Lecture 4: Announcements

- hw time spent
- office hours
- don't forget: talk topic + SSH key
- talk time assignment
Case Study: Streaming Workloads

Q: Estimate expected throughput for saxpy on an architecture with caches. What are the right units?

\[ z_i = \alpha x_i + y_i \quad (i = 1, \ldots, n) \]

Demo: [https://github.com/lcw/stream_ispc](https://github.com/lcw/stream_ispc)
Special Store Instructions

At least two aspects to keep apart:

- Non-temporal locality
- Spatial locality

What hardware behavior might result from these aspects?

- Non-temporal: writing past cache, hint at fast store
- Spatial: don’t bring in data, mark rest of line

Comment on what a compiler can promise on these aspects.

 Might these ‘flags’ apply to loads/prefetches?

(see also: [McCalpin ‘18])
Case study: Matrix-Matrix Mult. (’MMM’): Code Structure

\[ O(n^2) - O(n^3) \]

- How would you structure a high-performance MMM?
- What are sources of concurrency?
- What should you consider your working set?
\( \frac{ab}{c} + (d + e) \)
Case study: Matrix-Matrix Mult. ('MMM') via Latency

Come up with a simple cost model for MMM in a two-level hierarchy based on latency:

\[
\text{Avg. latency per access} = (1 - \text{miss ratio}) \cdot \text{cache latency} + \text{miss ratio} \cdot \text{miss cost}
\]

Total accesses: \(4N_b^3\)

Misses: \(3N_b^2\)

Miss ratio: \(\frac{3N_b^2}{4N_b^3} \cdot \text{cache line size} = \frac{3}{4N_b \cdot \text{CLS}}\)

[Yotov et al. ’07]
Case study: Matrix-Matrix Mult. (’MMM’) via Bandwidth

```
Come up with a cost model for MMM in a two-level hierarchy based on bandwidth:

\[ 16 \times 2 \text{ FMA} s \]
\[ \frac{512/32}{12} \text{ FMA-capable core units} \]
\[ \text{Cycle count for whole calc: } \frac{2N^3}{(2 \times 32)} \]
\[ 4 \frac{N^3}{(N^3/32)} = 128 \text{ Flops/cycle} \]
\[ \text{Total mem < cache dark matter: } \]
\[ \text{# blocks} \cdot (\text{block size}) \cdot 4 = \left(\frac{N}{N_3}\right)^3 \cdot N_8^2 \cdot 4 \]
\[ \text{Required mem bw: } \]
\[ 4 \frac{N^3}{N_3} / (N^3/32) = \frac{128}{N_8} \text{ floats/cycle} \]

[Yotov et al. ’07]
```
Case study: Matrix-Matrix Mult. (’MMM’): Discussion

Discussion: What are the main simplifications in each model?

[Yotov et al. ’07]

General Q: How can we analyze cache cost of algorithms in general?