Diski in datoteke

Iztok Savnik, FAMNIT

Prosojnice & učbenik

- Učbenik:
  - Raghu Ramakrishnan, Johannes Gehrke, *Database Management Systems, McGraw-Hill, 3rd ed., 2007.*
- *Prosojnice:*
  - *From „Cow Book“: R.Ramakrishnan, http://pages.cs.wisc.edu/~dbbook/*

Diski in datoteke

- SPUB shranjuje podatke na trdih diskih.
- To ima veliko posledic na zasnovo SUPB !
  - READ: prenos podatkov med trdim diskom in RAM.
  - WRITE: prenos med RAM na disk.
  - Obe vrsti operacij sta časovno potratni, zato se mora njihova uporaba planirati pazljivo!

Zakaj se vsega ne shrani v dinamični spomin?

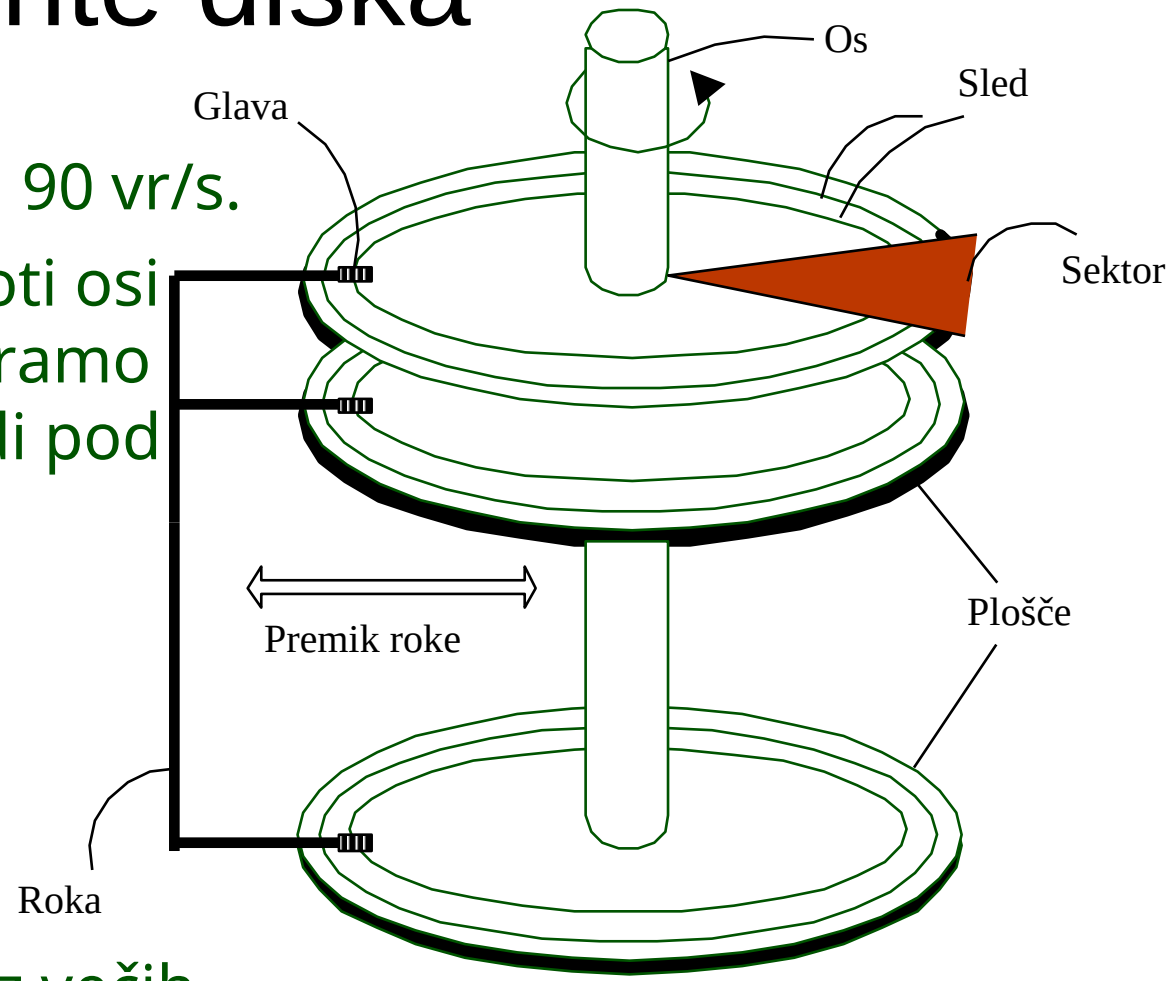
- *Prevelika cena*.
  - $100 ≈ 16GB RAM *ali* 4TB disk.
- *Dinamični spomin ne ohranja podatkov*. Želimo imeti shranjene podatke med večimi sejami s SUPB. (Očitno!)
- Tipična hierarhija pomnilnikov:
  - Dinamični spomin (RAM) za direktno delo s podatki.
  - Disk za glavno podatkovno bazo.
  - Trakovi za arhiviranje starejših verzij podatkov.

Diski

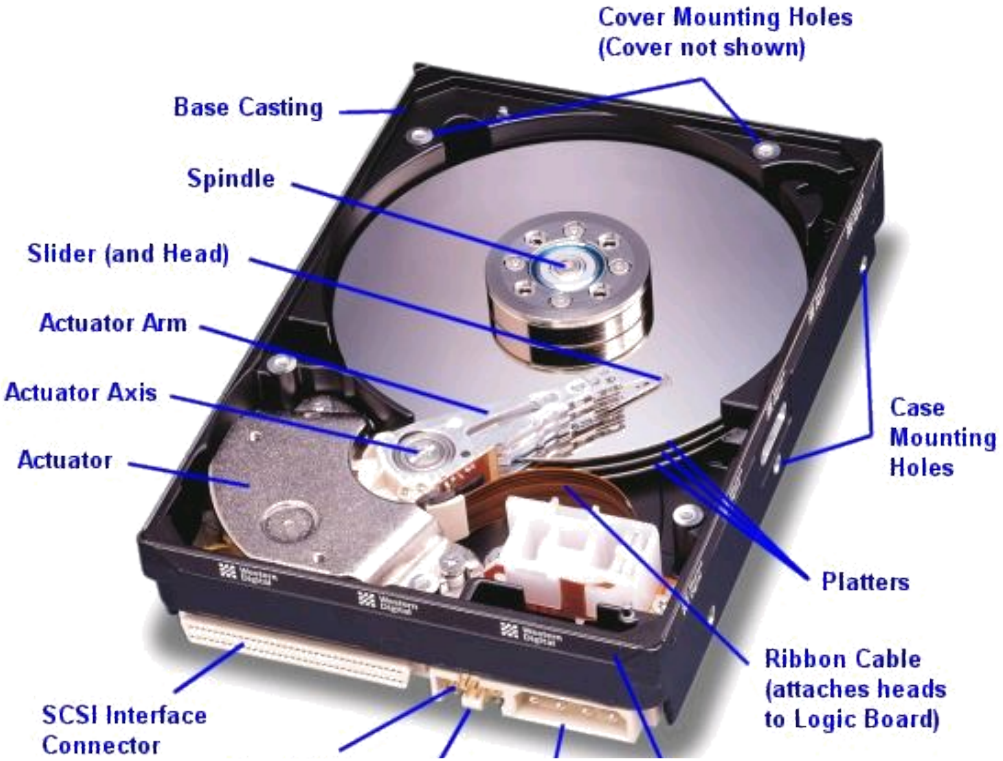
- Persistentno shranjevanje podatkov.
- Glavna prednost pred trakovi: *direkten dostop* vs. *sekvenčni*.
- Podatki se shranjujejo in prenašajo v enotah, ki jih imenujemo *diskovni bloki* ali *strani*.
- Čas potreben za dostop do podatkov je odvisen od lokacije podatkov na disku.
  - Položaj podatkov na disku ima vpliv na hitrost delovanja SUPB!

Komponente diska

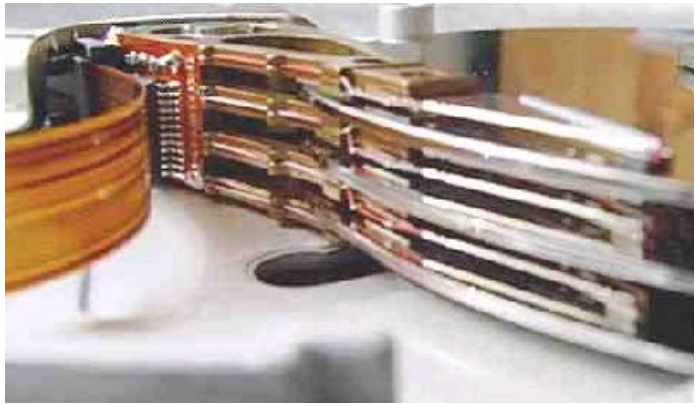
- Plošče se vrtijo npr. 90 vr/s.
- Roka se premika proti osi in navzven: pozicioniramo glavo nad sledmi. Sledi pod glavami imenujemo *cilinder.*
- Branje/pisanje poteka samo preko ene glave v danem trenutku.
- *Bloki* so sestavljeni iz večih *sektorjev* (ki so fiksni).



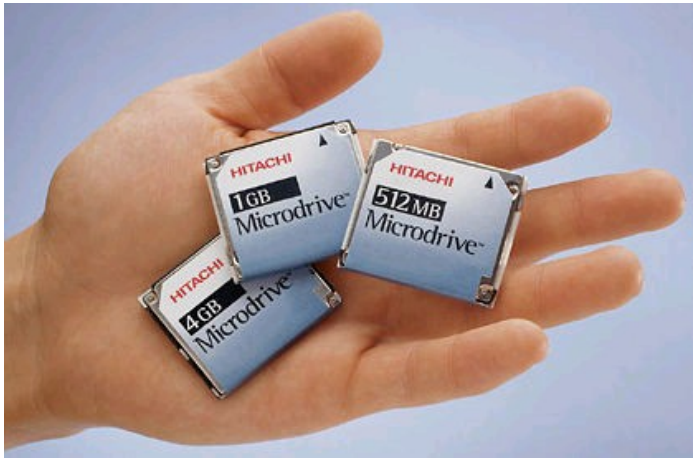
Trdi disk



Western Digital Drive
http://www.storagereview.com/guide/



Bralno/pisalna glava,
Stranski pogled



IBM/Hitachi Microdrive

IBM osebni računalnik/AT (1986)
- 30 MB trdi disk - $500
- 30-40ms čas dostopa
- 0.7-1 MB/s (est.)

Dostop do diskovne strani

- Čas dostopa (read/write) do bloka:
  - *čas iskanja* (premik roke/glave na željeno sled)
  - *rotacijska zakasnitev* (čakanje na blok pod glavo)
  - *čas prenosa* (prenos podatov iz/na površje plošče)
- Čas iskanja in rotacijska zakasnitev prevladujeta.
  - Čas iskanja se spreminja od 1 do 20msec (tipično 4-9ms)
  - Rotacijska zakasnitev se spreminja od 0 to 10msec (tipično 2ms)
  - Čas prenosa je tipično 1msec za 4KB stran
- Ključ za manjšo I/O ceno:
  - zmanjšati čas iskanja / rot. zakasnitev!
  - Hardware vs. software rešitev?

Barracuda®

The Power of One

Seagate®

Specifications	3TB[1]	2TB[1]	1.5TB[1]	1TB[1]	750GB[1]	500GB[1]	320GB[1]	250GB[1]
Model Number	ST3000DM001	ST2000DM001	ST1500DM003	ST1000DM003	ST750DM003	ST500DM002[2]	ST320DM000[2]	ST250DM000[2]
Interface Options	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ	SATA 6Gb/s NCQ
Performance								
Spindle Speed (RPM)	7200	7200	7200	7200	7200	7200	7200	7200
Cache, Multisegmented (MB)	64	64	64	64	64	16	16	16
SATA Transfer Rates Supported (Gb/s)	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5
Seek Average, Read (ms)	<8.5	<8.5	<8.5	<8.5	<8.5	<11	<11	<11
Seek Average, Write (ms)	<9.5	<9.5	<9.5	<9.5	<9.5	<12	<12	<12
Average Data Rate, Read/Write (MB/s)	156	156	156	156	156	125	125	125
Max Sustained Data Rate, OD Read (MB/s)	210	210	210	210	210	144	144	144
Configuration/Organization								
Heads/Disks	6/3	6/3	4/2	2/1	2/1	2/1	2/1	1/1
Bytes per Sector	4096	4096	4096	4096	4096	4096 or 512[2]	4096 or 512[2]	4096 or 512[2]

WD Red™ Pro

Specifications

Model Number[4]	WD221KFGX	WD201KFGX	WD181KFGX	WD161KFGX	WD141KFGX	WD121KFBX
Formatted capacity[1]	22TB	20TB	18TB	16TB	14TB	12TB
Recording technology	CMR	CMR	CMR	CMR	CMR	CMR
Interface	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s
Form factor	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch
Native command queuing	Yes	Yes	Yes	Yes	Yes	Yes
OptiNAND™ technology	Yes	Yes	No	No	No	No
Advanced Format (AF)	Yes	Yes	Yes	Yes	Yes	Yes
RoHS compliant[5]	Yes	Yes	Yes	Yes	Yes	Yes
Performance						
Interface speed (max)	6 Gb/s	6 Gb/s	6 Gb/s	6 Gb/s	6 Gb/s	6 Gb/s
Internal transfer rate[6]	265 MB/s	268 MB/s	272 MB/s	259 MB/s	255 MB/s	240 MB/s
Cache (MB)[1]	512	512	512	512	512	256
RPM	7200	7200	7200	7200	7200	7200
Reliability/Data Integrity						
Load/unload cycles[7]	600,000	600,000	600,000	600,000	600,000	600,000
Non-recoverable errors per bits read	<10 in 10^{14}	<10 in 10^{14}	<10 in 10^{14}	<10 in 10^{14}	<10 in 10^{14}	<10 in 10^{14}
MTBF (hours)[8]	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
Workload rate (TB/year)[2]	300	300	300	300	300	300
Limited warranty (years)[3]	5	5	5	5	5	5

- Kaj pa RAM?
  - Kakšna je razlika z HD?
- DDR4
  - 12-15ns latenca
  - 12-15 GB/s pretok
- DDR5
  - Enaka latenca (kot DDR4)
  - 38-50 GB/s pretok

CAS = Column address strobe

Standard name	Memory clock (MHz)	I/O bus clock (MHz)	Data rate (MT/s)	Module name	Peak transfer rate (MB/s)	Timings CL-tRCD-tRP	CAS latency (ns)
DDR4-1600J* DDR4-1600K DDR4-1600L	200	800	1600	PC4-12800	12800	10-10-10 11-11-11 12-12-12	12.5 13.75 15
DDR4-1866L* DDR4-1866M DDR4-1866N	233.33	933.33	1866.67	PC4-14900	14933.33	12-12-12 13-13-13 14-14-14	12.857 13.929 15
DDR4-2133N* DDR4-2133P DDR4-2133R	266.67	1066.67	2133.33	PC4-17000	17066.67	14-14-14 15-15-15 16-16-16	13.125 14.063 15
DDR4-2400P* DDR4-2400R DDR4-2400T DDR4-2400U	300	1200	2400	PC4-19200	19200	15-15-15 16-16-16 17-17-17 18-18-18	12.5 13.32 14.16 15
DDR4-2666T DDR4-2666U DDR4-2666V DDR4-2666W	333.33	1333.33	2666.67	PC4-21333	21333.33	17-17-17 18-18-18 19-19-19 20-20-20	12.75 13.50 14.25 15
DDR4-2933V DDR4-2933W DDR4-2933Y DDR4-2933AA	366.67	1466.67	2933.33	PC4-23466	23466.67	19-19-19 20-20-20 21-21-21 22-22-22	12.96 13.64 14.32 15
DDR4-3200W DDR4-3200AA DDR4-3200AC	400	1600	3200	PC4-25600	25600	20-20-20 22-22-22 24-24-24	12.5 13.75 15

Negibljivi disk (SSDs)

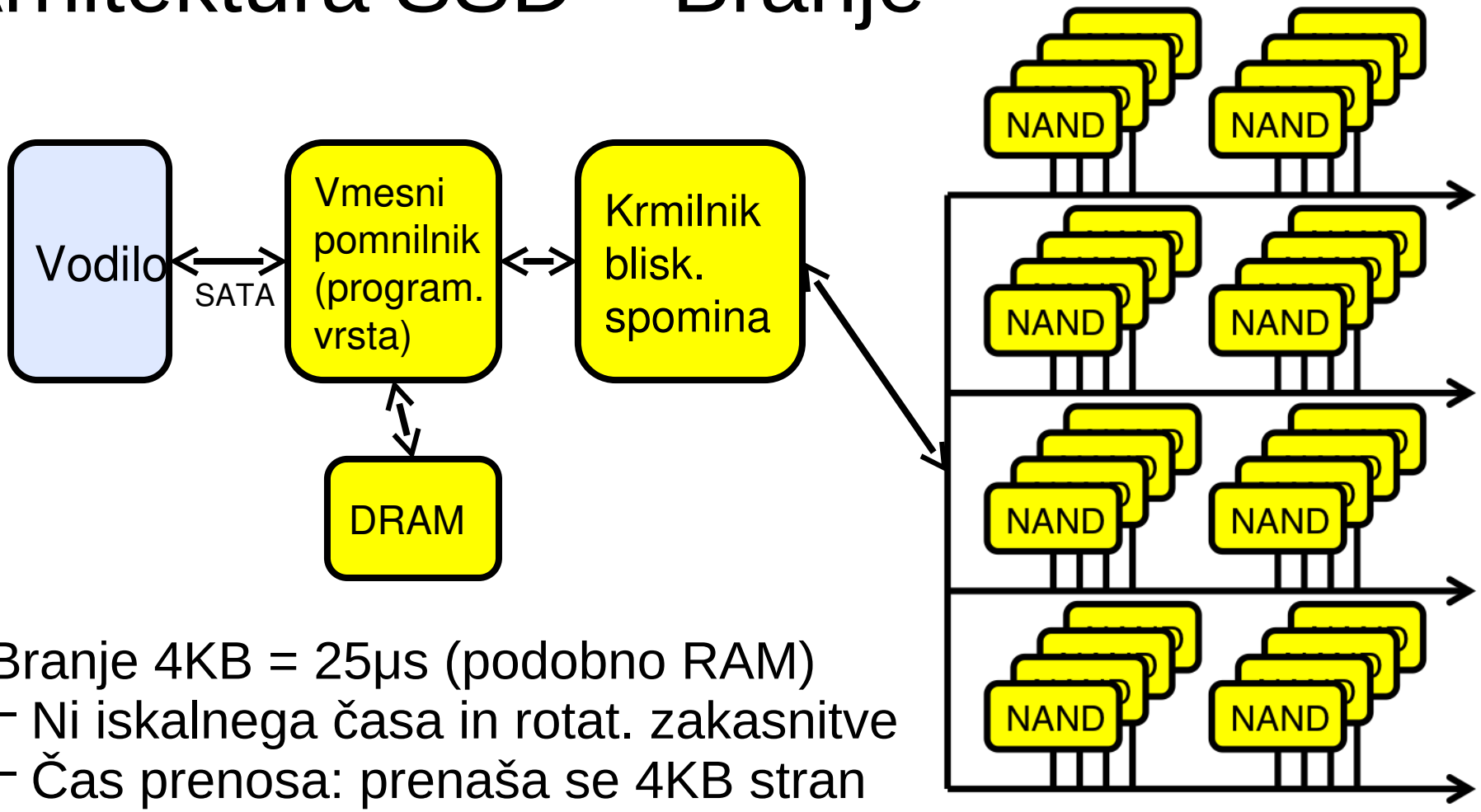
- Bliskovni pomnilnik
  - Elektronski obstojni pomnilnik je medij za shranjevanje podatkov, ki ga je možno elektronsko izbrisati in ponovno uporabiti za pisanje.
  - Dve osnovni vrsti bliskovnega spomina, NOR in NAND bliskovni spomin.
- Odkrit je bil v podjetju Toshiba leta 1980; osnovan je na EEPROM tehnologiji.
  - EPROM spomin je potrebno najprej kompletno izbrisati preden se lahko spet piše.
  - NAND bliskovni spomin se lahko briše in piše v blokih, ki so veliko manjši kot celotna naprava.
- Arhitektura NAND bliskovnega spomina
  - Hierarhična struktura: nizi, strani, bloki, plošče in čip.
    - Nis sestavlja 32-128 NAND celic
    - Še vedno je EEPROM: blok se najprej izbriše šele potem se lahko piše na strani!

Solid State Disk (SSD)

- 2009 – Bliskovni spomin (flash memory) iz večnivojskih NAND celic (2 bita/celico)
  - Naslovljiva stran (4 KB) se zlaga v bloke 4*64 strani
- Ni premičnih delov (ni motorjev za vrtenje/iskanje)
  - Eliminiran čas dostopa in rotacijsko zakasnitev (čas dostopa je 0.1-0.2ms)
  - Majhna poraba moči, lahek

Source: John Kubiatowicz, Lecture: Flash File Systems, UC Berkeley

Arhitektura SSD – Branje

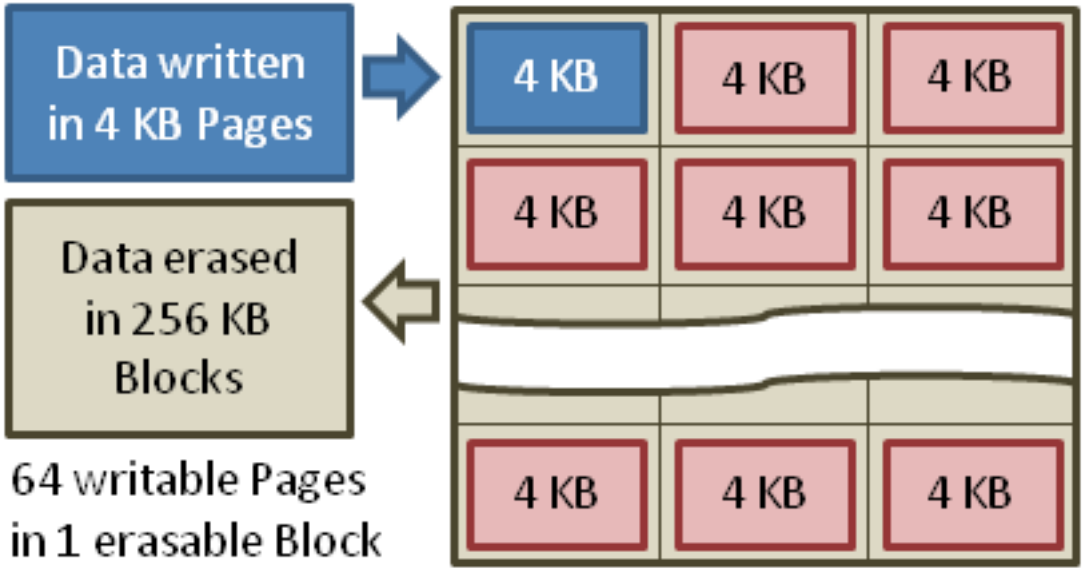


Branje 4KB = 25µs (podobno RAM)

- Ni iskalnega časa in rotat. zakasnitve
- Čas prenosa: prenaša se 4KB stran
  - SATA: 300-600MB/s => ~4 x10^3 B / 400 x 10^6 B/s => 10 µs
- Zakasnitev = Čas vrste + Čas krmilnika + Čas prenosa
- Največji pretok: Sekvenčno ali naključno branje!

Arhitektura SSD – Pisanje

- Pisanje podatkov je kompleksno! (~200µs – 1.7ms)
  - » Pišemo lahko samo na strani izbrisanega bloka
- Brisanje bloka ~1.5ms
- Krmilnik vzdržuje bazen prostih blokov; zlivanje neuporabljenih strani (read, erase, write); vzame nekaj % kapacitete.



Typical NAND Flash Pages and Blocks

Razvoj SSD

- Pomembne številke
  - Sekv. branje/pisanje
  - IOPS
  - Dostopni čas

SSD evolution

Parameter	Started with	Developed to	Improvement
Capacity	20 MB (Sandisk, 1991)	100 TB (Enterprise Nimbus Data DC100, 2018) (As of 2020 Up to 8 TB available for consumers)[16]	5-million-to-one[17] (400,000-to-one[17])
Sequential read speed	49.3 MB/s (Samsung MCAQE32G5APP-0XA, 2007)[18]	15 GB/s (Gigabyte demonstration, 2019) (As of 2020 up to 6.795 GB/s available for consumers)[19]	304.25-to-one[20] (138-to-one)[21]
Sequential write speed	80 MB/s (Samsung enterprise SSD, 2008)[22][23]	15.200 GB/s (Gigabyte demonstration, 2019) (As of 2020 up to 4.397 GB/s available for consumers)[19]	190-to-one[24] (55-to-one)[25]
IOPS	79 (Samsung MCAQE32G5APP-0XA, 2007)[18]	2,500,000 (Enterprise Micron X100, 2019) (As of 2020 up to 736,270 read IOPS and 702,210 write IOPS available for consumers)[19]	31,645.56-to-one[26] (Consumer: read IOPS: 9,319.87-to-one,[27] write IOPS: 8,888.73-to-one)[28]
Access time (in milliseconds, ms)	0.5 (Samsung MCAQE32G5APP-0XA, 2007)[18]	0.045 read, 0.013 write (lowest values, WD Black SN850 1TB, 2020)[29][19]	Read:11-to-one,[30] Write: 38-to-one[31]
Price	US$50,000 per gigabyte (Sandisk, 1991)[32]	US$0.10 per gigabyte (Crucial MX500, July 2020)[33]	555,555-to-one[34]

Seagate Nytro SSD

Specifications	Nytro 5550H 15 mm — Mixed Use		
Capacity	6.4TB	3.2TB	1.6TB
Standard Model[1]	XP6400LE70005	XP3200LE70005	XP1600LE70005
SED Model[1]	XP6400LE70015	XP3200LE70015	XP1600LE70015
FIPS 140-3/Common Criteria Model[1]	XP6400LE70025	XP3200LE70025	XP1600LE70025
Features			
Interface	PCIe® Gen4 ×4 NVMe	PCIe® Gen4 ×4 NVMe	PCIe® Gen4 ×4 NVMe
NAND Flash Type	3D eTLC	3D eTLC	3D eTLC
Form Factor	2.5 in × 15mm	2.5 in × 15mm	2.5 in × 15mm
Performance			
Sequential Read (MB/s) Sustained, 128 KB[2]	7,400	7,400	7,400
Sequential Write (MB/s) Sustained, 128 KB[2]	7,200	6,900	4,300
Random Read (IOPS) Sustained, 4 KB QD64[3]	1,700,000	1,700,000	1,700,000
Random Write (IOPS) Sustained, 4 KB QD64[3]	470,000	470,000	315,000
Average Read Latency (µs), 4 KB QD1	75	75	75
Average Write Latency (µs), 4 KB QD1	12	12	12

Nimbus ExaDrive DC

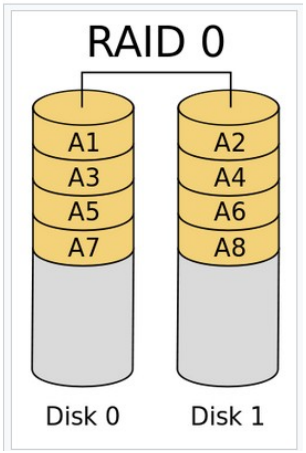
	EDDCT016	EDDCT032	EDDCT050	EDDCT100	EDDCS016	EDDCS032	EDDCS050	EDDCS100
Basics								
Capacity	16 TB	32 TB	50 TB	100 TB	16 TB	32 TB	50 TB	100 TB
Interface	SATA-3 (6.0 Gbps)				SAS-2 dual-port (for HA)			
Form Factor	3.5" (LFF)							
Reliability								
Endurance	Unlimited DWPD for 5 years							
MTBF (hours)	2.5 million hours							
Limited Warranty	5 years							
Performance								
Latency	0.1 ms	0.1 ms	0.1 ms	0.05 ms	0.2 ms	0.2 ms	0.2 ms	0.15 ms
Random Read (4 KB)	97K IOps	97K IOps	97K IOps	114K IOps	50K IOps	50K IOps	50K IOps	52K IOps
Random Write (4 KB)	91K IOps	91K IOps	91K IOps	106K IOps	25K IOps	25K IOps	25K IOps	26K IOps
Sequential Read	500 MBps	500 MBps	500 MBps	500 MBps	450 MBps	450 MBps	450 MBps	450 MBps
Sequential Write	460 MBps	460 MBps	460 MBps	460 MBps	260 MBps	260 MBps	260 MBps	260 MBps
Power								
Active Read Power	12.1 W	12.2 W	12.1 W	15.2 W	12.1 W	12.2 W	12.1 W	15.2 W
Active Write Power	13.1 W	13.2 W	13.8 W	16.8 W	13.1 W	13.2 W	13.8 W	16.8 W
Idle Power	6.8 W	7.2 W	7.2 W	11.1 W	7.0 W	7.4 W	7.4 W	11.3 W
Active Read Power / TB	0.76 W	0.38 W	0.24 W	0.15 W	0.76 W	0.38 W	0.24 W	0.15 W
Active Write Power / TB	0.82 W	0.41 W	0.28 W	0.17 W	0.82 W	0.41 W	0.28 W	0.17 W
Idle Power / TB	0.43 W	0.23 W	0.14 W	0.11 W	0.44 W	0.23 W	0.14 W	0.11 W

RAID

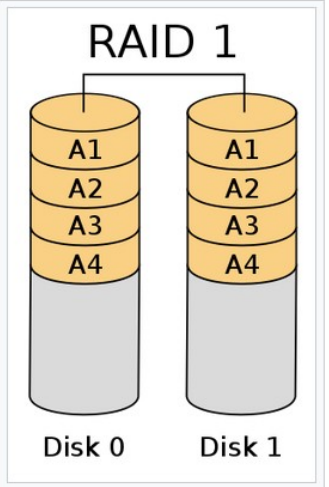
- Diskovno polje (disk array).
  - Več diskov urejenih v polje, ki omogoča abstrakcijo: polje je videti kot en sam velik disk.
- Cilji: povečati performanse in zanesljivost.
- Dve osnovni tehniki:
  - Podatkovni pasovi (strips): Podatki so razdeljeni po diskih po pasovih; velikost pasa se imenuje “pasovna enota”.
  - Redundanca: Več diskov => več napak. Redundantna informacija omogoča rekonstrukcijo podatkov, če se disk pokvari.

RAID Nivoji

- Nivo 0: Ni redundance.
  - Porazdeljeni podatki po pasovih
  - Ni redundance, paritete
  - Hitrost dostopa je glavni razlog

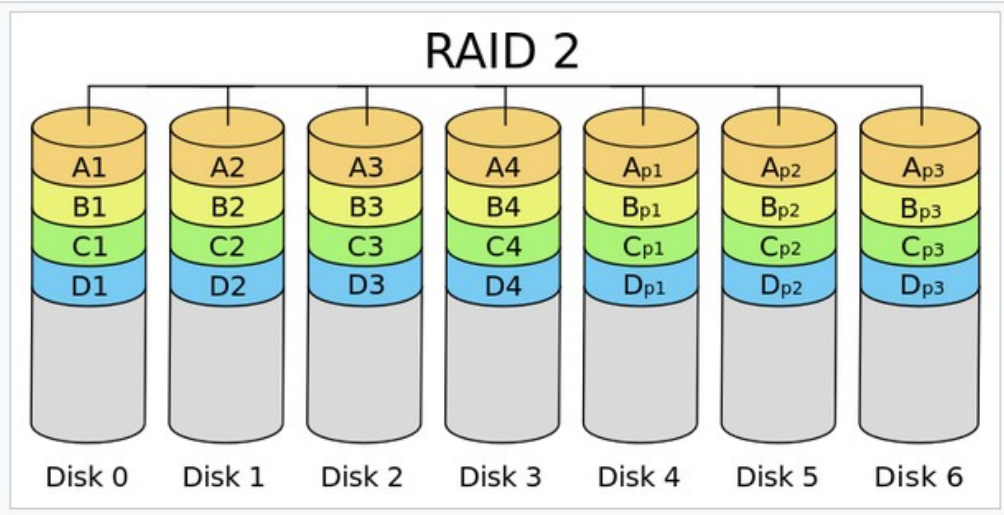


- Nivo 1: Zrcaljenje (dve identični kopiji).
  - Vsak disk ima zrcalno kopijo.
  - Paralelno branje; pisanje deluje nad dvemi diski.
  - Pohitritev branja = št. zrcaljenih diskov X

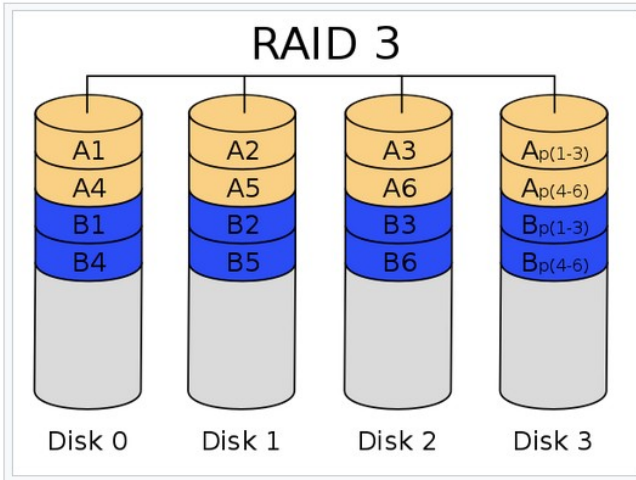


RAID Nivoji

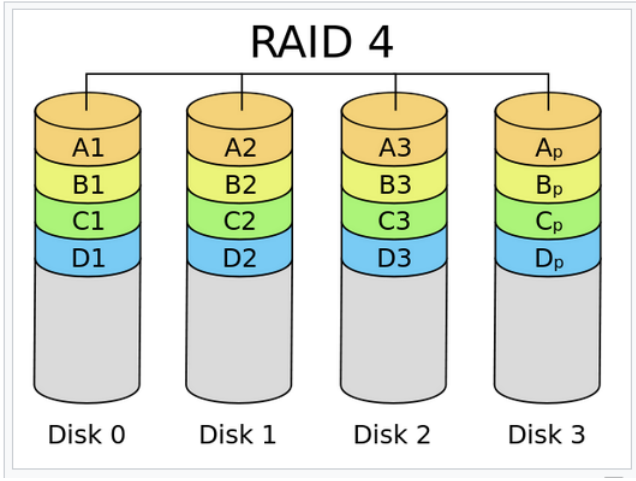
- Nivo 2: Pasovi in Hammingova koda.
  - Podatki se po bitih razporedijo po diskih.
  - Vzporedno branje; pisanje deluje nad več diski.
    - Maksimalen prenos = skupen pretok.
    - Redko se uporablja



RAID Nivoji



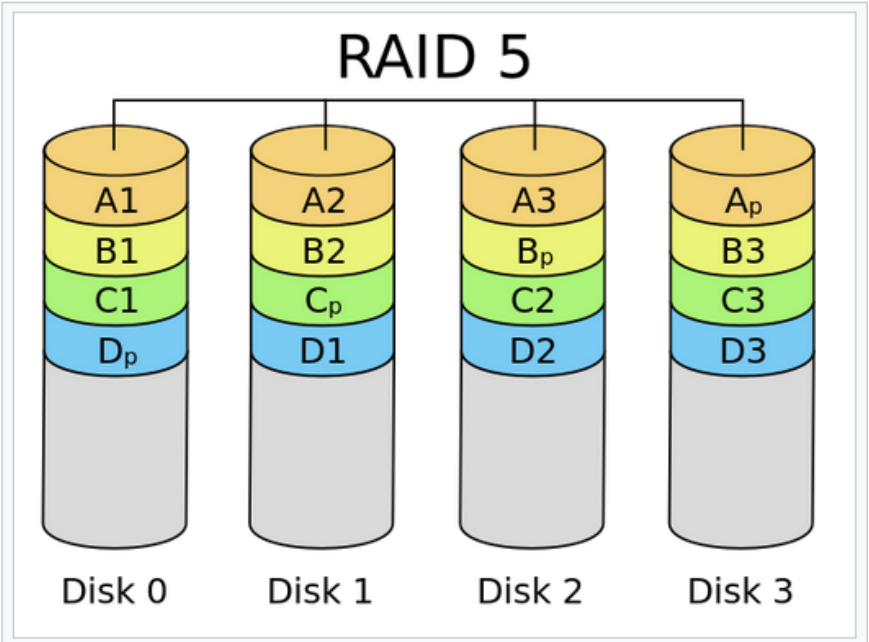
- Nivo 3: Pariteta “po bitih”.
  - Pasovna enota: en zlog (byte)
  - En disk za pariteto
  - Vsako pisanje/branje vključuje vse diske
  - Diskovno polje lahko procesira eno zahtevo v danem trenutku.



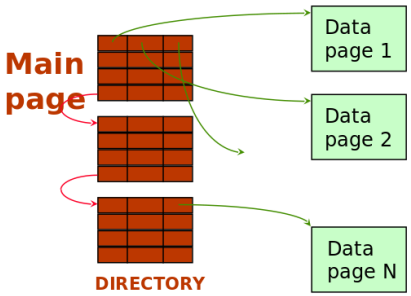
- Nivo 4: Pariteta “po blokih”.
  - Pasovna enota: En diskovni blok
  - Eden disk za pariteto
  - Paralelno branje za manjše zahteve, večje zahteve lahko uporabijo poln pretok.
  - Pisanje vključuje spremenjen blok in disk za preverjanje.

RAID Nivoji

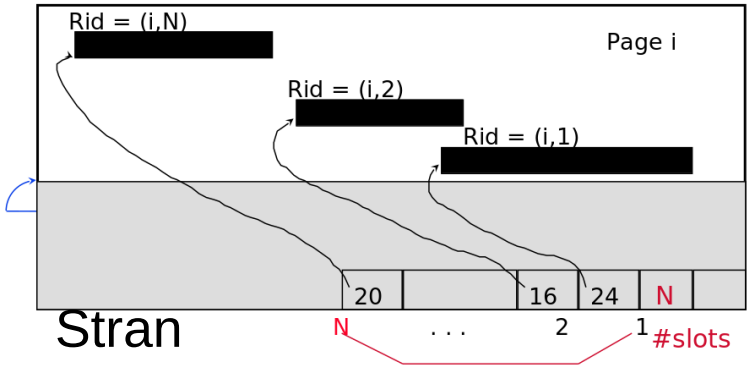
- Nivo 5: Porazdeljena pariteta “po blokih”
  - Podobno RAID 4; paritetni blok je porazdeljen po večih diskih



Pomnilnik SUPB

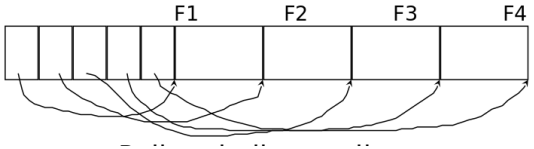


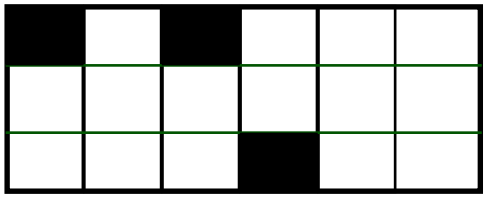
Datoteka



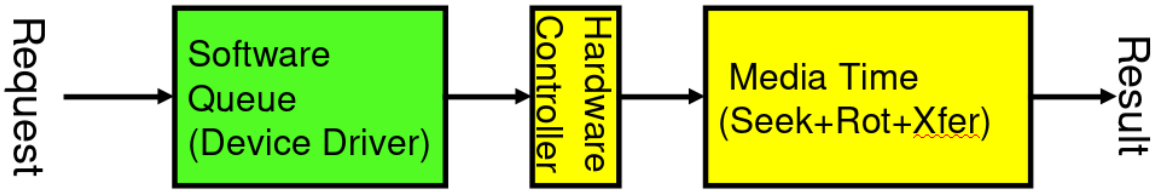
Stran

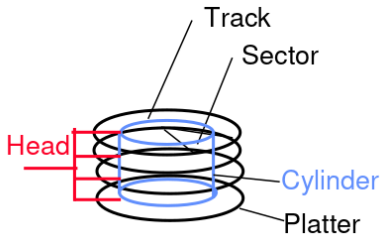
Zapis





Vmesni pomnilnik
Bazen strani





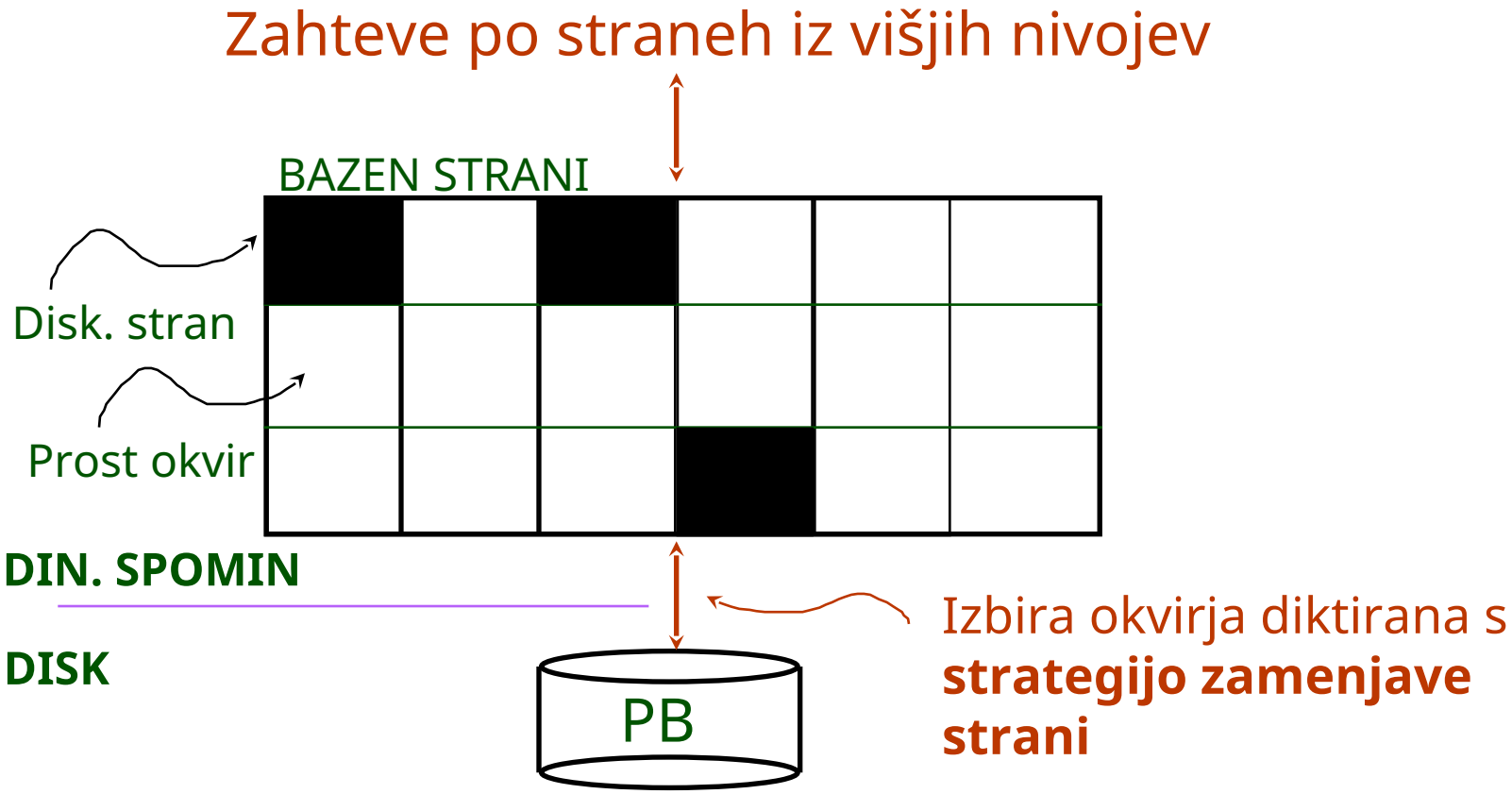
Ureditev blokov na disku

- Koncept: `*naslednja*' stran:
  - Bloki na isti sledi, ki sledijo
  - Bloki na istem cilindru, ki sledijo
  - Bloki na sosednem cilindru.
- Bloki ene datoteke naj bi bili organizirani sekvenčno na disku (glede na `naslednji'), da se minimizira čas iskanja in rotacijska zakasnitev.
- Sevenčno skeniranje: *branje vnaprej* tj. več sekvenčnih strani se prebere vnaprej; pridobitev na času!

Delo z diskovnim prostorom

- Najnižji nivo SUPB dela z diskovnim pomnilniškim prostorom.
- Višji nivoji zahtevajo od tega nivoja:
  - alokacijo/de-alokacijo strani
  - branje /pisanje strani
- Zahteve po *sekvenci* strani mora biti izvršena tako, da sistem alocira strani v enem sekvenčnem prostoru na disku!
  - Višji nivoji ne potrebujejo vedeti kako je to narejeno oz. kako se dela s prostim prostorom.

Vmesni pomnilnik v SUPB



- *Pod. morajo biti v RAM da lahko SUPB dela z njimi!*
- Uporablja se tabela parov: ***<okvir#, stranid>***

Zahteva po strani ...

- Če iskana stran ni v bazenu:
  - Izberi okvir za *zamenjavo*
  - Če je okvir “umazan” potem se mora zapisati na disk
  - Preberi izbrano stran v izbrani okvir
- *Pripni* stran in vrni njen naslov.

☛ *Če lahko predvidevamo zahteve (npr., sekv. pregled) potem lahko vnaprej preberemo več strani !*

Več o delu z vmes. pomnilnikom

- Zahtevana stran mora biti *odpeta.*
- Vidno mora biti, če je stran spremenjena ali *“umazana”*:
  - *Umazan* bit
- Stran v bazenu je lahko zahtevana večkrat,
  - Uporablja se *števec pripenjanj*. Stran je kandidat za zamenjavo če je *št. pripenjanj*= 0.
- Kontrola vzporednosti in okrevanje lahko povzroči dodatni I/O v primeru, da je stran izbrana za zamenjavo.
  - Protokol *Dnevnika;*
  - *Pisanje-vnaprej*; več kasneje.

Strategije za zamenjavo strani

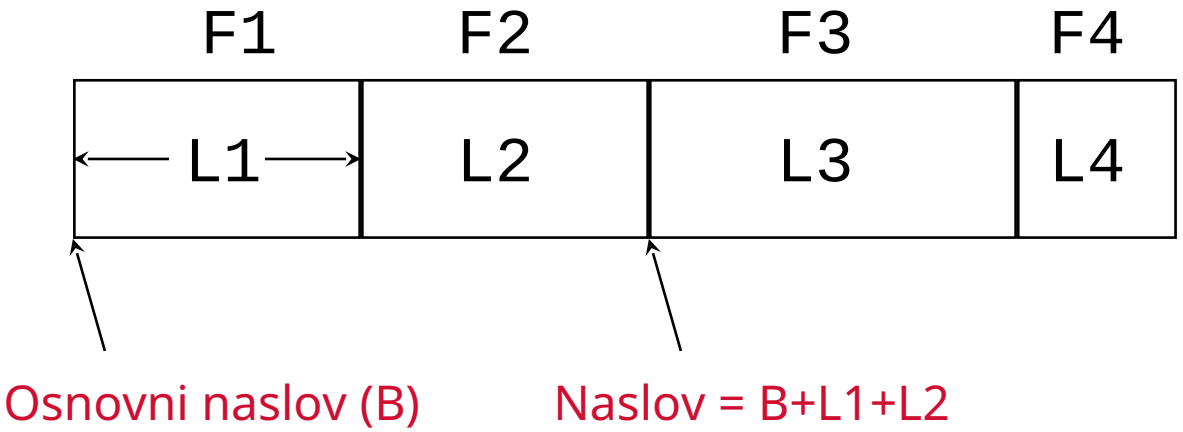
- Okvir izbran s *strategijo zamenjave:*
  - LRU, Ura (*Clock*), MRU, itd.
- Strategija ima lahko velik vpliv na # I/O operacij; odvisno od *vzorca dostopa*.
- *Sekvenčno prelivanje*:
  - Grda situacija povzročena z LRU + ponavljajoč sekvenčni pregled tabele.
  - # okvirjev < # strani datoteke vsaka zahteva povroči I/O. MRU je v tem primeru boljša (toda ne v vseh situacijah, seveda).

SUPB vs. OS Datotečni sistem

OS ureja diskovni prostor & vmesni pomnilnik: zakaj nebi OS za SUPB urejal ta opravila?

- Razlike v OS podpori: prenosljivost
- Nekatere omejitve, npr., datoteke se ne morajo raztezati preko velikosti diska.
- Delo z vmesnim pomnilnikom v SUPB zahteva zmožnost:
  - Pripeti stran v vmesni pomnilnik, prisiliti shranjevanje strani na disk (pomembno za SD & recovery),
  - Prilagoditev *strategije zamenjave strani in* branje strani vnaprej na osnovi obnašanja tipične operacije.

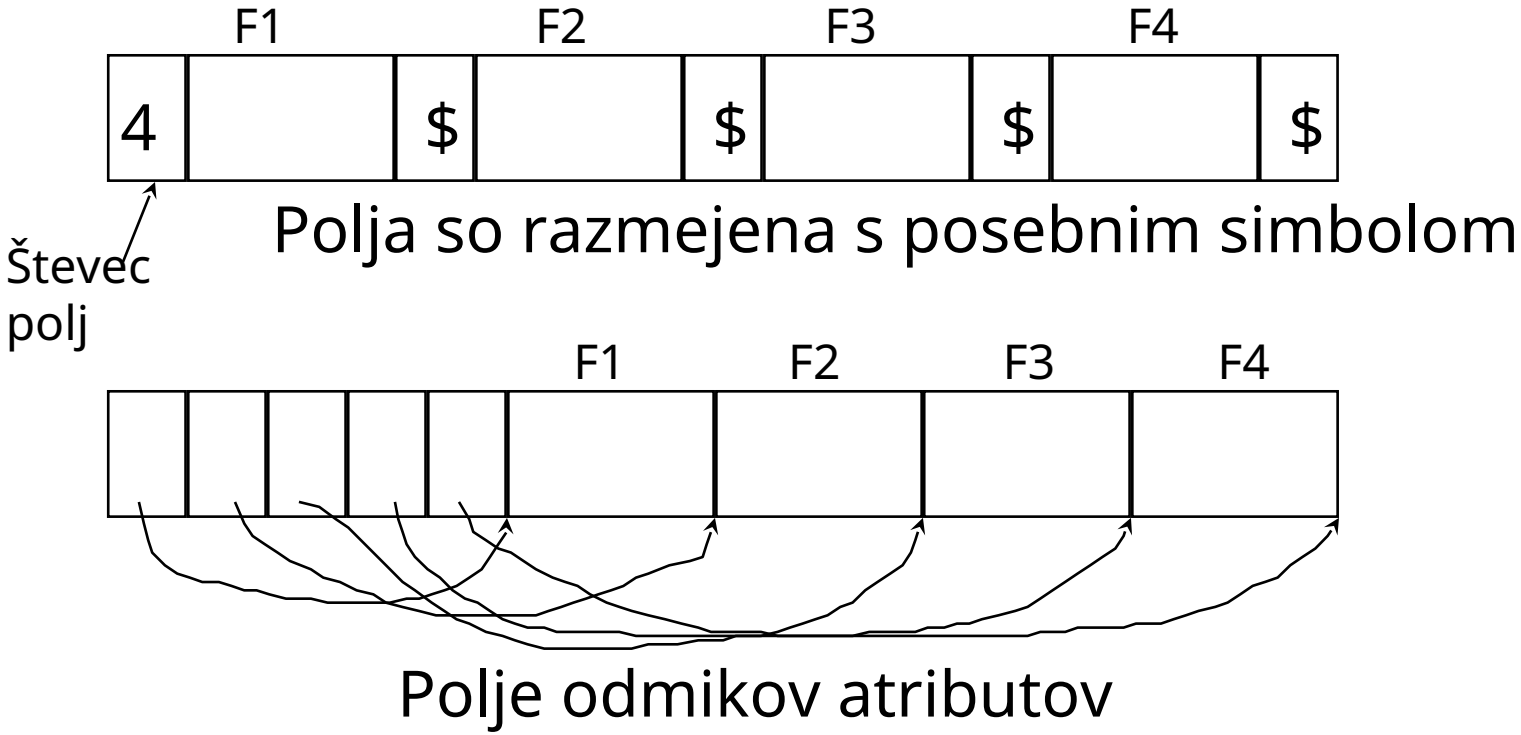
Format zapisa: Fiksna dolžina



- Informacija o tipih polja je enaka za vse zapise v datoteki; shranjuje se v *sistemskem katalogu.*
- Poišči *i-to* polje ne zahteva pregled vseh zapisov.

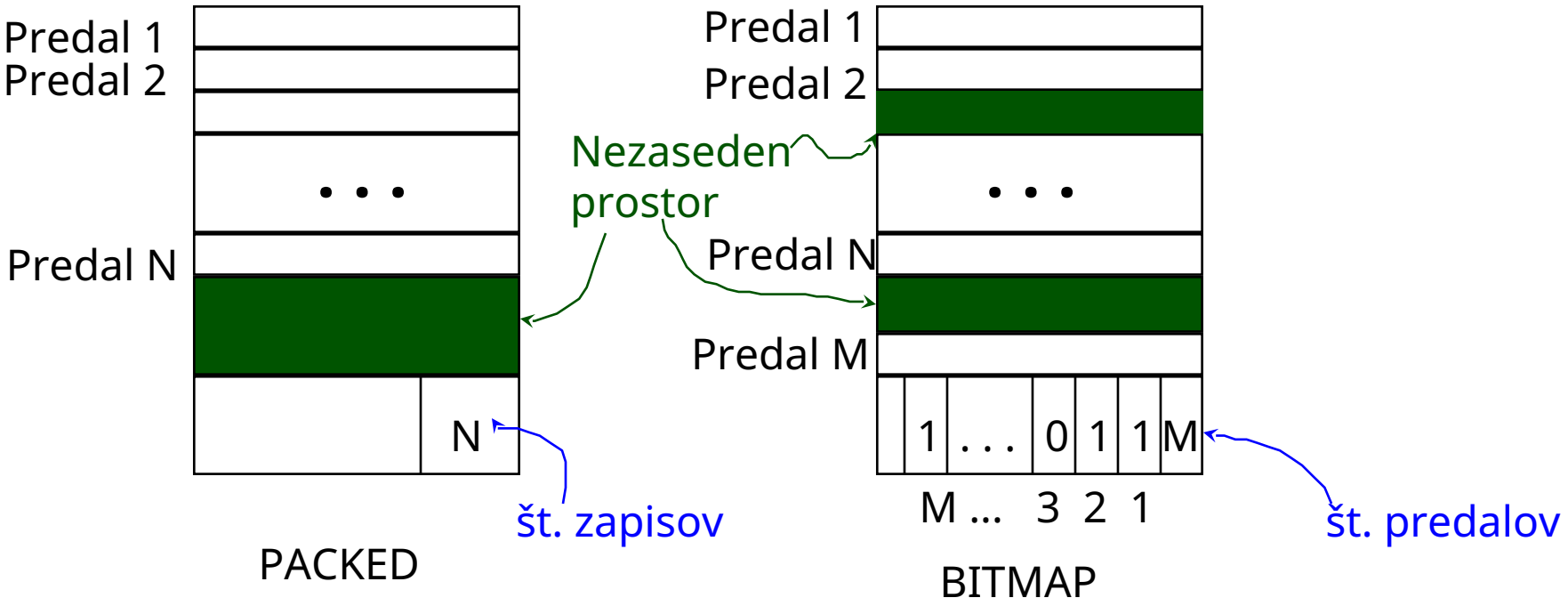
Format zapisa: Variabilna dolžina

- Dva alternativna formata (# polj je fiksno):



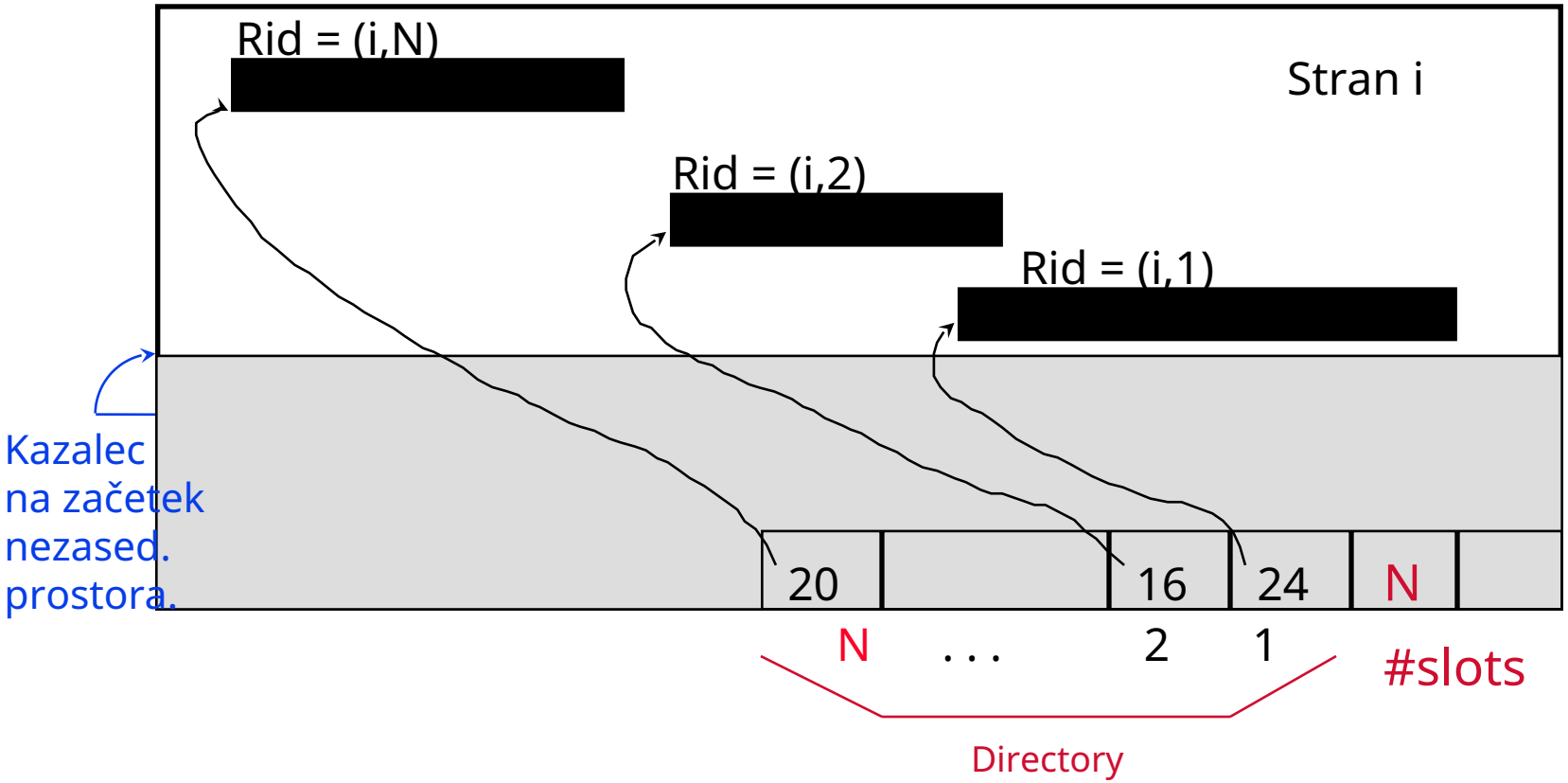
☛ Druga alternativa ponuja direkten dostop do i-tega polj; učinkovito shranjevanje *null*; dir. ne zaseda veliko prostora.

Format strani: Zapisi fiksne dolžine



- *Zapis id = <stran id, predal #>. V prvi alternativi zahteva premikanje zapisov za prazen prostor spremembo rid; ni vedno sprejemljivo.*

Format strani: Zapisi variabilne dolžine



- *Lahko premikamo zapise po strani ne da bi spremenili rid; privlačna predstavitev tudi za zapise s fiksno dolžino.*

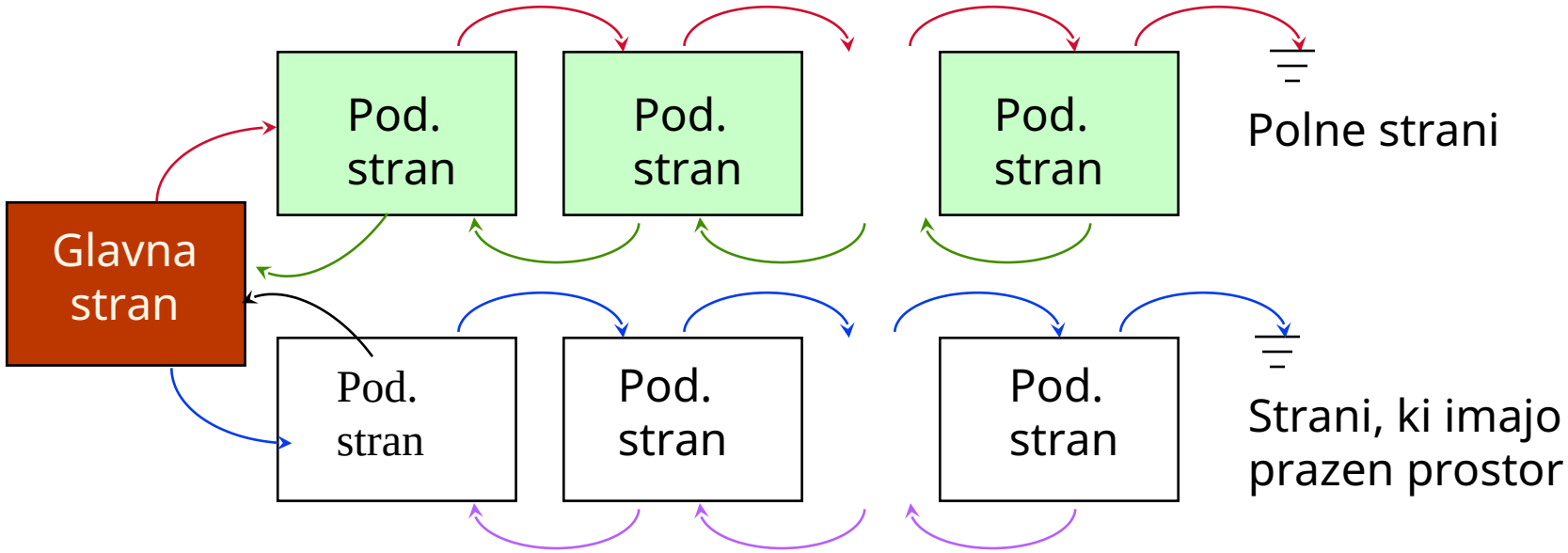
Datoteke zapisov

- Stran ali blok je OK za I/O
- Višji nivoji SUPB delujejo z *zapisi* ter z *datotekami zapisov*
- DATOTEKA: Zbirka strani; vsaka stran vsebuje množico zapisov. Operacije:
  - insert/delete/modify - zapis
  - read – zapis (uporaba *rid*)
  - pregled (scan) vseh zapisov (pogoji nad zapisi, ki naj jih sistem vrne)

Neurejene datoteke

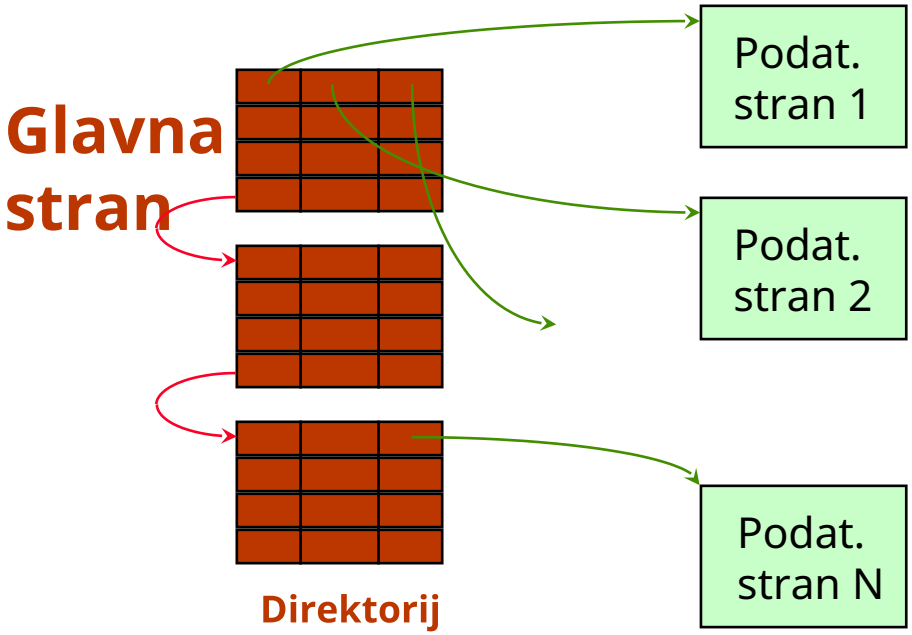
- Najenostavnejša datotečna struktura
  - Zapisi nimajo nobene urejenosti.
  - Tako kot se datoteka manjša in veča, tako se dodajajo in odvzemajo strani (bloki).
- Potrebno je shranjevati podatke o:
  - *straneh* datoteke
    - *neuporabljenem prostoru* na straneh
    - *zapisih* na strani
- Obstaja veliko alternativ za shranjevanje predstavljenih podatkov.

Neurejena datoteka implementirana s seznamom



- ID glavne strani in sama datoteka morata biti shranjena nekje na disku.
- Vsaka stran vsebuje 2 `kazalca' in podatke.

Neurejena datoteka z direktorijem strani



- Zapis pod. strani na glavni strani vsebuje lahko tudi št. prostih zlogov na pod. strani.
- Direktorij je zbirka strani; implementacija seznamov je ena alternativa.
  - *Seznam glavnih strani je precej manjši kot št. vseh strani!*

Sistemski katalogi

- Za vsak indeks:
  - struktura (npr., B+ drevo) in iskalni ključi
- Za vsako relacijo:
  - ime, ime datoteke, datotečna struktura
  - imena in tipi atributov
  - imena indeksov
  - integritetne omejitve
- Za vsak pogled:
  - ime pogleda in definicija
- Statistika, avtorizacija, velikost bazena strani, itd.

☛ *Katalogi so shranjeni kot relacije*!

Attr_Cat(attr_name, rel_name, type, position)

attr_name	rel_name	type	position
attr_name	Attribute_Cat	string	1
rel_name	Attribute_Cat	string	2
type	Attribute_Cat	string	3
position	Attribute_Cat	integer	4
sid	Studenti	string	1
ime	Studenti	string	2
login	Studenti	string	3
star	Studenti	integer	4
ocena	Studenti	real	5
fid	Fakulteta	string	1
fime	Fakulteta	string	2
placa	Fakulteta	real	3

Povzetek

- Disk je poceni persistentni pomnilniški medij.
  - Direkten dostop, hitrost je odvisna od lokacije strani; pomembno je urediti podatke na disku sekvenčno; minimizirata se *čas iskanja* in *zakasnitev rotacije*.
- Vmesni pomn.: stran prenešena iz diska v RAM.
  - Stran ostane v RAM tako dolgo, dokler je ne sprosti tisti, ki jo je zahteval.
  - Zapisana je na disk, ko je izbrana za zamenjavo (po tem, ko je sproščena).
  - Izbira okvirja za zamenjavo se izvede na osnovi *strategije za zamenjavo.*
  - Poskuša se prenesti več strani za kasnejšo uporabo.

Povzetek

- SUPB vs. OS Datotečni sistem
  - SUPB potrebuje lastnosti, ki jih OS ne nudi. Na primer, eksplicitno zapisovanje strani na disk, kontroliranje vrstnega reda strani na disku, datoteke preko diskov, možnost vnaprejšnjega branja večje količine strani, prilagajanje dela vmesnega pomnilnika na vzorce dostopa, itd.
- Zapisi variabilne dolžine s specificiranimi odmiki polj v direktoriju nudi direkten dostop do i-tega polja in omogoča null vrednosti.
- Format strani s predali omogoča shranjevanje zapisov variabline dolžine in omogoča premikanje zapisov po strani.

Povzetek

- Nivo datoteke omogoča delo s stranmi datoteke in nudi abstrakcijo “zbirke zapisov”.
  - Strani s praznim prostorom se identificirajo z dvojno povezanim seznamom ali preko direktorija.
- Indeksi omogočajo učinkovit dostop do zapisov na osnovi vrednosti nekega atributa.
- Katalog shranjuje podatke o relacijah, indeksih in pogledih.