a) Model A (8-bit fixed-point quantization) CPU + HW (Fixed-point) Tensor operation **DEQUANTIZE** SOFTMAX _CONNECTED FULLY 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 0.4 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 0.5 1.5 5 15 20 25 10 (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 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.2 0.3 0.4 0.5 0.6 (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 0.2 0.4 0.6 8.0 0.1 (s) (s)