String Matching Using Parallel Implementation

Presenter - Jiang Wu

CSE 633 2018 Spring

Instructor: Dr. Russ Miller

5/10/2018

Problem Description

- A short pattern P of length m
- A large text file contains a large amount of lines of strings of length n
- String matching is finding one or more exact occurrences of P in the file

- String searching perform important tasks in many applications
  - database operation
  - DNA sequencing
  - searching engine
  - library system

- Important to speed up the searching

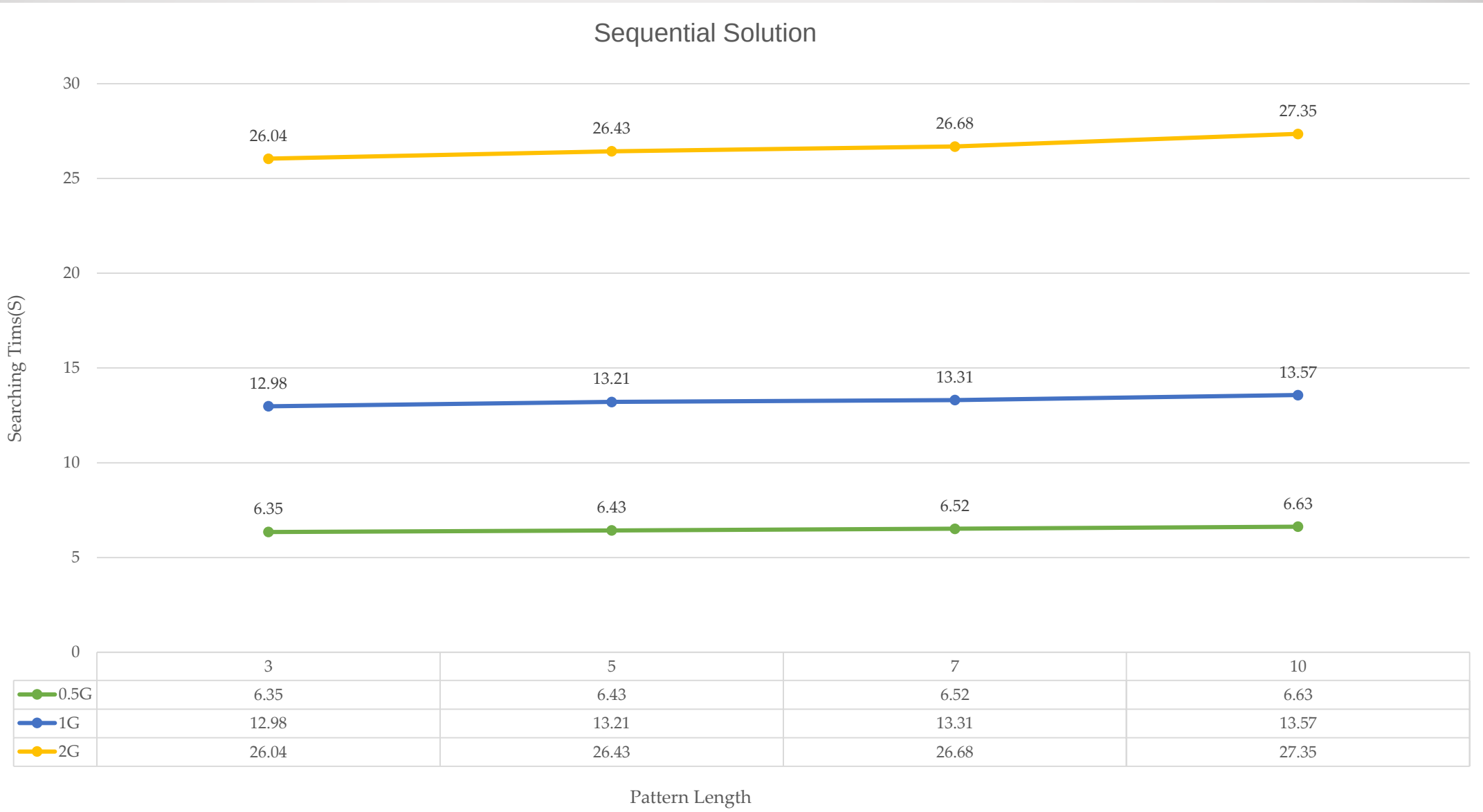
Knuth-Morris-Pratt Algorithm

- Pre-process the pattern to get longestPreSuf[i], the same as pattern length of m
- Using longestPreSuf[i] to skip matching unnecessary character in each windows, to reduce running time
- Worst case time complexity of naïve algorithm is O(m(n-m+1)), KMP algorithm is O(n)

Sequential Solution

- Using only one processor to search the whole file
- As baseline for comparison with parallel solution
- Pattern length 3, 5, 7, 10

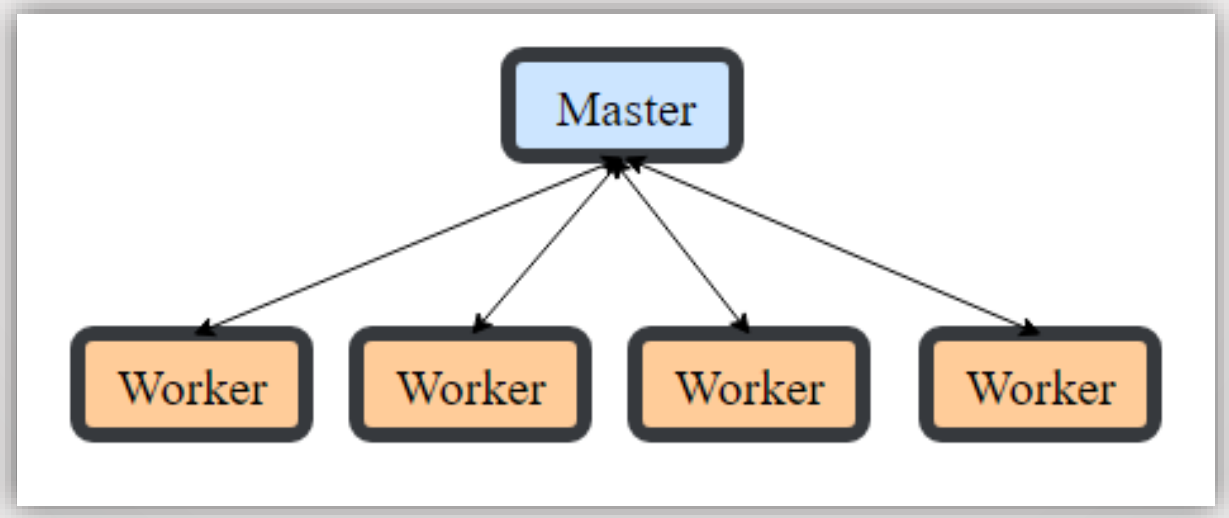
Size	Lines
0.5G	6,808,936
1G	13,944,700
2G	27,889,399



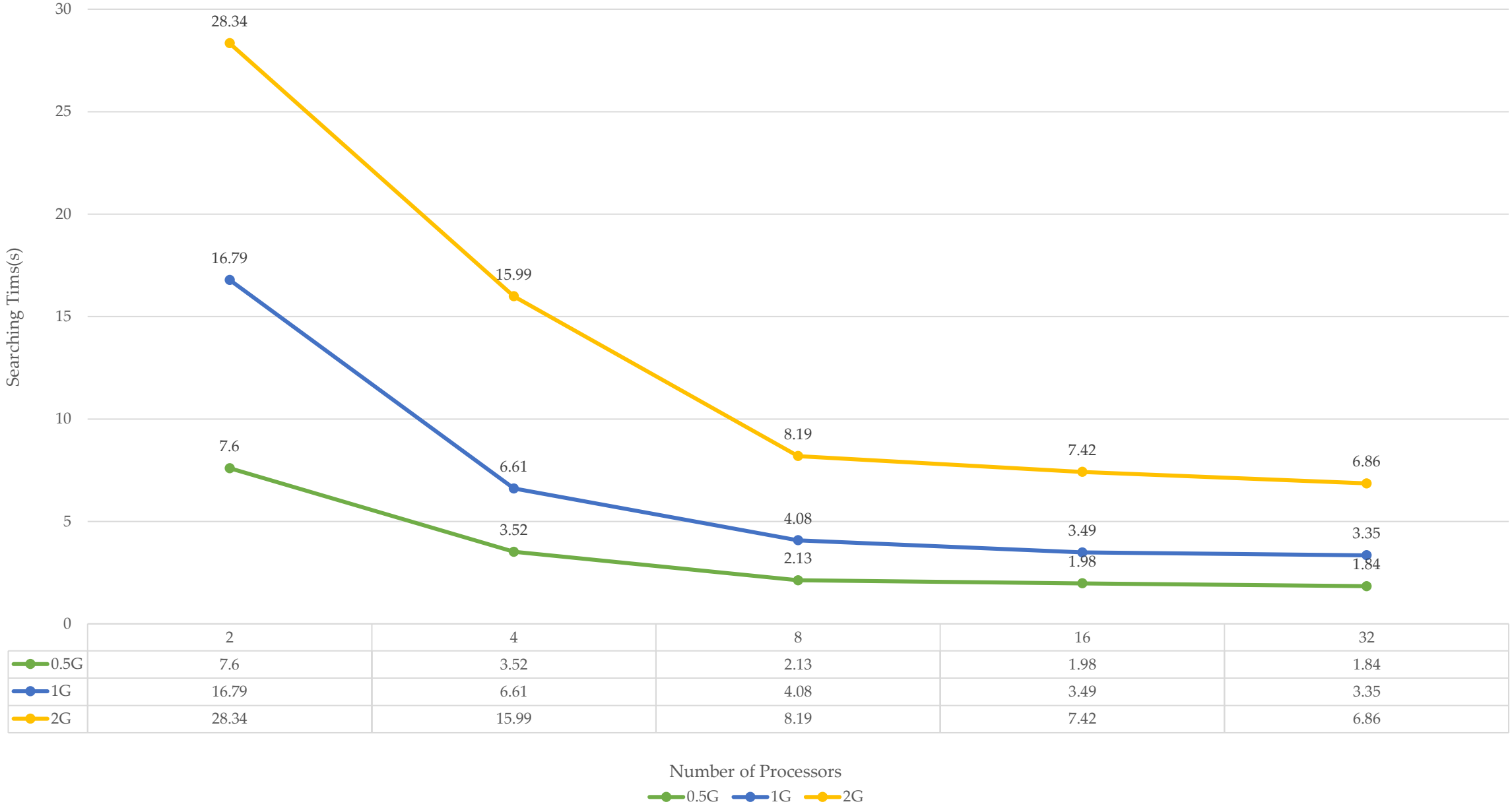
	3	5	7	10
0.5G	6.35	6.43	6.52	6.63
1G	12.98	13.21	13.31	13.57
2G	26.04	26.43	26.68	27.35

Using MPI

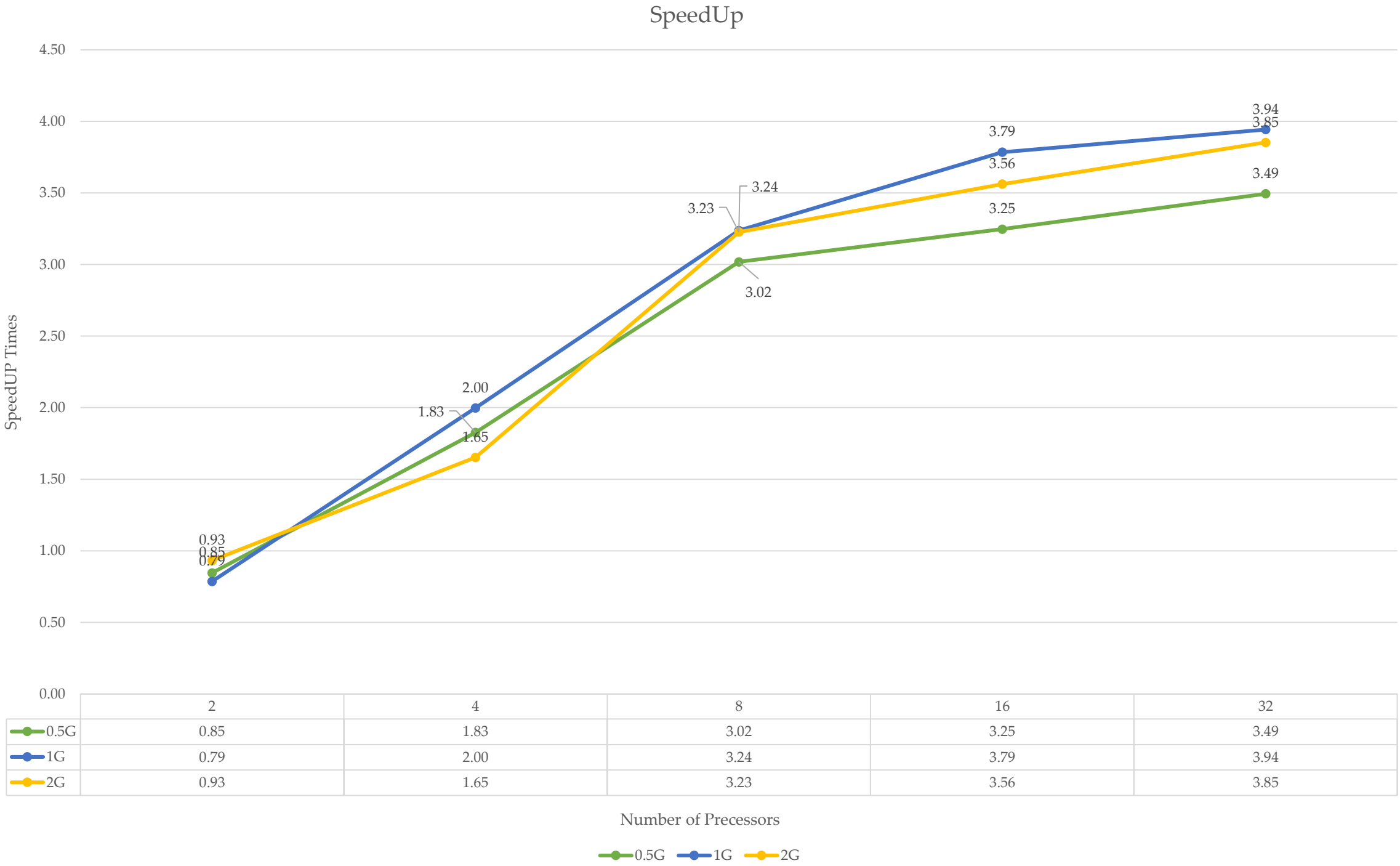
- Master node counts lines of the file and send messages to each worker node to equally assign lines. Then it waits for result messages from each worker node
- Each worker receives message from master node and focus on their assigned lines and do the search. Then send result message back to master node



Parallel Solution



	2	4	8	16	32
0.5G	7.6	3.52	2.13	1.98	1.84
1G	16.79	6.61	4.08	3.49	3.35
2G	28.34	15.99	8.19	7.42	6.86



	2	4	8	16	32
0.5G	0.85	1.83	3.02	3.25	3.49
1G	0.79	2.00	3.24	3.79	3.94
2G	0.93	1.65	3.23	3.56	3.85

Observation

- Parallel solution can speed up searching
- Parallel solution overhead may decrease the speedup
- Max speedup is 3.94, though 35 more processors are used

Future Work

- Testing larger files
- Finding best suitable configuration based on file size
- Considering the situation that arbitrary length assigned to workers, and pattern is longer than assigned lines
- Searching in a long whole line

Reference

- Al-Dabbagh, S.S.M., Barnouti, N.H., Naser, M.A.S. and Ali, Z.G. (2016) Parallel Quick Search Algorithm for the Exact String Matching Problem Using OpenMP. Journal of Computer and Communications, 4, 1-11. http://dx.doi.org/10.4236/jcc.2016.413001
- Implementing String Searching Algorithms on a Network of Workstations Using MPI, Panagiotis D Michailidis, Konstantinos G. Margaritis
- https://ubccr.freshdesk.com/support/solutions/articles/13000026245-tutorials-and-training-documents
- http://jacobmills.co.uk/university-high-performance-computing/
- https://bisqwit.iki.fi/story/howto/openmp/
- https://www.geeksforgeeks.org/searching-for-patterns-set-2-kmp-algorithm/

Thanks