Rate-distortion Optimized Coding for Efficient CNN Compression

Wang Zhe, Jie Lin, Mohamed Sabry Aly, Sean Young,
Vijay Chandrasekhar, and Bernd Girod

Institute for Infocomm Research, Singapore
Nanyang Technological University, Singapore
Stanford University, CA, USA

Outline

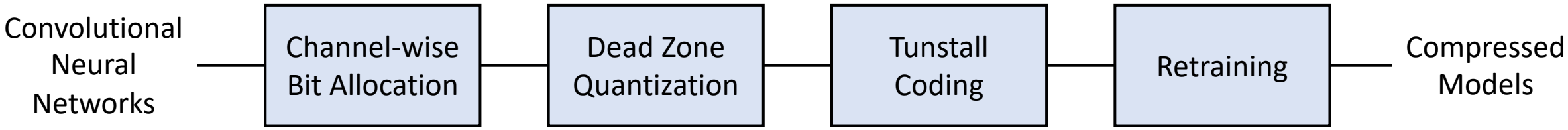
- Compressing Convolutional Neural Networks
- Rate-distortion Optimized Coding Framework
- Bit Allocation, Quantization, Entropy Coding, and Retraining
- Results on ResNet and MobileNet
- Performance on TPU and Eyeriss
- Conclusions

Deep Convolutional Neural Networks

Image Classification Accuracy on ImageNet

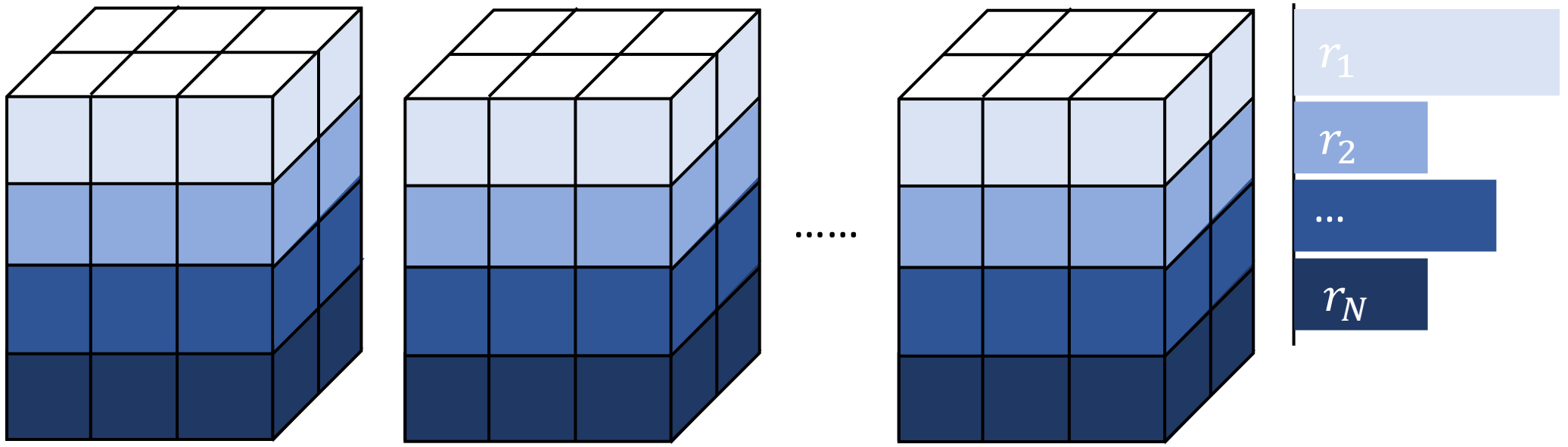
Model	Depth	Parameters	Size	Accuracy
AlexNet (Krizhevsky et al., 2012)	8	62.3×10^6	269 MB	57.2 %
VGG-16 (Simonyan et al., 2014)	16	138.4×10^6	528 MB	71.5 %
ResNet-50 (He et al., 2016)	50	25.6×10^6	102 MB	75.2 %
Inception-V3 (Szegedy et al., 2016)	159	23.9×10^6	96 MB	78.0 %
DenseNet-201 (Huang et al., 2017)	201	20.2×10^6	81 MB	77.3 %

Rate-distortion Optimized Coding Framework



20× compression ratio on deep CNNs, **4.3×**/**2.8×** speedups on TPU/Eyeriss

Channel-wise Bit Allocation



Total Rate $\sum_{i=1}^{L}\sum_{j=1}^{N_i} r_{i,j}$

Rate Distortion Optimization

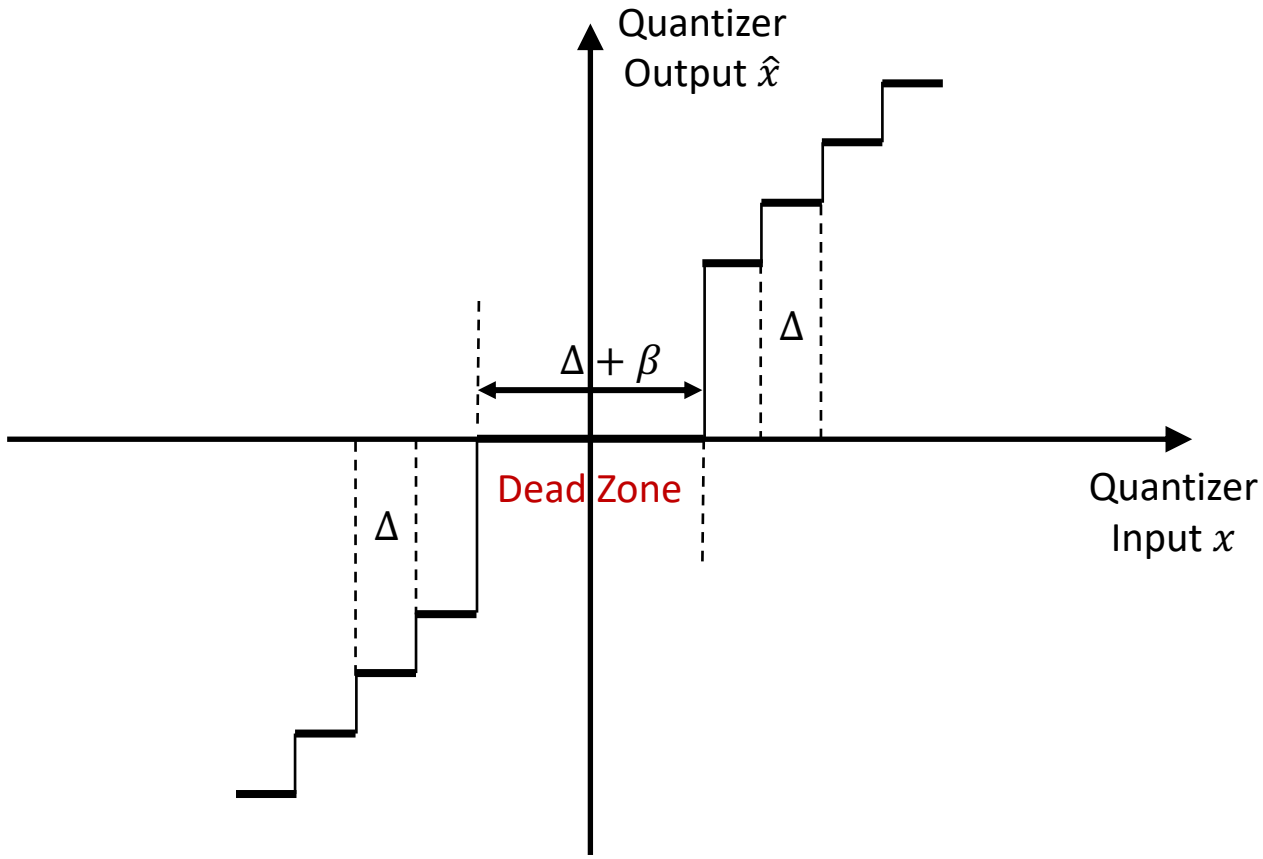
$$min\ \mathrm{D} = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \quad s.t. \sum_{\mathrm{i}=1}^{\mathrm{L}} \sum_{\mathrm{j}=1}^{\mathrm{N_i}} r_{i,j} \leq \mathrm{R}$$

$$\textit{Solution}\ \frac{\partial D_{i,1}}{\partial r_{i,1}} = \frac{\partial D_{i,2}}{\partial r_{i,2}} = \cdots = \frac{\partial D_{i,N_i}}{\partial r_{i,N_i}} = \lambda$$

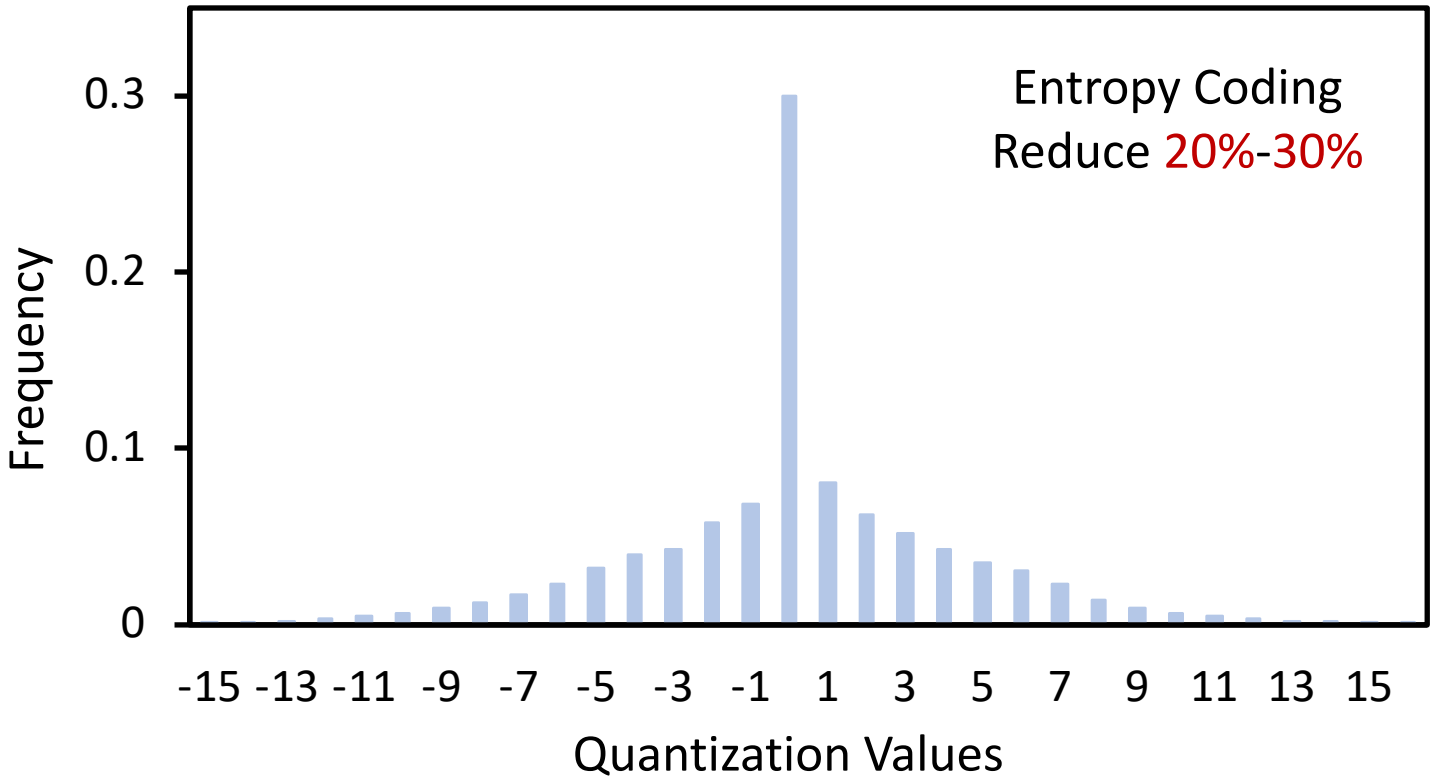
(*Additivity Property and Lagrangian Formulation**)

*Wang Zhe et al. Optimizing the Bit Allocation for Compression of Weights and Activations of Deep Neural Networks. ICIP 2019

Dead Zone Quantization

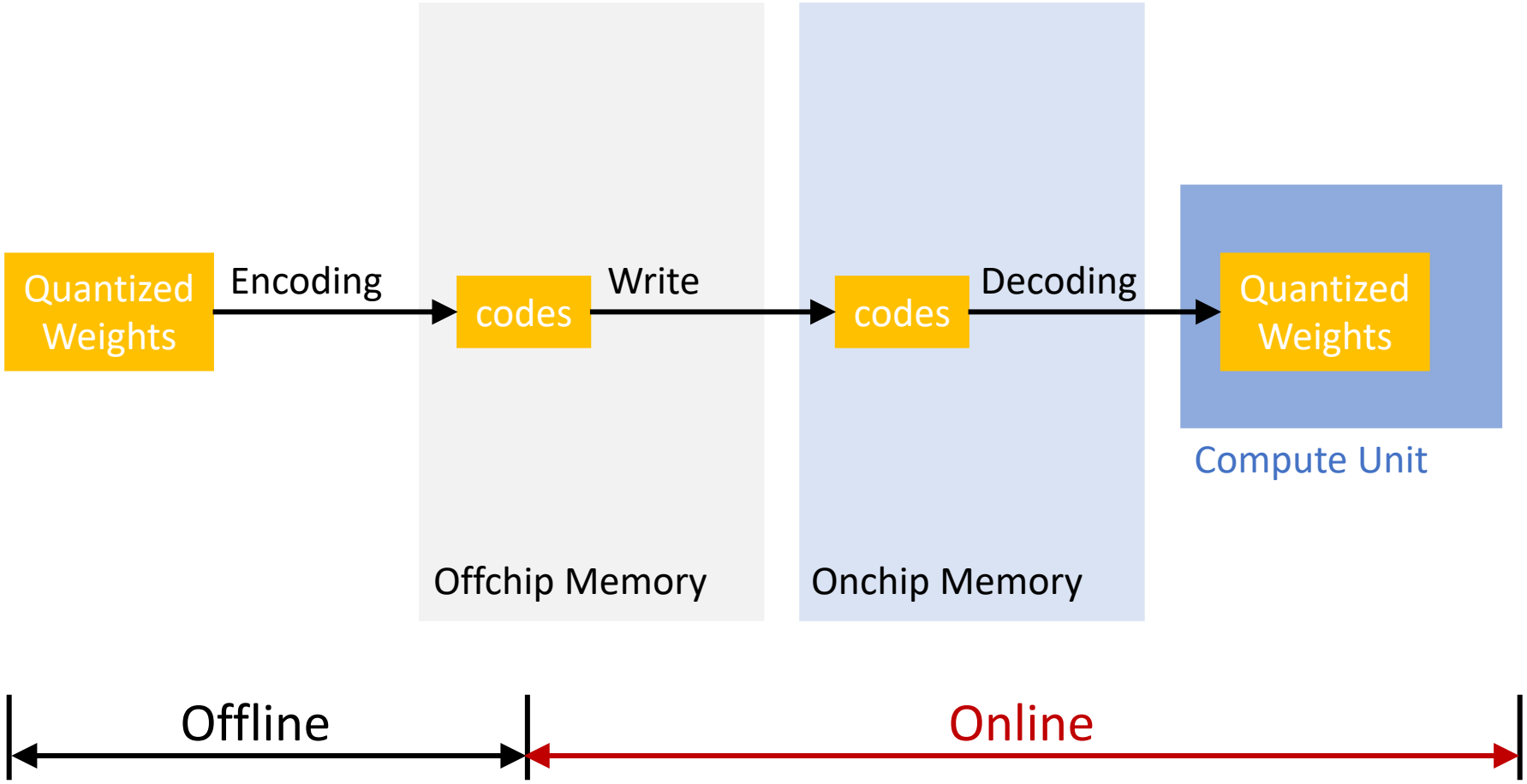


Distribution of Quantization Values



Dead zone quantization with 5 bits on ResNet-50

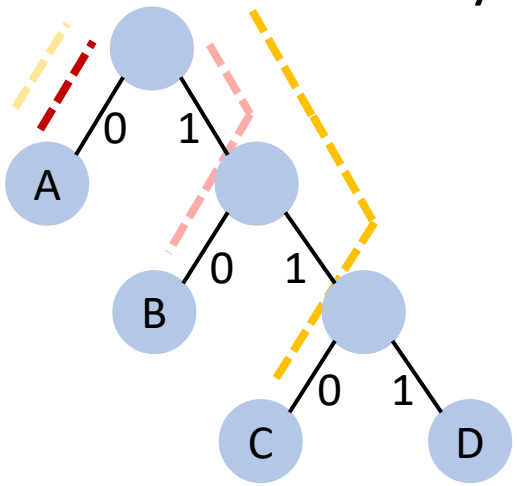
The Decoding Challenge



Huffman Coding vs Tunstall Coding

0 10 110 0

Bit-by-Bit

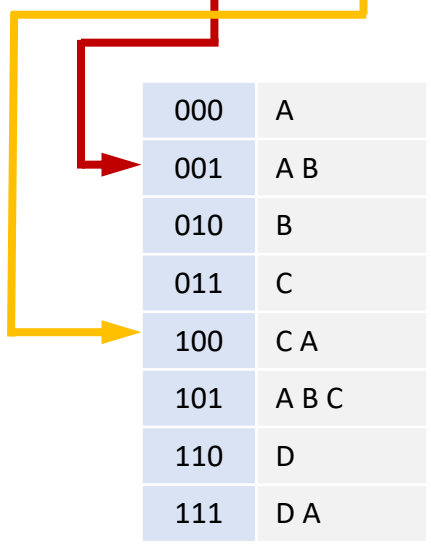


A + B + C + A

Huffman Decoding

001 100

Byte-by-Byte



000	A
001	A B
010	B
011	C
100	C A
101	A B C
110	D
111	D A

AB + CA

Tunstall Decoding

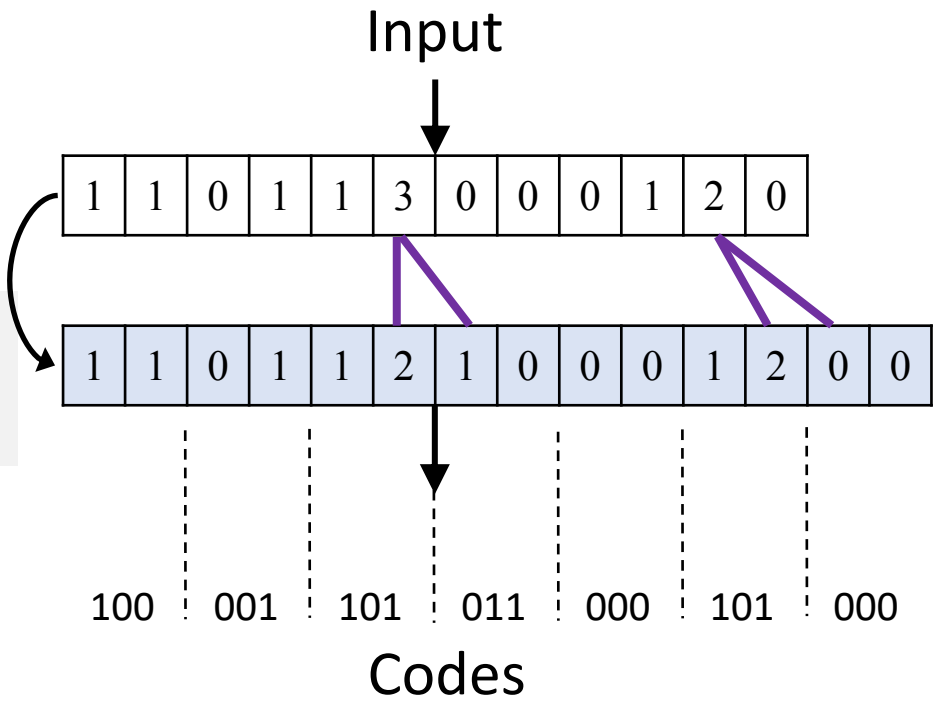
Modified Tunstall Coding

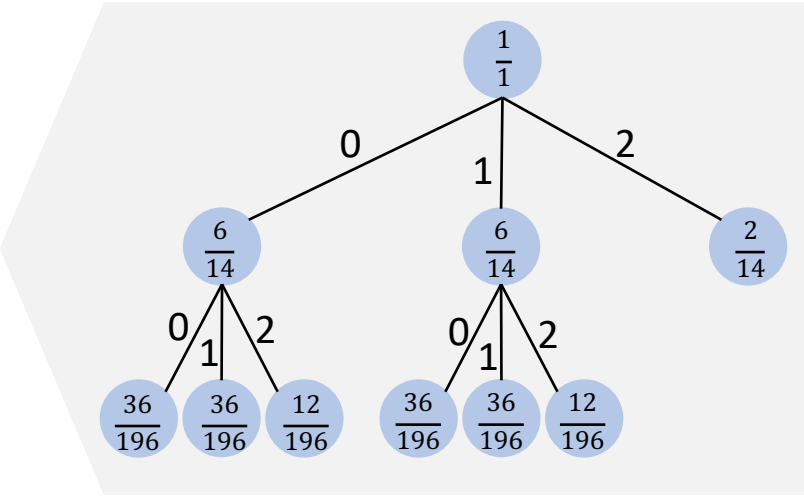
Preprocessing

$K_R = 2$

If $x \geq K_R$:

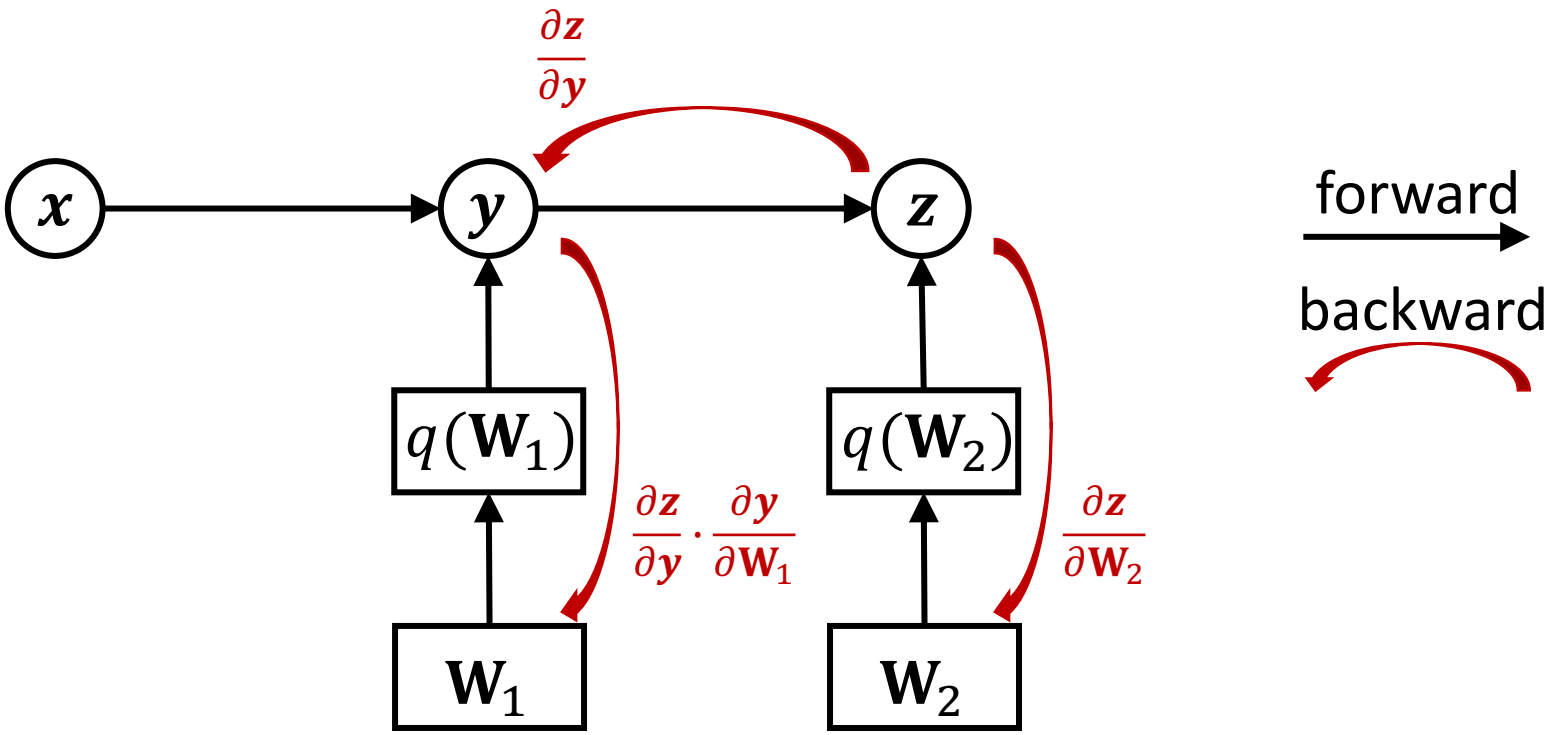
$x \rightarrow K_R, x - K_R$





Ramon Llados-Bernaus and Robert L. Stevenson. Fixed-length entropy coding for robust video compression. TCSVT 1998

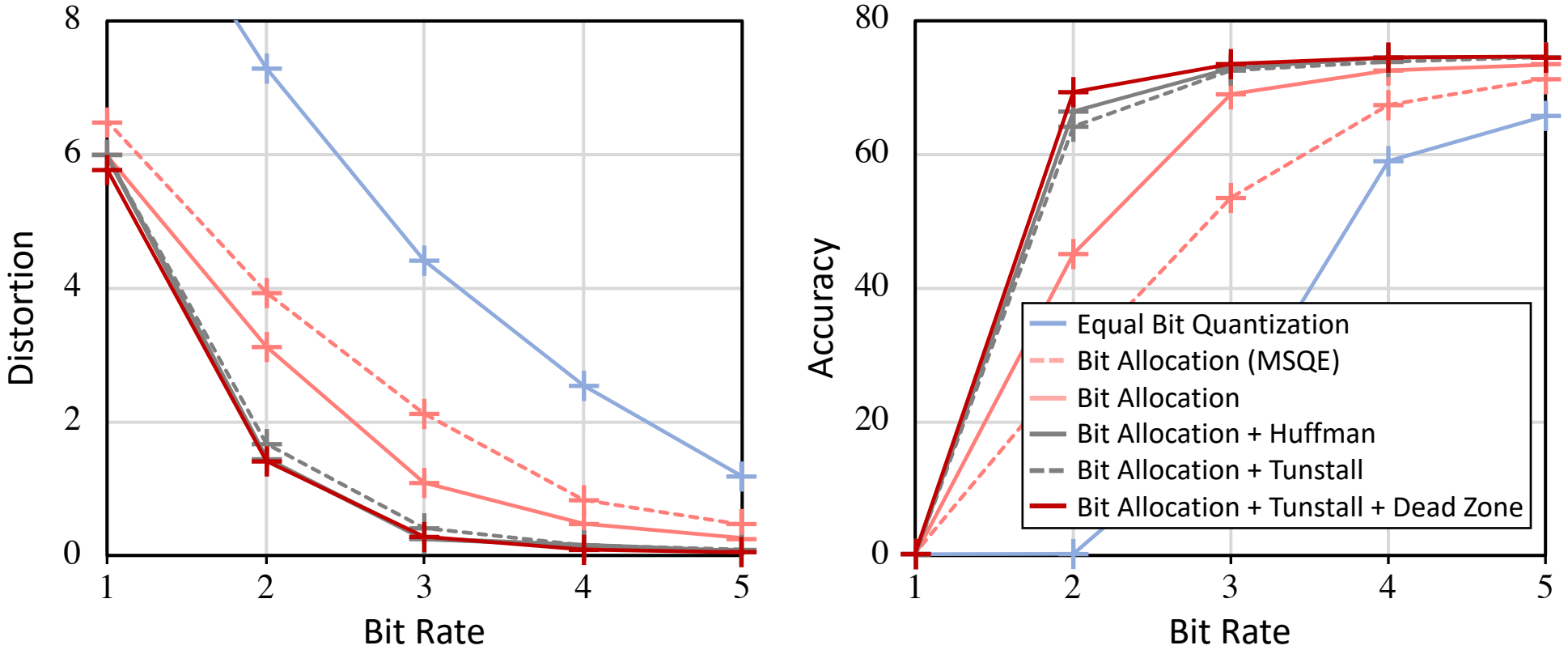
Retraining with STE



Straight-through Estimator (STE)

Bengio et al., Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv 2013

Effectiveness of the Coding Ingredient



Comparison with State-of-the-Arts on ImageNet

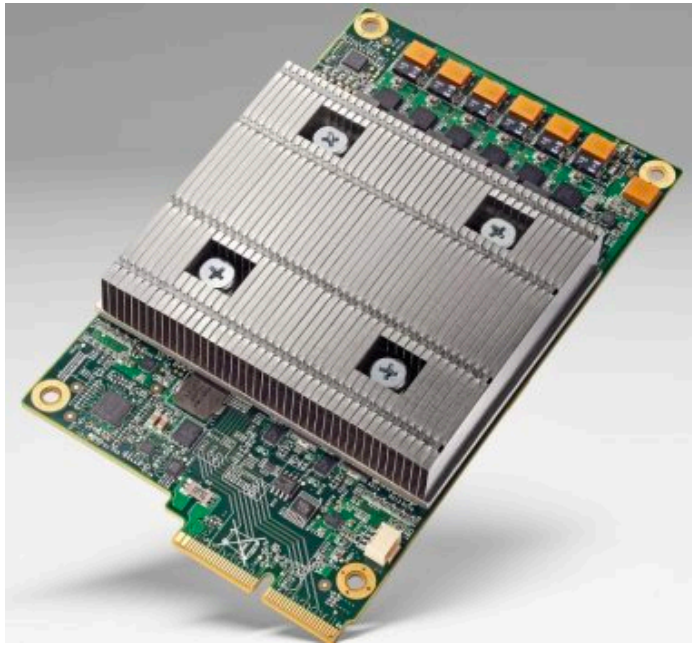
Method	Bit Rate	Comp. Ratio	Size	Accuracy
ResNet-50 Original	32-Bit	-	102 MB	75.5 %
INQ (Zhou et al., 2017)	5-Bit	6.4 ×	-	74.7 %
LQ-Nets (Zhang et al., 2018)	2-Bit	16 ×	-	75.1 %
Coreset (Dubey et al., 2018)	-	16 ×	6.5 MB	74.0 %
DeepCABA (Wiedemann et al., 2019)	-	17 ×	6.1 MB	74.1 %
EPR (Okay et al., 2020)	-	17 ×	5.9 MB	73.5 %
Ours	**1.6-Bit**	**20 ×**	**5.1 MB**	**75.1 %**

Results on ResNet-18, ResNet-34, and MobileNet-v2

Model	Bit Rate	Comp. Ratio	Size	Accuracy
ResNet-18	32-Bit	-	46.7 MB	69.0 %
Compressed	**1.6-Bit**	**20 ×**	**2.3 MB**	**68.7 %**
ResNet-34	32-Bit	-	83.0 MB	73.0 %
Compressed	**1.6-Bit**	**20 ×**	**4.9 MB**	**72.3 %**
MobileNet-v2	32-Bit	-	13.2 MB	71.0 %
Compressed	**3.2-Bit**	**10 ×**	**1.3 MB**	**70.2 %**

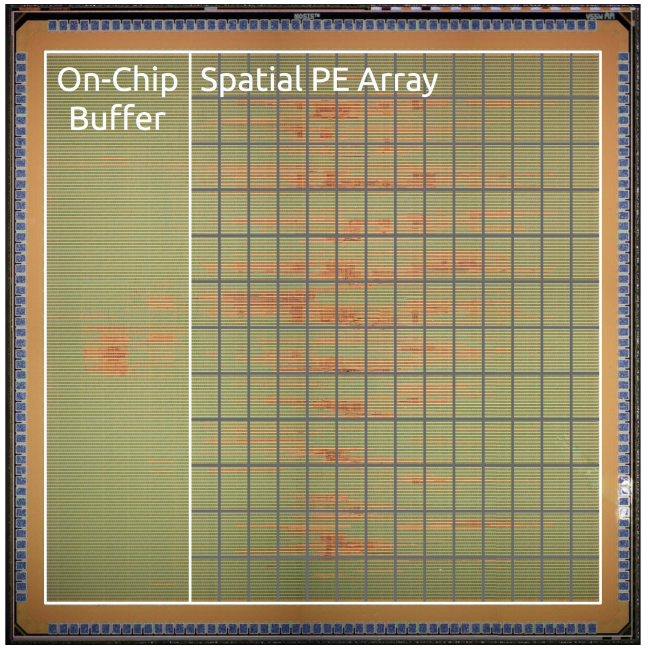
Evaluation on TPU and Eyeriss

TPU



Memory Bandwidth **12.5** GBytes/sec
Computing Performance **96** TOPs

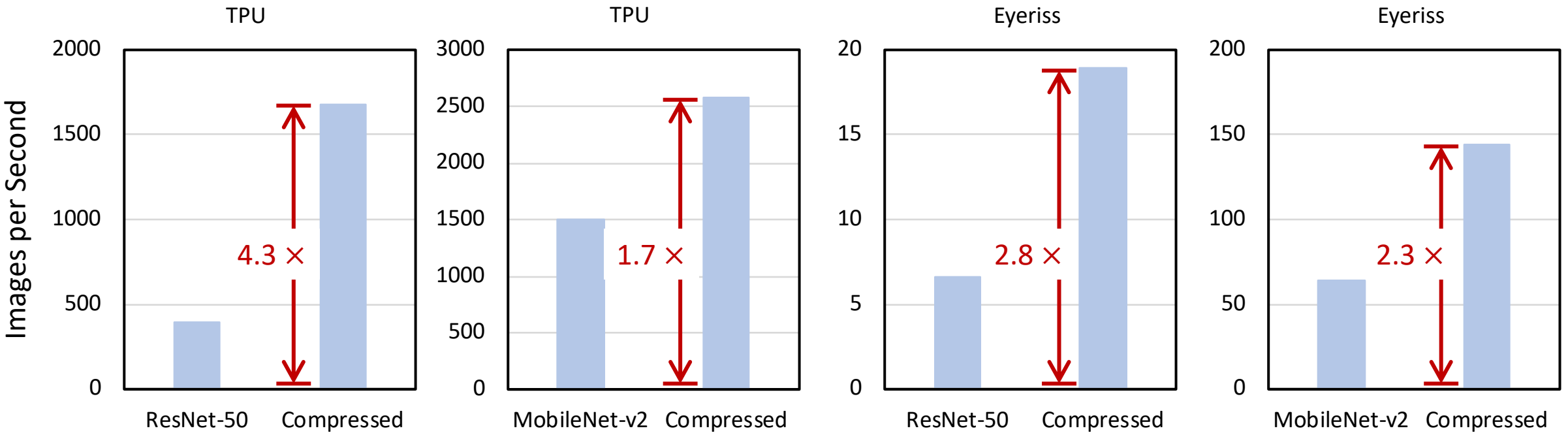
Eyeriss



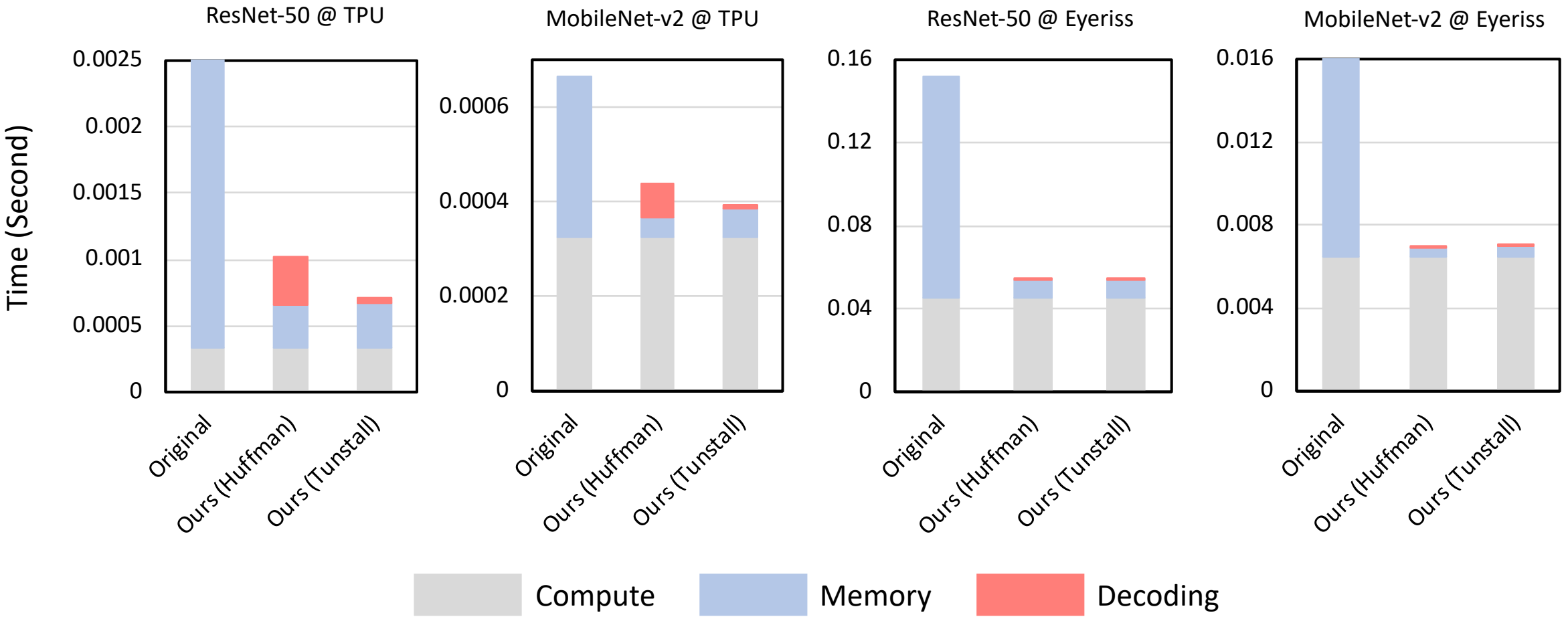
Memory Bandwidth **1** GBytes/sec
Computing Performance **34** GOPs

https://eyeriss.mit.edu/

Inference Speed



Time Breakdown



Conclusions

- Bit allocation, dead zone, and Tunstall coding all benefit the compression
- Decoding may slow down the processing if not efficient
- Tunstall decoding is about 10x faster than Huffman decoding
- Our approach obtains 20x/10x compression on ResNet/MobileNet
- The compressed models can bring up to 4.3x/2.7x speedups on TPU/Eyeriss

Thank You

Contact: mark.wangzhe@gmail.com