University of California Irvine

EECS 221: Languages and Compilers for Hardware Accelerators (Winter 2022)

Homework 3

Due: Thursday, March 17, 2022

1. *(80 points)* **Roofline Model.**

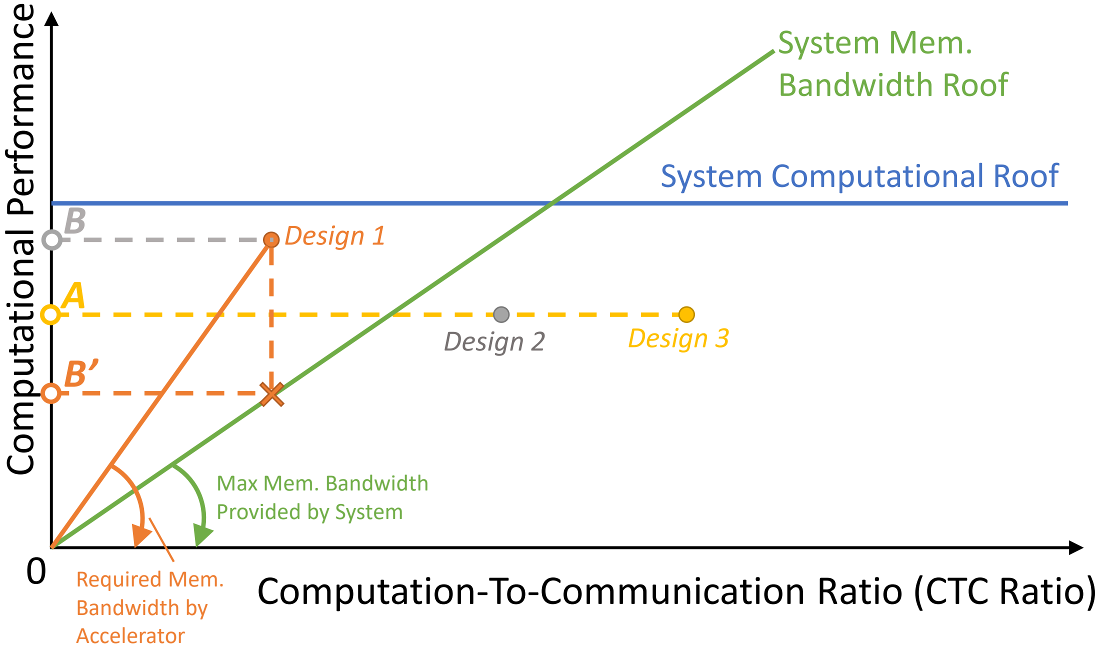
The roofline model is an intuitive visual performance model that combines memory bandwidth and computational bandwidth into a single figure. The y-axis of the figure is the *computational performance*, while the x-axis is the *computation-to-communication ratio*, or *CTC ratio*.

- Computational Performance = $\frac{\text{Total Number of Operations}}{\text{Execution Time}}$ (FLOPs/second, or FLOPS)
- Computation-to-Communication Ratio = $\frac{\text{Total Number of Operations}}{\text{Amount of External Data Access}}$ (FLOPs/Bytes)

Note that if you divide *Computational Performance* by *CTC ratio*, you will get

$$\frac{\text{Computational Performance}}{\text{CTC Ratio}} = \frac{\text{Amount of External Data Access}}{\text{Execution Time}},$$

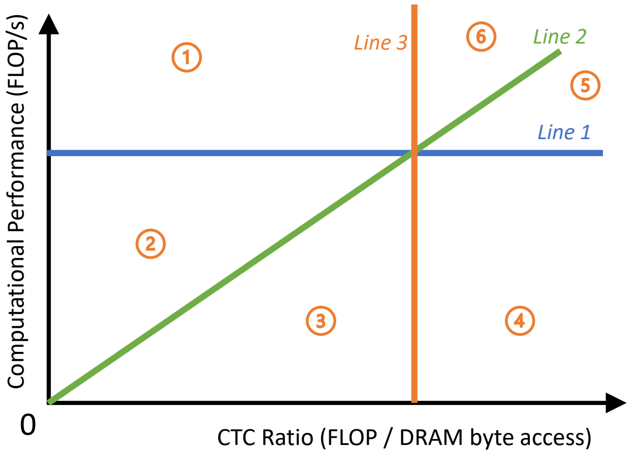
which is the memory bandwidth. The figure below shows an example of roofline model that illustrates the performance of the system and a few accelerator designs.



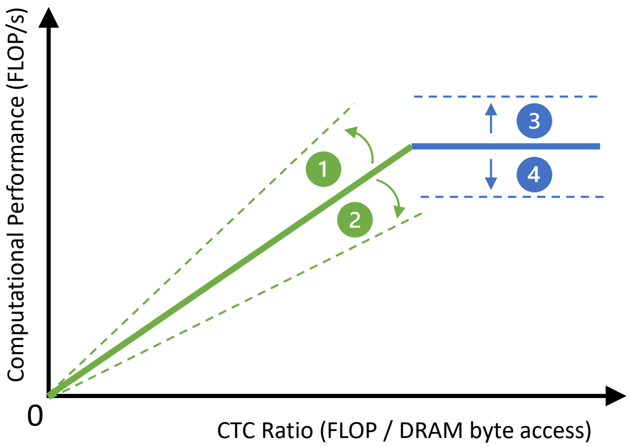
Given a specific computing system, it has a specific maximum computational throughput and a specific maximum memory access bandwidth that it can provide. In the roofline model figure, we can use two lines *“System Computational Roof”* and *“System Memory Bandwidth Roof”* to represent the maximum computational throughput (limited by the available hardware resource) and the maximum memory bandwidth of this system (limited by memory interface). Note that the slope of the system memory bandwidth roof represents the maximum memory bandwidth provided by the system.

Assume that we have a few accelerator designs, 1, 2, and 3. Each design has its own theoretical computational throughput and CTC ratio. Therefore we can plot the design in the roofline model figure, as shown in the figure. If draw a line between a design point and the origin, the slope represents the memory bandwidth required by that design. In our example above, design 1 has a higher theoretical computational throughput (“*B*”) than design 2 and design 3 (“*A*”). However, if we implement all these three design points with the system specified by the blue and green rooflines, design 2 and 3 can acheive the theoretical performance (computational and memory access throughput), but design 1 cannot. Since design 1 is requiring more memory bandwidth than what the system can provide, the performance of design 1 will be limited by the system memory access bandwidth. The attainable computational throughput of design 1 is “*B′*”, which has the same CTC ratio as the original design 1 and the memory bandwidth provided by the system.

Assume that you are building a DNN inference accelerator on a PYNQ FPGA board for your final course project. You decide to use a roofline model to analyze whether your design is compute bound or memory bound. Please answer the following questions.


Computational Performance (FLOP/s)
Line 3
⑥
Line 2
①
⑤
Line 1
②
③
④
0
CTC Ratio (FLOP / DRAM byte access)

(A)


Computational Performance (FLOP/s)
3
1
4
2
0
CTC Ratio (FLOP / DRAM byte access)

(B)

(a) *(10 pts)* What does compute bound and memory bound mean respectively? Briefly explain.

(b) *(10 pts)* In Figure (A), which lines (from line 1, line 2, and line 3) can separate compute-bounded and memory bounded designs?

(c) *(20 pts)* In Figure (A), please indicate all compute-bounded areas and memory-bounded areas (from area ① through ⑥).

(d) *(30 pts)* Assume that the maximum computational throughput and the memory bandwidth of your FPGA board are 100 GFLOP/s and 12.5 GB/s respectively. The theoretical peak computational throughput of your current design is 50 GFLOP/s and the CTC ratio is 1/12. Which area is your design located in (from area ① through ⑥)? What is the actual computational throughput (“attainable performance”) if you implement your design on this FPGA board?

(e) *(10 pts)* Assume that you switch to a larger FPGA board, which provides more compute units, higher archievable clock frequency, and more advanced memory system, how will the rooflines change? Please select all possible changes from ① through ④ in Figure (B).

2. *(20 points)* **Your Opinions.** (Note: you will get full points as long as you answer the questions)

(a) List any new topics/discussions that you would like to see in the future offerings of this course.

(b) What can be improved in this course? (e.g. course format, teaching style, course materials, etc.)