"AI on the Fly" - training and inferencing on a local platform

The advantages of "AI on the Fly" and how local inference influences deep learning for autonomous driving

Rapid technological advances, groundbreaking innovations and an actual need from an economic perspective ensure that artificial intelligence is no longer restricted to research laboratories and supercomputers. Artificial intelligence is the technological driver for the future and is now at the forefront of the further development of companies and entire industries.
AI-based content generation is already available, including for e-commerce platforms and multidimensional pattern recognition based on camera and sensor data. Although edge and cloud computing have been proven advantageous for such appliances and scalable concepts have already been implemented, there are fields of applications that rely on even lower latency.

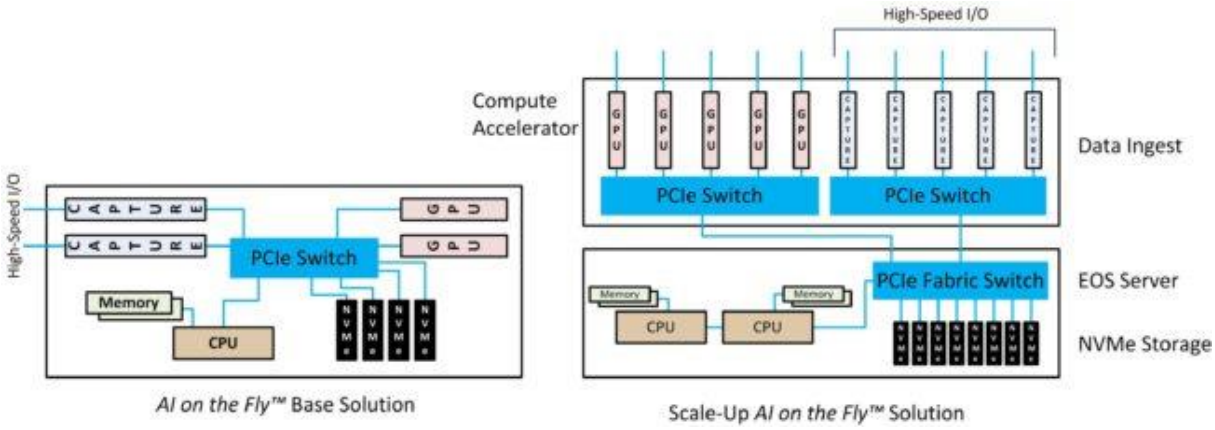
Computing and storage resources for the entire AI workflow

Companies such as BRESSNER Technology and its parent company One Stop Systems are therefore pursuing novel approaches for training AI algorithms. "AI on the Fly" mixes high-performance GPU-based computing processes with traditional edge computing. While a local edge device sends the data to a decentralized inference machine, training and inference for "AI on the Fly" takes place on a local level.
"AI on the Fly" therefore provides computing and storage resources for the entire AI workflow, not in the data center, but at the edge near the data sources. Applications for this new AI paradigm are emerging in various areas. These include autonomous vehicles, preventive personalized medical technology, control of defense solutions and industrial automation. The basic similarities of these solutions are the acquisition of high data rates, the storage with high speed and low latency as well as efficient high-performance AI training and inference computing. All of these building block elements are seamlessly connected to the PCI Express system with appropriate memory allocation. It is interconnected and adapted to meet the specific environmental requirements of on-site installations.

Flexible subsystems for autonomous driving and data acquisition

"AI on the Fly" consists of three modular subsystems. Data acquisition, data storage and computing engines. These subsystems support high-speed components such as data acquisition hardware, NVMe SSD storage as well as GPU and FPGA compute accelerators. PCIe interfaces ensure flexible scaling with high bandwidth and low latency. The data acquisition system must be able to absorb the huge amount of data which flows continuously from the sensors. In addition, the data must be processed for efficient transmission, both to the solid-state memory and to the computing modules. The PCIe range of functions enables multiple transfer of data to other subsystems simultaneously using RDMA transfers in order to avoid a system memory bottleneck

without additional network protocol latency. The computing features include machine learning tasks, data analysis, deep learning training tasks using neural network frameworks and inference engines using trained models based on new data. Special GPU resources are usually required for each of these elements. "AI on the Fly" offers all these elements in flexible building block components that can be easily adapted to the specific requirements of a vehicle developer for autonomous transport. The following figure shows an example of "AI on the Fly" configurations for autonomous vehicles.



AI on the Fly™ Base Solution

Scale-Up *AI on the Fly™* Solution

High-speed data acquisition technology is at the front end of these systems. Depending on the application, the data can be generated by a large number of sensors. In autonomous vehicles, data is generated via arrays of video and LIDAR sensors. Radar, sonar, FLIR (infrared) and HF sensors are used in field operations. Medical applications use MRI or CT sensors. In security applications, surveillance camera networks generate large amounts of video data. Industrial automation, on the other hand, includes telemetry data from IoT sensors and video feeds with high frame rates.

Multi-GPU performance that fits in the trunk

BRESSNER Technology works with One Stop Systems and industry leaders to provide technology for autonomous vehicle development programs and high-speed data collection. The development partners rely on the experience of OSS in the development of scalable PCI Express-based systems. These combine sensor data subsystems with high bandwidth, NVMe memory with low latency and high-performance multi-GPUs in special robust form factors. OSS recently announced a joint design win for a large international network transportation company. For its development fleet for autonomous driving, "AI on the Fly" components were used in 150 vehicles. This fleet is used to collect the data needed to develop and test the artificial intelligence algorithms that will eventually be used in thousands of commercial vehicles. In this case, the "AI on the Fly" data acquisition system is linked to a large number of video, radar and LIDAR sensors in the car. These are combined via redundant 50 Gbit/s Ethernet connections to the storage subsystem and are then connected directly to multi-GPU machine learning training systems and inference machines. The entire system is implemented in the trunk of the respective automobile.