Performance–Reliability Trade-offs for AI Acceleration on GPUs and FPGAs

Fernando Fernandes dos Santos
Univ Rennes, IRISA, Inria, Rennes, France

Abstract

DNN optimizations improve performance and resource use, but they also change how computations map to hardware and how faults propagate. This work studies GPUs and FPGAs under radiation, showing that precision, accelerator size, and reuse parameters can significantly change failure rates and fault criticality. Results show that reliability must be assessed together with performance, since faster configurations may produce more correct executions over time even when their FIT does not improve.

1 Motivation

DNN accelerators are increasingly deployed in systems where transient faults can compromise application correctness. Fault injection is therefore essential to identify vulnerable resources and track error propagation across hardware and software layers. However, conclusions from fault injection strongly depend on the selected abstraction level and fault model [1]. Gate-level, microarchitectural, instruction-level, and application-level simulations can expose different failure mechanisms, but all rely on user-defined fault models. In contrast, radiation experiments provide complementary evidence by exposing the complete system to realistic particle-induced effects and enabling application-level failure-rate estimation, although they usually cannot identify the exact fault site [2].

Prior radiation experiment studies have shown that GPUs and FPGAs running DNNs are susceptible to radiation induced faults [3], [4], [5]. GPUs are affected by their large amount of active resources and by the possibility of corrupting several output elements [6], [7]. FPGAs are additionally exposed through their programmable fabric, where faults may alter configuration-dependent circuit behavior [4], [8]. In this work we discuss practical reliability trends from radiation experiments on both platforms, with emphasis on the effect of mixed precision and hardware size. This study is a summary of recent works at the TARAN team [3], [4]

2 Evaluation and Main Results

2.1 Experimental Setup

Facility and method. Experiments were performed at the ChipIR facility of the Rutherford Appleton Laboratory, UK. ChipIR provides an atmospheric-like neutron spectrum [9] with flux up to approximately 3.5×10^6 n/(cm^{2}s), more than eight orders of magnitude above sea-level terrestrial flux [10]. To preserve a single-event interpretation, the campaigns were configured to keep observed error rates below one failure per 2,000 executions. Each DNN configuration was exposed for at least three effective beam hours, excluding setup, checking, initialization, and recovery time.

Devices and workloads. The GPU experiments used an NVIDIA Tesla V100 based on the 12 nm Volta microarchitecture. We evaluated cuBLAS GEMM in FP16, FP32, and FP64 because GEMM dominates many DNN computations. As an application-level case study, YOLOv3 object detection was evaluated in FP16 and FP32. The FPGA experiments used a TUL PYNQ-Z2 board with a 28 nm Xilinx Zynq-7000 SoC. Four HLS4ML-generated ANN accelerators were implemented for MNIST classification [11] (Opt_1 to Opt_4). The configurations vary the number of multipliers per layer, trading area for execution time.

Irradiation experiments. A software watchdog, running on a server outside the beam room, controlled the device under test, launched the workloads, monitored execution, logged failures, and restarted the application after hangs or crashes. The GPU board was managed by an x86 Intel i3 host, while the FPGA SoC used the embedded ARM processor. After each execution, the host compared the accelerator output with a golden reference. Output mismatches were classified as silent data corruptions (SDCs), while hangs or crashes were classified as detected unrecoverable errors (DUEs). The analysis only considers errors visible at kernel or application output. We consider an SDC critical when it changes the application-level decision, e.g., a wrong classification or misdetection.

Metrics. We measure the FIT rate, it scales the measured beam failure rate, $(\#SDC \text{ or } \#DUE)/fluence$, to the target neutron flux, such as the atmospheric flux of approximately 13 $n/cm^2/h$. Because FIT alone ignores performance differences, we also use the Mean Executions Between Failures (MEBF), which is defined as $\text{MEBF} = (\sigma \cdot flux)^{-1}/\text{ExecutionTime}$. The MEBF therefore captures how many correct executions can be expected before a visible failure, which is essential when comparing implementations with different execution times. We compute MEBF considering only critical SDCs and DUEs.

2.2 Experimental Results

Figure 1 reports relative FIT and MEBF with 95% Poisson confidence intervals. Values are normalized to the worst-performing configuration for each workload: FP64 GEMM,

FP32 YOLOv3, and Opt_1 for the FPGA ANN.

Mixed precision changes GPU reliability. Lower precision reduces the measured GPU FIT rate. For GEMM, FP32 reduces FIT by 21.50% compared with FP64, while FP16 reduces FIT by 31.51% compared with FP32. For YOLOv3, FP16 reduces FIT by 46.29% compared with FP32. This trend is consistent with lower-precision execution using fewer hardware resources, reducing the probability that a neutron-induced event reaches the application output. Precision also changes failure criticality: FP16 YOLOv3 produces fewer critical SDCs than FP32, with critical misdetections representing 21.43% and 27.27% of SDCs, respectively. This suggests that mixed precision is not only a performance and energy optimization, it also changes the accelerator's reliability profile.

FPGA area and execution time must be evaluated together. The FPGA FIT trend differs from the GPU trend. Opt_2 reduces FIT to 82.81% of Opt_1, while Opt_3 and Opt_4 reach 98.35% and 107.74% of Opt_1, respectively. Larger accelerators expose more critical resources to radiation and therefore increase the chance that a bit flip affects critical logic causing a SDC or DUE. Without scrubbing or ECC, smaller accelerators may also execute longer and reuse hardware more extensively, giving faults more probability to propagate as SDCs. Critical SDCs represent 18.26%, 21.92%, 28.52%, and 28.81% of the SDC FIT for Opt_1 to Opt_4, showing that larger FPGA accelerators tend to fail more critically when an SDC occurs.

MEBF exposes the performance–reliability trade-off. FIT alone is insufficient because it does not include execution time on the analysis [12]. On GPUs, lower precision improves both performance and reliability: GEMM MEBF is 2.52× higher for FP32 and 7.33× higher for FP16 than for FP64, while YOLOv3 FP16 performs 1.89× more correct predictions between failures than FP32. For the FPGA ANN, MEBF improves from Opt_1 to Opt_4: Opt_2, Opt_3, and Opt_4 increase MEBF by 2.72×, 6.35×, and 26.25×, respectively. Thus, faster and larger accelerators may deliver many more correct inferences before failure, even if each failure is more likely to be critical.

3 Main Takeaways

Mixed precision changes the reliability of DNN accelerators, not only their throughput and energy efficiency. In the evaluated GPU workloads, FP16 consistently improves FIT and MEBF relative to higher precision. For FPGA accelerators, larger designs increase critical-resource exposure, but their shorter execution time can still yield more correct executions between failures. Therefore, accelerator reliability should be reported using both FIT and MEBF, while separating SDCs by application-level criticality. Ongoing work connects beam experiments with fault simulation to identify the hardware structures behind these precision- and architecture-dependent trends, with the final goal of guiding architectural fault-tolerance mechanisms for GPUs, systolic arrays, and HLS-generated accelerators operating in harsh environments.

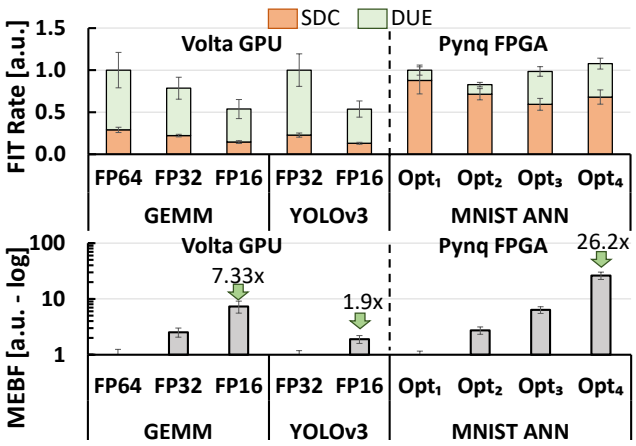

Figure 1: Relative SDC FIT, DUE FIT, and MEBF. Values are normalized to the worst configuration of each workload.

References

[1] G. Papadimitriou et al., "Demystifying the System Vulnerability Stack: Transient Fault Effects across the Layers," in *IEEE ISCA*, 2021.

[2] P. R. Bodmann et al., "Soft error effects on arm microprocessors: Early estimations versus chip measurements," *IEEE Trans. on Comp.*, 2021.

[3] F. Fernandes dos Santos et al., "Reliability evaluation of mixed-precision architectures," in *2019 IEEE HPCA*, 2019.

[4] M. Traiola et al., "Impact of high-level-synthesis on reliability of artificial neural network hardware accelerators," *IEEE Trans. on Nucl. Science*, 2024.

[5] R. L. Rech Junior et al., "High energy and thermal neutron sensitivity of google tensor processing units," *IEEE Trans. on Nucl. Science*, 2022.

[6] M. B. Sullivan et al., "Characterizing And Mitigating Soft Errors in GPU DRAM," in *IEEE Micro*. 2021.

[7] F. F. dos Santos et al., "Analyzing and increasing the reliability of convolutional neural networks on gpus," *IEEE Transactions on Reliability*, 2018.

[8] F. Libano et al., "Selective hardening for neural networks in fpgas," *IEEE Trans. on Nucl. Science*, 2018.

[9] C. Cazzaniga et al., "Progress of the scientific commissioning of a fast neutron beamline for chip irradiation," *Journal of Physics: Conference Series*, 2018.

[10] JEDEC, "Measurement and Reporting of Alpha Particle and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices," Tech. Rep., 2006.

[11] J. Duarte et al., "Fast Inference of Deep Neural Networks for Real-Time Particle Physics Applications," *ACM/SIGDA FPGA*, 2019.

[12] G. Reis et al., "Design and evaluation of hybrid fault-detection systems," in *32nd ISCA*, 2005.