Sistemi za čuvanje podataka
Baze podataka 2

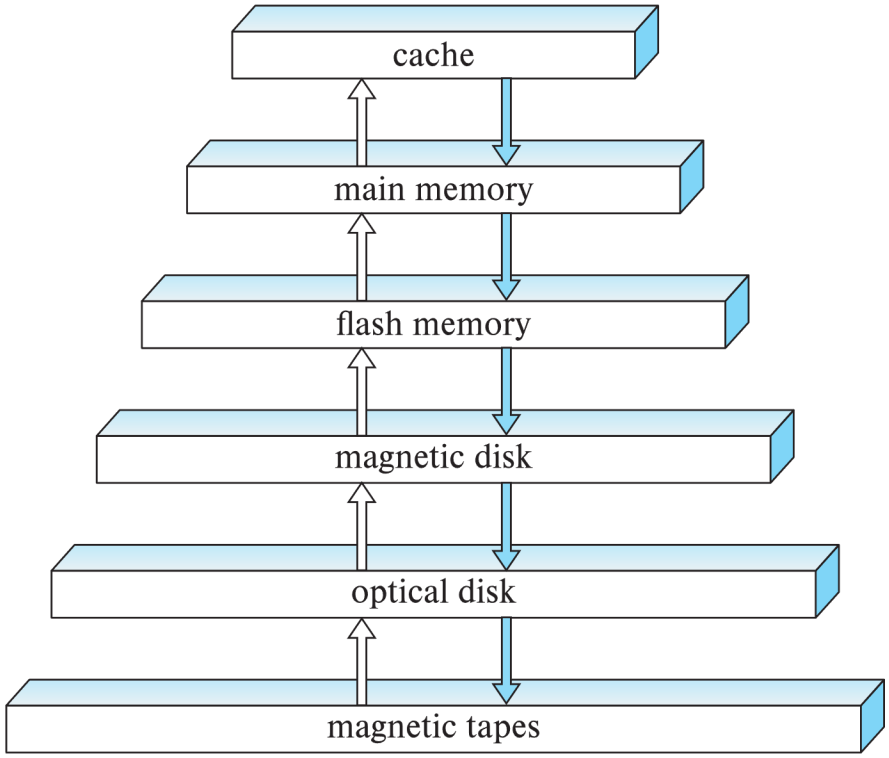
dr Miloš CVETANOVIĆ

Klasifikacija medijuma za čuvanje podataka (storage systems media)

- Gube sadržaj prilikom nestanka napajanja (**volatile storage**) – nestalni
  - Koriste se kod primarnih sistema za čuvanje podataka
- Čuvaju sadržaj (perzistitiraju) i nakon isključenja napajanja (**non-volatile storage**) – stalni
  - Koriste se kod sekundarnih i tercijalnih sistema za čuvanje podataka ili primarnih sistema sa baterijskim napajanjem (npr. battery-backed up main-memory)
- Faktori koji utiču na izbor medija za čuvanje uključuju:
  - Brzinu kojom se može pristupati podacima
  - Cenu po jedinici čuvanja podataka
  - Pouzdanost

Hijerarhija sistema za čuvanje podataka

- Primarni sistemi: Najveća brzina pristupa, nestalan medijum (volatile)
  - Praktično neophodan za funkcionisanje računarskog Sistema (**cache**, **main memory**)
- Sekundarni sistemi: Umerena brzina pristupa, stalan medijum (non-volatile)
  - Često se naziva on-line storage (**flash memory**, **magnetic disks - HDD**)
- Tercijalni sistemi: Spor pristup, stalan medijum (non-volatile) za arhiviranje podataka
  - Često se naziva off-line storage (**magnetic tape**, **optical storage**)



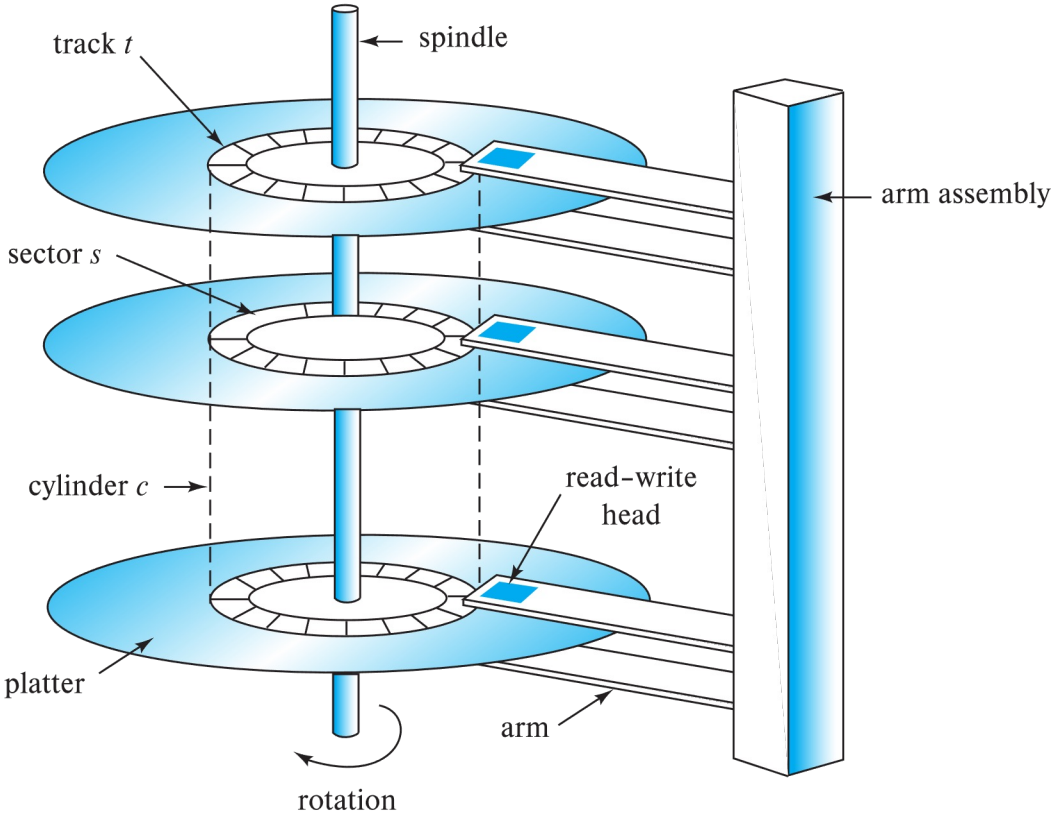
Magnetna traka
Sekvencijalni pristup
Nekoliko drajva sa mnogo traka
Formati: Digital Audio Tape – nekoliko GB
Digital Linear Tape – 10-40GB
Ultrium – 330GB
Brzina prenosa podataka nekoliko 10-tina MB/s
Džuboks (juke box) sa petabajtima podataka

Interfejsi sistema za čuvanje podataka (storage interfaces)

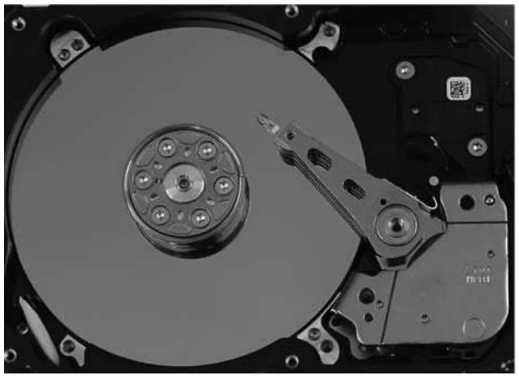
- Familije standarda za disk interfejse
  - **SATA** (Serial ATA) (npr. SATA 3 – brzine prenosa podataka do 6Gb/s
  - **SAS** (Serial Attached SCSI) (npr. SAS ver. 3 – brzine prenosa podataka do 12Gb/s
  - **NVMe** (Non-Volatile Memory Express) (preko PCIe konektora) – brzine do 24Gb/s
- Diskovi uglavnom povezani direktno na računarski sistem – **DAS** (Direct-Attached Storage)
- **SAN** (Storage Area Network) – više diskova povezanih brzom mrežom zasnovanoj na FC (Fibre Channel) ili iSCSI (Internet SCSI) sa više servera koji im pristupaju na nivou blokova (disk sistem interfejs)
- **NAS** (Network-Attached Storage) – više diskova povezanih lokalnom mrežom zasnovanoj na Ethernet (TCP/IP) sa više servera/računara koji im pristupaju na nivo fajlova (fajl sistem interfejs)

Magnetni disk (Hard disk HDD) – prikaz

Šematski prikaz magnetnog diska



Fotografija magnetnog diska



Magnetni disk (Hard disk HDD) – opis

- Glava (**head**) za čitanje/pisanje
- Površina disk ploča (**platers**) podeljena na cirkularne staze (**tracks**)
  - Oko 50K-100K staza po ploči na tipičnom HDD
- Svaka staza podaljena na sektore (**sectors**)
  - Sektor je najmanja jedinica podataka koja može biti pročitana/upisana
  - Sektor je tipično 512B
  - Broj sektora po stazi: 500-1000 (za unutrašnje staze) i 1000-2000 (za spoljne staze)
- Da bi se pročitao ili upisao sektor
  - Disk ruka (**arm**) mora da pozicionira glavu na odgovarajuću stazu
  - Ploča se konstantno okreće (upis/čitanje se dešava dok sektor prolazi ispod glave)
- Konstrukcija diska
  - Više disk ploča na zajedničkom vretenu (**spindle**) – tipično 1-5 ploča
  - Jedna glava po ploči, montirana na zajedničku ruku
- Cilindar (**cylinder**) na poziciji ***i*** predstavlja skup svih ***i-tih*** staza na svim pločama

Magnetni disk (Hard disk HDD) – logika

- Disk kontroler (**disk controller**) – interfejs između računarskog sistema i diska
  - Trenutno najčešće integrisan sa diskom
  - Prihvata komande za čitanje i pisanje po sektorima
  - Inicira akcija za pozicioniranje ruke diska na određenu stazu i početak čitanja/pisanja
  - Izračunava i pridružuje kontrolni zbir (**checksum**) na svaki sektor – radi verifikacije
  - Pri čitanju: pročitani kontrolni zbir se verifikuje izračunatim kontrolnim zbirom
  - Pri upisu: nakon upisa, radi se čitanje upravo upisanog sektora radi verifikacije
  - Radi remapiranje oštećenih sektora (**bad sectors**)

Magnetni disk (Hard disk HDD) – pristup i prenos

- Vreme pristupa (**access time**) – vreme od trenutka pristizanja zahteva za čitanje/pisanje do trenutka kada počne prenos podataka, a obuhvata:
  - Vreme traženja (**seek time**) – vreme pozicioniranje ruke diska iznad tražene staze
    * Prosečno vreme traženja je 1/2 od najgoreg vremena traženja
      (bilo bi 1/3 kada bi sve staze imale isti broj sektora
      i ako bi se ignorisalo vreme pokretanja i zaustavljanja ruke diska)
    * Kod tipičnih diskova iznosi 4 – 10ms
  - Rotaciono kašnjenje (**rotational latency**) – vreme pristizanja sektora ispod glave diska
    * Prosečno rotaciono kašnjenje je 1/2 vremena rotacije
    * Kod tipičnih diskova vreme rotacije iznosi 4 – 11ms (odnosno 5400 – 15000 rpm)
  - Tipično vreme pristupa je 5 – 20ms u zavisnosti od modela diska
- Brzina prenosa podataka (**data-transfer rate**) – brzina kojom podaci mogu biti preneti sa diska odnosno sačuvani na disku
  - Tipično 25 – 200 MB/s, odnosno sporije za unutrašnje staze

Magnetni disk (Hard disk HDD) – načini pristupa

- Disk blok (**disk block**) – je logička jedinica za čitanje/pisanje
  - Tipično veličine 4 – 16KB
    * Manji blok: više transfera sa diska
    * Veći blok: više utrošenog prostora usled parcijalne popunjenosti blokova
- Sekvencijalni pristup (**sequential access patern**)
  - Sukcesivni zahtevi za pristup se odnose na sukcesivne disk blokove
  - Vreme traženja potrebno samo za prvi blok
- Proizvoljni pristup (**random access patern**)
  - Sukcesivni zahtevi za pristup se odnose na disk blokove bile gde na disku
  - Vreme traženja potrebno za svaki blok
  - Brzina prenosa je niska zbog mnogo utrošenog vremena na traženje blokova
- Broj ulazno/izlaznih operacija (**IOPS**)
  - Broj čitanja proizvoljnih blokova koje disk može da obavi u jednoj sekundi
  - Tipično 50 – 200 IOPS

Magnetni disk (Hard disk HDD) – pouzdanost

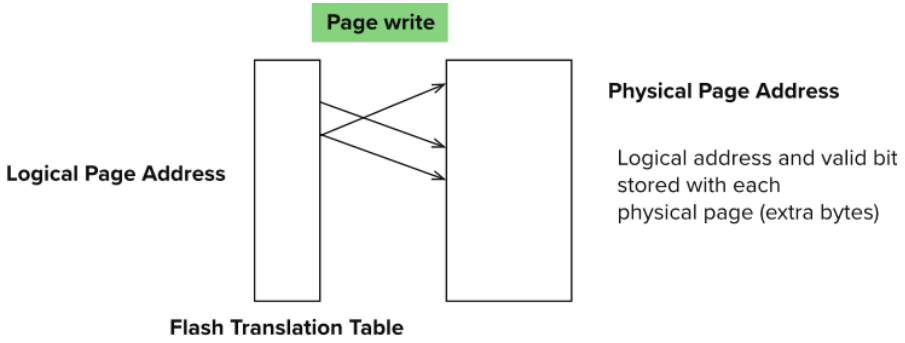
- Prosečno vreme do otkaza (**Mean time to failure - MTTF**) – je prosečno vreme koje se očekuje da će disk moći kontinualno da radi bez kvara
  - Tipično je 3 – 5 godina
  - Verovatnoća kvara novih diskova je poprilično mala i odgovara “teoretskom MTTF“ od 500 000 – 1 200 000 sati (57 – 136 godina?)
  - Treba razumeti da, na primer, MTTF od 1 200 000 sati za novi disk znači da na svakih 1000 takvih diskova, u proseku jedan otkaže na svakih 1200 sati
- MTTF opada sa starenjem diska

Fleš memorija (Flash memory) – opis

- NOR fleš i NAND fleš memorija (NOR brže čitanje, sporiji upis, manji kapacitet od NAND)
- NAND fleš
  - Široko rasprostranjena / jeftinija od NOR
  - Zahteva čitanje stranu-po-stranu (**page**) tipično veličine 512B – 4KB
    * Tipično 20 – 100ms da se pročita strana
    * Nema razlike u brzini između sekvencijalnog i proizvoljnog pristupa
  - Strana može biti upisana samo jednom
    * Strana mora biti obrisana da bi se omogućio ponovni upis
- Diskovi sa fleš memorijom (**Solid state disks - SSD**)
  - Koristi standardni disk sistem interfejs (pristup na nivou bloka)
  - Čuva podatke na više fleš memorija interno
  - Brzina prenosa podataka daleko veća nego kod magnetnih HDD
    * Tipično do 400MB/s koristeći SATA3, odnosno 3GB/s koristeći NVMe PCIe

Fleš memorija (Flash memory) – logika

- Brisanje se obavlja u jedinicama blokova-brisanja (**erase block**)
  - Traje 2 – 5ms
  - Blok-brisanja je tipično veličine 256KB – 1MB (odnosno 128 – 256 strana)
- Remapiranje logičkih adresa strana na fizičke adrese strana
  kako bi se izbeglo čekanje usled potrebe za brisanjem pre upisa
- Mapiranje se prati pomoću tabele translacija (**flash translation table**)
  - Takođe, nova adresa sačuvana u polju za labelu u okviru strane
  - Remapiranje obavlja poseban sloj (**flash translation layer**)



- Nakon 100 000 do 1 000 000 brisanja, blok-brisanja postaje nepouzdan i ne može se više koristiti, te je neophodno uravnotežiti habanje (**wear leveling**)

Fleš memorija (Flash memory) – performanse

- Proizvoljno (random) čitanje/pisanje
  - Tipično 4KB čitanje: 10 000 čitanja po sekundi (10 000 IOPS)
  - Tipično 4KB pisanje: 40 000 IOPS
  - SSD podržava paralelna čitanja
    * Tipično 4KB čitanje:
      100 000 IOPS ukoliko se rade 32 zahteva paralelno (QD-32) na SATA
      350 000 IOPS takođe sa QD-32, ali na NVMe PCIe
    * Tipično 4KB upis:
      100 000 IOPS sa QD-32
- Brzina prenosa podataka pri sekvencijalnom čitanju/pisanju
  - 400 MB/s na SATA3, odnodno 2 – 3 GB/s na NVMe PCIe
- Hibridni diskovi (**Hybrid disks**) – kombinuju manju fleš memoriju zbog brzine sa većim magnetnim diskom zbog kapaciteta

Memorija klase za smeštanje (Storage Class Memory) – opis

- Tehnologija 3D-XPoint razvijena od strane Intel
- Dostupna na Intel Optane
  - Manja kašnjenja nego klasični fleš SSD
  - Podržava direktni pristup na nivou reči
  - Brzina uporediva sa brzinom operativne memorije

Nizovi nezavisnih redundantnih diskova – uvod

- Nizovi nezavisnih redundantnih diskova (Redundant Arrays of Independent Disks - RAID)
  - Tehnika organizacije diskova koja upravlja većim brojem diskova, obezbeđujući privid upotrebe jedinstvenog diska:
    * Velikog kapaciteta i velike brzine koristeći više diskova u paraleli
    * Velike pouzdanosti čuvajući podatke redundantno, tako da podaci mogu biti oporavljeni čak i u slučaju otkaza nekog od diskova
- Verovatnoća da bar jedan disk u skupu od N diskova otkaže je mnogo veća nego verovatnoća da jedan određeni disk otkaže
  - Na primer, sistem od 100 diskova, svaki sa MTTF od 100 000 sati (približno 11 godina) bi imao MTTF čitavog sistema od 1000 sati (približno 41 dan)
  - Tehnike za upotrebu redundanse su od kritičnog značaja kako bi se izbegao gubitak podataka prilikom upotrebe sistema sa velikim brojem diskova

Nizovi nezavisnih redundantnih diskova – poboljšanje pouzdanosti pomoću redundanse

- Redundansa – čuvanje dodatnih informacija koje se mogu iskoristiti za oporavak prilikom otkaza nekog od diskova
- Čuvanje kopije ili preslikavanje (**mirroring**, **shadowing**)
  - Dupliciranje svakog diska. Logički disk se sastoji od dva fizička diska
  - Svaka operacija upisa se sprovodi na oba diska (a čitanje sa bilo kog od dva diska)
  - Ukoliko jedan od dva diska otkaže, podaci su i dalje dostupni na drugom
    * Do gubitka podataka dolazi samo ukoliko jedan od diskova otkaže, a potom i drugi otkaže pre popravke prvog
      Verovatnoća kombinovanih događaja je mala (izuzev u slučaju požara, poplave itd.)

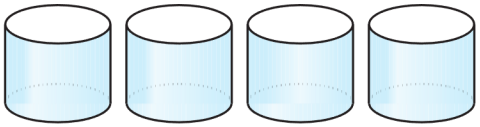
- Srednje vreme do gubitka podataka (mean time to data loss - MTTDL) zavisi od MTTF (mean time to failure) i MTTR (mean time to repair)
  - Na primer, MTTF od 100 000 sati, i MTTR od 10 sati, daju MTTDL od 57 godina za par diskova (mirroring) – ukoliko se ignorišu zavisni otkazi

Nizovi nezavisnih redundantnih diskova – poboljšanje performansi pomoću paralelizma

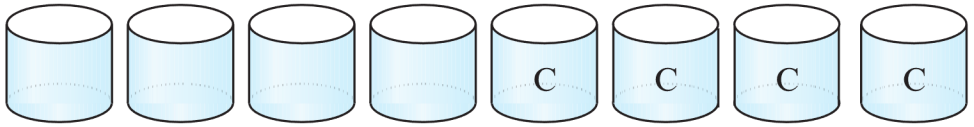
- Dva glavna cilja prilikom uvođenja paralelizma kod sistema diskova su:
  1. Balansiranje opterećenja više manjih pristupa da bi se povećala propusnost (**throughput**)
  2. Paralelizacija velikih pristupa kako bi se smanjilo vreme odgovora (**response time**)
- Poboljšanje brzine prenosa "raspoređivanjem delova" (**striping**) podataka na više diskova
- Delovi na nivou bitova (**bit-level striping**)
  - Sa nizom od 8 diskova, *i-ti* bit svakog bajta se čuva na *i-tom disku*
  - Operacija čitanja jednog bajta je 8 puta brže od čitanja bajta na jednom disku
  - Međutim, traženje podatka je lošije nego na jednom disku
    * Ovaj pristup se više ni ne koristi
- Delovi na nivou bloka (**block-level striping**)
  - Sa nizom od N diskova, *i-ti* blok nekog fajla se čuva na disku (*i* mod N)+1
  - Zahtevi do različitih blokova se mogu izvršavati u paraleli ukoliko se blokovi nalaze na različitim diskovima
  - Zahtev za dužim nizom blokova može da iskoristi sve diskove u paraleli

Nizovi nezavisnih redundantnih diskova – RAID nivoi

- Šeme koje obezbeđuju redundansu po nižem trošku kombinujući raspoređivanje blokova i bitova parnosti
  - Šeme predstavljaju RAID organizacije, tj. RAID nivoe i razlikuju se po ceni, performansama i pouzdanosti
- RAID nivo 0: Raspoređivanje blokova (block striping), bez redundanse
  - Koristi se za aplikacije visokih performansi gde gubitak podataka nije kritičan
- RAID nivo 1: Preslikani diskovi (mirroring disks) sa raspoređivanjem blokva (block striping)
  - Obezbeđuje najbolje performanse upisa
  - Popularan kod aplikacija za smeštanje dnevnika (tj. log fajlova) sistema baza podataka



(a) RAID 0: nonredundant striping



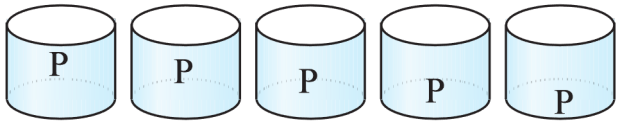
(b) RAID 1: mirrored disks

Nizovi nezavisnih redundantnih diskova – RAID nivoi

- Blokovi parnosti (**parity blocks**)
  - Blok parnosti *j* čuva XOR bitova iz bloka *j* svakog od diskova
  - Kada se upisuje podatak u blok *j*, blok parnosti *j* se takođe mora izračunati i upisati
    * To se može uraditi koristeći stare vrednosti bloka parnosti *j*, stare vrednosti bloka *j* i nove vrednosti bloka *j* (odnosno potrebna su čitanja 2 bloka i upisa 2 bloka)
    * To se može uraditi koristeći nove vrednosti svih blokova koji odgovaraju bloku parnosti (efikasnije kada se upisuje veća količina podataka sekvencijalno)
- U slučaju kvara, da bi se oporavio oštećeni podatak nekog bloka potrebno je izračunati XOR odgovarajućih bitova sa svih ostalih blokova uključujući i blok parnosti

Nizovi nezavisnih redundantnih diskova – RAID nivoi

- RAID nivo 5: Preplitanje blokova sa distribuiranom parnošću (**block-interleaved distrubuted parity**)
  - Particioniše podatke i parnost na svih N+1 diskova,
    umesto da podatke smešta na N diskova, a parnost uvek na 1 istom disku



(c) RAID 5: block-interleaved distributed parity

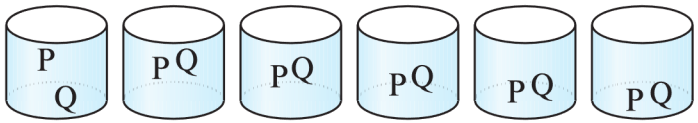
  - Na primer, sa 5 diskova, blok parnosti za *n-ti* skup blokova
    se smešta na disk (*n* mod 5)+1, dok se blokovi sa podacima smeštaju na ostala 4 diska

P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

  - Upis blokova se obavlja u paraleli ukoliko se svi blokovi i njihov blok parnosti
    nalaze na različitim diskovima

Nizovi nezavisnih redundantnih diskova – RAID nivoi

- RAID nivo 6: P+Q šema redundanse (**P+Q Redundacy scheme**)
  - Slično RAID nivou 5, ali umesto jednog bloka parnosti čuva dva bloka za korekciju grešaka (P, Q) i štiti od otkaza više jednog diska
  - Bolja pouzdanost nego kod nivoa 5, ali uz veću cenu
  - Dobija na značaju kako se kapacitet povećava



(d) RAID 6: P + Q redundancy

Nizovi nezavisnih redundantnih diskova – RAID nivoi

- Ostali RAID nivoi (koji se ne koriste u praksi):
  - Nivo 2 – Raspoređivanje na nivou bitova sa ECC (Memory style Error Corecting Codes) čuva dva bloka za korekciju grešaka (P, Q) i štiti od otkaza više jednog diska
  - Nivo 3 – Preplitanje bitova sa parnošću
  - Nivo 4 – Preplitanje blokova sa parnošću
    * Blok parnosti se čuva na posebnom disku parnosti (parity disk) za svih N diskov
    * RAID 5 je bolji od RAID 4 jer prilikom upisa proizvoljnih blokova disk parnosti dobija mnogo veće opterećenje i postaje usko grlo sistema

Nizovi nezavisnih redundantnih diskova – izbor RAID nivoa

- Faktori prilikom izbora RAID nivoa:
  - Cena implementacije
  - Performanse – broj potrebnih IOPS kao i protoka prilikom normalnog rada
  - Performanse prilikom otkaza delova sistema
  - Performanse prilikom oporavka otkazali diskova
    * Uključujući i vreme potrebno da se ponovo formira (restaurira)otkazali disk

Nizovi nezavisnih redundantnih diskova – izbor RAID nivoa

- RAID 0 – samo kada bezbednost podataka nije važna, a potrebne visoke performanse
  * Oporavak podataka na osnovu drugih izvora)
- RAID 1 – obezbeđuje mnogo bolje performanse upisa od RAID 5, ali ima veću cenu
  * Za upis jednog bloka RAID 5 zahteva čitanje 2 bloka i upis 2 bloka dok RAID 1 zahteva samo upis 2 bloka
  * RAID 1 je preferiran za aplikacije sa mnogo malim upisa proizvoljnih blokova
- RAID 5 – je preferiran za aplikacije sa mnogo velikih sekvencijalnih upisa i zahtevaju čuvanje velike količine podataka
- RAID 6 – obezbeđuje bolju zaštitu podataka nego RAID 5 jer može da toleriše otkaz dva diska (ili disk blokova) istovremeno
  * Dobija na značaju usled pojave tzv. skrivenog (latentnog) otkaza blokova koji u kombinaciji sa otkazom nekog drugog diska rezultira u gubitku podataka kod RAID 1 i RAID 5

Nizovi nezavisnih redundantnih diskova – implementacija

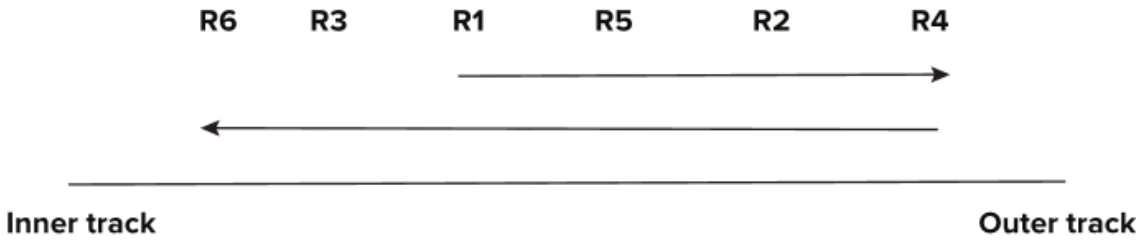
- Softverski RAID – implementacija u potpunosti u softveru, bez specifičnog hardvera
- Hardverski RAID – implementacija u vidu specifičnog hardvera
  - Upotreba stalnog (non-volatile) RAM da se sačuvaju upisi
    (a njihov pravi upis, odloži i optimizuje)
  - Nestanak struje tokom upisa može dovesti do oštećenja diska
    * Nestanak struje nakon upis jednog bloka pre nego se upiše kopija (mirror)
    * Oštećenja diska treba da budu detektovana nakon povratka struje
      Oporavak je sličan oporavku prilikom oporavka otkazalog diska
      NV-RAM pomaže u efikasnom detektovanju potencijalno oštećenih blokova
      jer u suprotnom svi blokovi diska moraj biti pročitani i upoređeni
      sa kopijom odnosno blokom parnosti

Nizovi nezavisnih redundantnih diskova – implementacija

- Skriveni otkazi (**latent failures**) – podaci koji su davno upisani uspešno, kasnije postanu oštećeni
- Čišćenje podataka (**data scrubing**) – kontinulno skeniranje/traganje za skrivenim otkazima, a potom i automatski oporavak na osnovu kopije odnosno parnosti
- Vruća zamena (**hot swapping**) – zamena diska bez isključivanja sistema
  - Podržano od strane hardverskih RAID sistema
  - Smanjuje vreme oporavka i značajno popravlja dostupnost (**availability**)
- Mnogi sistemi poseduju rezervne diskove sve vreme tako da mogu zameniti neki otkazali disk trenutno, odnosno odmah nakon detekcije otkaza nekog od diskova
  - Smanjuje vreme oporavka
- Mnogi hardverski RAID sistemi garantuju da nema jedne tačke otkaza (**single point of failure**) koja bi zaustavila funkcionisanje celog sistema
  - Redundanta napajanja sa baterijskom rezervom
  - Više kontrolera i više interkonkcija (redudansa kontrolera i veze)

Optimizacija pristupa blokovima diska

- Baferisanje (**buffering**) – keširanje disk blokova u operativnoj memoriji
- Čitanje unapred (**read-ahead**) – Prilikom čitanje jednog bloka, čitaju se i naredni blokovi zbog očekivanja da će i oni uskoro biti potrebni
- Raspored kretanja ruke diska (**disk-arm-scheduling**) – algoritmi za preuređivanje redosleda pristupa blokovima kako bi se minimiziralo (optimizovalo) kretanje ruke diska
  - Algoritam lifta (elevator algorithm)



- Organizacija fajlova (**file organization**) – sa ciljem što manje defragmentacije diska
  - Za fajl se alocira što više kontinulanih blokova (koliko god da je to moguće)
  - Alociranje u jedinicima proširenja (**extents**)