Question 1:

How long does it take to load 200B bf16 model into the systolic array of 32 TPU v4p from HBM

TPU v4p HBM to Tensorcore bandwidth

$= 1.2 \times 10^{12}$

Model size $= 2 \times 10^{9}$

$\text{Sharded\_model} = \frac{2 \times 2 \times 10^{9}}{32} \times 100$

$$\frac{\text{Sharded\_model}}{\text{HBM to Tensorcore BW}} = \frac{\frac{2 \times 2 \times 10^{9} \times 100}{32 \times 16}}{1.2 \times 10^{12}} = \frac{2 \times 1 \times 100}{1.2 \times 16 \times 10^{3}} \text{ sec}$$

$$= \frac{2 \times 1 \times 100}{19.2 \times 10^{3}}$$

$$= 0.052 \times 10^{-3} \times 100$$

$$= 2 \times 5.2 \times 10^{-5} \text{ sec}$$

$$= 10.4 \times 10^{-5} \times 100$$

$$= 10.4 \times 10^{-3}$$

$$= 10 \text{ ms}$$

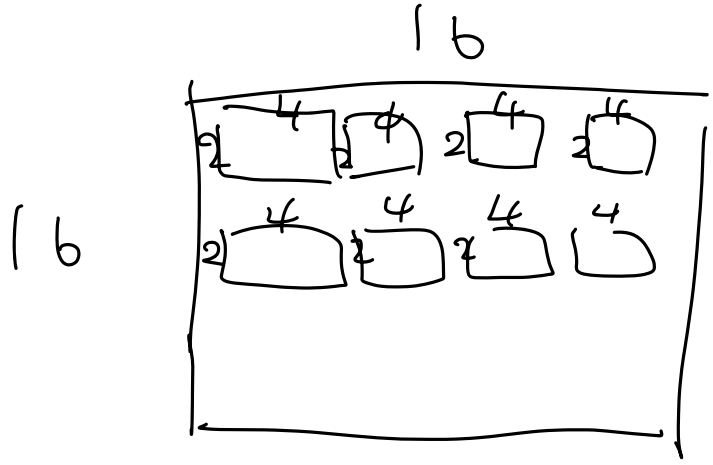
Question 2: TPU v5e pod

① How many CPU hosts are there?

TPU v5e pod size = 16 x 16

Every 4X2 contains a CPU

So = $\frac{16}{4} \times \frac{16}{2} = 4 \times 8 = 32$ CPU hosts ✓



② How many TPU Tensor Cores?

16 x 16 x 1 = 256

TPU v5e has 1 TPU core per chip

③ Total FLOPs/sec for the whole pod

$1.97 \times 10^{14} \times 16 \times 16 = 504.32 \times 10^{14} = 5.04 \times 10^{16}$

④ Total HBM = 16 GB x 16 x 16

= 4096 GB = 4 TB

Question 3: PCIe operational intensity