HATHI TRUST
a shared digital repository

Scaling Full-text Search

Cory Snavely
Library IT Core Services manager
University of Michigan

September 2011

HathiTrust project profile

- Launched October 2008
- 48 ~~29~~ member institutions and growing
- primarily Google-scanned materials but also other sources (e.g. Internet Archive, and increasingly content digitized by partners themselves)
- 9.6 ~~6.7~~ million volumes, 350 pages average
- 430 ~~250~~ terabytes in two US instances

Full-text search overview

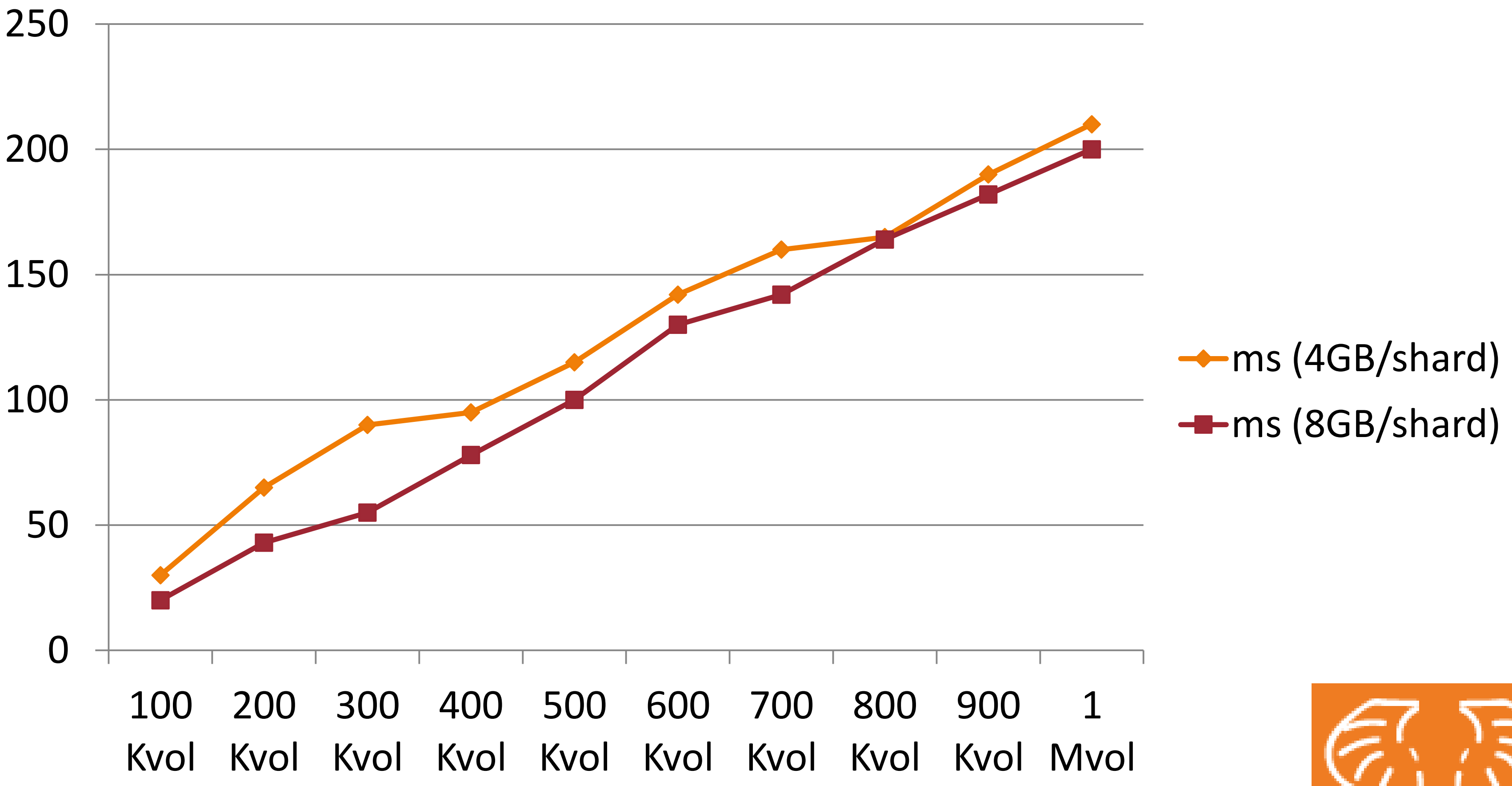
Determine, through experimentation, the optimal…

- …number and size of index “shards”.
- …amount of server memory for IO cache and JVM.

Evaluate whether rotational media can provide, given the above, …

- …acceptable query throughput.
- …a manageable update process, assuming continuous availability at two sites.

Query response on rotational media



www.hathitrust.org

Cache is king!

More memory improves response time linearly with index size, so…

- …choose a shard size that provides acceptable response time, and divide up the index evenly.
- …buy as much memory as possible!

Current configuration: 5TB index, 10 shards, ~3 shards and 72GB RAM per server.

Snapshots make daily releases a snap

- Index new materials throughout the day.
- When queue is empty, optimize, check, and quiesce.
- 3am: take snapshot of index and begin synchronizing from Michigan to Indiana.
- 6am: check that synchronization is complete – it should be - and release simultaneously in both sites or calmly notify staff.
- Rinse and repeat.

Boosting performance with SSD

Given that…

- …99th percentile of query response is > 1 second,
- …rotational media has nothing more to give, and
- …some parts of the index are repeatedly read,

we are looking into a DRAM- or SSD-based NFS cache as potentially the simplest way to reduce response time of edge cases without perverting our simple and elegant indexing and release workflow.

HATHI TRUST
a shared digital repository

me: Cory Snavely
csnavely@umich.edu

primary IR research: Tom Burton-West
tburtonw@umich.edu

http://www.hathitrust.org/large_scale_search