Reiknirit á straumum

Páll Melsted
HÍ

HÁSKÓLI ÍSLANDS
IÐNAÐARVERKFRÆÐI-, VÉLAVERKFRÆÐI- OG TÖLVUNARFRÆÐIDEILD

Einfalt vandamál

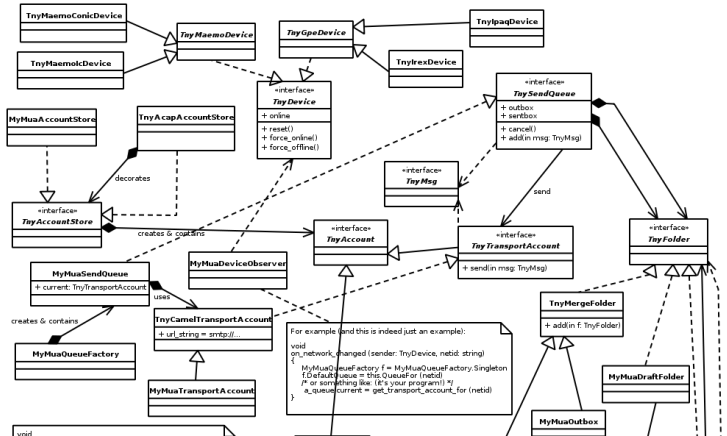
Gefin inntak með stökum

Finnum fjölda ólíkra staka

Inntakið 1, 3, 5, 10, 3, 1
4 ólíkar tölur

Einnar línu lausn

- Bash: `sort | uniq | wc –l`
- Python: `len(set(int(x) for x in sys.stdin))`
- C++: `cout << set<int>((istream_iterator<int>(cin)),istream_iterator<int>()).size();`
- Java:



Vandamál?

Allar lausnirnar geyma inntakið í minni

Hvað ef minni er af skornum skammti?

Þurfum við að geyma allt í minni?

Já

Dæmi

- IPS (Intrusion Prevention/Detection System)
  - Situr á netinu og skoðar pakka
  - 10 Gbps hraði, ca 40 ns per pakka
  - venjulegt minni, 100 ns sóknartími

- Höfum kannski 128 Mb af minni á lausu

Dæmi

- IPS (Intrusion Prevention/Detection System)
  - Situr á netinu og skoðar pakka
  - 10 Gbps hraði, ca 40 ns per pakka
  - venjulegt minni, 100 ns sóknartími

- Höfum kannski **128 Kb** af minni á lausu

- H
- C n
- K

Hv a?

Straumar – nýjar reglur

Inntakið er straumur af n stökum

Takmarkað innra minni m.v. n,

Fáum hvert stak einu sinni, afgreiðum áður en við skoðum næsta stak.

Getum aldrei farið aftur á bak í straumnum

Slæmar fréttir

- Ekki hægt að finna rétt svar
- Hvað ef við sættum okkur við nálgun?
  - í mesta lagi 10% frá réttu svari?
  - ekki hægt!
- Verðum að gefa nálgun með góðum líkum

Hakkaföll

- Hakkafall tekur sem inntak bitastreng og skilar tölu sem úttaki
  - úttakið á að vera “ófyrirsjáanlegt”
  - slembið og jafndreift
- Þurfum ekki örugg hakkaföll eins og md5 eða sha-1
  - Einföld föll $H(x) = ax + b \pmod{2^{64}}$
  - MurmurHash3, CityHash, SipHash

Slembin nálgunarlausn

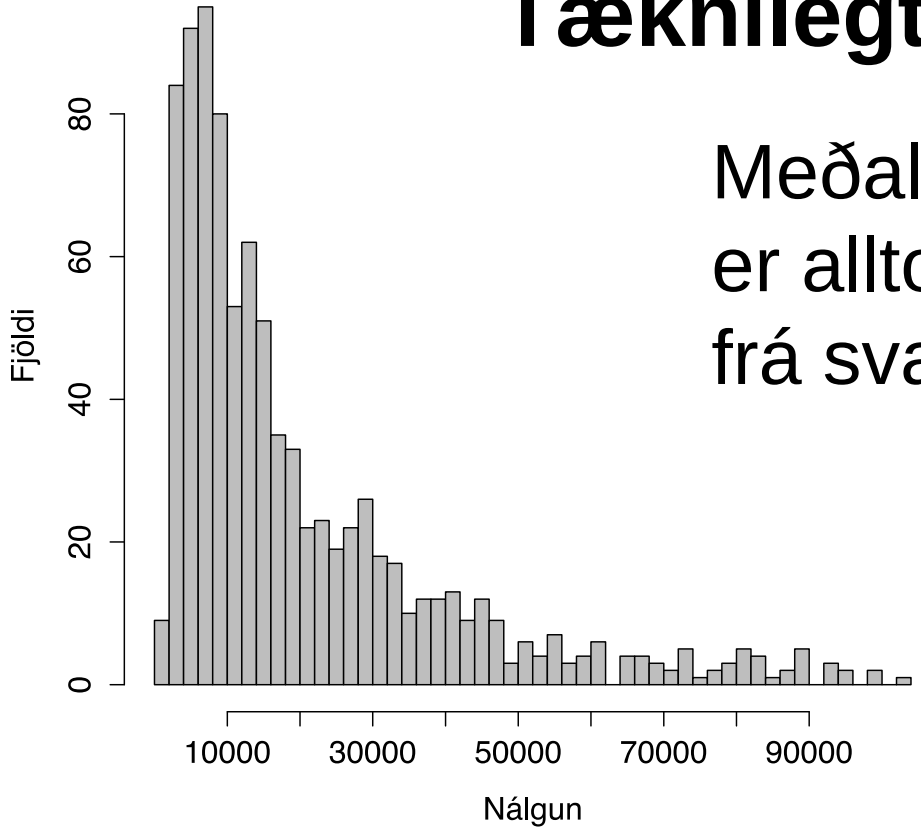
- Veljum hakkafall $h : [2^{64}] \to [0, 1]$
  úthlutar 64-bita heiltölum double tölu frá 0 til 1
- Fyrir hvert inntak x í straumnum
  - reiknum $h(x)$
  - geymum minnsta hakkagildið sem við höfum séð
- Skilum svarinu $\frac{1}{\min_x h(x)} - 1$

Slembin nálgunarlausn

- Hakkafallið gerir tvo hluti
  - Ef x er endurtekið þá fæst sama hakkagildi og svarið breytist ekki
  - Ef hakkafallið er gott þá eru öll h(x) jafnlíkleg
  - Fyrir n ólík gildi verður meðaltalið af minnsta gildinu

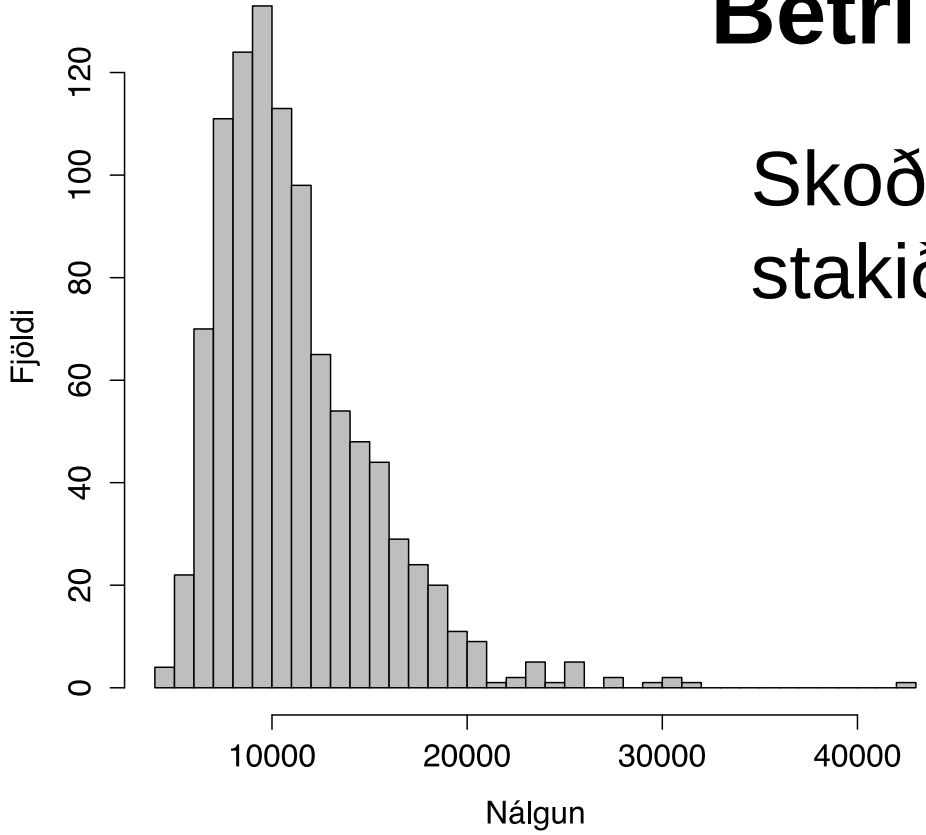
$$\frac{1}{n+1}$$

Tæknilegt vandamál



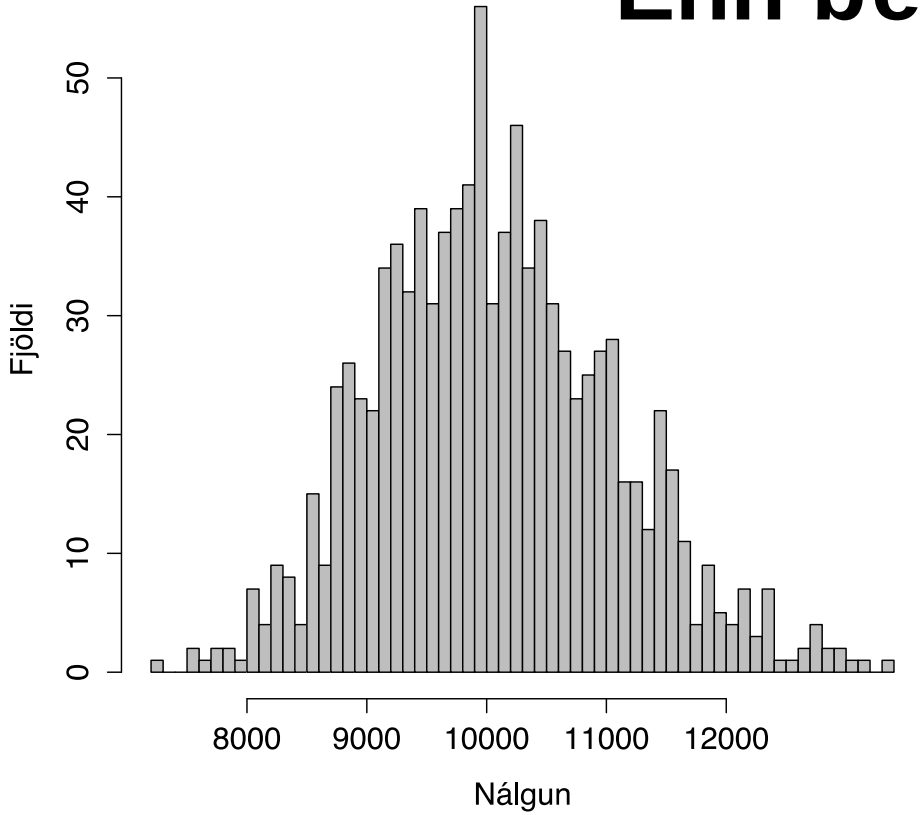
Meðaltalið er rétt en dreifingin er alltof mikil. Getur verið langt frá svarinu.

Betri lausn



Skoðum tíunda minnsta stakið, meðaltal $\frac{10}{n+1}$

Enn betri lausn



Skoðum hundraðasta minnsta stakið

69% tilrauna lenda innan 10% skekkjumarka

Almenn lausn

- Til að fá $1 \pm t$ nákvæmni þarf að finna $\frac{1}{t^2}$ minnstu stökin.
- Einföld útfærsla: notum tvíleitartré til að halda utan um k minnstu h(x) gildin
- Svarið breytist lítið eftir fyrstu stökin
- Langmestur tími fer í að reikna h(x)

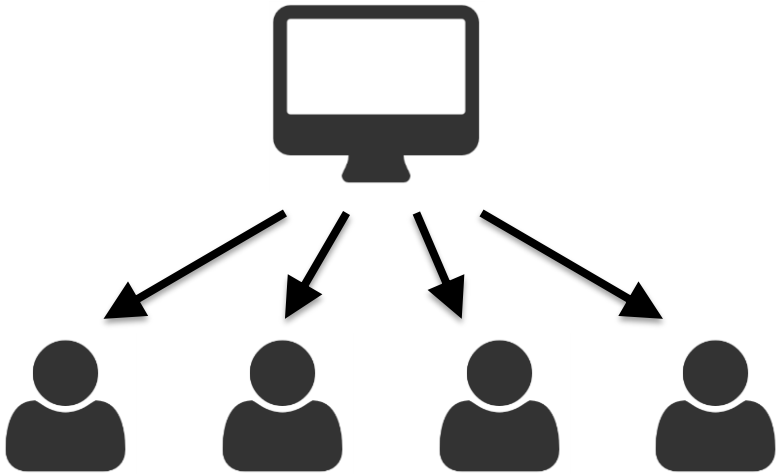
Útfærsla

- Útfært í C++, straumar og einföld útgáfa með BST
- Inntakið 100M tölur, 10M ólíkar
  - strauma útfærsla: **21 sek**, **1.2M** minni, 0.4% skekkja
  - nákvæm útfærsla: 180 sek, 465M minni
  - sort | uniq | wc –l : 95 sek, 238M minni

Aðrar hagnýtingar

- Query optimization í gagnagrunnum
- Þarf mat á hversu margar niðurstöður fást fyrir hverja leit
  - 1-2% skekkja í svari mun ekki breyta niðurstöðum
  - Þarf að vera hraðvirkt og má ekki taka mikið auka minni

Aðrar hagnýtingar



Dreifð kerfi

- Látum “notendur” telja fyrir okkur
- Skila aðeins samantekt
- Þurfum ekki að sjá gögnin
- Hægt að útfæra í map-reduce

Fleiri straumareiknirit

Miðgildi, 10% gildi og histogram

Heavy hitters

Frequency moments

Samantekt

- Hakkaföll líkja eftir slembibreytum en gefa alltaf sömu niðurstöðu
- Þurfum ekki alltaf nákvæmt svar
  - Með því að leyfa nálgun fæst betra reiknirit
- C++ kóði er á http://github.com/pmelsted/streaming-talk