Experiment Results for Tiny Google

Mehrnoosh Raoufi and Alireza Samadianzakaria

In this document we are showing the results of the experiments on the tiny google system. In the first section we present the results for the socket based implementation and on the second part we are measuring the differences for the hadoop implementation.

Socket Based Implementation

We did two experiments for this part. We used different processes in the same machine and changed the number to see the effect, we also used in a real distributed system and increased the number of elements.

The following table is the result of running using element machines.

Number of Helper	Run1 (ms)	Run2 (ms)	Run3 (ms)	Avg (ms)
1	12940	12640	12551	12710
2	2518	2205	2098	2273
3	1247	1118	1158	1174
4	1204	1392	733	1109
5	688	474	458	540

Clearly, by increasing the number of machines, the speed significantly increases. For instance using only these the trials the result of two samples t test between two machine and five machines is 0.000446.

The following table is the result of running on a single machine (personal laptop).

Helpers #	Docs #	Response Time (ms)
2	11	29223
3	11	37131
4	11	35292
5	11	33364
6	11	34481

As the table indicate, response time increases at some point by increasing the number of helpers (from 2 helpers to 3). The reason is that the overhead of communication and distributing the documents among the helpers is relatively high in comparison to computation time of indexing. However, at some points that we have 5 helpers for example, this overhead cost is amortized. The average response time for 34s for running on a single machine.

Hadoop

These are the results for running the hadoop indexer using combiner vs not using combiner.

Trial	Vanilla (ms)	Combiner (ms)
1	22412	28614
2	19278	28822
3	19341	50306
4	22385	22258
5	31675	31539
6	34753	22426
7	47237	24239
Avg	28154	29743

Unfortunately due to high overhead of the hadoop cluster bad configuration; the result without combine is better than the one with combiner however it is not significantly better and the variance is a lot. We expect this result becomes more significant in case of having larger dataset.

- *Both group members have contributed equally to this project.*