Performance of Gensim Word2vec

------detailed version

Minmei Wang
mwang107@ucsc.edu

https://github.com/RaRe-Technologies/gensim

fppt.com

Outline of Project

- Throughput benchmark
  - Measure IPC, branch MPKI, L1 MPKI, LLC MPKI for gensim
  - InputSize: X=6,678,526,848,423 (6T), ~X/4,~X/16,~X/64,~X/256
  - Threads: 1, 2, 4, 8
- Latency benchmark
  - Measure latency for gensim Word2vec
  - Total run 200 times for the input

fppt.com

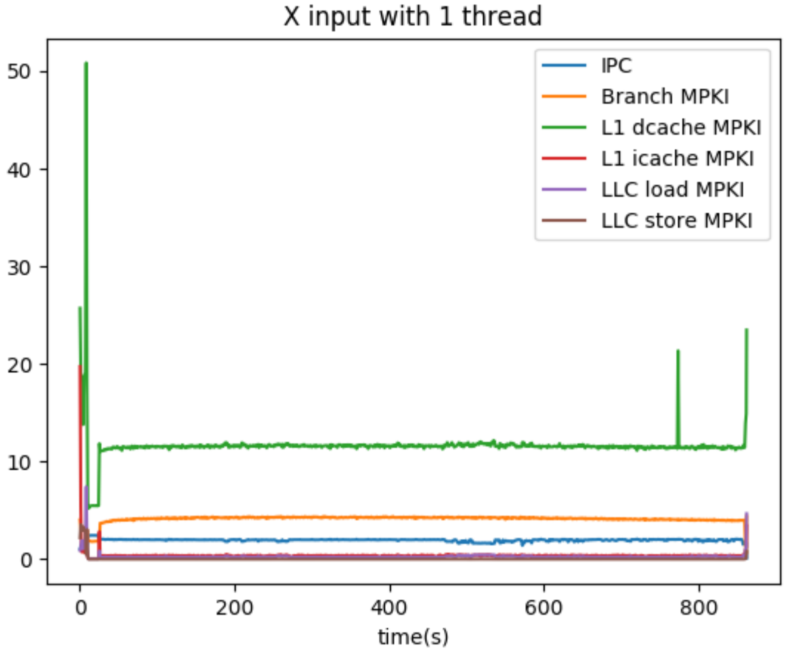
Part I

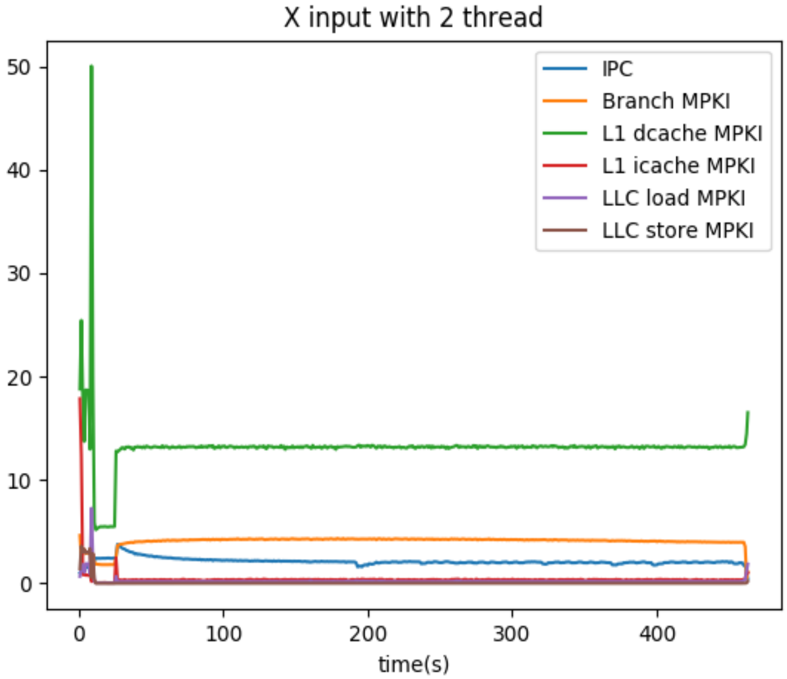
Throughput Benchmark

fppt.com

Throughput benchmark

- Input Size: X (6,678,526,848,423) (1,2,4,8 threads)

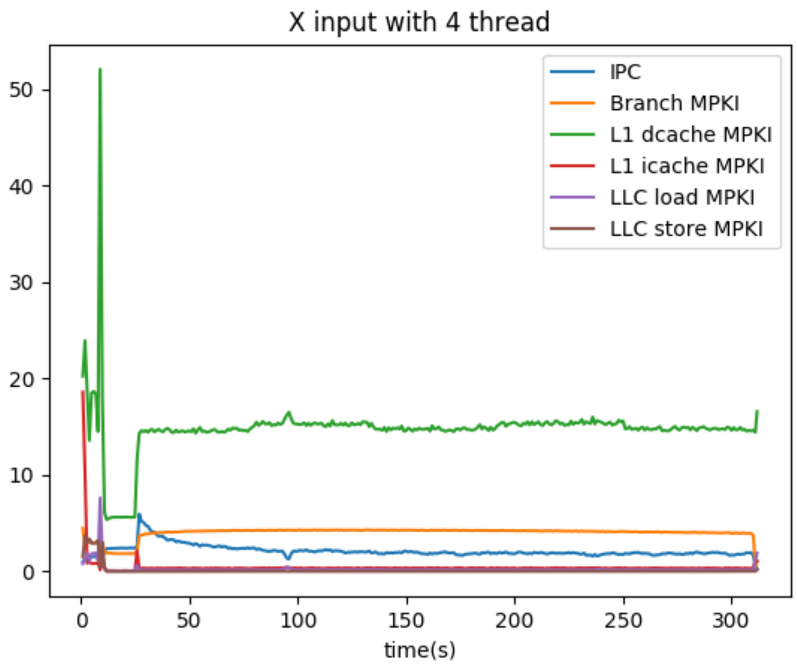


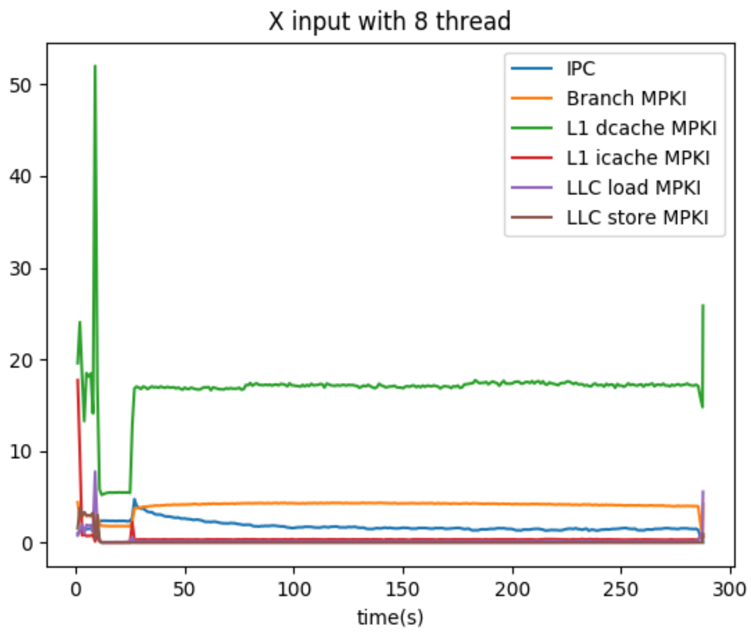


fppt.com

Throughput benchmark

- Input Size: X (1,2,4,8 threads)
- It has good speedup when threads enlarges to 4.

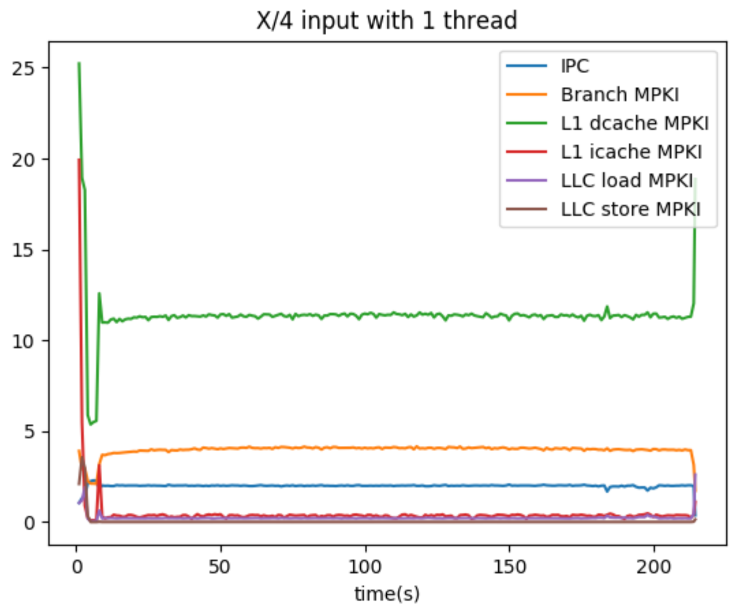


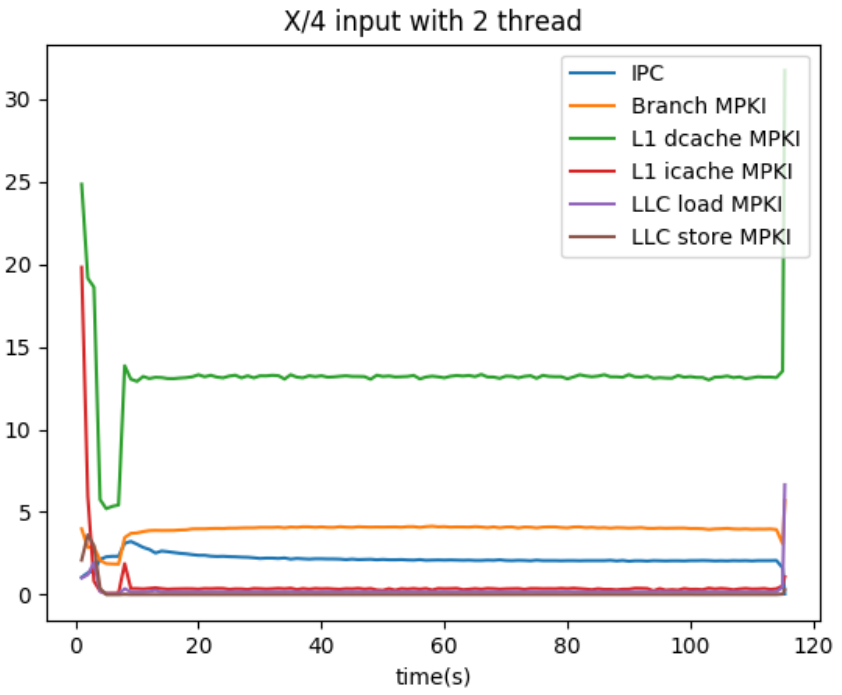


fppt.com

Throughput benchmark

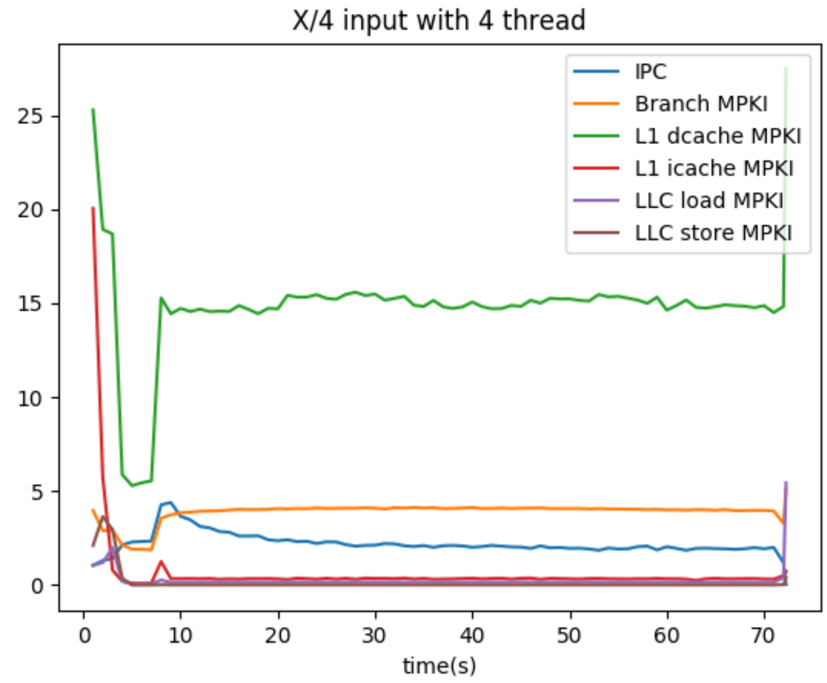
- Input Size: X/4 (1,709,851,631,548)(1,2,4,8 threads)

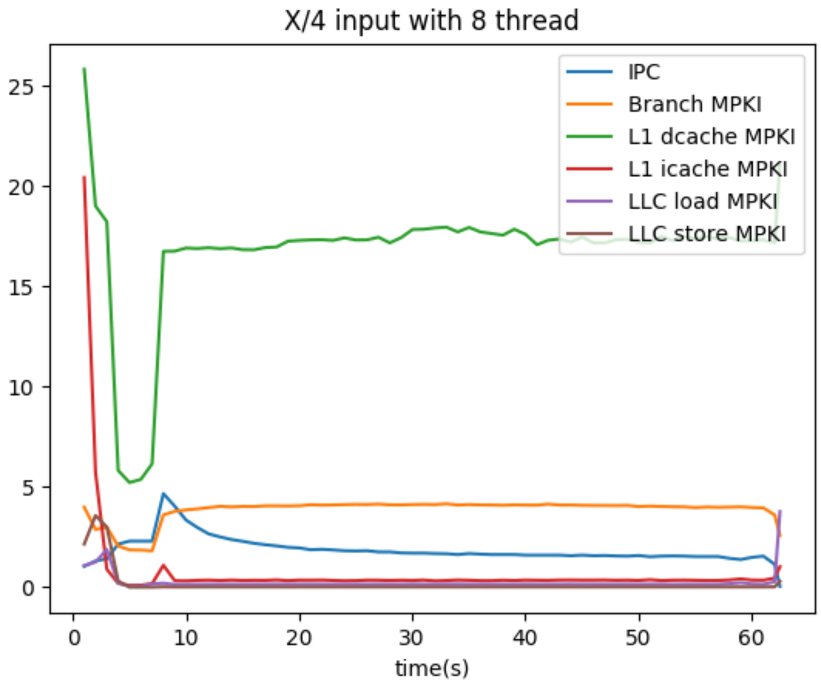




Throughput benchmark

- Input Size: X/4 (1,2,4,8 threads)

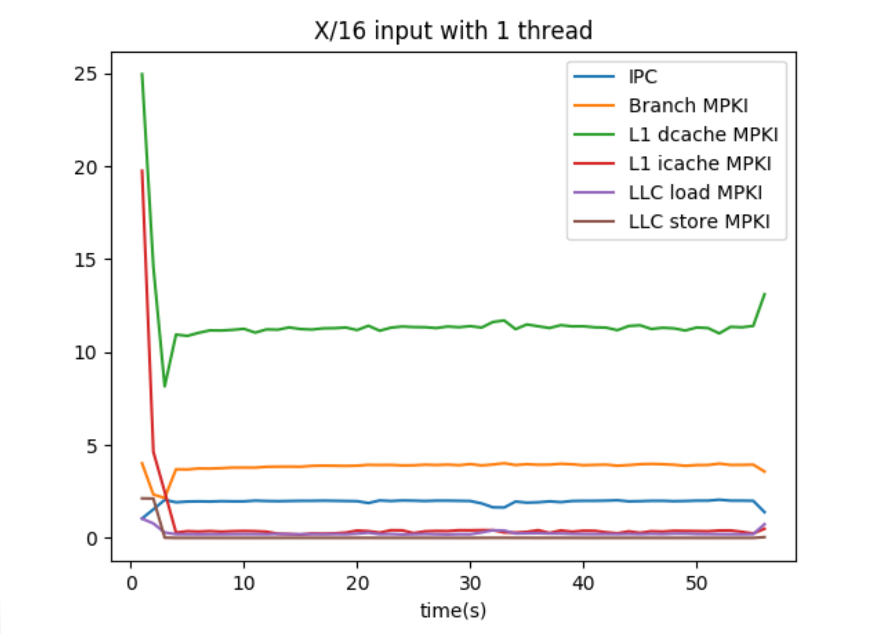


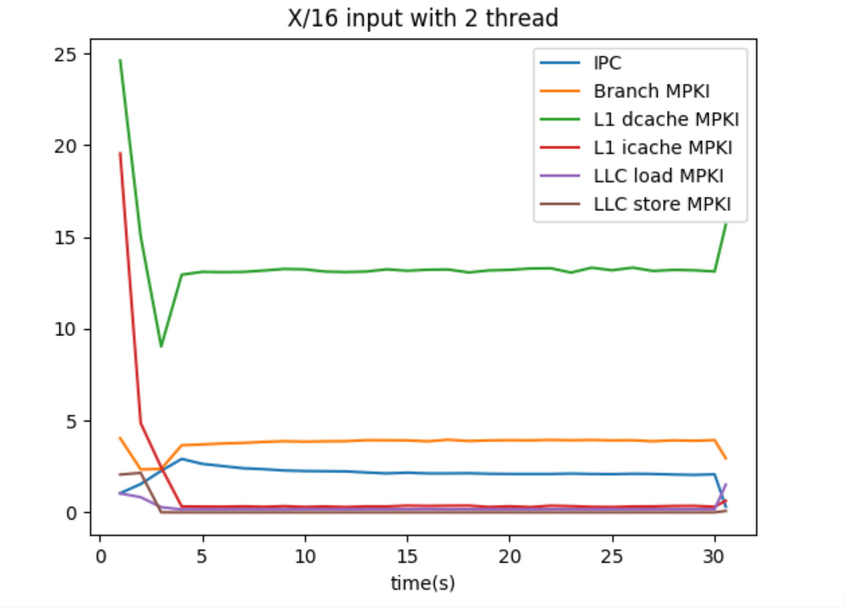


fppt.com

Throughput benchmark

- Input Size: X/16 (437,717,377,815)(1,2,4,8 threads)

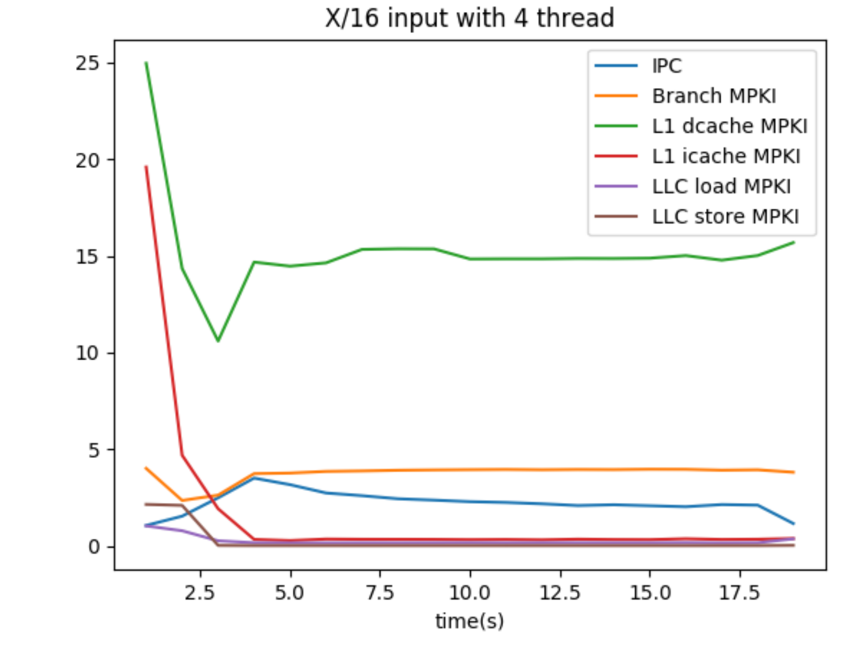


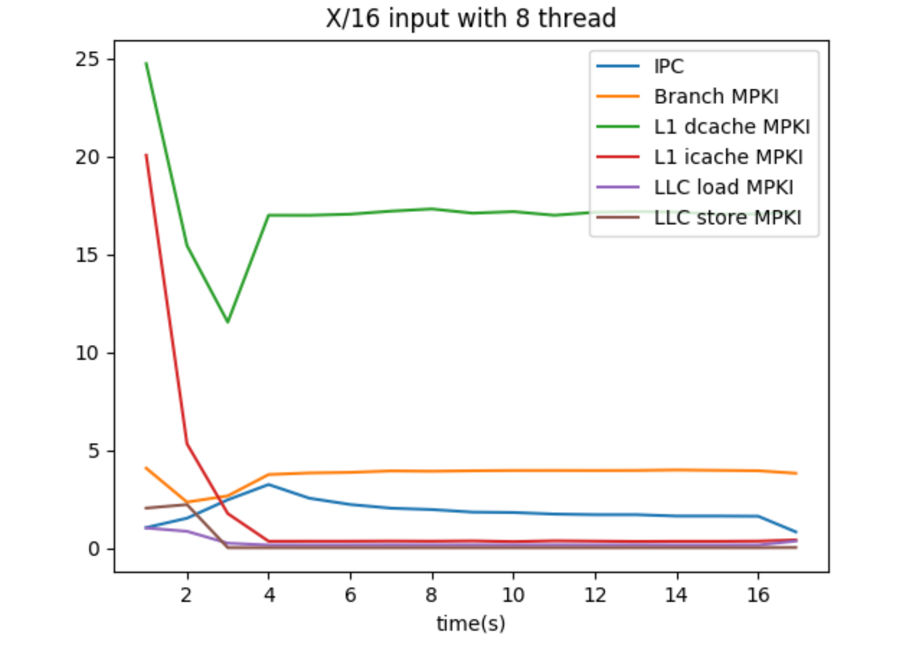


fppt.com

Throughput benchmark

- Input Size: X/16 (1,2,4,8 threads)

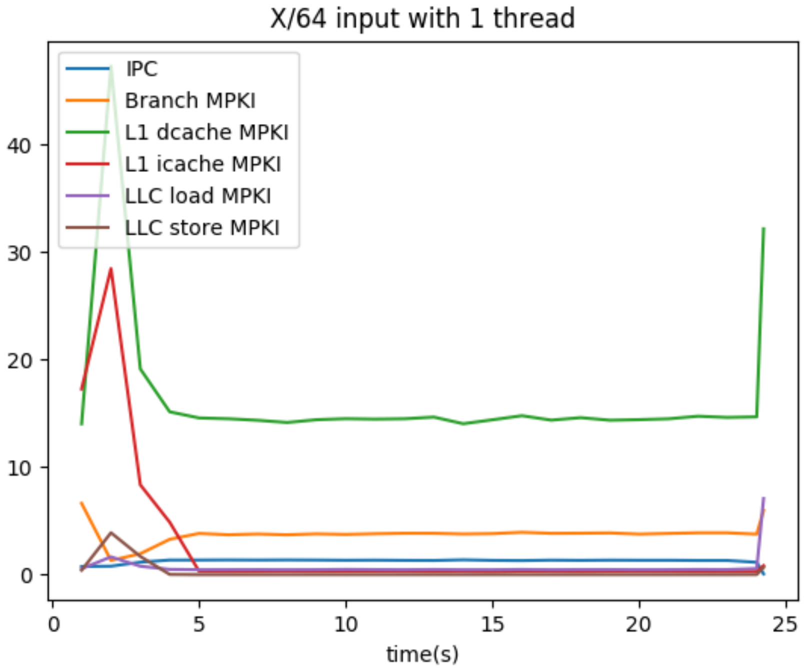


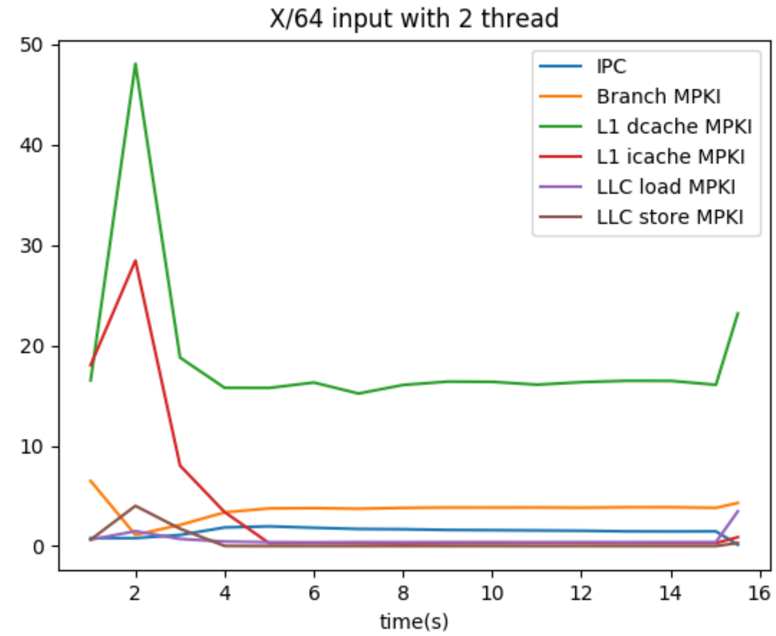


fppt.com

Throughput benchmark

- Input Size: X/64 (113,684,901,010)(1,2,4,8 threads)

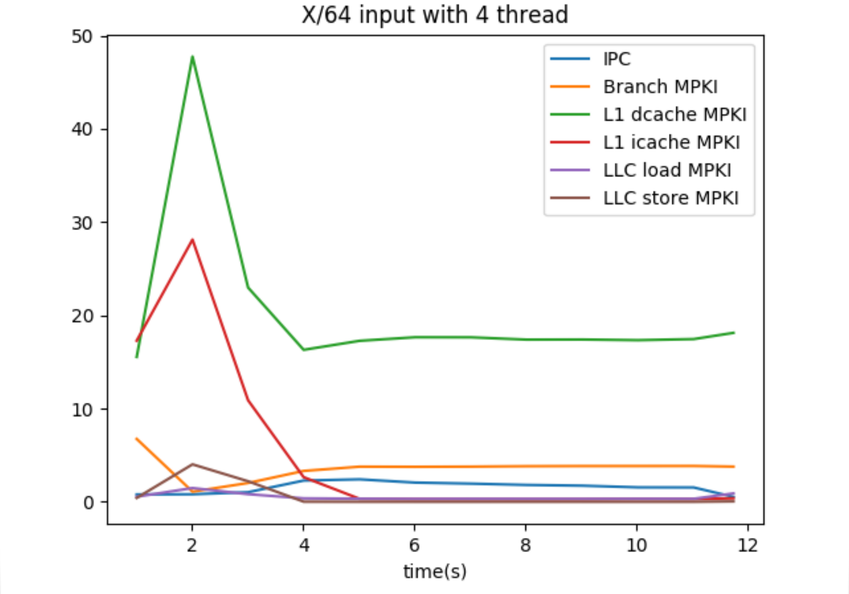


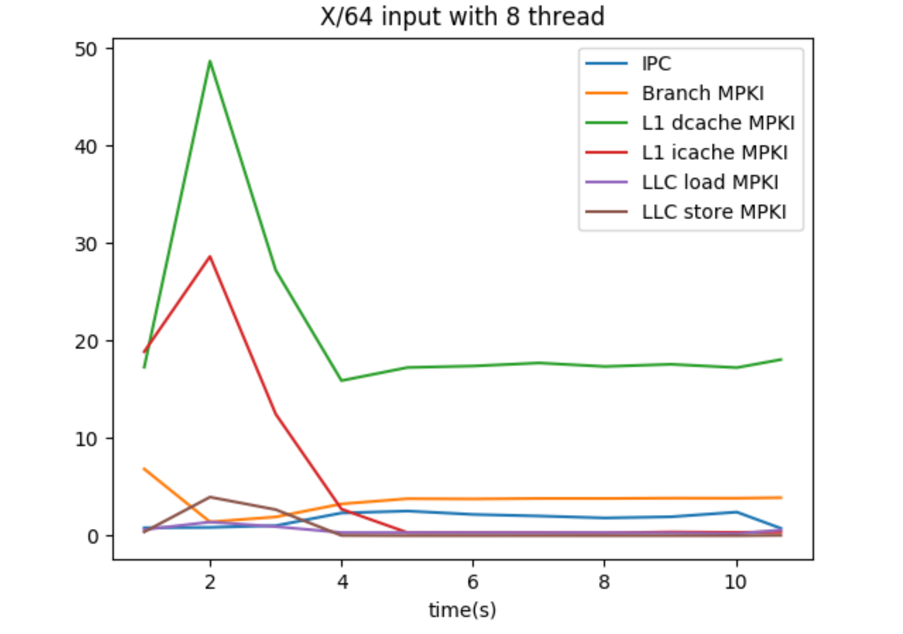


fppt.com

Throughput benchmark

- Input Size: X/64 (1,2,4,8 threads)

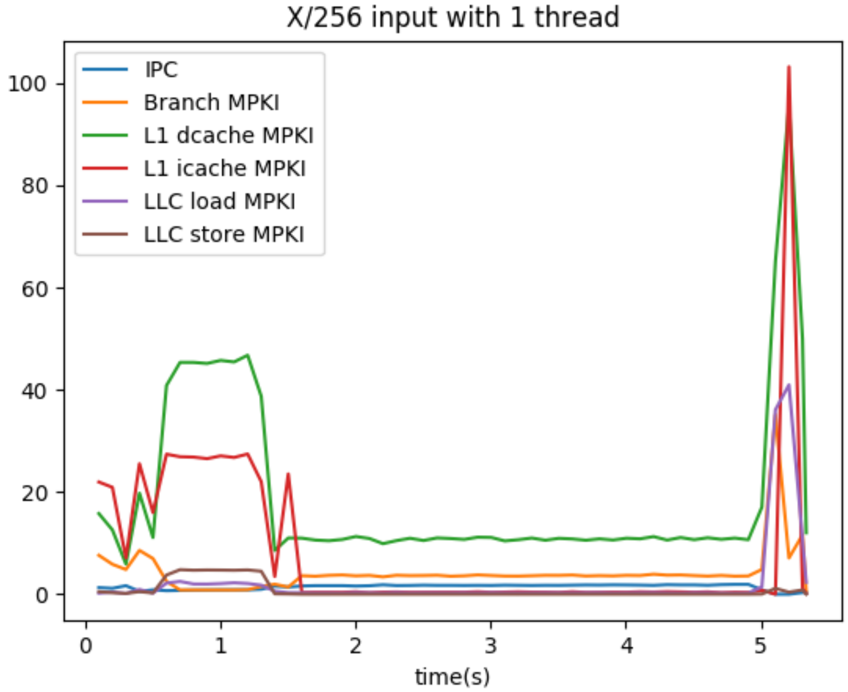


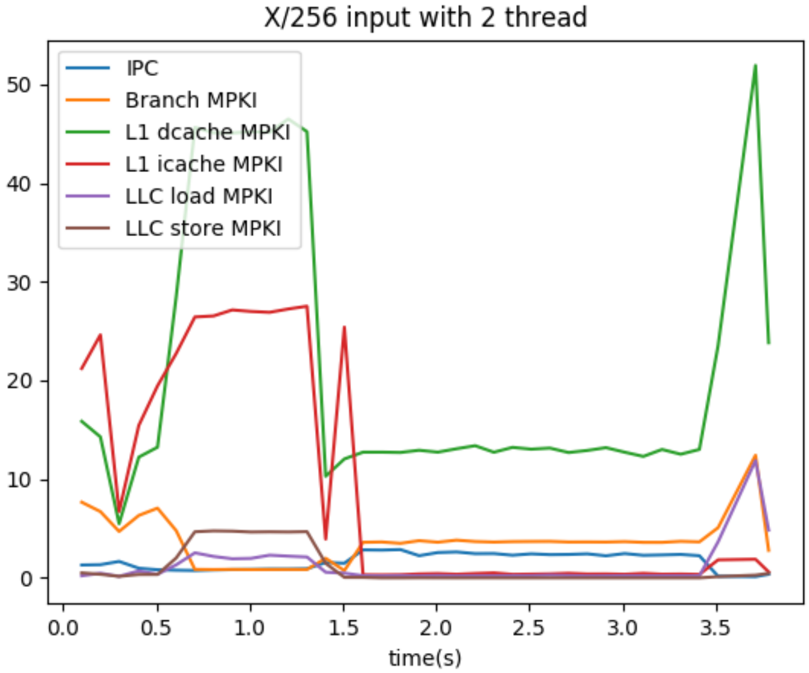


fppt.com

Throughput benchmark

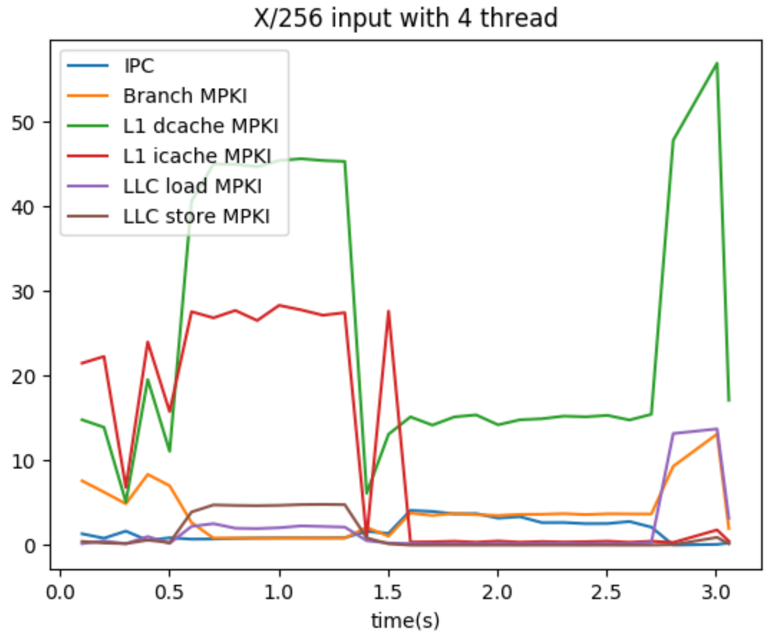
- Input Size: X/256(30,460,388,166) (1,2,4,8 threads)

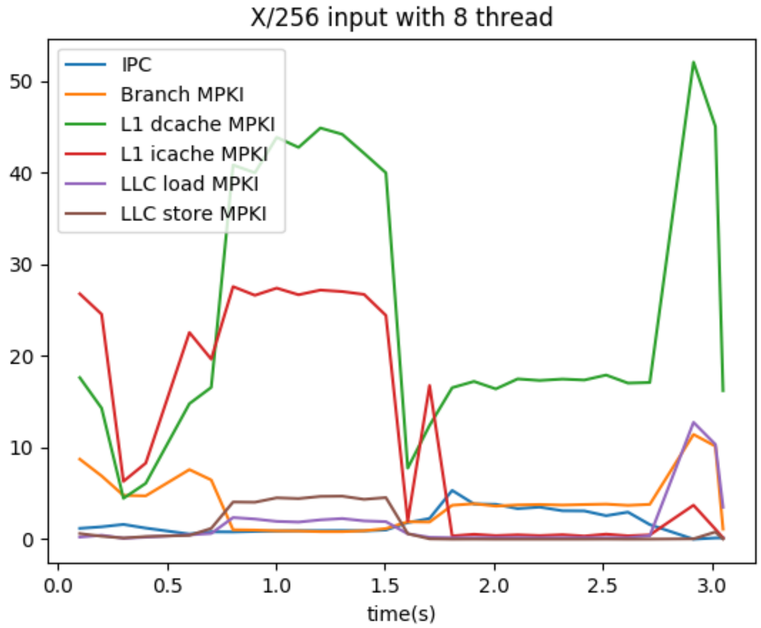




Throughput benchmark

- Input Size: X/256 (1,2,4,8 threads)





fppt.com

Conclusions

- The runtime is proportional to the size of instructions
- Multithreads
  - Multithreads is useful for speed-up, especially when threads is from 1 to 2, 4.
  - It has limitations, It has little speedup when the number of threads is from 4 to 8. It is the reason of data dependencies in the code.
- L1 dcache MPKI is relatively higher in the code
- LLC load and store MPKI performance is good in the code

fppt.com

Part II

Latency Benchmark

fppt.com

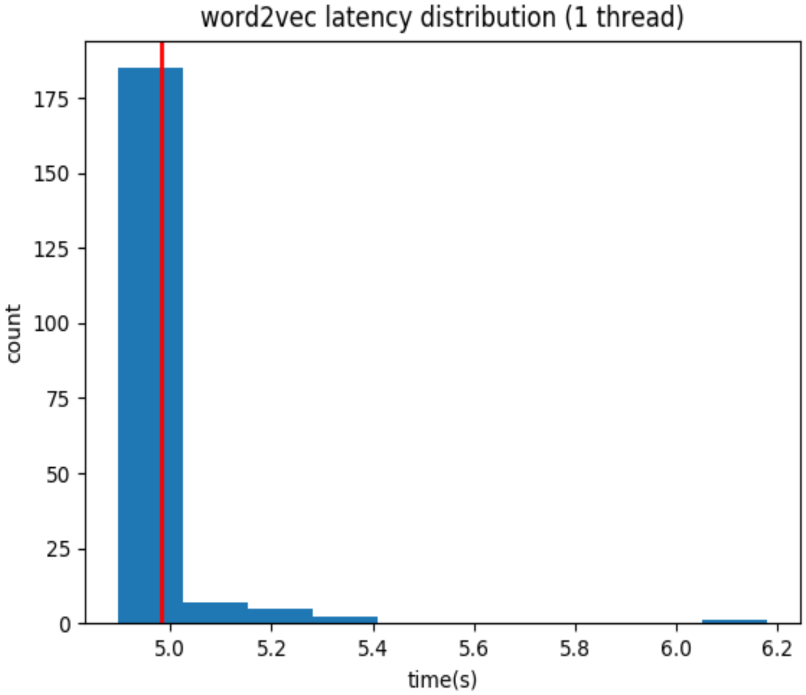
Latency Benchmark

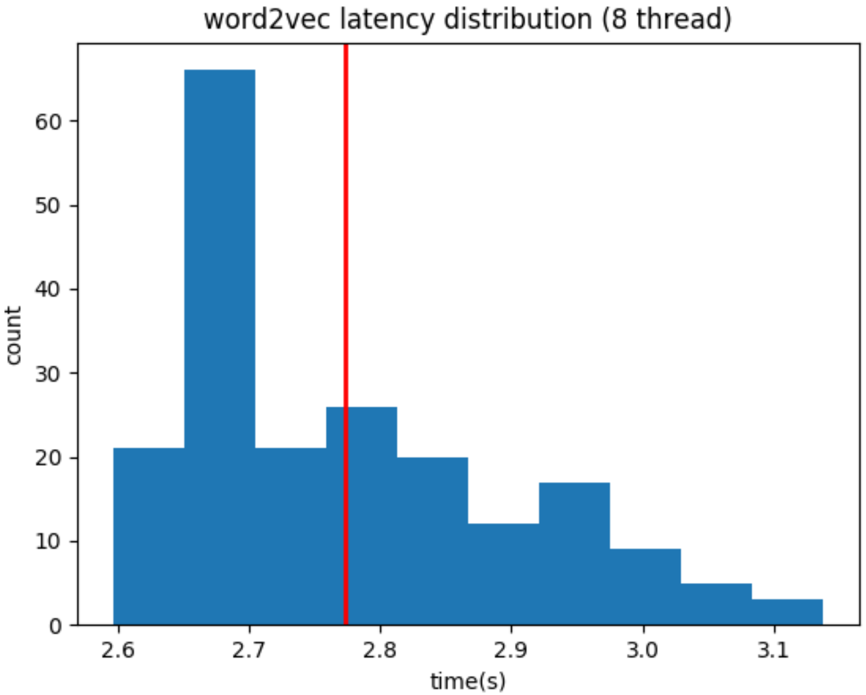
- Latency is the time between giving the input and generating the output
  - Input: A list of files
  - Output: A trained word2vec model

fppt.com

Latency Benchmark

- I run 200 times for benchmark with almost 30,460,388,166 instructions. The read line is the average.
- The latency distribution with 1 thread is more concentrated may be because the code has strong data dependency during the running process





fppt.com