IEEE Future NETWORKS

Enabling 5G and Beyond

International Network Generations Roadmap (INGR) Virtual Workshop

Edge Automation Platform

Prakash Ramchandran
Sujata Tibrewala
TK Lala

22nd October 2020

IEEE

10-year Vision

Edge will become the Key service delivery vehicle for systems and applications rather than cloud, which would mean ubiquitous thin clients have access to all the processing power edge can offer

- Micro Data Centers will spawn over different locations
- Distributed Access –
  - Edge Agnostic
  - Access Agnostic
  - Latency, Bandwidth and Coverage are three salient features
- Always available- Resilient Fault Tolerant platforms
- Leverage AI functionalities to facilitate the automation in Edge infrastructure operation and dynamic adjustment to improve the target application performance
- Delivery vehicles for the edge using microservices and container deployment

Scope

Service Based Edge Architecture:

- **Diverged use cases or workloads** (Platform Automation, Connected driving, Edge as a Service,Offloads from Cloud to Edge and vice versa, Intelligent Edge- audio/video/networking, Automated Manufacturing - IIoT)
- **Specify Declaratively Best Known Configuration** (BKC)   to build and automate both H/W & S/W
- Look at latest Frameworks/Architecture – ONAP, ORAN, MEC , Open Edge Infrastructure, OpenNESS & oneAPI
- **Identify common objectives and gaps** (for Consider BM & Containers) in microservices based objects
- Define the Reference Model for Edge, Interfaces, APIs
- Use case flows to formulate KPIs (3GPP, MEC, ORAN etc are defining KPIs)
- **Challenges to Achieve the KPIs & define edge required matrix of capabilities .**
- Automation/Orchestration tools
- And thoughts on future, innovations

Working Group Team

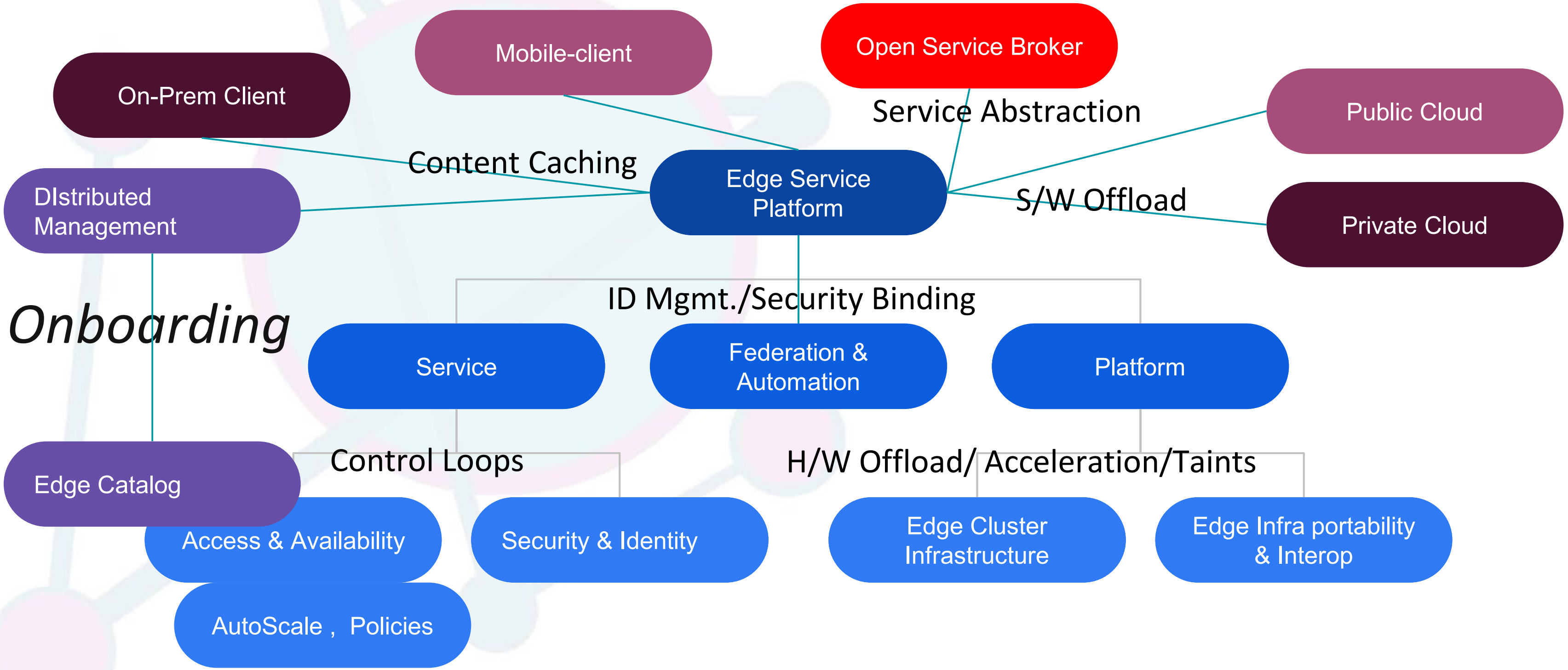
NAME	COMPANY
Sujata Tibrewala	Intel
Cagatay Buyukkoc	AT&T
Prakash Ramchandran	Dell
Liangkai Liu	Wayne State
TK Lala	ZcureZ
Frederick Kautz	Doc.ai, CNCF-TUG
Estefanía Coronado	i2CAT
Roberto Riggio	i2CAT
Someswar Ganugapati	AT&T
Sunku Ranganathan	Intel

CROSSTEAM GROUPS	COMPANY
Mohamad Patwary	Burmigham City School of Computing
Sunku ranganath	Intel

Top Needs for 10-year Vision

DRIVER NAME	METRIC (CURRENT STATE)	PROJECTED METRIC 3-YEAR (2023)	PROJECTED METRIC 5-YEAR (2025)	PROJECTED METRIC 10-YEAR (2030)
Standardization of Edge Platform	Evolving (converging to Kubernetes for containers & OpenStack for VMs over any edge Clusters & Nodes)	Interoperability between the workload VMs & Containers on Logical Hosts(Nodes)	Disaggregated workloads both Centralized & DIstributed Service Delivery compete for innovation to meet KPIs	Edge Services Becomes commoditized and Quantum computing emerges as disruption.
Standardization of appn / service via containers	Service Delivery getting better defined with VMs or Containers as hybrid approach is becoming more common	Common Orchestration Framework is emerging as Hybrid Service Gateways use Kubernetes as base	Hardware Security combined with Software Security	Commoditized (New Services using Quantum computing standards emerge)
Characterization (QoE)	10-20 ms, 1Gbps	5ms, 10 Gbps	1ms, 50 Gbps	100µs, ~100Gbps[DM4]
Security & Identity	Cyber Physical Security Service specific security embedded with multi-interface	AI powered security, Slice based and workload based Security, Privacy Preserving AI & Data	Multi-Layer Multi Modal automation for unified infrastructure and resource dependent security algorithms ?	Personalised decentralised security based on workload, device, interface etc.
Support of Heterogeneous Hardware	Multiple vendor platforms in compute, storage and networking with programming options such as P4, SYCL, DPC++	**A few parallel programming paradigm will emerge among the existing systems with Bare Metal/Thin real time OS suppor**t	Uniform support of heterogeneous computing irrespective of operating systems, drivers, debugging tools, etc without a hit on performance	True Cloud native computing using heterogeneous hardware, based on workload demand, via automated intelligent orchestration
Hybrid Cloud -Edge Service Platform (HC-ESP) cloud)	Uniform Infrastructure Edge Service Platform (ESP) is picking up steam.	With more managed edge clouds HC-ESP will become the norm.	HC-ESP may lead to more modular with Intelligent Infrastructure elements like GPU. FPGA, SmartNICS	More value adds expected like Quantum Solution Networks with Intelligent Infrastructure,

Hybrid Cloud (HC) -Edge Service Platform (ESP)



Mobile-client

Open Service Broker

On-Prem Client

Service Abstraction

Public Cloud

Content Caching

Edge Service Platform

DIstributed Management

S/W Offload

Private Cloud

ID Mgmt./Security Binding

Onboarding

Service

Federation & Automation

Platform

Control Loops

H/W Offload/ Acceleration/Taints

Edge Catalog

Access & Availability

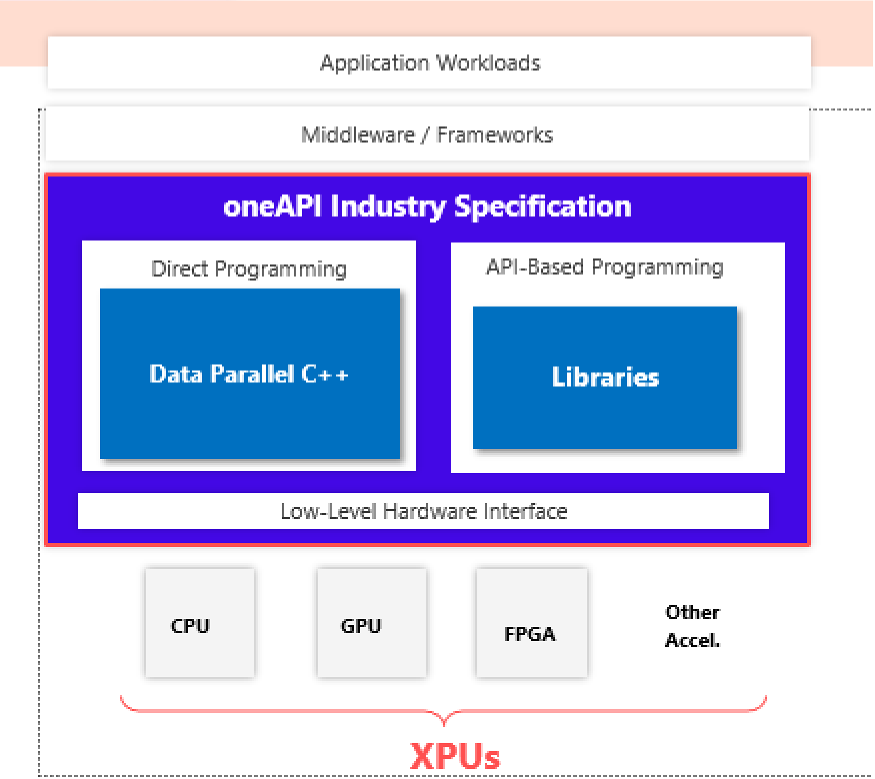
Security & Identity

Edge Cluster Infrastructure

Edge Infra portability & Interop

AutoScale , Policies

Heterogeneous Programming Support- DPC++



Application Workloads

Middleware / Frameworks

oneAPI Industry Specification

Direct Programming

Data Parallel C++

API-Based Programming

Libraries

Low-Level Hardware Interface

CPU

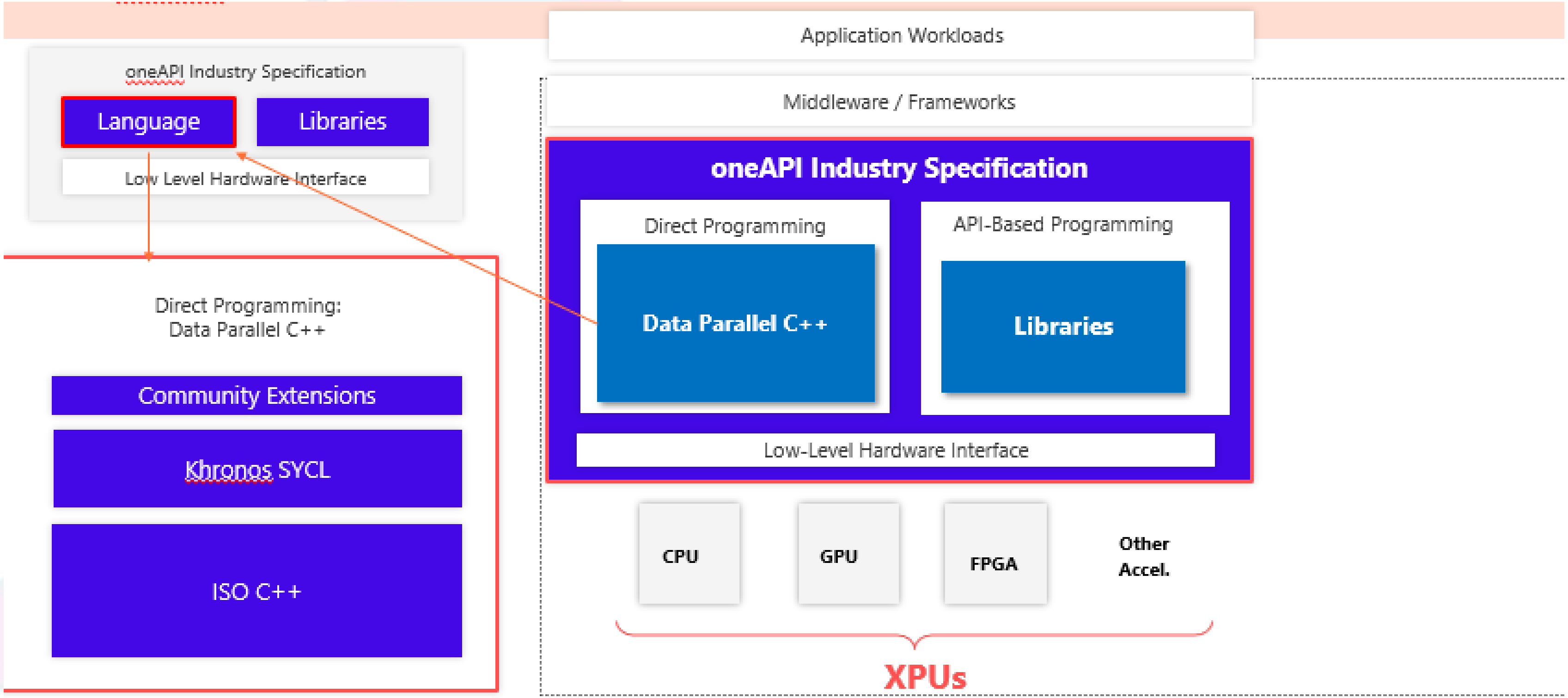
GPU

FPGA

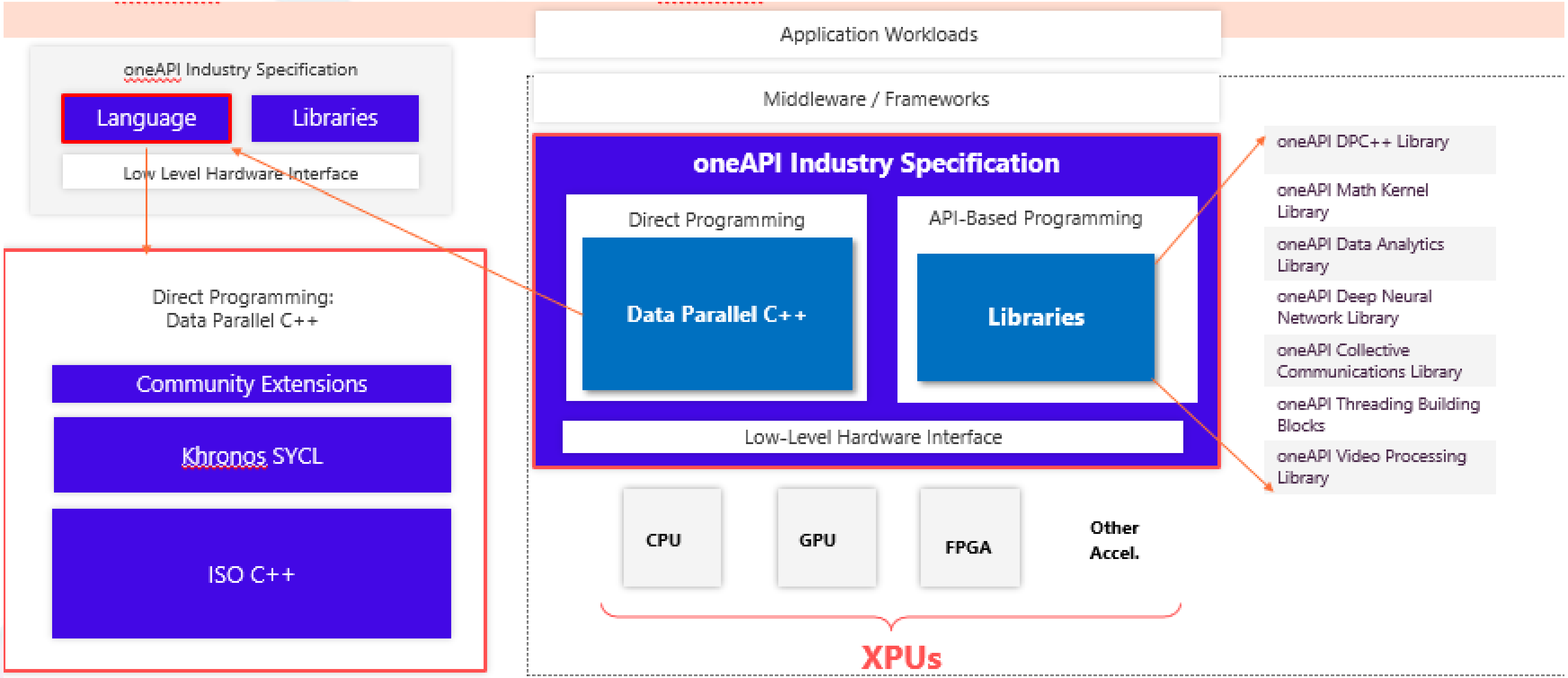
Other Accel.

XPUs

Heterogeneous Programming Support- DPC++



Heterogeneous Programming Support- DPC++



oneAPI Industry Specification

Language

Libraries

Low Level Hardware Interface

Direct Programming:
Data Parallel C++

Community Extensions

Khronos SYCL

ISO C++

Application Workloads

Middleware / Frameworks

oneAPI Industry Specification

Direct Programming

Data Parallel C++

API-Based Programming

Libraries

Low-Level Hardware Interface

CPU

GPU

FPGA

Other Accel.

XPUs

oneAPI DPC++ Library

oneAPI Math Kernel Library

oneAPI Data Analytics Library

oneAPI Deep Neural Network Library

oneAPI Collective Communications Library

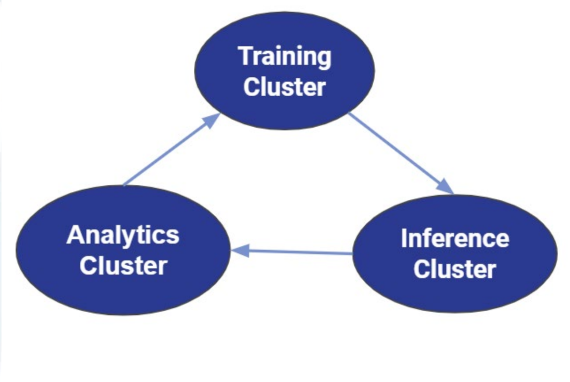
oneAPI Threading Building Blocks

oneAPI Video Processing Library

Visit for more details

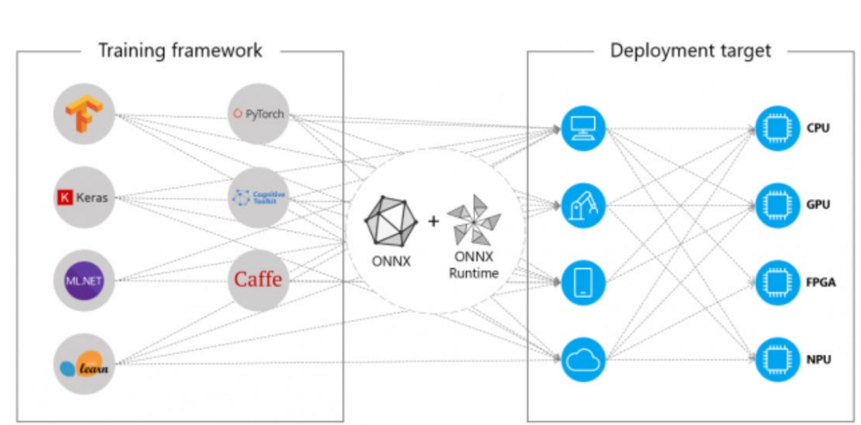
Overcoming Constraints of Edge & ML

Distribute Computation



- Training Cluster –Cloud (prefer high computes)
- Production Cluster –Edge (Prefer Inferencing and pretrained models)
- Analytics Cluster – Local/Global (Prefer time series aggregations)

Interop, portability & Device Acceleration



Open Neural Networking Exchange (ONNX) allows use of appropriate Model , Tools and Framework at right location and allow separation of concerns for optimal use of AIML model for both Training in Cloud and Inferencing at the Edge

Security for EAP

- **EAP Security Objectives and challenges**
  - Adapt and Optimize Business and Risk Management in Edge and Hybrid context
  - Higher, Broader, Faster services with 5G and next gen network with real time URLLC constraints
  - Numerous distributed smart and semi-smart devices and federated workload identity with wider attack surfaces
  - Sustain Core (foundational) Security at all time: IAM (Identity and Authentication) and Access Management (RBAC), Content Integrity management at rest and transit, Security confirmation, verification, attestation and conformity certification, Service Availability, Privacy (Data, Identity)
- **Security Strategic phases:**
  - Cradle to grave protection: Lifecycle DevOps protection of EAP services and resources,
  - Pre Attack protection:
    - Threat and Attacker modeling, Smart Threat anticipation, Scanning (SAST, DAST, vul scan), real time threat(anomalies) monitoring, DevOps
  - During Attack protection:
    - Contain and isolate blast radius, fall back resources and lower , initiator and trigger tracing
  - Post Attack restoration:
    - Threat thwart and stopping, real time Incident response including threat trap, Gradual shutdown, healing restoration.
  - Future Attack Prevention:
    - Analyze and adapt security armor with updated threat anticipation
- **EAP Security Key Focus:**
  - Security to adapt and support Complex Scalability and Limitless context
    - federated chain of trust framework (Blockchain etc.) to support Zero Trust mindset
    - cradle to grave (Full DevOps) security
  - Real time continuous security monitoring and actionable visibility

Security for EAP (cont’d)

- **EAP Security Key focus: (cont’d)**
  - Security in Next generation networks (5G and beyond):
    - Control plane and user plane: control plane centralized or federated (at the edge node) with high security, user plane distributed
    - Ultra low latency and higher and broader bandwidth:
      - Security seeded at the lowest level: TEE, Hardware TPM, High perf confidential computing
      - Use case, context and slice service cognizant and adaptive selective security
  - Security for Virtual environment and infrastructure:
    - Networking and connectivity
      - SDN, NFV : (security for control , User/data plane separation, management)
      - API servers and Gateways (server client context and need scrutiny and access adjustment)
    - Compute, Storage
      - Modern and Future Apps and services:
        - » VMs, Containers, FaaS: Workload Orchestrators (K8s- KubeEdge), Distributed Microservices microsegmentation, serverless, service mesh (e.g.,Linkerd, istio) : image integrity , east west traffic control, policies including Anti-affinity
        - » Blast radius minimization and containment
  - Security fueled through Policy and AI/ML driven Automation:
    - Threats: Continuous detection, smart anticipation prediction real time as well as history based
    - Auto incident response including Self isolation and restoration
    - Static and Dynamic Policy driven Auto adaptation of tailored security (one size does not fit all)

FCAPS for EAP

- Sidecar for ‘faster’ infrastructure
  - Sidecar proxies play a crucial role in data plane performance scalability as they increasingly act as TLS endpoints
  - Side cars take on bulk of network management, configuration and scale decisions
  - Hardware offloads of side car functionality would be differentiating factor for faster infrastructure

- Distributed & localized availability
  - Emergent systems don’t have properties of whole system, thus requiring role-based identity controls
  - Decentralized models with likes of Kubernetes requires tracking of persistent state of end to end service availability
  - Localized service mesh control plane decisions reduces latency for availability scaling across network slices

FCAPS for EAP (Contd…)

- Rethink microservice based reliability
  - Fault management mechanisms would leverage service mesh end points improving application resiliency
  - Service mesh control plane need to be aware of network slicing requirements to configure & control data plane side cars
- Observability driving assurance
  - Cloud native edge deployments require complete observability solution on top of monitoring – logging & tracing
  - Microservice based deployments would require decentralizing metrics endpoints across clusters to consume observability data & take localized decisions
  - Streaming telemetry would help in localized analytics to feed the decentralized closed loop implementations
  - Distributed monitoring & tracing methodologies would help ensure slicing requirements are met across microservice deployment

Get involved!

Please reach out to any of the EAP members or email 5GRM-eap@ieee.org

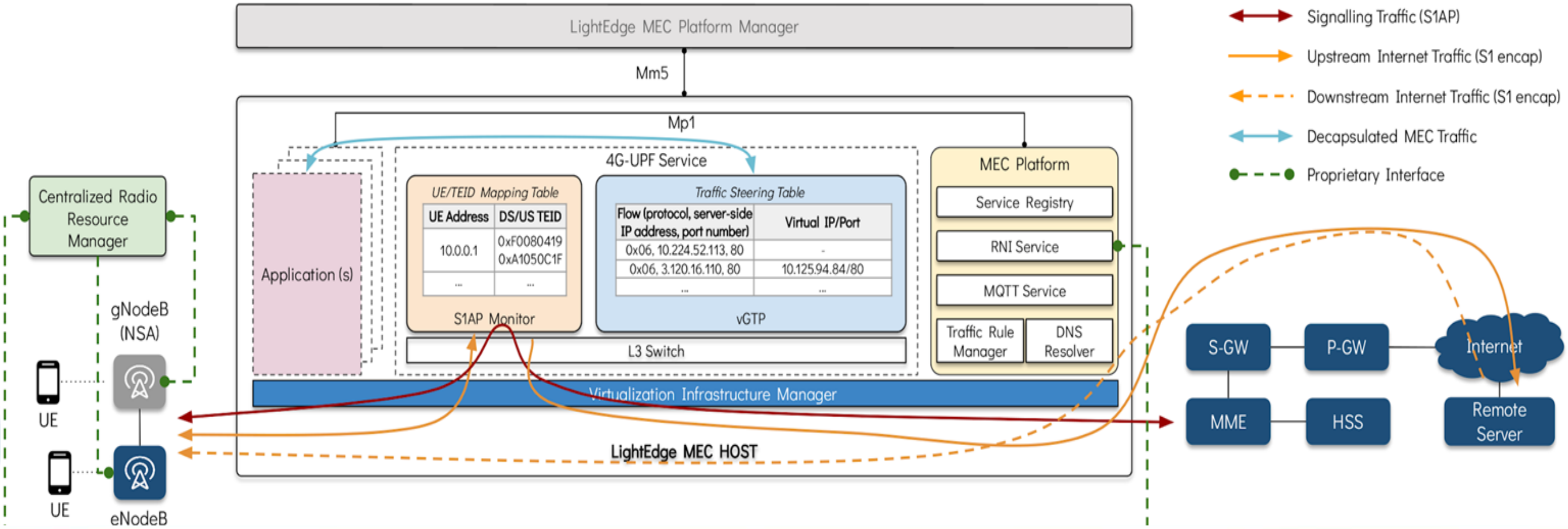
QUESTIONS?

Oct-9 - Light Edge (4G-5G) Estefania/Roberto/Somesh

- Higher:
  - Higher frequencies means smaller cells and thus more handovers.
  - If MEC is at the cell site -> a handover could trigger also a service migration. This may not always be the case because cell sites may share a backhaul which could make the service migration unnecessary (to some extent)
  - This is not an issue if MEC is at the aggregation site or the central office (to some extent since at some point also the best aggregation site for a service could change).
- Broader:
  - Marginally related to edge.
  - More bitrate could imply the necessity to have more computational power at the MEC site for some very specific use cases.
  - Could be a challenge if MEC is deployed at the cell site because many powerful MEC servers could be needed in this case
- Faster:
  - This could have a significant impact on data plane of the edge segment.
  - Since the radio now has latencies below 1ms the delay introduced by the edge data plane and the edge computing starts to become comparable to the access latency
  - Computational power at the cell site could also become a bottleneck (the end-to-end latency includes transmission and processing)
  - To achieve maximum throughput and ultra low latency the MEC is to be located at the cell sites along with the UPF closer to the end device.
- Overall the location of the MEC (deployment site - cell or aggregation or central office) and the use case drives the above tehcnology options.

LightEdge Reference System Architecture

LightEdge is a lightweight ETSI-compliant MEC solution for 4G and 5G networks, aiming at immediately bringing the advantages of edge computing to current 4G users and enabling a seamless transition from the 4G towards a full 5G architecture



UE/TEID Mapping Table

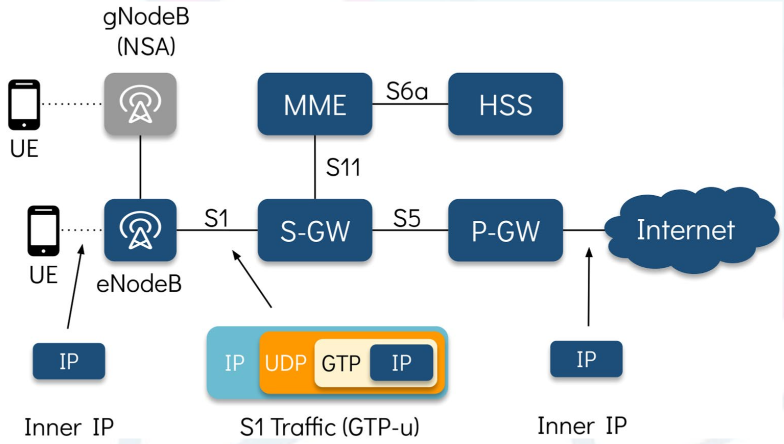
UE Address	DS/US TEID
10.0.0.1	0xF0080419 0xA1050C1F
...	...

Traffic Steering Table

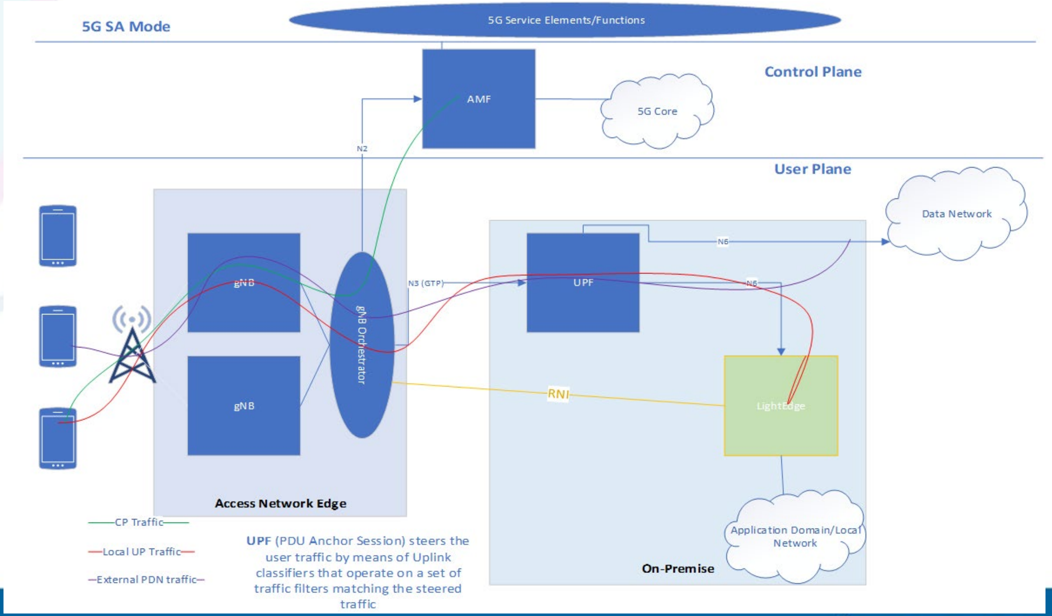
Flow (protocol, server-side IP address, port number)	Virtual IP/Port
0x06, 10.224.52.113, 80	-
0x06, 3.120.16.110, 80	10.125.94.84/80
...	...

LightEdge MEC support for 4G & 5G

Heterogeneous 4G/5G network (Non-StandAlone) setup



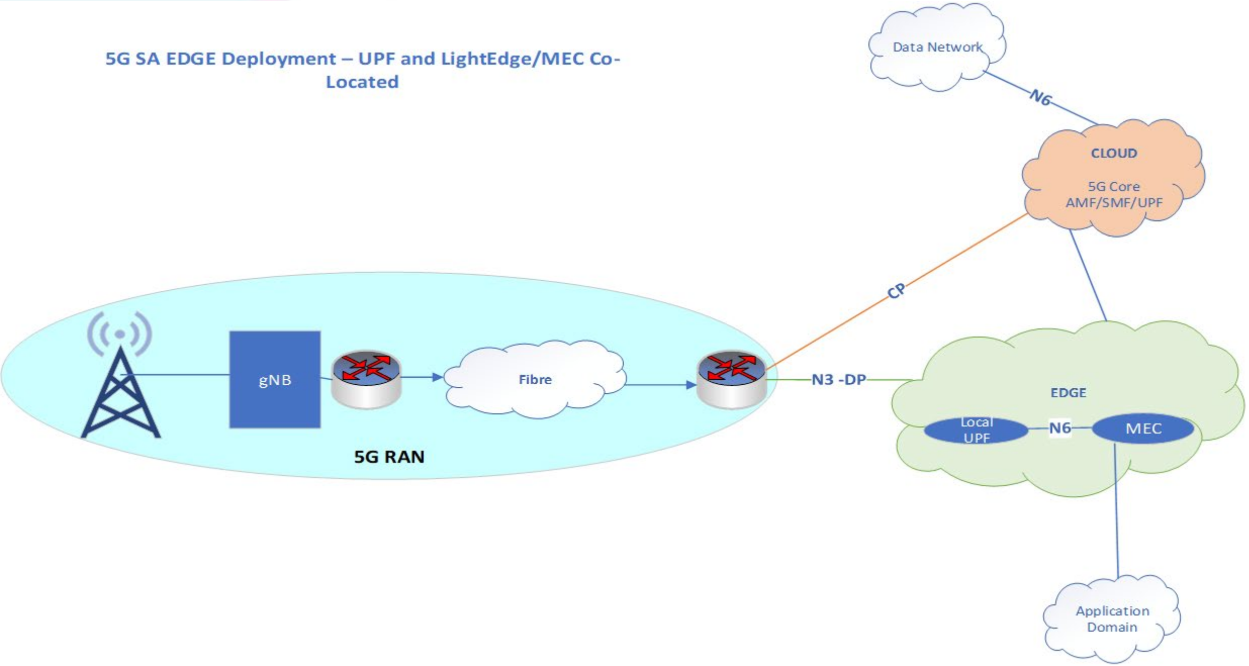
5G SA - CP/UP seperation/Uplink filter at SMF /UPF and MEC collocation / additional Role of PCF/AF in traffic classes



UPF (PDU Anchor Session) steers the user traffic by means of Uplink classifiers that operate on a set of traffic filters matching the steered traffic

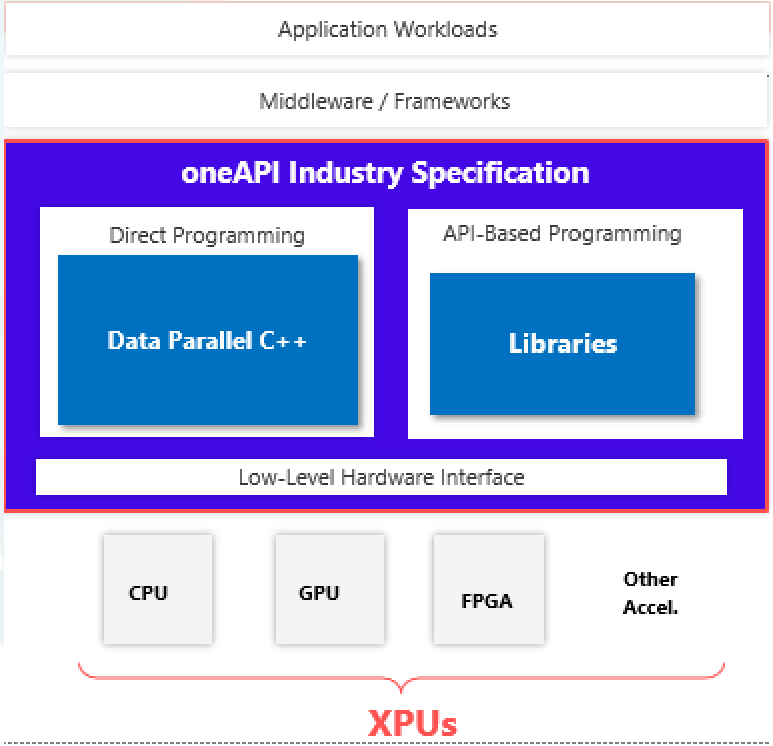
5G SA - Edge Deployment

UPF & LightEdge -MEC Co-located UPF routing local to colataed MEC & Applications



5G SA EDGE Deployment – UPF and LightEdge/MEC Co-Located

Heterogenous Programming Support-oneAPI



Application Workloads

Middleware / Frameworks

oneAPI Industry Specification

Direct Programming

Data Parallel C++

API-Based Programming

Libraries

Low-Level Hardware Interface

CPU

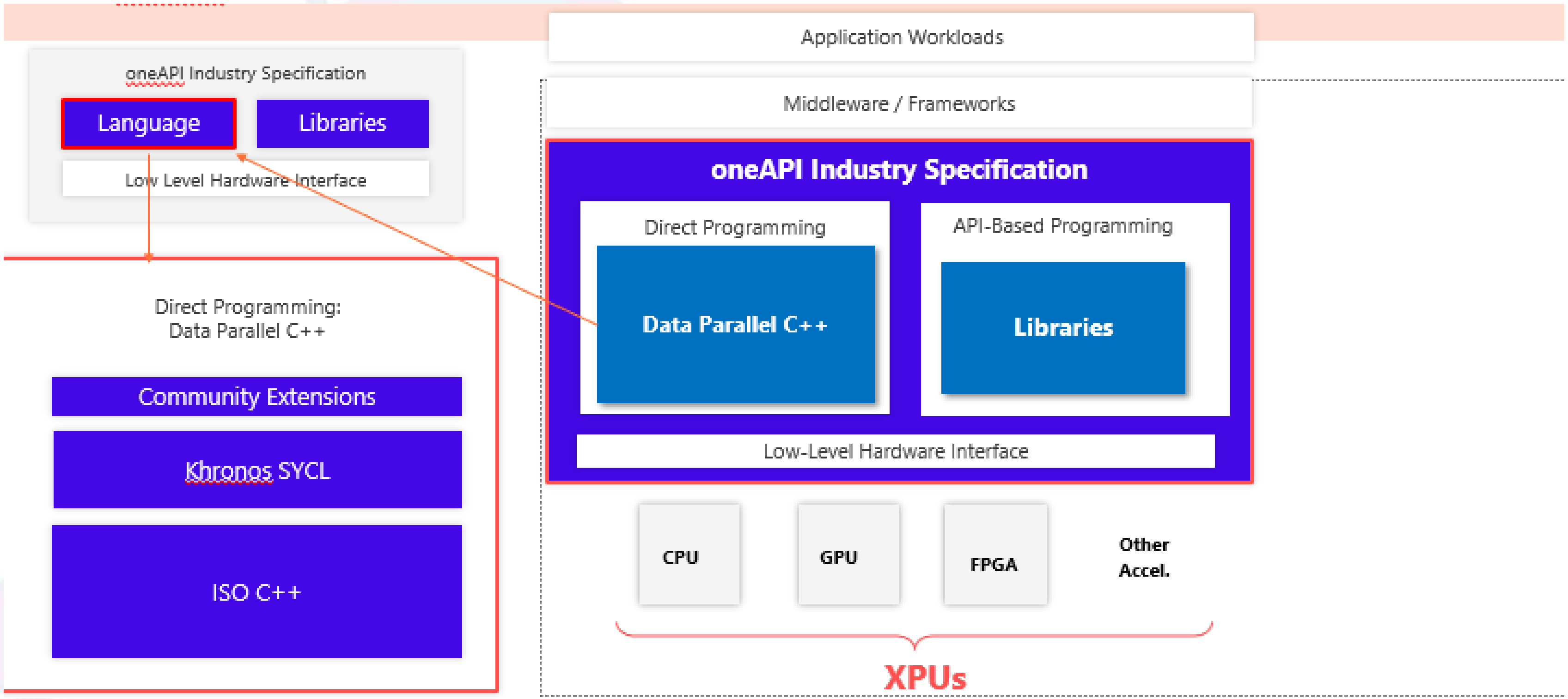
GPU

FPGA

Other Accel.

XPUs

Heterogenous Programming Support- DPC++



oneAPI Industry Specification

Language

Libraries

Low Level Hardware Interface

Direct Programming:
Data Parallel C++

Community Extensions

Khronos SYCL

ISO C++

Application Workloads

Middleware / Frameworks

oneAPI Industry Specification

Direct Programming

Data Parallel C++

API-Based Programming

Libraries

Low-Level Hardware Interface

CPU

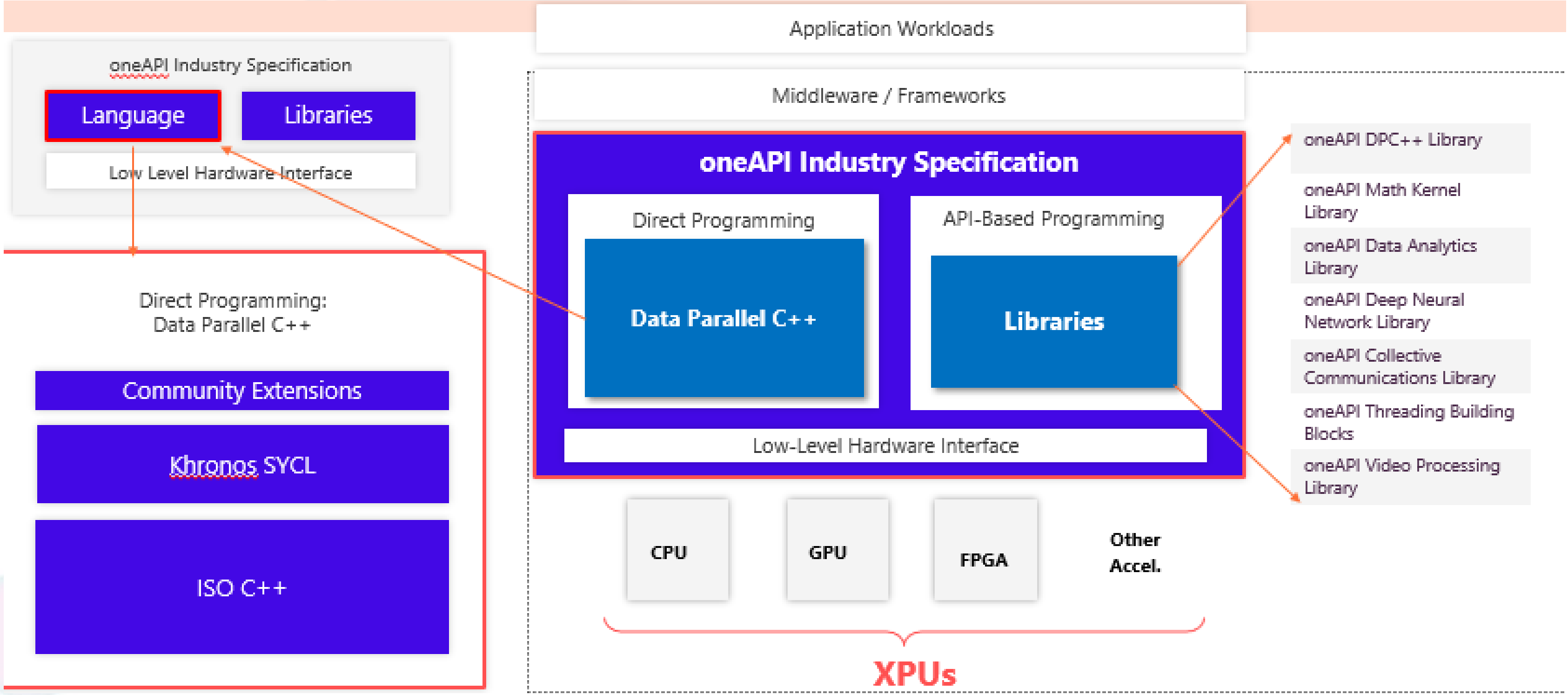
GPU

FPGA

Other
Accel.

XPUs

Heterogenous Programming Support- Libraries



oneAPI Industry Specification

Language

Libraries

Low Level Hardware Interface

Direct Programming:
Data Parallel C++

Community Extensions

Khronos SYCL

ISO C++

Application Workloads

Middleware / Frameworks

oneAPI Industry Specification

Direct Programming

Data Parallel C++

API-Based Programming

Libraries

Low-Level Hardware Interface

CPU

GPU

FPGA

Other Accel.

XPUs

- oneAPI DPC++ Library
- oneAPI Math Kernel Library
- oneAPI Data Analytics Library
- oneAPI Deep Neural Network Library
- oneAPI Collective Communications Library
- oneAPI Threading Building Blocks
- oneAPI Video Processing Library

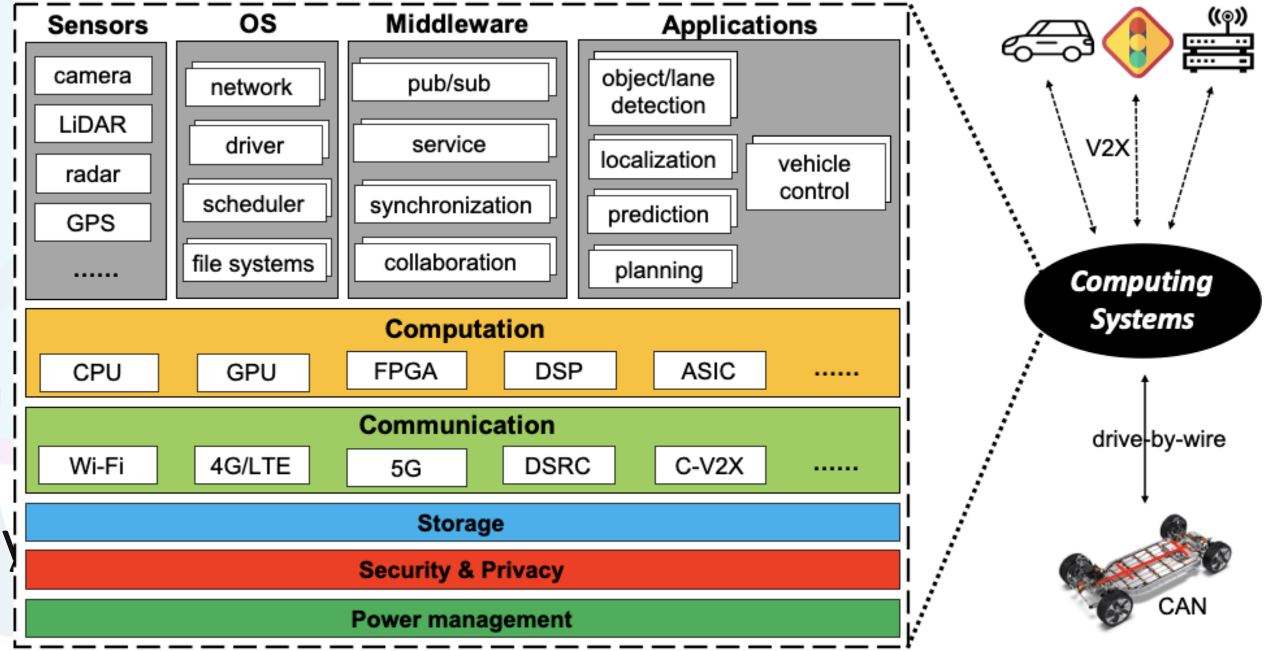
Autonomus Vehicle Edge Computing

End-to-End

Mostly for Learning

Modular with focus

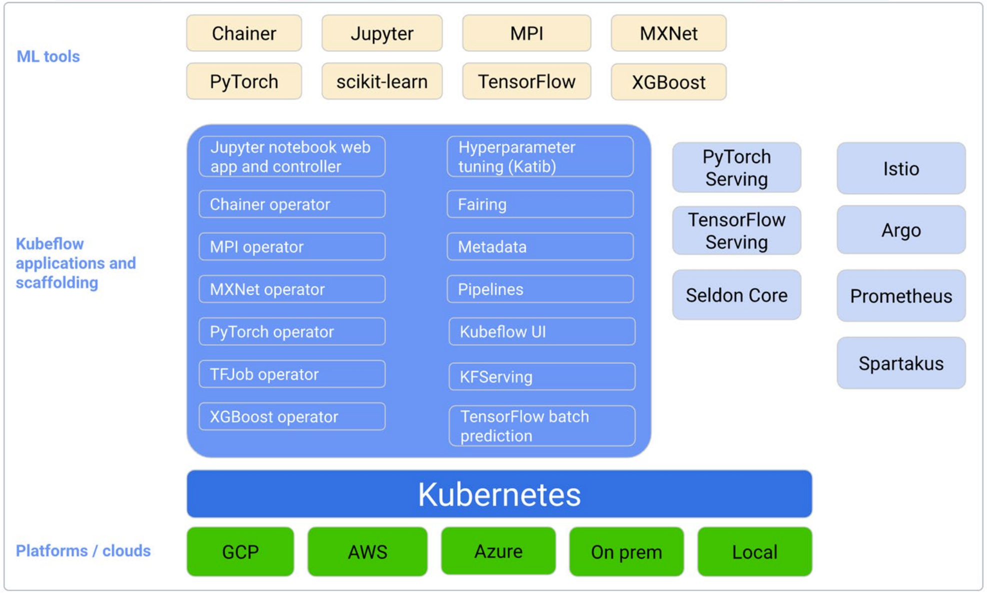
- Computation
- Communication
- Storage
- Security&Privacy
- Power Mgmt



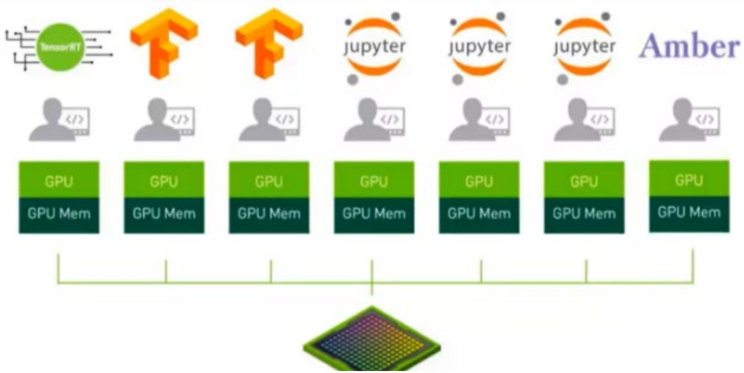
Model based on Sensor Data & Uses Edge Acceleration, All V2X/V2V apps use Middleware & OS to leverage resources

Edge Platform and AIML workload Distribution

KubeFlow Platform

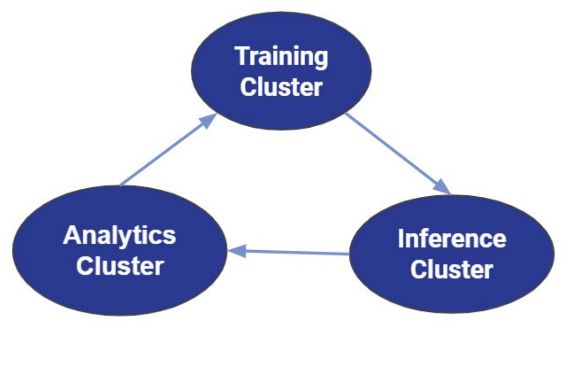


Non-optimal, Complex to Manage, Unpredictable to Scale



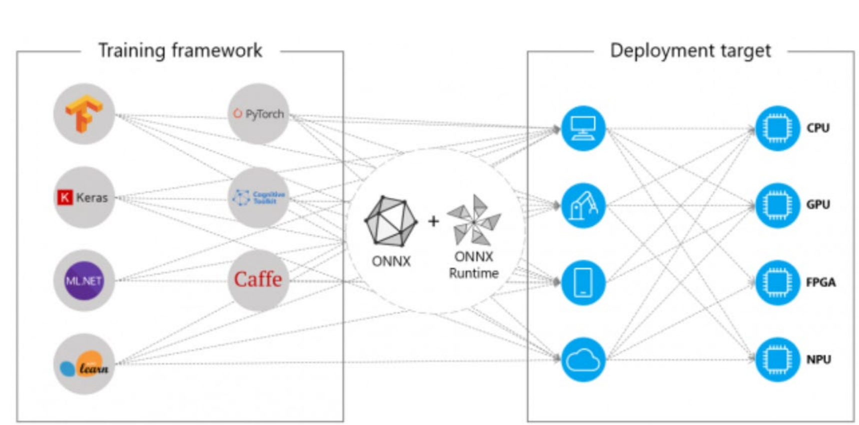
Overcoming Constraints of Edge & ML

Distribute Computation



- Training Cluster –Cloud (prefer high computes)
- Production Cluster –Edge (Prefer Inferencing and pretrained models)
- Analytics Cluster – Local/Global (Prefer time series aggregations)

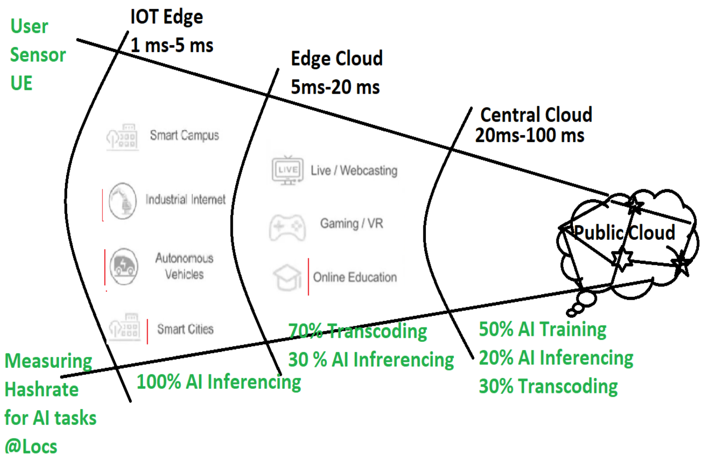
Interop, portability & Device Acceleration



Open Neural Networking Exchange (ONNX) allows use of appropriate Model , Tools and Framework at right location and allow separation of concerns for optimal use of AIML model for both Training in Cloud and Inferencing at the Edge

Infra for Deep Learning Evaluation (latency vs Proximity)

- cloud and Cloud application resources are at **latency** range of 20ms -100 ms in Massive Data Centers.
- Edge can be closer to user device or App clients depending on edge location plus constraints on resource.
- Edge Cloud in 5-20 ms range can support Applications like Online Education, AR/VR and Live Webcasting
- Even closer at IOT Edge with 1ms-5ms one can use for Autonomous Vehicles, Smart City applications etc.
- Compute Power of Crypto networks measured as Hashrate (usually used in Crypto calculations)can be applied to different Edge locations as shown.
- Measuring Compute Intensity is difficult considering the use cases.
- Example for Object recognition the Pruning & Fusing for Image Objects and Similarly Tiling and Vectorization for Graphs
- Goal is to improve Classification or Prediction Accuracy while reducing both millions of both computations and Hyper Parameters
- In this picture we are showing Hash rate at different edges showing the need for computational intensity optimally available to use for various processes in an ML pipeline for Edge delivery



Open Tracing at Edge & FCAPS Matrix + Logs

Closed Loop automation

Local - Edge Triggered

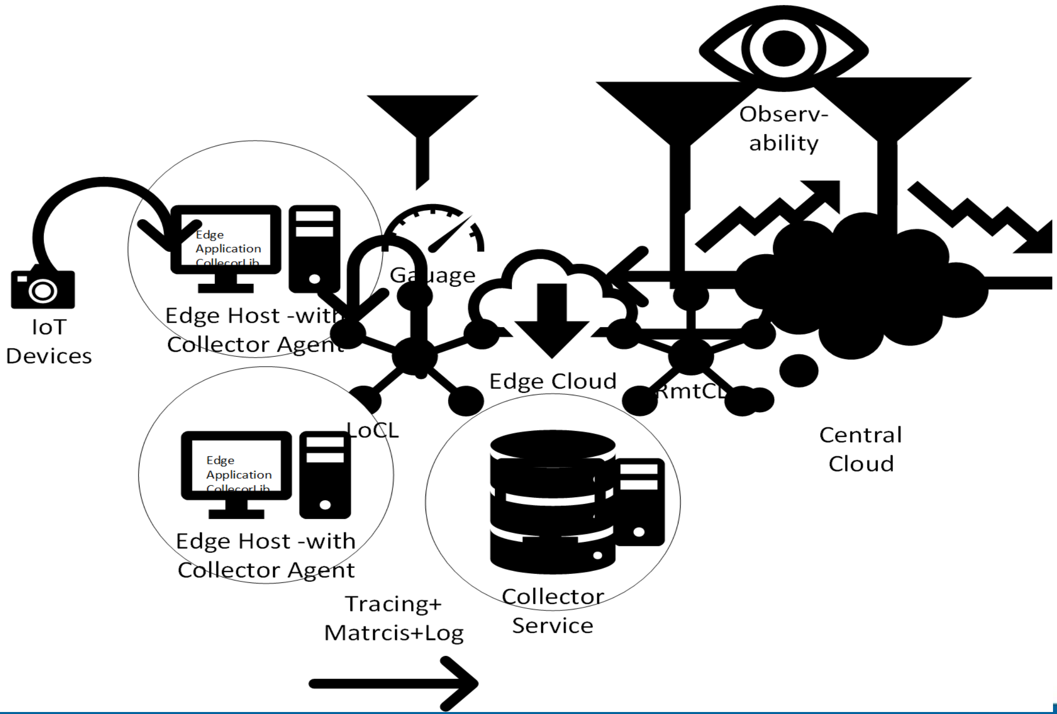
& Edge Managed (Faster)

(e.g.Anomaly Detection)

Global - Edge Triggered

& Cloud Managed(Slow)

Open Telemetry for Edge Nodes



Edge Cloud with Container as a Service (CaaS)

Edge as Service goes well with 5G Service Based Architecture

The same has been described as Edge Service Platform (ESP) in the earlier slides. Caching Service for Content Distribution has been used for long. However 4G/LTE Breakouts followed it and now the User Plane Function filtering and local routings to MEC & application Services at Edge between RAN Aggregation & any Front Haul Micro DCs or Cloud Provider Zones close to Cell site routers. CaaS is well suited and experiments in Tele-health and Autonomous driving are under lab testing UK.

Need #1, Standard Interfaces

Near-term Challenges: 2020-2023	*Description*
Multiple Edge Locations	Far Edge (Device, On Prem), Radio Edge, Near edge (Provider/network edge), Network Core
Multiple Standards & E Edge Stacks	Edge Service *Providers happy to point to MEC, CNTT, O-RAN, Open Edge Stack but prefer practically Kubernetes Automated Infra and* Open *Stack at edge with Vendor hardening*
MEC	MEC is not acceptable to Industry as hard to deploy. Radio API over MEC is relatively mature. %G O-RAN / OpenRAN have disrupted vRAN & C-RAN with Constrained Micro DC's
CNTT	Multiple CNTT architectures which are still evolving, RA1 is matures, RA2 is still evolving, plus OPNFV-CNTT merger is in offing
O-RAN	Multiple splits e.g. 7.2, 8, etc. Different standards apply to different use cases due to optimisation. Edge delivery at low latency needs more innovation.
Variety of Workload Requirements	Standards need to meet the requirements of different workloads HCI, AI/ML workloads and these may differ as per the geo requirements as well e.g. 5G Asia , EU, NA
Mid-term Challenges: 2024-2025	*Mapping between workload type and defining configurations to enable the functions required to execute certain popular use cases.*
Interoperability	Interoperability challenges amongst the emerging workloads
Long-term Challenges: 2026-2030	*New Market shifts in the Edge and last mile delivery due to availability of 5G & Fiber to the premise*
Integration of new technology	Emergence of newer Quantum computing shifting the edge application platforms

Need #2, Automation and Orchestration of Edge Platform

Near-term Challenges: 2020-2023	*Can the devices and platforms being used for Edge be programmed automatically?*
Device	Evolving devices for Edge (V2X, IIoT, Edge as a service etc)
Shared data	In Memory or shared memory processes e.g; GRPC
Device Discovery	Make the resources seen by an automated orchestration and scheduling framework so resource allocation can be effective
Monitoring	Health of devices, Bandwidth allocation and Utilisation, Performance indicators such as Latency and RTT (Round Trip Time)
Management	Various solutions to Solution Provisioning and Life Cycle Management including automation exist, e.g. Puppet, Ansible Playbook, OpenStack, Kubernetes, ad Kubernetes appears to lead the needs of end-to-end automation for edge using Operator framework besides using Custom Resource Definitions (CRDs)
Mid-term Challenges: 2024-2025	*Workloads, devices and standards will evolve based on Edge adoption*
Remote Provisioning and Management	Distributed deployment of edges means remote provisioning and management becomes more important as edge becomes more and more distributed
Long-term Challenges: 2026-2030	*Automating recognition of new types of devices and access technology*
Unknown devices and access technology will appear	EAP platforms should be able to handle any new type of devices and technologies as they become available instead of being reactive and adding support after the appearance, e.g. Quantum and related technology.

Need #3, Edge Automation Platform

Near-term Challenges: 2020-2023	*Description*
Cyber Physical Security	Addressing security for infrastructure & service heterogeneity; Hardware & Software adaptation for application specific security settings.
Security by design	Application specific security embedded with multi-interface involvement.
More challenges	Add more rows for each challenge
Mid-term Challenges: 2024-2025	*Description*
Multi-Layer Multi Modal automation requirement	Within a unified infrastructure usage, sensitivity and resource demand (and availability) dependent security algorithms
Long-term Challenges: 2026-2030	*Description*
AI powered centralized and decentralized accessibility	Within a unified infrastructure usage, sensitivity and resource demand (and availability), location & time dependent security.

Need #4 Support of Heterogeneous Platforms

Near-term Challenges: 2020-2023	*Description*
Multi vendor edge devices	Multiple vendor platforms in compute, storage and networking with programming options such as P4, SYCL, DPC++
Interoperability	Edge Apps and systems running on devices from multiple vendors should be able to interoperate
Management and Orchestration of Data Plane	Management of devices (which will be combination of number of system on a chip, connected with THz links) to be controlled from edge platforms
Mid-term Challenges: 2024-2025	*Description*
Just in time Compilation	workloads can run on available hardware and code can be compiled for that hardware at run time based on availability, even in a cloud native environment
Performance	Support of heterogeneous platform should not affect performance
Long-term Challenges: 2026-2030	*Description*
Support of TeraHetz Devices	Plug and play with THz link with local spectrum management from edge supporting Clock detection and synchronization for Device discovery and plugin to system

Need #5 Support of Hybrid Cloud (HC) - Edge Service Platform(ESP)

Near-term Challenges: 2020-2023	*Description*
Uniform Edge Service Platform	With Kubernetes!!! movement , Uniform Edge Service Platform is taking off with both Containers and VM as means using concept of mortal Pods.
Interoperability	With more vendors moving two Uniform ESP the goal the divergence and differentiation will appear as Hybrid Cloud (HC) ESP as evident from emerging Managed HC-ESP
End to Em]nd Management and Orchestration of Service	Intent based declarative Edge service with automated kubernetes and Operator framework for Hybrid Cloud will enable end to end slicing of workloads including transport. The HC-ESP will evolve and Software & Intelligence will rule.
Mid-term Challenges: 2024-2025	*Description*
Dynamic ESP	ESP will be provisioned and de-provisioned on need basis as automation becomes reliable with Open Service Brokers (OSB) leveraging the HC-ESP
Performance	
Long-term Challenges: 2026-2030	*Description*
Support Ultra low latency	Achieving low latencies will emerge as RTOS & parallel computing with AI deep learning (DL) take root at the edge beyond simple ML.

Today’s Landscape

- ETSI MEC- Is the standard Multi Access Edge Computing Reference Architecture
- OpenStack Edge Group- Is the deployment reference based on Hybrid, VMs and Containers
- LFN ONAP and OPNFV-CNTT is focused on VMs and Containers RA/RI/RC for Telco edge cloud platforms
- O-RAN- They are focused on mobile edge services for real time applications
- TIP/TUG/OpenRAN -Telecom Infra Project and Telecome users group are focused on migrating VM to container based Multi-Access edge/ platform deployment
- CNCF- The application of microservices in using “app containers” along with “OS containers” for edge services

Stakeholders

- ❖ Telco (MNO/MVNO)
- ❖ Cloud Service Providers
- ❖ Infrastructure providers
- ❖ Consumers
- ❖ System Integrators / Equipment Providers /Vendors
- ❖ Enterprise/Government/Cities
- ❖ App developers
- ❖ Education