

Operating Systems

Timothy Bourke
Notes by Antoine Groudiev

19th February 2024

Contents

1	Virtual Memory	1
1.1	Introduction	1
1.2	Physical and virtual addressing	2
1.3	Virtual memory address translation	2
1.4	VM as a tool for memory management	2
1.5	Address translation	3
1.6	Kernel and User Mode	5
1.7	Multi-level page tables	5

Abstract

This document is Antoine Groudiev's class notes while following the class *Systèmes d'exploitation* (Operating Systems) at the Computer Science Department of ENS Ulm. It is freely inspired by Timothy Bourke's classes, and especially its slides.

Introduction

An operating system is made of three main parts: a *kernel*, *libraries*, and *applications*. It is useful for sharing memory, computation time, processors, and devices (keyboards, disks, graphics cards, ...). It constitutes an abstraction layer used as a base for building bigger systems. For example, the OS allows hardware independence for applications, and provides common services (such as file systems with access control) and allows concurrency and communication with protection and access control.

The kernel contains lots of data structures and functions used for bookkeeping, as well as abstract modules such as interfaces, data structures, and functions for services. It allows a low-level control of hardware, and manages the concurrency and the events.

1 Virtual Memory

1.1 Introduction

We will start by studying the concept of *virtual memory*. An OS must share finite resources among multiple users and applications. It provides an abstraction for building such applications, which do not need to worry about the complexity of memory management with other applications. Indeed, physical memory must be shared: but what happens if it runs out, or if one process tries to read or write the memory of another? This is where virtual memory comes in handy: it allows to run each process in a virtual address space, and selectively share memory between

processes for them to communicate. Furthermore, it allows processes to use faster physical memory as a cache for files on slower disks.

Virtual memory is an abstraction provided by a sophisticated and elegant mix of hardware and software. Hardware capabilities include exceptions (synchronous interrupts), address translation (which we will study in this section), main memory and caching of files on disks. Software-wise, we will study the kernel memory system. Hardware is needed to intervene at the lowest-level – each individual `mov` instruction – and for speed. Software is needed to implement sophisticated, flexible algorithms and for pervasive integration within a kernel.

For application programmers, virtual memory is largely invisible. Only very few programmers ever have to deal directly with this low-level hardware and software. Nevertheless, we will study it because it pervades all levels of a computer system; understanding how it works gives a deeper understanding of how the system works. It also provides powerful capabilities that can be exploited in applications. OS programmers cannot avoid knowing about virtual memory.

1.2 Physical and virtual addressing

Physical addressing is used in "simple" systems like embedded microcontrollers, in devices like cars, elevators, digital picture frames, ...

Conversely, *virtual addressing* is used in all modern servers, desktops, laptops and high-end mobile phones. It uses an MMU (Memory Management Unit), which translated the virtual address (VA) to a physical address (PA).

1.3 Virtual memory address translation

A *virtual address space* is a set V of $|V| =: N = 2^n$ virtual addresses used within programs. These addresses are mapped to a *physical address space*, which is a set P of $|P| =: M = 2^m$ physical addresses of DRAM.

The address translation can be formalised by a function $\text{VMAP} : V \rightarrow P \text{ option}$. For some virtual address a , we have

$$\text{VMAP}(a) = \begin{cases} \text{Some } a' & \text{if data at virtual address } a \text{ is at physical address } a' \text{ in } P \\ \text{None} & \text{if data at virtual address } a \text{ is not in physical memory} \end{cases}$$

We can represent this simple model of the function VMAP by a data structure accessed by hardware (the MMU), and manipulated by software (the OS).

We could think of individually mapping each byte or word, but it is too complicated in terms of software, and too expensive in terms of hardware and data structure. Thus, the address spaces are divided into numbered *pages*, which are blocks of a specific size, e.g. 4 KiB. They are always aligned on *page size*, i.e. they cannot overlap. The mapping is therefore represented by a *page table*. In first approximation, it is a list of entries, one for each virtual page, specifying either `None` (when there is no mapping), or `Some p` where p is a physical page number.

1.4 VM as a tool for memory management

The key idea is that each process has its own virtual address space. It can view memory as a simple linear array of bytes. Mapping function scatters addresses through physical memory. Well-chosen mappings simplify memory allocation and management. Each virtual page can be mapped to any physical page. A virtual page can be stored in different physical pages at different times. This allows to share code and data among processes, by mapping virtual pages to the same physical page.

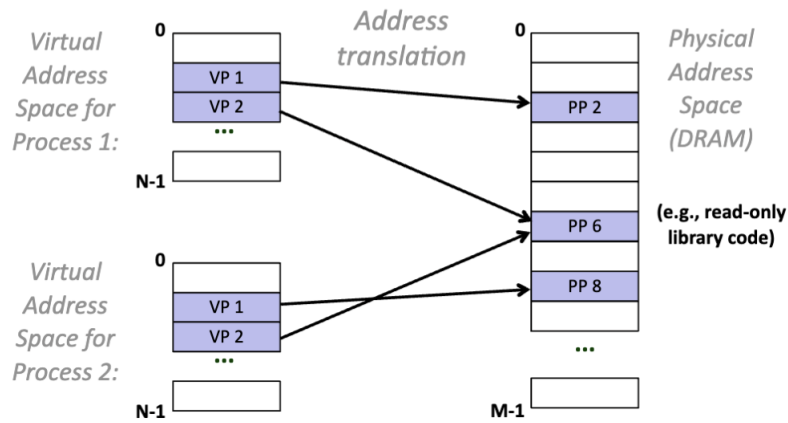


Figure 1: Address translation

1.5 Address translation

The following diagram shows the translation of addresses with a page table:

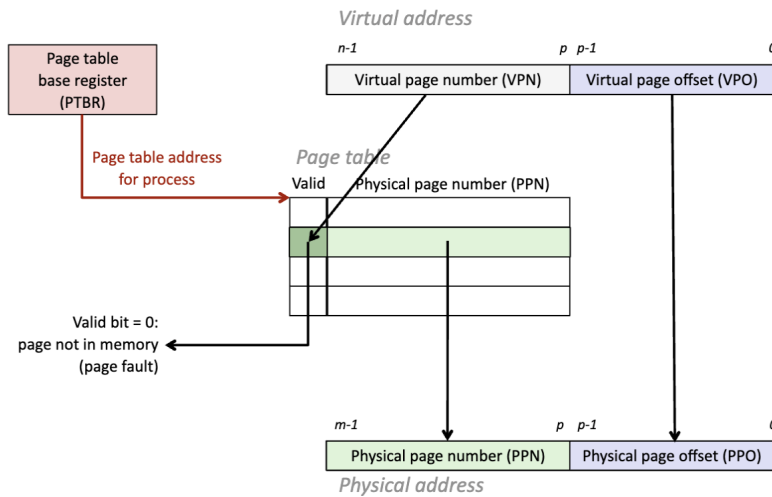


Figure 2: Use of a page table

The processor starts by sending a virtual address to the MMU, which requests a page table entry address from memory. The page table entry is fetched from page table in memory, and the MMU sends the physical address to cache or memory. Finally, the cache or memory sends the data word back to the processor. This is summarized in the following diagram:

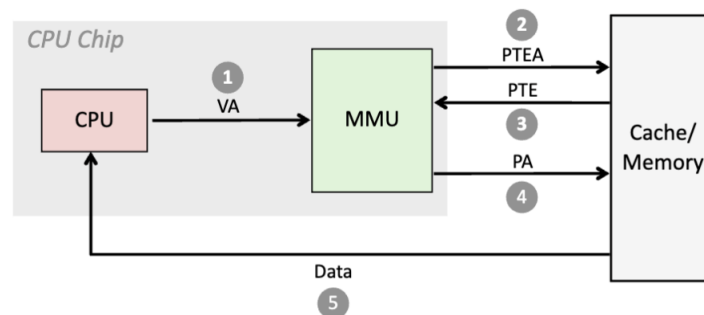


Figure 3: Page hit

Translation can be sped up using a *Translation Lookaside Buffer* (TLB). Page table entries (PTEs) are cached in L1 like any other memory word. PTE hit still requires a small L1 delay, and PTEs may be evicted by other data references. The solution to this is a *Translation Lookaside Buffer* (TLB). It is a small hardware cache in the MMU, which maps virtual page numbers to physical page numbers. It contains complete page table entries for a small number of pages.

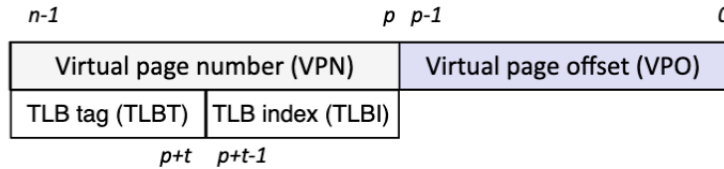


Figure 4: Use of a Translation Lookaside Buffer (TLB)

The modified hit procedure using TLB is therefore of the form:

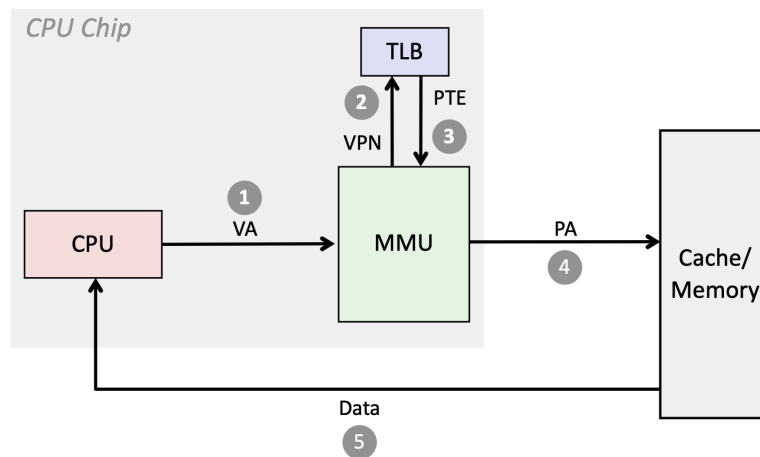


Figure 5: TLB hit

Notice that a TLB hit eliminates a memory access. A TLB miss incurs an additional memory access (the PTE). Fortunately, TLB misses are rare.

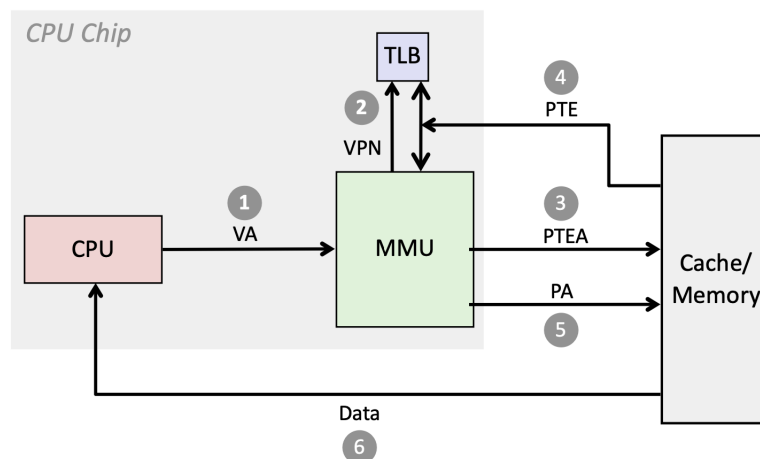


Figure 6: TLB miss

1.6 Kernel and User Mode

Each user process will have its own virtual address space defined by a distinct page table. But should user code be allowed to update page tables and virtual memory mappings directly? No! That would not provide any real protection. Therefore, a process has at least 2 operating modes: a *Kernel/Supervisor mode*, in which all instructions and registers are available, and a *User mode*, in which some instructions and registers are restricted. Special instructions or events (interrupts) allow to switch from one operating mode to the other. Virtual memory mappings can depend on the mode.

In practice, PTEs are extended with permission bits. Page permissions are checked for every move instruction to or from memory. If violated, the OS sends the process the error SIGSEGV (segmentation fault).

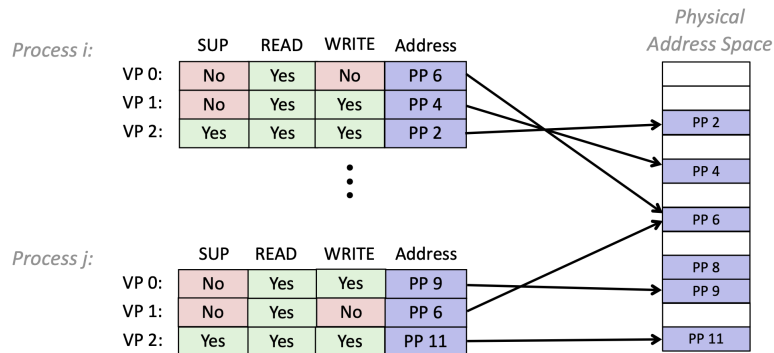


Figure 7: PTEs extended with permission bits

1.7 Multi-level page tables

Suppose that we maintain a 48-bit virtual address space, with a page size of 4 KiB (2^{12}), with an 8-byte PTE. We would need a 512 GiB page table:

$$2^{48} \times 2^{-12} \times 2^3 = 2^{39} \text{ bytes}$$

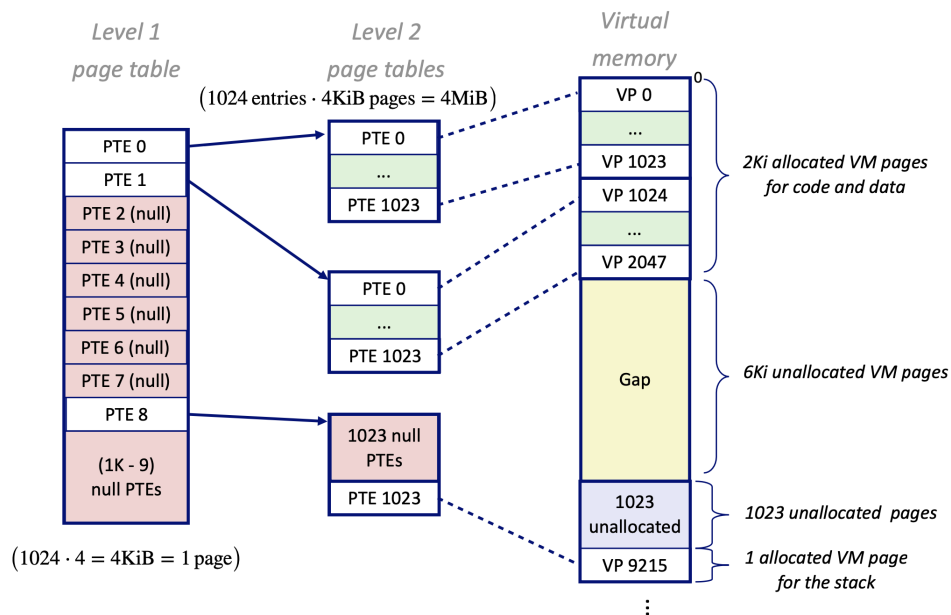


Figure 8: A two-level page table hierarchy

A solution to this is to use multi-level page tables. For example, we could use a 2-levels page table, where in the first level, each PTE points to a page table, and in the second level, each PTE points to a page. This reduces memory requirements. Usually, most of the virtual address space is unallocated. If a level 1 PTE is null, there is no corresponding level 2 page table. The level 2 page tables can be created as needed, reducing pressure on physical memory.

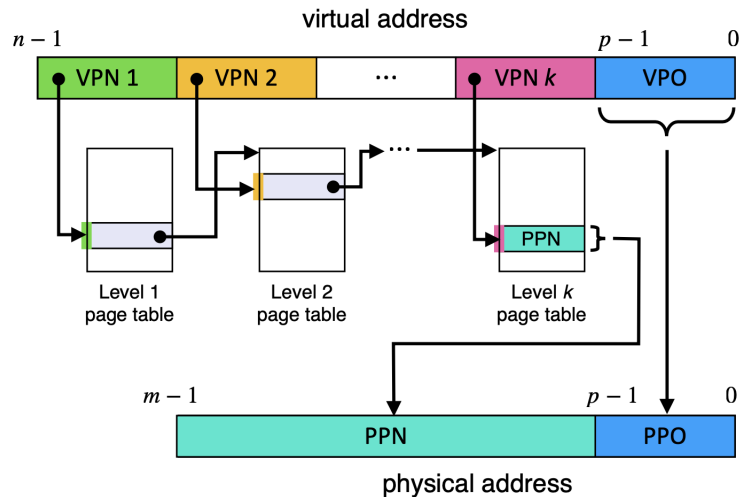


Figure 9: Address translation with a k -level page table

This may seem expensive and impractical, but the TLB helps by catching PTEs from the page tables at different levels, and in practice the lookup is not much slower than with a single level.