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 (podzim 2008)
    - Pavel Rychlý
    - Fakulta informatiky, Masarykova Univerzita

Optimalizace přístupu na disk

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

Omezení náhodných přístupů

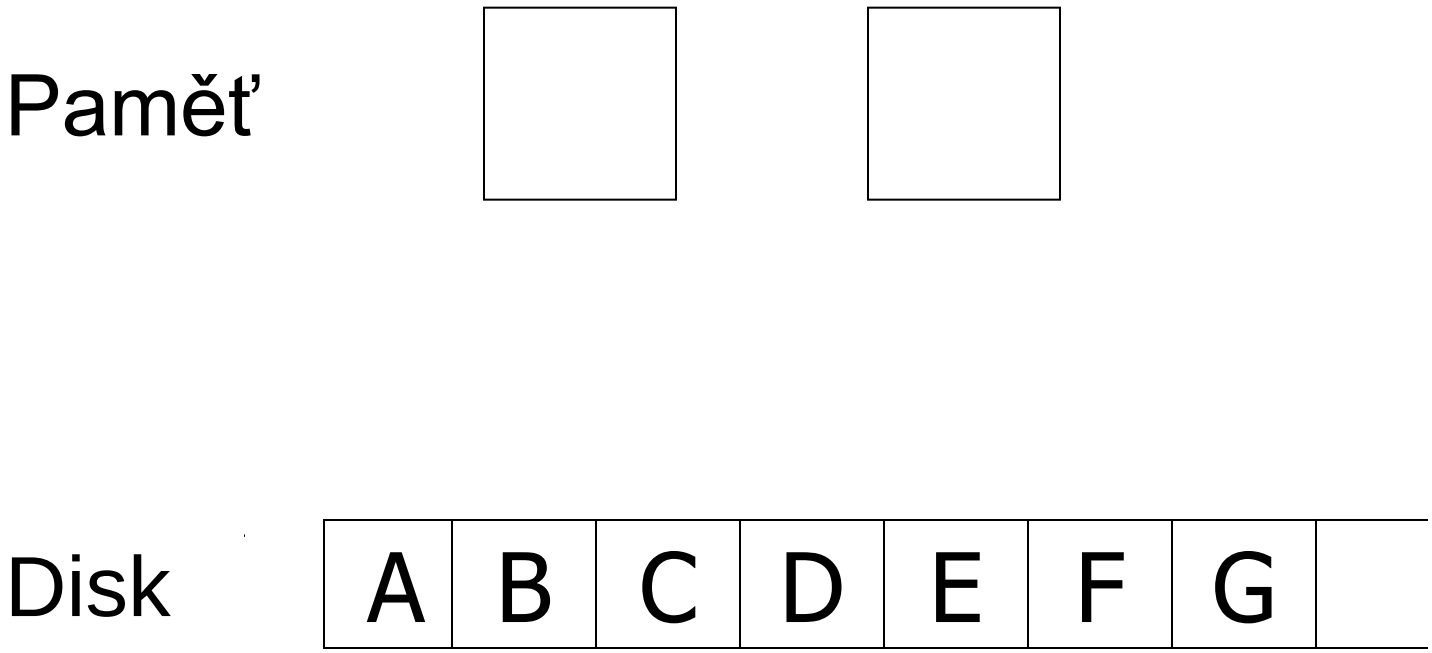
- Defragmentace
  - Uspořádání bloků do pořadí jejich zpracování
  - Souborový systém
    - Řeší na úrovni souborů
    - Alokace více bloků naráz, nástroje pro defragmentaci
- Plánování přístupů (výtah)
  - Pohyb hlavičky pouze jedním směrem
  - Přeuspořádávání požadavků na disk
    - Při zápisu použití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

Single Buffer

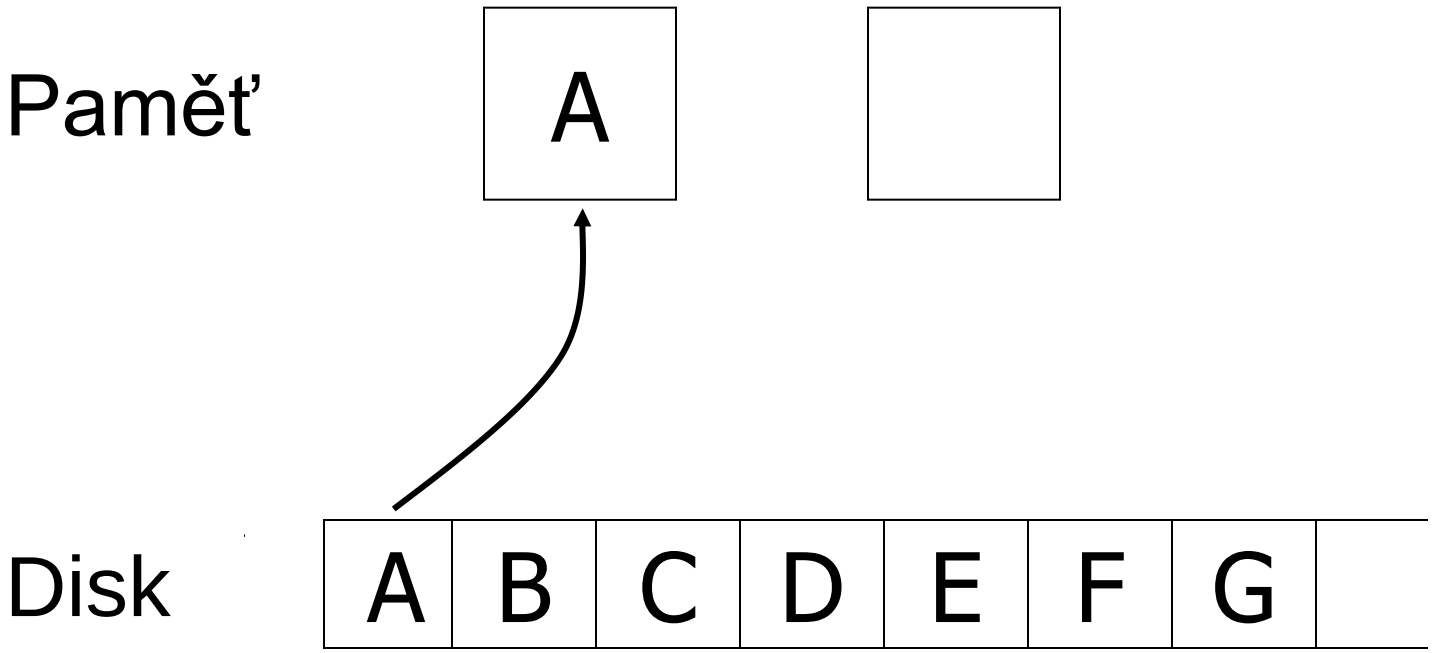
- Úloha:
  - Čti blok B1 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - Čti blok B2 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - …
- Náklady:
  - P = čas zpracování bloku
  - R = čas k přečtení 1 bloku
  - n = počet bloků ke zpracování
- Single buffer time = n(R+P)

Double Buffering

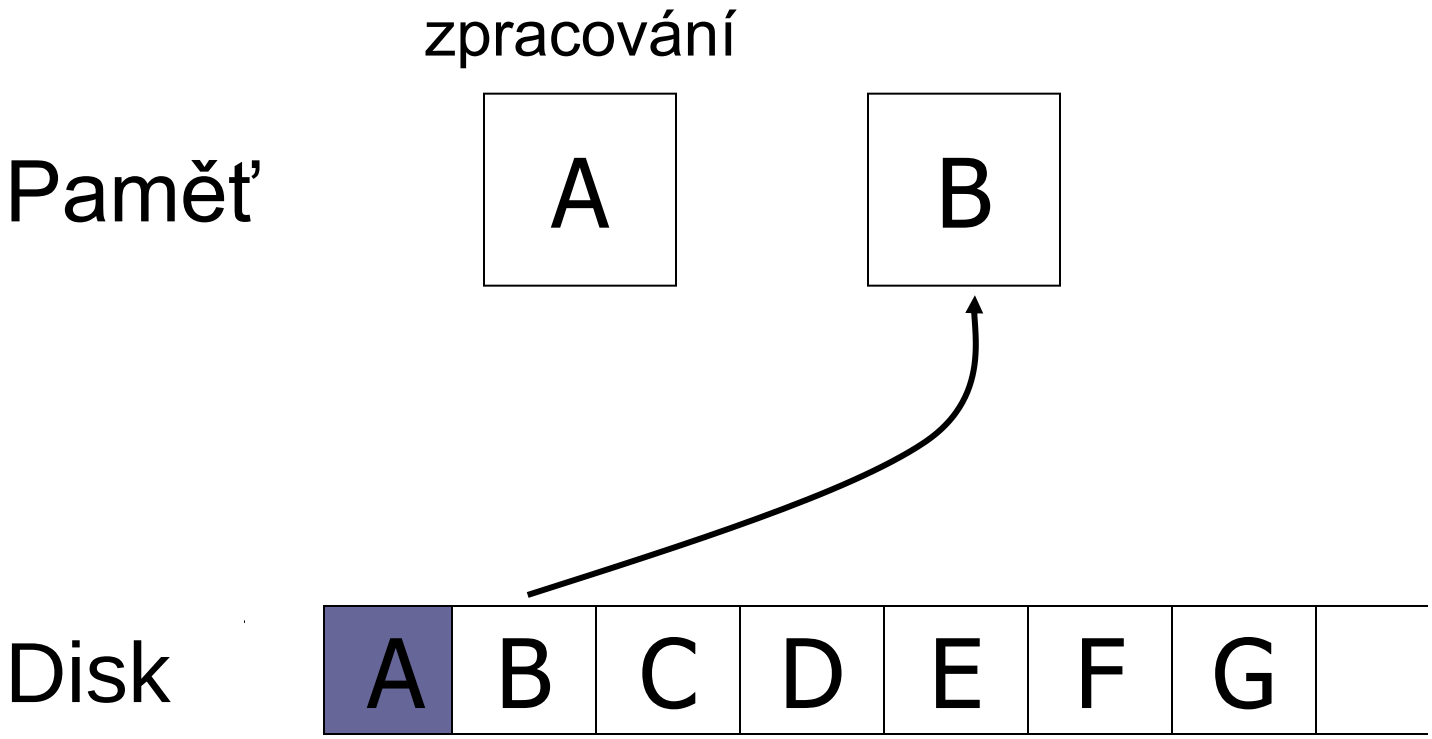
- Dva buffery v paměti, používané střídavě



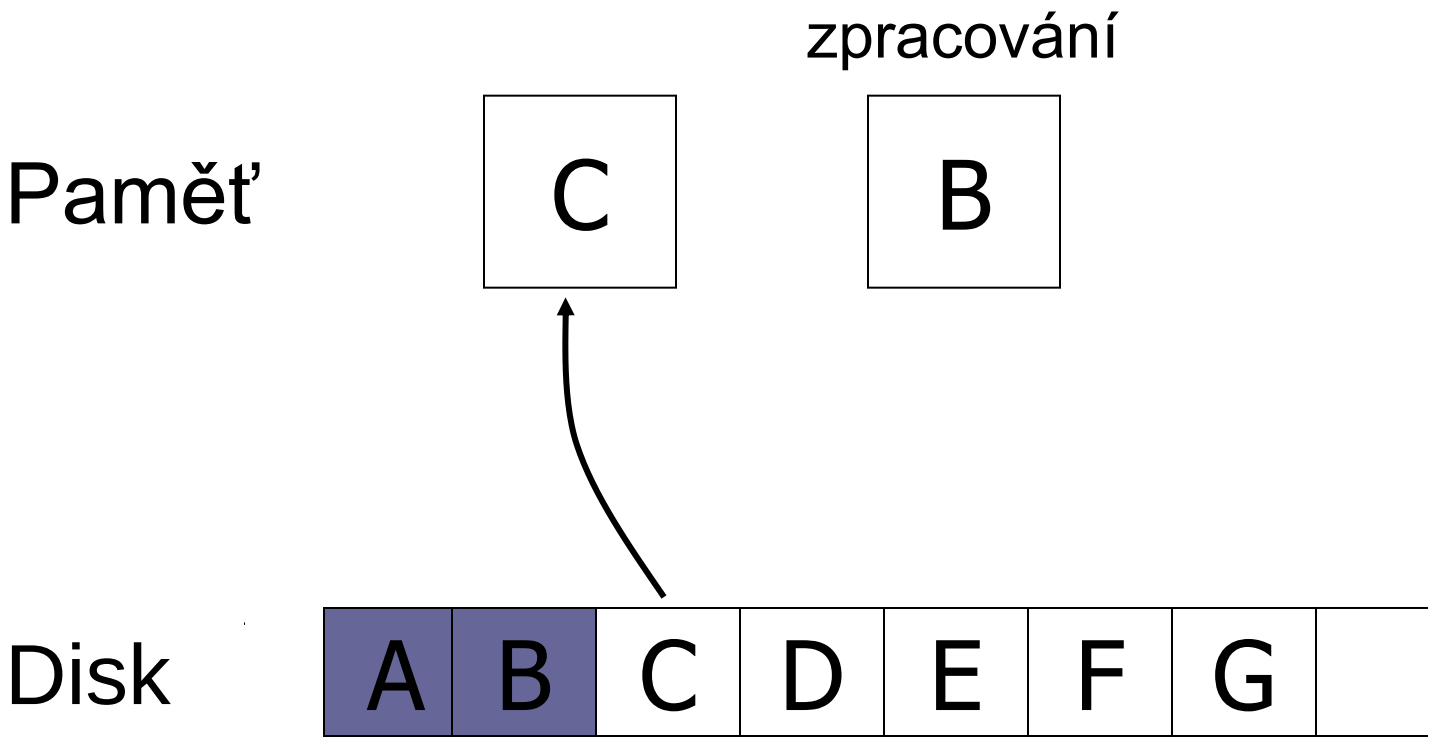
Double Buffering



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
  - Jinak
    - = nR + P

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

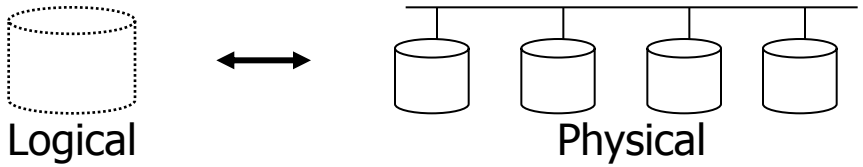
- Trend:
  - cena pamětí klesá, bloky se zvětšují

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 → málo pravděpodobné
- Pozor na závislé výpadky
  - Požár, elektrický zkrat, zničení HW řadiče pole, …

Rozdělování dat

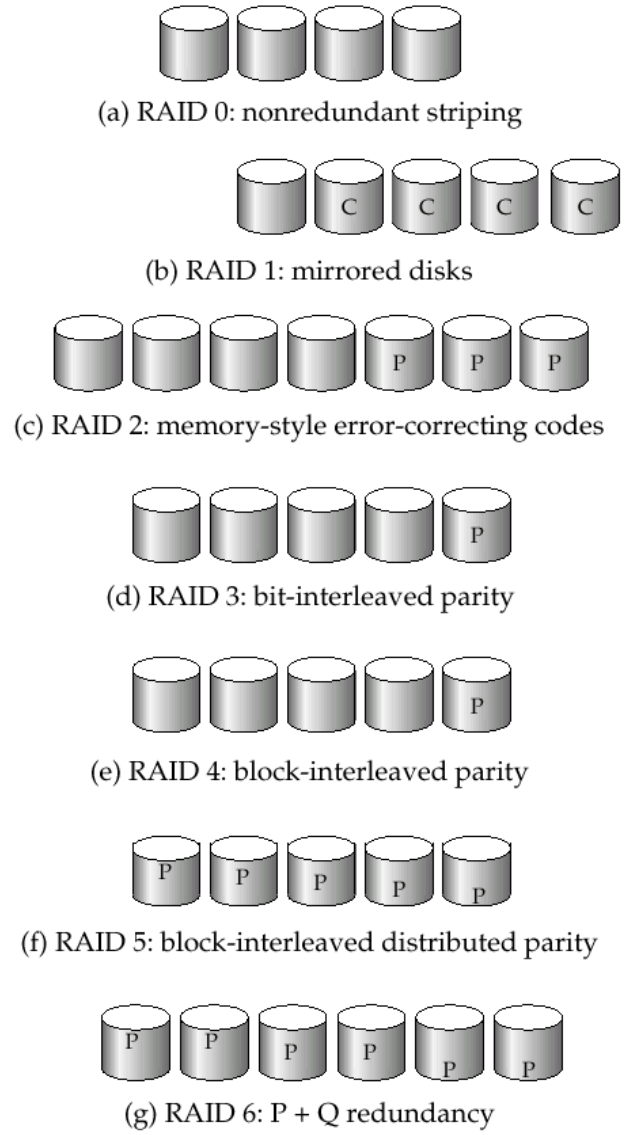
- Cíle:
  - Zvýšení přenosové rychlosti rozdělením na více disků
  - Paralelizace „velkého“ čtení ke snížení odezvy
  - Vyrovnání zátěže $\rightarrow$ zvýšení propustnosti
- Bit-level striping
  - Rozdělení každého bajtu na bity mezi disky
  - Přístupová doba je horší než u jednoho disku
  - Málo používané

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
  - RAID1+0
    - = RAID1, pak RAID0



(a) RAID 0: nonredundant striping

(b) RAID 1: mirrored disks

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

(d) RAID 3: bit-interleaved parity

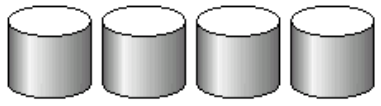
(e) RAID 4: block-interleaved parity

(f) RAID 5: block-interleaved distributed parity

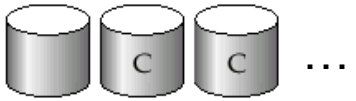
(g) RAID 6: P + Q redundancy

RAID0, RAID1

- RAID0
  - Block striping, neredundantní
  - Velmi vysoký výkon, žádné zabezpečení dat
  - Nesnížená kapacita
- RAID1
  - Zrcadlení disků
    - někdy omezeno na 2 disky
  - Kapacita 1/n, rychlé čtení, zápis jako 1 disk
  - Vhodné pro databázové logy, atp.
    - Zápis logů je sekvenční



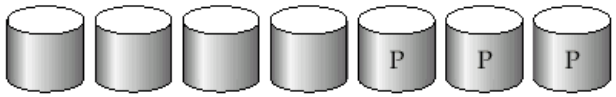
(a) RAID 0: nonredundant striping



(b) RAID 1: mirrored disks

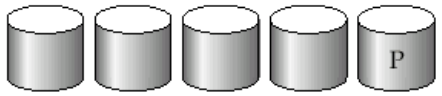
RAID2, RAID3

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



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

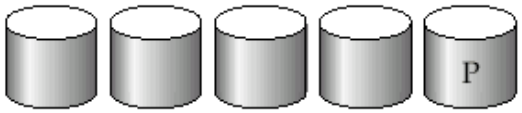
- RAID3
  - Bit-Interleaved Parity
  - 1 paritní disk
  - Zápis: spočítání a uložení parity
  - Obnova jednoho disku
    - XOR bitů z ostatních disků



(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, RAID4 – minimálně 3 disky (2+1)
  - Kapacita snížena o paritní disk

RAID5

- Block-Interleaved Distributed Parity
  - Rozděluje data i paritu mezi *n* 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

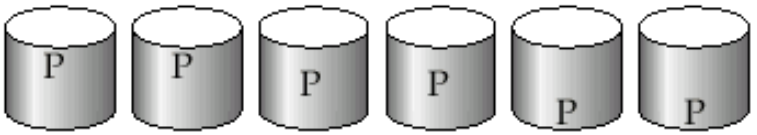
P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

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

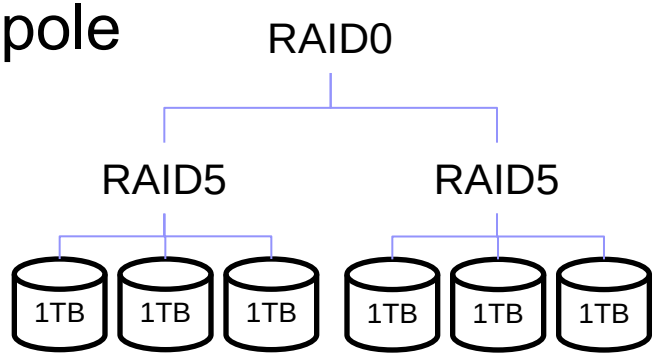
- P+Q Redundancy scheme
  - Podobné RAID5, ale ukládá extra informace pro obnovu při výpadku více disků
  - Více disků pro paritu (dual distributed parity)
    - Min. 4 disky v poli (kapacity snížena o 2 disky)
  - Samoopravné Hammingovy kódy
    - Opraví výpadek 2 disků
  - Není příliš používaný



(g) RAID 6: P + Q redundancy

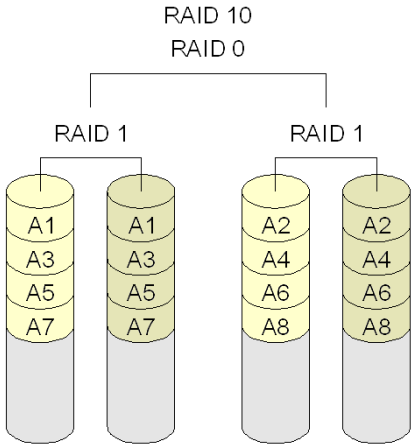
RAID – kombinace

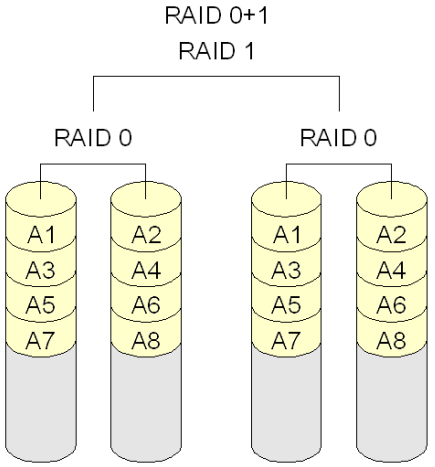
- Jednotlivé varianty polí lze kombinovat
  - Z fyzických disků se vytvoří pole
  - Pak z polí se vytvoří výsledné pole
- Vhodné k zvýšení výkonu / spolehlivosti
- Příklad:
  - RAID5+0 z 6 fyzických disků
    - Po třech vytvoříme dvě RAID5 pole
    - RAID5 pole spojíme do RAID0



RAID1+0 vs. RAID0+1

- RAID1+0
  - odolnější proti výpadkům
  - výpadek 1 disku v libovolném RAID1 ok
- RAID0+1
  - výpadek disku v prvním RAID0
  - výpadek lib. disk v druhém RAID0
  - ⇒ data ztracena





Zdroj: Wikipedia

RAID shrnutí

- RAID0 – bezpečnost dat není podstatná
  - Data lze snadno a rychle obnovit (ze záloh,…)
- RAID2 a 4 jsou nahrazeny RAID3 a 5
  - RAID3 se nepoužívá – bit-striping vede k využití všech disků při zápisu/čtení 1 bloku
- RAID6 – nepoužívaný
  - RAID1 a 5 poskytují dostatečnou spolehlivost
  - Spíše kombinace – RAID1+0, RAID5+0
- Vybíráme mezi RAID1 a RAID5

RAID shrnutí (pokrač.)

- RAID1+0
  - 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+0 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)
    - Nejlépe RAID1+0

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
  - Existence náhradních disků v poli

Výpadky disků

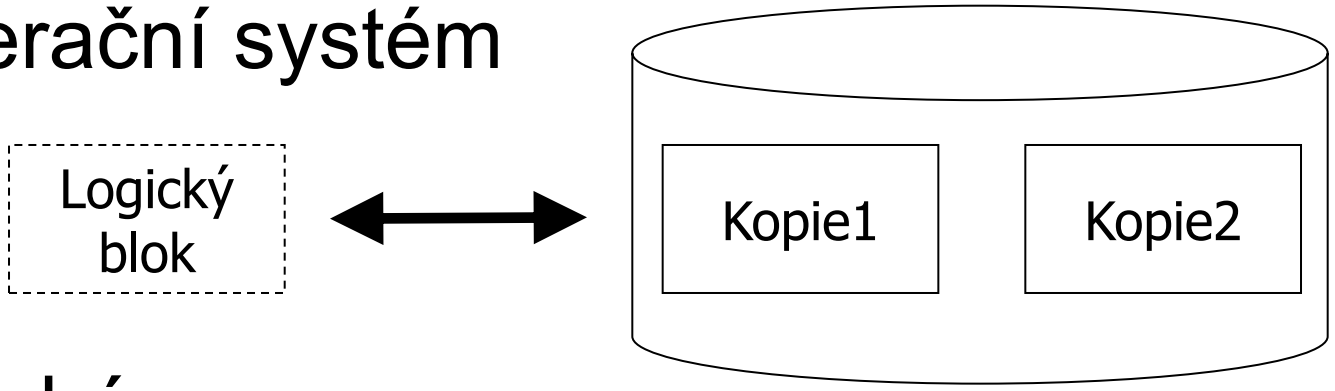
- Občasný výpadek
  - Chyba při čtení/zápisu → opakování → OK
- Vada média
  - Trvalá chyba nějakého sektoru
  - Moderní disky samy detekují a opraví
    - z vlastní rezervní kapacity
- Zničení disku
  - Totální výpadek → výměna disku

Ošetření výpadků disků

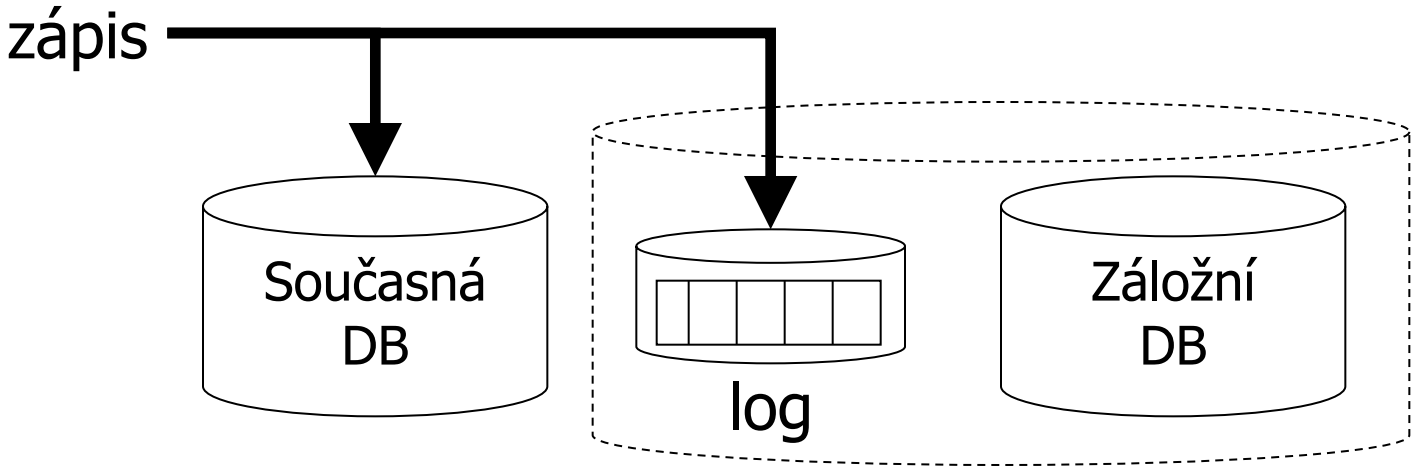
- Detekce
  - Kontrolní součty
- Opravy
  - Samoopravné kódy (ECC)
    - Hammingovy kódy, …
  - Stabilní uložení
    - Uložení na více místech stejného disku
      - Např. super-blok
    - Žurnálování (log změn)
    - Diskové pole

Stabilní uložení v databázích

- Operační systém



- Databáze



Zotavení ze zničení

- Mean Time To Failure (MTTF)
  - Někdy také: Mean Time Between Failures (MTBF)
  - odpovídá pravděpodobnosti výpadku
  - průměrná doba fungování mezi výpadky
    - polovina disků má výpadek během této doby
    - předpokládá se rovnoměrné rozložení výpadků
  - snižuje s věkem disku
  - obvykle 1 000 000 a více hodin
    - $\Rightarrow$ 114 let
    - tj. za 228 let vypadne 100% $\Rightarrow$ $P_{\text{výpadku za rok}}$=0,44%

Výpadky – pokračování

- Příklad:
  - MTTF 1 000 000 hours
  - $\Rightarrow$ v populaci 2 000 000 disků
    - vypadne každou hodinu jeden disk
    - tj. 8 760 disků ročně
    - $\Rightarrow$ pravděpodobnost výpadku za rok = 0,44%
      - = **Annualized Failure Rate (AFR)**

- Annual Replacement Rate (ARR)

  nebo Annualized Return Rate
  - Ne všechny výpadky jsou způsobeny disky
    - vadné kabely, atp.
  - Uvádí se: 40% z ARR je “No Trouble Found” (NTF)
  - AFR = ARR*0.6 ARR = AFR / 0.6

Výpadky – praxe

- Seagate http://www.seagate.com/docs/pdf/whitepaper/drive_reliability.pdf
  (November 2000)
  - Savvio® 15K.2 Hard Drives – 73 GB
    - AFR = 0,55%
  - Seagate estimates MTTF for a drive as the number of power-on hours (POH) per year divided by the first-year AFR.
  - AFR is derived from reliability-demonstration tests (RDT)
    - RDT at Seagate = hundereds of disks operating at 42ºC ambient temperature

Výpadky – praxe (2)

- Seagate
  - Vliv teploty na MTTF pro první rok

Temp (°C)	Acceleration Factor	Derating Factor	Adjusted MTBF
25	1.0000	1.00	232,140
26	1.0507	0.95	220,533
30	1.2763	0.78	181,069
34	1.5425	0.65	150,891
38	1.8552	0.54	125,356
42	2.2208	0.45	104,463
46	2.6465	0.38	88,123
50	3.1401	0.32	74,284
54	3.7103	0.27	62,678
58	4.3664	0.23	53,392
62	5.1186	0.20	46,428
66	5.9779	0.17	39,464
70	6.9562	0.14	32,500

Výpadky – praxe (3)

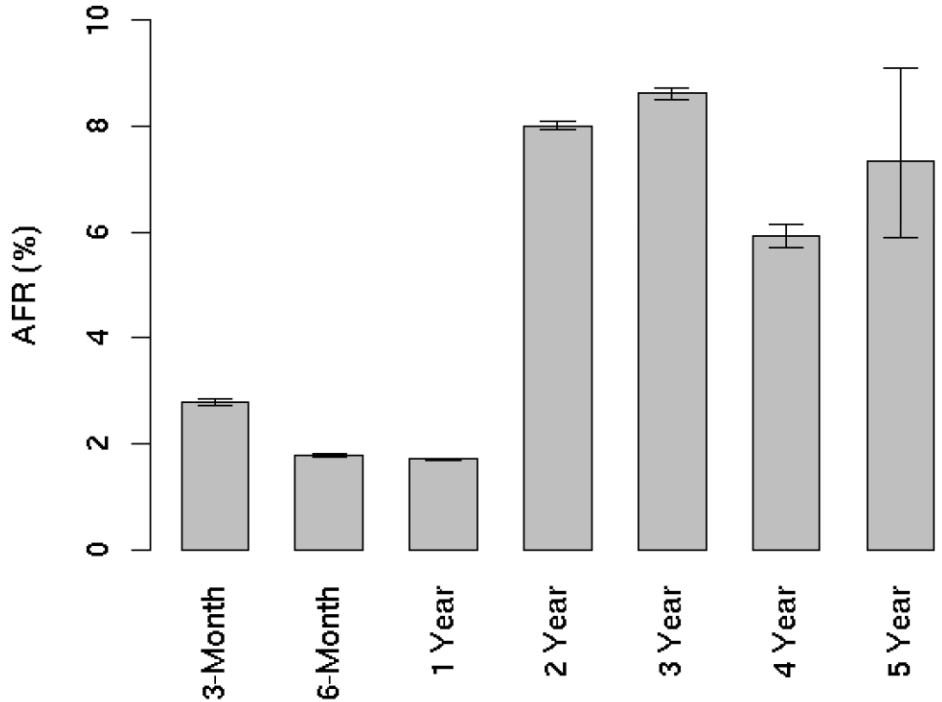
- Seagate

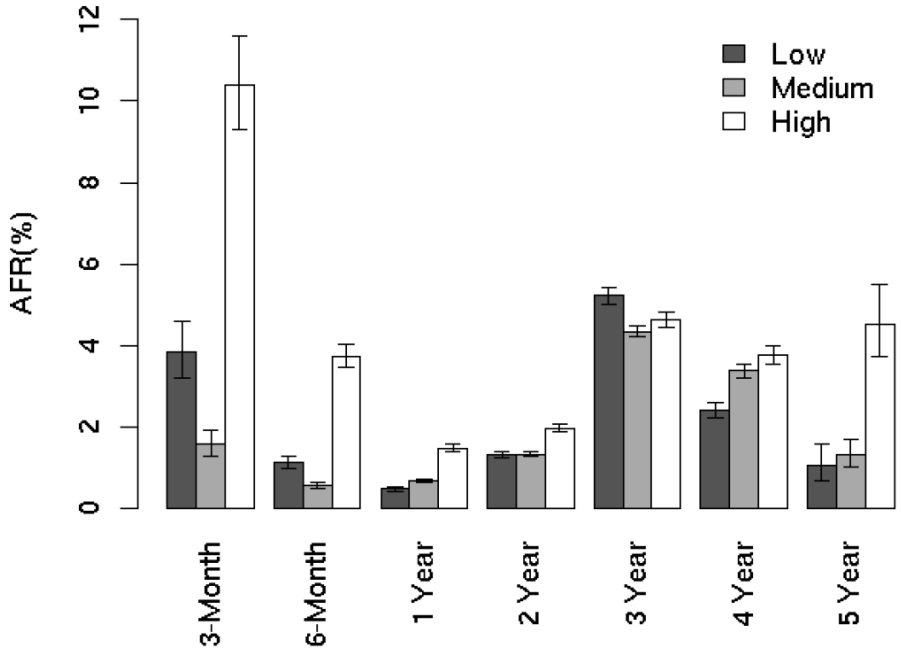
		MODEL:					
		Weibull		Warranty Data (OEM only)		Flatline Model	
Year	Cumulative power-on hours	Yearly failure rate	Cumulative failure rate	Yearly failure rate	Cumulative failure rate	Yearly failure rate	Cumulative failure rate
1	2,400	1.20%	1.20%	1.20%	1.20%	1.20%	1.20%
2	4,800	0.55%	1.75%	0.78%	1.98%	0.55%	1.75%
3	7,200	0.43%	2.18%	0.39%	2.37%	0.55%	2.30%
4	9,600	0.37%	2.55%			0.55%	2.86%
5	12,000	0.33%	2.88%			0.55%	3.41%
6	14,400	0.30%	3.18%			0.55%	3.96%
7	16,800	0.28%	3.46%			0.55%	4.51%
8	19,200	0.26%	3.72%			0.55%	5.06%
9	21,600	0.24%	3.96%			0.55%	5.62%
10	24,000	0.23%	4.19%			0.55%	6.17%

Weibull – SW pro modelování průběhu chybovosti

Výpadky - realita

- Google http://research.google.com/archive/disk_failures.pdf (Konference FAST 2007)
  - Test na 100 000 discích
    - SATA, PATA disky; 5400-7200 rpm; 80-400 GB

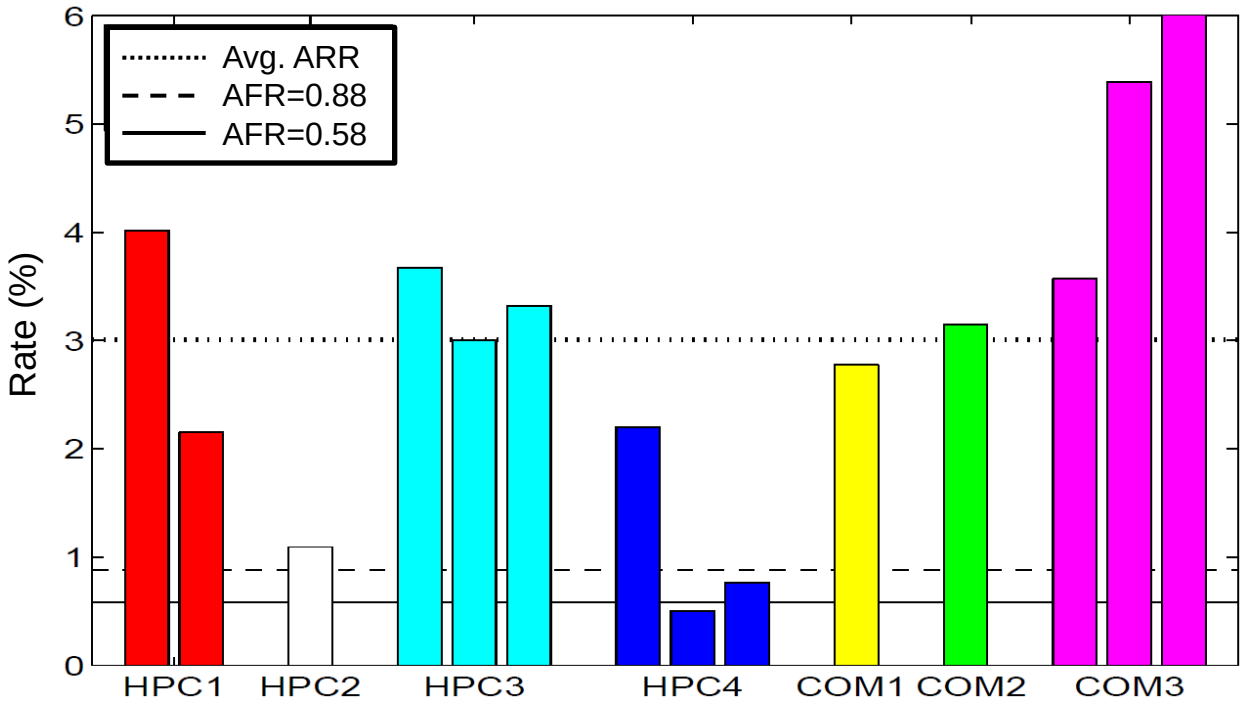




Výpadky - realita

- Studie 100 000 disků SCSI, FC, SATA

  http://www.cs.cmu.edu/~bianca/fast07.pdf (Konference FAST 2007)



HPC3: 3064x SCSI disk, 15k rpm, 146GB
11000x SATA disk, 7200 rpm, 250GB
144x SCSI disk, 15k rpm, 73GB

Výpadky - realita

- Závěry:
  - Obvykle se AFR zvyšuje s teplotou prostředí
    - Data z Google to nepotvrzují
  - SMART parameters are well-correlated with higher failure probabilities
    - Google
      - After the first scan error, a drive is 39 times more likely to fail within 60 days.
      - First errors in reallocations, offline reallocations, and probational counts are strongly correlated to higher failure probabilities.
  - Vhodné ve výpočtech používat AFR 3-4%
    - If you plan on AFR that is 50% higher than MTTF suggests, you'll be better prepared.
  - Po 3 letech provozu disku být připraven na jeho výměnu.

Oprava chyby

- Mean Time To Repair (MTTR)
  - Čas od výpadku do obnovení činnosti
  - = čas výměny vadného disku + obnovení dat
- Mean Time To Data Loss (MTTDL)
  - Závisí na MTTF a MTTR
  - Průměrná doba mezi ztrátou dat
  - Pro jeden disk (tj. data ukládám na jednom disku)
    - MTTDL = MTTF

Oprava chyby – sada disků

- Sada disků
  - Předpoklad
    - chyba každého disku je stejně pravděpodobná
    - a nezávislá na ostatních
- Příklad
  - Jeden disk
    - MTTF = 114 let, AFR = 0,44%
  - Systém se 114 disky
    - MTTF = 1 rok, AFR = 50%
    - Tj. průměrně každý rok jeden z disků vypadne (50% pravděpodobnost)

Příklad výpadku RAID1

- 2 zrcadlené disky
  - každý AFR=3%
- Výměna vadného a obnova pole do 3 hodin
  - MTTR = 3 hodiny
- Pravděpodobnost ztráty dat:
  - $P_{\text{výpadku 1 disku}} = AFR = 0.03$, $P_{\text{výpadku 1 ze 2}} = 0.06$
  - MTTR = 3 hod = 3/24 dne = 1/2920 roku
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} * MTTR * P_{\text{výpadku 1 disku}}$
    $= 1 / 1\,622\,222 = 0{,}000\,000\,616$
  - **MTTDL** $= 0.5 / P_{\text{ztráty dat}} = 0.5 / (1/1622222) = 811\,111$ let

Příklad výpadku RAID0

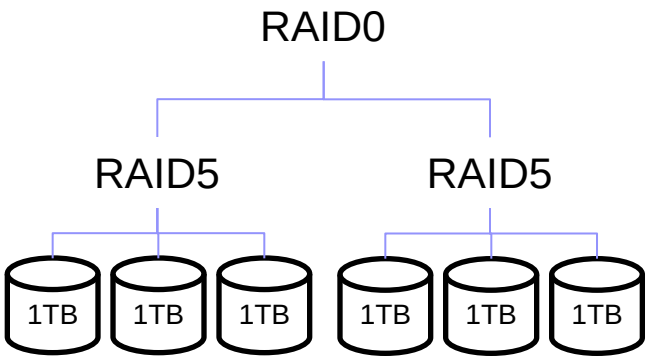
- AFR disku 3% ($P_{\text{výpadku 1 disku}}$)
- RAID0 – dva disky, striping
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} = 6\%$
  - MTTDL = 0.5 / (0.06) = 8,3 roku

Příklad výpadku RAID4

- AFR disku 3% ($P_{\text{výpadku 1 disku}}$)
- RAID4 – opravuje výpadek 1 disku
  - 4 disky (3+1), MTTR = 3 hodiny
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 4}} * \text{MTTR} * P_{\text{výpadku 1 ze 3}}$
  - $P_{\text{ztráty dat}} = 4*0{,}03 * 1/2920 * 3*0{,}03$
    $= 108 / 2\,920\,000 = 0{,}000\,003\,698$
  - $\text{MTTDL} = 0.5 / P_{\text{ztráty dat}} = 135\,185$ let

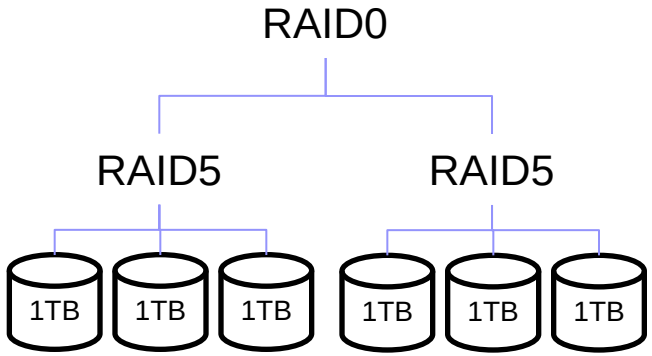
Příklad výpadku – kombinace RAID

- Kombinace polí
  - Spočítám MTTDL pro složky
    - Toto použiji v dalším jako MTTF „virtuálního disku“
  - Pak vypočítám výsledné MTTDL

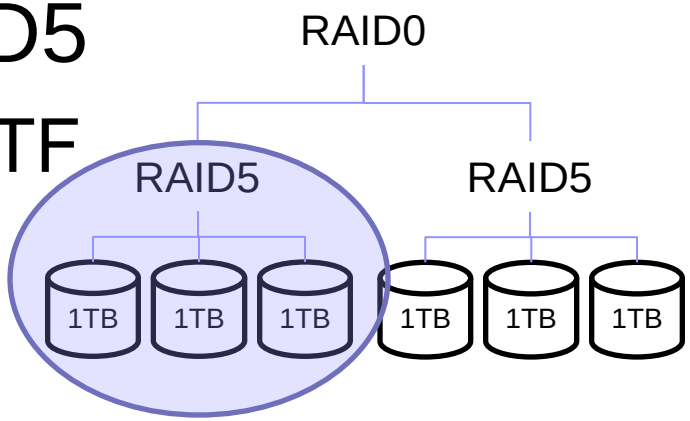


Příklad výpadku – kombinace RAID

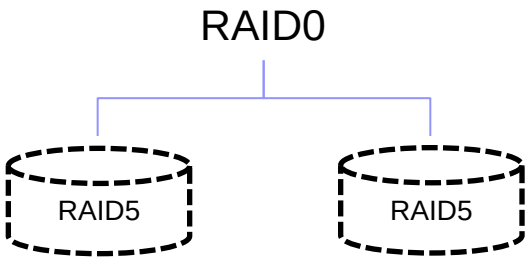
- RAID5+0
  - 1 disk má AFR_{disk} a $MTTF_{disk}$



1) Vypočítej MTTDL pro RAID5
   - To dále použijeme jako MTTF
     - $AFR_{RAID5} = 8760/(2*MTTF)$



2) Vypočítej MTTDL pro RAID0
   - $MTTDL = 0.5 / P_{\text{výpadku 1 ze 2}}$
     $= 0.5 / (2*AFR_{RAID5})$



Příklad výpadku – kombinace RAID

- RAID4+0 z 8 disků
  - 1 disk AFR=3%
  - Vždy ze 4 disků vyrobíme 1x RAID4
    - $MTTDL_{RAID4}$ = 135 185 let, tj. AFR_{RAID4} = 1 / 270 370
  - 2x RAID4 spojíme pomocí RAID0
    - $P_{ztráty\ dat\ RAID4+0} = P_{výpadku\ 1\ ze\ 2\ RAID4\ polí}$ = 2 * 1/270370
    - MTTDL = 0.5 / (2 * 1/270370) = 67 592,5 let