

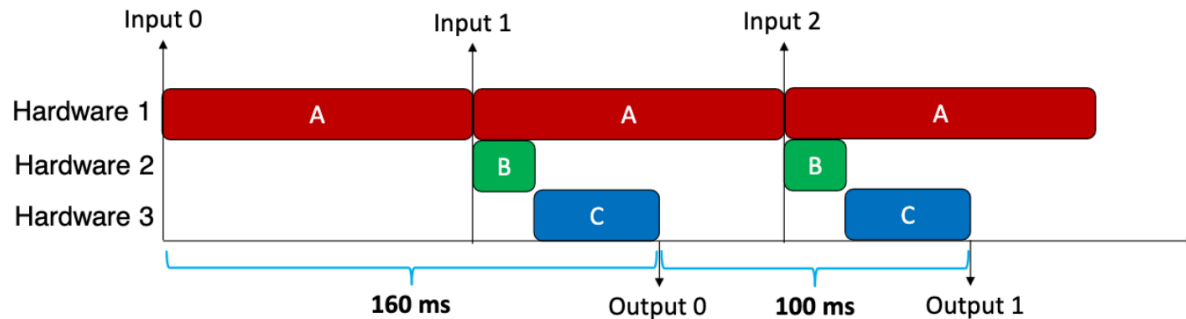
Serial Operations

- the operations all took very different amounts of time
- **latency** Time between the input arriving and the output being produced
- **throughput** Total number of inputs that can be processed per unit time.

Pipelining

- pipelining latency doesn't change
- throughput increases

Pipelining – Latency & throughput



- Pipeline latency: 160 ms
 - We are doing the same operations, so this does not change.
- Pipeline throughput: $1 \div 100 \text{ ms} = 10 \text{ inputs per second}$
 - We are producing each result faster than the serial case.

Limitations

- requires OP to use different hardware units
- always has overhead (setup individual hardware units, may override benefit from pipelining)
- requires extra memory to support double buffering
 - avoid reading and writing into the same array at the same time
 - can also use circular buffer

Example