



# Project Supernova

Common ML Inference Service from Edge to Cloud

Tiejun Chen, VMware OCTO

December 18, 2019

# Project Supernova

## Problem

With machine learning is widely used in enterprises, big data are trained in the cloud, inference services go to production either in the cloud or on the edge.

- On the edge
  - Edge devices have limited resources, space and power supply
  - Edge servers cost much higher than devices
  - Hardware accelerators are heterogeneous in architecture and various on interfaces and performance
- In the cloud
  - Accelerator market is dominated by Nvidia GPU
  - Other options come as AMD GPU, Intel Nervada NNP-I, AWS Inferentia, Xilinx FPGA etc.
- Common inference interfaces from cloud to edge doesn't appear generally
- Limitation on specific hardware accelerators or cloud leads to new vendor lock-in

# Project Supernova

## Why Now?

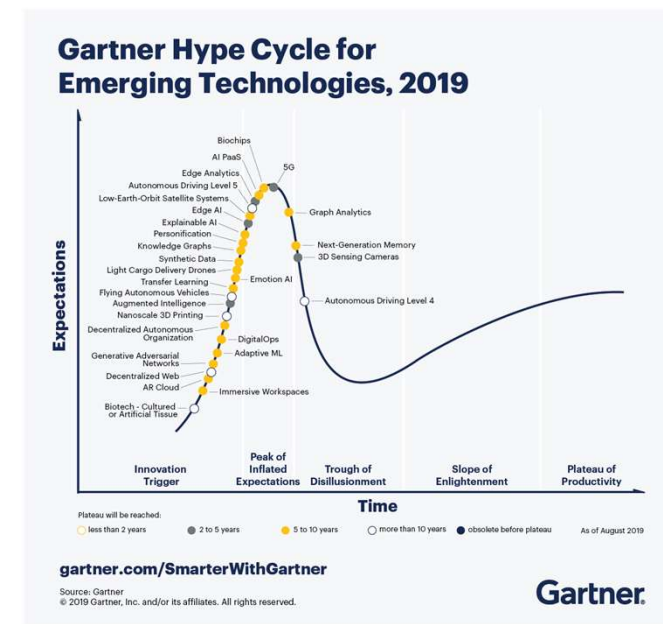
ML training accelerators are well used in data centers and clouds, while we see a set of different accelerators for inferences appear on both edge and cloud, to save cost and improve productivity.

### ➤ In the cloud

- AWS announced Greengrass ML interface, SageMaker Neo and Inferentia chip.
- IBM announced Edge analytics with Watson IoT Platform.
- Intel announced Nervana NNP-I and nGraph.
- Nvidia announced TensorRT, with Tesla T4.

### ➤ On the edge

- Google announced IoT Edge, TensorFlow Lite and Edge TPU.
- Microsoft announced IoT Edge ML module for Azure.
- Intel announced Movidius VPU, OpenVINO Toolkit.
- Nvidia announced TensorRT, with embedded GPU in SoC.
- Xilinx announced AI Edge Platform on FPGA and DNNDK.
- Many startups build IP/ASIC for ML inferences.



# Project Supernova

## Mission

Project Supernova is to build a common machine learning inference service by enabling machine learning inference accelerators across edge endpoint devices, edge systems and cloud.

## Project Supernova

### Solution

Implement a common ML inference service from the edge to cloud, with or without hardware accelerators.

- Micro-service based architecture with Restful API
- Support heterogenous system architectures from leading vendors
- Support accelerator compilers
- Neutral to ML training framework file formats
- Work on both edge devices and clouds

# Project Supernova

## Objectives

Build a common ML inference service from edge to cloud, with or without hardware accelerators.

### ➤ Hardware CPU support







- x86-64, ARM64
- Hardware accelerator support
  - P0: Nvidia GPU, Intel IPU/VPU, Google (Edge) TPU
  - P1: Xilinx FGPA, AMD GPU
  - P2: Intel FPGA

### ➤ Software

- Inference toolkit support: OpenVINO, nGraph, TensorRT, Tensorflow Lite, DNNDK, RoCm
- Training framework data format: Tensorflow, Caffe, ONNX, MxNet











# Project Supernova

## Targeted Accelerators

Priority	Interface (PCIe, M.2)	Accelerator	Vendor	Toolkit	Framework
P0		GPU	Nvidia	CUDA, TensorRT	Caffe, Tensorflow, MxNet, Pytorch, ...
P0		IPU, NNP-I, FPGA	Intel	nGraph	Caffe, Tensorflow, MxNet, ONNX, Kaldi, Pytorch
P1		FPGA	Xilinx	DNNDK, OpenCL	Caffe, Tensorflow, MxNet
P1		GPU	AMD	MIOpen, ROCm	Caffe, Torch, Tensorflow ...
P2		TPU	Google	Tensorflow Lite	Tensorflow
P2		Inferentia	AWS	TVM	Tensorflow, MxNet, PyTorch, ONNX

From Cloud

To Edge

Priority	Interface (mPCIe, MXM, PCIe, USB)	Accelerator	Vendor	Toolkit	Framework
P0	 	GPU	Nvidia	CUDA, TensorRT	Caffe, Tensorflow, MxNet, Pytorch, ...
P0	 	IPU, VPU, FPGA	Intel	OpenVINO, nGraph	Caffe, Tensorflow, MxNet, ONNX, Kaldi, Pytorch
P1	 	FPGA	Xilinx	DNNDK, OpenCL	Caffe, Tensorflow, MxNet
P1	 	GPU	AMD	MIOpen, ROCm	Caffe, Torch, Tensorflow ...
P2	 	Edge TPU	Google	Tensorflow Lite	Tensorflow

## Q & A



Tiejun Chen

Twitter: @Tiejun\_Chen

Linkedin: <https://www.linkedin.com/in/tiejun-chen-41a6654b/>

tiejunc@vmware.com







Thank You