



Supernova项目概述

边缘设备上的通用ML推理加速方案

路广

VMware 研发总监

April 12, 2019

Agenda

使命与专注

行业趋势

典型使用案例

用户痛点

竞争技术比较

总体方案

技术实现

使用场景

演示

未来计划

VMware边缘计算与智能试验室

使命与专注

VMware边缘计算与智能试验室隶属于VMware CTO办公室，在2017年创设，为了探索边缘计算与工业物联网领域内潜在的颠覆性技术挑战和商业机会，为现有或新产品和方案增加或构建独特价值，帮助企业客户成功实现数字化转型。

探索方向：

- 水平方向：基于开放生态系统的多层通用平台；
- 垂直方向：数据中心管理的特定行业。

焦点

- 非现有产品，而是邻接领域
- 例如：虚拟化边缘设备、边原生应用、边缘侧机器学习推理、机器人应用、智慧数据中心等。

“物理” 的软件试验室

边缘设备



加速器件 (VPU/GPU/FPGA)



机器人与激光雷达



行业趋势

中心式



大型机



小型机



云



分布式

客户服务器



互联网



物联网/边缘



典型使用案例

涌现而出的边缘应用和技术融合要求低延时与加速运算



用户痛点

边缘设备上的资源有限、空间狭小、功率低

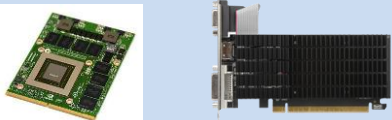
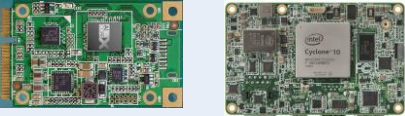



边缘服务器成本较高

各种加速器件API异构、性能也参差不齐

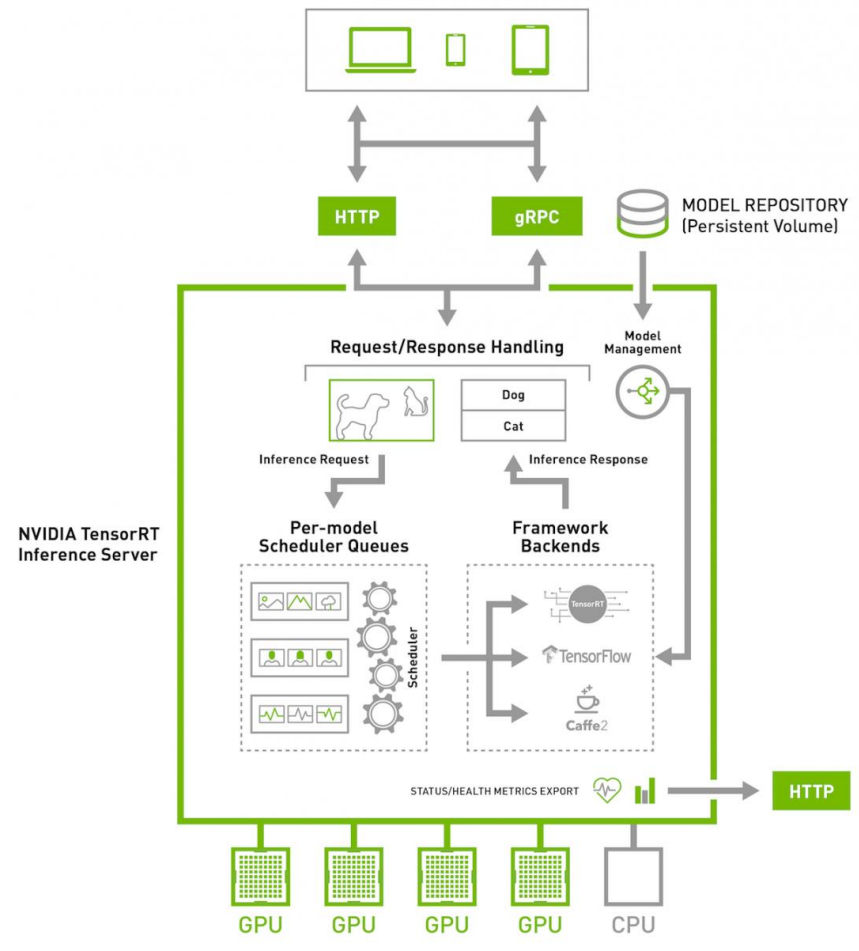
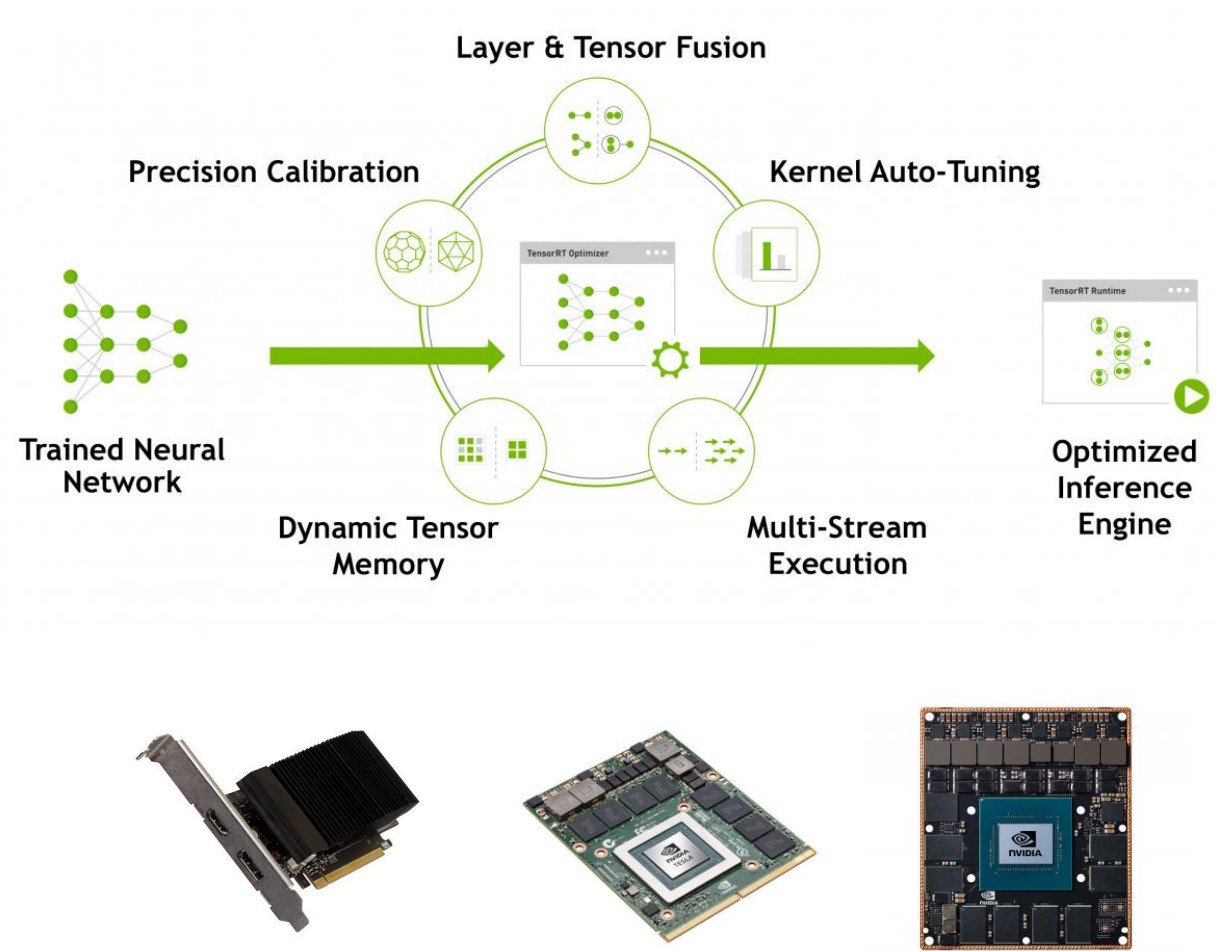
云-边协同的计算框架尚未普遍形成

若受限于特定加速器件/云，则会产生新的锁定

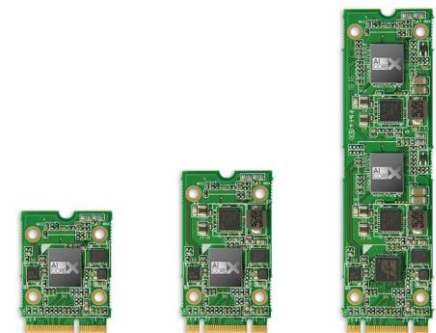
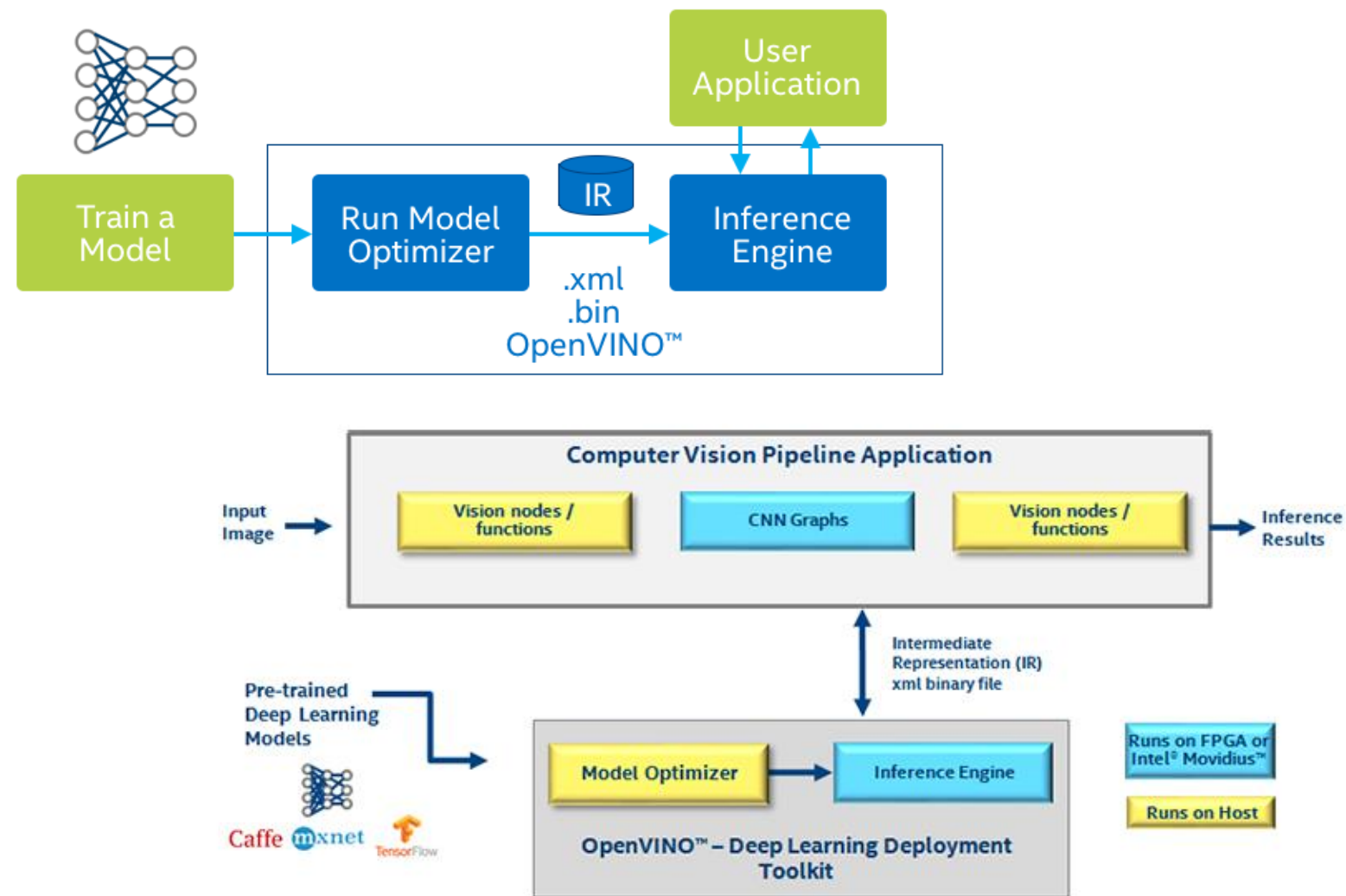
竞争技术比较

优先级	接口(mPCIe, MXM, PCIe, USB)	加速器件	厂商	工具包	机器学习框架
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

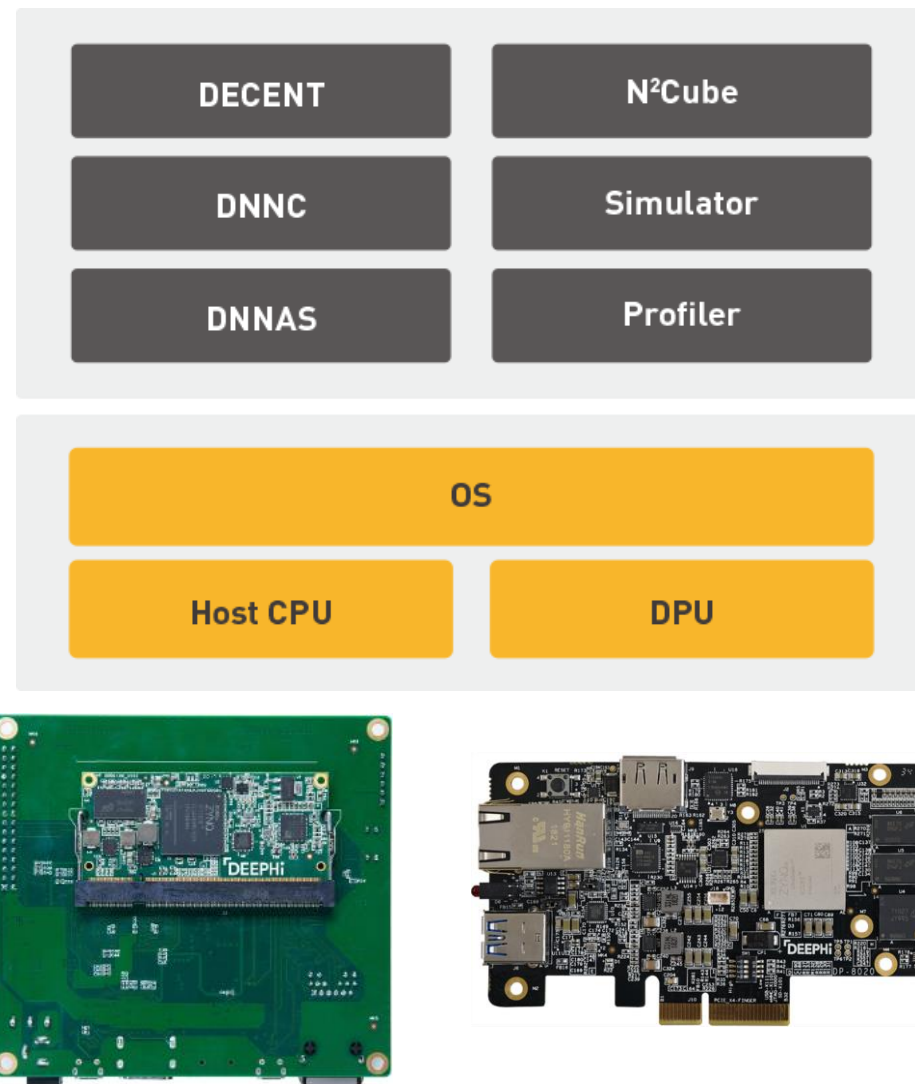
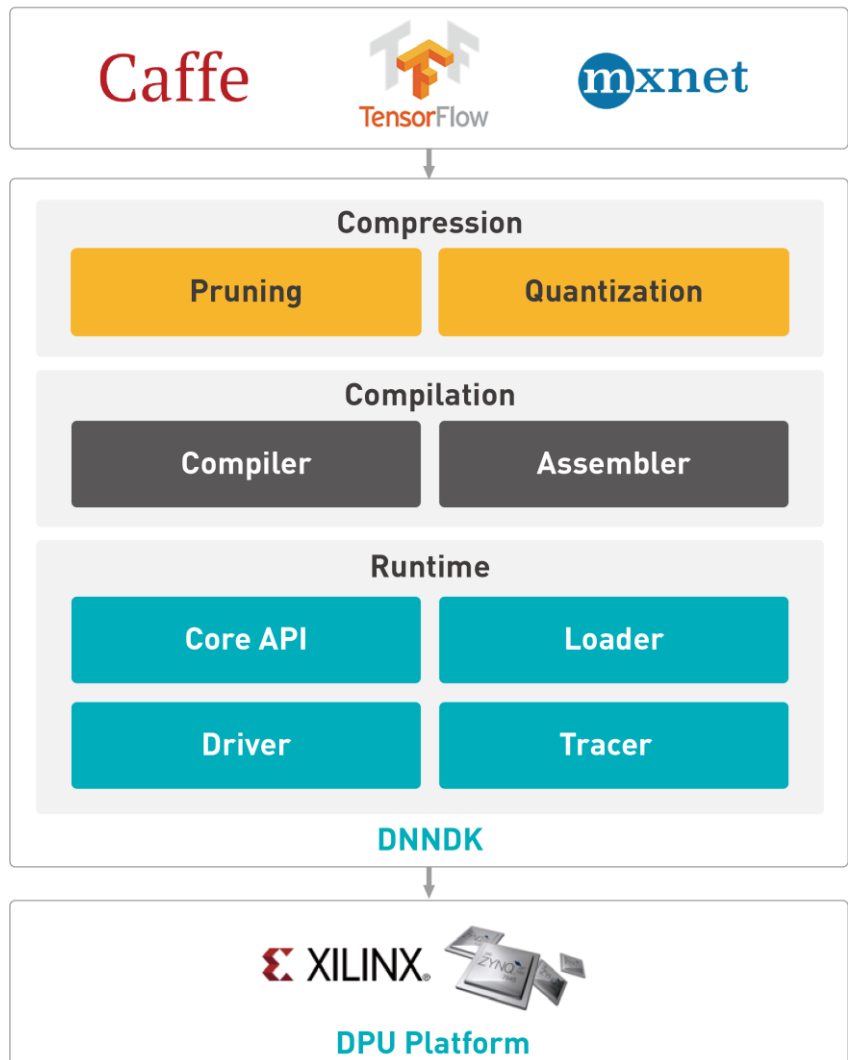
Nvidia TensorRT、嵌入式GPU与Jetson开发套件



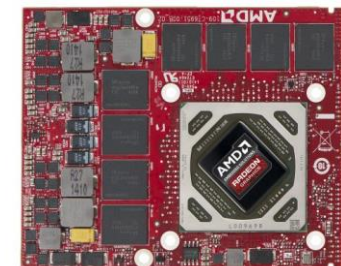
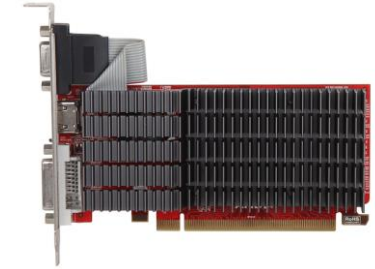
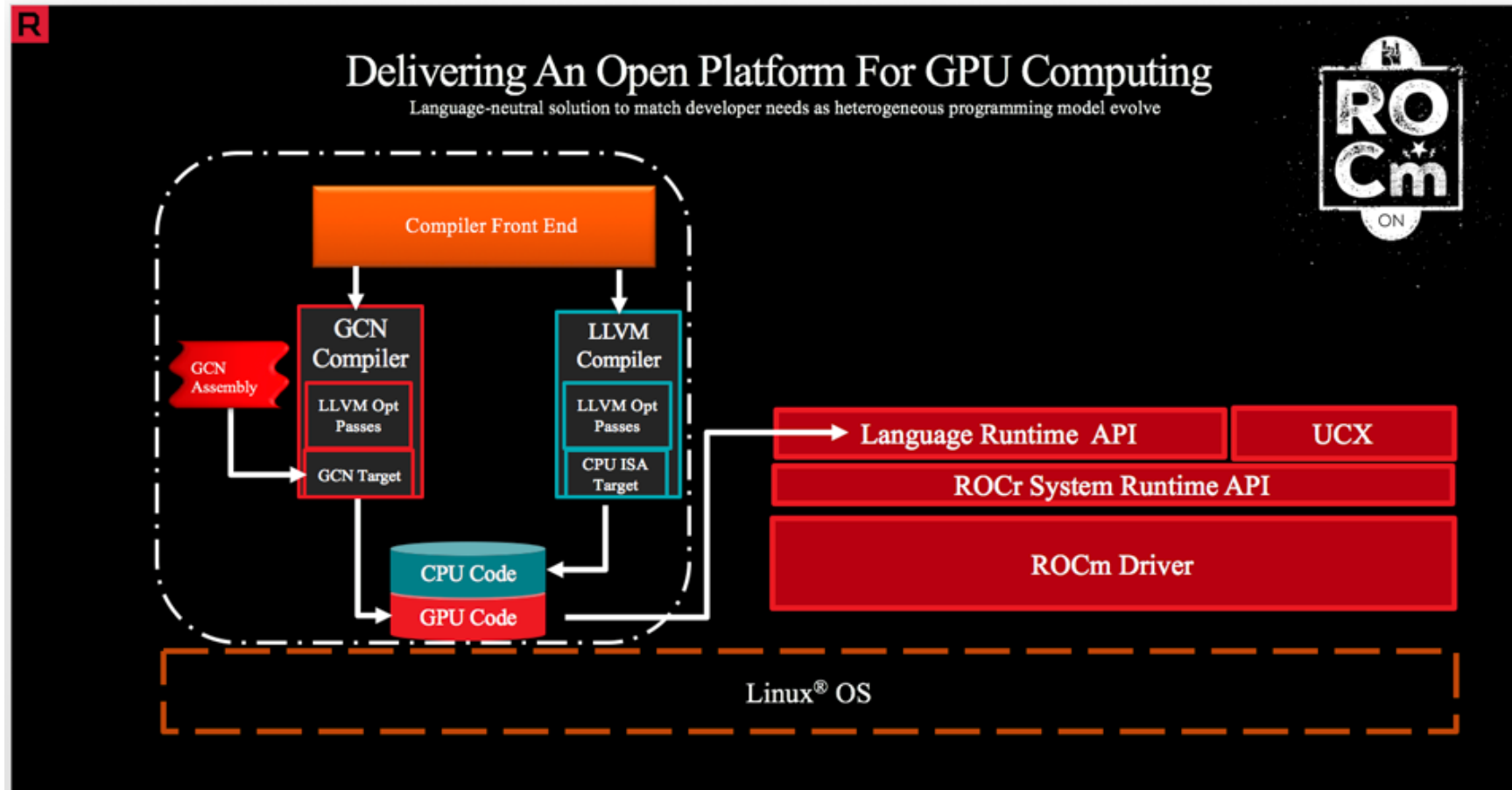
Intel OpenVINO工具包与VPU/FPGA



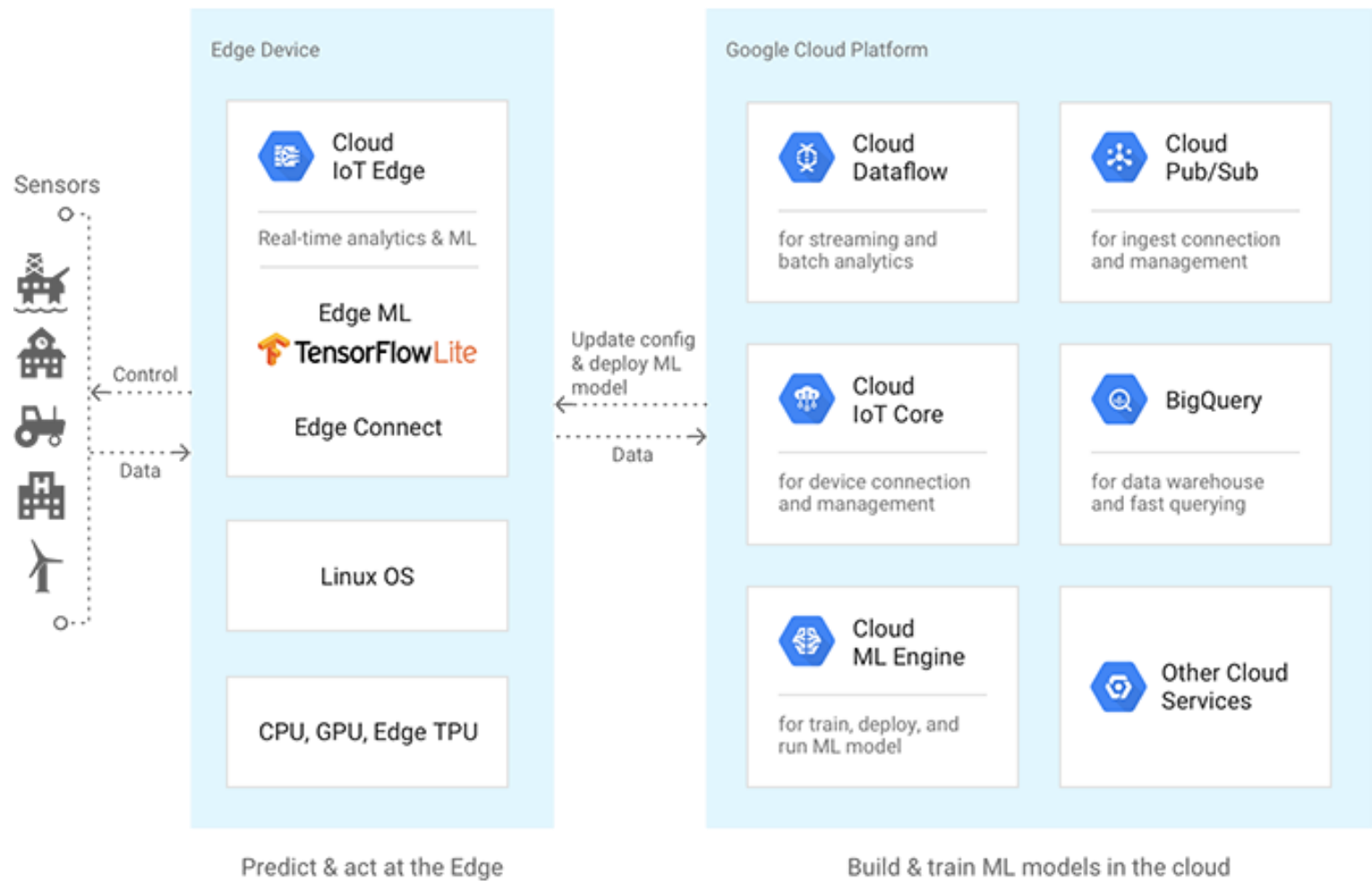
Xilinx的DNNDK与FPGA



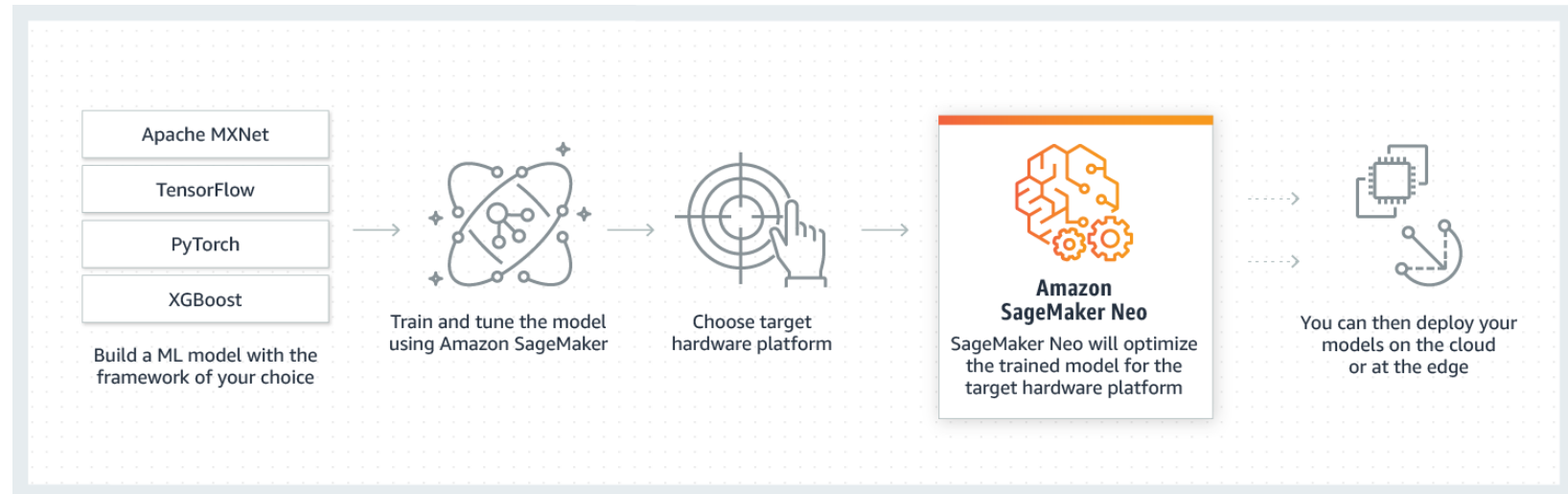
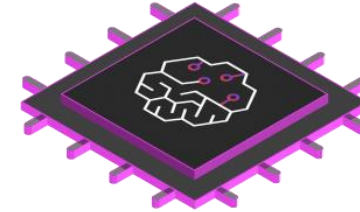
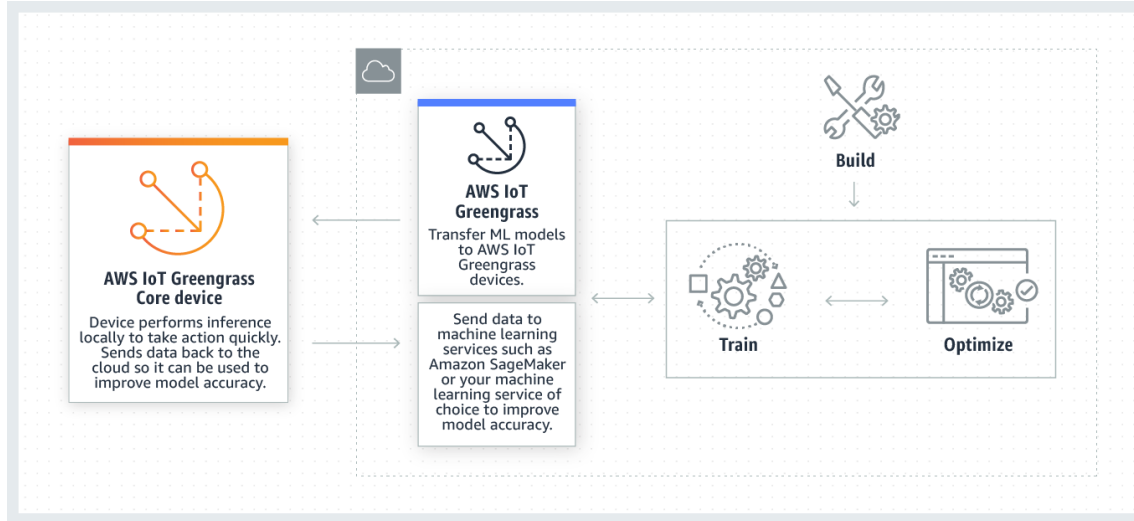
AMD的ROCm平台与嵌入式GPU



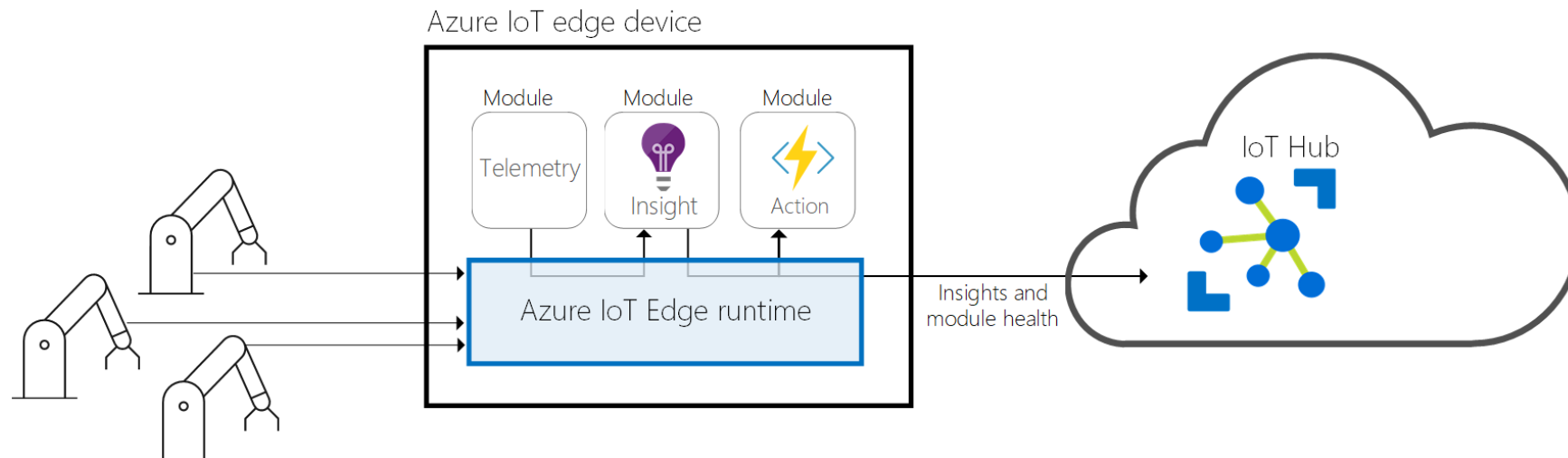
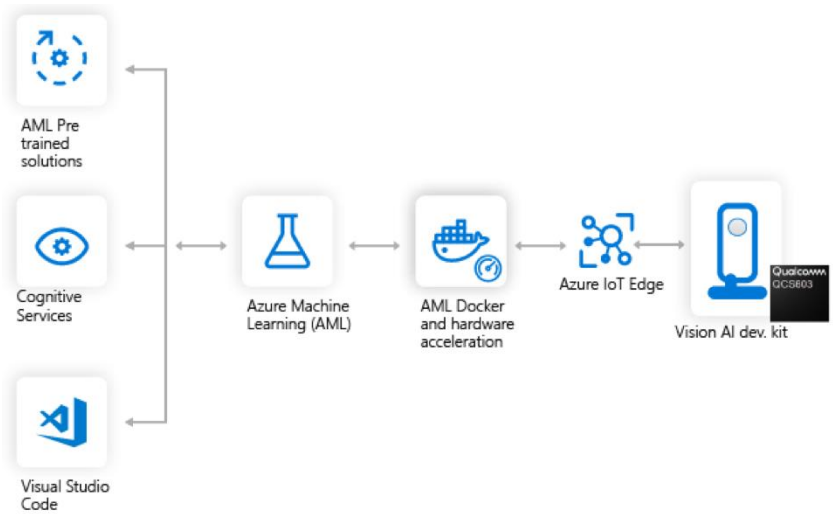
Google的TensorFlow Lite与Edge TPU



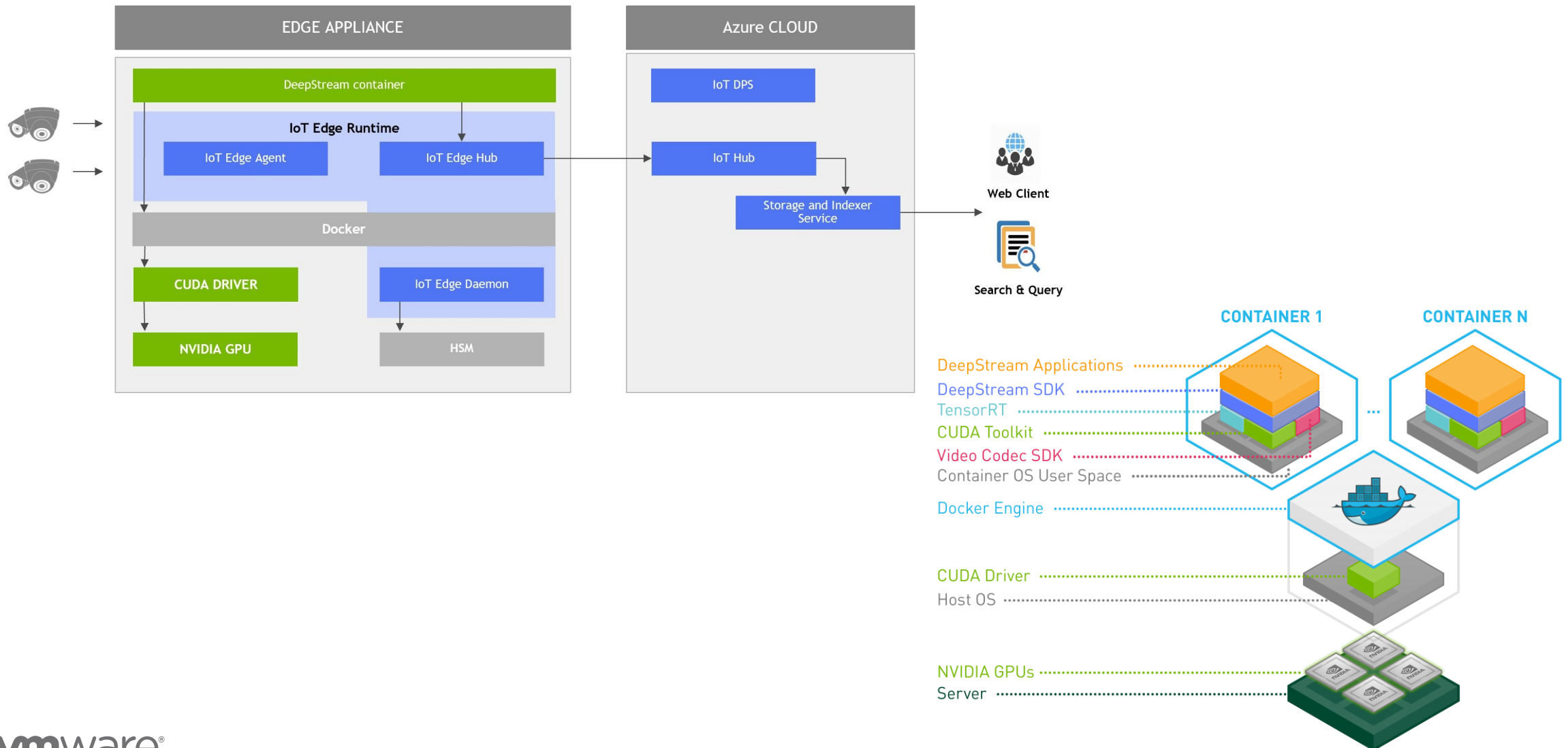
AWS IoT Greengrass、SageMaker Neo和Inferentia芯片



Azure IoT Edge机器学习与Vision AI开发套件

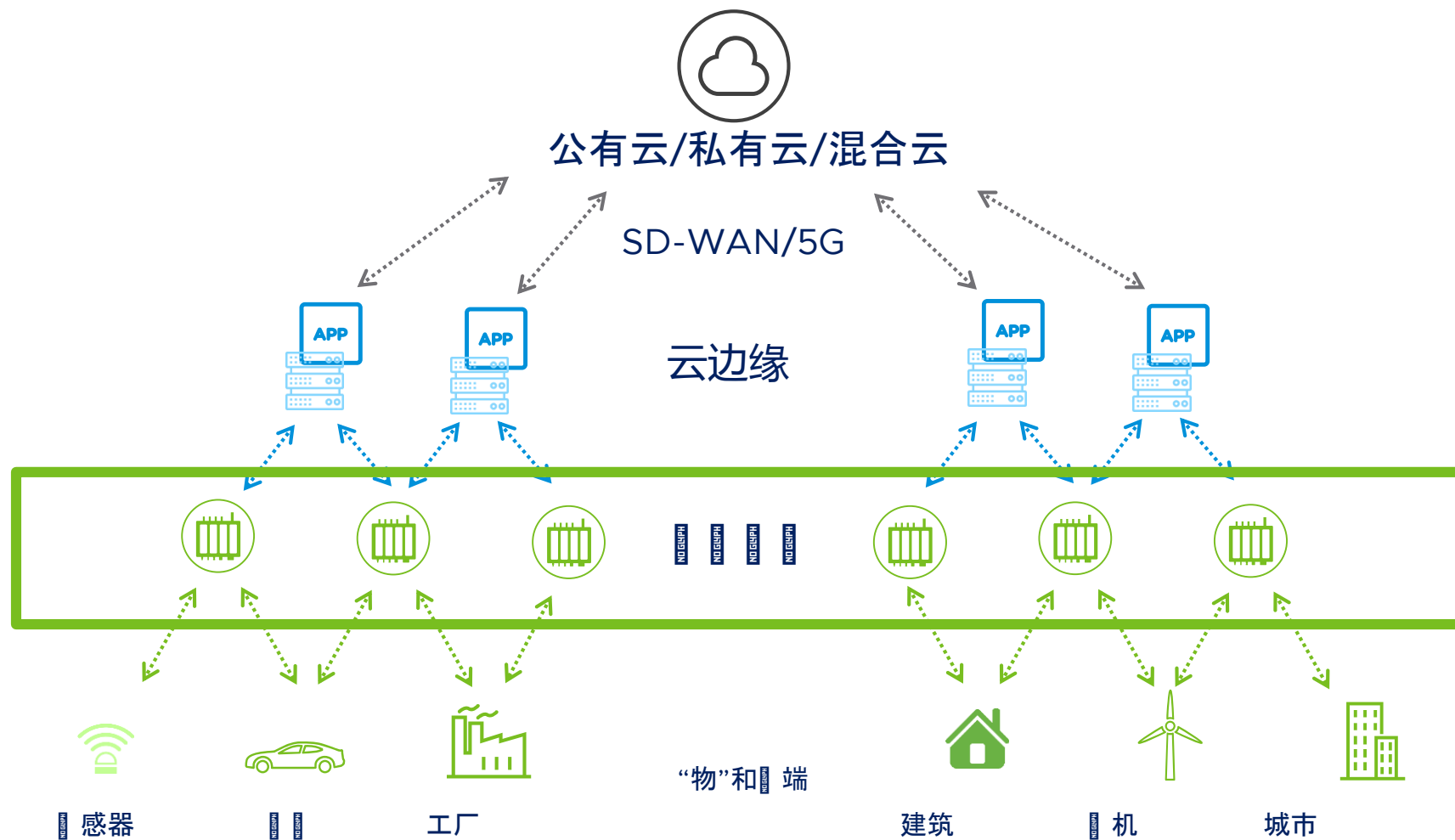


Azure IoT Edge与Nvidia深度流SDK



总体方案

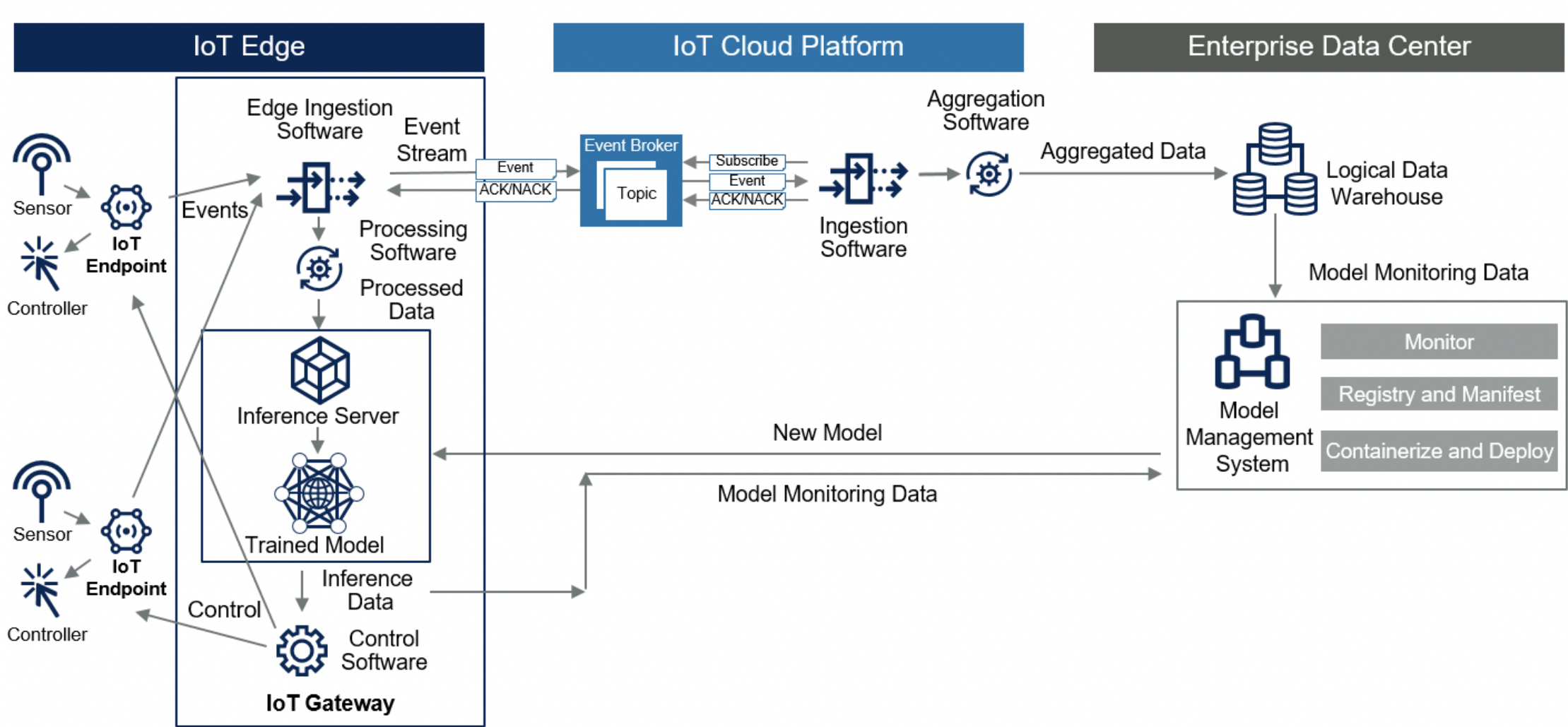
从数据中心迁移到“数据的中心”



深度学习

制定决策

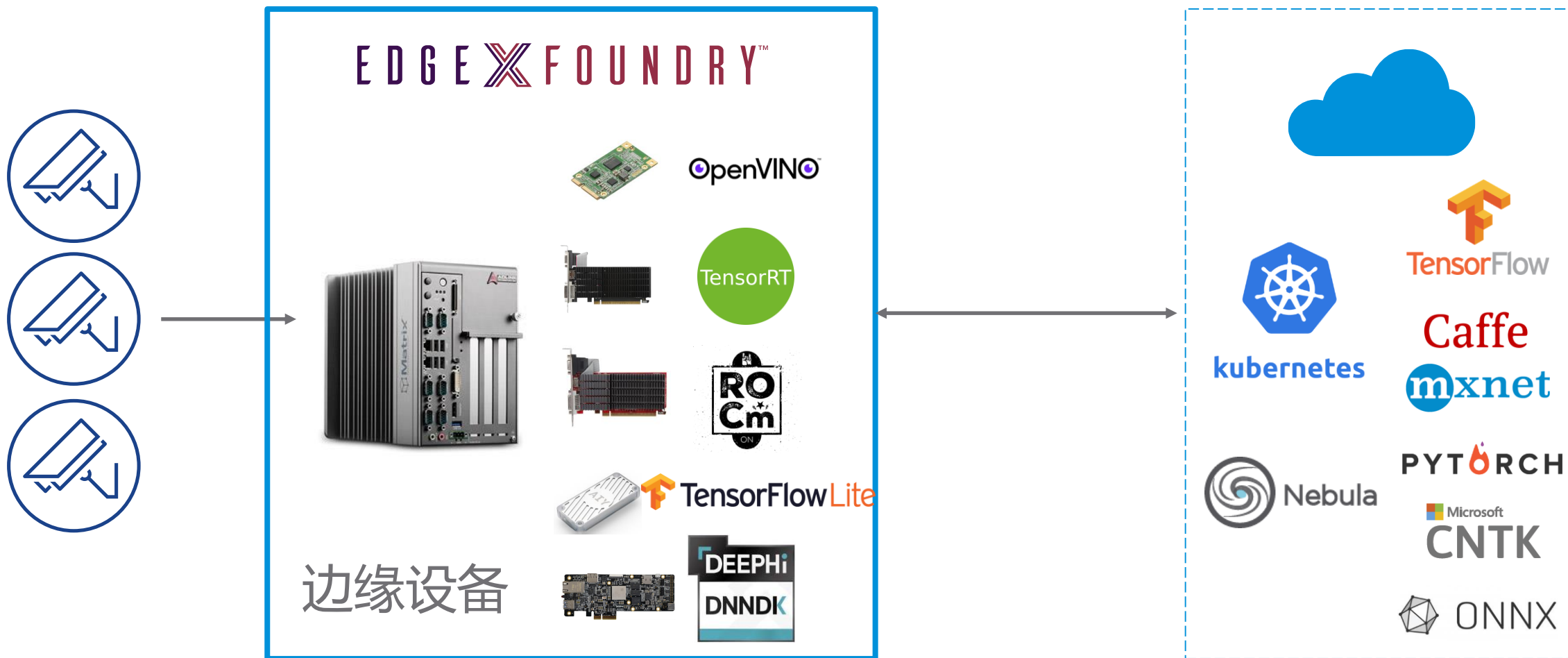
Gartner边缘设备上的ML推理参考架构



ID: 354956

© 2019 Gartner, Inc.

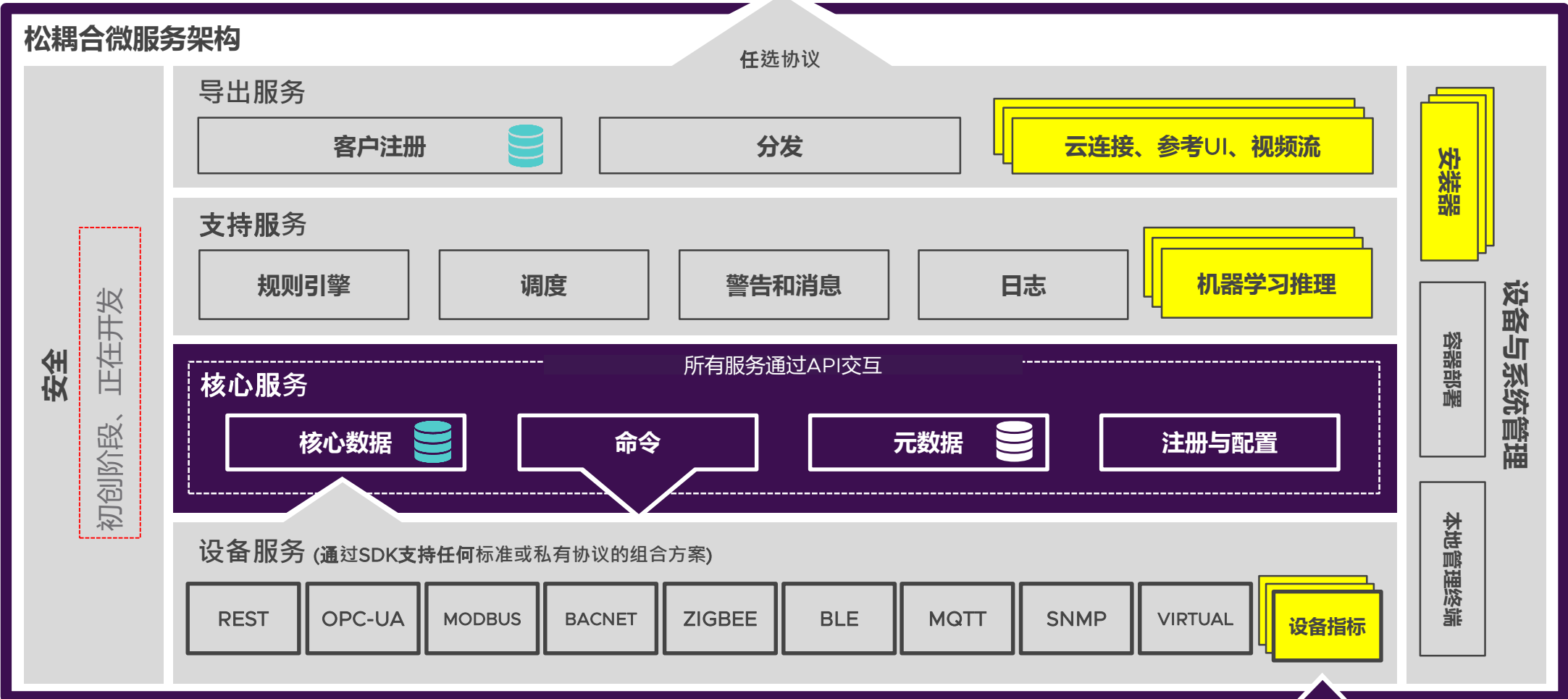
Supernova架构设计



技术实现



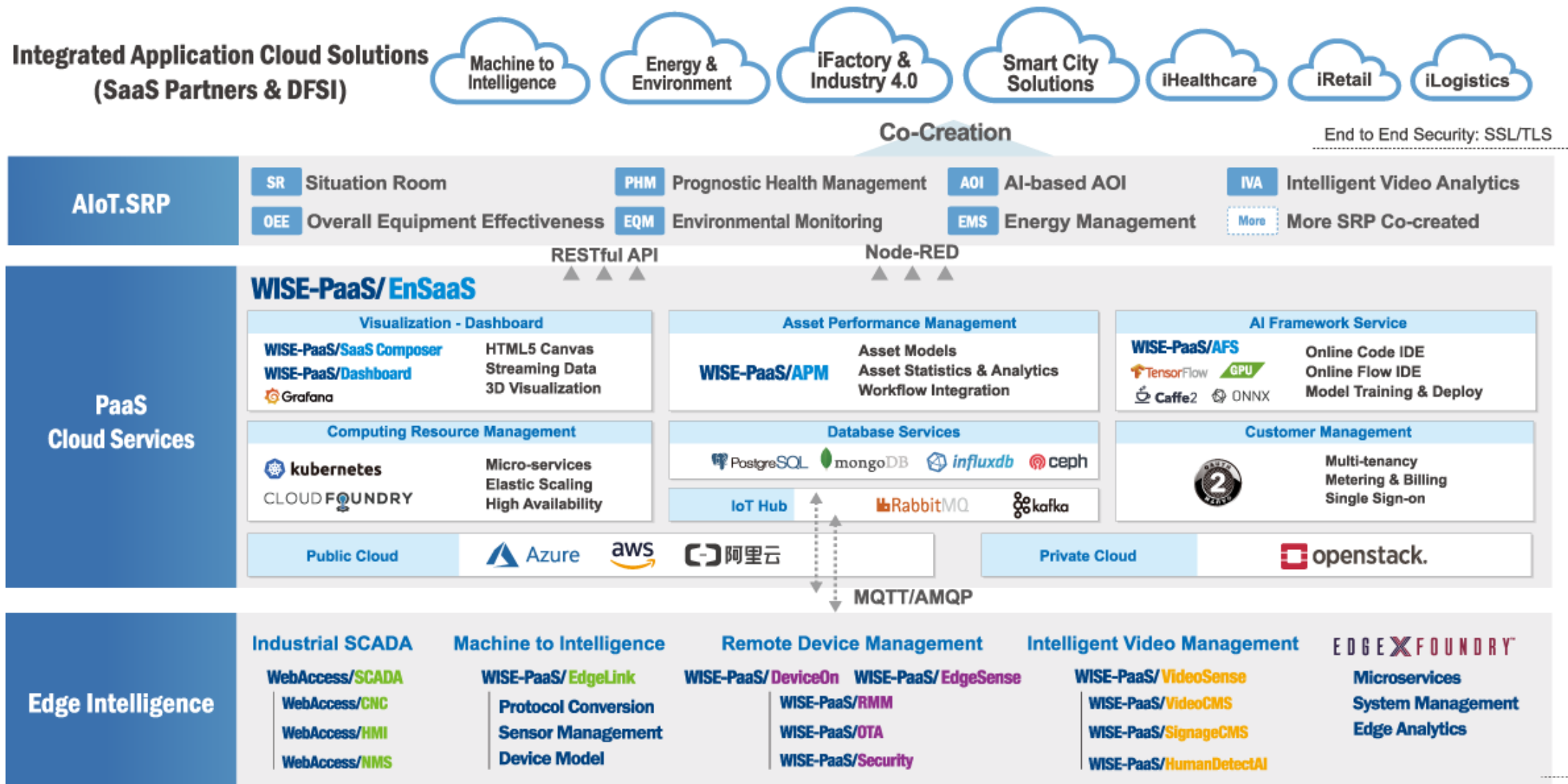
北向基础设施及应用



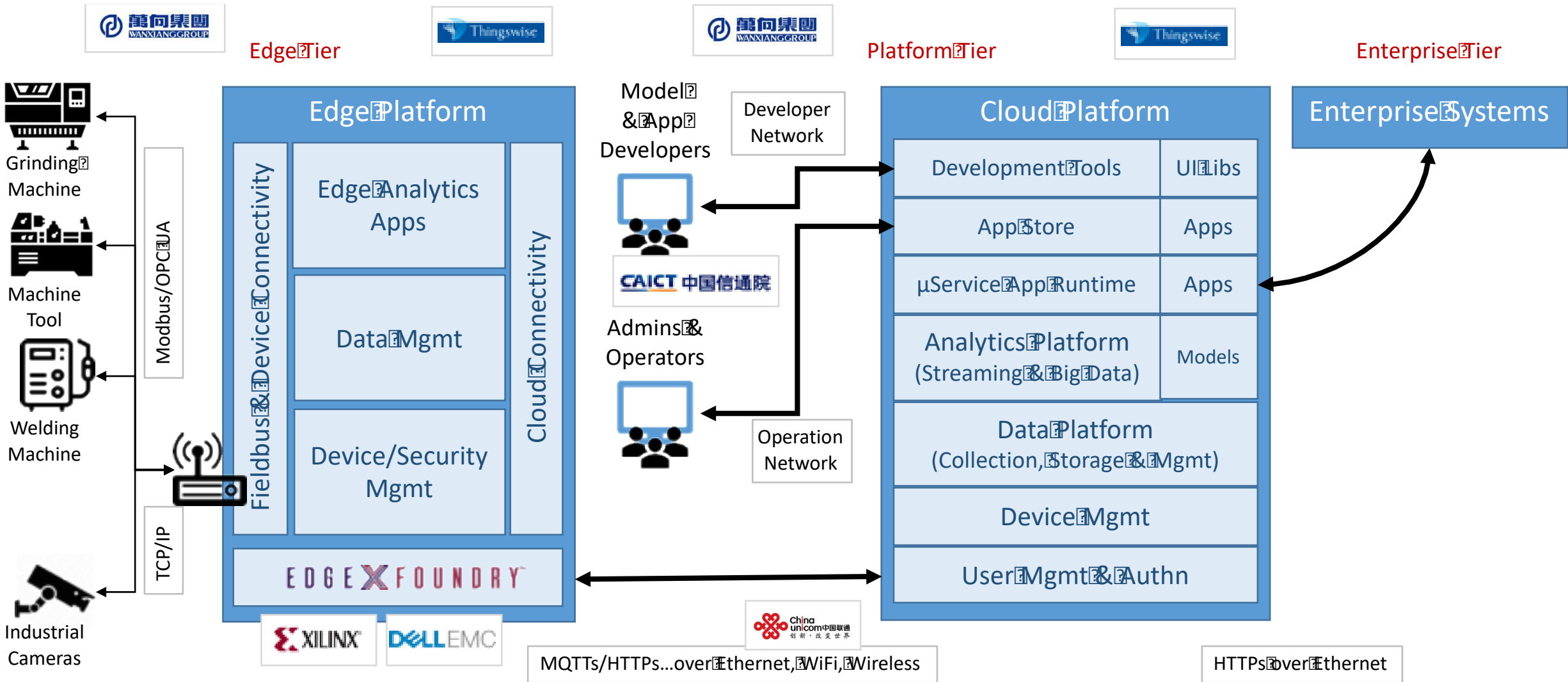
南向物、传感器和反应器

使用场景#1: 研华WISE-PaaS AIoT平台

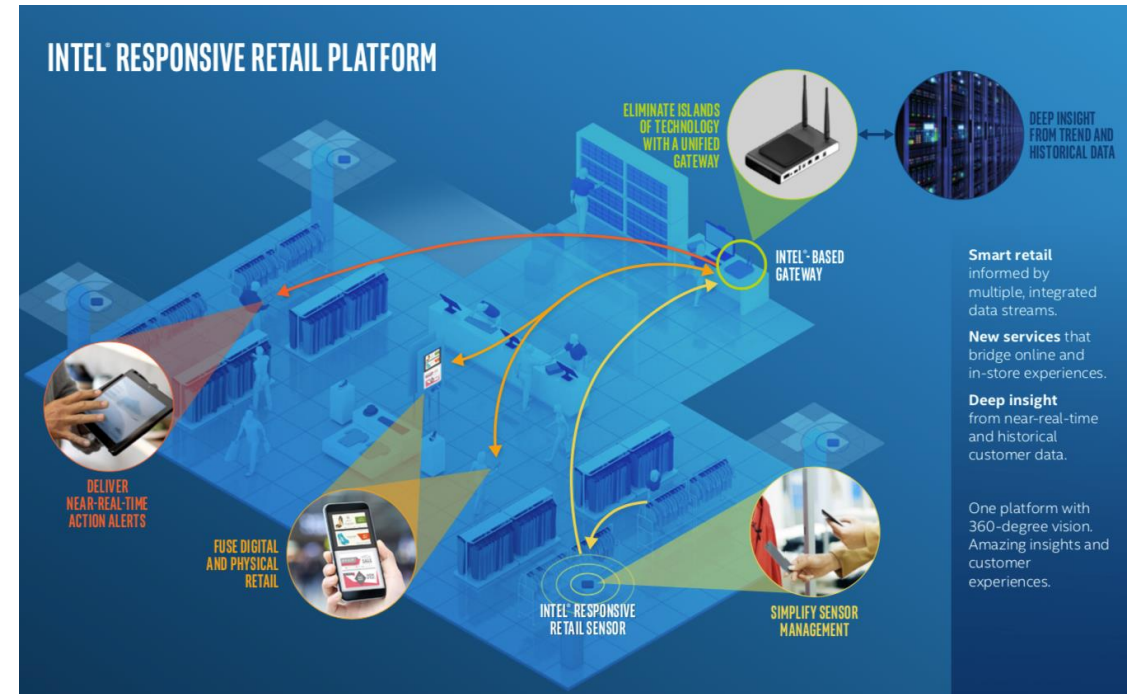
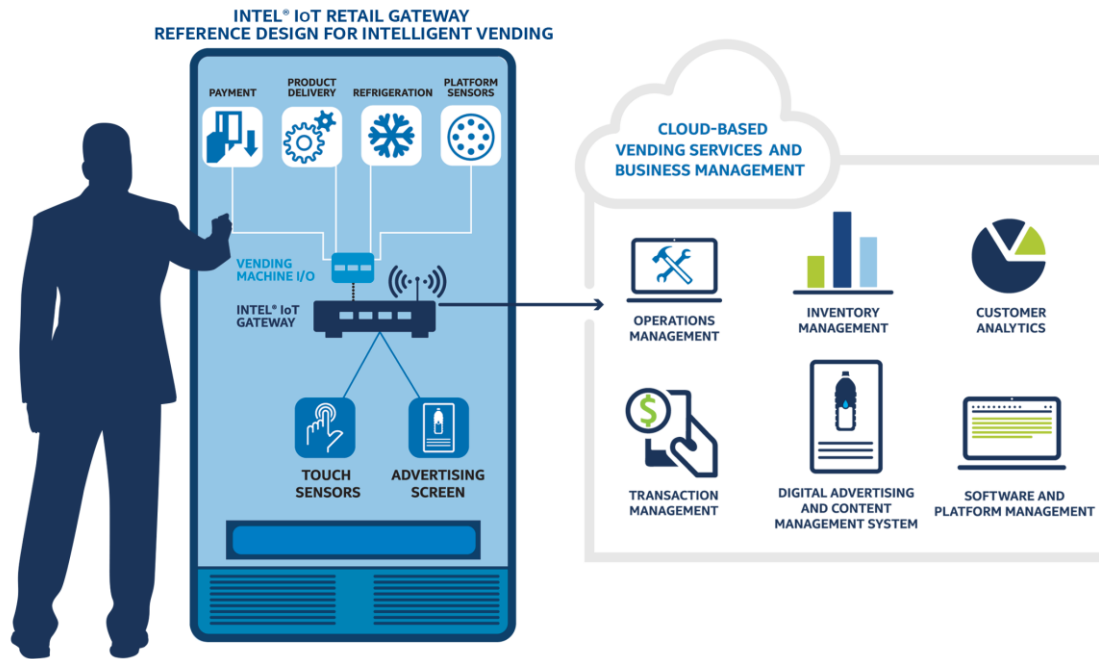
从边到云的架构



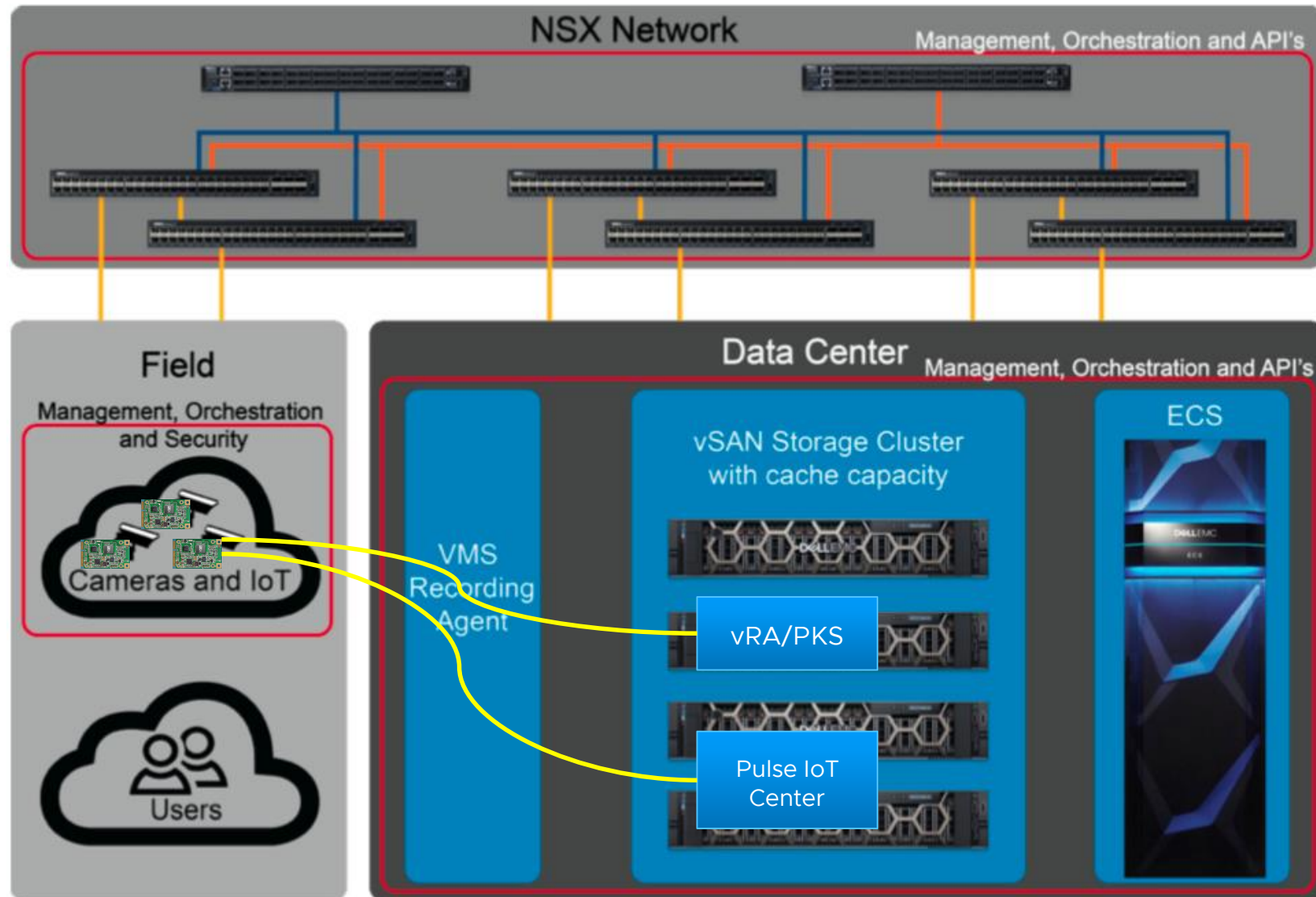
使用场景#2: 万向集团智慧制造试验床



使用场景#3: Intel智慧零售计划



使用场景#4: Dell智慧监控



演示

智慧监控：EdgeX Foundry上的通用ML推理加速服务，集成OpenVINO/VPU

未来计划

开放源码

- 机器学习推理服务，贡献到EdgeX Foundry社区

云-边协同

- 与Nebula项目集成，加强免费的EdgeX Foundry应用市场
- 与Kubernetes集成，建立云到端的ML CI/CD流程框架

Linux基金会边缘计划/EdgeX Foundry中国社区

欢迎你的加入!

正式成员



贡献者及用户





Thank You