



PA152: Efektivní využívání DB
2. Datová úložiště

Vlastislav Dohnal

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
- Přeuspořádání požadavků na disk
 - Při zápisu použití zálohované cache (nebo žurnálu)

■ Prefetching, double buffering

Single Buffer

■ Úloha:

- Čti blok B1 → buffer
- Zpracuj data v bufferu
- Čti blok B2 → buffer
- Zpracuj data v bufferu
- ...

■ Náklady:

- P = čas zpracování bloku
- R = čas k přečtení 1 bloku
- n = počet bloků ke zpracování

■ Single buffer time = $n(R+P)$

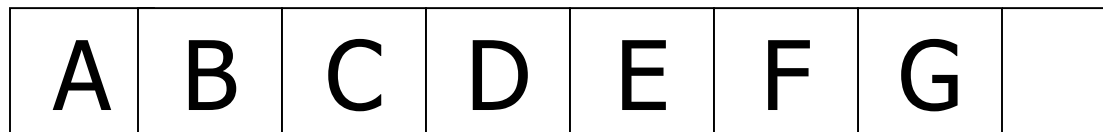
Double Buffering

- Dva buffery v paměti, používané střídavě

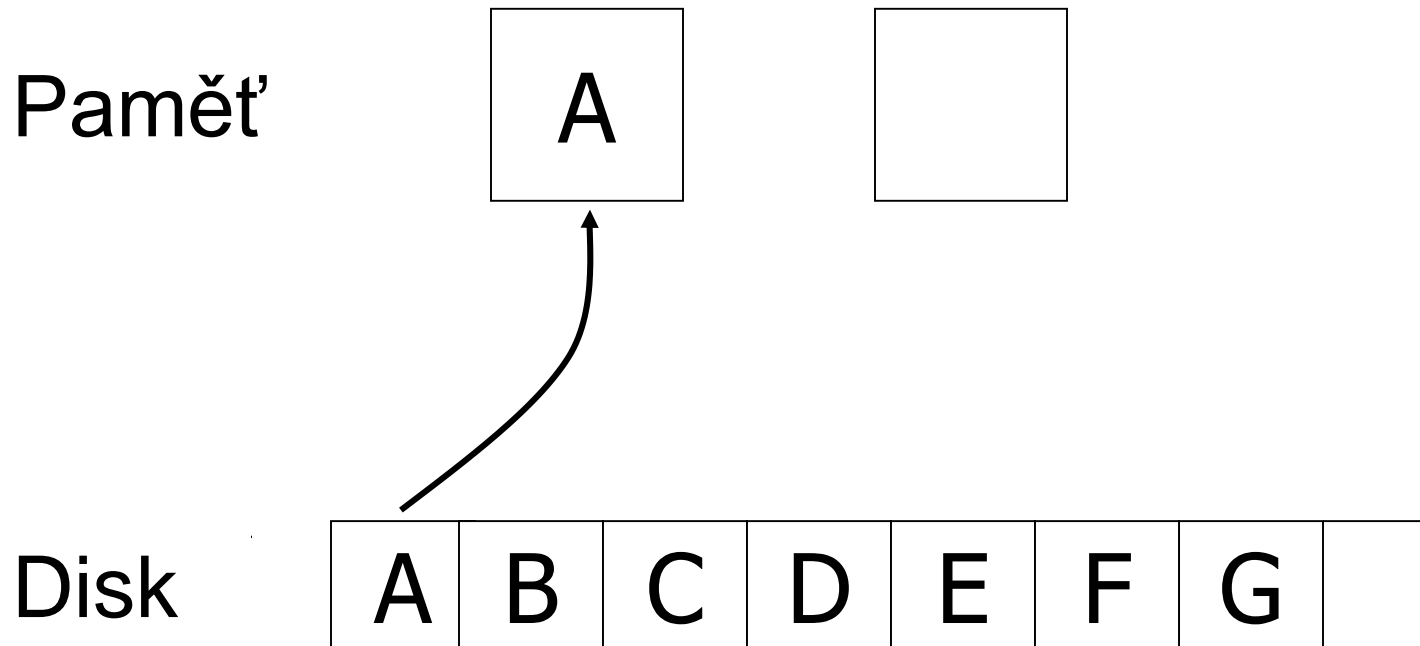
Paměť



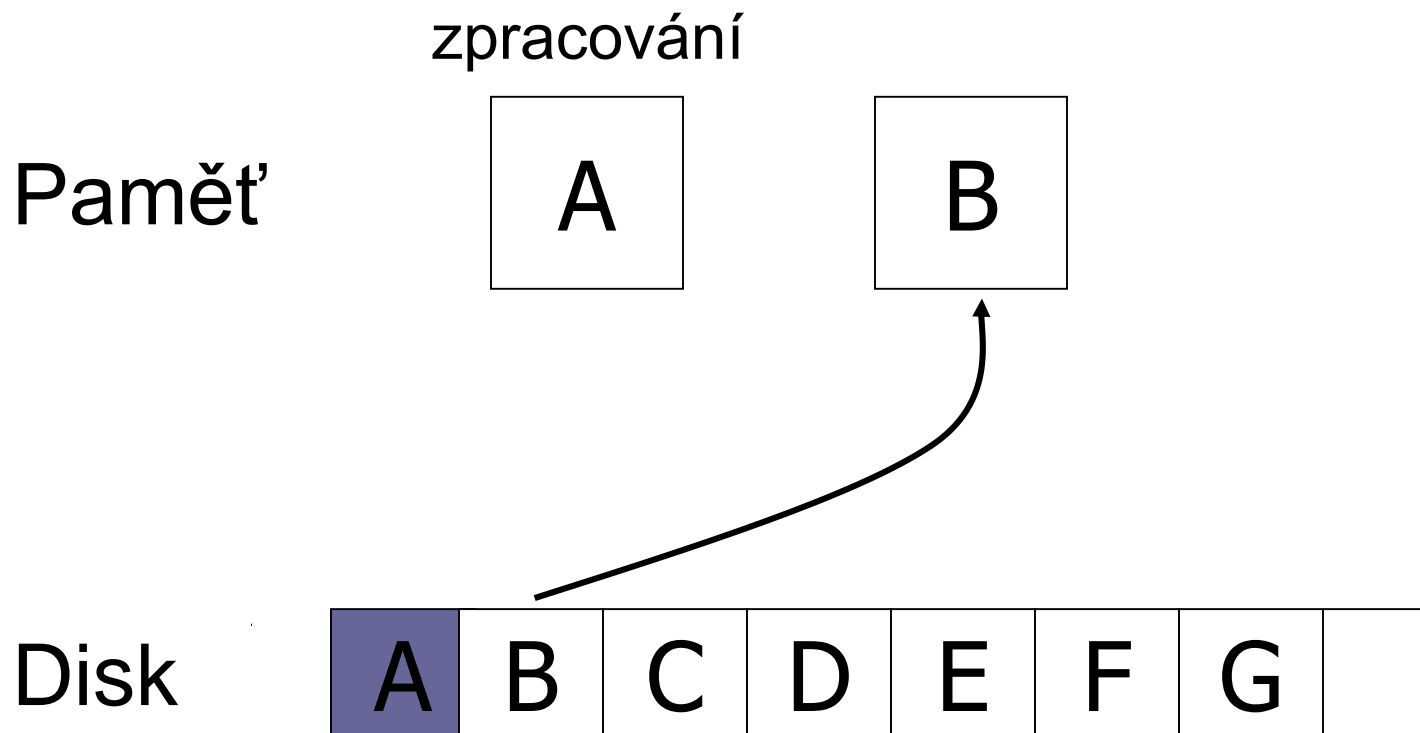
Disk



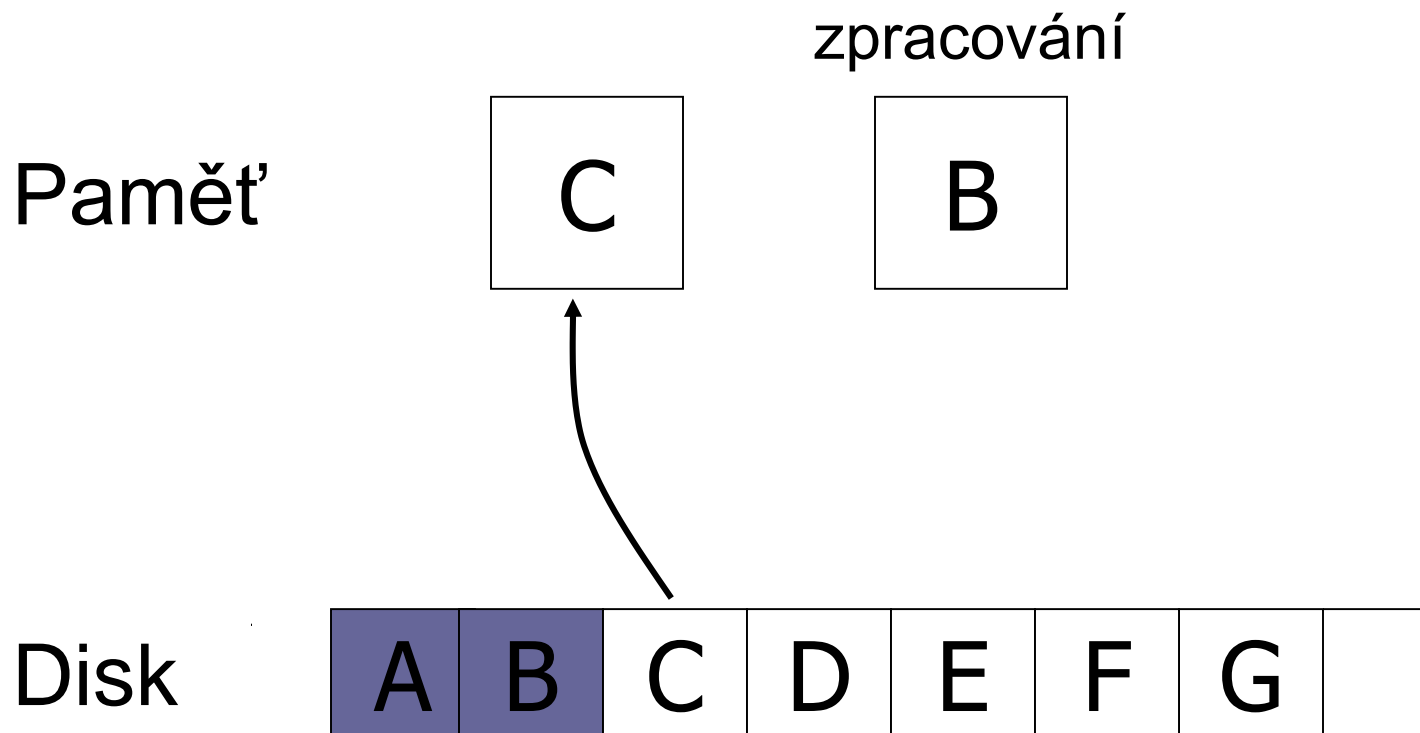
Double Buffering



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 $P \geq R$
- Jinak

- = $nR + P$

Optimalizace přístupu na disk

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

Velikost bloku

- Velký blok → amortizace I/O nákladů

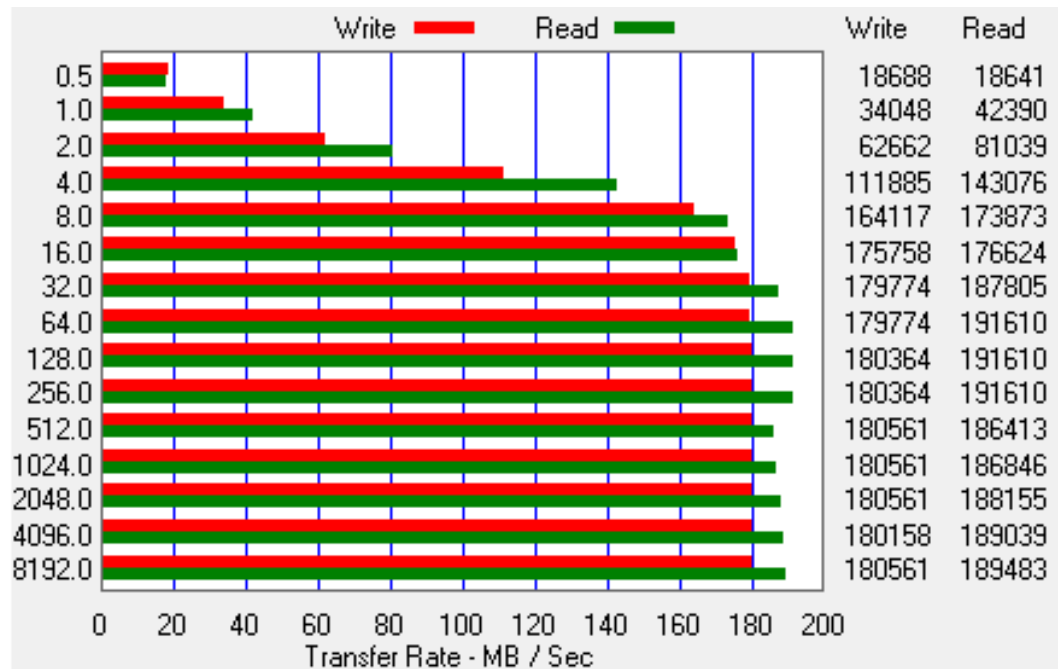
ALE

- Velký blok → čtení více „nepotřebných“ dat, čtení trvá déle
- Trend:
 - cena paměti klesá, data rostou, bloky se zvětšují

Velikost bloku

■ ATTO Disk Benchmark

- 256MB read sequentially block by block
- No caching
- Queue length 4



Western Digital 10EZEX 1TB, SATA3, 7200 RPM, sustained transfer rate 150 MB/s

IO za sekundu

- IOPS dle HD Tune Pro 5.50
 - Reading 4KiB blocks

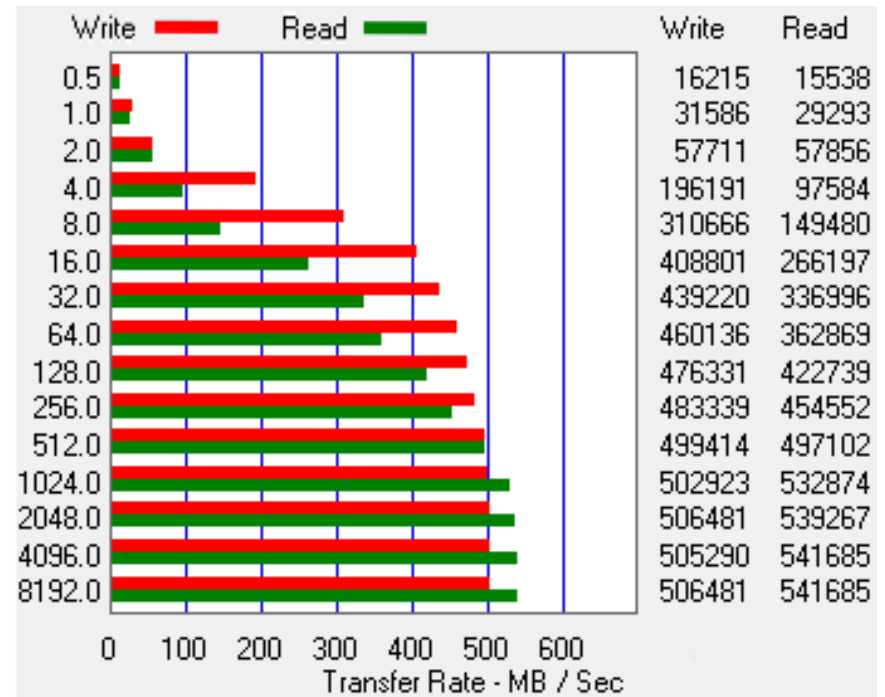
Test	I/O	Time	Transfer
<input checked="" type="checkbox"/> Random seek	65 IOPS	15.269 ms	0.032 MB/s
<input checked="" type="checkbox"/> Random seek 4 KB	65 IOPS	15.427 ms	0.253 MB/s
<input checked="" type="checkbox"/> Butterfly seek	57 IOPS	17.645 ms	0.028 MB/s
<input checked="" type="checkbox"/> Random seek / size 64 KB	64 IOPS	15.522 ms	0.991 MB/s
<input checked="" type="checkbox"/> Random seek / size 8 MB	21 IOPS	47.036 ms	86.209 MB/s
<input checked="" type="checkbox"/> Sequential outer	2798 IOPS	0.357 ms	174.870 MB/s
<input checked="" type="checkbox"/> Sequential middle	2295 IOPS	0.436 ms	143.454 MB/s
<input checked="" type="checkbox"/> Sequential inner	1351 IOPS	0.740 ms	84.414 MB/s
<input checked="" type="checkbox"/> Burst rate	5106 IOPS	0.196 ms	319.134 MB/s

Western Digital 10EZEX 1TB, SATA3, 7200 RPM, sustained transfer rate 150 MB/s

Bloky a IOPS

■ Stejné testy pro SSD

Kingston V300 120GB



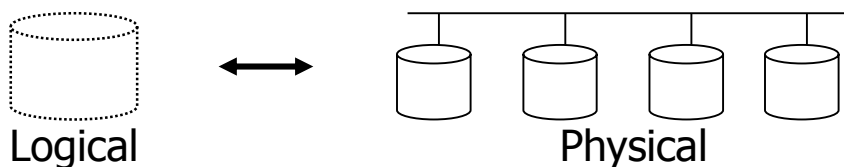
Test	I/O	Time	Transfer
<input checked="" type="checkbox"/> Random seek	5398 IOPS	0.185 ms	2.636 MB/s
<input checked="" type="checkbox"/> Random seek 4 KB	5316 IOPS	0.188 ms	20.764 MB/s
<input checked="" type="checkbox"/> Butterfly seek	5149 IOPS	0.194 ms	2.514 MB/s
<input checked="" type="checkbox"/> Random seek / size 64 KB	4292 IOPS	0.233 ms	65.999 MB/s
<input checked="" type="checkbox"/> Random seek / size 8 MB	118 IOPS	8.461 ms	479.246 MB/s
<input checked="" type="checkbox"/> Sequential outer	3894 IOPS	0.257 ms	243.389 MB/s
<input checked="" type="checkbox"/> Sequential middle	5747 IOPS	0.174 ms	359.183 MB/s
<input checked="" type="checkbox"/> Sequential inner	5816 IOPS	0.172 ms	363.506 MB/s
<input checked="" type="checkbox"/> Burst rate	4194 IOPS	0.238 ms	262.126 MB/s

Optimalizace přístupu na disk

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

Diskové pole

- Více disků uspořádaných do jednoho logického



- Zvětšení kapacity
 - Paralelní čtení / zápis
 - Průměrná doba vystavení hlaviček typicky zachována
- Metody
 - rozdělování dat (data 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 → 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 $(i \bmod n)+1$
- Čtení různých bloků lze paralelizovat
 - Pokud jsou na různých discích
- „Velké“ čtení může využít všechny disky

RAID

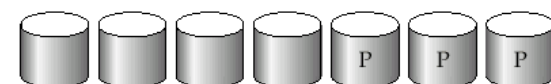
- Redundant Arrays of Independent Disks
- Různé varianty pro různé požadavky
 - Různá výkonnost
 - Různá spolehlivost
- Kombinace variant
 - RAID1+0 (nebo RAID10)
 - = RAID1, pak RAID0



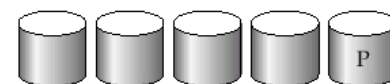
(a) RAID 0: nonredundant striping



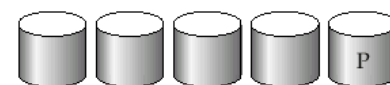
(b) RAID 1: mirrored disks



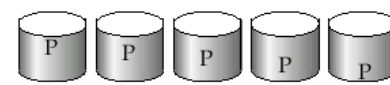
(c) RAID 2: memory-style error-correcting codes



(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, nezvýšená spolehlivost
- 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

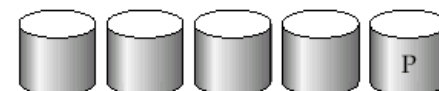
- Bit-striping, Hamming Error-Correcting-Code
- Zotavení z výpadku 1 disku



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

■ RAID3

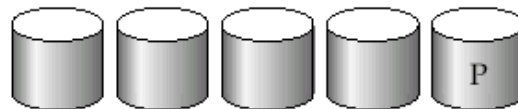
- Byte-striping with parity
- 1 paritní disk
- Zápis: spočítání a uložení parity
- Obnova jednoho disku
 - XOR bitů z ostatních disků



(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
 - Disky nemusí být plně synchronizované



(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 $(n \bmod 5) + 1$
- 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

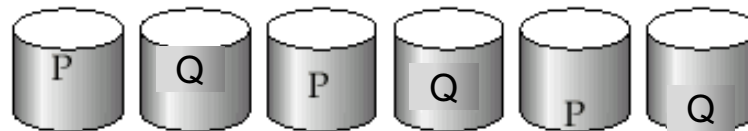
RAID5 (pokrač.)

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

RAID6

■ P+Q Redundancy scheme

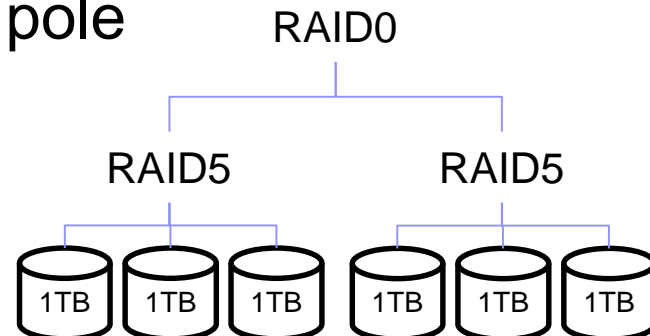
- Podobné RAID5, ale ukládá extra informace pro obnovu při výpadku více disků
- Více disků pro paritu (dual distributed parity)
 - Min. 4 disky v poli (kapacita snížena o 2 disky)
- Samoopravné Hammingovy kódy
 - Opraví výpadek 2 disků
- Vhodný pro vysokokapacitní disky



(g) RAID 6: P + Q redundancy

RAID – kombinace

- Jednotlivé varianty polí lze kombinovat
 - Z fyzických disků se vytvoří pole
 - Pak se z více polí vytvoří jedno 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



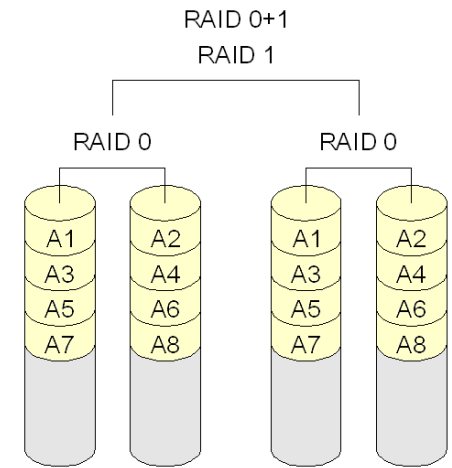
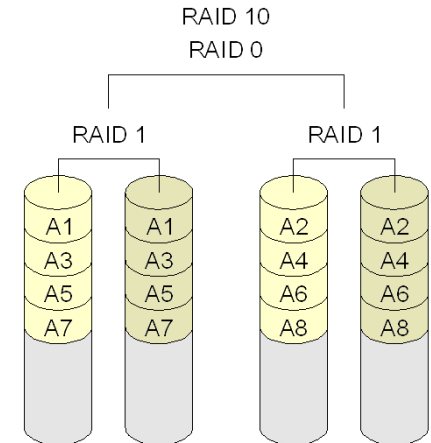
RAID1+0 vs. RAID0+1

■ 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 – dostupnost dat není podstatná
 - Data lze snadno a rychle obnovit (ze záloh,...)
- RAID2, 3 a 4 jsou nahrazeny RAID5
 - bit/byte-striping vede k využití všech disků při zápisu/čtení 1 bloku; nedistribuovaná parita
- RAID6 – méně používaný než RAID5
 - RAID1 a 5 poskytují dostatečnou spolehlivost
 - RAID6 spíše pro velmi velké disky
- 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 typicky 2 čtení a 2 zápisy
 - RAID1+0 vyžaduje pouze 2 zápisy
- Vhodný pro aplikace s menším množstvím zápisů
- Pozor na velikost „stripy“

■ Nároky dnešních aplikací na počet I/O

- Velmi vysoké (např. WWW servery, DB, ...)
- 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, popř. BIOS
 - 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
 - Přítomnost 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

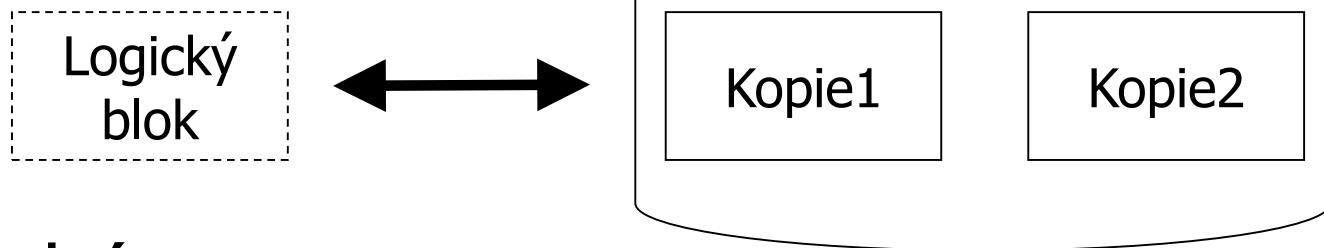
- super-blok; pro data ZFS

- Žurnálování (log/záznam změn)

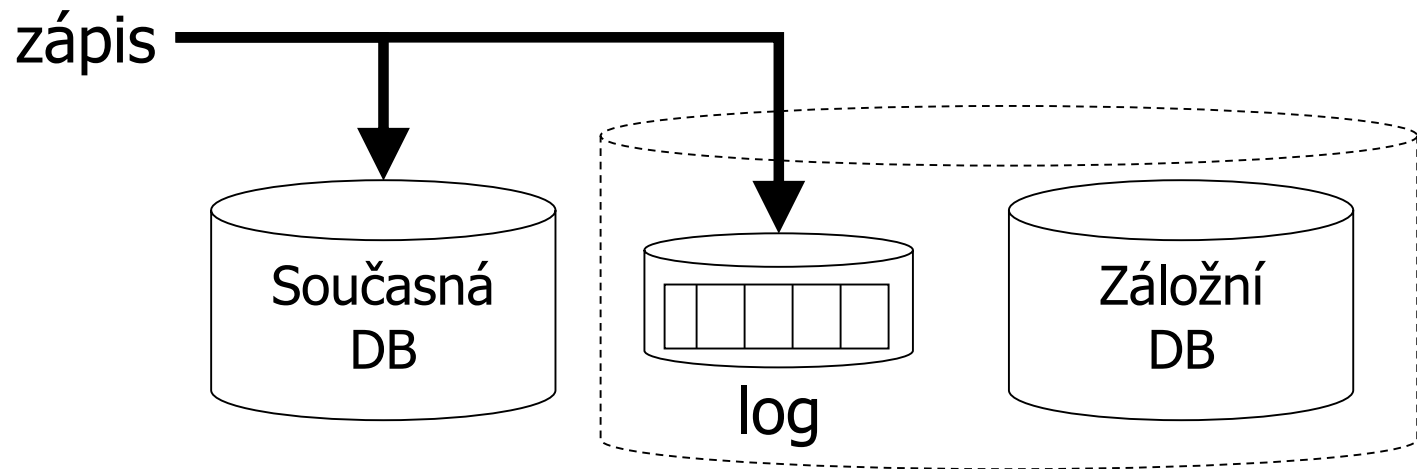
- Diskové pole

Stabilní uložení v databázích

■ Operační systém



■ Databáze



Výpadky

■ 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% $\Rightarrow P_{\text{výpadku za rok}} = 0,44\%$
 - \Rightarrow **Annualized Failure Rate (AFR)**

Výpadky – pokračování

■ Příklad:

- MTTF 1 000 000 hours
- \Rightarrow v populaci 2 000 000 disků
 - každou hodinu vypadne jeden disk
 - tj. 8 760 disků ročně
 - \Rightarrow pravděpodobnost výpadku za rok = 0,44%

Výpadky – pokračování

- Alternative measure (e.g. Western Digital)

- Annualized Failure Rate (**AFR**)

- Component Design Life

- 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 a výrobci

■ 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 = hundreds of disks operating at 42°C ambient temperature

Výpadky a výrobci

- Vliv teploty na MTTF pro první rok

- Seagate:

Temp (°C)	Acceleration Factor	Derating Factor	Adjusted MTTF
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 a výrobci

■ Seagate Barracuda ES.2 Near-Line Serial ATA drive

		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%

Note1: Weibull – SW pro modelování průběhu chybovosti

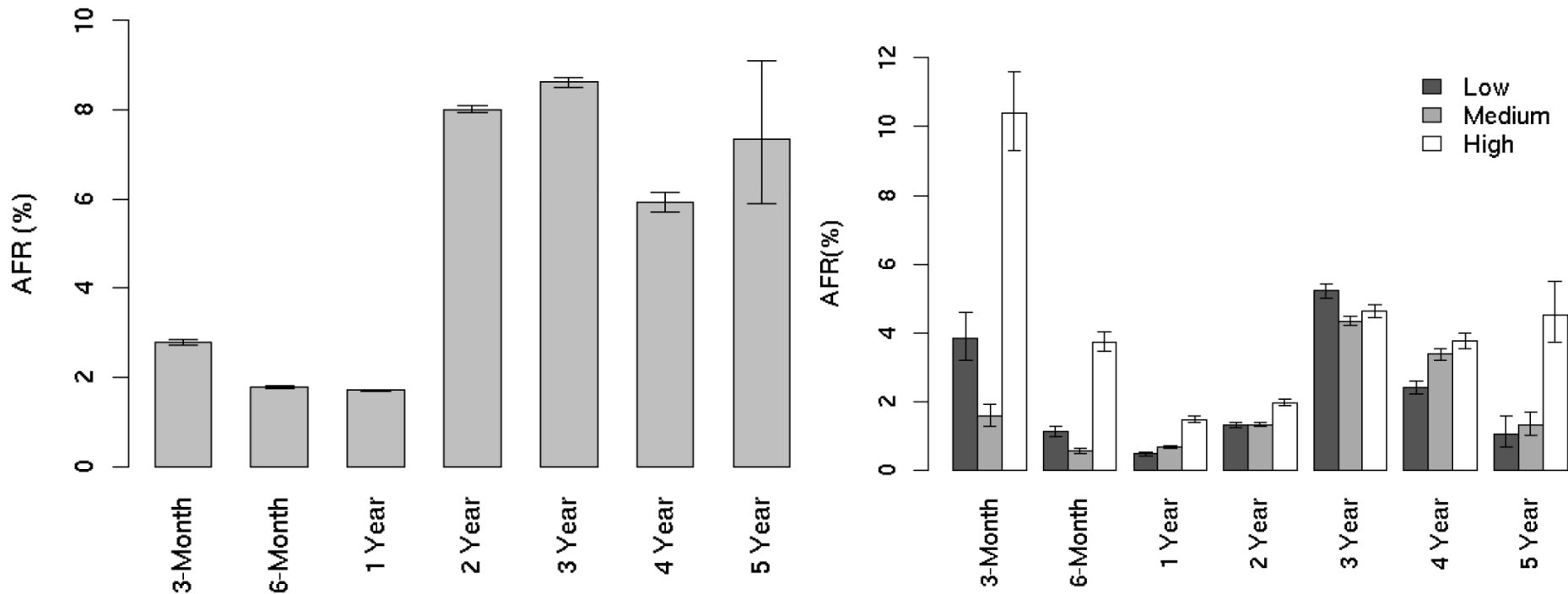
Note2: 2400 hours/yr => 6.5 hrs a day!

Výpadky – realita

■ Google http://research.google.com/archive/disk_failures.pdf (Konference FAST 2007)

□ Test na 100 000 discích

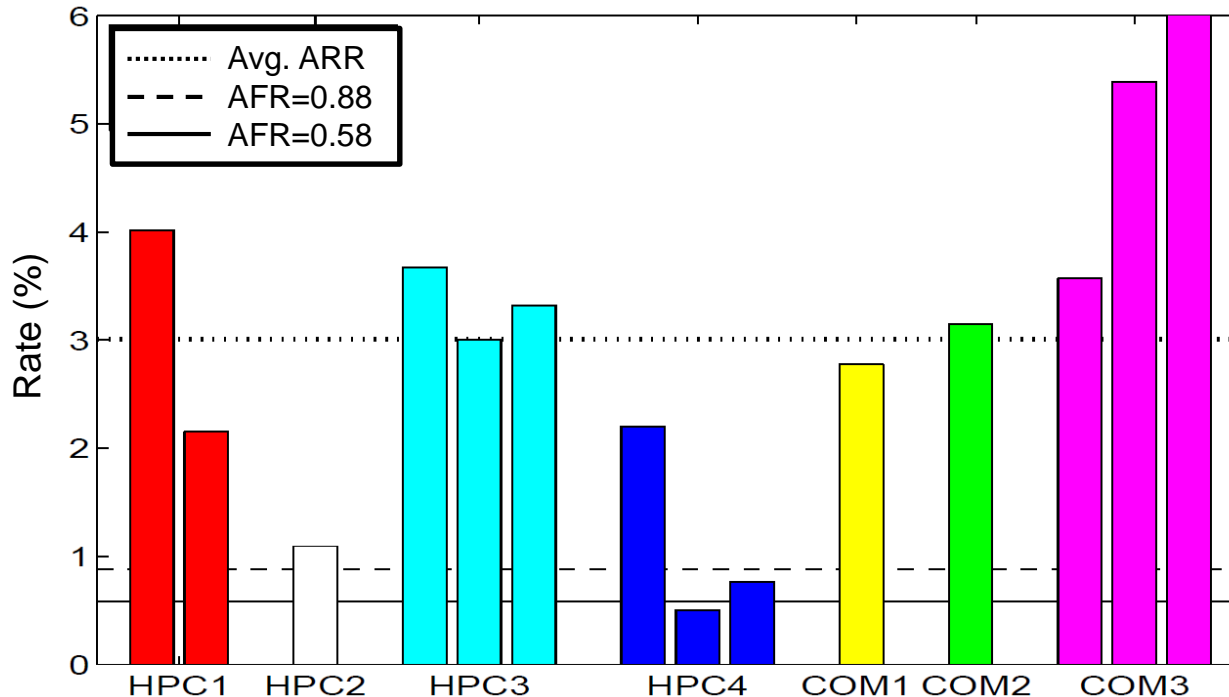
■ SATA, PATA disky; 5400-7200 rpm; 80-400 GB



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

- We know $AFR = 1 / (2 * MTTF)$
- Mean Time To Repair (MTTR)
 - Čas od výpadku do obnovení činnosti
 - = čas výměny vadného disku + obnovení dat
 - $P_{\text{Failure During Repair}} = P_{\text{FDR}} = (2 * MTTR) / 1 \text{ rok}$
 - Předpoklad: velmi krátká doba
- Mean Time To Data Loss (MTTDL)
 - Závisí na AFR a MTTR
 - Průměrná doba do ztráty dat
 - Pro jeden disk (tj. data ukládám na jednom disku)
 - $MTTDL = MTTF = 0.5 / AFR$

Pozor na jednotky! Roky vs. hodiny

Oprava chyby – sada disků

■ Předpoklad

- Chyba každého disku je stejně pravděpodobná a nezávislá na ostatních

■ Příklad

□ Jeden disk

- $AFR_{1 \text{ disk}} = 0,44\%$ (MTTF = 1,000,000 hrs. = 114 yrs.)

□ Systém se 100 disky (MTTF_{100 disků} = MTTF_{1 disk} / 100)

- $AFR_{100 \text{ disků}} = 44\%$ (MTTF = 10,000 hrs. = 1.14 yrs.)

- *Průměrně* každý rok cca jeden z disků vypadne

- Pravděpodobnost (alespoň 1 z n a právě 1 z n havaruje)

- $P_{\text{výpadek právě 1 ze 100}} = 28,43\%$ $P_{\text{výpadek alespoň 1 ze 100}} = 35,66\%$

- $P_{\text{výpadek právě 1 z 10}} = 4,23\%$ $P_{\text{výpadek alespoň 1 z 10}} = 4,31\%$

- $AFR_{n \text{ disků}} = AFR_{1 \text{ disk}} * n$

Příklad výpadku RAID1

- 2 zrcadlené disky
 - každý AFR=3%
- Výměna vadného a obnova pole do 3 hodin
 - MTTR = 3 hodiny
- Pravděpodobnost ztráty dat:
 - $P_{\text{výpadku 1 disku}} = \text{AFR} = 0.03$
 - $P_{\text{výpadku 1 ze 2}} = 0.06$
 - $P_{\text{FDR}} = 2 * \text{MTTR} / 1 \text{ rok} = 2*3 / 8760 = 0,000 685$
 - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} * P_{\text{FDR}} * P_{\text{výpadku 1 disku}}$
 $= 0,000 001 233$
 - **MTTDL** = $0.5 / P_{\text{ztráty dat}} = 405 515 \text{ let}$

Příklad výpadku RAID0

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

Příklad výpadku RAID4

- AFR disku 3% ($P_{\text{výpadku 1 disku}}$)
- RAID4 – opravuje výpadek 1 disku
 - 4 disky (3+1)
 - MTTR = 3 hodiny
 - $P_{\text{FDR}} = 2 \cdot 3 / 8760 = 0,000\ 685$
 - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 4}} * P_{\text{FDR}} * P_{\text{výpadku 1 ze 3}}$
 - $P_{\text{ztráty dat}} = 4 \cdot 0,03 * 1/2920 * 3 \cdot 0,03$
 $= 108 / 2\ 920\ 000 = 0,000\ 003\ 698$
 - což je AFR tohoto pole
 - $\text{MTTDL} = 0.5 / P_{\text{ztráty dat}} = 135\ 185 \text{ let}$

Spolehlivost pole

- n disků
 - celkem disků v poli (včetně paritních)
- 1 paritní disk
 - zajišťují redundanci dat
- $AFR_{\text{pole}} = n * AFR_{1 \text{ disku}} * P_{\text{FDR}} * (n-1) * AFR_{1 \text{ disku}}$
- $MTTDL = 0.5 / AFR_{\text{pole}}$

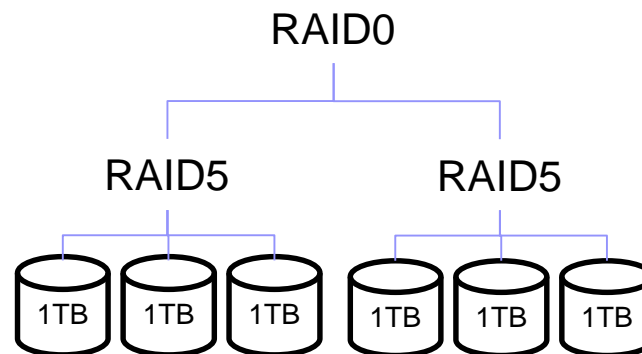
Příklad výpadku – kombinace RAID

■ Kombinace polí

- Spočítám AFR pro složky

- Toto použiji v dalším jako AFR „virtuálního disku“

- Pak vypočítám výsledné MTDDL



Příklad výpadku – kombinace RAID

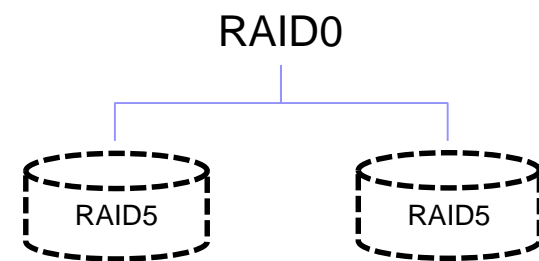
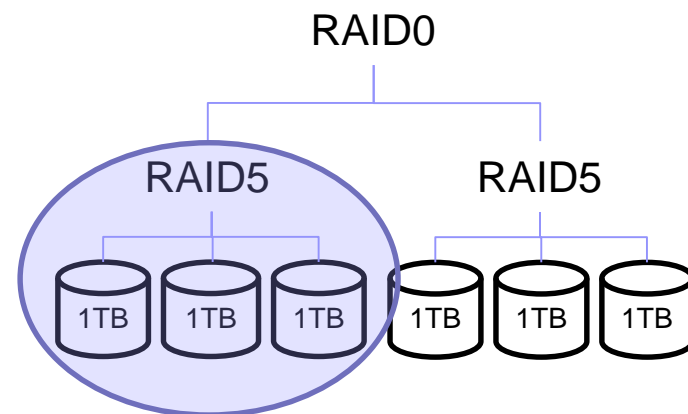
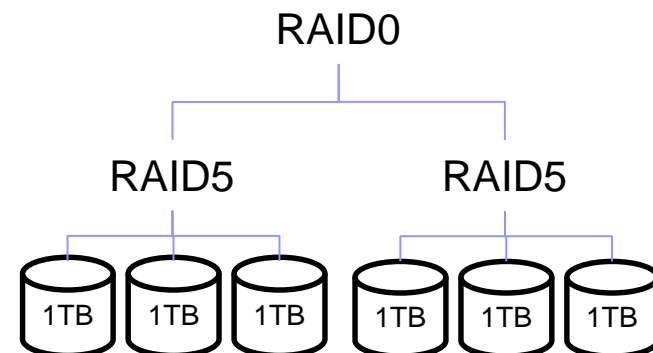
■ RAID50

□ 1 disk má AFR_{disk}

1) Urči AFR_{RAID5}

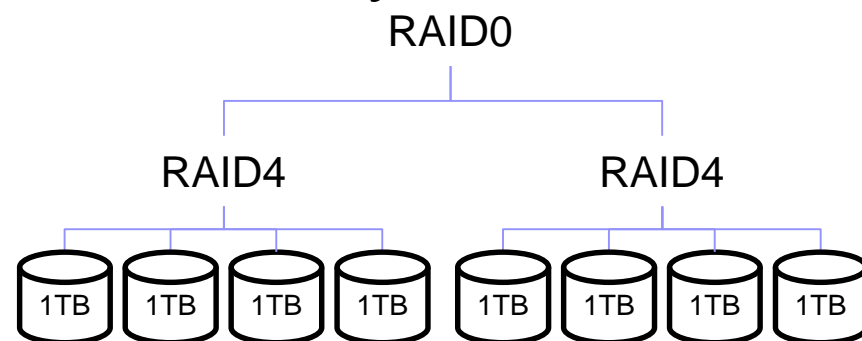
2) Urči $AFR_{\text{RAID0}} = 2 * AFR_{\text{RAID5}}$

3) $MTTDL_{\text{RAID50}} = 0.5 / AFR_{\text{RAID0}}$



Příklad výpadku – kombinace RAID

- RAID40 z 8 disků v konfiguraci níže
 - 1 disk AFR=3%, MTTR = 3 hodiny



- Vždy ze 4 disků vyrobíme 1x RAID4
 - $AFR_{RAID4} = 4 * AFR * P_{FDR} * 3 * AFR = \dots = 3.7 * 10^{-6}$
- 2x RAID4 spojíme pomocí RAID0
 - $AFR_{RAID40} = 2 * AFR_{RAID4} = 7.4 * 10^{-6}$
 - $MTTDL = 0.5 / AFR_{RAID40} = 67\ 567$ let

Failures: „Write Hole“ Phenomenon

- = *Data is not written to all disks.*
- Severity
 - Can be unnoticed
 - Discoverable during array reconstruction
- Solution
 - UPS
 - Synchronize the array
 - Journaling
 - but with “data written” commit message (-:
 - Special file system (ZFS)
 - uses "copy-on-write" to provide write atomicity
 - provides continuous integrity checking

File Systems

- Storing a data block:
 1. Add data block to list of used space
 2. Write data block
 3. Write file metadata referencing that data block
- More modern – use journaling
 - Start transaction in journal
 - Store info about steps 1.-3. to journal
 - Do steps 1.-3.
 - End transaction in journal

File System Tuning

- FS block size \leq DB block size
 - ZFS has 128KB by default!
- DB journal (WAL in PostgreSQL)
 - ext2; ext3/4 with data=writeback
- DB data
 - ext3/4 with data=ordered (only metadata journaled)
- Switch off *file access times* (noatime)
- Eliminate swapping (vm.swappiness = 0)
- Process memory allocation (vm.overcommit_memory = 2)
- ...

RAID nad disky SSD

■ SSD – problém opotřebení

- Omezený počet zápisů se řeší přesuny do jiných oblastí, tzv. wear-leveling
- Důsledek: po jisté době dojde k totálnímu výpadku

■ RAID nad SSD

- Zabezpečení dostupnosti dat horší
 - Téměř jistota, že SSD vypadnou naráz
- Diff-RAID
 - Distributes parity unevenly
 - After replace a failed SSD with a brand new one, parity is moved primarily to the most worn-out drive.