NEWS RELEASE

Penguin Computing Announces Open Compute-based Deep Learning Solution with NVIDIA Tesla P40 and P4 GPU Accelerators

2016-10-27

Computationally Intensive Requirements Drive Need for Exponentially Higher Throughput and Real Time Responsiveness Provided by GPUs

WASHINGTON, D.C. - October 26, 2016 - Penguin Computing, a provider of high performance, enterprise data center and cloud solutions, today announced availability of Open Compute- based Tundra Relion X1904GT and Tundra Relion 1930g servers, powered by NVIDIA® Tesla® P40 and P4 GPU accelerators, respectively, for deep learning training and inference requirements. Tundra Relion X1904GT, which already supports NVIDIA Tesla P100 and M40, now adds Tesla P40 to its accelerated computing platform offering design innovation, allowing customers to choose from a variety of topologies to accelerate applications. The Tundra Relion X1904GT also incorporates NVIDIA NVLink™ high-speed interconnect technology for best peer-to-peer performance. Tundra combines the operational benefits and efficiency of Open Computing with a high density and high performance architecture needed to meet the requirements of the most demanding compute intensive applications. Penguin Computing's Tundra platform demonstrates the company's leadership in providing flexible, open platforms for the high performance computing market. "Pairing Tundra Relion X1904GT with our Tundra Relion 1930g, we now have a complete deep learning solution in Open Compute form factor that covers both training and inference requirements," said William Wu, Director of Product Management at Penguin Computing. "With the ever evolving deep learning market, the X1904GT with its flexible PCI-E topologies eclipses the cookie cutter approach, providing a solution optimized for customers' respective applications. Our collaboration with NVIDIA is combating the perennial need to overcome scaling challenges for deep learning and HPC." NVIDIA Tesla GPU accelerators enable real-time responsiveness for the most complex deep learning models. Responsiveness is critical to user adoption for services such as interactive speech, visual search and video recommendations. As models increase in accuracy and complexity, CPUs are no longer capable of delivering interactive user experience. The NVIDIA Tesla P40 and P4

deliver more than 40X faster inference performance with INT8 operations for real-time responsiveness, even for the most complex models. Tesla GPU accelerators also meet exploding higher throughput requirements based on ever-growing data needs, in the form of sensor logs, images, videos and transactional records. “Penguin Computing’s Open Compute servers, powered by NVIDIA Pascal architecture Tesla accelerators, are leading-edge deep learning solutions for the data center,” said Roy Kim, Director of Accelerated Computing at NVIDIA. “Tesla P40 and P4 GPU accelerators deliver 40x faster throughput for training and 40x responsiveness for inferencing compared to what was previously possible.” Visit Penguin Computing’s booth at NVIDIA’s GPU Technology Conference, October 26-27, in Washington, D.C.