Operační systémy 2

Implementace souborových systémů II.

Petr Krajča

Katedra informatiky
Univerzita Palackého v Olomouci

22. listopad, 2011

NTFS: Úvod

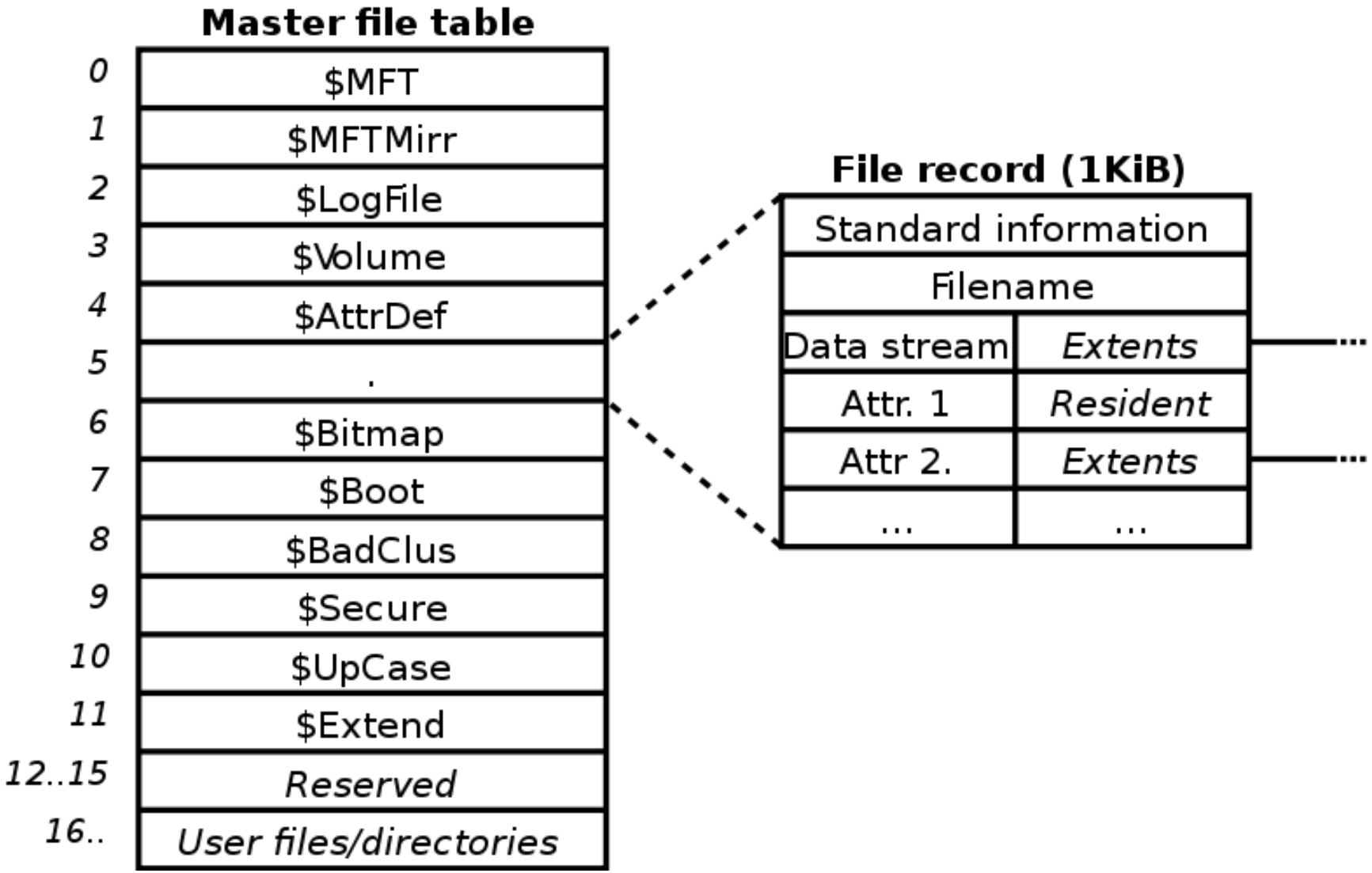
- hlavní souborový systém Windows NT
- kořeny v OS/2 a jeho HPFS (vyvíjen od roku 1993)
- velikost clusteru podle velikosti svazku (512 B–4 KB) $\Longrightarrow$ max. velikost disku 256 TB max. velikost souboru 16 TB
- oproti FAT (souborovému systému W9x) ochrana před poškozením + práva
- žurnálování a transakce
- podpora více streamů v jednom souboru
- dlouhé názvy (255 znaků) + unicode
- podpora standardu POSIX; hardlinky, symlinky
- komprese a řídké soubory

Adresáře

- opět technicky soubory; jména v B+ stromech
- některá metadata souborů jsou součástí adresáře

NTFS: Struktura disku (1/2)

- na začátku disku: boot sector
- 12 % MFT (Master File Table); 88 % data souborů
- MFT je soubor popisující všechny soubory na FS (MFT je taky soubor)
- MFT se skládá ze záznamů o velikosti 1 KB
- každý soubor je popsán tímto záznamem
- 32 prvních souborů má speciální určení ($MFT, $MFTMirr, $LogFile, $Volume, $Bitmap, $Boot, $BadClus, . . .)
- informace o souborech včetně jména, časů, atd. uloženy jako záznam v MFT jako dvojice *atribut-hodnota*
- tělo souboru je taky atribut $\Longrightarrow$ uniformní přístup; možnost uložit malé soubory přímo do MFT
- alternativní proudy $\Longrightarrow$ opět atributy
- v případě potřeby může jeden soubor zabrat víc záznamů v MFT
- případně lze použít místo mimo MFT (rezidentní a nerezidentní atributy)


Master file table
0
$MFT
1
$MFTMirr
2
$LogFile
3
$Volume
4
$AttrDef
5
.
6
$Bitmap
7
$Boot
8
$BadClus
9
$Secure
10
$UpCase
11
$Extend
12..15
Reserved
16..
User files/directories
File record (1KiB)
Standard information
Filename
Data stream
Extents
Attr. 1
Resident
Attr 2.
Extents
...
...

NTFS: Struktura disku (2/2)

- data v souboru jsou popsána pomocí (atributu) tabulky mapující VCN (virtual cluster number) na LCN (logical cluster number)
- VCN – číslo clusteru v souboru (indexováno od nuly)
- LCN – číslo clusteru ve svazku
- každý záznam v tabulce je ve tvaru: VCN, LCN, počet clusterů, např.

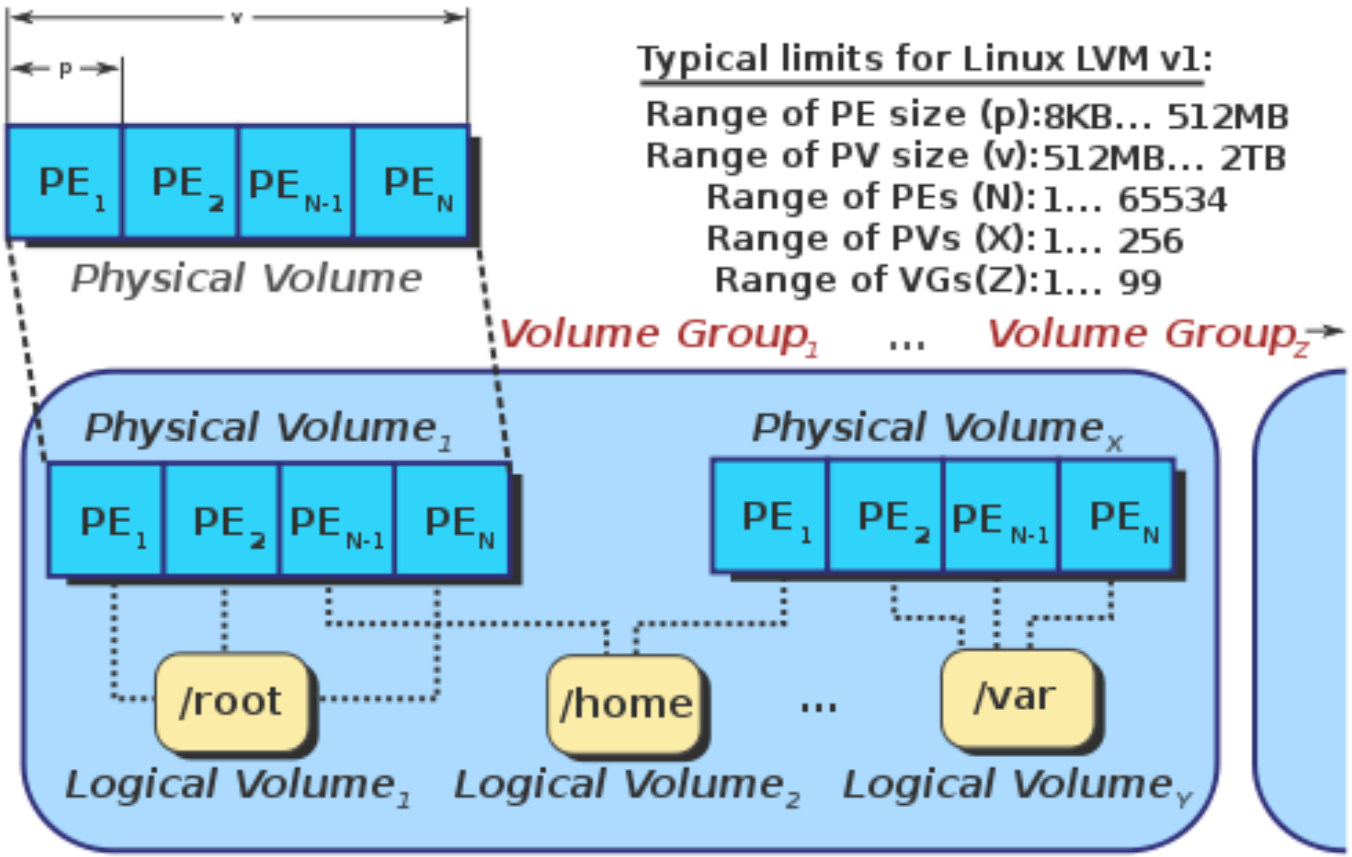
VCN	LCN	počet
0	42	4
4	123	8
32	456	15

Komprese

- řídké soubory
- možnost transparentně komprimovat obsah (vždy po 16 clusterech) $\Longrightarrow$ bloky dat zarovnány na 16 clusterů; pokud zabírá míň místa je komprimován
- čtení i zápis provádí (de)kompresi (LZ77) $\Longrightarrow$ dopad na výkon

LVM: Logical Volume Management

- problém: svazky mají pevnou velikost (rozdělení disku je pevně dané)
- řešení: logical volume management—vrstva mezi blokovým zařízením a FS
- fyzické disky (PV: physical volumes) rozdělen na rozasahy (PE: physical extents)
- jednotlivé PE poskytnuty do společné Volume Group
- odtud jsou pak přidělovány jednotlivým logickým svazkům $\Longrightarrow$ možnost dynamicky měnit velikost svazku $\Longrightarrow$ nutná podpora FS
- možnost emulovat RAID
- možnost vložit vrstvu, která se bude starat o snapshoty/klony (CoW)
- možnost transparentně provádět kódování
- ve Windows implementace podobná: Logical Disk Manager & Volume Snapshot Service (umožňují SW RAID); spolupráce s FS
- někdy dodáván jako software třetích stran


v
p
PE1
PE2
PEN-1
PEN
Physical Volume
Typical limits for Linux LVM v1:
Range of PE size (p):8KB... 512MB
Range of PV size (v):512MB... 2TB
Range of PEs (N):1... 65534
Range of PVs (X):1... 256
Range of VGs(Z):1... 99
Volume Group1
...
Volume GroupZ
Physical Volume1
Physical VolumeX
/root
/home
/var
Logical Volume1
Logical Volume2
Logical VolumeY

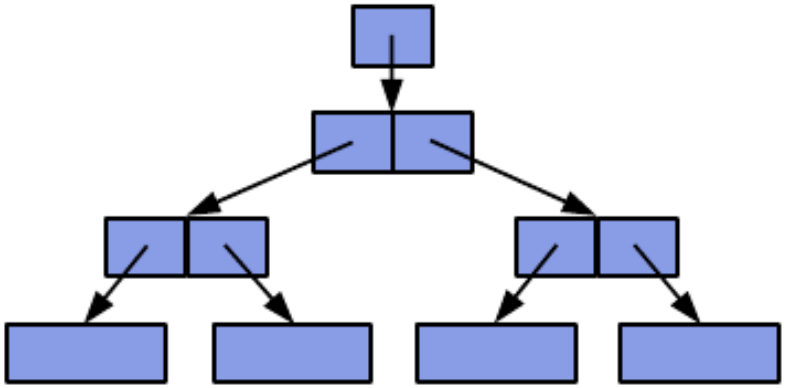
ZFS (1/2)

- moderní souborový systém (r. 2005); SUN (Oracle)
- podpora (open)Solaris, FreeBSD, NetBSD, MacOS X?, Linux (licenční problémy)
- kombinuje prvky LVM, RAID
- interně 128 bitová adresace (max. kapacita 256 ZB, ostatní limity kolem 16 EB)
- disky jsou spojeny do *poolu*, FS dělá automatický stripping $\Longrightarrow$ rozprostře se přes všechny disky
- bloky dat různých velikostí
- little- a big-endian (podle aktuální situace)
- ditto blocks (zdvojené zápisy)
- deduplikace
- podpora komprese

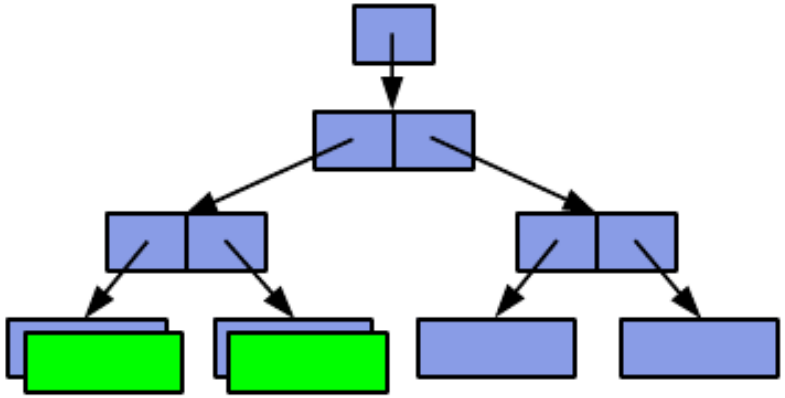
ZFS: Konzistence (2/2)

- RAID-Z: podobný RAID-5, ale má různě velké bloky (odpovídají logickým blokům) $\Longrightarrow$ např. 3 bloky dat + 1 paritní, atd.
- u dat jsou evidovány kontrolní součty $\Longrightarrow$ ochrana proti tichému poškození (chyba HW i SW)
- konzistence založena na metodě Copy-on-Write
- používaná data nikdy nejsou přepsána $\Longrightarrow$ nejdřív jsou zapsána data a pak jsou (atomicky) změněna metadata
- $\Longrightarrow$ výhodné slučovat operace do transakcí
- $\Longrightarrow$ FS je vždy v konzistentním stavu
- $\Longrightarrow$ infrastruktura pro vytváření snapshotů/klonů souborového systému

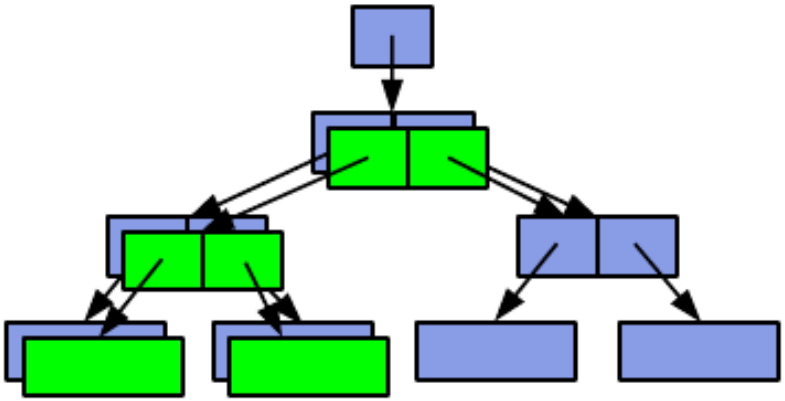
1. Initial block tree



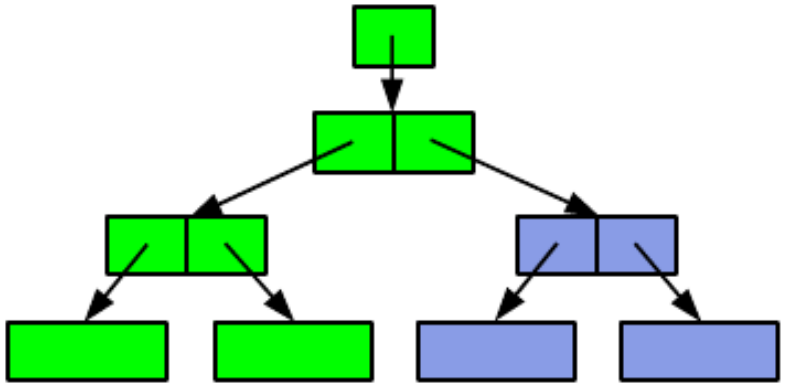
2. COW some blocks



3. COW indirect blocks



4. Rewrite uberblock (atomic)



ISO-9660

- souborový systém pro CD-ROM; podpora všech OS
- zápis jen jednou; sekvenční čtení $\Longrightarrow$ není potřeba dělat kompromisy
- logický sektor 2048 B (může být i větší)
- na disku může být víc logických svazků; svazek může být na více discích
- na začátku 16 rezervovaných bloků + 1 blok (Primary Volume Descriptor) $\Longrightarrow$ informace o disku; odkaz na kořenový adresář
- adresář popsán pomocí záznamů proměnlivé délky (viz Tan. 432)
  - textová data v ASCII
  - binární 2× (little- i big-endian)
- možnosti formátu určeny úrovněmi a rozšířeními
- **Level 1** – soubory 8.3; všechny soubory spojité; 8 úrovní adresářů
- **Level 2** – jména až 31 znaků
- **Level 3** – nespojité soubory (jednotlivé souvislé bloky se mohou opakovat)

ISO-9660: Rozšíření

Rock Ridge

- kompatibilita s unixy
- přidává dlouhá jména
- neomezené zanoření adresářů
- unixová oprávnění
- podpora symbolických odkazů; možnost mít na disku soubory zařízení

Joliet

- kompatibilita s Windows
- přidává dlouhá jména + podporu Unicode
- neomezené zanoření adresářů;