An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network Acceleration

Behzad Salami  Baturay Onural  Ismail Yuksel
Fahrettin Koc  Oguz Ergin  Adrian Cristal
Osman Unsal  Hamid Sarbazi-Azad  Onur Mutlu

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

TOBB ETÜ
Ekonomi ve Teknoloji Üniversitesi

ETH Zürich

50th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 30th June, 2020
Executive Summary

- **Motivation**: Power consumption of neural networks is a main concern
  - **Hardware acceleration**: GPUs, FPGAs, and ASICs

- **Problem**: FPGAs are at least 10X less **power-efficient** than equivalent ASICs

- **Goal**: Bridge the **power-efficiency gap** between ASIC- and FPGA-based neural networks by **Undervolting below nominal level**

- **Evaluation Setup**
  - 5 Image classification workloads
  - 3 Xilinx UltraScale+ ZCU102 platforms
  - 2 On-chip voltage rails

- **Main Results**
  - Large voltage guardband (i.e., 33%)
  - >3X power-efficiency gain
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Motivation and Background</td>
</tr>
<tr>
<td>• Our Goal</td>
</tr>
<tr>
<td>• Methodology</td>
</tr>
<tr>
<td>• Results</td>
</tr>
<tr>
<td>- Overall Voltage Behavior</td>
</tr>
<tr>
<td>- Power-Reliability Trade-off</td>
</tr>
<tr>
<td>- Frequency Underscaling</td>
</tr>
<tr>
<td>- Environmental Temperature</td>
</tr>
<tr>
<td>• Prior Works</td>
</tr>
<tr>
<td>• Summary, Conclusion, and Future Works</td>
</tr>
</tbody>
</table>
## Outline

- **Motivation and Background**
  - Our Goal
  - Methodology
  - Results
    - Overall Voltage Behavior
    - Power-Reliability Trade-off
    - Frequency Underscaling
    - Environmental Temperature
  - Prior Works
- **Summary, Conclusion, and Future Works**
Motivation and Background

• Motivation
  ✓ **Power consumption** of neural networks is a main concern
  ✓ **Hardware acceleration**: GPUs, FPGAs, and ASICs
  ✓ **FPGAs**: Getting popular but **less power-efficient** than equivalent ASICs
  ✓ **Large voltage guardbands** (12-35%) for CPUs, GPUs, DRAMs
  ✓ *Any potential of “Undervolting FPGAs” for power-efficiency of neural networks?*

• Background
  ✓ **Neural Networks**: Widely deployed with an **inherent resilience** to errors
  ✓ **FPGAs**: Higher throughput than **GPUs** and better flexibility than **ASICs**
  ✓ **Undervolting**: Reduces **power cons.**, may incur **reliability** or **performance** issues
Outline

• Motivation and Background
• Our Goal
• Methodology
• Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
• Prior Works
• Summary, Conclusion, and Future Works
Our Goal

• **Primary Goal**
  ✓ Bridge the *power-efficiency gap* between ASIC- and **FPGA-based neural networks** by:
    ➢ **Undervolting** (i.e., underscaling voltage *below nominal level*)

• **Secondary Goals**
  ✓ Study the *voltage behavior* of real FPGAs (e.g., *guardband*)
  ✓ Study the *power-efficiency gain* of undervolting for neural networks
  ✓ Study the *reliability overhead*
  ✓ Study the *frequency underscaling* to prevent the accuracy loss
  ✓ Study the effect of *environmental temperature*
# Outline

- Motivation and Background
- Our Goal
- Methodology
- Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
- Prior Works
- Summary, Conclusion, and Future Works
Overall Methodology

- **CNN image classification workloads**, i.e., VGGNet, GoogleNet, AlexNet, ResNet50, Inception.

- Xilinx DNNDK to map CNN into FPGA
  - By default optimized for INT8

- **3** identical samples of Xilinx ZCU102
  - ZYNQ Ultrascale+ architecture
  - Hard-core ARM for data orchestration
  - FPGA for CNN acceleration

- **2** on-chip voltage rails, via PMBus
  - $V_{CCINT}$: DSPs, LUTs, buffers, ...
  - $V_{CCBRAM}$: BRAMs
  - $V_{nom} = 850\text{mV}$ (set by manufacturer)

Vast majority (>99.9%) of the power is dissipated on $V_{CCINT}$
Outline

- Motivation and Background
- Our Goal
- Methodology
- Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
- Prior Works
- Summary, Conclusion, and Future Works
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Motivation and Background</td>
</tr>
<tr>
<td>• Our Goal</td>
</tr>
<tr>
<td>• Methodology</td>
</tr>
<tr>
<td>• Results</td>
</tr>
<tr>
<td>- Overall Voltage Behavior</td>
</tr>
<tr>
<td>- Power-Reliability Trade-off</td>
</tr>
<tr>
<td>- Frequency Underscaling</td>
</tr>
<tr>
<td>- Environmental Temperature</td>
</tr>
<tr>
<td>• Prior Works</td>
</tr>
<tr>
<td>• Summary, Conclusion, and Future Works</td>
</tr>
</tbody>
</table>
Overall Voltage Behavior

- **Guardband**: Large region below nominal level ($V_{nom} = 850mV$)
- **Critical**: Narrower region below guardband ($V_{min} = 570mV$)
- **Crash**: FPGA crashes below critical region ($V_{crash} = 540mV$)

- Slight variation of voltage behavior across platforms and benchmarks
Outline

• Motivation and Background
• Our Goal
• Methodology
• Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
• Prior Works
• Summary, Conclusion, and Future Works
Power-Reliability Trade-off

Power-efficiency (GOPs/W) gain
- >3X power saving (2.6X by eliminating guardband and further 43% in critical region)
- Slight variation across 3 platforms and 5 workloads

Reliability overhead (i.e., CNN accuracy loss)
- No accuracy loss in the guardband, accuracy collapse in the critical region
- Slight variation across 3 platforms and 5 workloads
**Outline**

- Motivation and Background
- Our Goal
- Methodology
- Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
- Prior Works
- Summary, Conclusion, and Future Works
**Frequency Underscaling**

- **Simultaneous** frequency underscaling to prevent CNN accuracy collapse in the **critical voltage region**
- For each voltage level below $V_{min}$, we found the $F_{max}$, the **maximum operating frequency** at which there is **no accuracy loss**
- Leads to **performance and energy-efficiency loss**

Best setting for **High-performance** and **Energy-efficiency**

Best setting for **Power-efficiency**

<table>
<thead>
<tr>
<th>VCCINT (mV)</th>
<th>Fmax (Mhz)</th>
<th>GOPs (Norm)</th>
<th>Power (W) Norm</th>
<th>GOPs/W (Norm)</th>
<th>GOPs/J (Norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>570</td>
<td>333</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>565</td>
<td>300</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
<td>0.87</td>
</tr>
<tr>
<td>560</td>
<td>250</td>
<td>0.83</td>
<td>0.84</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td>555</td>
<td>250</td>
<td>0.83</td>
<td>0.78</td>
<td>1.06</td>
<td>0.8</td>
</tr>
<tr>
<td>550</td>
<td>250</td>
<td>0.83</td>
<td>0.75</td>
<td>1.1</td>
<td>0.83</td>
</tr>
<tr>
<td>545</td>
<td>250</td>
<td>0.83</td>
<td>0.74</td>
<td>1.12</td>
<td>0.84</td>
</tr>
<tr>
<td>540</td>
<td>200</td>
<td>0.7</td>
<td>0.56</td>
<td>1.25</td>
<td>0.75</td>
</tr>
</tbody>
</table>

*(Voltage steps= 5mV, Frequency steps= 50Mhz)- shown for GoogleNet*
# Outline

- Motivation and Background
- Our Goal
- Methodology
- Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
- Prior Works
- Summary, Conclusion, and Future Works
Environmental Temperature

- Effects of **environmental temperature** on power-reliability
  - Use **fan speed** to test temperature in [34 °C, 50 °C]
  - On-board temperature monitored by **PMBus**

- Temperature effects on **power consumption**
  - \( \downarrow \text{Temp} \rightarrow \downarrow \text{Power} \) (direct relation of power and temp)
  - By undervolting, the impact of temperature on power consumption **reduces**.

- Temperature effects on **reliability**
  - \( \downarrow \text{Temp} \rightarrow \uparrow \text{Accuracy loss} \) (indirect relation of reliability and temp)
  - In our temperature range, \( V_{\text{min}} \) and \( V_{\text{crash}} \) do **not** change significantly.
Outline

• Motivation and Background
• Our Goal
• Methodology
• Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
• Prior Works
• Summary, Conclusion, and Future Works
Prior Works

**Undervolting**
- Studies for off-the-shelf real CPUs, GPUs, ASICs, DRAMs
- Large voltage guardband (from 12% to 35%) for many devices
- **This work** extends such studies for off-the-shelf FPGAs especially for neural network acceleration and **confirms large guardbands** (i.e., 33%)

**Power-Efficient Neural Networks**
- Studies on architectural-, hardware-, and software-level techniques
- Undervolting in neural network ASIC accelerator (e.g., GreenTPU-DAC’19)
- **This work** proposes a **hardware-level undervolting** for further power-saving (>3X) in FPGAs.

**Reliability in Neural Networks**
- Analytical and simulation-based studies (e.g., Thundervolt-DAC’18)
- Some studies on real hardware (e.g., EDEN-MICRO’19)
- **This work** studies the reliability of neural networks on **real FPGAs** when operating at reduced voltage levels.
Outline

• Motivation and Background
• Our Goal
• Methodology
• Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
• Prior Works
• Summary, Conclusion, and Future Works
Summary, Conclusion, and Future Works

• Summary
  ✓ We improve the **power-efficiency (>3X)** of off-the-shelf FPGAs via **undervolting** for neural network accelerators:
    ➢ 2.6X by eliminating the **guardband (i.e., 33%)** without any cost
    ➢ 43% by further undervolting below the guardband **with the cost of**
      ❖ either **accuracy loss**, when the **frequency is not underscaled**
      ❖ or **performance loss**, when the **frequency is underscaled**

• Conclusion
  ✓ **Undervolting** is an effective way to achieve significant power-saving for FPGA-based neural network accelerators

• Future Works
  ✓ **HW & SW extension of our undervolting** for FPGA clusters and other neural network models and tools
An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network Acceleration

Behzad Salami  Baturay Onural  Ismail Yuksel
Fahrettin Koc  Oguz Ergin  Adrian Cristal
Osman Unsal  Hamid Sarbazi-Azad  Onur Mutlu

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

LEGaTO

TOBB ETÜ
Ekonomi ve Teknoloji Üniversitesi

ETH Zürich

50th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 30th June, 2020