Can We Trust Undervolting in FPGA-based Deep Learning Designs at Harsh Conditions?

Fahrettin Koc
TOBB University of Economics and Technology

Behzad Salami
Barcelona Supercomputing Center

Oguz Ergin
TOBB University of Economics and Technology

Osman Unsal
Barcelona Supercomputing Center

Adrian Cristal Kestelman
Barcelona Supercomputing Center

***Abstract*—As more Neural Networks on Field Programmable Gate Arrays (FPGAs) are used in a wider context, the importance of power efficiency increases. However, the focus on power should never compromise application accuracy. One technique to increase power efficiency is reducing the FPGAs' supply voltage (”undervolting”), which can cause accuracy problems. Therefore, careful design-time considerations are required for correct configuration without hindering the target accuracy. This fact becomes especially important for autonomous systems, edge-computing, or data-centers.**

This study reveals the impact of undervolting in harsh environmental conditions on the accuracy and power efficiency of convolutional neural network benchmarks. We perform comprehensive testing in a calibrated infrastructure at controlled temperatures (between -40℃ and 50℃) and four distinct humidity levels (50%, 60%, 70%, 80%) for off-the-shelf FPGAs. We show that the voltage guard-band shift with temperature is linear and propose new reliable undervolting designs providing a 65% increase in power efficiency (GOPS/W).

■ ARTIFICIAL INTELLIGENCE and image processing applications are widely deployed in various modern use-cases, from edge computing and data-centers to end-user products. Image or video processing (recognition, analysis, and classification) applications often use deep learning algorithms, which are based on multi-layered neural networks. Currently, the most effective and preferred deep learning algorithm is Convolutional Neural Networks (CNNs), which employ filtering or abstraction of data, weight sharing among layers, and pooling to increase the performance. Usually, different types of hardware (e.g., GPUs, FPGA, ASICs) are leveraged to accelerate CNN applications [1], [2]. Among them, GPUs are relatively more flexible to support different types of CNNs but are less-efficient, while ASIC accelerators are relatively higher-throughput but less

DOI 10.1109/MM.2022.3153891

configurable. FPGAs provide the best flexibility-efficiency trade-off, and accordingly, are becoming a popular hardware substrate for CNNs, especially for embedded systems and edge/IoT devices. For such power-critical scenarios, it is necessary to reduce power consumption. Undervolting, where the voltage supply is decreased below the nominal value while the frequency is unchanged, is one of the effective power-efficiency techniques ([2], [3], [4]) and has been recently proposed for FPGA-based CNN accelerators.

There is significant potential for energy reduction through the undervolting of compute fabrics such as FPGAs. This is because the vendors adopt a one size fits all approach and include large voltage margins or guard-bands to ensure that the hardware can work even in the worst case for any FPGA. Undervolting can provide much better power efficiency than the baseline design, but the system needs to be designed carefully by considering the trade-off between power efficiency and accuracy/reliability [2], [5].

Power efficiency and reliability are both important for CNNs. For example, the battery life of an autonomous electric car is vital. However, no one wants to encounter an accident or a failure in pedestrian detection due to loss of accuracy or performance degradation in that vehicle's FPGA-based CNN application [6]. On the other side, environmental conditions can affect the reliability and power-efficiency of such a system, as well [7], [8], [9]. These conditions can be harsh, especially if the deployment is for a critical use-case such as an autonomous car.

Based on the above motivations, and for the first time, this paper comprehensively performs an experimental study of such power-reliability trade-offs under harsh thermal and humidity environments for FPGA-based CNN accelerators. We present best practices for optimizing the supply voltage in harsh environmental conditions where the reliability of the FPGA-accelerated CNN applications needs to be preserved. We apply undervolting and reduce the core supply voltage of FPGA (i.e., Vccint), which the on-chip components (Digital Signal Processors, Look-Up Tables, buffers, and other internal components) use commonly. We perform our experiments on three well-known CNN image classification benchmarks: VGG Net, LeNet, GoogleNet [10].

This paper expands our previous conference paper [2], with the following new contributions. We utilize a precise and calibrated test infrastructure to characterize the voltage and accuracy relationship as well as the optimal power efficiency (i.e., the maximum power efficiency possible without compromising accuracy) over a wide temperature range (-40 to 50°C) for multiple CNN applications. We establish that different CNN applications have unique minimal voltage curves over the temperature range, and we show that different FPGA boards have different minimal voltage scalability over this temperature range. Additionally, we experimentally show the impact of humidity on CNN accuracy in undervolted FPGAs. We show how each voltage level at different temperatures affects accuracy at four distinct humidity levels (50%, 60%, 70%, 80%).

In [2], we performed thermal tests at only six points between 32°C and 50°C by adjusting the fan speed without a controlled infrastructure. The tests were conducted for only one CNN application and using only one board. In addition, this former characterization study didn't present any undervolting designs which optimizes power-efficiency.

Additionally in this work, we propose the following three new designs, which help to answer the question in the title; we can indeed trust undervolting in FPGA-based CNN accelerators at harsh conditions:

- We observe that the minimum voltage that does not affect the accuracy depends on the temperature. For instance, while the critical voltage is 560mV at room temperature, it becomes 590mV at the lowest temperature. Based on this, we propose the Worst-case Undervolted FPGA (WUF) design that is reliably applicable across the temperature range. We improve the power-efficiency of FPGA-based CNN accelerators by over 1.65X through WUF design.
- If there is a typical lowest temperature attributed to the use case (e.g., in data-centers), we propose the Application-specific Undervolted FPGA (AUF) design where further efficiency is possible. For instance, AUF design based on 15°C, which is the typical minimum for data-centers, provides 1.8X better power efficiency without compromising accuracy.

- To enhance the efficiency of WUF and AUF designs, we propose Smart Undervolted FPGA (SUF) design approach that relies on a surrogate model developed to represent the relationship between accuracy and voltage in a wide temperature range. SUF aims to implement an adaptive undervolting scheme that does not compromise the accuracy (or reliability) at the current temperature. This approach ensures an even higher power efficiency due to its smart adaptation capability to temperature variations instead of being optimized for a single worst temperature value.

EXPERIMENTAL METHODOLOGY

CNNs encompass two stages, i.e., training and classification [11]. FPGAs are typically utilized in the field for the classification stage. For this reason, we implement the classification phase of the GoogleNet, VGG NET and ResNet benchmarks with the specifications in Table 1. The targeted accuracy in the literature and this study is the same. To implement these benchmarks on FPGA, we chose the DNN Development Kit (DNNDK) platform, a CNN framework offered by Xilinx. DNNDK provides CNN implementation on the Zynq based ZCU102 board by Xilinx [12]. Table 2 gives the configuration summary for our experiments. In the experiments, INT8 quantization is used as architectural optimization, and pruning optimization is not applied.

To set the voltage and monitor the power on the FPGA, we use the Inter-Integrated Circuit (I^2C) interface which allows access to the Power Management Bus (PMBus) power controllers and monitors. Through the I^2C interface and commands, we apply the desired voltage to the selected voltage rail. We get and save the power (both average and max) and accuracy results at runtime [12]. It is possible to get and report power on the BRAM and VCCINT rails. However, BRAM implementations on the latest generation of FPGAs feature very efficient static and dynamic power minimization; therefore, the relative BRAM power consumption is very low compared to the power dissipated over the blocks connected to the VCCINT [2], [12]. Thus, the application and observations in undervolting tests are carried out on the VCCINT rail. The baseline voltage is 850 mV.

Table 1. Specifications of the Benchmarks

Benchmark	GoogLeNet	VGG NET	ResNet
Data Set	Cifar-10	Cifar-10	ILSVRC2012
In	32*32	32*32	224*224
Out	10	10	1000
# of Layers	21	6	50
Size	6.6 MB	8.7 MB	102 MB
Accuracy	91%	87%	76%

The tests within the context of this study are performed at the Xilinx default frequency (300 MHz) with the ready-to-use boot image. We change the voltage at different environmental conditions step by step, but we don't apply any frequency scaling at any time.

We conducted detailed experiments on more than one board to verify that experimental variability across multiple boards is minimal and that aging and board-specific issues do not skew the results [2].

To show the impact of undervolting under harsh environmental conditions, we utilize our sealed test cabinet precisely adjusted to each step's environmental condition. We collected the accuracy and power consumption data during undervolting at the runtime of the CNN benchmarks. The test setup, the calibrated test cabinet, and sample views from different test steps are presented in Figure 1.

We also perform humidity tests subject to the operating range of our test cabinet (illustrated in Figure 2). The region marked as ”1” on this figure shows our cabinet's applicable humidity conditions, which vary depending on the temperature (Example: 70% humidity and above is applicable for temperatures above 10°C). We perform our thermal tests in the range of -40°C to 50°C, but the range where humidity tests can be performed starts from 20°C. The applicable humidity range with these thermal conditions varies from 40% to 98%.

Table 2. Preferred configuration for the experiments

Experimental Configuration Item	Configuration
Deep Learning Algorithm	Convolutional Neural Networks
CNN Implementation Platform	Xilinx DNN Development Kit
DNNDK Hardware Configuration	B4096 Deep learning Processing Units
FPGA Supported by DNNDK	Xilinx ZCU102 FPGA (16nm)
Processessing System (PS)	Zynq UltraScale+ XCZU9EG-2FFVB1156E MPSoC
Undervolting Interface Standards	PMBus Power Controllers and Monitor I2C connected to the PMBus interface
Undervolting Applied Voltage Rail	VCCINT (PMBus: 0x13, INA226: 0x40)
Architecture Level Optimization	DEep ComprEssioN Tool (DECENT)
Quantization Precision / Pruning	INT8 Precision / No Pruning

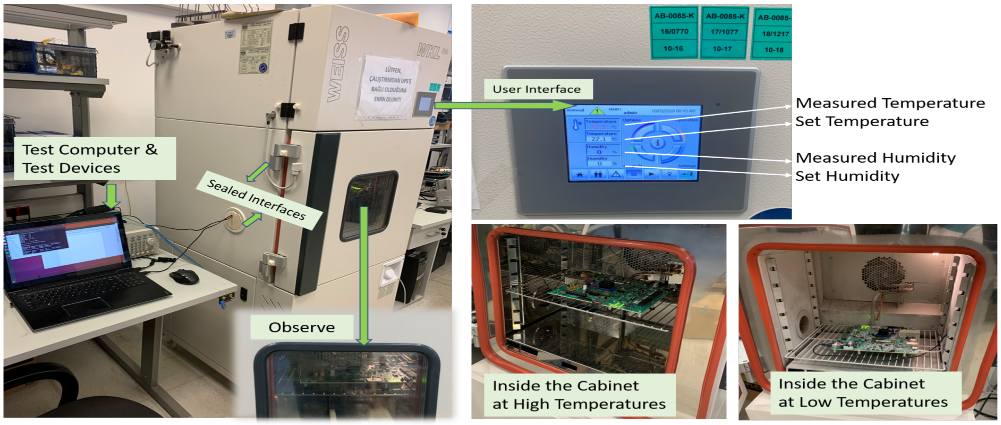

Figure 1. Test Infrastructure: Test setup, sensitive test cabinet and sample views from test steps

For humidity, we perform tests over four different temperatures (20°C, 30°C, 40°C, 50°C) and four distinct humidity levels (50%, 60%, 70%, 80%). We perform undervolting experiments at each combination (indicated by orange points in Figure 2) of these environmental conditions. For each such data point, we wait until both temperature and humidity settle before running the NN application.

RESULTS AND DISCUSSION

We present the impact of undervolting in different temperature and humidity levels on the accuracy and power efficiency of CNN benchmarks. We can reduce the voltage until the guard-band limit without decreasing the accuracy. However, this limit changes with the ambient temperature or humidity condition.

Reliable Undervolting at Varying Temperatures

Based on our observations, we propose three reliable undervolting designs:

1) WUF (Worst-case Undervolted FPGA)
2) AUF (App-specific Undervolted FPGA)
3) SUF (Smart Undervolted FPGA)

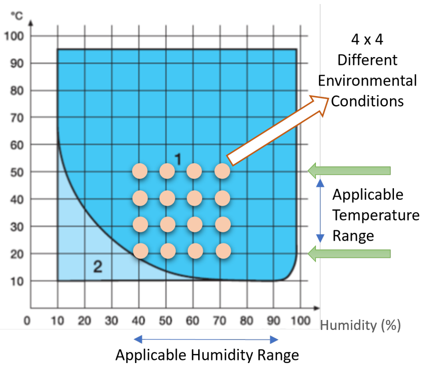

Figure 2. Applicable humidity based on temperature

In the WUF design, we apply the lowest voltage to prevent the accuracy from being compromised in the worst environmental condition, i.e., -40°C for this study. For the AUF design, the required minimum voltage for a specific condition is applied to the FPGA. If there is a typical lowest temperature attributed to the use case, then we propose to utilize the AUF technique. For instance, most of the data centers prefer keeping the ambient temperature higher than 15°C [13]. The proposed AUF design is applicable for CNN applications running in such data centers and it is adjusted to work based on this minimum temperature. However, there is still a chance to encounter lower temperatures than expected for an application. For that case, we recommend applying the SHE (Self Heating Element) method [14] - when the ambient temperature drops below the expected value the FPGA starts the control the temperature based on FPGA thermal sensors with the help of SHEs-, or employing our last proposed design (SUF) for getting the FPGA to the required temperature.

For our adaptive undervolting design, i.e., SUF, we propose to dynamically change the voltage level depending on the ambient temperature. Based on the characterization of the relationship between accuracy and undervolting at varying temperatures, the FPGA decides the voltage level. SUF implements this adaptive undervolting scheme to provide power efficiency

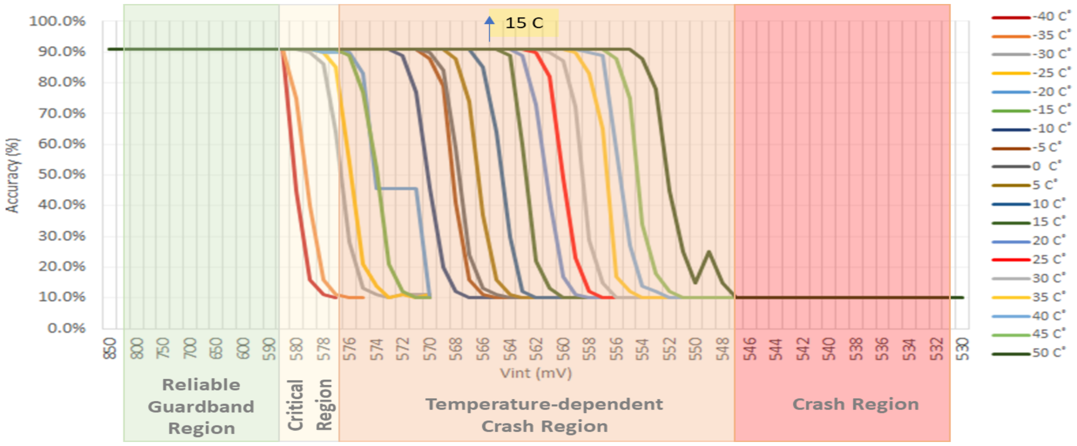

Figure 3. Effect of undervolting on accuracy of the CNN benchmark at different thermal conditions

while not compromising the accuracy at the current temperature. The optimized voltage levels at different temperatures may need application-specific derivation. We explain SUF further in the next subsection.

Accuracy and Reliability at Varying Temperature

Figure 3 shows how the accuracy of the GoogleNet benchmark changes with temperature at the different VCCINT voltage levels.

The Minimum voltage (the *Reliable Guardband Region* voltage) that can be applied safely under all thermal environmental conditions was observed at 590mV. Below 590mv, there is a narrow *Critical Region* where the accuracy starts to decrease without a crash. The range is similar for all temperatures, and this short-range is approximately 5mV. Voltage levels below 578mV cause a decrease in accuracy for some temperatures and eventually lead to a crash. We define this region as *Temperature-dependent Crash Region* because we could avoid the crash by manipulating temperature. To realize the AUF design, we select the sample scenario with the temperature of 15°C, and we apply the voltage (565mV) within the temperature-dependent crash region. Voltages below 548mV cause errors at all temperatures (*Crash Region*). Our design considers the cases that the specified regions may vary depending on the application requirements or variations in tolerance specifications for different environments.

Next, we formally define the *guard-band voltage* as the minimum voltage at a given temperature that does not have any accuracy decrease due to undervolting. From Figure 3, we can see that guard-band voltage changes with temperature. In particular the guard-band voltage shifts right, i.e. decreases, as the temperature increases. This is because the CNN accuracy increases with increasing temperature due to Inverse Thermal Dependence (ITD) [9] of deep-sub-micron technology nodes. Because of ITD, circuit latency decreases with increasing temperature, thus leading to fewer undervolting-related faults at higher temperatures. Consequently, Figure 4 directly plots the results for the guard-band voltage decrease with temperature. For our SUF design, i.e., the adaptive undervolting implementation, we model the effect of temperature on voltage and accuracy by assigning a 'best fit' function with the entire range given in Figure 4 (red dotted line presents the function curve) and the function is given by Equation 1.

$$AAV = \alpha T + \beta = -1.483T + 583.4 \qquad (1)$$

This equation indicates the Adaptively Applied Voltage ("AAV") that can be applied at a given temperature ("T") without changing the

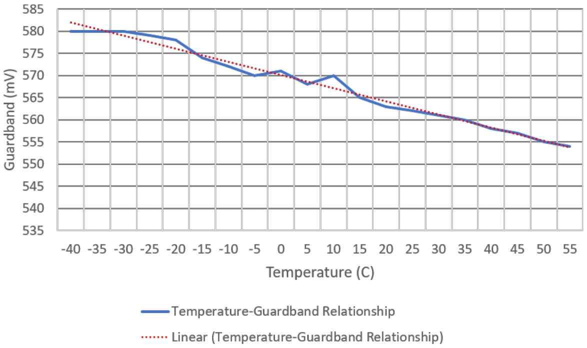

Figure 4. Guard-band voltage vs. the temperature

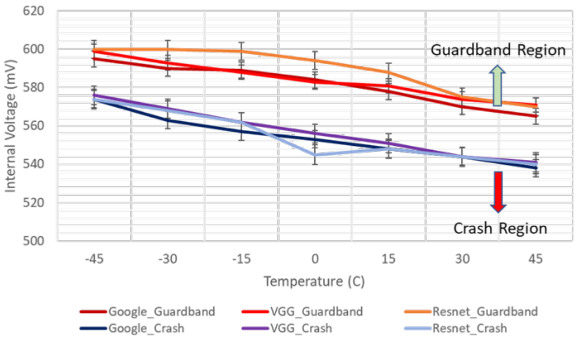

Figure 5. Undervolting with different benchmarks

accuracy. Using this equation, the adaptive design point for temperature dependent reliable undervolting is achieved. Based on the experimental results in this study, the α equals to -1.483 and β equals to 583.4. The model parameter coefficients need to be recomputed when the adaptive design is applied to different applications or platforms.

The variation of accuracies at different voltage levels and temperatures for distinct benchmarks (GoogleNet, VGG Net, ResNet) is introduced in Figure 5. We present both Guardband (the lines above) and Crash (the lines below) Voltages for three benchmarks. We observe only a 10 mV difference between any two benchmarks at any temperature for guardband voltages, and this conclusion also holds for the crash region limits. Figure 6 shows the undervolting results of the tests at different temperatures by using two different FPGAs (denoted B1 and B2) with the same specifications. For any given temperature, the difference in guard band voltages (indicated by the lines) that can be applied between the two boards is at most 15 mV. The highest difference between observed crash voltages (indicated with columns) is 13mV.

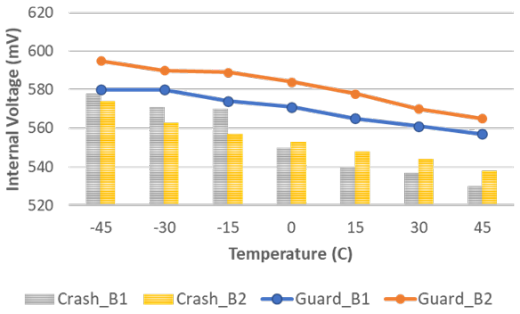

Figure 6. Undervolting with different FPGAs

Power Analysis at Varying Temperature

Instantaneous maximum current values and average power consumptions are collected and recorded during the tests. We present the change of average and maximum power consumption at different voltage levels and the temperature effects on this change in Figure 7. The variation of voltage and power consumption forms a similar trend at all temperatures. However, the effect of temperature on power consumption (for both average and maximum) is not linear.

The results presented in this study reveal that the WUF design achieves at least the power efficiency of 1.65X by applying 590mV supply voltage instead of 850mV, which is the nominal voltage under all conditions. The AUF design provides 1.78X power efficiency by applying 565mV, which is the minimum voltage that can be applied without affecting the accuracy at 15 °C, instead of 850 mV. SUF design ensures even higher power efficiency than WUF due to its adaptive undervolting scheme instead of being optimized for a single worst temperature condition. Depending on the pattern of the ambient temperatures, SUF may provide higher power efficiency than AUF.

As a result, if an FPGA used to accelerate CNN in an autonomous vehicle is to be undervolted, WUF design may be preferred because it should give error-free results in different seasonal conditions and geographical regions. If used in a IoT use-case, SUF design may be preferred because the battery is critical, but the accuracy is also sensitive. AUF design can be used in more controlled environmental conditions, for example, in air-conditioned data centers. There is a power impact (dominated by the dynamic variety) of SHEs and adaptive circuits; however this impact is minimal since these circuits are only activated in the rare cases of sudden temperature swings. In summary, our proposed reliable undervolting designs at varying temperatures provide at least a 1.65X power efficiency.

Humidity Effects on Accuracy of the CNNs

Humidity conditions and temperature are variable and challenging for products and systems in which CNNs are used. It is also necessary to look at how the proposed techniques change under humidity conditions.

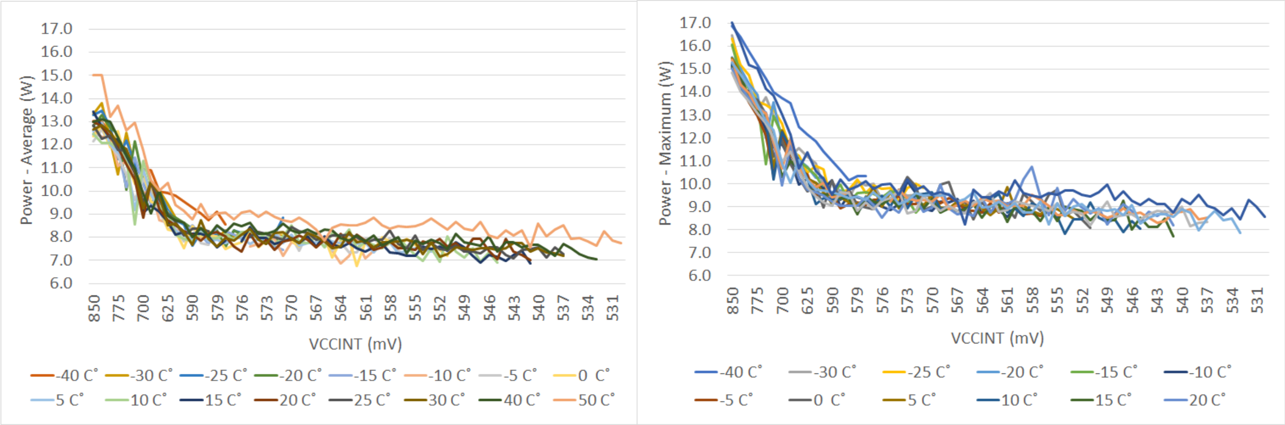

Figure 7. Effect of undervolting on average and maximum power consumption at different temperatures

Figure 8 shows how the CNN accuracy changes under different undervolting voltages for four distinct humidity levels at each selected temperature. We observe that the accuracy decreases slower with undervolting at high humidity. This experimental result can be explained by the fact that humidity creates heat capacity on the FPGA and circuits. As the humidity increases, the thermal capacities of the components within the FPGA increase. This causes less heat generated inside to be transferred to the outside, and the temperature rises. As mentioned, guard-band voltage decreases as the temperature increases.

As a result, guard-band voltage indirectly decreases as humidity increases. Since the effect of changing the humidity is limited, and it is costly to apply, it can be only used as a limited mitigation mechanism.

Impact of Frequency on Undervolting Experiments

In this work, we perform undervolting experiments at each harsh environmental condition without changing the frequency. To examine the impact of frequency separately, we synthesize a different instance for each target frequency value. We apply undervolting at three different frequencies; low (200 MHz), middle (250 MHz), and default frequency (300 MHz) under room temperature conditions with GoogleNet. We observed that the minimum guardband voltage shifts minimally with frequency (535 mV, 545 mV, and 570 mV for 200 MHz, 250 MHz, and 300 Mhz, respectively). Note that the voltage guardband increases with decreasing frequency. Thus, the worst-case frequency for the guardband is at the default value of 300 MHz used in this study.

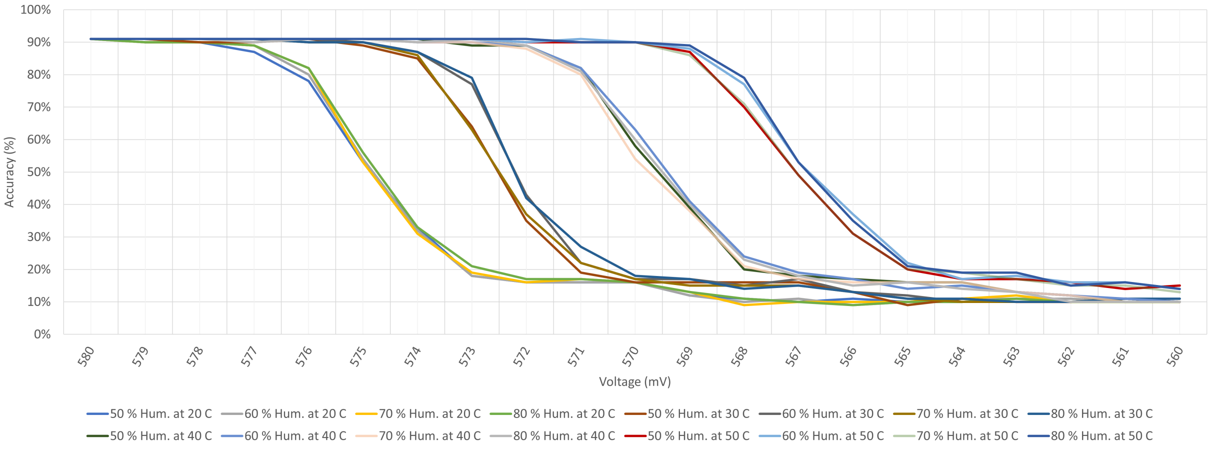

Figure 8. Voltage and accuracy relationship at different thermal and humidity conditions

CONCLUSION AND FUTURE WORK

In this study, we present the impact of undervolting in severe environmental conditions on the CNN benchmarks' accuracy and power efficiency. This paper comprehensively performs an experimental study of the power-reliability trade-offs under harsh thermal (-40°C to 50°C) and humidity conditions (40%, 50%, 70%, 80%) for FPGA-based CNN accelerators. Based on the experimental results, we propose new undervolting designs optimizing the supply voltage in harsh environmental conditions where the FPGA-accelerated CNN applications' reliability needs to be preserved. Compared to the baseline FPGA design, our proposed undervolting designs that are considered reliable at varying temperatures provide at least a 65% increase in power efficiency without compromising the accuracy.

The different applications show different characteristics in terms of the undervolting effects at different conditions. Future research should be devoted to the development of reliable undervolting techniques effective in various applications. Future studies could also investigate undervolting under other environmental conditions.

■ REFERENCES

1. V. Sze *et al.*, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
2. B. Salami *et al.*, "An experimental study of reduced-voltage operation in modern fpgas for neural network acceleration," in *IEEE/IFIP DSN*, 2020, pp. 138–149.
3. P. N. Whatmough *et al.*, "14.3 a 28nm soc with a 1.2ghz 568nj/prediction sparse deep-neural-network engine with 0.1 timing error rate tolerance for iot applications," in *IEEE ISSCC*, 2017, pp. 242–243.
4. N. Chandramoorthy *et al.*, "Resilient low voltage accelerators for high energy efficiency," in *IEEE HPCA*, 2019.
5. B. Khaleghi and T. Rosing, "Thermal-aware design and flow for fpga performance improvement," in *(DATE)*, 2019, pp. 342–347.
6. Y. Tian *et al.*, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *IEEE/ACM Int. Conf. on Software Eng. (ICSE)*, 2018, pp. 303–314.
7. K. K. Chang *et al.*, "Understanding reduced-voltage operation in modern dram devices: Experimental characterization, analysis, and mechanisms," *Proc. ACM Meas. Anal. Comput. Syst.*, 2017.
8. D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," in *HPCA 2001*, pp. 171–182.
9. K. Neshatpour *et al.*, "Enhancing power, performance, and energy efficiency in chip multiprocessors exploiting inverse thermal dependence," *IEEE Transac. on VLSI Systems*, vol. 26, no. 4, pp. 778–791, 2018.
10. C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.
11. Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
12. "Xilinx. zynq ultrascale+ mpsoc zcu102 evaluation kit https://www.xilinx.com/products/boards-and-kits/ek-u1-zcu102- g.html, 2019."
13. Ashrae technical committee 9.9. "thermal guidelines for data processing environments", fourth edition, 2015.
14. A. Amouri *et al.*, "Self-heating thermal-aware testing of fpgas," in *IEEE VLSI Test Symposium*, 2014, pp. 1–6.

Fahrettin Koc works as Chief Senior Researcher at TUBITAK Defense Industries Research And Development Institute (TUBITAK SAGE), Ankara, Turkiye. He is currently working toward a Ph.D. with the Department of Computer Engineering, TOBB University of Economics and Technology (TOBB ETU), Ankara, Turkiye. He obtained his B.S. degree in Electrical and Electronics Engineering and M.S. degree in Computer Engineering from TOBB ETU. He has four patent applications, two of which are already granted, in addition to his publications, and has received several grants for his researches. His research interests include VLSI design for adaptive memory and reconfigurable systems, energy-efficient, reliable, high-performance hardware accelerators & computer architectures. Contact him at fahrettin.koc@etu.edu.tr.

Behzad Salami is a post-doctoral researcher in the Computer Science department of Barcelona Supercomputing Center (BSC) and an affiliated research member of SAFARI Research Group at ETH Zurich. He received his Ph.D. with honors in Computer Architecture from Universitat Politecnica de Catalunya (UPC) in 2018. Also, he obtained MSc and BSc degrees in Computer Engineering from Amirkabir University of Technology (AUT) and Iran University of Science and Technology (IUST), respectively. He has received multiple awards and grants for his research. His research interests are heterogeneous systems, low-power & fault-resilient hardware accelerators, and near-data processing systems. Contact him at behzadsalami@gmail.com

Oguz Ergin is a professor in the computer engineering department at TOBB University of Economics and Technology. He received his B.S. in electrical and electronics engineering from Middle East Technical University, M.S., and Ph.D. in computer science from the State University of New York at Binghamton. He was a senior research scientist in Intel Barcelona Research Center prior to joining TOBB ETU. He is currently leading a research group in TOBB ETU, working on energy-efficient, reliable, and high-performance computer architectures. Contact him at oergin@etu.edu.tr.

Osman Unsal is co-manager of Computer Architecture for Parallel Paradigms research group at Barcelona Supercomputing Center. His research interests include computer architecture, reliability, and low-power computing. He received the B.S. degree from Istanbul Technical University, Istanbul, Turkey, the M.S. degree from Brown University, Providence, RI, USA, and the Ph.D. degree from the University of Massachusetts Amherst, MA, USA, all in electrical and computer engineering. Contact him at osman.unsal@bsc.es.

Adrian Cristal Kestelman received the Licenciatura degree in Computer Science from the Faculty of Exact and Natural Sciences, Universidad de Buenos Aires, Buenos Aires, Argentina, in 1995, and the Ph.D. degree in Computer Science from the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain. Since 2006, he has been a co-manager of the Computer Architecture for Parallel Paradigms Research Group at Barcelona Supercomputing Center (BSC). His current research interests include the areas of microarchitecture, multicore, and heterogeneous architectures and programming models for multicore architectures. Currently, he is leading the architecture development of the vector processor unit in the European Processor Initiative. Contact him at adrian.cristal@bsc.es.