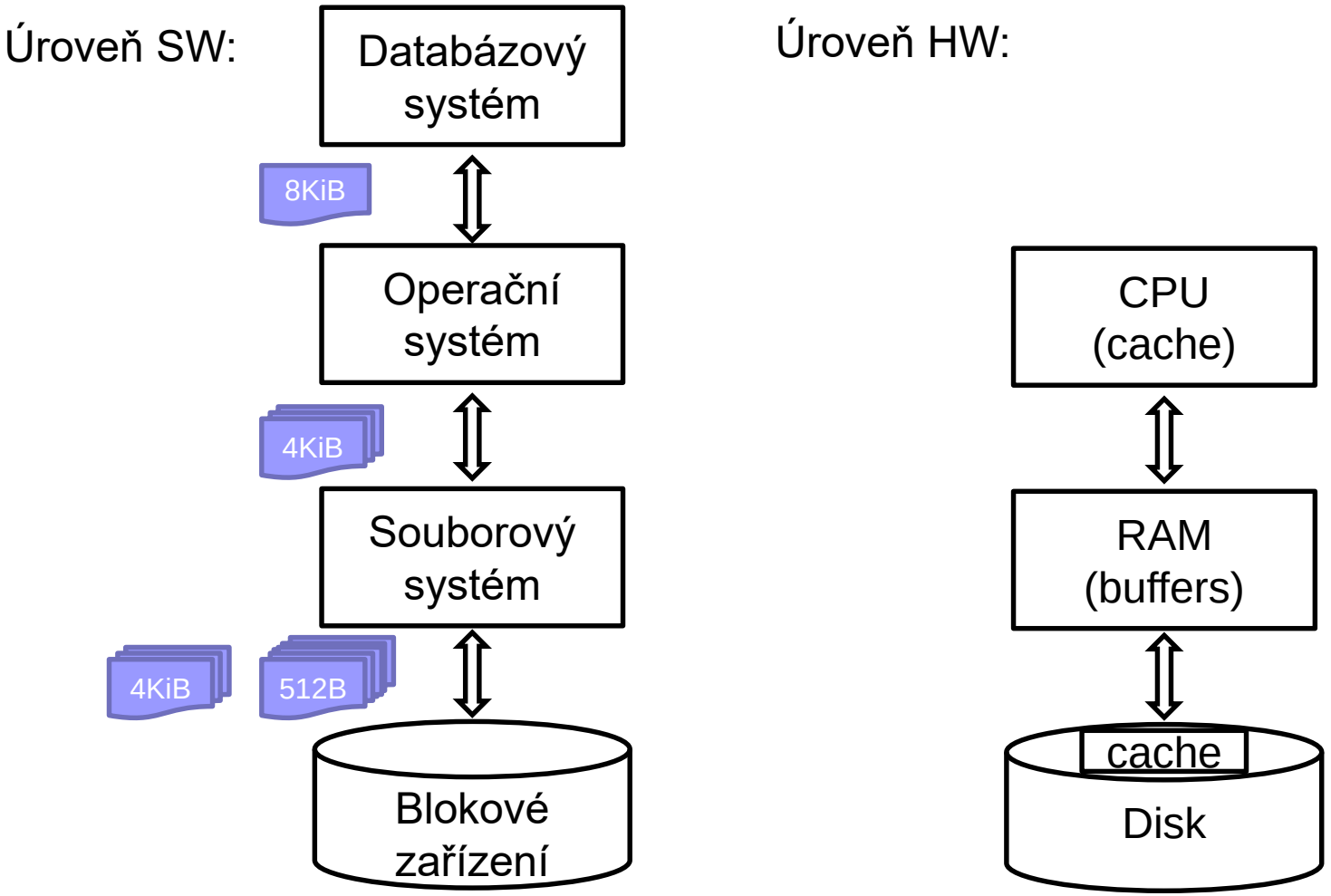
PA152: Efektivní využívání DB

2. Datová úložiště

Vlastislav Dohnal

Pohyb dat – přehled



Optimalizace přístupu na disk

- Techniky přístupu
  - Eliminace náhodných přístupů,…
- Objem dat
  - Velikost bloku
- Organizace úložiště
  - Diskové pole

Techniky přístupu k datům

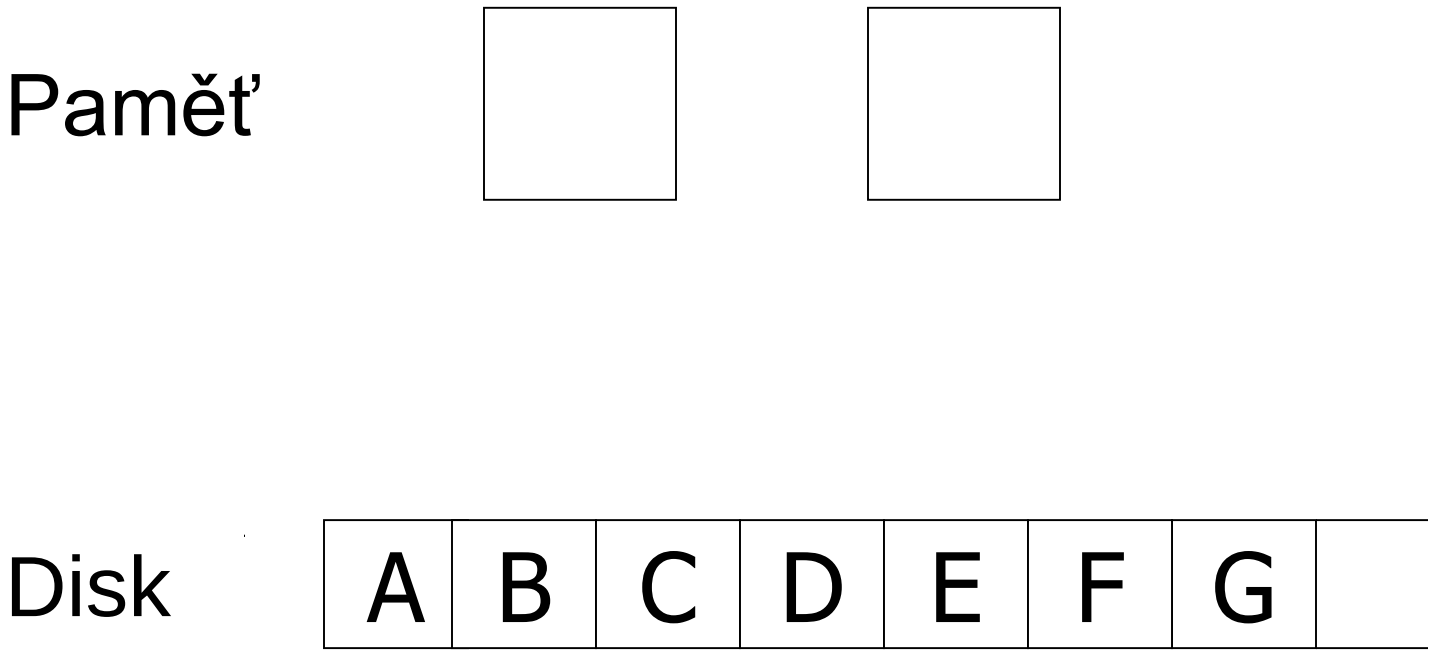
- App: Double buffering
- OS: Prefetching
- OS: Defragmentace
  - Uspořádání bloků do pořadí jejich zpracování
  - Souborový systém
    - Alokace více bloků naráz, nástroje pro defragmentaci
- HW: Plánování přístupů (výtah)
  - Pohyb hlavičky pouze jedním směrem
  - Přeuspořádání požadavků na disk
    - Při zápisu použití zálohované cache (nebo žurnálu)

Single Buffering

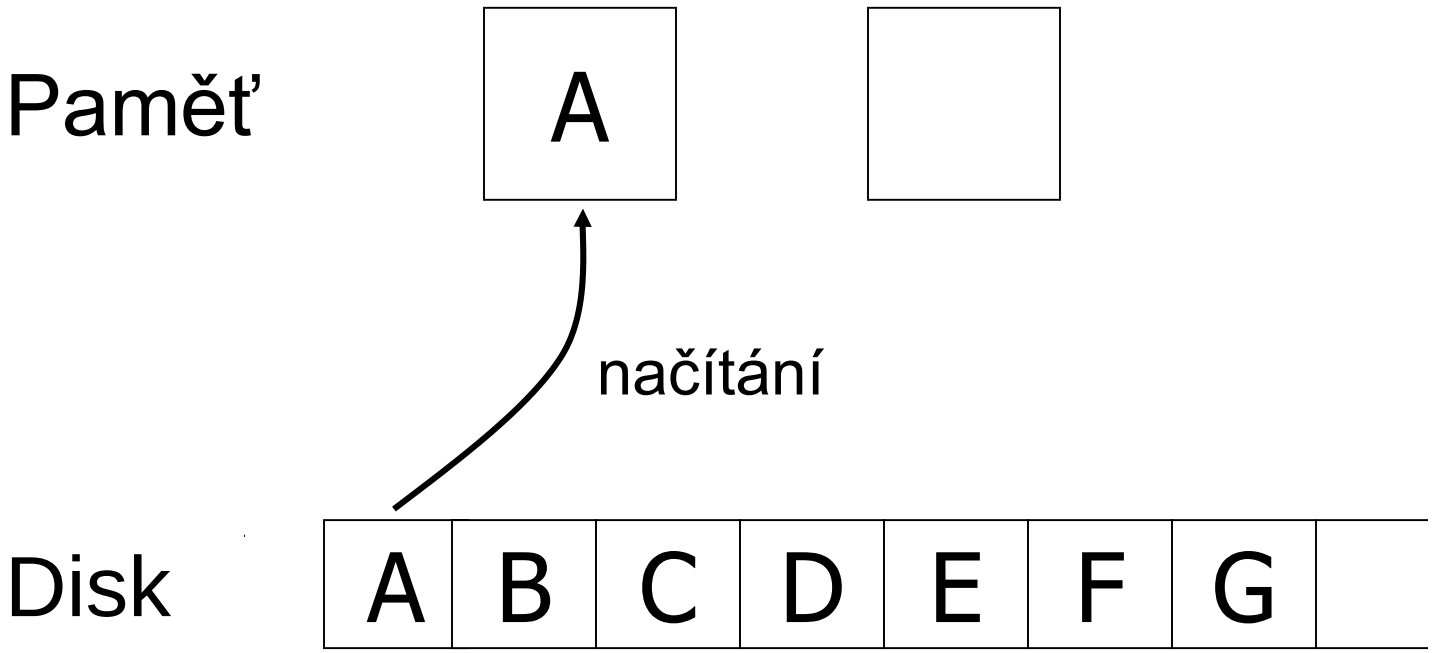
- Úloha:
  - Čti blok B1 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - Čti blok B2 $\rightarrow$ buffer
  - Zpracuj data v bufferu
  - …
- Náklady:
  - P = čas zpracování bloku
  - R = čas k přečtení 1 bloku
  - n = počet bloků ke zpracování
- Single buffer time = n(R+P)

Double Buffering

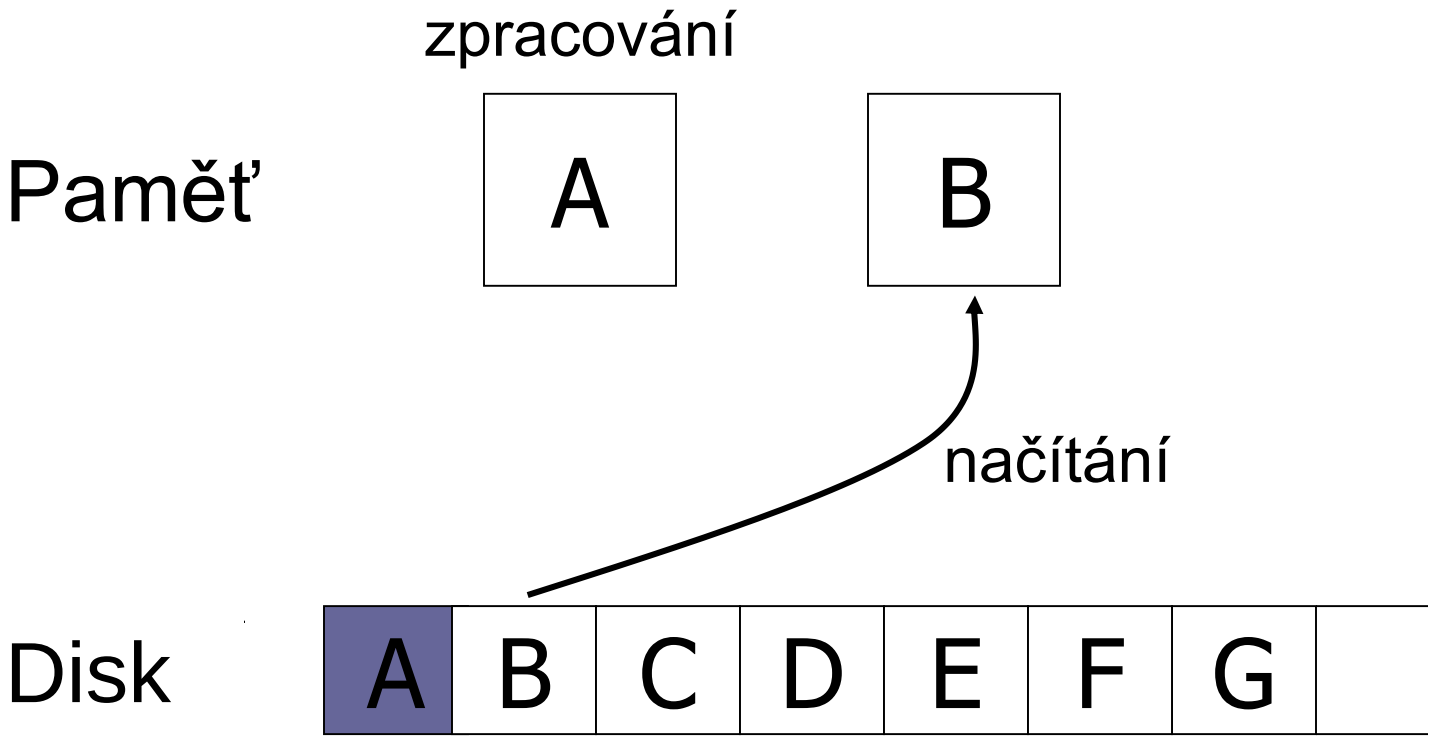
- Dva buffery v paměti, používané střídavě



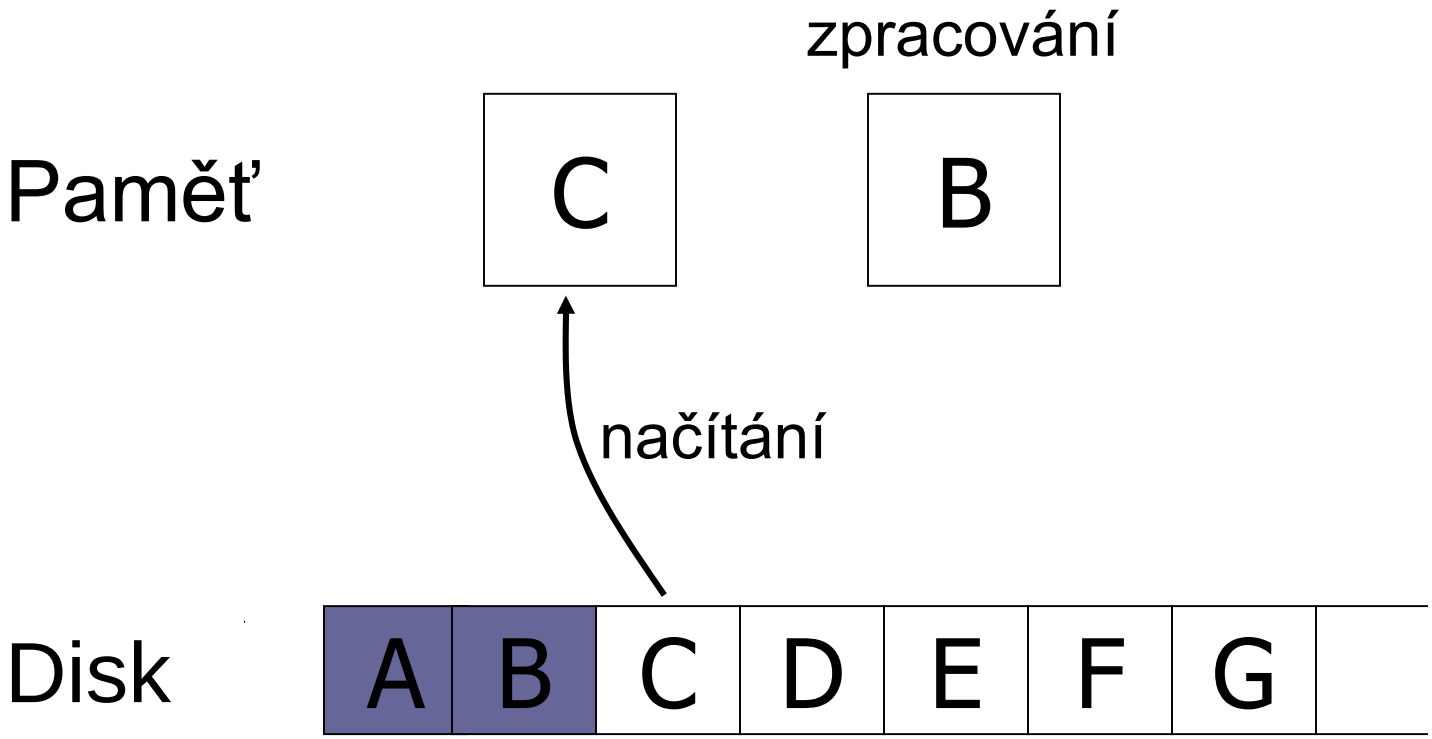
Double Buffering



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

- Techniky přístupu
  - Eliminace náhodných přístupů,…
- *Objem dat*
  - *Velikost bloku*
- Organizace úložiště
  - Diskové pole

Velikost bloku

- Velký blok → amortizace I/O nákladů

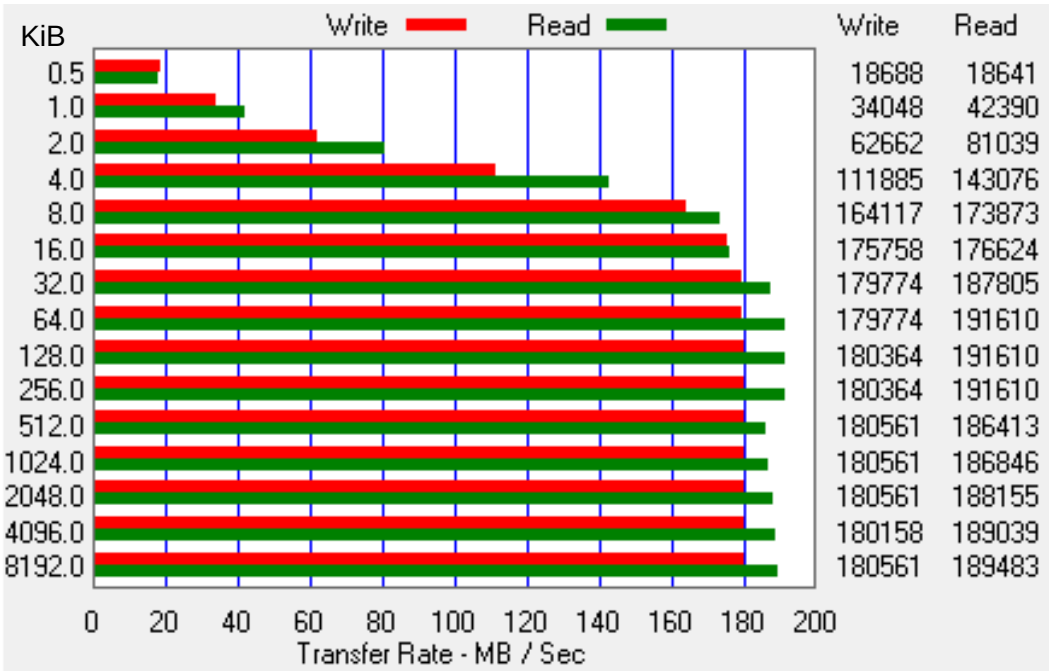
  ALE

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

- Trend:
  - cena pamětí 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

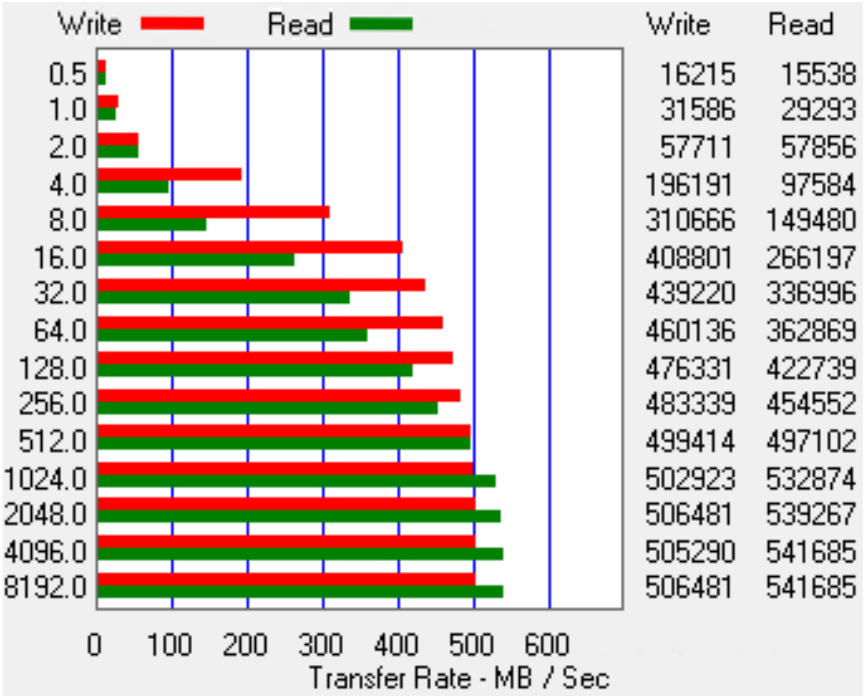
Test	I/O	Time	Transfer
Random seek	65 IOPS	15.269 ms	0.032 MB/s
Random seek 4 KB	65 IOPS	15.427 ms	0.253 MB/s
Butterfly seek	57 IOPS	17.645 ms	0.028 MB/s
Random seek / size 64 KB	64 IOPS	15.522 ms	**4.096 MB/s**
Random seek / size 8 MB	21 IOPS	47.036 ms	**168.000 MB/s**
Sequential outer	2798 IOPS	0.357 ms	174.870 MB/s
Sequential middle	2295 IOPS	0.436 ms	143.454 MB/s
Sequential inner	1351 IOPS	0.740 ms	84.414 MB/s
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



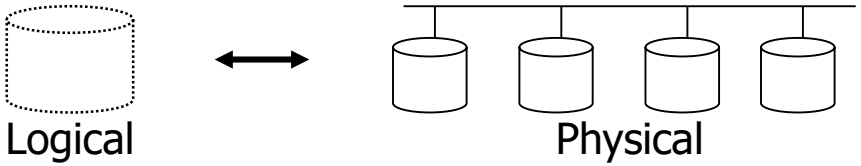
Test	I/O	Time	Transfer
Random seek	5398 IOPS	0.185 ms	2.636 MB/s
Random seek 4 KB	5316 IOPS	0.188 ms	20.764 MB/s
Butterfly seek	5149 IOPS	0.194 ms	2.514 MB/s
Random seek / size 64 KB	4292 IOPS	0.233 ms	**268.250 MB/s**
Random seek / size 8 MB	118 IOPS	8.461 ms	**944.000 MB/s**
Sequential outer	3894 IOPS	0.257 ms	243.389 MB/s
Sequential middle	5747 IOPS	0.174 ms	359.183 MB/s
Sequential inner	5816 IOPS	0.172 ms	363.506 MB/s
Burst rate	4194 IOPS	0.238 ms	262.126 MB/s

Optimalizace přístupu na disk

- Techniky přístupu
  - Eliminace náhodných přístupů,…
- Objem dat
  - Velikost bloku
- *Organizace úložiště*
  - *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

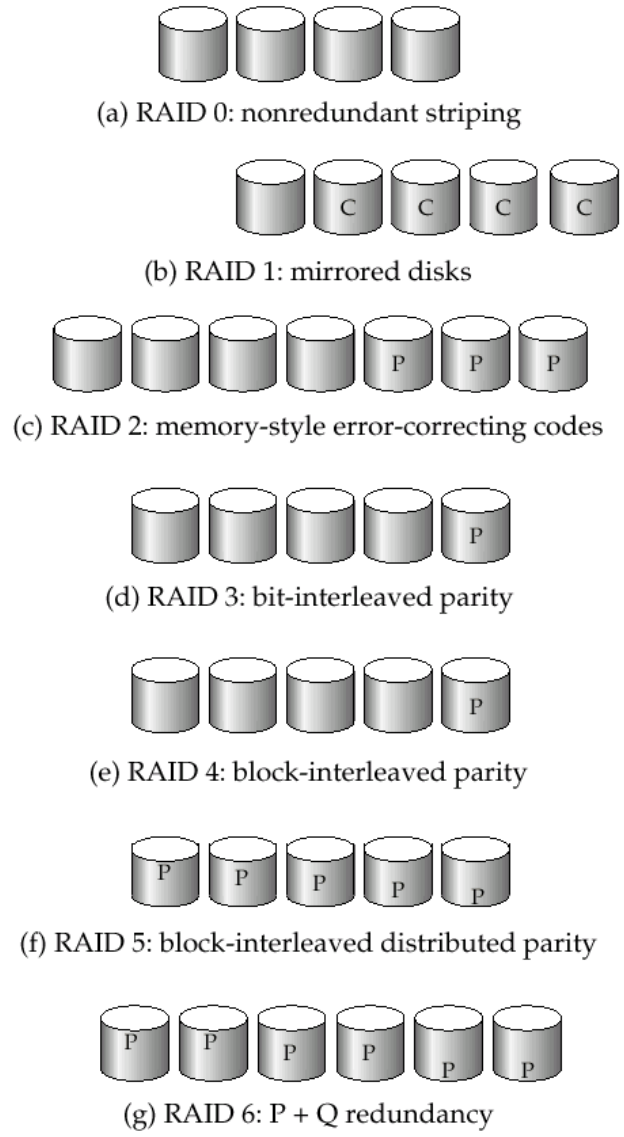
- Cíle:
  - Zvýšení přenosové rychlosti rozdělením na více disků
  - Paralelizace „velkého“ čtení ke snížení odezvy
  - Vyrovnání zátěže → zvýšení propustnosti
  - Snížení spolehlivosti

Rozdělování dat

- Bit-level striping
  - Rozdělení každého bajtu na bity mezi disky
  - Přístupová doba je horší než u jednoho disku
  - Málo používané
- 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

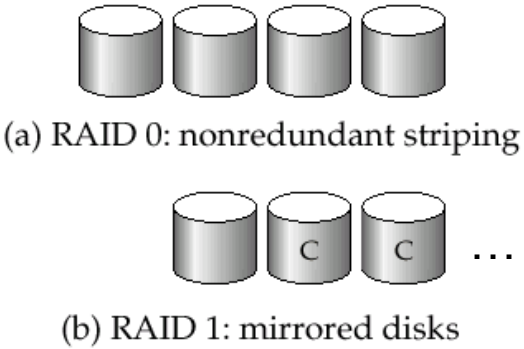
RAID

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



RAID0, RAID1

- RAID0
  - Block striping, neredundantní
  - Velmi vysoký výkon, snížená 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
  - RAID1E – kombinuje zrcadlení a dělení

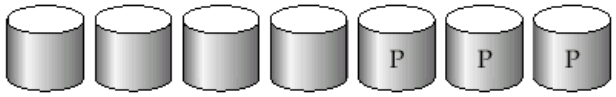


(a) RAID 0: nonredundant striping

(b) RAID 1: mirrored disks

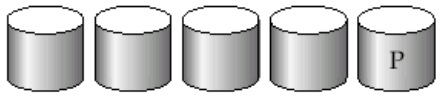
RAID2, RAID3

- RAID2
  - Bit-striping, Hamming Error Correcting Code
  - Zotavení z výpadku 1 disku
  - Nespoléhá na detekci chyby diskem



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

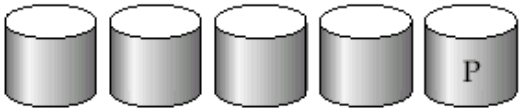
- RAID3
  - Byte-striping with parity
  - 1 paritní disk, chyby detekovány diskem
  - Zápis: spočítání a uložení parity
  - Obnova jednoho disku
    - XOR bajtů z ostatních disků



(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 bloků 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

  - Paritní blok pro *i*-tý blok je uložen na disku $\lfloor i/_{n-1} \rfloor \bmod n$
- Příklad (5 disků)
  - Parita pro blok *i* je na: $\lfloor i/_{4} \rfloor \bmod 5$

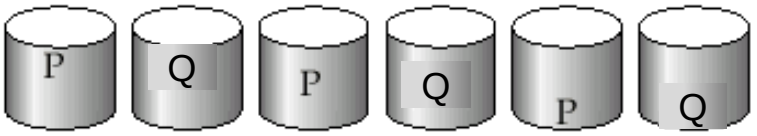
P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

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

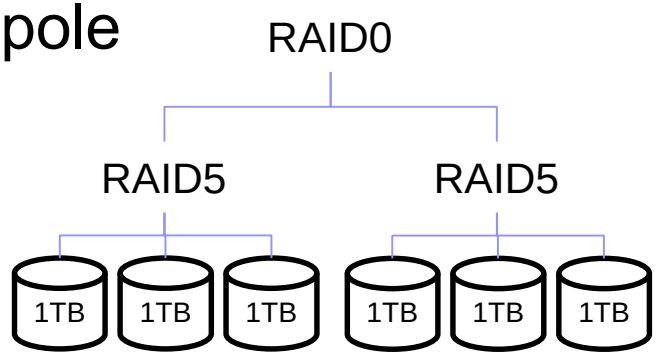
- P+Q Redundancy scheme
  - Podobné RAID5, ale ukládá extra informace pro obnovu při výpadku více disků
  - Dva disky pro paritu (dual distributed parity)
    - Min. 4 disky v poli (kapacita snížena o 2 disky)
  - Hammingovy samoopravné kódy
    - Opraví výpadek 2 disků
  - Vhodný pro vysokokapacitní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



(g) RAID 6: P + Q redundancy

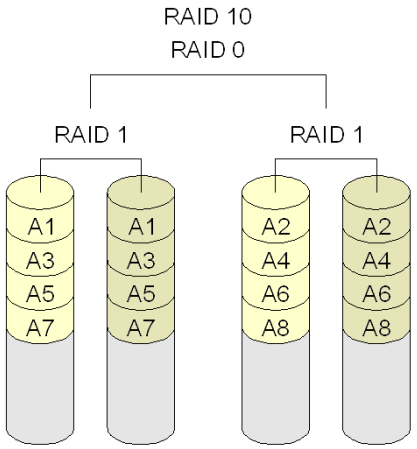
RAID – kombinace

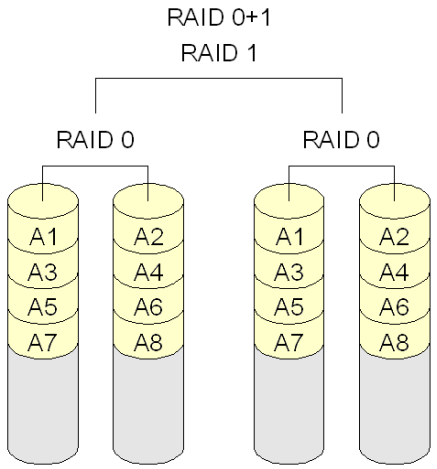
- Jednotlivé varianty polí lze kombinovat
  - Z fyzických disků se vytvoří pole
  - Pak se z více polí vytvoří jedno výsledné pole
- Vhodné ke zvýšení výkonu / spolehlivosti
- Příklad:
  - RAID5+0 z 6 fyzických disků
    - Po třech vytvoříme dvě RAID5 pole
    - RAID5 pole spojíme do RAID0



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
$\Rightarrow$ 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 využívá všechny disky při 1 zápisu/čtení; resp. nedistribuovaná parita
- RAID6 – stále častěji používaný
  - 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, DBMS, …)
  - 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

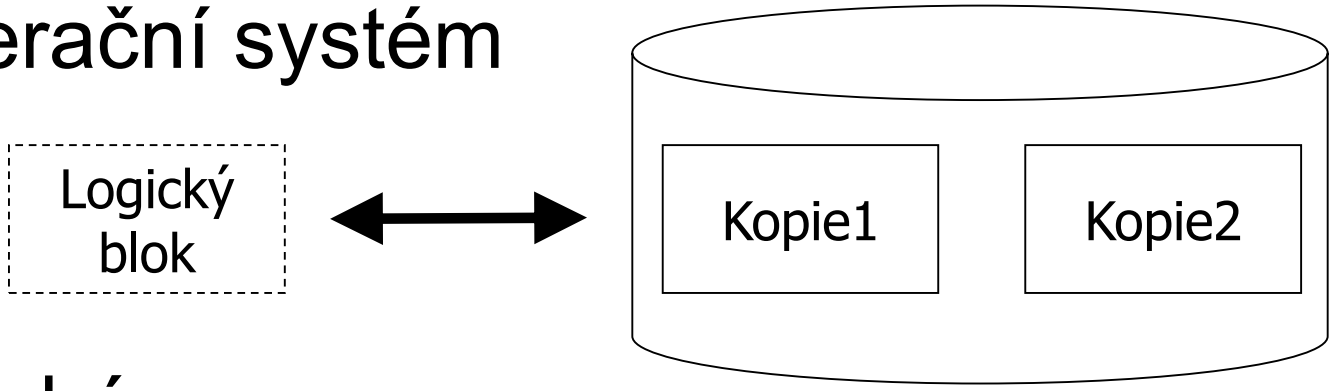
- Občasný výpadek
  - Chyba při čtení/zápisu → opakování → OK
- Vada média
  - Trvalá chyba nějakého sektoru
  - Moderní disky samy detekují a opraví
    - z vlastní rezervní kapacity
- Zničení disku
  - Totální výpadek → výměna disku

Ošetření výpadků disků

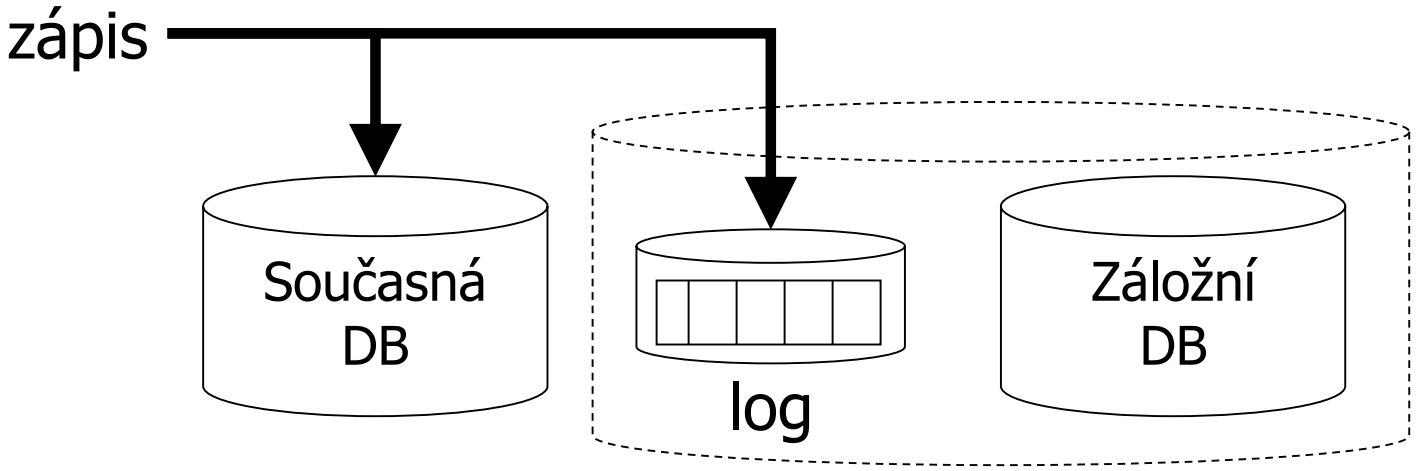
- Detekce
  - Kontrolní součty
- Opravy
  - Stabilní uložení
    - Diskové pole
    - Uložení na více místech stejného disku
      - super-blok; pro data ZFS
    - Žurnálování (log/záznam změn)
  - Samoopravné kódy (ECC)
    - Hammingovy kódy, …

Stabilní uložení v databázích

- Operační systém



- Databáze



Samoopravné kódy

- Paritní bit = sudá/lichá parita
  - Použito v RAID3,4,5
- Příklad sudá parita
  - RAID4 se 4 disky, blok č. 1:

```
Disk 1: 11110000…        Disk 1: 11110000…
Disk 2: 10101010…        Disk 2: ????????…
Disk 3: 00111000…        Disk 3: 00111000…
Disk P: 01100010…        Disk P: 01100010…
```

Samoopravné kódy

- Algebra s operací součtu modulo-2
  - Sudá parita, tj. dopočtení na sudý počet jedniček
  - $\vec{x} \oplus \vec{y} = \vec{y} \oplus \vec{x}$
  - $\vec{x} \oplus (\vec{y} \oplus \vec{z}) = (\vec{x} \oplus \vec{y}) \oplus \vec{z}$
  - $\vec{x} \oplus \vec{0} = \vec{x}$
  - $\vec{x} \oplus \vec{x} = \vec{0}$
- Když $\vec{x} \oplus \vec{y} = \vec{z}$, pak $\vec{y} = \vec{x} \oplus \vec{z}$
  - Přidáním $\vec{x}$ na obě strany…

Samoopravné kódy

- Hamming code
  - Example to recover from 2 crashes
    - 7 disks, four are data disks
  - Redundancy schema:
    - Parity disk contains even parity
    - Parity computed from data disks denoted 1

	Data				Parity		
Disk No:	1	2	3	4	5	6	7
	1	1	1	0	1	0	0
	1	1	0	1	0	1	0
	1	0	1	1	0	0	1

Samoopravné kódy (pokr.)

- Hamming code
  - Content sample and writing

Disk No:	Data 1	2	3	4	Parity 5	6	7
	1	1	1	0	1	0	0
	1	1	0	1	0	1	0
	1	0	1	1	0	0	1

```
Disk 1: 11110000…
Disk 2: 10101010…
Disk 3: 00111000…
Disk 4: 01000001…
Disk 5: 01100010…
Disk 6: 00011011…
Disk 7: 10001001…
```

```
Disk 1: 11110000…
Disk 2: 00001111…
Disk 3: 00111000…
Disk 4: 01000001…
Disk 5: 11000111…
Disk 6: 10111110…
Disk 7: 10001001…
```

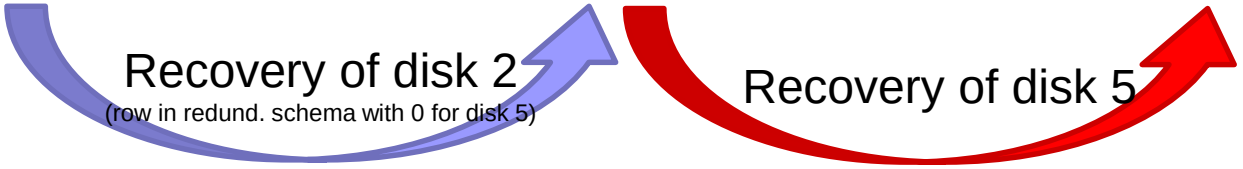
Writing to disk 2:
00001111…

Samoopravné kódy (pokr.)

- Hamming code
  - Disk failure

	Data				Parity		
Disk No:	1	2	3	4	5	6	7
	1	1	1	0	1	0	0
	1	1	0	1	0	1	0
	1	0	1	1	0	0	1

```
Disk 1: 11110000…        Disk 1: 11110000…
~~Disk 2:~~ ????????…    Disk 2: 00001111…
Disk 3: 00111000…        Disk 3: 00111000…
Disk 4: 01000001…        Disk 4: 01000001…
~~Disk 5:~~ ????????…    Disk 5: 11000111…
Disk 6: 10111110…        Disk 6: 10111110…
Disk 7: 10001001…        Disk 7: 10001001…
```



Samoopravné kódy (pokr.)

- Definition of Hamming Code
  - A *code* of length *n* is a set of *n*-bit vectors (*code words*).
  - *Hamming distance* is the count of different values in two n-bit vectors.
  - *Minimum distance of a code* is the smallest Hamming distance between any different code words.
  - *Hamming code is a code with min. dist. "3"*
    - Up to 2 bit flips can be detected (but not corrected).
    - 1 bit flip is detected and corrected.

Samoopravné kódy (pokr.)

- Generating Hamming Code (n,d); p=n-d
  - Number bits from 1, write them in cols in binary
  - Every column with one ‘X’ (one bit set) is parity
  - Row shows the sources for parity computation
  - Column shows which parity bits cover data bit.

Bit position		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Data/parity bits		p1	p2	d1	p3	d2	d3	d4	p4	d5	d6	d7	d8	d9	d10	d11	p5	d12
Parity bit coverage	p1 2^0	X		X		X		X		X		X		X		X		X
	p2 2^1		X	X			X	X			X	X			X	X		
	p3 2^2				X	X	X	X					X	X	X	X		
	p4 2^3								X	X	X	X	X	X	X	X		
	p5 2^4																X	X

Samoopravné kódy (pokr.)

Bit position		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Data/parity bits		p1	p2	d1	p3	d2	d3	d4	p4	d5	d6	d7	d8	d9	d10	d11	p5	d12
Parity bit coverage	p1 2^0	X		X		X		X		X		X		X		X		X
	p2 2^1		X	X			X	X			X	X			X	X		
	p3 2^2				X	X	X	X					X	X	X	X		
	p4 2^3								X	X	X	X	X	X	X	X		
	p5 2^4																X	X

- Store data bits 1010 in Hamming Code (7,4)
- To correct errors in data read from storage:
  - Check all parity bits.
  - Sum the positions of bad ones to get address of the wrong bit.
- Examples:
  - 1111010 → 1 bit was flipped, so it can be corrected.
  - 1011110
  - 1001110 → 2 bits were flipped, so it cannot distinguish 2-bit and 1-bit error.
  - 1110000 → 3 bits were flipped here

Samoopravné kódy (pokr.)

- Extended Hamming Code
  - Add an extra parity bit over all bits
    - To tell even or odd number of error

Bit position		0	1	2	3	4	5	6	7
Data/parity bits		pX	p1	p2	d1	p3	d2	d3	d4
Parity bit coverage	p1 2^0		X		X		X		X
	p2 2^1			X	X			X	X
	p3 2^2					X	X	X	X
	pX 0	X	X	X	X	X	X	X	X

- Store data bits 1010 in Extended Hamming Code (7,4)
- Detect/correct error if any:
  - 01111010
  - 01011110
  - 01001110 → 2 bits were flipped; but no clue which.
  - 01110000 → odd number **(>1)** bits were flipped; but no clue which.

Samoopravné kódy (pokr.)

- Reed-Solomon Code (n,d)
  - ECC adding *t* check bits to data bits
  - Can detect up to *t* bit errors
  - Can correct up to $\lfloor t/2 \rfloor$ errors
  - So, min. Hamming distance is $\mathrm{t} = n - d + 1$.

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% $\Rightarrow$ $P_{\text{výpadku za rok}}$=0,438%
    - $\Rightarrow$ **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 = 0,438%

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), copy available in supplemental study materials
  - 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 42ºC 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

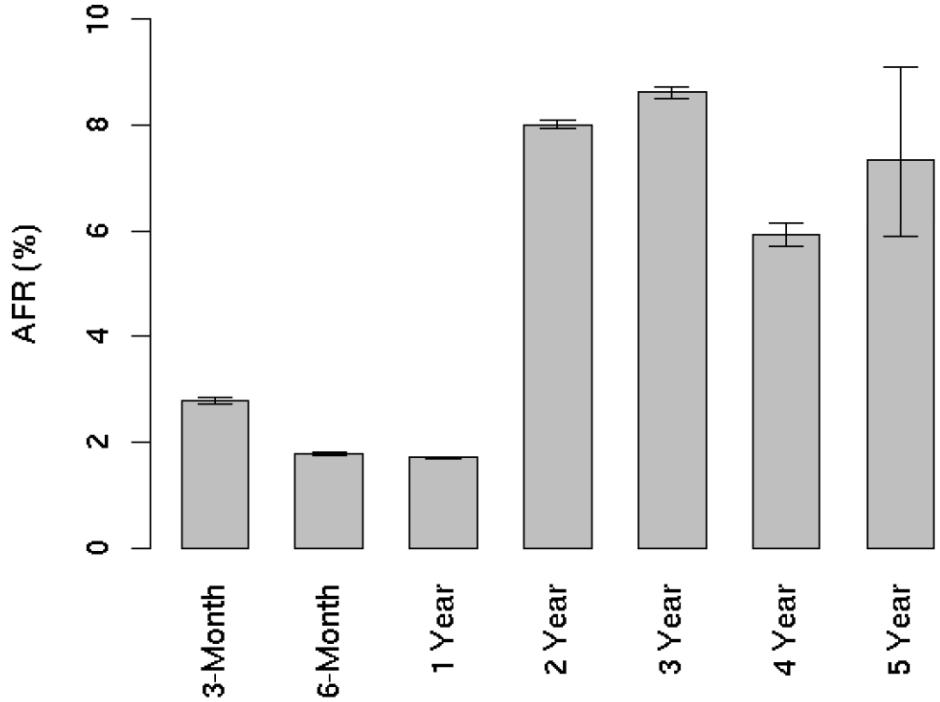
- Seagate Barracuda ES.2 Near-Line Serial ATA drive

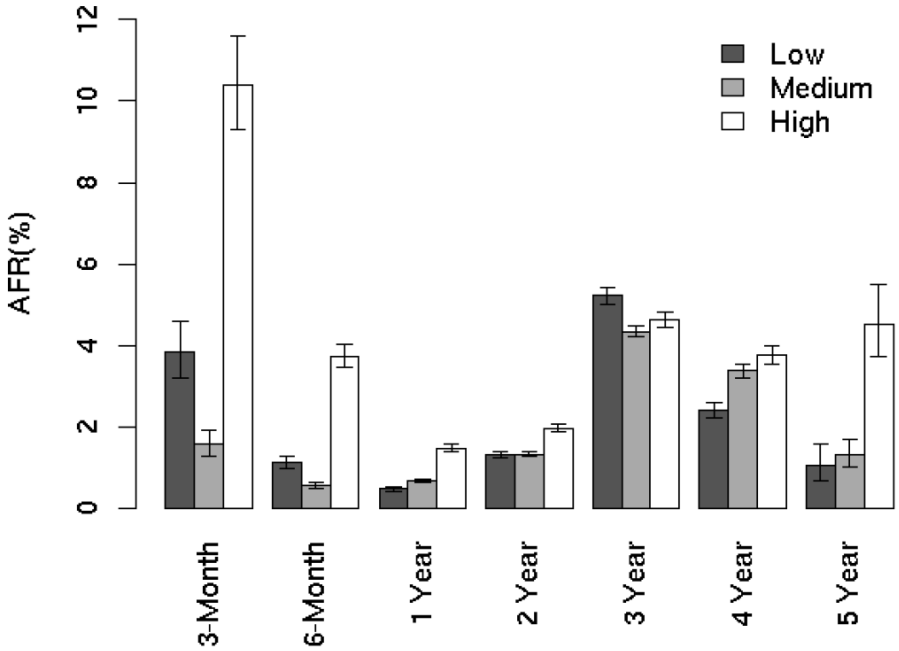
		MODEL:					
		Weibull		Warranty Data (OEM only)		Flatline Model	
Year	Cumulative power-on hours	Yearly failure rate	Cumulative failure rate	Yearly failure rate	Cumulative failure rate	Yearly failure rate	Cumulative failure rate
1	2,400	1.20%	1.20%	1.20%	1.20%	1.20%	1.20%
2	4,800	0.55%	1.75%	0.78%	1.98%	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 – stat. metoda pro modelování průběhu chybovosti
Note2: power-on-hours: 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

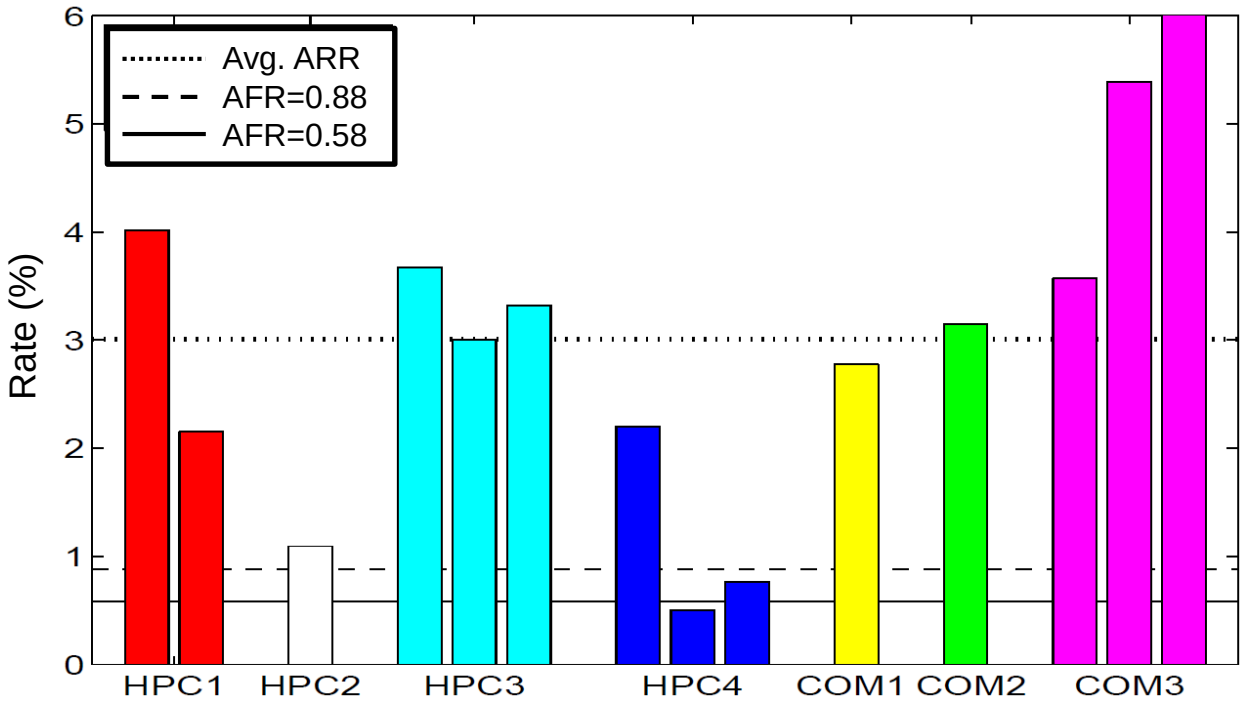




Výpadky – realita

- 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
144x SCSI disk, 15k rpm, 73GB
11000x SATA disk, 7200 rpm, 250GB

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
  - $P_{Failure\ During\ Repair} = P_{FDR} = (2*MTTR) / 1\ rok$
    - 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
    - $AFR_{1\ disk}$ = 0,44% (MTTF = 1,000,000 hrs. = 114 yrs.)
  - Systém se 100 disky ($MTTF_{100\ disků} = MTTF_{1\ disk} / 100$)
    - $AFR_{100\ disků}$ = 44% (MTTF = 10,000 hrs. = 1.14 yrs.)
      - *Průměrně* každý rok cca jeden z disků vypadne
    - Pravděpodobnost (právě *1 z n* vs. alespoň *1 z n* havaruje)
      - $P_{výpadek\ právě\ 1\ ze\ 100}$ = 28,43% **$P_{výpadek\ alespoň\ 1\ ze\ 100}$ = 35,66%**
      - $P_{výpadek\ právě\ 1\ z\ 10}$ = 4,23% **$P_{výpadek\ alespoň\ 1\ z\ 10}$ = 4,31%**
- $AFR_{n\ disků} = AFR_{1\ disk} * n$

Příklad spolehlivosti RAID1

- 2 zrcadlené 500GB disky
  - každý AFR=3%
- Výměna vadného a obnova pole do 3 hodin
  - MTTR = 3 hodiny (při 100MB/s je kopírování za 1,5h)
- Pravděpodobnost ztráty dat:
  - $P_{\text{výpadku 1 disku}} = AFR = 0.03$
  - $P_{\text{výpadku 1 ze 2}} = 0.06$
  - $P_{FDR} = 2 * MTTR / 1\ \text{rok} = 2*3 / 8760 = 0{,}000\ 685$
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 2}} * P_{FDR} * P_{\text{výpadku 1 disku}}$
    $= 0{,}000\ 001\ 233$
  - **MTTDL** $= 0.5 / P_{\text{ztráty dat}} = 405\ 515$ let

Příklad spolehlivosti 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
    - Tj. $AFR_{\text{pole}} = 6\%$

Příklad spolehlivosti RAID4

- AFR disku 3% ($P_{\text{výpadku 1 disku}}$)
- RAID4 – opravuje výpadek 1 disku
  - 4 disky (3+1)
  - MTTR = 3 hodiny
    - $P_{FDR} = 2*3 / 8760 = 0{,}000\ 685$
  - $P_{\text{ztráty dat}} = P_{\text{výpadku 1 ze 4}} * P_{FDR} * P_{\text{výpadku 1 ze 3}}$
  - $P_{\text{ztráty dat}} = 4*0{,}03 * 2/2920 * 3*0{,}03$
    $= 0{,}000\ 007\ 397$
    - což je AFR tohoto pole
  - $\text{MTTDL} = 0.5 / P_{\text{ztráty dat}} = 67\ 592$ let

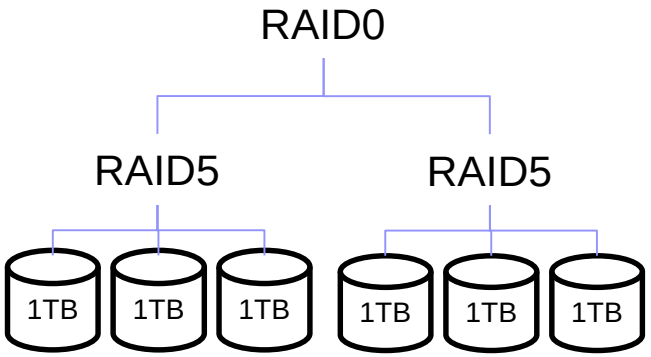
Spolehlivost pole

- n disků
  - celkem disků v poli (včetně paritních)
- 1 paritní disk
  - zajišťují redundanci dat

- $AFR_{pole} = n*AFR_{1\ disku} * P_{FDR} * (n-1)*AFR_{1\ disku}$
- $MTTDL = 0.5 / AFR_{pole}$

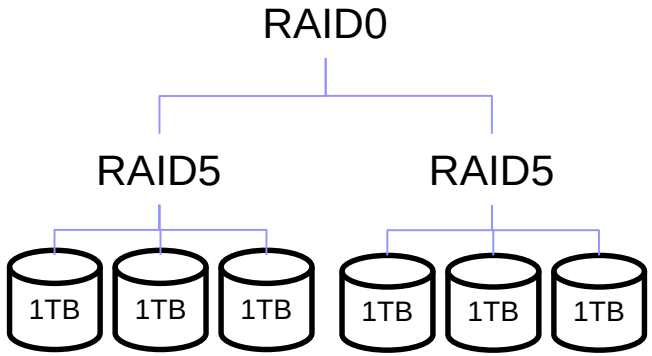
Příklad spolehlivosti – kombinace RAID

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

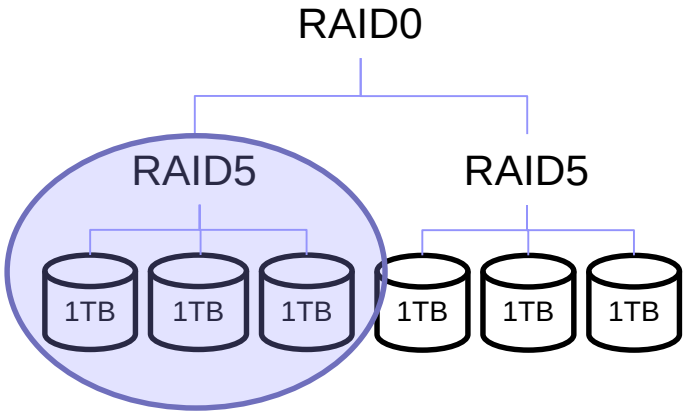


Příklad spolehlivosti – kombinace RAID

- RAID5+0
  - 1 disk má AFR_{disk}

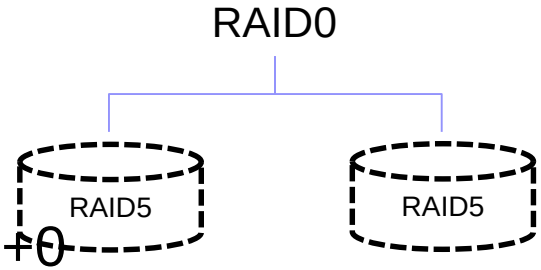


1) Urči AFR_{RAID5}



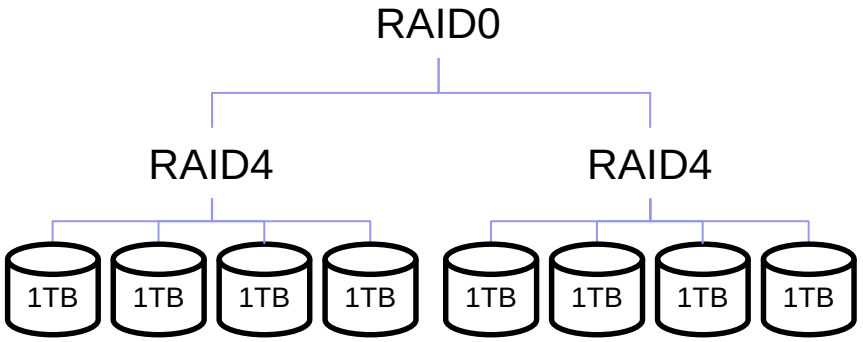
2) Urči $AFR_{RAID5+0} = 2 * AFR_{RAID5}$

3) $MTTDL_{RAID5+0} = 0.5 / AFR_{RAID5+0}$



Příklad spolehlivosti – kombinace RAID

- RAID4+0 z 8 disků v konfiguraci níže
  - 1 disk AFR=3%, MTTR = 3 hodiny



  - Vždy ze 4 disků vyrobíme 1x RAID4
    - $AFR_{RAID4} = 4*AFR * P_{FDR} * 3*AFR = \ldots = 7.4*10^{-6}$
  - 2x RAID4 spojíme pomocí RAID0
    - $AFR_{RAID4+0} = 2 * AFR_{RAID4} = 1.48*10^{-5}$
    - $MTTDL = 0.5 / AFR_{RAID4+0}$ = 33 796 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 an unused block to list of used space
  2. Write data block
  3. Write file metadata referencing that data block
- Modern FS uses journaling
  - Start transaction in journal
  - Store info about steps 1.-3. to journal
  - Do steps 1.-3.
  - End transaction in journal

File System & Memory Tuning

- FS block size ≤ DB block size
  - ZFS has 128KB by default!
- DB journal (WAL in PostgreSQL)
  - ext2; ext3/4 with data=writeback (journal off)
- 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 řešení 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 replacing a failed SSD with a brand new one, parity is moved primarily to the most worn-out drive.

Recommended Reading

- Dual parity
  - https://www.kernel.org/pub/linux/kernel/people/hpa/raid6.pdf
- Software RAID in SUSE
  - https://www.suse.com/documentation/sles10/stor_admin/data/raidevms.html
  - Sections:
    - **Managing Software RAIDs with EVMS**
    - **Managing Software RAIDs 6 and 10 with mdadm**
- SSD on Wikipedia
  - https://en.wikipedia.org/wiki/Solid-state_drive
- Živě.cz: Ze světa: kolik reálně zapsaných dat vydrží moderní SSD?
  - http://m.zive.cz/ze-sveta-kolik-realne-zapsanych-dat-vydrzi-moderni-ssd/a-177557/?textart=1
- Chunk size and performance
  - https://raid.wiki.kernel.org/index.php/Performance