

# 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) **ADVANTECH** Enabling an Intelligent Planet (intel (intel®) EUROTECH G R O U P Google SOFTIRON **E** XILINX® **AMD** LATTEPANDA 机器人与激光雷达

**ADLINK** 



**Y** ULIDAR

**PAXIOMTEK** 

## 行业趋势

中心式



大型机



小型机



굸



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

| NFV边缘基础设施 | ( <b>(</b> )) | 无线<br>AN,vEPC) |          | 有线<br>(PON) |       | uCPE<br>(SD-WAN) |     | IP 企业服务     |
|-----------|---------------|----------------|----------|-------------|-------|------------------|-----|-------------|
| 自主设备      |               | 无人机            |          | 自主车辆        | *     | 工业机器人            | ô   | 医药          |
| 沉浸式体验     |               | 虚拟现实           | <b>4</b> | 增强现实        |       | 全息影像             |     | 可穿戴<br>认知助理 |
| 物联网与分析    | <b>4.</b> ))  | 工业传感器          | <b>^</b> | 家庭设备        | \$    | 零售               | *** | 保健          |
| 随需<br>NFV | 2             | 硬件加速           | <b>©</b> | 人工智能        | ((•)) | 微服务              |     | 5G          |

#### 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

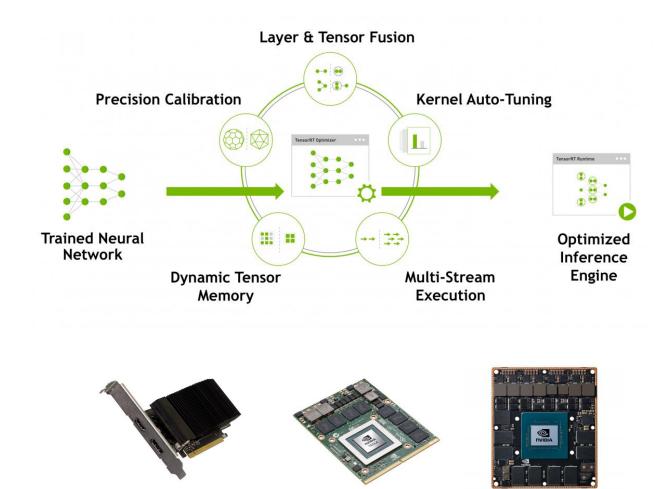


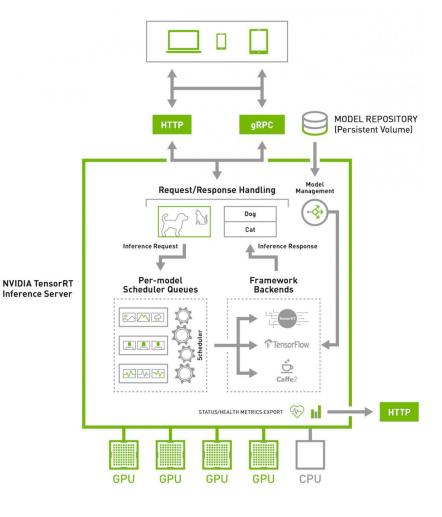
# 竞争技术比较

| 优先级 | 接口(mPCle, MXM,<br>PCle, USB) | 加速器件              | 厂商     | 工具包                 | 机器学习框架   |
|-----|------------------------------|-------------------|--------|---------------------|--|
| PO  |                              | GPU               | Nvidia | CUDA,<br>TensorRT   | Caffe, Tensorflow, MxNet, Pytorch,             |
| PO  |                              | IPU, VPU,<br>FPGA | Intel  | OpenVINO,<br>nGraph | Caffe, Tensorflow, MxNet, ONNX, Kaldi, Pytorch |
| P1  |                              | FPGA              | Xilinx | DNNDK,<br>OpenCL    | Caffe, Tensorflow, MxNet                       |
| P1  |                              | GPU               | AMD    | MIOpen,<br>ROCm     | Caffe, Torch, Tensorflow                       |
| P2  |                              | Edge TPU          | Google | Tensorflow<br>Lite  | Tensorflow                                     |



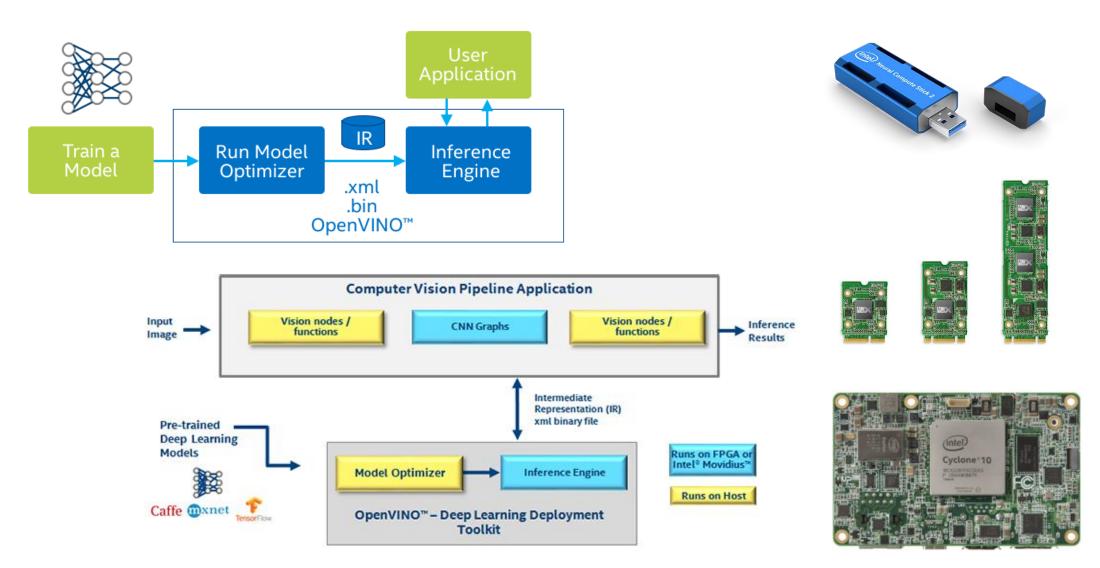
#### Nvidia TensorRT、嵌入式GPU与Jetson开发套件





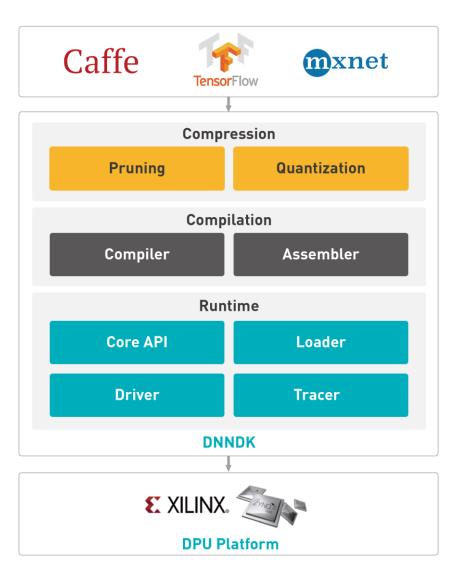


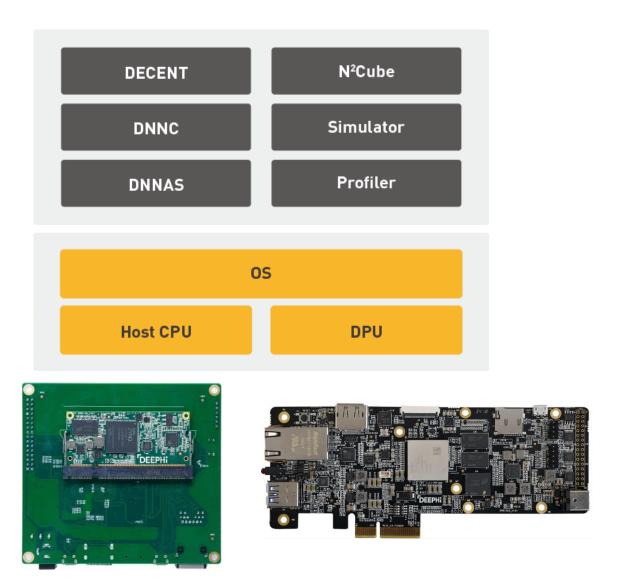
#### Intel OpenVINO工具包与VPU/FPGA



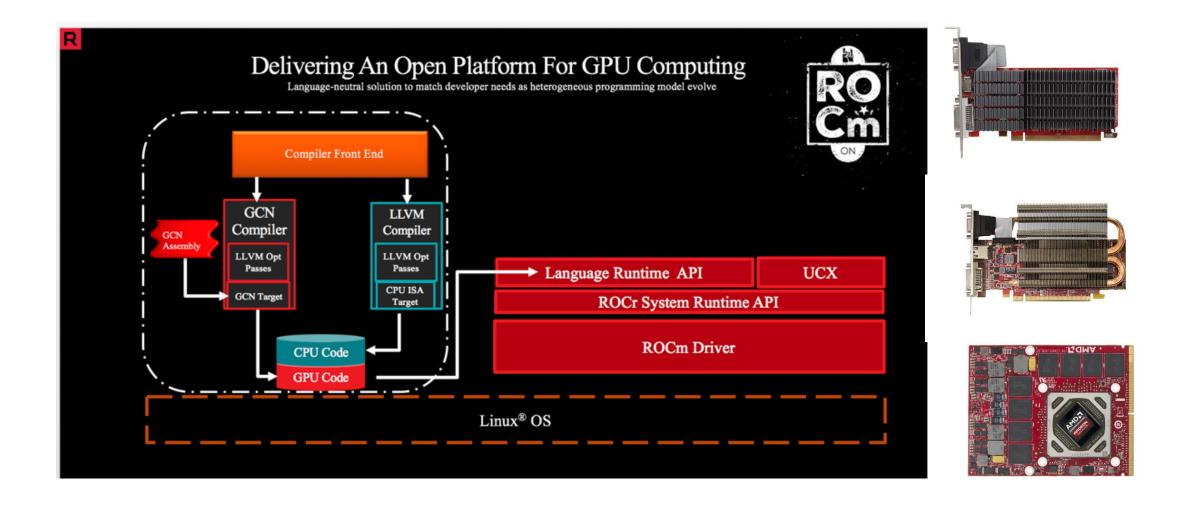


#### Xilinx的DNNDK与FPGA



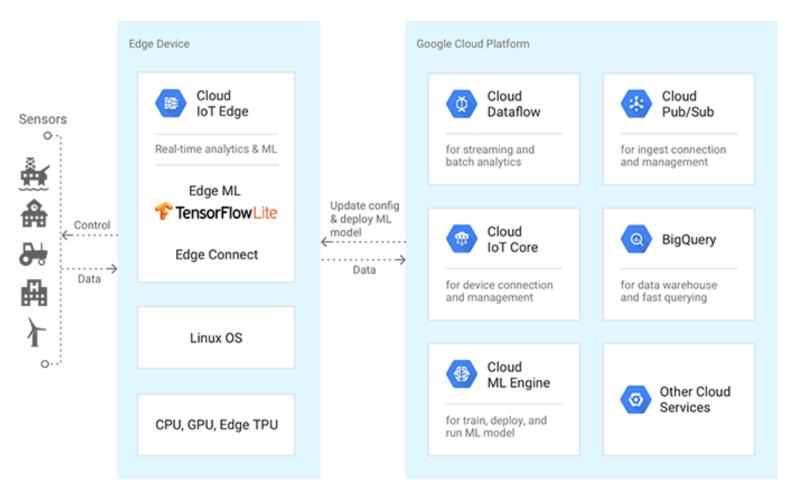


### AMD的ROCm平台与嵌入式GPU





#### Google的TensorFlow Lite与Edge TPU





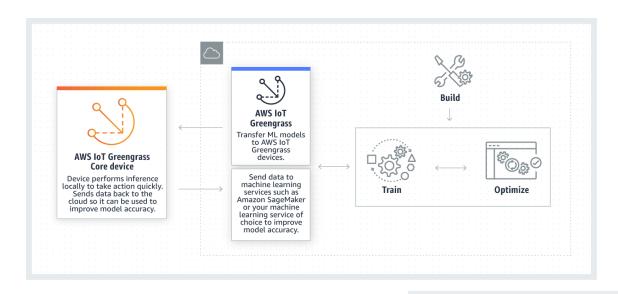


Predict & act at the Edge

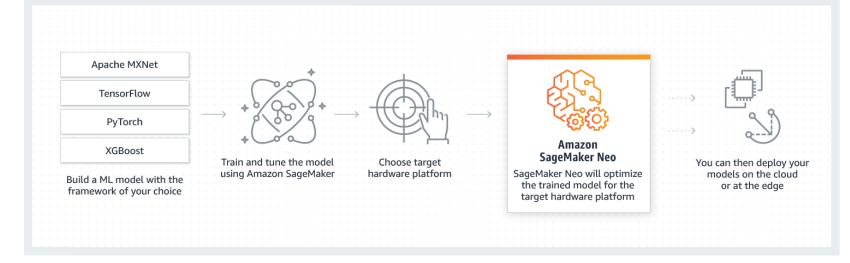
Build & train ML models in the cloud



#### AWS IoT Greengrass、SageMaker Neo和Inferentia芯片



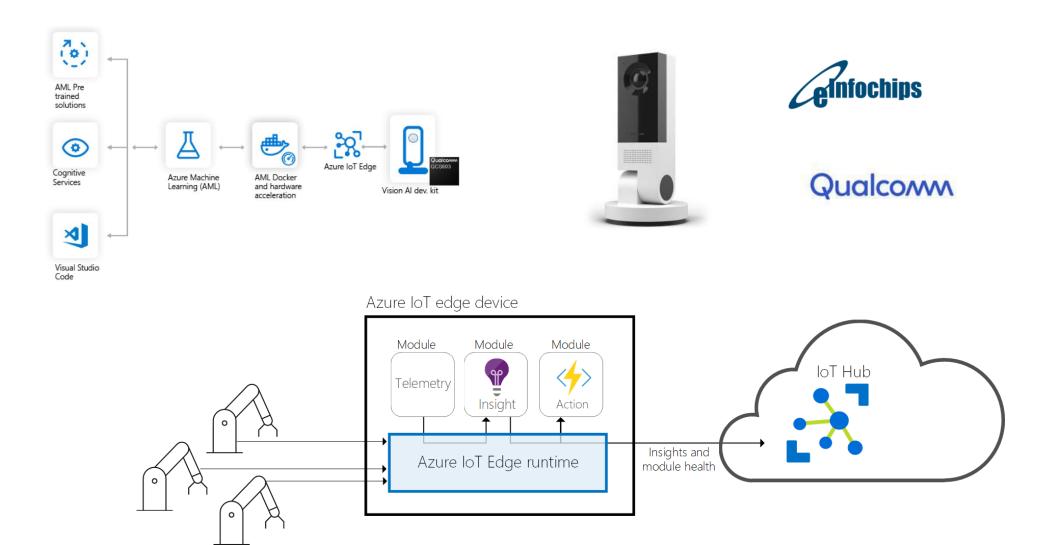






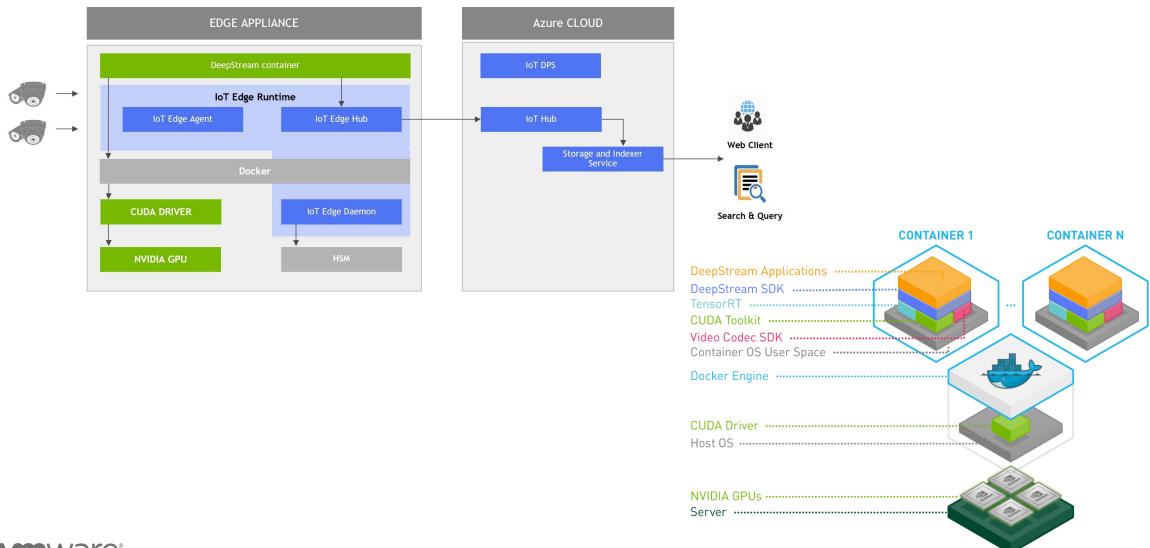
14

### Azure IoT Edge机器学习与Vision AI开发套件

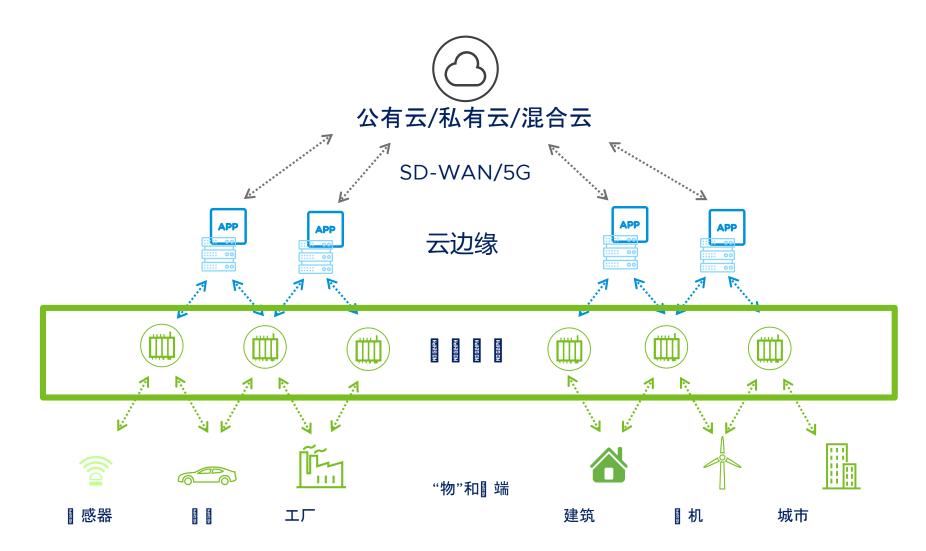




### Azure IoT Edge与Nvidia深度流SDK

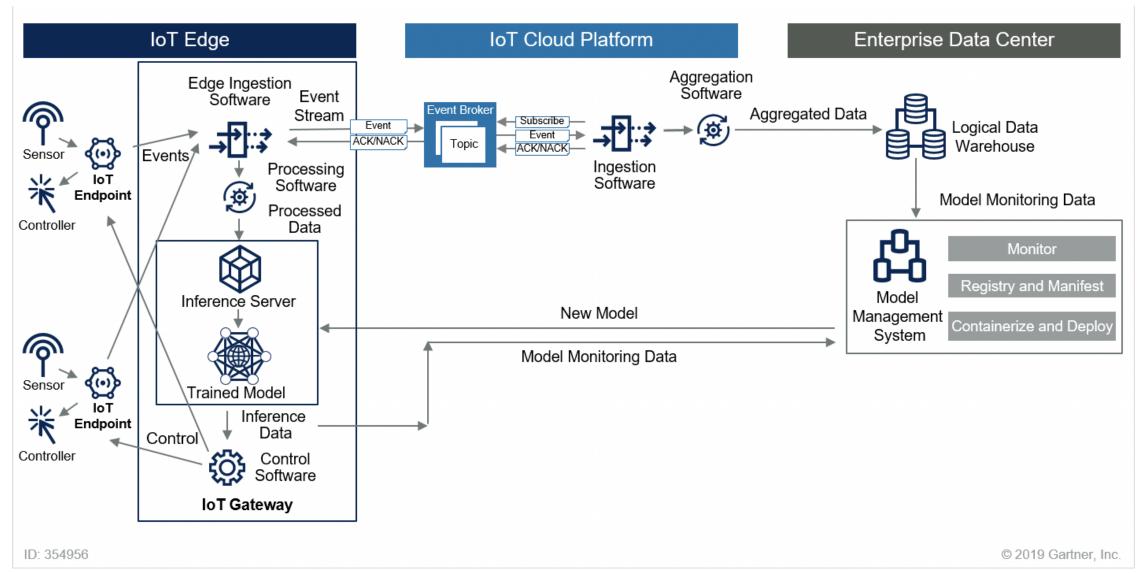


# **总体方案** 从数据中心迁移到"数据的中心"



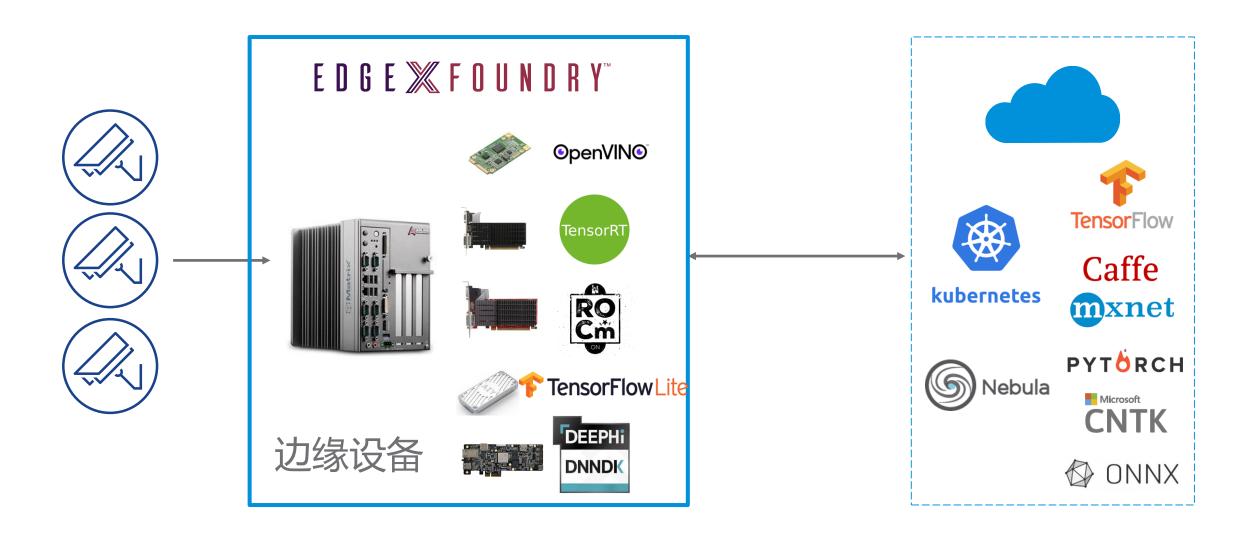


#### Gartner边缘设备上的ML推理参考架构





### Supernova架构设计

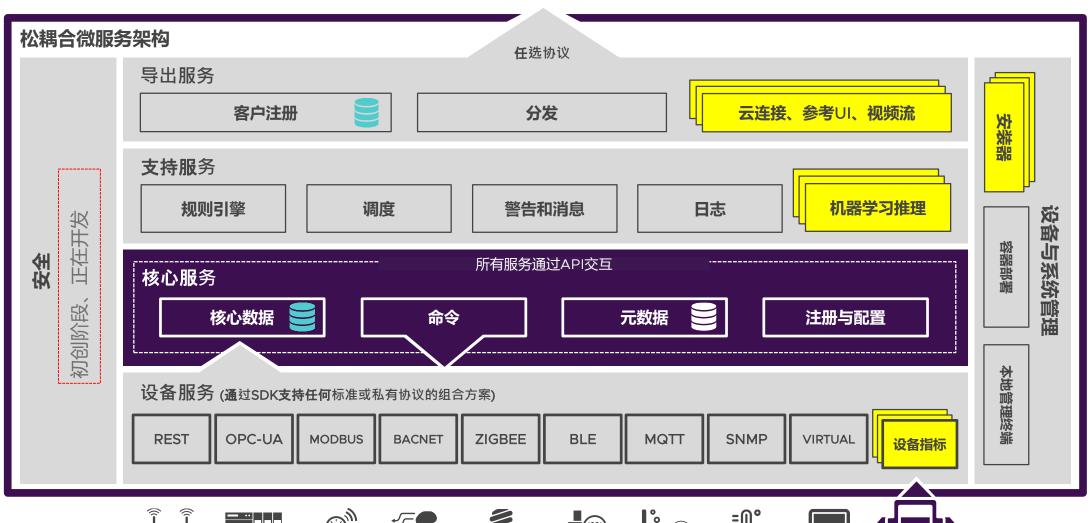




#### 技术实现



北向基础设施及应用





















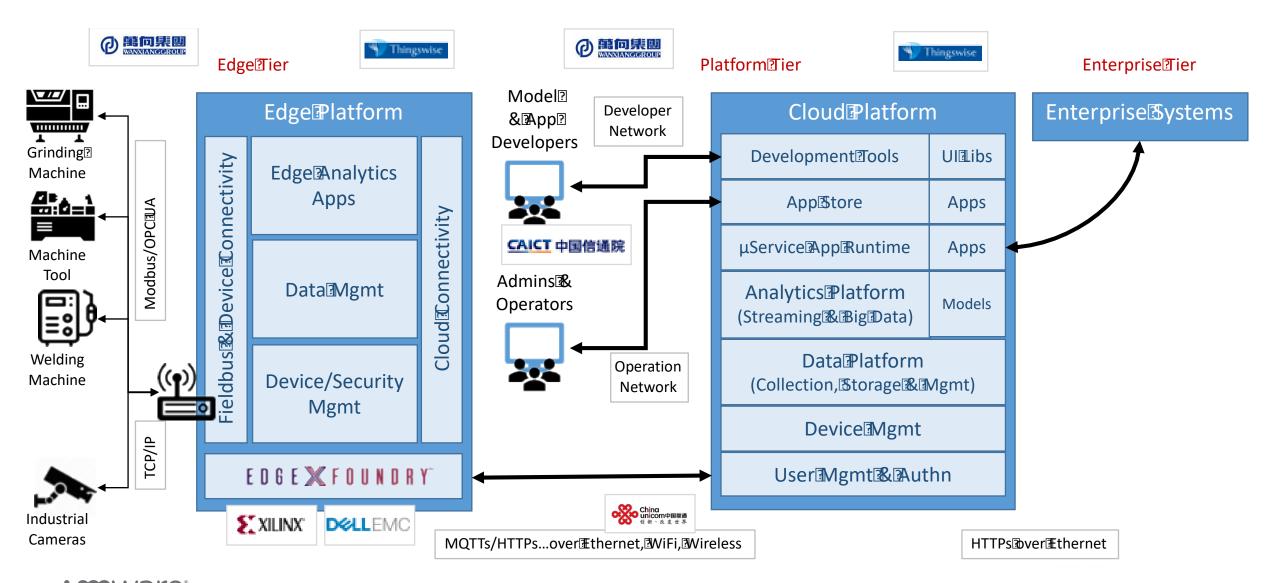


#### 使用场景#1: 研华WISE-PaaS AloT平台 从边到云的架构

**Integrated Application Cloud Solutions** iFactory & Machine to Smart City Energy & Intelligence Industry 4.0 iHealthcare iRetail iLogistics Solutions Environment (SaaS Partners & DFSI) Co-Creation End to End Security: SSL/TLS SR Situation Room Prognostic Health Management A01 Al-based AOI Intelligent Video Analytics AloT.SRP OEE Overall Equipment Effectiveness EQM Environmental Monitoring Energy Management More SRP Co-created Node-RED RESTful API A A A WISE-PaaS/EnSaaS Visualization - Dashboard **Asset Performance Management** Al Framework Service WISE-PaaS/AFS WISE-PaaS/SaaS Composer **HTML5 Canvas** Asset Models Online Code IDE Streaming Data **Asset Statistics & Analytics** TensorFlow GPU WISE-PaaS/APM Online Flow IDE WISE-PaaS/Dashboard 3D Visualization Workflow Integration Model Training & Deploy 6 Grafana Caffe2 ONNX PaaS **Computing Resource Management Database Services Customer Management** Cloud Services PostgreSQL mongoDB (a) influxdb (a) ceph Multi-tenancy Micro-services 0 kubernetes Metering & Billing Elastic Scaling CLOUD FOUNDRY Single Sign-on ‱kafka **High Availability** ■RabbitMQ IoT Hub aws (-) 阿里云 Azure openstack. **Public Cloud Private Cloud** MQTT/AMQP Industrial SCADA Machine to Intelligence **Remote Device Management** Intelligent Video Management EDGE X FOUNDRY WISE-PaaS/ DeviceOn WISE-PaaS/ EdgeSense WebAccess/SCADA WISE-PaaS/ EdgeLink WISE-PaaS/V Microservices **Edge Intelligence** WISE-PaaS/RMM System Management WebAccess/CNC WISE-PaaS/VideoCMS Protocol Conversion **Edge Analytics Sensor Management** WISE-PaaS/OTA WebAccess/HMI WISE-PaaS/SignageCMS **Device Model** WISE-PaaS/Security WebAccess/NMS WISE-PaaS/HumanDetectAl

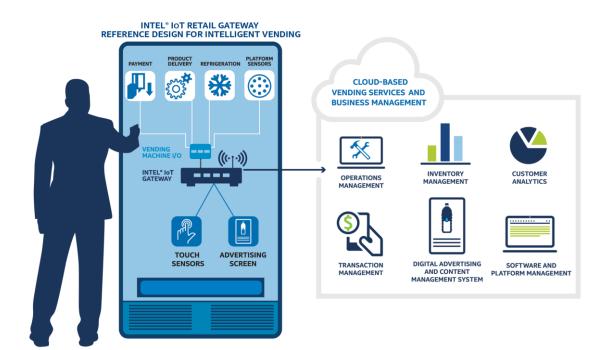


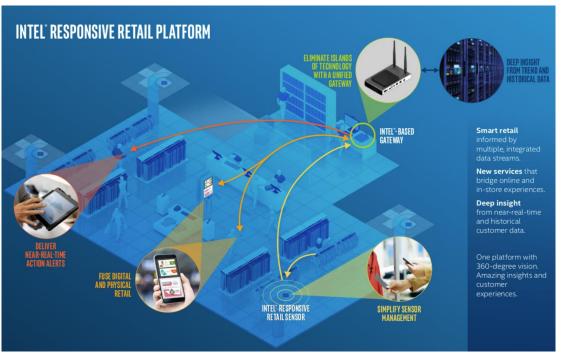
#### 使用场景#2: 万向集团智慧制造试验床



Confidential | ©2019 VMware, Inc.

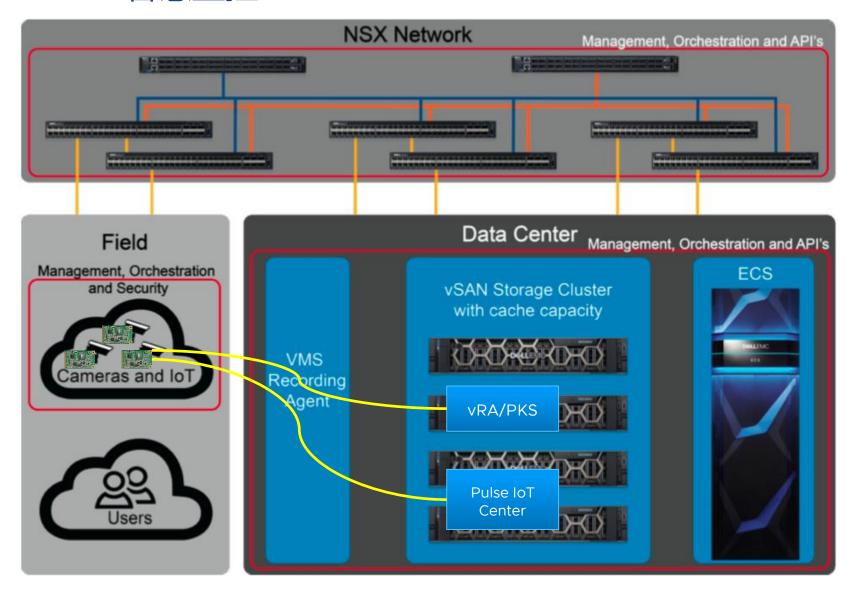
## 使用场景#3: Intel智慧零售计划







#### 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

