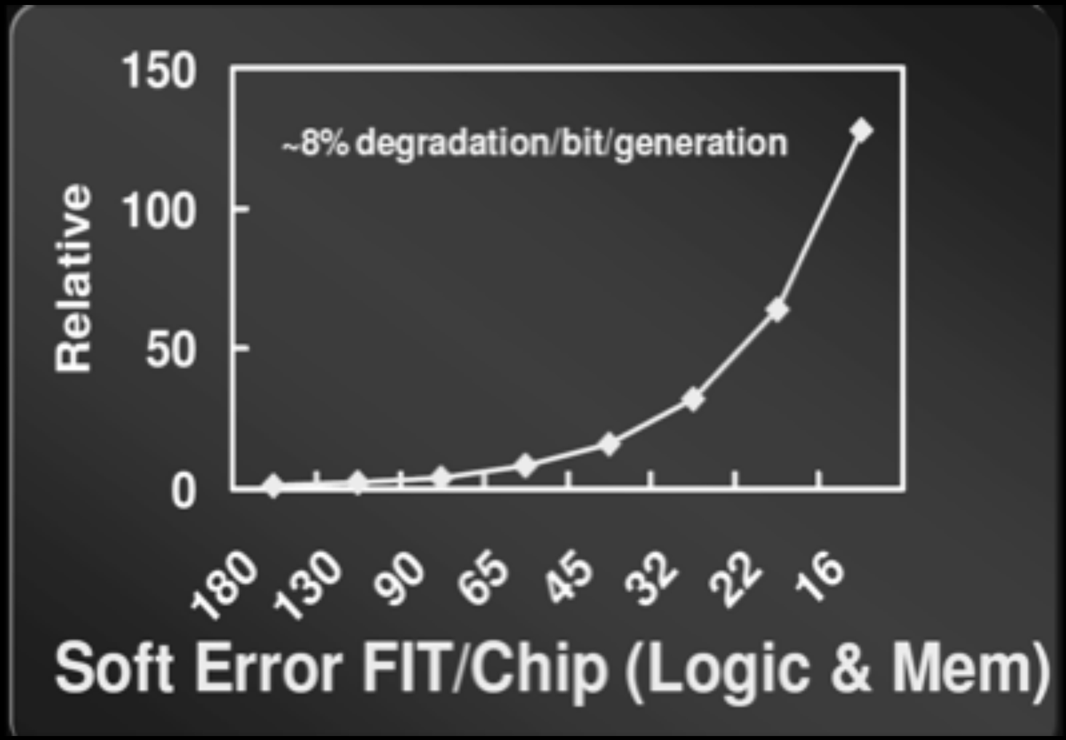
Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications

Guanpeng (Justin) Li,
Karthik Pattabiraman

Siva Kumar Sastry Hari,
Michael Sullivan, Tim Tsai,
Joel Emer, Stephen Keckler

Soft Error Problem

- Soft errors are increasing in computer systems



Source: Shekar Borkar (Intel) - Stanford talk

DNNs

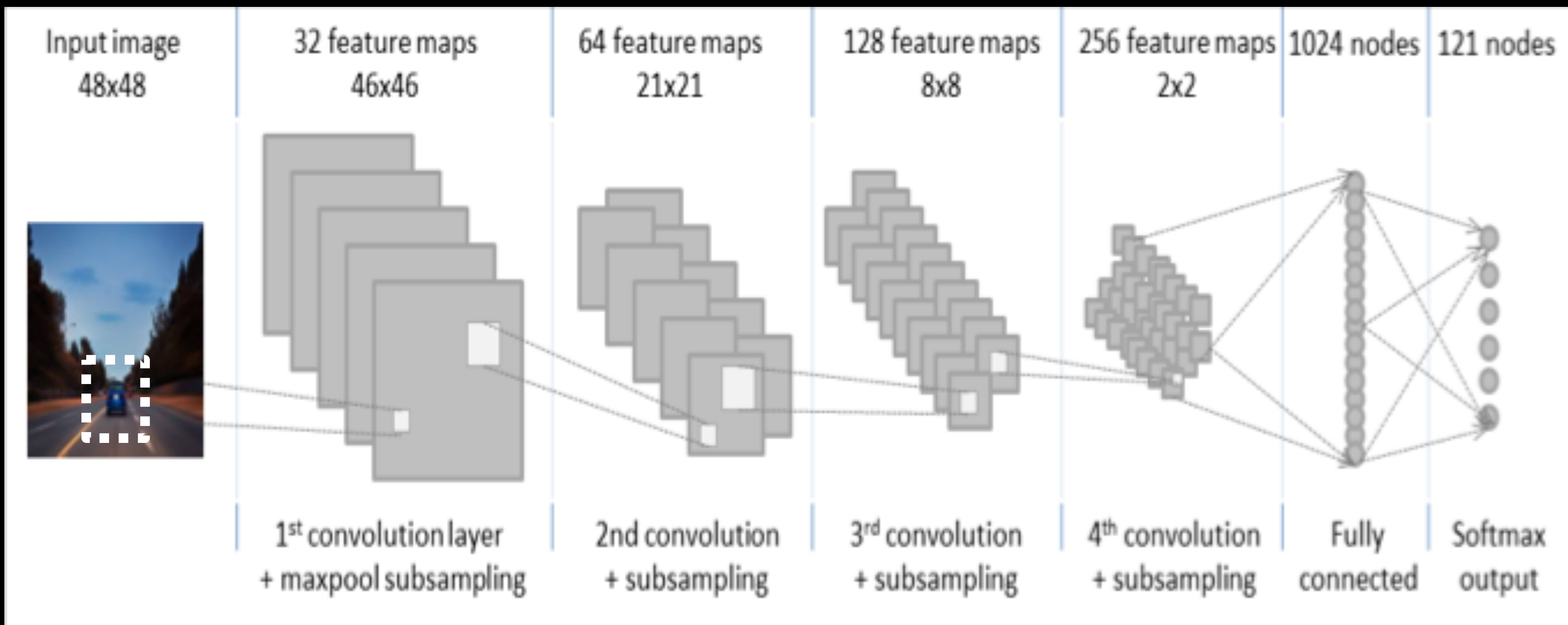
- **DNN applications are widely deployed in safety critical applications**
- **Specialized accelerators for real-time processing (e.g., Nvidia NVDLA and Google TPU)**
- **Silent Data Corruptions (SDCs)**
  - Results in wrong prediction of DNN application
  - Safety standard requires SoC FIT<10 overall (ISO 26262)

Goals

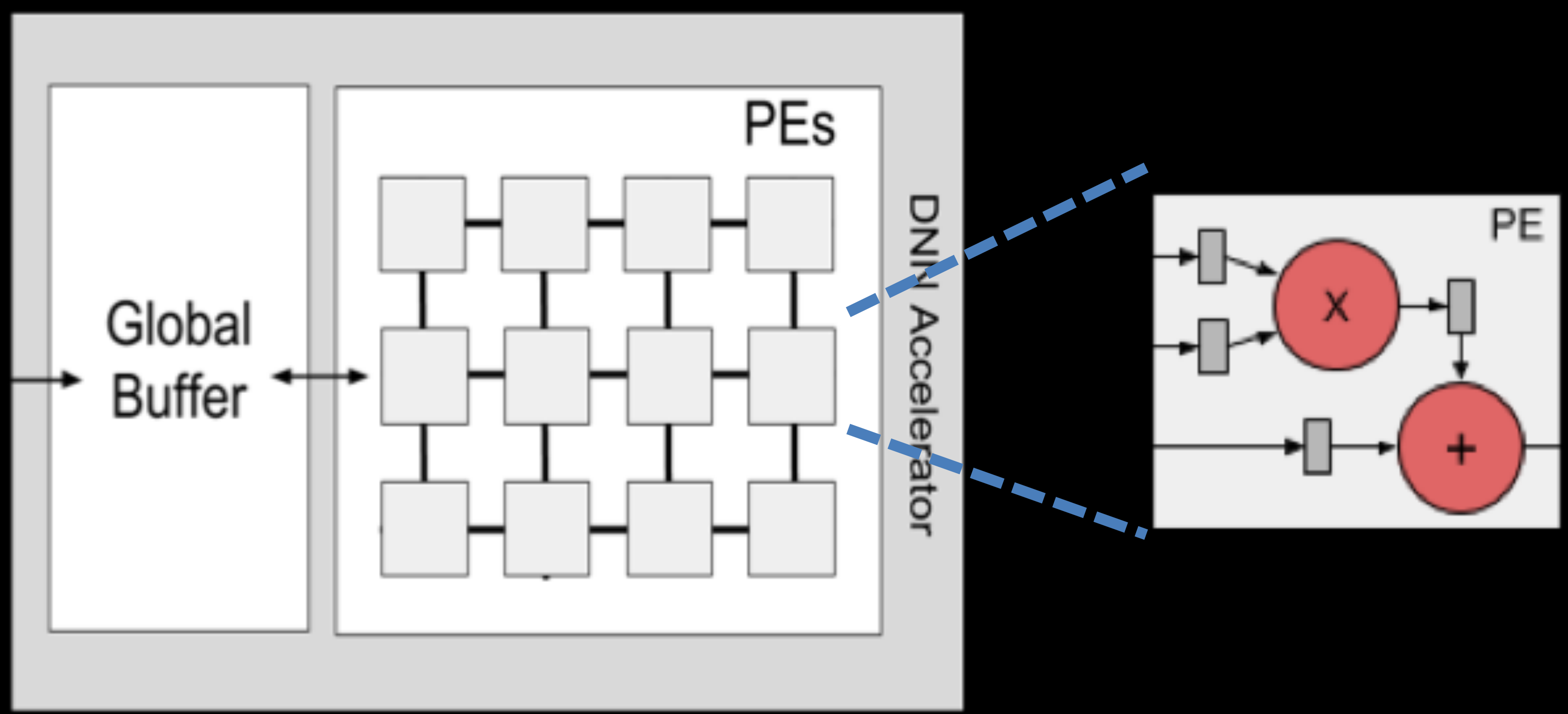
- **Understand error propagation in DNN accelerators - fault injection**
  - Quantification
  - Characterization

- **Based on the insights, mitigate failures:**
  - Efficient way to detect errors
  - Hardware: Selective duplication
  - Software: Symptom-based detection

Deep Neural Network (DNN)



DNN Accelerator Architecture (e.g., Eyeriss – MIT)



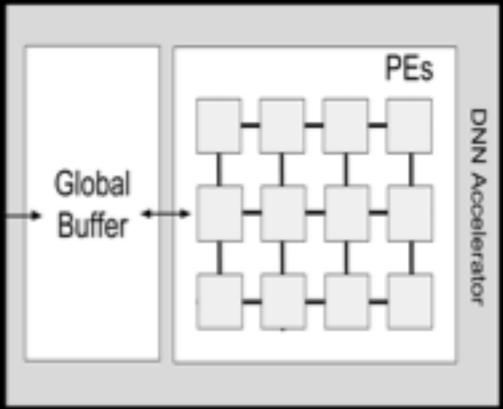
Fault Injection Study: Setup

- **Fault Injection**
  - 3,000 random faults per each latch in each layer

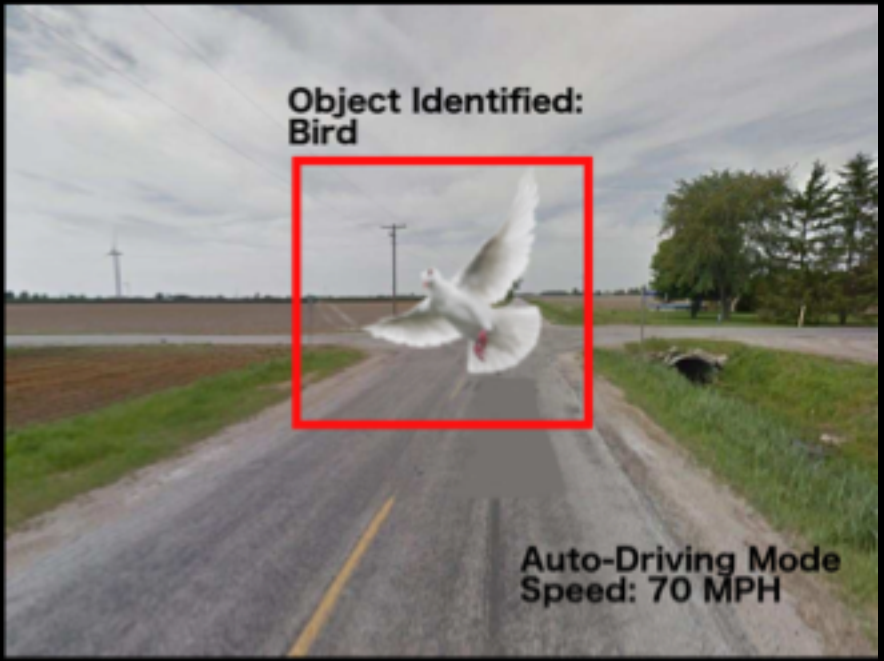
- **Simulator**
  - DNN simulation in Tiny-CNN in C
  - Fault injections at C line code

```
...
foreach layer:
  ...
  foreach weight:
    ...
    foreach input:
      ...
      R_L2.2 = inject_fault(R_L2.2)
      R_L3 = R_L2.2 + R_L5
      ...
...
```

- **Fault Model**
  - Transient single bit-flip
  - Execution Units: Latches
  - Storage: buffer SRAM, scratch pad, REG



Silent Data Corruption (SDC) Consequences



A single bit-flip error → misclassification of image by the DNN

SDC Types

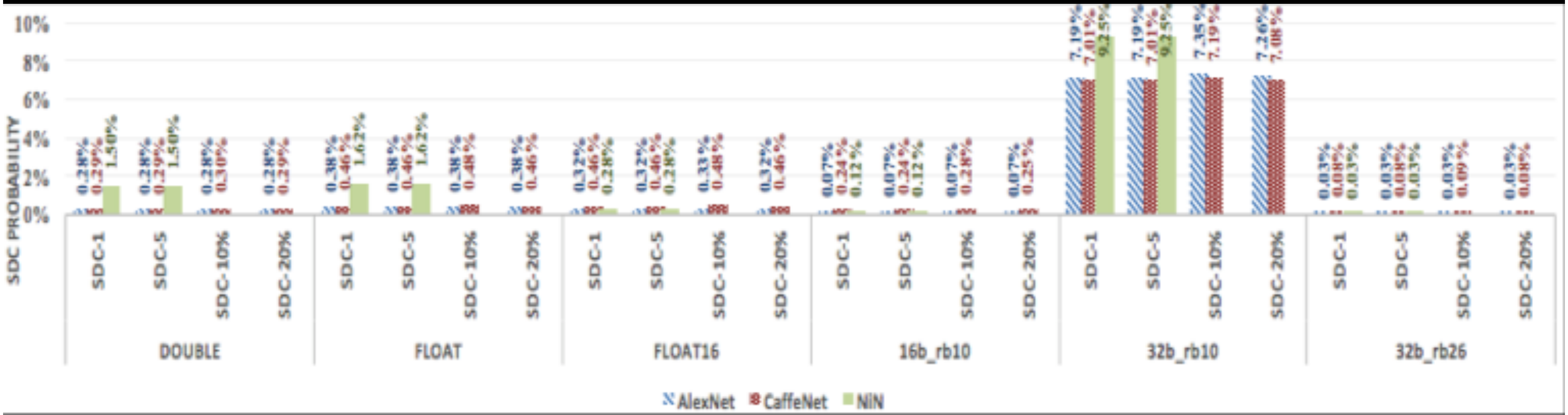
SDC1:
- Mismatch between winners in faulty and fault-free execution

SDC5:
- Winner is not in top 5 predictions in the faulty execution

SDC10%:
- Confidence of the winner drops more than 10%

SDC20%:
- Confidence of the winner drops more than 20%

Finding 1: SDC in DNNs

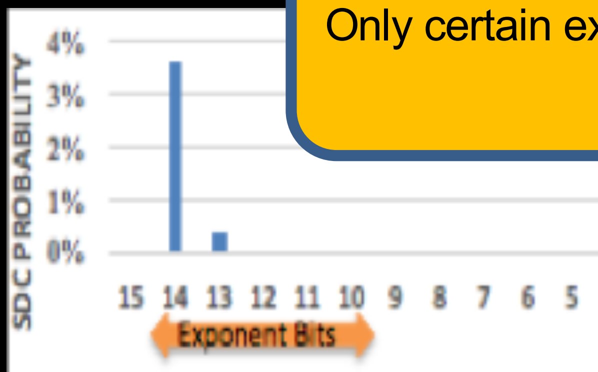


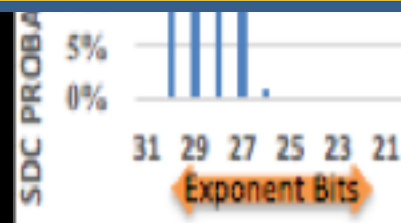
1. All SDCs defined have similar SDC probabilities
2. SDC probabilities are different in different DNNs
3. SDC probabilities vary a lot using different data types

Finding 2: Bit Sensitivity

FP data types:

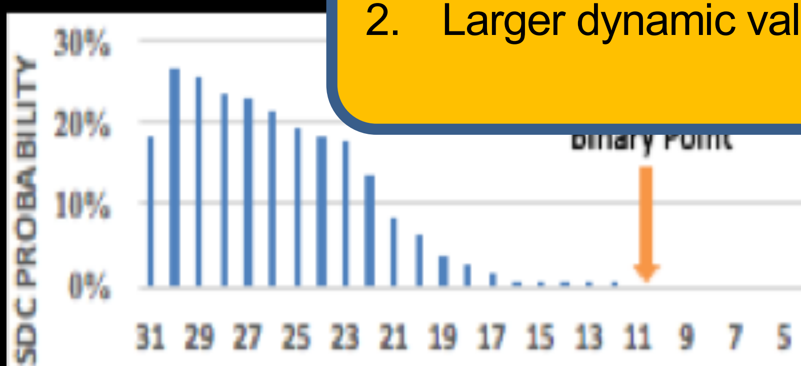
Only certain exponent bits are vulnerable to SDCs

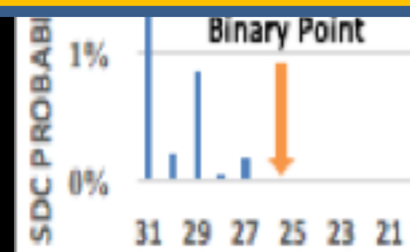




FxP data types:

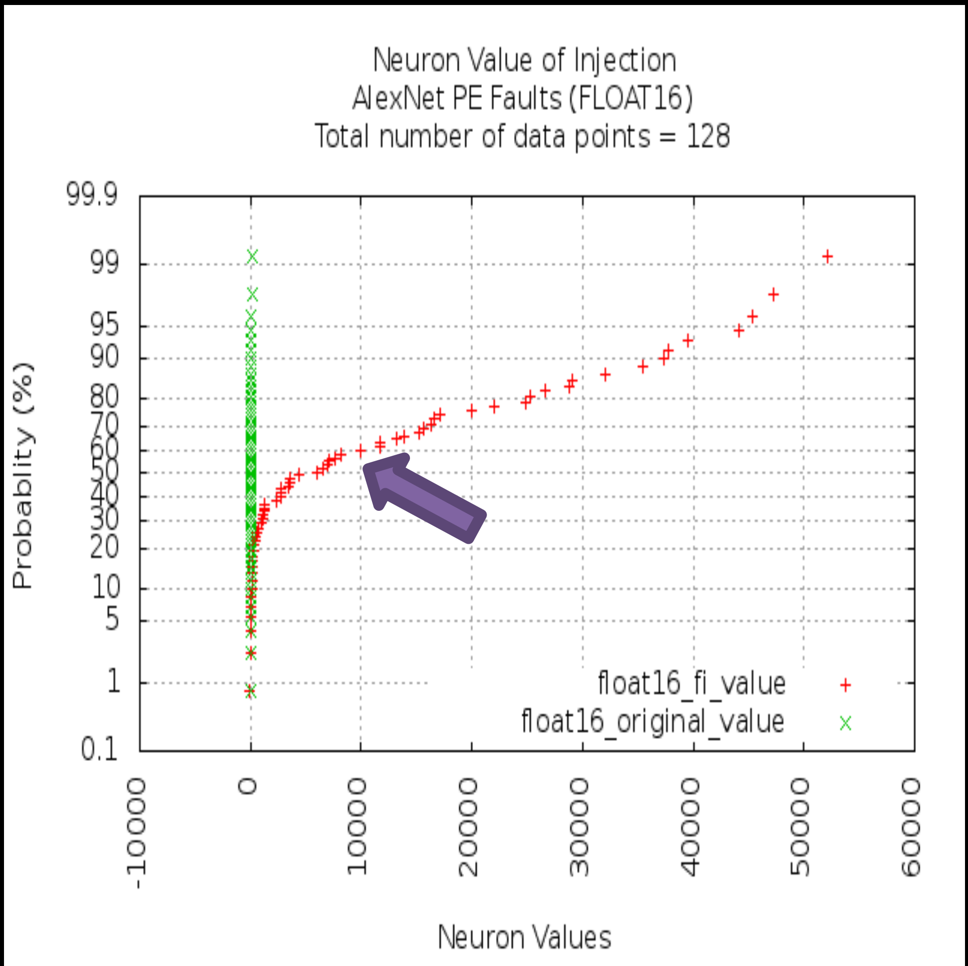
1. High-order bits are vulnerable
2. Larger dynamic value range allows more vulnerable bits



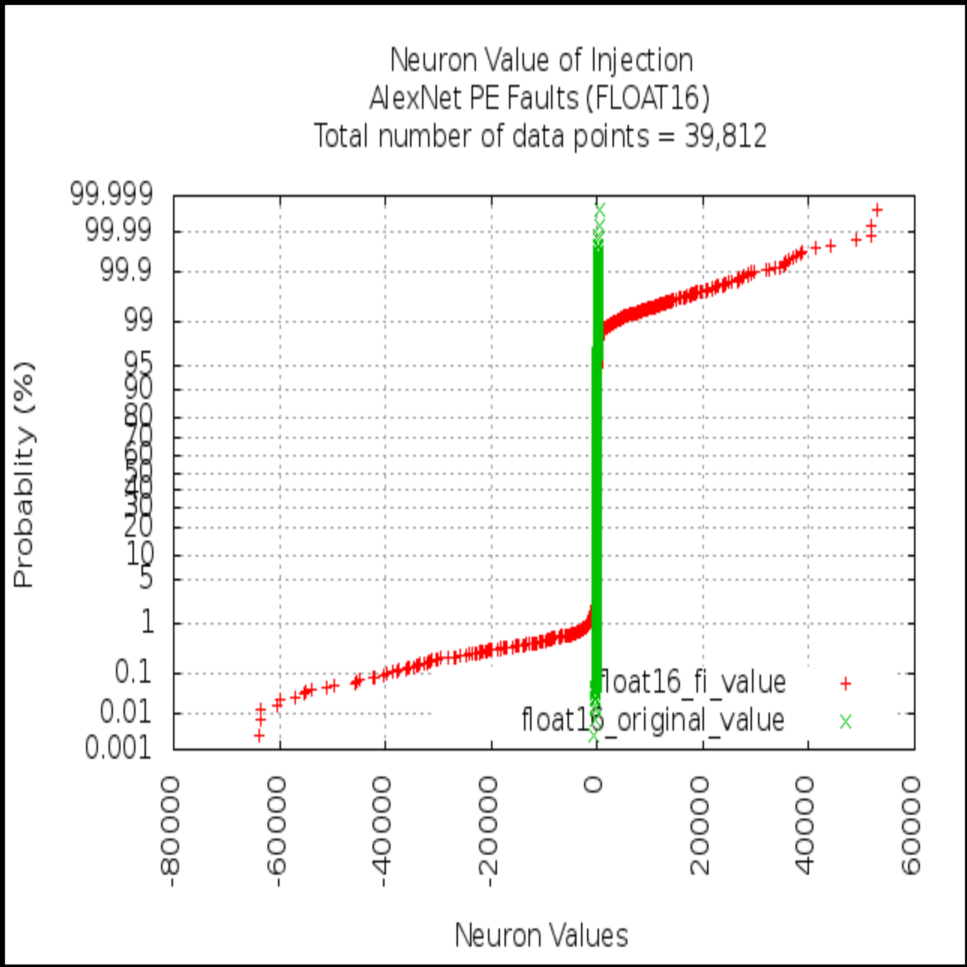


Finding 3: Value Changes

AlexNet, PE Errors, Float16



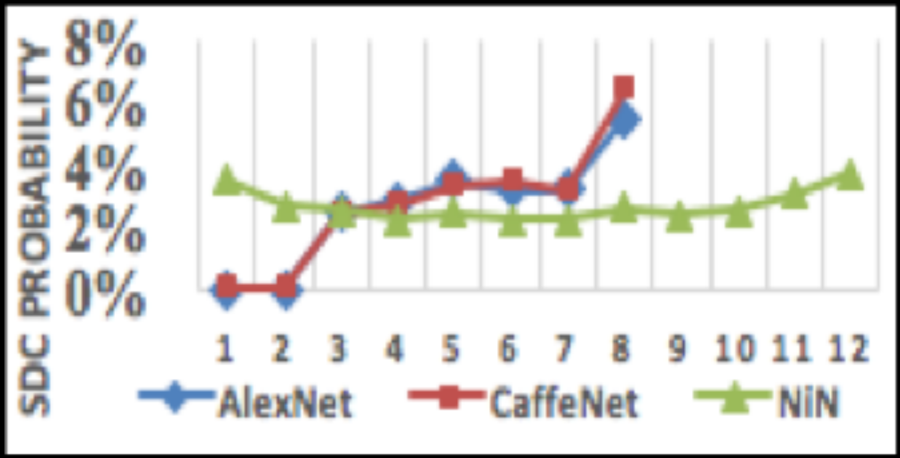
SDC

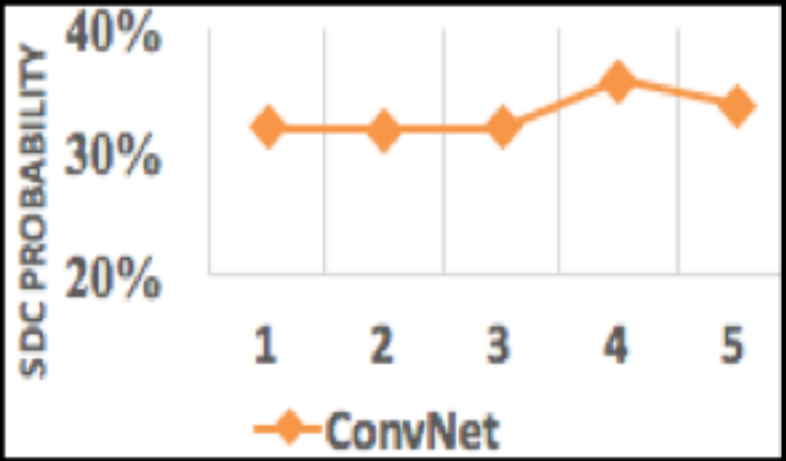


Benign

If a neuron value is changed to be a large value under a fault, it likely causes SDC

Finding 4: Different Layers





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

2.SDC probability increases as layer numbers increase

Mitigation Techniques

- Data type choice (Programmer)
- Symptom-based Error Detection (Software)
- Selective Latch Hardening (Hardware)
- Algorithmic Error Resilience (Ongoing)

Conclusions

Characterized error propagation in DNN accelerators based on data types, layers, value types & topologies

Key Findings:

- Different CNN structures have different resilience
- Higher order bits are more vulnerable to SDCs
- Correct values in each layer are close to zero
- Later layers have higher impact on SDC rates

More details in our SC'17 Paper: "Understanding Error Propagation in Deep-Learning Neural Networks (DNN) Accelerators and Applications"