FPGA processors in AI edge applications

by Firgan Feradov | Angel Marinov | Technical University - Varna | Technical University - Varna

In the field of AI technologies, the term “Edge” is used to define a broad range of devices which allow for data gathering, processing, storage and communication outside of a data center or the cloud. The group of edge applications encompasses a large number of varying devices such as smart watches, cameras, wearable devices, robots, industrial sensors, drones and many more. As such, these devices are closest to the specific object or target of the AI system and interact most dynamically with it. As a result, Gartner predicts that by 2025 75% of all enterprise data will be generated outside traditional data centers, and business should consider more decentralized AI solutions.

The implementation of AI systems on edge devices has two main advantages. The first one is that it facilitates the emergence of new real-time use cases, as edge devices allow for the gathering and processing of a wide range of different data which opens up the possibility for interactive applications. The second major advantage is the improvement of the data security parameters of the AI system as edge processing prevents sensitive data to be sent to data centers, cloud services or be transmitted through the network.

The key elements for the implementation of a successful edge AI application are: high performance, high capacity, low latency and robust security. These requirements are set by the operational conditions of edge systems – they generate large amounts of computationally expensive data which has to be processed in a limited time, in accordance with the specifications of the real-time task which they perform. In addition to that, the gathered data must be securely processed and stored, as most of the modern AI systems require large volumes of sensitive data.

Field Programmable Gate Array (FPGA) processors offer a technological solution to the challenges faced in the design of edge applications. FPGAs offer bit level dynamic programmability for the logical circuit design of the implemented functional blocks, as well as the connections between the both the different functional blocks within the processor and the functional blocks and the physical pins of the FPGA. This allows designers to create custom solutions, tailored specially for the operational parameters of the system in which they will be used. Specifically, FPGA processors can be used both as a sole processor on which the entire AI system is implemented as well as in combination with other processors (CPU, GPU) or implementation of Systems of Chip (SoC). In the first case the computational load handled by the FPGA processor directly through silicon level solutions which decreases the latency of the system. In the second case the FPGA can be used both as a main processor communicating with GPU or VPU, or it can be configured as an application-specific device for.