

a)

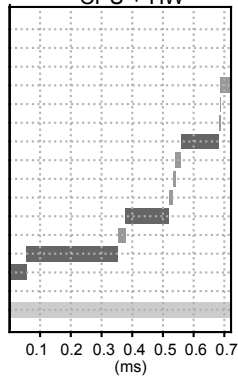
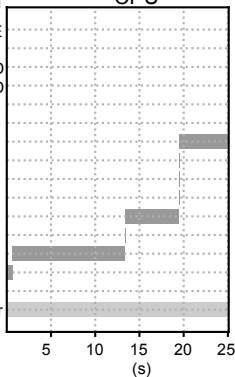
Inference schedule

Tensor operation

CPU

CPU + HW

DEQUANTIZE
 SOFTMAX
 FULLY_CONNECTED
 FULLY_CONNECTED
 RESHAPE
 MAX_POOL_2D
CONV_2D
 ADD
 MUL
 MAX_POOL_2D
CONV_2D
 MAX_POOL_2D
CONV_2D
CONV_2D
 QUANTIZE
 Interpreter



b)

