www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

On the Resilience of RTL NN Accelerators: Fault Characterization & Mitigation

Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman

High Performance Machine Learning Workshop (HPML), 24-Sept-2018, Lyon, France.




Outline

- **Motivation**
  - **Why accelerators for NNs? Why Register-Transfer Level (RTL) model?**
  - **Why to study resilience in NNs Accelerators?**
- Fault Characterization of RTL NN
  - Empirically vulnerability analysis of different components of RTL NN
- Fault Mitigation of RTL NN
  - An efficient technique to mitigate faults
- Summary and Future Works

Motivation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Why Hardware Accelerators for NNs?

- NNs are inherently compute- and power-intensive applications.
- Hardware accelerators i.e., FPGAs and ASICs are commonly used. On the accelerators, NN computations (matrix multiplications) can be performed ***in parallel*** and with ***streaming mode***.
- Register-Transfer Level (RTL) is a hardware design level can be used for both ASICs and FPGAs. It is *accurate-enough* like hardware and *straightforward-enough* like software. Thanks to High-Level Synthesize (HLS) Tools.

Why Resilience in NNs?

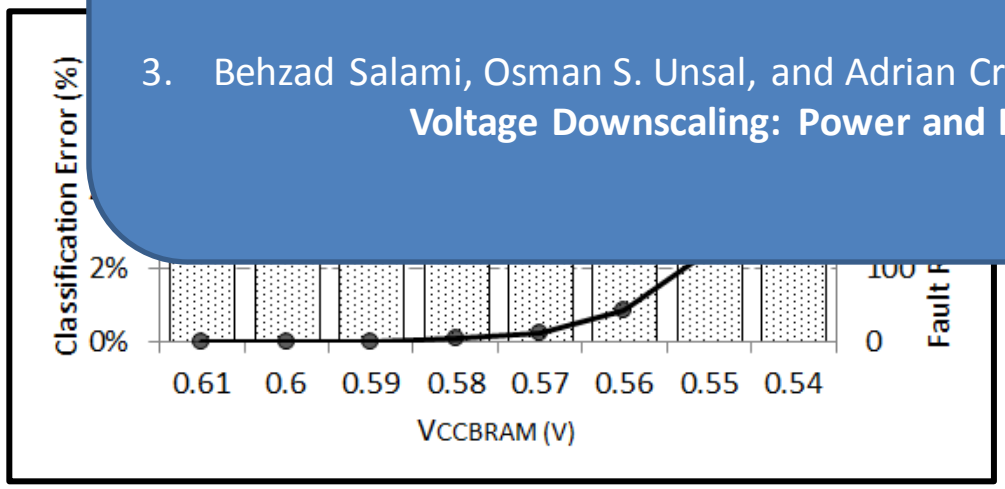
- Continually increasing the fault rate stemming from ***aggressive Undervolting***, manufacturing defects, aging issues, etc, specially in nano-scale technology nodes.
- The accuracy of NN can be significantly affected.

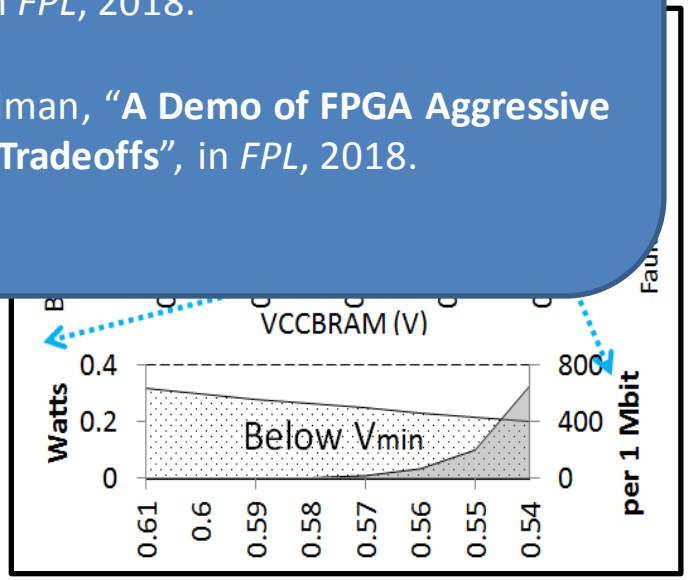
Motivation: Aggressive Undervolting Effects on NNs

Underscaling the supply voltage *below the nominal level* :

- **Power/Energy Efficiency**: Reduces dynamic and static power; quadratically and linearly, respectively.
- **Reliability**: Increases the circuit delay and in turn, causes timing faults.

1. Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman, "**A Comprehensive Evaluation of Supply Voltage Underscaling in FPGA on-chip Memories**", in *Micro51*, 2018.
2. Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman, "**Fault Characterization Through FPGA Undervolting**", in *FPL*, 2018.
3. Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman, "**A Demo of FPGA Aggressive Voltage Downscaling: Power and Reliability Tradeoffs**", in *FPL*, 2018.





Aggressive Undervolting below the voltage guardband

Outline

- Motivation
  - Why accelerators for NNs? Why Register-Transfer Level (RTL) model?
  - Why to study resilience in NNs Accelerators?

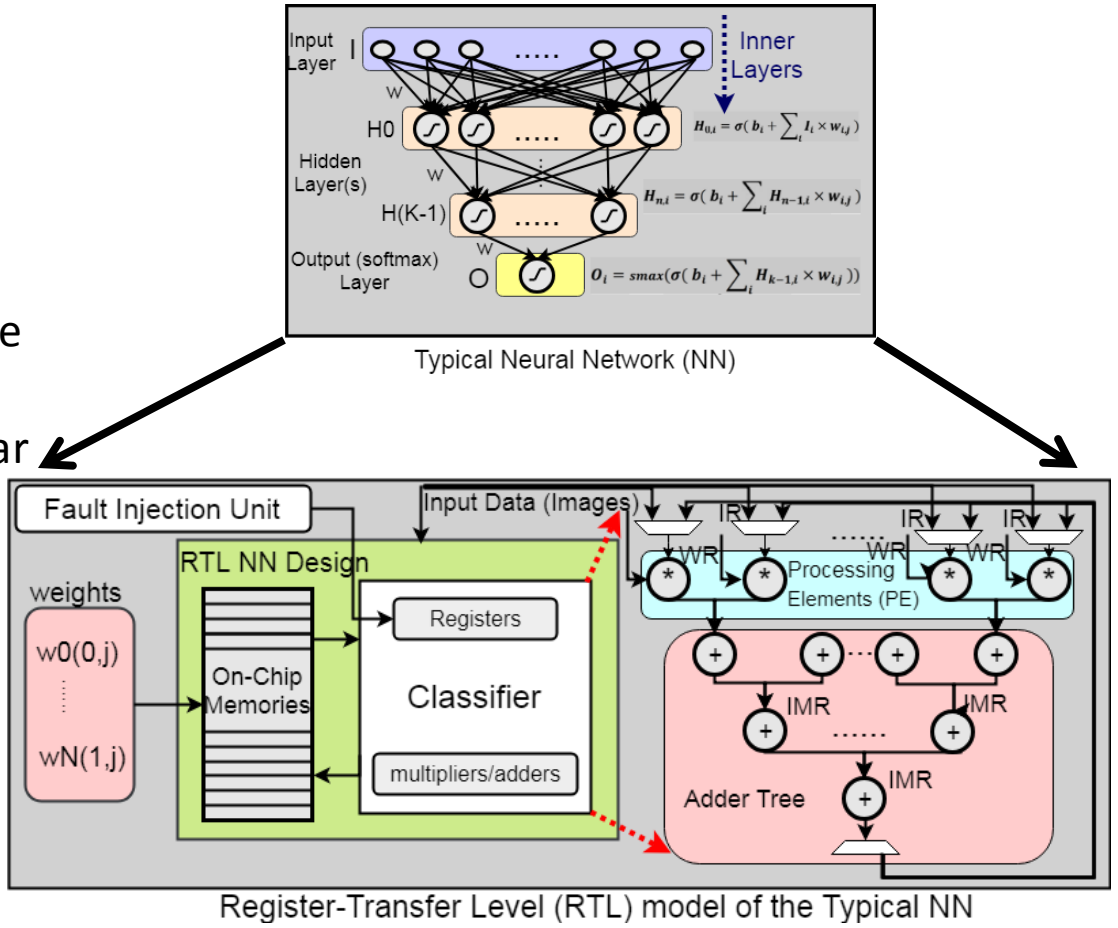
- **Fault Characterization of RTL NN**
  - **Empirically vulnerability analysis of different components of RTL NN**

- Fault Mitigation of RTL NN
  - An efficient technique to mitigate faults

- Summary and Future Works

Overall Methodology

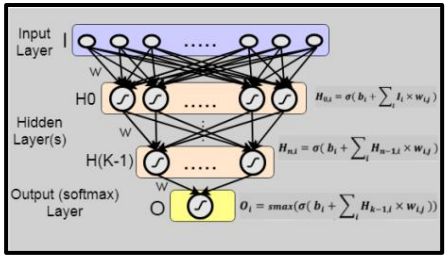
- Register-Transfer Level (RTL) is a hardware design model.
- Advantages of the RTL design:
  - Accurate-enough (similar to the on-silicon design)
  - Straightforward-enough (similar to the software code).
- With the rise of High-Level Synthesize (HLS) tools, RTL models are increasingly being common models.

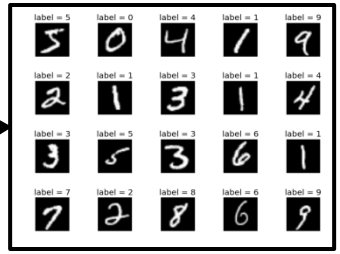


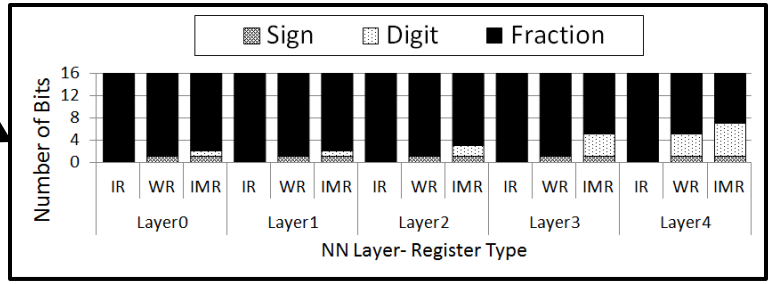
To build the RTL model of the NN , we use **Bluespec** (a cycle-accurate HLS tool).

Details of the Methodology

Neural Network (NN)	
Type	Fully-Connected Classifier
Topology (number of layers)	6L (1L input, 4L hidden, 1L output)
Per Layer Size (number of neurons)	(784, 1024, 512, 256, 128, 10)= 2714
Total Number of Weights	~1.5 million
Activation Function	Logarithmic Sigmoid (logsig)
Major Benchmark	
Name-Type	MNIST [12]- Handwritten Digits
Number of Images	Training: 60000, Inference: 10000
Number of Pixels per Image	28*28= 784
Number of Output Classes	10
Additional Benchmarks	
1. Forest	[13]
2. Reuters	[14]
Data Representation Model	
Type	16-bits Fixed-Point (FP)
Precision	Min sign and digit per layer (Fig. 2)
An Example Synthesize of RTL NN on FPGA	
FPGA Platform-Chip	VC707-Virtex7
Operating Frequency	100Mhz
BRAM Usage (Total: 2060)	70.8%
DSP Usage (Total: 2800)	8.6%
FF Usage (Total: 303,600)	3.8%
LUT Usage (Total: 607,200)	4.9%
Number of *PE*s	64

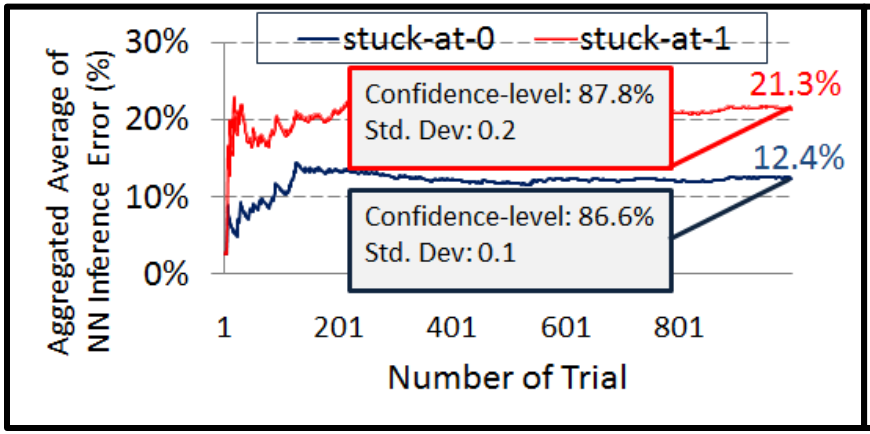






Fault Model

- Where to inject fault?
  - A set of bits is fully randomly selected among all available NN data.
- Supported type of faults:
  - Permanent (stuck-at-0 or stuck-at-1): stuck to 0 or 1 for the whole execution cycles.
  - Transient: bit-flip for a single cycle
- Statistically significant results:
  - Due to high number of possibilities to inject faults, it is more practical to randomly-select a subset of these possibilities. But how many?



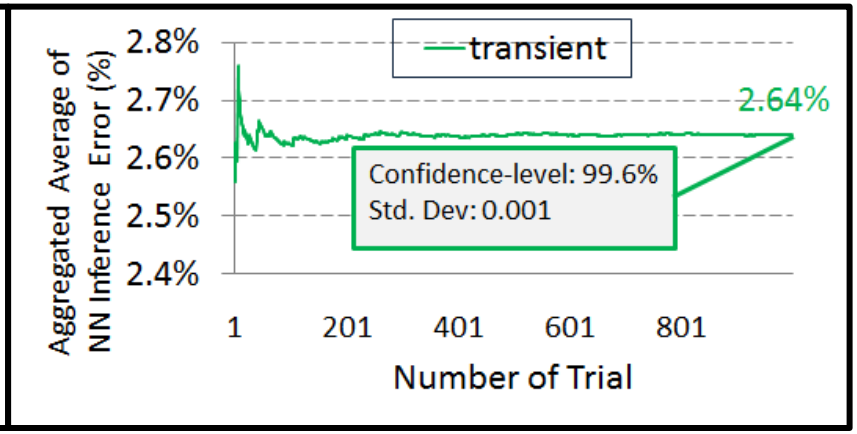
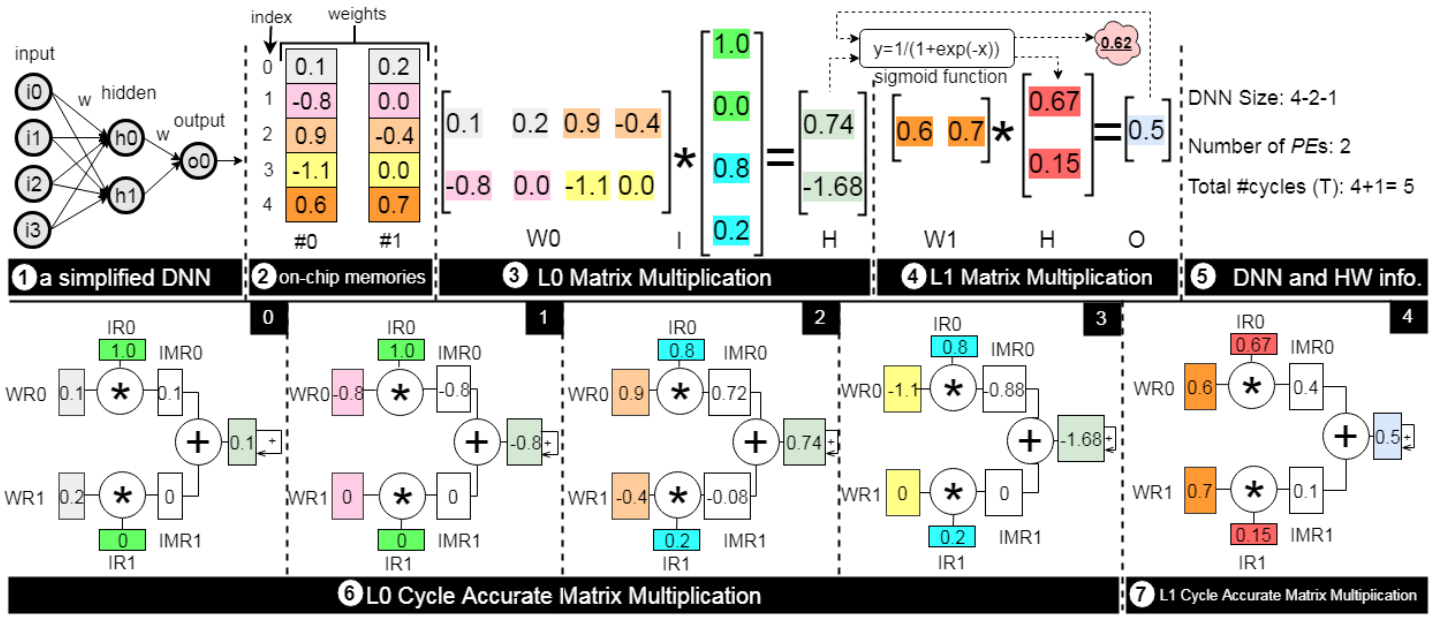
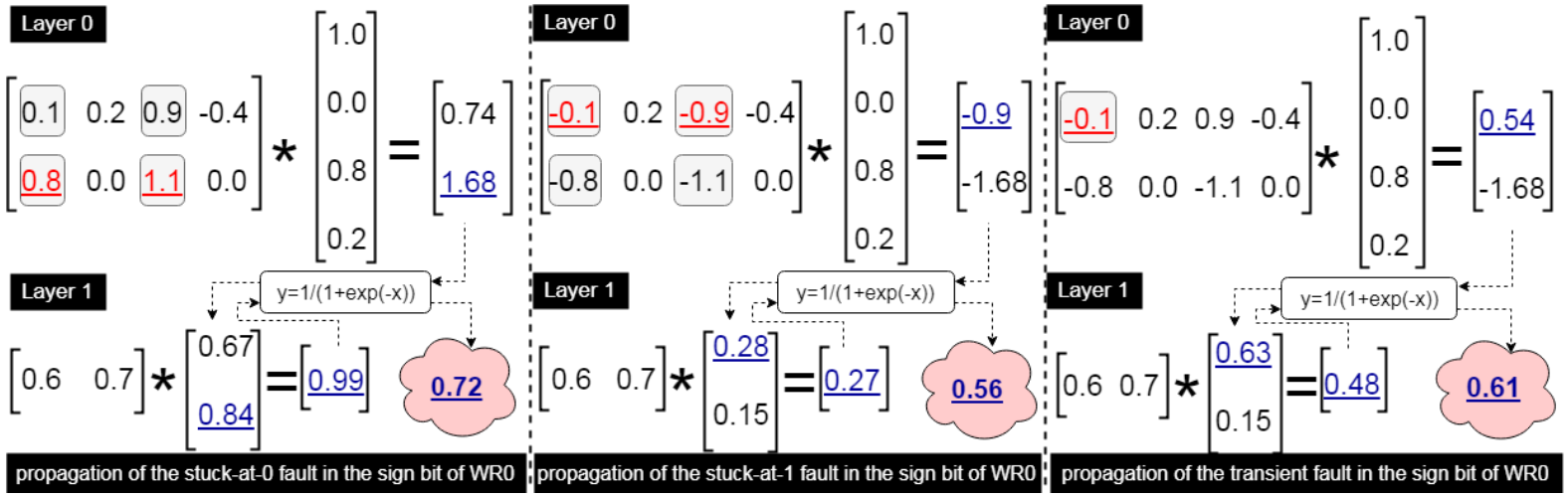


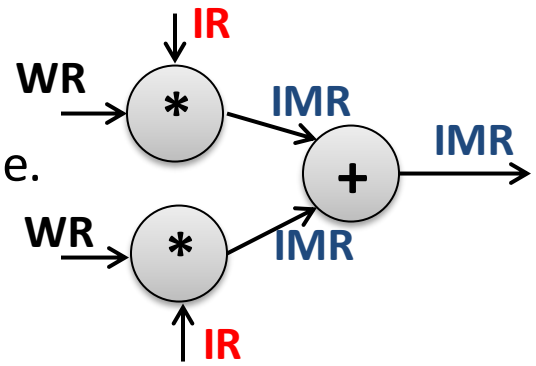
Illustration of Methodology

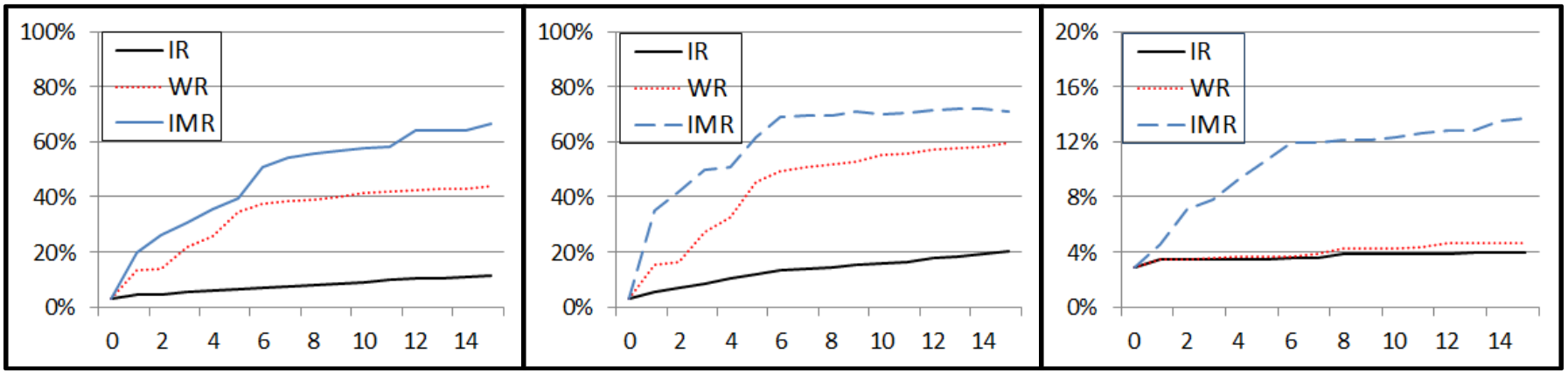




Vulnerability of Data Types of NN

- Three main data types of a typical NN:
  - **Weights or WRs** (parameters of the NN, uploaded from the offline training stage)
  - **Inputs or IRs** (images in MNIST, …)
  - **InterMediate or IMRs** (the internal NN data, result of multiply-add computations)
- Methodology: Injecting faults in individual data types
  - Select ***random bits to inject faults*** among individual data types
- Results: Inputs/Intermediate are the least/most vulnerable.
  - Intermediate has the longest digit component.
  - They are in the adder part (not multiplier).

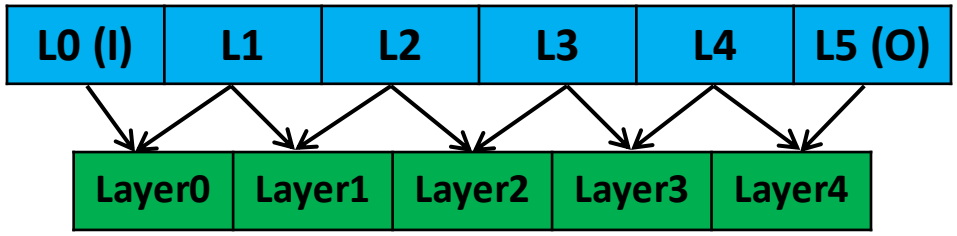


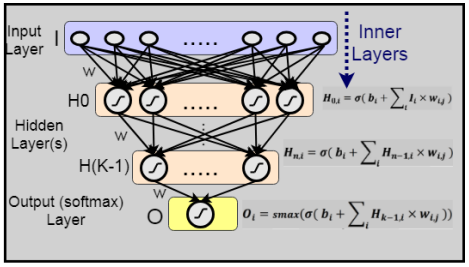


Stuck-at-0 Stuck-at-1 Transient

Vulnerability of Layers of NN

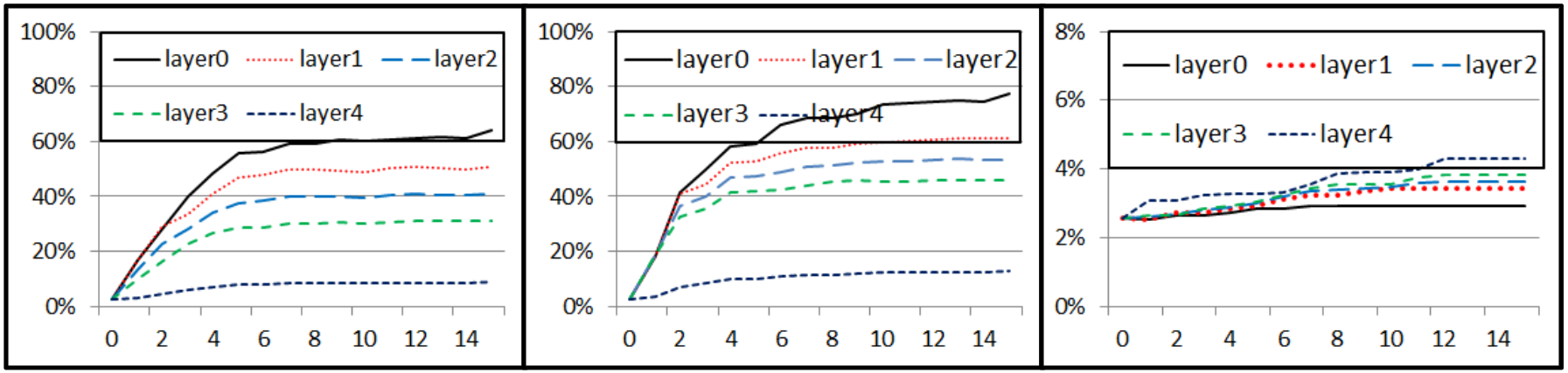
- There is an activation function between consecutive NN layers.





Typical Neural Network (NN)

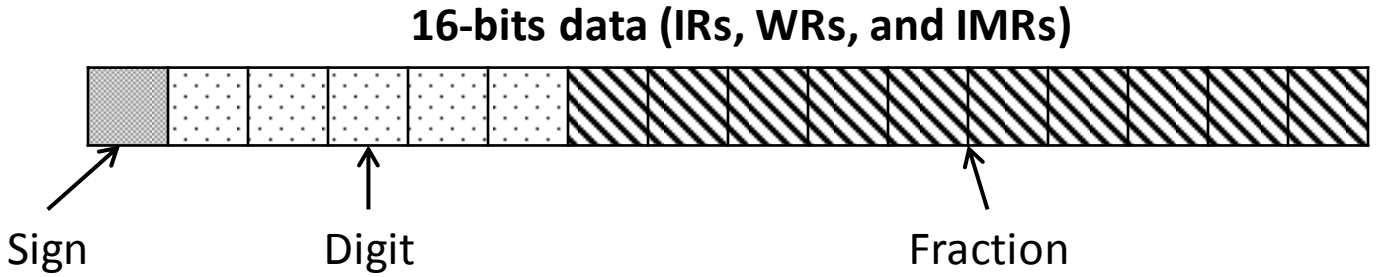
- Methodology: Injecting faults in individual NN layers
  - Select ***random bits to inject faults*** among individual NN layers
- Results:
  - Inner layers (closer to the output) are relatively more vulnerable, as the result of the less thresholding by activation functions.



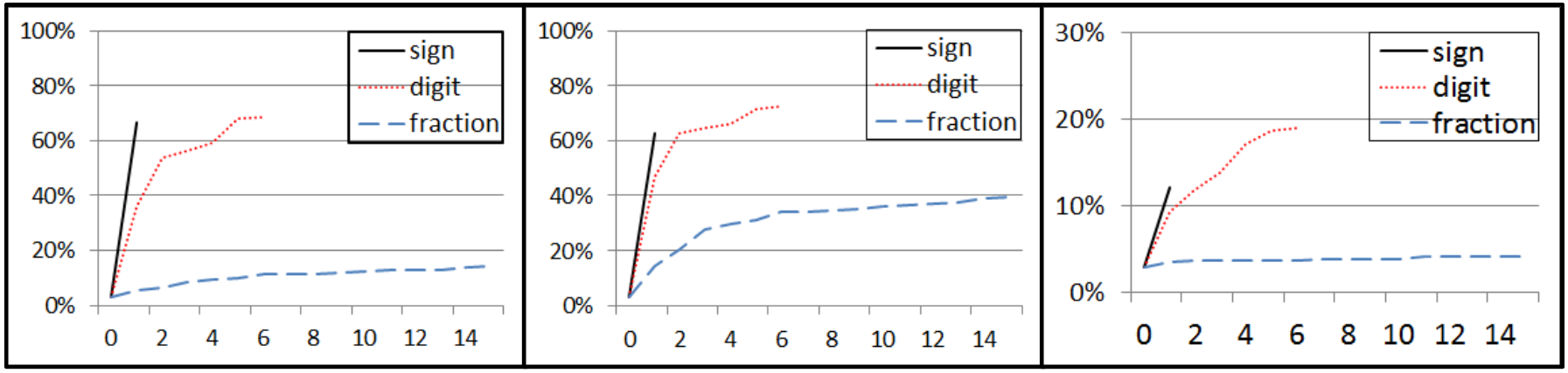
Stuck-at-0 Stuck-at-1 Transient

Vulnerability of Fixed-point Components

- Low-precision fixed-point data representation model:
  - More energy-efficient than full-precision floating point
  - 16-bits composed of Sign, Digit, and Fraction Components (minimum for sign and digit and the rest for fraction)



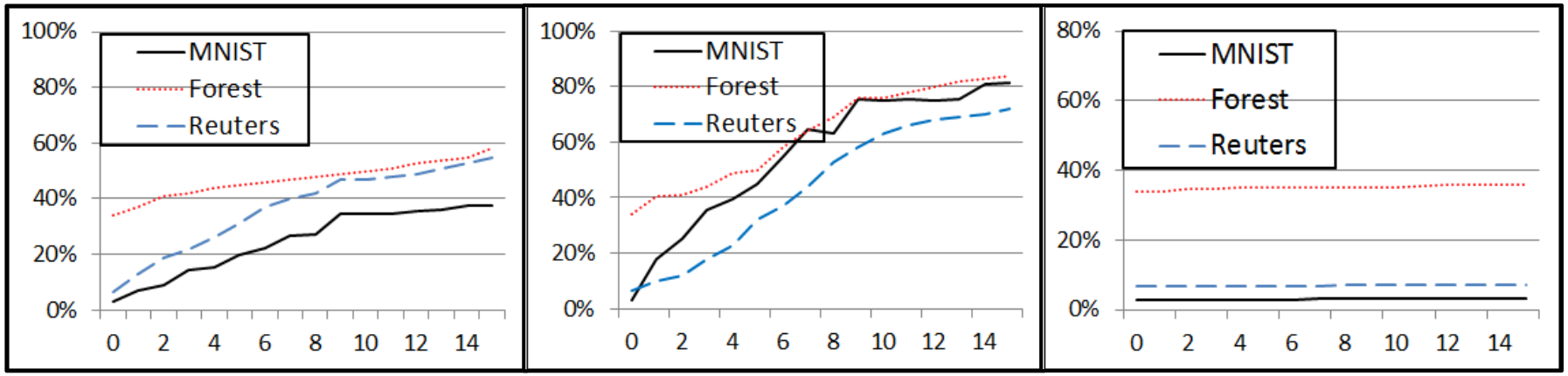
  - Methodology: Injecting faults in individual components
    - Select ***random bits to inject faults*** among individual data components, i.e., sign, digit, and fraction.
  - Results: As expected, sign, digit, and fraction components are more vulnerable in order.



Stuck-at-0 Stuck-at-1 Transient

Multiple NN Benchmarks

- Validating the generality of results by more benchmarks:
  - **MNIST**: Handwritten digit black-and-white images
    - *(|Input|= 784, |Output|= 10)*
  - **Forest**: Cartographic observations for classifying the forest cover type
    - *(|Input|= 54, |Output|= 8)*
  - **Reuters**: News articles for text categorization
    - *(|Input|= 2837, |Output|= 52)*
- Discussion on Results:
  - Inherent error rate (without fault): MNIST (2.56%), Forest (5.6%), and Reuters (37.8%)
  - Most of the findings on MNIST are valid for new two benchmarks too, e.g., data sparsity.
  - Reuters is relatively less-sparse so less-effected by stuck-at-1 faults.

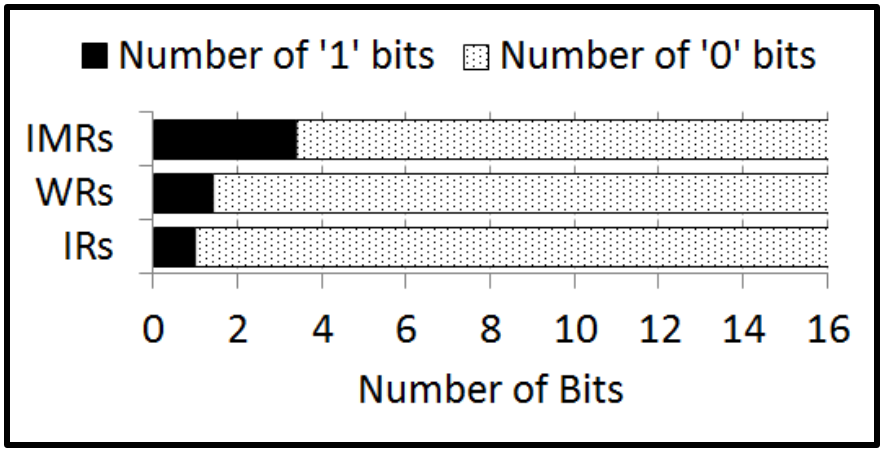


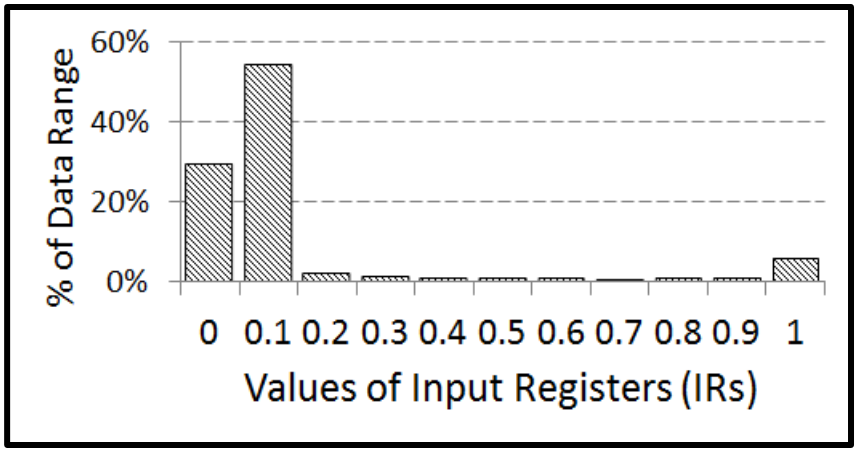
Stuck-at-0 Stuck-at-1 Transient

Sparsity of NN Benchmarks

- Data of studied benchmarks are sparse, i.e., more number of '0' than '1'.
  - Previous papers show similar feature for other state-of-the-art benchmarks, .e.g., ImageNet and AlexNet.

- Due to the inherent data sparsity of NNs:
  - Stuck-at-1 faults are more destructive than stuck-at-0 faults.
  - Good for aggressive undervoting faults, as primarily experimented.





Outline

- Motivation
  - Why accelerators for NNs? Why Register-Transfer Level (RTL) model?
  - Why to study resilience in NNs Accelerators?

- Fault Characterization of RTL NN
  - Empirically vulnerability analysis of different components of RTL NN

- **Fault Mitigation of RTL NN**
  - **An efficient technique to mitigate faults**

- Summary and Future Works

State-of-the-art Fault Mitigation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Studied case:

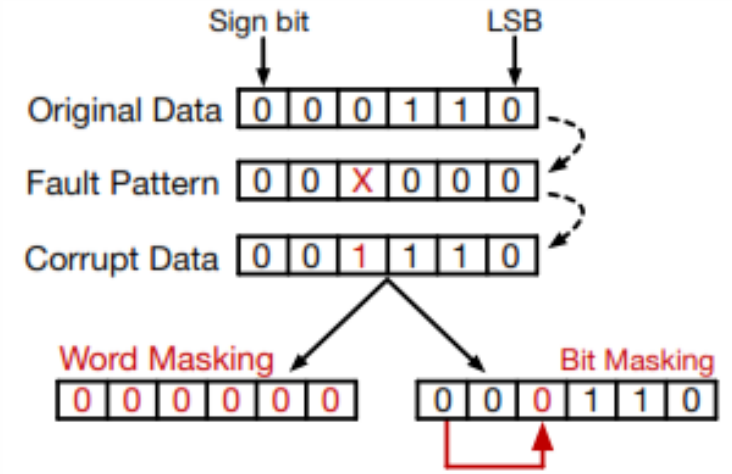
- *Brandon Reagen, et. al. Minerva: Enabling Low-power, Highly-Accurate DNN Accelerators* (ISCA-2016).

Fault Detection Assumptions:

- There is no limit on the number of faults that can be detected.
- Information is available on which bits are affected.
- Razor shadow register is a feasible solution to achieve above goals.

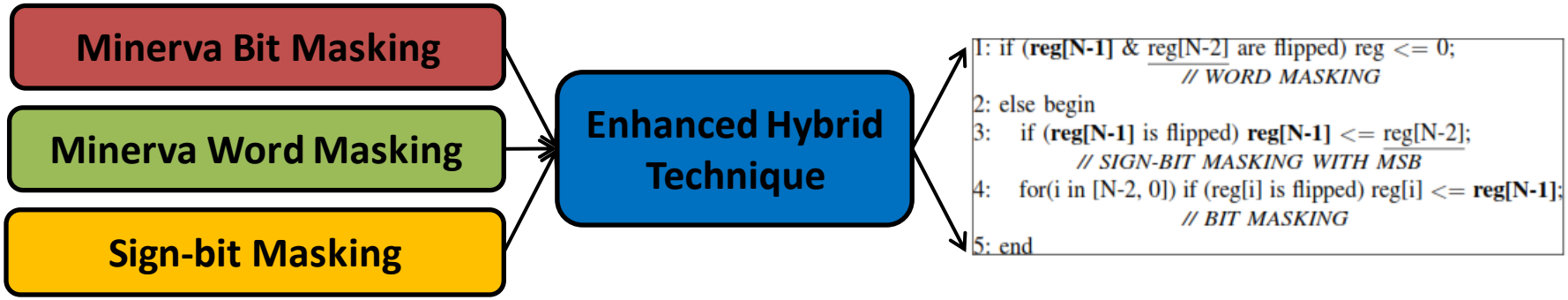
Fault Mitigation Techniques:

- **Bit Masking**: any bit that experiences fault is replaced with the sign-bit.
- **Word Masking:** when a fault is detected all bits of the word are reset to '0'.
- **Results:** The combination of Razor with Bit Masking allows the NN weights to tolerate **44X** more faults than Word Masking.

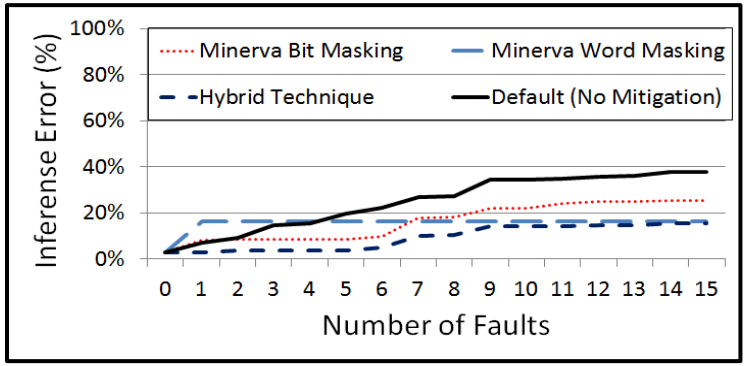


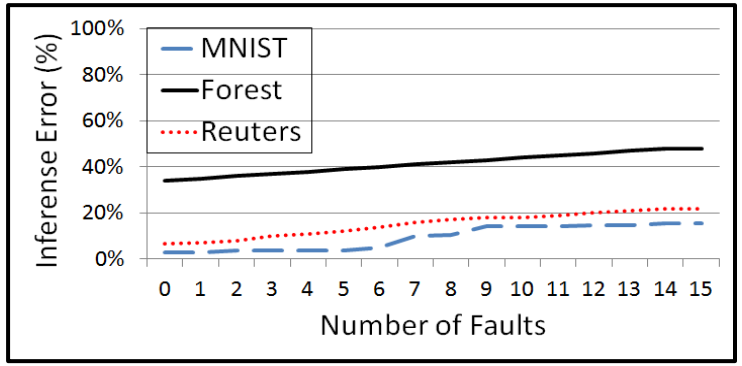
An Enhanced Fault Mitigation Technique

- A combination of **Bit Masking**, **Word Masking**, and **Sign-bit Masking** (if a fault in sign-bit is detected, mask it with MSB).
- It relies on the "sparsity of NN data" and "sign-bit and MSB have same logic".



- Experimental Results:
  - Hybrid technique is 47.3% better than Word Masking.
  - Bit Masking is not efficient when sign-bit is corrupted.





Outline

- Motivation
  - Why accelerators for NNs? Why Register-Transfer Level (RTL) model?
  - Why to study resilience in NNs Accelerators?

- Fault Characterization of RTL NN
  - Empirically vulnerability analysis of different components of RTL NN

- Fault Mitigation of RTL NN
  - An efficient technique to mitigate faults

- **Summary and Future Works**

Summary & Future Works

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Summary

- We showed that NN accelerators are susceptible to faults, e.g., Undervolting faults.
- For a more comprehensive analysis, we analyzed the Resilience of NN accelerators in RTL that is a close model to hardware.
- We extracted the severity of different components of the NN accelerator against faults (Fault Characterization).
- We evaluated an efficient technique to minimize the effect of faults on NN accuracy (Fault Mitigation).

Future Works

- Advanced Neural Network models like CNNs, LSTMs, etc.
- Evaluate the mitigation technique on the silicon.
- Confirming the experimental results by the analytical analysis.

Thanks!

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Contact:
Behzad Salami
behzad.salami@bsc.es

The research leading to these results has received funding from the European Union's Horizon 2020 Programme under the LEGaTO Project (www.legato-project.eu), grant agreement n° 780681.

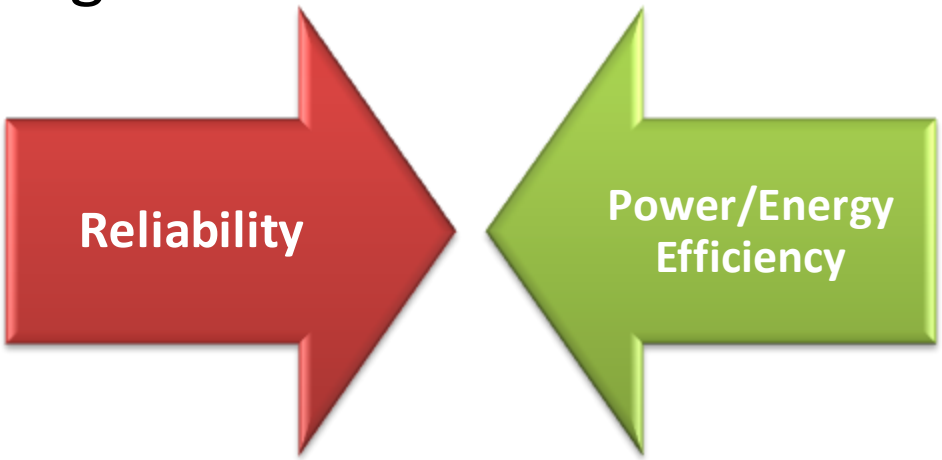



Backup

Undervolting

Underscaling the supply voltage *below the nominal level* :

- **Power/Energy Efficiency**: Reduces quadratic ally dynamic and linearly static power.
- **Reliability**: Increases the circuit delay and in turn, causes timing faults.



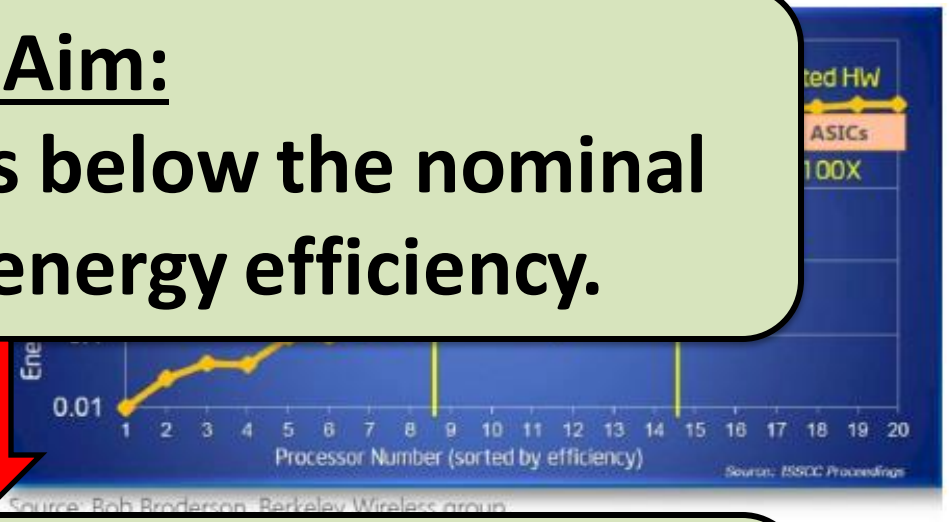
Aggressive Undervolting is not DVFS!

Motivation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

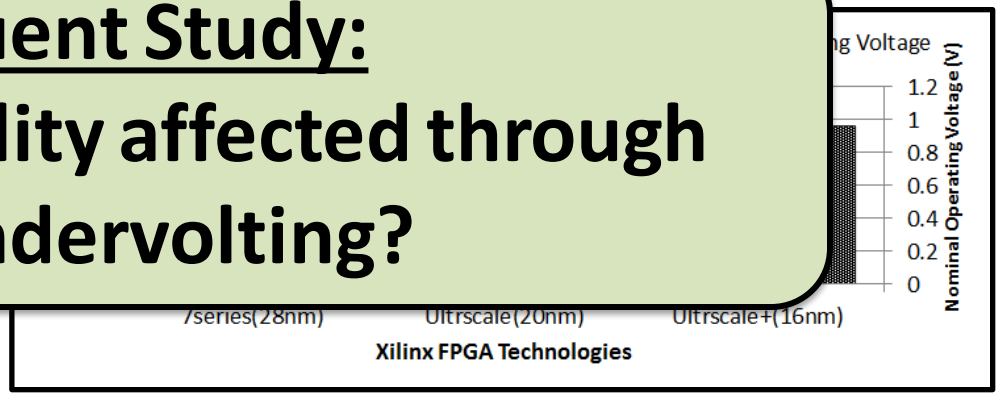
Contribution of FPGAs in large data centers is growing, expected to be in 30% of datacenter servers by 2020 (Top500 news).

Our Aim:
Undervolting FPGAs below the nominal level to achieve energy efficiency.



Source: Bob Broderson, Berkeley Wireless group

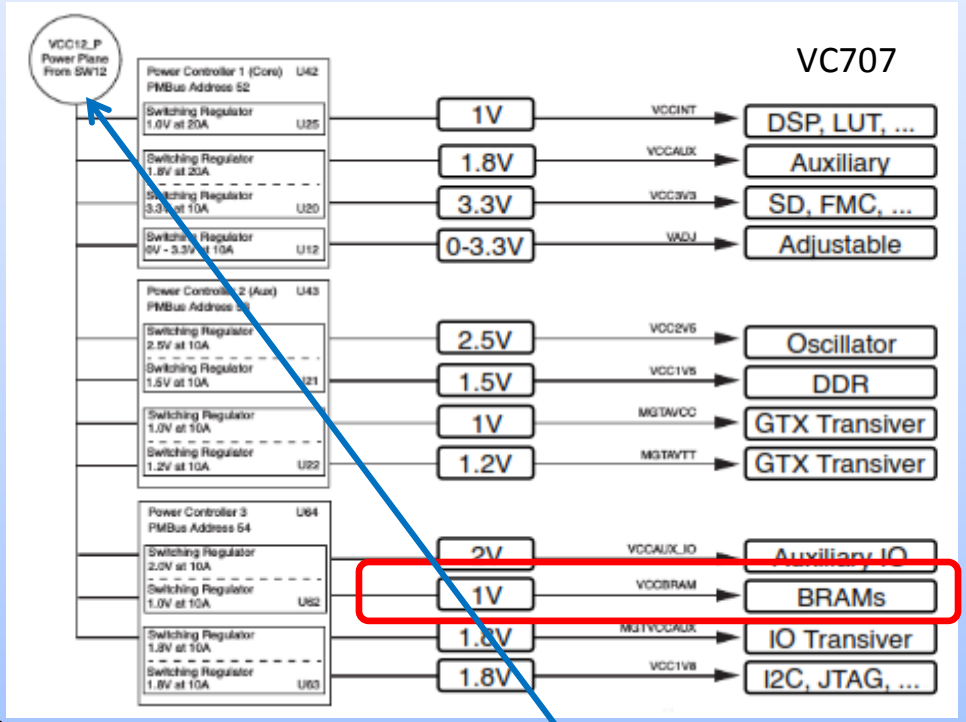
Subsequent Study:
How is the reliability affected through FPGAs Undervolting?



Voltage Scaling Capability in Xilinx

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Voltage Distribution on Xilinx Platforms



Evaluated Xilinx Platforms

VC707: performance-efficient design

KC705: power-efficient design

Voltage Regulator

- Power Management Bus (PMBus).
- Hardwired to the host.

Influencing the Direction of Power Management

PMBus® A·V·S

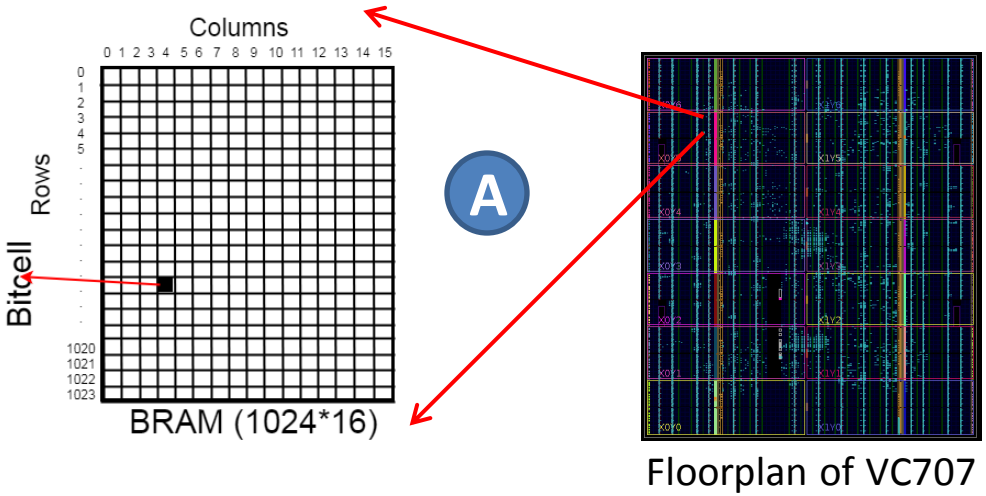
Experimental Methodology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A Detailed study on FPGA BRAMs, which are a set of bitcells in the row-column format.

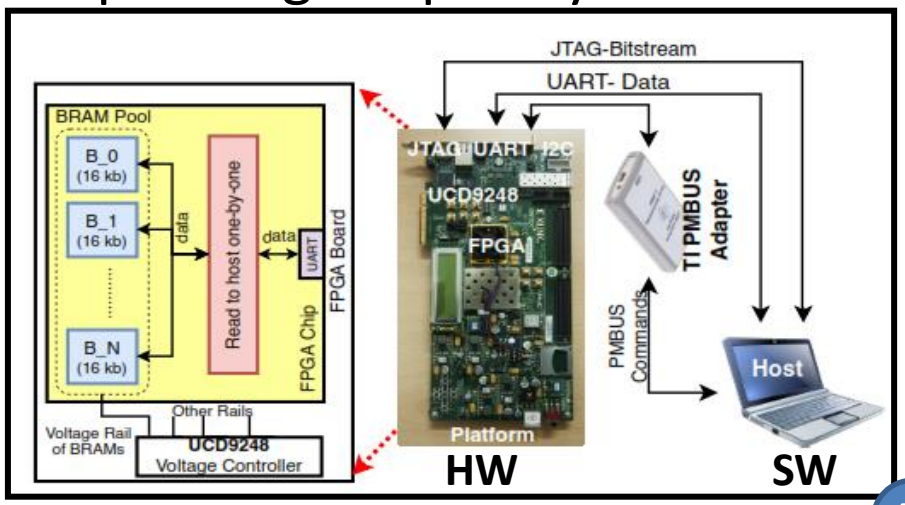
B Experimental Methodology:

1. **HW**: Transfer content of BRAMs to the host.
2. **SW**: Analyze data, and adjust voltage of BRAMs.



Floorplan of VC707

Operating frequency is set to the maximum, i.e., ~500mhz.




```
1: V_CCBRAM = V_min;
2: while(V_CCBRAM >= V_crash) begin
3:     while(numRun <= 100) begin
4:         delay(1sec);
5:         Transfer content of BRAMs to the host;
6:         Analyse faulty data (rate and location);
7:         numRun++;
8:     end
9:     V_CCBRAM -= 10(mV);
10: end
```

B

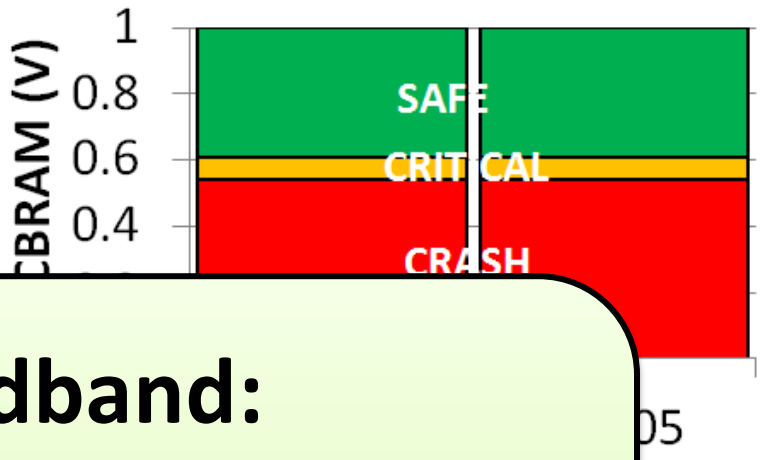
Overall Behavior- Power & Reliability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SAFE
- No observable fault
- Voltage Guardband Below Vnom

CRITICAL
- Faults manifest
- Below Vmin, min safe voltage

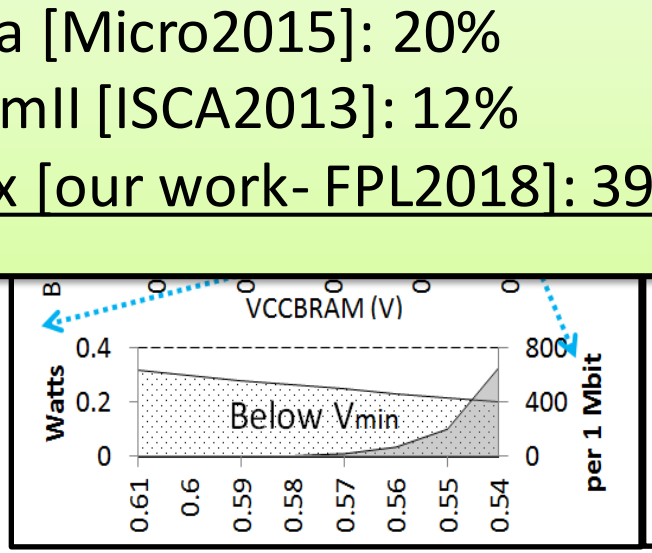
C



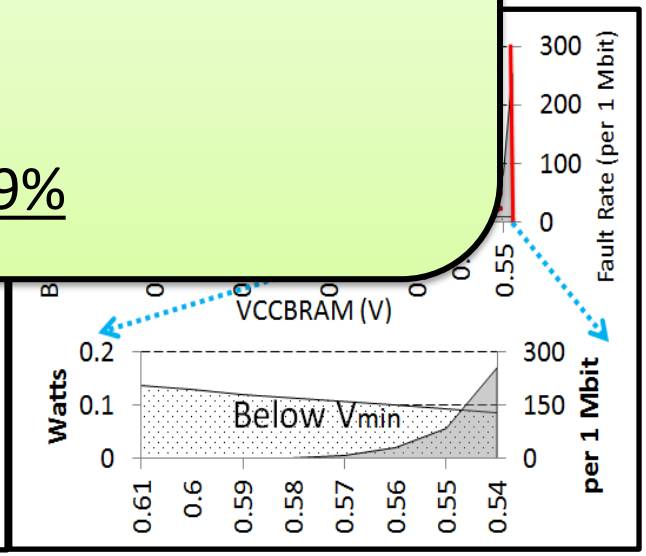
Voltage Guardband:

1- DRAM- MultipleVendors [Sigmetrics2017]:16%
2- GPU- NVidia [Micro2015]: 20%
3- CPU- ItaniumII [ISCA2013]: 12%
4- FPGA- Xilinx [our work- FPL2018]: 39%

1. V
2. V
   s
3. M
   ef
4. Exponential fault rate increase.
5. VC707 experiences relatively more fault rate.



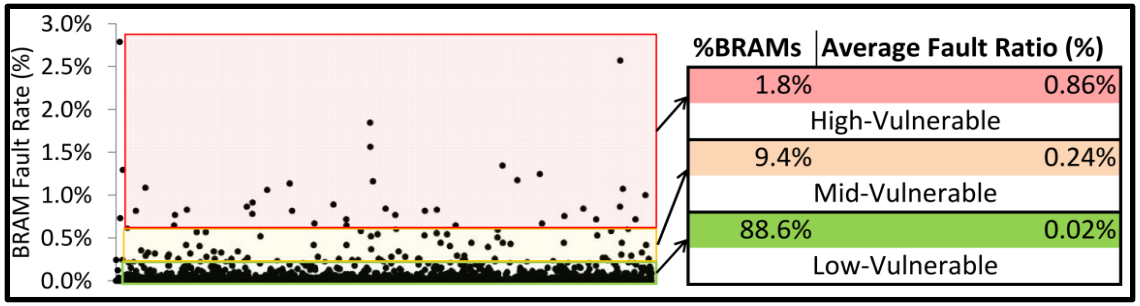
VC707



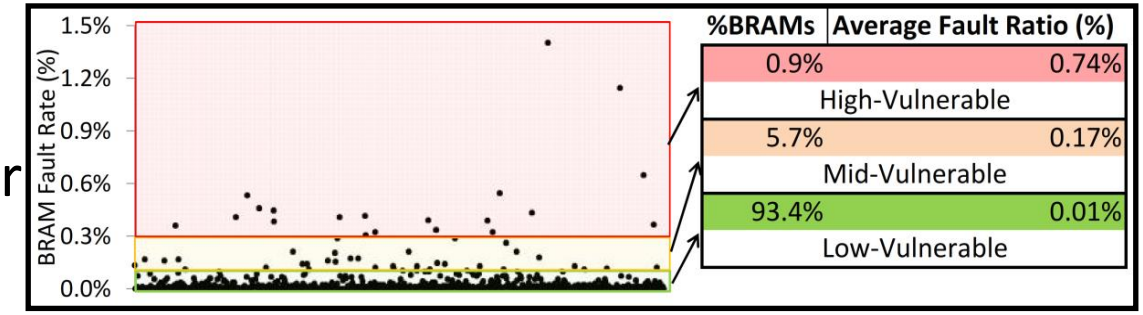
KC705

Fault Variability between BRAMs

- BRAMs clustering using K-Mean clustering.
- Majority of BRAMs are low-vulnerable.
- ~36% of BRAMs never experience faults.
- Fully non-uniform fault distribution.



VC707



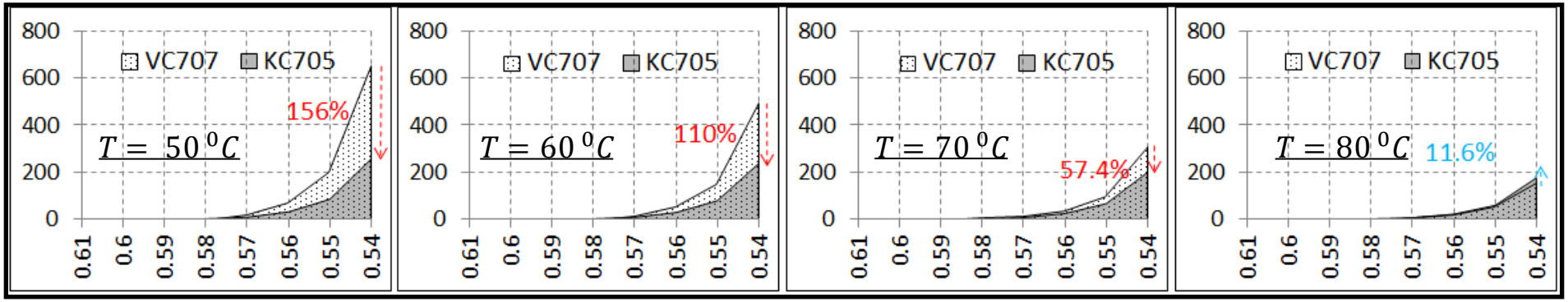
KC705

VCCBRAM= Vcrash

* Different scales in y-axis * *Pattern= 18'h3FFFF *

Environmental Temperature

- **Methodology:** Adjusting environmental temperature, monitoring on-board temperature via PMBus.
- **Experimental Observation:**
  - At higher temperatures, fault rate is significantly reduced.
  - The rate of this reduction is highly platform-dependent (VC707 > KC705).
- **Inverse Temperature Dependency (ITD):**
  - For nano-scale technologies, under ultra low-voltage operations, the circuit delay reduces at higher temperatures since supply voltage approaches the threshold voltage.



* y-axis: VCCBRAM (V), y-axis: fault rate (per 1Mbit) *

Summary & Future Works

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Summary

- We experimentally showed how Xilinx FPGAs work under aggressive low-voltage operations.
- There is a conservative voltage guardband below the nominal level.
- BRAMs power is significantly reduced through Undervolting; however, reliability degrades below min safe voltage.
- We characterized the behavior of Undervolting faults at the critical region.

Future Works

- Dynamic Vmin scaling, adapted by frequency and temperature.
- More advanced designs, where other components such as I/O, DDR, DSP are undervolted.
- Efficient Fault Mitigation Techniques.
- Profiling applications such as Deep Neural Networks (DNNs), among others.
- Extending Undervolting for other commercial FPGAs such as Intel/Altera.

Outline

- Background
  - What does Undervolting mean?
  - Motivation: FPGAs Undervolting
- First Contribution: Undervolting Xilinx FPGAs
  - Experimental Methodology
  - Overall Power and Reliability Trade-off
- Second Contribution: Fault Characterization
  - Fault Variability
  - Fault Types
  - Impact of the Environmental Temperature
- Related Work
- Summary and Future Works

Fault Characterization at CRITICAL Region

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Permanent ‘1’ to ‘0’ bit-flips

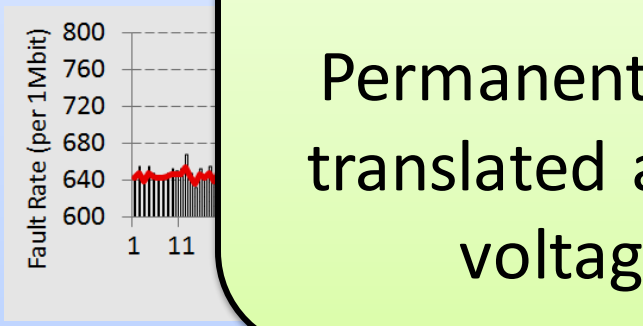
Permanent:

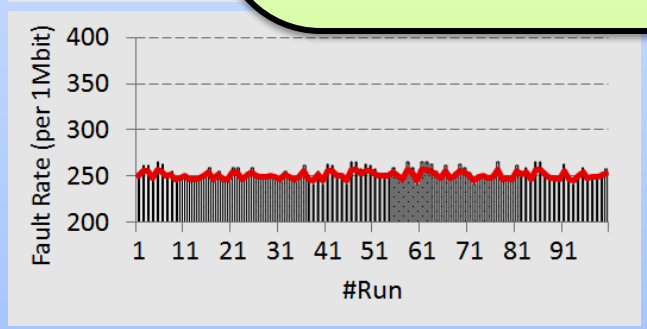
- There is no c… majority of *rate* and *loc*…
- Validated by … mutations. 100 times.

‘1’ to ‘0’ bit flips:

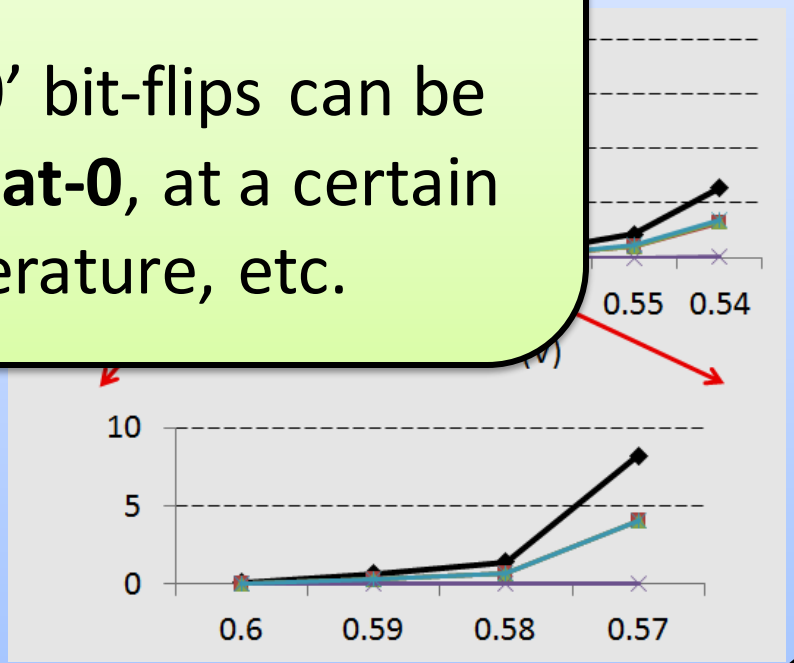
Conclusion:

Permanent ‘1’ to ‘0’ bit-flips can be translated as **stuck-at-0**, at a certain voltage, temperature, etc.





KC705



VC707

Related Works of Undervolting

- **Simulation-based**: **(Lack of precise information of the real hardware.)**

Focus of Previous Works:

(1) **Covered** in our work for FPGAs

- Voltage Guardband
- Fault Characterization at Critical Region
- Impact of Environmental Conditions

(2) **Not-covered** in our work on FPGAs (Future Work)

- Dynamic Vmin Prediction
- Fault Mitigation at Critical Region
- Application Profiling

Constraints of Xilinx FPGAs

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Future of FPGA Undervolting needs more advanced voltage designs, by vendors:

1. Many FPGA platforms, e.g., Zynq are not equipped with voltage scaling capability.
2. There is no standard about the voltage distribution among platform components.
3. Voltage regulators are hardwired to the host through PMBus interface.
4. In many cases, several components on the FPGA platform share a single voltage rail.
5. Vendors set unnecessarily conservative voltage guardbands that increase the energy.
6. There is no publicly-available circuit-level information of FPGAs.