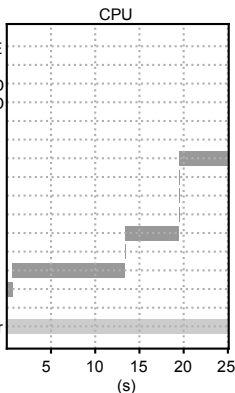


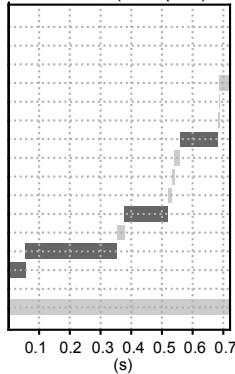
a) Model A (8-bit fixed-point quantization)

Tensor operation

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

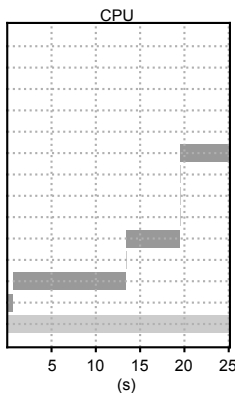


CPU + HW (Fixed-point)

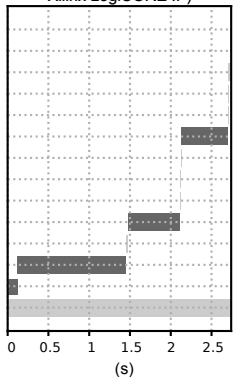


b) Model A (Floating-point)

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
Interpreter

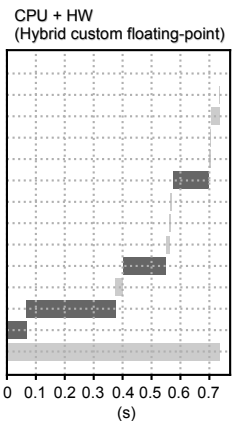


CPU + HW
(Floating-point
Xilinx LogiCORE IP)

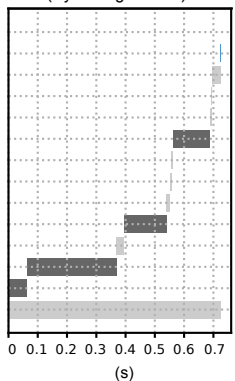


b) Model A (Floating-point)

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
Interpreter

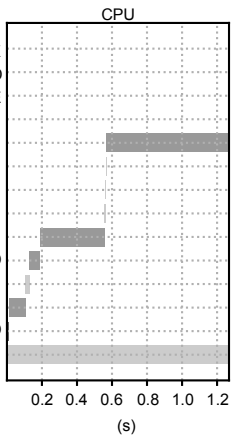


CPU + HW
(Hybrid logarithmic)



b) Model B (Floating-point)

SOFTMAX
FULLY_CONNECTED
RESHAPE
MAX_POOL_2D
CONV_2D
ADD
MUL
MAX_POOL_2D
CONV_2D
DEPTHWISE_CONV_2D
MAX_POOL_2D
CONV_2D
DEPTHWISE_CONV_2D
Interpreter



CPU + HW
(Hybrid custom floating-point)

