# FPGA synthesis: resources and latency

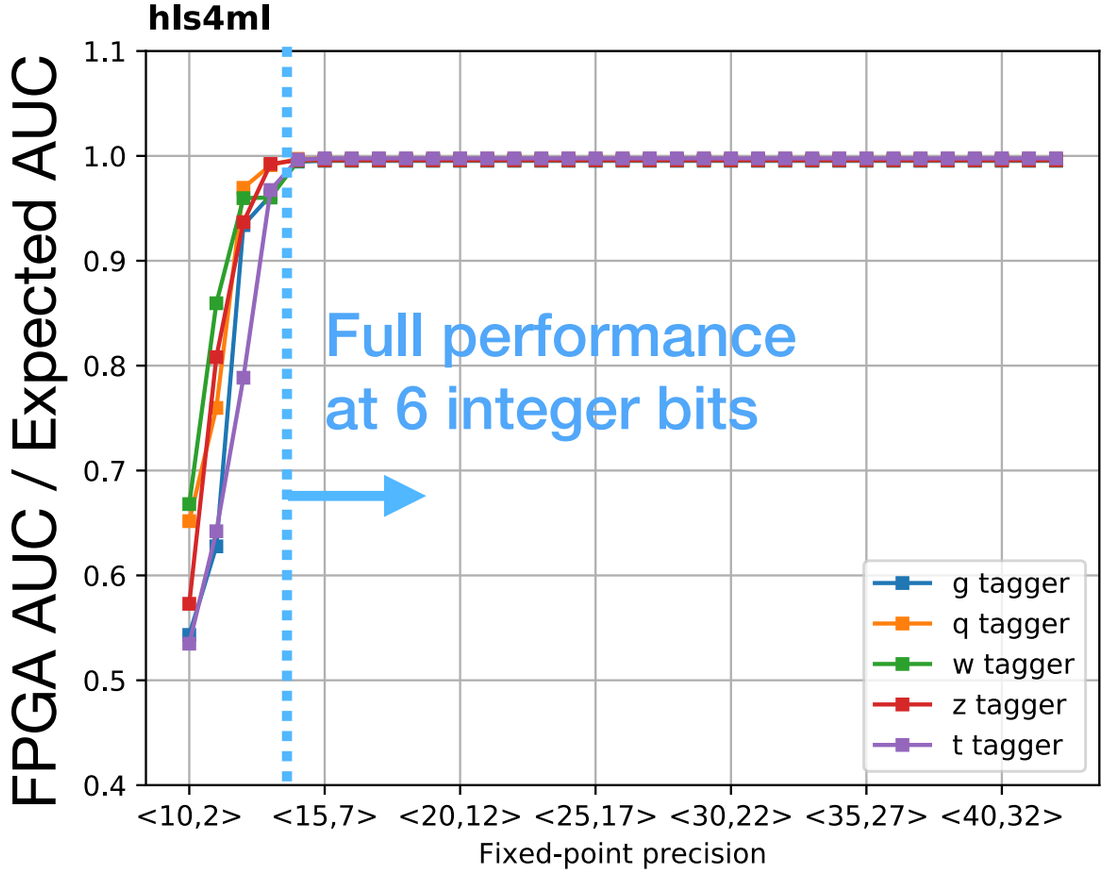
# Efficient NN design: quantization

ap_fixed<width,integer>


- Quantify the performance of the classifier with the AUC
- Expected AUC = AUC achieved by 32-bit floating point inference of the neural network

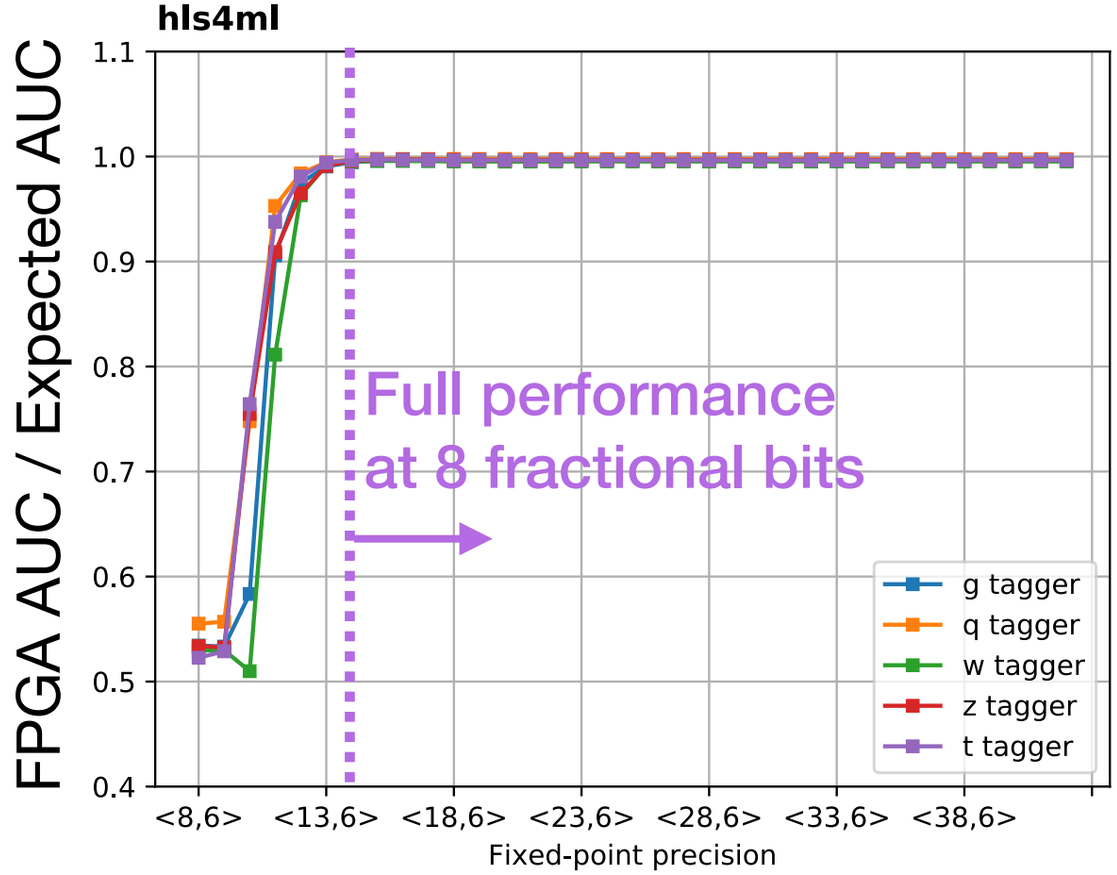
## Scan integer bits

Fractional bits fixed to 8



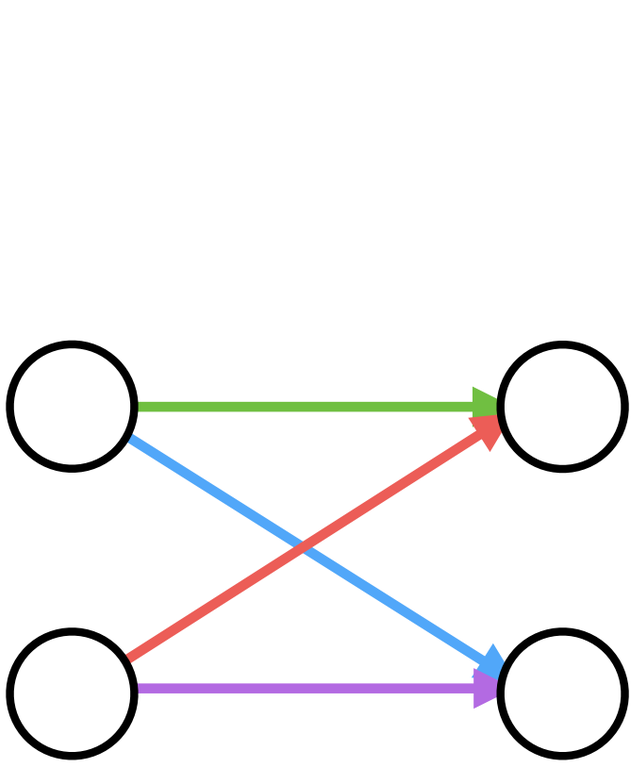

## Scan fractional bits

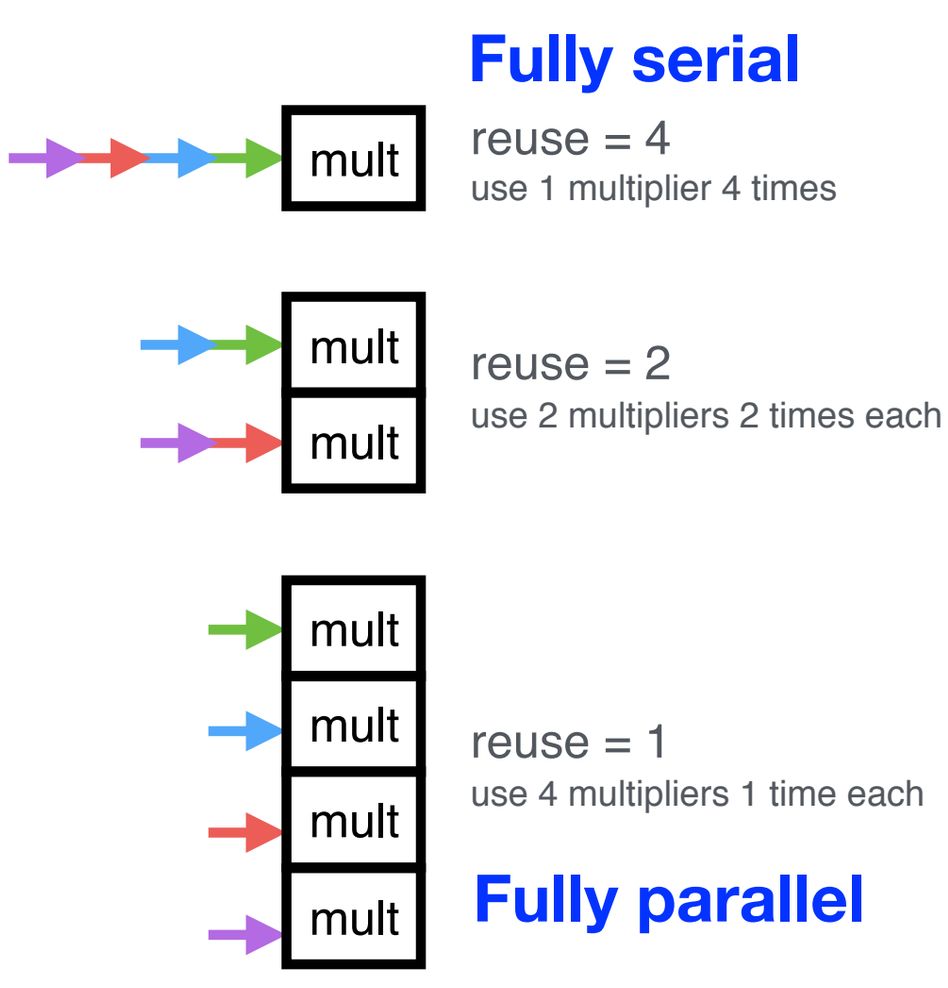
Integer bits fixed to 6



# Efficient NN design: parallelization

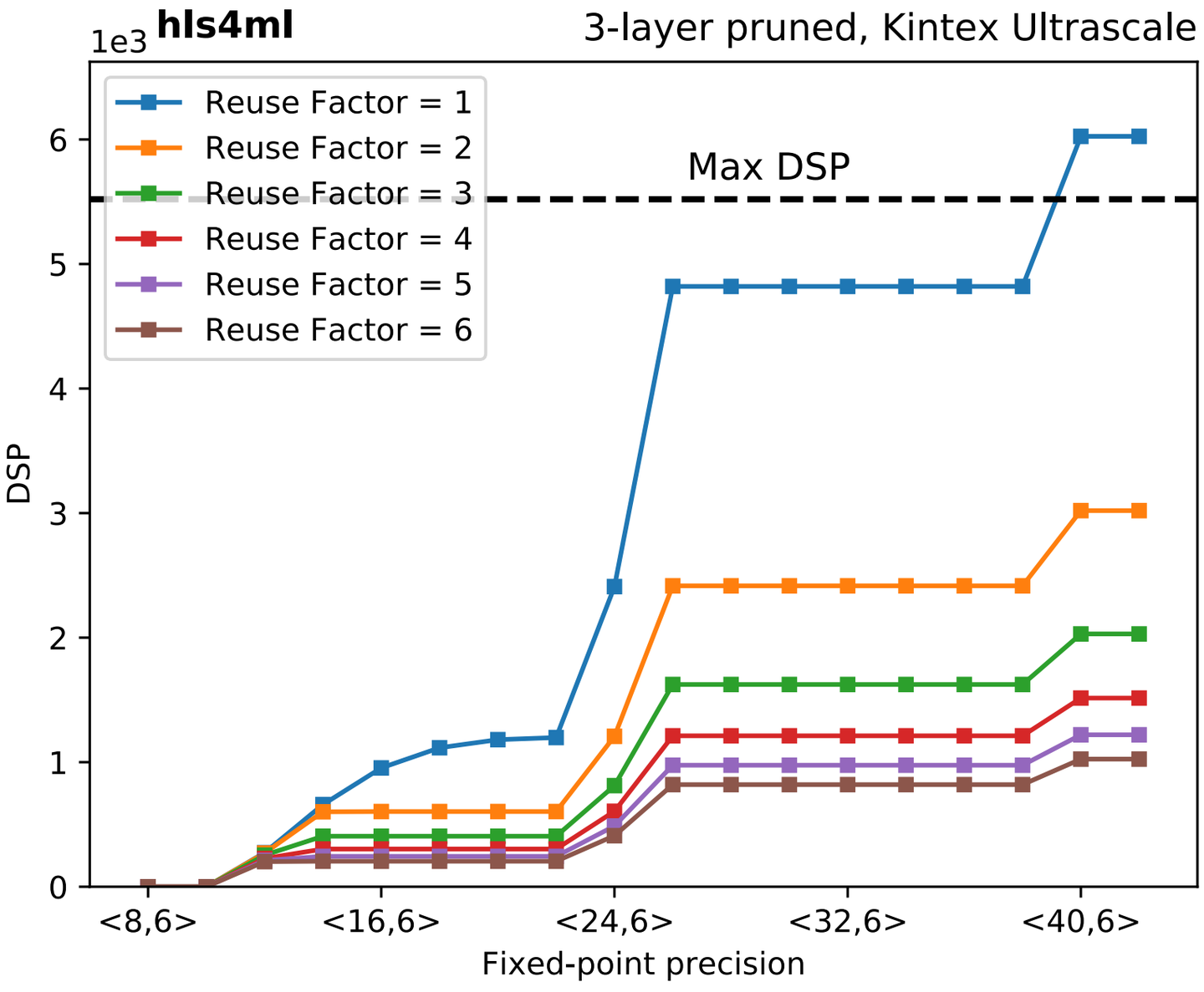
- Trade-off between latency and FPGA resource usage determined by the parallelization of the calculations in each layer
- Configure the **“reuse factor”** = number of times a multiplier is used to do a computation






**Reuse factor**: how much to parallelize operations in a hidden layer

# Parallelization: DSPs usage




Fully parallel
Each mult. used 1x

Each mult. used 2x

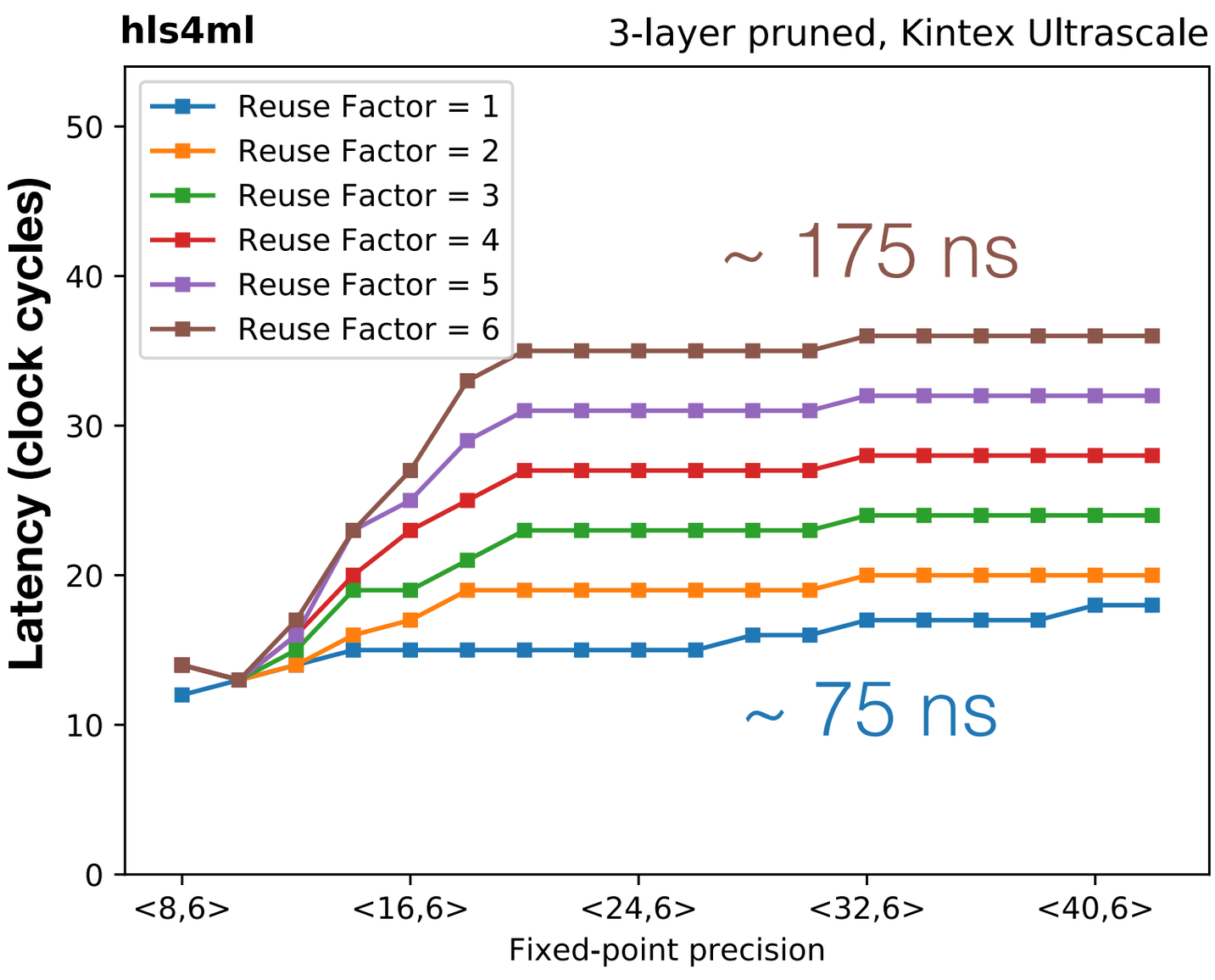
Each mult. used 3x

⋮

# Parallelization: Timing

**Latency of layer m**

$$L_m = L_{\text{mult}} + (R-1) \times II_{\text{mult}} + L_{\text{activ}}$$




Longer latency

Each mult. used 6x

Each mult. used 3x

Fully parallel
Each mult. used 1x

More resources