latency vs bandwidth vs throughput

latency vs bandwidth vs throughput

- **latency:**
  length of the pipe
- **bandwidth:**
  max amount of water flow in unit time
- **throughput:**
  actual amount of water flow in unit time

latency and bandwidth are independent parameters

Assume that you are delivering pizzas:

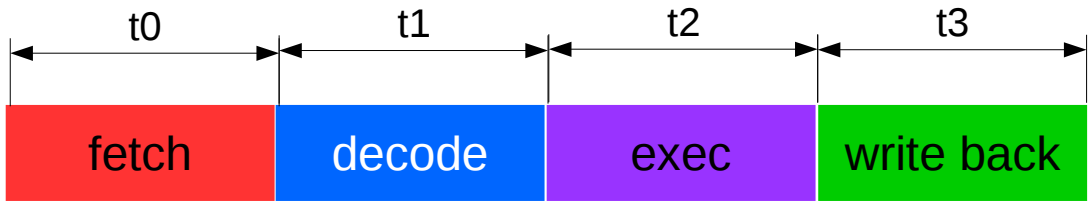
1) You can deliver pizzas as soon as an order comes
   - low latency (and low throughput)
2) You can wait till there are enough order to fill your bag
   - high throughput (and high latency)

*this is a simplified example and has many implicit assumptions

latency vs bandwidth vs throughput

- **Latency:** the amount of time between when the instruction is issued and when it completes
- **Throughput:** the number of instructions completed in unit amount of time

4 stage pipeline

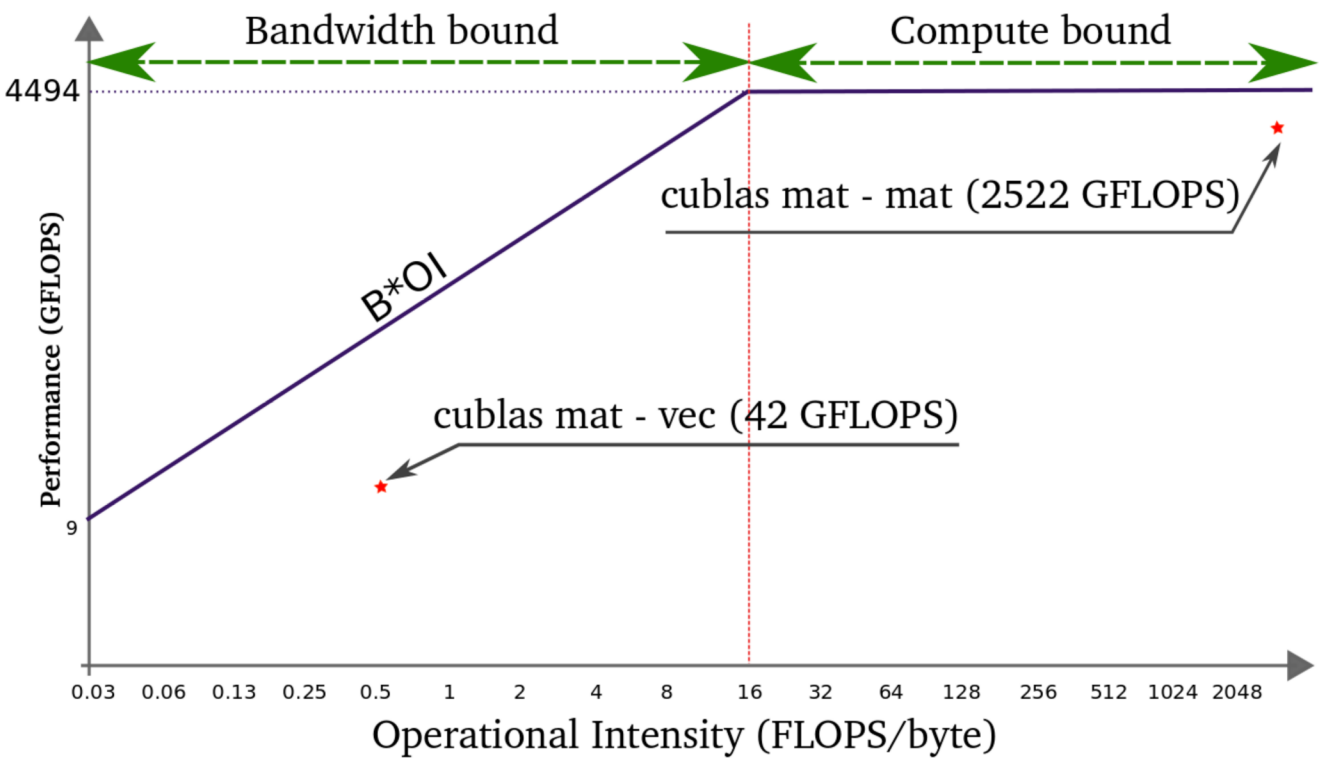


- latency
  - t0 + t1 + t2 + t3
- max throughput
  - max (t0, t1, t2, t3)

Roofline model

- **Operation Intensity**: The number of ops per data element(or bytes) moved (OI = Total ops / Total data movement)

- **Roofline model**: A performance model based on two platform specific ceilings: i) peak compute capacity ii) Max Memory bandwidth



Nvidia Tesla K80 (float)

(on one GK210)