



## Open Source Programování

<http://rtime.felk.cvut.cz/osp/>

Pavel Píša

<pisa@fel.cvut.cz>

<http://cmp.felk.cvut.cz/~pisa>

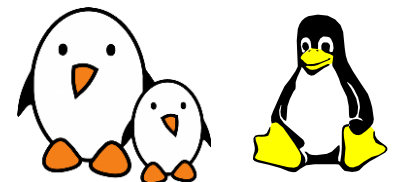
Michal Sojka

František Vacek

**DCE FEL ČVUT**



© Copyright 2004-2010, Pavel Píša, Michal Sojka, František Vacek,  
Free-Electrons.com, GNU.org, kernel.org, Wikipedia.org  
Creative Commons BY-SA 3.0 license Latest update: 7. V. 2015



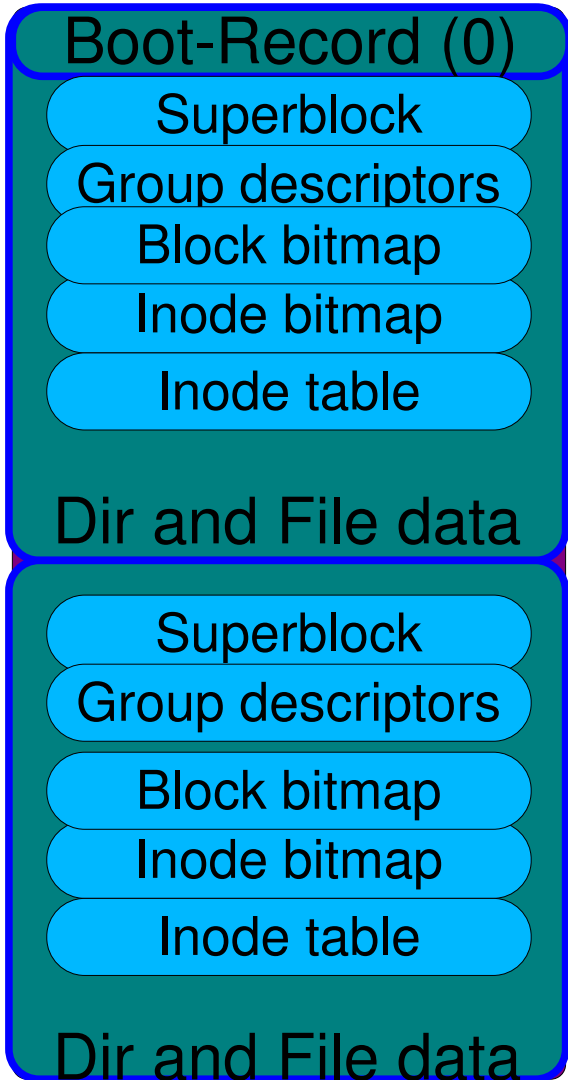
- ▶ Jednotné, přenositelné API
- ▶ Jednotný přístup k datům, periferiím, komunikaci
- ▶ Organizace dat a periferií v (hierarchických) jmenných prostorech
  - ▶ Souborový systém (FS) – open, read, write, close
  - ▶ Znaková a bloková zařízení většinou přístupná přes globální jmenný prostor (Unix/Linux /dev/\*, Windows NT+ \\.\, ...)
  - ▶ IPC (FIFO, Unix Socket, Shared MEM, Semaphores), privátní i sdílené, POSIX nevyžaduje pro vše jeden jmenný prostor s FS
  - ▶ Síťování a sockety – BSD jiný jmenný prostor ale ostatní operace shodné, Plan 9 vše jednotný prostor, Windows přes IOCTL
- ▶ Správa paměti, stránkování – Linux pagecache a swapcache
- ▶ Správa procesů, oddělení paměťových prostorů, ochrana paměti a hierarchie oprávnění - capabilities



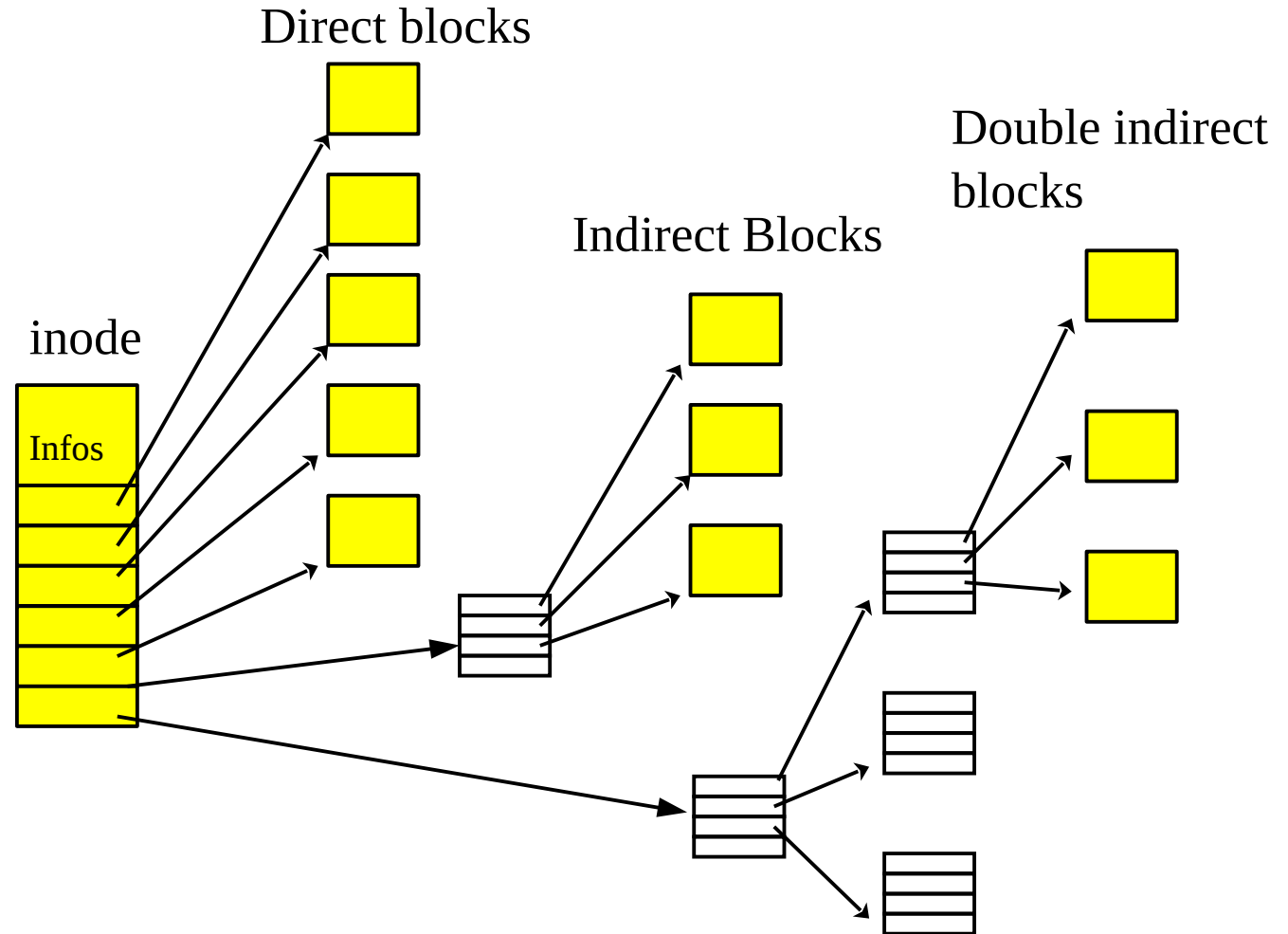
# Základní operace nad „souborem“



Unix/Linux	file_operations	Windows	C library	Java
open, creat O_NONBLOCK	open	CreateFile FILE_FLAG_OVE RLAPPED	fopen	java.nio.channels. FileChannel
close	release*	CloseHandle	fclose	Channel.close
read, aio_read	read, aio_read	ReadFile	fread, fprintf	ByteChannel.read
write, aio_read	write, aio_write	WriteFile	fwrite, fscanf	ByteChannel.write
ioctl	ioctl, unlocked_ioctl	DeviceIoControl	ioctl, tcgetattr, ...	No portable way
fsync	flush	FlushFileBuffers	fflush	
select, poll, (epoll, kevent)	poll	WaitForMultipleO bjects, select		java.nio.channels. Selector
opendir, fdopendir, readdir, closedir	readdir	FindFirstFile, FindNextFile		
fcntl, flockfile, funclockfile	lock	LockFile, UnlockFile		
mmap	mmap	CreateFileMappin g MapViewOfFile		java.nio.channels. FileChannel



## EXT2/3/4



Množství dalších FS pro bloková zařízení– FAT, NTFS, BTRFS, SquashFs, ISO9660  
 ale i virtuálních – dočasné TmpFS nebo správa systému sysfs, proc, cgroup, debugfs,  
 usbfs, fuse

- ▶ Na vyšší úrovni Java nebo `fopen("/home/franta/dopis.txt", "w+")`
- ▶ Po zpracování/alokaci stavových informací na vyšších úrovních dojde k volání `open("/home/franta/dopis.txt", O_RDWR|O_CREAT, S_IRWXU)`
- ▶ Volání dále prochází (G)LIBC/CRT `glibc/sysdeps/unix/sysv/linux/open64.c`  
`INLINE_SYSCALL (open, 3, file, oflag | O_LARGEFILE, mode);`
- ▶ Makro nahradí mnemonické určení služby „open“ za číslo systémového volání a předá řízení jádru OS. Číslo je nadefinováno v souboru `linux-3.x/include/uapi/asm-generic/unistd.h`  

```
#define __NR_open 1024
__SYSCALL(__NR_open, sys_open)
```
- ▶ Makro `INTERNAL_SYSCALL` je již architekturně závislé, pro 32-bit Intel x86 systém je definováno v souboru `glibc/sysdeps/unix/sysv/linux/i386/sysdep.h`  
`INTERNAL_SYSCALL(name, err, nr, args...)`  
`... LOADARGS_##nr; movl %1, %%eax; int $0x80; RESTOREARGS_##nr`
- ▶ Protože je volání přes IDT na x86 velmi neefektivní, preferuje se **SYSENTER**
- ▶ CPU přejde do systémového režimu (x86 Ring 3 → Ring 0) a volá vstupní bod jádra `system_call` v `linux-3.x/arch/x86/kernel/entry_32.S`, v této funkci je podle čísla systémového volání (předáno v EAX na x86) nalezena položka v tabulce **sys\_call\_table**

- ▶ **open** náleží do skupiny **souborových služeb** → implementace v adresáři *fs linux-2.6.32/fs/open.c*

```
SYSCALL_DEFINE3(open, const char __user *, filename, int,
flags, int, mode)
    ...; ret = do_sys_open(AT_FDCWD, filename, flags, mode);
```

- ▶ Protože není dopředu jasné, na jakém disku a souborovém (nebo virtuálním) filesystemu se dané adresáře nacházejí, jsou veškeré souborové služby nejdříve zpracovávány obecně - **VFS** (virtual filesystem). Když je cesta nalezena ve vyhledávacích strukturách (directory) **dcache**, je již znám filesystem. Je zkontrolována existence/vytvoření souboru, založena stavová struktura se stavem k jako konkrétnímu otevření (`struct file`) a je zavolaná „metoda“ daného filesystemu *linux-2.6.x/fs/ext3/file.c*:

```
struct file_operations ext3_file_operations = { ..., .open =
generic_file_open, ...};
```

- ▶ Protože specializace FS není v tomto případě potřeba je zavolaná obecná implementace
- ▶ Skutečná specializace většiny metod běžných filesystemů je totiž řešena až na úrovni souborových uzlů `inode` `static const struct address_space_operations ext3_{ordered|writeback|journalled}_aops`

- ▶ U znakových zařízení a virtuálních filesystemů dané „metodami“ `read` a `write` ve `struct file_operations`.
- ▶ U běžných souborů na disku je potřeba zajistit **paralelní přístup** přes **více otevření** souboru k jeho datům, která musí být nejdříve načtena někde do paměti. Zároveň paralelně běžící přístupy musí výsledek vidět tak, jako by přímo po bytech měnily data na disku.
- ▶ Úplně přímý **přístup** na disk **po bytech není možný**, i načtení, úprava a uložení bloku je extrémně pomalé, nejvíce času zabírá vystavování hlaviček. Je potřeba spravovat vyrovnávací paměť, přitom musí být zajištěné, že v paměti ke každým datům existuje jen jedna kopie a její modifikace mohou probíhat v podstatě paralelně a změny musí být okamžitě vidět pro ostatní přístupy (i `mmap`).
- ▶ Existují dvě logické možnosti, kam vložit vyrovnávací paměti – diskové buffery (buffer heads mezi diskem a FS) a **pagecache** mezi FS a aplikacemi.
- ▶ Původně Linux používal především diskové buffery, ale vzhledem k požadavku na možnost mapování souborů do paměťových prostorů (`mmap`) se data mohla objevit ve dvou kopiích (pro moderní VMM nepřijatelné) nebo se složitě po `mmapu` upravovaly buffer heads tak, aby směřovaly do stránek. Postupem času se proto zcela přešlo na **pagecache**.

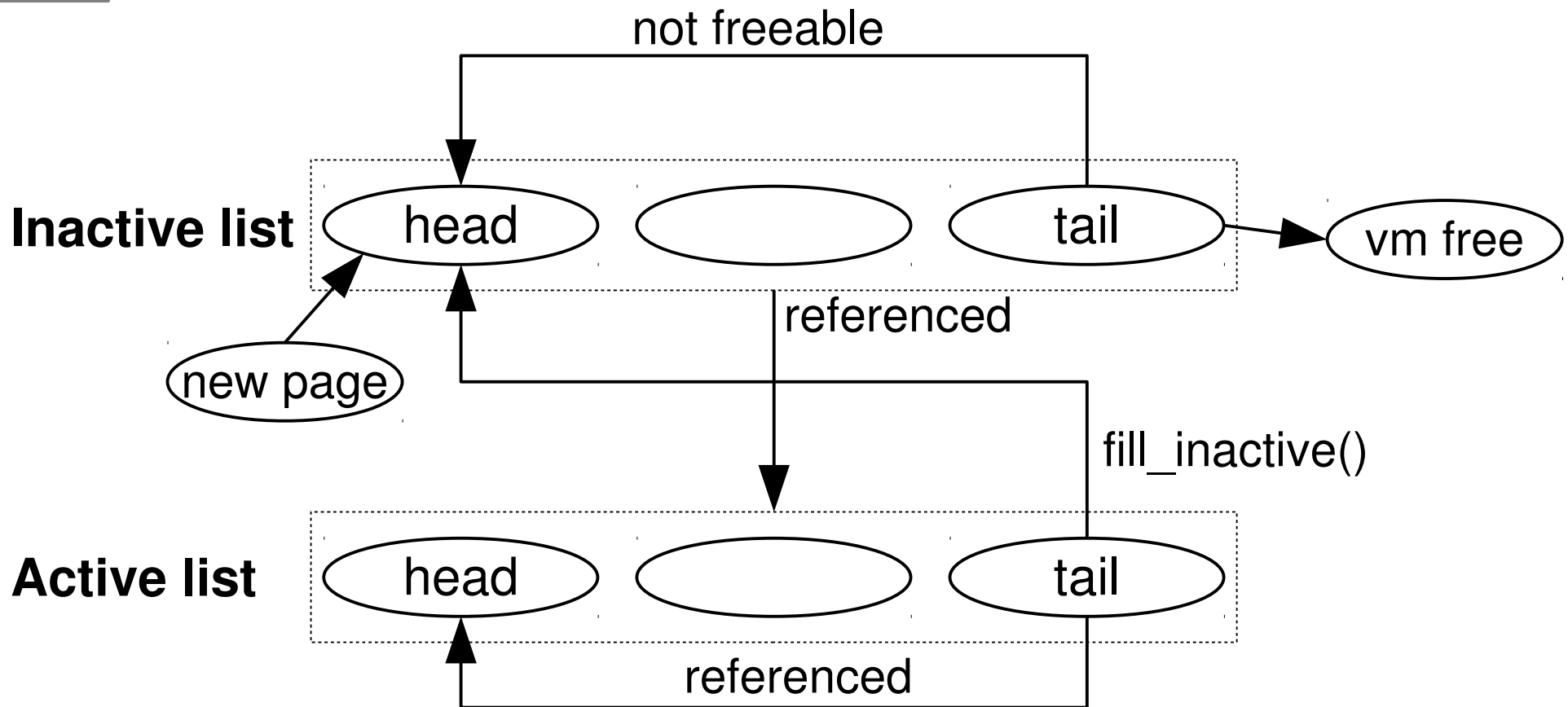
- ▶ Data nejsou v paměti udržovaná podle **pozice na disku**, ale podle **příslušnosti a offsetu** v daném **souboru (inode)**. Přitom jsou offsety a data zarovnány tak aby odpovídaly násobkům velikosti **paměťové stránky** (4 kB na x86).
- ▶ Protože buffer heads jsou používané opravdu jen pro přenosy dat z a na disk, musí být i data adresářů a množství dalších rozšiřujících informací filesystemem udržované v pagecache v interních stránkových sadách, které přísluší buď k adresáři nebo **superbloku** FS (globálnímu stavu filesystemu)
- ▶ Na první pohled to vypadá nevýhodně, ale umožňuje to úplnou koherenci/**konsistenci** mezi daty čtenými a modifikovanými voláními **read** a **write** a stránkami **mapovanými do adresních prostorů** (mm context) procesů. Téměř veškerá paměť může podle potřeby sloužit jako cache disků/souborů a užití se přizpůsobuje zátěži.
- ▶ Práce souborových systémů a základních operací je silně provázaná se správou paměti



- ▶ Informace o stavu obsazení fyzické paměti a odkládacích (Swap) oddílů a souborů
- ▶ Alokace volných stránek pro účely jádra a všech ostatních subsystémů
- ▶ V Linuxu slouží v podstatě většina fyzické paměti jako
  - ▶ stránková/pagecache pro zpřístupnění a mapování souborů, které se nalézají na souborových systémech, FS pak vlastně slouží k rozdělení kapacity blokových zařízení mezi soubory
  - ▶ pro anonymní privátní a sdílené stránky, které se odkládají na swap zařízení – obecně swapcache
- ▶ Hierarchie paměti (připomenutí látky z architektur počítačů):
  - zápisníková paměť – registry, běžná transparentní L1 až Lx cache, hlavní/fyzická paměť, soubory a data na disku

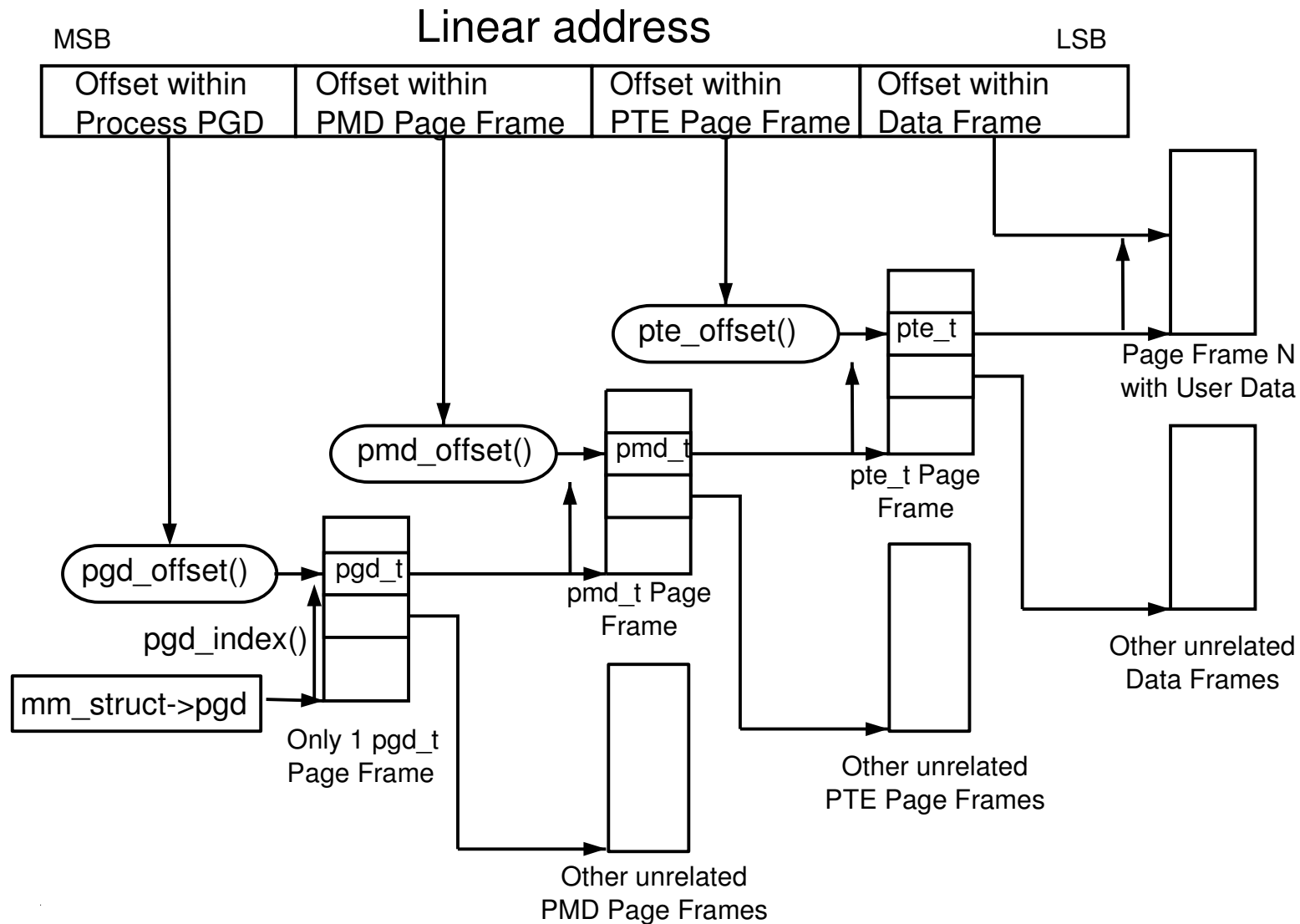
- ▶ Žádný úsek (stránky) spustitelného souboru/programu ani sdílených knihoven (.so, DLL) není „nikdy“ načten do paměti pokud ho není potřeba – do paměti je pouze mapován
- ▶ I pokud je stejná stránka potřeba ve více paměťových prostorech, je ve skutečnosti mapována do všech prostorů stejná stránka fyzické paměti nebo je ve všech prostorech položka stránkovacích tabulek invalidována a zpracování výpadku zajistí její natažení ze souboru (inode, backing store)
- ▶ Celkový součet paměťových požadavků může řádově přesahovat množství fyzické paměti. Nepoužívané (LRU) stránky, které odpovídají obsahu backing store jsou vyřazeny, modifikované sdílené stránky zapsány zpět a modifikované anonymní nebo privátně mapované stránky jsou odkládány na swap úložiště
- ▶ Pro výkonnost systému je kritický dobře navržený algoritmus pro výměnu stránek – nahrání stránek které jsou potřeba (způsobí výjimku nebo jsou odhadnuté pro readahead), které vyžaduje zajištění volného místa realizované vyhledáním a vyjmutím těch fyzických stránek, která již nejsou (s největší pravděpodobností) potřeba
- ▶ Kvůli sdílení stránek mezi paměťovými kontexty je potřeba nalézt vhodný algoritmus na invalidaci všech odpovídajících položek v stránkových tabulkách 10

- ▶ Existuje víc algoritmů pro správu stránek. Linux používá algoritmus založený na seznamech referencovaných a nereferencovaných stránek.
- ▶ Uvolnění stránky je odloženo až do doby kdy nastane potřeba alokovat stránky za jiným účelem.
- ▶ Je potřeba vyhledat stránku, která má být uvolněna. Vyhledává se stránka u které je nejmenší pravděpodobnost, že data v ní uložená by mohla být ještě potřeba. To se zjišťuje na základě toho, jestli byla daná stránka v nedávné době používána (nastavený bit 'accessed/referenced').
- ▶ Aby se nemusely prohledávat všechny stránky, jsou stránky organizovány do seznamů a prohledávají se jen stránky z konce „unreferenced“ seznamu.



Klíčovým bodem je ve spodní úrovni funkce `__get_free_pages()/alloc_pages()`, která alokuje volné stránky a vrací jejich adresu/ukazatel na struct `page` (stav fyzické stránky).

Skutečná implementace je složitější, neboť paměť je z důvodů omezenosti některých architektur a sběrnic dělena na zóny – `GFP_DMA`, `GFP_DMA32`, `__GFP_HIGHMEM`, `NORMAL`, a dále je definován způsob plnění požadavku – `GFP_NOWAIT`, `GFP_ATOMIC`, `GFP_NOIO`, `GFP_NOFS`, `GFP_KERNEL`, `GFP_TEMPORAR`.



PGD Page Global Directory  
PMD Page Middle Directory  
PTE Page Table Entry

Z výpisu je vidět, které paměťové oblasti VMA přísluší danému paměťovému kontextu/procesu

```
> cat /proc/1/maps (init process)
```

start	end	perm	offset	major:minor	inode	mapped file name
00771000	0077f000	r-xp	00000000	03:05	1165839	/lib/libselinux.so.1
0077f000	00781000	rw-p	0000d000	03:05	1165839	/lib/libselinux.so.1
0097d000	00992000	r-xp	00000000	03:05	1158767	/lib/ld-2.3.3.so
00992000	00993000	r--p	00014000	03:05	1158767	/lib/ld-2.3.3.so
00993000	00994000	rw-p	00015000	03:05	1158767	/lib/ld-2.3.3.so
00996000	00aac000	r-xp	00000000	03:05	1158770	/lib/tls/libc-2.3.3.so
00aac000	00aad000	r--p	00116000	03:05	1158770	/lib/tls/libc-2.3.3.so
00aad000	00ab0000	rw-p	00117000	03:05	1158770	/lib/tls/libc-2.3.3.so
00ab0000	00ab2000	rw-p	00ab0000	00:00	0	
08048000	08050000	r-xp	00000000	03:05	571452	/sbin/init (text)
08050000	08051000	rw-p	00008000	03:05	571452	/sbin/init (data, stack)
08b43000	08b64000	rw-p	08b43000	00:00	0	
f6fdf000	f6fe0000	rw-p	f6fdf000	00:00	0	
fefd4000	ff000000	rw-p	fefd4000	00:00	0	
ffffe000	ffffff00	---p	00000000	00:00	0	

`mmap` může být vhodnou alternativou k systémovým voláním `read`, `write`. A to především, pokud přístupy nejsou sekvenční, vlastní délka dat je malá – např. položky databáze, operace nad RAW obrázky nebo vykreslování bodů do grafického subsystému

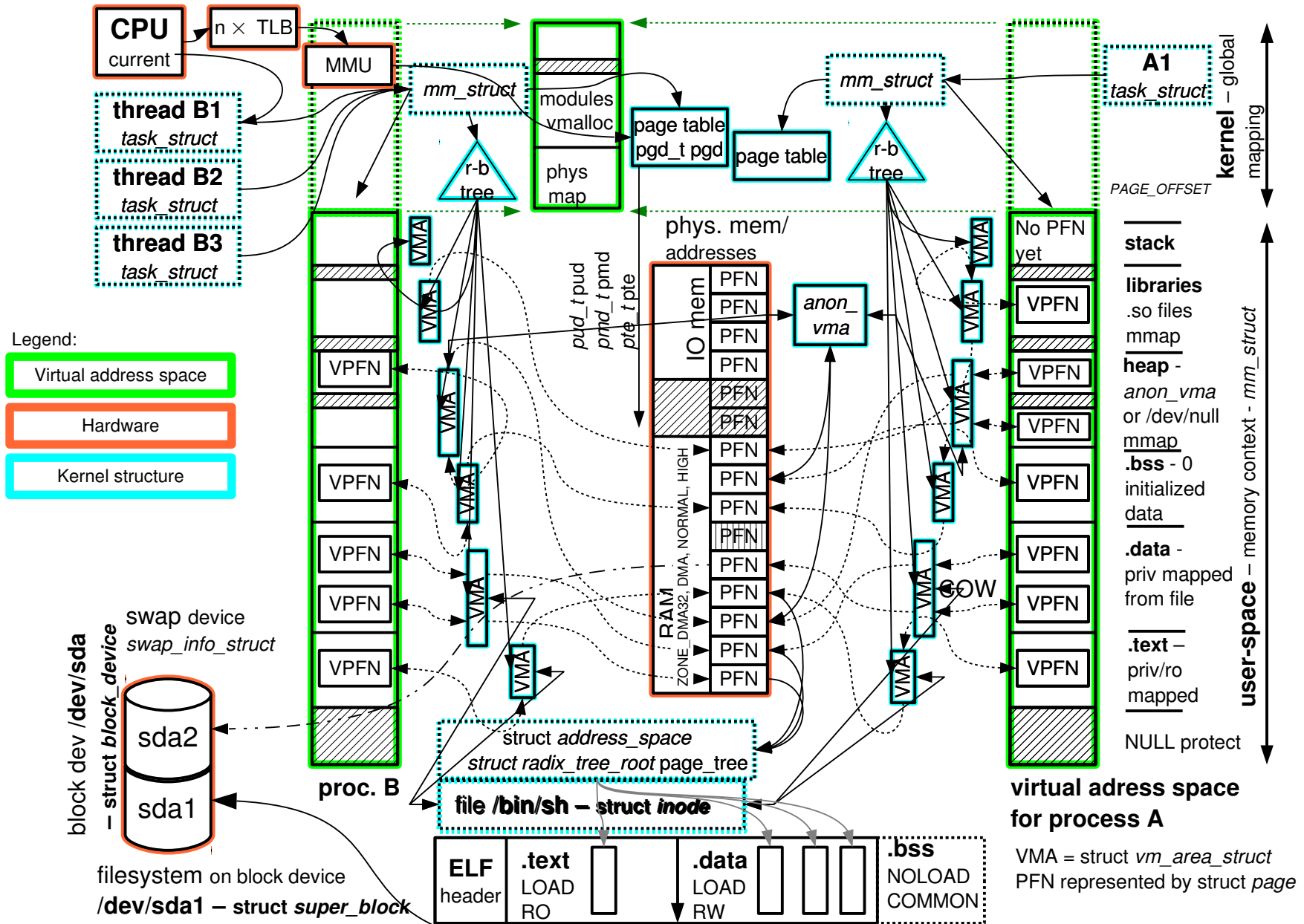
`X server` využívá `mmap` především k přístupu ke grafickým kartám

```
fd = open("/dev/mem",O_RDWR|O_SYNC);
pagesize=getpagesize();
mm = mmap(NULL, length, PROT_WRITE|PROT_READ, MAP_SHARED, fd, start_pfn*pagesize);
```

start	end	perm	offset	major:minor	inode	mapped file name
08047000	081be000	r-xp	00000000	03:05	310295	/usr/X11R6/bin/Xorg
081be000	081f0000	rw-p	00176000	03:05	310295	/usr/X11R6/bin/Xorg
...						
f4e08000	f4f09000	rw-s	e0000000	03:05	655295	/dev/dri/card0
f4f09000	f4f0b000	rw-s	4281a000	03:05	655295	/dev/dri/card0
f4f0b000	f6f0b000	rw-s	e8000000	03:05	652822	/dev/mem
f6f0b000	f6f8b000	rw-s	fcff0000	03:05	652822	/dev/mem

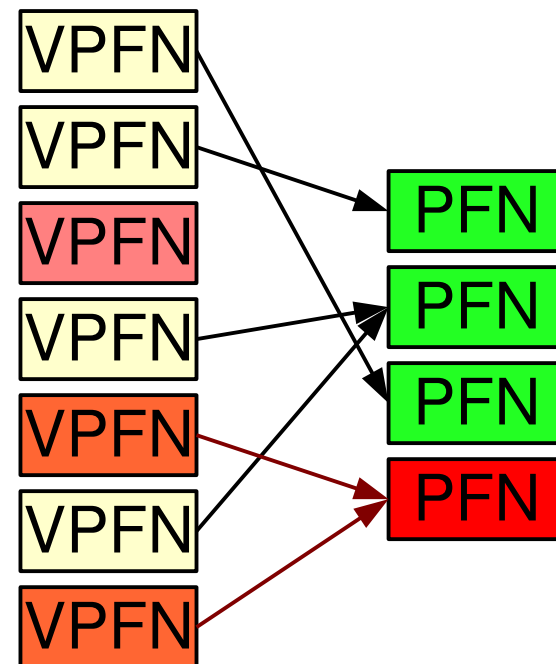
Pro výpis je možné použít i příkaz: `pmap <pid>`

See the description in the notes of the OpenDocument format





- ▶ Virtuální stránky (VPFN) z více a dokonce i z jednoho paměťového kontextu mohou odkazovat na jednu fyzickou stránku (PFN)
- ▶ Stránky virtuální paměti stojí pouze položku (PTE) ve stránkových tabulkách (4/8 byte) + něco málo ve vyšších úrovních tabulek (PGD, PUD, PMD) + popis rozsahu (*vm\_area\_struct*)

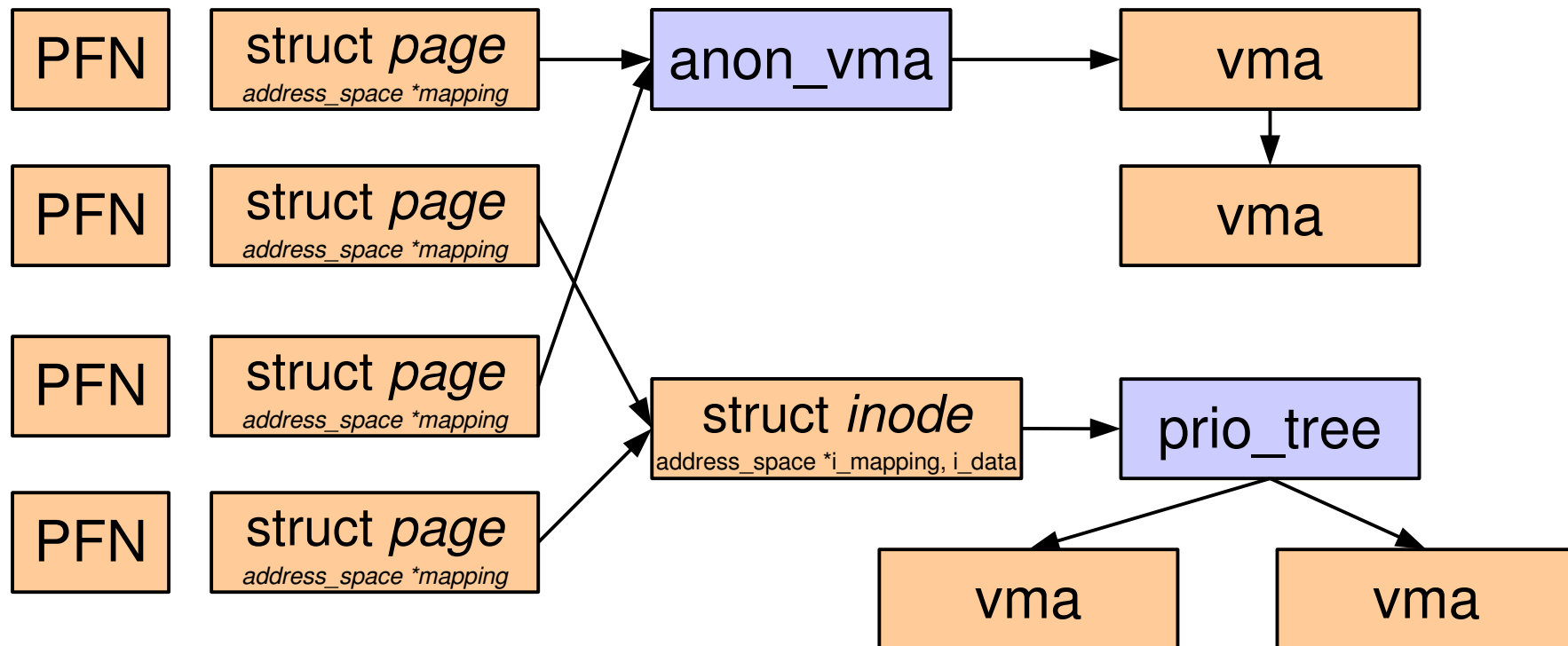


- ▶ Fyzické stránky jsou kritickým zdrojem, každá je popsána svojí struct *page* a referencovaná svým umístěním *page-frame-number*

$PFN = \text{virtual\_address} \gg \text{PAGE\_SHIFT}$

- ▶ Těžká operace je nalezení všech PTE k danému PFN. To je potřeba ke sledování referencí a při uvolnění stránky k zneplatnění všech PTE.
- ▶ Pokud by se seznam držel ke každé stránce, tak vyjde 8 byte na položku seznamu, přitom na x86 32-bit je k dispozici je ~900M lowmemory
- ▶ Při 1000 proc. mapujících 2G sdílené (shared) paměti by seznamy vyšly na  $1000 * 2 * 1024 * 1024 * 1024 / 4096 * 8 / 1024 / 1024 / 1024 = 3.9 \text{ GiB}$

- ▶ Rešení: objrmap + anon\_vma + prio\_tree



- ▶ Dohledání konkrétních PTE položek v stránkových tabulkách je již formalita, když již známe VMA a tím i *mm\_struct*, která odkazuje na stránkovou tabulku konkrétního procesu

```
struct vm_area_struct {
    struct mm_struct * vm_mm;          /* The address space we belong to. */
    unsigned long vm_start;           /* Our start address within vm_mm. */
    unsigned long vm_end;             /* The first byte after our end address
                                       within vm_mm. */

    /* linked list of VM areas per task, sorted by address */
    struct vm_area_struct *vm_next;

    pgprot_t vm_page_prot;           /* Access permissions of this VMA. */
    unsigned long vm_flags;          /* Flags, see mm.h. */

    struct rb_node vm_rb;

    union {
        struct {
            struct list_head list;
            void *parent; /* aligns with prio_tree_node parent */
            struct vm_area_struct *head;
        } vm_set;

        struct raw_prio_tree_node prio_tree_node;
    } shared;

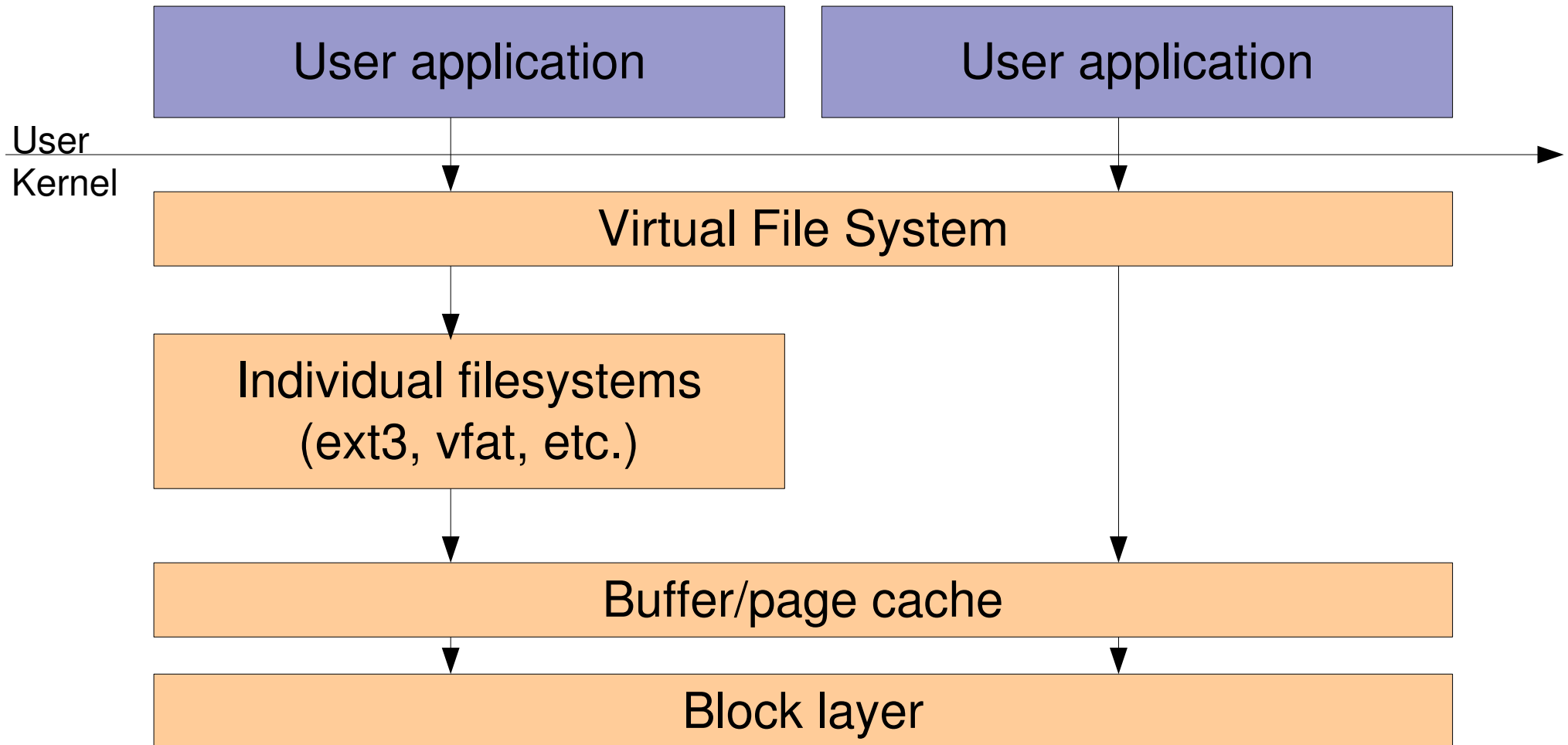
    struct list_head anon_vma_node; /* Serialized by anon_vma->lock */
    struct anon_vma *anon_vma;     /* Serialized by page_table_lock */

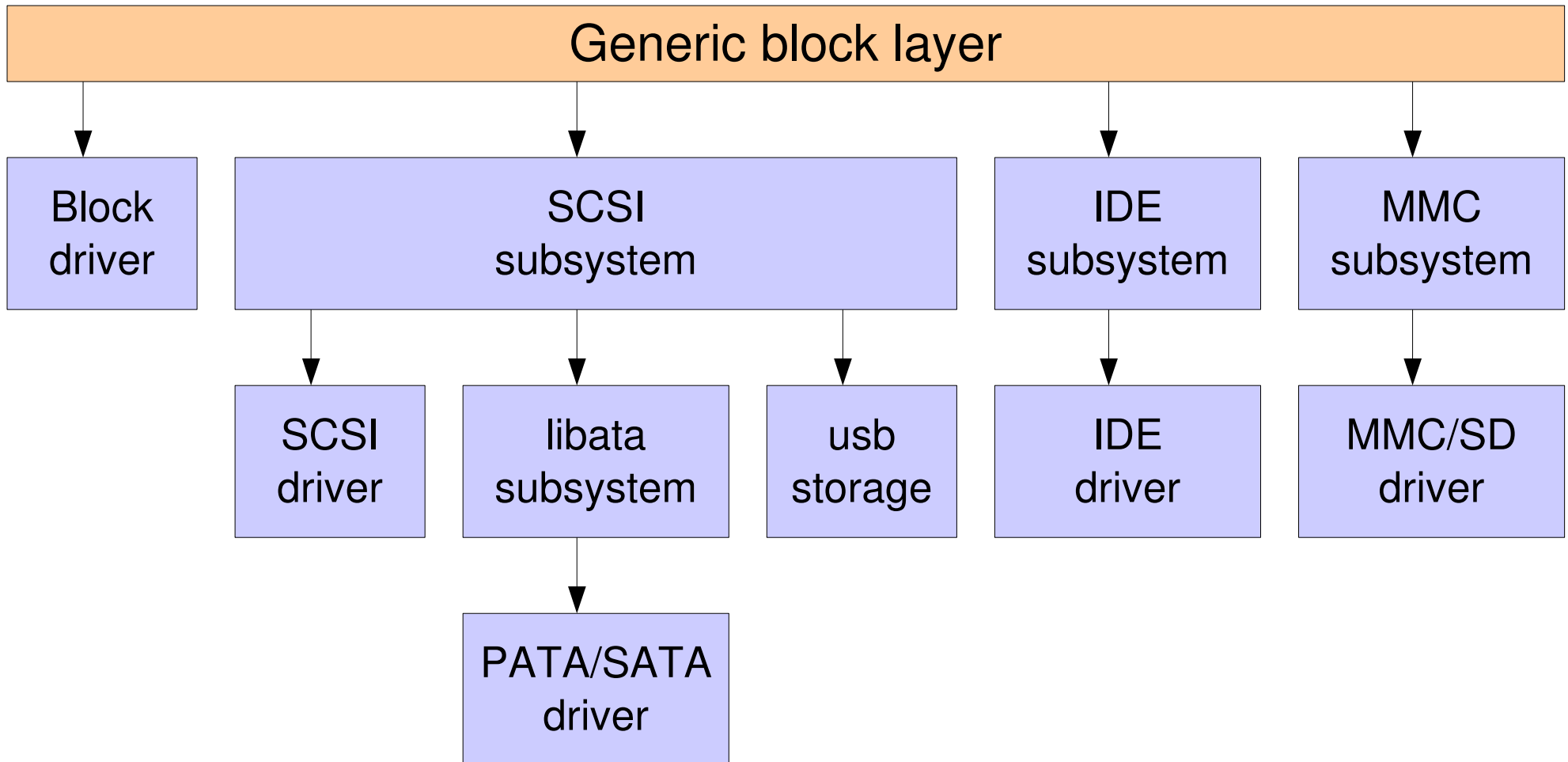
    /* Function pointers to deal with this struct. */
    const struct vm_operations_struct *vm_ops;

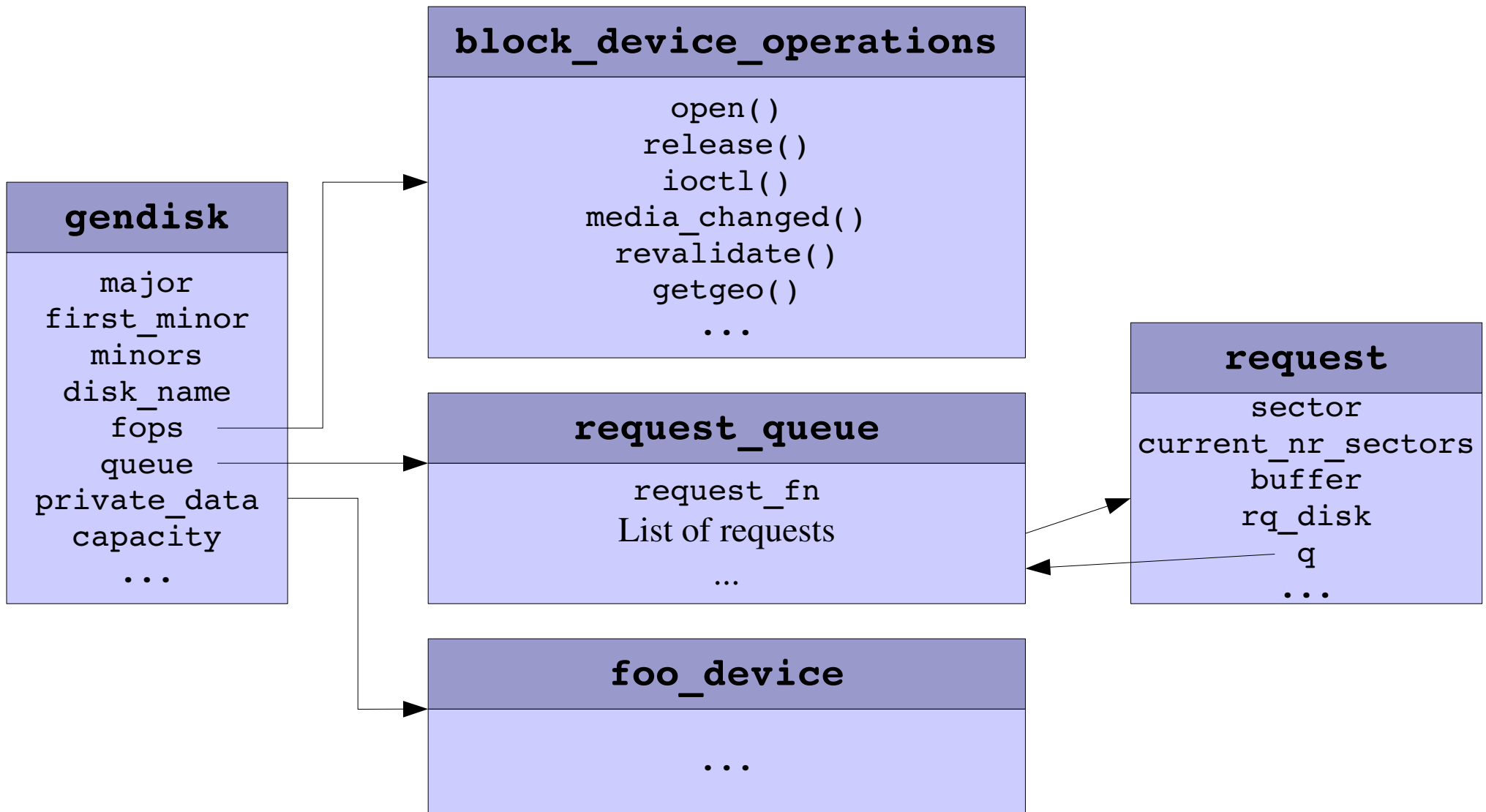
    /* Information about our backing store: */
    unsigned long vm_pgoff;         /* Offset (within vm_file) in PAGE_SIZE
                                       units, *not* PAGE_CACHE_SIZE */
    struct file * vm_file;          /* File we map to (can be NULL). */
    void * vm_private_data;         /* was vm_pte (shared mem) */
    unsigned long vm_truncate_count; /* truncate_count or restart_addr */
};
```

A file's MAP\_PRIVATE vma can be in both i\_mmap tree and anon\_vma list, after a COW of one of the file pages. A MAP\_SHARED vma can only be in the i\_mmap tree. An anonymous MAP\_PRIVATE, stack or brk vma (with NULL file) can only be in an anon\_vma list.

For areas with an address space and backing store, linkage into the address\_space->i\_mmap prio tree, or linkage to the list of like vmAs hanging off its node, or linkage of vma in the address\_space->i\_mmap\_nonlinear list.







Současné výkonné systémy, jejichž veškerá paměť je pod správou jednoho jádra operačního systému (single system image)

	SGI UV 2000	SGI UV 20
CPU Speed (Cores)	Intel® Xeon® processor E5-4600 product family 2.4GHz-3.3GHz	Intel® Xeon® processor E5-4600 product family 2.4GHz-3.3GHz
Min/Max Sockets	4/256	2/4
Min/Max Cores (Threads)	32/2048 (4096)	8/48
Max Memory	64TB	1.5TB
Interconnect	NUMALink® 6	Intel® Quickpath
Enclosure	10U rackmount	2U rackmount
Rack Size	Standard 19" Rack	Standard 19" Rack



Je možné, že současné klasické souborové systémy, založené na koncepci blokových zařízení a přenosu části dat do pagecache budou pro budoucí počítačové systémy zcela nevhodné

Již v současné době jsou vyvíjené a připravené nové paměťové technologie, které umožňují přístup po bytech, mohou být přímo mapované do fyzického adresního prostoru CPU a jsou téměř tak rychlé jako klasické SDRAM čipy. Pro přístup z aplikací tedy není nutné data kopírovat. Ovšem použití jako klasických PFN je problematické, protože paměti je příliš mnoho a pokud se použije i pro servisní struktury, tak bude docházet k jejím fyzickému opotřebení. Zajímavý pohled na problematiku je popsán například v článku

**XFS: There and back ... and there again?** <https://lwn.net/Articles/638546/>

But we also need to be thinking a little further ahead. Looking at the progression of capacities and access times for "spinning rust" shows 8GB, 7ms drives in the mid-1990s and 8TB, 15ms drives in the mid-2010s. That suggests that the mid-2030s will have 8PB (petabyte, 1000 terabytes) drives with 30ms access times.

The progression in solid-state drives (SSDs) shows slow, unreliable, and "damn expensive" 30GB drives in 2005. Those drives were roughly \$10/GB, but today's rack-mounted (3U) 512TB SSDs are less than \$1/GB and can achieve 7GB/second performance. That suggests to him that by 2025 we will have 3U SSDs with 8EB (exabyte, 1000 petabytes) capacity at \$0.1/GB.

persistent memory

NVDIMM battery-backed DIMMs 8 GB

400GB

Memristors



- ▶ Mel Gorman: Understanding The Linux Virtual Memory Manager 2004, ISBN 0-13-145348-3, <http://www.skynet.ie/~mel/projects/vm/guide/pdf/understand.pdf> již relativně staré. Především informace o začátku 2.6.x
- ▶ LinuxMM <http://linux-mm.org/>
- ▶ Linux Weekly News <http://lwn.net/Kernel/Index/>
- ▶ Free Electrons <http://free-electrons.com/docs/>