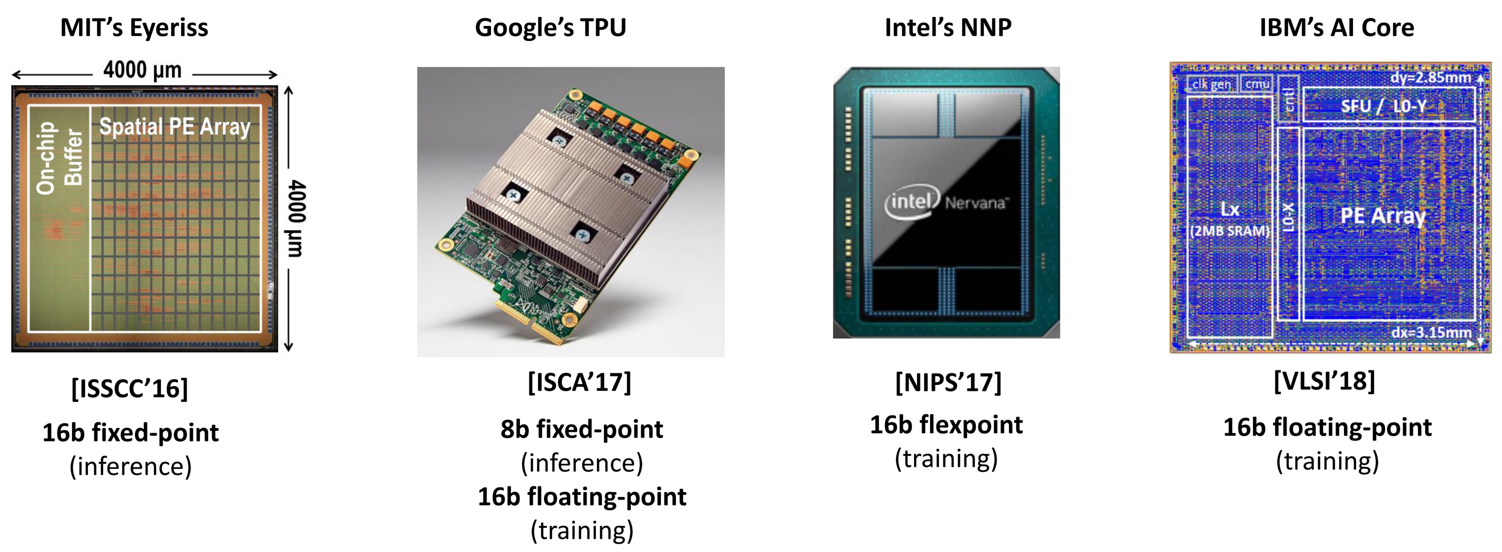
Per-Tensor Fixed-Point Quantization of the Back-Propagation Algorithm

Charbel Sakr & Naresh Shanbhag
University of Illinois at Urbana-Champaign
{sakr2,shanbhag}@Illinois.edu

Motivation

Machine Learning in Reduced Precision



Are these the minimum precisions required?
Can minimum precision requirements be determined analytically?
Specifically for training

Current Approaches

Reducing Complexity of Inference

Forward path quantization:
- Fixed-point via SQNR analysis [Lin et al., ICML'17]
- Extreme quantization, e.g., BinaryNet, via training [Courbariaux et al., NIPS'15 – Rastegari et al., ECCV'16 – Hubara et al., NIPS'16]

Structural methods:
- Pruning [Han et al., NIPS'15]
- Parameter clustering [Wu et al., ICML'18]

No theoretical guarantees on accuracy

Reducing Complexity of Training

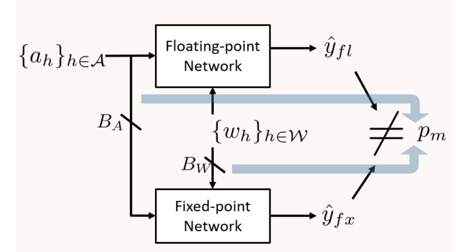
Quantized back-propagation:
- Fixed-point via stochastic rounding [Gupta et al., ICML'15]
- Fixed-point/floating-point hybrid [Koester et al., NIPS'17]
- Finite precision floating-point [Wang et al., NIPS'18]

Gradient Compression:
- Extreme gradient quantization, e.g., TernGrad [Wen et al., NIPS'17]
- Gradient sparsification [Lin et al., ICLR'18]

Hybrid fixed/floating-point training

Largely based on heuristics

Fixed-point inference with theoretical guarantees

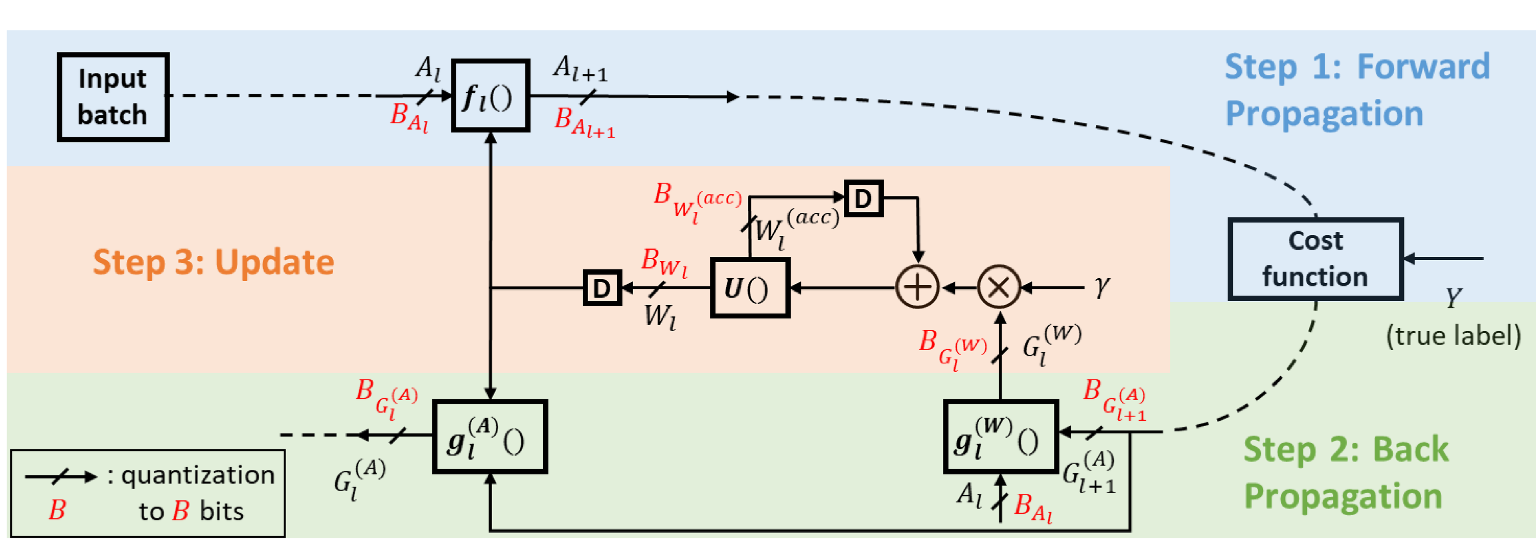


$$p_m \leq \Delta_A^2 E_A + \Delta_W^2 E_W$$

Sakr, Kim, Shanbhag
ICML 2017
Sakr & Shanbhag
ICASSP 2018

What about training?

Problem Setup and Challenges



- ➢multiple forward quantization noise sources
- ➢unknown gradient dynamic range
- ➢instability due to quantization noise bias in updates
- ➢back-propagation of quantization noise in activation gradients
- ➢risk of premature stoppage of convergence

Criteria-based Approach

Criterion 1: equalization of quantization noise gains

$$B_{W_l} = rnd\left(\log_2\left(\sqrt{\frac{E_{W_l\to p_m}}{E^{(\min)}}}\right)\right) + B^{(\min)} \qquad B_{A_l} = rnd\left(\log_2\left(\sqrt{\frac{E_{A_l\to p_m}}{E^{(\min)}}}\right)\right) + B^{(\min)}$$

Criterion 2: proper gradient clipping

$$r_{G_l^{(W)}} \geq 2\sigma^{(\max)}_{G_l^{(W)}} \qquad r_{G_{l+1}^{(A)}} \geq 4\sigma^{(\max)}_{G_{l+1}^{(A)}}$$

Criterion 3: quantization bias elimination

$$\Delta_{G_l^{(W)}} < \frac{\sigma^{(\min)}_{G_l^{(W)}}}{4}$$

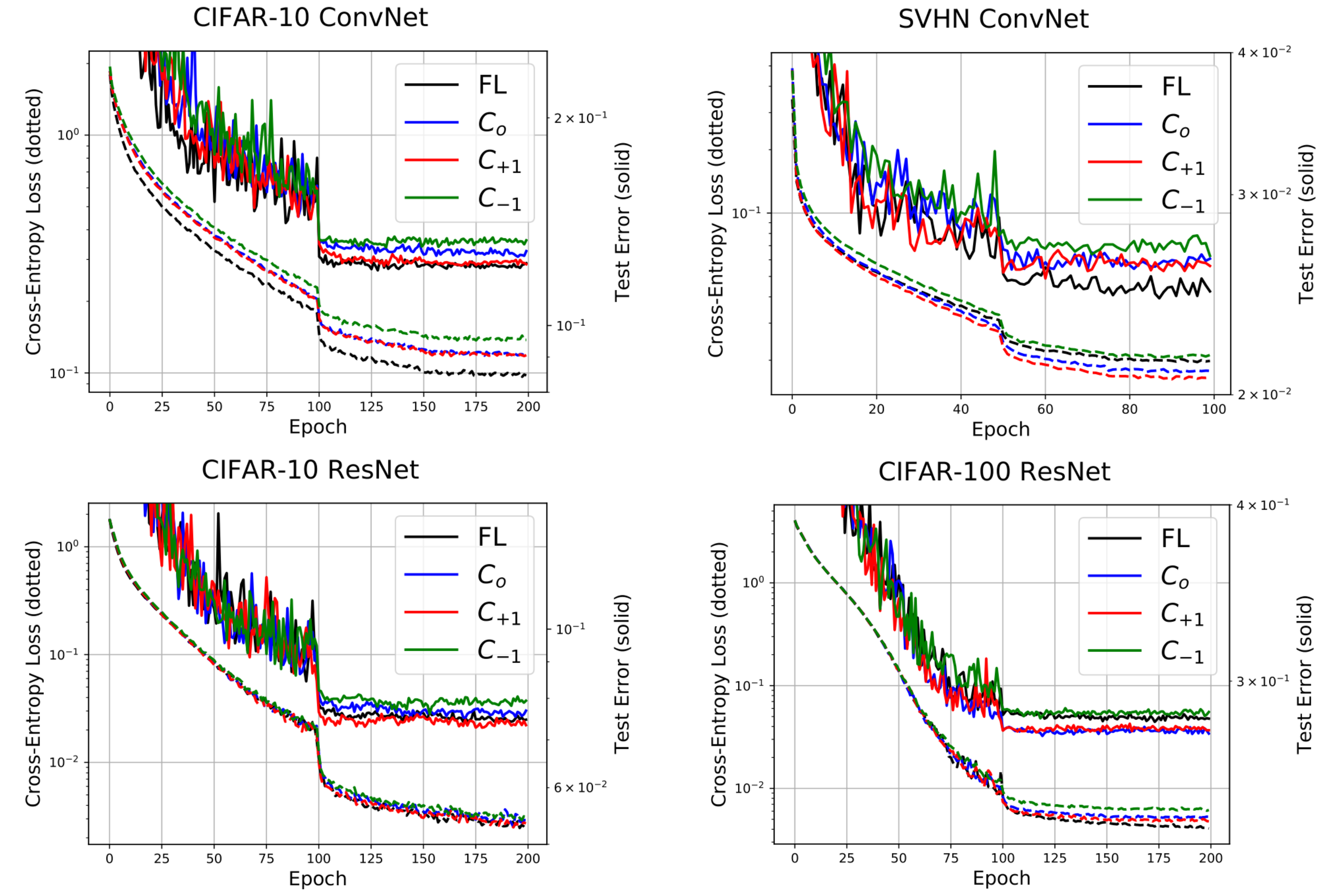
Criterion 4: back-propagated noise bound

$$\Delta_{G_{l+1}^{(A)}} < \frac{\Delta_{G_l^{(W)}}}{\sqrt{\lambda^{(\max)}_{G_{l+1}^{(A)}\to G_l^{(W)}}}}\left(\frac{\left|G_l^{(W)}\right|}{\left|G_{l+1}^{(A)}\right|}\right)^{1/4}$$

Criterion 5: accumulator stopping condition

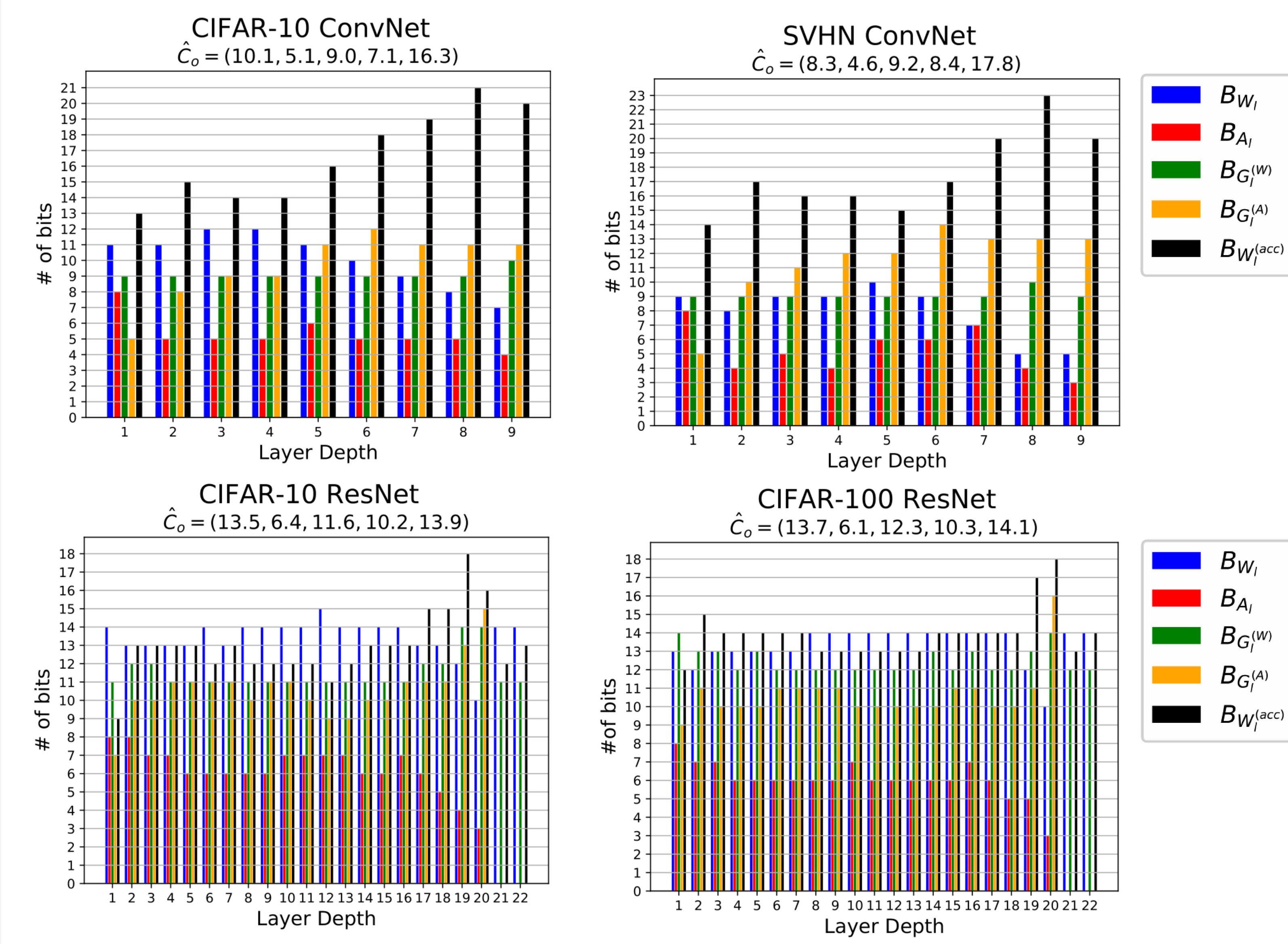
$$r_{W_l^{(acc)}} \geq 2^{-B_{W_l}} \qquad \Delta_{W_l^{(acc)}} < \gamma^{(\min)}\Delta_{G_l^{(W)}}$$

Convergence with Close-to-Minimal Precision



- ➢FX training was believed to be impossible due to dynamic range issues [Koester et al. – NIPS'2017]
- ➢proposed FX training is able to match FL training accuracy
- ➢precision assignment found to be nearly minimal

Per-Layer Precision Trends



- ➢weight precision decreases from network input to output
- ➢precisions of activation gradients and weight accumulators increase
- ➢ResNets have uniform precision requirements per tensor type

Hyper-Precision Reduction is Inefficient

	$\mathcal{C}_W$ (10^6b)	$\mathcal{C}_A$ (10^6b)	$\mathcal{C}_M$ (10^9FA)	$\mathcal{C}_C$ (10^6b)	Test Error	$\mathcal{C}_W$ (10^6b)	$\mathcal{C}_A$ (10^6b)	$\mathcal{C}_M$ (10^9FA)	$\mathcal{C}_C$ (10^6b)	Test Error
	CIFAR-10 ConvNet					**SVHN ConvNet**				
FL	148	9.3	94.4	49	12.02%	148	9.3	94.4	49	**2.43%**
FX (C_o)	**56.5**	**1.7**	11.9	14	12.58%	**54.3**	**1.9**	10.5	14	2.58%
BN	100	4.7	**2.8**	49	18.50%	100	4.7	**2.8**	49	3.60%
SQ	78.8	**1.7**	11.9	14	**11.32%**	76.3	**1.9**	10.5	14	2.73%
TG	102	9.3	94.4	**3.1**	12.49%	102	9.3	94.4	**3.1**	3.65%
	CIFAR-10 ResNet					**CIFAR-100 ResNet**				
FL	1784	96	4319	596	7.42%	1789	97	4319	597	28.06%
FX (C_o)	**726**	**25**	785	216	7.51%	**750**	**25**	776	216	**27.43%**
BN	1208	50	**128**	596	**7.24%**	1211	50	**128**	597	29.35%
SQ	1062	**25**	785	216	7.42%	1081	**25**	776	216	28.03%
TG	1227	96	4319	**37.3**	7.94%	1230	97	4319	**37.3**	30.62%

- ➢feedforward binarization (BN) and gradient ternarization (TG) fail to match FL accuracy for same topology
- ➢stochastic quantization (SQ) provides marginal returns
- ➢BN, TG, SQ do not address the fundamental problem of realizing true FX training

Acknowledgement

This work was supported in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.