UnderVolt_FNN: An Energy-Efficient and Fault-Resilient Low-Voltage FPGA-based DNN Accelerator

Behzad Salami
Barcelona Supercomputing Center
Universitat Politcnica de Catalunya
behzad.salami@bsc.es

Osman S. Unsal
Barcelona Supercomputing Center
osman.unsal@bsc.es

Adrian Cristal Kestelman
Barcelona Supercomputing Center
Universitat Politcnica de Catalunya
IIIA - Articial Intelligence
Research Institute CSIC
adrian.cristal@bsc.es

ABSTRACT

There is continually growing interest to accelerate DNNs using FPGAs, e.g., Microsoft Catapult [3], thanks to FPGAs inherently power-efficient architecture and their capability for streaming-fashion data marshalling and computation on the massively parallel FPGA structure. However, in comparison to ASIC-based accelerators, e.g. Google TPU [2] and IBM TrueNorth [1], their energy efficiency is still a key concern. In order to bridge this gap, we propose UnderVolt_FNN, a novel FPGA-based accelerator that leverages aggressive undervolting for improving the energy efficiency of DNNs in the inference phase, without compromising the inference accuracy and performance. Toward this goal, we first extensively characterize the behavior of undervolting faults on real FPGAs, and accordingly, propose a fault mitigation technique to prevent the DNNs accuracy loss. In consequence, UnderVolt_FNN achieves 25.2% of the total power savings gain without compromising the DNNs accuracy and performance. We believe that the proposed technique has potential to improve the energy-efficiency of enterprise systems such as Microsoft Catapult.

KEYWORDS

FPGA, DNN, Voltage Underscaling, Energy, Resilience

1 BACKGROUND

Recently, there are many FPGA-based designs proposed for accelerating DNNs, while several application-level ideas, e.g., binarizing and pruning are deployed to improve the energy-efficiency of such designs. However, to the best of our knowledge, FPGAs undervolting as an orthogonally architecture-level energy-saving approach is not experimentally studied yet. Although, undervolting has been shown to deliver significant energy-saving in ASIC-based DNNs such as Thundervolt and YodaNN. However, these studies are mostly conducted on simulated frameworks, which is apparent that reproducing these results on real platforms remains as a key concern. We leverage the potential of undervolting for FPGA-based DNNs.

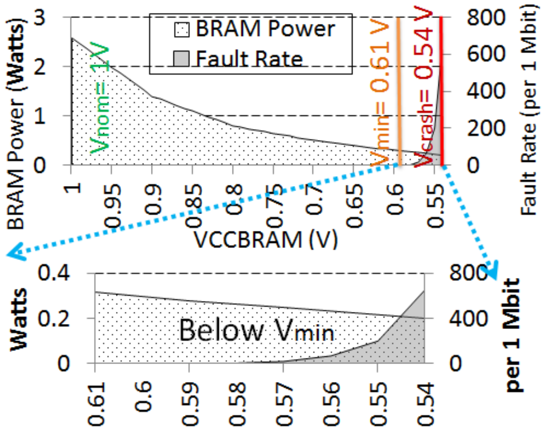

Figure 1: The overall power and reliability behavior of FPGAs through aggressive undervolting (example shown for VC707).

2 CONTRIBUTIONS

UnderVolt_FNN aims an energy-efficient and resilient FPGA-based DNN accelerator through aggressive undervolting on real commercial FPGAs from Xilinx, a main vendor, confirmed on VC707, KC705, and ZC702. Main contributions in UnderVolt_FNN are:

(1) **Improving the Energy-Efficiency of FPGA-based DNN Accelerator through Aggressive Undervolting:** Our concentration is on the FPGA-based on-chip memories, or Block RAMs (BRAMs), since BRAMs play crucial role in the FPGA-based DNNs structure and also, according to the studied FPGAs architectures, the supply voltage of BRAMs are allowed to be independently regulated. Through experimental analysis, we found that undervolting until a certain conservative level, V_{min}, does not introduce any observable fault, see Figure 1. We measured this voltage guardband to be on average 39% of the nominal level ($V_{nom} = 1V, V_{min} = 0.61V$), although it slightly varies among studied FPGAs. In consequence, more than an order of magnitude BRAMs power is saved that leads to 24.1% of the design's energy reduction. Further undervolting below the voltage guardband causes timing faults.

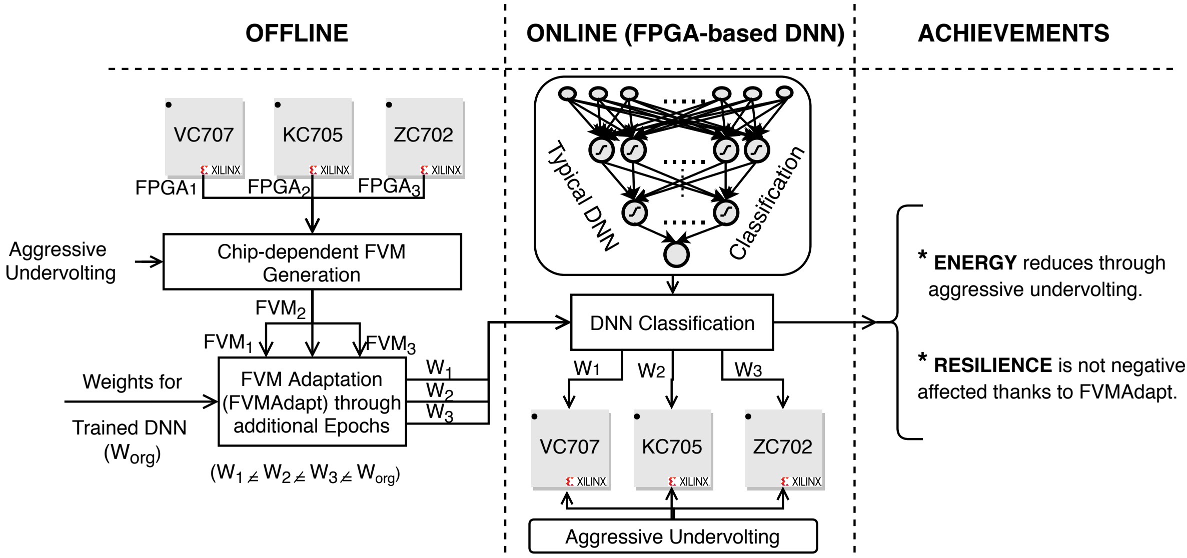
OFFLINE
ONLINE (FPGA-based DNN)
ACHIEVEMENTS
VC707
KC705
ZC702
XILINX
FPGA1
FPGA2
FPGA3
Aggressive Undervolting
Chip-dependent FVM Generation
FVM2
FVM1
FVM3
Weights for Trained DNN (Worg)
FVM Adaptation (FVMAdapt) through additional Epochs
W1
W2
W3
(W1 ≉ W2 ≉ W3 ≉ Worg)
Typical DNN
Classification
DNN Classification
Aggressive Undervolting
* ENERGY reduces through aggressive undervolting.
* RESILIENCE is not negatively affected thanks to FVMAdapt.

Figure 2: The overall methodology to FVMAdapt technique to generate DNN weights adapted by the chip-dependent Fault Variation Map (FVM) through aggressive undervolting below the voltage guardband, i.e., V_{mib}, with the aim of achieving energy efficiency and preventing the DNN accuracy loss.

(2) **Fault Characterization of FPGAs under Low-voltage Operations:** By further undervolting below V_{min} and until a system crash voltage level, and without accompanying the frequency underscaling to achieve better energy efficiency, the rate of these faults exponentially increases, up to 0.06% (652 per 1 Mbit) in the platform with the worst reliability behavior, i.e., VC707, see Figure 1. More specifically, for all platforms that we study, we observe that faults are non-uniformly distributed over various BRAMs and by on average 99.9% of these faults manifest themselves as '1' to '0' bit flips, while, locations of faults do not change over time. This deterministic fault behavior allows to extract a chip-dependent Fault Variation Map (FVM) that is effectively exploited within the proposed fault mitigation technique.

(3) **Improving the Resilience of FPGA-Based DNN Accelerator under Low-voltage Operations:** In UnderVolt_FNN, DNN weights are located in BRAMs, input data items are streamed through the off-chip DDR-3s, and the required computations are performed in parallel on DSPs, as usual for state-of-the-art FPGA-based DNNs. Note that weights are imported after a compile-time training on the software application. By undervolting BRAMs below the V_{min}, further BRAM power reduces that leads to in total 25.2% of the total power savings gain; however, with up to 4.5% of accuracy loss for MNIST, a common image classification benchmark, for instance for VC707. In order to prevent this accuracy loss, we propose a novel fault mitigation technique, i.e., FVM-aware Weights Adaptation (FVMAdapt) that relies on the characterized behavior of faults, see Figure 2. FVMAdapt is a post-training phase that exploits FVM and generates adapted weights to chip-dependent fault variation map by additional epochs. In consequence, weights that are used in the inference phase are adapted to the fault map of the corresponding FPGA. Toward this goal, we set the corresponding faulty locations to 0 for post-training weights since vast majority of undervolting faults are '1' to '0' bit-flips, and allows to continue for additional training epochs. We experimentally observed a significant performance of FVMAdapt for the validation dataset of MNIST with the overhead of only a single additional epoch, which in turn, leads to a similar validation and test error rate as the original.

3 ONGOING DEVELOPMENTS

We are extending UnderVolt_FNN on other FPGA components, e.g., I/O and internal resource such as DSPs, and evaluating it with more DNN benchmarks, .e.g., ImageNet and AlexNet.

REFERENCES

[1] Hsu, J. (2014). Ibm's new brain [news]. *IEEE spectrum*, **51**, 17–19.

[2] Jouppi, N.P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A. *et al.* (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 1–12, ACM.

[3] Putnam, A., Caulfield, A.M., Chung, E.S., Chiou, D., Constantinides, K., Demme, J., Esmaeilzadeh, H., Fowers, J., Gopal, G.P., Gray, J. *et al.* (2014). A reconfigurable fabric for accelerating large-scale datacenter services. *ACM SIGARCH Computer Architecture News*, **42**, 13–24.