A New Storage-Less Hardware Compression Technique for CNNs

Mohamed Ayoub Neggaz[1], Smail Niar[1],

Ihsen Alouani[1] and Fadi. J. Kurdahi[2]

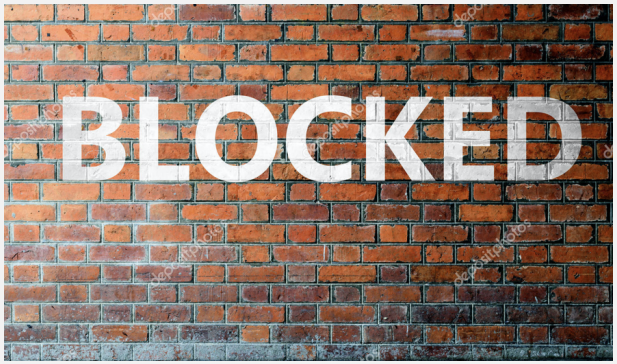
[1] Université Polytechnique Hauts de France

[2] University of California Irvine

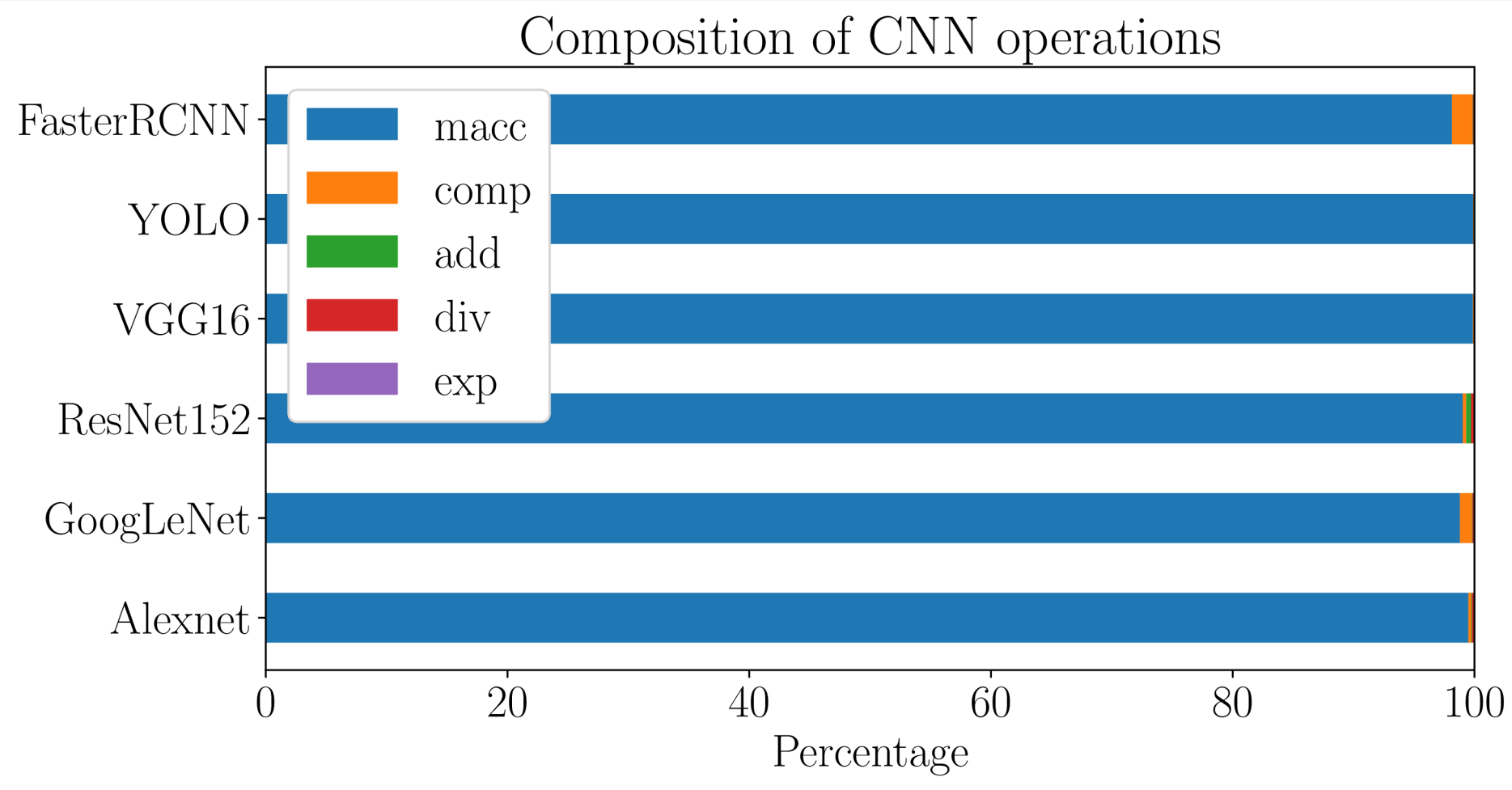
Motivations

- Machine Learning is about multiplication!

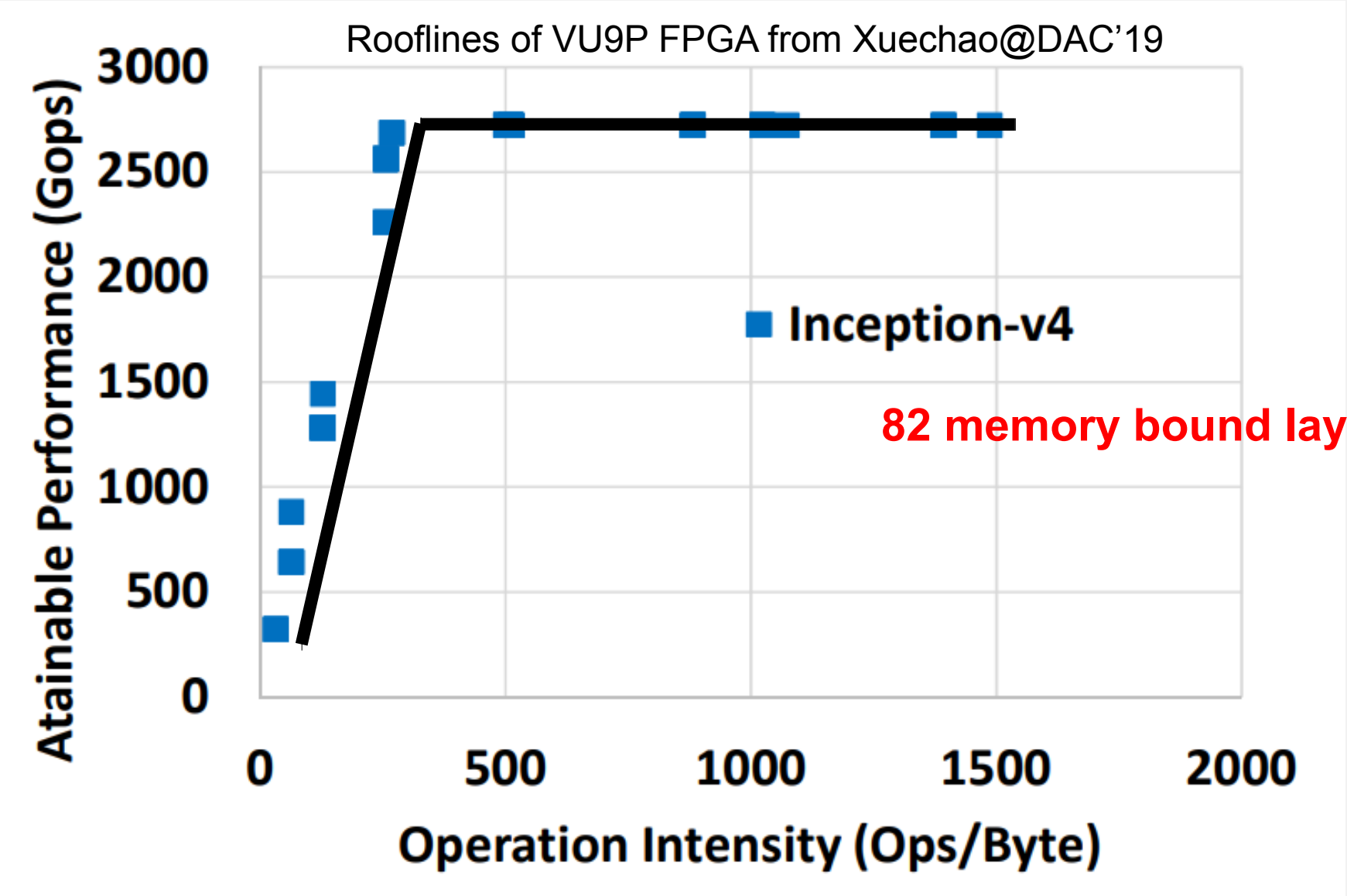
- Memory bandwidth is the limit
- Quantization is key



Operation Types

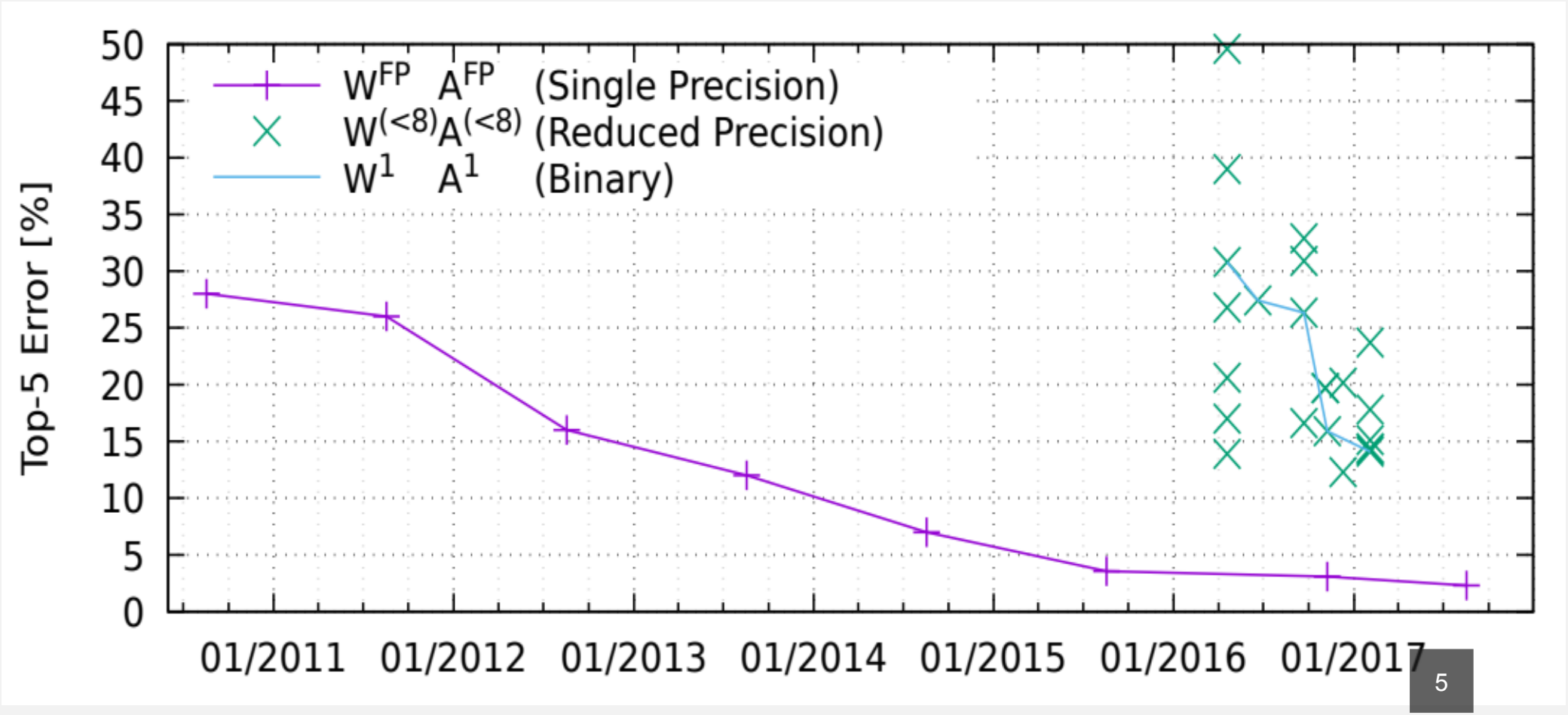


Memory Bandwidth

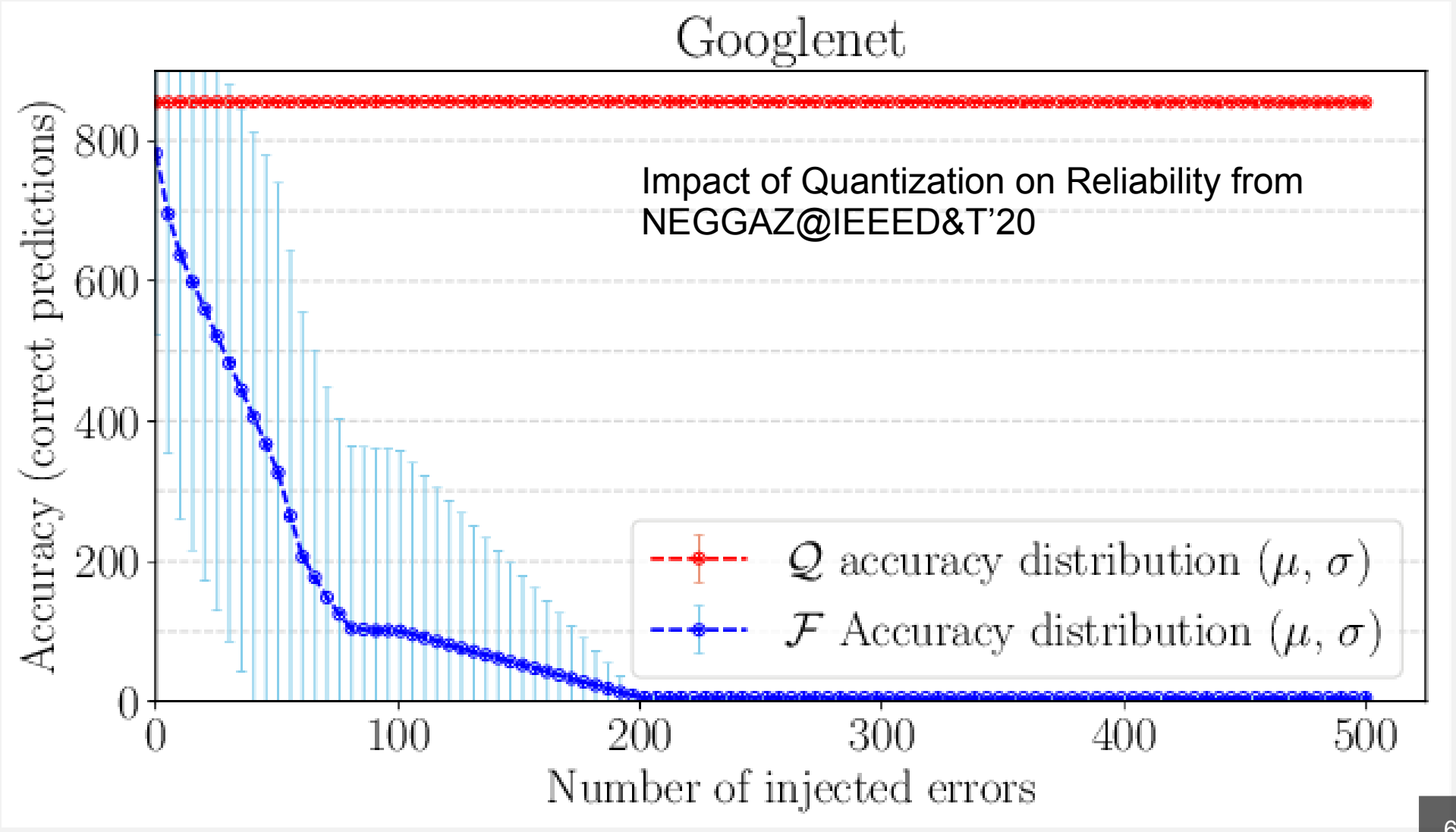


Quantization Results

Accuracy over time for ImageNet classification from BLOTT@TRETS'18



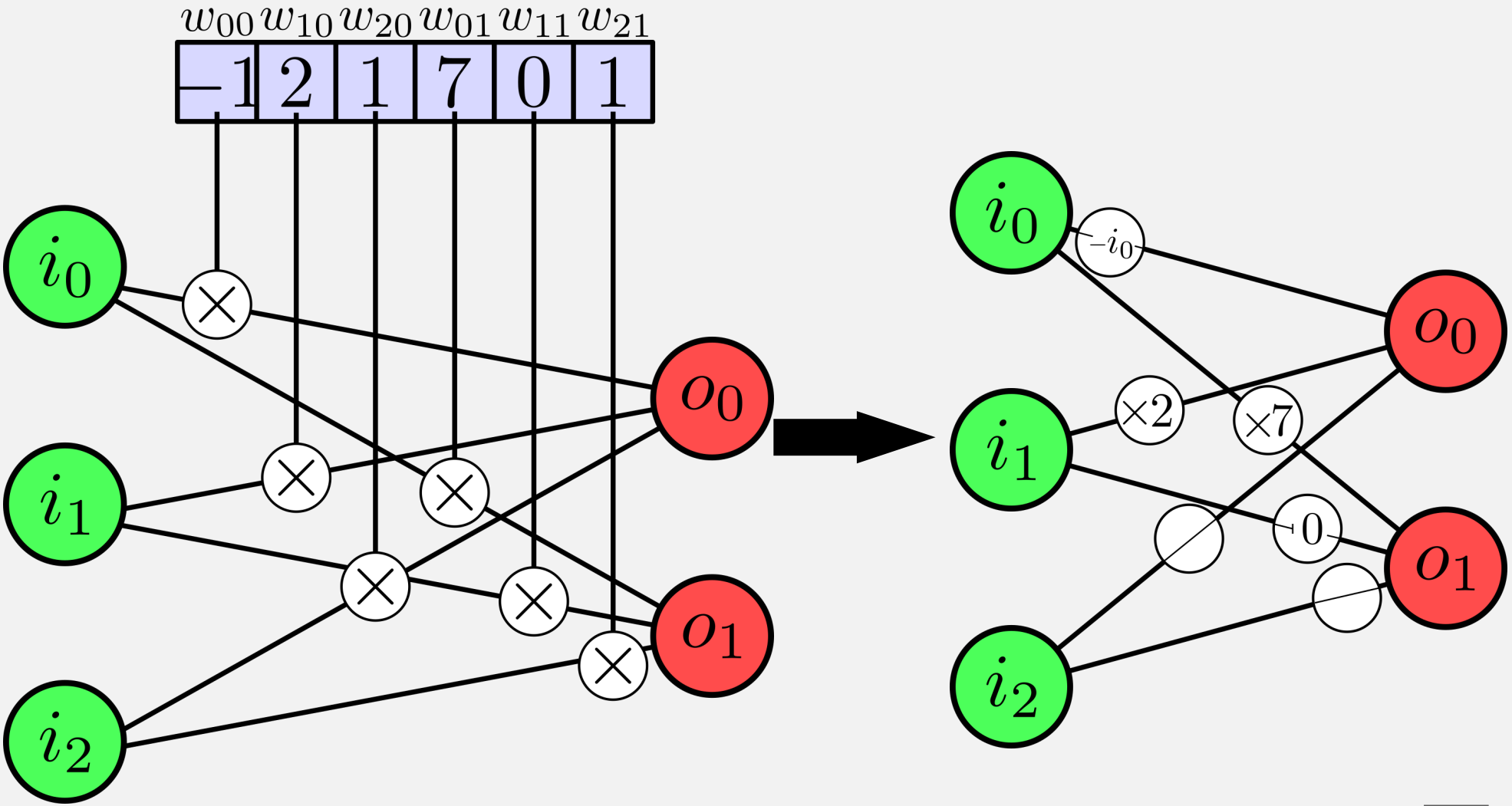
Quantization even for Reliability!



Idea

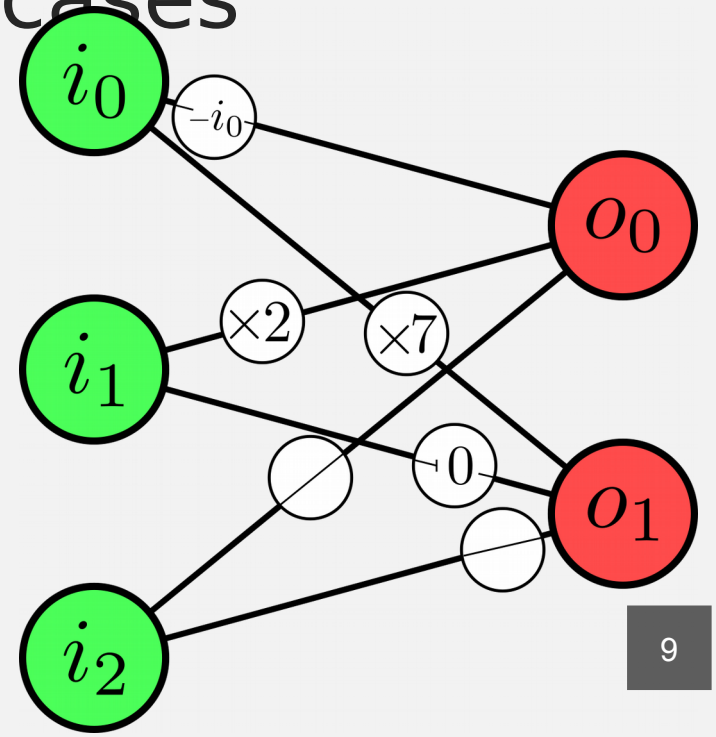
- Weights do not change after training => Constant Multipliers
- Memory is problematic => Compute units will store values
- Multiplications are dominant => Use lighter multipliers

FC example

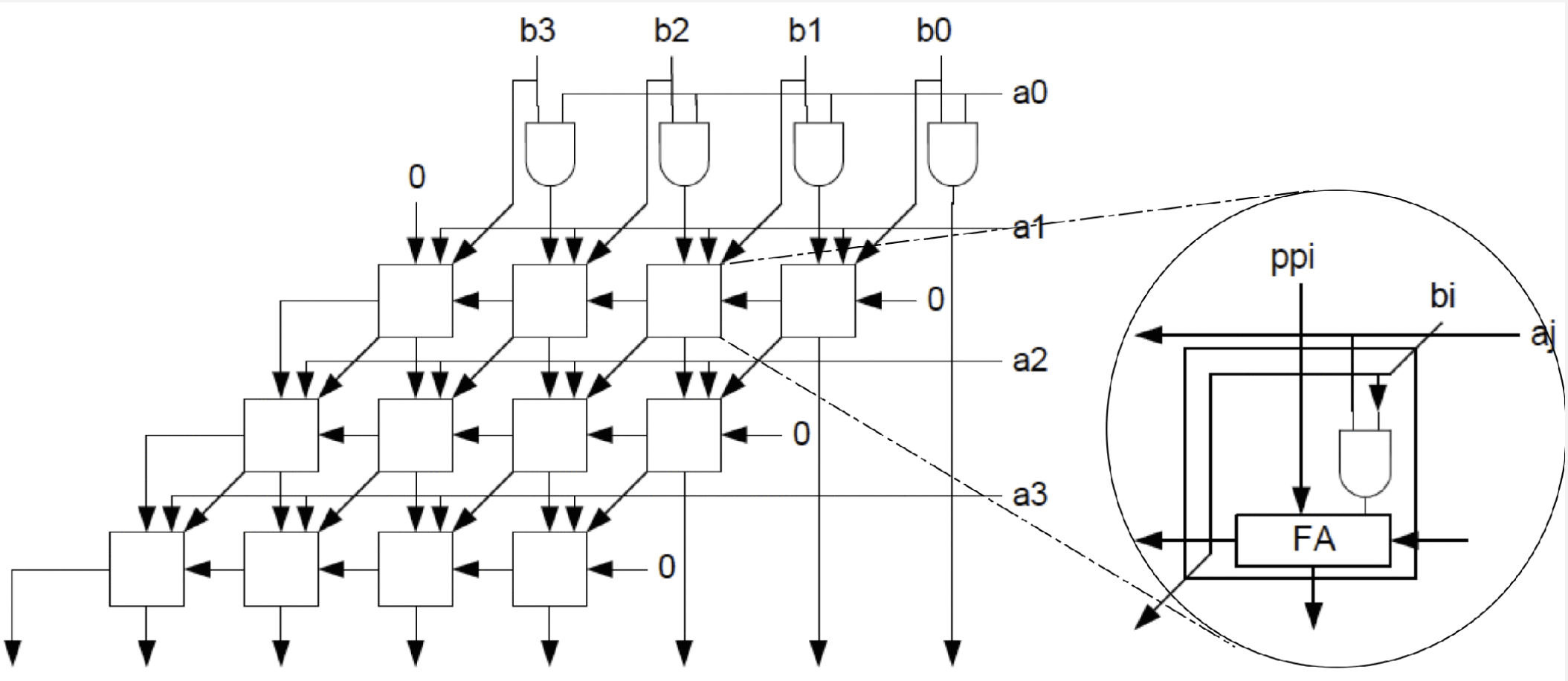


How does it works?

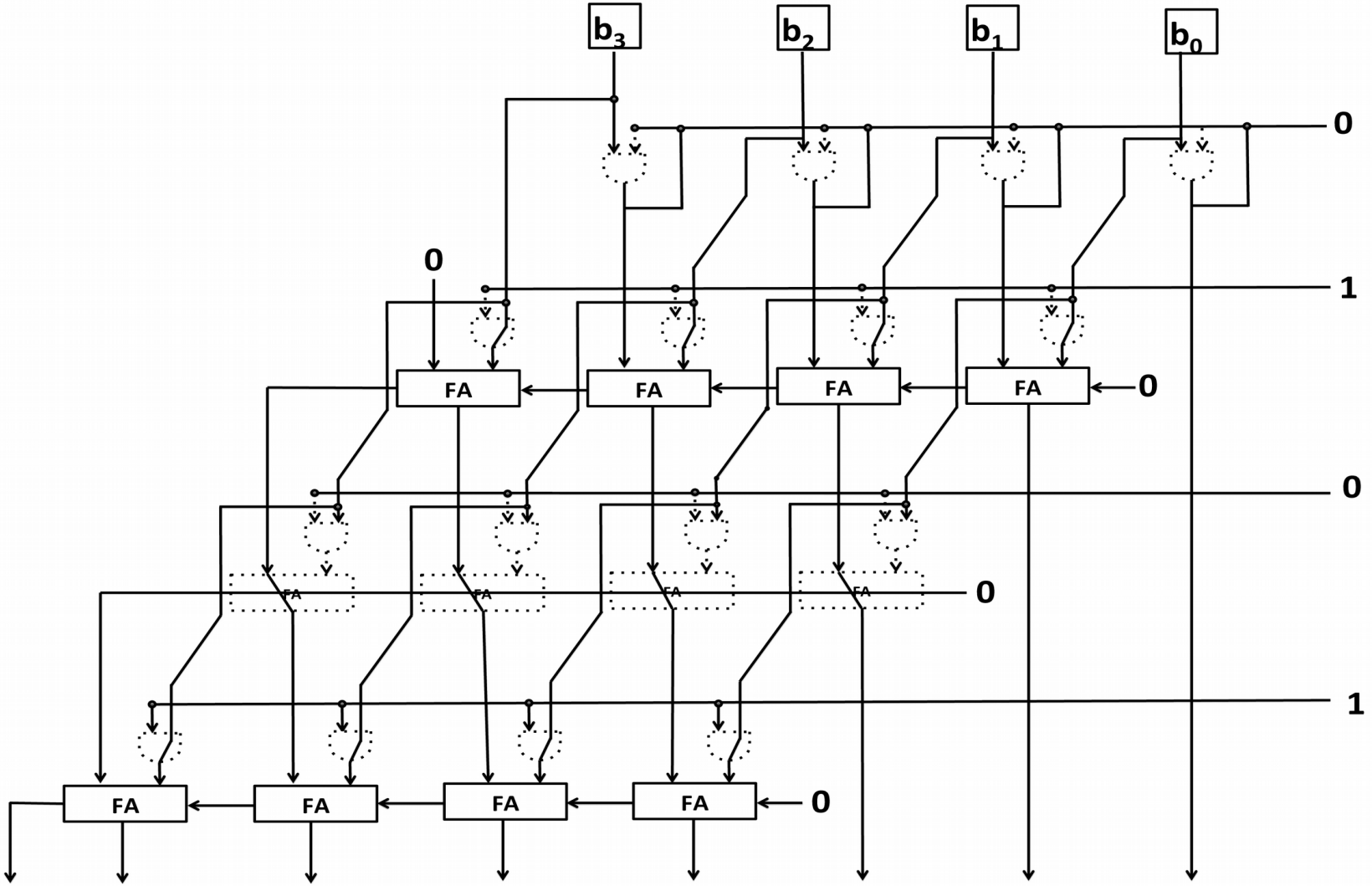
- Powers of 2 are replaced with shifts
- Zeros are eliminated (ground the circuit)
- The sign bit is applied separately
- Custom circuit for other cases



Array Multiplier



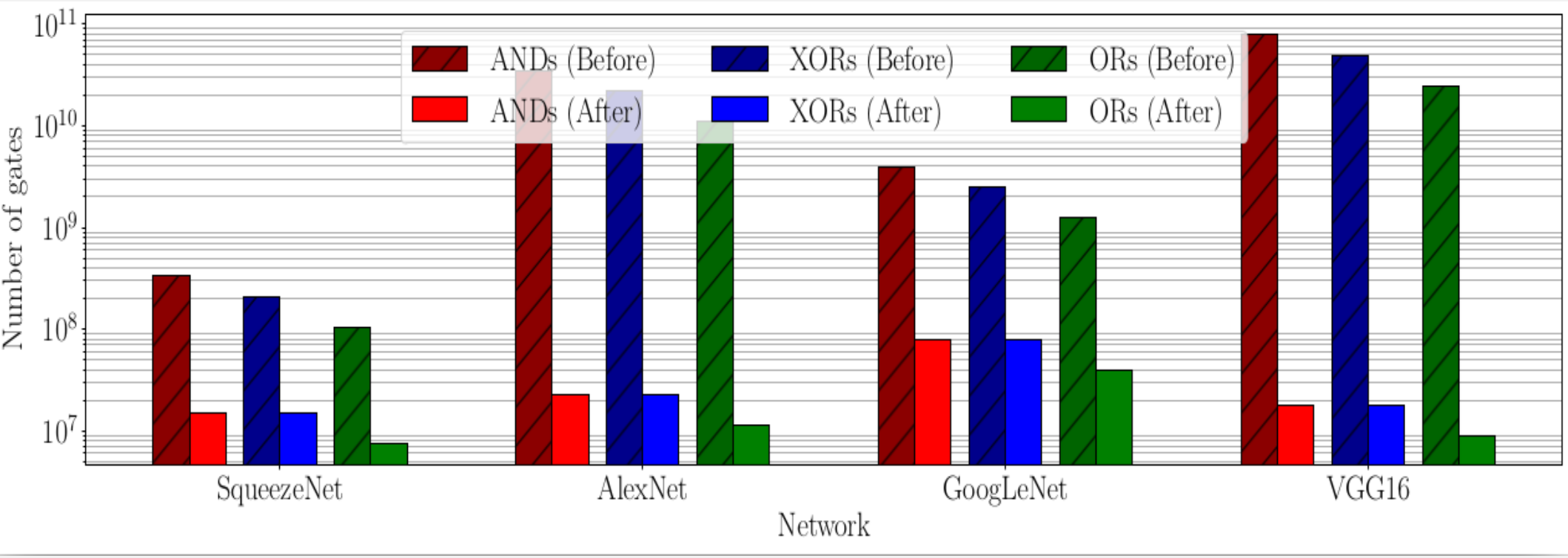
Multiplication Example



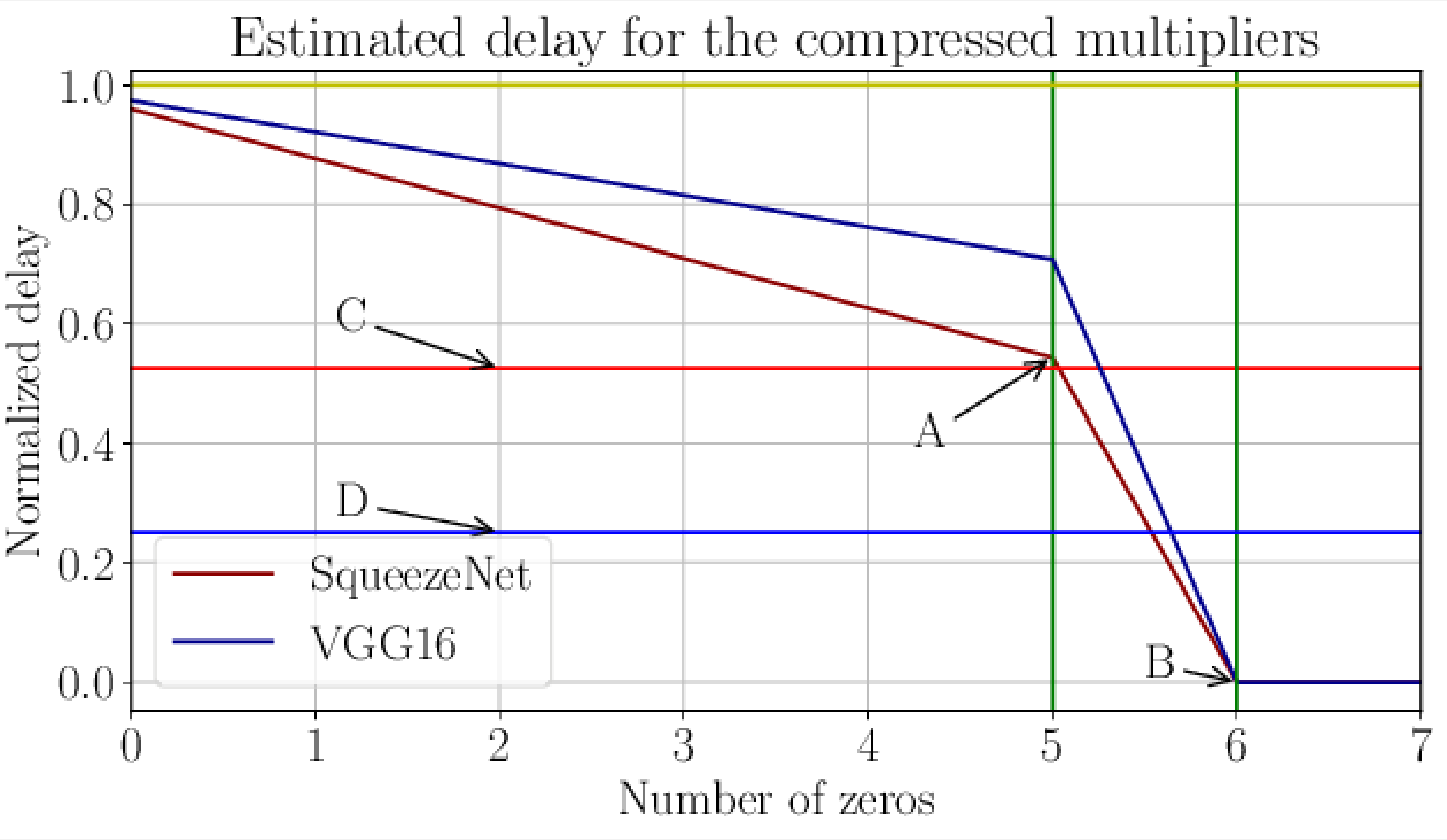
Network Results

- 8 bit fixed-point quantized networks: Sign+7 bits
- Only consider Inference multiplication circuit
- Fully unroll the network i.e. as much multipliers as weights.
- Shared weights in conv layers are used sequentially.

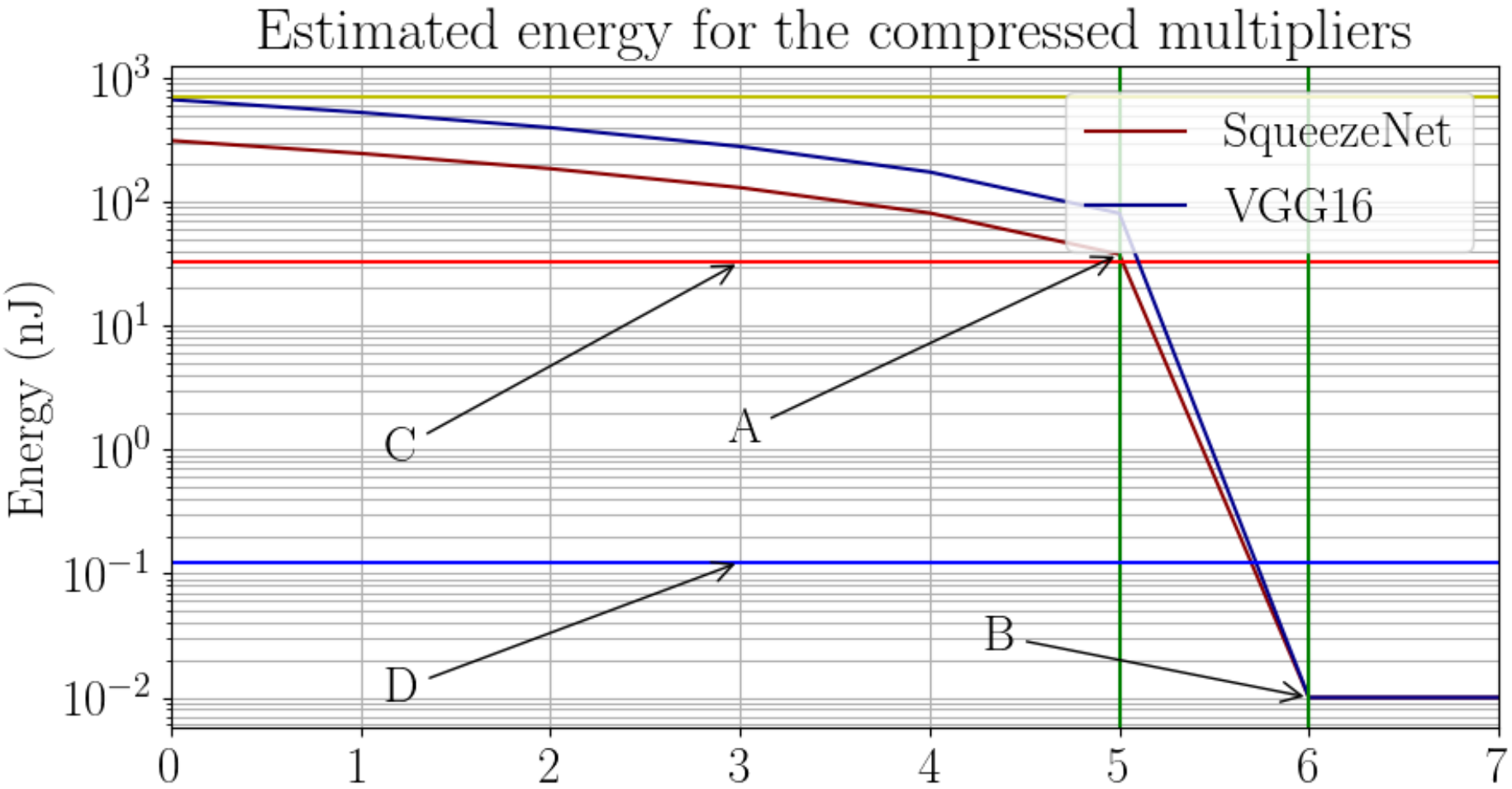
Resource Savings



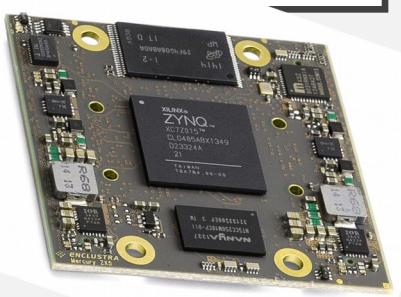
Delay


Estimated delay for the compressed multipliers
Normalized delay
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
Number of zeros
C
A
D
B
SqueezeNet
VGG16

Energy

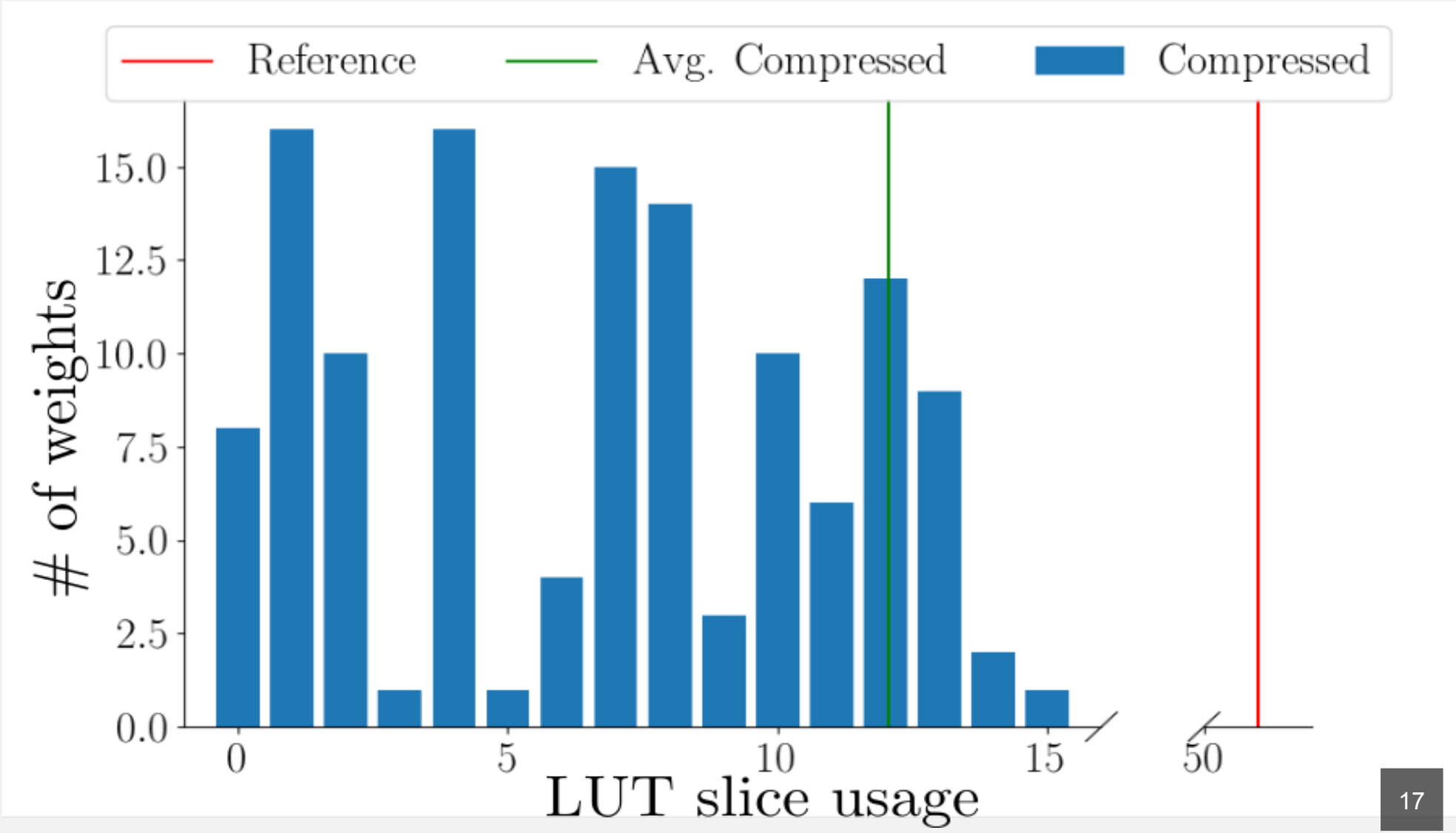

Estimated energy for the compressed multipliers
Energy (nJ)
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
0
1
2
3
4
5
6
7
SqueezeNet
VGG16
A
B
C
D
Number of Zeros

RTL-level Analysis

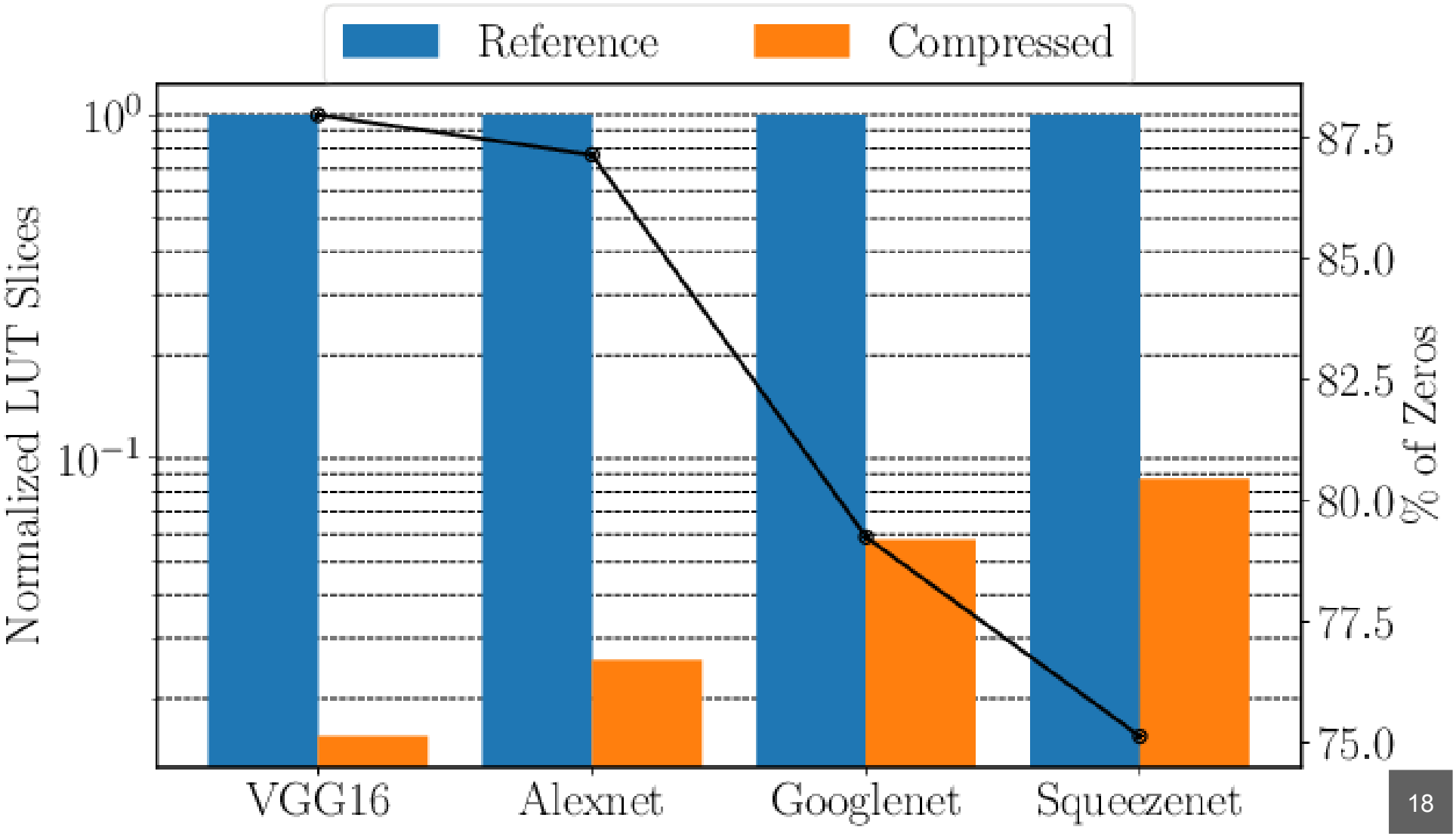


- Impact depends on hardware architecture
  - Number, type and location of slices
- Results from Xilinx’ ZCU102
- Same setup as before:
  - Inference only
  - 8 bit quantized networks: sign + 7 bits
  - Fully unroll layers
  - Conv layers are executed sequentially

Individual comparisons



Network level



Conclusions

- Exploit an under-explored area
- Huge energy and resource savings
- Trading flexibility for efficiency

Perspectives

- “In theory there is no difference between theory and practice. But practice shows that there actually is.”
- Flexibility for continuously trained networks (RL)
- CSM-HLS tool!
  - Network description and weights as input
  - Hardware as output