NEWS RELEASE

# NVIDIA Tesla V100-Based Servers Drive Deep Learning and AI

9/27/2017

PENGUIN COMPUTING™ ANNOUNCES NVIDIA® TESLA® V100-BASED SERVERS TO DRIVE DEEP LEARNING, ARTIFICIAL INTELLIGENCE

FREMONT, CA – September 28, 2017 – Penguin Computing, provider of high performance computing, enterprise datacenter and cloud solutions, today announced strategic support for the field of artificial intelligence through availability of its servers based on the highly-advanced NVIDIA® Tesla® V100 GPU accelerator, powered by the NVIDIA VoltaTM GPU architecture. “Deep learning, machine learning and artificial intelligence are vital tools for addressing the world’s most complex challenges and improving many aspects of our lives,” said William Wu, Director of Product Management, Penguin Computing. “Our breadth of products covers configurations that accelerate various demanding workloads – maximizing performance, minimizing P2P latency of multiple GPUs and providing minimal power consumption through creative cooling solutions.” NVIDIA Tesla V100 GPUs join an expansive GPU server line that covers Penguin Computing’s Relion® servers (Intel®-based) and Altus® servers (AMD-based) in both 19” and 21” TundraTM form factors. Penguin Computing will debut a high density 21” Tundra 1OU GPU server to support 4x Tesla V100 SXM2, and 19” 4U GPU server to support 8x Tesla V100 SXM2 with NVIDIA NVLink™ interconnect technology optional in single root complex. The NVIDIA Volta architecture is bolstered by pairing NVIDIA CUDA® cores and NVIDIA Tensor Cores within a unified architecture. A single server with Tesla V100 GPUs can replace hundreds of CPU servers for AI. Equipped with 640 Tensor Cores, Tesla V100 delivers 125 TeraFLOPS of deep learning performance. That’s 12X Tensor FLOPS for deep learning training, and 6X Tensor FLOPS for deep learning inference when compared to NVIDIA Pascal™ GPUs. “Penguin Computing continues to

demonstrate leadership by providing Volta-based systems to support critical AI research," said Paresh Kharya, Group Product Marketing Manager, NVIDIA. "Tesla V100 systems will enable their customers to create innovative AI products and services by accelerating their AI research and deployments." Today's announcement reinforces Penguin Computing's philosophy and broader capabilities as a full-spectrum provider offering complete solutions. This includes tailored, custom designs that are supportable and scale to large deployments, and fully engineered and architected designs.

About Penguin Computing

Penguin Computing is one of the largest private suppliers of enterprise and high-performance computing solutions in North America and has built and operates the leading specialized public HPC cloud service Penguin Computing On-Demand (PODTM). Penguin Computing pioneers the design, engineering, integration and delivery of solutions that are based on open architectures and comprise non-proprietary components from a variety of vendors.

Penguin Computing is also one of a limited number of authorized Open Compute Project (OCP) solution providers leveraging this Facebook-led initiative to bring the most efficient open data center solutions to a broader market, and has announced the Tundra product line which applies the benefits of OCP to high performance computing.

Penguin Computing has systems installed with more than 2,500 customers in 40 countries across eight major vertical markets. Visit **www.penguincomputing.com** to learn more about the company and follow **@PenguinHPC** on Twitter.

---