Operativni sistemi 1

Sekundarne memorije

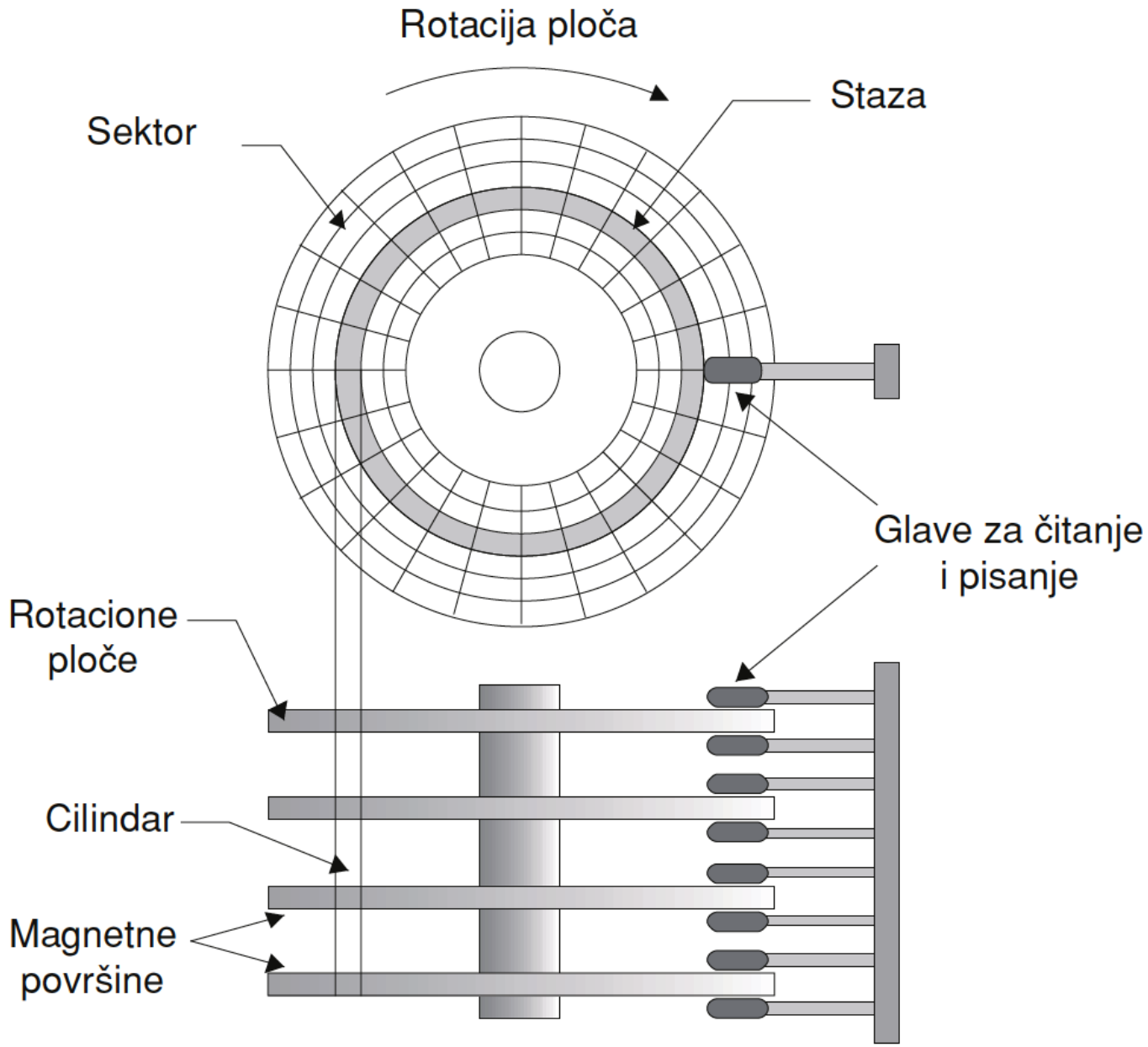
Nemanja Maček

Sadržaj

- Struktura diskova
- Priprema diskova za rad
- Nivoi keširanja diskova
- Raspoređivanje zahteva za rad sa diskom
- RAID strukture - realizacija stabilnih sistema
- Priključivanje diskova

Šta je hard disk a šta disk kontroler?

- Skup **rotacionih kružnih ploča** koje rotiraju oko zajedničke ose.
- Površine ploča su presvučene **magnetnim materijalom**.
- Svaka površina ima pridruženu **glavu za čitanje i pisanje**.
  - Čitaju ili upisuju podatke sa magnetnih ploča.
  - Linearno se pokreću pomoću sopstvenog servo-sistema.
  - Na taj način im je uz rotaciju ploča omogućen pristup svim delovima magnetne površine.
- Računar i disk komuniciraju putem **disk kontrolera** (engl. *disk controller*).
  - Disk kontroleri pružaju **interfejs ka ostatku računara**.
    - Računar ne mora da zna način rada niti da kontoliše elektro-mehaniku diska.
    - Dodatne funkcije kontrolera:
      - baferovanje podataka koje treba upisati na disk,
      - keširanje diskova,
      - automatsko obeležavanje neispravnih sektora diska.


Rotacija ploča
Staza
Sektor
Glave za čitanje
i pisanje
Rotacione
ploče
Cilindar
Magnetne
površine

Geometrija diskova

- Površina diska je podeljena u koncentrične prstenove – **staze** (engl. *tracks*).
- Svaka staza je dalje podeljena na **sektore** (engl. *sectors*).
- Tipična količina podataka koja se može upisati u jedan sektor je **512 bajtova**.
  - To je najmanja količina podataka koja se može upisati na disk ili pročitati sa diska!
- Sve površine magnetnih ploča su jednako podeljene na staze i sektore.
  - To znači da se glave za čitanje i pisanje na svim pločama diska u jednom vremenskom trenutku nalaze na istim stazama.
- Ekvidistantne staze svih ploča čine jedan **cilindar** (engl. *cylinder*).
- Datoteke koje nisu smeštene u okviru jednog cilindra su **fragmentisane**.
  - Pomeranje glava sa jedne staze na drugu prilikom čitanja ovakvih datoteka unosi **kašnjenje**.
  - Performanse diska se mogu uvećati smeštanjem datoteke **u okviru jednog cilindra** kad god je to moguće.

Geometrija diskova

- Geometrija diska je u opštem slučaju određena:
  - brojem magnetnih površina (odnosno glava za čitanje i pisanje),
  - brojem cilindara,
  - brojem sektora.
- **Trodimenzionalnim adresiranjem** (engl. *head-cylinder-sector addressing*) može se pristupi svim delovima diska.
- Primer:
  - Podatak koji je upisan na drugu površinu, u stazu 3, na sektoru 5.
  - (*head*, *cylinder*, *sector*)=(2,3,5).

Smernice razvoja savremenih disk uređaja

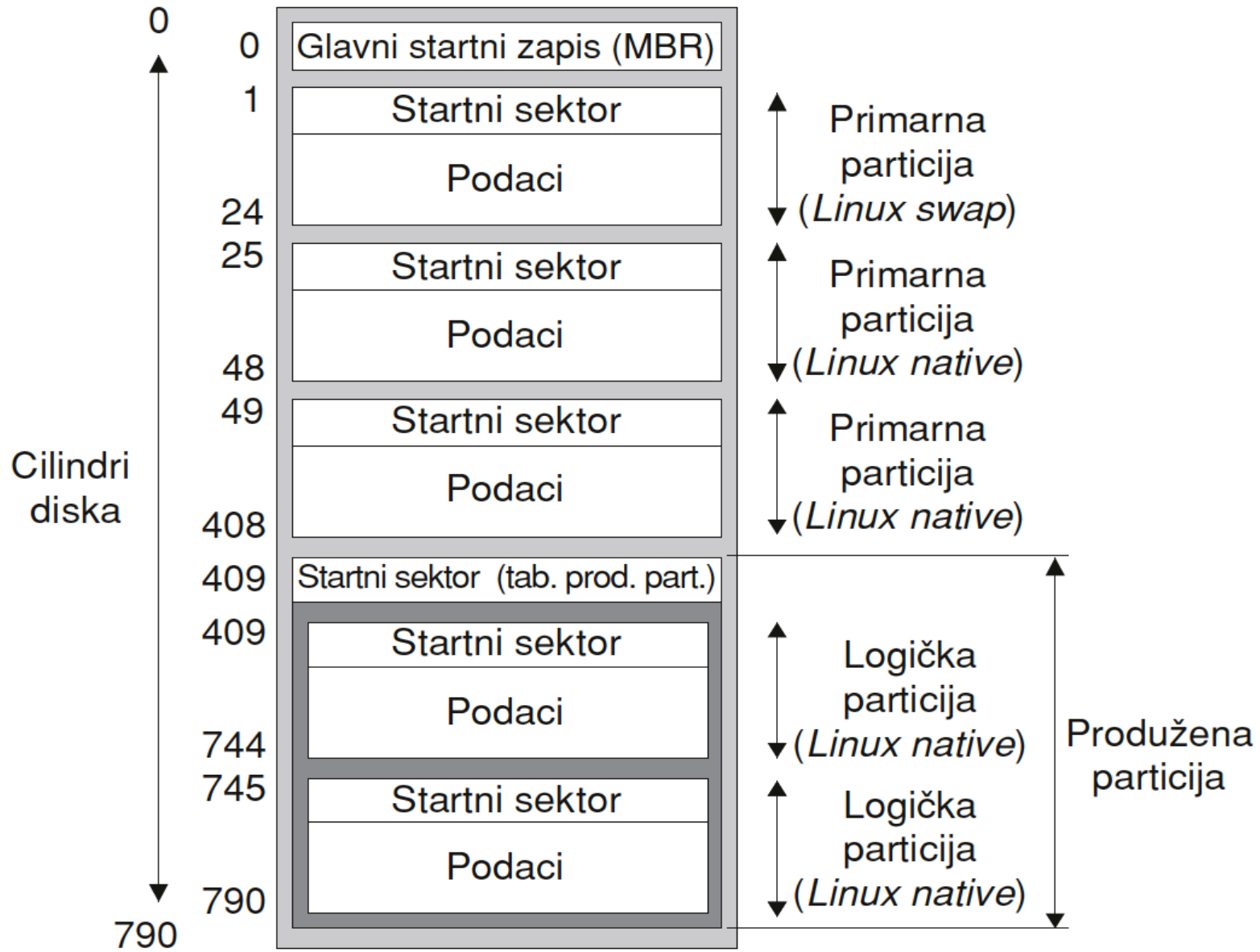
- Smanjivanje vremena poziociniranja (engl. *seek time*).
- Povećanje rotacione brzine.
- Povećanje gustine magnetnog mediuma.
- Zonska tehnika (engl. *Bit Zone Recording*).
  - Unutrašnje staze (bliže centru) imaju manju površinu pa samim tim i manje magneta.
  - Cilindri se grupišu u zone iste gustine.
    - Povećava se kapacitet diska i brzina čitanja sa medijuma.
    - Upravljanje promenljivom gustinom staza je složenije.
- Rezervni regioni na mediumu (engl. *spare regions*) za zamenu defektnih blokova.
- Upotreba procesora solidne snage na disku.
- Keširanje na disk uređaju (engl. *On-Board Cache*).
- Logičko-fizičko mapiranje blokova na disku (LBN $\rightarrow$ PBN).
  - *Bit Zone Recording* i upravljanje defektima utiču na složenost mapiranja.
- …

IDE, SATA, SCSI

- **IDE** (engl. *Integrated Drive Electronics*) tj. **ATA** (engl. *Advanced Technology Attachment*).
  - Kontroler integrisan na matičnoj ploči.
  - /dev/hda, /dev/hdb, …
  - Dva kanala: *primary* i *secondary*.
  - Na svaki se mogu vezati do dva uređaja u odnosu *master-slave*.
- **SATA** (engl. *Serial* ATA)
  - Kontroler integrisan na matičnoj ploči.
  - /dev/sda, /dev/sdb, …
- **SCSI**
  - Kontroler NIJE integrisan na matičnoj ploči.
  - Na kontroler je moguće vezati od 7 do 15 uređaja.
  - SCSI uređaji se ne nalaze u master-slave odnosu već se vezuju prema prioritetima.
    - Prioritet uređaja određen je njegovim ID koji se postavlja preko džampera.
    - ID=0 (najviši prioritet), ID=15 (najniži prioriter), ID=7 (rezervisan za SCSI kontroler).

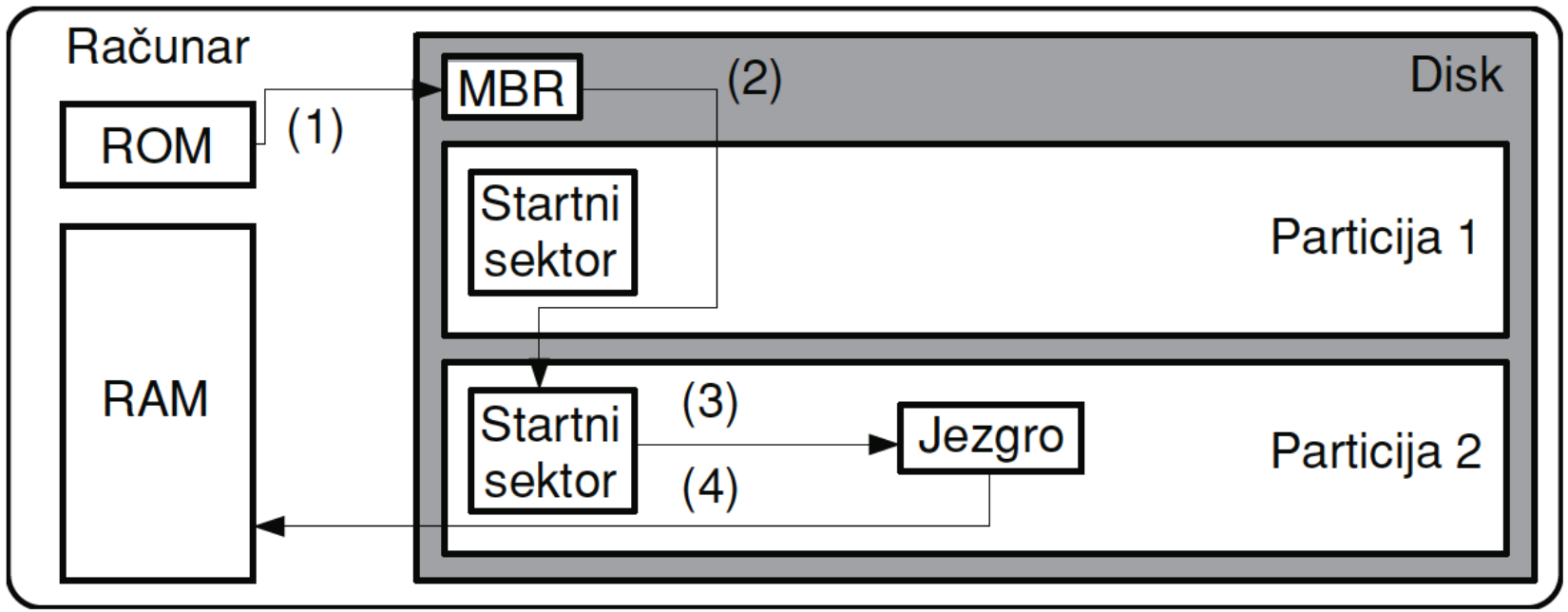
Priprema diskova za rad

- **Formatiranje diskova niskog nivoa** (stvar prošlosti).
- Kreiranje particija.
  - **MBR** (engl. *Master Boot Record*).
    - Najviše četiri particije po jednom disku.
    - Primarne particije (*primary*).
    - Produžena particija *(*extended) – okvir u kome se mogu kreirati nekoliko logičkih particija.
    - Logičke particije se ponašaju kao primarne, ali se razlikuju po načinu kreiranja.
  - **GPT** (engl. *GUID Partition Table*).
    - *Unified Extensible Firmware Interface* (UEFI) standard.
    - 64 bita za logičke adrese.
    - Najveća veličina diska je 2^{64} sektora.
    - Ako je sector 512 bajta, najveća veličina diska je 2^{64} x 2^{9} bajta = 9.4ZB (9.4×10^{21} bajta).
- **Kreiranje sistema datoteka**.


0
0
Glavni startni zapis (MBR)
1
Startni sektor
Podaci
24
25
Startni sektor
Podaci
48
49
Startni sektor
Podaci
408
409
Startni sektor (tab. prod. part.)
409
Startni sektor
Podaci
744
745
Startni sektor
Podaci
790
790
Cilindri diska
Primarna particija (Linux swap)
Primarna particija (Linux native)
Primarna particija (Linux native)
Logička particija (Linux native)
Logička particija (Linux native)
Produžena particija

Bootstrap rutina

- Kada se računar uključi BIOS izvršava **POST rutinu** (engl. *Power On Self Test*).
- POST = serija testova hardvera.
- **Podizanje sistema** (engl. *boot*) je procedura koja se izvršava u cilju dovođenja sistema u operativno stanje.
- Primer (MBR):
  - Kod upisan u prvom MBR najpre identifikuje **aktivnu particiju** u particionoj tabeli.
  - Zatim se izvršava **kod upisan u boot sektoru** aktivne particije.
  - Program u boot sektoru je zadužen da **pokrene punjenje RAM memorije OS-om**.
  - Napomena:
    - Delovi koda u toj fazi nalaze se na fiksnim područjima diska, a ne u sistemima datoteka.
    - Zašto?
      - U toj fazi nema kernela, pa nemamo podršku za sistem datoteka.
  - Rana faza podizanja operativnog sistema se završava učitavanjem jezgra.


Računar
ROM
(1)
RAM
MBR
(2)
Disk
Startni sektor
Particija 1
Startni sektor
(3)
(4)
Jezgro
Particija 2

Nivoi keširanja diskova

- **Keširanje na nivou OS-a** (engl. *built-in **file** caching*).
- **Keširanje na nivou disk kontrolera** (engl. *HBA level caching*).
  - Ne koristi se za klasične disk kontrolere.
  - Nezaobilazni deo u najkvalitetnijim i najsloženijim disk kontrolerima.
- **Keširanje na nivou disk uređaja** (engl. *disk drive level caching*).
  - Diskovi imaju solidne procesore i veliku količinu memorije koja služi za keširanje.
  - Disk keš memorija optimalno je mesto za tehniku predikovanog čitanja!
    - Disk najbolje poznanje svoj servo-sistem i raspored podataka na medium!
- **RAID keširanje** (engl. *caching in RAID*).
  - Svaki RAID kontroler pored RAID funkcionalnosti predstavlja i potpuni keš kontroler.
- **Keširanje na nivou aplikacije** (engl. *application level caching*).
  - Svi pomenuti keš nivoi su po prirodi opšte namene, generalni.
  - Zahvaljući dobrom poznavanju sopstvenih potreba u radu sa diskom, kvalitena aplikacija kešira disk saglasno svojim potrebama (znatno bolje nego generalni keš na nivou OS-a).

Raspoređivanje zahteva za rad sa diskom

- **Vreme pristupa kod diska** zavisi od:
  - **Vremena pozicioniranja glava** sa tekuće pozicije na zahtevani cilindar.
  - **Vremena rotacionog kašnjenja** (zavisi od brzine okretanja rotacionih površina diska).
  - **Brzine transfera podataka** sa magnetnog medijuma (zavisi od gustine medijuma i brzine okretanja rotacionih površina diska).
- **Brzina disk transfera** (engl. *bandwidth*): količnik ukupnog broja prenetih bajtova i ukupnog vremena.
- U višeprocesnoj okolini u jednom trenutku postoji veliki broj zahteva za rad sa diskom.
- Pravilnim **raspređivanjem ovih zahteva** (engl. *disk scheduling*) ukupno vreme pozicioniranja ili rotacionog kašnjenja se može smanjiti.
- Napomena:
  - Relativne performance algoritama izrazićemo ukupnim brojem cilindara koje glave za čitanje i pisanje prelaze pri opsluživanju zahteva.

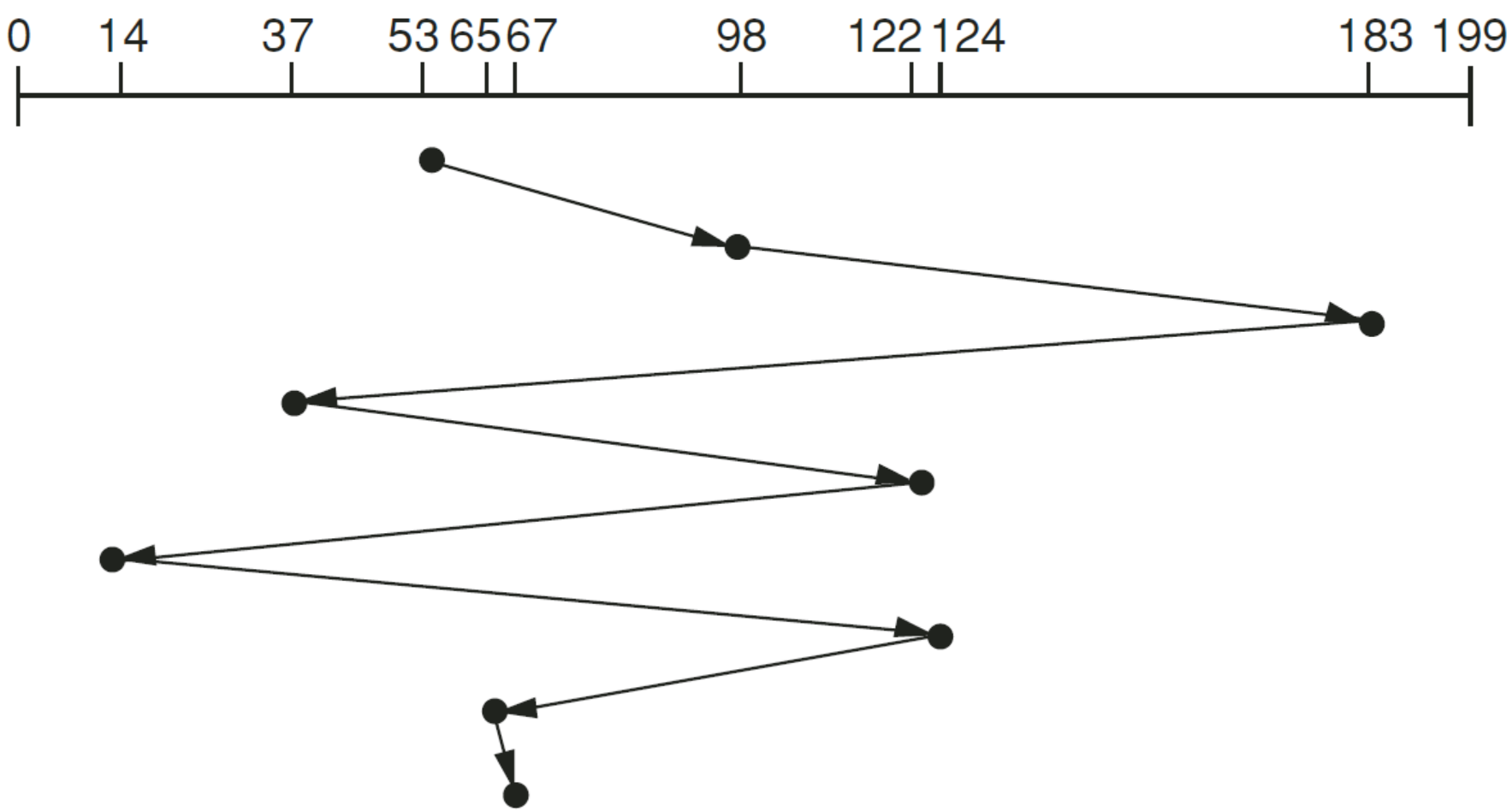
Raspređivanje zahteva za rad sa diskom

- Svaki zahtev koji je upućen disku sadrži **sledeće informacije**:
  - da li se zahteva operacija čitanja ili pisanja,
  - adresu bloka na disku,
  - adesu memorijskog bafera,
  - broj bajtova koje treba preneti.
- Više zahteva može stići istovremeno.
  - U jednom trenutku disk može obraditi samo jedan.
- Postoje više algoritama koji će obaviti raspoređivanje zahteva za rad sa diskom.
  - FCFS,
  - SSTF,
  - SCAN,
  - C-SCAN,
  - LOOK,
  - C-LOOK.

Raspređivanje zahteva za rad sa diskom

- **Algoritam FCFS** (engl. *first come, first served*).
- Najjednostavniji algoritam.
- Zahteve prosleđuje u onom redosledu u kome su stigli.
- Obezbeđuje krajnje fer odnose prema prispelim zahtevima, ali i znatno loše performanse.
- Primer:
  - Trenutna pozicija glava za čitanje i pisanje je na cilindru 53.
  - U red čekanja za disk zahtevi pristižu po sledećem redu: 98, 183, 37, 122, 14, 124, 65, 67.
  - Ukupni pomeraj glava diska iznosi 640 cilindara.

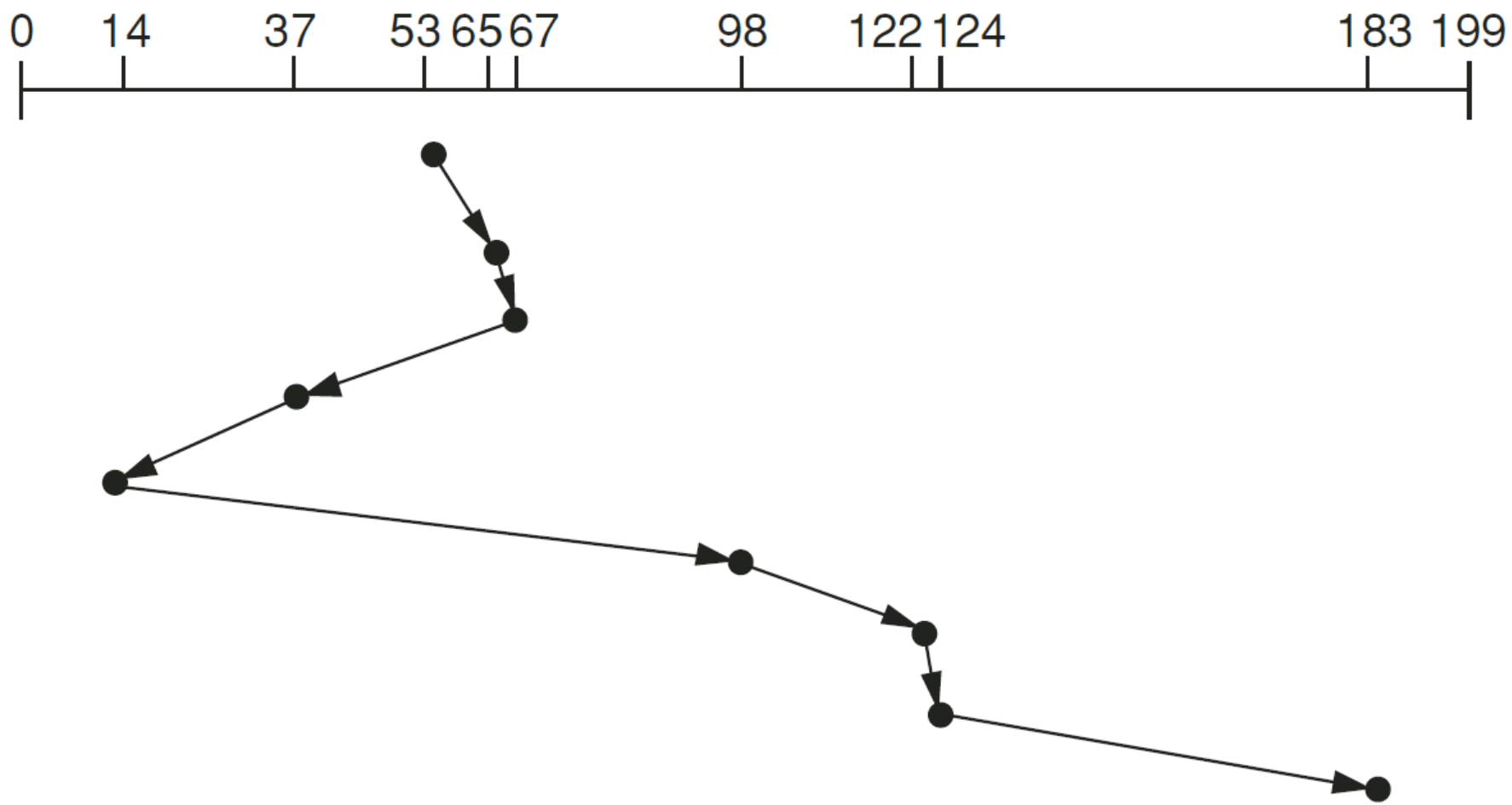
Trenutna pozicija glave: cilindar 53
Red čekanja: 98, 183, 37, 122, 14, 124, 65, 67



Raspređivanje zahteva za rad sa diskom

- **Algoritam SSTF** (engl. *shortest seek time first*).
- Od prispelih zahteva najpre se uzima onaj koji će izazvati **najmanji pomeraj glava** (engl. *seek time*).
- Primer:
  - Trenutna pozicija glava za čitanje i pisanje je na cilindru 53.
  - U red čekanja za disk zahtevi pristižu po sledećem redu: 98, 183, 37, 122, 14, 124, 65, 67.
  - Ukupni pomeraj glava diska iznosi 236 cilindara.
- Napomene:
  - Podseća na SJF (engl. *shortest job first*) algoritam za raspoređivanje procesora.
  - Optimalan je po pitanju **vremena pozicioniranja**.
  - Kod SSTF algoritma prisutan je problem **zakucavanja** (engl. *starvation*).
    - Glave mogu ostati veoma dugo u jednoj zoni opslužujući zahteve koji unose male pomeraje.
    - Zahtevi čiju su cilindri daleko od trenutne pozicije mogu dugo čekati u redu.

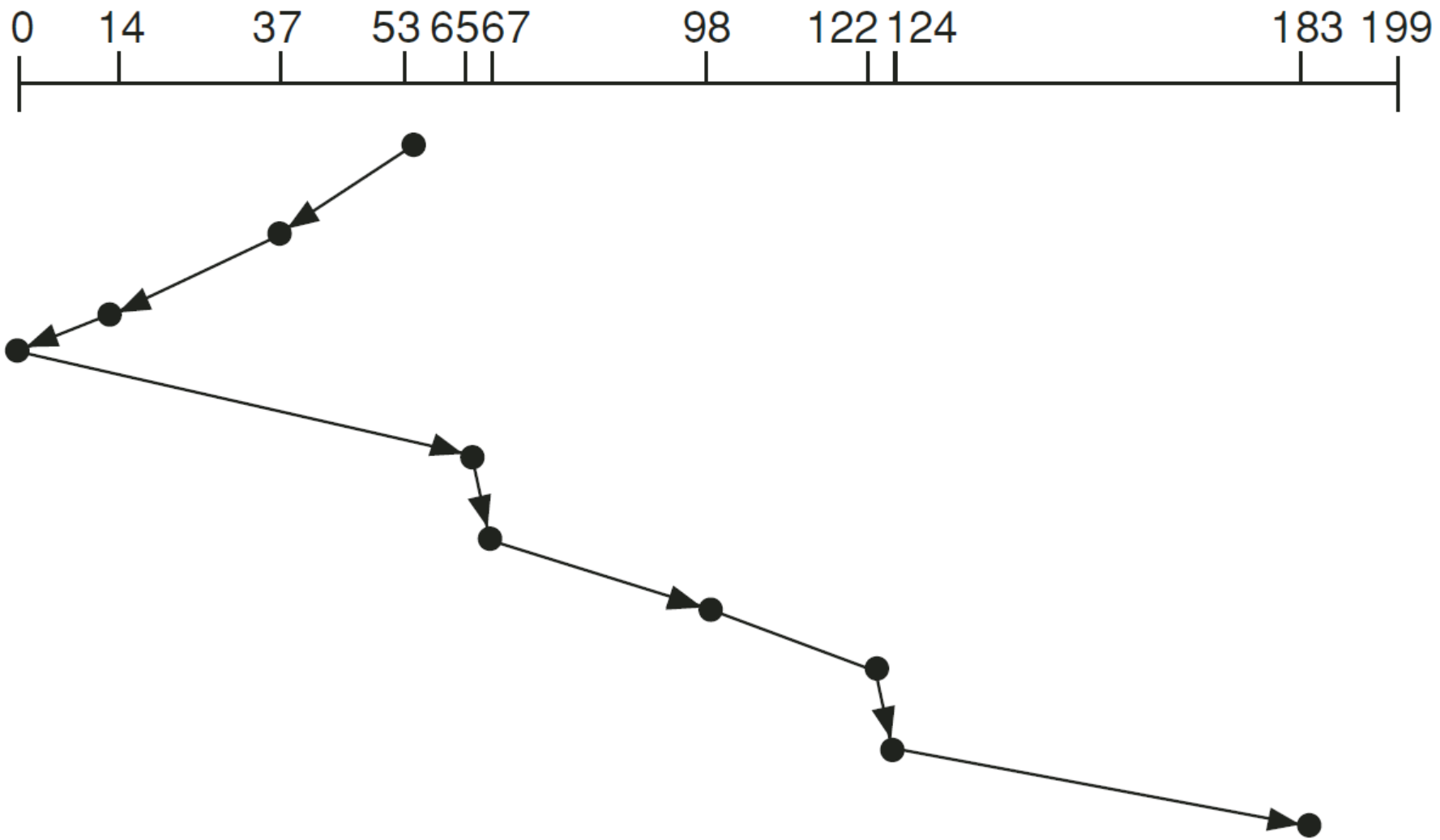
Trenutna pozicija glave: cilindar 53
Red čekanja: 98, 183, 37, 122, 14, 124, 65, 67



Raspređivanje zahteva za rad sa diskom

- **Algoritam SCAN**.
- Radi na principu lifta koji se naizmenično kreće od prizemlja do vrha zgrade.
- Algoritam naizmenično pomera glave od početka do kraja diska i unazad i opslužuje zahteve koji se nalaze na tekućem cilindru.
- Na ovaj način se rešava problem zakucavanja.
- Primer:
  - Trenutna pozicija glava za čitanje i pisanje je na cilindru 53.
  - U red čekanja za disk zahtevi pristižu po sledećem redu: 98, 183, 37, 122, 14, 124, 65, 67.
  - Ukupni pomeraj glava diska iznosi 208 cilindara.
- Napomene:
  - Prilikom obrade zahteva, SCAN daje prednost unutrašnjim cilindrima u odnosu na periferne.

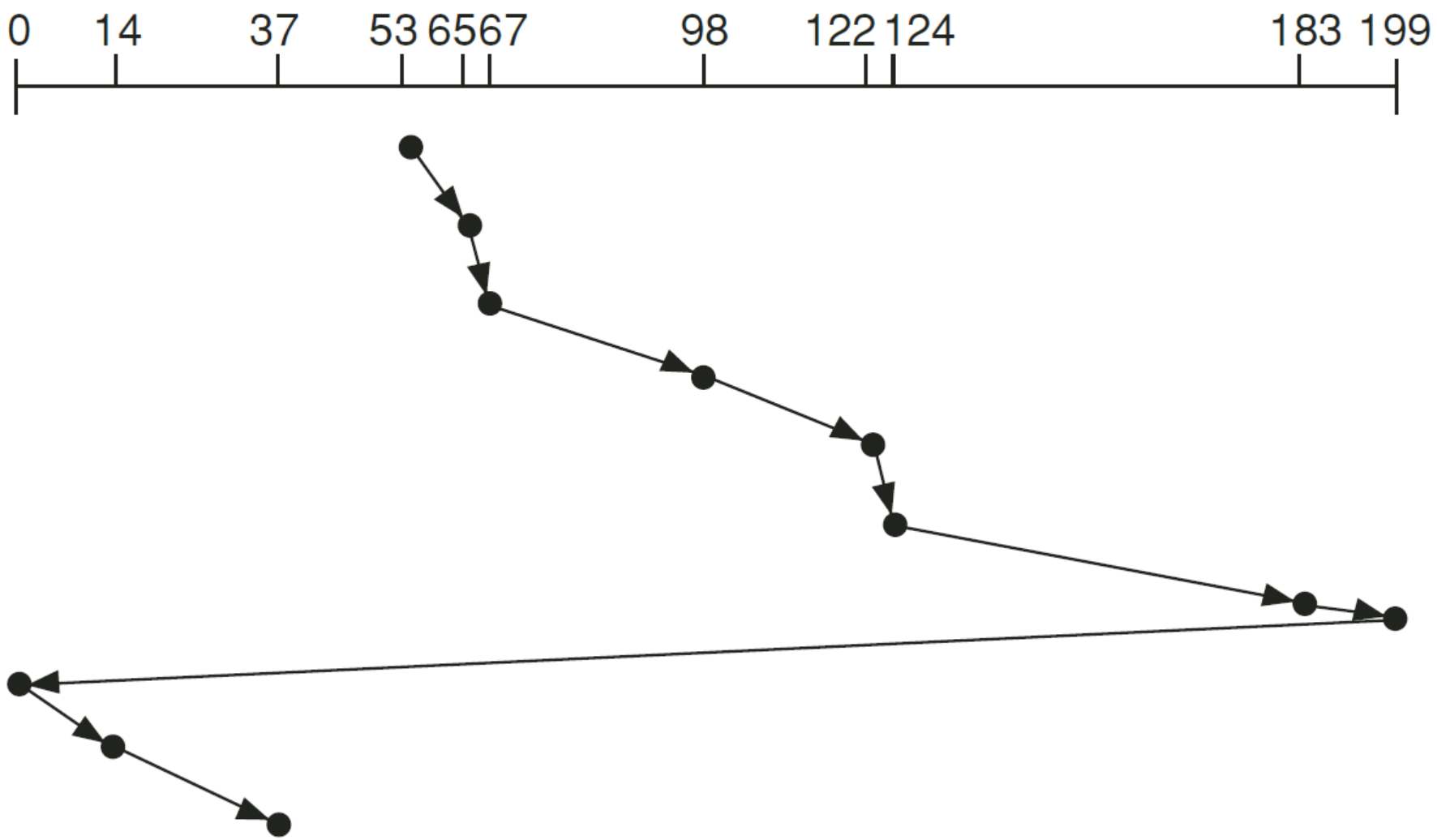
Trenutna pozicija glave: cilindar 53
Red čekanja: 98, 183, 37, 122, 14, 124, 65, 67



Raspređivanje zahteva za rad sa diskom

- **Algoritam C-SCAN** (engl. *Circular SCAN*).
- Varijanta SCAN algoritma koja **razrešava problem favorizovanja unutrašnjih cilindara**.
- Izmena se sastoji u tome da se zahtevi opslužuju samo u jednom smeru.
  - Kada glave dođu do poslednjeg cilindra, pomeraju se na početak, ne opslužujući zahteve na tom putu.
  - Posle toga se nastavlja opsluživanje zahteva od početnog do krajnjeg cilindra.
- Primer:
  - Trenutna pozicija glava za čitanje i pisanje je na cilindru 53.
  - U red čekanja za disk zahtevi pristižu po sledećem redu: 98, 183, 37, 122, 14, 124, 65, 67.

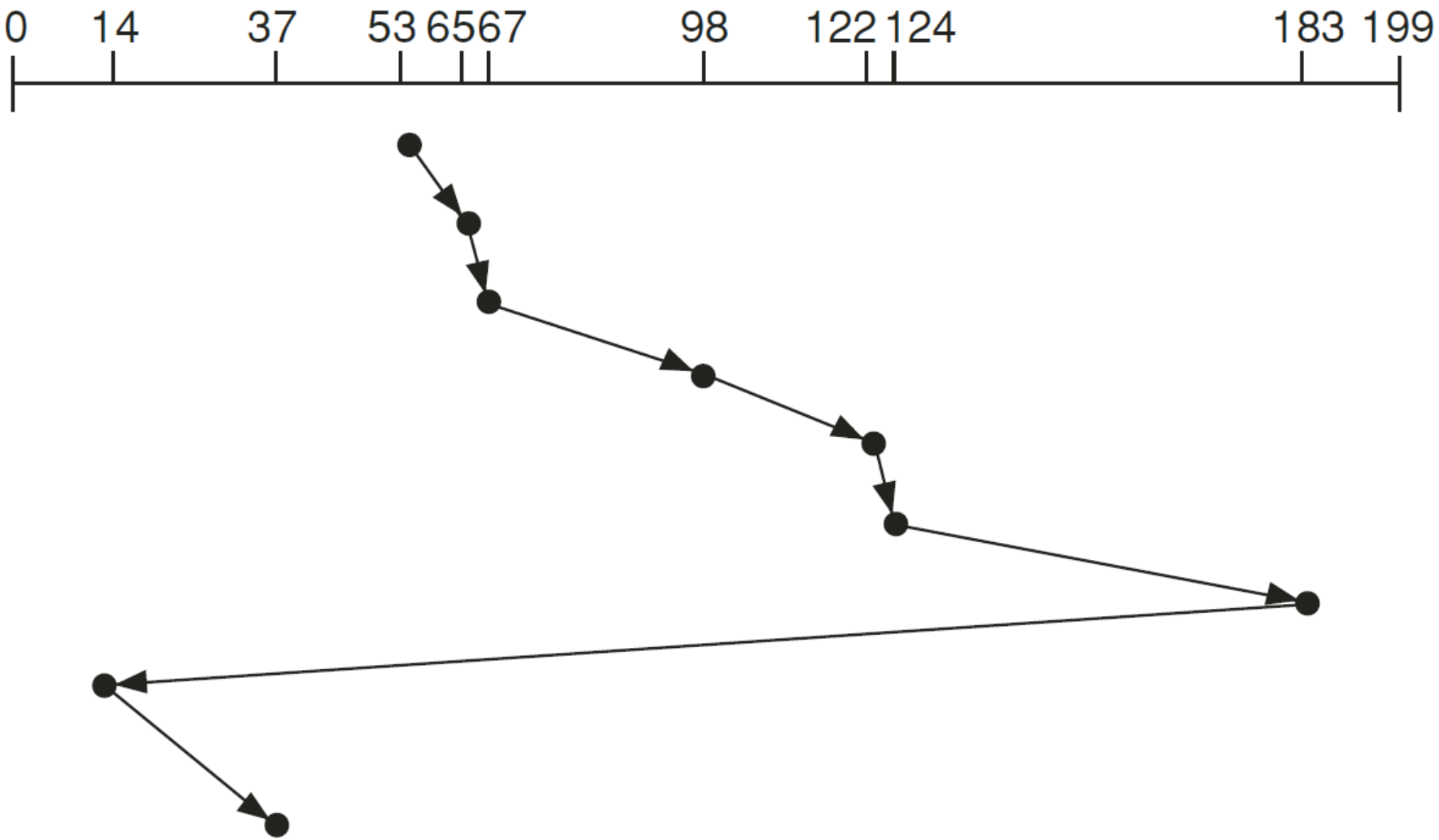
Trenutna pozicija glave: cilindar 53
Red čekanja: 98, 183, 37, 122, 14, 124, 65, 67



Raspređivanje zahteva za rad sa diskom

- **Algoritmi LOOK i C-LOOK** (engl. *Circular LOOK*).
- Modifikacije SCAN i C-SCAN algoritama.
- Glave se ne pomeraju do poslednjeg zahteva koji se nalazi u redu čekanja u tom smeru.
- LOOK opslužuje zahteve u oba smera.
- C-LOOK opslužuje zahteve samo u rastućem smeru do poslenjeg zahteva u redu nakon čega se vraća na zahtev najbliži početku diska.
- Primer (C-Look):
  - Trenutna pozicija glava za čitanje i pisanje je na cilindru 53.
  - U red čekanja za disk zahtevi pristižu po sledećem redu: 98, 183, 37, 122, 14, 124, 65, 67.
- Napomena:
  - Ime su dobili po tome što “gledaju” na kom se cilindru nalazi poslednji zahtev u tom smeru.
  - U praksi se umesto SCAN algoritama uvek koriste LOOK algoritmi.

Trenutna pozicija glave: cilindar 53
Red čekanja: 98, 183, 37, 122, 14, 124, 65, 67



Kako izabrati najbolji algoritam za rapoređivanje disk zahteva?

- Svi algoritmi su bolji od FCFS, ali je teško odrediti koji je od njih najbolji!
- Performanse samih algoritama **zavise od prispelih zahteva za rad sa diskom**.
  - Kružne varijante SCAN i LOOK algoritama imaju mnogo bolju raspodelu opsluživanja i nemaju problem zakucavanja (engl. *starvation*).
  - C-LOOK je najbolje rešenje za jako opterećene sisteme.
- Modernije varijate ovih algoritama **minimiziraju i pozicioniranje i rotaciono kašnjenje**.
  - Rotaciono kašnjenje ima dominantan uticaj na performanse savremenih diskova.
  - Jedan takav algoritam je **SATF** (engl. *shortest access time first*).
  - Radi na principu SSTF algoritma ali pri odabiru sledećeg zahteva iz reda računa obe mehaničke komponente.
- Najsavremeniji algoritmi uzimaju u obzir i keširanje na samom disk uređaju.
  - C-LOOK u kombinaciji sa ugrađenim disk keširanjem daje najbolje rezultate, što potvrđuju brojne simulacije iz otvorene literature.

Oporavak sistema od otkaza upisa

- Upis na disk može da se završi na 3 načina:
  - **Uspešno okončan upis**.
    - Svi sektori su uspešno upisani na disk.
  - **Delimični otkaz**.
    - Otkaz je nastupio u sredini transfera.
    - Neki su sektori dobro upisani, a neki su oštećeni.
  - **Potpuni otkaz**.
    - Ništa nije upisano jer je ciklus upisa odmah otkazao.
- Kako se sistem može oporaviti od otkaza?
  - Koristimo:
    - Tehnike vođenja dnevnika transakcija (engl. *journalling*).
    - RAID strukture koje koriste princip ogledala i parnosti.

Osnovne karakteristike RAID sistema

- **Paralelizam i konkurentnost operacija**.
  - Tehnika deljenja podataka između različitih diskova.
  - Paralelno izvršenje više nezavisnih disk operacija u istom trenutku (konkurentnost operacija).
- **Povećanje pozdanosti uvođenjem redundanse**.
  - Svaki disk ima svoju pouzdanost koja se meri srednjim vremenom otkaza.
  - Verovatnoća otkaza RAID strukture koju čini N diskova uvećava se N puta.
  - Problem pouzdanosti se rešava uvođenjem redundanse.
    - Čuvaju se dodatne informacije koje obezbeđuju mogućnost potpunog povratka podatak u slučaju otkaza jednog diska.
  - Tehnike:
    - **Ogledala** (engl. *mirroring*).
    - **Parnosti** (engl. *parity*).

RAID 0

- Najbolje performance čitanja i upisa.
- Nema redundanse: otkaz jednog diska znači gubitak svih podataka!

Disk 1	Disk 2	Disk 3	Disk 4
A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P
↓	↓	↓	↓

RAID 1

- Svaki disk ima svoje ogledalo!
- Najveći utrošak prostora, najveći stepen redundanse.
- Najgore performance upisa.

RAID 4

- Deljenja podataka na blok nivou.
- Redundansa: parnost za diskove na nivou bloka (engl. *block-interleaved parity*).
  - Za N blokova (sa N diskova) dovoljan je jedan blok parnosti.
- Dobre osobine: kombinacija paralelizma i konkurentnosti.
- Mana: u svakom ciklusu upisa učestvuje i disk parnosti (postaje usko grlo u sistemu).

Disk 1	**Disk 2**	**Disk 3**	**Disk 4 (par)**
A	B	C	par (A, B, C)
D	E	F	par (D, E, F)
G	H	I	par (G, H, I)
J	K	L	par (J, K, L)
↓	↓	↓	↓

RAID 5

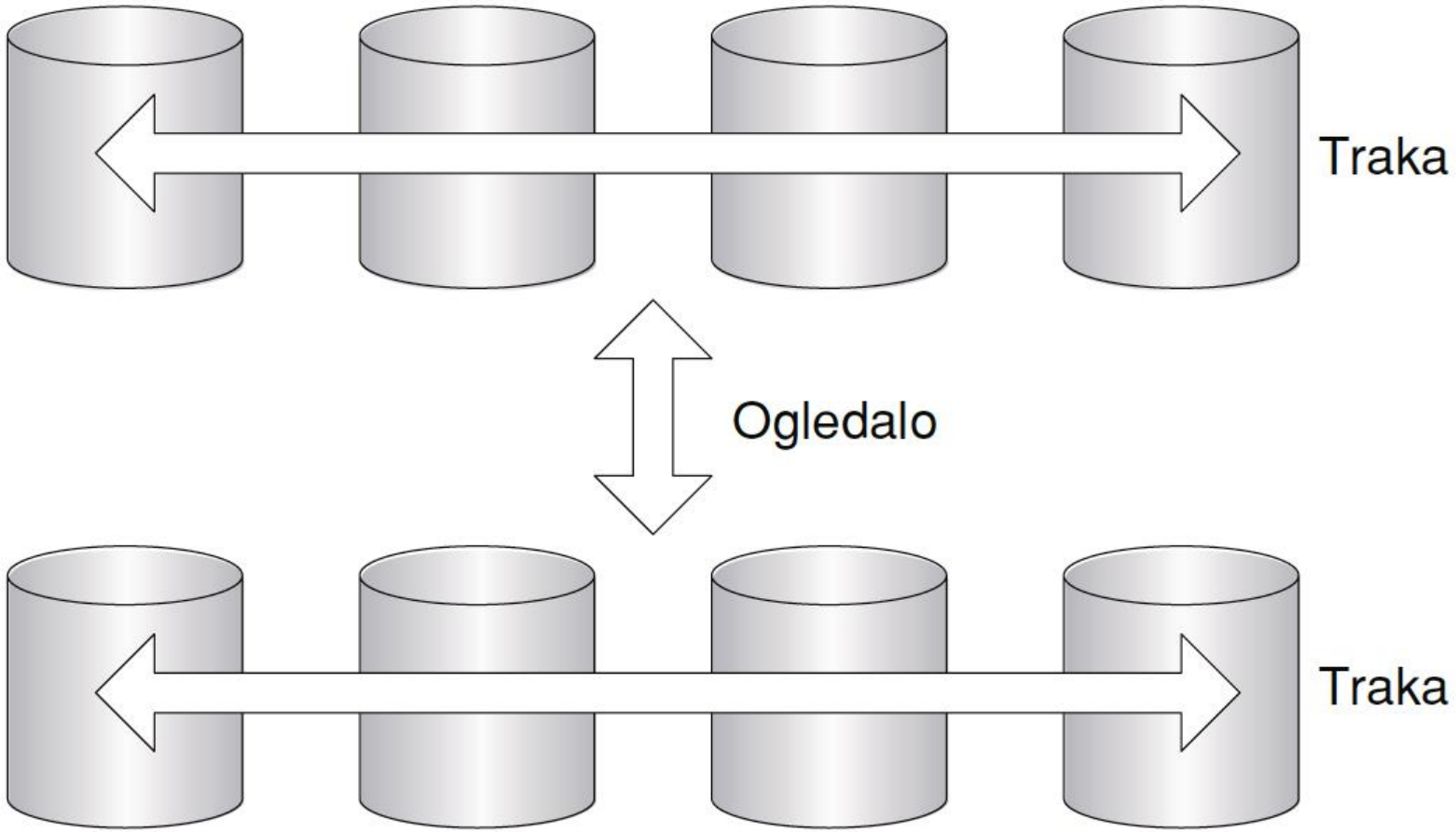
- RAID 5 (engl. *block-interleaved distributed parity*, *rotating parity array*).
- Svih N+1 diskova predstavljaju i diskove podataka i diskove parnosti.
- Parnost se upisuje u levom simetričnom rasporedu.
- Najbolja kombinacija: paralelizam, konkurentnost, diskovi su ravnomerno opterećeni.

Disk 1	**Disk 2**	**Disk 3**	**Disk 4**
A	B	C	par (A, B, C)
D	E	par (D, E, F)	F
G	par (G, H, I)	H	I
par (J, K, L)	J	K	L
↓	↓	↓	↓

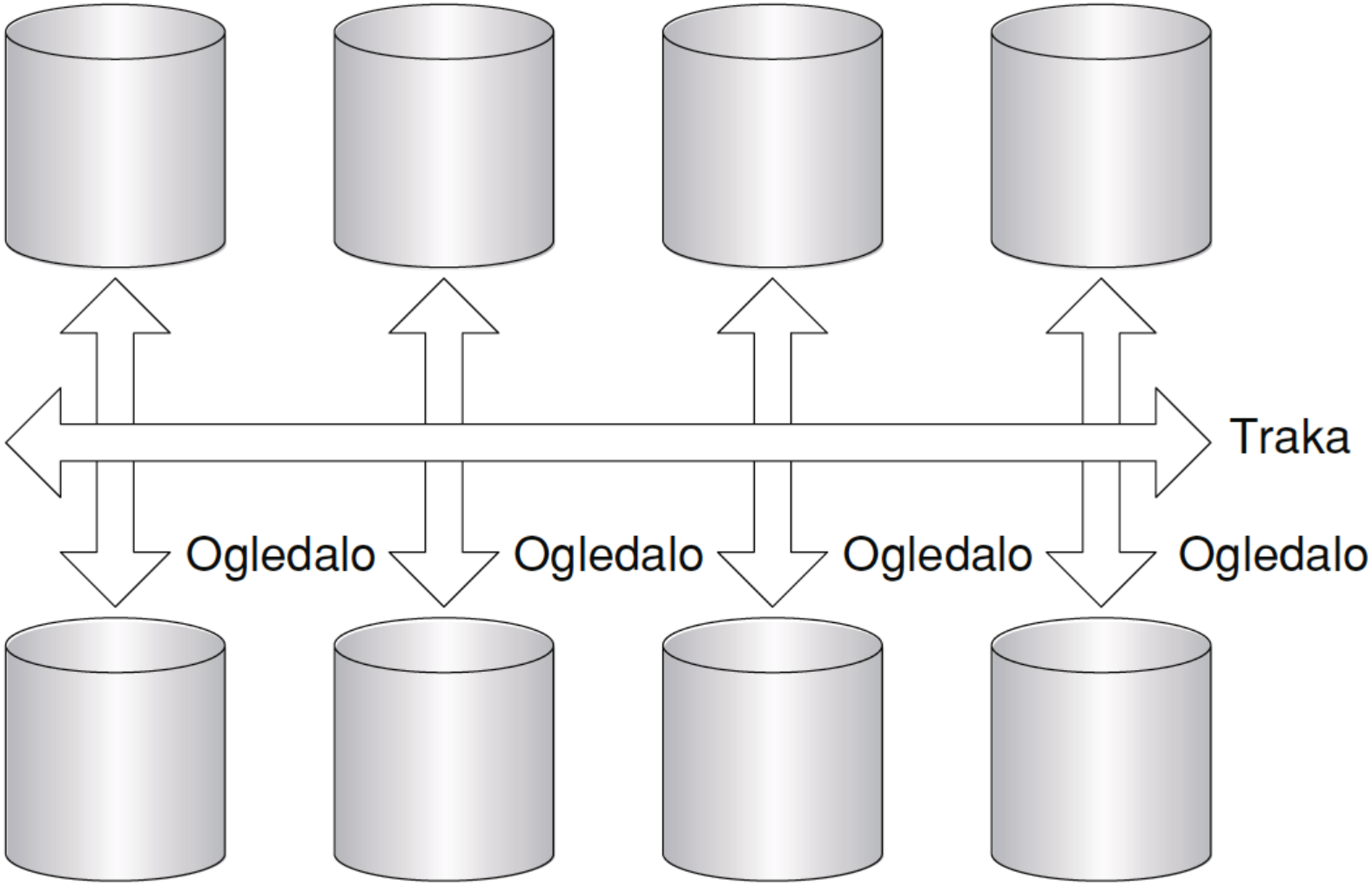
RAID 0+1 / 1+0

- Kvalitetne kombinacije tehnike 0 i 1.
  - RAID 0 obezbeđuje visoke performance.
  - RAID 1 obezbeđuje visoku pouzdanost.
  - Kombinacija ostaje i dalje skupa jer udvostručava broj diskova!
- Kombinacija 0+1:
  - Skup diskova deli podatke.
  - Potom se sve stripe jedinice u celini kopiraju u svoje ogledalo.
- Kombinacija 1+0:
  - Svaki disk ima svoje ogledalo.
  - Podaci se dele između ogledala.

RAID 0+1



RAID 1+0



Literatura

1. B. Đorđević, D. Pleskonjić, N. Maček (2005): Operativni sistemi: teorija, praksa i rešeni zadaci. Mikro knjiga, Beograd.
2. R. Popović, I. Branović, M. Šarac (2011): Operativni sistemi. Univerzitet Singidunum, Beograd.

Hvala na pažnji

Pitanja su dobrodošla.