$$T_{comm} = \frac{\text{Comm bytes}}{\text{bandwith bytes/sec}}$$

depend on how you write code

$$T_{math} = \frac{\text{computation FLOPs}}{\text{Accelerator FLOPs/sec}}$$

Constant, determined by hardware

$$T_{lower} = \max(T_{math}, T_{comms})$$

(assuming overlap computation)

$$\text{Arithmetic intensity} = \frac{\text{FLOPs}}{\text{bytes}}$$

When

$T_{math} > T_{comm}$, this means communication can be overlapped with T_{math},

so actual $\frac{\text{FLOPs}}{\text{sec}} = \frac{\text{Computation FLOPs}}{T_{math}}$

$= \text{Accelerator FLOPs/sec}$

(constant)

$T_{math} > T_{comm}$ also means

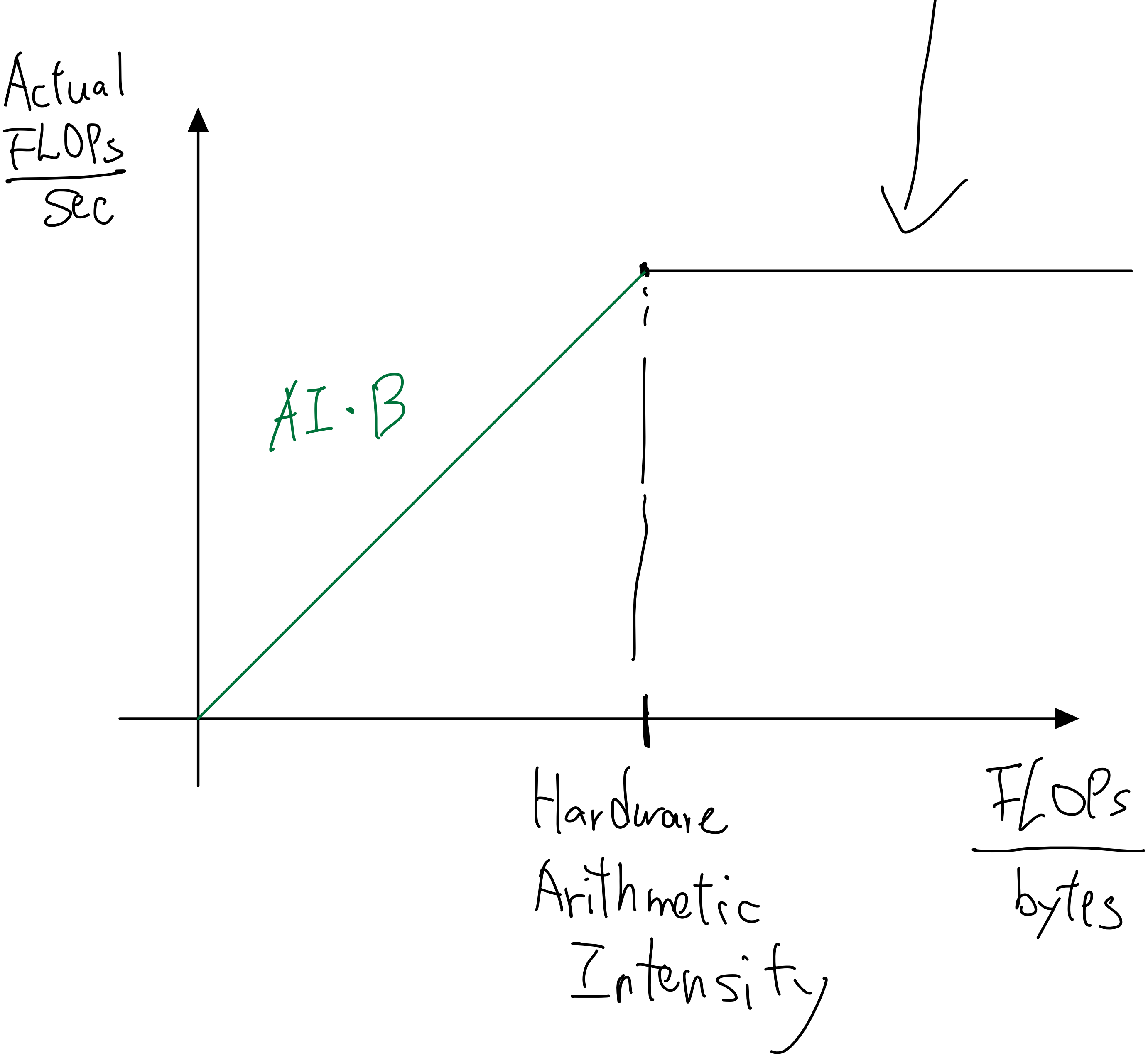
$$\Leftrightarrow \frac{\text{Computation FLOPs}}{\text{Accelerator FLOPs}} > \frac{\text{Communication bytes}}{\text{bandwith } \frac{\text{bytes}}{\text{sec}}}$$

$$\Leftrightarrow \underset{(\text{compute})}{\text{Arithmetic Intensity}} > AI_{(\text{hardware})}$$

Summary: when $AI_{\text{compute}} > AI_{\text{hardware}}$

actual $\frac{\text{FLOPs}}{\text{sec}}$ realized is

always Accelerator FLOS/sec



Actual FLOPs / sec
AI · B
Hardware Arithmetic Intensity
FLOPs / bytes

When

$$T_{math} < T_{comm}$$

$$\Leftrightarrow AI_{compute} < AI_{hardware}$$

This also means the actual throughput is

$$\frac{FLOPs}{T_{comm}} = \frac{FLOPs}{\left(\frac{\text{communication bytes}}{\text{bandwith } \frac{\text{bytes}}{\text{sec}}}\right)}$$

$$= \frac{FLOPs}{\text{communication bytes}} \cdot \left(\text{bandwith bytes/sec}\right)$$

$$= AI_{compute} \cdot B$$

This means, as $AI_{compute}$ increase

the realized $\frac{FLOPs}{sec}$ increases linearly

Combining results from:

① $T_{math} < T_{comm}$

② $T_{math} > T_{comm}$

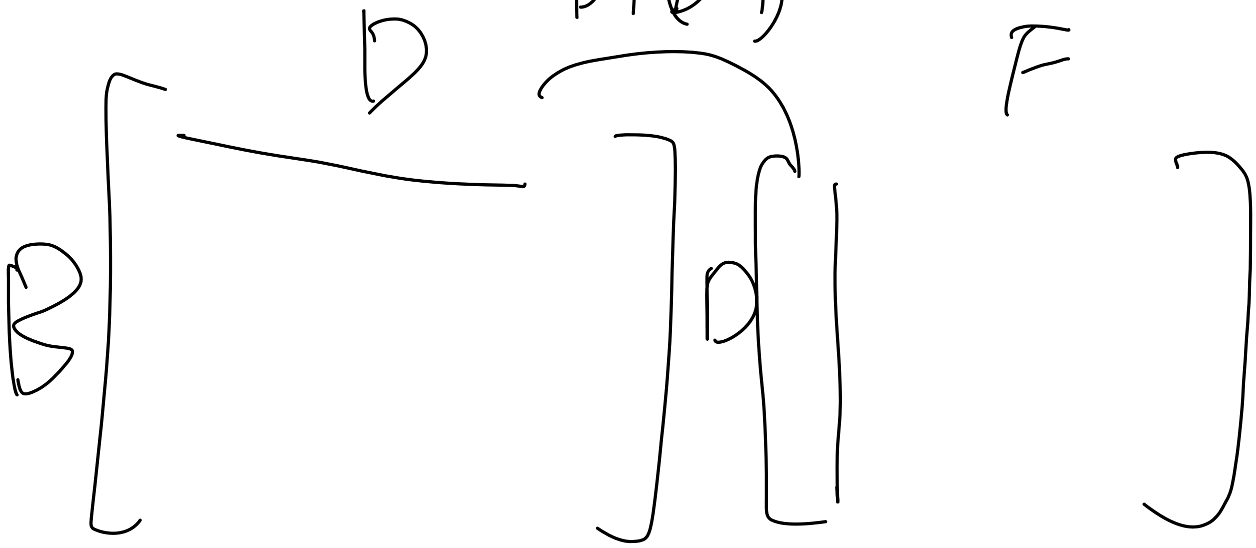
observations, we get the roofline model

Matrix multiplication

$X : (B, D)$ bf16 $\rightarrow$ 16 bit $\rightarrow$ 2 byte

$Y : (D, F)$ bf16 $\longrightarrow (B, F)$

$$\overset{*}{D} + \overset{+}{(D-1)}$$



Load $= 2\text{byte} \cdot B \cdot D + 2 \cdot D \cdot F$

#FLOPS $= (2D-1) \cdot B \cdot F$

$\simeq 2BDF$

Write $= \underset{\text{bytes}}{2} BF$

$$\text{Intensity} = \frac{2BDF}{2BD + 2DF + 2BF}$$

$$= \frac{BDF}{BD + DF + BF}$$

$$\simeq \frac{BDF}{DF}$$

(assume D, F are dominating in Transformer cases which is hidden dimensions)

$$= \boxed{B}$$

Network communication rooflines

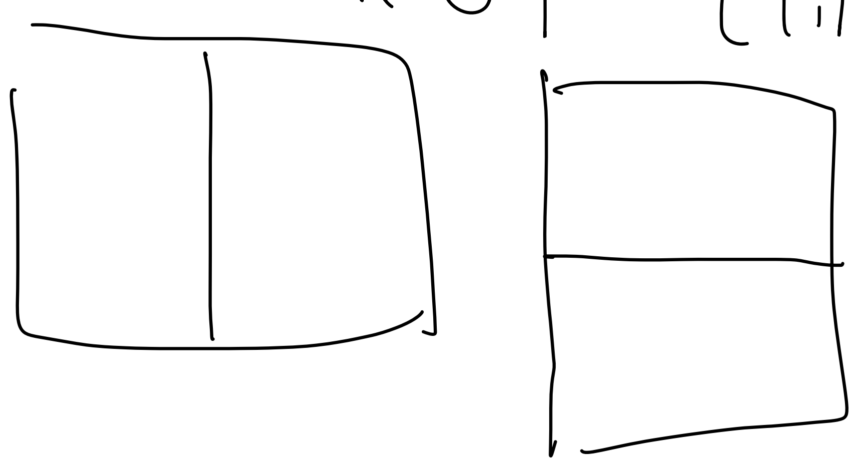
X : bfloat16[B, D] matmul → bfloat[B, F]
Y : bfloat16[D, F]
across 2 TPUs

Consider: $X[:, :D//2]$ @ $Y[:D//2, :]$

$+ \; X[:, D//2:]$ @ $Y[D//2:, :]$

$= X$ @ Y (tiling)

X



Assume: Network bandwith = 4.5×10^{10} bytes/sec
FLOPs/sec each chip = 1.97×10^{14} FLOPs/sec

$$T_{math} = \frac{2BDF}{\underset{\substack{\uparrow \\ \text{2 chips}}}{2} \cdot 1.97 \times 10^{14}} = \frac{BDF}{1.97 \times 10^{14}}$$

Assume the best case when we fully utlize a

Single TPU

$T_{comms} = \frac{BF + BF}{4.5 \times 10^{10}}$ (left TPU send BF to right and right send BF to left)

When

$T_{math} > T_{comms}$ happens

$\Leftrightarrow \frac{D}{2} = \frac{\cancel{B}D\cancel{A}}{2\cancel{B}\cancel{A}} > \underbrace{\frac{1.97 \times 10^{14}}{4.5 \times 10^{10}}}_{\text{Arithmetic intensity}}$

$\Leftrightarrow D > 8755$

$\Leftrightarrow$ When $D > 8755$, we know that we know 2 TPU is compute-bound (means good), assuming each TPU is fully utlizing its FLOPs/sec (the best case)

$\Rightarrow$ so this means the previous B needs to be > 240 (hardware arithmetic intensity)

Question 1:

$X[B, D]$ @ $Y[D, F] \to Z[B, F]$ in int8 (1 byte) instead of bfloat16

1. Loaded $= BD + DF$

   Write $= BF$

2. $\text{FLOPs} = B \cdot F \cdot (\overset{\text{mul}}{D} + \overset{\text{+ op}}{(D-1)})$

   $= BF(2D - 1)$

3. $AI = \dfrac{BF(2D-1)}{BD+DF+BF} \simeq \dfrac{2BDF}{BD+DF+BF}$

   assume DF dominates

   $= 2B$

4. $T_{math} = \dfrac{2BDF}{3.94 \times 10^{14}}$

   $T_{comm} = \dfrac{BD+DF+BF}{8.1 \times 10^{11}}$

   if $2B > \dfrac{3.94 \times 10^{14}}{8.1 \times 10^{11}}$

   $\Leftrightarrow B > 2.43 \times 10^{2}$

   $\Leftrightarrow B > 243$

Lower Bound = $\max(T_{math}, T_{comm})$

Upper Bound = $T_{math} + T_{comm}$

(assume we cannot overlap them)

Question 2: int8 + bf16

$$\text{bfloat16}[B, D] \times \text{int8}[D, F] \rightarrow \text{bfloat16}[B, F]$$

Load = $2BD + 1 \cdot DF$

Write = $2BF$

$$\text{Compute FLOPs} = BF \cdot (\underbrace{D}_{\text{mul}} + \underbrace{D-1}_{+})$$

$$\cong 2BDF$$

$$AI = \frac{2BDF}{2BD + DF + 2BF} \cong \frac{2BDF}{\underbrace{DF}_{\text{Assume } DF \gg BF}} = 2B$$

$DF \gg BD$

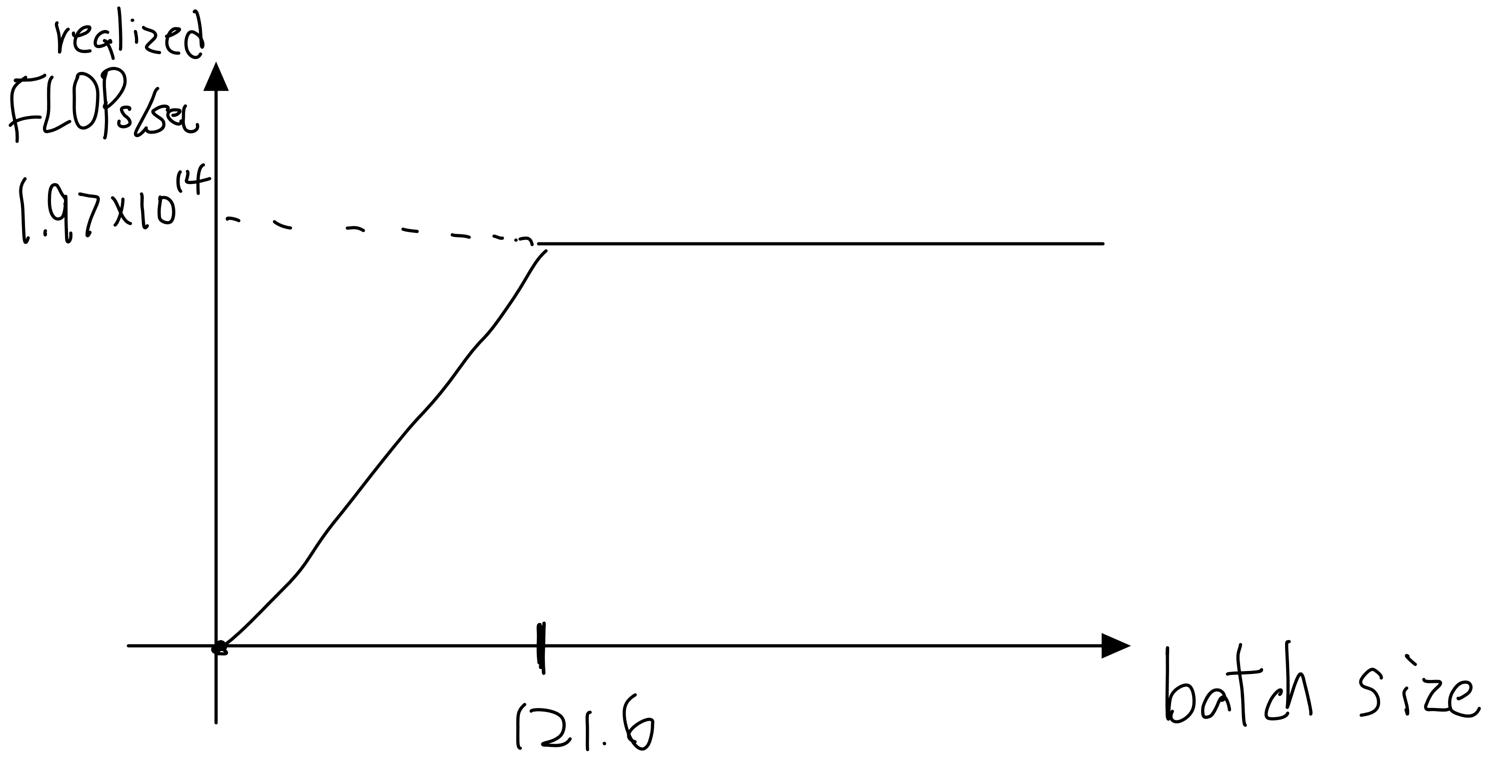
$$\text{AI hardware} = \frac{1.97 \times 10^{14}}{8.1 \times 10^{11}} = 243.20$$

$$\Rightarrow \quad 2B > 243.20$$

$$\Rightarrow \quad B > 121.6$$

Observation: Easier to get to compute-bound when [D, F] matrix is in int8

Question 3:



(Assume $D, F \gg B$ when $AI \approx \frac{2BDF}{DF} = 2B$

which means D, F do not matter)

If $D, F \gg B$ is not true, we need to consider all terms in $AI = \frac{2BDF}{2BD + DF + 2BF}$

Pick some random numbers like $D = F = 128$

$$AI = \frac{2B \cdot 128 \times \cancel{128}}{2B \times \cancel{128} + \cancel{128} \times 128 + 2B \times \cancel{128}}$$

$$= \frac{256B}{2B + 128 + 2B} = \frac{256B}{4B + 128} = \frac{64B}{B + 32}$$

This means: $\lim_{B \to \infty} \frac{64B}{B+32} = 64 < \underbrace{243}_{AI_{hardware}}$

This means when $D = F = 128$, it will always be memory-bound no matter how big is B

But when $D = F = 8000$, AI becomes

$$AI = \frac{2B \times 8000 \times 8000}{2B \times 8000 + 8000^2 + 2B \times 8000}$$

$$= \frac{2B \times 8000}{2B + 8000 + 2B}$$

$$= \frac{B \times 8000}{B + 4000 + B}$$

$$= \frac{8000B}{2B + 4000} = \frac{4000B}{B + 2000}$$

When $AI > 243$

$$\Leftrightarrow \frac{4000B}{B + 2000} > 243$$

$$\iff 4000B > 243B + 243 \times 2000$$

$$\iff 3757B > 243 \times 2000$$

$$\iff B > 129.35$$

This means when $D = F = 8000$,

$B > 129.35 \iff$ it is compute-bound

If $D = F = 16{,}000$

$$\frac{B \times 16000}{B + 8000 + B} = \frac{8000B}{B + 4000} > 243$$

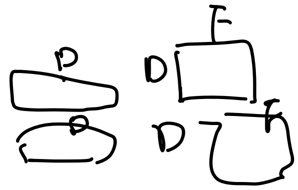
$$\Rightarrow 8000B > 243B + 243 \times 4000$$

$$\Rightarrow B > 125.30$$

So this matches that if $D, F \gg B$, should be converging to $B > 121$

Question 4:

$$\text{int8}[B, D] *_D \text{int8}[B, D, F] \rightarrow \text{int8}[B, F]$$



Answer:

There are B $[1, D] \times [D, F]$ matmul

so $FLOPs = B \cdot (D + (D-1)) \cdot F$

$$\cong 2BDF$$

$$\text{Load} + \text{Write} = BD + BDF + BF$$

$$\text{Arithmetic Intensity} = \frac{2BDF}{BD + BDF + BF}$$

$$\overset{D, F \gg B}{\cong} 2$$

This means 2 will always < 243

so always memory-bound.

Question 5:

$$\frac{1.979 \times 10^{15} \text{ FLOPs}}{2} \cong 1 \times 10^{15} \text{ FLOPs/sec}$$

memory bandwith $= 3.35 \times 10^{12}$ bytes/sec

$$AI = \frac{1 \times 10^{15}}{3.35 \times 10^{12}} = \frac{1000}{3.35} = 298$$

$$AI_{\text{bfloat matmul}} = \frac{2BDF}{2DF} = B$$

So $B > 298$ is when it becomes compute-bound