Lucene/Solr samsøgning og skalering

BibTekKonf 2013

Toke Eskildsen, Statsbiblioteket

Overblik

- TF-IDF-rankering
- Samsøgning
- Hverdagshardware
- Mediestream (belastning)
- Avisdigitalisering (processering)
- Netarkivet (størrelse)

Open Source inverterede index

Elasticsearch

TF-IDF og rankering

- Dokumentvægt (<doc boost="2.5">)
- Feltvægt (qf=author^4.0)
- Termvægt (foo^1.5)

$$score(q,d) = coord(q, d) * queryNorm(q) * \sum(tf(t \text{ in } d) * idf(t)^2 * t.getBoost() * norm(t, d))$$

https://wiki.apache.org/solr/UpdateXmlMessages#Optional_attributes_on_.22doc.22
https://wiki.apache.org/solr/ExtendedDisMax#qf_.28Query_Fields.29
https://lucene.apache.org/core/2_9_4/api/core/org/apache/lucene/search/Similarity.html

Samsøgning

```
score(q,d) = coord(q, d) * queryNorm(q) *
∑( tf(t in d) * idf(t)² * t.getBoost() * norm(t, d) )
```

- Udregn fælles IDF for hver term på tværs af alle index
- For hver søgeterm, sæt boost til $\frac{idf(t)^2}{IDF(t)^2}$

“the economist” → “the"^0.139 "economist"^0.373”
→ “the"^1.336 "economist"^1.085”

summa
delightful search

Relateret: https://issues.apache.org/jira/browse/SOLR-1632

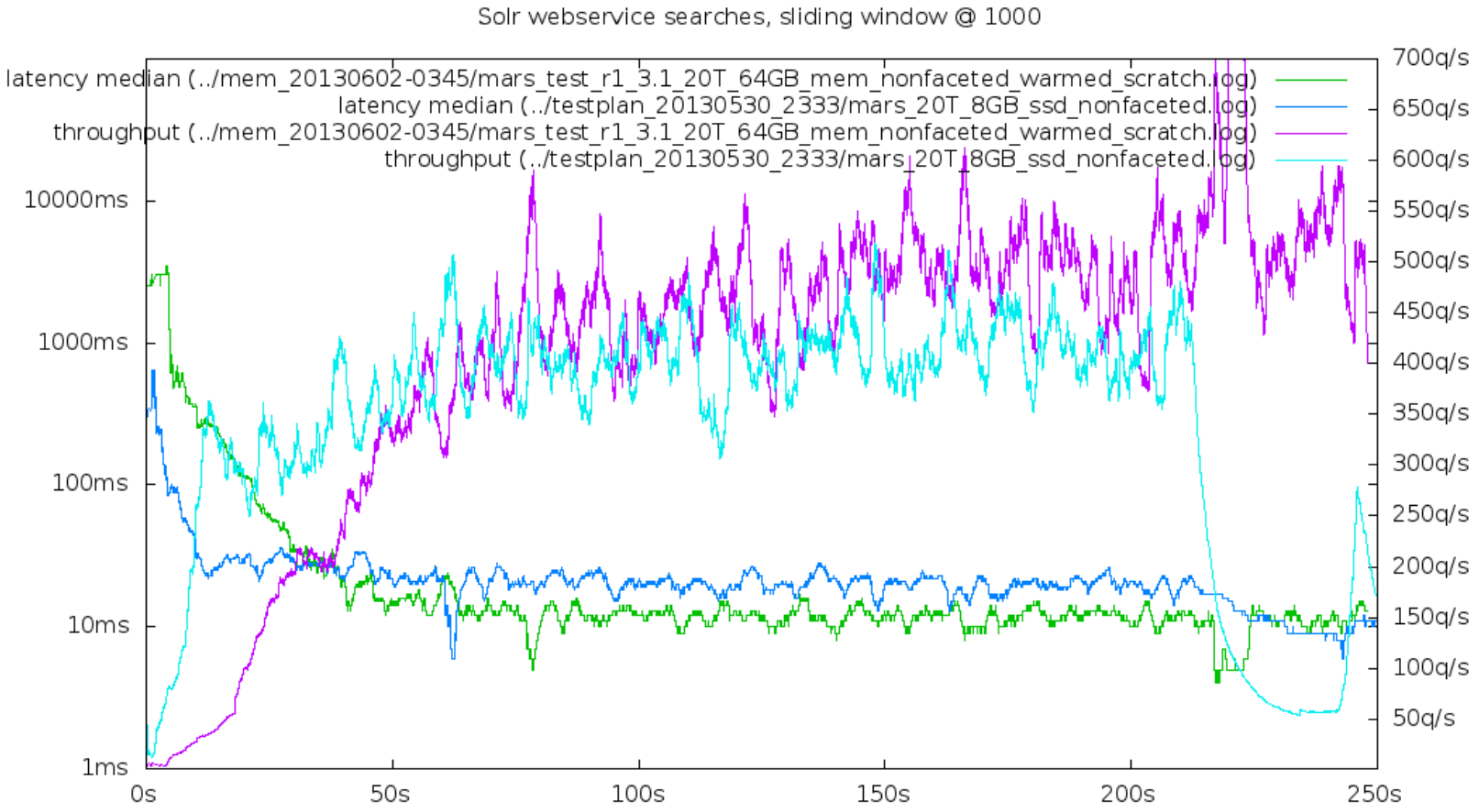
De nemme korpus

- Veldefinerede felter
- Autoritativt indhold
- Få millioner poster
- Lav mængde forespørgsler
- Simple søgninger
- Natlige opdateringer

Hverdagshardware

- Index på hver server
  - 3 stk, samlet 156 GB / 24 M poster
  - 40.000 søgninger/dag, 3-4/sek i peak
  - Facetter: 16 felter, 16 M termer, 144 M referencer
- Hardware
  - 4 core 2.5 GHz Xeon L5420
  - 16 GB RAM, heraf 6 GB fri til cache
  - 500 GB SSD

16 GB RAM til 156 GB index?

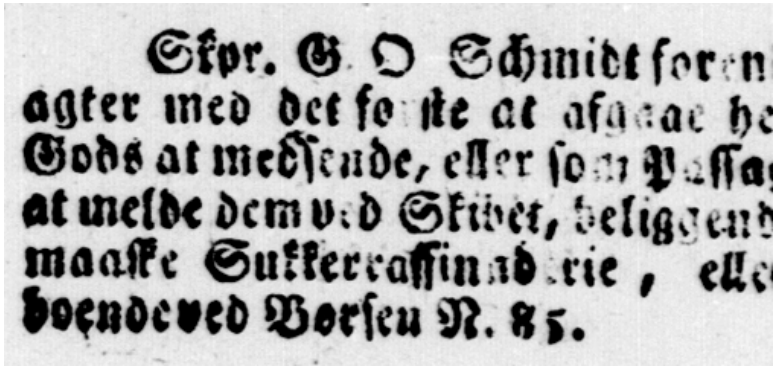


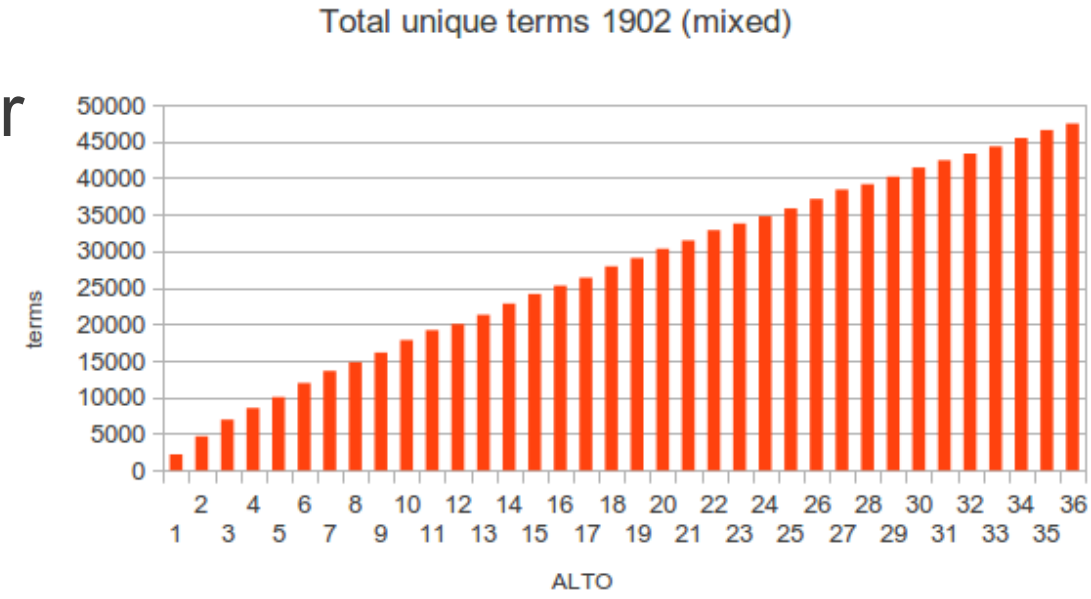
MEDIESTREAM

- 2.2 GB index, 1.3 M simple poster
- Simple søgninger
- Simple metadata
- Finkornede rettigheder
- Potentielt højt peak

Avisdigitalisering

- 32 M avissider
  - 480 M artikler
- Anslået 1.5 TB index
- Ringe fraktur OCR
  - 240 K - 40 G termer
- Ordbog?
- Fuzzy search?





Fraktur OCR

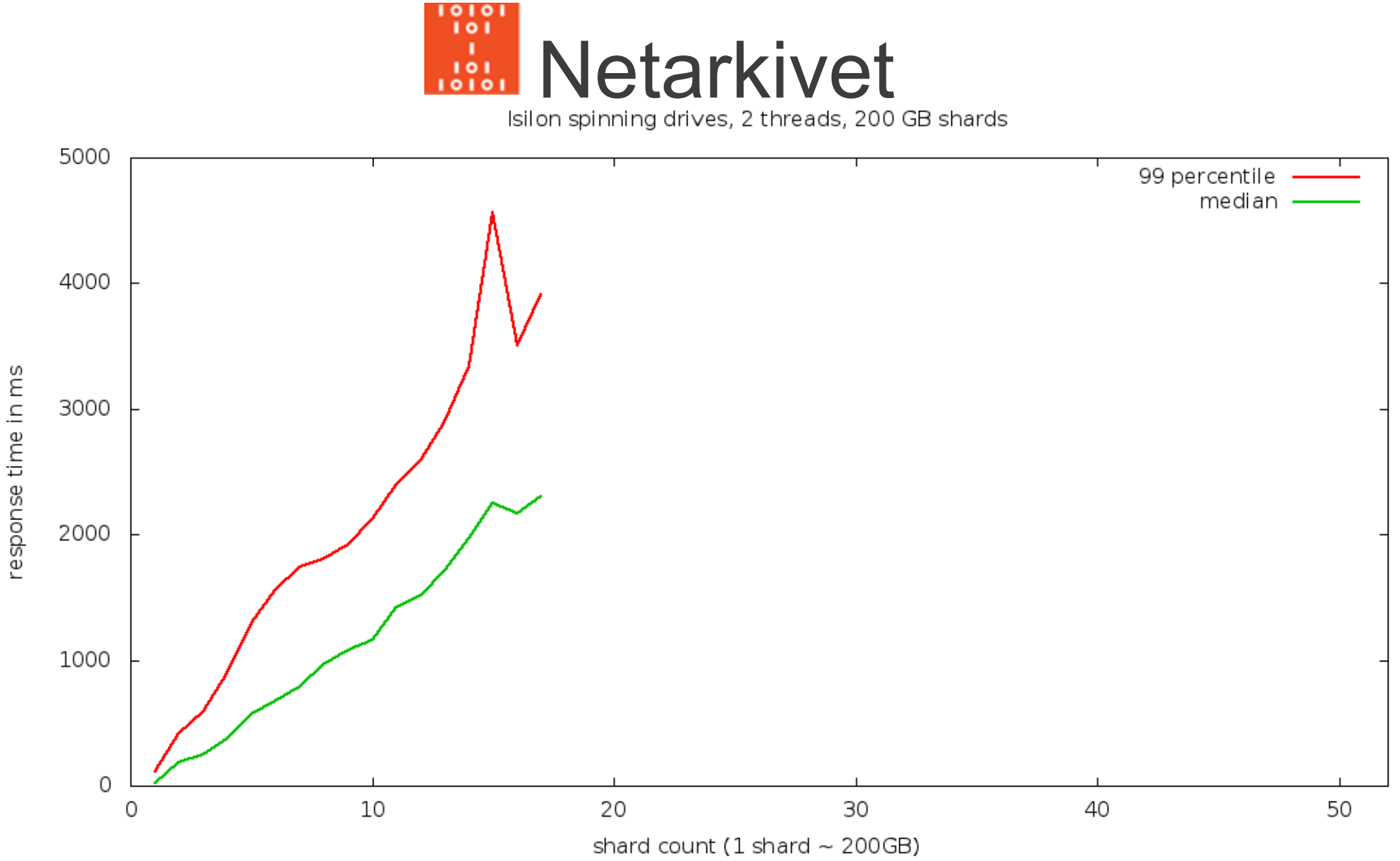
Zebeb«z«r. nozle R<^t?'5LSsqer. it« ved lld^tanten bl!>rkoin?>e,Mahoni« N rskrin, en Vask,mask,ne m d id«ekk N, il Tro med sortstribet H stehaarS Betrak, Piedestlaer to tyik-sk- Guvttpper, endeel S«n,e« kl-der, to Madftr, 2 Punseb.ller, 6 stor« Fade, ig blaae oq hvide Tallerkener, z blaae vg hvid« Ter, in.r. Hvem scw t>ar imodtaqet ommødre Toi eller om s mm« kan giv- nogen

Netarkivet

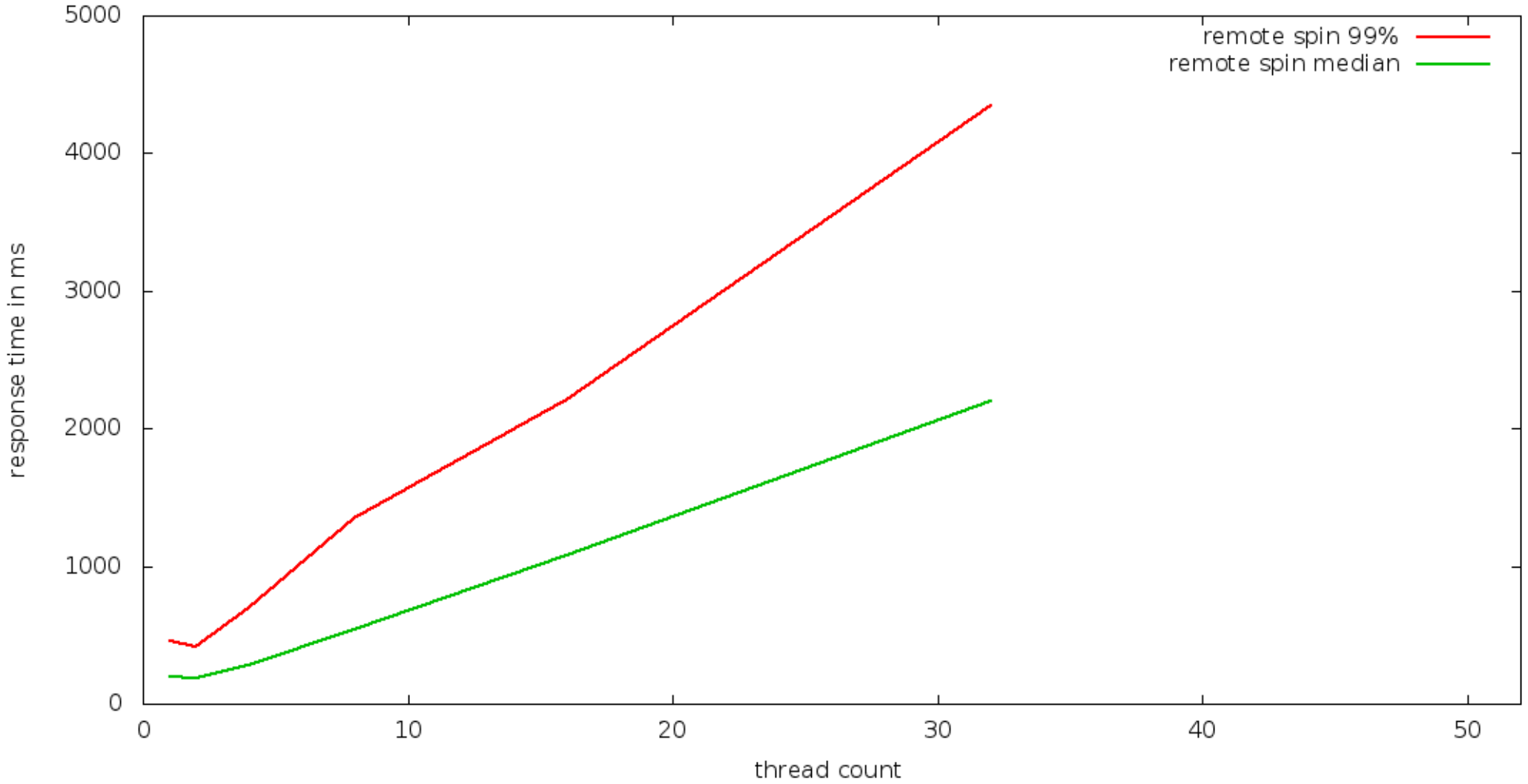
- 10 G poster (10.000.000.000)
- 20 TB index (fra 372 TB rå data)
- +4 TB index/år
- Få brugere
- Gruppering, facetering

Under Construction CC http://bestandworstever.blogspot.dk

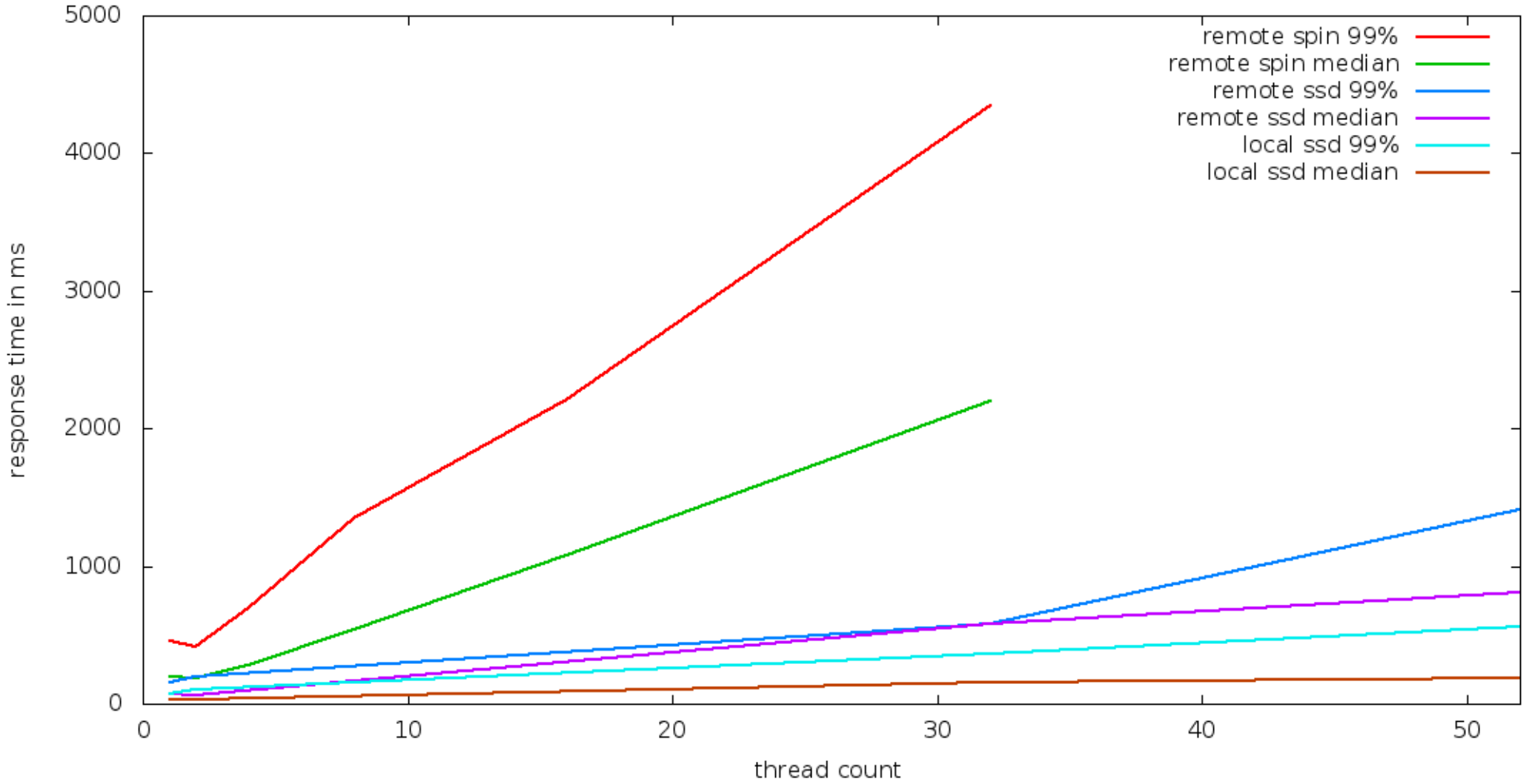
Netarkivet


Isilon spinning drives, 2 threads, 200 GB shards
99 percentile
median
response time in ms
shard count (1 shard ~ 200GB)
0
1000
2000
3000
4000
5000
0
10
20
30
40
50

Netarkivet


Response times for 2 shards of 200GB
remote spin 99%
remote spin median
response time in ms
thread count
0
1000
2000
3000
4000
5000
0
10
20
30
40
50

Netarkivet


Response times for 2 shards of 200GB
remote spin 99%
remote spin median
remote ssd 99%
remote ssd median
local ssd 99%
local ssd median
response time in ms
thread count
0
1000
2000
3000
4000
5000
0
10
20
30
40
50

Spørgsmål?

Toke Eskildsen, Statsbiblioteket
http://sbdevel.wordpress.com
te@statsbiblioteket.dk