ZEM: Zero-cycle Bit-masking Module for Deep Learning Refresh-less DRAM

Duy-Thanh Nguyen[1], Nhut-Minh Ho[2], Minh-Son Le[1],
Weng-Fai Wong[2], Ik-Joon Chang[1]
[1]{dtnguyen,sonlm,ichang}@khu.ac.kr, [2]{minhhn,wongf}@comp.nus.edu.sg

경희대학교
KYUNG HEE UNIVERSITY

I. Abstract

In sub-20nm technologies, DRAM cells suffer from poor retention time. With the technology scaling, this problem tends to be worse, significantly increasing refresh power of DRAM. This is more problematic in memory heavy applications such as deep learning systems, where a large amount of DRAM is required, DRAM refresh power contributes to a considerable portion of total system power. With the growth in deep learning workloads, this is set to get worse. In this work, we present a *zero-cycle bit-masking* (ZEM) scheme that exploits the asymmetry of retention failures, to eliminate DRAM refresh in the inference of convolution neural networks, natural language processing, and the image generation based on generative adversarial network. Through careful analysis, we derive a bit-error-rate (BER) threshold that does not affect the accuracy of inference. Our proposed architecture, along with the techniques involved, are applicable to all types of DRAMs. Our results on 16Gb devices show that ZEM can improve the performance by up to 17.31% while reducing the total energy consumed by DRAM by up to 43.03%, dependent on the type of DRAM.

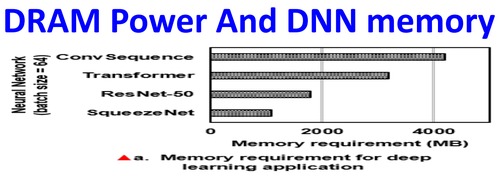
II. Introduction

Motivation

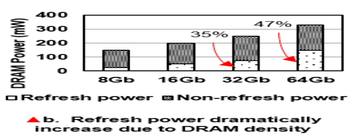
- DNNs become deeper and wider-> Large memory
- DNN can be tolerant to the errors
- DRAM refresh can consume up to 46% of total system power

–> DRAM refresh power can be reduced

DRAM Power And DNN memory requirements



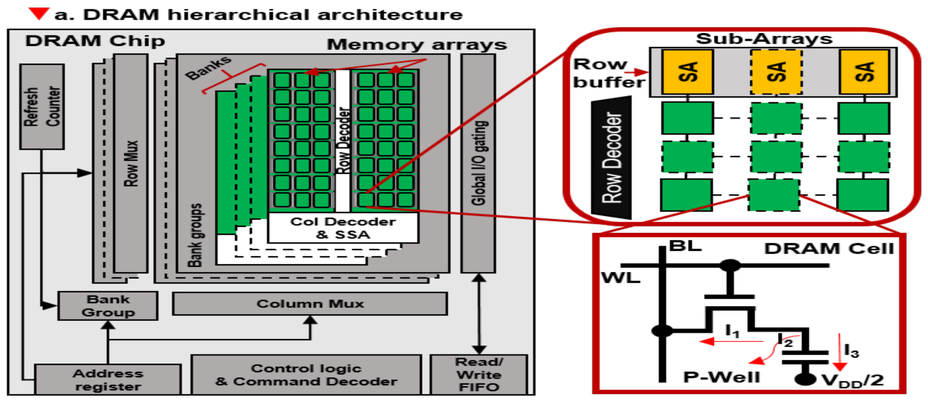
▲a. Memory requirement for deep learning application



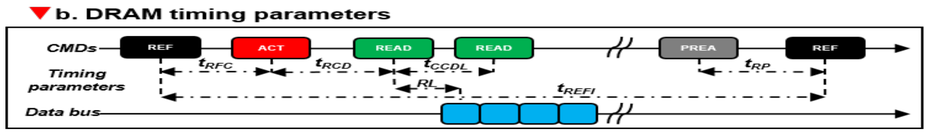
▲b. Refresh power dramatically increase due to DRAM density

DRAM architecture

▼a. DRAM hierarchical architecture



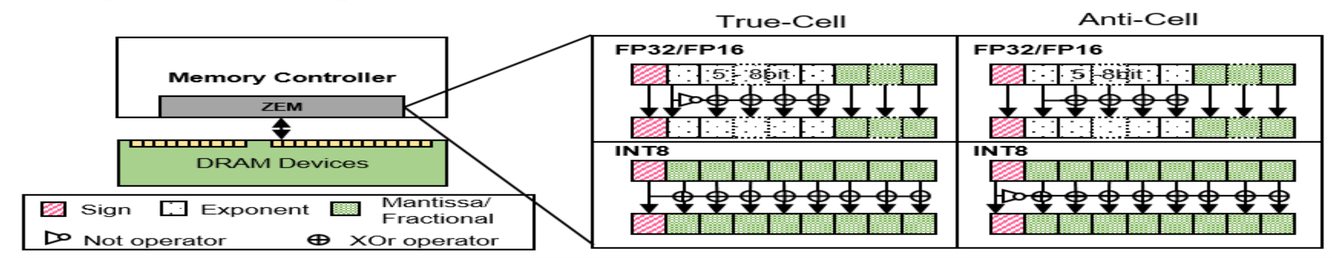
▼b. DRAM timing parameters



DRAM errors rarely occur on one bits

III. Our Approaches

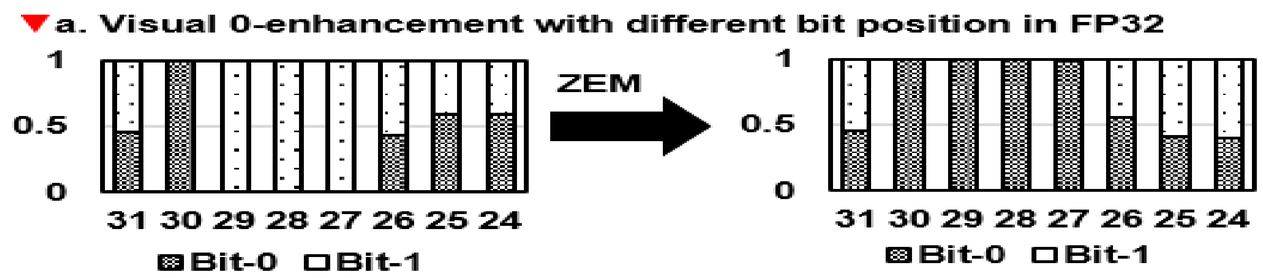
Zero-cycle bit-Masking



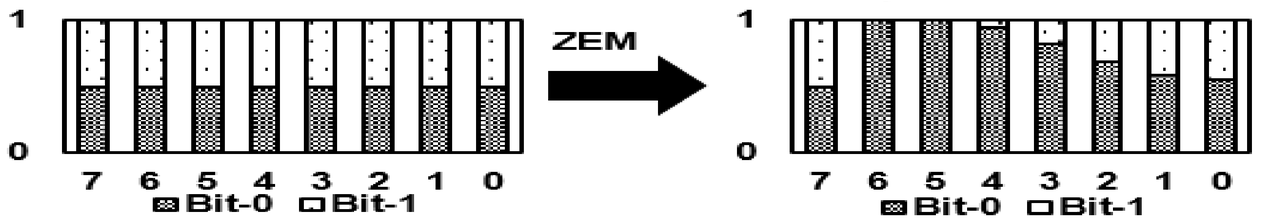
FP32 Decimal (Binary) Exponent Only	Normal		1(1000000)	2(10000001)	3(10000010)	4(10000011)
		0(01111111)	-1(0111110)	-2(01111101)	-3(01111100)	-4(01111001)
	ZEM	0(00000000)	-1(0000001)	-2(00000010)	-3(00000011)	-4(00000110)
FP16 Decimal (Binary) Exponent Only	Normal		1(10000)	2(10001)	3(10010)	4(10011)
		0(01111)	-1(01110)	-2(01101)	-3(01100)	-4(01011)
	ZEM	0(00000)	-1(00001)	-2(00010)	-3(00011)	-4(00100)
INT8 Decimal (Binary)	Normal	0(0000000)	1(00000001)	2(00000010)	3(00000011)	4(00000100)
			-1(11111111)	-2(11111110)	-3(11111101)	-4(11111100)
	ZEM	0(0000000)	-1(10000000)	-2(10000001)	-3(10000010)	-4(10000011)

Asymmetry-enhancement with ZEM

▼a. Visual 0-enhancement with different bit position in FP32



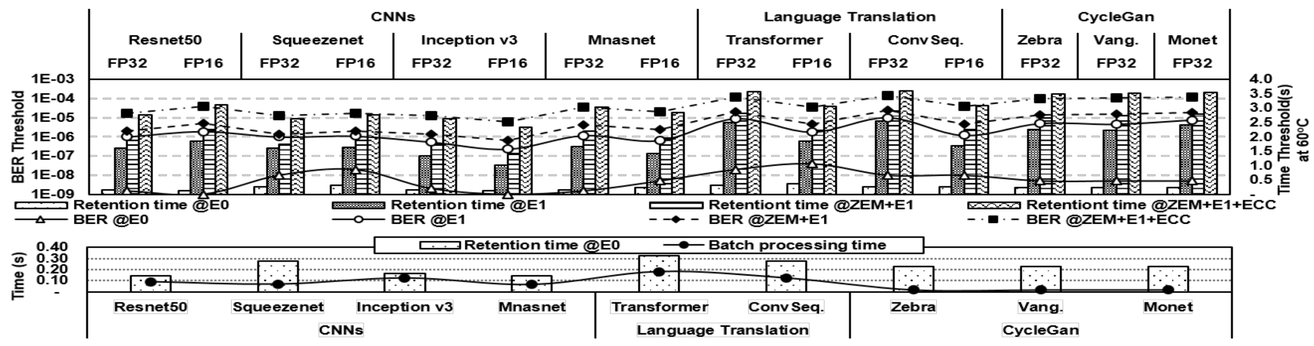
▼b. Visual 0-enhancement with different bit position in INT8



IV. Validation

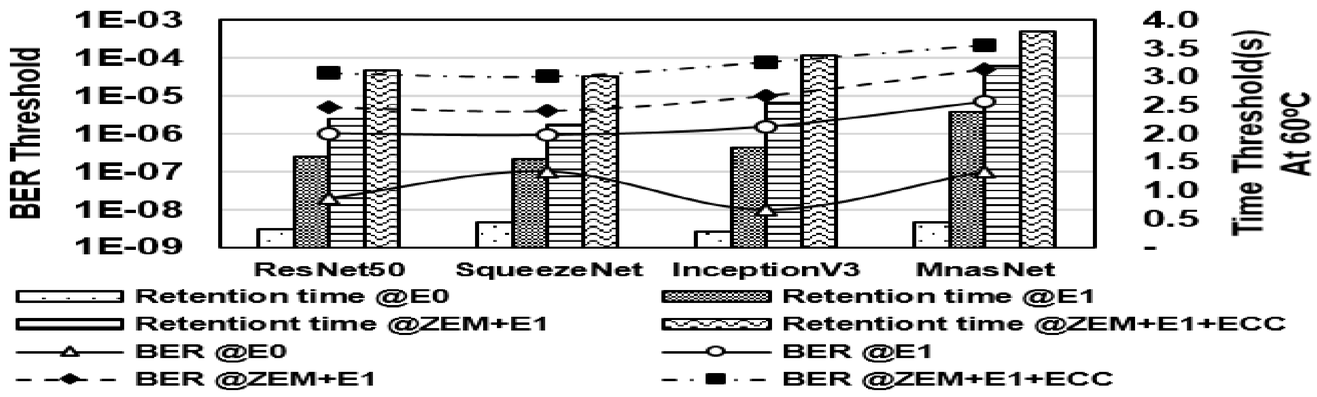
Floating-point 32/16bit

▼a. Bit error rate threshold and acceptance time threshold at 60°C of different deep learning applications



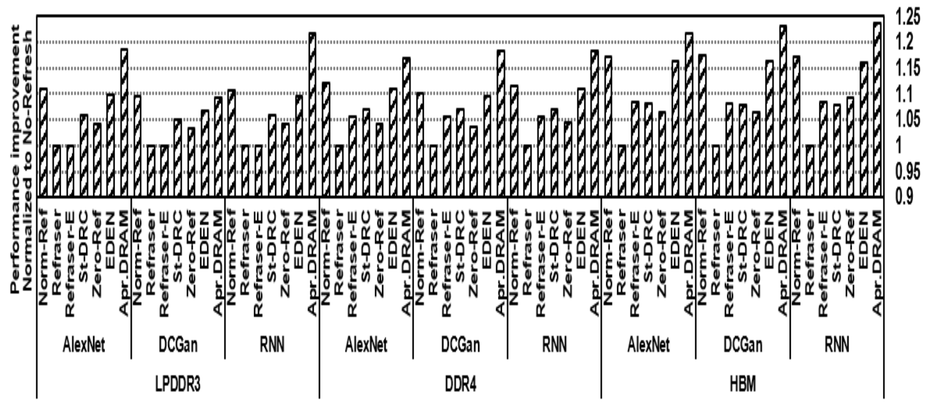
▲b. Comparison between original model time threshold at 60°C and batch processing time of floating-point 32bit

Integer 8bit



V. Energy/Performance Simulation Results

Performance Improvement



Energy Saving