HW2

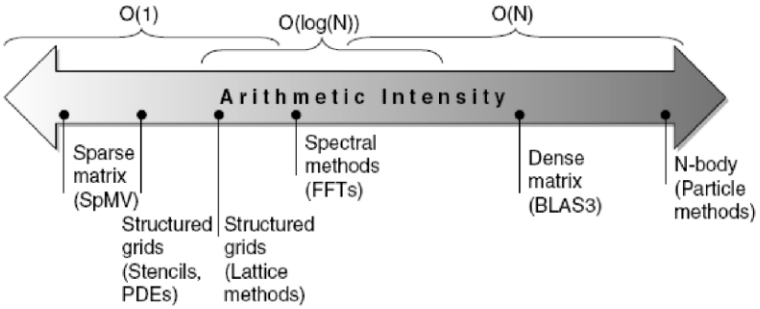
Speaker：Alan

Advisor：Prof. An-Yeu Wu

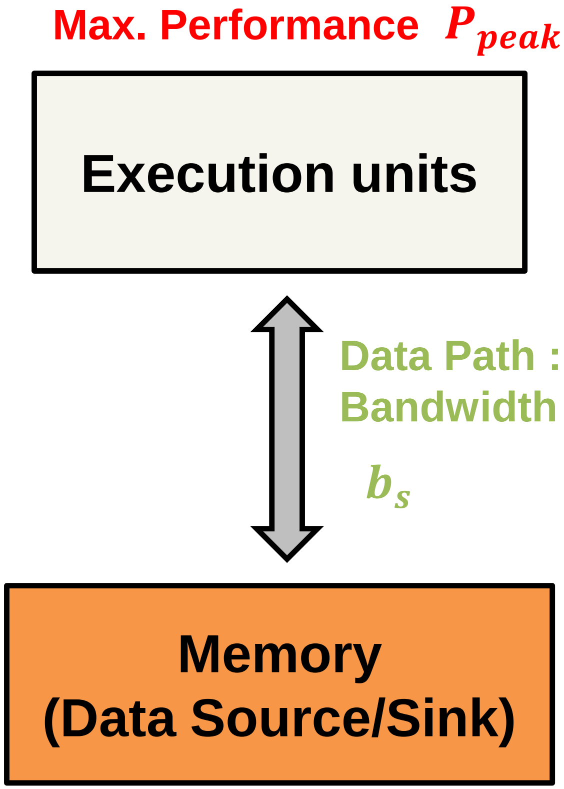
Date：2022/09/20

Roofline Visual Performance Model

- A simple performance model : Execution vs. Data Transfer
- Three key factors
  - System Spec (Hardware Level)
    - Computation : peak floating point performance
      - Floating-point ops /sec
    - Memory: peak memory bandwidth
      - Bytes per sec
  - Program characteristics (Algorithm Level)
    - Arithmetic intensity : Floating-Point Ops/ byte
      - Ratio of floating-point operations in a program to the number of data types accessed by a program from main memory



Simplistic view


Max. Performance P_{peak}

Execution units

Data Path : Bandwidth

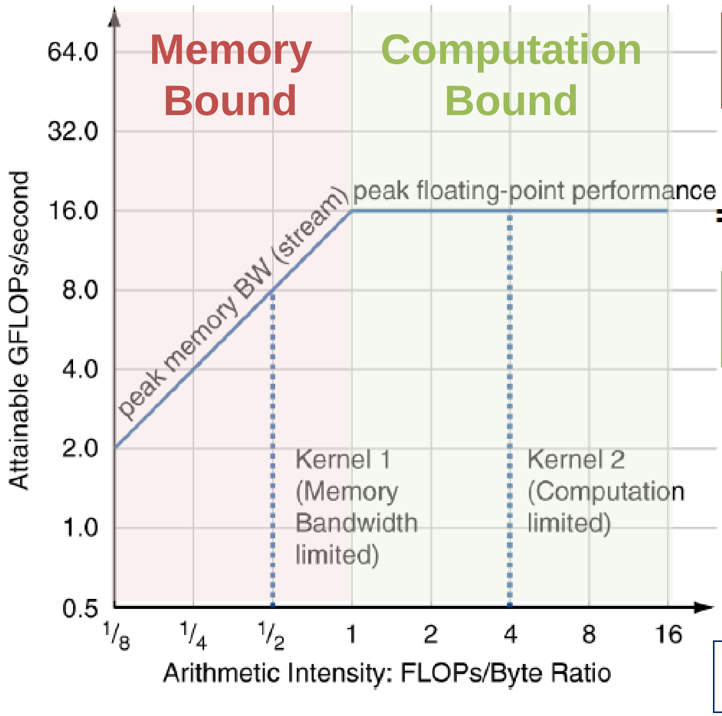
b_s

Memory (Data Source/Sink)

- How fast can tasks be processed?
- The Bottleneck:
  - The execution of Work:
    - P_{peak} [flops : flop/s]
  - The data path
    - $I * b_s$ [flop/byte * byte/s]
    - I : Arithmetic Intensity
- Roofline Model Equation

$$P = \min\left(P_{peak}, I * bs\right)$$

Roofline Diagram



Memory Bandwidth Limited

Floating-Point Ops/sec

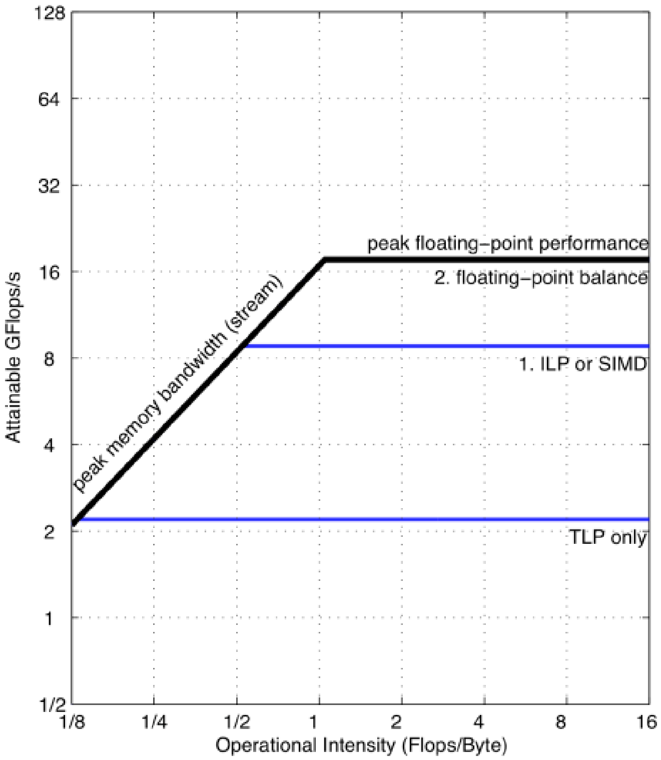
= Peak Bytes/sec x Floating-Point Ops/ byte

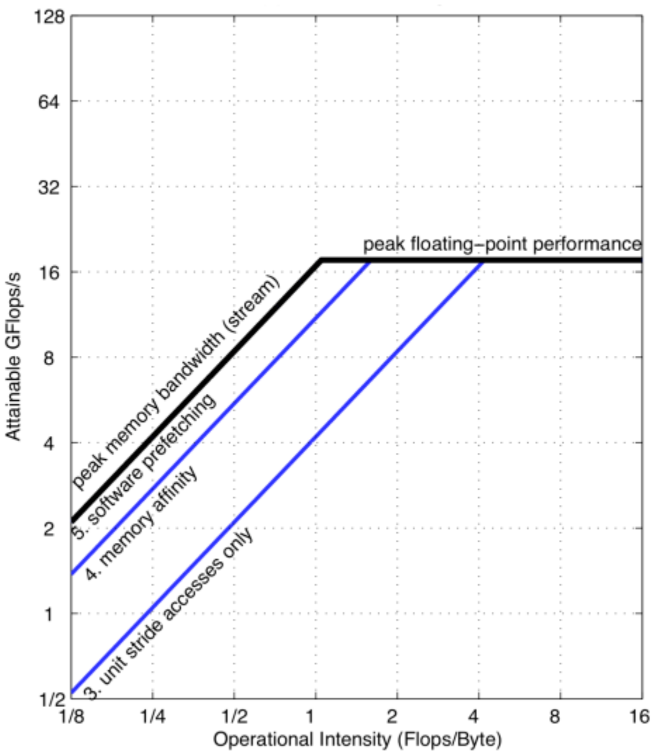
Computation Limited

Peak Floating-Point Ops/sec

(assuming 16GB/sec peak bandwidth)

Roofline Analysis



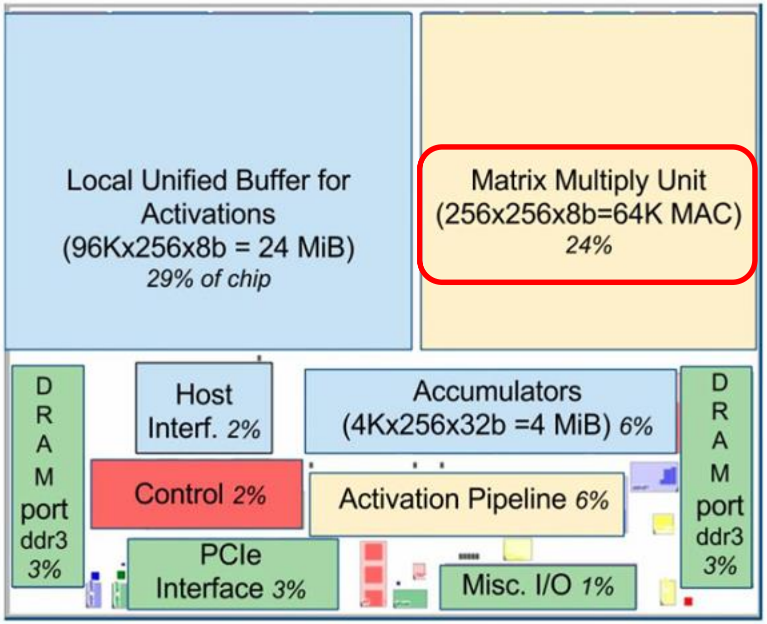


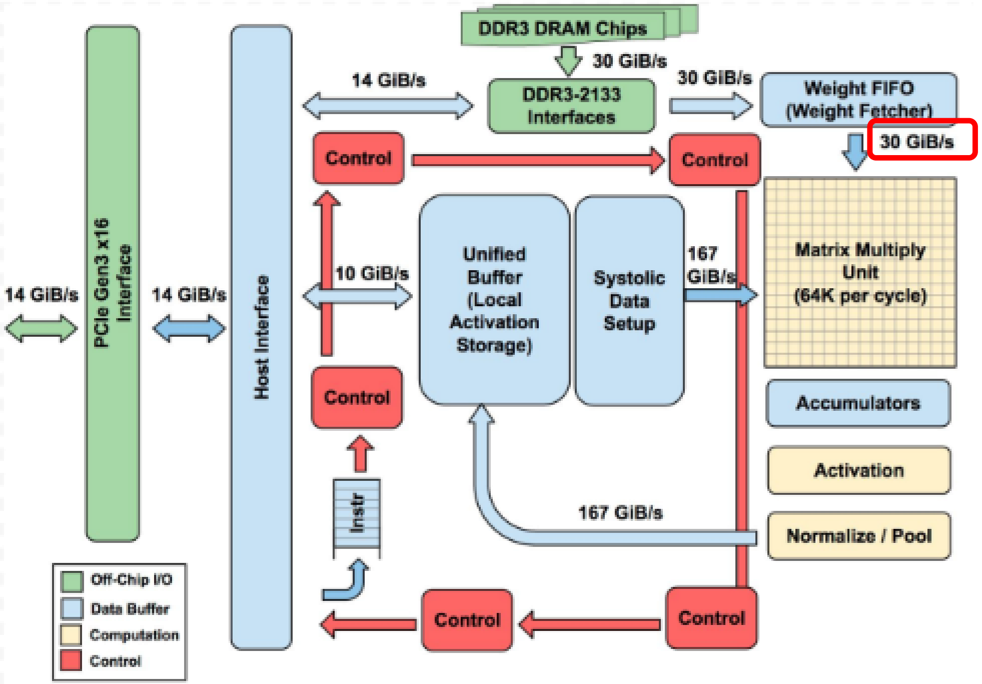
- The Roofline model gives an upper bound to performance.
  - Need some techniques to achieve the ceiling

*TLP : Thread-Level Parallelism
ILP : Instruction-Level Parallelism

HW2 – Roofline Model

- TPU Example


Local Unified Buffer for Activations (96Kx256x8b = 24 MiB) 29% of chip
Matrix Multiply Unit (256x256x8b=64K MAC) 24%
D R A M port ddr3 3%
Host Interf. 2%
Accumulators (4Kx256x32b =4 MiB) 6%
D R A M port ddr3 3%
Control 2%
Activation Pipeline 6%
PCIe Interface 3%
Misc. I/O 1%



Peak Performance :

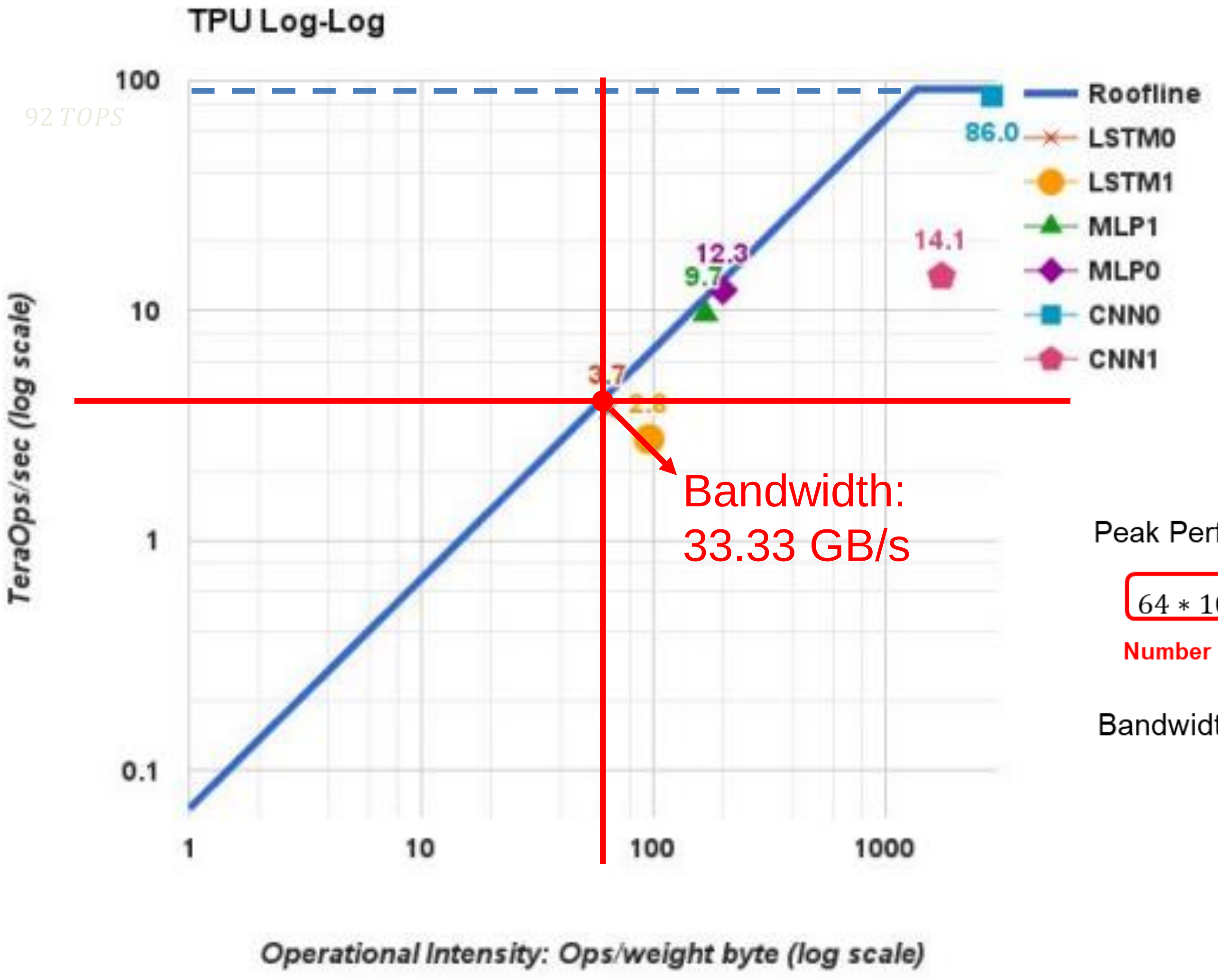
$$\boxed{64 * 1024} * \boxed{2} * \boxed{700 * 10^6} = 91.7504\ Tops$$

Number of PE MAC Frequency

Bandwidth:

$$30 * \frac{2^{30}}{10^9} = 32.2GB/s$$

HW2 – Roofline Model



Bandwidth:
33.33 GB/s

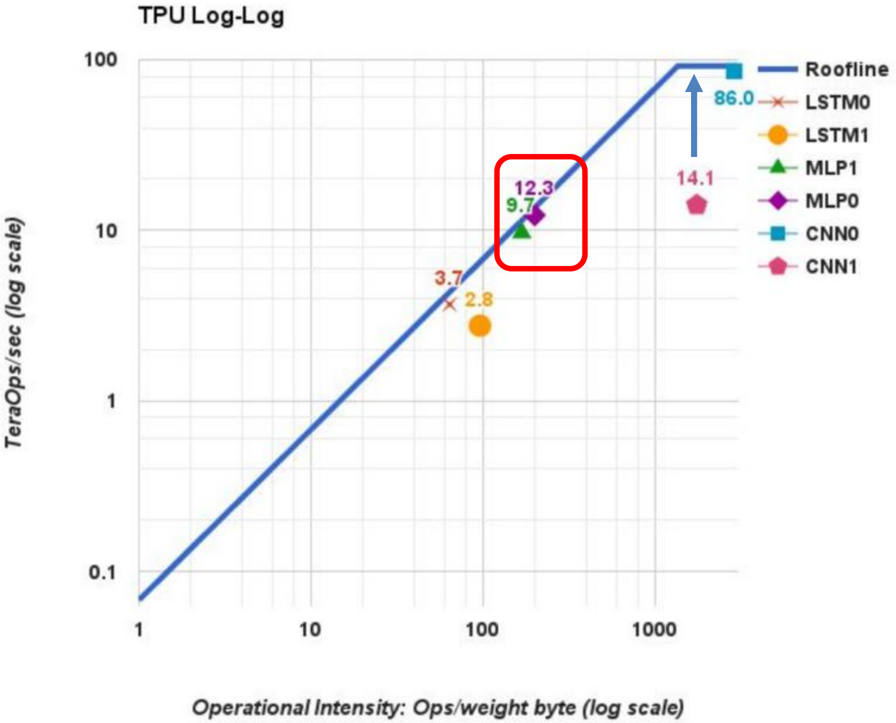
Peak Performance :

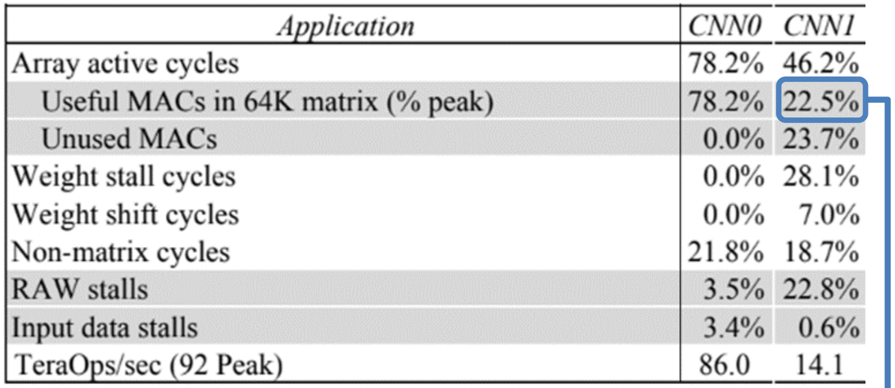
$$\underbrace{\boxed{64 * 1024}}_{\text{Number of PE}} * \underbrace{\boxed{2}}_{\text{MAC}} * \underbrace{\boxed{700 * 10^6}}_{\text{Frequency}} = 91.7504\, Tops$$

Bandwidth:

$$30 * \frac{2^{30}}{10^9} = 32.2GB/s$$

Roofline Model : TPU





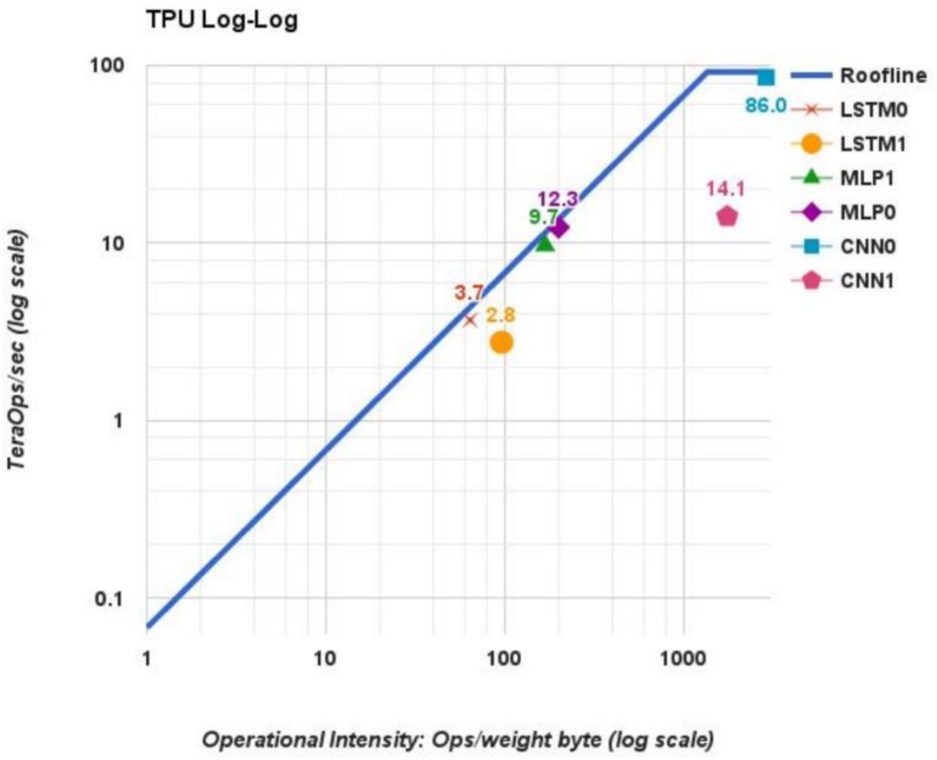
Application	*CNN0*	*CNN1*
Array active cycles	78.2%	46.2%
Useful MACs in 64K matrix (% peak)	78.2%	22.5%
Unused MACs	0.0%	23.7%
Weight stall cycles	0.0%	28.1%
Weight shift cycles	0.0%	7.0%
Non-matrix cycles	21.8%	18.7%
RAW stalls	3.5%	22.8%
Input data stalls	3.4%	0.6%
TeraOps/sec (92 Peak)	86.0	14.1

Low Utilization

- Fully connected layer is less operation-intensive than convolution layer.
- CNN1 has some layers with shallow feature depths.
  - Utilization is not high
    - The actual efficiency is far away from ceiling

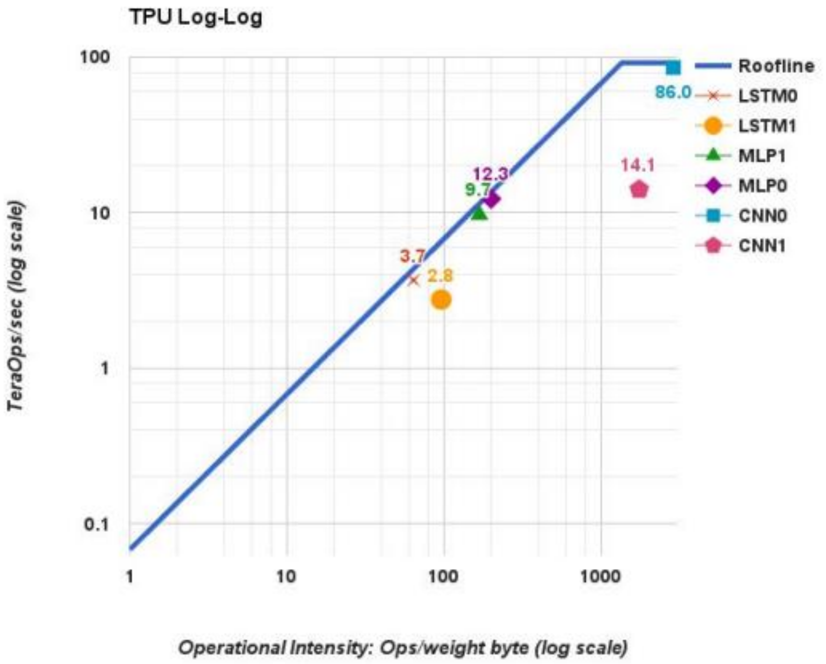
Problem 1 (20 points)

- Recap the concept of arithmetic intensity (AI)
  - Compare and discuss the arithmetic intensity of *LSTM*, *MLP*, and *CNN*



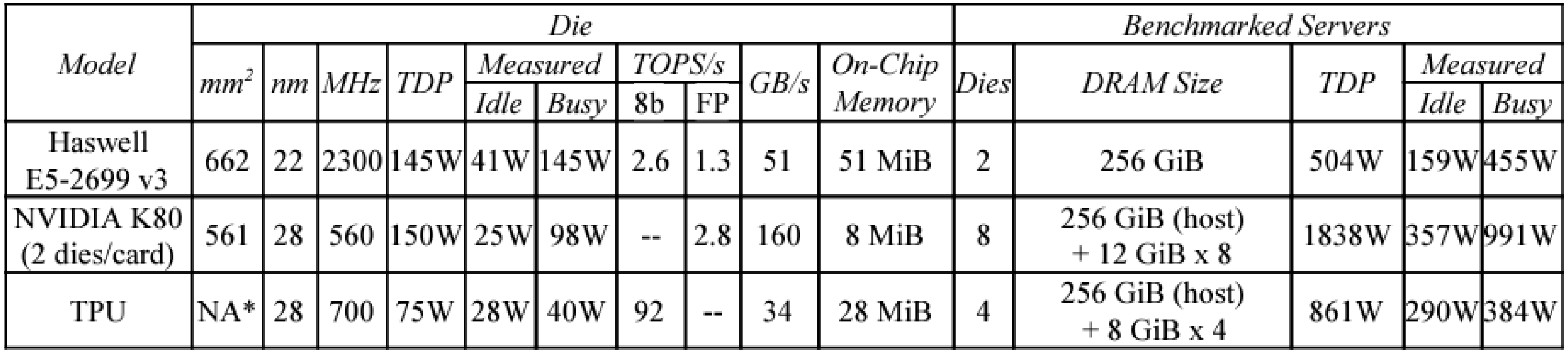
Problem 2 (20 points)

- Goal
  - Understand the meaning of the roofline model
  - Adjust the roofline according to different specification
- Plot the roofline curve if the TPU has upgraded its PE array
  - From 256x256 to 320x320



Problem 3 (60 points)

- Goal
  - Understand the meaning of the roofline model
  - Adjust the roofline according to different specification
- Plot the roofline model if we change the hardware
  - CPU (Haswell), GPU (Nvidia K80), and TPU



Model	*Die*										*Benchmarked Servers*				
	mm^2	*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

Requirements

- 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
- Deadline: 2022/09/26 23:59
  - Late submission will only get half score (deadline: 2022/09/30 23:59)