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**

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**

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

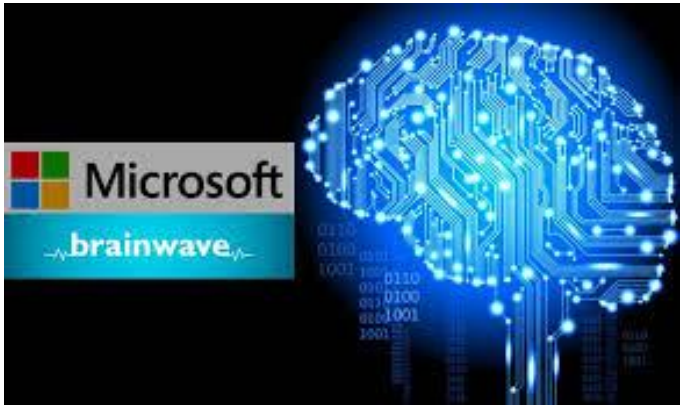
Outline

- 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

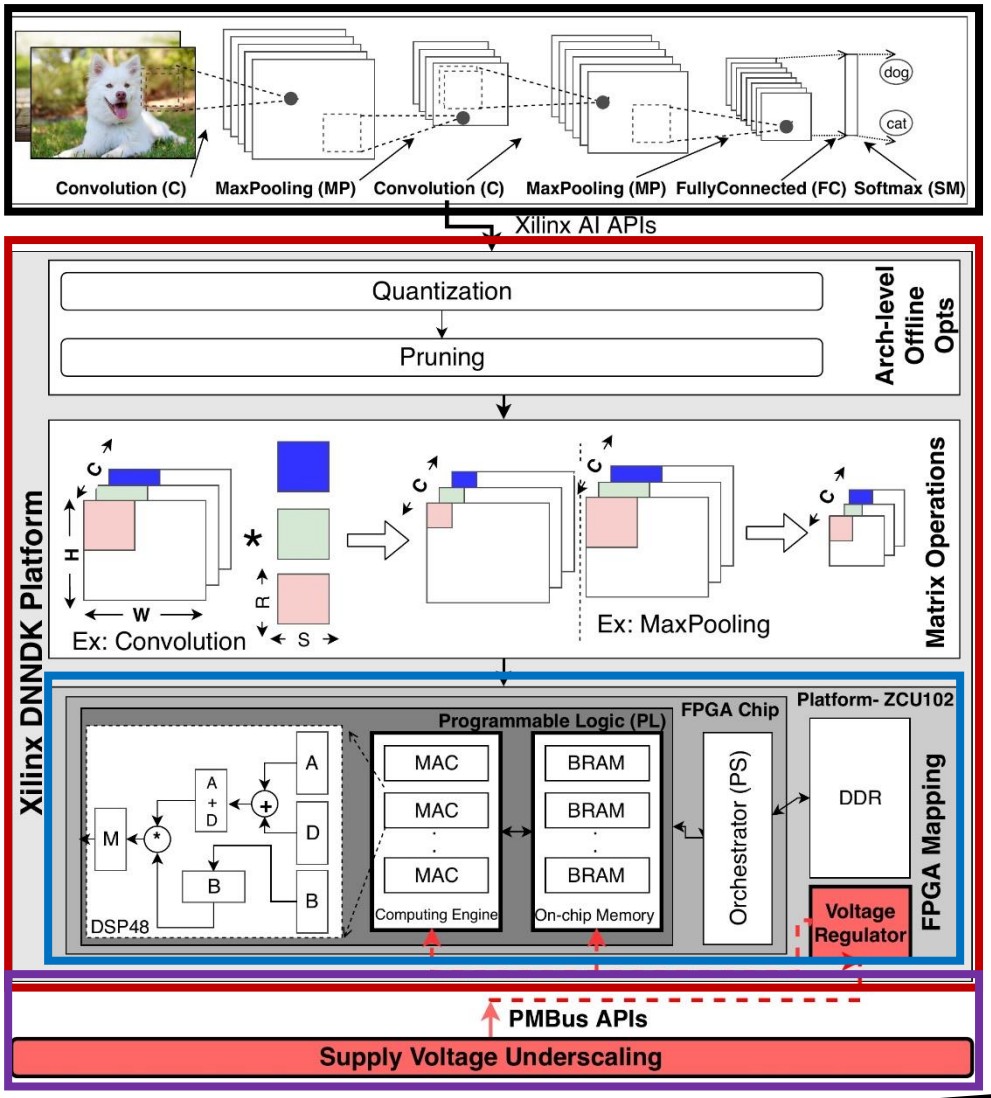
- **Primary Goal**
  - ✓ Bridge the **power-efficiency gap** between ASIC- and **FPGA-based neural networks** by:
    - ➢ **Undervolting** (i.e., underscaling voltage **below nominal level**)

- **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



- **5** 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 Ultrscale+ 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} = 850mV (set by manufacturer)

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

Outline

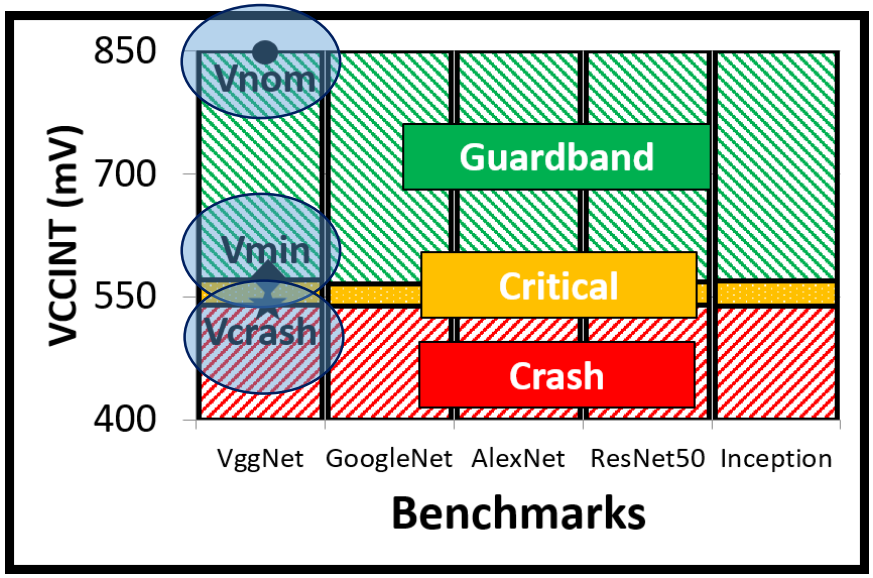
- Motivation and Background
- Our Goal
- Methodology
- Results
  - Overall Voltage Behavior
  - Power-Reliability Trade-off
  - Frequency Underscaling
  - Environmental Temperature
- Prior Works
- Summary, Conclusion, and Future Works

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 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 ($V_{crash} = 540mV$)



Guard band

- **No** performance or reliability loss
- Added by the **vendor** to ensure the worst-case conditions
- **Large** guardband, average of **33%**

Critical

- A **narrow** voltage region
- Neural network **accuracy collapse**

Crash

- FPGA **stops operating**

Slight variation of voltage behavior across platforms and benchmarks

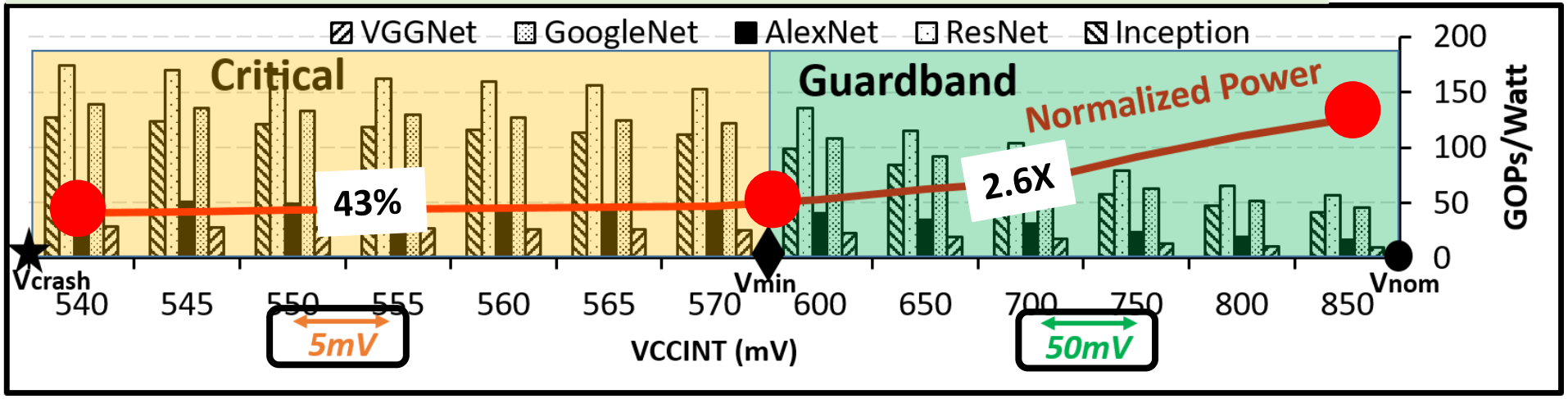
Outline

- 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

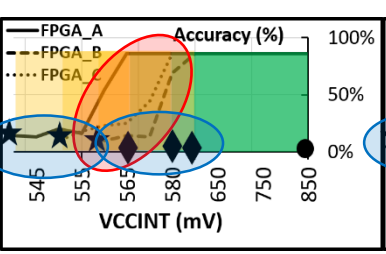
Power-efficiency (GOPs/W) gain

- **>3X** power saving (**2.6X** by eliminating guardband and further **43%** in critical region)
- **Slight variation** across 3 platforms and 5 workloads

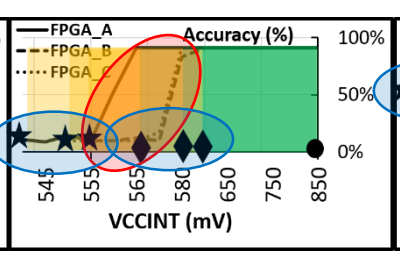


Reliability overhead (i.e., CNN accuracy loss)

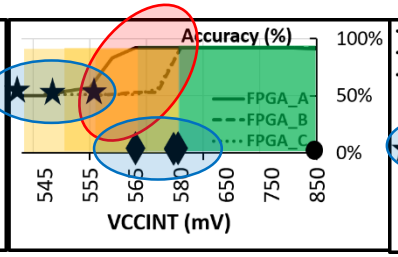
- **No accuracy loss** in the guardband**, accuracy collapse** in the critical region
- **Slight variation** across 3 platforms and 5 workloads



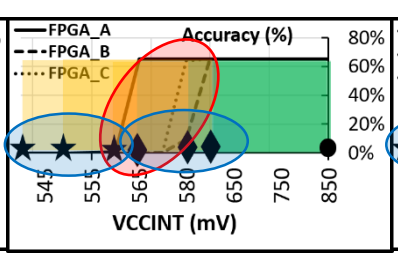
VGGNet



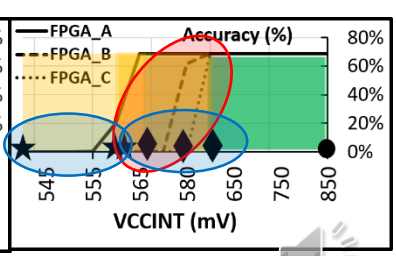
GoogleNet



AlexNet



ResNet



Inception

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**

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

(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

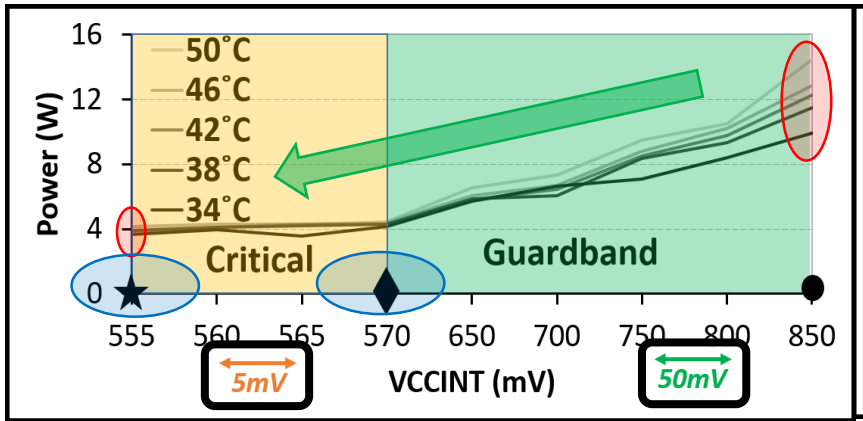
Environmental Temperature

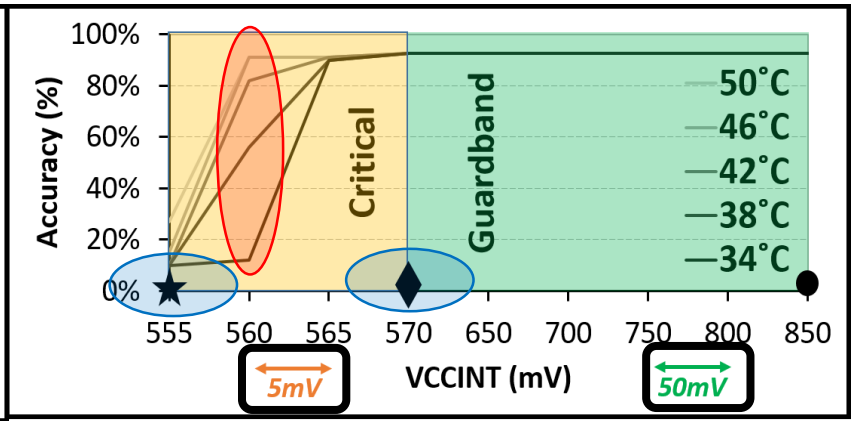
- Effects of **environmental temperature** on power-reliability
  - ✓ Use **fan speed** to test temperature in [34 ℃, 50 ℃]
  - ✓ On-board temperature monitored by **PMBus**

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

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

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

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**

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