Operační systémy

Datová úložiště

Jan Trdlička

České vysoké učení technické v Praze, Fakulta informačních technologií
Katedra počítačových systémů

https://courses.fit.cvut.cz/BI-OSY

Obsah přednášky

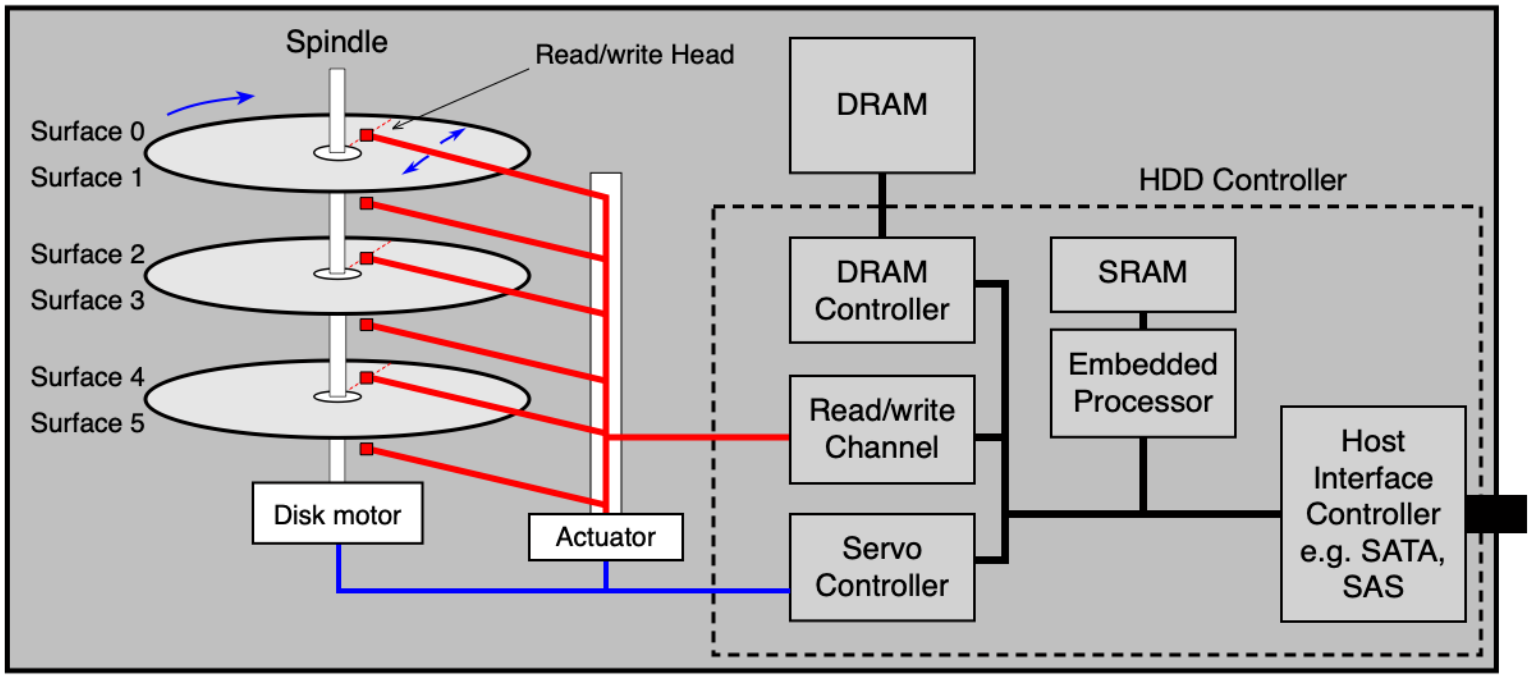
1. Datové uložiště
2. HDD (Hard Disk Drive)
3. SSD (Solid State Drive)
4. RAID (Redundant Array of Independent Disks)
   - RAID 0 (concatenation, striping)
   - RAID 1 (mirroring)
   - RAID 1 + 0 (mirroring + striping)
   - RAID 2, 3, 4
   - RAID 5
   - RAID 6
5. Typy připojení datového úložiště k výpočetnímu systému
   - DAS (Direct-Attached Storage)
   - NAS (Network-Attached Storage)
   - SAN (Storage Area Network)

Datové uložiště

- Hardware, který slouží k dlouhodobému uložení informací.
- Někdy je také označován jako "sekundární paměť" (secondary storage) nebo externí paměť (external memory).
- Úložný prostor, který datové uložiště nabízí,
  - se skládá ze sektorů (nejměnší adresovatelná jednotka, 512B/4KB).
  - je obvykle rozdělen na menší oblasti (např. diskové oblasti v případě disků nebo "volumes" v případě RAID), které jsou spravovány prostřednictvím
    - ★ OS: oblast obsahuje systém souborů (FS), data jsou zde uložena ve formě souborů, OS poskutuje aplikacím rozhraní na úrovni FS.
    - ★ aplikace: data jsou zde uložena v proprietárním formátu definovým konkrétní aplikací (např. databáze), OS poskutuje pouze rozhraní na úrovni sektorů.
- V současné době mezi nejběžnější typy datových úložišť patří
  - HDD (Hard Disk Drive) neboli pevný disk,
  - SSD (Solid State Drive),
  - RAID (Redundant Array of Independent Disks).
- Datové úložiště může být připojeno k systému několika způsoby
  - DAS (Direct Attached Storage),
  - NAS (Network Arrea Storage),
  - SAN (Storage Arrea Network).

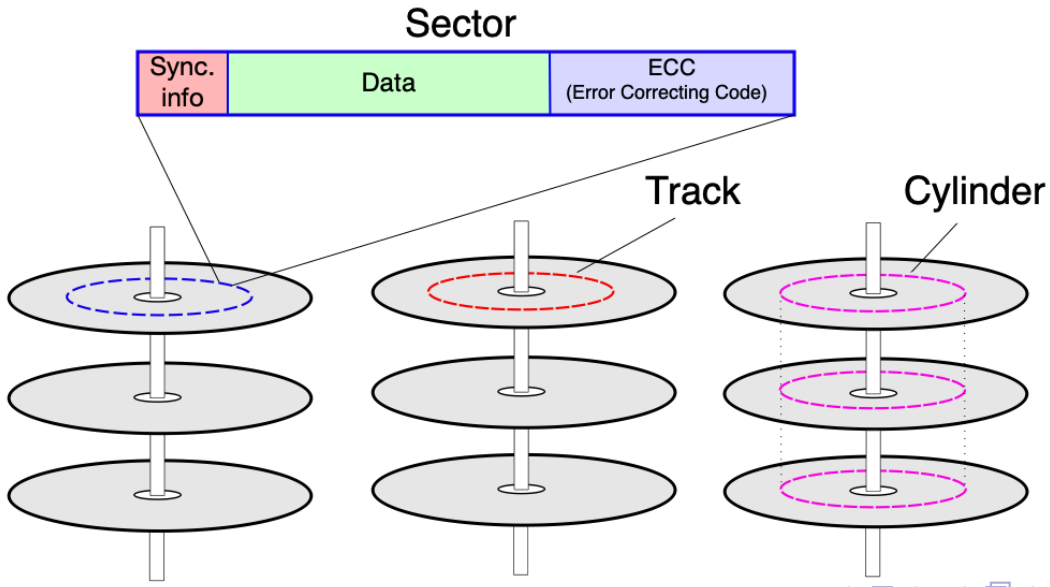
HDD (Hard Disk Drive): Architektura

- **Mechanické části**
  - Několik ploten, které mají dva povrchy a otáčí stejnou rychlostí.
  - Pohyblivé hlavičky, které umožňují čtení/zápis z příslušného povrchu a nacházejí se všechny vždy ve stejné vzdálenosti od středu povrchu.
- **Elektrické části**
  - Řadič disku: obsahuje procesor, paměť s firmwarem a příslušné řadiče.
  - Vyrovnávací paměť (DRAM): obsahuje čtená/zapisovaná data a může mít velikost několik stovek MB.



HDD: Geometrie

- **Sektor (Sector)**
  - Nejmenší adresovatelná jednotka na disku, která obsahuje
    - ★ synchronizační informace sloužící pro řadič disku,
    - ★ data ("nově" 4096 B, dříve 512 B),
    - ★ ECC (Error Correction Code) pro detekci a opravu chyb.
- **Stopa (Track)**
  - Množina sektorů na jednom povrchu ve stejné vzdálenosti od středu.
  - Počet sektorů ve stopě se může lišit v závislosti na poloměru.
  - Stopy se číslují od vnějšího okraje povrchu.
- **Cylindr (Cylinder)**
  - Množina všech stop o daném poloměru na všech površích.



HDD: Geometrie

- **Adresování sektorů**
  - CHS (Cylinder Head Sector)
    - Starší adresování podle geometrie disku.
    - Například sektor $[1, 2, 3]$ představuje data na 1. cylindru, 2. povrchu a ve 3. sektoru.
  - LBA (Logical Block Addressing)
    - "Novější" adresování.
    - Sektory jsou číslovány sekvenčně od nuly, začíná se od vnějšího okraje cylindru: $[0, 0, 0] \rightarrow 0$, $[0, 1, 0] \rightarrow 1$, $[0, 2, 0] \rightarrow 2$, ...
- **Zone Bit Recording (ZBR)**
  - Stopy rozděleny do zón.
  - V rámci zóny je konstantní počet sektorů na stopu.
  - Vnější zóny mají větší počet sektorů na stopu než zóny blíže ke středu.

HDD: Host Interface Controller

- HDD může být připojen k systému pomocí různých typů sběrnic s různými vlastnostmi.
- **Menší systémy (osobní počítače)**
  - SATA (Serial ATA)
    - Sériová sběrnice, rychlosti: 1.5 Gb/s, 3 Gb/s, 6 Gb/s,
    - vzdálenost: 1 m.
  - Thunderbolt
    - Seriová sběrnice spojující PCI-Express a Display port,
    - rychlosti: 40 Gb/s.
- **"Enterprise" systémy (velké dražší servery)**
  - SAS (Serial Attached SCSI)
    - Sériová sběrnice, rychlosti: 22.5 Gb/s, vzdálenost: 10m.
  - FC (Fibre Channel)
    - Sériová sběrnice, rychlosti: 128 Gb/s
    - vzdálenosti: 10km.

HDD: Rychlost přístupu k datům

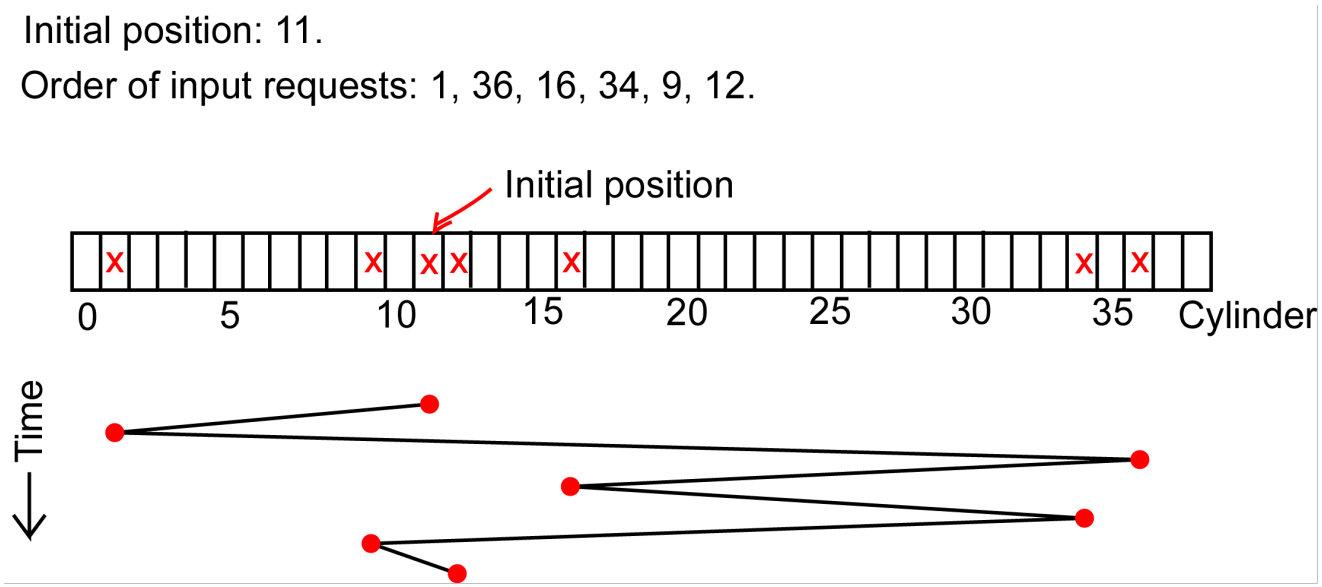
- **Na čem závisí rychlost čtení/sápisu jednoho sektoru?**
  - Doba vystavení (seek time)
    - ★ Čas nastavení hlaviček nad správný cylindr (cca 1-10ms).
  - Průměrné rotační zpoždění (rotational delay)
    - ★ Čas posunutí správného sektoru pod hlavičku,
    - ★ Při rotaci disku 5 000-15 000 rpm (rotations per minute)
      $\Rightarrow$ průměrné rotační zpoždění je 6-2ms.
  - Čas přenosu dat.
- **OS/aplikace je odpovědný/á za efektivní používání disku.**
  - minimalizování doby vystavení,
  - maximalizace počtu přenesených bytů za čas.
- **Algoritmy plánování přístupu na disk**
  - Dříve implementované pouze v OS, nyní v řadičích disků.
  - Určují pořadí zpracování jednotlivých požadavků.
    - ★ Tagged Command Queuing (TCQ) pro SCSI a ATA disky,
    - ★ Native Command Queuing (NCQ) pro SATA.

HDD: Příklad sekvenční x náhodný přístup k datům

- **Mějme disk s těmito parametry**
  - HDD má pouze jeden povrch, který se otáčí rychlostí 10000 rpm.
  - Velikost sektoru 512 B.
  - Každá stopa má 320 sektorů.
  - Průměrný čas vystavení hlaviček (seek time) je 10 ms.
  - Vystavení hlaviček nad sousední stopu (track-to-track seek time) je 1ms.
- **Jaké je průměrné rotační zpoždění?**
  - $\frac{0+(60/10000)}{2} = 3$ ms
- **Kolik stop potřebuji na uložení 2560 sektorů dat?**
  - $2560/320 = 8$ stop
- **Jak dlouho bude trvat přečíst 2560 sektorů uložených na sousedních stopách?**
  - Načtení sektorů na první stopě: $10 + 3 + 6 = 19$ ms.
  - Načtení sektorů na následující sousední stopě: $1 + 3 + 6 = 10$ ms.
  - Celková doba čtení všech sektorů: $19 + 7 \times 10 = 89$ ms.
- **Jak dlouho bude trvat přečíst 2560 sektorů uložených náhodně na disku?**
  - Načtení jedno sektoru: $10 + 3 + 6/320 = 13.01875$ ms.
  - Celková doba čtení všech sektorů: $2560x13.01875 = 33.328$ s.

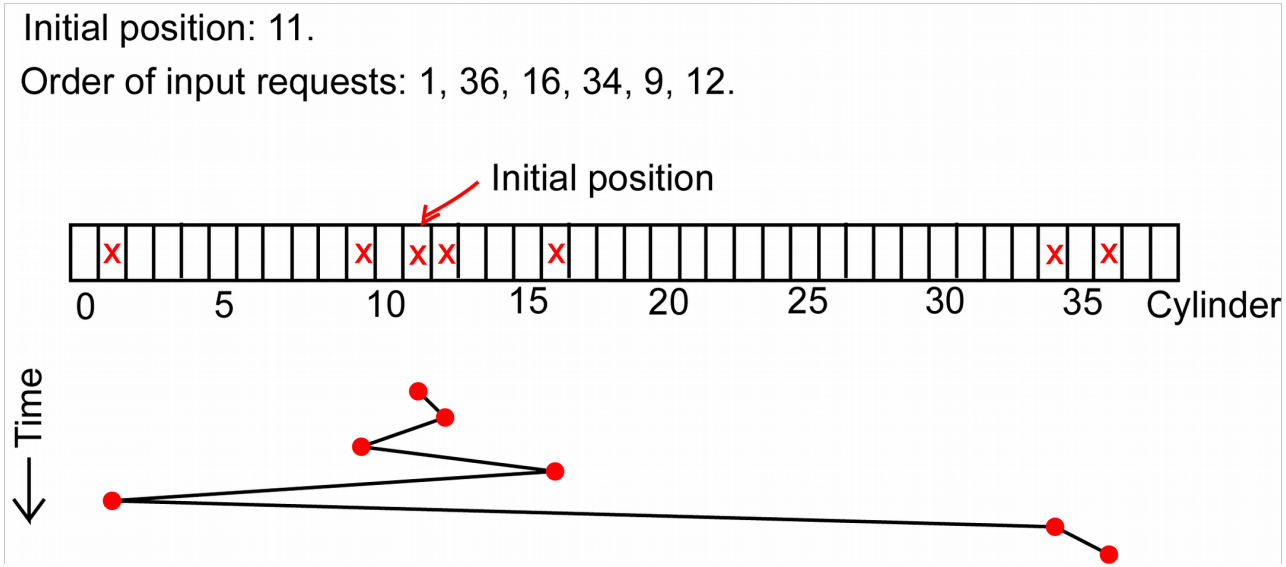
HDD: Algoritmy plánování přístupu na disk

- **First-In-First-OUT (FIFO)**
  - Požadavky (čtení/zápis) jsou řazeny do fronty.
  - Požadavky budou obslouženy v pořadí v jakém přišly.
  - Výhody: spravedlnost.
  - Nevýhody: horší výkon.



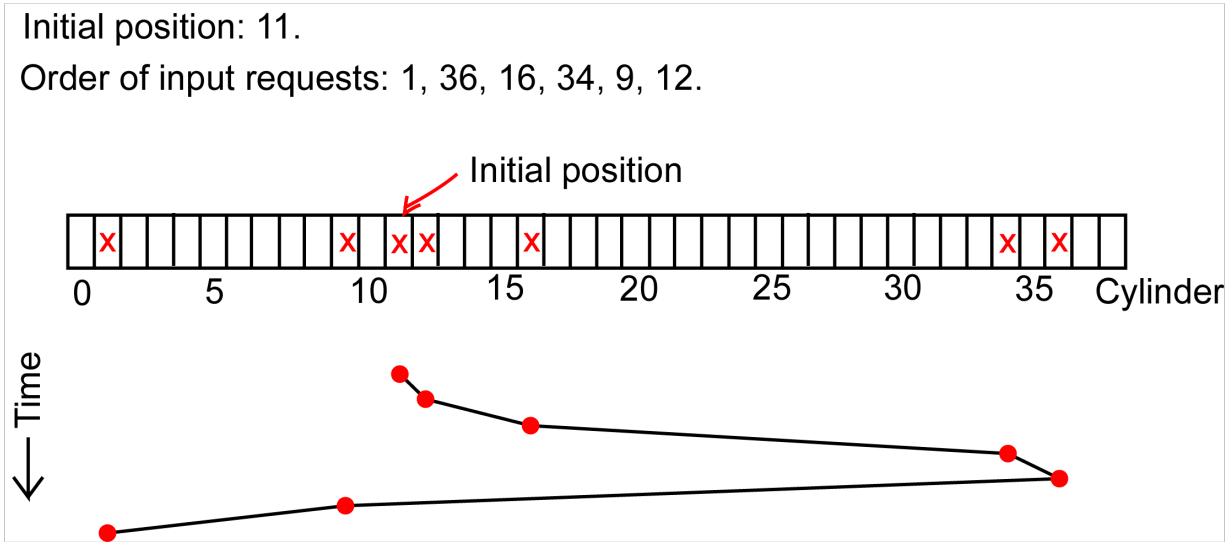
HDD: Algoritmy plánování přístupu na disk

- **Shortest Service Time First (SSTF)**
  - Požadavky (čtení/zápis) jsou řazeny do fronty.
  - Nejdříve jsou obslouženy požadavky, které vyžadují nejmenší pohyb hlaviček z aktuální pozice.
  - Výhody: lepší výkon než u FIFO.
  - Nevýhody:
    - Hlavičky mají tendenci setrvávat uprostřed disku.
    - Vzniká problém stárnutí u požadavků z krajních pozic.



HDD: Algoritmy plánování přístupu na disk

- **SCAN Algorithm (elevator alg.)**
  - Požadavky (čtení/zápis) jsou řazeny do fronty.
  - Hlavičky se pohybují nejdříve jedním směrem a uspokojí se všechny požadavky v daném směru. Pokud už není žádný požadavek v daném směru, směr se změní a uspokojí se všechny požadavky v druhém směru. Toto postup se opakuje.
  - Výhody: částečně se omezil problém stárnutí požadavků.
  - Nevýhody
    - Trochu horší výkon než SSTF algoritmus.
    - Neřeší stárnutí při velkém počtu požadavků v úzké oblasti cylindrů.

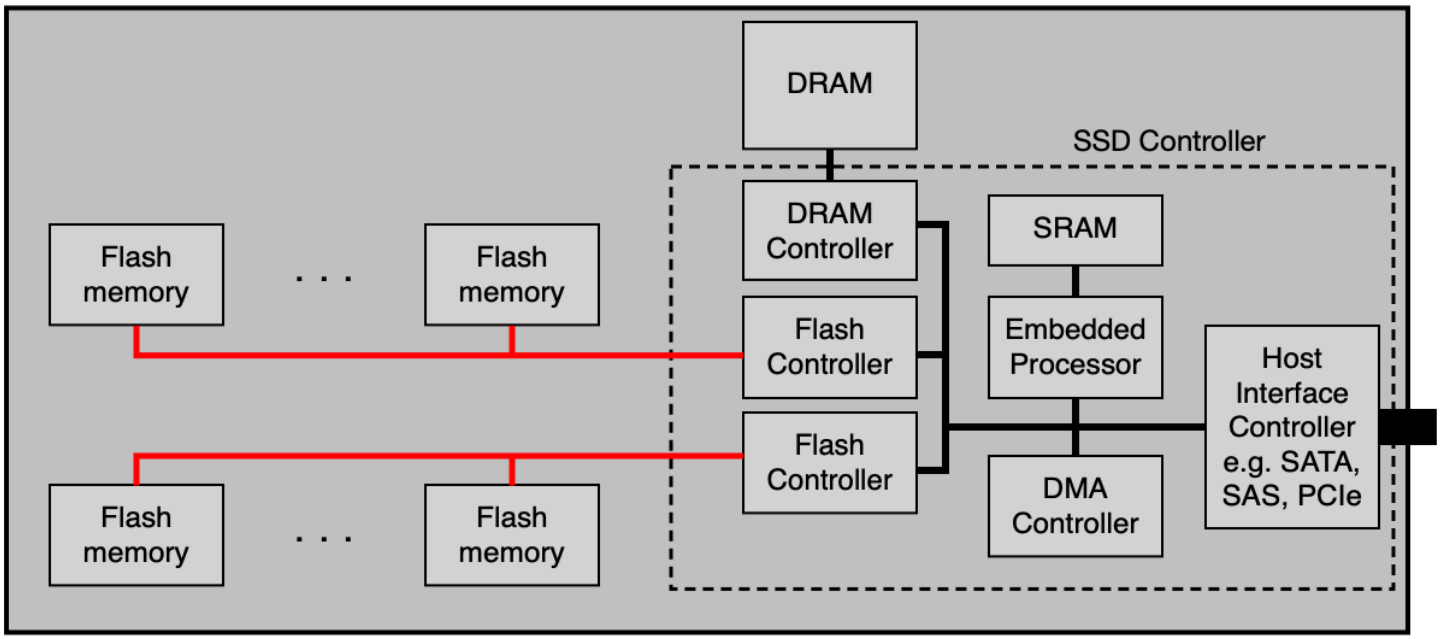


HDD: Algoritmy plánování přístupu na disk

- ***N*-step SCAN**
  - Vylepšená verze SCAN algortimu, který odtraňuje problém, strárnutí požadavků.
  - Původní fronta požadavků je rozdělena na několik front délky N, které se postupně plní požadavky.
  - Jednotlivé fronty jsou zpracovány postupně. Požadavky z jedné fronty jsou obslouženy pomocí SCAN algoritmu.
  - Tento algoritmus je zobecněním předchozích algoritmů.
    - Pokud bude $N = 1$, pak se bude chovat jako FIFO algoritmus.
    - Pokud bude $N \to \infty$, pak se bude chovat jako SCAN algoritmus.
  - Výhody:
    - Omezil se problém strárnutí požadavků, protože je garantováno, že požadavek může být předbehnut maximálně $N - 1$ jinými požadavky.
  - Nevýhoda: trochu horší výkon než SCAN algoritmus.

SSD (Solid State Drive): Architektura

- Neobsahuje žádné "mechanické" části jako HDD
  $\Rightarrow$ přístup k datům, který není závislý na umístění dat.
- Výhody
  - Rychlý přístup k datům.
  - Menší spotřeba, menší rozměry.
- Nevýhody
  - Menší kapacita.
  - Vyšší cena za byte.



Porovnání HDD a SSD

Parametr	HDD [5]	SSD [6]	
I/O interface	SATA/SAS	SATA/SAS	NVMe
Seuential Read (MB/s)	250	555	6 800
Sequential Write (MB/s)	250	520	6 000
Random Read (4KB) IOPS	240	98 000	1 000 000
Random Write (4KB) IOPS	240	75 000	180 000
Latency Read (ms)	Average 4.16	Max 0.4	0.1
Latency Write (ms)	Average 4.16	Max 4.0	0.03
Capacity	14 TB	3.84 TB	3.2 TB

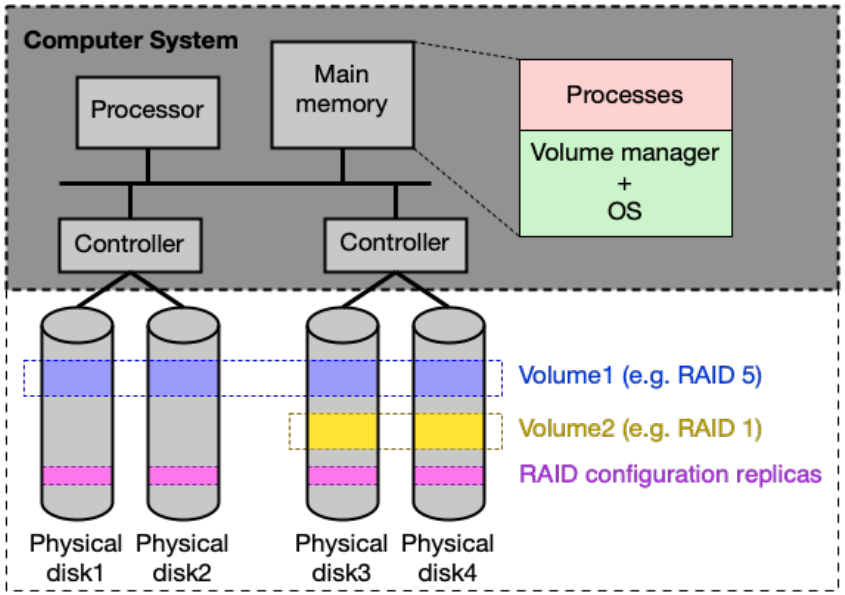
- Pokud potřebujeme
  - větší kapacitu datového úložiště,
  - větší rychlost sekvenčního čtení/zápisu,
  - větší počet R/W operací za sekundu (IOPS),
  - větší spolehlivost (dostupnost dat při výpadku mechaniky disku/flash paměti, diskového řadiče, připojení,...),
    $\Rightarrow$ pak je řešením RAID.

RAID (Redundant Array of Independent Disks)

- Myšlenka publikována v 1988 na univerzitě California Berkeley.
- Datové úložiště se skládá z množiny fyzických disků (v ideálním případě stejných disků HDD/SSD) a data jsou na ně ukládána (mapována) různými způsoby (různé typy RAID: 0, 1, 1+0, 5, 6, ...) a tím je dosahováno různých vlastností datového úložiště.
- Kromě RAID 0 jsou všechny ostatní typy redundantní $\Rightarrow$ část kapacity datového úložiště obsahuje redundantní informace, což umožňuje, že data jsou dostupná i v případě výpadku jednoho nebo více fyzických disků.
- Výhodou RAID oproti HDD/SSD je, že většina jeho vlastností lze efektivně konfigurovat prostřednictvím typu RAID a jeho parametrů.
  - Kapacita.
  - Rychlost sekvenčního čtení/zápisu.
  - Počet R/W operací za sekundu.
  - Spolehlivost.
- RAID je v praxi implementován dvěma základním způsoby
  - softwarový RAID,
  - hardwarový RAID.

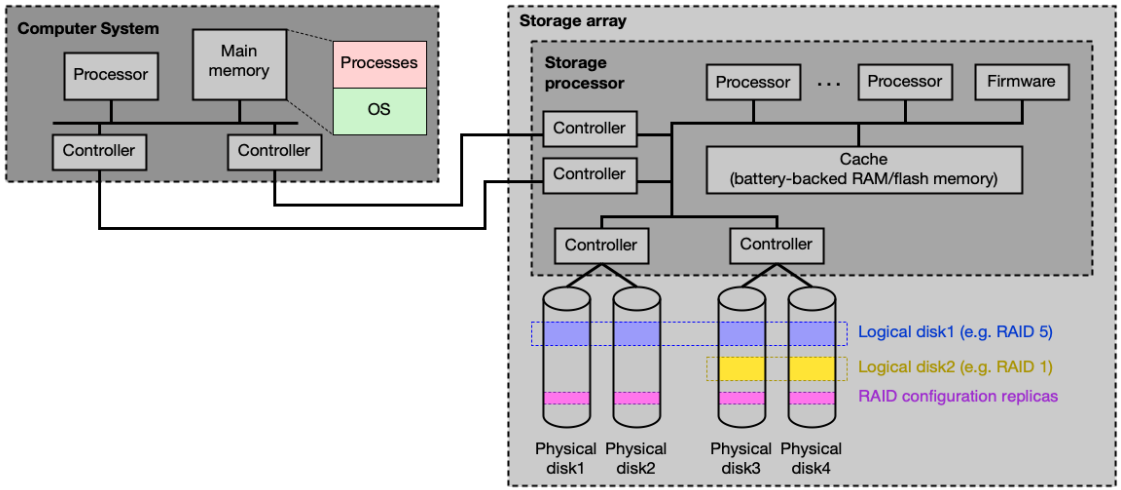
Softwarový RAID

- Fyzické disky HDD/SSD (disk1,..., disk4) jsou k systému připojeny přes příslušné sběrnice a jsou standardně spravované prostřednictvím OS .
- **Volume manager (VM)**
  - Software, který ukládá (mapuje) data na jednotlivé fyzické disky a provádí nutné výpočty (např. výpočet parity,...).
  - Spravuje logické disky (volumes), které představují konkrétní typ RAIDu a poskytuje interface, přes který k nim může přistupovat OS a jednotlivé aplikace.
  - Konfigurace celého VM je uložena na fyzických discích.
  - Příklady reálných VM
    - ★ Logical VM (Linux), Solaris VM (Solaris),...
    - ★ Veritas VM (MS Windows, Linux, Solaris, AIX,...).



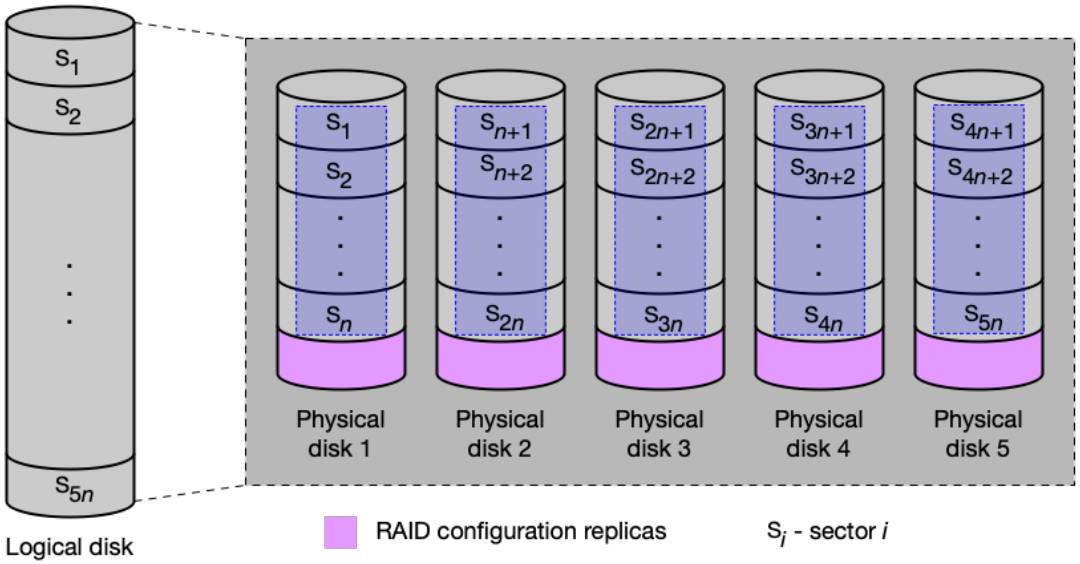
Hardwarový RAID

- Reprezentován speciálním hardwarem, který obsahuje
  - jeden nebo několik procesorů,
  - skrytou paměť, která je chráněna proti výpadku napájení a slouží pro dočasné uložení dat,
  - paměť s firmwarem,
  - fyzické disky (HDD/SSD), které jsou připojeny příslušnými sběrnicemi k systému.
- **Firmware**
  - Stará se o ukládání (mapování) dat na jednotlivé fyzické disky a provádí příslušné výpočty (např. výpočet parity,...).
  - Spravuje logické disky, které představují konkrétní typ RAIDu a poskytuje k nim interface pro OS/aplikace, které běží na připojeném výpočetním systému (OS přímo nevidí fyzické disky HW RAIDu).
  - Poskytuje interface pro konfigurování a monitorování HW RAIDu.



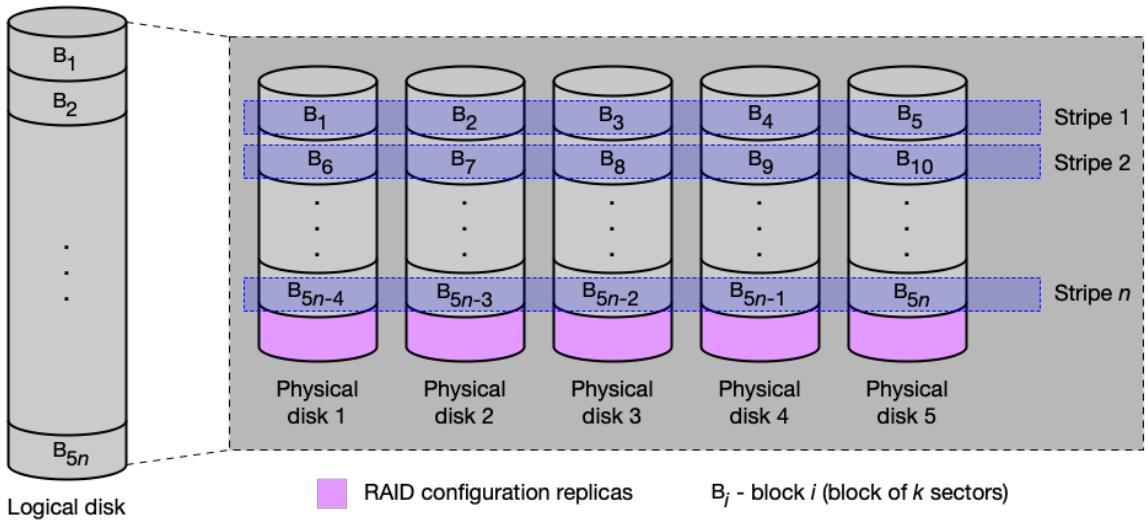
RAID 0 – zřetězení (concatenation)

- Někdy také označováván jako JBOD (Just a Bunch Of Disks).
- **Princip**
  - Data jsou ukládána/mapována postupně na jednotlivé fyzické disky (jakmile se zaplní první disk, data se začnou ukládat na druhý disk, ...).
- **Vlastnosti**
  - Redundance je 0%
    $\Rightarrow$ výpadek jednoho disku způsobí ztrátu všech dat.
  - Výkon logického disku je skoro stejný jako výkon jednoho fyzického disku.
- **Použití**
  - Navýšení kapacity diskového úložiště.



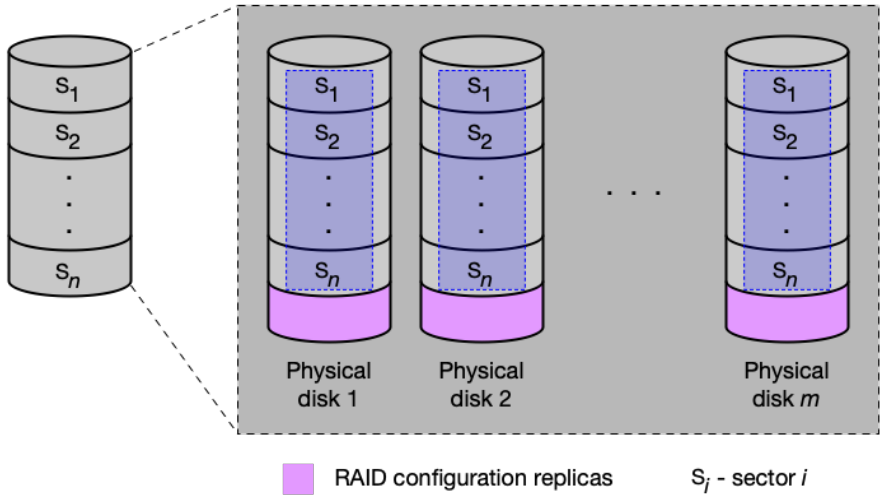
RAID 0 – prokládání (striping)

- **Princip**
  - Při vytváření RAIDu administrátor definuje blok jako k sousedních sektorů.
  - Data jsou ukládána/mapována cyklicky po blocích na jednotlivé fyzické disky (jakmile se zaplní první "stripe", data se začnou ukládat na druhý "stripe", ...).
- **Vlastnosti**
  - Redundance je 0% $\Rightarrow$ výpadek jednoho disku způsobí ztrátu všech dat.
  - Nechť m je počet fyzických disků.
  - R/W operace se zrychlí až m krát, pokud velikost dat bude m bloků.
  - Počet R/W operací za sekundu se zvýší až m krát, pokud velikost dat bude jeden blok.
- **Použití**
  - Navýšení kapacity a výkonu diskového úložiště.



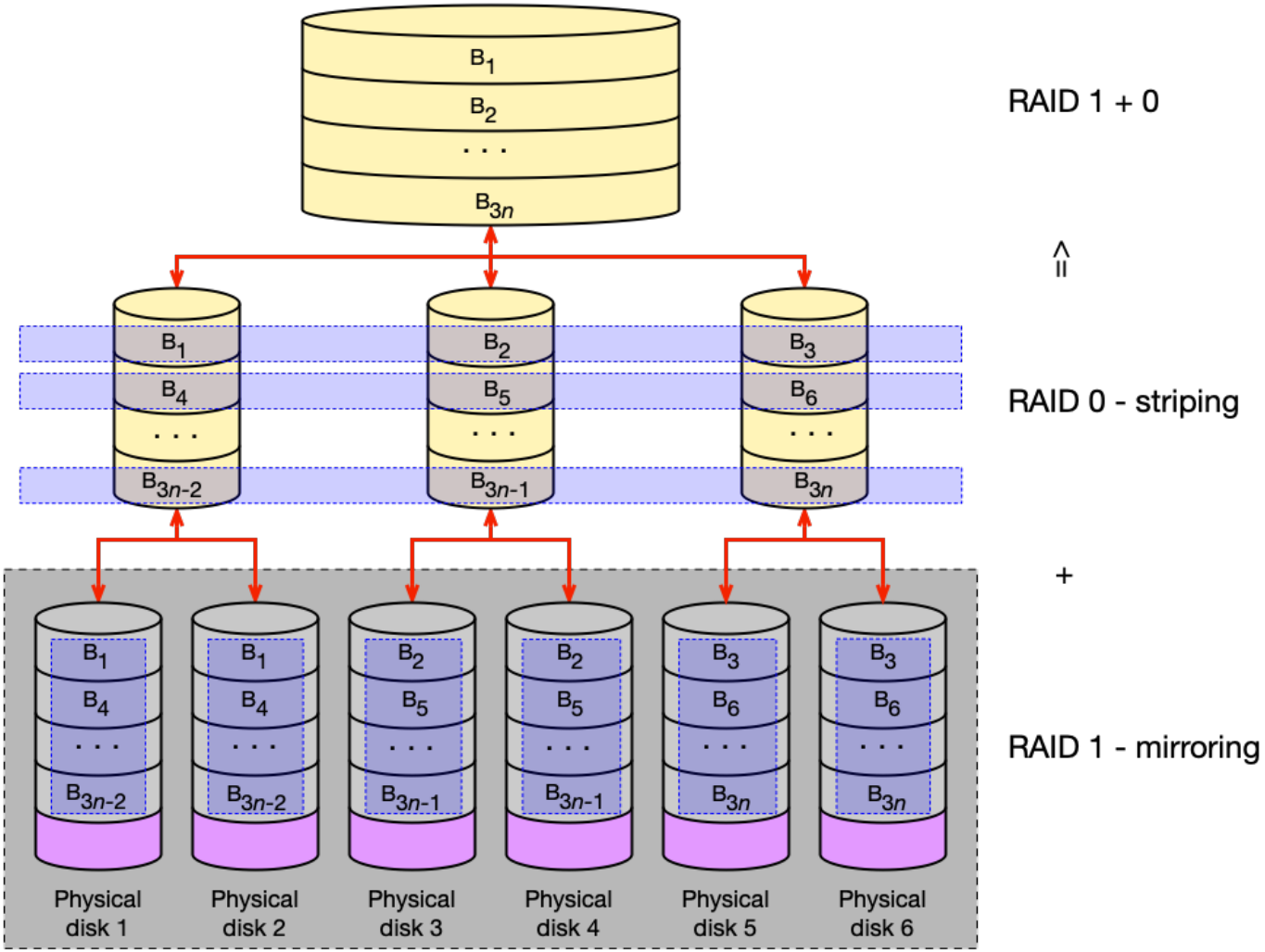
RAID 1 – zrcadlení (mirroring)

- **Princip**
  - Stejná data jsou ukládána/mapována na všechny fyzické disky (každý disk obsahuje stejnou kopii dat). RAID 1 běžně obsahuje dvě kopie dat/dva fyzické disky, ale obecně může obsahovat m kopií dat/fyzických disků.
- **Vlastnosti**
  - Redundance je $100 \times (m-1)/m$ %
    $\Rightarrow$ data přežijí výpadek $m-1$ disků a výkon nebude degradován.
  - R/W operace bude přibližně stejně rychlá jako u fyzického disku.
  - Počet Read operací za sekundu se zvýší až m krát a počet Write operací za sekundu bude přibližně stejný jako u fyzického disku.
- **Použití**
  - Zabezpečení dat na datovém úložišti.



RAID 1+0 (stripe), RAID 10

- RAID 1+ 0 je kombinací RAIDu 1 (zrcadlení) a RAIDu 0 (prokládání), tak aby výsledný RAID získal dobré vlastnosti z obou typů RAIDů.

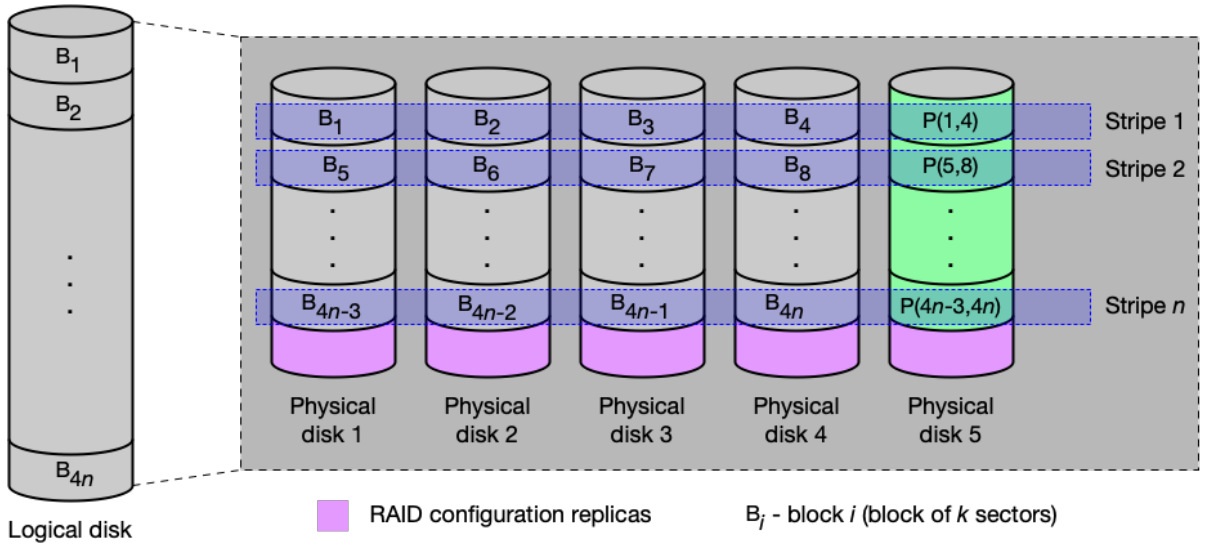


RAID 1+0 (stripe), RAID 10

- **Vlastnosti**
  - Redundance je 50 % (pokud zrcadlení bude mít pouze dvě kopie dat)
    $\Rightarrow$ data přežijí teoreticky výpadek $m/2$ disků a výkon nebude degradován.
  - R/W operace se zrychlí až $(m/2)$ krát, pokud velikost dat bude $m/2$ bloků.
  - Počet Read operací za sekundu se zvýší až m krát, pokud velikost dat bude jeden blok.
  - Počet Write operací za sekundu se zvýší až $(m/2)$ krát, pokud velikost dat bude jeden blok.
  - Při výpadku jednoho fyzického disku bude doba obnovy úměrná kapacitě jednoho fyzického disku.

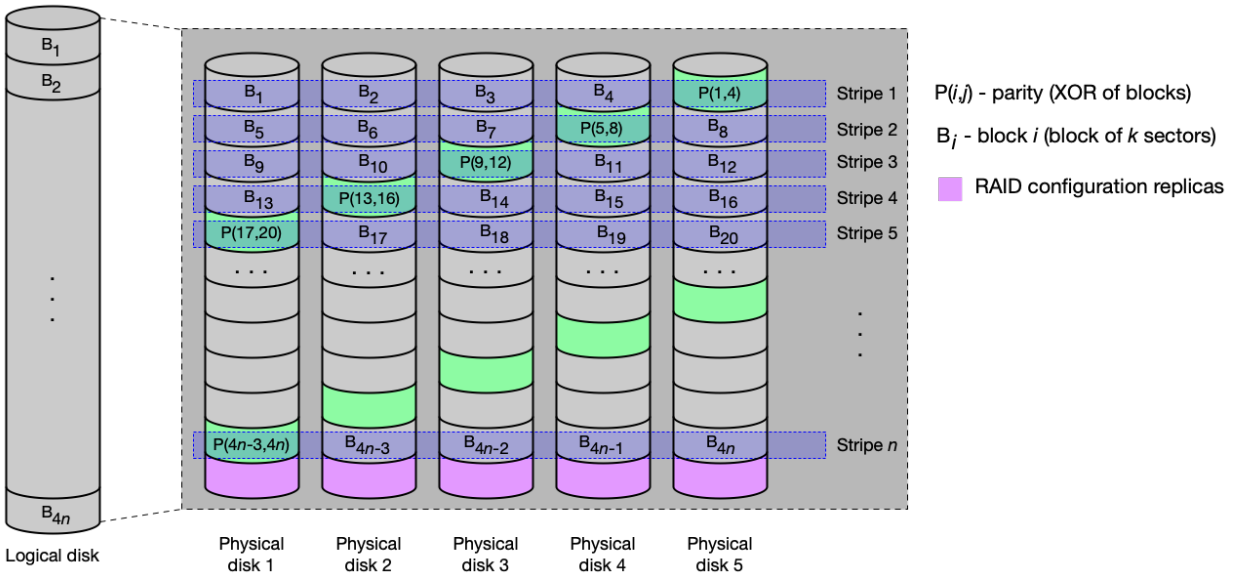
Raid 2, 3, 4

- RAID 2, 3 a 4 se běžně v datových úložištích nepoužívají.
- **RAID 2**
  - Prokládání po bitech + zabezpečení pomocí Hammingova kódu.
- **RAID 3**
  - Prokládání po bytech + zabezpečení pomocí parity uložené na jednom fyzickém disku.
- **RAID 4**
  - Prokládání po blocích + zabezpečení pomocí parity uložené na jednom fyzickém disku. Parita je definována jako $P(i,j) = B_i \; XOR \; B_{i+1} \; XOR \; \ldots \; XOR \; B_j$.
  - Problém: paritní disk může být při větším počtu zápisů přetížen.



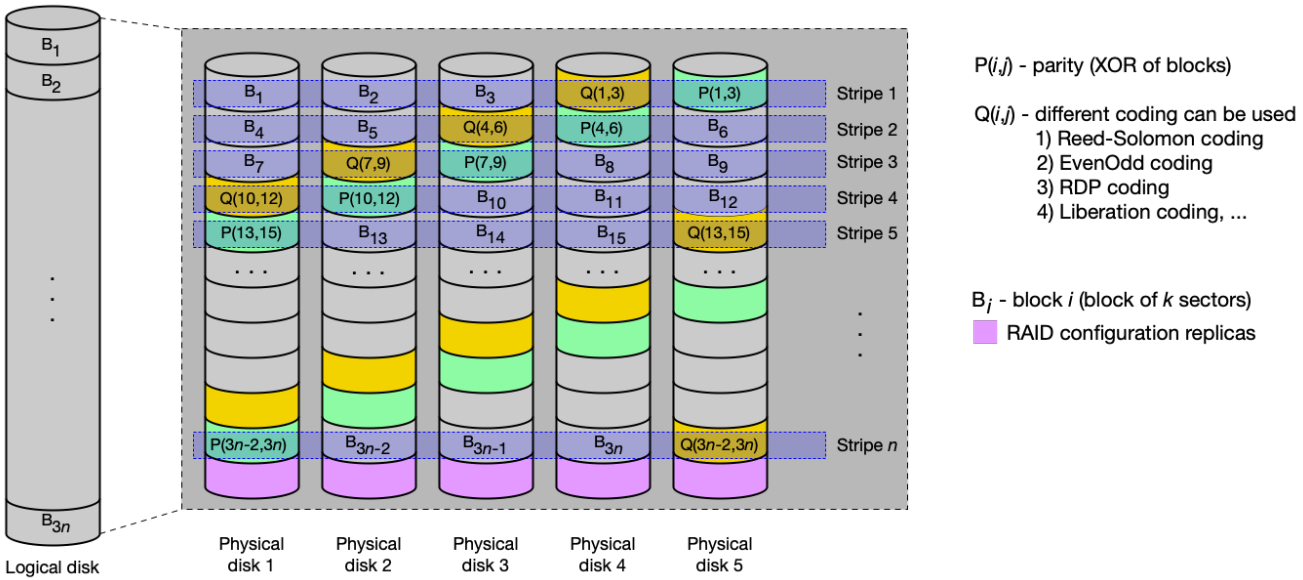
Raid 5 – prokládání s distribuovanou paritou

- **Princip**
  - Prokládání po blocích na m fyzických discích + zabezpečení pomocí parity cyklicky ukládané na jednotlivých fyzických discích.
- **Vlastnosti**
  - Redundance je $100/m$ % $\Rightarrow$ při výpadku jednoho disku budou data ještě dostupná, ale bude degradován výkon.
  - Read operace se zrychlí až $(m-1)$ krát, pokud velikost dat bude $(m-1)$ bloků.
  - Write operace pomalejší vzláště u SW RAIDu!
  - Počet R operací za sekundu se zvýší až m krát, pokud velikost dat bude jeden blok.
- **Použití**
  - Navýšení kapacity, zabezpečení dat a navýšení výkonu u Read operací.



Raid 6 – prokládání s distribuovanou paritou

- **Princip**
  - Prokládání po blocích na m fyzických discích + zabezpečení pomocí dvojí parity cyklicky ukládané na jednotlivých fyzických discích.
- **Vlastnosti**
  - Redundance je $200/m$ % $\Rightarrow$ při výpadku dvou disků budou data ještě dostupná, ale bude degradován výkon.
  - Read operace se zrychlí až $(m-2)$ krát, pokud velikost dat bude $(m-1)$ bloků.
  - Write operace pomalejší vzláště u SW RAIDu!
  - Počet R operací za sekundu se zvýší až m krát, pokud velikost dat bude jeden blok.
- **Použití**
  - Navýšení kapacity, zabezpečení dat a navýšení výkonu u Read operací.

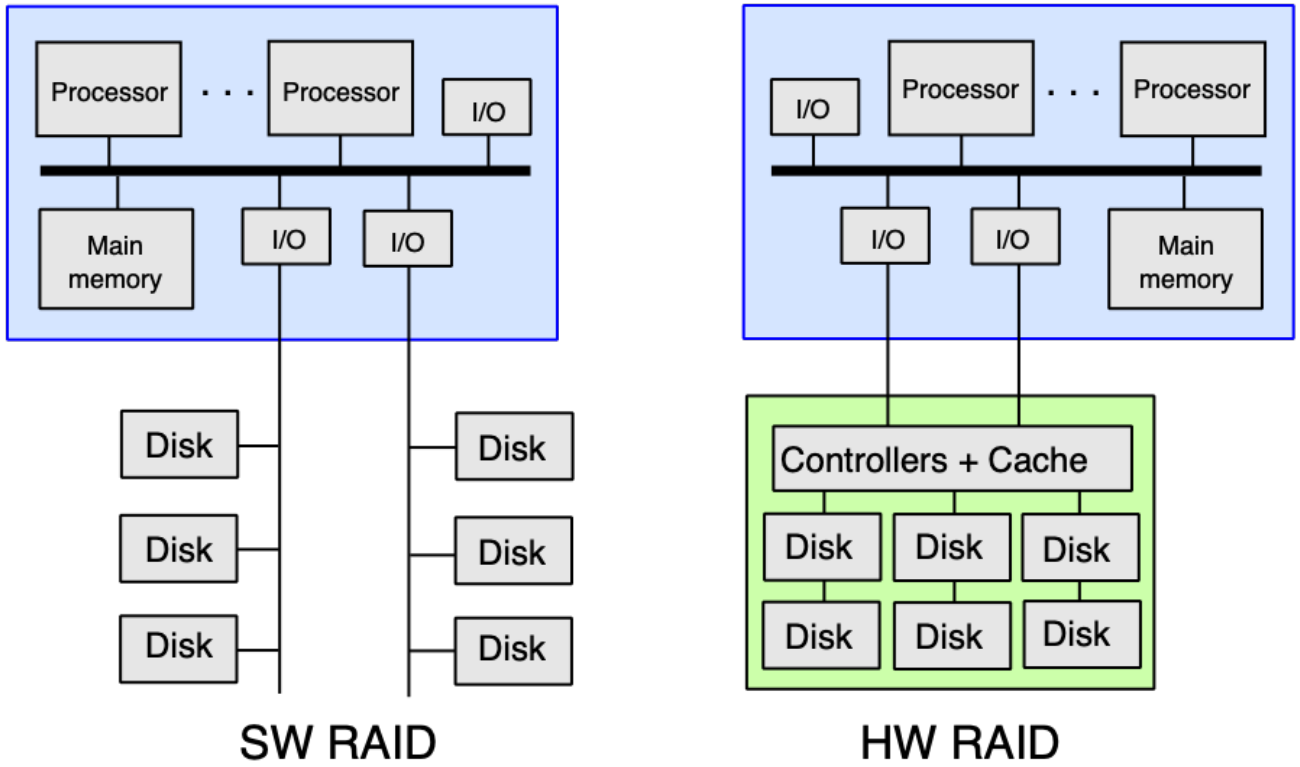


Porovnání různých typů RAID

Vlastnosti	RAID 0 striping	RAID 1	RAID 10	RAID 5	RAID 6
Minimální počet disků	2	2	4	3	4
Ochrana dat	Žádná ochrana	Výpadek jednoho disku	Výpadek jednoho disku	Výpadek jednoho disku	Výpadek dvou disků
Výkon čtení	Vysoký	Vysoký	Vysoký	Vysoký	Vysoký
Výkon zápisu	Vysoký	Střední	Střední	Nízký	Nízký
Výkon čtení (při výpadku disku)	–	Střední	Vysoký	Nízký	Nízký
Výkon zápisu (při výpadku)	–	Vysoký	Vysoký	Nízký	Nízký
Využitá kapacita	100%	50%	50%	67%-94%	50%-88%

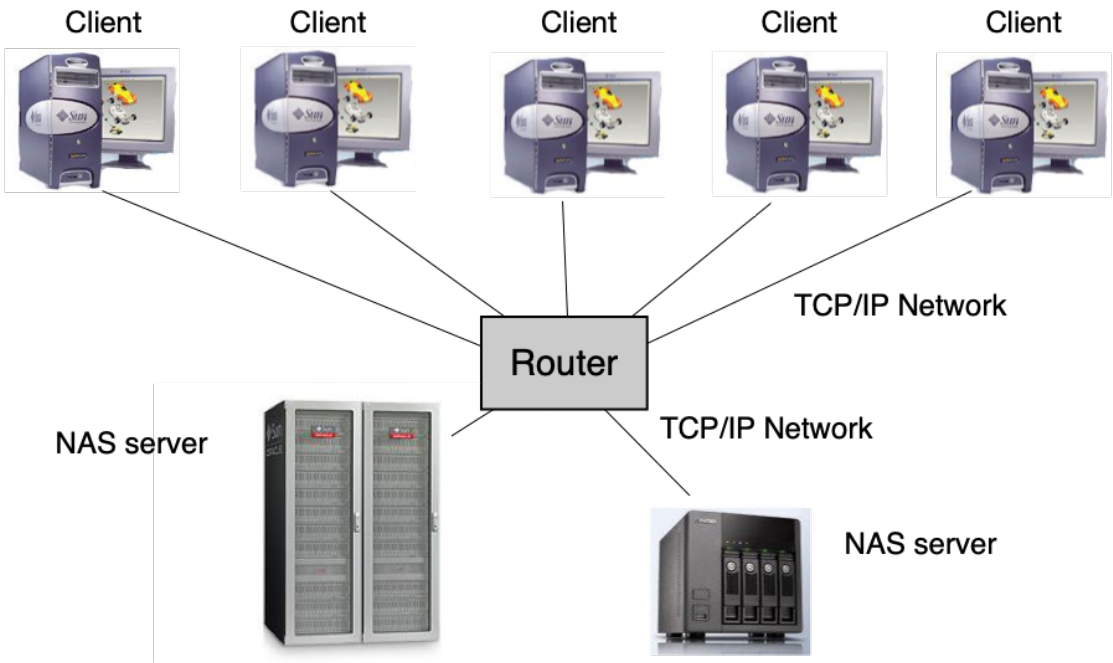
DAS (Direct-Attached Storage)

- Úložiště (HDD/SSD/RAID) je k systému připojeno přímo přes V/V porty systému.
- OS/aplikace vidí jednotlivé sektory v úložišti.
- **Technologie připojení**
  - SATA, SAS, Ethernet, Fibre channel,...
- **Vlastnosti**
  - Některé technologie (např. SCSI) umožňují "multihosting" (jedno úložišti lze připojit k více systémům současně).



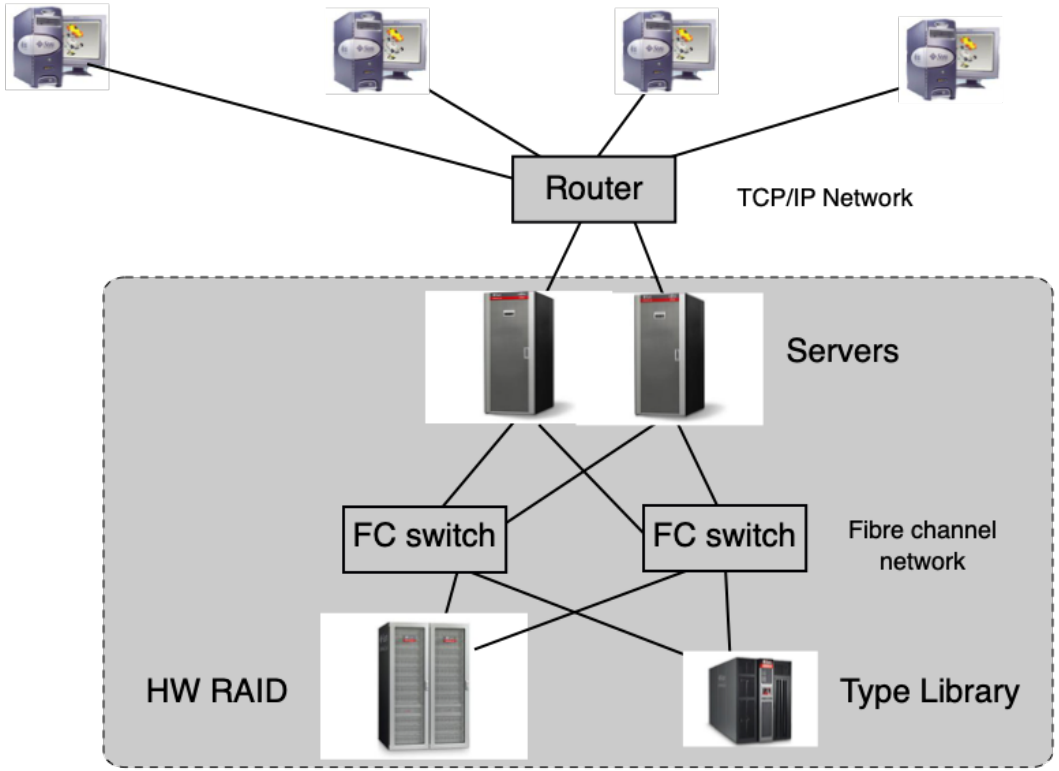
NAS (Network-Attached Storage)

- **NAS server**
  - Výpočetní systém + OS.
  - Datové úložiště (HDD/SSD/RAID).
- **Technologie připojení**
  - Data jsou přístupná na úrovni distribuovaného systému souborů přes příslušné protokolysíťové protokoly
    - ★ NFS (Network File System),
    - ★ SMB (Server Message Block), Samba, CIFS (Common Internet File System).



SAN (Storage Area Network)

- Datová uložiště jsou na oddělěné síti.
- OS/aplikace vidí jednotlivé sektory v úložišti.
- **Technologie připojení**
  - Ethernet, Fibre channel,...
- **Vlastnosti**
  - Multihosting(jedno úložiště lze připojit k více systémům současně).
  - Multipathing (mezi datovým úložištěm a systémem existuje více nezávislých cest).



Použité zdroje

1. A. S. Tanenbaum, H. Bos: *Modern Operating Systems (4th edition)*, Pearson, 2014.
2. W. Stallings: *Operating Systems: Internals and Design Principles (9th edition)*, Pearson, 2017.
3. A. Silberschatz, P. B. Galvin, G. Gagne: *Operating System Concepts (9th edition)*, Wiley, 2012.
4. R. McDougall, J. Mauro: *Solaris Internals: Solaris 10 and OpenSolaris Kernel Architecture (2nd edition)*, Prentice Hall, 2006.
5. *Seagate HDD.* [Online]. Available: https://www.seagate.com/www-content/datasheets/pdfs/ironwolf-pro-14tb-DS1914-7-1807US-en_US.pdf. [Accessed: 23-Apr-2019].
6. *Kingston SSD.* [Online]. Available: https://www.kingston.com/en/ssd/enterprise/DC500R. [Accessed: 23-Apr-2019].