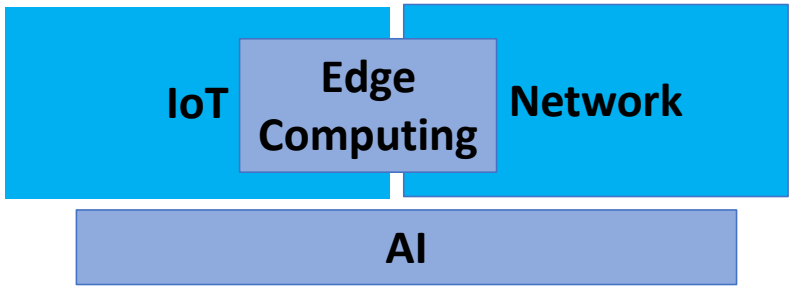
Converged AI for IoT and Network Edge

Hassnaa Moustafa, PhD
Principal Engineer, Intel

IEEE ComSoc Talk – Nov 3rd, 2022

IoT & Network Convergence



IoT generate huge amount of data

- 79.4 Zettabyte by 2025 Statista 2022
- By 2025, 75% of the data will be created and processed out-side of central data centers Gartner 2018

IoT services need expanded edge resources for data processing

- Enabling automation and intelligent functions through AI services
- Edge computing with 5G play key role Edge Computing Forecast, 2022-2027
- Software-defined Network functions scale to serve IoT workload needs

IoT application development follow cloud native approach

- Microservices-based approach for modularity, re-use, and ease of on-boarding, management, and orchestration

Private and Public 5G Network Infrastructure

- Private 5G is catalyst for IoT services Market Size & Segments Forecast 2022-2030
- Public 5G enables Cloud and Network Infrastructure intercept Cloud Telco 5G Edge and creates a revolution in media services (XDN and immersive media)

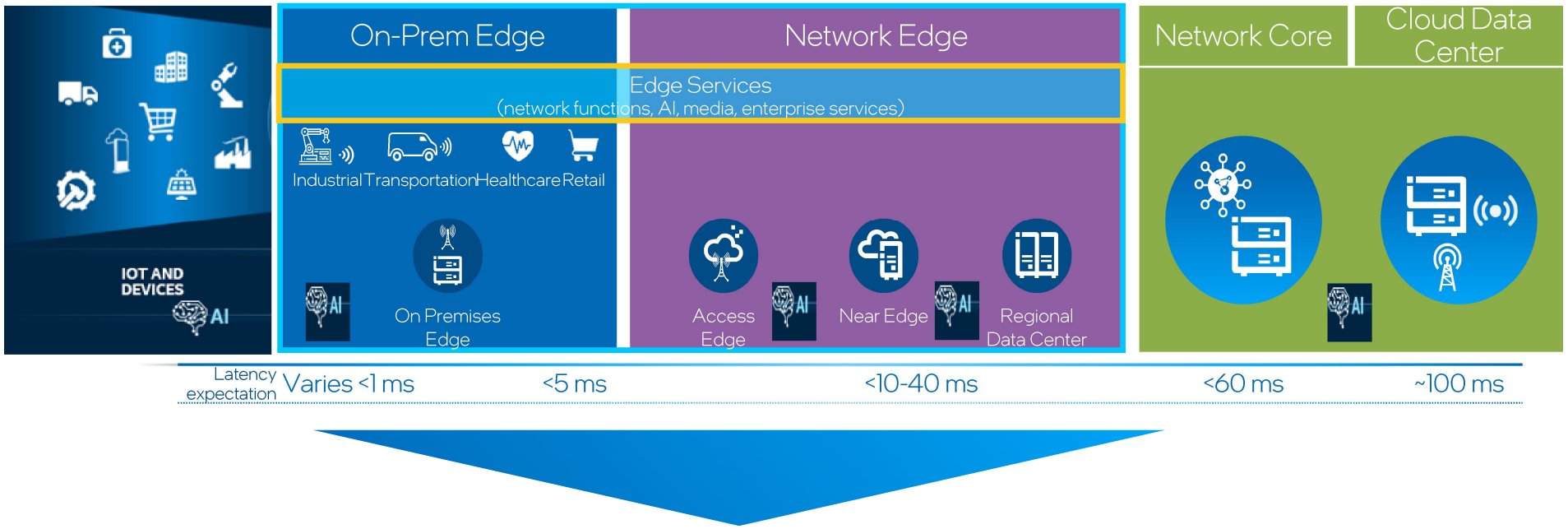
AI for Network functions automation (efficiency &better QoS/QoE)

- Service quality increase and Real-time services footprint expansion by > 50% (E///)
- Operation & maintenance cost decrease by up to 60% (E///)
- Energy saving increase by 238% over baseline solar (Huawei)

Cloud-Native Infrastructure as a Service emerging trend by Telcos

- Telcos infrastructure enables multi-tenant services and provide capabilities for microservices on-boarding, management, and orchestration

IoT & Network Edge – Big Picture

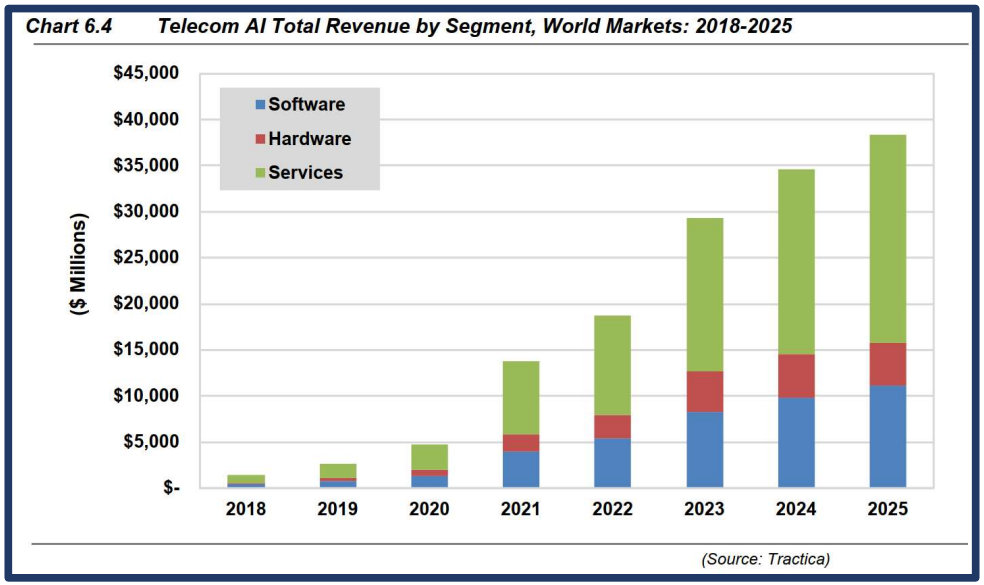


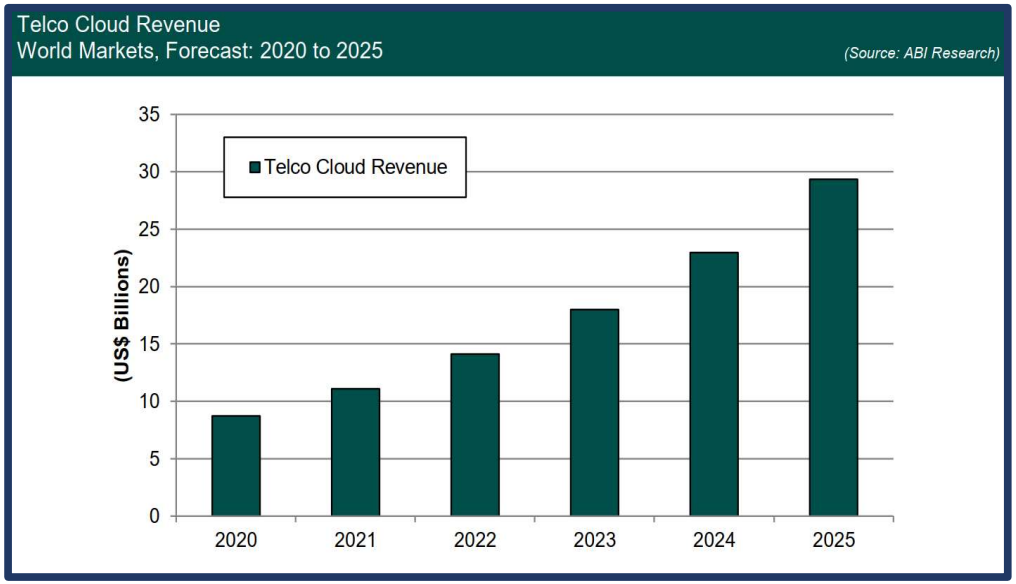
AI Services Convergence across IoT and Network Infrastructure

IoT AI-based services scale cross diverse edge locations expanding the network infrastructure footprint to on-premise edge

AI enables optimize the network functions to meet stringent KPIs for each category of service

AI Opportunity in Network Infrastructure

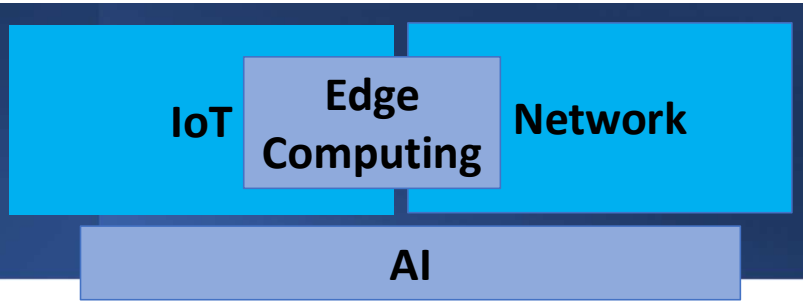




IoT and Network Converged Use Cases

- IoT services trigger the need for Edge Computing expanding the network infrastructure to the on-premise edge location
- AI is key feature in most IoT services and is needed in the network for better QoS/QoE and meeting SLAs for different use cases

IoT & Network Convergence



AI-based services are the most representative use cases and services that triggered/leverage the IoT & Network Convergence

Use Cases/Services Examples

- Smart Manufacturing
- Retail Services (smart stores)
- Safety and Security Applications
- Smart Cities
- Connected and Autonomous Vehicles
- eHealth
- Immersive Media (AR/VR)

Requirements

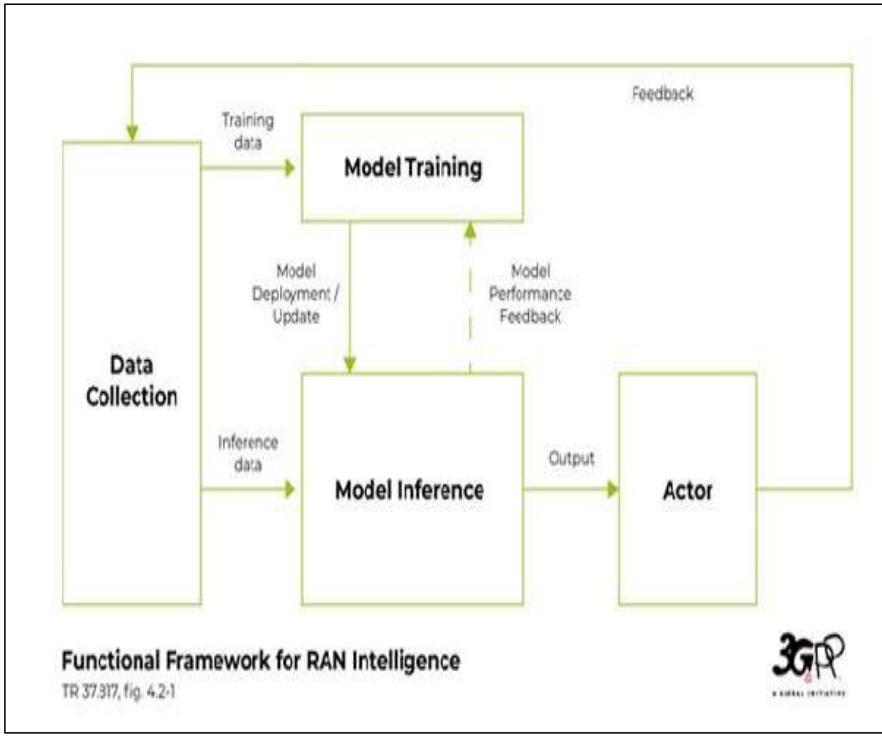
- **5G & Edge Computing**
- **AI**
  - **Image/Video Analytics**
  - **Time Series Data Analytics**
  - **Audio Analytics**
  - **Network Analytics for QoS/QoE**

Converged AI Consideration in Standards and Opensource Communities

- Besides AI/ML for automation and intelligent services, AI/ML is key for telcos for predictive analytics, anomaly detection, trend analysis, and clustering to enable customer experience management, personalized marketing or data monetization in addition to network management
- With 5G, advanced prescriptive analytics are considered to enable closed-loop automation and self-healing networks

AI Consideration in 3GPP – 3GPP RAN3

- 3GPP RAN3 RAN studies the support of **AI/ML** through a **functional framework**
- Study on enhancement for data collection for NR and ENDC is approved **TR 37.817**
- **Deployment options for AI/ML functions**
  - AI/ML Model Training in OAM and AI/ML Model Inference in the gNB (gNB-CU for split RAN).
  - AI/ML Model Training and AI/ML Model Inference are both in the gNB (gNB-CU for split RAN).
- **Main Use Cases**
  - **Network Energy Saving**: improve energy through automated features such as traffic offloading, coverage modification, and cell deactivation
  - **Load Balancing**: improve performance through intelligent distribution of load and load prediction
  - **Mobility Optimization**: maintain performance level during mobility based on UE connectivity prediction



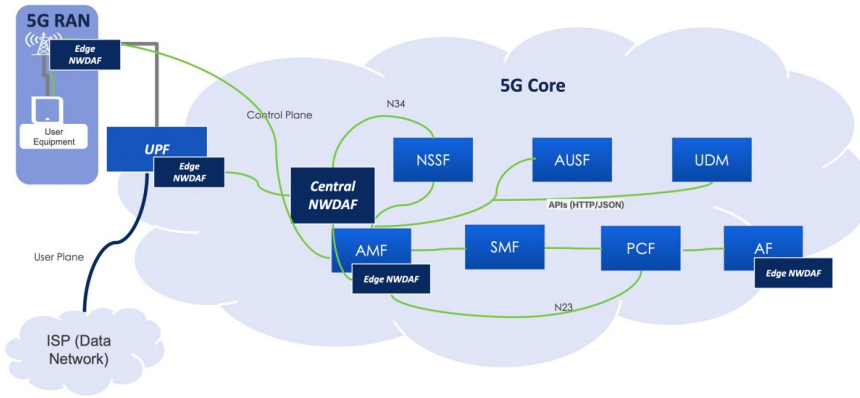
Functional Framework for RAN Intelligence
TR 37.317, fig. 4.2-1

AI Consideration in 3GPP – 3GPP SA2

- 3GPP SA2 specifies Network Data Analytics Function (NWDAF) integrating analytics into the network to derive actionable insights
  - Considers AI/ML to process in real time the vast streams of Key Performance Indicators (KPIs) from the network
  - Leverages advanced traffic classification and deep packet inspection techniques
  - Insights are applied to 5GC to enhance functionality

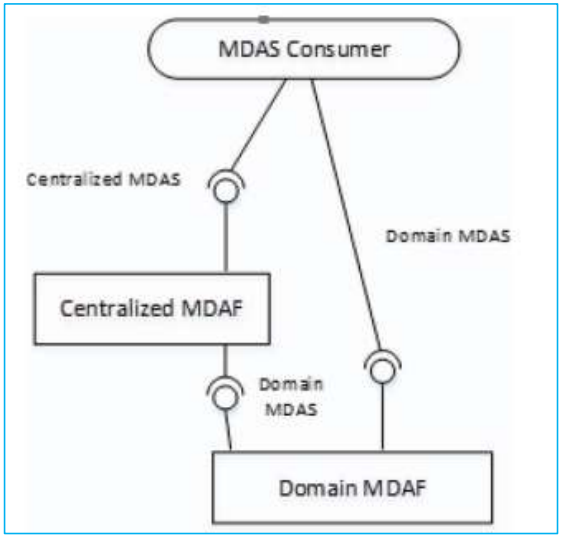
- AI-ML analytics use cases for 5G using NWDAF 3GPP TR 23.791
  - Load-level computation and prediction for a network slice instance
  - Service experience computation and prediction for an application/UE group
  - Load analytics information and prediction for a specific NF
  - Network load performance computation and future load prediction
  - UE behavior prediction, Abnormal behavior/anomaly detection, Mobility & Communication Pattern Prediction
  - Congestion information with prediction for a specific location
  - Quality of service (QoS) sustainability

5G Architecture with Distributed NWDAF



AI Consideration in 3GPP – 3GPP SA5

- 3GPP SA5 defines Management Data Analytics Function (MDAF) to enable service assurance
- MDAF provides Management Data Analytics Service (MDAS) to support management and orchestration
  - Centralized PLMN-wide MDAFs for E2E slice assurance for example
  - Domain specific MDAF deployment (RAN, CN, NSSI)
- Closed-loop assurance between domain MDAF and Centralized MDAF



Management Data Analytics Services

AI Consideration in ETSI – ZSM & ENI

ETSI Zero-touch Service Management (ZSM)

- Architecture to support zero-touch fully automated management and operations
- Goal is to provide self-configuration, self-monitoring, self-healing and self-optimization
- Recognizes different management domains and describe the services for these domains
- Considers AI/ML closed-loop control (collect data, analysis, decide, and act)

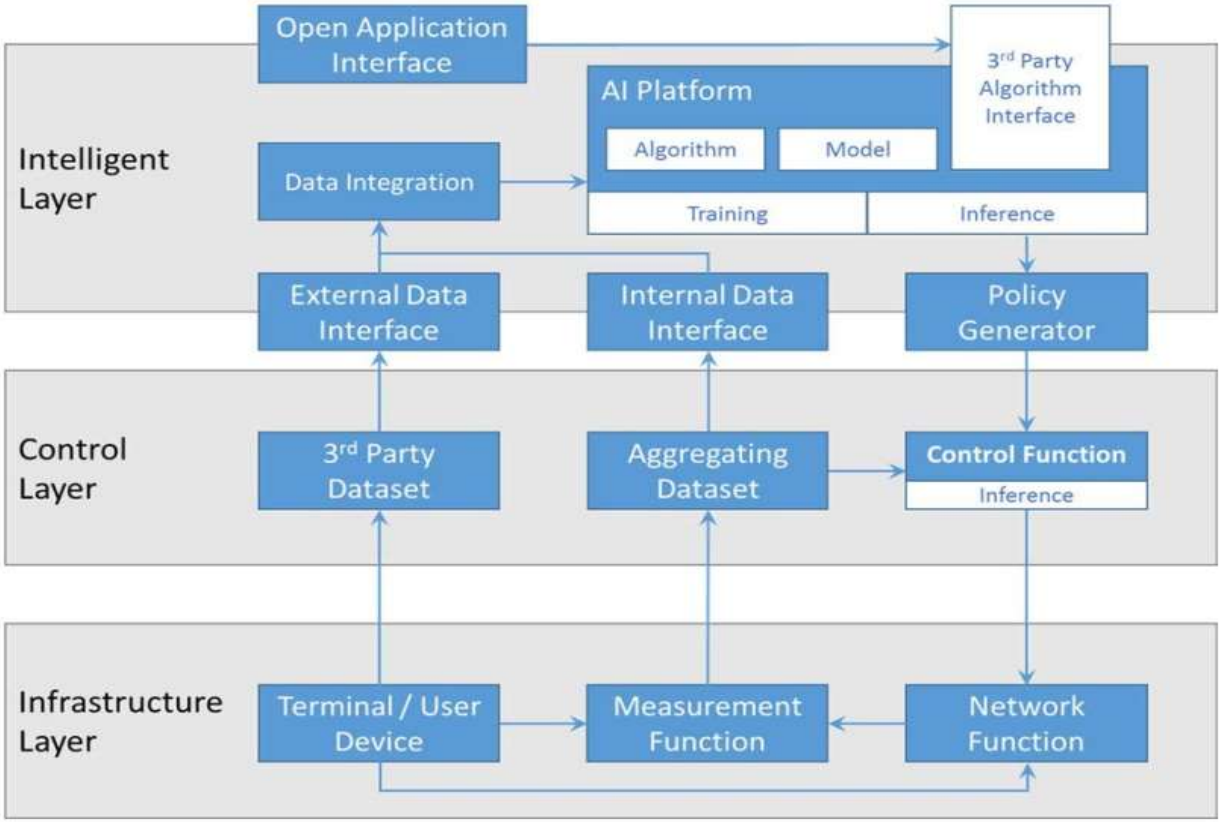
ETSI Experiential Networked Intelligence (ENI)

- Cognitive Network Management Architecture
- Cognitive layer for the telco industry
- Use AI to adjust the offered services based on dynamic user needs and business goals
- Add intelligence on top of legacy systems

AI Consideration in ETSI - IDN

ETSI Intelligence Defined Network (IDN)

- IDN Architecture integrating with various network nfrastuctures
- Learning from historical and new data and make predictions and intelligent decisions
- Support human-based decision by pre-processing data and rendering insights to users through advanced UIs



AI Consideration in IEEE

IEEE Standards for Activities Intelligent Systems (AIS)

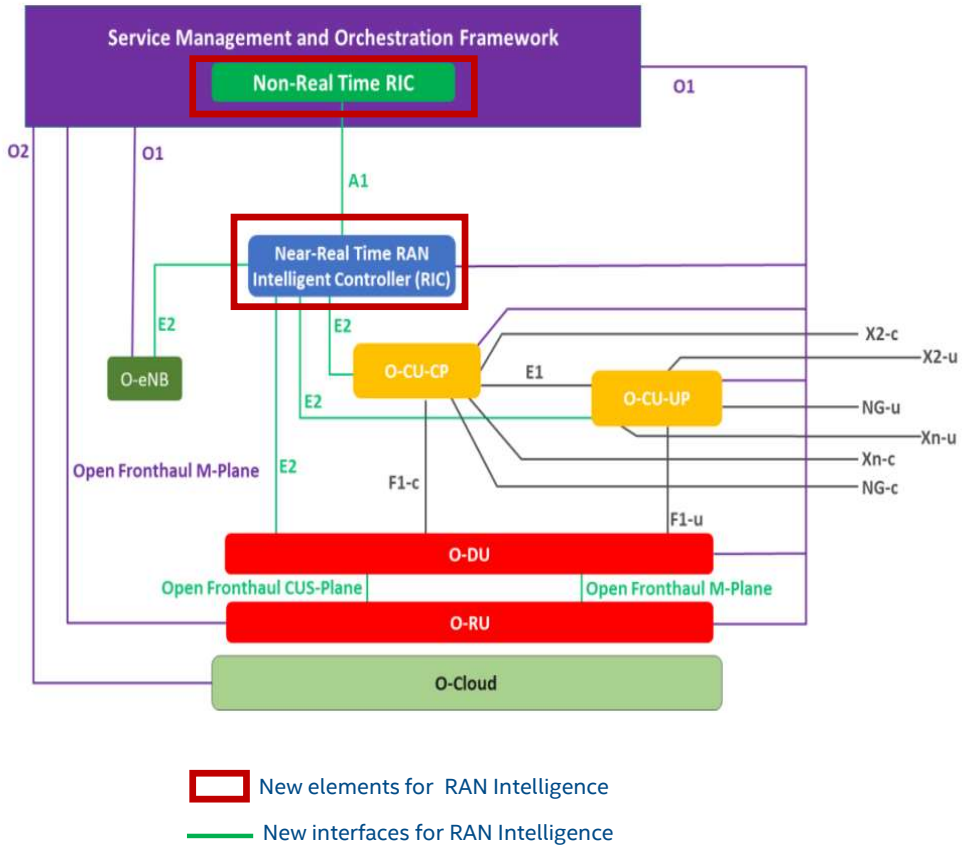
- Autonomous Robotics Ontology IEEE P1872.2™
- Augmented Reality Learning Experience Model IEEE 1589-2020™
- Taxonomy and Definitions for Connected and Automated vehicles P2040/P2040.1™
- Framework and Process for Deep Learning Evaluation IEEE P2841™
- Standard for Responsible AI Licensing IEEE P2840™

IEEE Standards for AI Affecting Human Well Being

- Child and Student Data Governance IEEE P7004™
- Transparent Employer Data Governance IEEE P7005™
- Personal Data AI Agent IEEE P7006™

AI Consideration in ORAN

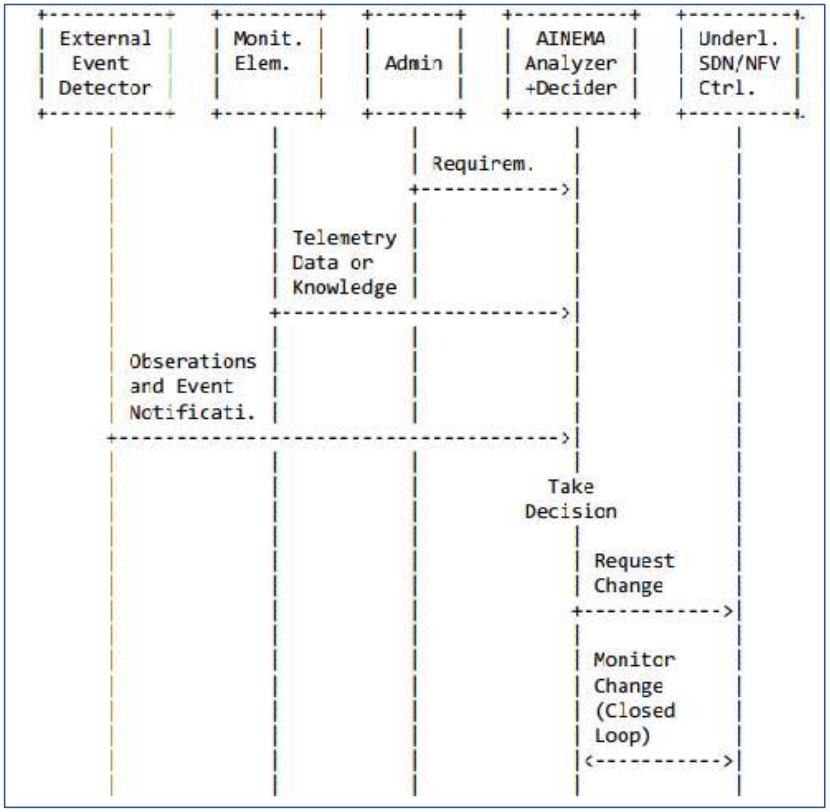
- Open-RAN Alliance (ORAN) evolves 3GPP access with Open Interfaces and Intelligence
- Definition and Specification of RAN Intelligence
  - Near-Real-time RAN Intelligent Controller (near-RT RIC)
  - Non-Real-time RAN Intelligent Controller (non-RT RIC)
- Open-source implementation
  - ORAN Open-source Community (OSC)
  - Open Network Foundation (ONF)



AI Consideration in IETF

IETF Network Management Research Group (NMRG)

- Evolution of Network Management with AI
- Evaluate the gap between Network Management Solutions and AI Solutions
- Identify the research challenges for coupling AI with Network Management
- AI Framework for Network Management (AINEMA)



AINEMA Workflow

AI Consideration in Linux Foundation (LF)

LF Networking (ONAP)

- Open Networking Automation Platform (ONAP)
- Provide comprehensive platform for orchestration, management, and automation of network and edge computing services
- Provide non-real-time RIC implementation

LF ORAN SW Community (OSC)

- Collaboration between ORAN Alliance and LF
- OSC SW projects include RAN Intelligent Controller (RIC) beside other open RAN components

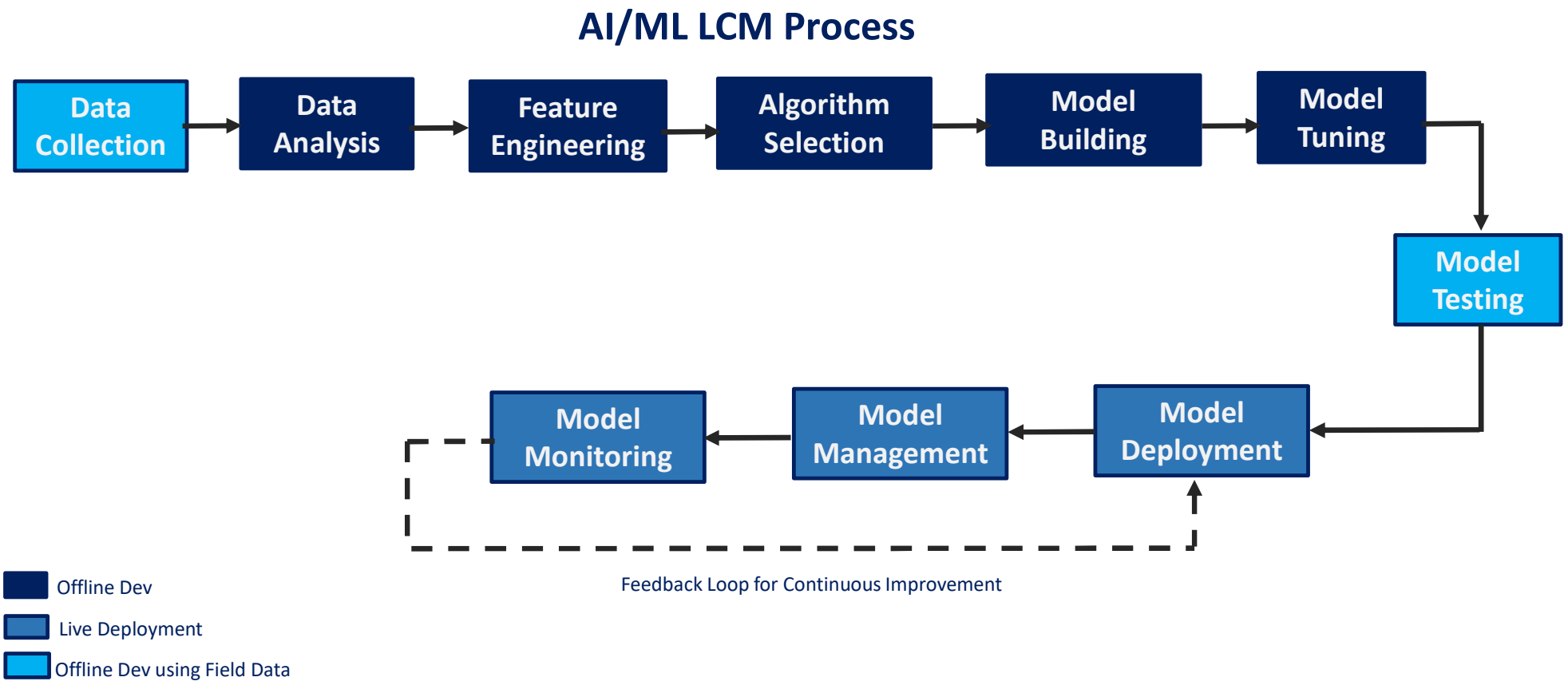
LF Edge (Akraino)

- Opensource Edge Stack tested with several edge platforms
- Deliver blueprints for new edge use cases (e.g., robotics, and AR/VR) combined with 5G Open RAN and 5G Core components)
- Wide participation across industry and research

AI/ML Life Cycle Management in the Network & Challenges

- AI/ML in the network is driven by multiple players in the ecosystem
- Besides standards alignment, a synergy is required between the AI/ML capabilities across the multiple ecosystem players

Life Cycle Management for AI/ML in the Network



Ecosystem Actors for AI/ML in the Network

Functional Distribution in an Open AI/ML Ecosystem

AI/ML Enabling Functionality	Ecosystem Player
Building AI/ML models/algos	• Network SW Vendor, TEM, Telco
AI/ML Model Inference	• Telco, TEM
Data collection	• Telco
AI/ML Model Training	• Telco, TEM, SMO vendor, CSP
AI/ML Model Update	• Telco, TEM, SMO Vendor, CSP
Regulatory Compliance (data anonymization, privacy, secure access, secure storage)	• Telco, TEM, SMO Vendor
AI/ML Model Secure Execution	• Telco, TEM, SMO Vendor

Challenges to Adopt AI/ML in the Network

The Life Cycle Management (LCM) of AI/ML models introduces new aspects beyond traditional software LCM processes

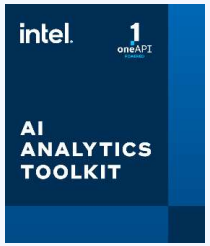
Challenges for LCM of AI/ML in the Network

- Lack of access to data (real data from operational network) due to regulations regarding privacy and ownership
- Training AI/ML is mostly done by SW vendors using opensource data from research
- Fragmentation and overlap in different standards and open- source initiatives
- Time for telcos to trust automation technologies
- Clear Return-of-Investment (ROI) for telco to add AI functions in the network
- Multiple actors in the network AI ecosystem (Telco, TEMs, SIs, ISVs, SMO Vendors, CSPs) and lack of full alignment

AI & Analytics Tools – Big Picture

Intel® oneAPI Software Tools for AI and Analytics

Intel® oneAPI Toolkits



Intel® oneAPI AI Analytics Toolkit (AI Kit)

Accelerate machine learning and data science pipelines with optimized deep learning frameworks and high-performing Python libraries

Data Scientists, AI Researchers, DL/ML Developers

Intel® oneAPI Base Toolkit (Base Kit)

Incl. Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), and Intel® oneAPI Data Analytics Library (oneDAL)

Optimize primitives for algorithms and framework development

DL Framework Developers - Optimize algorithms for Machine Learning and Analytics

Toolkit Powered by oneAPI

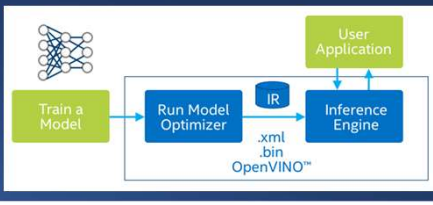
Intel® Distribution of OpenVINO™ Toolkit

Deploy high performance inference and applications from edge to cloud

AI Application, Media, and Vision Developers

OpenVINO™

Intel Distribution of OpenVINO™ Toolkit



Deep Learning

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

IR = Intermediate Representation file

Open Model Zoo

40+ Pretrained Models

Sample Apps

Model Downloader

Deep Learning Workbench

Calibration Tool

Model Analyzer

Benchmark App

Accuracy Checker

Aux. Capabilities

Traditional Computer Vision

Optimized Libraries & Code Samples

OpenCV*

OpenVX*

Samples

For Intel CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For GPU/Intel® Processor Graphics

Optimize Intel® FPGA (Linux* only)

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

OS Support: CentOS* 7.4 (64 bit), Ubuntu* 16.04.3 LTS (64 bit), Microsoft Windows* 10 (64 bit), Yocto Project* version Poky Jethro v2.0.3 (64 bit), macOS* 10.13 & 10.14 (64 bit)

Intel® Architecture-Based Platforms Support

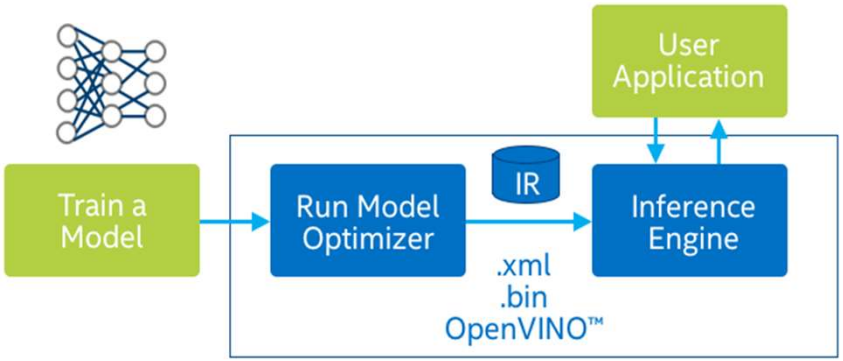
Intel® Vision Accelerator Design Products & AI in Production/ Developer Kits

An open source version is available at 01.org/openvinotoolkit (deep learning functions support for Intel CPU/GPU/NCS/GNA).

Inference Workload Optimization through OpenVINO™

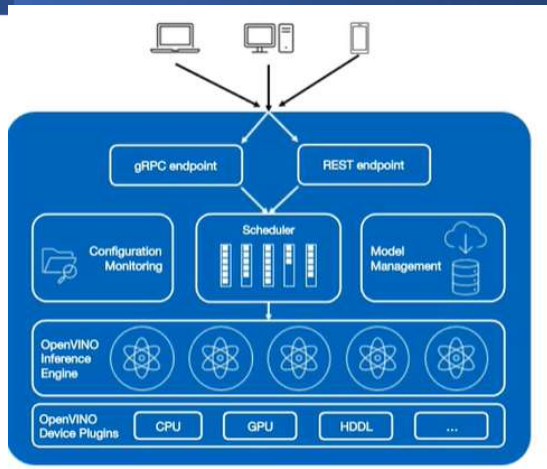
OpenVINO
- AI Inference Optimization on Intel Platforms
- Deep Learning models with sample apps (detection/classification/segmentation)

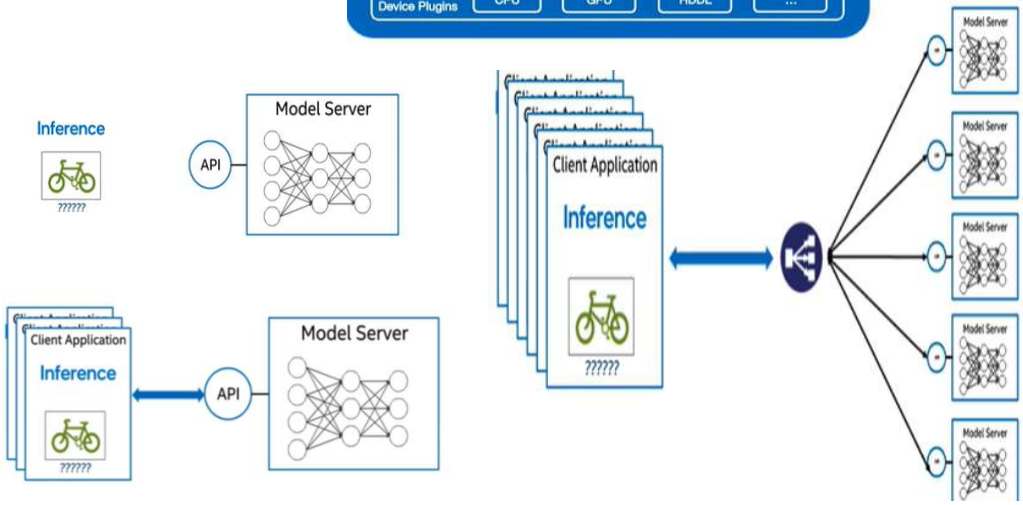
OpenVINO Model Server (OVMS)
- Production grade inference server
- Ease edge AI applications deployment & scaling



OpenVINO Inference Optimization

OpenVINO Model Server (OVMS)



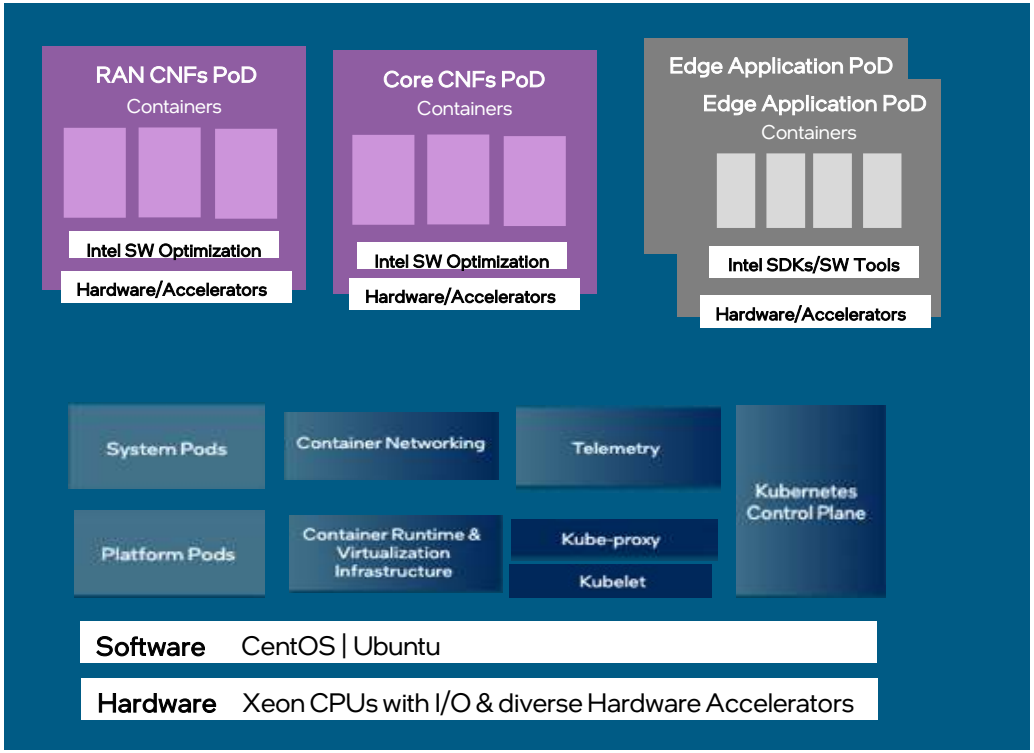


Converged IoT and Network Edge - Frameworks

Intel® Smart Edge

Edge-native Kubernetes Certified Distributed Computing Platform

- Enables deployment and management of IoT and Network workloads at the with cloud-like ease, resiliency and security
- Runs demanding workloads like AI, media, and software-defined networking functions to enable 5G and network services
- Intelligent workload management and distribution with power-consumption awareness and resources awareness



Q&A

Thank you ….