National Aeronautics and Space Administration

Algorithm-based Fault Tolerance: Handling Bit-flips in Neural Network computations on Snapdragon

Principal Investigator: Lini Mestar (174); Co-Investigators: Hamsa Shwetha Venkataram (174), Yutao He (349), Jack Kosaian (Carnegie Mellon University), Rashmi Korlakai Vinayak (Carnegie Mellon University)

Program: FY22 R&TD Innovative Spontaneous Concepts

Objectives:

The overall objective of the 4-month task is to demonstrate feasibility and effectiveness of employing Algorithm-based Fault Tolerance (ABFT) to handle bit flips in neural network models deployed on Qualcomm Snapdragon SoC for the task of image classification, while preserving the additional bits when passed through the Snapdragon Neural Processing Engine (SNPE) optimizers to advance the spacecraft avionics system flight software to TRL 5, and to demonstrate system performance capabilities. Figure 1. demonstrates the preliminary design of ABFT-as-a-service.

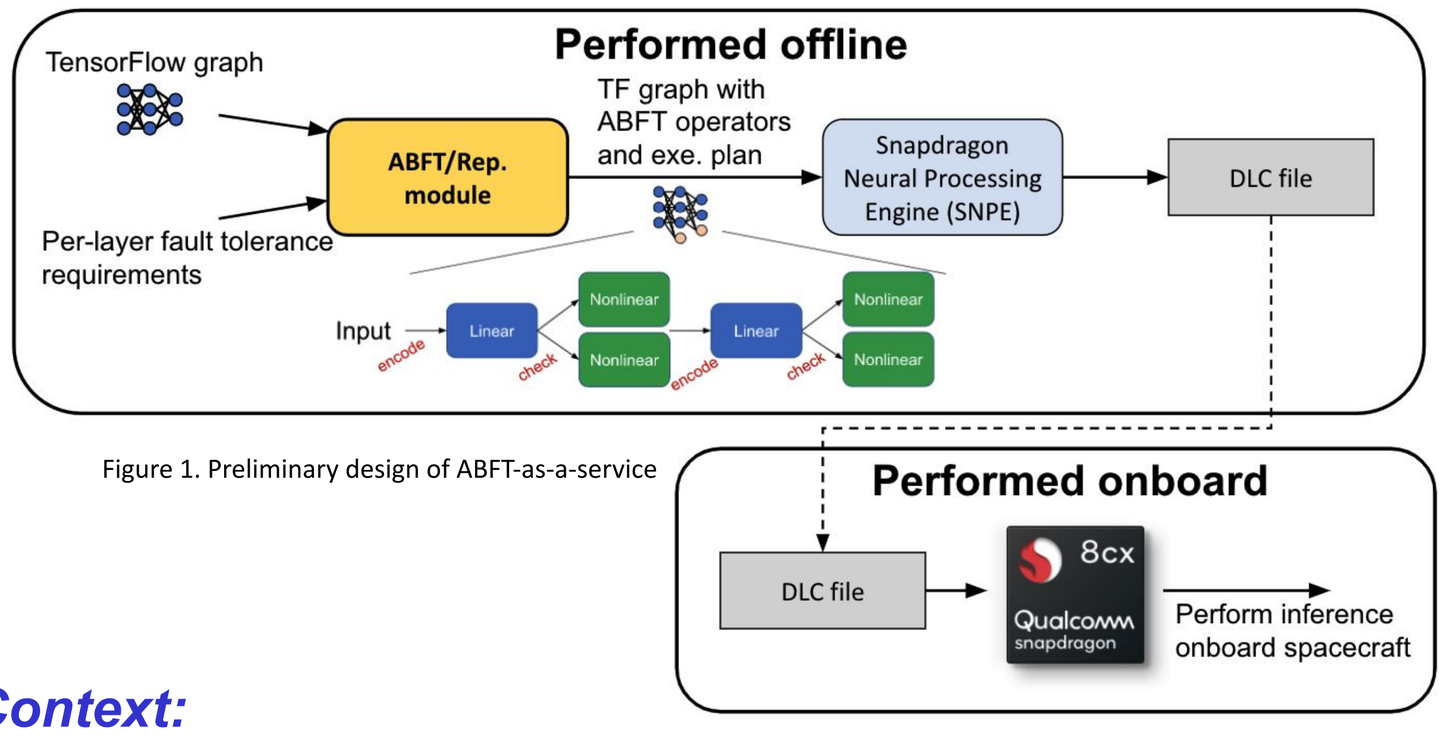

Figure 1. Preliminary design of ABFT-as-a-service

Technical Approach:

The technical approach involved the development of avionics hardware and key software, as well as system integration and testing. We used a remote testbed that was linked to the Snapdragon 8155 development kit. This is the cutting-edge development kit used by JPL for Snapdragon development, namely by the Qualcomm Co-Processor (QCP) team.

To achieve the objectives, we divided the task into three well-defined subtasks to:
(a) Demonstrate that ABFT can enable neural network inference to run on COTS SoC with less overhead than Triple-modular redundancy
(b) Complete and demonstrate the ABFT port to Snapdragon utilizing Qualcomm's optimization tools while retaining all ABFT layers
(c) Demonstrate ABFT metrics for overhead (e.g. execution-time, energy-consumption, and memory).

Context:

Enabling autonomy using neural network models requires highly reliable, fault-tolerant processor platforms that operate under effects of radiation. To ensure reliability under these operational constraints, it is increasingly important to have error detection and correction (EDAC). COTS products like Qualcomm Snapdragon, however, lack such features. ABFT will add robustness and reliability in deep neural networks (DNN) by handling bitflips with minimal overhead as opposed to TMR. In a previously funded Innovative Spontaneous Concept R&TD, our team was able to demonstrate porting an ABFT enabled CNN model onto the Snapdragon using SNPE, gathering system metrics, and maintaining the same resilience as the standard TMR with less overhead. Our long-term goal is to fly autonomous spacecraft using DNN models aboard SoCs like Snapdragon. To that end, ABFT for DNN is a critical but mostly unexplored component.

Innovation:

Employing ABFT for fault-tolerant neural network inference on the Snapdragon SoC in space environments significantly decreases the amount of hardware required in state-of-the-art methods i.e Triple Modular Redundancy (TMR), resulting in considerable cost savings, minimum overhead, and reduction in the time necessary for hardware replication.

Results:

We were able to demonstrate, using various experiments, that employing **ABFT in neural network inference** results in ~65% decrease in model size, ~75% decrease in energy and power consumption and reduced overhead (~1.03x) when compared with state-of-the-art methods i.e TMR.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

www.nasa.gov

Clearance Number: CL# URS311477
Poster Number: RPC#R22212

PI/Task Mgr. Contact Information:
Email: Lini.Mestar@jpl.nasa.gov