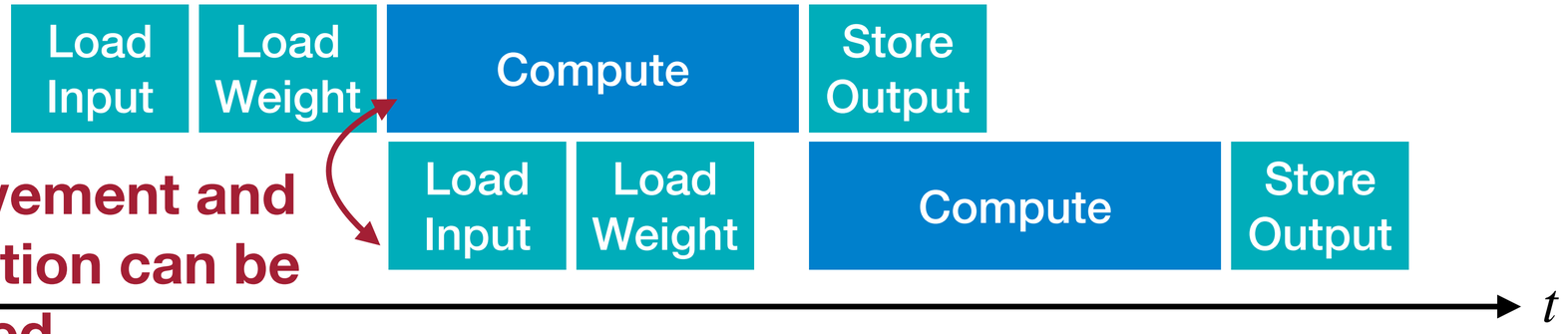


Data Movement and Computation can be overlapped.

$$Latency \approx \max\left(T_{computation}, T_{memory}\right)$$

$$T_{\text{computation}} \approx \frac{\text{Number of Operations in Neural Network Model}}{\text{Number of Operations that Processor can Process Per Second}}$$

Number of Operations in Neural Network Model — **NN Specification**

Number of Operations that Processor can Process Per Second — **Hardware Specification**

$$T_{\text{memory}} \approx T_{\text{data movement of activations}} + T_{\text{data movement of weights}}$$

$$T_{\text{data movement of weights}} \approx \frac{\text{Neural Network Model Size}}{\text{Memory Bandwidth of Processor}}$$

Neural Network Model Size — **NN Specification**

Memory Bandwidth of Processor — **Hardware Specification**

$$T_{\text{data movement of activations}} \approx \frac{\text{Input Activation Size + Output Activation Size}}{\text{Memory Bandwidth of Processor}}$$

Input Activation Size + Output Activation Size — **NN Specification**

Memory Bandwidth of Processor — **Hardware Specification**