OPEN Compute Project | inspur | intel

OCP CHINA DAY | June 25th 2019 Beijing

The Practices and Learning on AI Edge Computing

Gang Chen | Network System Architect

25th June, 2019

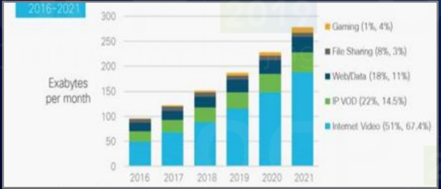
CONTENTS

01 Overall Considerations on AI Edge Computing

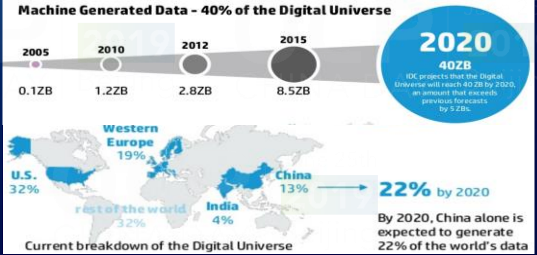
02 The Practices on AI Edge Computing

Streaming and Computation Schema

Streaming: growing steadily at 45% per year



Computation: most data will be computed on Edges


Machine Generated Data - 40% of the Digital Universe

By 2020, China alone is expected to generate 22% of the world's data

Current breakdown of the Digital Universe

Definition : 1080P -> 4K -> 8K

Fluency : 30fps -> 60fps ->120fps

Multi-Flows: Single Flows -> 360° immersive exp

Model: on-demand -> Live Broadcast

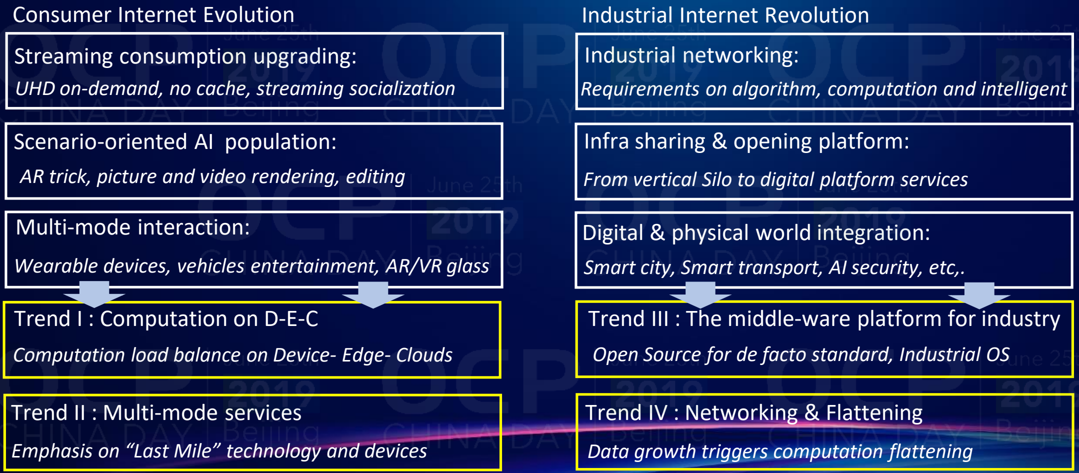
Industrial devices equipped with edge computing board

Smart Phone

Edge Computing@5G & CDN

OCP CHINA DAY June 25th 2019 Beijing

Consumer && Industrial Internet



Consumer Internet Evolution

Streaming consumption upgrading:

UHD on-demand, no cache, streaming socialization

Scenario-oriented AI population:

AR trick, picture and video rendering, editing

Multi-mode interaction:

Wearable devices, vehicles entertainment, AR/VR glass

Trend I : Computation on D-E-C

Computation load balance on Device- Edge- Clouds

Trend II : Multi-mode services

Emphasis on “Last Mile” technology and devices

Industrial Internet Revolution

Industrial networking:

Requirements on algorithm, computation and intelligent

Infra sharing & opening platform:

From vertical Silo to digital platform services

Digital & physical world integration:

Smart city, Smart transport, AI security, etc,.

Trend III : The middle-ware platform for industry

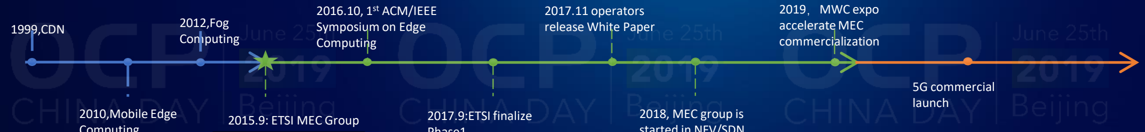
Open Source for de facto standard, Industrial OS

Trend IV : Networking & Flattening

Data growth triggers computation flattening

OCP CHINA DAY June 25th 2019 Beijing

5G Motivates Edge Computing Progress



1999,CDN

2010,Mobile Edge Computing

2012,Fog Computing

2015.9: ETSI MEC Group

2016.10, 1st ACM/IEEE Symposium on Edge Computing

2017.9:ETSI finalize Phase1

2017.11 operators release White Paper

2018, MEC group is started in NFV/SDN alliance

2019， MWC expo accelerate MEC commercialization

5G commercial launch

ETSI
World Class Standards

ETSI White Paper No. 11

Mobile Edge Computing
A key technology towards 5G

First edition – September 2015

ISBN No. [illegible]

2012,10: ETSI, NFV White paper

Network Functions Virtualisation – Introductory White Paper Issue 1

Network Functions Virtualisation

An Introduction, Benefits, Enablers, Challenges & Call for Action

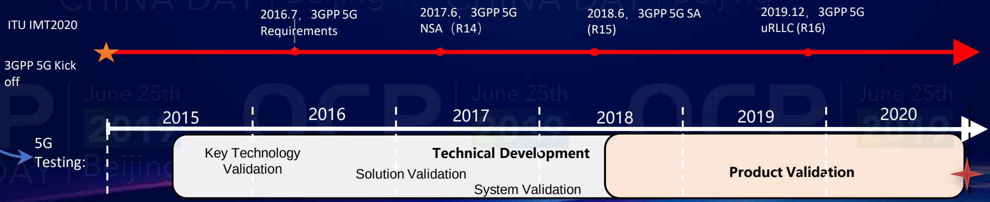
OBJECTIVES

This is a non-proprietary white paper authored by network operators.

The key objective for this white paper is to outline the benefits, enablers and challenges for Network Functions Virtualisation (as distinct from Cloud/SDN) and the rationale for encouraging an international collaboration to accelerate development and deployment of interoperable solutions based on high volume industry standard servers.

CONTRIBUTING ORGANISATIONS & AUTHORS

AT&T:	Margaret Chiosi.
BT:	Don Clarke, Peter Willis, Andy Reid.
CenturyLink:	James Feger, Michael Bugenhagen, Waqar Khan, Michael Fargano.
China Mobile:	Dr. Chunfeng Cui, Dr. Hui Deng.
Colt:	Javier Benitez.
Deutsche Telekom:	Uwe Michel, Herbert Damker.



ITU IMT2020

3GPP 5G Kick off

2016.7， 3GPP 5G Requirements

2017.6， 3GPP 5G NSA （R14）

2018.6， 3GPP 5G SA (R15)

2019.12， 3GPP 5G uRLLC (R16)

5G Testing:

Key Technology Validation

Technical Development

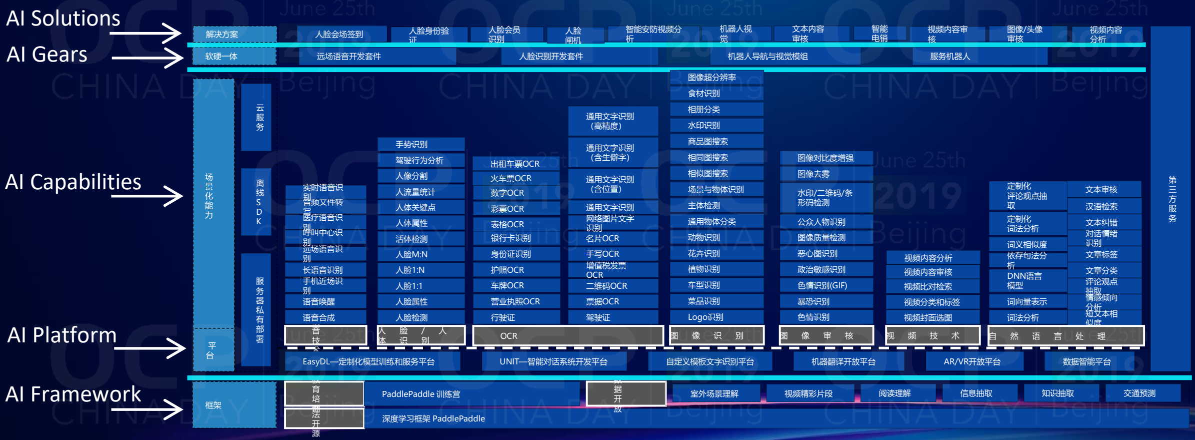
Solution Validation

System Validation

Product Validation

OCP CHINA DAY | June 25th 2019 Beijing

Edge Computing Powered by BAIDU AI Full-stack



AI Solutions → 解决方案：人脸会场签到 | 人脸身份验证 | 人脸会员识别 | 人脸闸机 | 智能安防视频分析 | 机器人视觉 | 文本内容审核 | 智能电销 | 视频内容审核 | 图像/头像审核 | 视频内容分析

AI Gears → 软硬一体：远场语音开发套件 | 人脸识别开发套件 | 机器人导航与视觉模组 | 服务机器人

AI Capabilities → 场景化能力：云服务 | 离线SDK | 服务器私有部署

语音技术：实时语音识别 | 音频文件转写 | 医疗语音识别 | 呼叫中心识别 | 远场语音识别 | 长语音识别 | 手机近场识别 | 语音唤醒 | 语音合成

人脸/人体识别：手势识别 | 驾驶行为分析 | 人像分割 | 人流量统计 | 人体关键点 | 人体属性 | 活体检测 | 人脸M:N | 人脸1:N | 人脸1:1 | 人脸属性 | 人脸检测

OCR：出租车票OCR | 火车票OCR | 数字OCR | 彩票OCR | 表格OCR | 银行卡识别 | 身份证识别 | 护照OCR | 车牌OCR | 营业执照OCR | 行驶证 | 通用文字识别（高精度） | 通用文字识别（含生僻字） | 通用文字识别（含位置） | 通用文字识别 | 网络图片文字识别 | 名片OCR | 手写OCR | 增值税发票OCR | 二维码OCR | 票据OCR | 驾驶证

图像识别：图像超分辨率 | 食材识别 | 相册分类 | 水印识别 | 商品图搜索 | 相同图搜索 | 相似图搜索 | 场景与物体识别 | 主体检测 | 通用物体分类 | 动物识别 | 花卉识别 | 植物识别 | 车型识别 | 菜品识别 | Logo识别

图像审核：图像对比度增强 | 图像去雾 | 水印/二维码/条形码检测 | 公众人物识别 | 图像质量检测 | 恶心图识别 | 政治敏感识别 | 色情识别(GIF) | 暴恐识别 | 色情识别

视频技术：视频内容分析 | 视频内容审核 | 视频比对检索 | 视频分类和标签 | 视频封面选图

自然语言处理：定制化评论观点抽取 | 定制化词法分析 | 词义相似度 | 依存句法分析 | DNN语言模型 | 词向量表示 | 词法分析 | 文本审核 | 汉语检索 | 文本纠错 | 对话情绪识别 | 文章标签 | 文章分类 | 评论观点抽取 | 情感倾向分析 | 短文本相似度

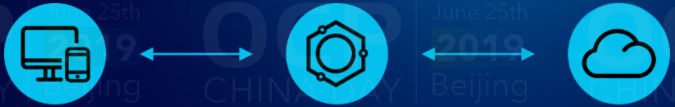
第三方服务

AI Platform → 平台：EasyDL—定制化模型训练和服务平台 | UNIT—智能对话系统开发平台 | 自定义模板文字识别平台 | 机器翻译开放平台 | AR/VR开放平台 | 数据智能平台

AI Framework → 框架：教育培训 | PaddlePaddle 训练营 | 数据开放 | 室外场景理解 | 视频精彩片段 | 阅读理解 | 信息抽取 | 知识抽取 | 交通预测 | 算法开源 | 深度学习框架 PaddlePaddle

OCP CHINA DAY June 25th 2019 Beijing

Computation Deployment: DEC Model



Device

Edge

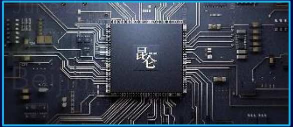
Cloud

D-E-C Computation Model

	Device	Edge	Cloud
算力	1-10 TOPS	10-100 TOPS	100+ TOPS
功耗	0.1-10 W	10-100 W	100+ W
延时	10-100 ms	ms~s	ms~s

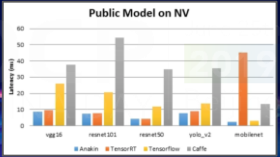
AI HW Accelerators

Up to 260Tops



AI IF Accelerators

Support Intel-CPU, NV-GPU, AMD-GPU and etc,.



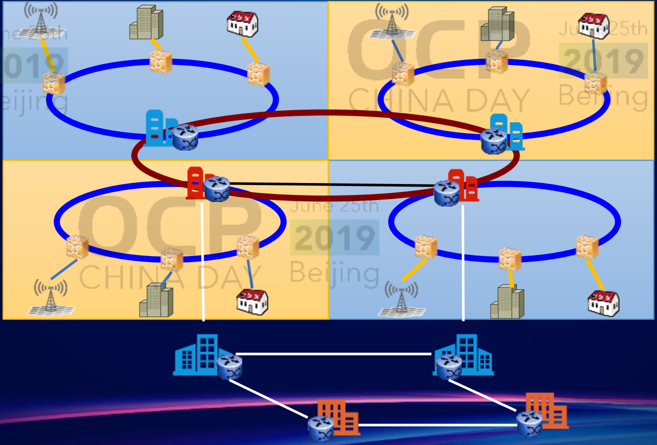
Network Deployment: Grid Model

Campus， 60k- 70k

Metro-access， 5k-7k

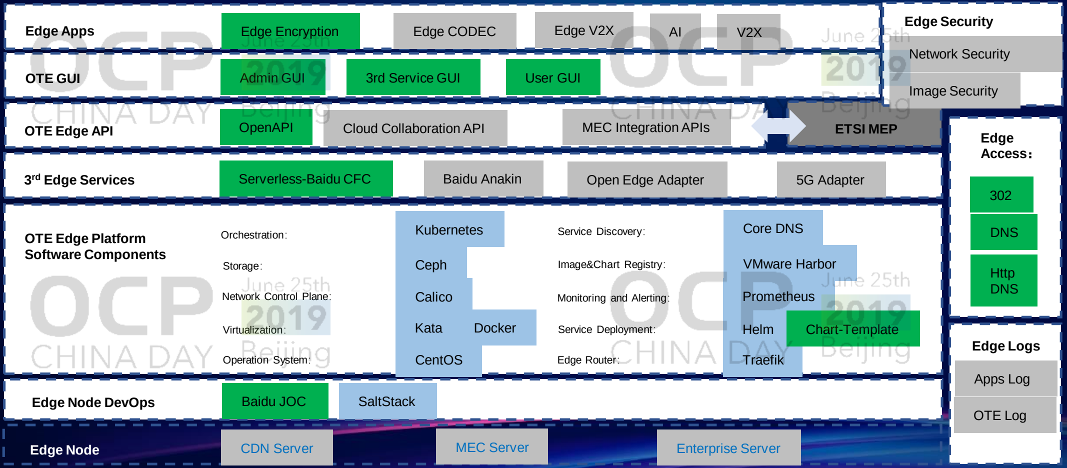
Metro-core， 600-700

Provincial ， 20-30



OCP CHINA DAY June 25th 2019 Beijing

OTE: Over The Edge



Edge Apps: Edge Encryption, Edge CODEC, Edge V2X, AI, V2X

OTE GUI: Admin GUI, 3rd Service GUI, User GUI

OTE Edge API: OpenAPI, Cloud Collaboration API, MEC Integration APIs ↔ ETSI MEP

3rd Edge Services: Serverless-Baidu CFC, Baidu Anakin, Open Edge Adapter, 5G Adapter

OTE Edge Platform Software Components

Orchestration:	Kubernetes	Service Discovery:	Core DNS
Storage:	Ceph	Image&Chart Registry:	VMware Harbor
Network Control Plane:	Calico	Monitoring and Alerting:	Prometheus
Virtualization:	Kata Docker	Service Deployment:	Helm Chart-Template
Operation System:	CentOS	Edge Router:	Traefik

Edge Node DevOps: Baidu JOC, SaltStack

Edge Node: CDN Server, MEC Server, Enterprise Server

Edge Security: Network Security, Image Security

Edge Access: 302, DNS, Http DNS

Edge Logs: Apps Log, OTE Log

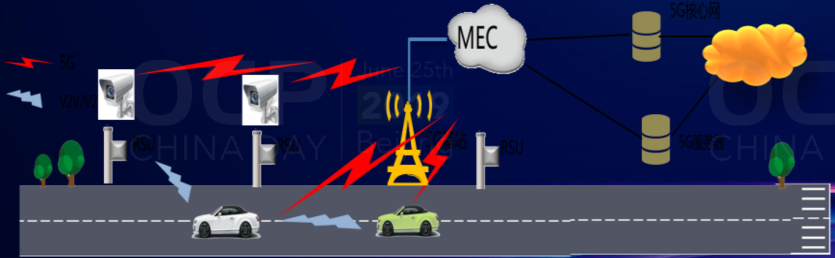
CONTENTS

01 Overall Considerations on AI Edge Computing

02 The Practices on AI Edge Computing

V2X @ MEC

C-V2X	Access	Bandwidth	Latency	Request MEC	Mobility
perception	5G/Fiber	8Mbps（上行）	5ms	Y	N
RSU	5G/Fiber	1Mbps（下行）	5ms	Y	N
Automotive	5G	50-100Mbps	10ms	Y	Y
Vehicle Entertainment	5G	20-50Mbps	20~30ms	Y	Y

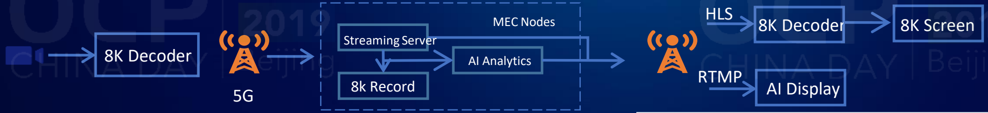


Smart Vehicles

Intelligent Roads

UHD 8K Live Broadcast @ MEC

- **8K UHD Broadcast based on 5G SA**

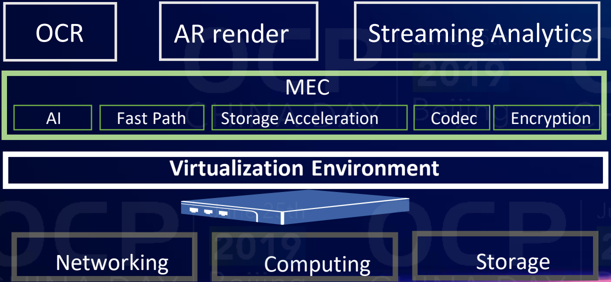


- >100Mbps symmetric streaming
- Live AI analytic processing
- Multiply streaming protocols supports

AR Rendering @ MEC

- Online Translations and AR rendering are completed on the edges

	4G	5G
Delay	24ms	5.8~23ms
BW	2.32Mbps	600Mbps



OCR

AR render

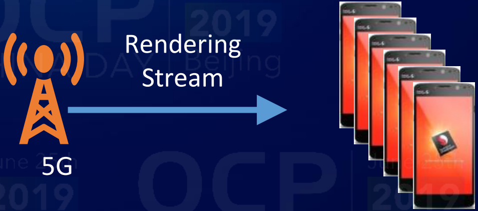
Streaming Analytics

MEC

AI | Fast Path | Storage Acceleration | Codec | Encryption

Virtualization Environment

Networking | Computing | Storage



Rendering Stream

5G

Thank you