Moderní souborový systém - XFS

Jaroslav Velíšek

Struktura XFS

Allocation Groups

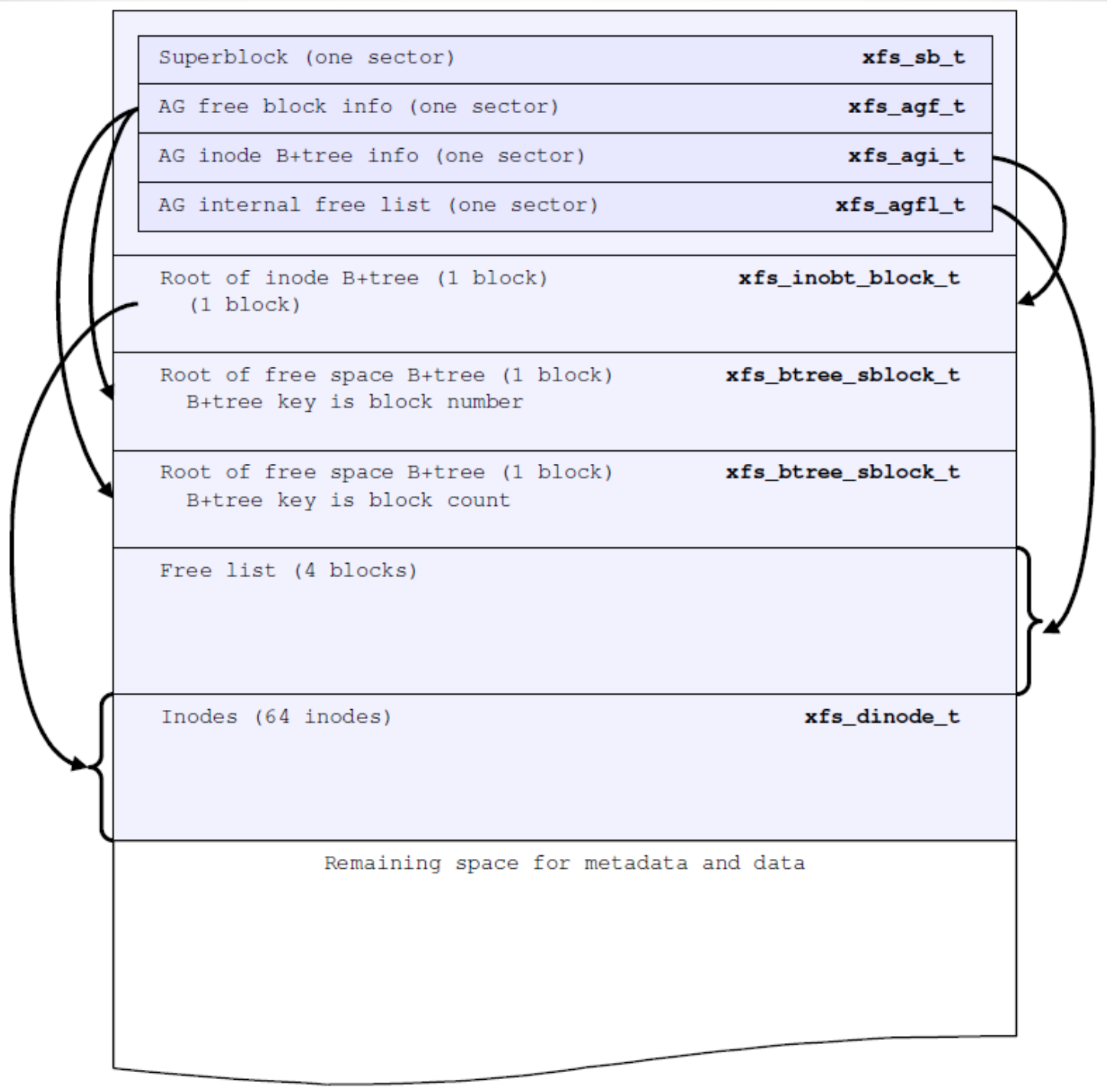
AG Free Space Management

AG I-node Management

On-disk I-node

Allocation Groups (AG)

- Lze téměř považovat za individuální souborový systém
- Řídí samostatně svou vlastní skupinu i-uzlů a volné místo
- Poskytují škálovatelnost a paralelismus
- Velikost AG musí být minimálně 16 MB a maximálně necelý 1TB
- V prvním (primárním) AG je volné místo v souborovém systému a celkový počet i-uzlů.


Superblock (one sector)
xfs_sb_t
AG free block info (one sector)
xfs_agf_t
AG inode B+tree info (one sector)
xfs_agi_t
AG internal free list (one sector)
xfs_agfl_t
Root of inode B+tree (1 block)
(1 block)
xfs_inobt_block_t
Root of free space B+tree (1 block)
B+tree key is block number
xfs_btree_sblock_t
Root of free space B+tree (1 block)
B+tree key is block count
xfs_btree_sblock_t
Free list (4 blocks)
Inodes (64 inodes)
xfs_dinode_t
Remaining space for metadata and data

Superblock

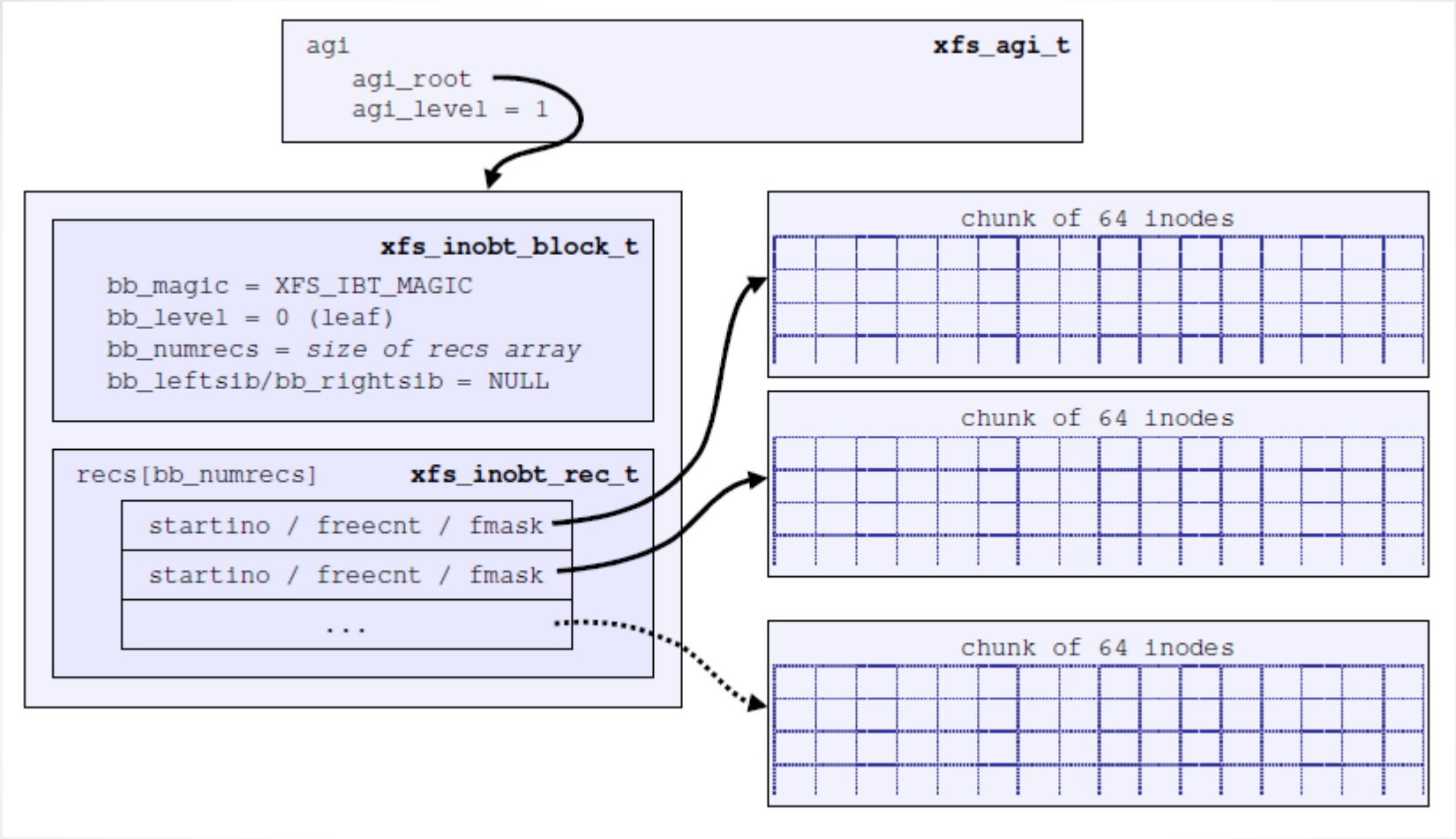
- První sektor AG
- Každý AG začíná superblokem
- Ukládá souhrnné informace o AG

AG Free Space Management

- XFS sleduje volné místo v AG pomocí páru B+stromů
  - 1. sleduje prostor číslem bloku (index startovního bloku volného regionu)
  - 2. sleduje prostor podle délky volného regionu
- AG Free Space Block (AGF)
  - informace o dvou Free Space B+trees a přidružuje tyto informace pro AG
- AG Free Space B+trees
  - ukládá seřazené pole offset bloku a počet bloků do listů B+Stromu
  - První B+strom řadí dle offsetu, druhý dle počtu nebo velikosti
- AG Internal Freelist (AGFL)
  - jedná se o pole relativních ukazatelů bloku na vyhrazený prostor pro nárůst Free Space B+trees

AG I-node Management

- Dynamické alokování i-uzlů
- AG inode B+tree (AGI)
  - Informace o uzlech alokační skupiny
- Root of Inode B+tree
  - I-uzly jsou alokovány ve skupinách po 64 i-uzlů
  - B+stromy se používají ke sledování těchto skupin i-uzlů (jak jsou přidělené resp. uvolněné)


agi
xfs_agi_t
agi_root
agi_level = 1
xfs_inobt_block_t
bb_magic = XFS_IBT_MAGIC
bb_level = 0 (leaf)
bb_numrecs = size of recs array
bb_leftsib/bb_rightsib = NULL
recs[bb_numrecs]
xfs_inobt_rec_t
startino / freecnt / fmask
startino / freecnt / fmask
...
chunk of 64 inodes
chunk of 64 inodes
chunk of 64 inodes

On-disk I-node

- I-node core
  - Obsahuje tradiční UNIX i-node metadata – vlastník, skupina, počet bloků, časové údaje a několik specifických XFS doplňků (např. projekt ID)
- Data fork
  - Obsahuje běžné údaje vztahující se i i-uzlu (Regular Files, Directories, Symbolic Links, Other File Types)
- Extended attribute fork
  - Obsahuje rozšířené atributy – není součástí rozhraní POSIX, ale je podporováno všemi moderními OS

Specifikace a rozšiřující technologie XFS

Snapshot
Žurnálování
Kvóty
Limity
Kontrola konzistence
Backup & Recovery
Další

Snapshot

- XFS neposkytuje přímou podporu pro snapshot
- XFS pracuje se standardním LVM
- Při vytváření LVM snapshotu není nutné provádět freeze/unfreeze

Žurnálování

- pouze metadata
- Max velikost žurnálu je 128MB
- Oprava - automaticky při mount FS
- Oprava trvá v řádu sekund a není závislá na vel. fs
- Při neočekávaném vypnutí stroje se ze souborů v tu dobu otevřených pro zápis stanou prázdné soubory
- Ve spojení s XFS se doporučuje používat pole RAID a UPS

Kvóta

- Kvóty poskytují metody pro správu disku
- Kvóty řídí tyto prostředky dvěma způsoby:
  - Správou místa na disku (bloky)
  - Správou počtu souborů (i-uzly)
- Tyto zdroje mohou být spravovány na jednotlivé:
  - Uživatele (User quotas)
  - Skupiny uživatelů (Group quotas)
  - Adresář (Project quotas)

  - PQ a GQ nemohou být použity zároveň, ale PQ a UQ mohou

Limity XFS

- Maximální velikost souboru
  - 9 EB
- Maximální velikost souborového systému
  - 18 EB
- Maximální délka názvu souboru
  - 255B
- Povolené znaky v názvu souboru
  - Všechny kromě null

Kontrola konzistence

- Pro kontrolu konzistence se provede příkaz xfs_check
- xfs_check zkontroluje všechny struktury metadat a zjistí nekonzistentnost
- xfs_repair skenuje souborový systém a opravuje vzniklé problémy
  - diskový oddíl nesmí být připojen pro zápis
  - může být připojen v režimu „pouze pro čtení“
- Poškozené soubory jsou ukládány do /lost+found

Backup & Recovery

- XFS nabízí dvě utility – xfsdump a xfsrestore
- xfsdump umožňuje zálohovat pouze „live“ filesystem
- Pokud dojde k přerušení (náhodné/úmyslné) zálohy lze ji opět obnovit (xfsdump i xfsrestore)
- Archivuje všechny typy unixových souborů
- xfsdump nabízí vysoký výkon zálohování
- Podpora inkrementální zálohy
- xfsrestore podporuje interaktivní operace – lze vybrat část souborů/adresářů pro obnovu

Další

- GROW/SHRINK
  - zmenšit partition (online) není v XFS možné
  - souborový systém může být rozšířen (online) pomocí xfs_growfs
- Delayed allocation
  - XFS podporuje „zpožděnou alokaci“
  - Systém se snaží fyzický zápis co nejvíce oddálit a ukládá data do bufferu
  - Pokud se soubor uzavře nebo dojde prostor v bufferu, data se zapíší na disk do bloků jdoucích za sebou ➔ snížení fragmentace
- Striped allocation
  - rovnoměrné rozprostření i-uzlů, žurnálu i samotných dat na všechna zařízení

- Sparse files
  - XFS podporuje řídké soubory (potřebné např. v databázových serverech)
- Direct I/O
  - přímý přístup k I/O operacím, kde se nebude využívat cache
  - aplikace je tak napojena přímo na disk používající DMA, což dává aplikaci přístup k plné šířce pásma
- Defragmentace
  - XFS nabízí možnost za běhu defragmentovat celý souborový systém pomocí nástroje XFS_FSR

Shrnutí

- Výhody XFS
  - 64-bitový souborový systém
  - Žurnálovací souborový systém – rychlá obnova při havárii systému
  - Rychlá práce s velkými soubory
  - Vysoký výkon u paralelních operací
  - Podpora multiproc. počítačů a extrémně velkých diskových farem
  - Podpora DMAPI (pro HSM)
  - Kompatibilní s NFS
  - Robustnost, důvěryhodnost
- Nevýhody XFS
  - Pomalý při mazání velkého počtu malých souborů
  - Příliš velký kód
  - Žurnáluje pouze metadata

Děkuji za pozornost …