New Techniques for STL generation for Vector Processing Units

Gustavo Vilar de Farias*, Josie E. Rodriguez Condia*, Matteo Sonza Reorda*
*Politecnico di Torino - Department of Control and Computer Engineering (DAUIN)
{gustavo.vilarde, josie.rodriguez, matteo.sonzareorda}@polito.it

***Abstract*—[1] Advances in technology nodes, activity stress caused by data-intensive workloads, and harsh operating conditions increase the occurrence of faults in edge devices used in Artificial Intelligence domains, severely impacting their reliability. Among diverse accelerators, Vector Processing Units (VPUs) offer a versatile and flexible architecture for accelerating workload execution by exploiting data-level parallelism. The extensive parallelism and their customizable architecture significantly increase the complexity of reliability analysis and fault effect mitigation, a problem incompletely addressed in the literature.**

This paper describes a method for generating Software Test Libraries (STLs) to detect permanent faults arising during the in-field operational phase. The method not only generates an STL for the baseline accelerator, but also customizes it for any VPU configuration. The STL for the baseline VPU configuration is built by first exploiting application code, identifying suitable data patterns to improve fault detectability, and then adding a few Ad Hoc routines. The resulting test, once the target hardware parameters are known, is customized using the proposed framework to ensure its effectiveness on the final VPU configuration. The strict time constraints of in-field tests are taken into account by also minimizing the test duration.

The experimental results, obtained on a configurable VPU for edge AI applications (Klessydra-T), show that our framework supports the development of effective STLs (up to 96% stuck-at Fault Coverage (FC) for large size configurations). Remarkably, we observed that using a generic STL without taking into account the specific hardware configuration may cause an FC drop of about 25%, highlighting the need for a framework that automatically customizes the test according to the VPU's parameters. The framework is particularly attractive for scalable devices where long-term reliability is essential, and tests must be designed for different hardware configurations.

***Index Terms*—Edge AI, Stress maximization, Vector accelerators**

[1]This work was partially supported by the TIRAMISU project under the EU Grant#101169378 — HORIZON-MSCA-2023-DN-01.