Problem 3: Research is Fun! (26 pts)

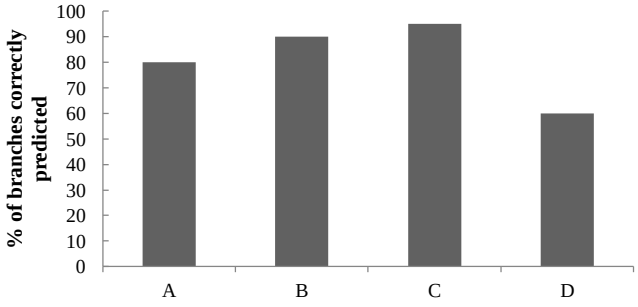
A researcher is studying compression in on-chip caches. You may assume it's very similar to Base-Delta-Immediate compression studied in this course, with an implementation that allows one base per cache line. She's considering four workloads: A, B, C, and D. She builds a baseline System X, without compression, and compares this against System Y, which is employing on-chip compression in the last-level cache (the modifications required to support compression, such as doubling the number of tags, are the only differences between System X and System Y).

- Figure 0 shows the branch prediction accuracy in System X of the four workloads.
- Figure 1 shows the accuracy of System X's stream prefetcher. As a reminder, a stream prefetcher identifies that the processor is continuously accessing consecutive cache lines (i.e., streaming) and prefetches future cache lines into the cache before the processor requests them.
- Figure 2 shows the misses per thousand instructions (MPKI) in the last-level cache (LLC) of System X.
- Figure 3 shows the effective cache capacity in the compressed LLC in System Y.
- Figure 4 shows the instructions per cycle (IPC) of System Y normalized to the IPC of System X.
- Figure 5 shows the normalized LLC MPKI of System Y normalized to the LLC MPKI of System X.

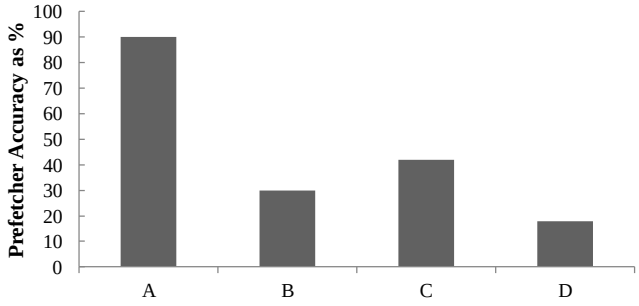
Answer the following questions, providing the most likely explanation for each considering the information provided in the figures.

Continued . . .

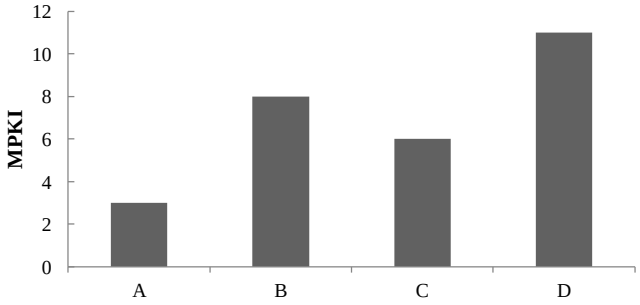
0. Branch prediction accuracy in System X



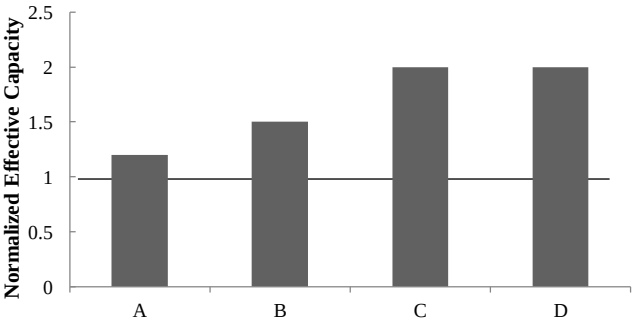
1. Prefetcher accuracy in System X



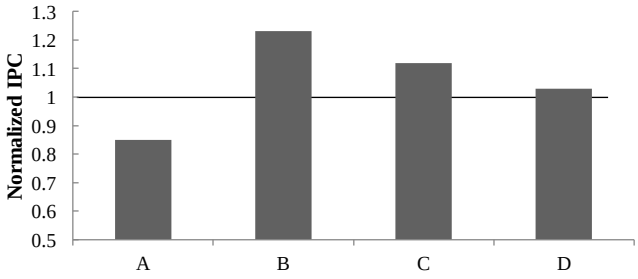
2. LLC MPKI in System X



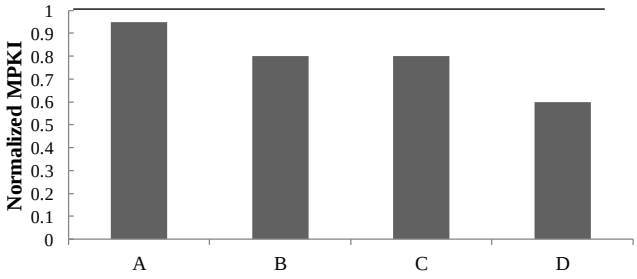
3. Effective LLC cache capacity in System Y



4. Performance (IPC) of System Y normalized to System X IPC



5. LLC MPKI in System Y normalized to LLC MPKI in System X



A) [4 pts] Why might the normalized IPC of workload A be less than 1.0? (15 words or less)

Decompression latency

B) [4 pts] Why might the normalized IPC of workload B be greater than the normalized IPC of workload C? (15 words or less)

B is more sensitive to cache size

C) [4 pts] Why might the normalized IPC of workload C be greater than the normalized IPC of workload D? (15 words or less)

C is more sensitive to cache size

Assume just for the next two subquestions that most of Workload C's data is of type Flow and most of Workload D's data is of type Node:

```
struct Flow { // defined in Workload C
  long flow_time;
  Pipe * inlet;
  Pipe * outlet;
  char identifier;
  float flow_rate;
}

struct Node { // defined in Workload D
  Node * right_sibling;
  Ancestor * parent;
  Descendent * child;
  Node * left_sibling;
  Node * metadata;
}
```

D) [4 pts] Just from looking at the above code, which workload's data is likely more compressible?

CIRCLE ONE: **C** **D**

Why? (15 words or less)

Pointers often have low dynamic range, so D has more compressible data.

E) [5 pts] For a new workload E, is it possible that the LLC MPKI in System Y is greater than the LLC MPKI in System X?

CIRCLE ONE: **YES** **NO**

Why or why not? (15 words or less)

Yes, depends on replacement policy.

Continued . . .

F) [5 pts] The results in the figures were determined through simulation. To increase simulation speed, the researcher was using a fixed 300 cycle latency for all memory requests. Now, she decides to model the DRAM system accurately. Across all workloads, the average memory access latency with this new model is 300 cycles. Which workload's performance in System X do you expect to change the most, compared to the old simulation results? State your assumptions for partial credit (15 words or less).

CIRCLE ONE: **A** **B** **C** **D**

A, high streaming prefetcher accuracy may indiciate good row buffer locality (average latency will be less than 300 cycles)