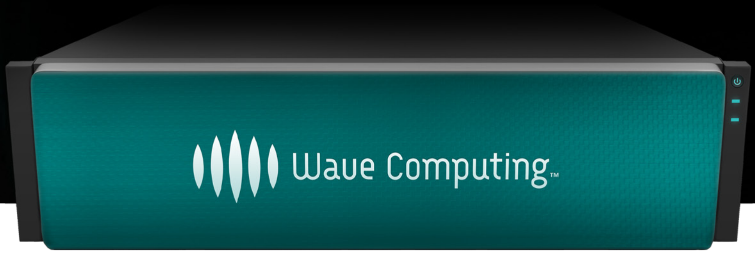
Wave Computing™



Introducing the World's Fastest Dataflow Computer for Machine Learning

Wave's Family of Machine Learning Computers Delivers up to 1000x the Performance for Neural Network Training

Finally, data scientists and researchers have the performance they've been asking for: Wave's family of machine learning computers that are ideal for training both deep and shallow neural networks.

Wave's computers are based on a revolutionary dataflow architecture that improves training performance up to 1000x compared to CPUs, GPUs and FPGAs, giving the data scientist faster results and improved accuracy. Each Wave machine learning computer comes in a 3U form factor that easily fits into existing data center environments. All come with Wave's software, programming tools and dataflow agent libraries.

A Future Proof Machine Learning Computer

Initially supporting TensorFlow, Wave's computers can support a range of frameworks such as CNTK, MXNet and more. Also, the Dataflow Processing Unit (DPU)-based boards within each Wave computer are upgradable, allowing for faster high-bandwidth memory clusters and future generations of Wave DPUs to be added over time.

Specifications (for each Wave Machine Learning Computer)

Performance	Performance/computer (peak)	2.9 PetaOPS/second
	Performance/node (peak)	11.6 PetaOPS/second
	Dataflow Processing Elements (PE's)	Up to 256,000 (16,000 PE's per Wave DPU chip)
Scalability	Wave machine learning computers per data center node	Up to 4 computers delivering 1,000,000 PE's
Memory	High-speed memory	128 GB HMC DRAM
	SSD storage	16 TB
	Bulk storage	2 TB DDR4 DRAM
Connections	Data center backbone connection	10 GbE or 40 GbE
	High-speed inter-computer communication within a single data center node	Wave's proprietary communication system that connects up to 4 computers within a single data center node
Physical	Data center form factor	Each Wave computer comes in a 3U form factor; up to 4 computers can be added per data center node
	Dimensions per each 3U computer	866D x 444W x 131H (mm)
	Operating temperature	10° – 35° C
Software	Machine learning framework	TensorFlow (initially)
	Operating system for Wave Session Manager server	Linux Server
	Library	WaveFlow Agent Library
	Development toolkit	WaveFlow SDK
	Data runtime	WaveFlow Execution Engine

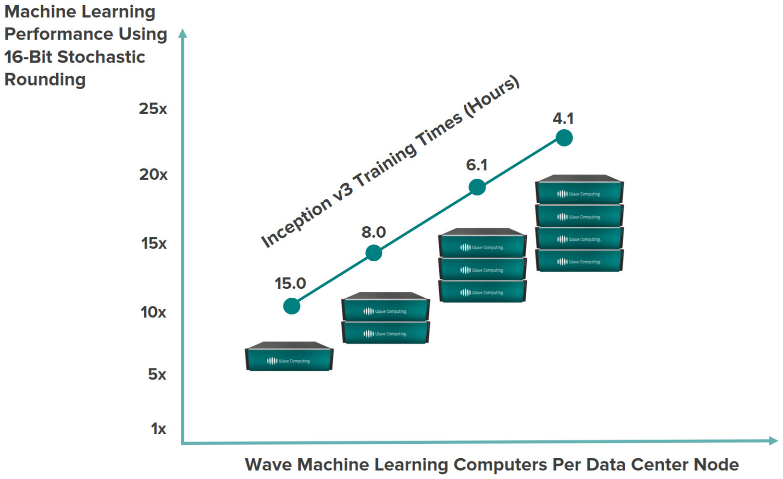
Wave's Revolutionary Dataflow Architecture

Wave's native dataflow architecture is the fundamental technology behind each of Wave's machine learning computers. It is built upon a revolutionary dataflow computing technology that eliminates the traditional CPU/GPU co-processor structure and associated memory bottlenecks, as well as inter-node communication between Wave's systems. This allows Wave's dataflow solutions to exploit data and model parallelisms present in deep learning models, such as convolutional and recurrent neural networks.

Wave's dataflow systems utilize Dataflow Processing Units (DPUs), which contain thousands of interconnected dataflow Processing Elements (PE's). The performance and scalability of Wave's systems make them ideal for organizations using machine learning to easily develop, test and deploy their deep learning models for frameworks such as TensorFlow. To help the data scientist speed time to results, Wave's systems include complete software solutions and stacks: the WaveFlow SDK, the WaveFlow Agent Library, WaveFlow Execution Engine, and the Wave Machine Learning Framework Interface.

Training a Deep Convolutional Neural Network

Google's Inception v3 model can be trained by a Wave dataflow computer in 15 hours using a WaveFlow Agent library that supports 16-b fixed point computations with stochastic rounding. This represents a 10X performance improvement over current GPU-based systems. Training times can be further reduced to as little as four hours by distributing the computations across multiple Wave computers a single node.



Training a Shallow Recurrent Neural Network

Using a single DPU within a single Wave machine learning computer, Wave's solution can deliver more than 500x faster training for Word2Vec compared to CPU- or GPU-based systems. Wave uses a mix of 16-b, 32-b and 64-b fixed point arithmetic.

Platform (skip-gram Hierarchical Softmax)	Time to Train 17M Words
TensorFlow on AWS using 4 CPUs	76 min
TensorFlow on AWS using 4 CPUs and 1 K80 GPU	69 min
TensorFlow on Google Cloud using 4 CPUs	68 min
Single Wave Machine Learning Computer (current generation; single DPU only)	**6.75 sec**

Unmatched Scalability within the Data Center

Wave's dataflow architecture enables the scaling to 4 computers within a single data center node. Wave's high-speed, proprietary communication system eliminates the need to use the data center backbone for intra-node communication, reduces network congestion, and enables a single data center rack to accommodate up to 12U machine learning computers.

2.9 PetaOPS /sec	5.8 PetaOPS /sec	8.7 PetaOPS /sec	11.6 PetaOPS /sec
16 DPUs	32 DPUs	48 DPUs	64 DPUs
128GB High Speed Memory	256GB High Speed Memory	384GB High Speed Memory	512GB High Speed Memory
16TB SSD Storage	32TB SSD Storage	48TB SSD Storage	64TB SSD Storage
2TB Bulk Storage	4TB Bulk Storage	6TB Bulk Storage	8TB Bulk Storage

Up to Four 3U Wave Computers Per a Single Data Center Node

Wave's Early Access Program

Qualified third party developers and customers can participate in Wave's Early Access Program. Please contact *info@wavecomp.com* for more information.

About Wave Computing

Wave Computing is a VC-backed startup that is revolutionizing the machine learning industry. After years of dataflow architecture, hardware, software and tools development, Wave's world-class team has developed its patented, native dataflow solution that outperforms any other machine learning training product available today. Based in Campbell, California, Wave is providing its solutions to customers globally.

www.wavecomp.ai