St-DRC: Stretchable DRAM Refresh Controller with No Parity-overhead Error Correction Scheme for Energy-efficient DNNs

Duy-Thanh Nguyen[1], Nhut-Minh Ho[2], Ik-Joon Chang[1]

[1]Kyung Hee University, Republic of Korea

[2]National University of Singapore, Singapore

경희대학교

KYUNG HEE UNIVERSITY

Outline

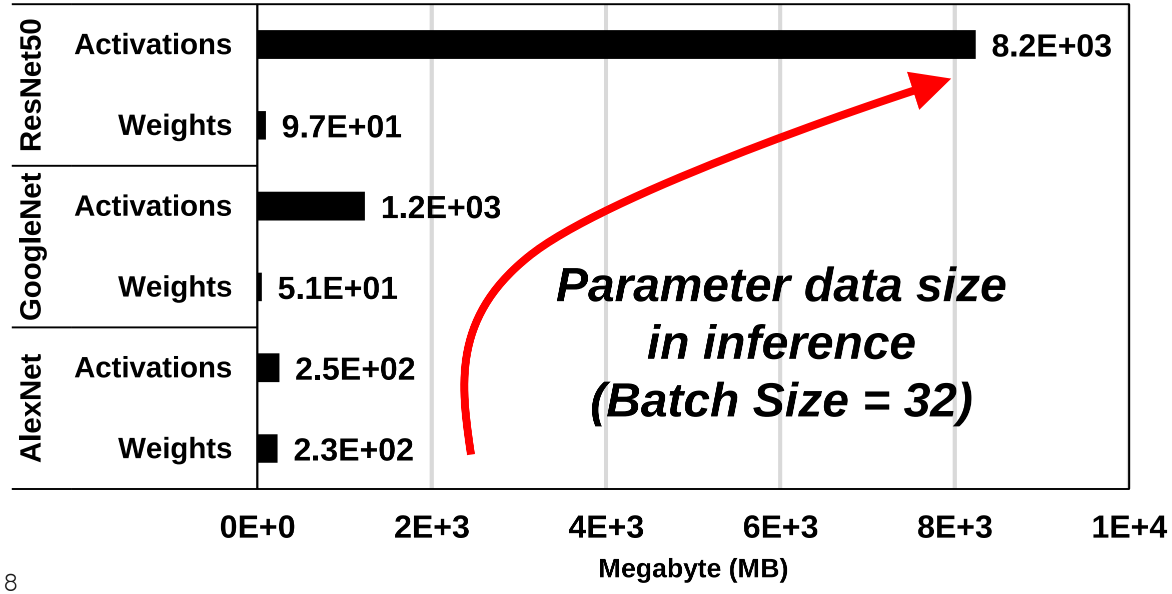
- Motivation
  - The Effect of DRAM Refresh Relaxation
- Major Challenge
  - Floating-point IEEE754 under Retention Errors
  - Characteristic of DNN’s Data
  - Large Sensitivity to Bit-errors of Some Exponent Bits
- Our Approach: Significant-Bit Protection
  - DRAM Controller with Stretchable Refresh Period
  - Validation
- Energy and Performance Simulation Results
- Conclusion

Motivation

- Large processing time for the training of DNNs
  - – Training speed needs to be improved
- Hybrid CPU-GPU platform is widely used for training DNNs
  - – Main Memory DRAM + GPU Memory DRAM → **DRAM power is very significant**

Nvidia Server	Main Memory (DDR4)	GPU Memory (HBM2)
DGX-1 (8X Tesla V-100)	512GB	32GB per GPU x 8 = 256GB
DGX-2 (16X Tesla V-100 /8X Tesla V-100)	1.5TB	32GB per GPU x 16 = 512GB/ 32GB per GPU x 8 = 256GB

- DNNs become deeper and wider
  - – The size of DNN Parameters tends to be larger
  - – DRAM size should be larger
  - – **DRAM power would be more significant in data-center**

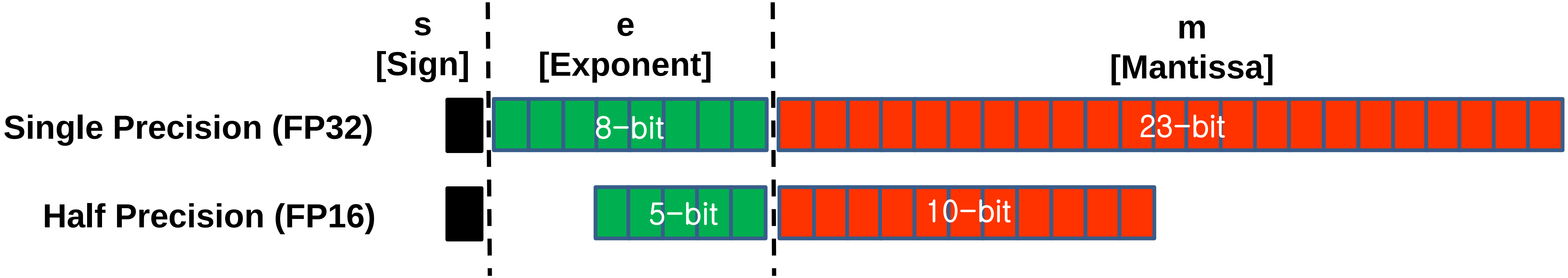


The Effect of DRAM Refresh Relaxation

	RAIDR @ISCA12	**REFLEX @ISCA15**	**Flikker @ASPLOS2011**	**Quality-aware DRAM @DATE2015**
DRAM Power Saving Rate	16.1%	20%	25~32% (standby), 20~25%(overall)	73% (But some quality-loss)
System Performance Improvement Rate	8.1%	7~10%	Not discussed	Not discussed
Application	Generic	Generic	Video	Video
Precise/Approximate	Precise	Precise	Approximate	Approximate
Challenge	Retention time characterization of full rows for all chips (Large Verification Overhead)	Retention time characterization of full rows for all chips (Large Verification Overhead)	Not applicable to DNN (Large Accuracy Degradation)	1. Retention time characterization of full rows for all chips (Large Verification Overhead) 2. Not applicable to DNN (Large Accuracy Degradation)

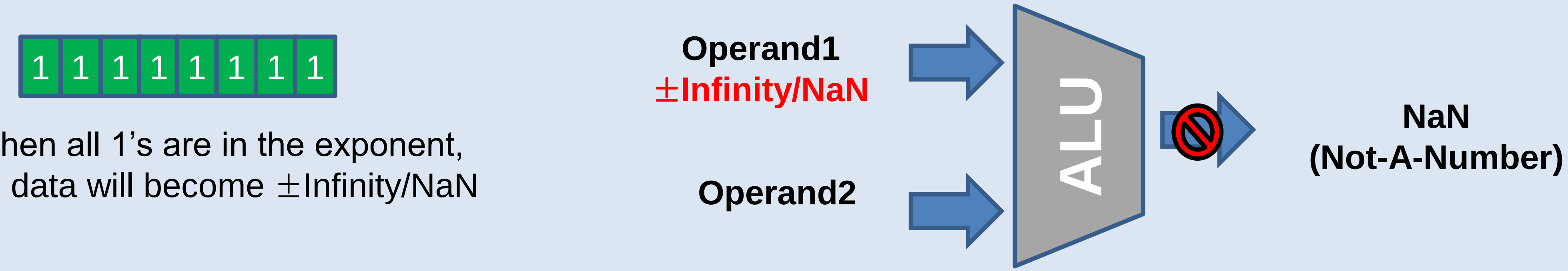
- Our Contribution: Negligible accuracy degradation in DNN in spite of some retention errors, Reasonable verification effort, Significant power saving, System performance improvement

Floating point IEEE 754 under Retention Errors



- Conversion {s,e,m} → {-1^{s} x M x 2$^{(e-Bias)}$ | Bias = 127 for FP32 or 15 for FP16, M = 1.m}

Significant Challenge Due to Retention Errors

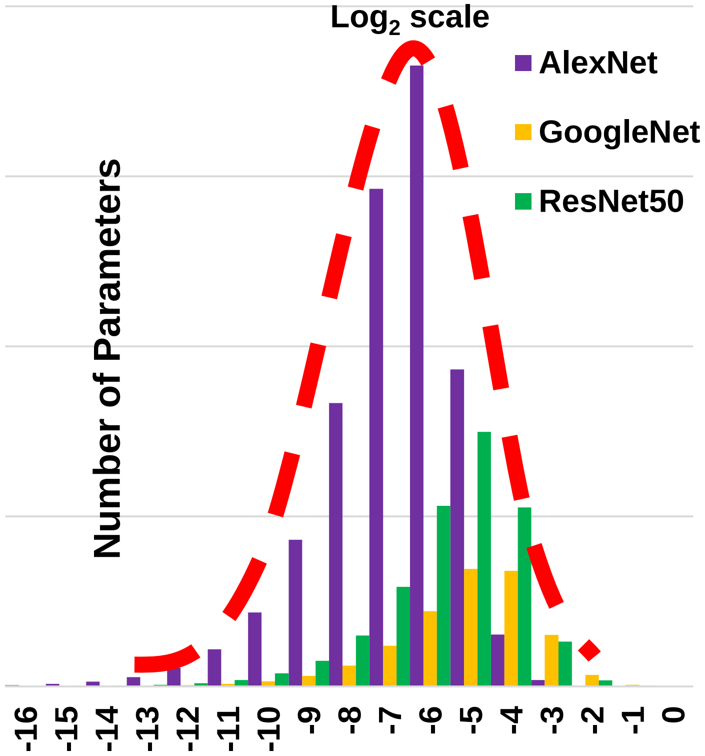


When all 1’s are in the exponent, the data will become ±Infinity/NaN

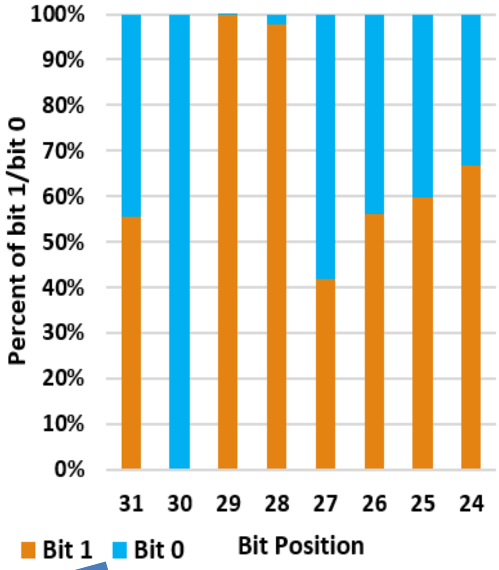
- Error propagation due to +/- Infinity and NaN → Catastrophic system errors

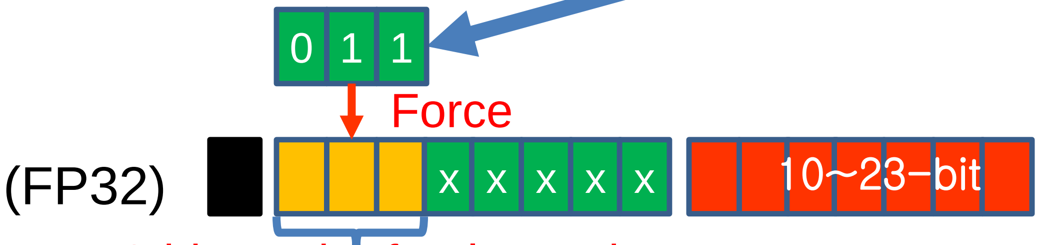
Characteristic of DNNs’ Data

Weight parameter



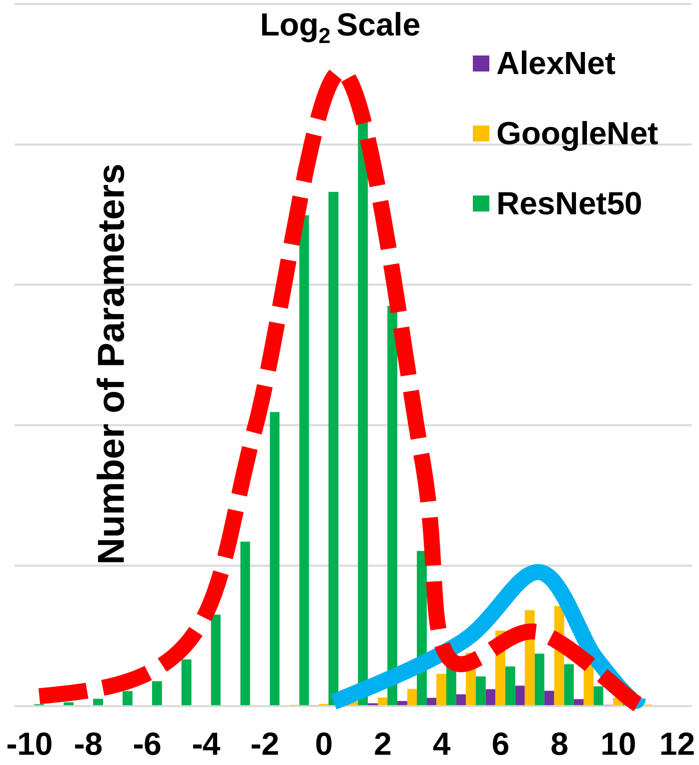
99% ~ $[2^{-13}:2^{-2}]$





3-bit can be freely used for another purpose

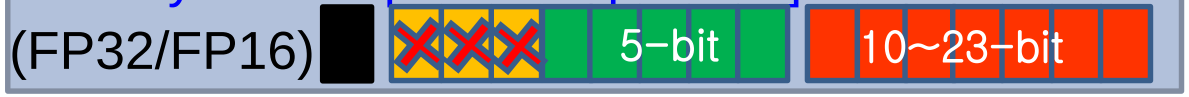
Activation parameter



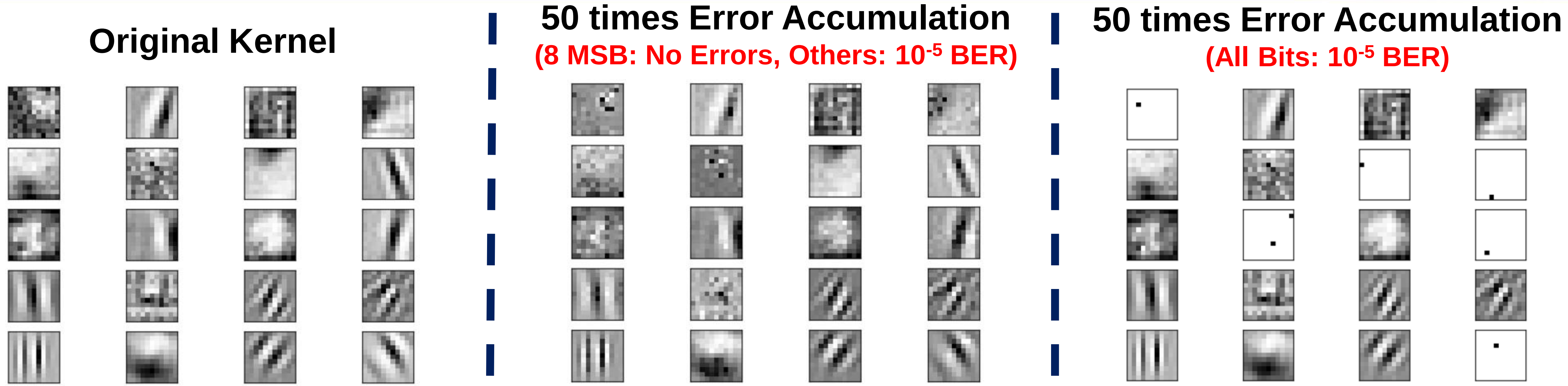
99% ~ $[2^{0}:2^{11}]$

99% ~ $[2^{-10}:2^{11}]$

- Weight and Activation are fully covered with only 5-bit exponent ~ $[2^{-15}:2^{15}]$



Large Sensitivity to Bit-errors of Some Exponent Bits (Inference)

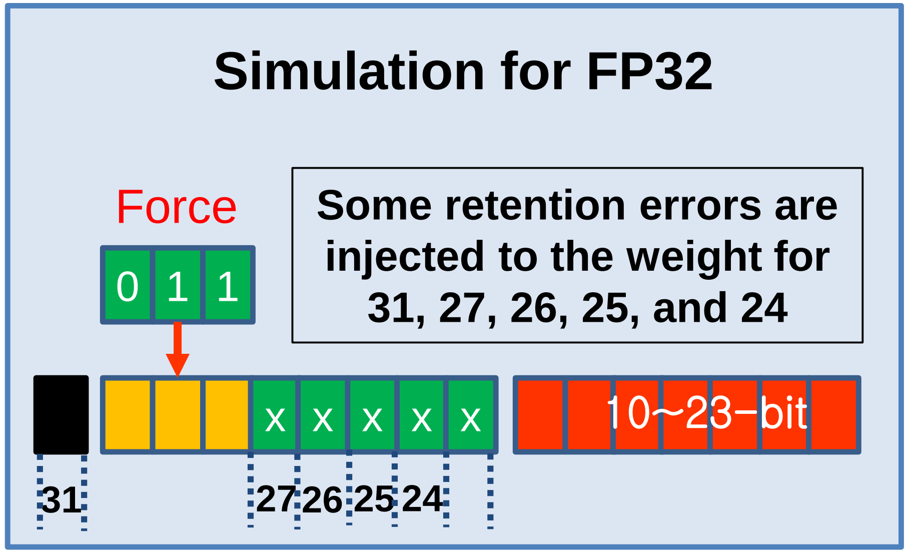


Original Kernel

50 times Error Accumulation (8 MSB: No Errors, Others: 10^{-5} BER)

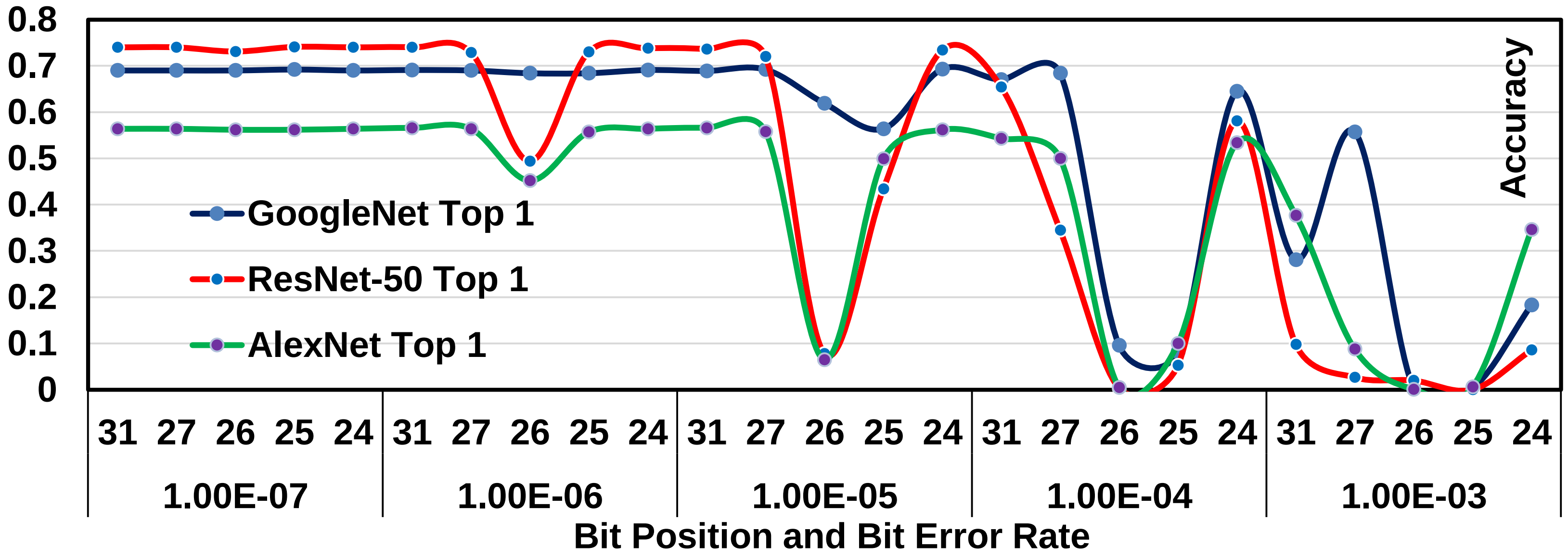
50 times Error Accumulation (All Bits: 10^{-5} BER)

- Exponent bits are extremely sensitive to the error

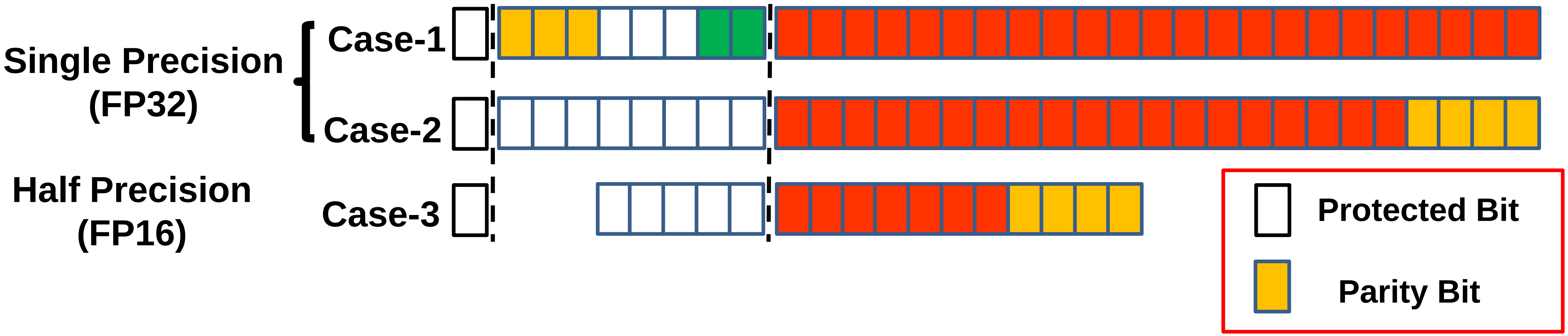


Simulation for FP32

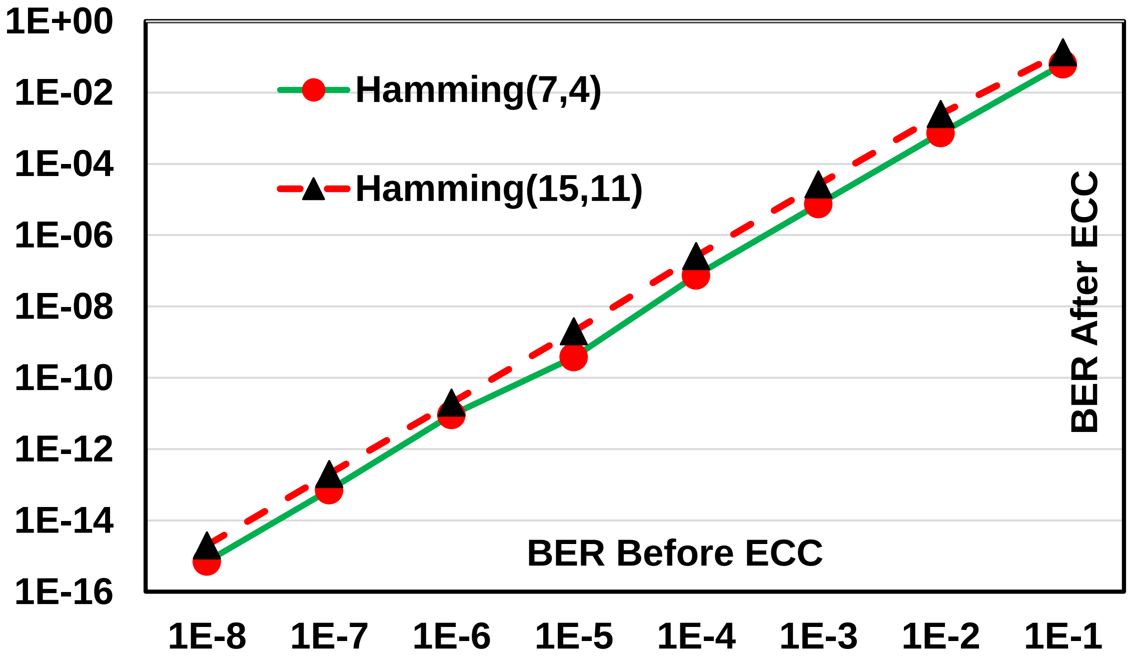
Some retention errors are injected to the weight for 31, 27, 26, 25, and 24



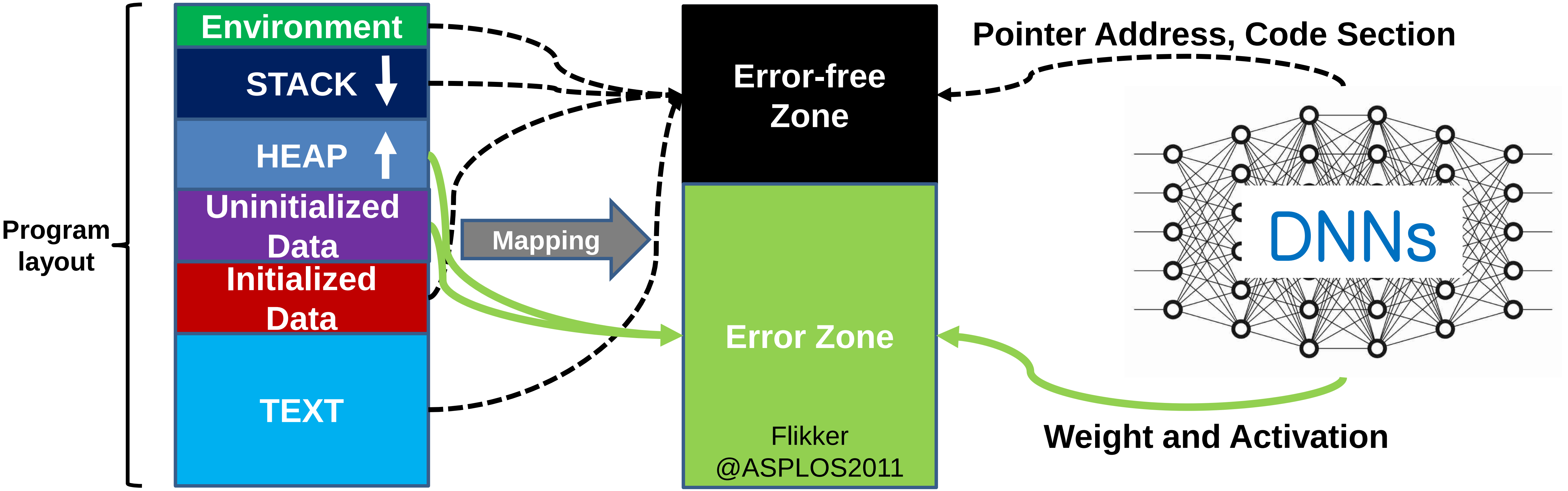
Our Approach: Significant-Bit Protection



- Utilize some non-critical bits as parity bits for ECC
  - No additional memory overhead
- Protect some critical bits from retention errors
- Hamming codes for ECC
  - Hamming(7,4) : Case-1
  - Hamming(15,11) : Case-2 and Case-3

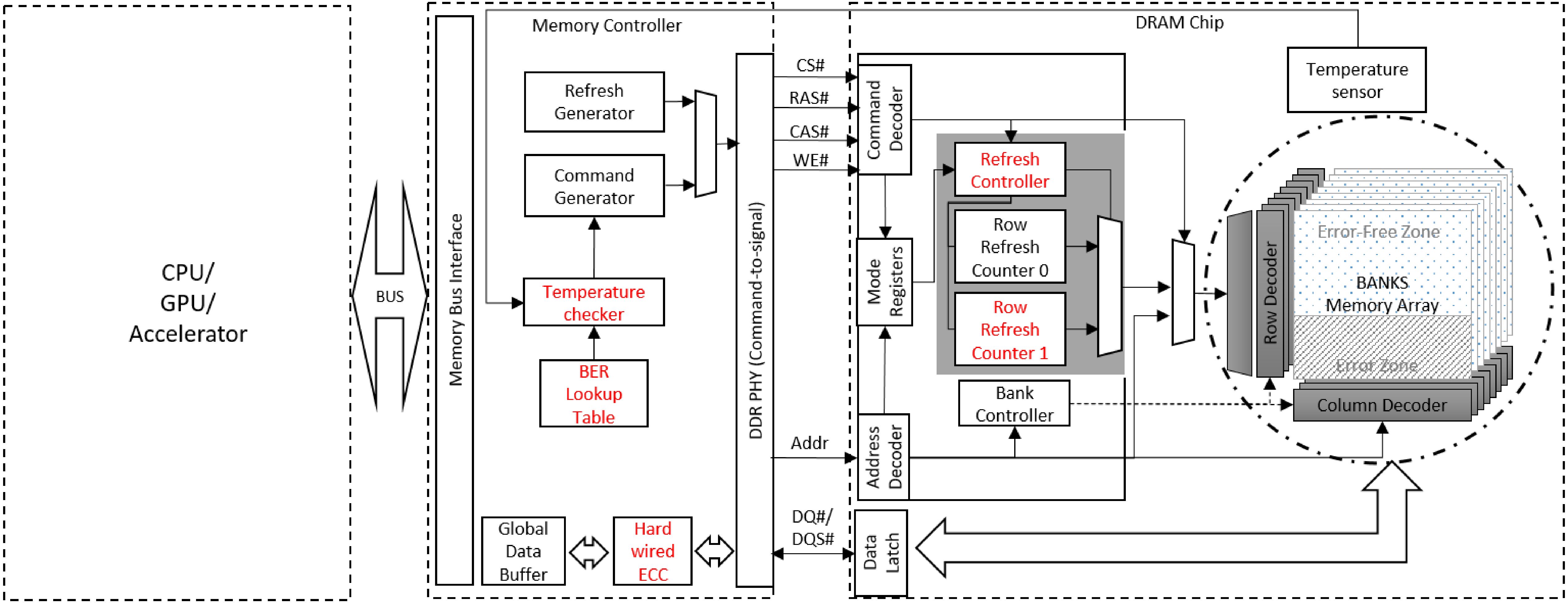


Data Mapping in Our Scheme



- Text, initialized data, environment, and stack sections are stored in the ‘Error-free Zone’.
  - **Prevent retention errors for control information**
- Weight and activation parameters (significantly dominant in DNN) are stored in the ‘Error Zone’.

DRAM Controller with Stretchable Refresh Period



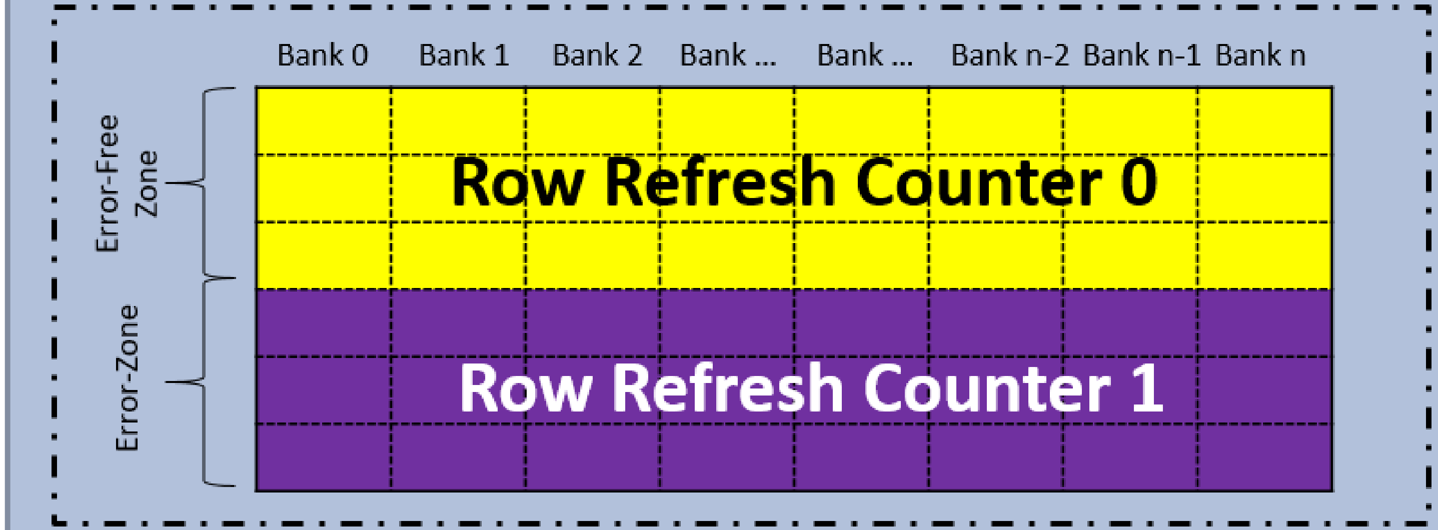
System level

Memory Controller & PHY

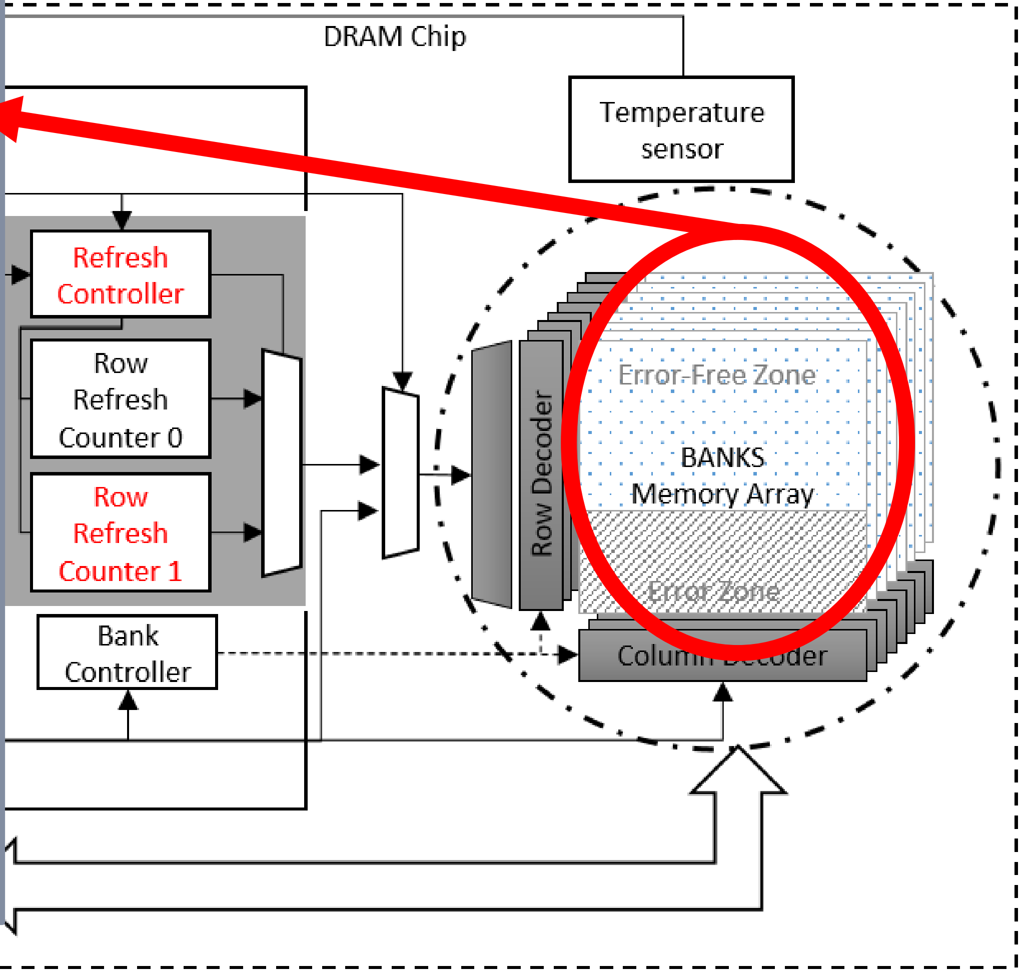
Physical DRAM Chip

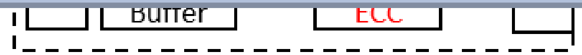
- We slightly modify the Memory Controller and the Physical DRAM Chip

DRAM Controller with Stretchable Refresh Period (Cont.)



- On DRAM Chip
  - –Error-free Zone (Row Refresh Counter 0)
  - –Error Zone (Row Refresh Counter 1)



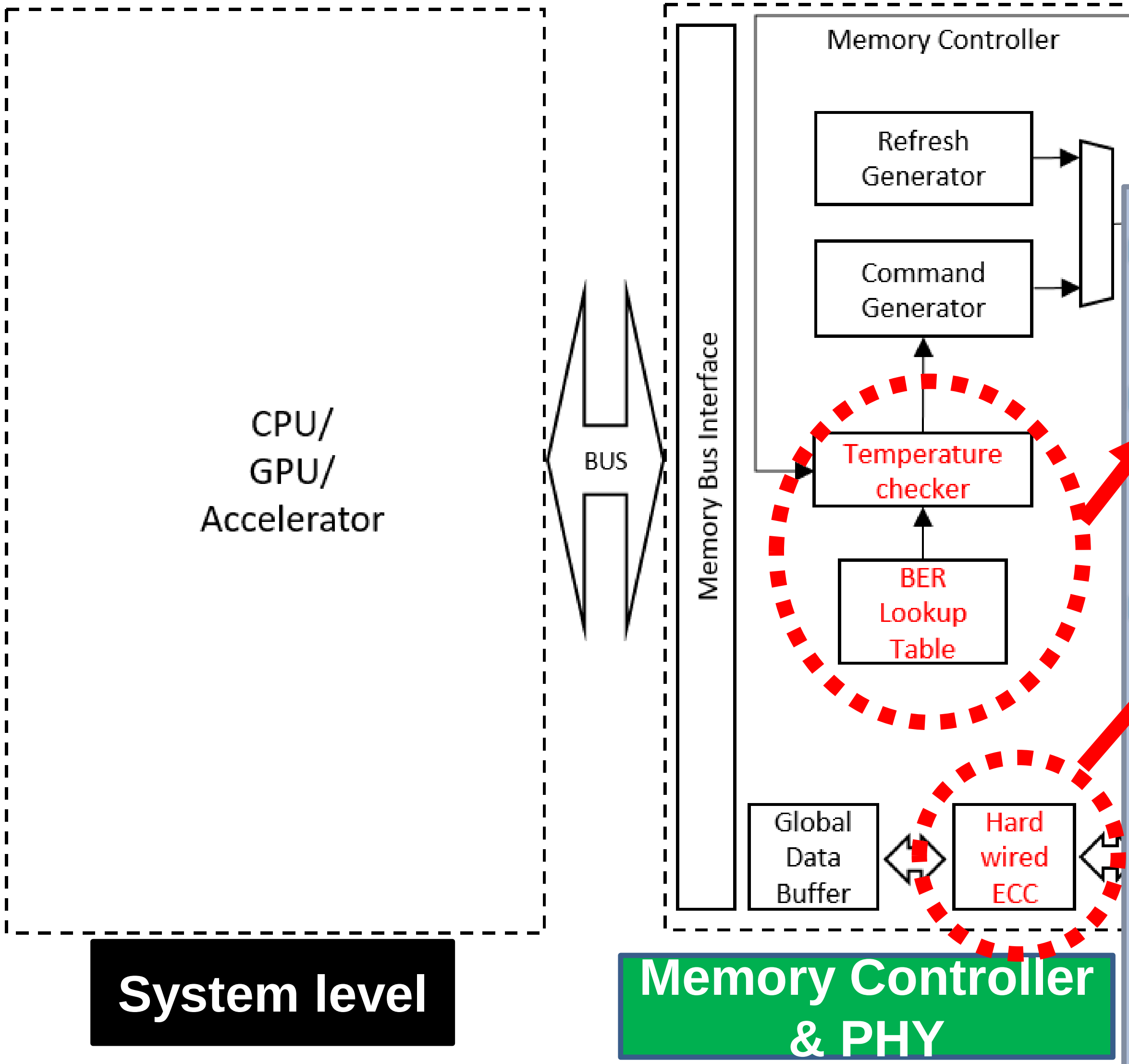


System level

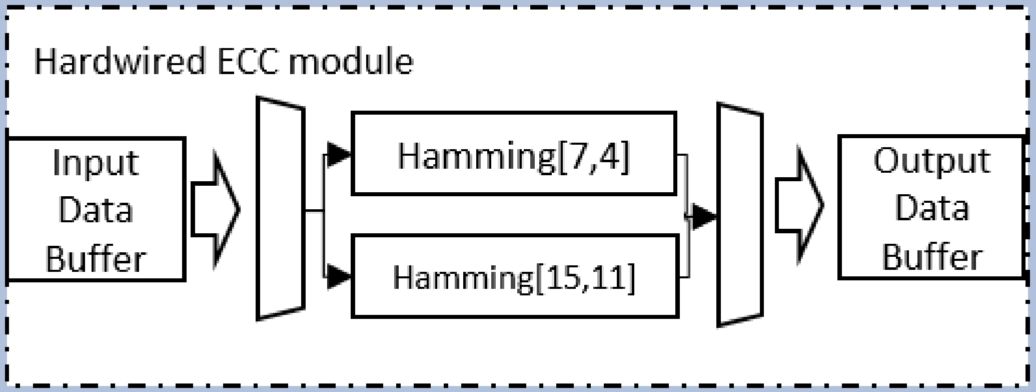
Memory Controller & PHY

Physical DRAM Chip

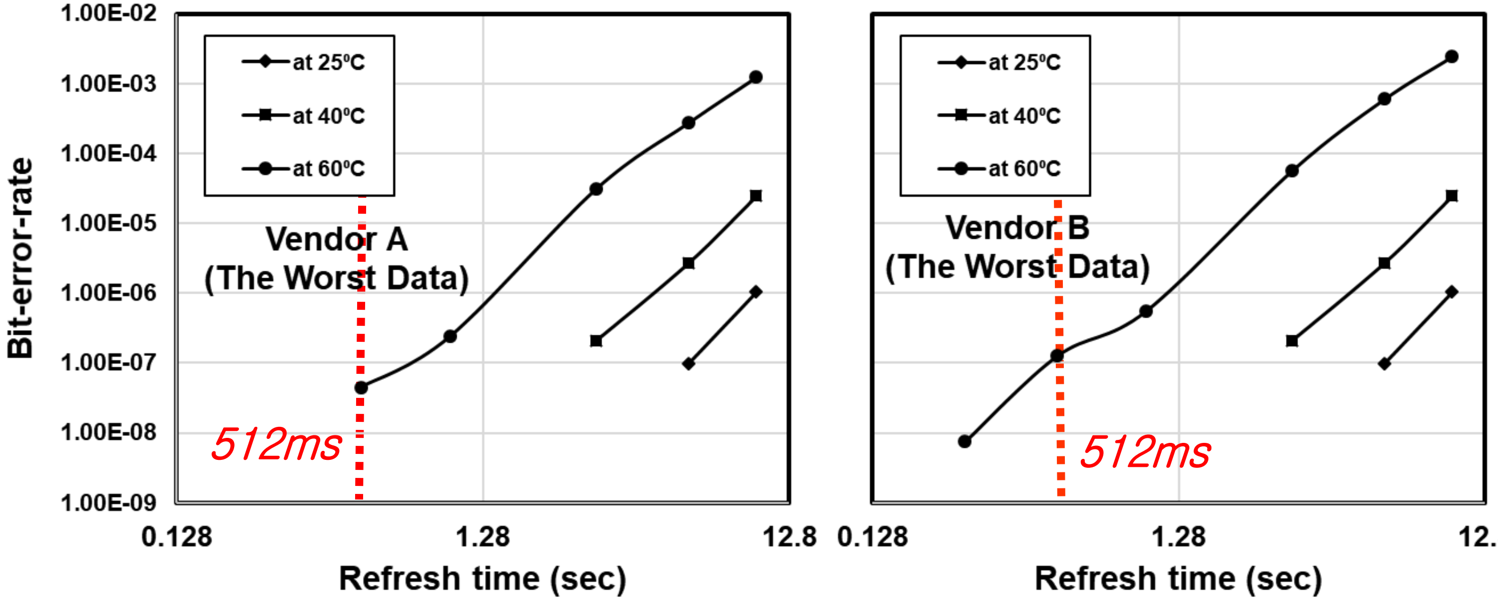
DRAM Controller with Stretchable Refresh Period (Cont..)



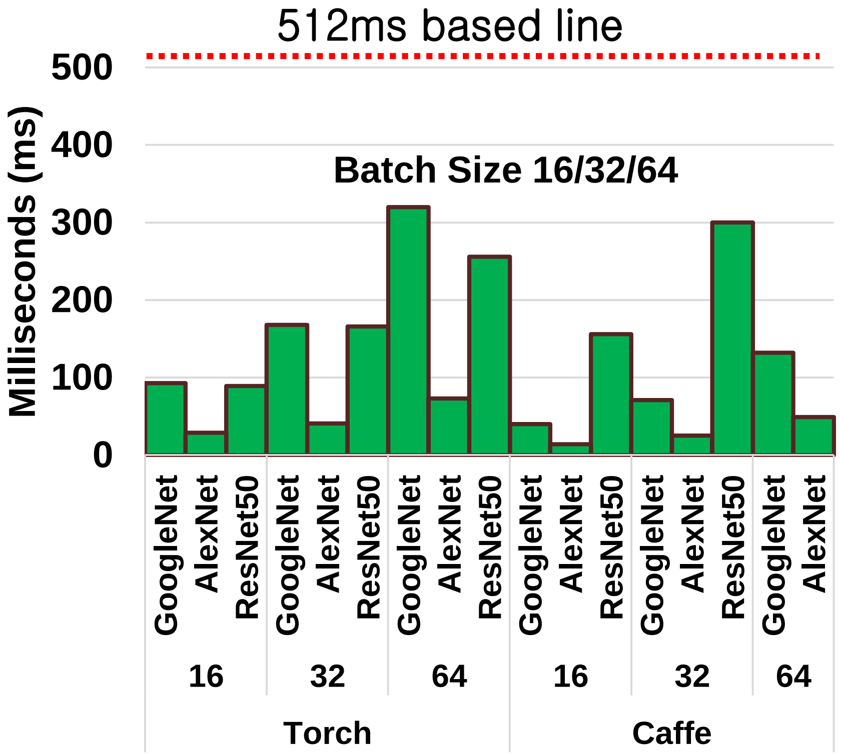
- The Register to control the size of the ‘Error-Zone’
  - – The size of the ‘Error-Zone’ is reconfigurable.
- BER Control
  - – Regulating BER by adjusting the refresh time of the ‘Error Zone’
- Hardwired ECC (optional)
  - – Software-level ECC can be used



Validation - Setup



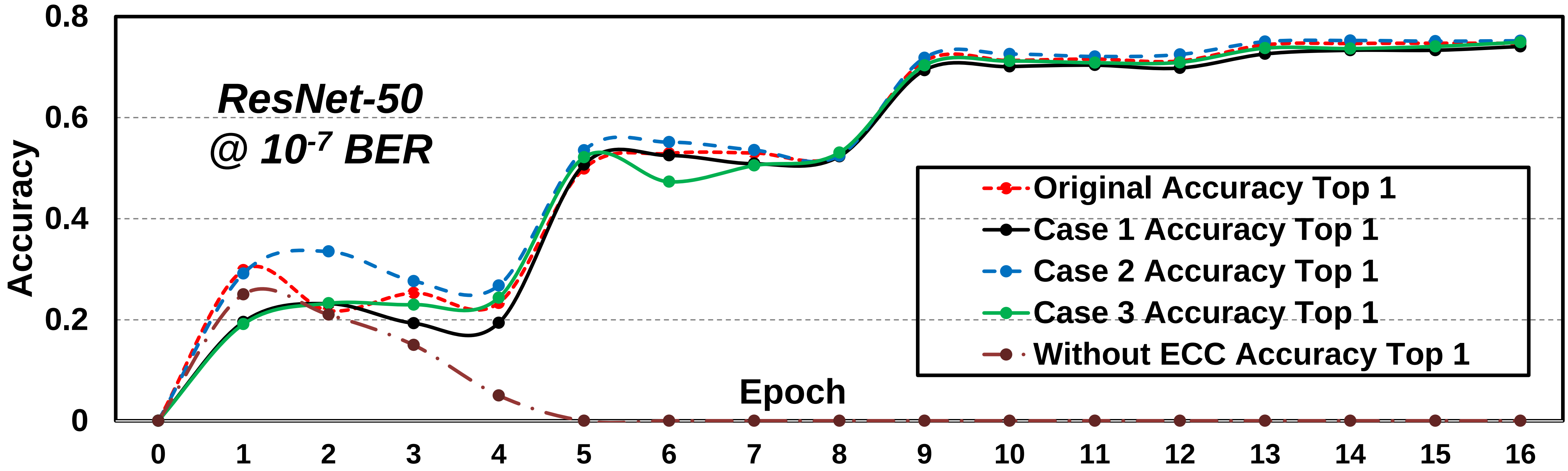
Retention Errors vs. Refresh Time (Measurement Results)



- The working temperature is less than 60ºC in the data-center (**)
  - – BER ~ 10^{-7} at 60°C for two major vendors (DDR3)
  - – Inject 10^{-7} BER to weights and activations during forward/backward phase
- (Forward time+ backward time)/iteration << 512ms (Measurement on Torch and Caffe)
- JEDEC requires 10^{-15} BER @ 32/64ms in the normal working temperature

(**)Donghyuk Lee, Adaptive-latency DRAM: Optimizing DRAM timing for the common-case. HPCA 2015

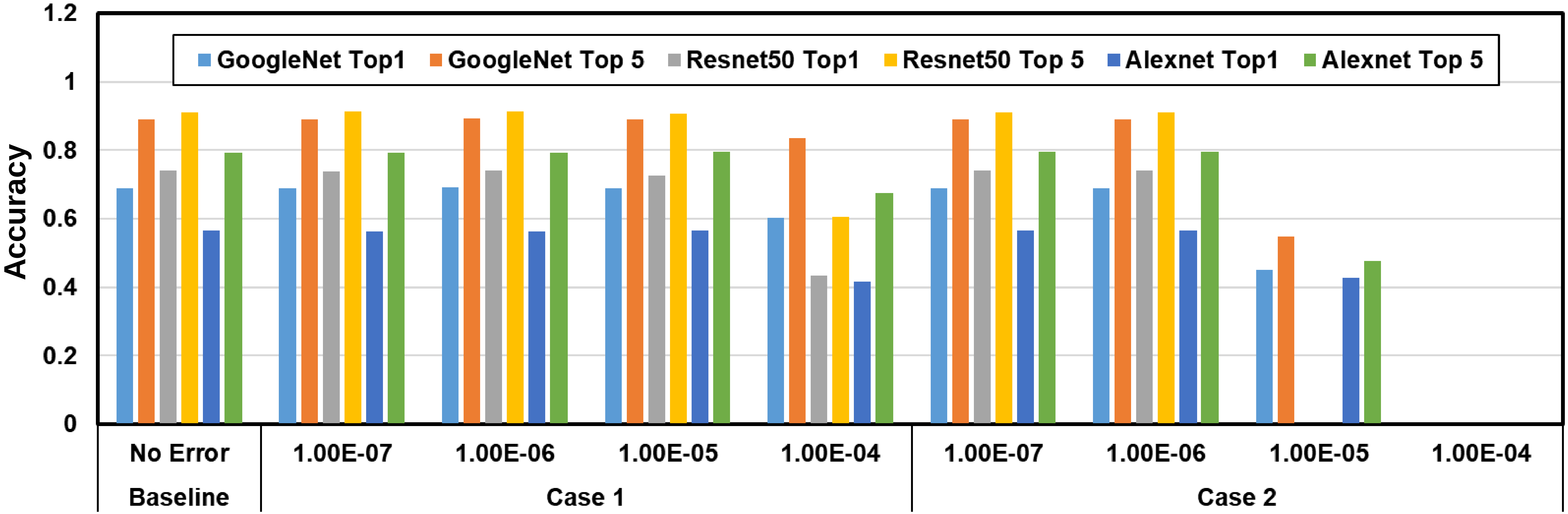
Validation (Cont.) – Training process



- Works accurately and more efficiently than JEDEC standard, which is 10^{-15} BER

CNNs	Last Trained Epoch	Original Acc. Top-1	Case-1 Acc. Top-1	Case-2 Acc. Top-1	Case-3 Acc. Top-1
LeNet	10	99.07	99.12	99.06	99.02
ConvNet	10	75.69	74.66	76.56	75.83
SqueezeNet	7	58.50	58.83	58.34	58.21
GoogleNet	22	70.05	69.55	69.73	70.07

Validation (Cont..) - Inference

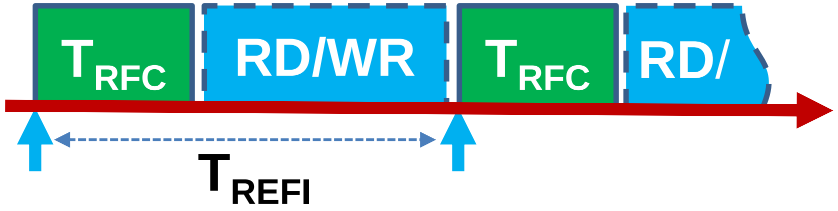


- We figure out the safe BER threshold for the inference of each DNN

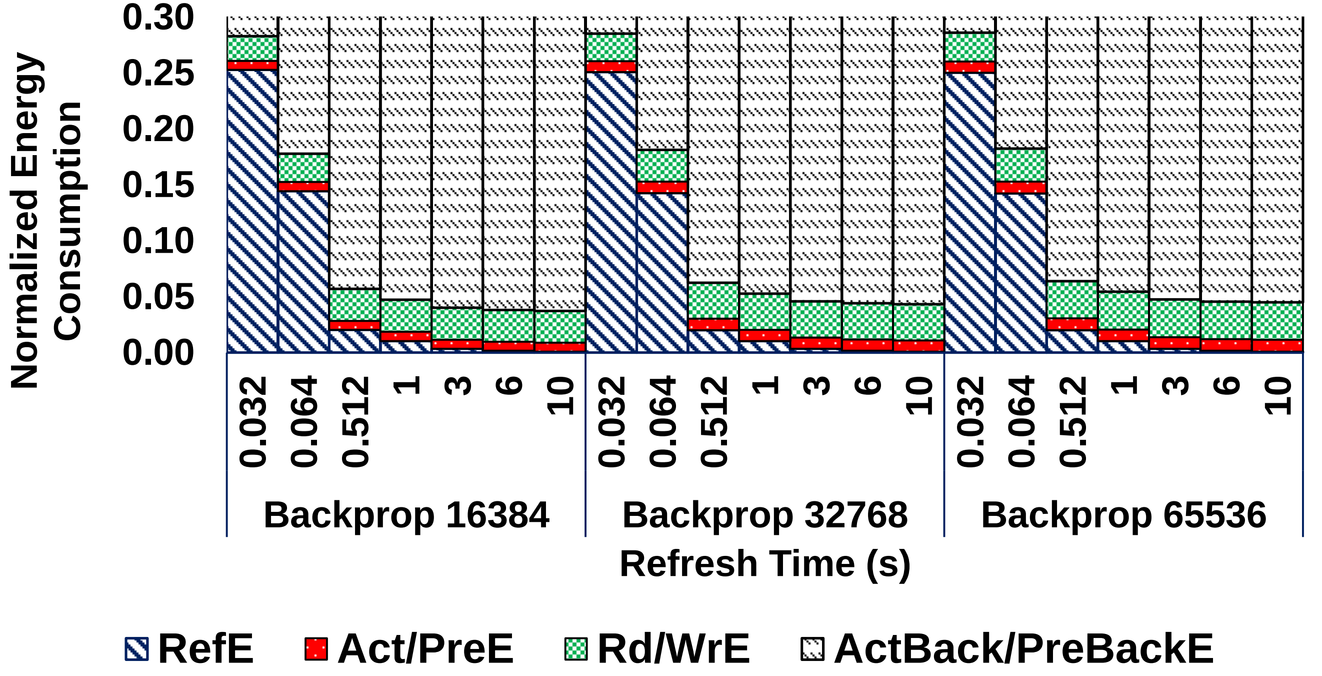
  → The Case-1 is a little robust than The Case-2

  → The Case-3 is almost similar to the Case-2

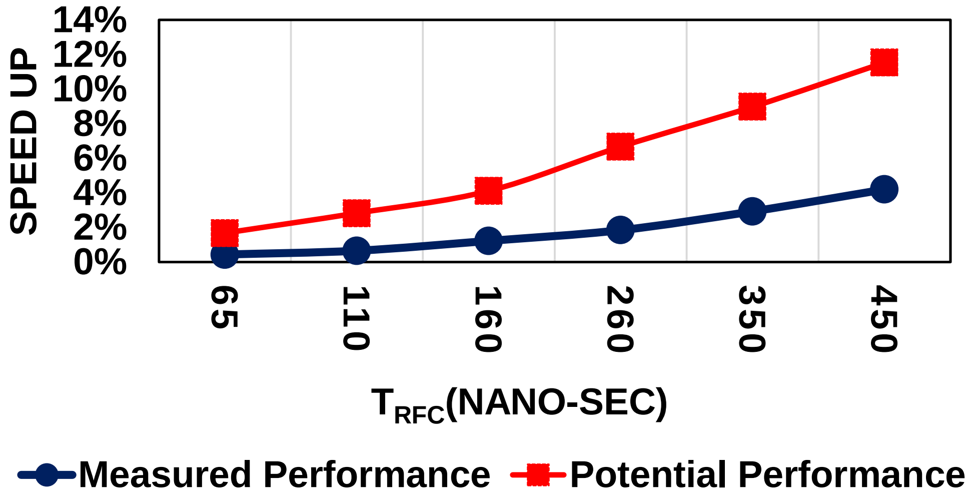
Energy and Performance Simulation Results



tRFC (ns)	65	110	160	260	350	450
tREFI (ns)	3900/7800 (GDDR5/DDR4)					



- Back-propagation test results with different tRFC/tREFI (Benchmark Set: Rodinia from IISWC 2009)
  - Hybrid CPU-GPU Platform (GPGPU + GEM5)
  - Refresh time is 512ms (10^{-7} BER), DRAM energy reduces:
    - 23% on graphic memories
    - 12% on main memories
  - Performance improves 0.43~4.12%



Conclusion

- Present the stretchable DRAM Refresh controller to control the BER according to
  - –Temperature
  - –User desired BER
- The proposed Error Correction Schemes can safeguard the important bit
- With only 512ms (10^{-7} BER), our proposed system can potentially help:
  - –Speed up the training process up to 4.12%
  - –Reduce 12% and 23% DRAM energy in main and graphic memories

Acknowledgment & Thank you

- Supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT1602-03

- Any question?