Experimental Evaluation of Supply Voltage Underscaling in FPGAs

Presentation by: Leonardo Bautista-Gomez
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.
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.

- Nominal voltage reduction of FPGAs is naturally applied for different generations.
FPGA BRAMs:
- Hierarchy of set of bit-cells distributed over the chip.
- Size of each BRAM: 16-kbits

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

Floorplan of VC707 (2060 BRAMs)
Overall Trade-offs on BRAMs - Power & Reliability

VC707

Fault Rate (per 1 Mbit)
### NN Use case: Experimental Methodology

**Neural Network (NN)**

<table>
<thead>
<tr>
<th>Type</th>
<th>Fully-connected classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of weights</td>
<td>~1.5 millions</td>
</tr>
<tr>
<td>Activation function</td>
<td>Logsig (logarithmic sigmoid)</td>
</tr>
</tbody>
</table>

**Major benchmark**

<table>
<thead>
<tr>
<th>Name-type</th>
<th>MNIST- handwritten digit images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of images</td>
<td>Training: 60000, Classification: 10000</td>
</tr>
<tr>
<td>Number of pixels per image</td>
<td>28*28=256</td>
</tr>
<tr>
<td>Number of output classes</td>
<td>10</td>
</tr>
</tbody>
</table>

**Additional benchmarks**

<table>
<thead>
<tr>
<th>Names</th>
<th>Forest and Reuters</th>
</tr>
</thead>
</table>

**Data representation model**

<table>
<thead>
<tr>
<th>Type</th>
<th>16-bits fixed-point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>Minimum sign and digit per layer</td>
</tr>
</tbody>
</table>

**An example implementation on VC707**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>100 Mhz</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAM usage (total: 2060)</td>
<td>70.8%</td>
</tr>
</tbody>
</table>
Low-Voltage FPGA-based NN

Power saving

- Significant power reduction until the minimum safe voltage, *i.e.*, $V_{\text{min}}$ (By eliminating the voltage guardband).
- Additional 40% power reduction below the voltage guardband.

NN accuracy loss

- The NN classification error exponentially increases from 2.56% (inherent classification error) to 6.74% through undervolting BRAMs beyond $V_{\text{min}}$.
- Fault mitigation techniques to prevent the accuracy loss:
  - Application-aware mechanism
  - Built-in ECC
Fault Mitigation: 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)
ECC for NN Accelerator

**Pros:**
- Significant accuracy loss prevention.
- Negligible power and performance overhead.

**Cons:**
- Requires larger data rows/lines.
- Not all FPGAs are equipped with this technique.

---

**Area Utilization (%)**

<table>
<thead>
<tr>
<th></th>
<th>BRAM</th>
<th>LUT</th>
<th>FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without ECC</td>
<td>96%</td>
<td>3%</td>
<td>0.25%</td>
</tr>
<tr>
<td>With ECC</td>
<td>100%</td>
<td>12%</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

**BRAM Power (W)**

<table>
<thead>
<tr>
<th></th>
<th>Vnom= 1V</th>
<th>Vmin= 0.61V</th>
<th>Vcrash= 0.54V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without ECC</td>
<td>2.4</td>
<td>0.31</td>
<td>0.198</td>
</tr>
<tr>
<td>With ECC</td>
<td>----</td>
<td>----</td>
<td>0.211</td>
</tr>
</tbody>
</table>

ECC efficiency to prevent NN accuracy loss

ECC area and power costs
Ongoing/Future Works

- Different **computing paradigms**, e.g., Heterogeneous Computing, Approximate Computing, Stochastic Computing, among others.

- Generalizing observations by extending the experiments to other FPGA vendors like *Intel/Altera*.

- Evaluation the undervolting in *noisy and harsh environments*.

- More advanced designs, where other components such as *I/O and DSP* are undervolted.

- Application profiling to analyze *workload-to-workload* variation.

- *Dynamic V_{min} prediction and scaling*, adapted by frequency and temperature.
For More Information ....


- 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 Processsng (PDP), 2019.

Thanks!

Contact:
behzad.salami@bsc.es
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.*, $V_{min}$ 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)
Voltage Scaling Capability in Xilinx

Voltage distribution on Xilinx platforms

- VC707
  - VCCINT: 1V, 1.8V, 3.3V, 0-3.3V, 2.5V, 1.5V, 1V, 1.2V
  - VCCBRAM: 2V, 1V

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.
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
Overall Voltage Behavior

- **SAFE**
  - No observable fault
  - Voltage Guardband below $V_{\text{nom}}$

- **CRITICAL**
  - Faults manifest
  - Below $V_{\text{min}}$, min safe voltage

- **CRASH**
  - FPGA stops operating below $V_{\text{crash}}$, 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)**.
Overall Trade-offs on BRAMs - Multiple Platforms

**ZC702**

- **VCCBRAM (V)**:
  - $V_{nom} = 1 V$
  - $V_{min} = 0.59 V$
  - $V_{crash} = 0.53 V$

- **Fault Rate (per 1 Mbit)**
  - 0
  - 50
  - 100
  - 200

- **BRAM Power (mWatts)**
  - 0
  - 2
  - 4

- **BRAM Power (Watts)**
  - 0.05
  - 0.1
  - 0.15

**VC707**

- **VCCBRAM (V)**:
  - $V_{nom} = 1 V$
  - $V_{min} = 0.59 V$
  - $V_{crash} = 0.54 V$

- **Fault Rate (per 1 Mbit)**
  - 80
  - 60
  - 40
  - 20

- **BRAM Power (Watts)**
  - 0.05
  - 0.1
  - 0.15

- **BRAM Power (mWatts)**
  - 0
  - 2
  - 4

**KC705-A**

- **VCCBRAM (V)**:
  - $V_{nom} = 1 V$
  - $V_{min} = 0.59 V$
  - $V_{crash} = 0.53 V$

- **Fault Rate (per 1 Mbit)**
  - 0
  - 50
  - 100
  - 200

- **BRAM Power (mWatts)**
  - 0
  - 2
  - 4

- **BRAM Power (Watts)**
  - 0.05
  - 0.1
  - 0.15

**KC705-B**

- **VCCBRAM (V)**:
  - $V_{nom} = 1 V$
  - $V_{min} = 0.57 V$
  - $V_{crash} = 0.54 V$

- **Fault Rate (per 1 Mbit)**
  - 80
  - 60
  - 40
  - 20

- **BRAM Power (Watts)**
  - 0.05
  - 0.1
  - 0.15

- **BRAM Power (mWatts)**
  - 0
  - 2
  - 4
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.*, VCCBRAM and VCCINT.

- **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.

**K-means clustering**

<table>
<thead>
<tr>
<th>%BRAMs</th>
<th>Average Fault Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8%</td>
<td>0.86%</td>
</tr>
<tr>
<td>9.4%</td>
<td>0.24%</td>
</tr>
<tr>
<td>52.3%</td>
<td>0.03%</td>
</tr>
<tr>
<td>36.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**VC707** (2060 BRAMs)
VCCBRAM@ \(V_{\text{crash}}=0.54\)V
Temperature@ Ambient
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.
- Validated by repeating experiments for 100 times.
- The physical location of BRAMs is extracted using Vivado.

**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:**

1. Voltage
2. Temperature
3. Chip

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
Fault Characterization (Temperature Impacts)

**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 (ITD)**
  - For nano-scale technology nodes, under ultra low-voltage operations, the circuit delay reduces at higher temperatures since supply voltage approaches the threshold voltage.

\[ T = 50 \, ^0\text{C} \quad T = 60 \, ^0\text{C} \quad T = 70 \, ^0\text{C} \quad T = 80 \, ^0\text{C} \]

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

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.

---

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.

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](image1.png) ![KC705-B](image2.png)
Detailed Fault Characterization

- **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
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.
Below voltage guardband level at **CRITICAL** voltage region, we present ICBP 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.
ICBP Evaluation

- **Pros:**
  - Significant accuracy loss prevention.
  - No power and performance overhead.

- **Cons:**
  - Needs the FVM as a pre-process step → Built-in ECC is evaluated without having this cost.

![Graph showing BRAMs Power and NN Classification Error](image)

- **Inherent NN Error:** 2.56%
- **NN Error by Default Placement**
- **NN Error by ICBP**
- **BRAM Power**
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
- Potential Next Steps
- ....
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.
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**.
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.
LEGaTO is a low energy toolset for heterogeneous computing

https://legato-project.eu/