a) Model A (8-bit fixed-point quantization) CPU + HW (Fixed-point) Tensor operation **DEQUANTIZE** SOFTMAX FULLY\_CONNECTED FULLY CONNECTED RESHAPE MAX\_POOL\_2D CONV\_2D MUL MAX\_POOL 2D CONV 2D MAX\_POOL\_2D CONV\_2D CONV\_2D QUANTIZE Interpreter 5 15 20 25 0.3 0.4 0.5 0.6 0.7 (s) (s) b) Model A (Floating-point) CPU + HW (Floating-point Xilinx LogiCORE IP) SOFTMAX **FULLY CONNECTED** FULLY\_CONNECTED RESHAPE MAX\_POOL\_2D CONV\_2D ADD MUI MAX\_POOL\_2D CONV\_2D MAX\_POOL\_2D CONV\_2D CONV\_2D Interpreter 1.5 2.5 5 10 15 20 25 0.5 (s) (s) b) Model A (Floating-point) CPU + HW CPU + HW (Hybrid logarithmic) (Hybrid custom floating-point) SOFTMAX CONNECTED **FULLY** FULLY\_CONNECTED RESHAPE MAX\_POOL\_2D CONV\_2D ADD MUL MAX\_POOL\_2D CONV\_2D MAX\_POOL\_ 2D CONV\_2D CONV\_2D Interpreter 0.5 0.6 0.7 0.1 0.2 0.3 0.4 0.5 0.6 0.7 (s) (s) b) Model B (Floating-point) CPU + HW (Hybrid custom floating-point) CPU 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 0.6 0.8 1.0 1.2 0.2 0.4 0.6 8.0 0.1 (s) (s)