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

depend on how you write code

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

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

$= \text{Accelerator } \frac{\text{FLOPs}}{\text{sec}}$

(constant)

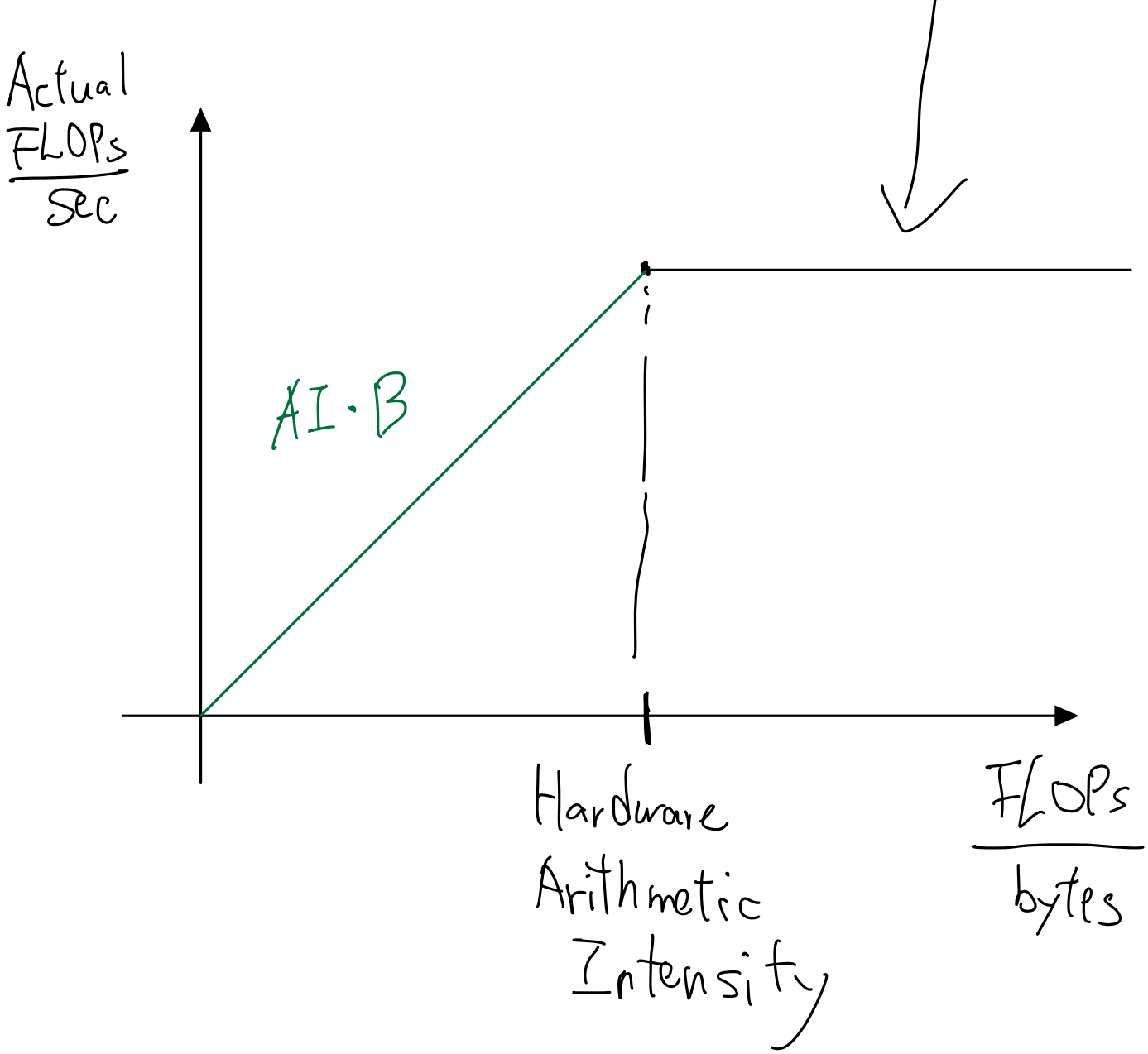
$T_{math} > T_{comm}$ also means

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

$$\Leftrightarrow \text{Arithmetic Intersity}_{\text{(compute)}} > \text{AI}_{\text{(hardware)}}$$

Summary: when $\text{AI}_{\text{compute}} > \text{AI}_{\text{hardware}}$

actual $\frac{\text{FLOPs}}{\text{sec}}$ realized is always $\frac{\text{Accelerator FLOS}}{\text{sec}}$


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

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

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

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

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