Task 2A Software Prefetching

Deliverables

1. Analyze the impact of embedding table size:

Following chart shows impact of different embedding table sizes on performance.

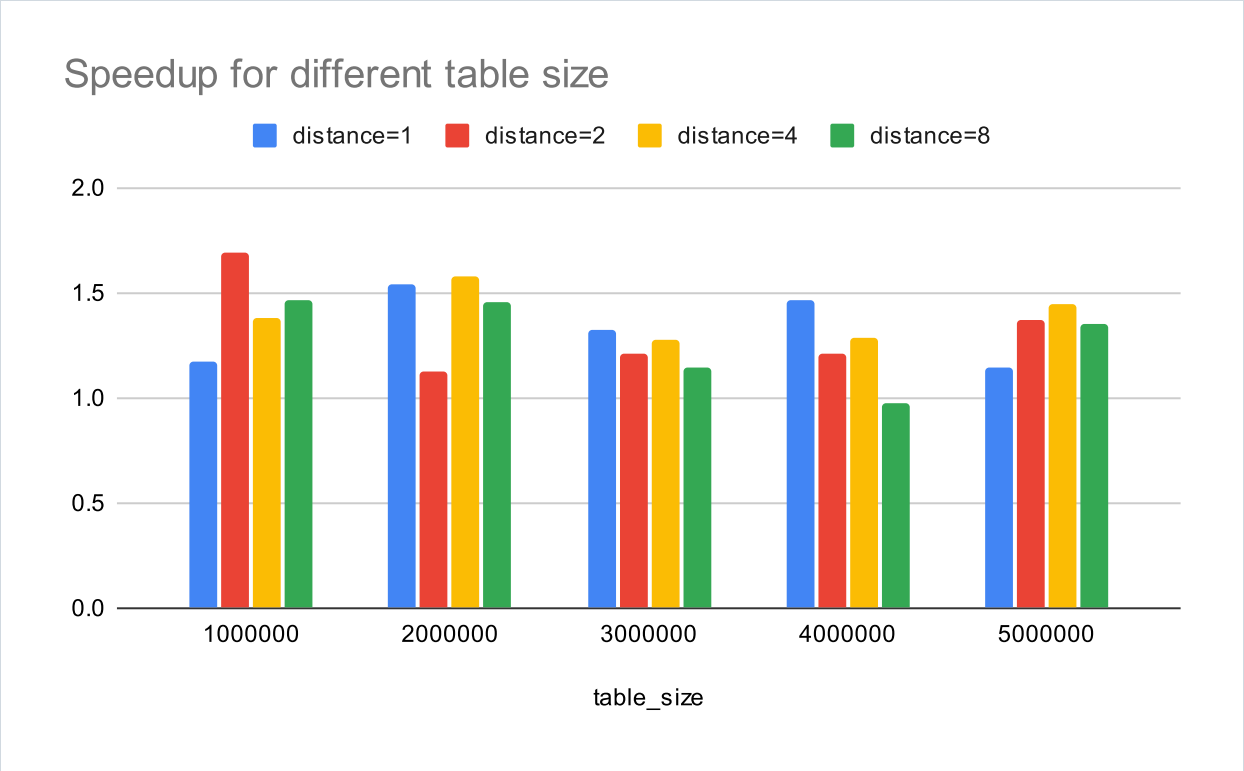

Fig 1. Impact of different embedding table sizes

2. Analyze CPU metrics:

We couldn't find an event to profile number of misses for L2 and LLC cache levels. In a chart below we see impact of different table sizes and prefetch distance on L1D cache misses.

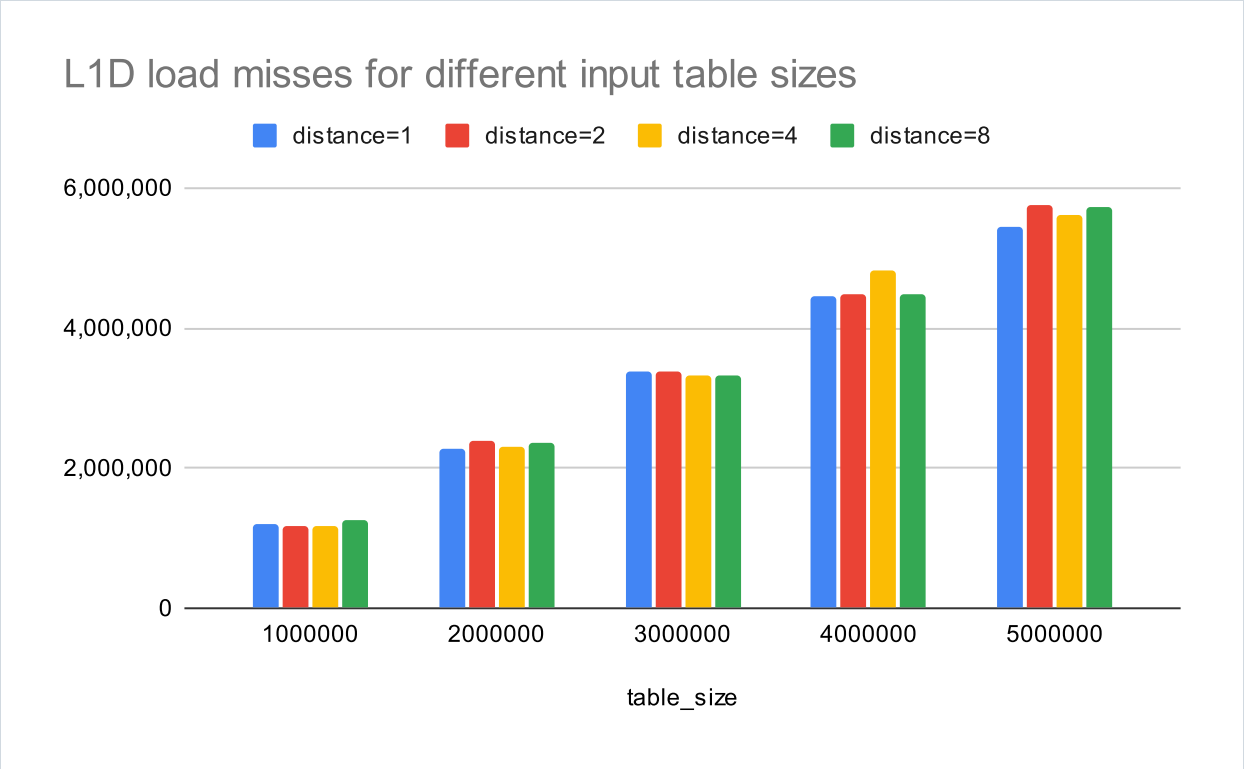

Fig 2. Impact of different embedding table sizes on L1D cache misses

3. Collect exectution time and compute speedup

Please refer to Fig 1.

4. Identify optimal parameters

However there is no clear optimal choice for prefetch distance we believe that `distance=4` works well. Because it has generally lower L1D misses and it provides more speedup in majority of table size.

Questions

Q1. What trend do you observe in speedup with different embedding table sizes?

- Unlike matrix multiplication, we do not see higher speedup with increased input size (Fig 1.). This is because we are not making access cache friendly but rather just prefetching blocks that will be used in future. Hence, table size becomes irrelevant.

Q2. What is the best prefetch distance?

- Our data suggest `distace=4` works best.

Q3. At what cache fill level do you achieve the maximum speedup?

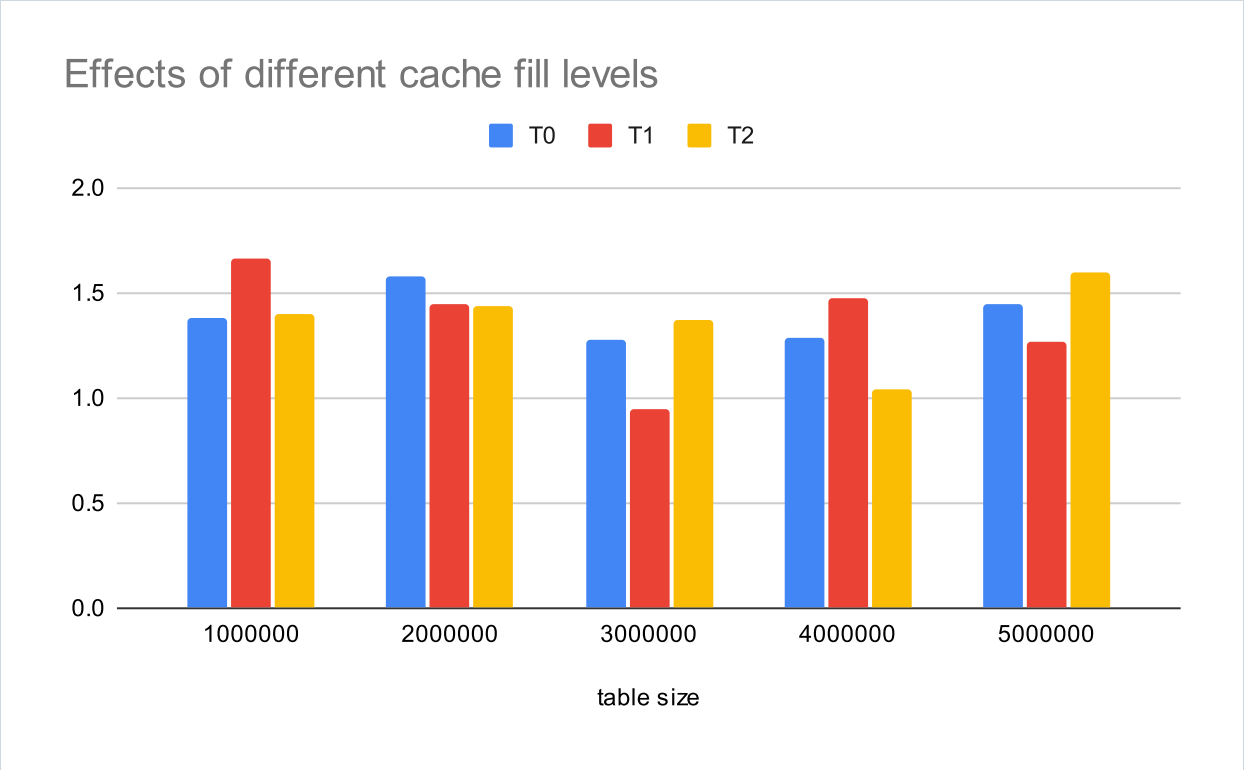

Fig 3. Effects of different cache fill levels

Task 2B SIMD (Single Instructor Multiple Deadline)

Deliverables

1. Analyze the impact of SIMD instruction width:

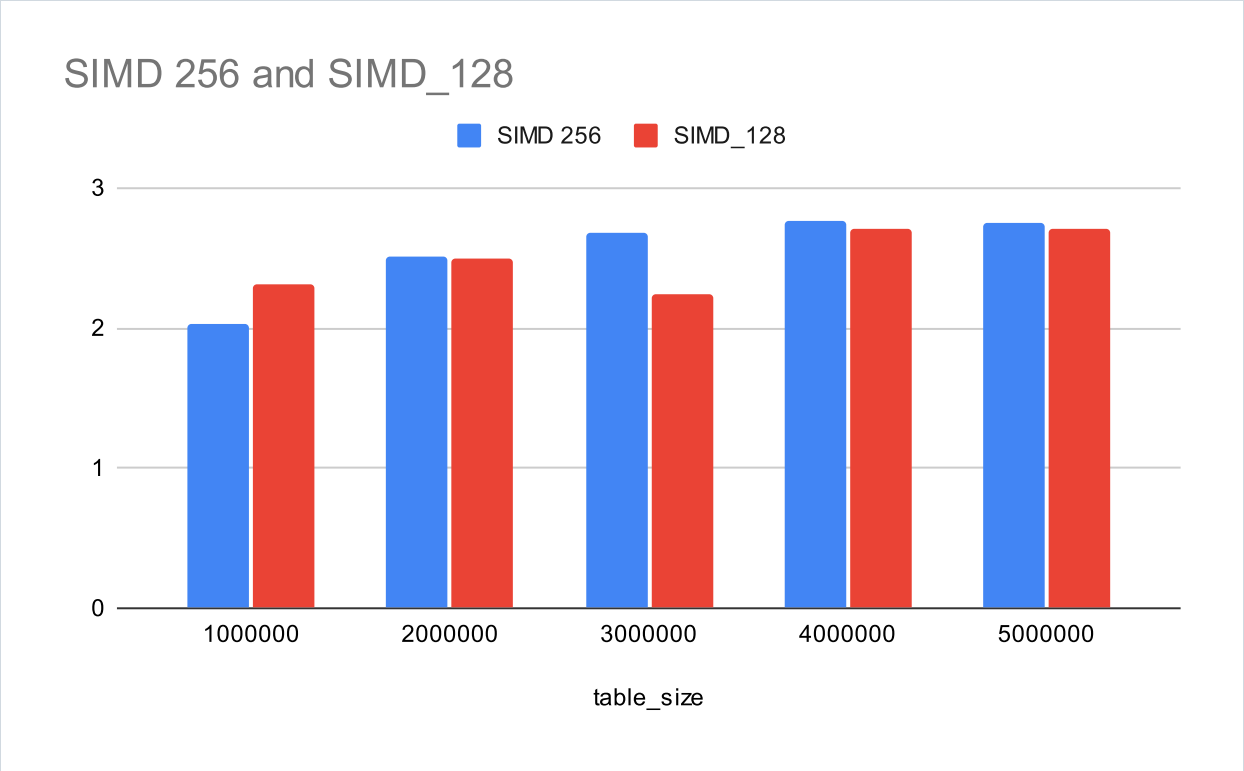

Fig 4. Impact of SIMD instruction width

We see higher speedup for 256 instruction width.

2. Analyze the impact of Embedding dimension

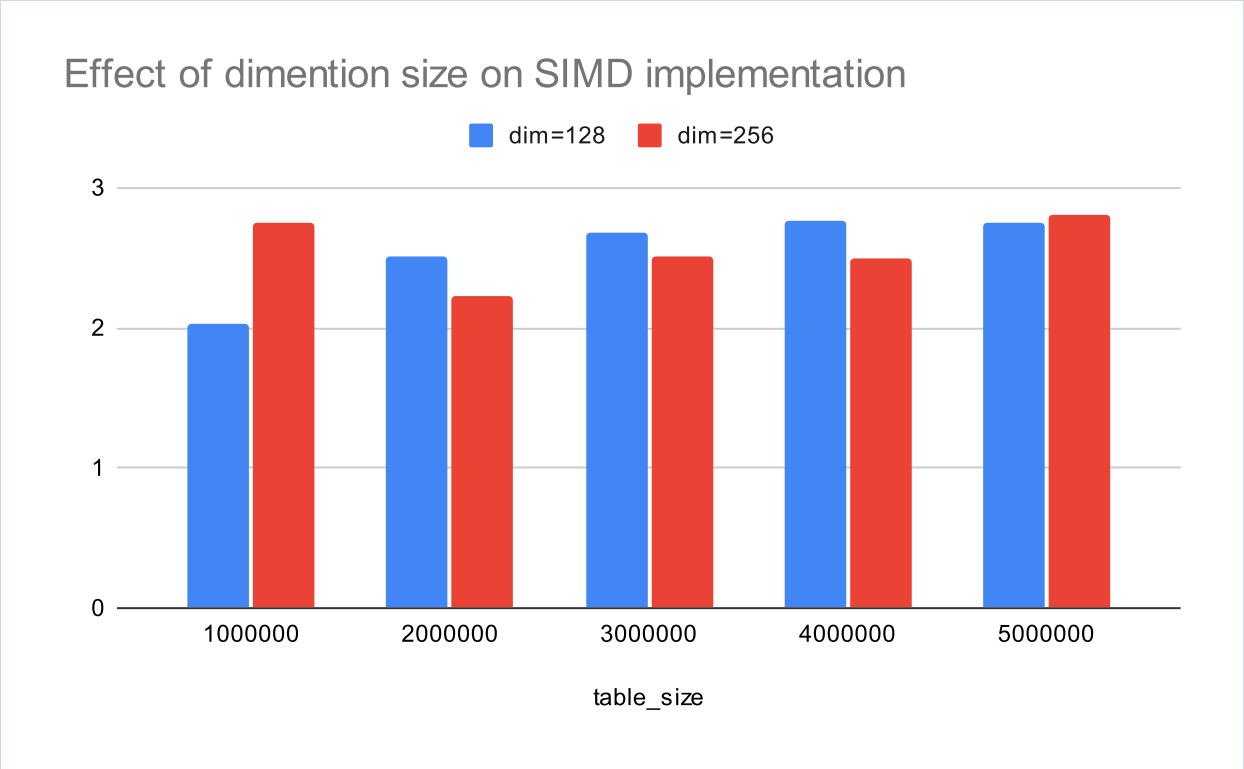

Fig 5. Effects of dimention size on SIMD

We see drop in speedup when increasing the number of dimention. This is expected becauuse we simply have more work to do.

3. Analyze instruction count

The effects of instruction count will be similar to one we saw on matrix multiplication. With higher instruction width register we will see more reduction in instruction count.

4. Collect execution time and compute speedup

Please refer to Fig 4.

Questions

Q1. What trends do you observe in speedup for different combinations of embedding dimensions and SIMD widths?

Higher SIMD width provides greater speedups. And with increased number of embedding dimensions we see decrease in speedup (because there is more work to do).

Q2. For which SIMD width do you achieve the maximum speedup?

For 256 width we achieve maximum speedup.

Task 2C Software Prefetching + SIMD

Here is the final summary of software prefetching + SIMD. We have used 256 bit wide instructions. And we have used prefetch distance of four. We use T0 fill level.

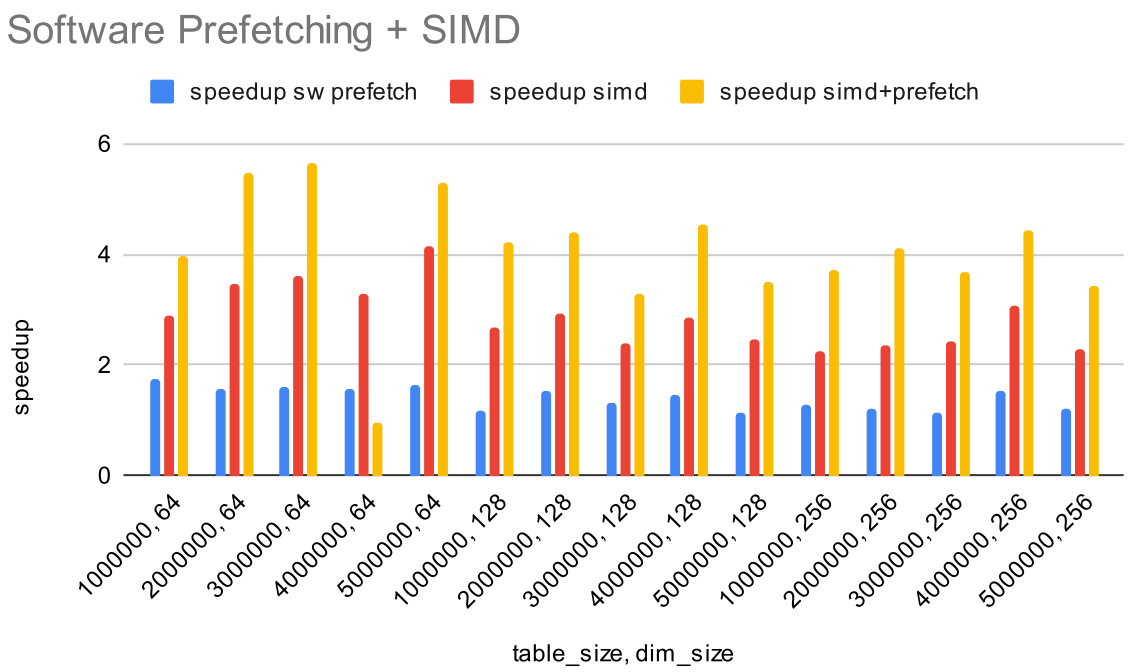

Fig 6. Final Analysis of Software Prefetching + SIMD