Demonstrating Reduced-Voltage FPGA-Based Neural Network Acceleration for Power-Efficiency

Erhan Baturay Onural
TOBB ETÜ

Ismail Emir Yuksel
TOBB ETÜ

Behzad Salami
BSC

This demo aims to demonstrate undervolting below the nominal level set by the vendor for off-the-shelf FPGAs running Deep Neural Networks (DNNs), to achieve power-efficiency. FPGAs are becoming popular [1-4], thanks to their higher throughput than GPUs and better flexibility than ASICs. To further improve the power-efficiency, we propose to employ undervolting below the nominal level (i.e., V_{nom}= *850mV* for studied platform). FPGA vendors usually add a voltage guardband to ensure the correct operation under the worst-case circuit and environmental conditions [5-9]. However, these guardbands can be very conservative and unnecessary for state-of-the-art applications. Reducing the voltage in this guardband region does not lead to reliability issues under normal operating conditions, and thus, eliminating it can result in a significant power reduction for a wide variety of real-world applications. We will experimentally demonstrate a large voltage guardband for modern FPGAs: an average of 33%. Eliminating this guardband leads to significant power-efficiency (*GOPs/W*) improvement, on average, 2.6X, see Figure 1.

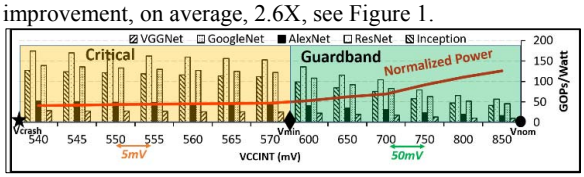

Figure 1: Power-efficiency of FPGA-based DNN via undervolting

With further undervolting below the minimum safe voltage level (i.e., V_{min}= *570mV*), the power-efficiency improves by an extra 43%, leading to a total improvement of more than 3X. This additional gain comes with the cost of DNN accuracy collapse, detailed reliability analysis in [10]. To prevent this accuracy loss, we will demonstrate a simultaneous frequency underscaling. The aim is to find F_{max} at each voltage level, where there is no DNN accuracy loss. This technique reduces the power-efficiency gain from 43% to 25%, see Table 1. Further undervolting, FPGA does not operate, and it crashes at V_{crash} (i.e., *540 mV*).

Table 1: Frequency underscaling in critical voltage region. Best results of each metric are highlighted.

V_{CCINT} (mV)	F_{max} (Mhz)	GOPs (Norm)	Power (Norm)	GOPs/W (Norm)	GOPs/J (Norm)
570	**333**	**1.00**	1.00	1.00	**1.00**
560	**250**	0.83	0.84	0.99	0.75
550	**250**	0.83	0.75	1.10	0.83
540	**200**	0.70	**0.56**	**1.25**	0.75

Figure 2 depicts the overall methodological flow. ***(i) Benchmark***: we evaluate undervolting with five commonly-used image classification benchmarks, as detailed them in [9]. This demonstration is based on AlexNet, with eight layers and a large parameter size of 233.2 MB. ***(ii) Software***: For our implementation, we leverage the Xilinx DNNDK tool in the classification phase of DNNs. ***(iii) Hardware***: Our prototype is based on the Xilinx Ultrascale+ ZCU102. We will demonstrate undervolting V_{CCINT} that is supplying the internal FPGA components.

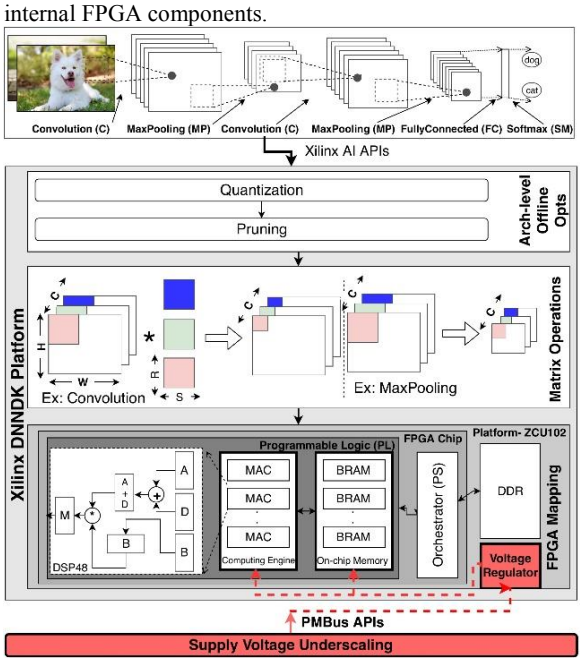

Figure 2: Overall methodology

ACKNOWLEDGMENT

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.

REFERENCES

[1] B. Salami, et al., "HATCH: Hash Table Caching in Hardware for Efficient Relational Join on FPGA", in FCCM, 2015.

[2] B. Salami, et al., "AxleDB: A Novel Programmable Query Processing Platform on FPGA", in MICPRO, 2017.

[3] B. Salami, et al., "Accelerating hash-based query processing operations on FPGAs by a hash table caching technique", in CARLA, 2016.

[4] O. Arcas-Abela, "Hardware Acceleration for Query Processing: Leveraging FPGAs, CPUs, and Memory", in CISE, 2016.

[5] B. Salami, et al., "Comprehensive Evaluation of Supply Voltage Underscaling in FPGA On-chip Memories", in MICRO, 2018.

[6] B. Salami, et al., "Fault Characterization Through FPGA Undervolting", in FPL, 2018.

[7] B. Salami, et al., "Evaluating built-in ECC of FPGA on-chip memories for the mitigation of undervolting faults", in PDP, 2019.

[8] K. Givaki, et al., "On the Resilience of Deep Learning for Reduced-voltage FPGAs", in PDP, 2020.

[9] B. Salami, et al., "An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network Acceleration" in DSN, 2020.

[10] B. Salami, et al., "On the Resilience of RTL NN Accelerators: Fault characterization and mitigation", in SBAC-PAD, 2018.