

# 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 ranking

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

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

[https://wiki.apache.org/solr/UpdateXmlMessages#Optional\\_attributes\\_on\\_.22doc.22](https://wiki.apache.org/solr/UpdateXmlMessages#Optional_attributes_on_.22doc.22)  
[https://wiki.apache.org/solr/ExtendedDisMax#qf\\_.28Query\\_Fields.29](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](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”<sup>0.139</sup> “economist”<sup>0.373</sup>  
→ “the”<sup>1.336</sup> “economist”<sup>1.085</sup>



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

# De nemme korpus

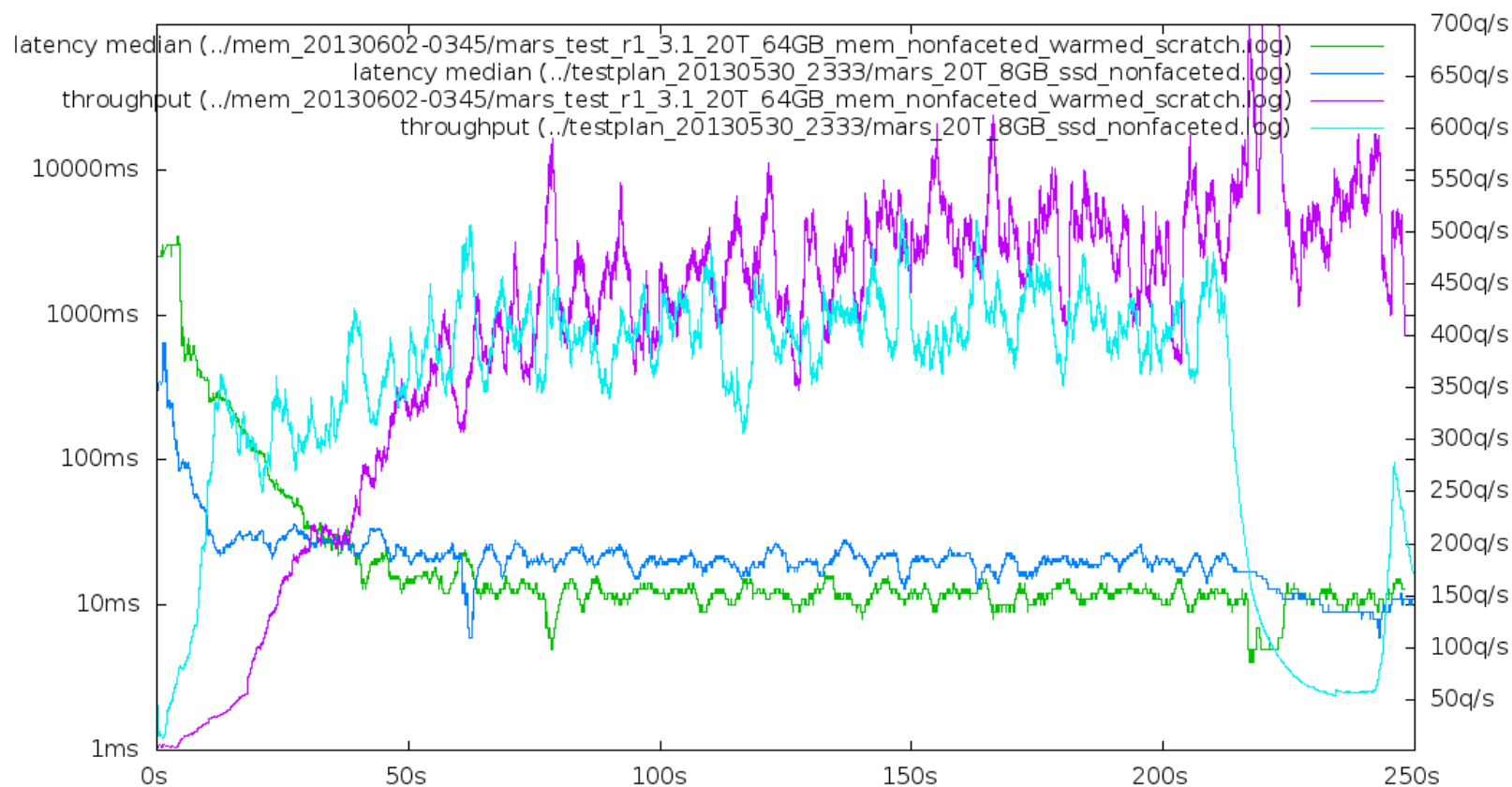
- Veldefinerede felter
- Autoritativt indhold
- Få millioner poster
- Lav mængde forespørgsler
- Simple søgninger
- Natlige opdateringer

# 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?

Solr webservice searches, sliding window @ 1000



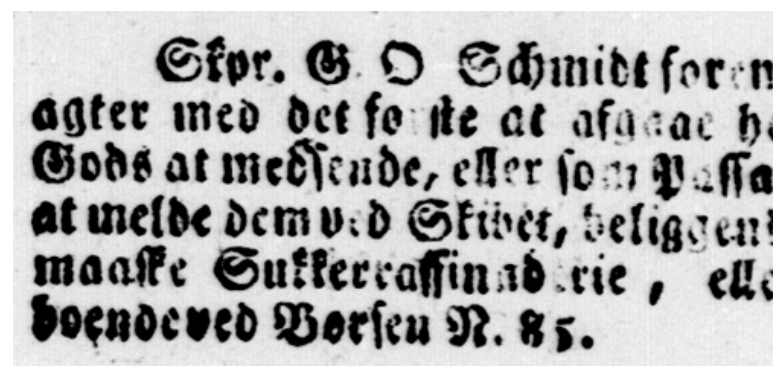


# M=DI\_STREAM

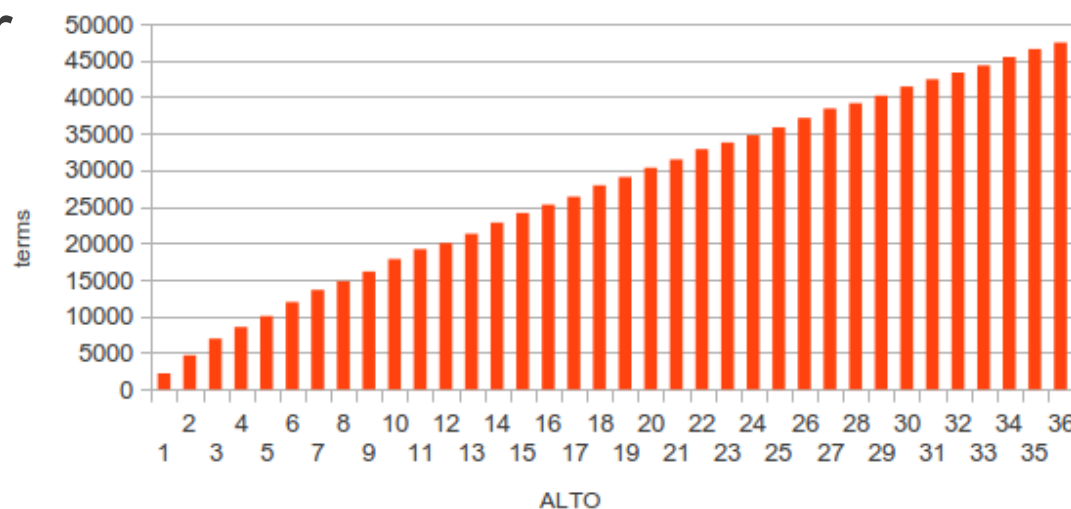
- 2.2 GB index, 1.3 M simple poster
- Simple søgninger
- Simple metadata
- Finkornede rettigheder
- Potentielt højt peak

# Avisdigitalisering

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



Total unique terms 1902 (mixed)



# Fraktur OCR

Zebeeb«z«r. nozle R&lt;^t?'5LSsqer. it« ved  
Ild^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.Iler, 6 stor« Fade, ig blaae  
og hvide Tallerkener, z blaae vg hvid« Ter, in.r.  
Hvem scw t>ar imodtaget 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

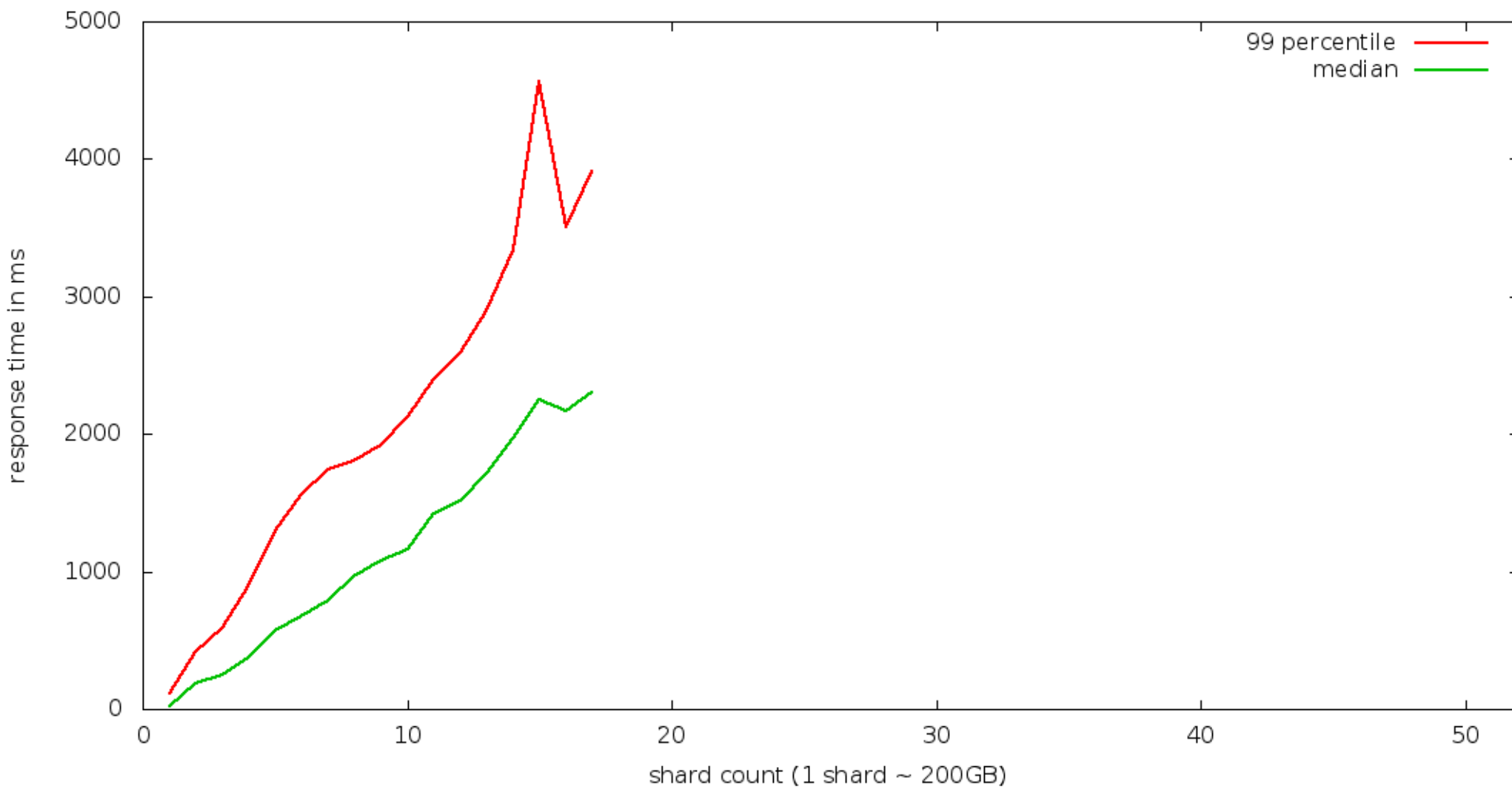


Under Construction CC <http://bestandworstever.blogspot.dk>



# Netarkivet

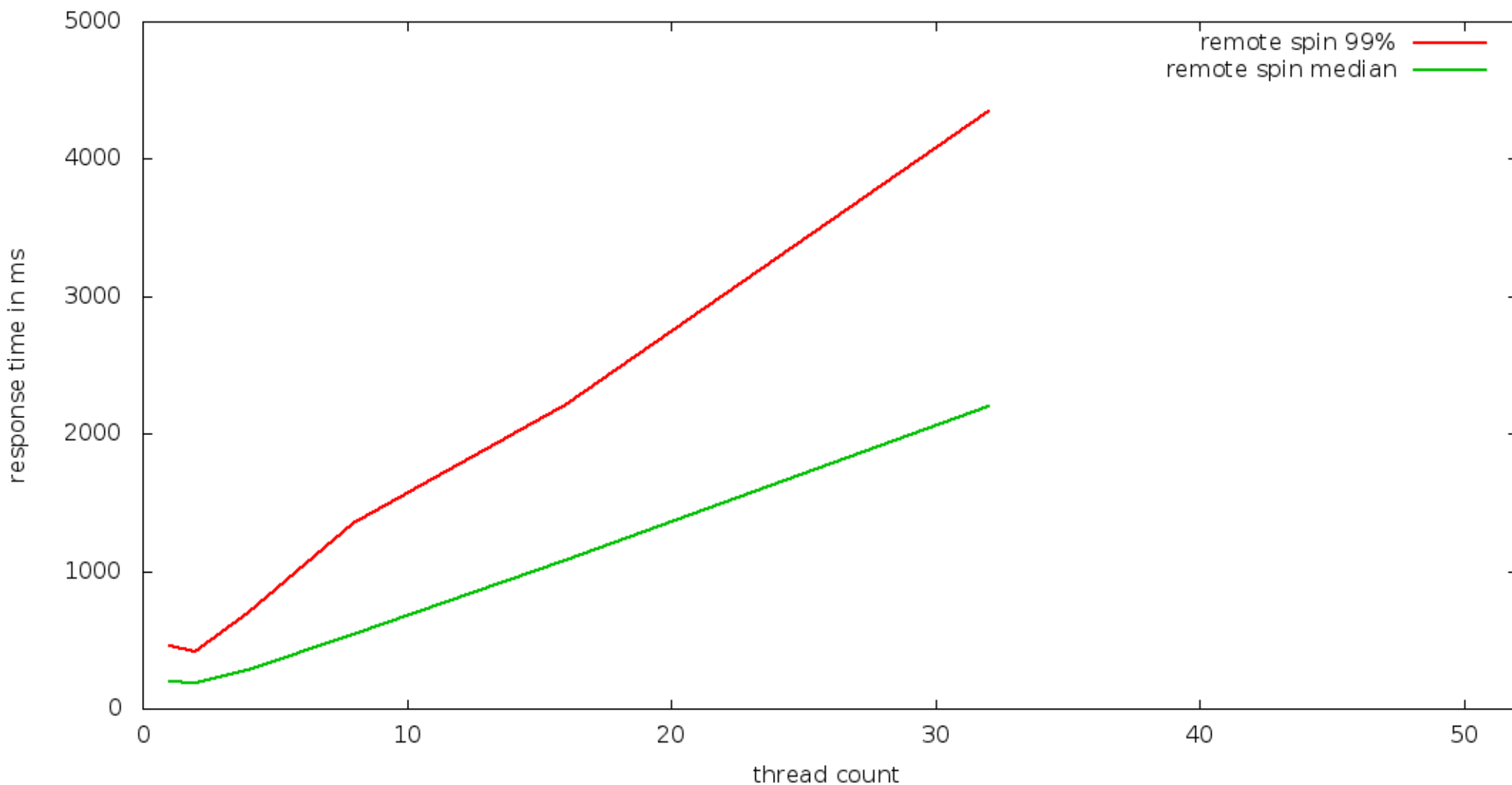
Isilon spinning drives, 2 threads, 200 GB shards





# Netarkivet

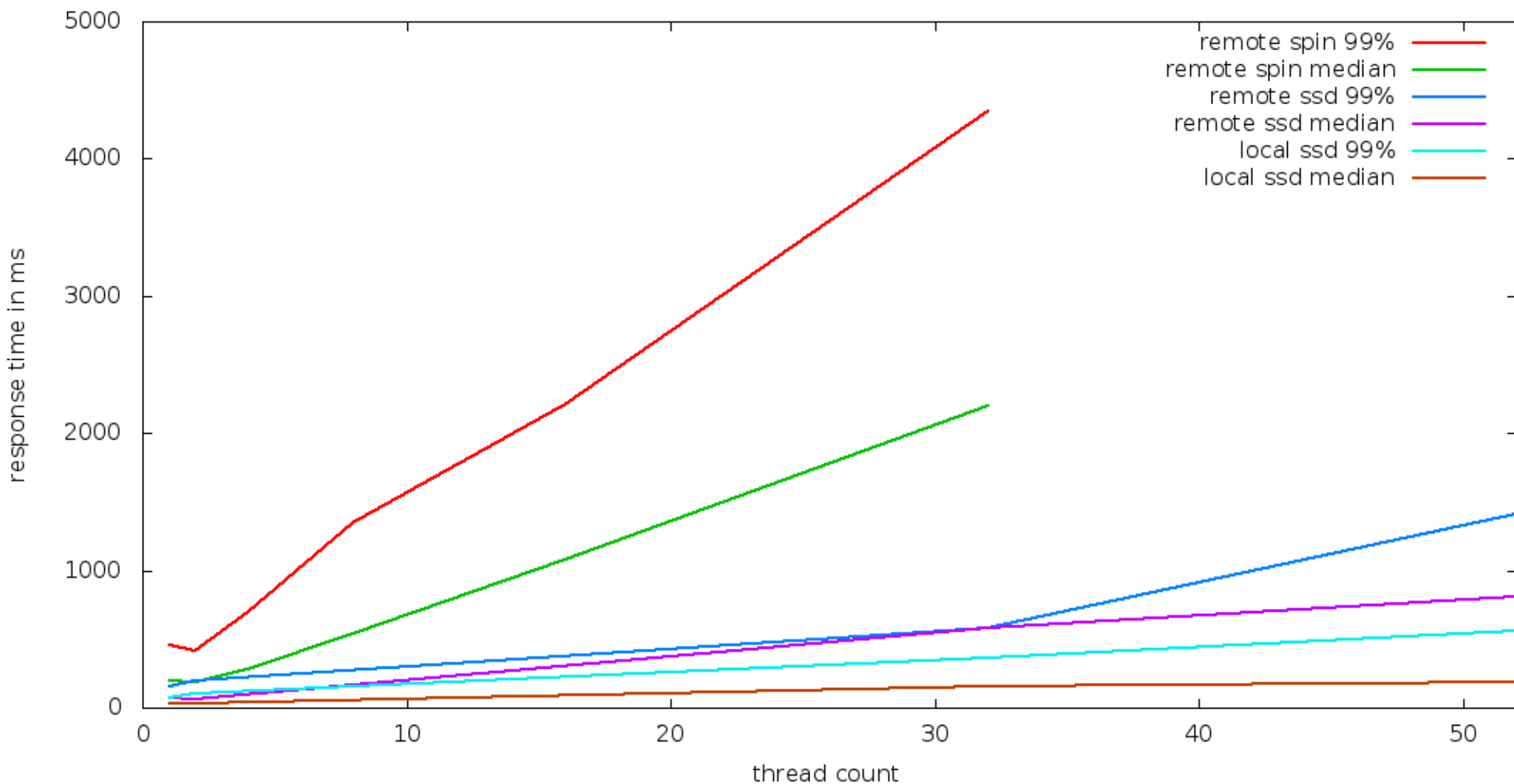
Response times for 2 shards of 200GB





# Netarkivet

Response times for 2 shards of 200GB



# Spørgsmål?

Toke Eskildsen, Statsbiblioteket  
<http://sbdevel.wordpress.com>  
[te@statsbiblioteket.dk](mailto:te@statsbiblioteket.dk)