Webscale for a small web

- Since 2005
- All *.dk websites
- 4 times/year
- ~100 special sites harvested daily
- Explicitely stated by law that everything public can and must be harvested

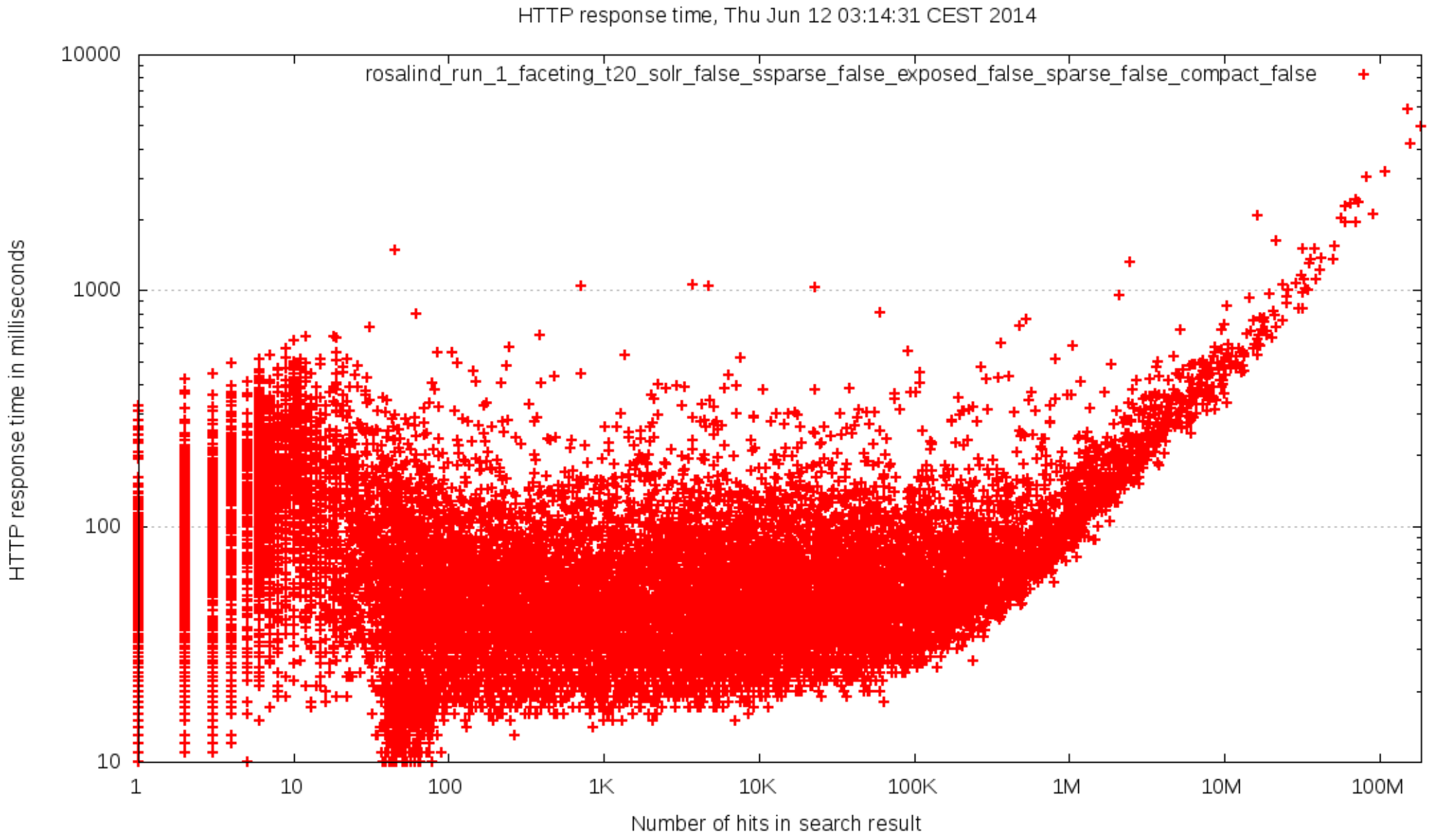
ELAG 2014 quickly hacked lightning talk, Toke Eskildsen, te@statsbiblioteket.dk, State and University Library, Denmark

Numbers

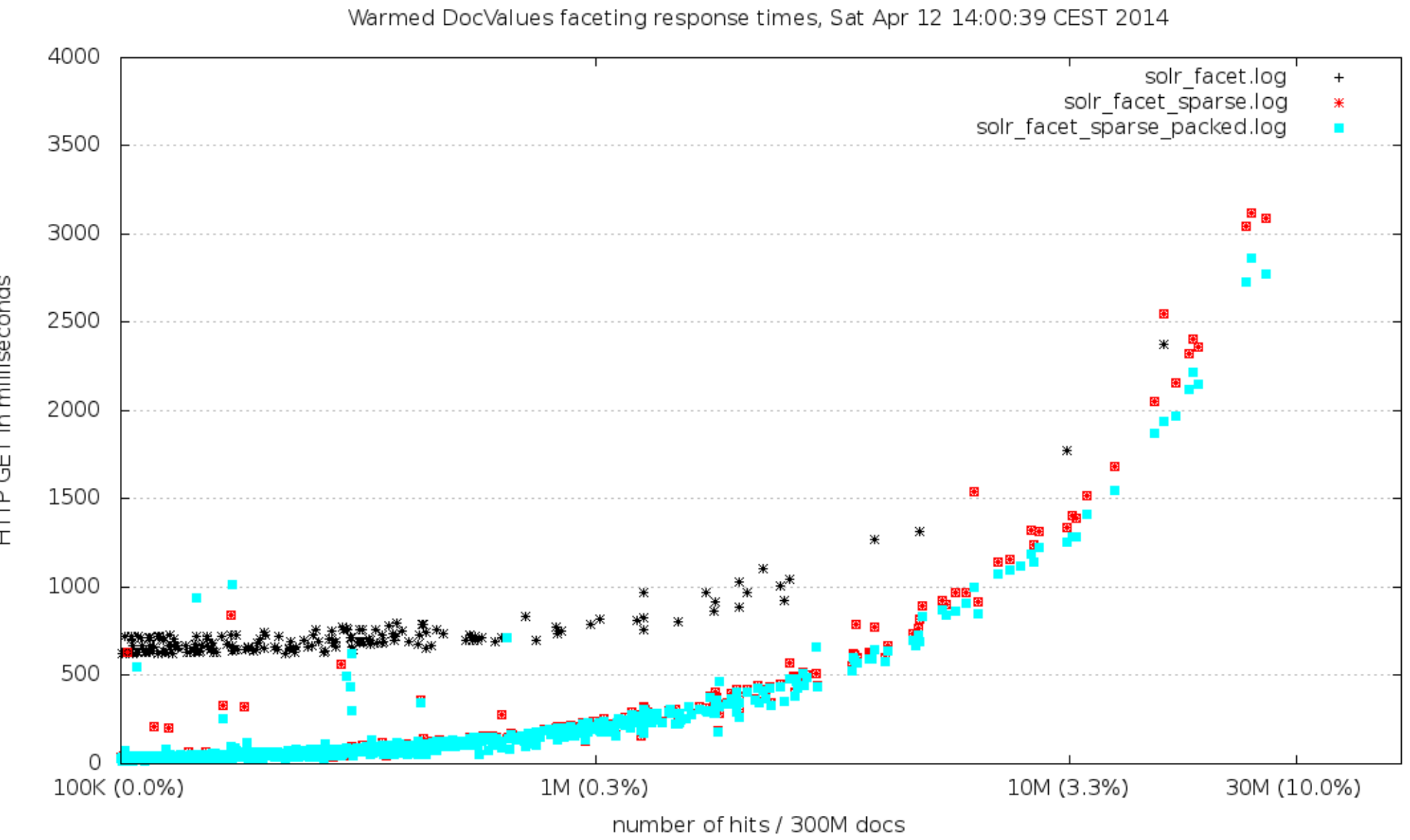
- Currently 370TB+ / 8-10B web resources
- Estimated final index: 20-24TB
- Indexing: 24 core / 256GB RAM / 5TB SSD
- Searching: 16 core / 256GB RAM / 24TB SSD
  - Cost: ~£12,000
- 1 optimized shard / SSD (900GB, 300M docs)
  - Build time / shard: ~8 days
- 1 solr / shard, connected with Solrcloud

https://github.com/netarchivesuite/netsearch
(based on https://github.com/ukwa/webarchive-discovery)

20 threads, 3.6TB, simple searches



1 thread, faceting, 900GB, URL-field



Remember the milk!