PA152: Efektivní využívání DB

2. Datová úložiště

Vlastislav Dohnal

Poděkování

- Zdrojem materiálů tohoto předmětu jsou:
  - Přednášky CS245, CS345, CS345
    - Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom
    - Stanford University, California
  - Přednášky dřívější verze PA152
    - Pavel Rychlý
    - Fakulta informatiky, Masarykova Univerzita

Optimalizace přístupu na disk

- *Omezení náhodných přístupů*
- Velikost bloku
- Diskové pole

Omezení náhodných přístupů

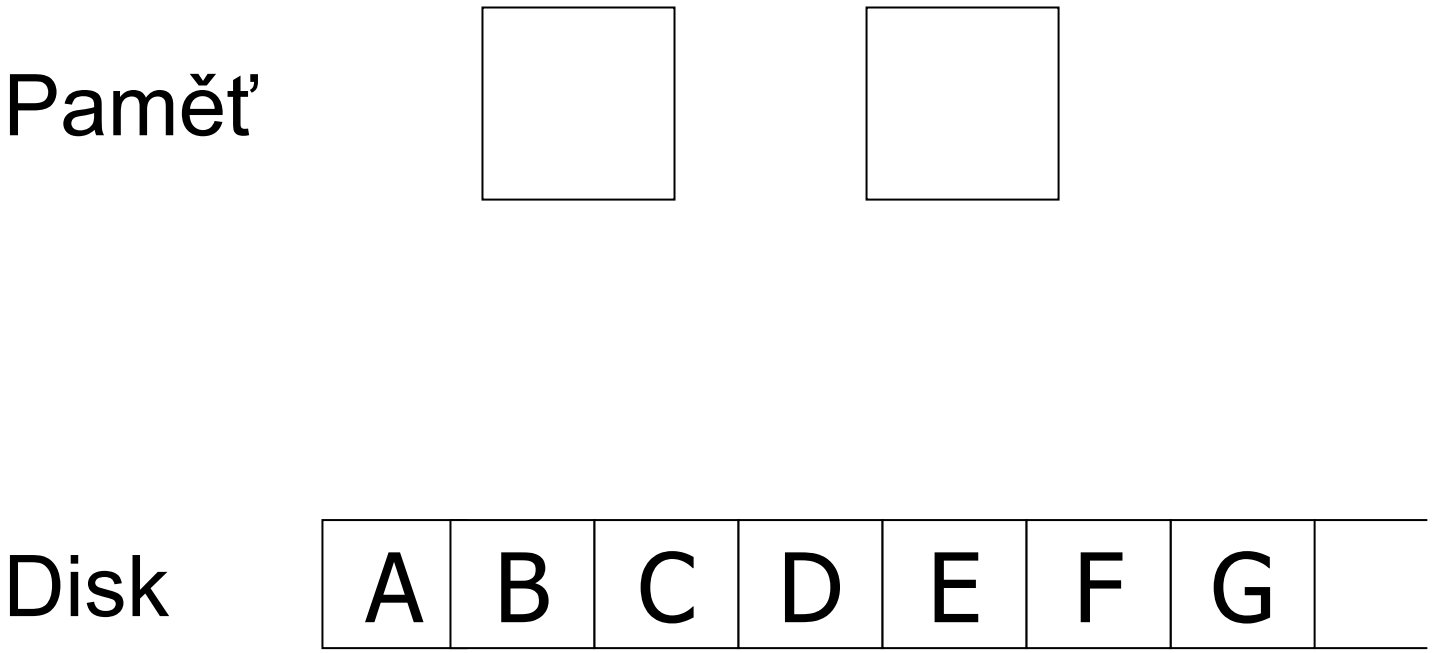
- Defragmentace
  - Uspořádání bloků do pořadí jejich zpracování
  - Souborový systém
    - Řeší na úrovni souborů
    - Alokace více bloků naráz, nástroje pro defragmentaci
- Plánování přístupů (výtah)
  - Pohyb hlavičky pouze jedním směrem
  - Přeuspořádávání požadavků na disk
    - Při zápisu použití zálohované cache (nebo logu)
- Prefetching, double buffering

Single Buffer

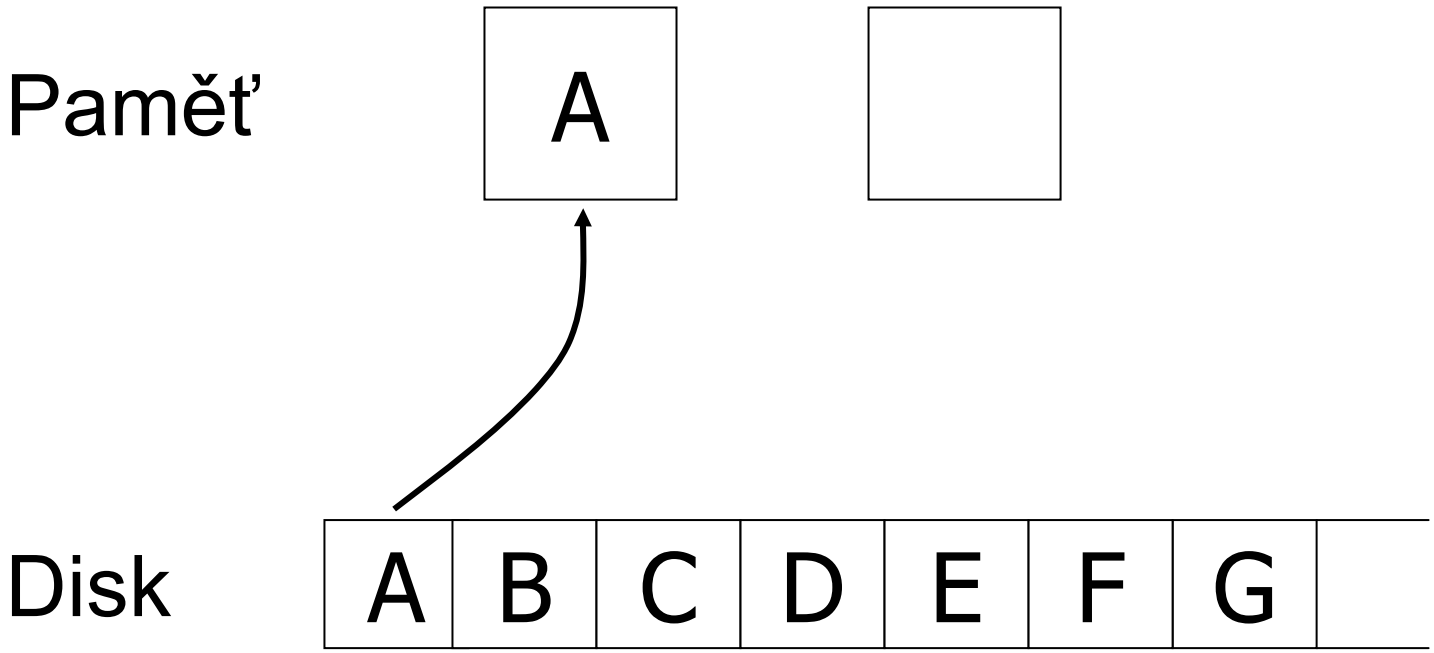
- Úloha:
  - Čti blok B1 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - Čti blok B2 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - …
- Náklady:
  - P = čas zpracování bloku
  - R = čas k přečtení 1 bloku
  - n = počet bloků ke zpracování
- Single buffer time = n(R+P)

Double Buffering

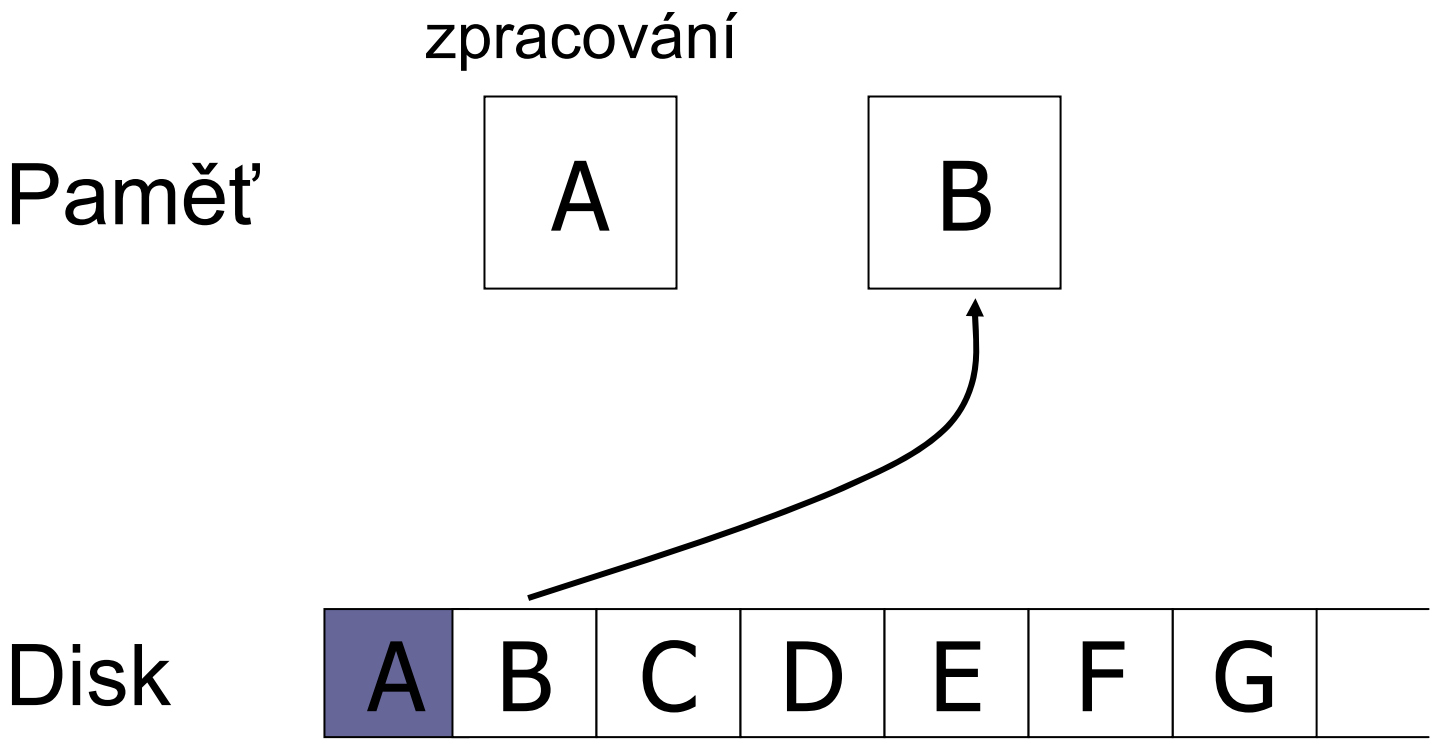
- Dva buffery v paměti, používané střídavě



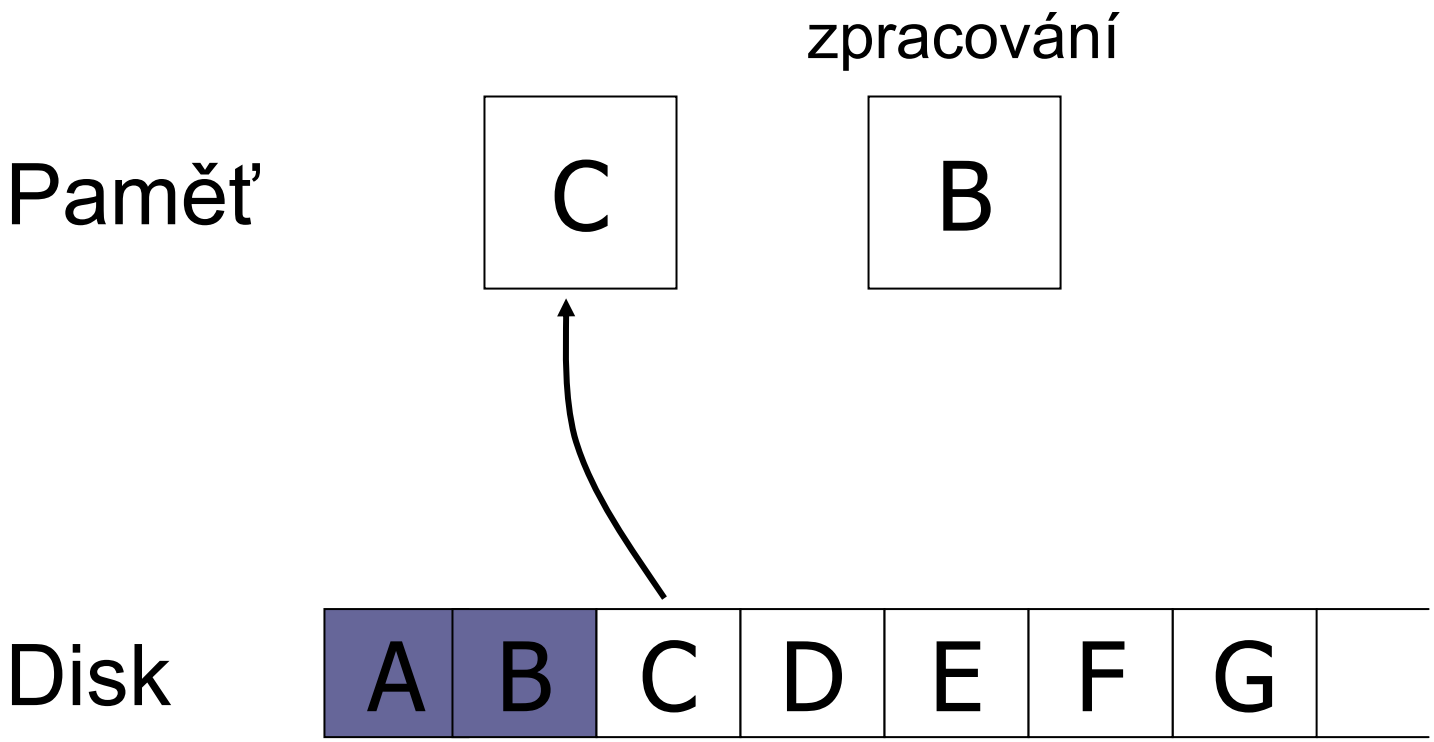
Double Buffering



Double Buffering



Double Buffering



Double Buffering

- Náklady:
  - P = čas zpracování bloku
  - R = čas k přečtení 1 bloku
  - n = počet bloků ke zpracování
- Single buffer time = n(R+P)
- Double buffer time = R + nP
  - Předpokládáme P ≥ R

Optimalizace přístupu na disk

- Omezení náhodných přístupů
- *Velikost bloku*
- Diskové pole

Velikost bloku

- Velký blok → amortizace I/O nákladů

  ALE

- Velký blok → čtení více „nepotřebných“ dat, čtení trvá déle

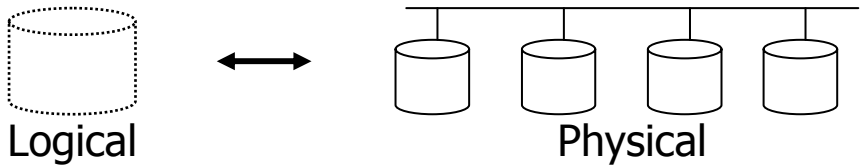
- Trend:
  - cena pamětí klesá, bloky se zvětšují

Optimalizace přístupu na disk

- Omezení náhodných přístupů
- Velikost bloku
- *Diskové pole*

Diskové pole

- Více disků uspořádaných do jednoho logického



  - Paralelní čtení / zápis
  - Snížení průměrné doby vystavení hlaviček
- Metody
  - rozdělování dat (block striping)
  - zrcadlení dat (mirroring)

Zrcadlení

- Zvýšení spolehlivosti pomocí replikace
  - Logický disk je sestaven ze 2 fyzických disků
  - Zápis je proveden na každý z disků
  - Čtení lze provádět z libovolného disku
- Data dostupná při výpadku jednoho disku
  - Ztráta dat při výpadku obou $\rightarrow$ málo pravděpodobné
- Pozor na závislé výpadky
  - Požár, elektrický zkrat, zničení HW řadiče pole, …

Rozdělování dat

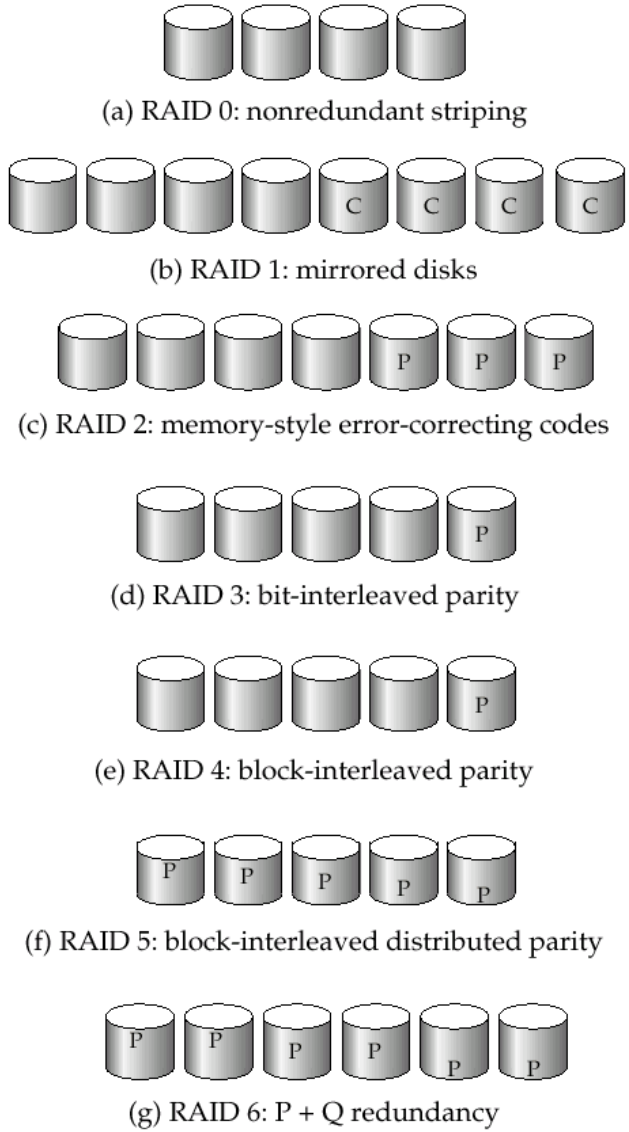
- Cíle:
  - Zvýšení přenosové rychlosti rozdělením na více disků
  - Paralelizace „velkého“ čtení ke snížení odezvy
  - Vyrovnání zátěže → zvýšení propustnosti
- Bit-level striping
  - Rozdělení každého bajtu na bity mezi disky
  - Přístupová doba je horší než u jednoho disku
  - Málo používané

Rozdělování dat

- Block-level striping
  - *n* disků, blok *i* je uložen na disk *(i mod n)+1*
  - Čtení různých bloků lze paralelizovat
    - Pokud jsou na různých discích
  - „Velké“ čtení může využít všechny disky

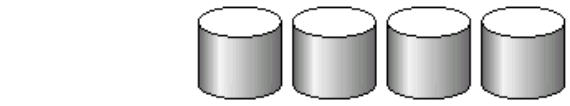
RAID

- Redundant Arrays of Independent Disks
- Různé varianty
  - Různé požadavky
  - Různá výkonnost
  - Různé vlastnosti
- Kombinace variant
  - RAID10 = RAID1, pak RAID0

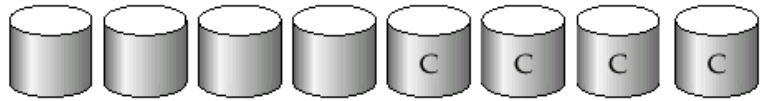


RAID0,1

- RAID0
  - Block striping, neredundantní
  - Velmi vysoký výkon, žádné zabezpečení dat
  - Nesnížená kapacita
- RAID1
  - Zrcadlení disků
    - často s block-striping
  - Poloviční kapacita, rychlé čtení
  - Vhodné pro databázové logy, atp.



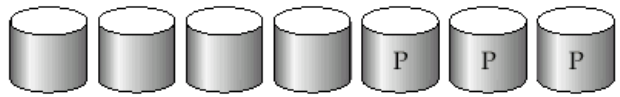
(a) RAID 0: nonredundant striping



(b) RAID 1: mirrored disks

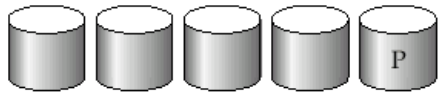
RAID2,3

- RAID2
  - Bit-striping, Hamming Error-Correcting-Code
  - Zotavení z výpadku 1 disku



(c) RAID 2: memory-style error-correcting codes

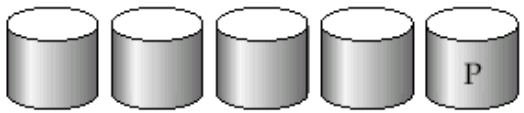
- RAID3
  - Bit-Interleaved Parity
  - 1 paritní disk
  - Zápis: spočítání a uložení parity
  - Obnova jednoho disku
    - XOR bitů z ostatních disků



(d) RAID 3: bit-interleaved parity

RAID4

- Oproti RAID3 používá block-striping
  - Paritní blok na zvláštním disku
  - Zápis: spočítání a uložení parity
  - Obnova jednoho disku
    - XOR bitů z ostatních disků
  - Vyšší rychlost než RAID3
    - Blok je čtený pouze z 1 disku → paralelizace



(e) RAID 4: block-interleaved parity

RAID4 (pokrač.)

- Zápis bloku → výpočet paritního bloku
  - Vezmi původní paritu, původní blok a nový blok (2 čtení a 2 zápisy)
  - Nebo přepočítej paritu ze všech bloků (n-1 čtení a 2 zápisy)
  - Efektivní pro sekvenční zápis velkých dat
- Paritní disk je úzké místo!
  - Zápis libovolného bloku vede k zápisu parity
- RAID3,4 – minimálně 3 disky (2+1)
  - Kapacita snížena o paritní disk

RAID5

- Block-Interleaved Distributed Parity
  - Rozděluje data i paritu mezi n+1 disků
  - Odstranění zátěže na paritním disku RAID4

(f) RAID 5: block-interleaved distributed parity

- Příklad (5 disků)
  - Paritní blok pro *n* bloků je uložen na disku (n mod 5) + 1
  - Datové bloky uloženy na ostatních 4 discích

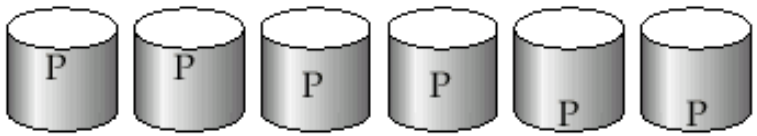
P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

RAID5 (pokrač.)

- Vyšší výkon než RAID4
  - Zápis bloků je paralelní, pokud jsou na různých discích
  - Nahrazuje RAID4
    - má stejné výhody a ruší nevýhodu paritního disku
- Často používané řešení

RAID6

- P+Q Redundancy scheme
  - Podobné RAID5, ale ukládá extra informace pro obnovu při výpadku více disků
  - Více disků pro paritu (dual distributed parity)
    - Min. 4 disky v poli (kapacity snížena o 2 disky)
  - Samoopravné Hammingovy kódy
    - Opraví výpadek 2 disků
  - Není příliš používaný



(g) RAID 6: P + Q redundancy

RAID shrnutí

- RAID0 – bezpečnost dat není podstatná
  - Data lze snadno a rychle obnovit (ze záloh,…)
- RAID2,4 jsou nahrazeny RAID3,5
  - RAID3 se nepoužívá – bit-striping vede k využití všech disků při zápisu/čtení 1 bloku
- RAID6 – nepoužívaný
  - RAID1,5 poskytují dostatečnou spolehlivost
  - Spíše kombinace – RAID10, RAID50
- Vybíráme mezi RAID1 a RAID5

RAID shrnutí (pokrač.)

- RAID1
  - Mnohem rychlejší zápis než RAID5
  - Použití pro aplikace s velkým množstvím zápisů
  - Dražší než RAID5 (má nižší kapacitu)
- RAID5
  - pro každý zápis vyžaduje min. 2 čtení a 2 zápisy
    - RAID1 vyžaduje pouze 2 zápisy
  - Vhodný pro aplikace s menším množstvím zápisů
- Nároky dnešních aplikací na počet I/O
  - Velmi vysoké (např. WWW servery, …)
  - Nákup množství disků pro splnění požadavků
    - Mají dostatečnou volnou kapacitu, pak RAID1 (nic nás dále nestojí)

RAID shrnutí (pokrač.)

- Nenahrazuje zálohování!!!
- Implementace
  - SW – téměř každý OS podporuje
  - HW – speciální řadič
    - Nutné zálohování cache bateriemi nebo non-volatile RAM
    - Pozor na výkonnost procesoru řadiče – může být pomalejší než SW!!!
- Hot-swapping (výměna za provozu)
  - HW implementace většinou podporují
  - SW není problém, pokud HW podporuje
- Spare disks
  - V poli jsou většinou přítomné náhradní disky

Výpadky disků

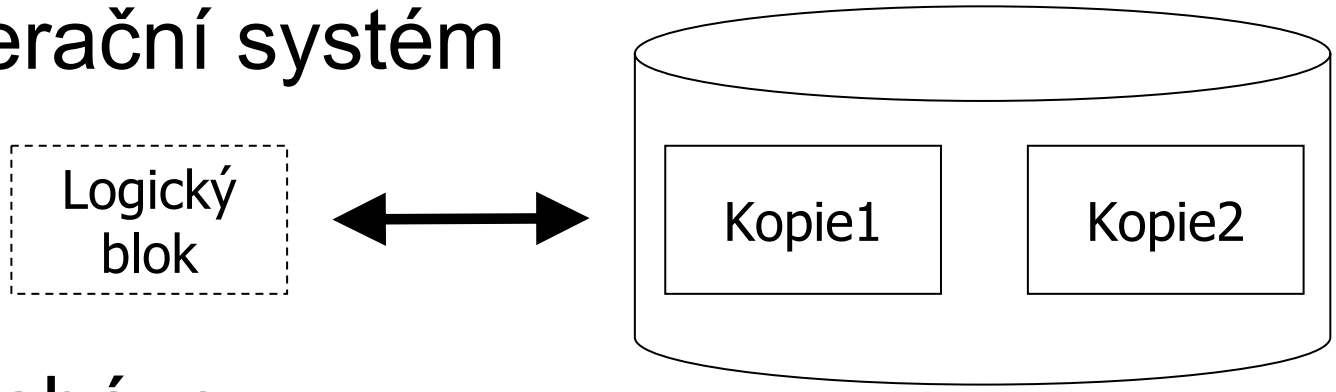
- Občasný výpadek
  - Chyba při čtení/zápisu → opakování OK
- Vada média
  - Trvalá chyba nějakého sektoru
  - Moderní disky samy detekují a opraví
    - z vlastní rezervní kapacity
- Zničení disku
  - Totální výpadek → výměna disku

Ošetření výpadků disků

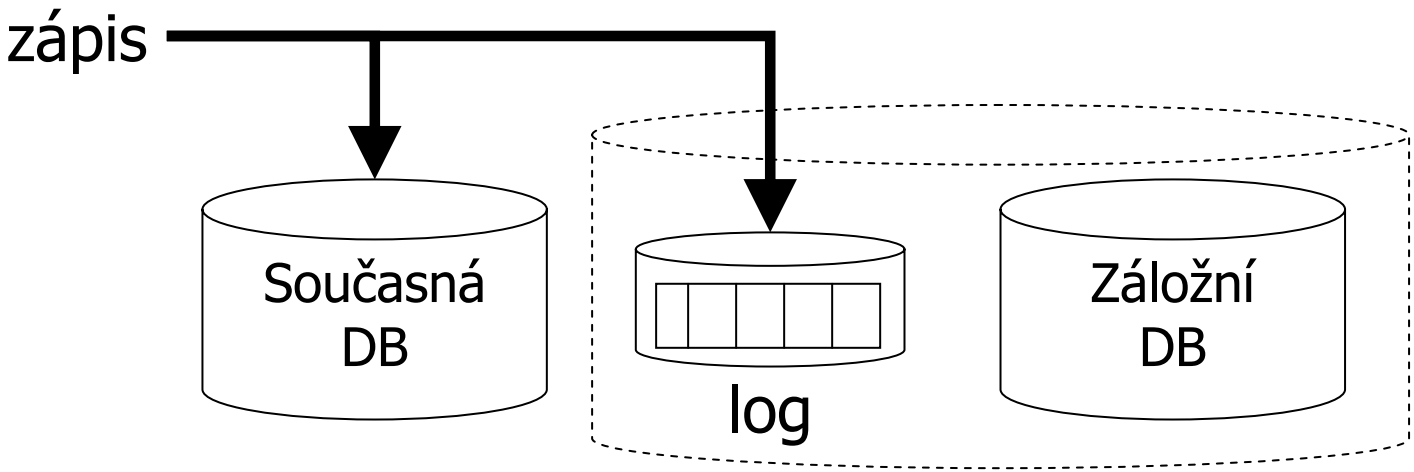
- Detekce
  - Kontrolní součty
- Opravy
  - Samoopravné kódy (ECC)
    - Hammingovy kódy, …
  - Stabilní uložení
    - Uložení na více místech stejného disku
    - Žurnálování (log změn)
    - Diskové pole

Stabilní uložení

- Operační systém



- Databáze



Zotavení ze zničení

- Pravděpodobnost výpadku
  - Mean time to failure (MTTF)
    - Někdy také Mean time between failures (MTBF)
  - Průměrná doba fungování mezi výpadky
    - Typicky 10 a více let
  - MTTF se snižuje s věkem disku
- Pozor:
  - Jeden disk má MTTF 10 let
  - Systém s 10 disky má MTTF 1 rok
    - Tj. průměrně každý rok jeden z disků vypadne

Pravděpodobnost výpadku

- Interpretace MTTF 10 let
  - Průměrně 50% disků vypadne za 10 let
  - 100% disků vypadne za 20 let
  - $P_{\text{výpadku za rok}} = 1/20 = 0.05 = 5\%$
  - Lineární interpolace výskytu chyb
    - Obvykle pro výpočty dostatečná
    - V praxi jinak
      - více chyb na začátku a pak na konci životnosti

Oprava chyby

- Mean time to repair (MTTR)
  - Čas od výpadku do obnovení činnosti
  - Tj. čas výměny vadného disku
- Mean time to data loss (MTTD)
  - Závisí na MTTF i MTTR
  - Průměrná doba mezi ztrátou dat

Příklad výpadku RAID1

- 2 zrcadlené disky
  - Každý MTTF 10 let
  - Každých 5 let vypadne nějaký disk (10/2)
- Výměna vadného do 3 hodin
  - MTTR = 3 hodiny
- Pravděpodobnost ztráty dat:
  - $P_{\text{výpadku 1 disku}}$ = 1/20 roku, $P_{\text{výpadku 1 ze 2}}$ = 1/10 roku
  - MTTR = 3/24 dne = 1/2920 roku
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} * \text{MTTR} * P_{\text{výpadku 1disku}} = 1/584000$
  - **MTTD** = 0.5 / $P_{\text{ztráty dat}}$ = 0.5 * 1/584000 = 292 000 let

RAID0 a RAID4

- MTTF disku 10 let ($P_{\text{výpadku 1 disku}}$=1/20=5%)
- RAID0 – dva disky, striping
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} = 1/10$
  - MTTD = 0.5 / (1/10) = 5 let
- RAID4 – opravuje výpadek 1 disku
  - 4 disky (3+1), MTTR = 3 hodiny
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 4}} * \text{MTTR} * P_{\text{výpadku 1 ze 3}}$
  - $P_{\text{ztráty dat}} = 4/20 * 1/2920 * 3/20 = 3/292000$
  - MTTD = 0.5 * 292000/3=292000/6=48 667 let