www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FPGAVolt: Low-power FPGA-based DNN Accelerator through Aggressive Undervolting

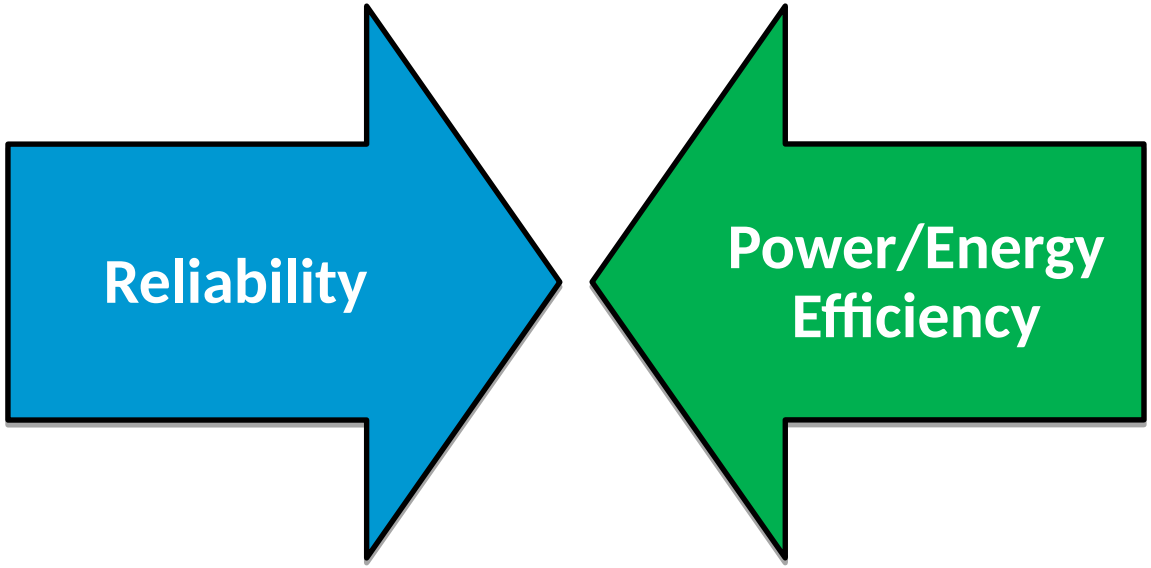
Presentation by: **Sanem Arslan**

Barcelona Supercomptuing Center (BSC)
Barcelona, Spain.

HiPEAC2020-WAICA: Workshop on Artificial Intelligence and Computer Architecture
22- January- 2020. Bologna, Italy.

Aggressive Undervolting

- ***Aggressive undervolting-*** Underscaling the supply voltage below the nominal and safe level:
  - ***Power/Energy Efficiency***: Reduces dynamic and static power quadratically and linearly, respectively.
  - ***Reliability***: Increases the circuit delay and in turn, causes timing faults.



- ***Dual/Multi-Vdd, DVS, and DVFS: Similar but different mechanisms to aggressive undervolting:***
  - ***Similarity***: Underscaling the supply voltage.
  - ***Difference***: Undervolting is until a certain **safe** level, usually constrained by vendors.

State-of-the-art

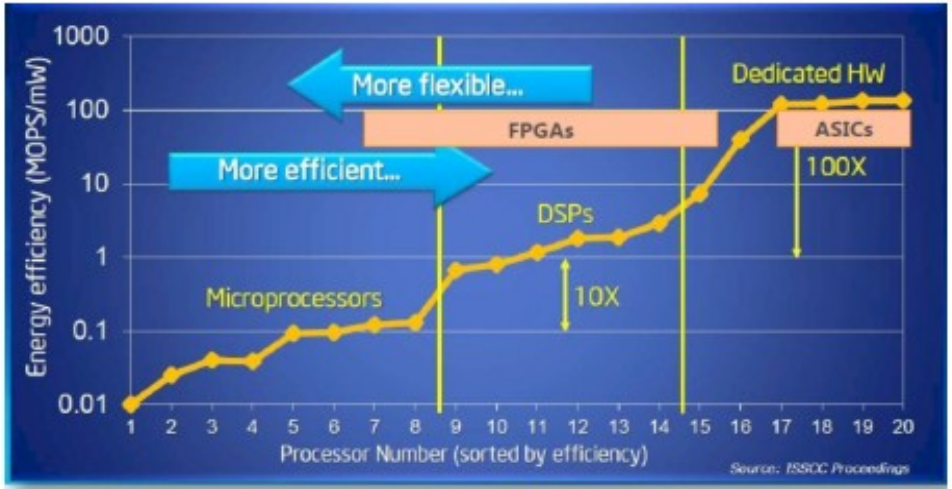
1. **Real hardware:** Aggressive undervolting has shown significant efficiency to reduce the energy consumption.
   - Devices:
     - CPUs: Itanium II (ISCA2014), X86 (IOLTS2017), ARM (HPCA2017)
     - GPUs: NVidia (Micro2015)
     - DRAMs: Multiple Brands (Sigmetrics2017)
     - **FPGA: This work**
   - Focus of the previous works:
     - Voltage guardband
     - Minimum safe voltage, *i.e.*, *Vmin* prediction
     - Fault characterization and mitigation
     - Chip-to-chip, core-to-core, and workload-to-workload variation
     - ....

2. **Simulation-based studies:** More straightforward and more parameters but less precise
   - ASIC DNN: Minerva (Micro2016), Thundervolt (DAC2018)
   - CPU: Bravo (HPCA2017)
   - Network On-Chip (HPCA2014)

Undervolting on FPGAs: Motivation

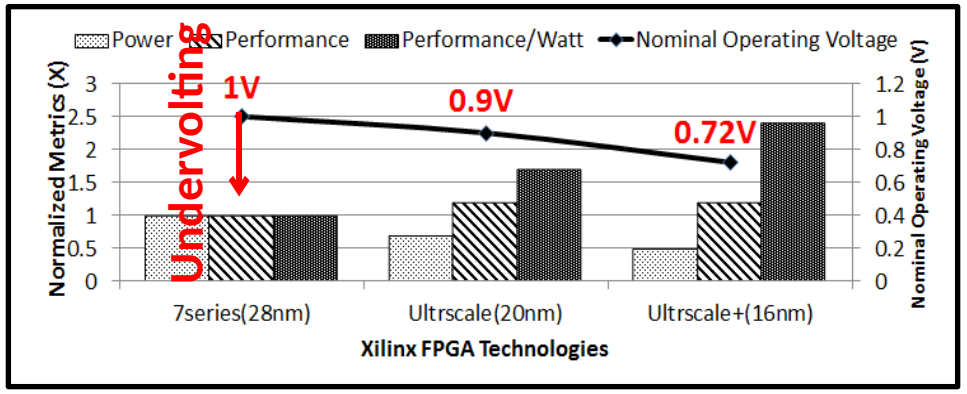
Contribution of FPGAs in large data centers is growing, expected to be in ***30%*** of datacenter servers by 2020 (Top500 news).

- In comparison to ASICs, energy efficiency of FPGAs is a serious concern, *i.e.*, 10X-100X less-efficient.



Source: Bob Broderson, Berkeley Wireless group [Intel/Altera]

- Nominal voltage reduction of FPGAs is naturally applied for different generations.



[Xilinx]

Contributions

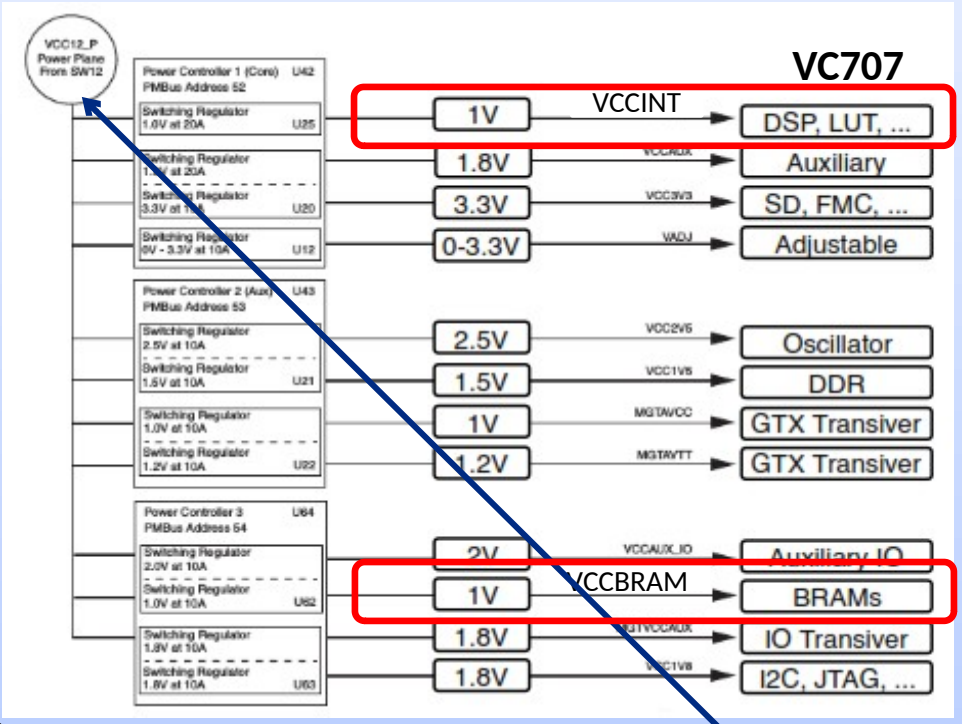
1. Undervolting FPGAs
   - ❑ Voltage guardband
   - ❑ Overall power and reliability trade-off

2. Fault characterization in FPGA on-chip memories
   - ❑ Fault type, location, and rate
   - ❑ Temperature, Chip

3. Low-voltage FPGA-based Neural Network (NN)
   - ❑ Power consumption and NN accuracy characterization
   - ❑ Fault mitigation techniques
     - ❖ Application-aware technique
     - ❖ Built-in ECC

Voltage Scaling Capability in Xilinx

Voltage distribution on Xilinx platforms



Evaluated Xilinx platforms

VC707: performance-efficient design

KC705: power-efficient design (A & B)

ZC702: ARM integrated with FPGA

Voltage regulator

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

Influencing the Direction of Power Management

PMBus® A·V·S

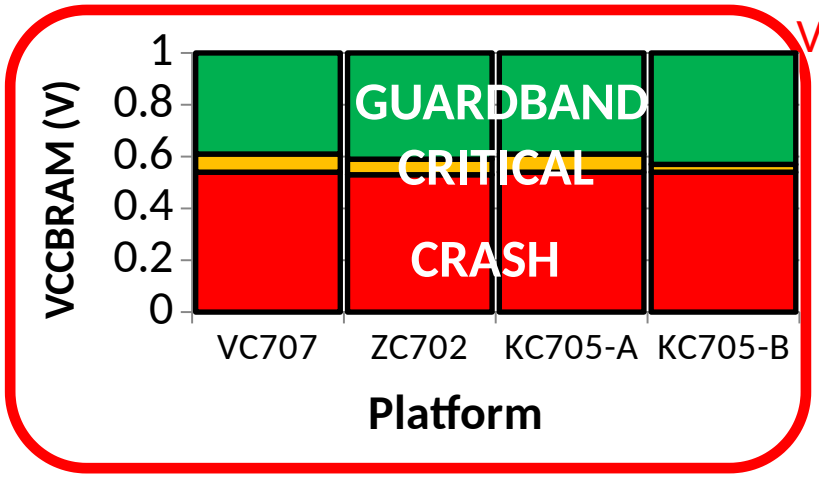
Overall Voltage Behavior

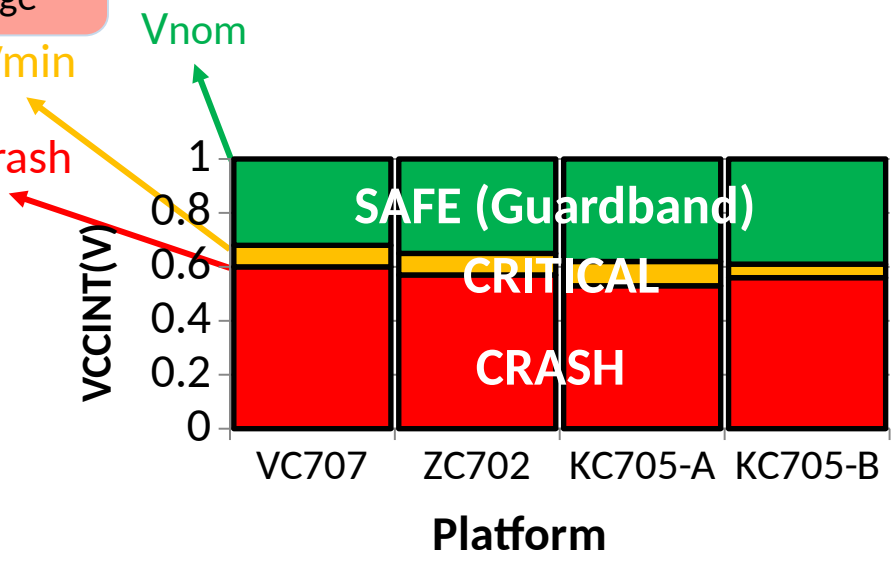
SAFE
- No observable fault
- Voltage Guardband below Vnom

CRITICAL
- Faults manifest
- Below Vmin, min safe voltage

CRASH
- FPGA stops operating below Vcrash, min operating voltage

- ***Voltage guardband***: to ensure the worst-case environmental and process technologies.
- ***Experimental conditions***: At ambient temperature and maximum operating frequency.

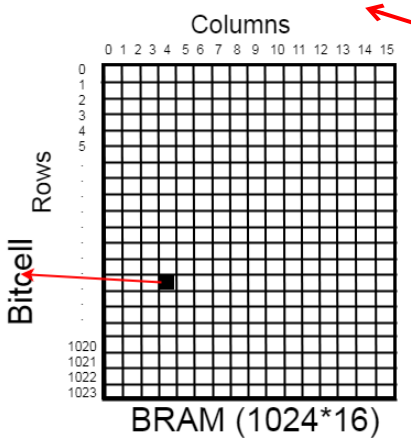


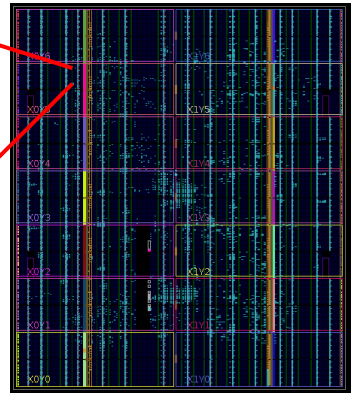


We performed more detailed studies on **FPGA on-chip memories (BRAMs).**

Experimental Methodology

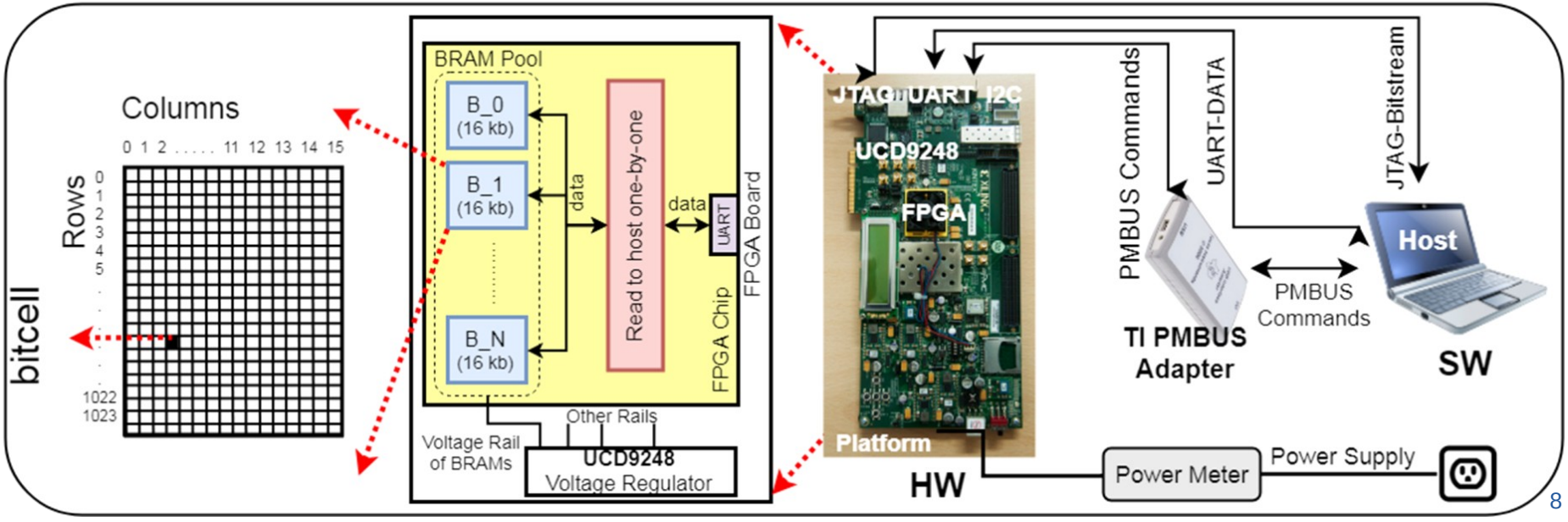
- ***FPGA BRAMs:***
  - Hierarchy of set of bit-cells distributed over the chip.
  - Size of each BRAM: 16-kbits





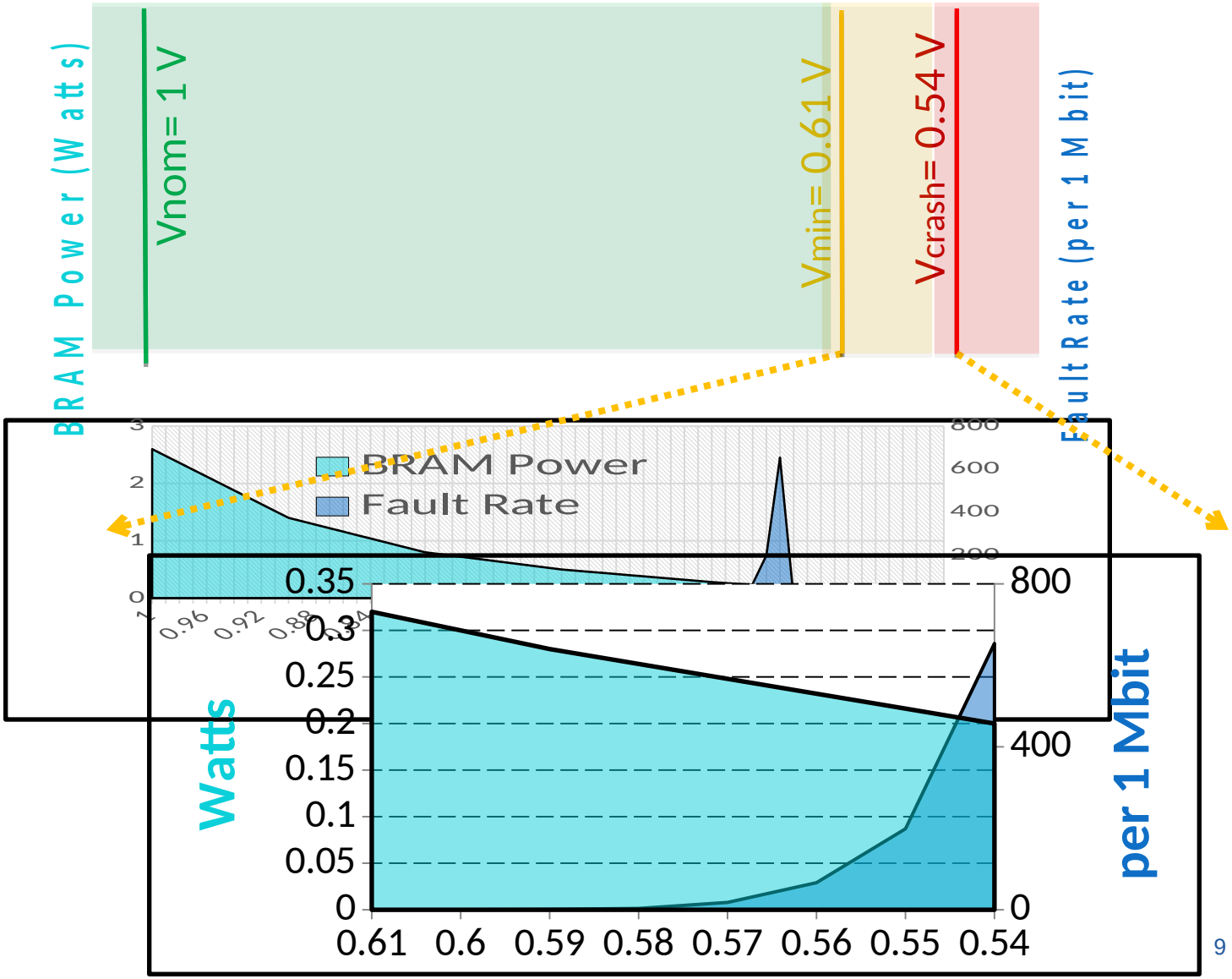
Floorplan of VC707 (2060 BRAMs)

- ***Experimental Methodology:***
  - **HW**: Transfer content of BRAMs to the host.
  - **SW**: Analyze data, and adjust voltage of BRAMs.



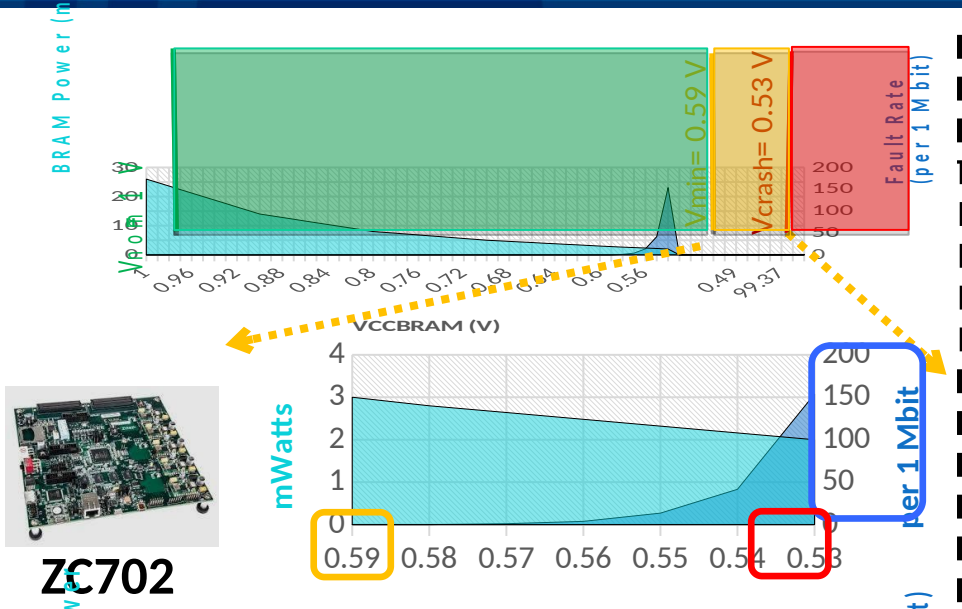
Overall Trade-offs on BRAMs- Power & Reliability

VC707

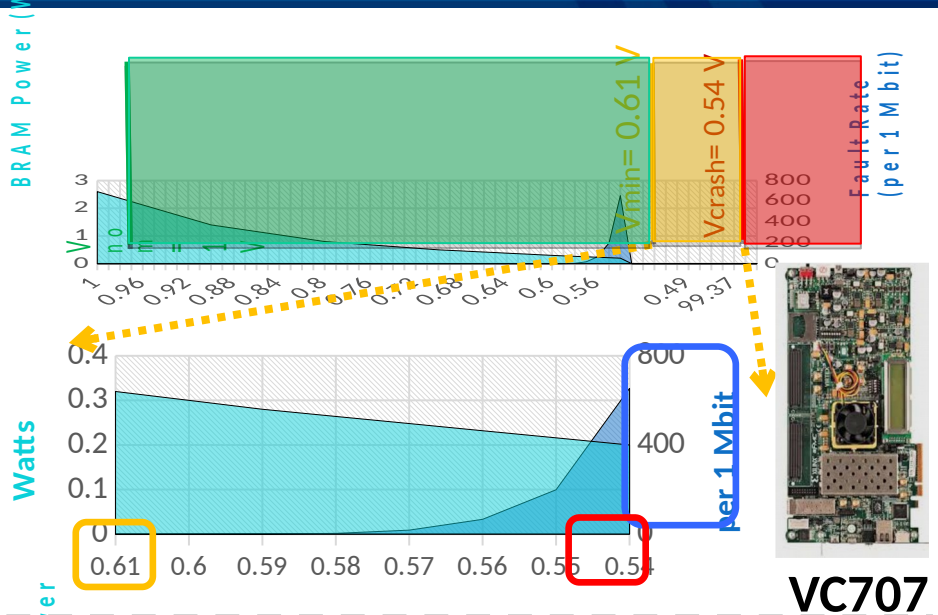


Overall Trade-offs on BRAMs- Multiple Platforms

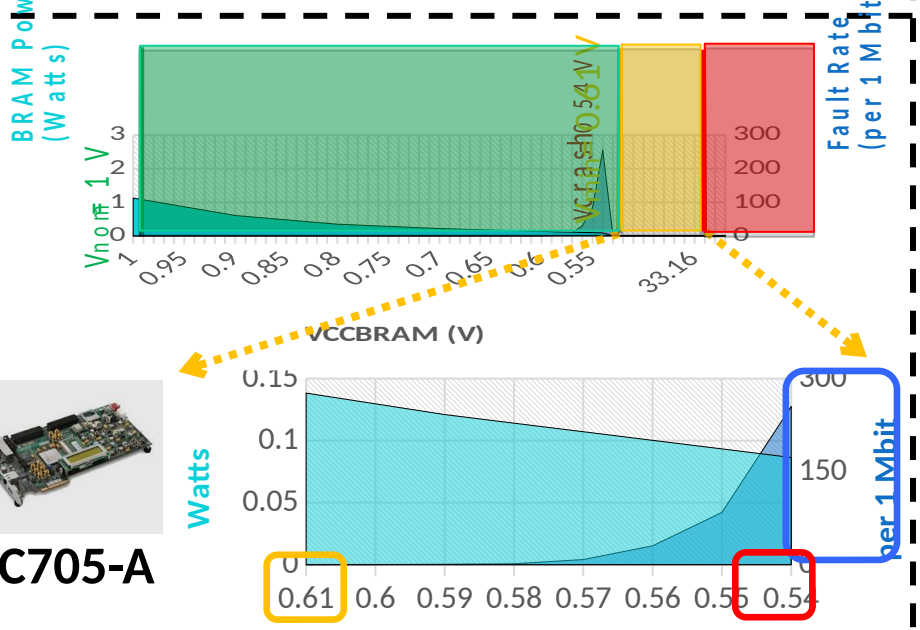
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



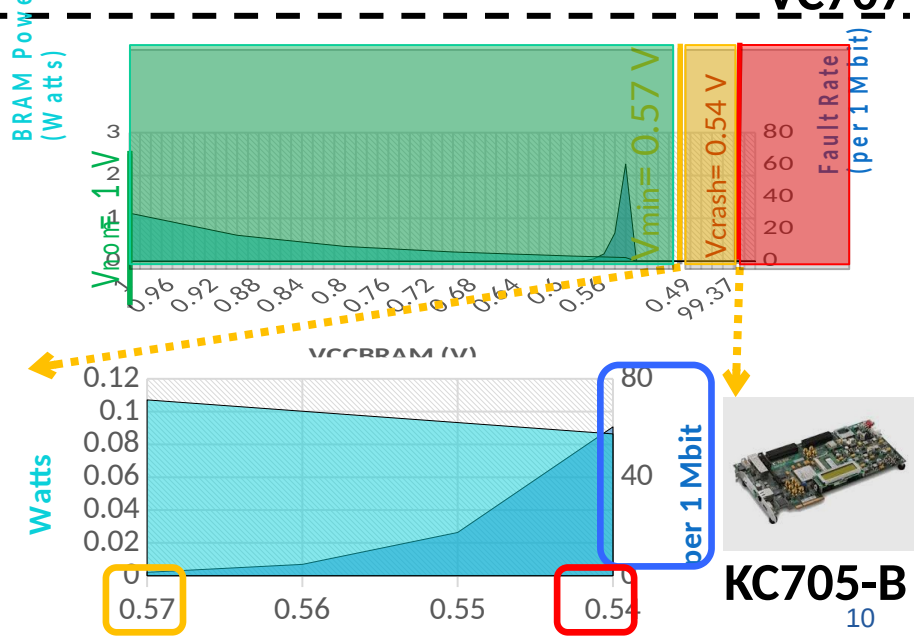
ZC702



VC707



KC705-A



KC705-B

Key Points of the First Contribution

- **Voltage regions:** Safe, Critical, and Crash voltage regions exist for all platforms, slightly different among studied platforms.
- **Voltage guardbands:** Large voltage guardband confirmed for all platforms on the studied voltage rails, *i.e.*, V_{CCBRAM} and V_{CCINT}.
- **Power reduction:** There is significant power reduction through aggressive undervolting, with more details studied for BRAMs.
- **Reliability costs:** Fault rates exponentially increase in the Critical voltage region.

Contributions

1. Undervolting FPGAs
   - ❑ Voltage guardband
   - ❑ Overall power and reliability trade-off

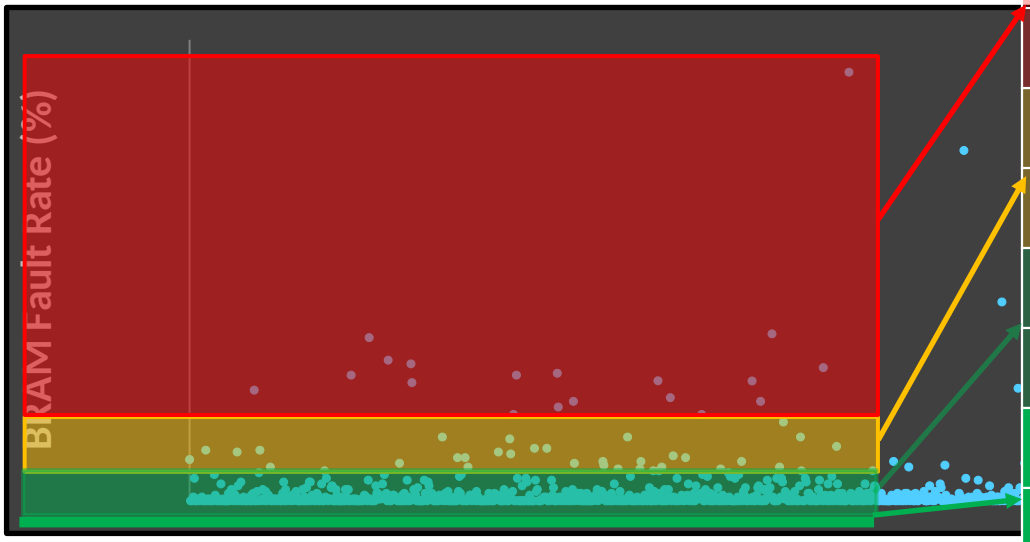
2. Fault characterization in FPGA on-chip memories
   - ❑ Fault type, location, and rate
   - ❑ Temperature, Chip

3. Low-voltage FPGA-based Neural Network (NN)
   - ❑ Power consumption and NN accuracy characterization
   - ❑ Fault mitigation techniques
     - ❖ Application-aware technique
     - ❖ Built-in ECC

Fault Characterization at CRITICAL Region

Fault variability among FPGA BRAMs:
Fully non-uniform fault distribution

- Fully non-uniform fault distribution.
- Majority of BRAMs do not experience many faults.



VC707 (2060 BRAMs)
VCCBRAM@ V_{crash}= 0.54V
Temperature@ Ambient

K-means clustering

%BRAMs	Average Fault Rate (%)
1.8%	0.86%
High-vulnerable	
9.4%	0.24%
Mid-vulnerable	
52.3%	0.03%
Low-vulnerable	
36.3%	0.0%
Zero-vulnerable	

Fault Characterization at CRITICAL Region

Type of undervolting faults:

Permanent faults at specific voltage

- There is no considerable change on the rate and location of faults over time.
- Val
- Th
- Fa

s.

Key observations discussed:

1. Fault rate exponentially increases by further undervolting.
2. BRAMs have fully different reliability behavior against undervolting faults.
3. The fault rate and location is deterministic over the time.

Three parameters orthogonally have significant impact on the rate and location of faults:

4. Voltage
5. Temperature
6. Chip

BRAM Fault Rate (%)

(VCCBRAM @Vcrash, T= ambient, chip= VC707)

FVM can be potentially used in fault mitigation techniques!

Location of undervolting faults:

Fault Inclusion Property (FIP)

- FIP: A corrupted bit at a specific voltage stays faulty in lower voltages as well.

- FIP can be used in mitigation techniques.

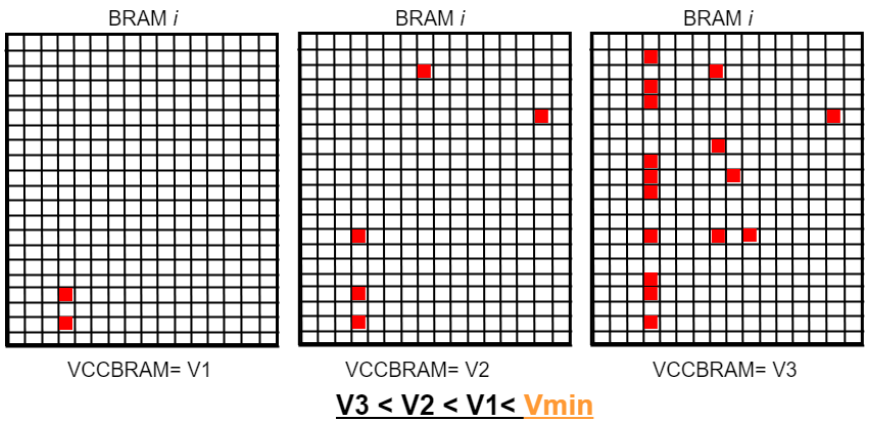
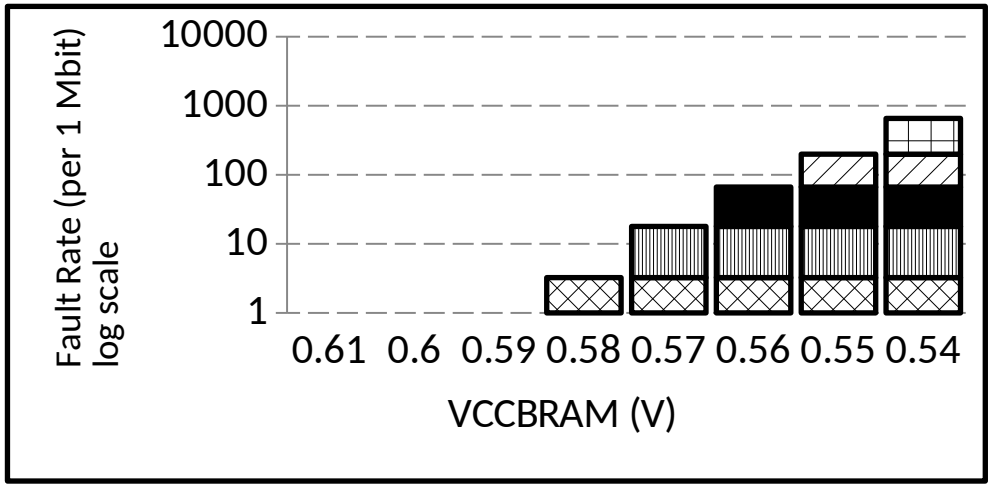


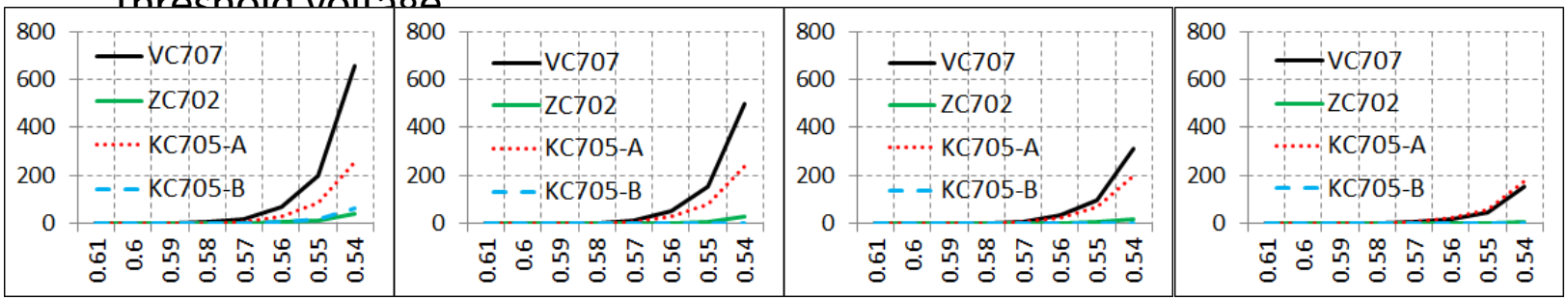
Illustration of FIP



FIP shown as fault rate for VC707

Practical confirmation of Inverse Temperature Dependency (ITD)

- ***Methodology*:** Adjusting environmental temperature, monitoring on-board temperature via PMBus.
- ***Experimental Observation*:**
  - At higher temperatures, fault rate is significantly reduced.
- ***Inverse Temperature Dependency*:**
  - For nano-scale technology nodes, under ultra low-voltage operations, the circuit delay reduces at higher temperatures since supply voltage approaches the threshold voltage.



$T = 50\,^{0}C$ $T = 60\,^{0}C$ $T = 70\,^{0}C$ $T = 80\,^{0}C$

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

(1) Neshatpour, K., Burleson, W., Khajeh, A., & Homayoun, H. (2018). Enhancing Power, Performance, and Energy Efficiency in Chip Multiprocessors Exploiting Inverse Thermal Dependence. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, (4), 778-791.

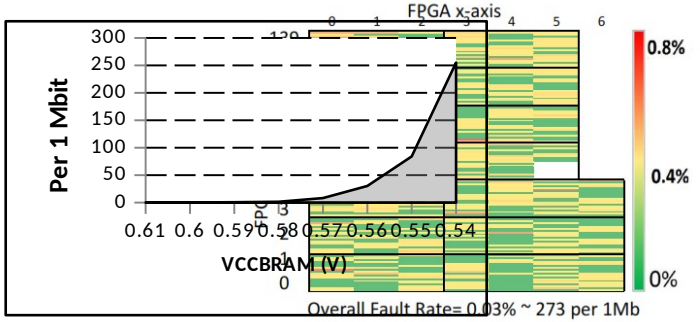
Fault Characterization (**Chip Impacts**)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Even identical samples of same chips have totally different reliability behavior, due to the process variation/aging effects.

- ***Methodology:*** Repeating experiments on two identical samples of KC705 (A&B).
- ***Observations:***
  - Fault rates significantly vary, more than 4X.
  - Fault Variation Maps (FVMs) are entirely different.

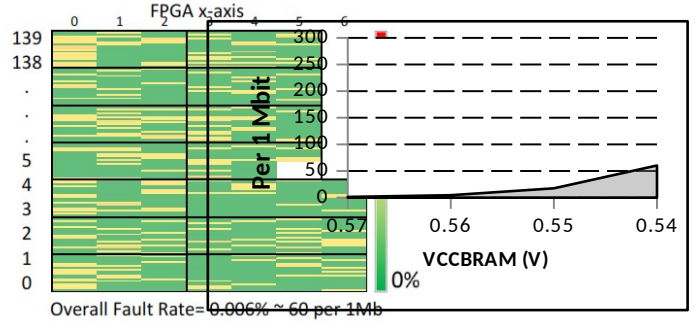
KC705-A



@VCCBRAM= Vcrash

Fault rate Fault location

KC705-B



@VCCBRAM= Vcrash

Fault location Fault rate

Key Points of the Second Contribution

- **Fault rate:** The increase of the fault rate by further undervolting is exponential.
- **Non-uniform fault distribution among BRAMs:** BRAMs do not have similar sensitivity against undervolting.
- **Deterministic behavior of faults:** The location of faults does not change over the time, at certain *voltage* and *temperature*, and for a certain *chip*.
- **Reliability behavior over different voltage levels:** There is Fault Inclusion Property (FIP).
- **Environmental temperature:** At higher temperatures, FPGA BRAMs shows better reliability behavior, *i.e.*, less fault rate.
- **Reliability differences for chips:** Even identical chips shows fully different reliability behaviors.

Contributions

1. Undervolting FPGAs
   - ❑ Voltage guardband
   - ❑ Overall power and reliability trade-off

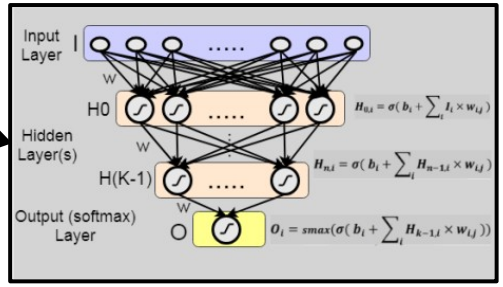
2. Fault characterization in FPGA on-chip memories
   - ❑ Fault type, location, and rate
   - ❑ Temperature, Chip

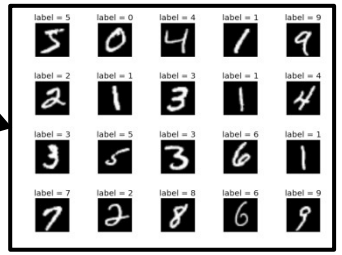
3. Low-voltage FPGA-based Neural Network (NN)
   - ❑ Power consumption and NN accuracy characterization
   - ❑ Fault mitigation techniques
     - ❖ Application-aware technique
     - ❖ Built-in ECC

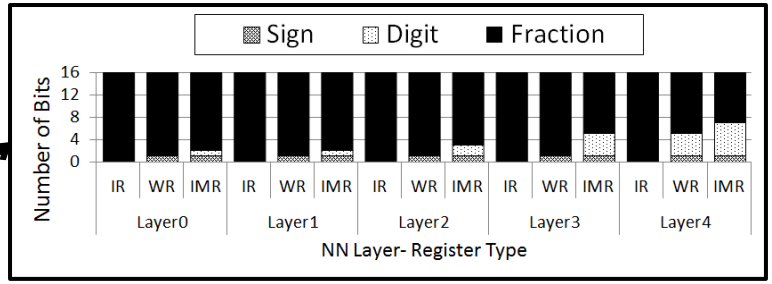
Experimental Methodology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Neural Network (NN)	
Type	Fully-connected classifier
Total number of weights	~1.5 millions
Activation function	Logsig (logarithmic sigmoid)
Major benchmark	
Name-type	MNIST- handwritten digit images
Number of images	Training: 60000, Classification: 10000
Number of pixels per image	28*28=256
Number of output classes	10
Additional benchmarks	
Names	Forest and Reuters
Data representation model	
Type	16-bits fixed-point
Precision	Minimum sign and digit per layer
An example implementation on VC707	
Frequency	100 Mhz
BRAM usage (total: 2060)	70.8%

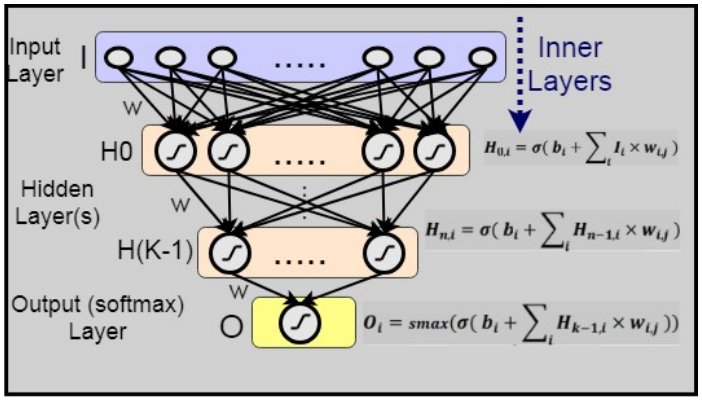




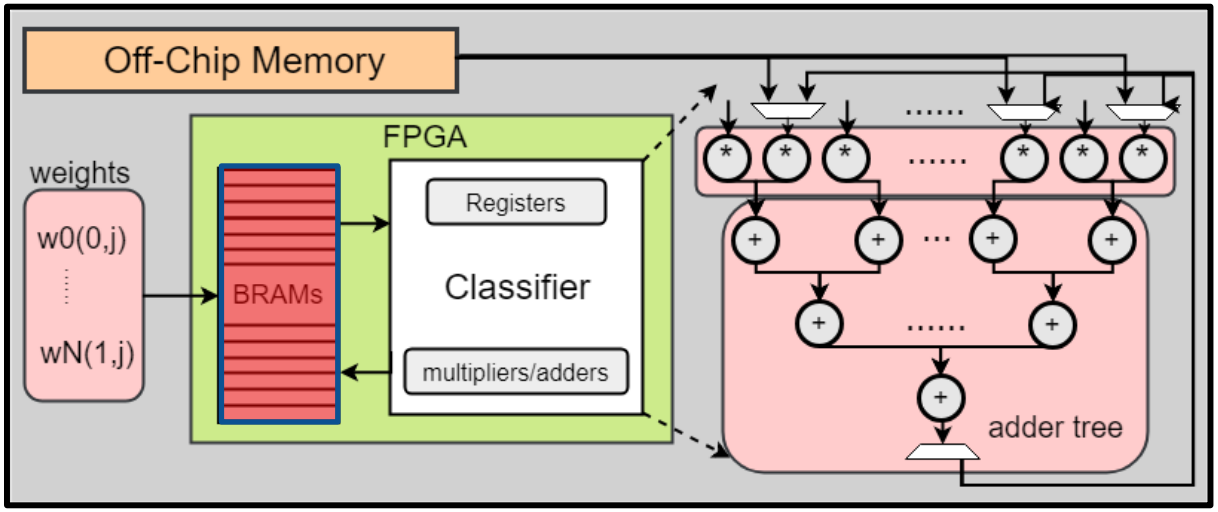


NN Implementation on FPGA

- Input data: off-chip DDR memory.
- Weights: on-chip FPGA BRAM.
- Computation: Streaming data onto DSPs and LUTs.
- We undervolt VCCBRAM:
  - Weights of the NN are potentially affected.



Typical Neural Network (NN)

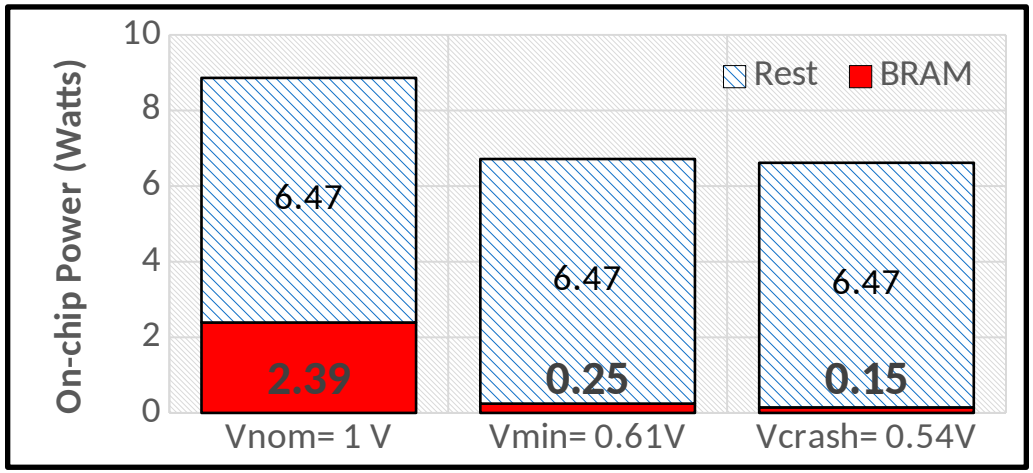


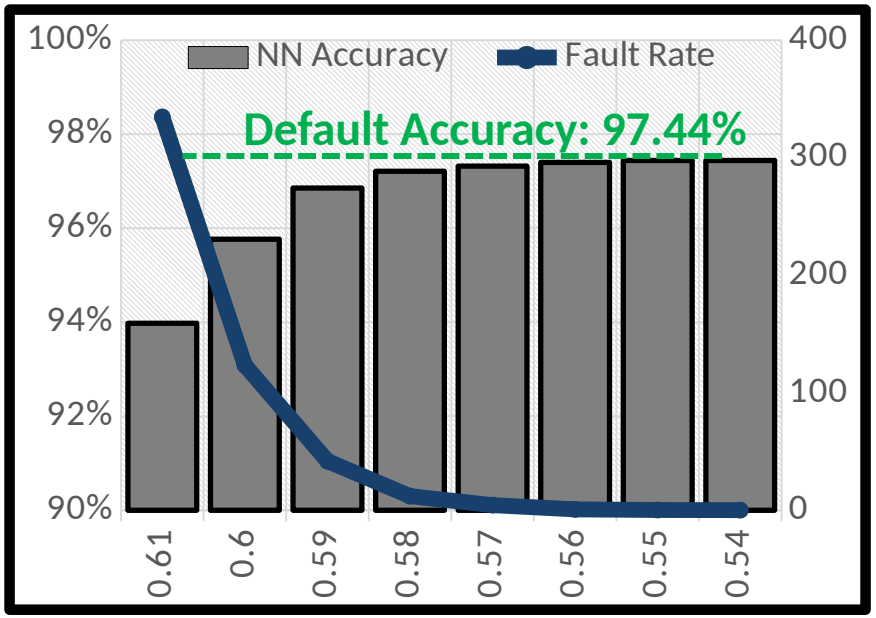
FPGA Implementation

Low-Voltage FPGA-based NN

Power saving

- Significant power reduction until the minimum safe voltage, *i.e.*, *Vmin* (By eliminating the voltage guardband).
- Additional 40% power reduction below the voltage guardband.



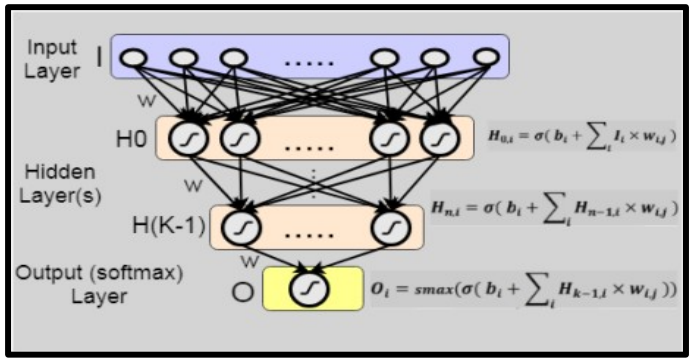


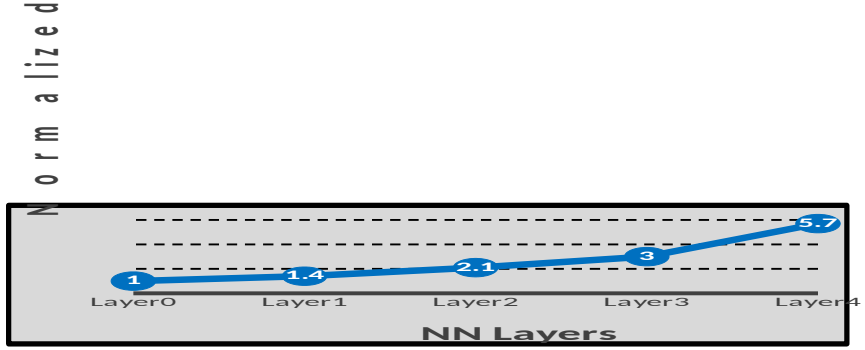
NN accuracy loss

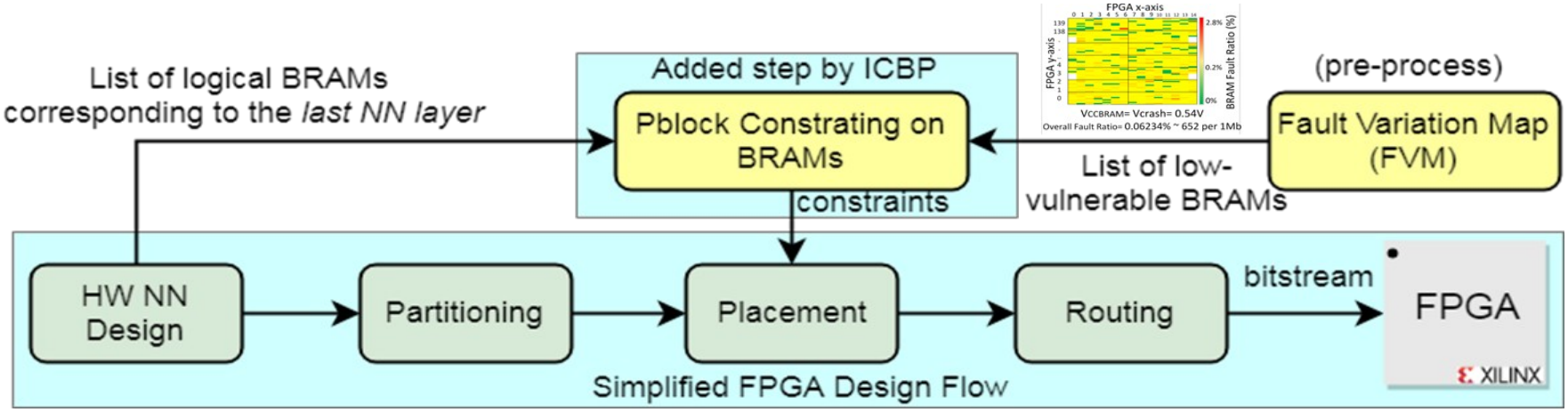
- The NN accuracy exponentially decreases from *97.44%* (inherent accuracy) to *93.86%* through undervolting BRAMs beyond *Vmin*.
- Fault mitigation techniques to prevent the accuracy loss:
  - Application-aware mechanism
  - Built-in ECC

Intelligently-Constrained Memory Mapping (IMM)

- Below voltage guardband level at CRITICAL voltage region, we present IMM to prevent NN classification error rate loss.
- ***Core Idea:*** Map most-sensitive weights to faults into robust BRAMs.
  - **Q:** Which are the most-sensitive NN weights? **A:** Deeper Layers.

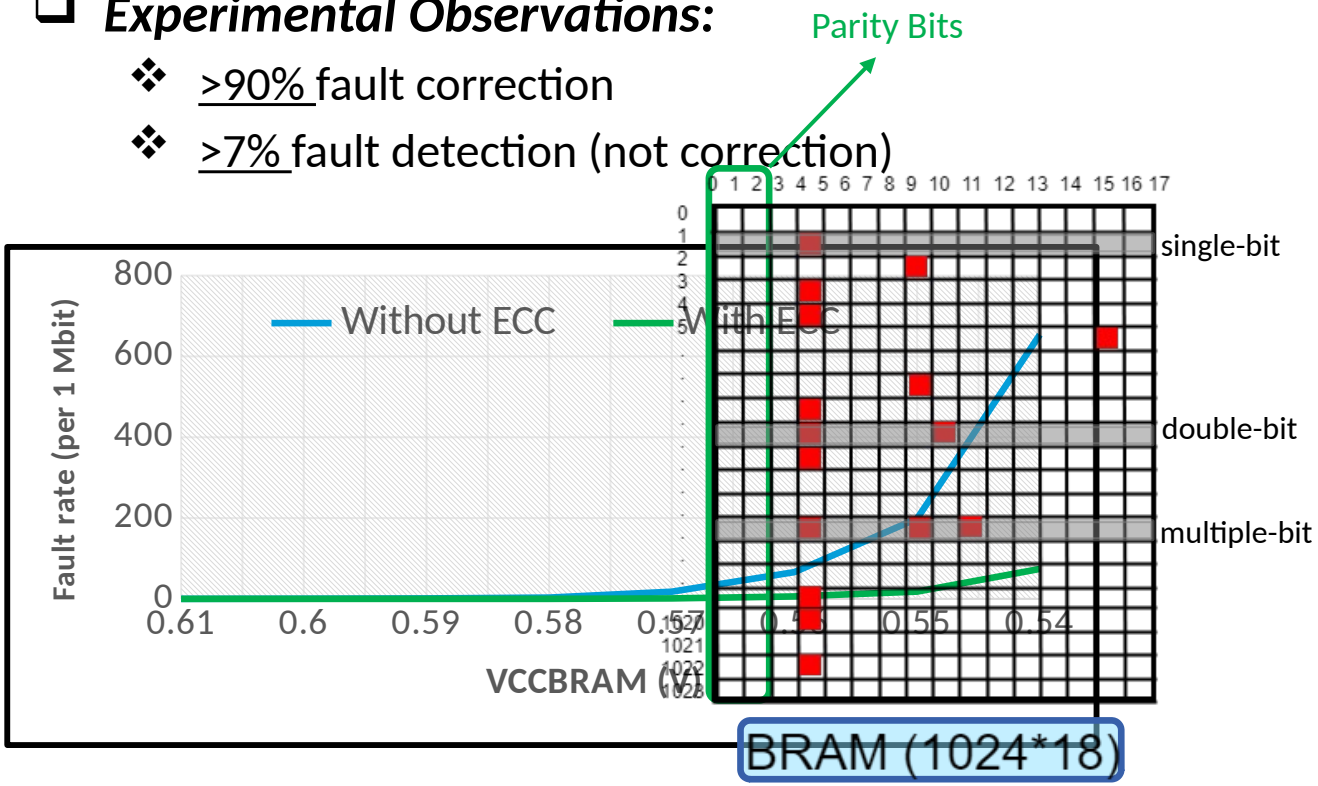






Built-in ECC

- ***Built-in ECC of FPGA BRAMs:***
  - Hamming-code.
  - Two (2) additional bits per row are reserved as parities.
  - SECDED (Single-Error Correction and Double-Error Detection).
- ***Experimental Methodology***:
  - Activate built-in ECC under low-voltage read operations.
- ***Experimental Observations:***
  - >90% fault correction
  - >7% fault detection (not correction)

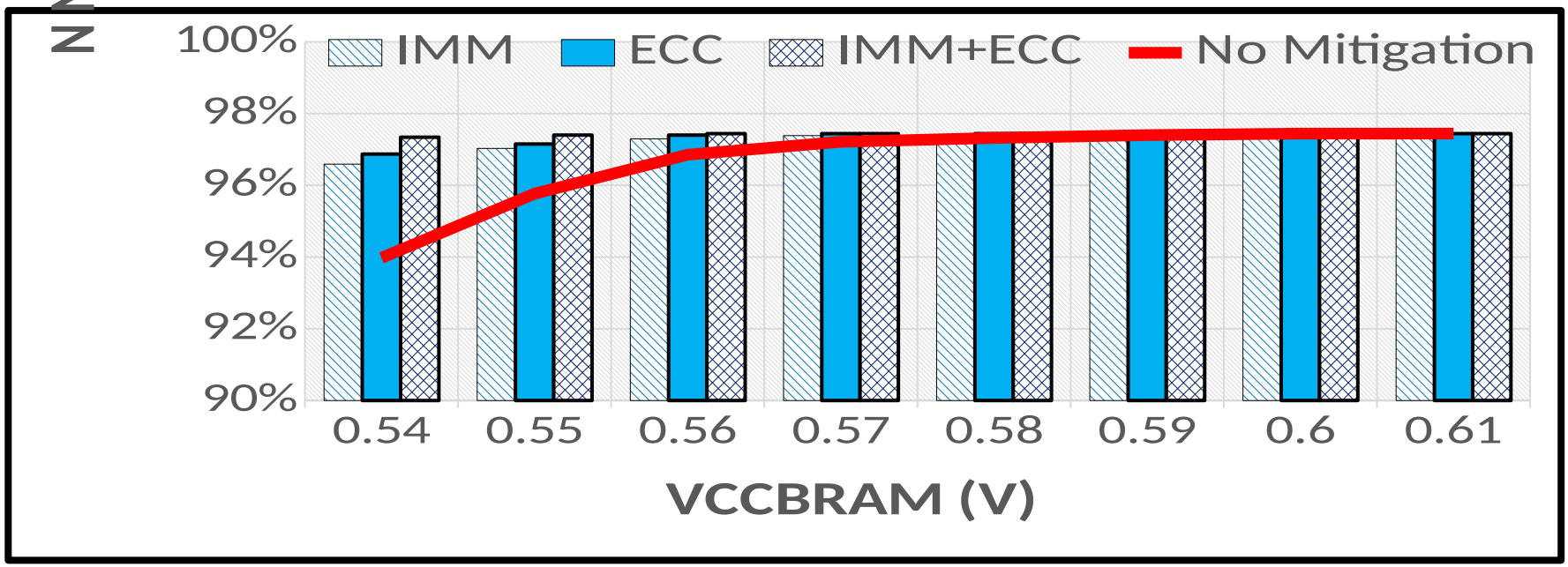


Evaluating Fault Mitigation Techniques

IMM: Exclude High-vulnerable BRAMs

ECC: Cover most of single-bit faults

IMM+ECC: Cover faults in non High-vulnerable BRAMs



Key Points of the Third Contribution

- **Power reduction for FPGA-based accelerates:** Significant energy improvement can be achieved for FPGA-based accelerators (studied for typical NN) through undervolting:
  - By eliminating the voltage guardband
  - By further undervolting in the critical voltage region

- **Cost of undervolting:** Accuracy loss is also significant but controllable at the critical voltage region.

- **Fault mitigation techniques:** According to the fault characterization study, efficient mitigation techniques can be deployed to prevent the NN accuracy loss.

Wrap up

- Summary
- Conclusion
- Ongoing Projects
- ….

Summary

- We ***experimentally*** showed how Xilinx FPGAs work under aggressive low-voltage operations.
- There is a ***conservative voltage guardband*** below the nominal voltage level, *i.e.*, V_{nom}.
- BRAMs ***power*** significantly reduces through undervolting; however, ***reliability*** degrades below the minimum safe voltage, *i.e.*, V_{min}.
- We ***characterized*** the behavior of undervolting faults at the critical region.
- We evaluated FPGA undervolting for a ***typical NN accelerator***.

Conclusion

- There is significant potential in commercial FPGAs to improve the energy efficiency through aggressive undervolting.
  - By eliminating the conservative voltage guardband
  - By further undervolting into the voltage critical region

- Undervolting faults manifest deterministic behaviors.

- Efficient fault mitigation techniques can be deployed which can allow to further energy saving.

- State-of-the-art FPGA-based accelerators can be adapted by undervolting approach.

Constraints of the Xilinx FPGAs for Undervolting

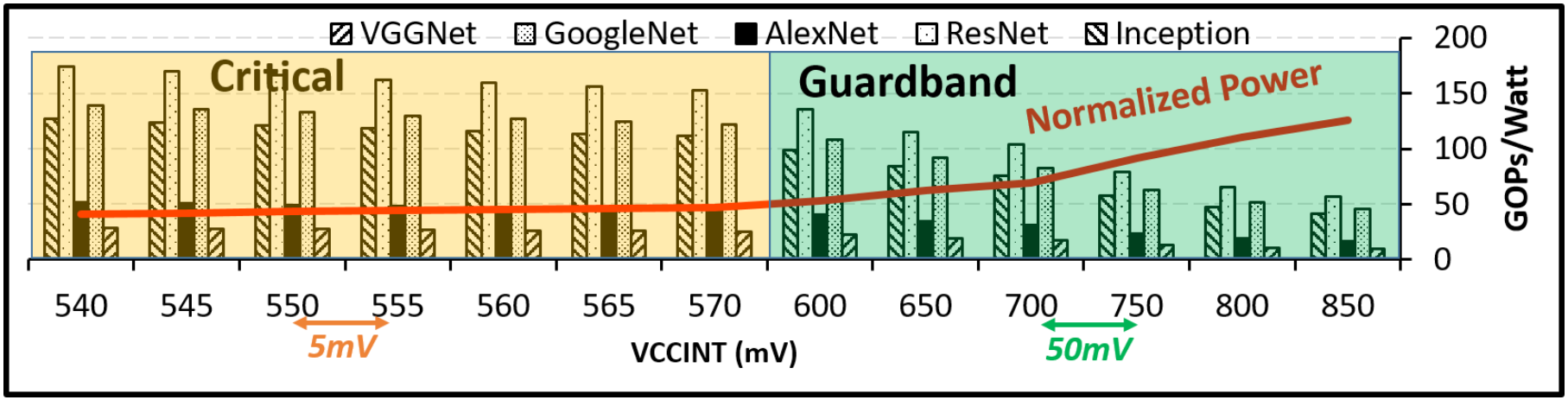
- Many FPGA platforms, *e.g.*, Zynq *are not equipped* with voltage scaling capability.
- There is *no standard* about the voltage distribution among platform components.
- In Xilinx products, voltage regulators are *hardwired* to the host through PMBus interface.
- In many cases, several components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

Ongoing Projects (1)

Title: Low-voltage and fault-resilient FPGA-based accelerators for Convolutional Neural Networks (CNNs), in training and inference.

Goals: Improving the energy-efficiency of FPGA-based CNNs through aggressive undervolting (*for both computing elements and on-chip memories*)

Preliminary Results: more than 2.3X improvement achieved!



Ongoing Projects (2)

Title: Run-time voltage underscaling for task-based programming models.

Goals: Extending OmpSs@FPGA framework (compiler and programming models) to support Dynamic Voltage Scaling (DVS).

Preliminary Experiments: Evaluations for the matrix multiplier on OmpSs@FPGA is ongoing.

Ongoing Projects (3)

Title: Evaluating voltage underscaling in modern DRAMs, *i.e.*, High-bandwidth Memory (HBM) and SRAMs, *i.e.*, UltraRAM.

Goals: Efficient implementation of DNN on UltraRAM and improving its energy-efficiency through aggressive undervolting on *Xilinx VCU128*.

Potentials for Collaborations around low-voltage FPGAs

- Real-world applications (Personalized medicine, Engineering simulations, …)
- FPGA clusters
- Harsh environments (Soft errors, Temperature, Humidity, …) and reliability techniques (TMR, ECC, Checkpointing, …)
- Stochastic and approximate computing

For More Information ....

- *Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman*, "Comprehensive Evaluation of Supply Voltage Underscaling in FPGA on-chip Memories.", in *51st Annual IEEE/ACM International Symposium on Microarchitecture (**MICRO**), 2018.*

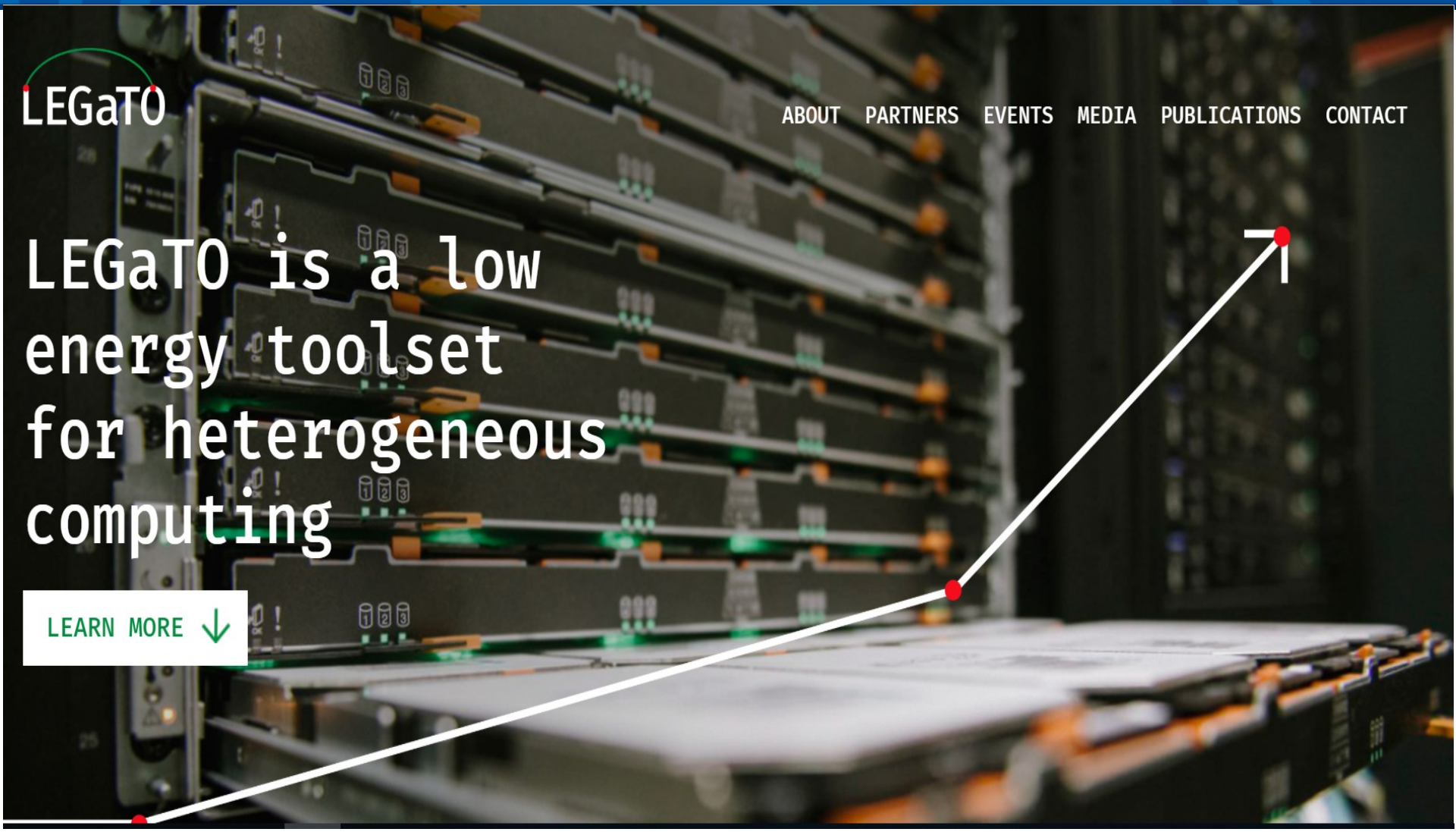
- *Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman*, "Fault Characterization Through FPGAs Undervolting.", in *28th International Conference on Field Programmable Logic & Applications (**FPL**), 2018.*

- *Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman*, "Evaluating Built-in ECC of FPGA on-chip Memories for the Mitigation of Undervolting Faults.", in *27st Euromicro International Conference of on Parallel, Distributed, and Network-based Processsing (**PDP**), 2019.*

- *Behzad Salami, Osman S. Unsal, and Adrian Cristal Kestelman*, "On the Resilience of RTL NN Accelerators: Fault Characterization and Mitigation.", in *30th International Symposium on Computer Architecture and High Performance Computing (**SBAC-PAD**), 2018.*

- *Dimitris Gizopoulos, George Papadimitriou, Athanasios Chatzidimitriou, Vijay Janapa Reddi, Behzad Salami, Osman S. Unsal, Adrián Cristal Kestelman, Jingwen Leng*, "Modern Hardware Margins: CPUs, GPUs, FPGAs Recent System-Level Studies." in *25th IEEE International Symposium on On-Line Testing and Robust System Design (**IOLTS**), 2019.*

LEGaTO



https://legato-project.eu/

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thanks!

Any Question/Comment?

Contact:
Behzad Salami
behzad.salami@bsc.es