

Univerza v Ljubljani  
Fakulteta za računalništvo  
in informatiko



# Predmet: Osnove podatkovnih baz

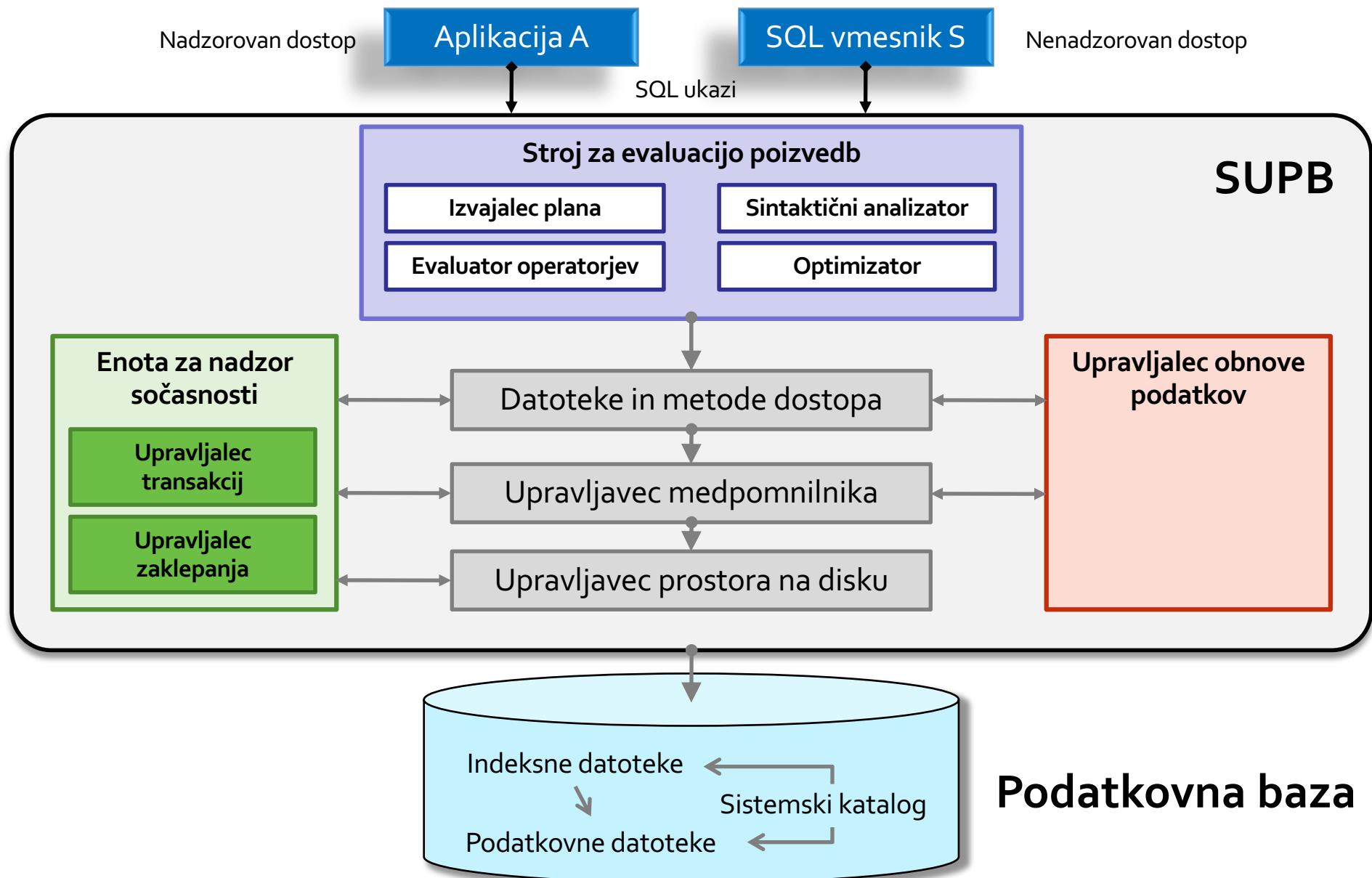
Modul:  
Diski in diskovna polja

Gradivo:  
v.2015



# Vsebina

- Zgradba SUPB, komponente SUPB za delo s podatki
- Primarni, sekundarni in terciarni pomnilnik
- Magnetni disk
- Polje diskov





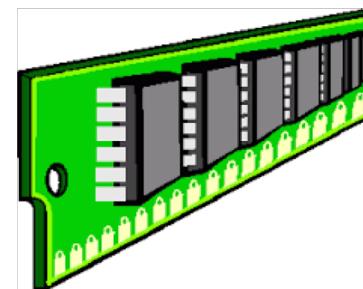
# Komponente SUPB za delo s podatki...

- Podatki iz PB se hranijo na diskih (in trakovih).
- **Upravljalec prostora na disku (*Disk Space Manager*):**
  - upravlja s prostorom na disku,
  - ukaze v zvezi z zaseganjem/sproščanjem prostora prejema od upravlјavca z datotekami.
- **Upravljalec z datotekami (*File Manager*):**
  - Posreduje zahteve za zaseganje/sproščanje prostora na disku v enotah – **straneh**.
  - odgovoren za upravljanje strani znotraj datoteke, za urejanje zapisov znotraj strani.
  - Velikost strani eden od parametrov SUPB (tipično 4 - 8 KB).



# Komponente SUPB za delo s podatki

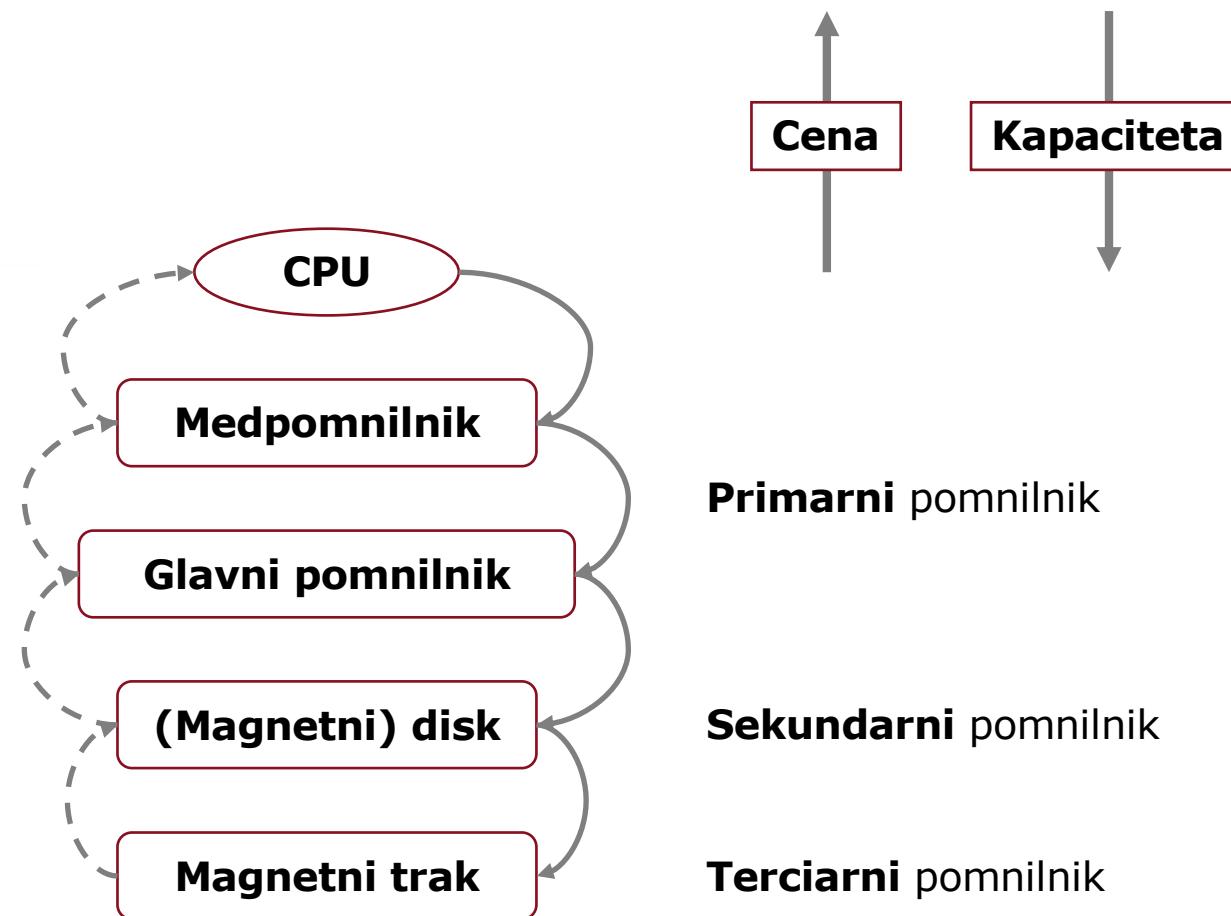
- Upravljalec medpomnilnika (*Buffer manager*):
  - prenos strani iz diska v medpomnilnik (*buffer pool*).
  - stran, kjer je zapis, poišče upravljalec z datotekami.
  - prenos v medpomnilnik izvede upravljalec medpomnilnika.





# Hierarhija pomnilnika...

Zahteva po podatkih →  
Prenos podatkov na zahtevo →



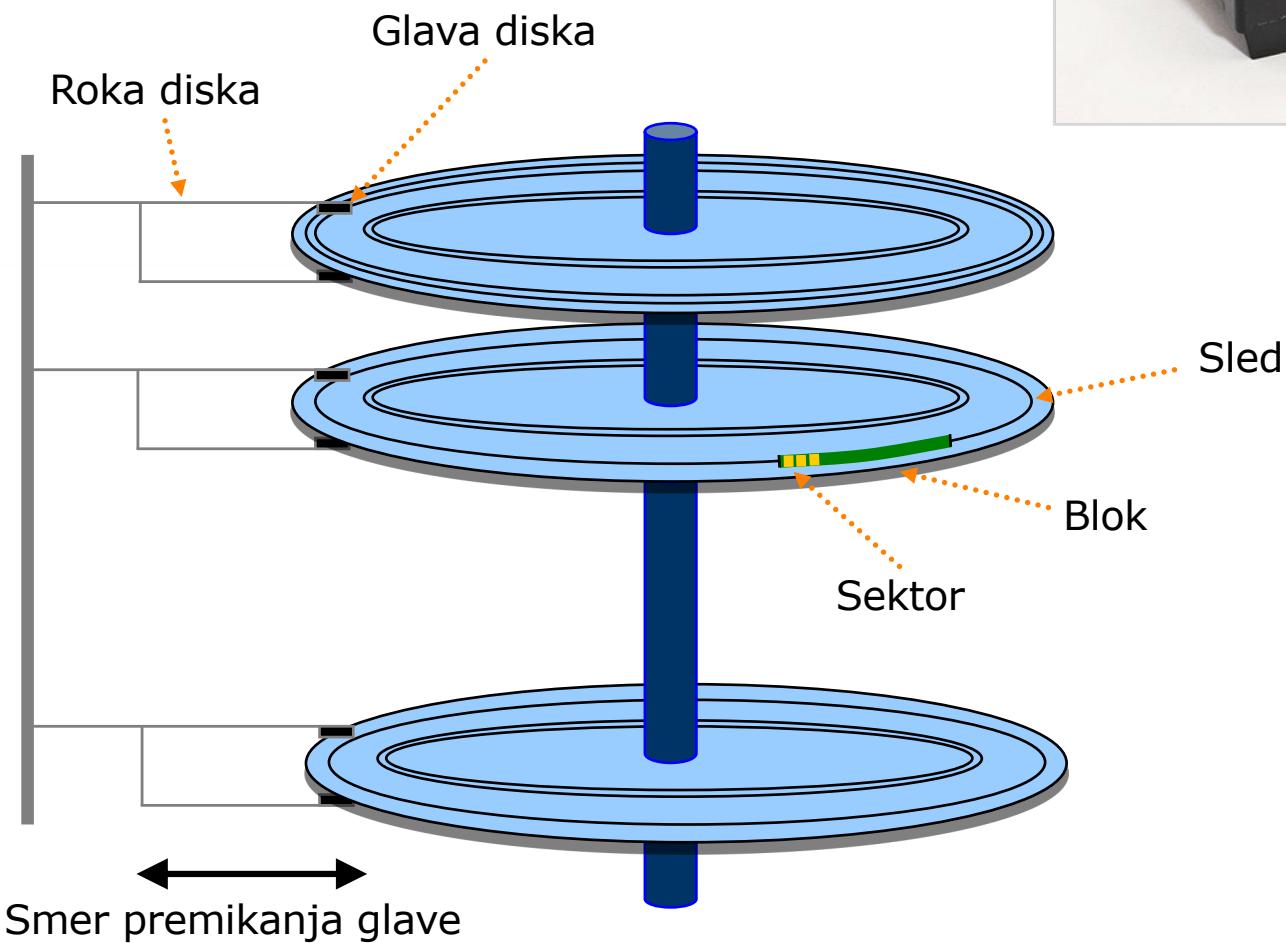


# Hierarhija pomnilnika

- Razlogi za shranjevanje podatkov v sekundarnem in terciarnem pomnilniku:
  - Obstojnost podatkov
  - Cena na enoto
  - Omejen naslovni prostor primarnega pomnilnika ( $2^{32}=4\text{Gb}$  podatkov...)



# Magnetni disk





# Magnetni disk

- Povprečni dostopni čas:
  - iskalni čas
  - rotacijska zakasnitev
  - čas prenosa
- Organizacija podatkov na disku vpliva na povprečni dostopni čas!
- Čas prenosa običajno večji od časa obdelave → pomembna organizacija strani...
- Dostopni čas RAM : disk  $\approx 1 : 1000$

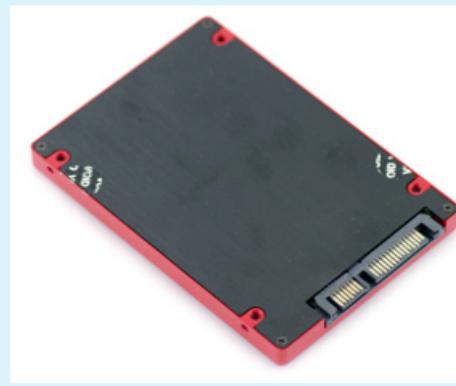


# HDD in SSD disk...

- HDD – *Hard Disk Drive*
- SSD – *Solid State Drive*



SSD Top Side



SSD Bottom Side



HDD Top Side



HDD Bottom Side



# HDD in SSD disk

Attribute	SSD (Solid State Drive)	HDD (Hard Disk Drive)
Power Draw / Battery Life	Less power draw, averages 2 – 3 watts, resulting in 30+ minute battery boost	More power draw, averages 6 – 7 watts and therefore uses more battery
Cost	Expensive, \$1.00 per gigabyte (based on buying a 240GB drive)	Only around \$0.075 per gigabyte, very cheap (buying a 4TB model)
Capacity	Typically not larger than 512GB for notebook size drives	Typically 500GB – 2TB for notebook size drives
Operating System Boot Time	Around 22 seconds average bootup time	Around 40 seconds average bootup time
Noise	There are no moving parts and as such no sound	Audible clicks and spinning can be heard
Vibration	No vibration as there are no moving parts	The spinning of the platters can sometimes result in vibration
Heat Produced	Lower power draw and no moving parts so little heat is produced	HDD doesn't produce much heat, but it will have a measurable amount more heat than an SSD due to moving parts and higher power draw
Failure Rate	Mean time between failure rate of 2.0 million hours	Mean time between failure rate of 1.5 million hours
File Copy / Write Speed	Generally above 200 MB/s and up to 500 MB/s for cutting edge drives	The range can be anywhere from 50 – 120MB / s
Encryption	Full Disk Encryption (FDE) Supported on some models	Full Disk Encryption (FDE) Supported on some models
File Opening Speed	Up to 30% faster than HDD	Slower than SSD
Magnetism Affected?	An SSD is safe from any effects of magnetism	Magnets can erase data



# Polje diskov...

- Disk potencialno **ozko grlo** za učinkovitost SUPB ... vpliva na zanesljivost delovanja sistema.
- Učinkovitost CPU/disk:
  - CPU: 50% na leto
  - Diski: 10% na leto
- Diski mehanske naprave → verjetnost za napake večja kot pri notranjem pomnilniku.
- Odpoved diska ... katastrofa.
- Možna rešitev: več diskov.



# Polje diskov

- Polje diskov: povezava več diskov z namenom:
  - povečanja **učinkovitosti** in/ali
  - izboljšanja **zanesljivosti**.
- Učinkovitost ... porazdelitev podatkov (*data striping*)
- Zanesljivost ... podvajanje podatkov - redundanca



# RAID

- RAID – *Redundant Arrays of Independent Disks* – diskovna polja, ki implementirajo porazdelitev/ podvajanje podatkov.
- Več vrst RAID ... razlika v kompromisu med **učinkovitostjo** in **zanesljivostjo**.





# RAID s porazdelitvijo podatkov

- Uporabniku se kaže kot velik disk.
- Podatki se razdelijo na enake **enote (*striping units*)**, ki se zapišejo na več diskov. Vsaka enota na en disk.
- Enote se po diskih porazdelijo po “***round robin***” algoritmu: če polje vključuje D diskov, se enota i zapiše na “ $i \bmod D$ ” disk.



## Primer 1

- RAID z D diskami v polju.
  - RAID enota = 1 bit.
-



## Primer 2

- RAID z D diskami v polju.
  - RAID enota = 1 **blok**.
-



# RAID z redundanco podatkov...

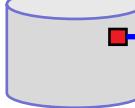
- Več diskov → večja učinkovitost (shranjevanja) → manjša zanesljivost.
- Primer:
  - MTTF (*mean-time-to-failure*) enega diska  $\approx 50.000$  ur (5,7 let). Pri 100 diskih v polju MTTF  $50.000/100 \approx 500$  ur (**21 dni**).
- Za večjo zanesljivost (večji MTTF) potrebna **redundanca**.
- Primer:
  - Če polju 100-ih diskov dodamo 10 diskov z redundantnimi podatki → MTTF > 250 let!!!



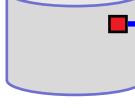
# RAID z redundanco podatkov...

- Vprašanje:
  - Kje hraniti redundantne podatke (na določenih/vseh diskih)?
- Kaj podvajati?
- Večinoma redundanten disk za **paritetni bit**...

# Paritetni bit

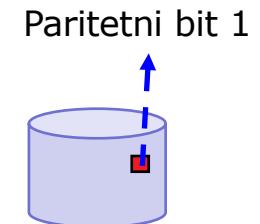
Disk 1  Bit 1 na disku 1 = 1

Disk 2  Bit 1 na disku 2 = 0

Disk 3  Bit 1 na disku 3 = 0

Disk D  Bit 1 na disku D = 1

Redundantni disk



$$\text{Paritetni\_bit}_i = \text{Odd}(\sum_{n=1}^D \text{disk}_n(\text{bit}_i))$$



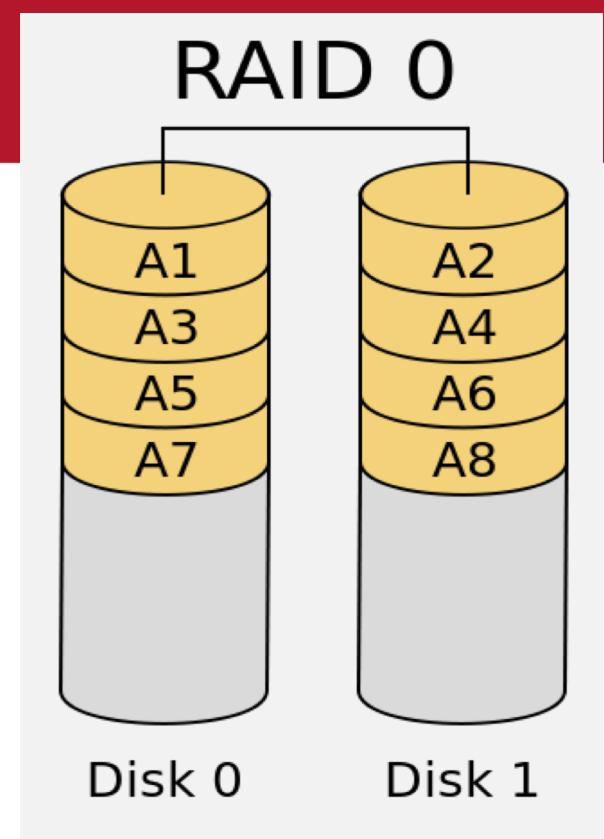
# Stopnje redundance...

- V RAID sistemu diskovno polje sestavljeni iz:
  - množice **podatkovnih diskov** in
  - množice **kontrolnih diskov**.
- Število kontrolnih diskov odvisno od stopnje redundance.
- Parametri za primere v nadaljevanju:
  - Količina podatkov za 4 diske
  - Ena sama kontrolna skupina



# RAID 0

- Porazdeljen, brez redundance (*nonredundant*)
- Uporablja porazdeljevanje podatkov za povečanje pasovne širine.
- Ne vzdržuje nobene redundantne informacije.
- PROBLEM: MTTF pada linearno s številom diskov v polju.
- PREDNOSTI:
  - najvišja učinkovitost → ni potrebno vzdrževati nobenih redundantnih podatkov.
  - 100% izraba prostora na disku. V našem primeru rabimo za svoje podatke 4 diske.

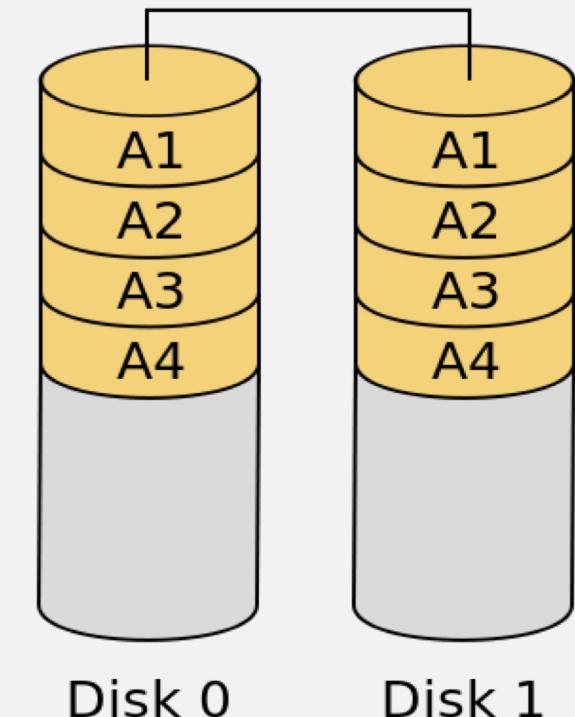




# RAID 1

- Zrcaljen (*mirrored*)
  - Najdražja rešitev → vzdržujeta se dve kopiji podatkov na dveh diskih.
  - Vsako pisanje bloka na disk pomeni pisanje na dva diska.
- 
- Pisanje eno za drugim (možnost nesreče med pisanjem).
  - Branje lahko paralelno (branje dveh različnih blokov iz dveh diskov; možno branje iz diska z minimalnim dostopnim časom).
  - Izraba prostora 50%. V našem primeru 8 diskov (4 + 4).

RAID 1



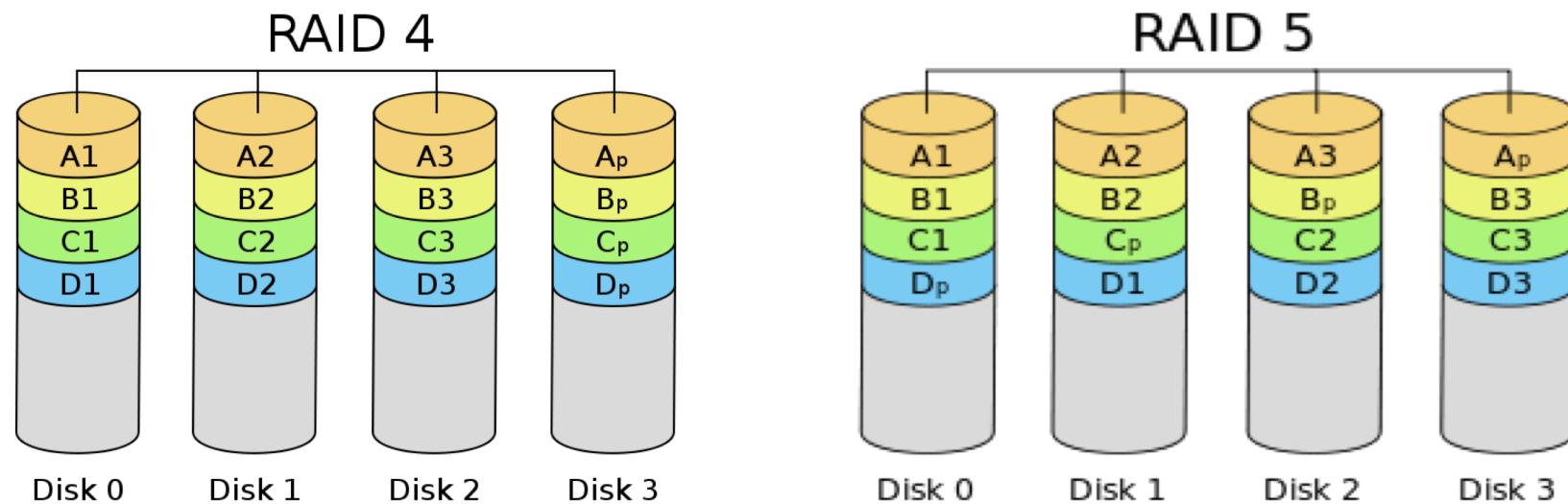


# Druge stopnje redundance

- RAID 2: *Error-Correction Codes*
- RAID 3: *Byte-Interleaved Parity*
- RAID 4: *Block-Interleaved Parity*
- RAID 5: *Block-Interleaved Distributed Parity*
- RAID 6: *Block-Interleaved Double Distributed Parity*
- Pomembni parametri:
  - *Space Efficiency* - Izkoriščenost prostora
  - *Fault Tolerance* – Št. diskov, ki lahko brezizgubno odpovedo
  - *Array failure rate* – Verjetnost, da odpove polje
  - *Read performance* – učinkovitost branja
  - *Write performance* – učinkovitost pisanja

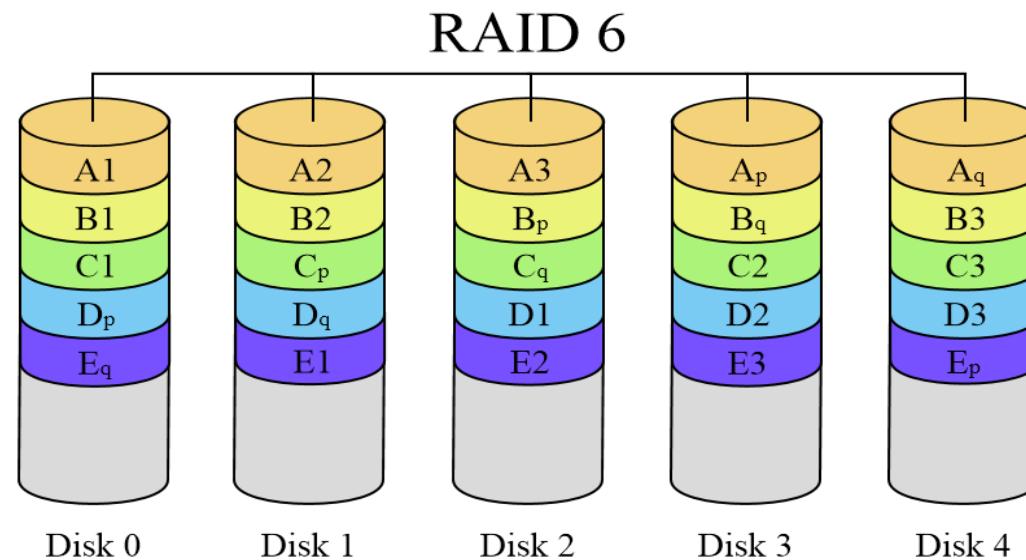
# Razlika med RAID 4 in RAID 5

- RAID 4: *Block-Interleaved Parity*
- RAID 5: *Block-Interleaved Distributed Parity*



# RAID 6

- Porazdeljen s porazdeljeno redundanco
- Enota porazdelitve je **blok**
- Redundanca je **paritetni bit** – uporablja dva paritetna bita, oba sta **porazdeljena**.
- Neobčutljiv na hkratno odpoved do dveh diskov.



# Gnezdena polja

- Primer: RAID 6 + 0

