Area-energy-time tradeoff with a low-power accelerator for reliable Edge AI efficiency under real-world radiation

Anuj Justus Rajappa*[1], Philippe Reiter[1], Paolo Rech[2], Siegfried Mercelis[1], Jeroen Famaey[1]

[1] *IDLab, University of Antwerp - imec, Sint-Pietersvliet 7, 2000 Antwerp, Belgium*
[2] *Università di Trento,via Calepina, 14 - I-38122 Trento, Italy*

With the advent of Edge AI, machine learning algorithms, such as neural networks (NN), can process data closer to data sources, including sensors using low-power, commercially available off-the-shelf (COTS) compute devices, as a result of real-time operability, reduced cost, state-of-the-art performance, etc. Edge AI thus finds itself used in space and remote terrestrial applications where it may not be ideally protected from environmental radiation, such as neutrons and protons, due to application constraints. This exposes these edge AI systems to single-event effects (SEE) that can lead to soft errors. These soft errors manifest as bit-flips and affect Edge AI reliability. However, resource limitations at the Edge require executing AI inferences more efficiently while maintaining reliability. Hence, we hypothesize that trading off additional runtime chip area for reduced execution time and energy with a hardware AI accelerator can improve Edge AI efficiency while maintaining similar levels of reliability. We tested our hypothesis by executing NN inferences in a low-power, ARM platform-based COTS device with a floating point unit under neutron radiation, with proton-like SEE effects, at ChipIr in the UK. The tests showed that the efficiency improved by 1.9 times while maintaining reliability against soft error-induced bit-flips.

1 Introduction

Autonomy requirements have increased for compute systems deployed anywhere from terrestrial to space environments in applications and domains ranging from Earth observation [1], smart production [2], automotive, smart cities, health care [3], nuclear power [4], avionics [5], space avionics [6], communication, constellation orbital control and guidance [7], space robotics, space debris removal [8], rendezvous and proximity operations (RPO) [9], Internet of Space Things (IoST) [10] and Internet of Things (IoT) [3] in general. This requires a variety of sensors onboard said systems, such as vision and multi-spectral cameras [8], to improve their sensing capabilities, which generate a lot of data [6]. Processing these data for real-time decision-making can increase autonomy. Aritifial Intelligence (AI) algorithms, such as Multilayer Perceptrons (MLP) or Deep Neural Networks (DNN) [11], can effectively process this data using powerful cloud servers [12] on Earth.

However, overheads associated with the communication between deployed systems (edge devices [1, 3]) and cloud servers (edge-cloud communication) can significantly limit the ability of the systems to adequately respond to space weather events, autonomous landing [8] and navigation [10] scenarios, etc. For instance, short and intense phases of landing for interplanetary rovers like Curiosity on Mars last for about 7 minutes [13, 14], but the radio signal-based communication round-trip time between Mars and Earth can take 8 to 48 minutes due primarily to the fundamental universal speed of light limit [15, 16], which leaves no room to respond with Earth-based cloud servers! While the speed of light could be less limiting as missions get closer to Earth, other limitations associated with uplink, downlink, ground station availability, power budget for system operation including communication, etc., remain [1]. Similar limitations due to edge-cloud communication overhead – e.g., data transfer delays, bandwidth saturation and delayed system response – exist in terrestrial applications, as well [3].

Such limitations associated with edge-cloud communication gave rise to Edge AI [17, 18], where AI algorithms are executed closer to the data source [3] to improve real-time decision-making and reduce communication overhead. Edge AI can be achieved using onboard computers, but it is limited by factors such as available power budget [1], compute performance per Watt [6], and thermal management capabilities [19]. Thus, Edge AI is generally resource-constrained [17, 18] and can support AI inference more readily than AI learning or training operations due primarily to the associated differences in compute requirements [3].

[*] First and corresponding author. ORCID: 0000-0001-8167-9171. E-Mail: anuj.justusrajappa@uantwerpen.be

Application constraints including form factor, computing efficiency, cost and payload weight [1] can result in onboard computers running AI algorithms to not be ideally shielded from environmental radiation particle strikes that cause single-event effects (SEE) [1]. SEEs can occur in any natural enviroment [20] as the radiation particles like protons originate from the Sun and radiation belts around the Earth [20], and neutrons can be predominantly found at terrestrial levels due to cosmic rays from outer space interacting with the Earth's atmosphere [21]. Neutrons can also infiltrate spacecrafts, habitats under the Martian surface, etc., due to cosmic rays interacting with their shielding materials [22, 23]. SEEs can lead to soft errors such as bit-flips [21] (soft error-induced bit-flips) that impact the reliability of Edge AI by causing AI algorithms to fail [24]. Therefore, the reliability of Edge AI should be maintained while optimizing for compliance with target constraints.

When performance per Watt improves for computing with hardware accelerators [6], the efficiency in terms of execution energy and time (i.e., throughput or FPS [6]) of a given task, such as DNN inference, also improves. While this allows Edge AI to operate under constraints such as reduced power budgets, the additional runtime chip area to which the Edge AI is exposed, resulting from the hardware AI accelerator and general purpose computing unit or CPU [25], might affect its reliability against soft error-induced bit-flips. This is because increasing runtime chip area increases the number of radiation particles that can potentially affect Edge AI for a given particle flux [21] within a given time. These particles affect both combinational circuits (processing) and sequential elements (memory storage), whose sensitivity to radiation-induced soft errors has increased due to factors such as drastic device shrinking, low operating voltages, and high operating frequencies [26].

Hardware AI acceleration consumes additional runtime chip area to increase AI efficiency in terms of execution time and energy. This increase in area can decrease reliability against soft error-induced bit-flips [27, 28]. We **hypothesize** that the reduction in reliability due to this area increase is countered by the reduction in the execution time, thus, maintaining the reliability.

Because of challenging size, weight, and power (SWaP) requirements [6]; project budget and schedule constraints [29]; and the state-of-the-art (SotA) performance benefits [6], commercially available off-the-shelf (COTS) parts are preferred to radiation-hardened or space-grade parts. This makes COTS devices more attractive for Edge AI execution under resource constraints. While integrating new application specific integrated circuit (ASIC)-based hardware AI accelerators, such as the AdAM [30], to existing COTS architectures is challenging to realize, existing accelerators like the Floating Point Unit (FPU) supported on many low-power COTS devices, such as ARM Cortx-M4 processors [31], can be leveraged for Edge AI.

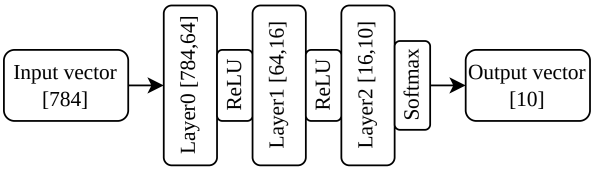

Figure 1: Fully connected DNN architecture used for tests.

Various methods for DNN reliability analysis involve Internal Fault Injection (IFI), where faults are injected internally by the test designer to simulate faults at the hardware, software, or platform level. However, real-world radiation can generate the faults externally, rendering the assessment realistic [32]. Therefore, to test our hypothesis, AI inferences, with the fully connected DNN architecture shown in Figure 1, were executed on the aforementioned ARM, low-power COTS device with FPU (test device) while being exposed to neutrons with atmospheric-like spectrum available at the ChipIr facility [33, 34] in Oxfordshire, UK. The SEE effect associated with this neutron beam is similar to that of protons and, thus, can be considered for space [23, 35] and terrestrial scenarios [21]. IFI is generally considered not accurate enough in the space community because it cannot take technological features into account. For this purpose, radiation test results are needed, which are costly to obtain and are seldom available [20]. Hence, we intend to publish the data associated with this radiation campaign post-curation. We also analysed the efficiency of DNN inference by measuring the execution time and energy.

2 Related works

Multiple DNN accelerators based on a range of computing technologies from ASICs, Gprahic Processing Units (GPUs) and Field-Programmable Gate Arrays (FPGAs) have been proposed for edge systems in domains such as space [1] to combat ever-increasing computing performance needs and power density limitations [36]. Various studies have been conducted to analyse the reliability of DNNs against soft errors. For instance, a methodology [25] for hybrid analytical and hierarchical reliability assessment of a systolic-array-based DNN hardware accelerator using IFI and the vulnerability value range for each neuron was proposed. Another work [37] proposed a framework, DeepAxe, to explore the tradeoff between approximate computing and reliability in FPGA-based DNN accelerators with IFI. An adaptive, fault-tolerant approximate multiplier, AdAM [30], was designed as a DNN accelerator tailored for an ASIC implementation and its reliability was assessed using IFI against other multiplier architectures.

Several AI inference engines use CPU-plus-FPU setups for acceleration and even have been optimized for breaking down matrix multiplication, a core function of a DNN, into smaller chunks for parallelization using lock-free programming techniques with pipelining [8], which may not be applicable for single core systems. Another work, [38], assessed the performance of various AI algorithms using a variety of hardware and found that the availability of an FPU heavily influenced performance. While the above two works consider a CPU with FPU for AI acceleration, they do not address the reliability against soft errors. The effect of precision used to represent floating point numbers (i.e., double or 64-bit, single or 32-bit and half or 16-bit) during DNN execution on DNN reliability was analyzed with both IFI and neutrons in [27]. They observed a 2-fold decrease in error rate as the precision was halved, probably due to reduced runtime chip area usage, and state that the FPU chip area grows quadratically with precision.

In our other work [28], we hypothesized reducing both the execution time and data transfers (thereby, reducing runtime chip area) using the SMART software technique for improving DNN reliability, which is different from the hypothesis proposed in this article. We could not identify other articles which propose a hypothesis similar to the one proposed in this article. Hence, to the best of our knowledge, we are the first to test the proposed hypothesis using real-world radiation experiments.

3 Methodology

The radiation experiment setup is shown in Figure 2. The test device used during the radiation experiment was an nRF52840 DK [39], which consists of a Cortex-M4 ARM-platform-based CPU operating at 64 MHz with an FPU. We developed a three-layer DNN architecture with 64, 16 and 10 neurons in the successive layers, as shown in Figure 1 and trained it with MNIST [40]. Two DNN versions, **CPU** and **CPU+FPU**, were generated by compiling the trained DNN with the Segger C-language compiler [41] using both soft and hard application binary interface (ABI) compiler options for floating point operations (cf., 'ARM FP ABI Type' in [42]). The generated executables for each version were compatible for executing on the test device.

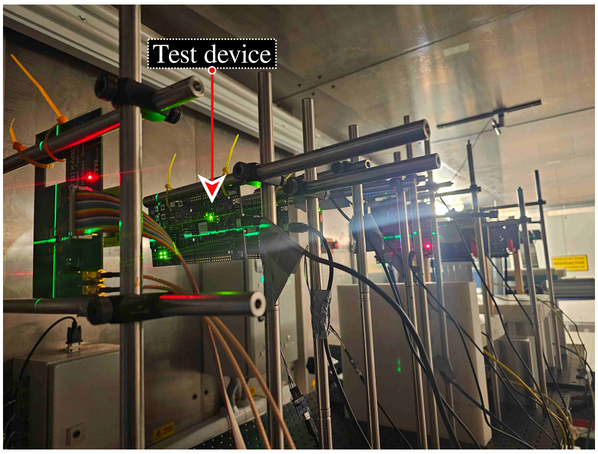

Figure 2: Aligning test device with the neutron beam path using laser markers inside the radiation chamber at ChipIr.

The CPU version was expected to have a higher execution time, exposing the inference to radiation particles for a longer duration while consuming less runtime chip area due to the lack of AI acceleration, compared to the CPU+FPU version using the FPU for AI acceleration. This contrast in runtime chip area and execution time requirement between the two versions is expected to maintain reliability against soft error-induced bit-flips as the number of radiation particles that pass through a surface under a given flux scales with both the area of the surface and the time duration of the exposure. Hence, by reducing one while increasing the other, we may be able to maintain the number of radiation particles that pass through the surface at a similar level and maintain reliability.

Input images for inferencing during the experiments were passed to the test device over USB and had an associated communication overhead. This required the number of inputs passed during the experiment to be limited to maximize the time spent on inference execution within the available beam time at ChipIr. Ten images were randomly selected from each of the 10 classes from the MNIST test dataset to generate 100 evaluation images, which were used for DNN inferences throughout this work. Multiple inferences were executed for each input transferred during the radiation experiment.

The execution time and energy of inferences were analysed outside the radiation chamber. Execution time for inferences was measured in CPU cycles using the Data Watchpoint and Trace Unit (DWT) [43]. Execution energy was measured using a Jouelscope [44], which monitors the current and voltage supplied to the test board along with the timing information and uses them to obtain the energy consumption value. The input voltage to the board was fixed at 3 V during the energy measurements.

4 Results and Discussion

The output of the inferences executed on the test device inside and outside of the radiation chamber are compared and any mismatch was considered an error. The errors with the same mismatches (same errors) in the successive inferences executed within the radiation chamber are counted together as one **unique error**, as the probability of them being caused by discrete bit-flips is considered negligible. Handling the same errors with techniques like scrubbing is out of the scope of this work.

Table 1 shows the overview of the results from the radiation experiment and the efficiency analysis, including the number of inferences executed under radiation with each DNN version and their corresponding unique error counts and fluence values (i.e., the total number of neutrons that passed through the test device per cm^2 during the radiation experiment) [21]. It should be noted that the CPU+FPU version was exposed to almost double the number of neutrons compared to the CPU version during the radiation experiment, leading to double the unique errors experienced. The reliability against soft error-induced bit-flips due to FPU-based AI acceleration was measured in soft error cross-section (SEC), which is the unique error count divided by the corresponding fluence [21] and the higher the SEC, the worse the reliability.

FPU-based AI acceleration is said to have: maintained the reliability if the SEC of CPU version is equal to that of the CPU+FPU version; increased the reliability if the SEC of CPU version is greater than that of the CPU+FPU version; and decreased the reliability if the SEC of CPU version is less than that of the CPU+FPU version.

Measurement	CPU	CPU+FPU
SEC $\left(10^{-10}\,cm^2\right)$	$1.66 \pm 5.6\%$	$1.73 \pm 2.8\%$
Exec. time $(cycles)$	$8.73e6$	$4.57e6$
Exec. energy (mJ)	3.51	1.80
Fluence $\left(cm^{-2}\right)$	$1.09e11$	$2.08e11$
Unique errors	18 ± 1	36 ± 1
Irradiated inferences	138314	360713

Table 1: Reliability and efficiency analysis overview for the two DNN versions with and without AI acceleration.

However, experimental measurements can have uncertainties. This uncertainty means that the actual SEC for CPU and CPU+FPU version can be any value within a range of values. If there is an overlap in this range measured for both these versions, then it suggests that the reliability can be maintained as there is a possibility that SEC for CPU and CPU+FPU versions can have the same value. Multiple experiments are required to obtain the probability of the possibility and it is out of scope of this work.

To understand the hypothesis, let us assume that CPU version uses an average runtime chip area a_{cpu} during the execution time t_{cpu}. For a given flux ϕ of radiation particles like neutrons, the total number of particles N that could influence the CPU version's reliability is given by $N = \phi \times a_{cpu} \times t_{cpu}$. Increasing N can increase the soft-error induced bit-flips encountered. In CPU+FPU version, the average runtime chip area increases due to FPU by a_Δ while decreasing the execution time by t_Δ. Thus, $a_{cpu} + a_\Delta = a_{cpu+fpu}$ and $t_{cpu} - t_\Delta = t_{cpu+fpu}$ We expect these changes to keep the N approximately constant, i.e., $\phi \times a_{cpu} \times t_{cpu} \approx \phi \times a_{cpu+fpu} \times t_{cpu+fpu}$. This means that the total number of neutrons that could affect the CPU and CPU+FPU versions are similar in quantity and thus the reliability against soft-error induced bit-flips could be maintained.

To prove that reliability is maintained, the unique error uncertainty and the corresponding SEC uncertainty, as shown in Table 1, were calculated. Let S be the actual SEC of the test device under the given experimental conditions. Then, the fluence F within which one unique error is observed is $1/S$, two unique errors are observed is $2/S$ and so on. As S and F are unknown before the experiment, the total fluence to which the test device is exposed during the experiment (F_e) may not be perfectly divided by F, leaving a remainder fluence value (f_r), which causes the ideal unique error count U that should be observed within F_e, while satisfying S, to become a fraction. As events cannot be counted in fractions, U cannot be observed in the real world. Let u_f be fractional part of U. As $f_r < F$, a unique error may or may not be observed within f_r. Let U_e be the unique errors observed during the experiment. Extreme uncertainties in U (i.e., $U \approx U_e \pm 1$) can happen when u_f approaches the minimum possible value (zero), then $U \approx U_e - 1$ if f_r caused a unique error, else $U \approx U_e$. Or, when u_f approaches the maximum possible value (one), then $U \approx U_e$ if f_r caused a unique error, else $U \approx U_e + 1$.

According to the first row in Table 1, the CPU+FPU version has $1.73e - 10\,cm^2$ SEC, which is 4% more than the CPU version. However, the unique error uncertainty (i.e., $U_e \pm 1$) associated with the CPU version introduces ±5.6% uncertainty in its SEC, ranging from $1.57e - 10\,cm^2$ to $1.75e - 10\,cm^2$. This shows that the CPU+FPU version's SEC can be within the range of SEC values of the CPU version, proving that the reliability against soft error-induced bit-flips can be maintained, despite the AI acceleration with the FPU,

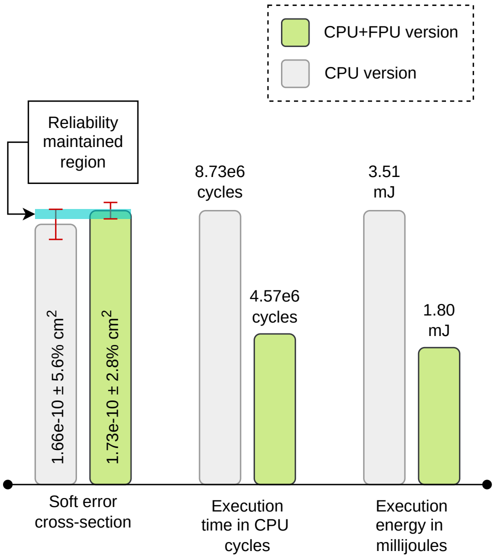

Figure 3: Overview of the radiation test results and efficiency measurements.

while reducing execution time by 48% and execution energy by 47%, as shown in Figure 3.

5 Conclusion and Future Work

We experimentally verified our hypothesis using real-world radiation experiments and efficiency analysis on a low-power COTS device with FPU. We observed reliability against soft error-induced bit-flips to remain at similar levels with and without FPU acting as an AI accelerator. Moreover, with the AI accelerator, the execution time and energy efficiency improved by *1.9* times.

Techniques such as SMART [28] or N-module redundancy [45] could be combined with DNN accelerators in COTS to improve the reliability and efficiency of such AI algorithms. Further experiments with different types of AI accelerator-based devices could be directed at this hypothesis to expand its relevance. An extended version of this article is planned to include the mathematical analysis of the hypothesis, detailed methodology and public dataset.

Future works will consider analysing results for more error types, and developing error models and more realistic software-level fault simulators for faster reliability estimates [32], which could be beneficial during design space exploration. While we consider the marginal increase in soft error cross-section by 4% as negligible due to the measurement uncertainty, future experiments could further add insights into this reliability difference as well as the soft error sensitivity of different chip areas used at runtime. Reducing the execution time further with an AI accelerator could increase the reliability and this hypothesis will be investigated in the future.

Acknowledgement

Beam experiments were provided by the ChipIR team thanks to Christopher Frost, Carlo Cazzaniga, and Maria Kastriotou (DOI: 10.5286/ISIS.E.RB2200004-1). This work has been supported by the MOVIQ (Mastering Onboard Vision Intelligence and Quality) project funded by Flanders Innovation & Entrepreneurship (VLAIO) and Flanders Space (VRI) and has received co-funding from the European Union NextGenerationEU; and the European Union's Horizon 2020 research and innovation programme under the grant agreement № 101008126.

References

1. Furano, G. *et al.* Towards the Use of Artificial Intelligence on the Edge in Space Systems: Challenges and Opportunities. *IEEE Aerospace and Electronic Systems Magazine* **35,** 44–56 (2020).
2. Um, J., Gezer, V., Wagner, A. & Ruskowski, M. *Edge Computing in Smart Production* in *Advances in Service and Industrial Robotics* (eds Berns, K. & Görges, D.) (Springer International Publishing, Cham, 2020), 144–152. ISBN: 978-3-030-19648-6.
3. Merenda, M., Porcaro, C. & Iero, D. Edge Machine Learning for AI-Enabled IoT Devices: A Review. *Sensors* **20.** ISSN: 1424-8220. https://www.mdpi.com/1424-8220/20/9/2533 (2020).
4. Ramos, A. *et al.* Artificial intelligence and machine learning applications in the Spanish nuclear field. *Nuclear Engineering and Design* **417,** 112842. ISSN: 0029-5493. https://www.sciencedirect.com/science/article/pii/S002954932300691X (2024).
5. Schweiger, A. *et al. Classification for Avionics Capabilities Enabled by Artificial Intelligence* in *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)* (2021), 1–10.
6. Leon, V. *et al.* Improving Performance-Power-Programmability in Space Avionics with Edge Devices: VBN on Myriad2 SoC. *ACM Trans. Embed. Comput. Syst.* **20.** ISSN: 1539-9087. https://doi.org/10.1145/3440885 (Mar. 2021).
7. Koktas, E. & Basar, E. Communications for the Planet Mars: Past, Present, and Future. *IEEE Aerospace and Electronic Systems Magazine*, 1–35 (2024).
8. Ghiglino, P. & Harshe, M. *A Low Power And High Performance Software Approach to Artificial Intelligence On-Board* in *2023 IEEE Space Computing Conference (SCC)* (2023), 63–70.

9. Lu, P. & Liu, X. Autonomous Trajectory Planning for Rendezvous and Proximity Operations by Conic Optimization. *Journal of Guidance, Control, and Dynamics* **36,** 375–389. eprint: https://doi.org/10.2514/1.58436. https://doi.org/10.2514/1.58436 (2013).

10. Oche, P. A., Ewa, G. A. & Ibekwe, N. Applications and Challenges of Artificial Intelligence in Space Missions. *IEEE Access* **12,** 44481–44509 (2024).

11. Hua, H. *et al.* Edge Computing with Artificial Intelligence: A Machine Learning Perspective. *ACM Comput. Surv.* **55.** issn: 0360-0300. https://doi.org/10.1145/3555802 (Jan. 2023).

12. Pasumarty, R., Praveen, R. & R, M. T. *The Future of AI-enabled servers in the cloud- A Survey* in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (2021), 578–583.

13. NASA. *Mission Timeline Summary.* Available online: https://science.nasa.gov/planetary-science/programs/mars-exploration/mission-timeline/. (accessed on 19 May 2024).

14. NASA-JPL. *Curiosity EDL timeline.* Available online: https://www.jpl.nasa.gov/infographics/curiosity-edl-timeline. (accessed on 26 May 2024).

15. ESA. *Time delay between Mars and Earth.* Available online: https://blogs.esa.int/mex/2012/08/05/time-delay-between-mars-and-earth/. (accessed on 19 May 2024).

16. Wynne, K. Causality and the nature of information. *Optics Communications* **209,** 85–100. issn: 0030-4018. https://www.sciencedirect.com/science/article/pii/S0030401802016383 (2002).

17. Lakrouni, S., Sebgui, M. & Bah, S. *Using AI and IoT at the Edge of the network* in *2022 8th International Conference on Optimization and Applications (ICOA)* (2022), 1–6.

18. Singh, R. & Gill, S. S. Edge AI: A survey. *Internet of Things and Cyber-Physical Systems* **3,** 71–92. issn: 2667-3452. https://www.sciencedirect.com/science/article/pii/S2667345223000196 (2023).

19. Lv, Y.-G. *et al.* Review on Thermal Management Technologies for Electronics in Spacecraft Environment. *Energy Storage and Saving.* issn: 2772-6835. https://www.sciencedirect.com/science/article/pii/S277268352400013X (2024).

20. Luza, L. M., Wrobel, F., Entrena, L. & Dilillo, L. *Impact of Atmospheric and Space Radiation on Sensitive Electronic Devices* in *2022 IEEE European Test Symposium (ETS)* (2022), 1–10.

21. JEDEC. *Measurement and Reporting of Alpha Particle and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices, Std. JESD89B, Sep. 2021.* Available online: https://www.jedec.org/system/files/docs/JESD89B.pdf. (accessed on 28 Jan 2024).

22. KÃ¶hler, J. *et al.* Measurements of the neutron spectrum in transit to Mars on the Mars Science Laboratory. *Life Sciences in Space Research* **5,** 6–12. issn: 2214-5524. https://www.sciencedirect.com/science/article/pii/S2214552415000164 (2015).

23. Cazzaniga, C., Bagatin, M., Gerardin, S., COSTANTINO, A. & FROST, C. D. First Tests of a New Facility for Device-Level, Board-Level and System-Level Neutron Irradiation of Microelectronics. *IEEE Transactions on Emerging Topics in Computing* **9,** 104–108 (2021).

24. Chen, Z., Li, G. & Pattabiraman, K. *A Low-cost Fault Corrector for Deep Neural Networks through Range Restriction* in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (2021), 1–13.

25. Hasan Ahmadilivani, M. *et al. Special Session: Reliability Assessment Recipes for DNN Accelerators* in *VTS 2024 - IEEE VLSI Test Symposium* **10** (IEEE, Tempe AZ USA, United States, Apr. 2024), 131788–131828. https://hal.science/hal-04572731.

26. Li, J. & Draper, J. Accelerated Soft-Error-Rate (SER) Estimation for Combinational and Sequential Circuits. *ACM Trans. Des. Autom. Electron. Syst.* **22.** issn: 1084-4309. https://doi.org/10.1145/3035496 (May 2017).

27. Dos Santos, F. F., Navaux, P., Carro, L. & Rech, P. *Impact of Reduced Precision in the Reliability of Deep Neural Networks for Object Detection* in *2019 IEEE European Test Symposium (ETS)* (2019), 1–6.

28. Rajappa, A. J. *et al.* SMART: Selective MAC zero-optimization for neural network reliability under radiation. *Microelectronics Reliability* **150,** 115092. issn: 0026-2714. https://www.sciencedirect.com/science/article/pii/S0026271423001920 (2023).

29. Hodson, Robert F. and et. al. *Technical Memorandum: Recommendations on Use of Commercial-Off-The-Shelf (COTS) Electrical, Electronic, and Electromechanical (EEE) Parts for NASA Missions.* Available online: https://ntrs.nasa.gov/citations/20205011579. (accessed on 21 May 2024).

30. Taheri, M. *et al.* AdAM: Adaptive Approximate Multiplier for Fault Tolerance in DNN Accelerators. http://dx.doi.org/10.36227/techrxiv.171502587.72983622/v1 (May 2024).

31. in. *Practical Microcontroller Engineering with ARM® Technology* 927–950 (John Wiley & Sons, Ltd, 2015). isbn: 9781119058397. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119058397.ch11. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119058397.ch11.

32. Ruospo, A. *et al.* A Survey on Deep Learning Resilience Assessment Methodologies. *Computer* **56,** 57–66 (2023).

33. Cazzaniga, C. & Frost, C. D. Progress of the Scientific Commissioning of a fast neutron beamline for Chip Irradiation. *Journal of Physics: Conference Series* **1021,** 012037. https://dx.doi.org/10.1088/1742-6596/1021/1/012037 (May 2018).

34. Chiesa, D. *et al.* Measurement of the neutron flux at spallation sources using multi-foil activation. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **902,** 14–24. issn: 0168-9002. https://www.sciencedirect.com/science/article/pii/S016890021830737X (2018).

35. Coronetti, A. *et al.* Radiation Hardness Assurance Through System-Level Testing: Risk Acceptance, Facility Requirements, Test Methodology, and Data Exploitation. *IEEE Transactions on Nuclear Science* **68,** 958–969 (2021).

36. Cong, J. *et al. Accelerator-Rich Architectures: Opportunities and Progresses* in *Proceedings of the 51st Annual Design Automation Conference* (Association for Computing Machinery, San Francisco, CA, USA, 2014), 1–6. isbn: 9781450327305. https://doi.org/10.1145/2593069.2596667.

37. Taheri, M. *et al. DeepAxe: A Framework for Exploration of Approximation and Reliability Trade-offs in DNN Accelerators* in *2023 24th International Symposium on Quality Electronic Design (ISQED)* (2023), 1–8.

38. Rupprecht, B., Hujo, D. & Vogel-Heuser, B. *Performance Evaluation of AI Algorithms on Heterogeneous Edge Devices for Manufacturing* in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)* (2022), 2132–2139.

39. Nordic semiconductors. *nRF52840 DK.* Available online: `https://www.nordicsemi.com/Products/Development-hardware/nrf52840-dk`. (accessed on 28 Jan 2024).

40. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).

41. SEGGER. *SEGGER compiler.* Available online: `https://wiki.segger.com/SEGGER_compiler`. (accessed on 28 Jan 2024).

42. SEGGER. *SEGGER Embedded Studio Reference Manual.* Available online: `https://www.segger.com/downloads/embedded-studio/EmbeddedStudio_manual.pdf`. (accessed on 23 May 2024).

43. ARM. *Chapter 9. Data Watchpoint and Trace Unit.* Available online: `https://developer.arm.com/documentation/ddi0439/b/Data-Watchpoint-and-Trace-Unit?lang=en`. (accessed on 28 Jan 2024).

44. Joulescope. *Joulescope™ JS110 User's Guide.* Available online: `https://download.joulescope.com/docs/JoulescopeUsersGuide/JoulescopeUsersGuide_v1_1.pdf`. (accessed on 28 Jan 2024).

45. Koren & Su. Reliability Analysis of N-Modular Redundancy Systems with Intermittent and Permanent Faults. *IEEE Transactions on Computers* **C-28,** 514–520 (1979).