EE5142/921 U9360

高等積體電路設計 Advanced Integrated Circuit Design Homework 2

馬咏治 kane@access.ee.ntu.edu.tw
羅翊誠 alan@access.ee.ntu.edu.tw

Preliminaries:

Roofline Model (https://en.wikipedia.org/wiki/Roofline_model)

The roofline model is an intuitive visual performance model used to provide performance estimates of a given compute kernel or application running on multi-core, many-core, or accelerator processor architectures, by showing inherent hardware limitations, and potential benefits and priority of optimizations. It helps you identify the bottleneck for multi-core architecture, guides you through optimization, and tells you when to stop.

The most basic roofline model can be visualized by plotting floating-point performance as a function of arithmetic intensity. The resultant curve is effectively a performance bound under which kernel or application performance exists, and includes two platform-specific performance ceilings: a ceiling derived from the memory bandwidth and one derived from the processor's peak performance (see Figure 1).

- The arithmetic intensity (AI), also referred to as operational intensity, is the ratio between the work $\boldsymbol{W}$ to the memory traffic $\boldsymbol{Q}$

$$I = \frac{W}{Q},$$

where the work $\boldsymbol{W}$ denotes the number of operations performed by a given kernel or application, and the memory traffic $\boldsymbol{Q}$ denotes the number of bytes of memory transfers incurred during the execution of the kernel or application.

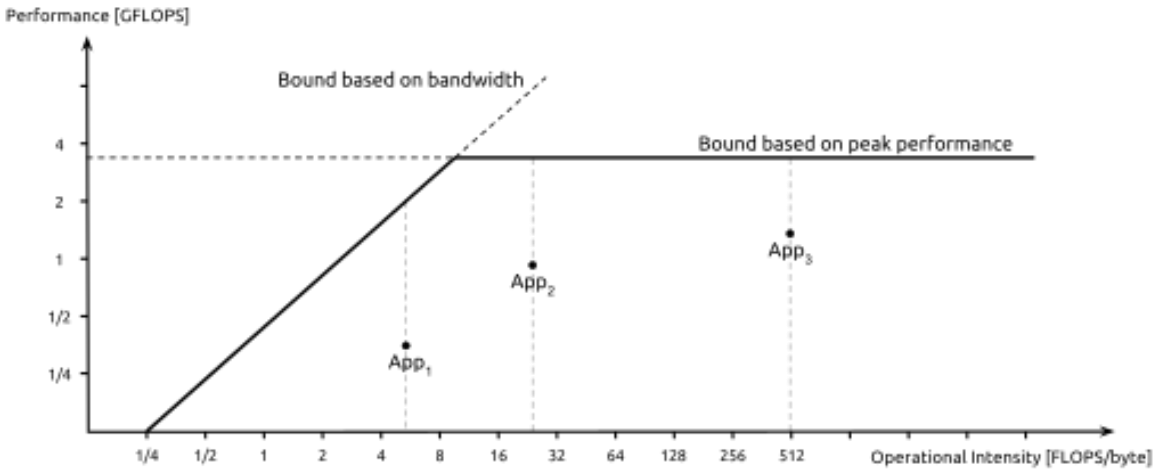


Figure 1: *An example of a Roofline model in its basic form. As the image shows, the curve consists of two platform-specific performance ceilings: the processor's peak performance and a ceiling derived from the memory bandwidth. Both axes are in logarithmic scale.*

How to read Roofline Performance Model? (Source: https://www.nersc.gov/assets/Uploads/Tutorial-ISC2019-Intro-v2.pdf)

For a multi-core architecture, the sustainable performance is bound by:

$$\text{GFLOP/s} = \min\begin{cases}\text{Peak GFLOP/s}\\ \text{AI} * \text{Peak GB/s}\end{cases},$$

where

Arithmetic Intensity (AI) = FLOPs/Bytes.

The "roofline" shown in Figure 2 indicates the theoretical performance of a multi-core architecture, which are *Bandwidth-bound* and *Compute-bound*.

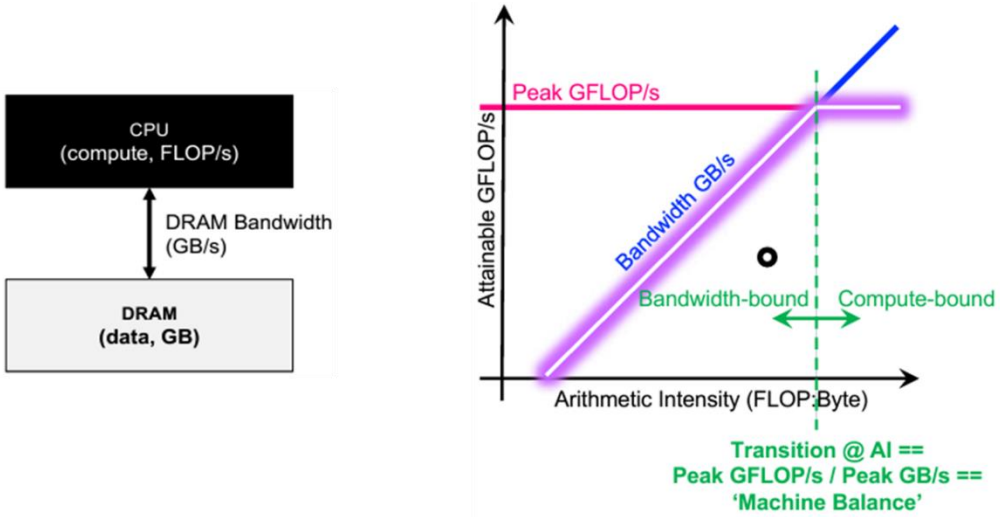


Figure 2: *Left:* a schematic diagram of a system, it consists of a computational engine and memory system. *Right*: a roofline model example. It illustrated the region of bandwidth-bound and compute-bound.

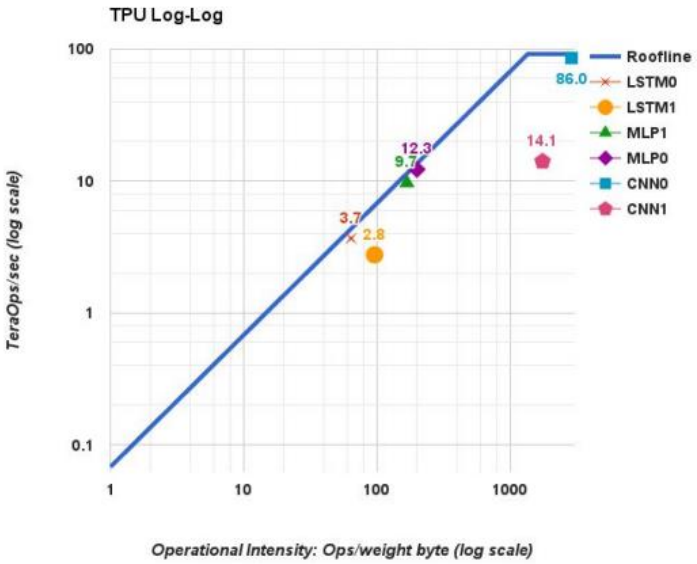
- *Bandwidth-bound* occurs on LOW arithmetic intensity program. These programs have low data locality, which means many data movement are involved for making certain amount of FLOP. Typically, it is bounded by DRAM of other data transition interface.
- *Compute-bound* occurs on HIGH arithmetic intensity program. These programs have high data locality, which means relatively few data movements are involved for making a certain amount of FLOP. Those data could be reused multiple times for computations (A huge amount of FLOP with few data movements). In this situation, the bottleneck of the system is no longer limited by memory bandwidth, but the computation power of its core.
- *Machine Balance* occurs when the AI is equal to $\frac{\text{Peak GFLOP/s}}{\text{Peak GB/s}}$.

Reading Material:

[1] Williams, S., Waterman, A., & Patterson, D. (2009). Roofline: an insightful visual performance model for multi-core architectures. *Communications of the ACM*, *52*(4), 65-76.

[2] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017, June). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture* (pp. 1-12).

HW 2 Questions:



1. (20%)
   The figure above compares the arithmetic intensity (AI) of three types of layers: CNN, LSTM, and MLP. Please sort these three layer types in ascending order regarding its AI. Discuss the difference between these three layer types, and explain why some of them have higher AI while others have lower AI.

2. (20%)
   The figure above shows the roofline curve regarding Google TPU, a matrix multiply unit (systolic array) of 256x256 PE array. Please plot the roofline curve if the TPU has upgraded its matrix multiply unit to a 320x320 PE array.

3. (60%)

Model	*Die*										*Benchmarked Servers*				
	mm²	*nm*	*MHz*	*TDP*	*Measured*		*TOPS/s*		*GB/s*	*On-Chip Memory*	*Dies*	*DRAM Size*	*TDP*	*Measured*	
					Idle	*Busy*	8b	FP						*Idle*	*Busy*
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	NA*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

   The table above shows the benchmark servers which use Haswell CPUs, K80 GPUs, and TPU. Please use the information provided by the table above to plot the roofline curve of these three benchmark servers (x-axis[log scale]: Ops/weight byte; y-axis[log scale]: TeraOps/sec), and find the value of arithmetic intensity for achieving *machine balance*.

* We assume that each parameter equals to 1 weight byte.

Submission requirement:

1. The report should be merged as a single **pdf file** and **uploaded to NTU COOL**.
   Example of filename: AVLSI_HW2_d09943011.pdf
   Note that you have to replace d09943011 with your student ID number.
2. **Deadline: 2022/09/26 23:59**

The homework will be graded ONLY IF the filename of your submission is correct!