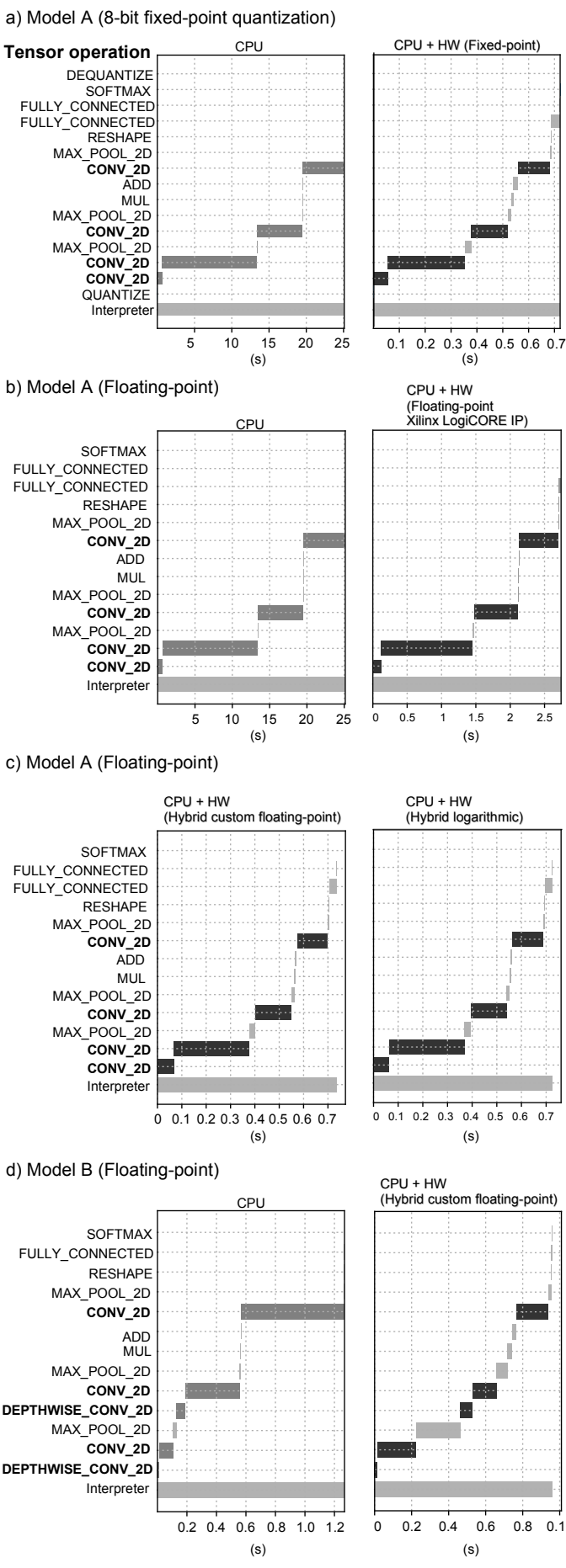
a) Model A (8-bit fixed-point quantization)
Tensor operation
CPU
CPU + HW (Fixed-point)
DEQUANTIZE
SOFTMAX
FULLY_CONNECTED
FULLY_CONNECTED
RESHAPE
MAX_POOL_2D
CONV_2D
ADD
MUL
MAX_POOL_2D
CONV_2D
MAX_POOL_2D
CONV_2D
CONV_2D
QUANTIZE
Interpreter
5 10 15 20 25
(s)
0.1 0.2 0.3 0.4 0.5 0.6 0.7
(s)
b) Model A (Floating-point)
CPU
CPU + HW
(Floating-point
Xilinx LogiCORE IP)
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
5 10 15 20 25
(s)
0 0.5 1 1.5 2 2.5
(s)
c) Model A (Floating-point)
CPU + HW
(Hybrid custom floating-point)
CPU + HW
(Hybrid logarithmic)
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
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
(s)
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
(s)
d) Model B (Floating-point)
CPU
CPU + HW
(Hybrid custom 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
0.2 0.4 0.6 0.8 1.0 1.2
(s)
0 0.2 0.4 0.6 0.8 0.1
(s)