用27位中的顶9位选择根page-table页中的PTE，用中间9位选择树中下一级page-table页中的PTE，用底9位选择最后的PTE。

如果转换一个地址所需的三个PTE中的任何一个不存在，分页硬件就会引发一个页面错误的异常，让内核来处理这个异常（见第4章）。这种三层结构允许pagetable在大范围虚拟地址没有映射的情况下省略整个页表页。

每个PTE包含标志位，告诉paging硬件如何允许使用相关的虚拟地址。PTE\_V表示PTE是否存在：如果没有设置，对该页的引用会引起异常（即不允许）。PTE\_R控制是否允许指令读取到页面。PTE\_W控制是否允许指令向页面写入。PTE\_X控制CPU是否可以将页面的内容解释为指令并执行。PTE\_U控制是否允许用户模式下的指令访问页面；如果不设置PTE\_U，PTE只能在supervisor模式下使用。图3.2显示了这一切的工作原理。标志和其他所有与页硬件相关的结构在(kernel/riscv.h)中定义了。

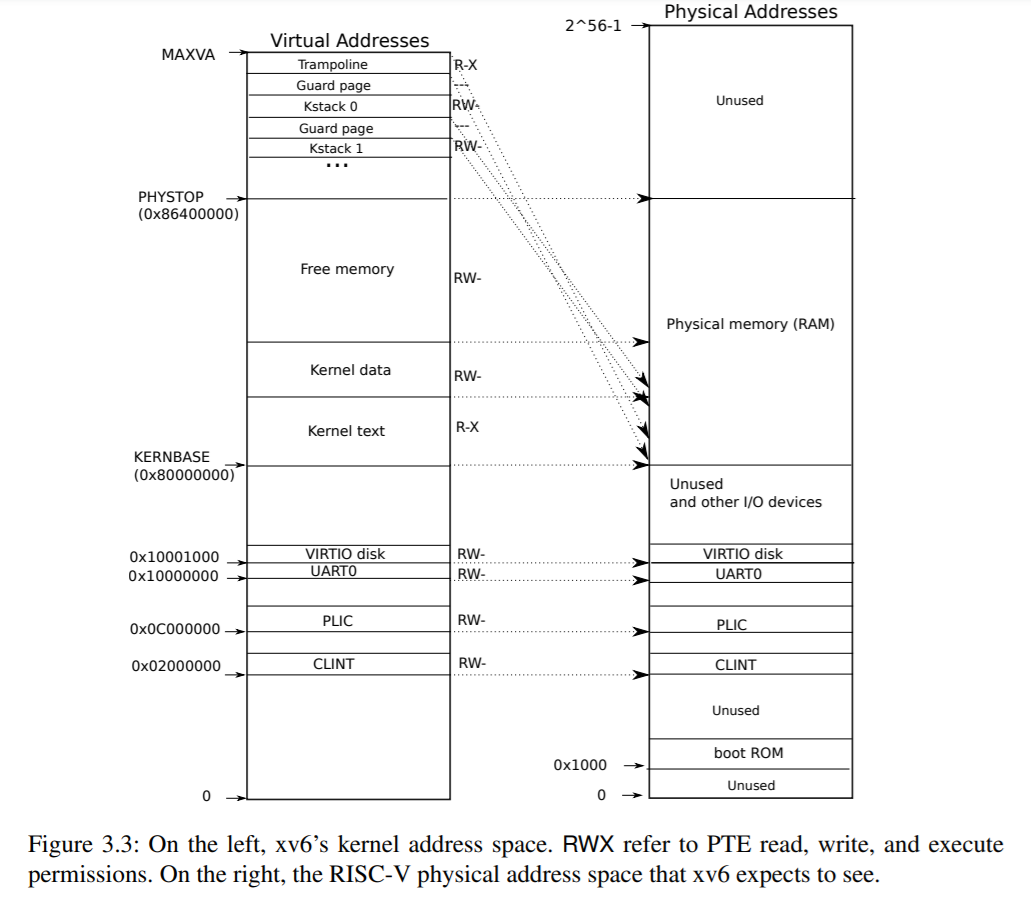


要告诉硬件使用page table，内核必须将根页表页的物理地址写入satp寄存器中。每个CPU都有自己的satp。一个CPU将使用自己的satp所指向的页表来翻译后续指令产生的所有地址。每个CPU都有自己的satp，这样不同的CPU可以运行不同的进程，每个进程都有自己的页表所描述的私有地址空间。

关于术语的一些说明。物理内存指的是DRAM中的存储单元。物理存储器的一个字节有一个地址，称为物理地址。指令只使用虚拟地址，分页硬件将其翻译成物理地址，然后发送给DRAM硬件，以读取或写入存储。不想物理内存和虚拟地址，虚拟内存不是一个物理对象，而是指内核提供的管理物理内存和虚拟地址的抽象和机制的集合。

### 3.2 Kernel address space

Xv6为每个进程维护页表，一个描述每个进程的用户地址空间，外加一个描述内核地址空间的单页表。内核配置其地址空间的设计，使其能够通过可预测的虚拟地址访问物理内存和各种硬件资源。图3.3显示了这个设计是如何将内核虚拟地址映射到物理地址的。文件(kernel/memlayout.h)声明了xv6内核内存布局的常量。



QEMU模拟的计算机包括RAM（物理内存），从物理地址0x80000000开始，一直到至少0x86400000，xv6称之为PHYSTOP。QEMU模拟还包括I/O设备，如磁盘接口。QEMU将设备接口作为内存映射的控制寄存器暴露给软件，这些寄存器位于物理地址空间的0x80000000以下。内核可以通过读取/写入这些特殊的物理地址与设备进行交互；这种读取和写入与设备硬件而不是与RAM进行通信。第4章解释了xv6如何与设备交互。

内核使用“直接映射”的方式获取RAM和内存映射的设备寄存器，也就是在虚拟地址上映射资源，这些地址与物理地址相等。例如，内核本身在虚拟地址空间和物理内存中的位置都是KERNBASE=0x80000000。直接映射简化了读或写物理内存的内核代码。例如，当fork为子进程分配用户内存时，分配器返回该内存的物理地址；fork在将父进程的用户内存复制到子进程时，直接使用该地址作为虚拟地址。

有几个内核虚拟地址不是直接映射的:

1．trampoline 页。它被映射在虚拟地址空间的顶端；用户页表也有这种映射。第4章讨论了trampoline 页的作用，但我们在这里看到了页表的一个有趣的用例；一个物理页（存放trampoline代码）在内核的虚拟地址空间中被映射了两次：一次是在虚拟地址空间的顶部，一次是直接映射。

2．内核stack页。每个进程都有自己的内核stacks，内核stacks被映射到地址高处，所以在它后面xv6可以留下一个未映射的防护页。守护页的PTE是无效的（即PTE\_V没有设置），这样如果内核溢出内核stack，很可能会引起异常，内核会报错。如果没有防护页，溢出的stack会覆盖其他内核内存，导致不正确的操作。报错是比较好的

当内核通过地址高处内存映射使用stack时，它们也可以通过直接映射的地址被内核访问。另一种的设计是只有直接映射，并在直接映射的地址上使用stack。然而，在这种安排中，保护页（stack物理地址后面的一页）将不映射虚拟地址，既无法通过PTE的标志位来设置是否允许访问，这些地址将指向物理内存，内核stack溢出时就会直接访问的物理地址。

内核为trampoline和text映射的page会有PTE\_R(读)和PTE\_X（执行）权限。内核从这些pages读取和执行指令。内核映射的其他page会有PTE\_R和PTE\_W（写）权限，以便内核读写这些页面的内存。守护页的映射是无效的（设置PTE\_V）；

### 3.3 Code: creating an address space

核心数据结构是pagetable\_t，它实际上是一个指向RISC-V根页表页的指针；pagetable\_t可以是内核页表，也可以是每个进程的页表之一。核心函数是walk和mappages，前者为虚拟地址寻找PTE，后者为新的映射安置PTE。以 kvm 开头的函数操作内核页表；以 uvm 开头的函数操作用户页表；其他函数同时用于这两个方面。copyout 可以将数据copy到用户虚拟地址，copyin可以将用户虚拟地址的数据copy到内核，用户虚拟地址由系统调用的参数提供；它们在 vm.c 中，因为它们需要显式转换这些地址，以便找到相应的物理内存。

在启动序列的前面，main调用kvminit(kernel/vm.c:22)来创建内核的页表。这个调用发生在xv6在RISC-V上启用分页之前，所以地址直接指向物理内存。Kvminit首先分配一页物理内存来存放根页表页。然后调用kvmmap来install内核所需的translations。这些translations包括内核的指令和数据，KERNBASE到PHYSTOP（0x86400000）的物理内存，以及实际上是设备的内存范围。

kvmmap (kernel/vm.c:118) 调用 mappages (kernel/vm.c:149)，它将一个虚拟地址范围映射到一个物理地址范围。它对范围内地址分割成多页（忽略余数），每次install一页的顶端地址。对于每个要映射的虚拟地址（页的顶端地址），mapages调用walk找到该地址的最后一级PTE的地址。然后，它初始化PTE，使其持有相关的物理页号、所需的权限(PTE\_W、PTE\_X和/或PTE\_R)，以及PTE\_V来标记PTE为有效(kernel/vm.c:161)。

walk (kernel/vm.c:72)模仿RISC-V分页硬件查找虚拟地址的PTE(见图3.2)。walk每次降低3-level page table 的9位。它使用每一级的9位虚拟地址来查找下一级页表或最后一级（kernel/vm.c:78）的PTE。如果PTE无效，那么所需的页还没有被分配；如果分配参数被设置，walk分配一个新的页表页，并把它的物理地址放在PTE中。它返回PTE在树的最低层的地址(kernel/vm.c:88)。

main 调用 kvminithart (kernel/vm.c:53) 来install内核页表。它将root page-table page的物理地址写入寄存器satp中。在这之后，CPU将使用内核页表翻译地址。由于内核使用identity映射，所以下一条指令的现在虚拟地址将映射到正确的物理内存地址。

procinit (kernel/proc.c:26)，它由main调用，为每个进程分配一个内核stack。kvmmap将映射的PTE添加到内核页表中，调用kvminithart将内核页表重新加载到satp中，这样硬件就知道新的PTE了。

每个RISC-V CPU都会在Translation Look-aside Buffer(TLB)中缓存页表项，当xv6改变页表时，必须告诉CPU使相应的缓存TLB项无效。如果它没有这样做，那么在以后的某个时刻，TLB可能会使用一个旧的缓存映射，指向一个物理页，而这个物理页在此期间已经分配给了另一个进程，结果，一个进程可能会在其他进程的内存上乱写乱画。RISC-V有一条指令sfence.vma，可以刷新当前CPU的TLB。xv6在重新加载satp寄存器后，在kvminithart中执行sfence.vma，也会在从内核空间返回用户空间前，切换到用户pagetable 的trampoline代码中执行sfence.vma(kernel/trampoline.S:79)。

### 3.4 Physical memory allocation

内核必须在运行时为页表、用户内存、内核堆栈和管道缓冲区分配和释放物理内存。 xv6使用内核地址结束和PHYSTOP之间的物理内存进行运行时分配。它每次分配和释放整个4096字节的页面。它通过保存free page 链表，来跟踪哪些page是空闲的。分配包括从链表中删除一个page；释放包括将释放的页面添加到链表中。

### 3.5 Code: Physical memory allocator

分配器在kalloc.c(kernel/kalloc.c:1)中。分配器的数据结构是一个可供分配的物理内存页的空闲链表，每个空闲页的链表元素是一个结构run（kernel/kalloc.c:17）。分配器从哪里获得内存来存放这个数据结构呢？它把每个空闲页的run结构存储在空闲页本身，因为那里没有其他东西存储。空闲列表由一个自旋锁保护(kernel/kalloc.c:21-24)。链表和锁被包裹在一个结构中，以明确锁保护的是结构中的字段。现在，请忽略锁以及获取和释放的调用；第6章将详细研究锁。

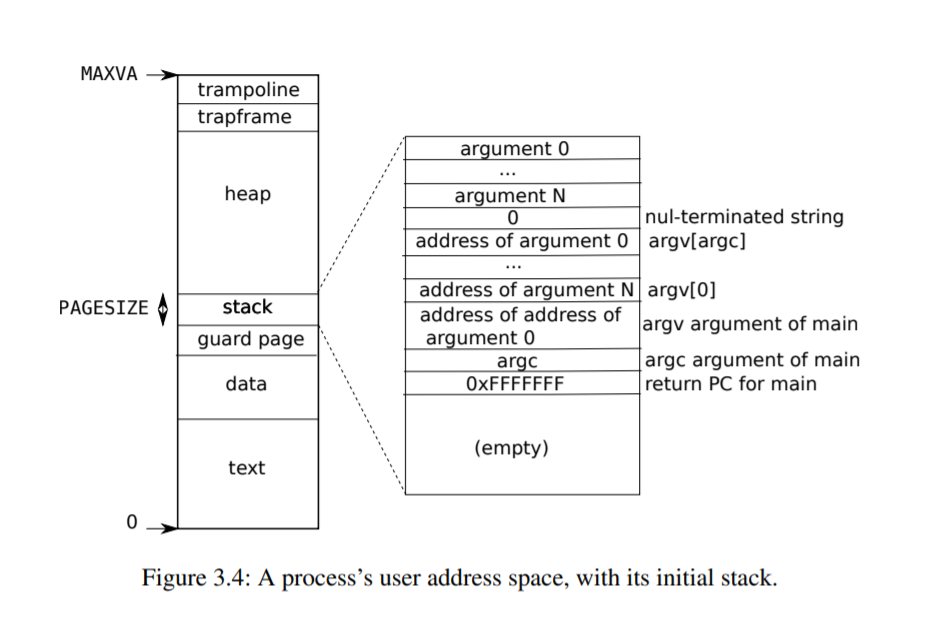
main函数调用kinit来初始化分配器(kernel/kallo .c:27)。kinit初始化free page链表，以保存内核地址结束和PHYSTOP之间的每个页面。xv6应该通过解析硬件提供的配置信息来确定有多少物理内存可用。但是它没有做，而是假设机器有128M字节的RAM。Kinit通过调用freerange来添加free page链表，freerange则对每一个page都调用kfree。PTE只能引用按4096字节边界对齐的物理地址9(是4096的倍数)，因此freerange使用PGROUNDUP来确保它只释放对齐的物理地址。分配程序开始时没有内存;这些给kfree的电话给了它一些管理的机会。

函数main调用kinit来初始化分配器(kernel/kalloc.c:27)。kinit初始化free page 链表，以保存内核地址结束到PHYSTOP之间的每一页。 xv6应该通过解析硬件提供的配置信息来确定有多少物理内存可用。kinit调用freerange通过每页调用kfree来增加内存到空闲列表中。一个PTE只能引用一个在4096字节边界上对齐的物理地址（是4096的倍数），所以freerange使用PGROUNDUP来确保它只释放对齐的物理地址。分配器开始时没有内存，这些对kfree的调用给了它一些内存。

分配器有时把地址当作整数来处理，以便对其进行运算（如遍历自由区的所有页），有时把地址当作指针来读写内存（如操作存储在每页中的运行结构）；这种对地址的双重使用是分配器代码中充满C类型转换的主要原因。另一个原因是，释放和分配本质上改变了内存的类型。

函数kfree (kernel/kalloc.c:47)将被free的内存中的每个字节设置为1。这将导致在释放内存后使用内存的代码(使用悬空引用)读取垃圾而不是旧的有效内容；希望这将导致这类代码更快地崩溃。然后kfree将页面预存入释放列表：它将pa（物理地址）转为指向struct run的指针，在r->next中记录释放列表的旧开始，并将释放列表设为r，kalloc移除并返回free 链表中的第一个元素。

### 3.6 Process address space

每个进程都有一个单独的页表，当xv6在进程间切换时，也会改变页表。如图2.3所示，一个进程的用户内存从虚拟地址0开始，可以增长到MAXVA(kernel/riscv.h:348)，原则上允许一个进程寻址256GB的内存。

当一个进程要求xv6提供更多的用户内存时，xv6首先使用kalloc来分配物理页，然后将指向新物理页的PTE添加到进程的页表中。然后它将指向新物理页的PTE添加到进程的页表中。Xv6在这些PTE中设置PTE\_W、PTE\_X、PTE\_R、PTE\_U和PTE\_V标志。大多数进程不使用整个用户地址空间；xv6使用PTE\_V来清除不使用的PTE。

我们在这里看到了几个不错的页表使用实例。首先，不同的进程页表将用户地址转化为物理内存的不同页，这样每个进程都有私有的用户内存。第二，每个进程都认为自己的内存具有从零开始的连续的虚拟地址，而进程的物理内存可以是不连续的。第三，内核会映射带有trampoline 代码的page，该trampoline处于用户地址空间顶端，因此，在所有地址空间中都会出现一页物理内存。

图3.4更详细地显示了xv6中执行进程的用户内存布局。栈是一个单页，图中显示的是由exec创建的初始内容。字符串的值，以及指向这些参数的指针数组，位于栈的最顶端。下面是允许程序在main启动的值，就像函数main(argc, argv)刚刚被调用一样。

为了检测用户stack溢出分配的stack内存，xv6会在stack的正下方放置一个无效的保护页。如果用户stack溢出，而进程试图使用栈下面的地址，硬件会因为映射无效而产生一个页错误异常。现实世界中的操作系统可能会在用户stack溢出时自动为其分配更多的内存。

### 3.7 Code: sbrk

Sbrk是一个进程收缩或增长内存的系统调用。该系统调用由函数growproc(kernel/proc.c:239)实现，growproc调用uvmalloc或uvmdealloc，取决于n是正数还是负数。 uvmdealloc 调用 uvmunmap (kernel/vm.c:174)，它使用 walk 来查找 PTE，使用 kfree 来释放它们所引用的物理内存。

xv6使用进程的page-table不仅是为了告诉硬件如何映射用户虚拟地址，也是做为分配给该进程的物理地址的唯一记录。这就是为什么释放用户内存（在uvmunmap中）需要检查用户page table的原因。

### 3.8 Code: exec

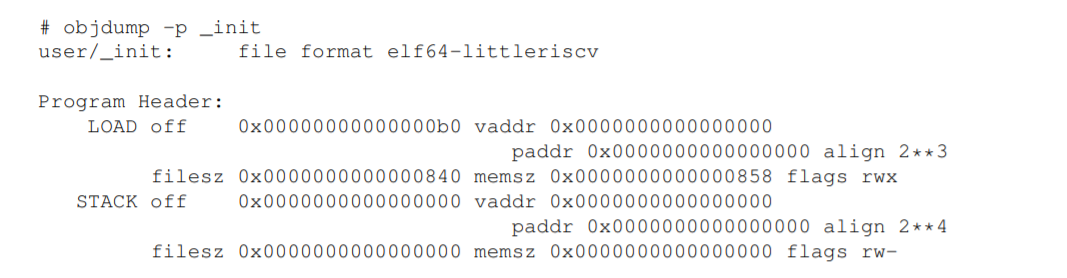
Exec是创建用户地址空间的系统调用。它读取储存在文件系统上的文件用来初始化用户地址空间。Exec (kernel/exec.c:13)使用namei (kernel/exec.c:26)打开二进制文件路径，这在第8章中有解释。然后，它读取ELF头。xv6应用程序用广泛使用的ELF格式来描述，它定义在(kernel/elf.h)。一个ELF二进制文件包括一个ELF头，elfhdr结构(kernel/elf.h:6)，后面是一个程序section headers序列，proghdr结构(kernel/elf.h:25)。每一个proghdr描述了一个必须加载到内存中的程序section；xv6程序只有一个程序section headers，但其他系统可能有单独的指令section和数据section需要加载到内存。

第一步是快速检查文件是否包含一个ELF二进制文件。一个ELF二进制文件以四个字节的”魔法数字“ 0x7F、`E`、`L`、`F`或ELF\_MAGIC(kernel/elf.h:3)开始。如果ELF头有正确的”魔法数字“，exec就会认为该二进制文件是正确的类型。

Exec 使用proc\_pagetable(kernel/exec.c:38)分配一个没有使用的page table，使用uvmalloc (kernel/exec.c:52)为每一个ELF segment分配内存，通过loadseg (kernel/exec.c:10)加载每一个segment到内存中。

Exec allocates a new page table with no user mappings with proc\_pagetable (kernel/exec.c:38), allocates memory for each ELF segment with uvmalloc (kernel/exec.c:52), and loads each segment into memory with loadseg (kernel/exec.c:10). loadseg uses walkaddr to find the physical address of the allocated memory at which to write each page of the ELF segment, and readi to read from the file.

用exec创建的第一个用户程序/init的程序部分header是这样的。



程序部分头的filesz可能小于memsz，说明它们之间的空隙应该用0（C语言全局变量）来填充，而不是从文件中读取。对于/init来说，filesz是2112字节，memsz是2136字节，因此uvmalloc分配了足够的物理内存来容纳2136字节，但只从文件/init中读取2112字节。

exec在stack页的下方放置了一个不可访问页，这样程序如果试图使用多个页面，就会出现故障。这个不可访问的页允许exec处理过大的参数；在这种情况下，exec用来复制参数到stack的copyout(kernel/vm.c:355)函数会注意到目标页不可访问，并返回-1。

During the preparation of the new memory image, if exec detects an error like an invalid program segment, it jumps to the label bad, frees the new image, and returns -1. Exec must wait to free the old image until it is sure that the system call will succeed: if the old image is gone, the system call cannot return -1 to it. The only error cases in exec happen during the creation of the image. Once the image is complete, exec can commit to the new page table (kernel/exec.c:113) and free the old one (kernel/exec.c:117).

Exec将ELF文件中的字节按ELF文件指定的地址加载到内存中。用户或进程可以将任何他们想要的地址放入ELF文件中。因此，Exec是有风险的，因为ELF文件中的地址可能会意外地或故意地指向内核。对于一个不小心的内核来说，后果可能从崩溃到恶意颠覆内核的隔离机制(即安全漏洞)。xv6执行了一些检查来避免这些风险。例如if(ph.vaddr + ph.memsz < ph.vaddr)检查总和是否溢出一个64位整数。危险的是，用户可以用指向用户选择的地址的 ph.vaddr 和足够大的 ph.memsz 来构造一个 ELF 二进制，使总和溢出到 0x1000，这看起来像是一个有效值。在旧版本的xv6中，用户地址空间也包含了内核（但在用户模式下不可读/写），用户可以选择一个对应内核内存的地址，从而将ELF二进制中的数据复制到内核中。在RISC-V版本的xv6中，这是不可能的，因为内核有自己独立的page table；loadseg加载到进程的页表中，而不是内核的页表中。

内核开发人员很容易忽略一个关键的检查，现实中的内核有很长一段缺少检查的history，用户程序可以利用缺少这些检查来获得内核特权。xv6很可能没有完成验证提供给内核的用户级数据的工作，恶意用户程序可能利用这些数据来绕过xv6的隔离。

### 3.9 Real world

像大多数操作系统一样，xv6使用分页硬件进行内存保护和映射。大多数操作系统对分页的使用要比xv6复杂得多，它将分页和分页错误异常结合起来，我们将在第4章中讨论。

Xv6的内核使用虚拟地址和物理地址之间的直接映射，

这样会更简单，并假设在地址0x8000000处有物理RAM，即内核期望加载的地方。这在QEMU中是可行的，但是在真实的硬件上，它被证明是一个糟糕的想法；真实的硬件将RAM和设备放置在不可预测的物理地址上，例如在0x8000000处可能没有RAM，而xv6期望能够在那里存储内核。更好的内核设计利用页表将任意的硬件物理内存布局变成可预测的内核虚拟地址布局。

RISC-V支持物理地址级别的保护，但xv6没有使用该功能。

在有大量内存的机器上，使用RISC-V对super page的支持可能是有意义的。当物理内存很小的时候，小页是有意义的，可以精细地分配和分页到磁盘。例如，如果一个程序只使用8千字节的内存，那么给它整整4兆字节的超级物理内存页是浪费的。更大的页面在有大量内存的机器上是有意义的，可以减少页表操作的开销。

xv6内核缺乏一个类malloc的分配器为小程序提供内存，这防止内核使用需要动态分配的复杂数据结构，从而简化了设计。

内存分配是一个常年的热门话题，基本问题是有效利用有限的内存和为未来未知的请求做准备[7]。如今人们更关心的是速度而不是空间效率。此外，一个更复杂的内核可能会分配许多不同大小的小块，而不是（在xv6中）只分配4096字节的块；一个真正的内核分配器需要处理小块分配以及大块分配。

### 3.10 Exercises

1、分析RISC-V的设备树（device tree），找出计算机有多少物理内存。

2、编写一个用户程序，通过调用sbrk(1)使其地址空间增加一个字节。运行该程序，研究调用sbrk之前和调用sbrk之后的程序页表。内核分配了多少空间？新内存的PTE包含哪些内容？

3、修改xv6为内核使用super page(4M)

4、修改xv6，使用户程序间接引用[[1]](#footnote-1)一个空指针时，会收到一个异常，即修改xv6，使用户程序的虚拟地址0不被映射。

5、Unix实现的exec传统上包括对shell脚本的特殊处理。如果要执行的文件以文本#!开头，那么第一行就被认为是要运行的程序来解释文件。例如，如果调用exec运行myprog arg1，而myprog的第一行是#!/interp，那么exec执行/interp myprog arg1。在xv6中实现对这个约定的支持。

1. dereference，就是对地址取值，例如\*p。 [↑](#footnote-ref-1)