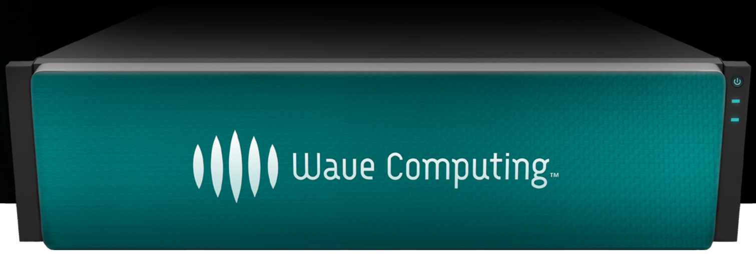
Wave Computing™



Introducing the World's Fastest Dataflow Computer for Machine Learning

Delivering Breakthrough Performance & Scalability for Machine Learning

Finally, data scientists and researchers have the solution they've been asking for: Wave's machine learning compute appliance that is ideal for training and inferencing of deep and shallow neural networks.

Wave's appliances are based on a revolutionary dataflow architecture that improves machine learning performance up to 1000x compared to CPUs, GPUs and FPGAs, giving the data scientist faster results and improved scaling in the data center. Each Wave machine learning system comes in a 3U form factor that easily fits into existing data center environments, as well as the needed software, programming tools and dataflow agent libraries to get up and running quickly.

A Future-Proof Machine Learning Solution

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

Specifications for each Wave Compute Appliance

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

www.wavecomp.ai

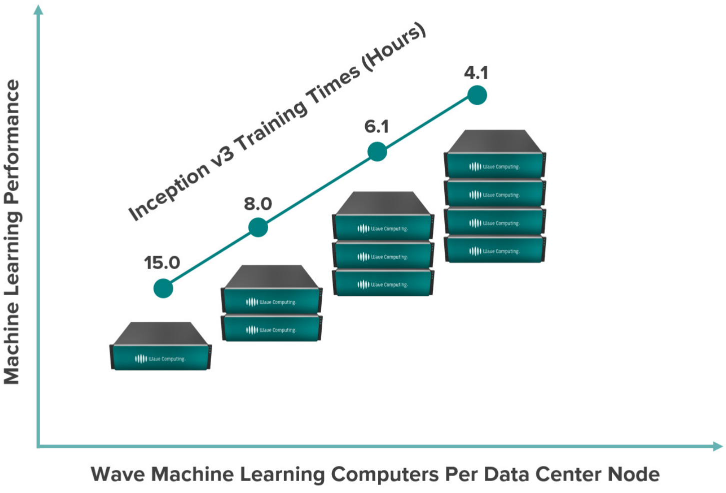
Wave's Revolutionary Dataflow Architecture

Wave's native dataflow architecture is the fundamental technology behind each of Wave's machine learning appliances. It is built upon a revolutionary dataflow computing technology that eliminates the traditional CPU/GPU co-processor structure and associated performance and memory bottlenecks. 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 in as little as four hours by Wave's dataflow solution, which can scale to as many as four compute appliances in a single data center node. This represents a significant performance and scalability improvement for the training of machine learning networks over state-of-the-art CPU- and GPU-based systems.



Training a Shallow Recurrent Neural Network

Using a single DPU within a single Wave machine learning appliance, Wave's solution can deliver more than 600x faster training for Word2Vec compared to current hardware acceleration systems.

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 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 four appliances 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, reducing network congestion and enabling a single data center rack to accommodate a 12U machine learning system configuration.

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

Wave's EAP gives qualified data scientists and developers Cloud-based access to a Wave compute appliance prototype before official sales begin. Availability is limited, so apply today. Please contact *info@wavecomp.com*.

About Wave Computing

Wave Computing is a VC-backed startup that is revolutionizing the machine learning industry. Wave's world-class team has developed and patented a native dataflow solution that outperforms any other machine learning product available today for the data center. Based in Campbell, California, Wave is providing its solutions to customers globally.

www.wavecomp.ai