one6G Open Lectures #7
In-Network Computing and Intelligent User Plane for 6G

In-network service-aware computing with hardware acceleration in 6G networks

21st March 2024

NTT Network Service System Laboratories
Hiroki Baba

Bio

Hiroki Baba

Senior research engineer at Network Service System Laboratories (NS-Lab) in NTT.

<Background>

Network architecture for more than 10 years.

- NFV commercial introduction at DOCOMO.
- Network slicing study and PoC in the MEF forum and ETSI ZSM.
- Currently, studying 6G mobile network architecture, especially on the theme of in-network computing.
- B.E. and M.E. degrees from the Tokyo Institute of Technology in 2006 and 2008, respectively.

Outline

1. NTT network architecture study toward 6G :

   “Inclusive core”

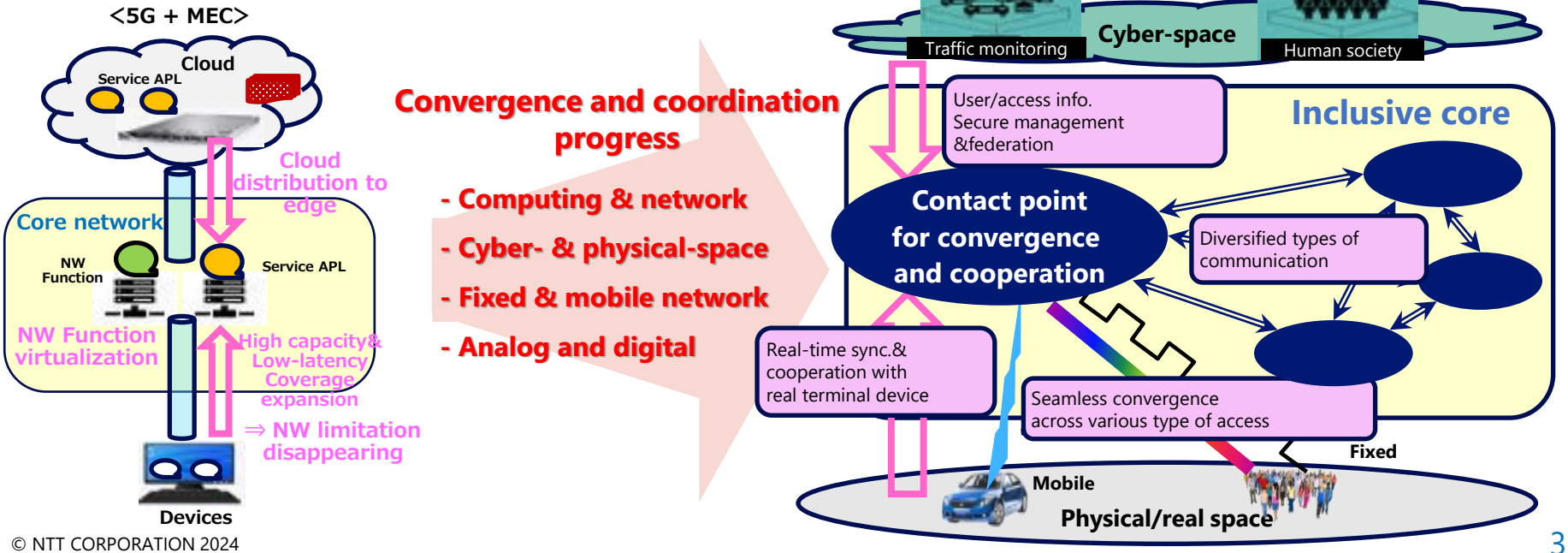
2. ISAP: In-network Service Acceleration Platform

3. ISAP related PoC activities

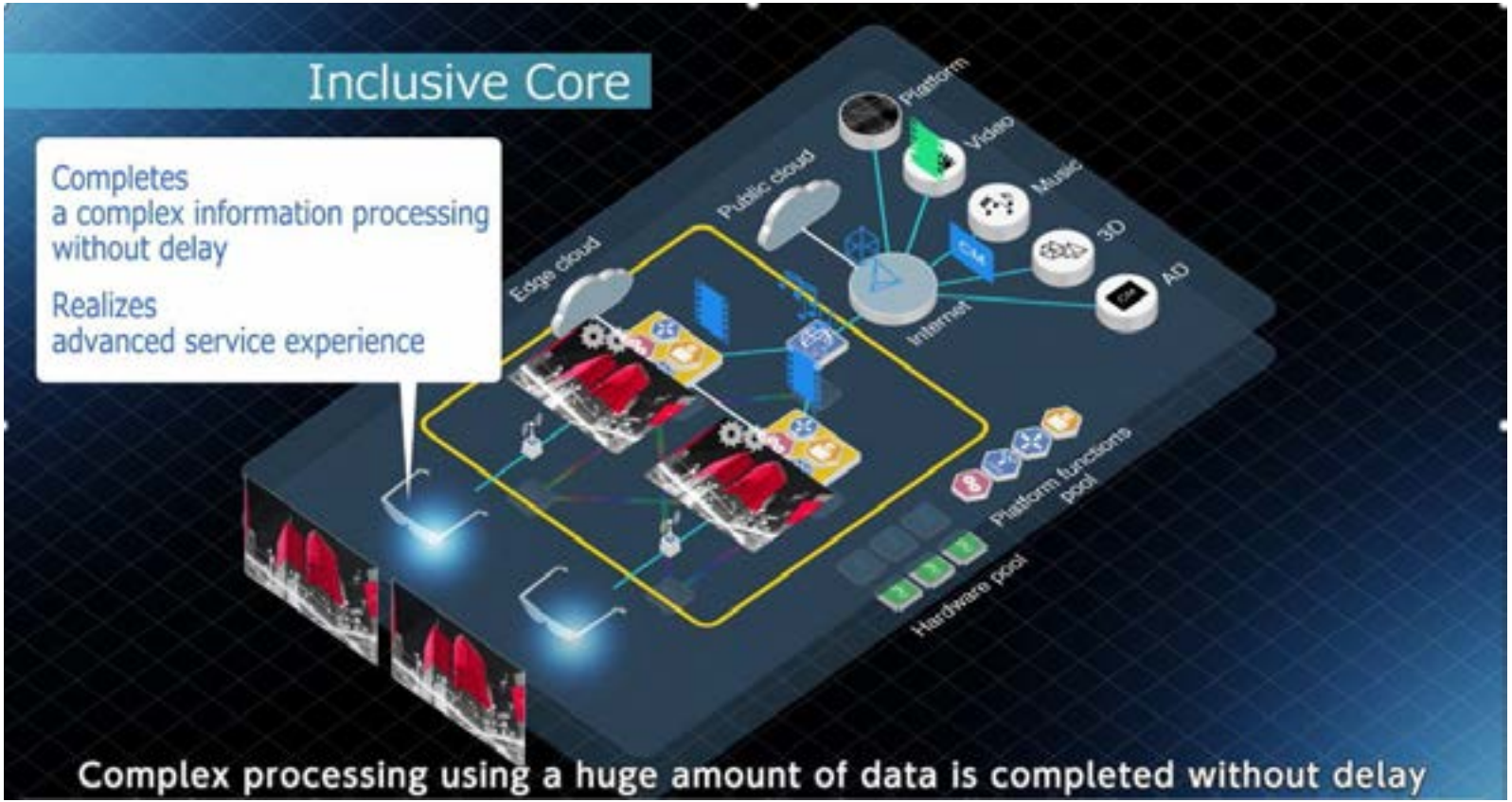
4. Concluding remarks

NTT's network concept: Inclusive Core

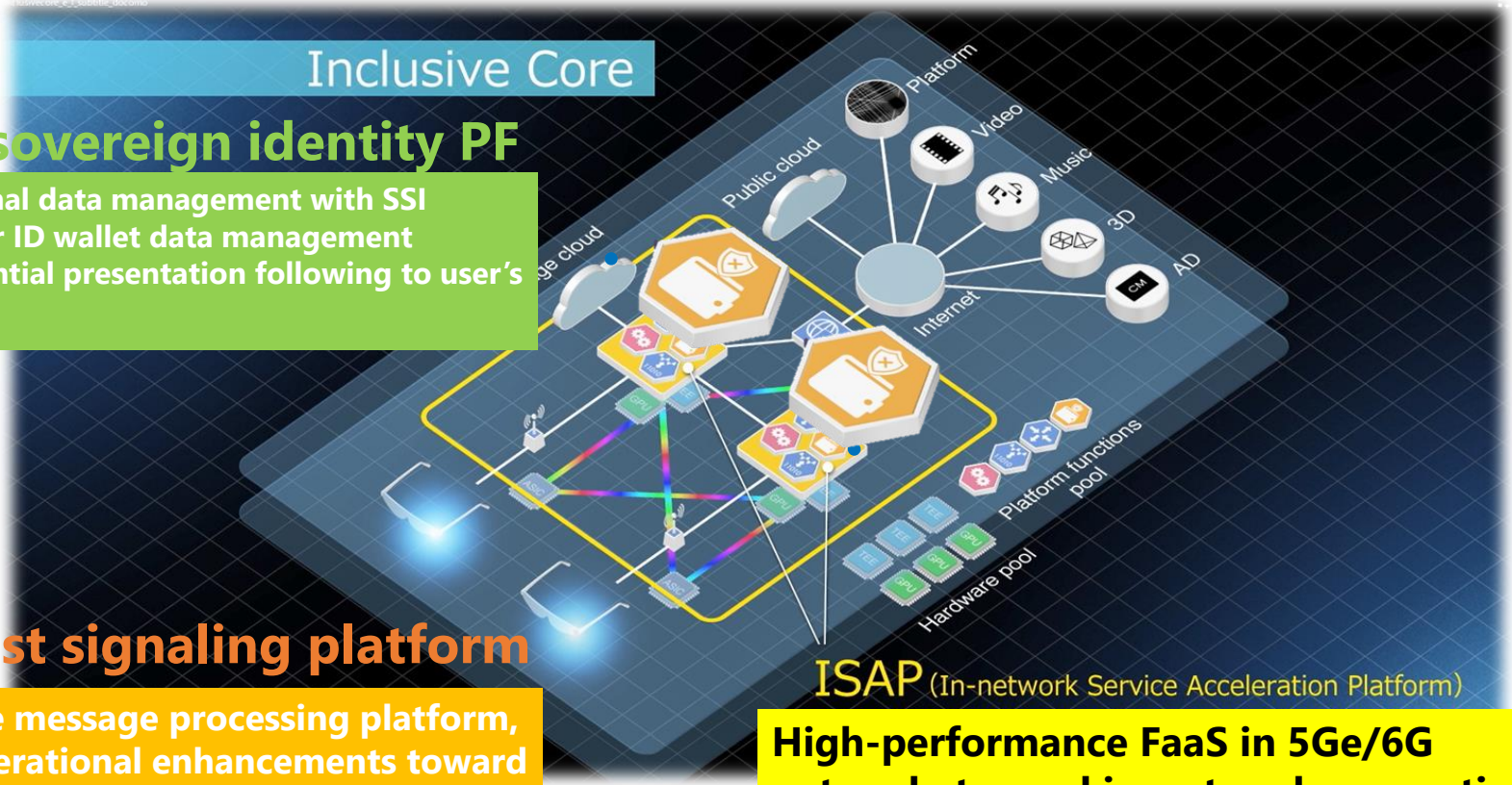
- **Multilateral convergence & cooperation outside/inside of network in progress**
- **"Inclusive Core" stands for "Core network for convergence and cooperation"**:
  provides means of communication between cyberspace/real space and across terminals/locations,
  and seamless communication services independent of physical terminal configurations and access networks.



Inclusive core architecture concept



Inclusive core technical components



Self-sovereign identity PF

1. Personal data management with SSI
2. Carrier ID wallet data management
3. Credential presentation following to user's policy

Robust signaling platform

C-plane message processing platform, and operational enhancements toward robust mobile core

High-performance FaaS in 5Ge/6G network toward in-network computing (INC)

Ref: Inclusive Core whitepaper

https://www.rd.ntt/e/ns/inclusivecore.html

Inquiry: inclusive-core@ntt.com

Outline

1. NTT network architecture study toward 6G :

   “Inclusive core”

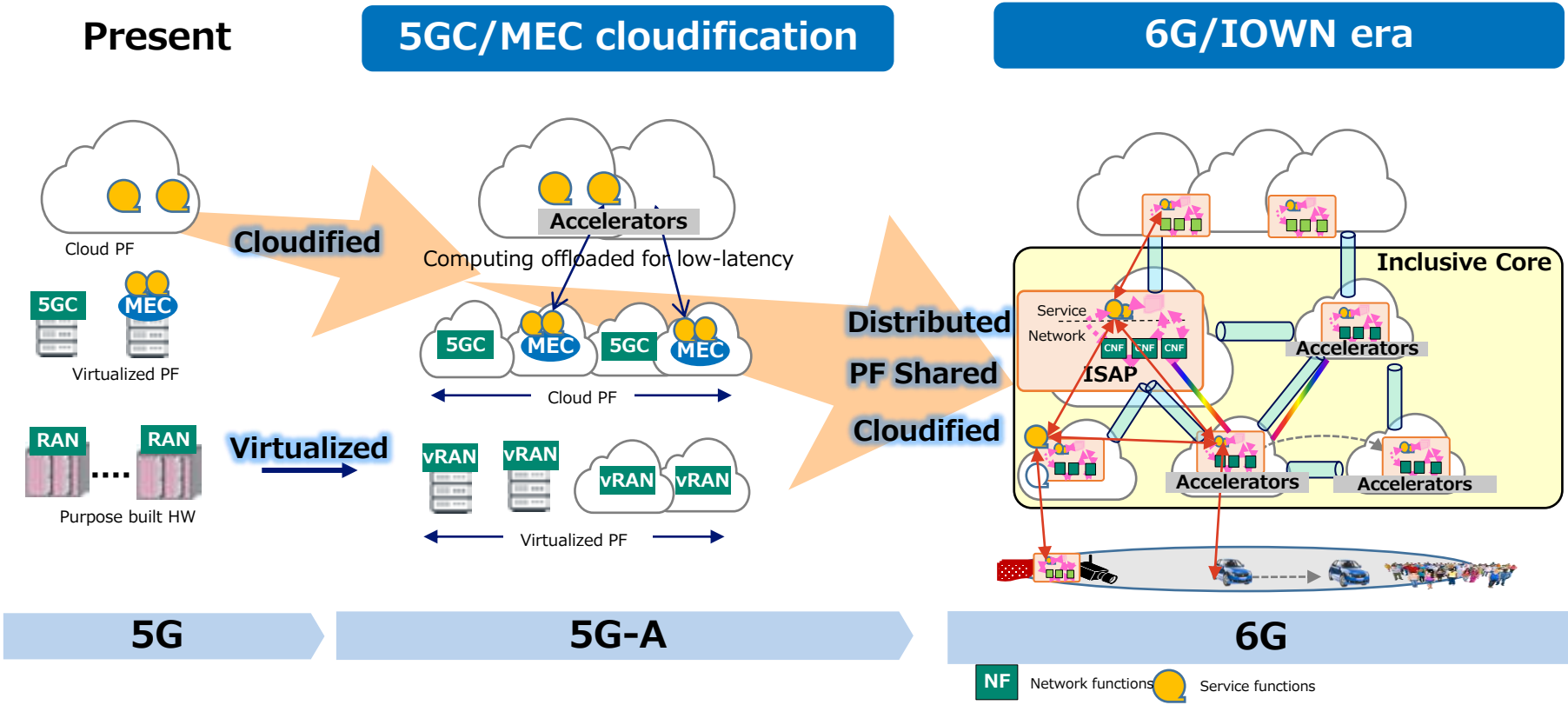
2. ISAP: In-network Service Acceleration Platform

3. ISAP related PoC activities

4. Concluding remarks

Computing & Network Convergence

Cloudified RAN will be widely deployed towards 5G-A/6G following 5GC/MEC infrastructure



Mobile NW and computing PF integration

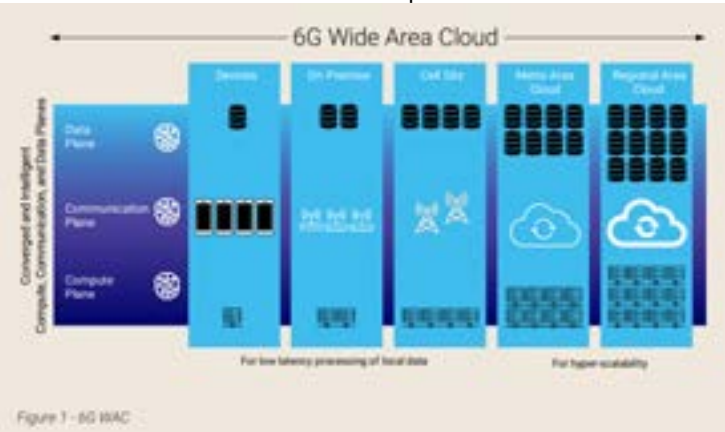
Recommendation ITU-R M.2160-0 “Framework and overall objectives of the future development of IMT for 2030 and beyond”

2.2.2 Ubiquitous computing

In addition to ubiquitous intelligence, it is expected that ubiquitous use of data computing resources would also expand throughout the IMT-2030. Emerging trends in this regard include expansion of data processing in the network infrastructure to the network cloud and devices that are closer to the origin of the data and support for proliferation of ubiquitous intelligence throughout the IMT-2030. This trend also contributes to the improvements for applications requiring real-time responses and data transport. Ubiquitous computing disseminated across IMT-2030 is expected to enable efficient utilization of resources and optimal placement of workloads, as well as scales and manages the infrastructure to run the applications.

6G wide area cloud

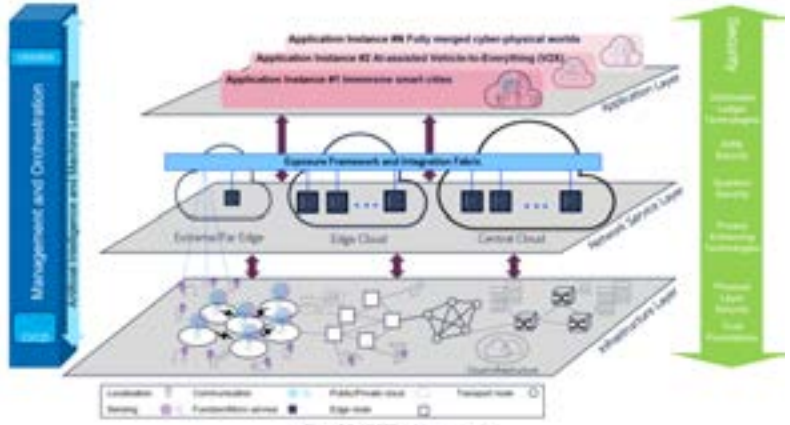
- Computing and data services in addition to communicationo
- Service dedicated transport and function chain



Next G Alliance Report: 6G Technologies for Wide-Area Cloud Evolution
https://nextgalliance.org/white_papers/6g-technologies-for-wide-area-cloud-evolution/

CaaS(Compute-as-a-Service)

- Compute service as 6G NW architecture
- Heterogeneous computing resource (e.g., CPU, GPU, FPGA 等) abstraction interface available

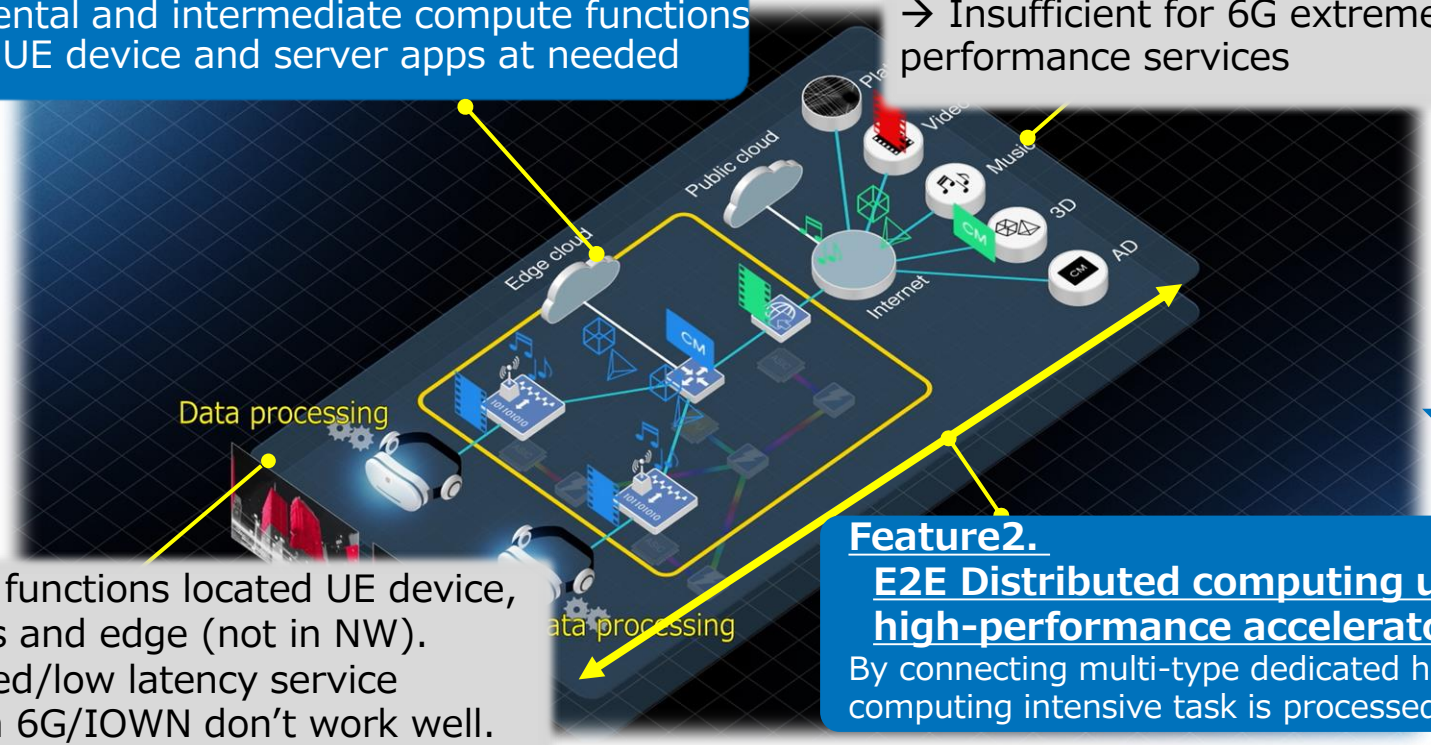


Hexa-X Deliverable D1.3, “Targets and requirements for 6G - initial E2E architecture”
https://hexa-x.eu/wp-content/uploads/2022/03/Hexa-X_D1.3.pdf

ISAP: target features and issues to be solved

Feature1.
Network and computing integrated platform
Complemental and intermediate compute functions in NW for UE device and server apps at needed

Compute resource shared by multiple-software functions (incl. NW and service)
→ Insufficient for 6G extremely high-performance services



Computing functions located UE device, app servers and edge (not in NW).
→High speed/low latency service expected in 6G/IOWN don't work well.

Feature2.
E2E Distributed computing using high-performance accelerators
By connecting multi-type dedicated hardware computing intensive task is processed immediately

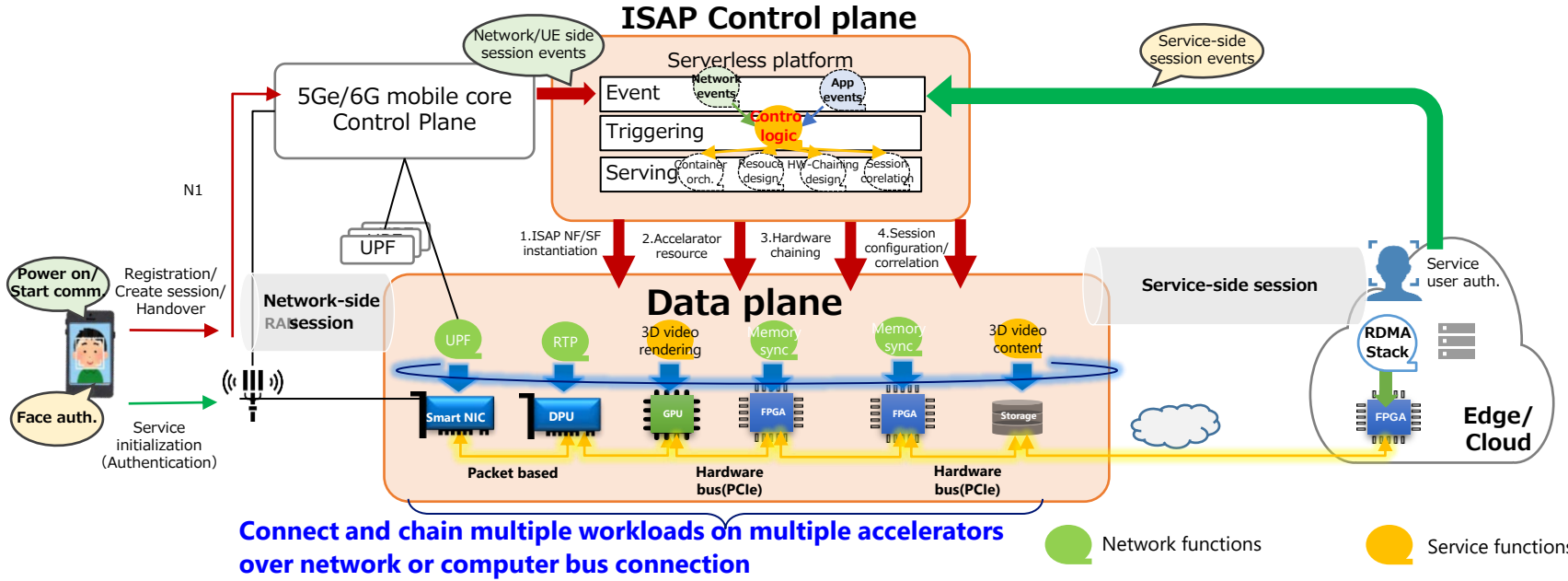
ISAP: Architecture

a. Control plane

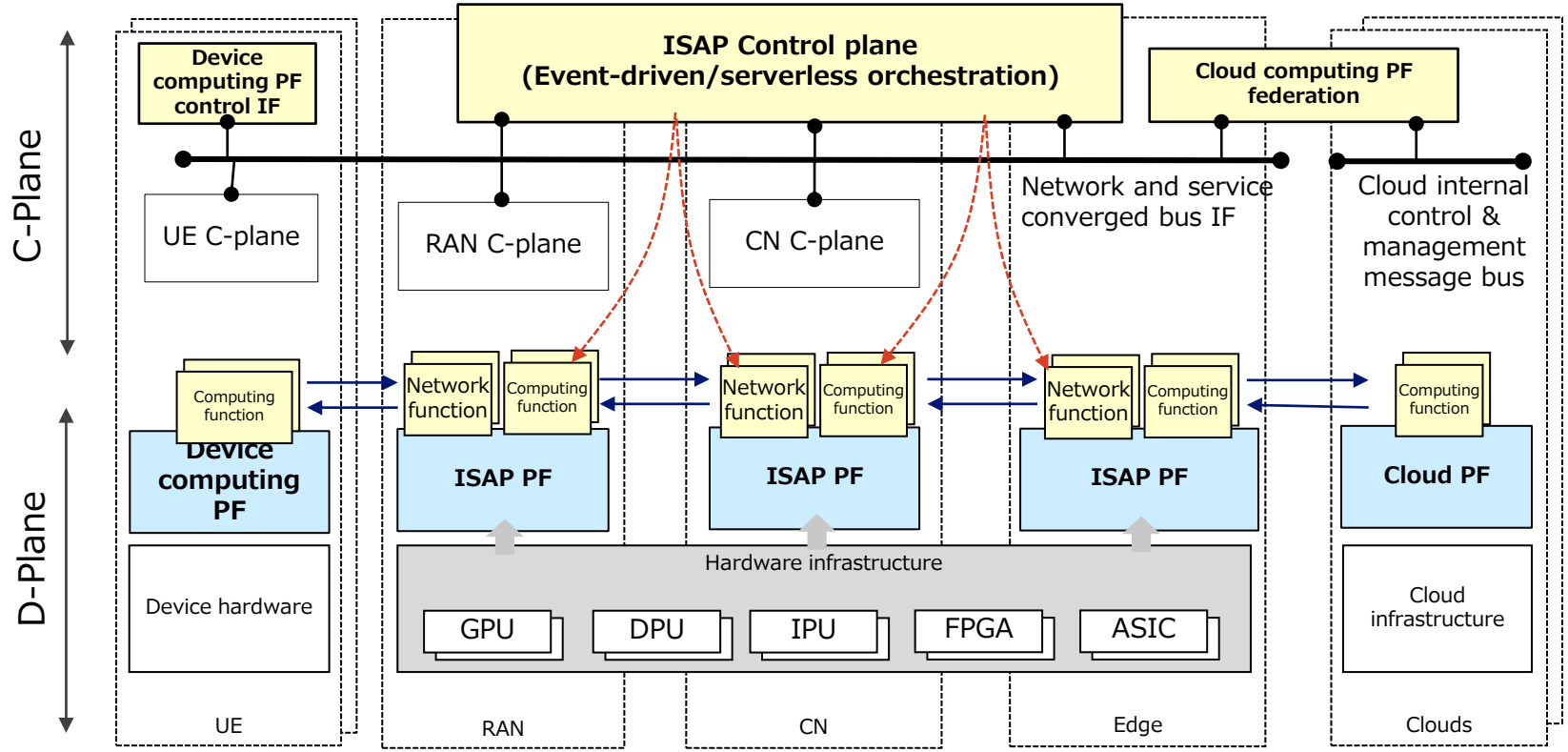
Federate in-network computing with network and services to high-performance CaaS for each users sessions

b. Data plane

HW-accelerated network and application function chaining



Ref: Architecture diagram



a. Service and network convergence

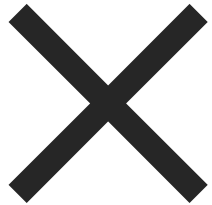
Network control linked with service event in addition to network events realizes network configuration suitable for service properties

e.g. Best fit data path for specific service being set up only when its communication duration

<Network event info.>

- Network registrations/auth
- Location updates
- Communication tunnel setup/delete
- Handovers
- VPN setup/delete
- Packet loss rates
- Data transfer delay

Etc.



<Service event info.>

- Service auth
- Content selection
- Video player start/stop
- Audio communication start/stop
- AI analsys start/stop
- Remote operation start/stop
- Inter-service transition

Etc.

a. Services and NW convergence control

Dedicated computing service provisioned by service property & requirements awareness

e.g. Metaverse APLs

Communication experience

Service space Transition

Immersive experience

Live experience

Service property

- Talk and chat based service
- Multiple users accesses common metaverse space

Service property change

- High resolution and highly composited 3D space
- Enjoy composited rich content

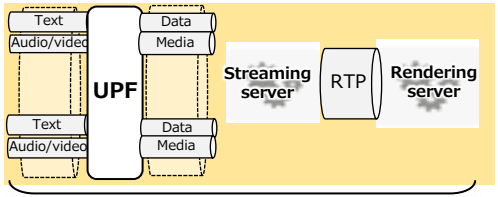
- High volume video and remote play
- Low-latency interaction

ISAP control mechanism

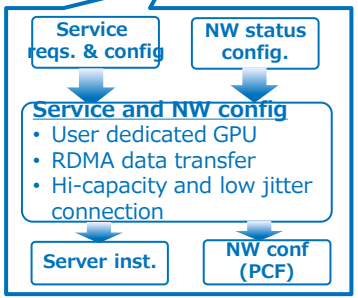
Service & network configuration

Configuration for communication centric service

E2E latency : <50ms, Normal H.O.



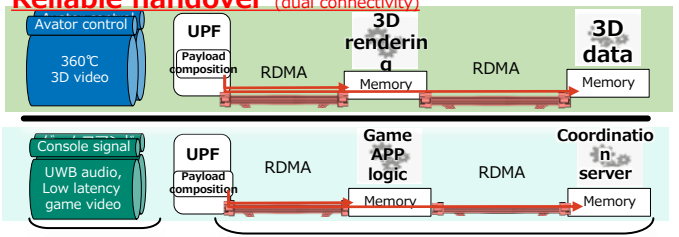
Best effort shared by multiple users



Configuration for Immersive/Live experience centric service

E2E latency: <5ms、deterministicity : 10Hz

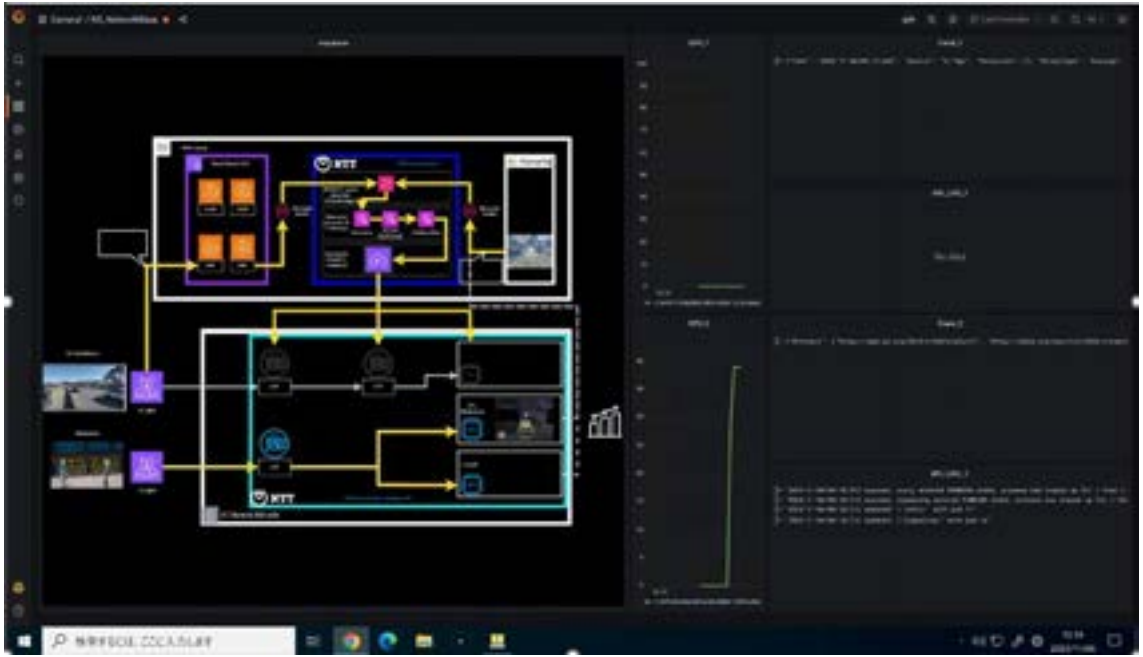
Reliable handover (dual connectivity)

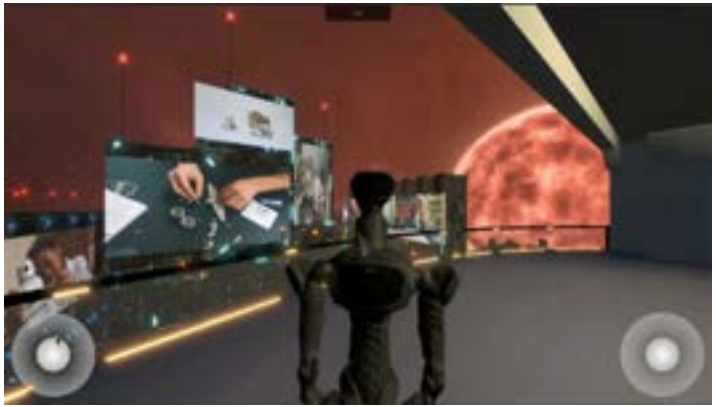


QoS enabled・Optimized route

Resource provisioning per user session basisi

a. Event driven orchestration of ISAP

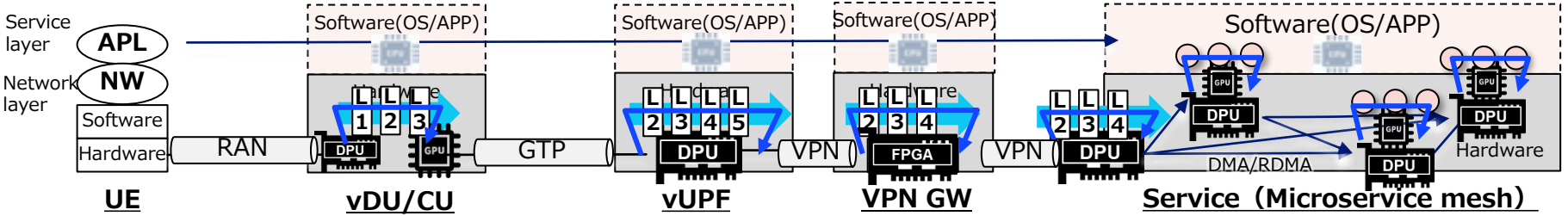




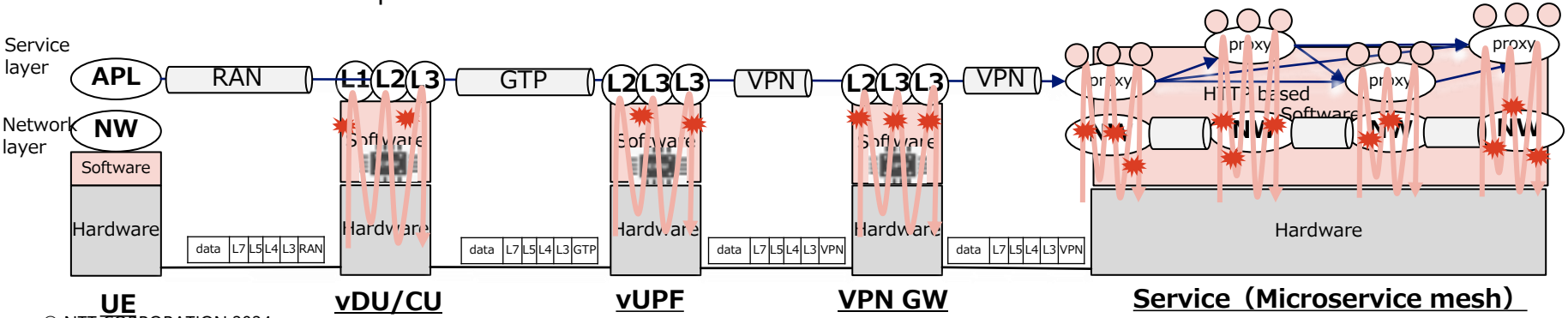
b. Accelerator chaining

ISAP user and computing

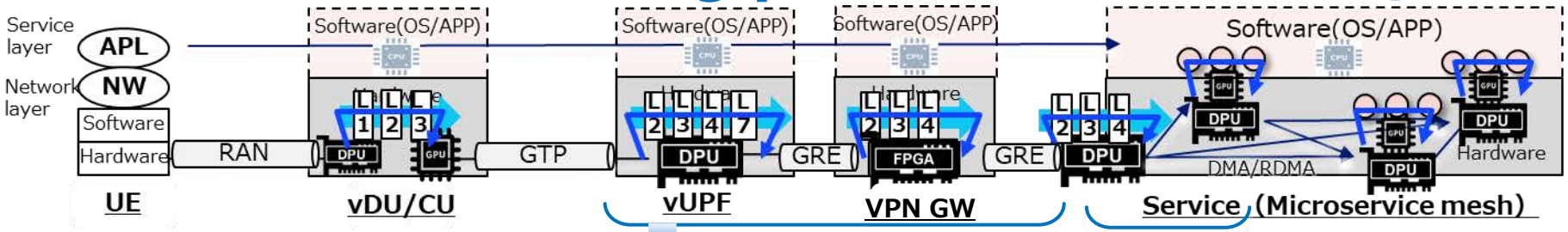
Both NW and service functions achieve deterministic performance through HW processing (no CPU intervention) and pipeline processing.



Conventional : Delays and fluctuations increase due to software processing and CPU-mediated interrupts in various parts of NW functions and service functions.

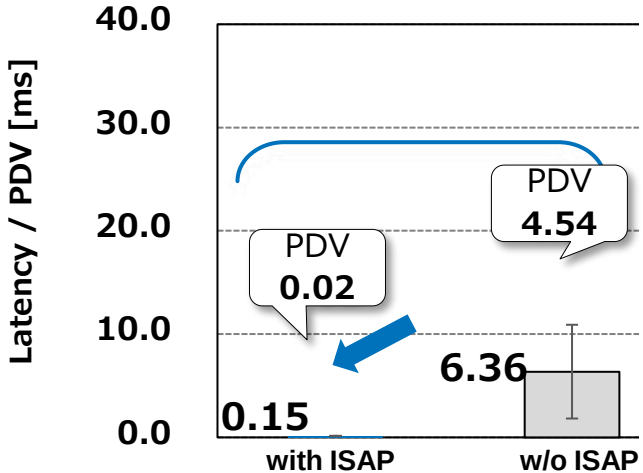


b. Accelerator chaining performance



Network related data transmission delay/PDV

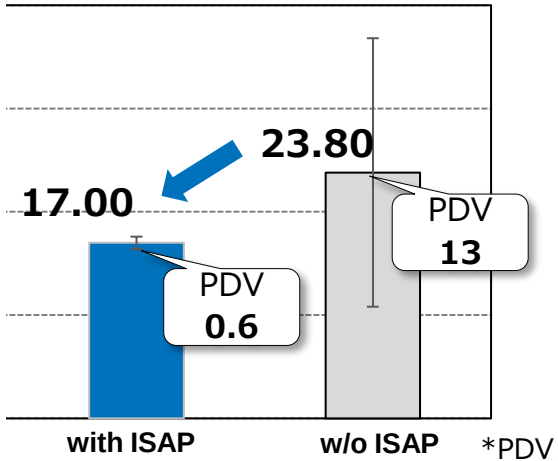
(Example of UPF on DPU and vGW on FPGA chain)



Communication delay/PDV
(Packet transfer)

Service related processing delay/PDV

(Example: Video data stream processing with AI analytics)



Computing processing delay/PDV
(AI video stream analysis)

*PDV: Packet Delay Variation

Outline

1. NTT network architecture study toward 6G :

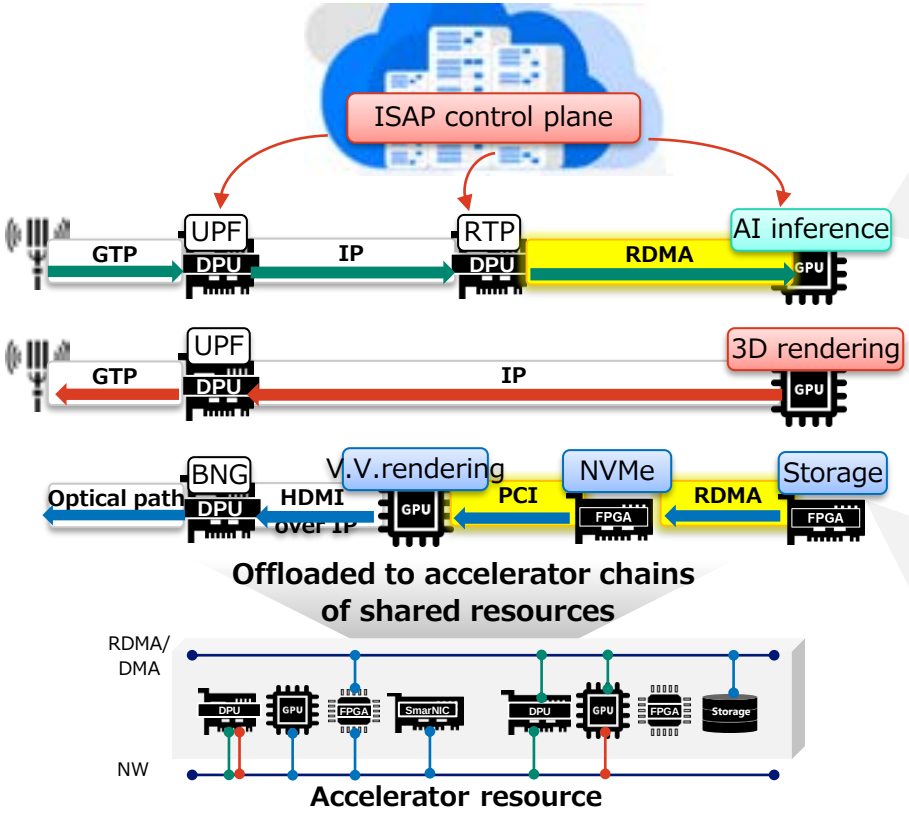
   “Inclusive core”

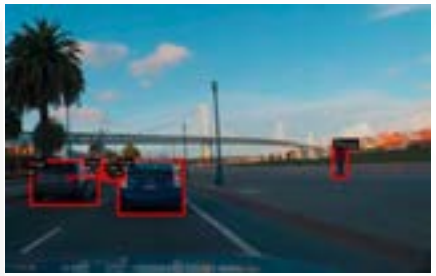
2. ISAP: In-network Service Acceleration Platform

3. ISAP related PoC activities

4. Concluding remarks

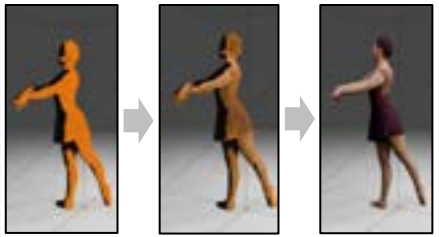
ISAP application example





Real-time AI video analytics

Metaverse cloud rendering & streaming

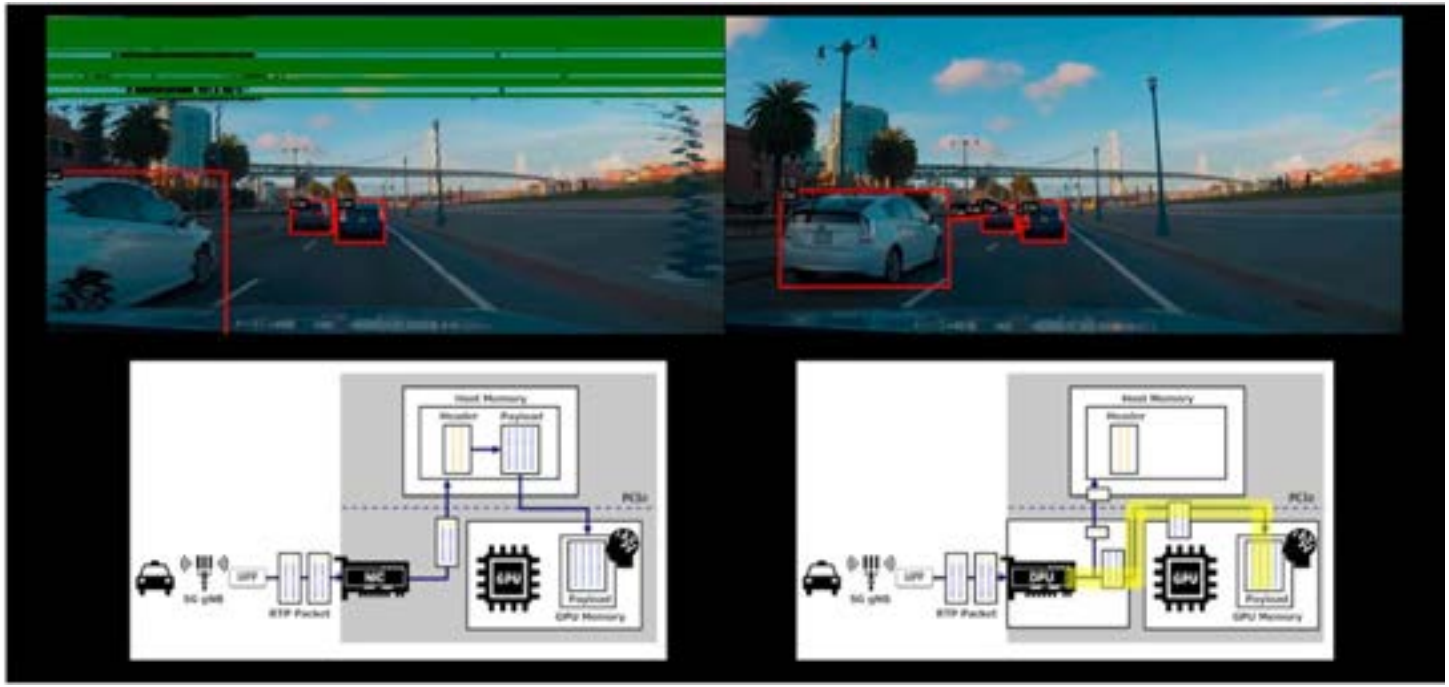


Volumetric video rendering & streaming

Ultra high-resolution real-time AI video analytics

NTT

Without ISAP hardware chaining **With ISAP hardware chaining**



4K, no-compressed video streams analysis using AI

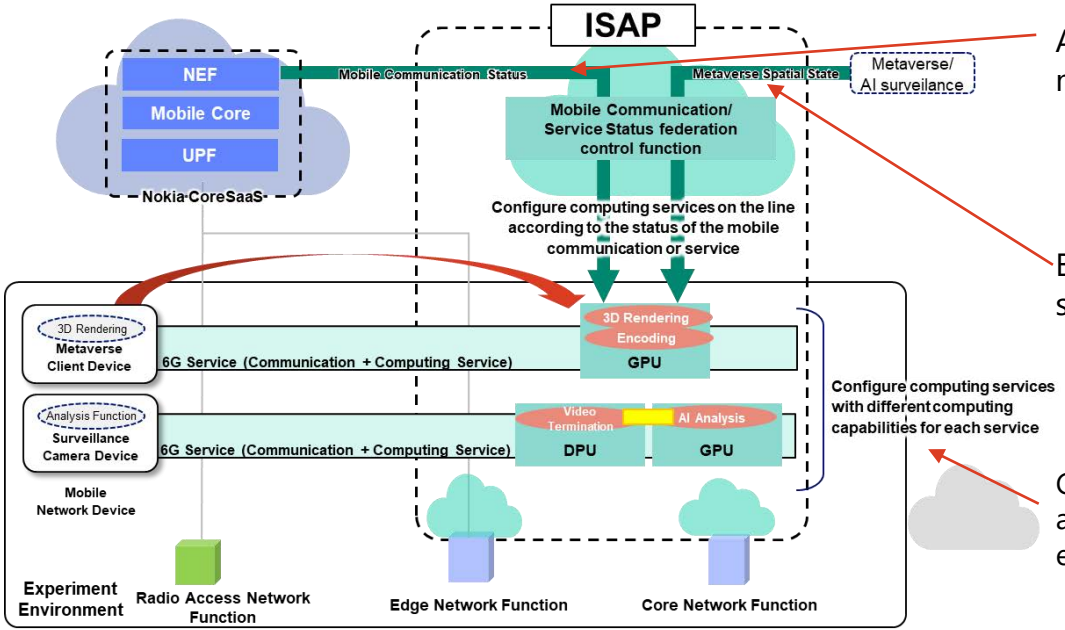
Metaverse 3D rendering application

ISAP coordinates 5G session status and metaverse app. session status, and deploy 3D rendering and streaming function on DPU and GPU in INC only when high resolution and composite required.

6G INC architecture and PoC with NOKIA

- **Joint arch. PoC of In-network computing(INC) for 6G with NOKIA**
- Successfully finished PoC and demonstration of INC at MWC:
  → Proved that ISAP and NOKIA CoreSaaS work well to control INC functions.

【PoC system architecture】



【Features confirmed through PoC】

A. Provide computing resources according to the mobile communication status of the device

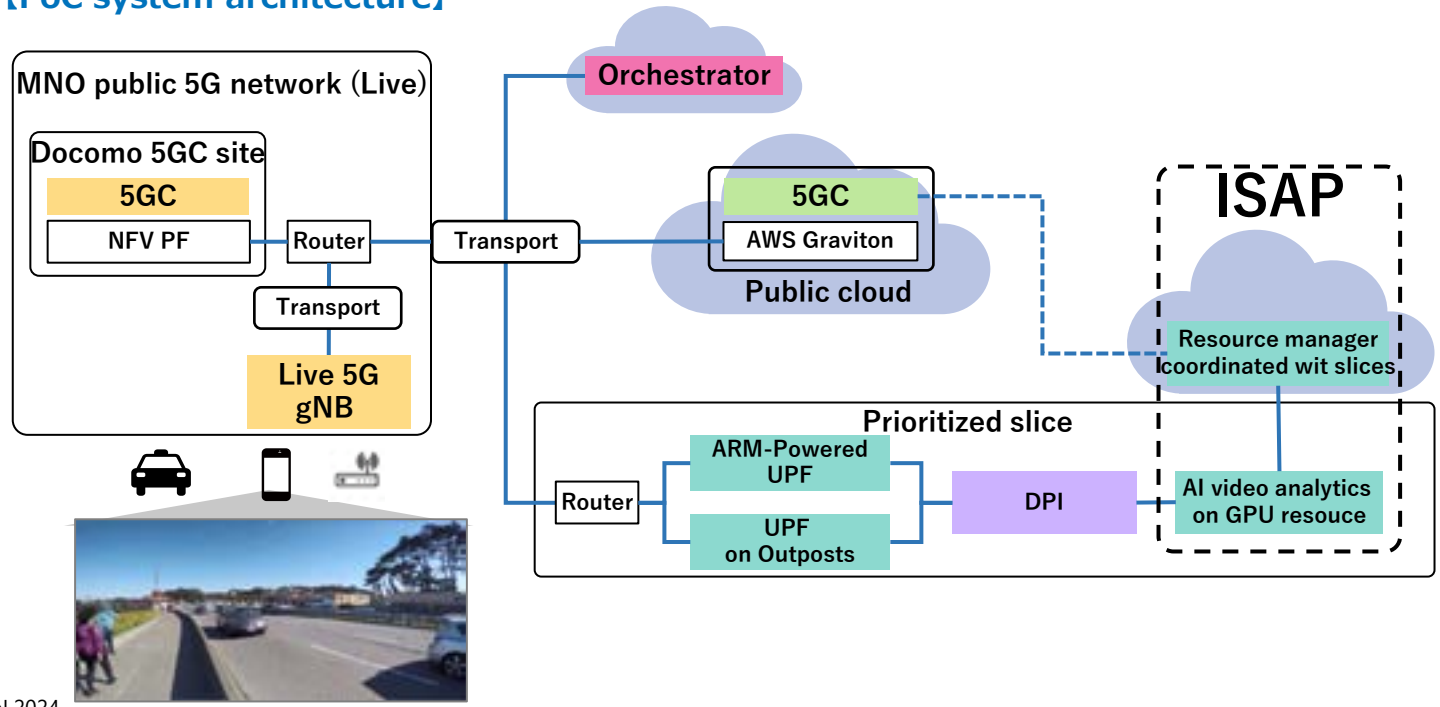
B. Provide computing resources according to the status changes in cloud service

C. Dynamic control of optimized computing resources according to the characteristics of mobile networks and each cloud service

Public 5G core slice PoC with ISAP

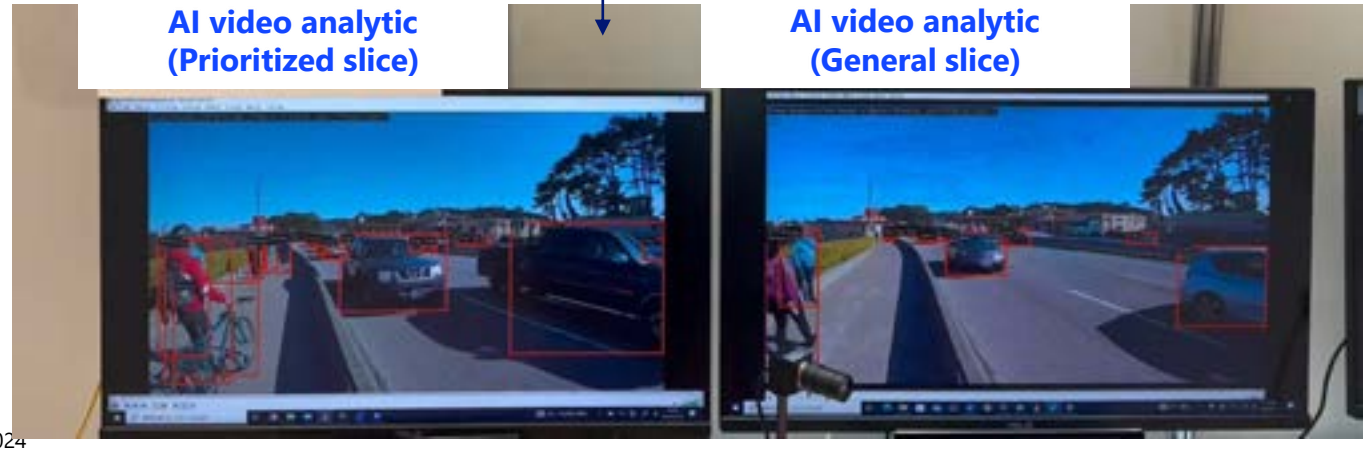
- Live 5G NW is used to demonstrate 5G slice equipped with ISAP controlled GPU resources
- Slice and session control by 5GC and GPU resource control by ISAP are coordinated.

【PoC system architecture】



Ref: 5G slice and ISAP integrated demo

NW slice and
ISAP collaborated
demonstration



Concluding remarks

Inclusive core

Next gen. core network concept proposed by NTT

- Core network for convergence and cooperation incl. computing and network
- INC is one of fundamental technologies
  (Other: robust/resilient signaling and self-sovereign identity management, etc.

In-network computing for 6G

ISAP: Integration network and compute for 6G/IOWN

- Resource sharing NFs and service application functions
- Serverless framework for efficient use of compute resource in NW
- Mobile core and computing PF inter-working and coordination

Accelerator HW

HW acceleration and chaining for NW & service acceleration

- Multi-tenancy and resource sharing and isolation of accelerators
- Open API/SDKs, common functional and performance spec.
  (Linux Foundation OPI project)
- HW layer function chaining IF with low-layer protocol (RDMA/DMA)

Your Value Partner