National Technical University of Athens, Greece
School of Electrical and Computer Engineering
Microprocessors Lab

EXECUTIVE SUMMARY

CAIRS21 4000135491/21/NL/GLC/ov

Prepared by: George Lentaris,
Vasileios Leon,
Chronis Sakos,
Panagiotis Minaidis,
Simon Vellas (OHB-Hellas)
Mathieu Bernou (OHB-Hellas)

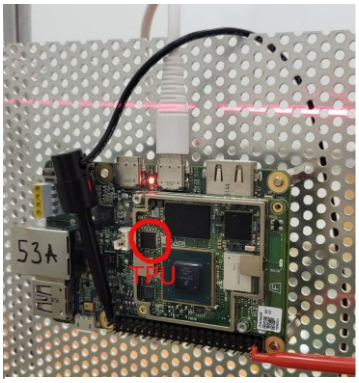
Approved by: George Lentaris
Authorized by: Prof. Dimitrios Soudris

Code: NTUA-CAIRS21-ES
Version: 1.0
Date: 25/11/2022
Contract no.: 4000135491/21/NL/GLC/ov

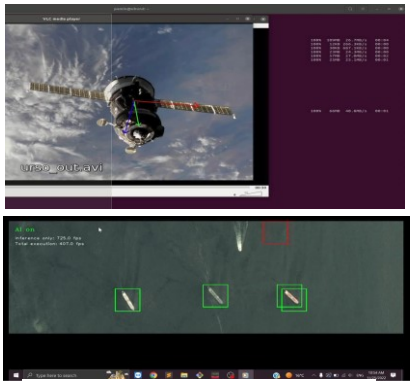
EXECUTIVE SUMMARY

The ESA OSIP activity CAIRS21 studied the performance and suitability of COTS AI accelerators for payload processing in space. CAIRS21's primary focus was the Google Coral Edge TPU and, in particular, AI USB accelerators that can be incorporated into mixed-criticality architectures together with FPGAs. Following the success of AI/ML in terrestrial applications and the urge to transfer such technology to space, the TPU was identified among all candidate devices for AI high-performance embedded computing. Due to the post-covid market disruption, the USB version of TPU was temporarily unavailable and CAIRS21 also involved the NCS2 Myriad USB accelerator to its architectural exploration. Overall, the CAIRS21 outcome is fourfold:

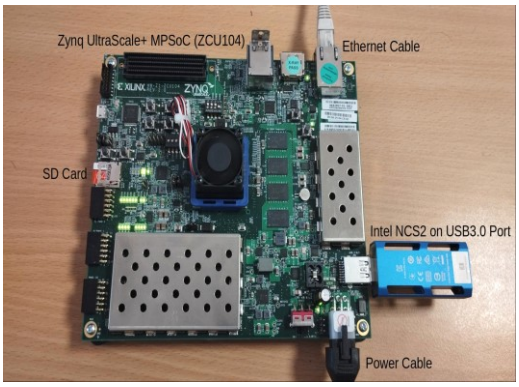
a) performed extensive benchmarking of TPU with respect to performance and programmability
b) evaluated the resilience of TPU to TID radiation
c) introduced an AI/ML USB accelerator to a mixed-criticality architecture based on SoC FPGA
d) analyzed the space market in terms of current HW & SW products and application requirements for EO



TPU under irradiation test



AI apps (pose est., ship det.)



Architecture with FPGA + AI accelerator

The CAIRS21 study was completed successfully by generating a variety of results on the aforementioned directions. In particular, we performed a thorough evaluation of the latest Edge TPU processor, including performance, power, size, programmability, and radiation tolerance. Furthermore, we compared TPU to other competitive COTS devices for space, i.e., ARM CPU, Zynq FPGA, Myriad VPU, and low-power NVIDIA GPU. The comparative study was based on hands-on evaluation of a variety of AI/ML benchmarks to cover virtually all mainstream AI/ML applications. For radiation tolerance, we performed our own TID test of the TPU chip and prepared a SEE proton test as the immediate next step after CAIRS21. Besides the TPU processing core itself, CAIRS21 studied its integration in mixed-criticality architectures, i.e., how TPU/NCS2 can be actually deployed in space to offload computationally intensive tasks in real-time. In this direction, besides the standalone TPU SOM, we considered a USB accelerator (NCS2 in absence of TPU) integrated with a highly complex MPSoC FPGA. The architectural study focused more on the SW side, i.e., on the development frameworks that need to be combined, the partitioning of the algorithms, and the virtualization methods that can ease the use of the proposed avionics by external users. On top of the above, CAIRS21 performed a market analysis to understand the potential use of such products in the space industry, i.e., we analyzed gaps, trends, and requirements of upcoming space missions to position AI/ML and TPU in future high-performance payload processors.

The results were promising for the majority of examined aspects. First of all, the performance of TPU for mid-sized CNN/MLP algorithms was unprecedented. In particular, when assuming AI applications within the context of embedded computing (with relatively limited algorithmic requirements, as opposed to server computing and as would be more reasonable for space missions), then TPU outperforms all other candidates: it provides almost two orders of magnitude acceleration versus CPU and almost one order of magnitude versus small GPU/VPU devices, while it can also become twice as fast as Zynq MPSoC FPGA (assuming high-level programming in all cases). When also taking into account the small power consumption of TPU, i.e., 2W for the chip or 5W for the SOM, then TPU improves performance/Watt by an order of magnitude versus its competitors (less for Zynq FPGA, for which however TPU is still more than 2x better). In nearby application domains, e.g., for large AI/ML algorithms or classical digital processing, the TPU becomes worse than VPU/GPU/FPGA, either due to its limited on-chip memory or its decreased set of supported functions (only few Python TF operations). Farther on the downside, TPU can act only as a co-

processor (requires a separate CPU), supports only 8bit variables, and cannot accelerate AI training. That is to say, TPU is useful only to a specific subset of applications. However, its very small form factor accommodates packing with other standalone avionics (TPU chip size is only 5x5 mm2 and SOM size is 48x48 mm2). Secondly, the TPU proved quite resilient to TID irradiation by withstanding almost 50 Krad(Si) in our own test. Therefore, it seems suitable for LEO missions at 1000+ Km altitude and 5+ year lifetime with relatively large safety margins. Thirdly, when considering some form of USB acceleration of AI/ML in mixed-criticality architectures, supported by high-level programming frameworks and virtualization methods, we achieved a complex combination of all desired HW & SW tools on the Zynq FPGA. Thus, as already tested, we enabled potential third-party users to easily deploy and accelerate their algorithm on our architecture without compromising its original accuracy/performance. Such a capability will be greatly appreciated in upcoming industrial use cases, as derived by our market analysis, which revealed today a relative lack of complete, mature HW+SW solutions, despite the need to provide space customers with well-understood development frameworks.