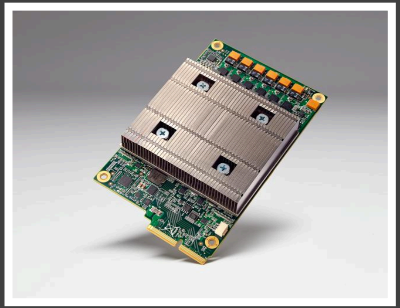
# Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications

Guanpeng(Justin) Li, Siva Kumar Sastry Hari, Michael Sullivan, Tim Tsai,

Karthik Pattabiraman, Joel Emer, Stephen W. Keckler

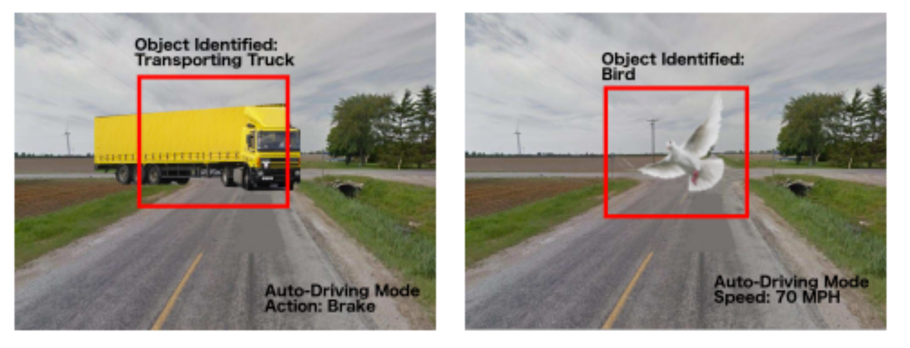
# Motivation

- Neural network applications are widely deployed nowadays
  - Deep learning neural network (DNN): Robots, satellites, cars etc
  - Safety-critical: Detecting cars and pedestrians in self-driving cars
- DNN accelerators are crucial
  - High throughput for real-time inferencing
  - Nvidia NVDLA and Google TPU

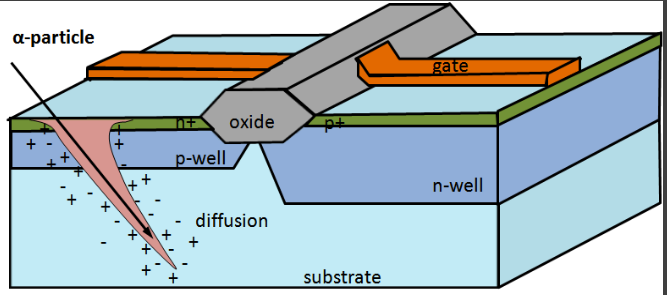


# Soft Error

- Transient hardware error
  - Caused by high-energy particles
  - Random single bit-flip
- Silent Data Corruptions (SDCs)
  - Results in wrong prediction of DNN application
- Safety standard requires low SoC FIT for cars
  - ISO26262




Observed SDC



# Current Solutions

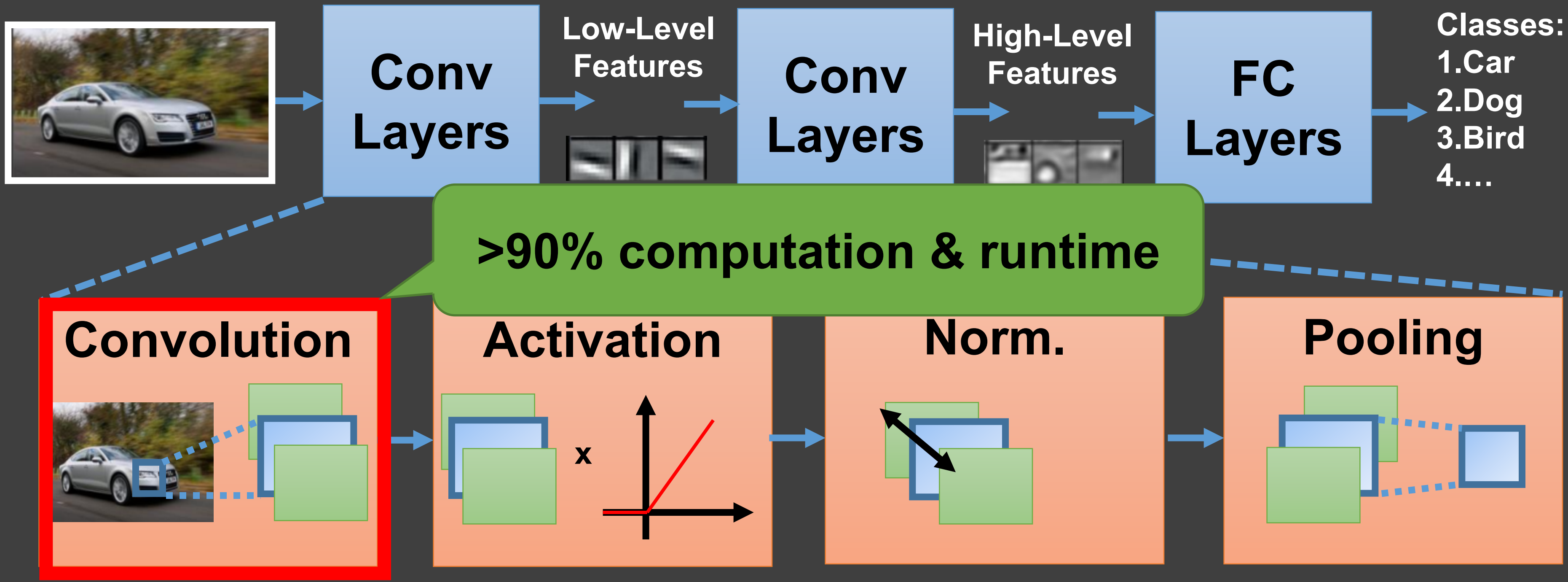
- Traditionally
  - Triple Modular Redundancy (TMR) for execution units
  - Error Correction Code (ECC) for DRAMs
- Other protection techniques
  - DNN-algorithm agnostic
  - Generic micro-arch

**Non-optimal for DNN applications & accelerators**

# Goal

- Understand error propagation in DNN applications & accelerators
  - Quantification
  - Characterization
- Based on the insights, mitigate SDC:
  - Efficient way to detect errors

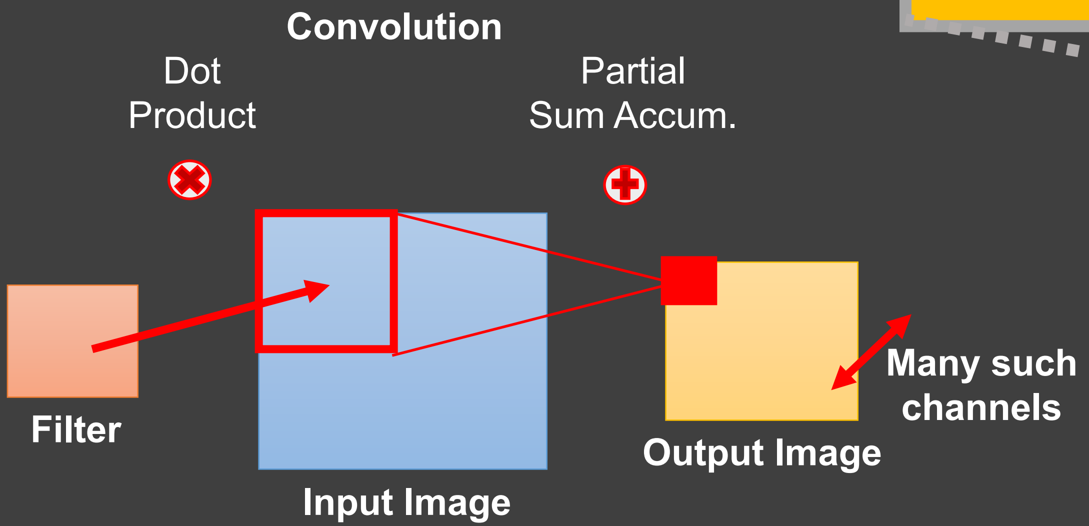
# DNN Explained

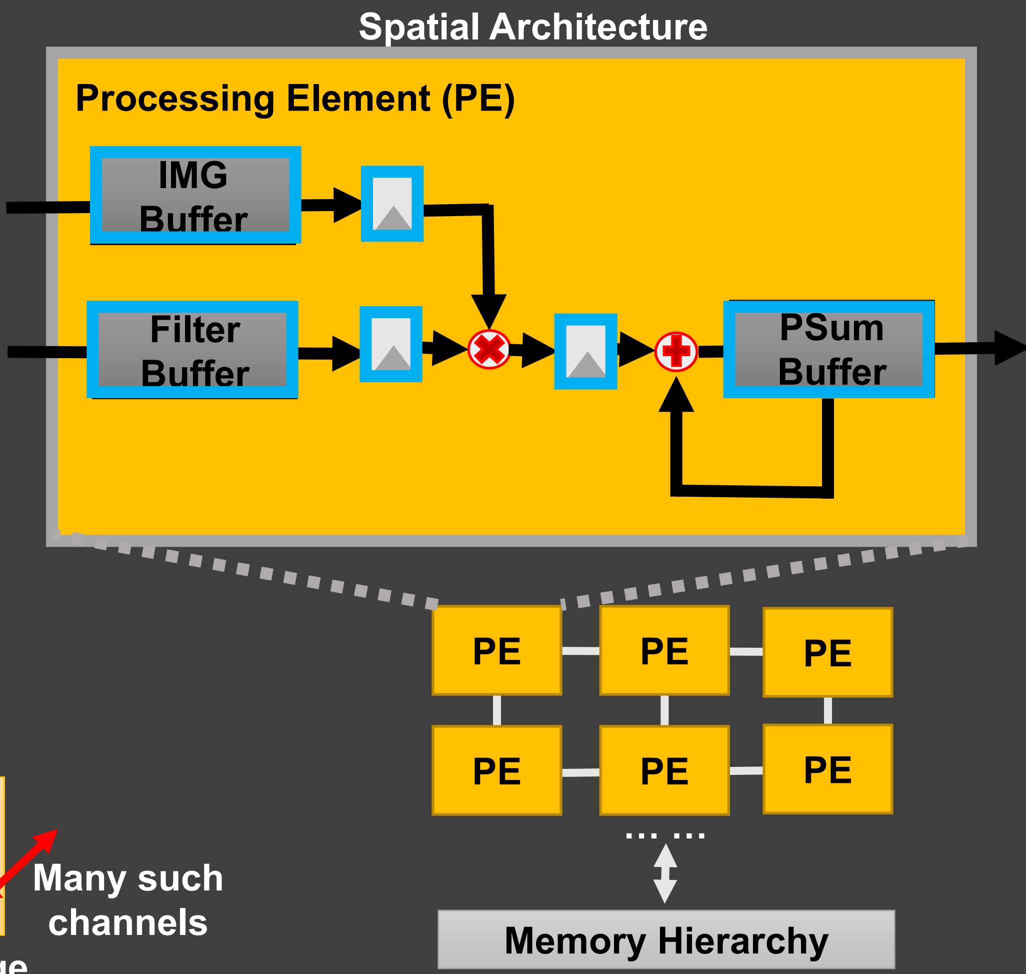



DNN for image classifications, also called CNN

# DNN Accelerator & Fault Model

**Fault Model:**
- **Latch Faults**
- **Buffer Faults**



# Experimental Setup

- Fault Injection
  - Map arch. component to C code in Tiny-CNN
  - 3,000 random faults per component per layer (error bar: 0.11%~0.34%)
  - 1 fault injected per run
  - Popular pre-trained DNNs with ImageNet / CIFAR-10

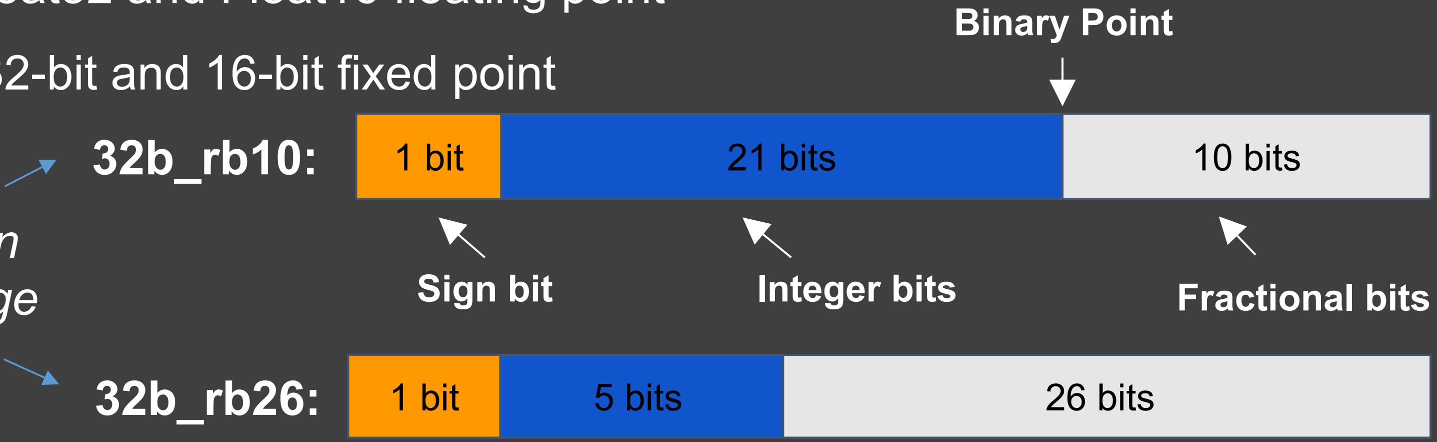
```
// Simulation in Tiny-CNN

function feed_forward(){
        …
        weight = inject_fault (weight);
        multiply = weight * img;
        …
}
```

# Experiment Setup

- Silent Data Corruption (SDC)
  - Mismatch between winner of fault-free execution
- Generic DNN Accelerator & Eyeriss
  - FP: Float32 and Float16 floating point
  - FxP: 32-bit and 16-bit fixed point

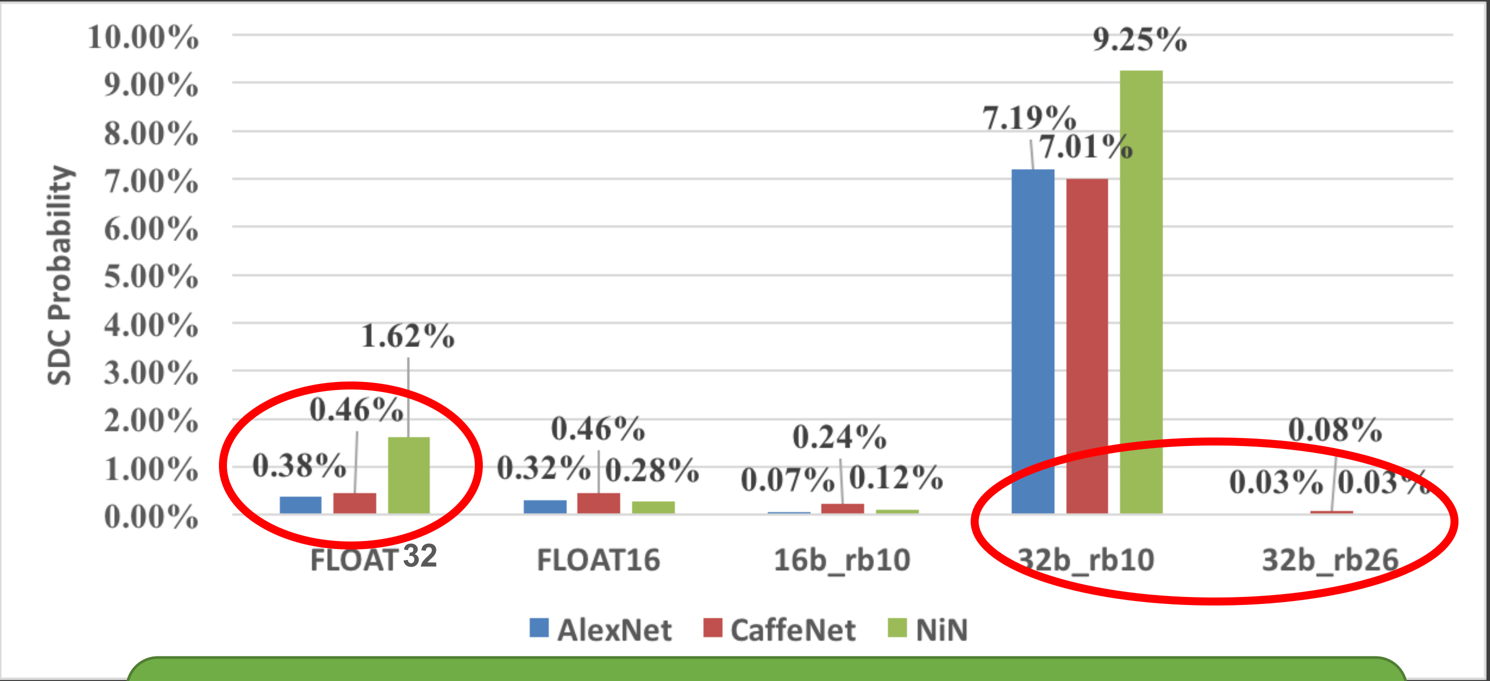
*Trade off between precision and range*



# Research Questions (RQs)

- RQ1: What are SDCs in different DNNs using different data types ?
- RQ2: Which bits are sensitive to SDCs in different data types ?
- RQ3: How do errors affect values that result in SDCs ?
- RQ4: How does error propagate layer by layer ?
- RQ5: What is the SDC sensitivity in different data reuse buffers ?

# SDC in DNNs




**1.SDC probabilities are different in different DNNs**

**2.SDC probabilities vary a lot using different data types**

# Research Questions (RQs)

- RQ1: What are SDCs in different DNNs using different data types ?
- RQ2: Which bits are sensitive to SDCs in different data types ?
- RQ3: How do errors affect values that result in SDCs ?
- RQ4: How does error propagate layer by layer ?
- RQ5: What is the SDC sensitivity in different data reuse buffers ?

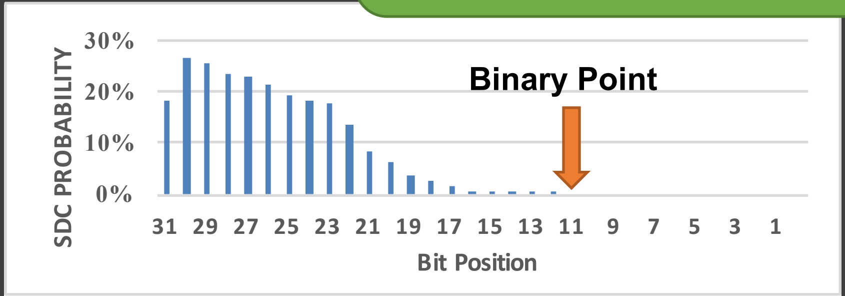
# Bit Positions

**Error Mitigation:**

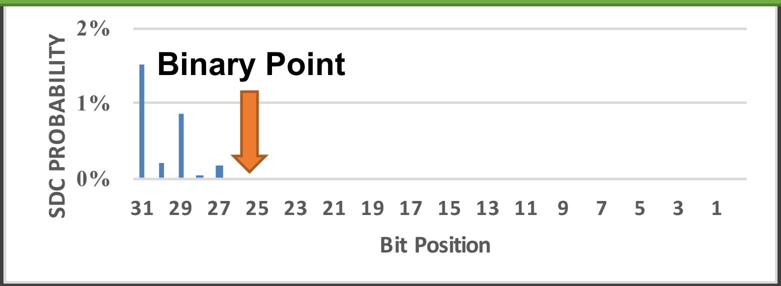
- **Restrain dynamic value range in data type**
- **Selective latch hardening**

## FxP data types:

1. **High-order bits are vulnerable**
2. **Larger dynamic value range leads to higher SDC probability**
3. **Sliding binary point controls the amount of vulnerable bits**



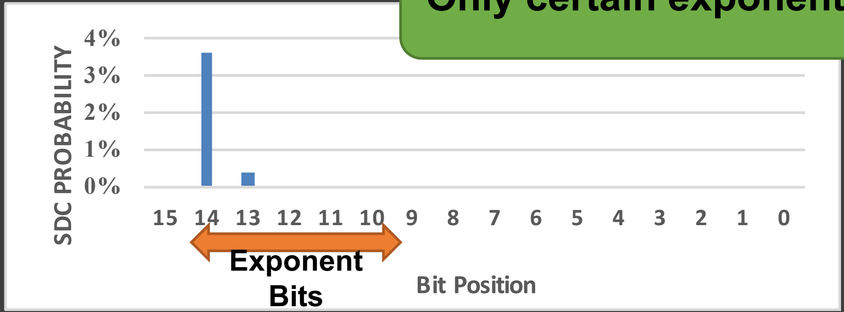

32b_rb10



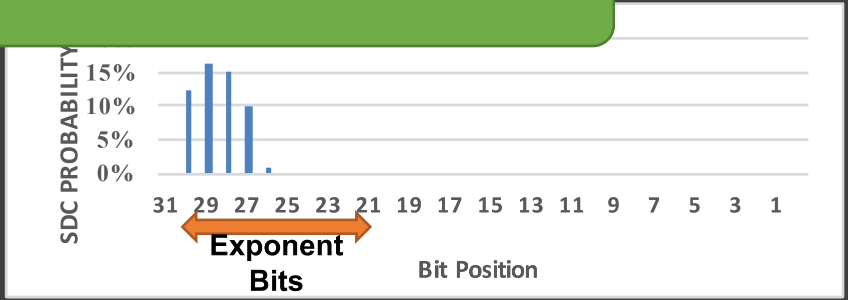

32b_rb26

## FP data types:

**Only certain exponent bits are vulnerable to SDCs**




FLOAT16




FLOAT32

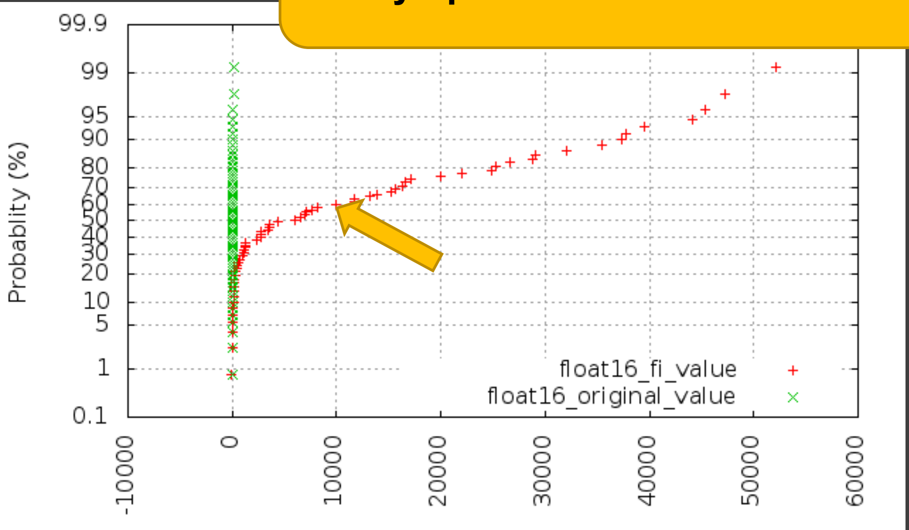
# Research Questions (RQs)

- RQ1: What are SDCs in different DNNs using different data types ?
- RQ2: Which bits are sensitive to SDCs in different data types ?
- RQ3: How do errors affect values that result in SDCs ?
- RQ4: How does error propagate layer by layer ?
- RQ5: What is the SDC sensitivity in different data reuse buffers ?

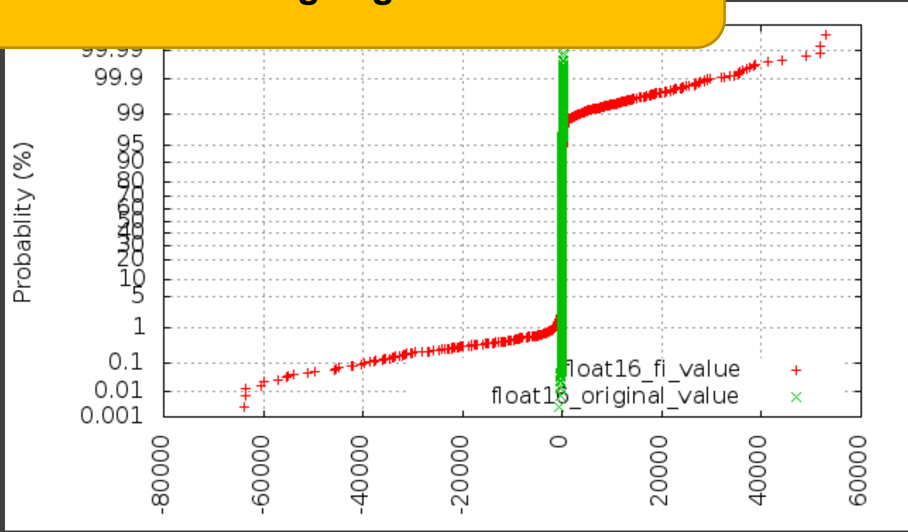
# Value Changes under Errors

**Error Mitigation**

- **Symptom-based error detector: Check value range against errors**




SDC



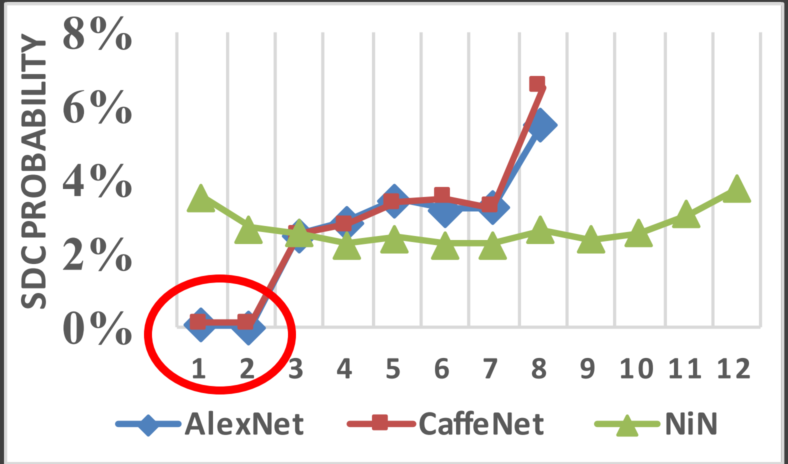

Benign

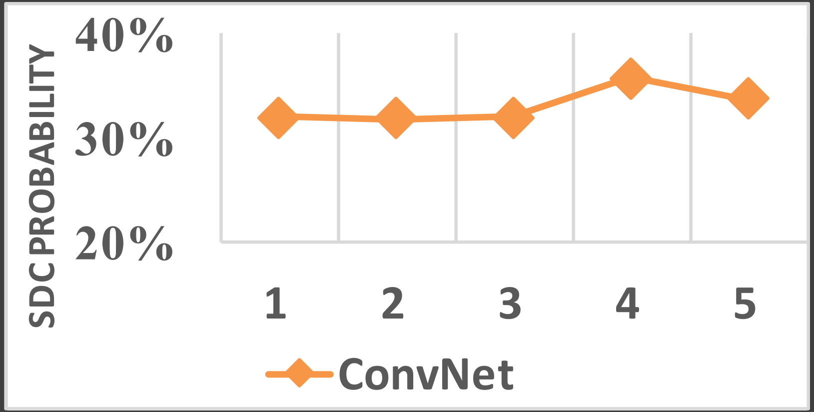
**If the numeric value is modified to be a large positive one by faults, it likely causes SDC**

# Research Questions (RQs)

- RQ1: What are SDCs in different DNNs using different data types ?
- RQ2: Which bits are sensitive to SDCs in different data types ?
- RQ3: How do errors affect values that result in SDCs ?
- RQ4: How does error propagate layer by layer ?
- RQ5: What is the SDC sensitivity in different data reuse buffers ?

# SDC in Different Layers



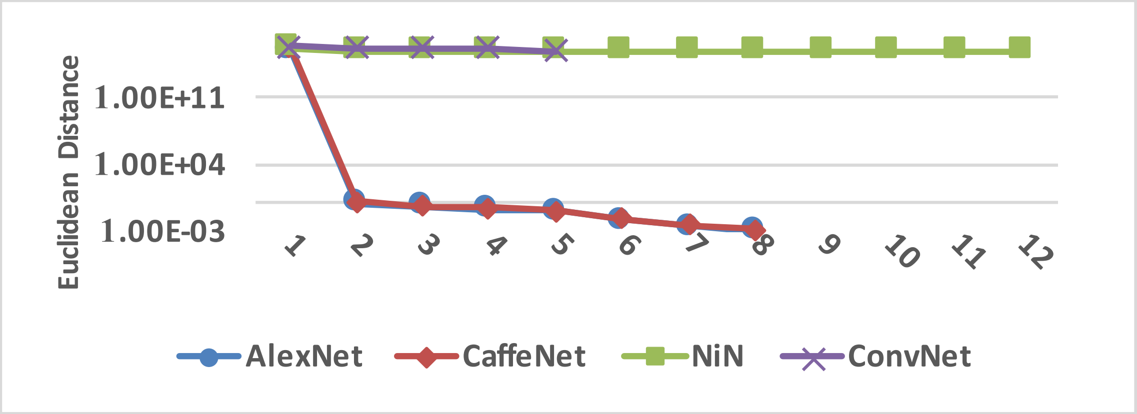

1.SDC probability increases by layers

2.Layer 1&2 have lower SDC probabilities in AlexNet and CaffeNet

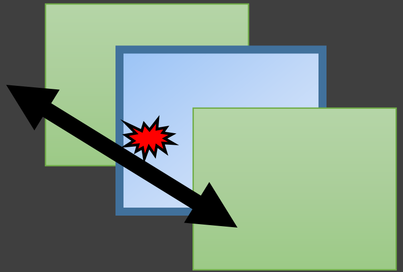
# Normalization Layer

**Error Mitigation**

- **Error detectors can be placed after LRN**




** Faults injected at first layer only*



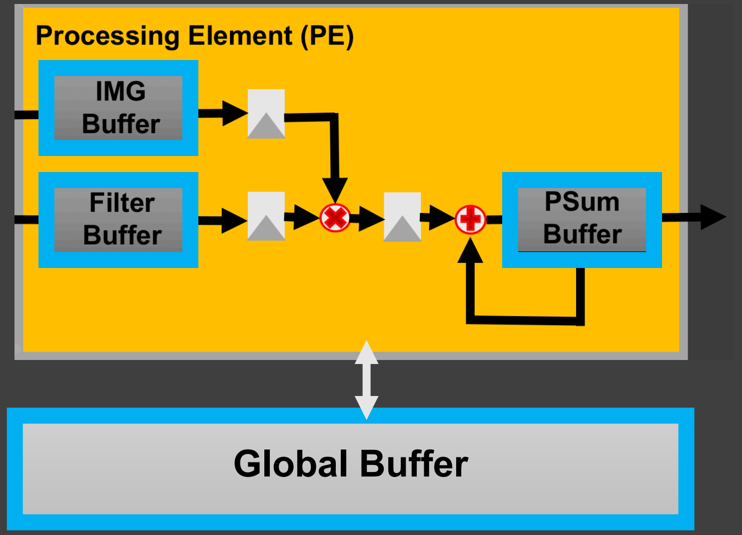
1. **Euclidean distance decreases by layers in AlexNet and CaffeNet**
2. **Local Response Normalization (LRN) in AlexNet & CaffeNet in Layer 1&2 re-normalizes values back towards normal range**

# Research Questions (RQs)

- RQ1: What are SDCs in different DNNs using different data types ?
- RQ2: Which bits are sensitive to SDCs in different data types ?
- RQ3: How do errors affect values that result in SDCs ?
- RQ4: How does error propagate layer by layer ?
- RQ5: What is the SDC sensitivity in different data reuse buffers ?

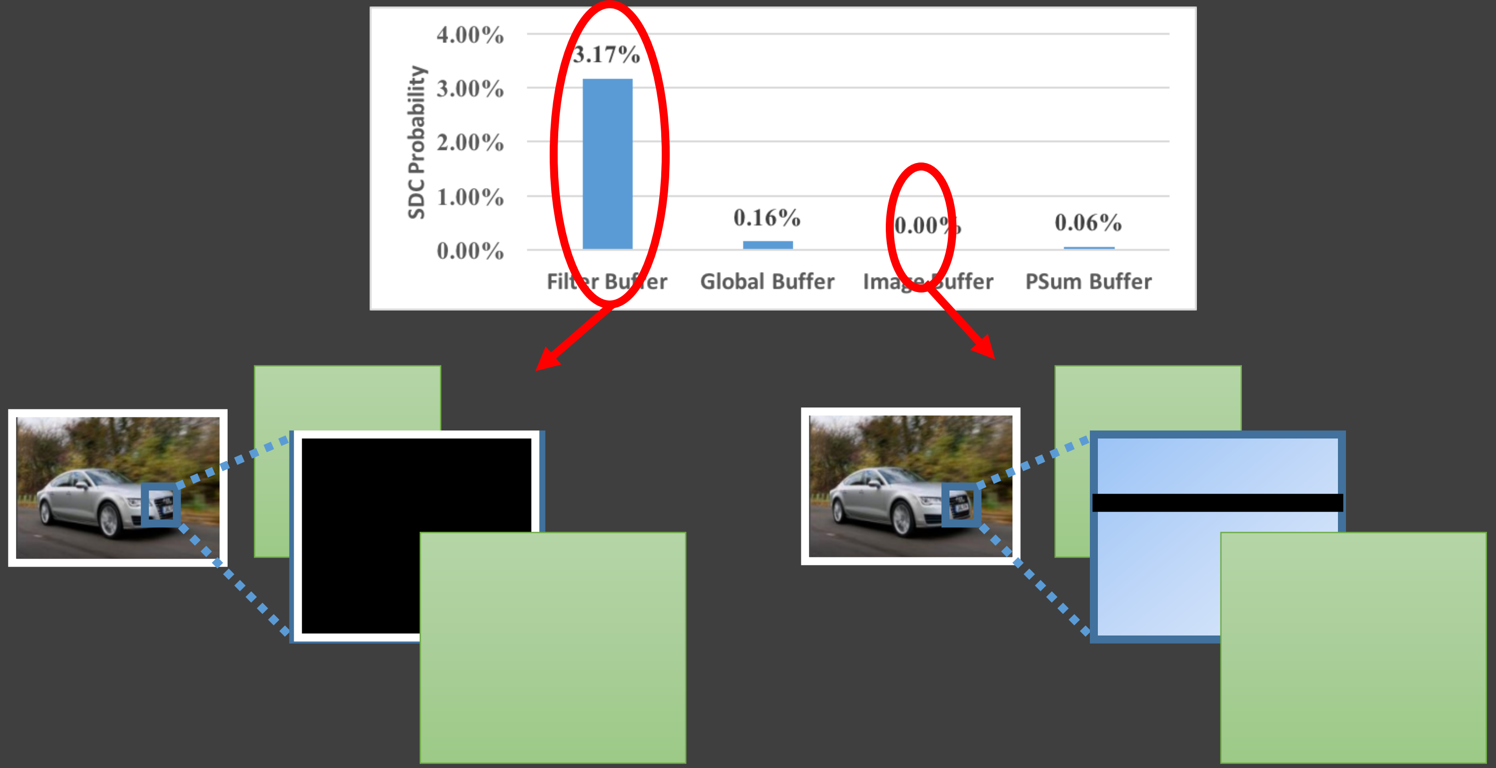
# Buffer Faults

- Eyeriss: Row-Stationary Dataflow
  - Global Buffer: Reuse image data
  - IMG Buffer: Reuse image data of a row
  - Filter Buffer: Reuse filter weights
  - PSum: Reuse partial sum results



# Buffer Faults

* *AlexNet, 16b_rb10 on Eyeriss*



# Summary

- Characterized error propagation which depends on data types, layers, values topologies etc
- Mitigation techniques including value range checker and selective latch hardening

Guanpeng(Justin) Li, PhD at University of British Columbia

http://www.ece.ubc.ca/~gpli

gpli@ece.ubc.ca