IAF SPACE SYSTEMS SYMPOSIUM (D1)
Technologies to Enable Space Systems (3)

Author: Mr. Nan Li
University of Chinese Academy of Sciences; Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, China, linan@csu.ac.cn

Mr. Aimin Xiao
Chinese Academy of Sciences, China, am.xiao@csu.ac.cn
Mr. Mengxi Yu
Technology and Engineering Center for Space Utilization,Chinese Academy of Sciences, China, yumengxi@csu.ac.cn
Dr. Jianquan Zhang
Chinese Academy of Sciences, China, jqzhang@csu.ac.cn
Dr. Wenbo Dong
Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, China, wbdong@csu.ac.cn

# APPLICATION OF GPU ON-ORBIT AND SELF-ADAPTIVE SCHEDULING BY ITS INTERNAL THERMAL SENSOR

**Abstract**

High-performance COTS components such as GPU and FPGA have been widely used in the applications of big data technology and artificial intelligence technology on the ground. In some on-orbit systems for high-performance applications where GPU components have to be used, the commercial components must be chosen and its strict requirement on power and thermal conditions must be considered, including power supply capability, thermal convection or thermal conduction conditions and so on.

This paper presents a method operating on GPU processors to achieve acceptable computing performance while keeping the temperature sampled from GPU internal thermal sensor within a reasonable range. The feedback control strategy is designed with the sampled GPU temperature as input and the recommended amount of parallel computing resources occupied as output. In other words, the heat of the running GPU is reduced by degrading the GPU performance properly. Furthermore, multiple GPU workloads are scheduled in a single processor and the number of concurrent threads is adjusted correspondingly in a fine-grained manner.

Fuzzy logical control theory is used to estimate the correlation between the sampled GPU temperature and the number of typical concurrent thread operations. With regard to a single workload, dependency graph is applied to guarantee the computing sequence of different parts inside the workload when it has to limit the degree of parallelism. As to multi-independent workloads, different priorities are assigned to these workloads in order to guarantee the performance of workloads with high priorities.

Because the GPU processor of Jetson TX2 module is a system on chip, a simple self-adaptive scheduling framework based on the above method is implemented in ARM cores of the processor. Moreover, ethernet communication interface is supported by the framework and it is easy to expand the framework to the circumstances in which several Jetson TX2 modules are interconnected or several docker containers are interconnected.

When the Jetson TX2 module is put into a thermal chamber, experiments show that this method can successfully complete the auto adjustment of scheduling strategy in accordance with the environment

temperature variation. As a result, it reduces the negative impact from the bad situations when thermal throttling is triggered, which incurs large fluctuations of the GPU frequency and power consumption.