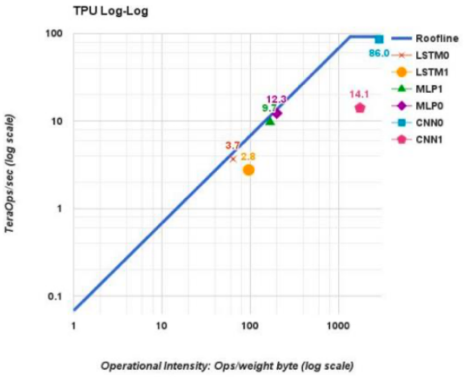
**高等積體電路設計 Advanced VLSI** **HW#2**

電子所 R10943170 連德宇

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.



Sol:

根據 Roofline Model 的定義，X 軸表示 Operational Intensity，Y 軸 W 表示計算平台的性能上限(work)，斜率是取決於 Bandwidth 的大小，而 arithmetic intensity (AI) 則是$I = \frac{W}{Q}$

故由 Arithmetic intensity 大小可以將圖上各點由小到大做排序如右 LSTM0, LSTM1, MLP1, MLP0, CNN 1, CNN0

CNN0，充分利用了 Bandwidth 且達到了平台 work 的上限，位於 Compute Bound 位置，故 AI 值最大

CNN1，位於 Compute Bound 位置，但卻未在 Roofline Model 的線上，故 Bandwidth 應該有達到上限，但資料量卻未達到了平台 work 的上限，故 AI 值次大

MLP0，位於 Memory Bound 位置，在 Roofline Model 的線上，Bandwidth 未達到平台上限，故 AI 值第三大

MLP1，位於 Memory Bound 位置，在 Roofline Model 的線上，Bandwidth 未達到平台上限，較 MLP0 小，故 AI 值第四大

LSTM1，位於 Memory Bound 位置，但並未在 Roofline Model 的線上，Bandwidth 未達到平台上限，較 MLP1 小，故 AI 值第五大

LSTM0，位於 Memory Bound 位置，在 Roofline Model 的線上，Bandwidth 未達到平台上限，較 LSTM1 小，故 AI 值第最小

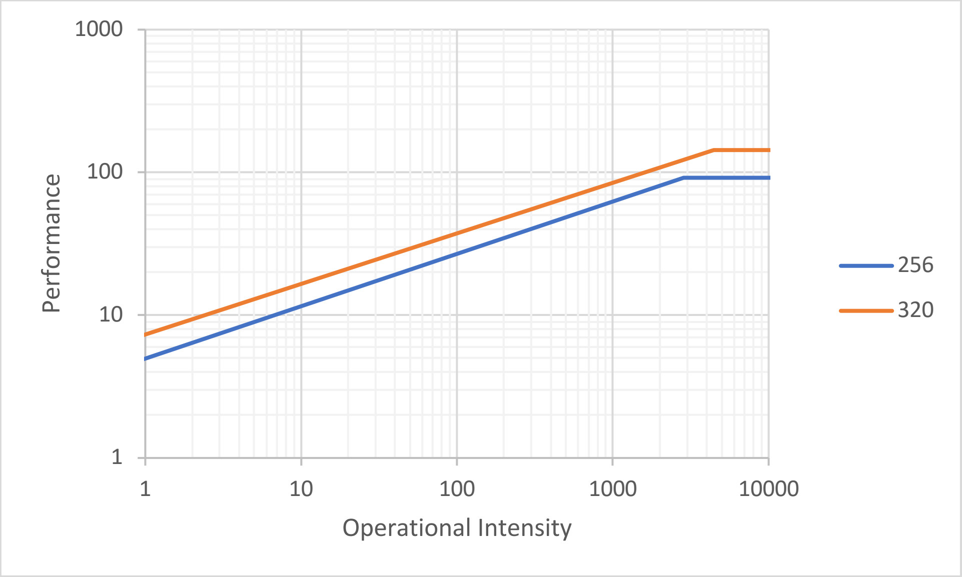
2. (20%)

   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.

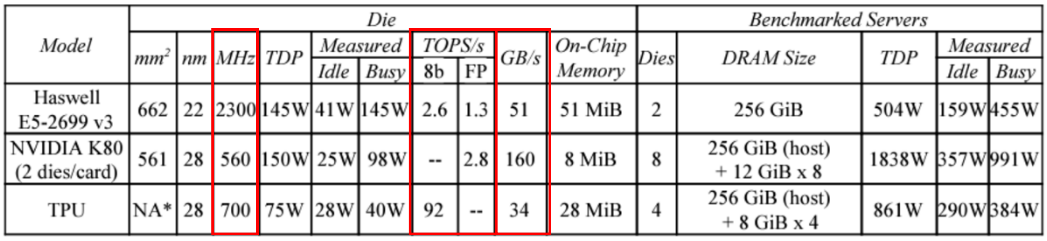
Sol:

320*320 PE array's Peak Performance: $320 * 320 * 8\text{b} * 1024 * 2 * 700 * 10^6 = 143.36$ Tops

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



3. (60%)



| 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 |

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.

Sol:

Haswell CPUs:

Peak Performance: $(2.6 + 1.3) * 2 = \ 7.8$ Tops

Bandwidth: $51 * \frac{2^{30}}{10^9} = 54.760833\ GB/s$

K80 GPUs:

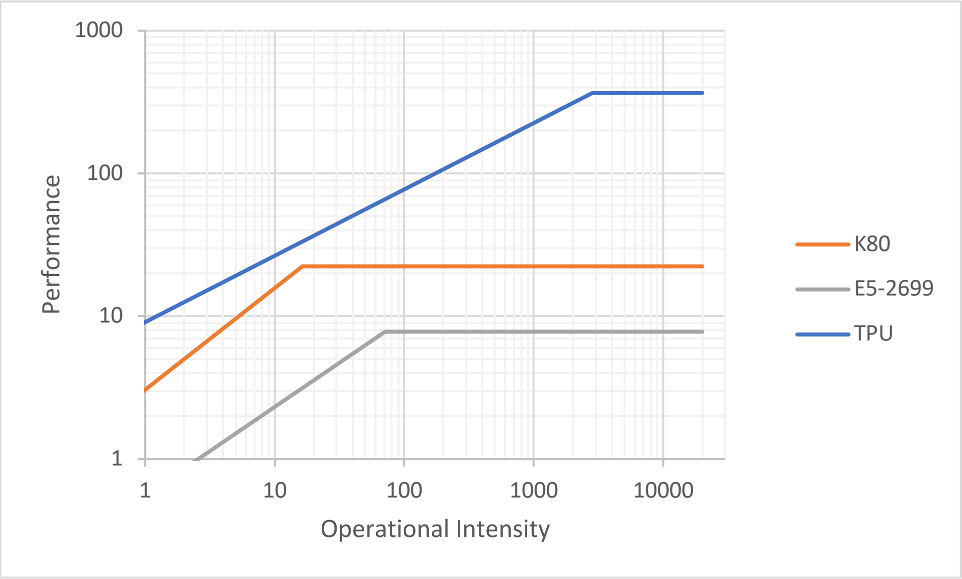
Peak Performance: $2.8 * 8 = 22.4$Tops

Bandwidth: $160 * \frac{2^{30}}{10^9} = 171.79869\ GB/s$

TPU:

Peak Performance: $92 * 4 = 368$ Tops

Bandwidth: $34 * \frac{2^{30}}{10^9} = 36.5\ GB/s$


Performance
Operational Intensity
K80
E5-2699
TPU