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 Tensor Core bandwidth

$= 1.2 \times 10^{12}$

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

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

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

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

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

$$= 100\times 2\times 5.2\times 10^{-5}\text{sec}$$

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

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

$$= 10\text{ ms}$$

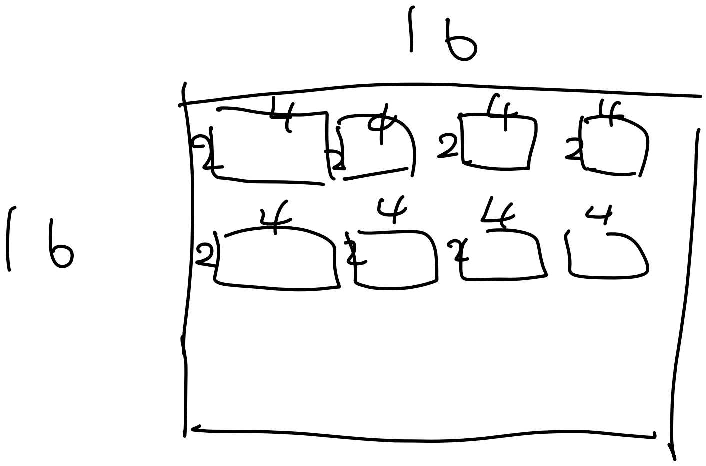
Question 2: TPU v5e pod

① How many CPU hosts are there?

TPU v5e Pod size = 16 x 16

Every 4X2 contains a CPU

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



② How many TPU TensorCores?

$16 \times 16 \times \underset{\sim}{1} = 256$

TPU v5e has 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 $= 16\,\text{GB} \times 16 \times 16$

$= 4096\ \text{GB} = 4\,\text{TB}$

Question 3 : PCIe operational intensity

$A = \text{bfloat}[D, F]$

x Activations $= \text{bfloat}[B, D]$

on DRAMs

$$\text{FLOPs} = BF(\underbrace{\overset{\text{mul}}{D}} + \underbrace{\overset{\text{add}}{D - 1}})$$

$$= BF(2D-1) = 2BDF - BF$$

$$\simeq \underline{2BDF}\ \checkmark$$

$$\text{FLOPs} = 2BDF \overset{F=4D}{=} 2BD \cdot 4D = 8BD^2$$

$$\text{Bytes} = \underline{\overset{\text{Load}}{2(DF + BD)}} + \underbrace{\overset{\text{write}}{2BF}}$$

$$\overset{F=4D}{=} 2(4D^2 + BD) = 8D^2 + 2BD \simeq 8D^2$$

$$\frac{\text{FLOPs}}{\text{Bytes}} = \frac{8BD^2}{8D^2} = B$$

Assume single chip

$$AI_{hardware} = \frac{\cancel{4\times 2}\times 9.2\times 10^{14}}{1.5\times 10^{10}}$$

$$= \frac{\cancel{8}\times 9.2\times 10^{4}}{1.5} = \cancel{49\times 10^{4}} \quad 61250$$

$$= \cancel{4.9\times 10^{5}}$$

Question 4: general matmul latency

Multiply int8[16384, 4096] by int8[B, 4096]

Assumes: 1 TPUv5e (chip)

1. ✓ Time = max(bytes load/write, FLOPs time)

$$= \max\left(\frac{16384\times 4096 + B\cdot 4096 + B\cdot 16384}{8.1\times 10^{11}}, \frac{B\times 16384\times 2\times 4096}{3.94\times 10^{14}} \right)$$

$$= \max\left(\frac{6.7\times 10^{7} + 20480B}{8.1\times 10^{11}} + \frac{1.3\times 10^{8}B}{3.94\times 10^{14}} \right)$$

Compute-bound when $\frac{1.3\times10^{8}B}{3.94\times10^{14}} > \frac{6.7\times10^{7}+2\times10^{4}B}{8.1\times10^{11}}$

$$\Leftrightarrow \frac{1.3B}{3.94\times10^{6}} > \frac{6.7\times10^{3}+2B}{8.1\times10^{7}}$$

$$\Leftrightarrow 1.3\times8.1\times10^{7}B > 6.7\times10^{3}\times3.94\times10^{6} + 2\times3.94\times10^{6}B$$

$$\Leftrightarrow 10.53\times10^{7}B - 7.88\times10^{6}B > 26.398\times10^{9}$$

$$\Leftrightarrow 105.3\times10^{6}B - 7.88\times10^{6}B > 2.6\times10^{10}$$

$$\Leftrightarrow 97.42\times10^{6}B > 2.6\times10^{10}$$

$$\Leftrightarrow B > \frac{2.6\times10^{4}}{97.42}$$

$$\Leftrightarrow B > 268$$

2. Assumes: bandwith $= 22\times8.1\times10^{11}$

Compute-bound when:

$$\frac{1.3\times10^{8}B}{3.94\times10^{14}} > \frac{6.7\times10^{7}+2\times10^{4}B}{22\times8.1\times10^{11}}$$

$\downarrow 10^{-8}$ $\qquad$ $\downarrow 10^{-4}$

$$\Leftrightarrow \frac{1.3\times B}{3.94\times10^{6}} > \frac{6.7\times10^{3}+2B}{22\times8.1\times10^{7}}$$

$$\Leftrightarrow \quad 1.3 \times 22 \times 8.1 \times 10^{7} B > 3.94 \times 6.7 \times 10^{9} + 2 \times 3.94 \times 10^{6} B$$

$$\Leftrightarrow \quad 231.66 \times 10^{7} B > 26.398 \times 10^{9} + 7.88 \times 10^{6} B$$

$$\Leftrightarrow \quad 2316.6 \times 10^{6} B - 7.88 \times 10^{6} B > 2.6398 \times 10^{10}$$

$$\Leftrightarrow \quad 2308.72 \times 10^{6} B > 2.6398 \times 10^{10}$$

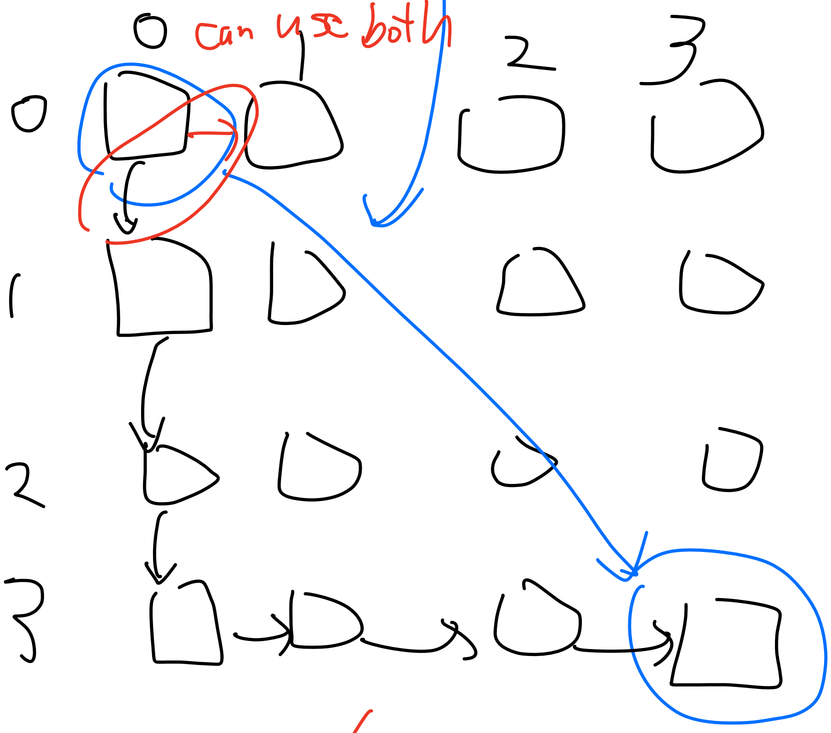
$$\Leftrightarrow \quad 2.30872 \times 10^{9} B > 2.6398 \times 10^{10}$$

$$\Leftrightarrow \quad B > \frac{2.6398 \times 10^{10}}{2.30872 \times 10^{9}} \simeq \underline{\underline{11.43}} \quad \#$$

Question 5: ICI bandwidth

TPU v5e : 4x4 slice

bfloat[8, 128, 8192]



✓

1. $\underbrace{6 \times 1\,\mu s}_{(hops)}$ = first byte latency

2. $$\text{Total time} = 6\,\mu s + \frac{2 \times 8 \times 128 \times 8192}{2 \times 4.5 \times 10^{10}}$$

$$= 6\,\mu s + \frac{16{,}777{,}216}{2 \times 4.5 \times 10^{10}}$$

$$= 6\,\mu s + \frac{1.6 \times 10^{7}}{2 \times 4.5 \times 10^{10}}$$

$$= 6\,\mu s + \frac{1.6}{2 \times 4.5 \times 10^{3}}$$

$$= 6 \times 10^{-6}\,\text{sec} + \underbrace{0.35}_{2} \times 10^{-3}\,\text{sec}$$

$$= 6 \times 10^{-6}\,\text{sec} + \underbrace{3.5}_{2} \times 10^{-4}\,\text{sec}$$

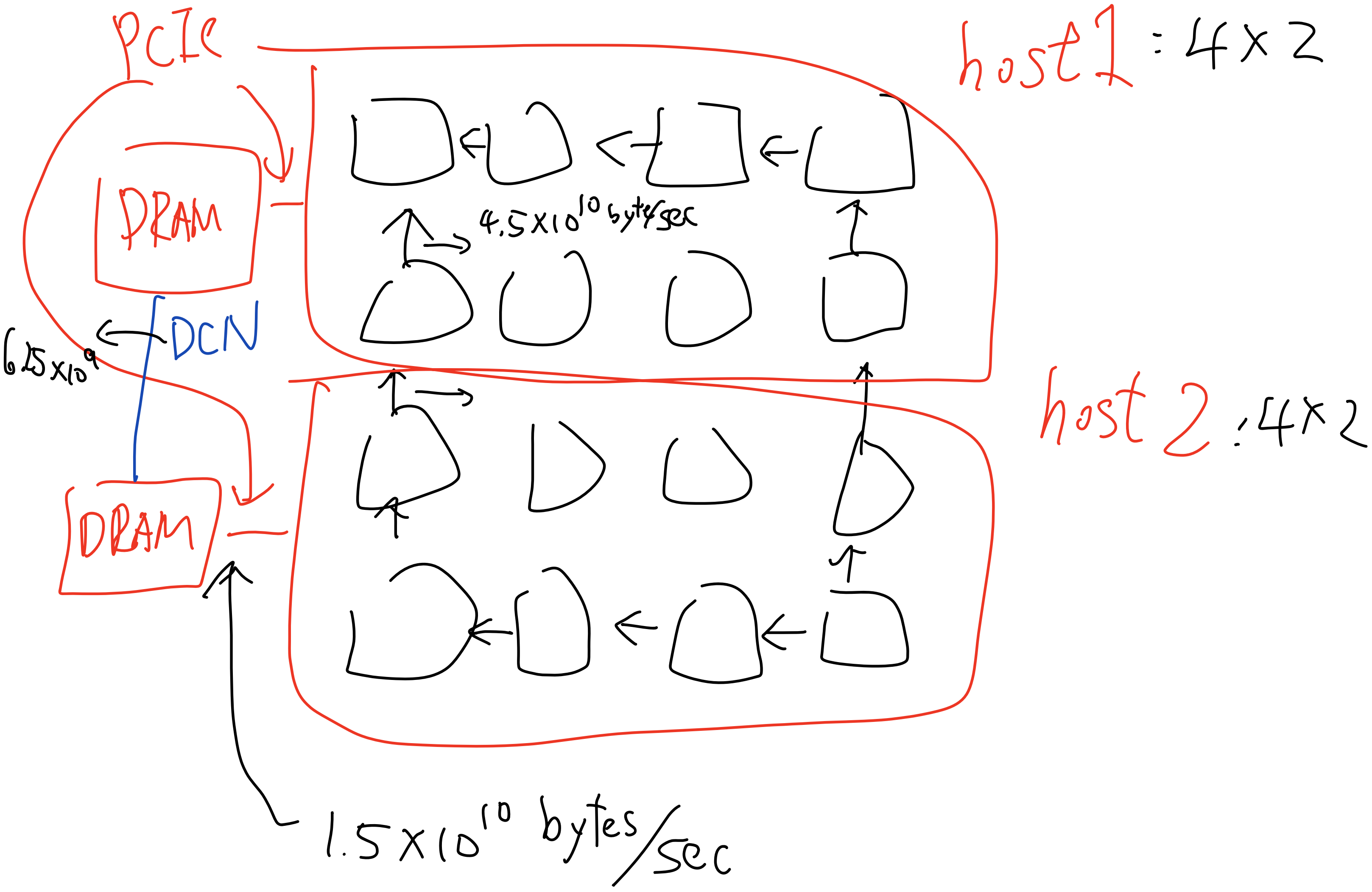
= ~~356~~ μs

181

Question 6:

Matrix A: int8[128x1024, 128x1024]

Sharded evenly across TPU v5e 4x4 slice but they are on host's DRAM on each chip.



First : We can choose either :

① Copy half of the data through DCN to host 1's DRAM and then copy things to HBM from DRAM

or ② Each host copy its data to TPU chips and then utilize ICI to copy data

Option ① : DCN throughput is slow, so use ②

Steps : (will take the max over those numbers)

1. Copy data from DRAM to HBM
2. Transfer data from all chips to TPU {0, 0}
3. Compute int8[$\overset{2^7}{\overset{\|}{128}}$ x $\overset{2^{10}}{\overset{\|}{1024}}$, 128 x 1024]
   bfloat16[8, 128 x 1024]

1. $128 \times 1024 \times 128 \times 1024 = 2^{34} = 2^{4}$ GB

$= 16$ GiB

$\frac{16 \text{ GiB}}{2} = 8 \text{ GiB}$ = data on each host

Time to load from DRAM to HBM to all TPU chips through PCIe

$$= \frac{8 \text{ GiB}}{16 \times 1.5 \times 10^{10} \text{ bytes/sec}} \checkmark = 0.035 \text{ sec} = 35 \text{ms}$$

2. Time to send data from HBM to TPU{0,0} through ICI;

15 GiB because TPU{0,0} needs to receive 15 GiB

$$\frac{15 \cancel{8} \text{ GiB}}{2 \times 4.5 \times 10^{10}} = \frac{\overset{15}{\cancel{8}} \text{ GiB}}{9 \times 10^{10} \text{ bytes/sec}}$$

$= 0.178$ sec

$= 178$ ms

3. Time to load and write to MXU

Load int8[128×1024, 128×1024] $= \frac{128 \times 1024 \times 128 \times 1024}{8.1 \times 10^{11}}$ $= 0.021$ $= 21$ms

Load bf16[8, 128×1024] = I think the problem assume it is already in TPU HBM

Write bf16[8, 128×1024] $= 2 \times (8 \times 128 \times 1024)$

$= 2 \times 2^{3} \times 2^{7} \times 10^{10}$

$= 2\text{MiB}$

$$\frac{2\text{MiB}}{8.1\times10^{11}\,(\text{HBM speed})} = \text{neglible}$$

3. Compute time :

$$\frac{2BDF}{1.97\times10^{14}\,\text{FLOP/sec}} = \frac{2\times(8\times128\times1024\times128\times1024)}{1.97\times10^{14}\ \text{FLOPs/sec}}$$

$$= \frac{2\times(2^3\times2^7\times2^{10}\times2^7\times2^{10})}{1.97\times10^{14}}$$

$$= \frac{2^{38}}{1.97\times10^{14}}$$

$$= \frac{256\ \text{GiB}}{1.97\times10^{14}} = 0.00139\text{sec}$$

$= 1.3\text{ms}$

Total = max stage

= 178ms