

Gagnasafnsfræði  
Geymsluaðferðir og flýtvísar

Hallgrímur H. Gunnarsson

# Mechanical sympathy

## **Hardware vs. software**

"The most amazing achievement of the computer software industry is its continuing cancellation of the steady and staggering gains made by the computer hardware industry." (Henry Peteroski)

## **Mechanical sympathy**

The name "Mechanical Sympathy" comes from the great racing driver Jackie Stewart, who was a 3 times world Formula 1 champion. He believed the best drivers had enough understanding of how a machine worked so they could work in harmony with it. (Martin Thompson)

# Mechanical sympathy: Gagnagrunnar

## Gagnagrunnar

Dýrustu aðgerðirnar í gagnagrunnum eru I/O aðgerðir á móti diskum

## Útfærsla

Til að útfæra gagnasafnskerfi sem tryggir ákveðna eiginleika (t.d. ACID eiginleikana) og gerir það á hagkvæman hátt, þarf að skilja hvernig diskar virka

## Almennt

Almennt mikilvægt í allri forritun að hafa tilfinningu fyrir ákveðnum grundvallar stærðum til að geta gróflega áætlað afkastagetu o.s.frv.

# Mechanical sympathy: Diskar og minni

## Diskar vs. minni

Mjög dýrt að sækja af disk miðað við að sækja úr aðalminni

## Diskar

Misjafnlega dýrt að lesa af disk. Minni er random-access en diskar eru það ekki.

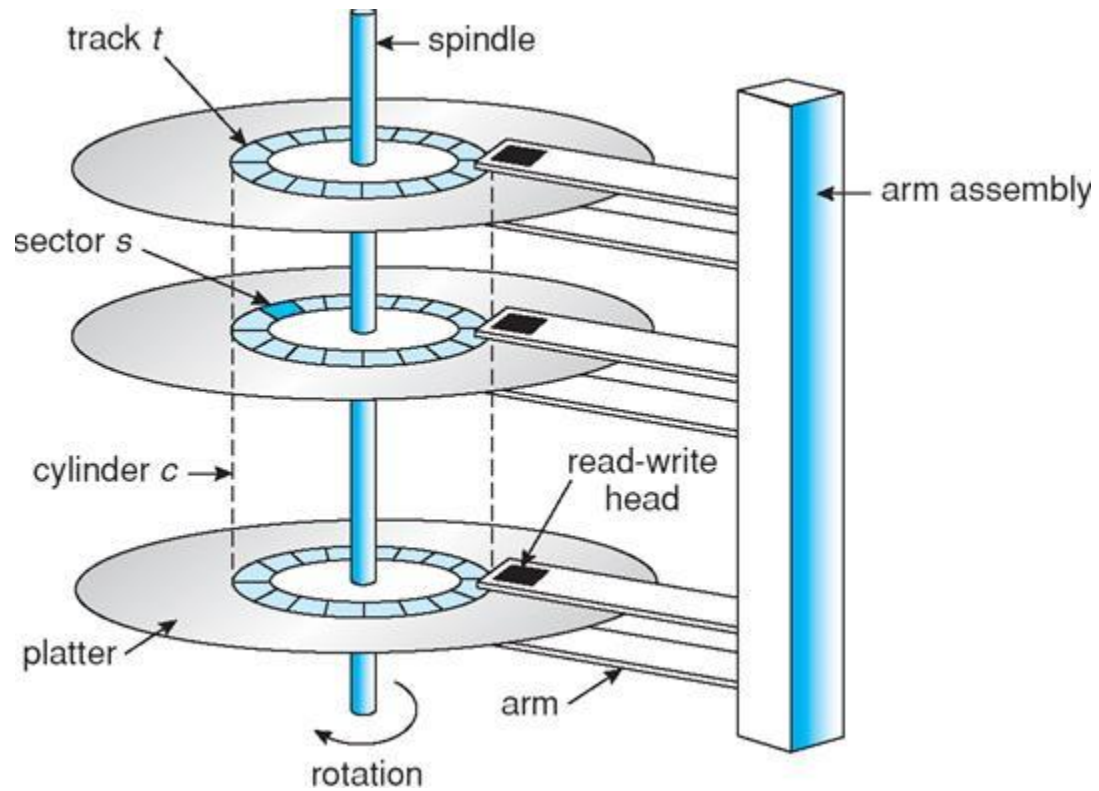
Diskar eru eins og segulbönd, ódýrt að lesa samliggjandi gögn, mjög dýrt að spóla milli staða

DRAM: 100ns to access, 20 GB/s, \$12/GB

Disk: 10ms to access, 200 MB/s, \$0.05/GB

SSD: 0.1ms to access, 250 MB/s, \$1.5/GB

# Seguldiskur



# Eiginleikar seguldiska

## **Snúningar**

Diskur snýst á föstum hraða, sem er mældur í snúningum á mínútu (RPM)

## **Plata (e. platter)**

Flöt plata með segulnæmu yfirborðslagi á annarri eða báðum hliðum

## **Rás (e. track)**

Plötu er skipt í margar sammiðja hringrásir (e. concentric rings)

## **Blokk (e. sector/block)**

Rás er skipt upp í margar jafnstórar blokkir (týpísk stærð: 512 bæti)

# Seguldiskur



# Blokkir

## Stærð á blokk (e. sector/block size)

Blokk geymir 512 bæti af gögnum (hefðbundin stærð) + bókhald um blokkina sjálfa

## Réttleiki gagna

Í bókhaldinu fyrir hverja blokk er ECC villukóði (e. error-correcting code)

ECC er uppfærður þegar blokk er skrifuð og athugaður þegar blokk er lesin

Þegar blokk er skrifuð þá eru þrjár mögulegar útkomur:

1. Allt skrifað og ECC uppfært
2. Ekkert skrifað
3. Blokk er skemmd (ECC athugun stemmir ekki)



# Blokkir

## Samskipti

Blokk er minnsta einingin í samskiptum disks og tölvu.

Til að breyta 1 bæti á disk, þá þarf að lesa viðkomandi blokk, breyta 1 bæti í henni og skrifa alla blokkina til baka.

## Vistfenging (e. addressing)

Hvernig er blokk auðkennd í samskiptum disks og tölvu?

Aðferðin sem er notuð í dag heitir LBA (Logical Block Addressing)

Blokkir eru einfaldlega númeraðar frá 0 upp í N

Reynt að númera blokkir í röð eftir því hversu "nálægt" hvor annarri þær eru

# Viðbragðstími seguldiska

## Aðgangstími (e. access time)

Seek + Rotation + Transfer

## Staðsetningartími (e. seek time)

Tíminn sem þarf til að færa leskaus frá einni rás yfir á aðra rás. Háð fjarlægð.

## Snúningstími (e. rotational delay)

Þegar haus er lentur á réttri rás, þá er *snúningstíminn* tíminn þar snúningur disksins hefur fært rétta blokk undir hausinn. Háð snúningshraða og fjarlægð.

## Flutningstími (e. transfer time)

Tíminn sem það tekur að lesa umbeðnar blokkir. Háð flutningshraða.

# Tölur í dag

## Meðal staðsetningartími

| Diskategund     | Meðaltími (ms) |
|-----------------|----------------|
| Netþjónadiskur  | 3ms            |
| Borðtölvudiskur | 9ms            |
| Fartölvudiskur  | 12ms           |

# Tölur í dag

## Meðan snúningstími

| RPM   | RPS | 1 snúningur (ms) | Meðaltími (ms) |
|-------|-----|------------------|----------------|
| 4200  | 70  | 14.28            | 7.14           |
| 5400  | 90  | 11.11            | 5.56           |
| 7200  | 120 | 8.33             | 4.17           |
| 10000 | 166 | 6.00             | 3.00           |
| 15000 | 250 | 4.00             | 2.00           |

# Tölur í dag

## Flutningstími (e. transfer time)

Flutningstími er háður flutningshraða og fjölda blokka sem á að lesa (og staðsetningu þeirra)

Týpískur hraði í dag er 800-1200 Mbit/s, þ.e. 100-150 MB/s

## Dæmi

Grf. flutningshraða upp á 100 MB/s og 512 bæta blokkstærð

Getum sótt  $100 \text{ MB} / 512 \text{ bæti} = 204800$  blokkir á sekúndu

Að lesa eina blokk tekur því  $1000 / 204800 = 0.0048 \text{ ms}$

# Svartími diska

## **Svartími (e. response time)**

Tími í biðröð (e. queue time) + aðgangstími (e. access time)

Með öðrum orðum: hversu langan tíma tekur frá því beiðni um I/O var lögð inn þar til beiðnin hefur verið afgreidd og henni svarað

## **I/O verkröðun (e. I/O scheduling)**

Hægt að spara tíma með því að raða I/O beiðnum rétt. Til mörg mismunandi reiknirit. Verður að passa upp á selti (e. starvation) og sanngirni (e. fairness)

Stundum framkvæmir stýrikerfið I/O verkröðun, en stundum er það gert í diskastýringunni (e. disk controller)

# Einfalt líkan af flæði

## I/O beiðni

Forrit  $\rightarrow$  Stýrikerfi  $\rightarrow$  Biðröð  $\rightarrow$  Diskur

## Gögn

Forrit  $\leftarrow$  Stýrikerfi  $\leftarrow$  Buffer cache í kjarna  $\leftarrow$  Cache á disk

## Blokkir vs. bæti

Stýrikerfið sér disk sem einvítt fylki af blokkum, en kjarninn býður forritum upp á skráarkerfis API þar sem skrár eru fylki af bætum

Hvernig er það útfært?

## Blokkir vs. bæti

### **Blokk sem er ekki í minni (cache miss)**

1. Forrit les 1 bæti úr skrá: `read(fd, buffer, 1);`
2. Diskur les blokkina sem inniheldur það bæti
3. Blokkin er geymd í flýtvistun kjarnans (e. kernel buffer pool)
4. Eitt bæti úr blokkinni er afritað yfir í forrit

### **Blokk sem er í minni (cache hit)**

1. Forrit les 1 bæti úr skrá: `read(fd, buffer, 1);`
2. Blokkin fannst í flýtvistun
3. Eitt bæti úr blokkinni er afritað yfir í forrit



# Skráarkerfi

## API

Skráarkerfi býður upp á abstraction þar sem skrár eru fylki af bætum. Býður einnig upp á nafnakerfi sem er tré, innri hnútar eru skráarsöfn, lauf eru vísar á skrár.

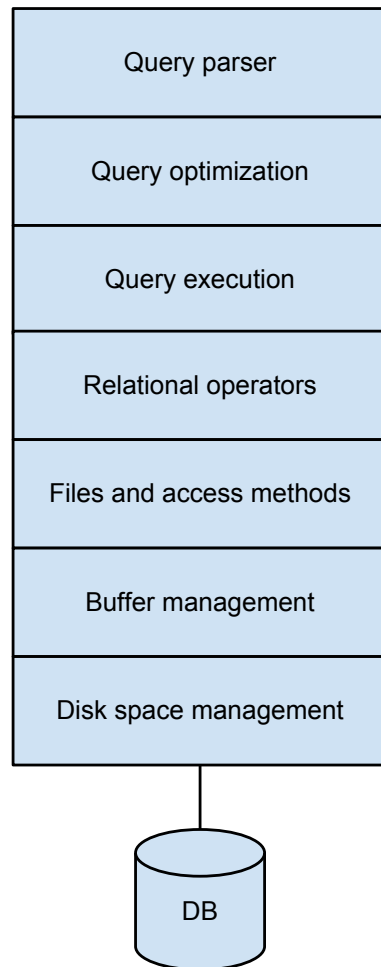
## Blokkir

Skráarkerfi notast við aðra stærð af blokkum en diskar. Ein blokk í skráarkerfi samanstendur af nokkrum samliggjandi diskablokkum. Típísk stærð á skráarkerfisblokk er 4KB eða 8KB.

## Flýtvistun í stýrikerfi

Flýtvistun stýrikerfisins byggir á skráarkerfisblokkum (ekki diskablokkum)

# Uppbygging gagnasafnskerfis



# Gagnasafnskerfi

## Skrár

Gagnasafnskerfi hefur sitt eigið hugtak yfir skrár. Skrá í gagnasafnskerfi samanstendur af blokkum, stundum sagt síðum (e. pages).

## Blokkir

Gagnasafnskerfi hefur einnig sína eigin stærð af blokkum. Oftast valið þannig að ein blokk í gagnasafnskerfinu rúmast fyrir í einni blokk í skráarkerfinu (hægt að lesa blokkina í einu lagi)

## Gagnasafnskerfi án skráarkerfis

Sum gagnasafnskerfi geta unnið beint á disk (raw device access) án þess að fara í gegnum skráarkerfi. Fara þá alveg fram hjá flýtvistun í kjarnanum, ákvörðunum

í skráarkerfi um staðsetningu á blokkum o.s.frv. Flest gagnasafnskerfi (t.d. PostgreSQL og MySQL) nota samt bara skrár í skráarkerfinu.

# Skrár í gagnasafnskerfi

## Gagnaskrá (e. data file)

Geymir raðir (e. rows) í töflu. Skránni er skipt í margar síður, hver síða geymir ákveðinn hluta af röðunum.

Röð er auðkennd með svokölluðu *Record ID*, oft bara (Page ID, Item ID)

Útfærsluatriði: raðir með fasta lengd (fixed-length) eða breytilega lengd (variable length)

## Flýtvísisskrá (e. index file)

Geymir einnig raðir (e. rows) en mismunandi skipulag eftir tegund flýtvísis (hakkavísir, trévísir)