Disková pole (RAID)

Architektury RAID

- Co architektury RAID řeší: 1) zvýšení požadavků na rychlost získávání dat (reakce na zvyšující se rychlost procesoru), 2) zvýšení požadavků na spolehlivost dat.
- Pozice diskové paměti v klasickém personálním počítači – vyhovuje pro aplikace s jedním uživatelem.
- Splnění požadavků 1) a 2) – požadavky např. v serverech.
- Řešení: data jsou distribuována na více disků, datová operace je realizována paralelně.
- Co to nabízí: kromě distribuování dat na více disků možnost zvýšení spolehlivosti – využití redundance (zdvojování disků nebo vygenerování a záznam informace, která umožní opravu).
- Paralelní přístup a detekci poruchy diskové paměti a příp. opravu.
- Fyzická realizace diskových polí: disky SCSI.

Architektury RAID

- Význam zkratky: **R**edundant **A**rray of **I**ndependent **D**isks.
- Jiné vysvětlení: **R**edundant **A**rray of **I**nexpensive **D**isks, první termín je výstižnější.
- Původně 7 úrovní architektur RAID (tzn. RAID 0 až RAID 6) pokrývajících různým způsobem požadavky 1) a 2), u novějších verzí RAID kombinace těchto architektur.
- Neexistují hierarchické úrovně.
- Sada fyzických disků (každý fyzický disk má svůj řadič), operační systém je vidí jako jeden disk.
- Data jsou distribuována do všech fyzických disků.
- Možnost využití dodatečné kapacity pro uložení informace o paritě.
- První zásady o RAID definovány v r. 1988.
- Disky RAID a jejich vlastnosti jsou často konfrontovány s architekturami SLED – Single Large Expensive Disk (samostatný velký nákladný disk).

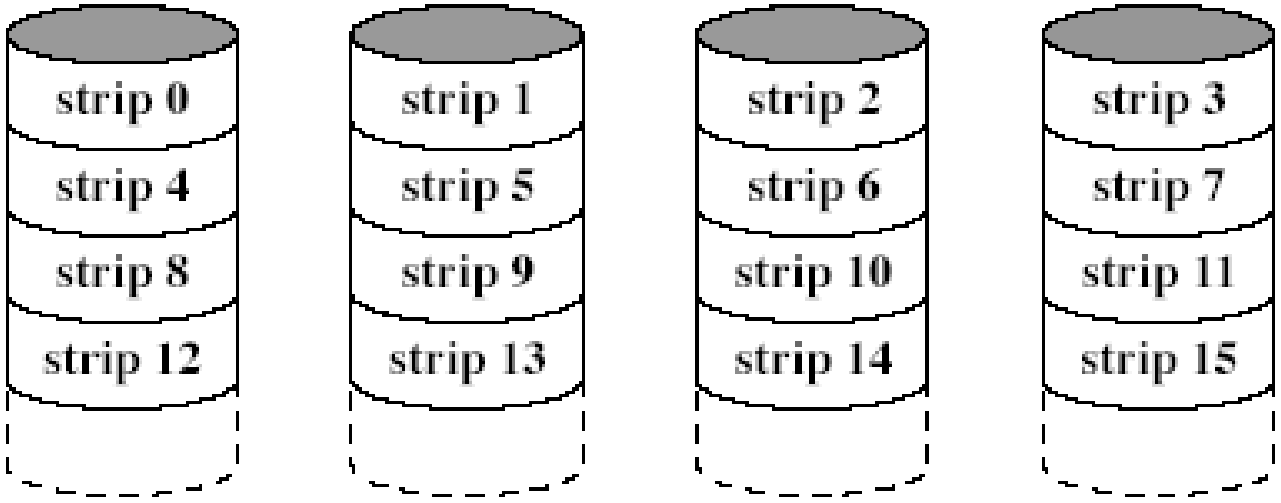
Architektury RAID – fyzická realizace

- V samostatném rámu mimo počítač nainstalována sada disků.
- V sestavě je dále server, řadič disků (anglický termín RAID controller) a SCSI řadič (fyzické disky jsou připojeny přes SCSI řadič).
- Počet fyzických disků – podle šířky sběrnice SCSI (adresace 1 z n).
- Operačnímu systému se RAID jeví jako SLED (jeden disk), má však větší výkon a vyšší spolehlivost.

RAID 0

- Žádná redundance kvůli zabezpečení (všechny disky jsou využity pro uložení dat), splnění pouze požadavku 1 – viz slide 2.
- Data jsou rozdělena na všechny disky.
- Zvýšení rychlosti - vysvětlení:
  - Požadovaná data jsou rozdělena na více disků.
  - Operace „vystavení“ (seek) je prováděna paralelně (současně na všech discích) – dostaneme se současně k více datům jednou operací vystavení prováděnou paralelně.
  - **Důležitý princip: všechny operace jsou prováděny paralelně**.

RAID 0

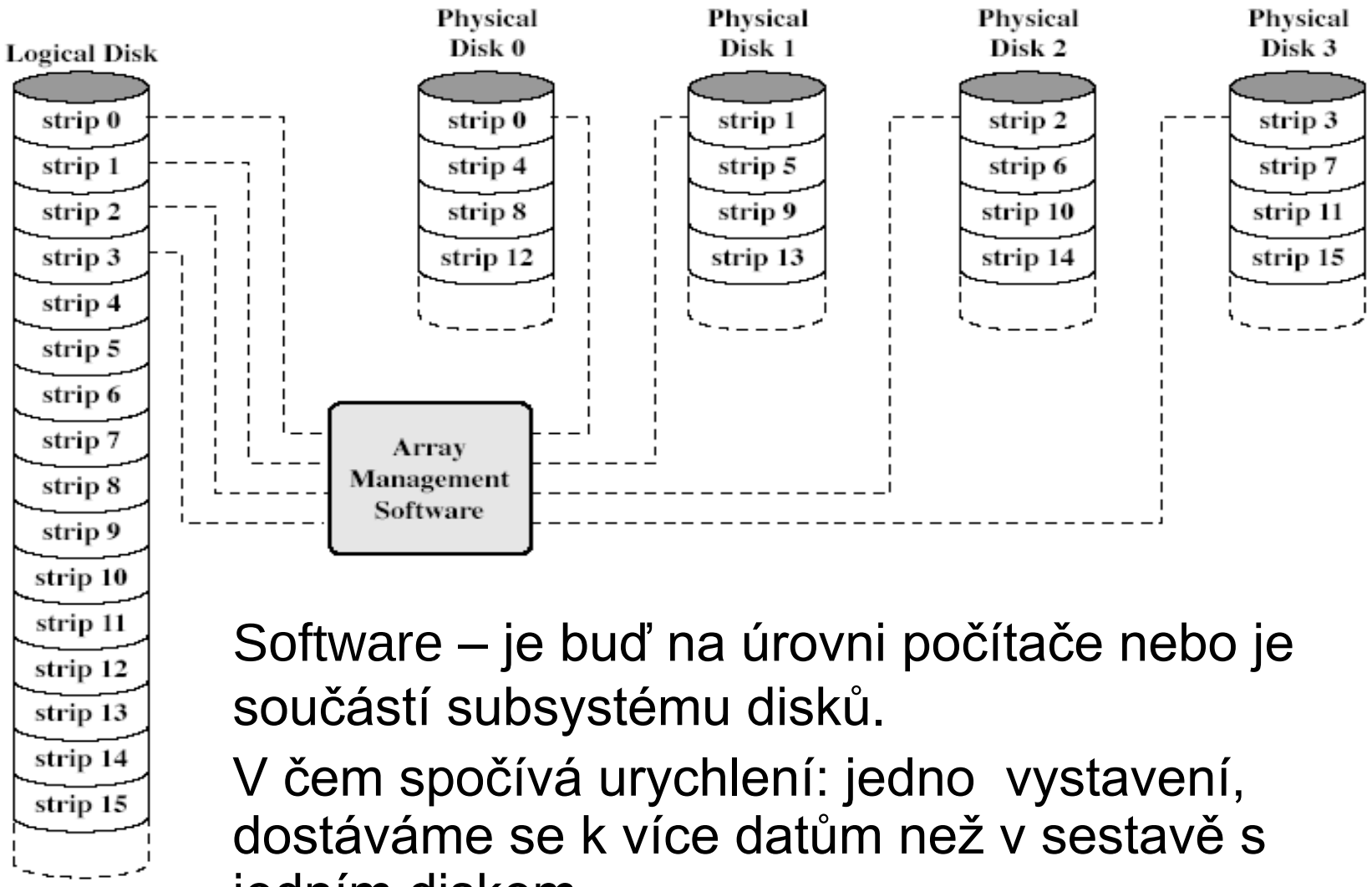


Důležité:
Splnění V/V požadavku – data jsou na různých discích.
V datech není žádná redundance s cílem dosáhnout vyšší spolehlivosti.

RAID 0

- Data jsou rozdělena na více disků (striped).
- Všechna uživatelská i systémová data jsou z hlediska OS uložena jakoby na jednom systémovém disku.
- Každý disk je rozdělen na **strips.**
- Příklad: pole *n* disků, prvky strip sestávají z *k* sektorů.
  první **strip** na všech discích tvoří první **stripe**.
  sektory jsou číslovány takto:
  strip 0: *0* až *k-1,* strip 1: *k* až *2k-1*
  Pokud *k=2*, pak strip je tvořen dvěma sektory.
- Výhoda: současně je možné zpracovávat *n* prvků typu strip.
- Data se zapisují na jednotlivé strip, tento proces se nazývá **striping.**
- Význam výrazů **strip** [strip] a **stripe** [strajp]: pruh, proužek.

RAID 0



Software – je buď na úrovni počítače nebo je součástí subsystému disků.

V čem spočívá urychlení: jedno vystavení, dostáváme se k více datům než v sestavě s jedním diskem.

RAID 0

- RAID 0 není redundantní.
- Ztráta jednoho disku znamená ztrátu celého pole.
- Důvodem použití je výkon, tedy zvýšení přenosové rychlosti nebo propustnosti dat tam, kde na spolehlivém uchování dat nezáleží tak, jako na rychlosti.

RAID 0

- Princip součinnosti operačního systému s disky RAID:

  operační systém vydá příkaz na čtení dat (předpokládá se, že data za sebou následují na jednotlivých strip),

  řadič RAID „rozdělí“ tento příkaz na 4 samostatné příkazy, ty předá řadičům disků,

  disky tyto příkazy provedou paralelně.

RAID 0 – podpora vysoké rychlosti přenosu

- Požadavek na vysokou rychlost diskové paměti (souvislost s technickou úrovní).
- Kvalitní spoj z diskové paměti (tvoří jeden celek s řadičem) do počítače.
- Rychlý řadič diskové paměti – každý disk v diskovém poli má svůj řadič, který autonomně řídí periferní operace – disk SCSI je takto vybaven.
- Rychlá systémová sběrnice.
- Rychlý procesor.
- **Toto vše platí obecně pro architektury RAID.**

RAID 0 – charakter V/V požadavků

- V/V požadavky – požadavky na data.
- Ideální organizace dat – data, která spolu souvisejí (např. soubor) by měla být uložena v sousedních strip (v jednom stripe) – pak přenos paralelně.
- Výsledek – výrazně efektivnější operace související s diskem – hlavně rychlost přenosu (možnost přenosu jednoho souboru paralelně – jeho jednotlivé části).

RAID 0 – využití v systému zaměřeném na řešení transakcí

- Podpora vysokého objemu V/V požadavků.
- Řešení těchto požadavků – diskové pole umožňuje tyto požadavky vyváženě rozložit na více fyzických disků.
- Dvě situace – **1)** větší počet nezávislých transakcí nebo **2)** transakce, které je možné rozdělit na jistý počet asynchronních činností (termín „počet" souvisí s počtem strip).
- Souvislost s velikostí strip – měl by být takový, aby řešení transakce nevyžadovalo více přístupů na disk, tzn. vystavení (eliminovat pomalou operaci vystavení).

RAID 0

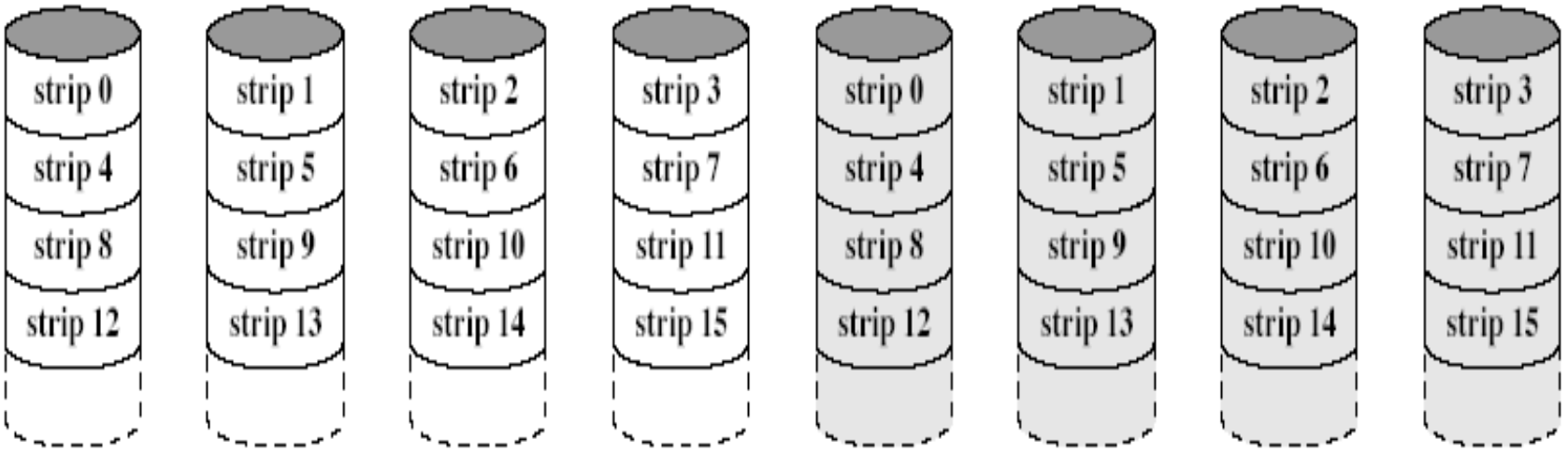
- Architektura RAID 0 je vhodná pro situace, kdy se vyžadují přenosy velkých objemů dat.
- Problém, pokud je požadavek na data větší než počet disků vynásobený velikostí strip, pak musí být násobné požadavky na data (nestačí jedno vystavení) – toto řeší řadič RAID.
- Pokud pouze požadavky na data velikosti jeden sektor – nevyužité možnosti RAID (není paralelismus diskových operací – není dosažen maximální výkon).
- Spolehlivost takových instalací je menší než spolehlivost sestav SLED.

  Vysvětlení:

  4 disky RAID s parametrem střední doba mezi poruchami (MTBF – Mean Time Between Failures) 20 000 hodin, pak každých 5 000 hodin má jeden z disků poruchu – data z tohoto disku jsou ztracena) – z hlediska spolehlivosti jde o sériový systém (porucha jediného disku, celý systém je nefunkční).

  Větší spolehlivost má jeden disk SLED (zůstane zachováno 20 000 hodin).

RAID 1



Architektura RAID 1:
Má v sobě rysy pravé architektury RAID, **disky jsou zdvojeny (redundance)**.

RAID 1

- Zrcadlené disky – snaha o zvýšení spolehlivosti.
- Každý strip je uložen na dva fyzické disky.
- Poznámka: v dalších typech řešeno způsobem, **který nepředstavuje řešení typu „hrubá síla“**, k datům se počítá parita (nevýhoda – při každé změně dat se musí parita znova počítat – režie).
- RAID 1 - výrazná redundance.
- Jednoduché zotavení při chybě.
- Nákladné.

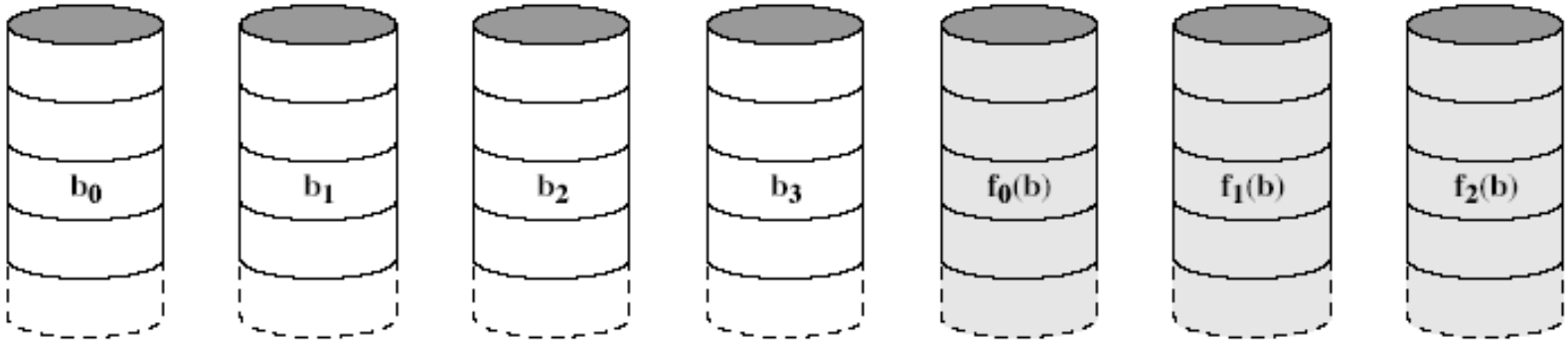
RAID 1

- **Čtení dat** – pouze z jednoho disku (z toho, kde to bude rychlejší – doba vystavení + rotační zpoždění).
- **Zápis dat** – provádí se paralelně na oba disky → výkon při zápise bude ovlivněn tím diskem, na němž to bude trvat déle (delší doba vystavení + rotační zpoždění).
- Zotavení po poruše – data se čtou z disku, který je funkční.
- Nevýhoda architektury RAID 1 – vysoké náklady.

RAID 1 - využití

- Z hlediska výkonu tam, kde podstatnou část transakcí tvoří transakce „čtení“ (např. systémové disky – systémové programové vybavení a data - zálohování).
- Transakčně orientovaná aplikace – výhoda, pokud výrazným počtem transakcí jsou transakce typu čtení, horší stav v případě transakcí typu zápis.

RAID 2

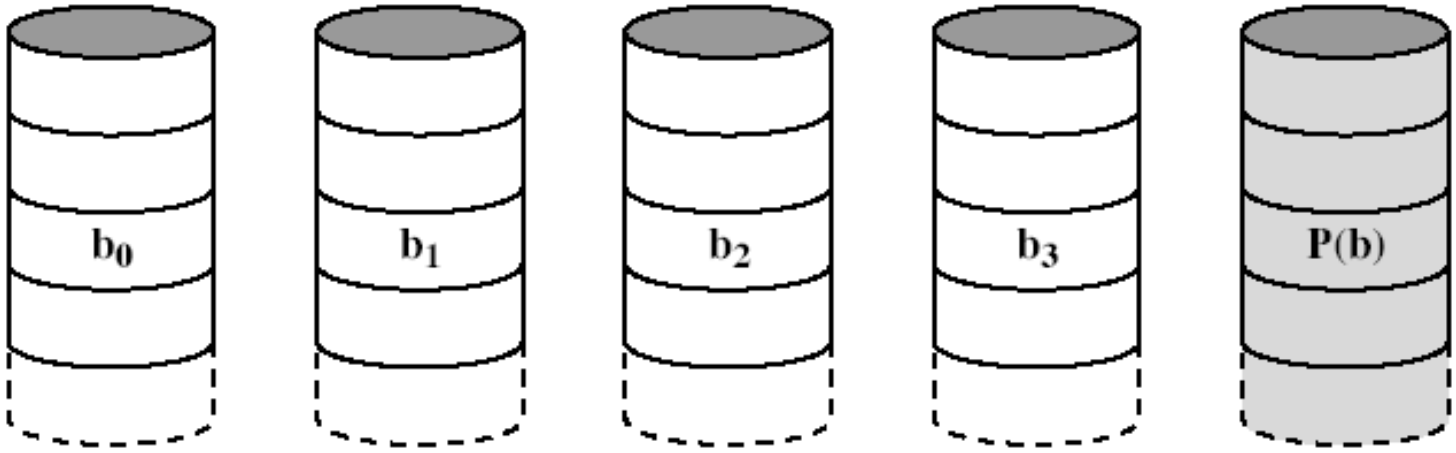


Čtyři disky – uložení dat, tři disky – informace pro opravu chyb.

RAID 2

- Disky jsou synchronizovány, takže na všech discích jsou hlavy ve stejné pozici – z hlediska otáčení disku a vystavení.
- Paralelní přístup – na vyřízení každého V/V požadavku se podílejí všechny paralelně pracující disky.
- Informace potřebná pro opravu chyb se počítá z odpovídajících bitů na discích.
- Na paritní disky se ukládají bity vygenerované jako Hammingův kód z odpovídajících datových bitů. Hammingův kód (7,4): 4 datové bity jsou zakódovány do 7 bitů – oprava 1 vadného bitu a detekce 2 vadných bitů.
- Velká redundance – nevýhodné (RAID 2 se příliš nevyužívaly).

RAID 3



Čtyři disky – uložení dat, jeden disk – parita.

RAID 3

- Organizováno podobně jako RAID 2.
- Pouze jeden redundantní disk bez ohledu na to, jak je rozsáhlé diskové pole.
- Jeden paritní bit pro každou sadu odpovídajících bitů.
- Data na disku, který má poruchu, mohou být rekonstruována z existujících dat a parity.
- Vysoké rychlosti přenosu.

RAID 3 - využití redundance

- Pokud má disk poruchu, tak se přečte paritní bit a data se zrekonstruují ze zbývajících bitů, pro rekonstrukci se použije i bit paritní.
- Po výměně vadného disku je možné data ze zbývajících bitů zrekonstruovat.

- **Rekonstrukce dat:**
  Uvažujme diskové pole sestávající z pěti disků.
  X0 – X3 obsahují data
  X4 – paritní bit

- Schéma tvorby paritního bitu:
  X4(i) = X3(i) xor X2(i) xor X1(i) xor X0(i)

RAID 3 – rekonstrukce dat (jeden z disků má poruchu)

- Předpokládejme, že disk X1 přestal fungovat, tzn. data z něj nejsou k dispozici.
- Výchozí vztah:

  X4(i) = X3(i) xor X2(i) xor X1(i) xor X0(i)
- K oběma stranám rovnice přidáme X4(i) xor X1(i)
- Pak dostaneme:

  X4(i) xor X4(i) xor X1(i) = X3(i) xor X2(i) xor X1(i) xor X0(i) xor X4(i) xor X1(i),

  X4(i) xor X4(i) je vždy 0, stejně tak X1(i) xor X1(i)

  X1(i) = X4(i) xor X3(i) xor X2(i) xor X0(i)

  Hodnota bitu z disku, který má poruchu, se vypočte ze zbývajících bitů.

  Závěr: bit z vadného disku se počítá ze zbývajících bitů, tento princip se používá pro RAID3 až RAID6 (tzv. redukovaný režim).

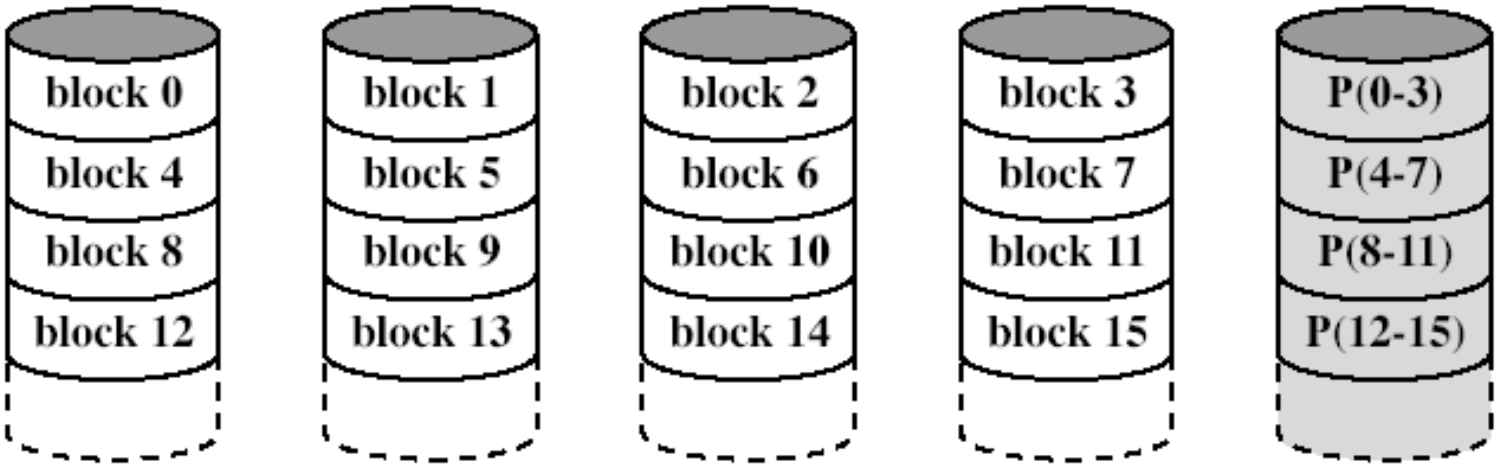
RAID 3

- Zápis v situaci, kdy je disk vadný:

  Ze všech bitů se vytvoří paritní bit, vše se zaznamená (bez zápisu do vadného bitu – disk je vadný), po výměně disku se patřičný bit zrekonstruuje.
- Výkon:

  Je vysoký, protože se čte z více disků současně.

RAID 4



Parita typu „block-level“.

Data se ukládají do větších „proužků“ s velikostí na úrovní bloků.

RAID 4

- Každý disk je nezávislý.
- Vhodný pro aplikace s vysokým objemem V/V požadavků.
- Velké bloky (stripes).
- Parita se počítá bit po bitu přes celé bloky (stripes) na každém disku.
- Parity se uloží na paritní disk.

RAID 4 - využití redundance

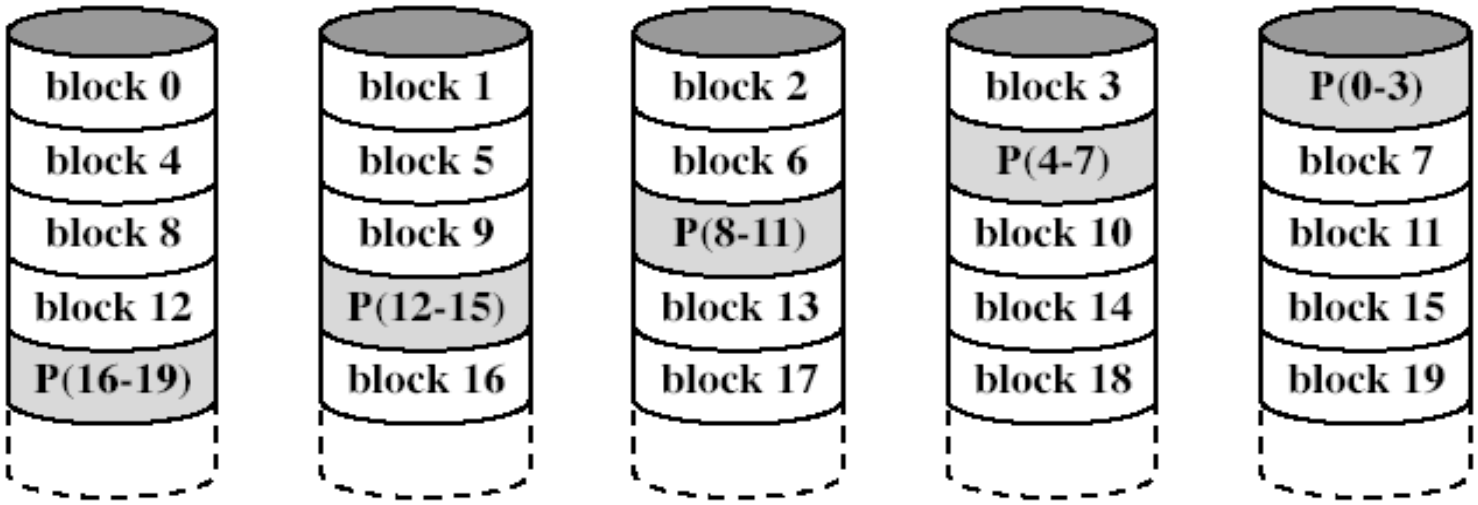
- Situace: potřebujeme zapisovat pouze na jeden disk – jak vypočteme paritu?
- Pole sestává z 5 disků: X0 – X3 – data, X4 – parita.
- X4(i) = X3(i) xor X2(i) xor X1(i) xor X0(i)
- Nová parita (změna v bitu X1(i) :

  X4´(i) = X3(i) xor X2(i) xor X1´(i) xor X0(i) = X3(i) xor X2(i) xor X1´(i) xor X0(i) xor X1(i) xor X1(i)

  = X3(i) xor X2(i) xor X1(i) xor X0(i) xor X1(i) xor X1´(i)

  = X4(i) xor X1(i) xor X1´(i)
- Má-li se vypočítat nová parita, tak se pro výpočet použije stará parita, původní hodnota bitu a nová hodnota bitu.
- Zápis: zapisuje se jak datový bit, tak i parita.

RAID 5

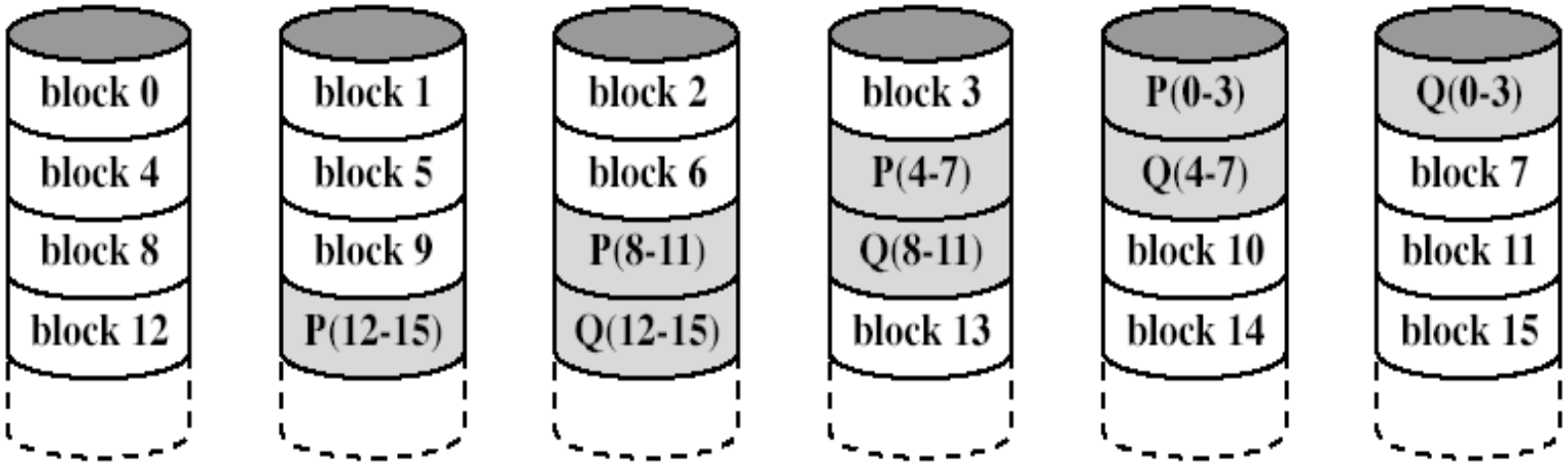


Parita typu „block-level distributed“

RAID 5

- Podobné s RAID 4.
- Parita je uložena na všech discích.
- Všechny dosavadní mechanismy byly schopny napravovat problém, pokud nastal na jednom disku.
- RAID 5 – zvýšení spolehlivosti.
- Stav, kdy poruchu mělo více disků – doposud neřešitelný.

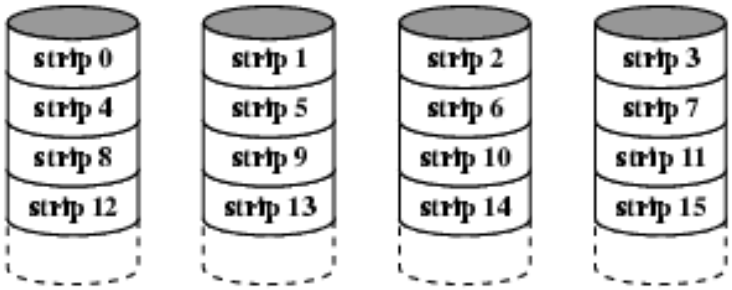
RAID 6


block 0
block 4
block 8
block 12
block 1
block 5
block 9
P(12-15)
block 2
block 6
P(8-11)
Q(12-15)
block 3
P(4-7)
Q(8-11)
block 13
P(0-3)
Q(4-7)
block 10
block 14
Q(0-3)
block 7
block 11
block 15

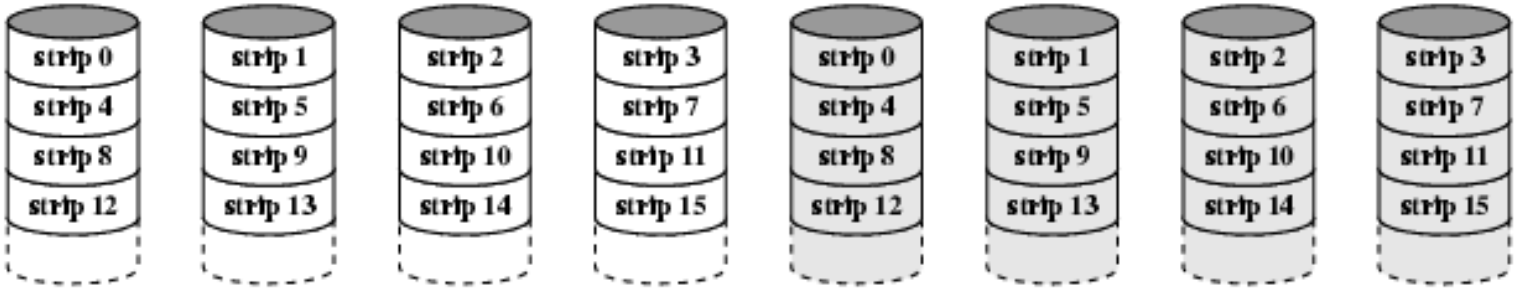
RAID 6

- Počítají se dvě parity.
- Parita se ukládá do samostatných bloků na různých discích.
- Je potřeba další dva disky navíc.
- Porucha dvou disků – je možná náprava dat.
- Porucha tří disků – neřešitelné.

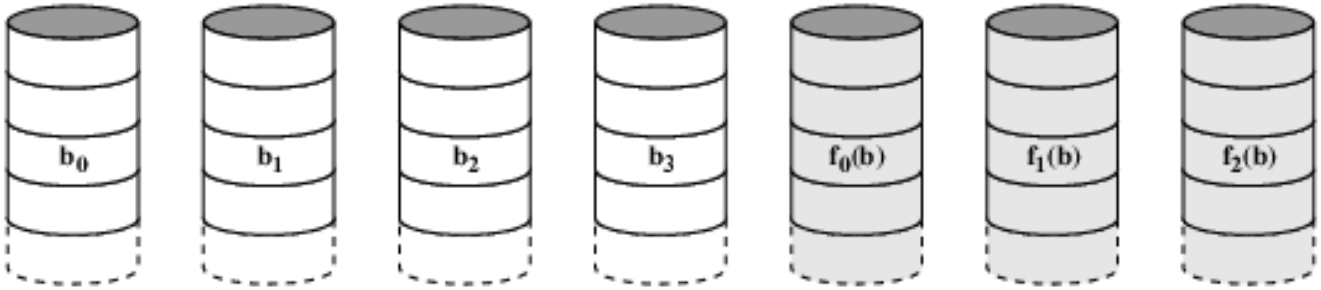
RAID 0, 1, 2



(a) RAID 0 (non-redundant)

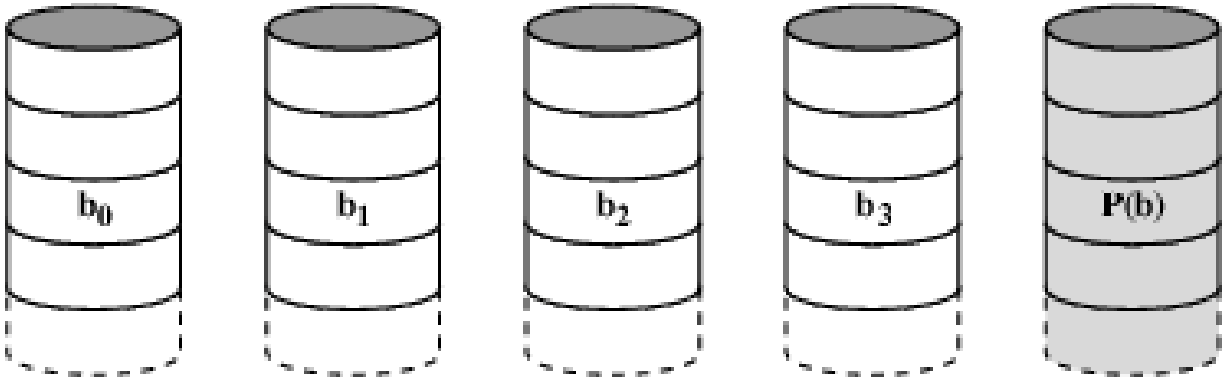


(b) RAID 1 (mirrored)

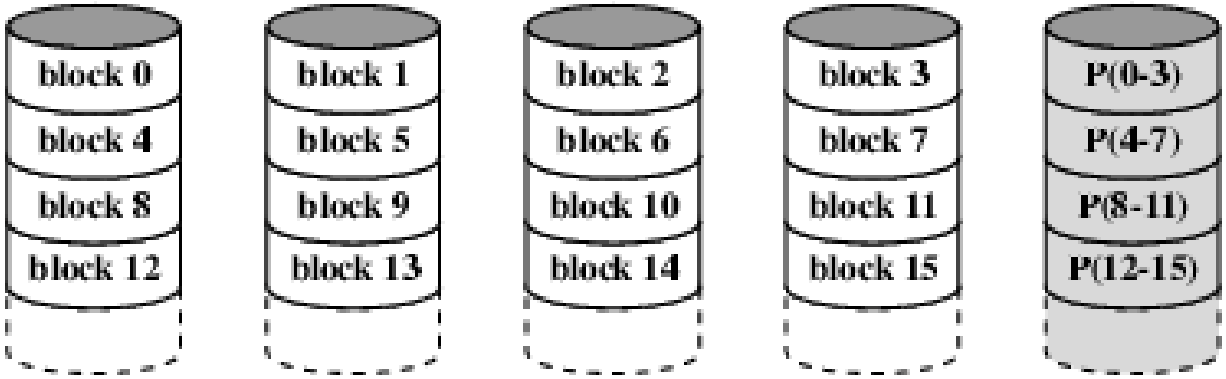


(c) RAID 2 (redundancy through Hamming code)

RAID 3 & 4

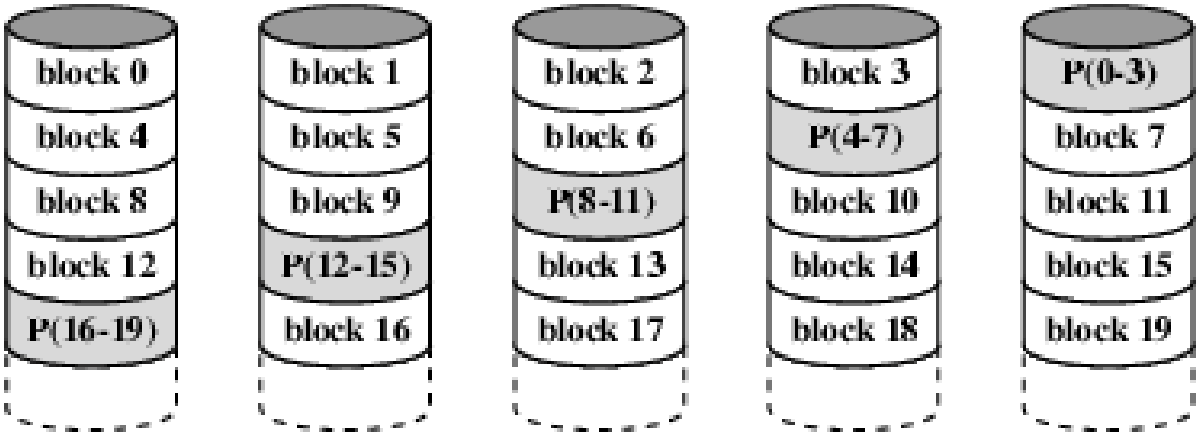


(d) RAID 3 (bit-interleaved parity)

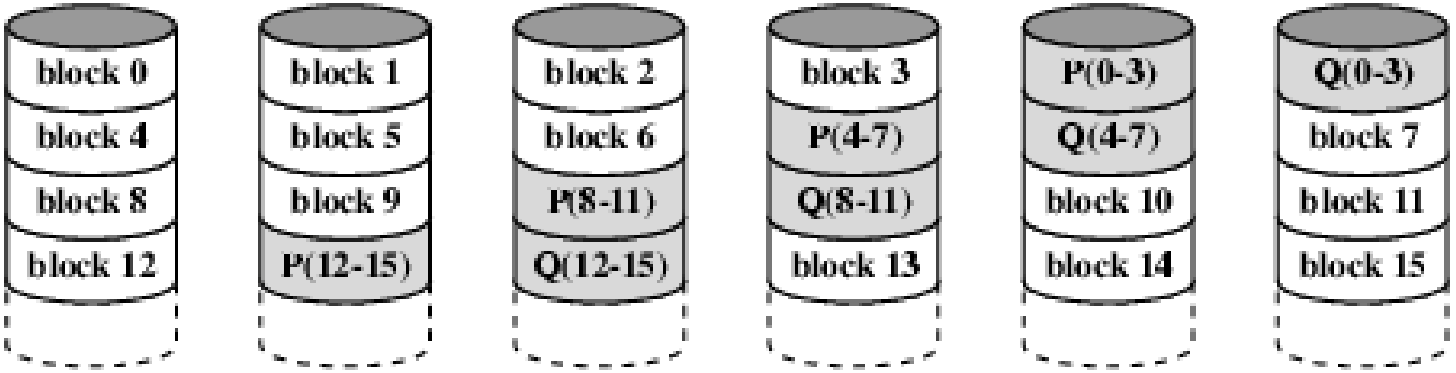


(e) RAID 4 (block-level parity)

RAID 5 & 6



(f) RAID 5 (block-level distributed parity)



(g) RAID 6 (dual redundancy)

RAID 7

- Každý disk je řízen nezávisle, má k dispozici vlastní datové cesty včetně cache – disk má vlastní technické prostředky (dedikované).
- Přenosy do centrální cache jsou nezávislé.
- Na systém lze napojit 48 disků a 12 hostitelských počítačů.
- Systém pracuje v reálném čase a má jak SCSI sběrnici, tak i interní vysokovýkonnou sběrnici (320MB/s) a sběrnici pro řízení.

RAID 7

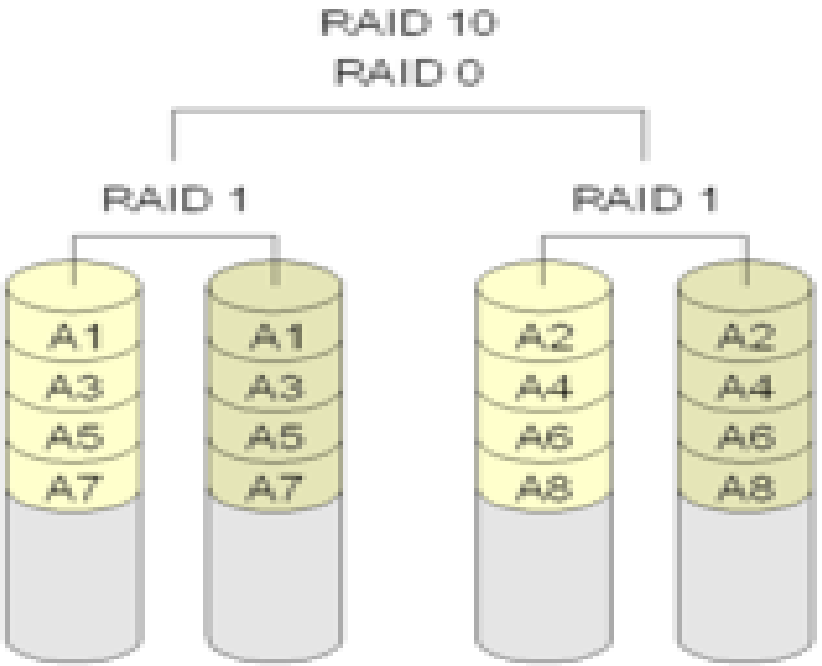
- Řadič – je zajištěna nezávislost jednotlivých cest a disků – příspěvek ke spolehlivosti.
- Soustava pracuje se 3 paritami.
- Umožňuje ochranu dat i při poruše více disků.
- Umí detekovat i sudý počet chyb, který se u předešlých systémů tváří jako v pořádku.
- Má zdvojeny napájecí zdroje a 3 úrovňovou cache.

RAID 7

- Diskové pole RAID 7 vlastně není typickým diskovým polem v porovnání s předchozími modely.
- Tento typ byl popsán a zároveň i patentován firmou Storage Computers.
- Díky patentovému chránění i technické náročnosti (ceně) se tento typ používá jen velmi málo.
- Celkový výkon při zápisu je o 25 až 90% vyšší než při zápisu na jeden disk, a o 50 až 500% vyšší v porovnání s ostatními úrovněmi RAID.

RAID 10

- Kombinace RAID 0 (stripe) a RAID 1 (zrcadlo).



RAID 10

- Jedná se vlastně o zrcadlený stripe. Minimální počet disků 4, režie 100% diskové kapacity navíc.
- Poskytuje nejvyšší výkon v bezpečných typech polí, podstatně rychlejší než RAID 5 zejména při zápisu.
- Další výhodou je odolnost proti ztrátě až 50% disků (naproti tomu RAID 5 odolává ztrátě pouze jednoho disku).

Architektury RAID - závěr

- Cíl architektur RAID – **zvýšení rychlosti přístupu k datům** a **zvýšení spolehlivosti** se u různých architektur RAID řeší s různým výsledkem.
- První typy architektur – především zvýšení rychlosti přístupu k datům.
- U dalších typů architektur pak naplňování požadavku na zvýšení spolehlivosti.