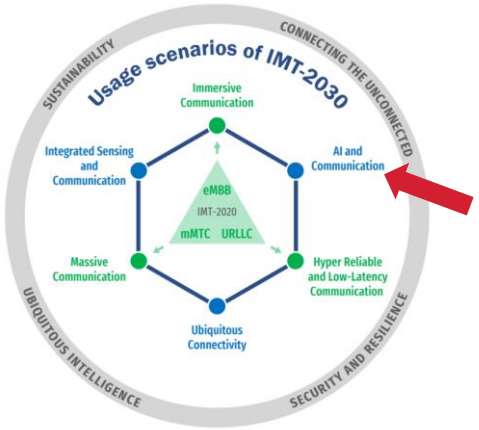
6G Fueling the AI Compute Continuum

CELTIC-NEXT Proposers Brokerage Day
Aveiro, Sept. 11 2025

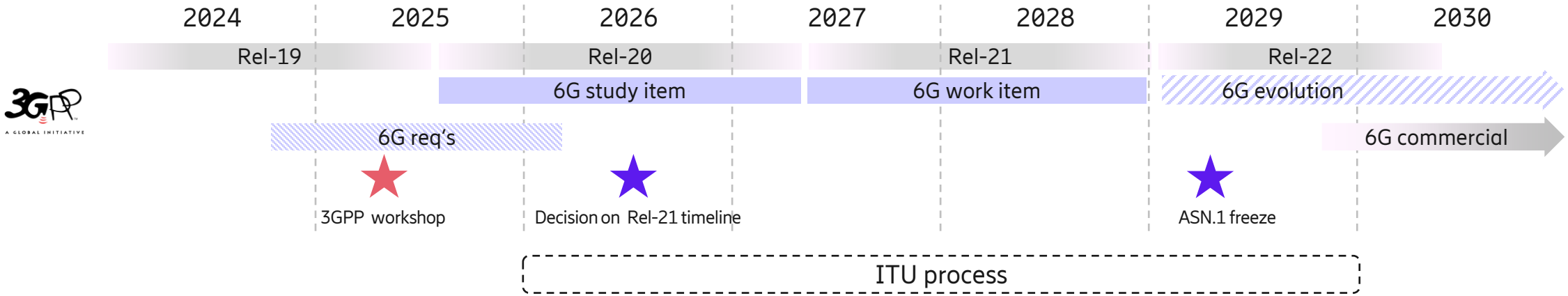
Wolfgang John
Ericsson Research, Sweden
Compute & Software

6G is starting to happen

- 6G standardization has started in 3GPP
- Study item, followed by a work item – ~3 years in total
- 6G fundamentals will be set very soon, but 6G will continue to evolve…

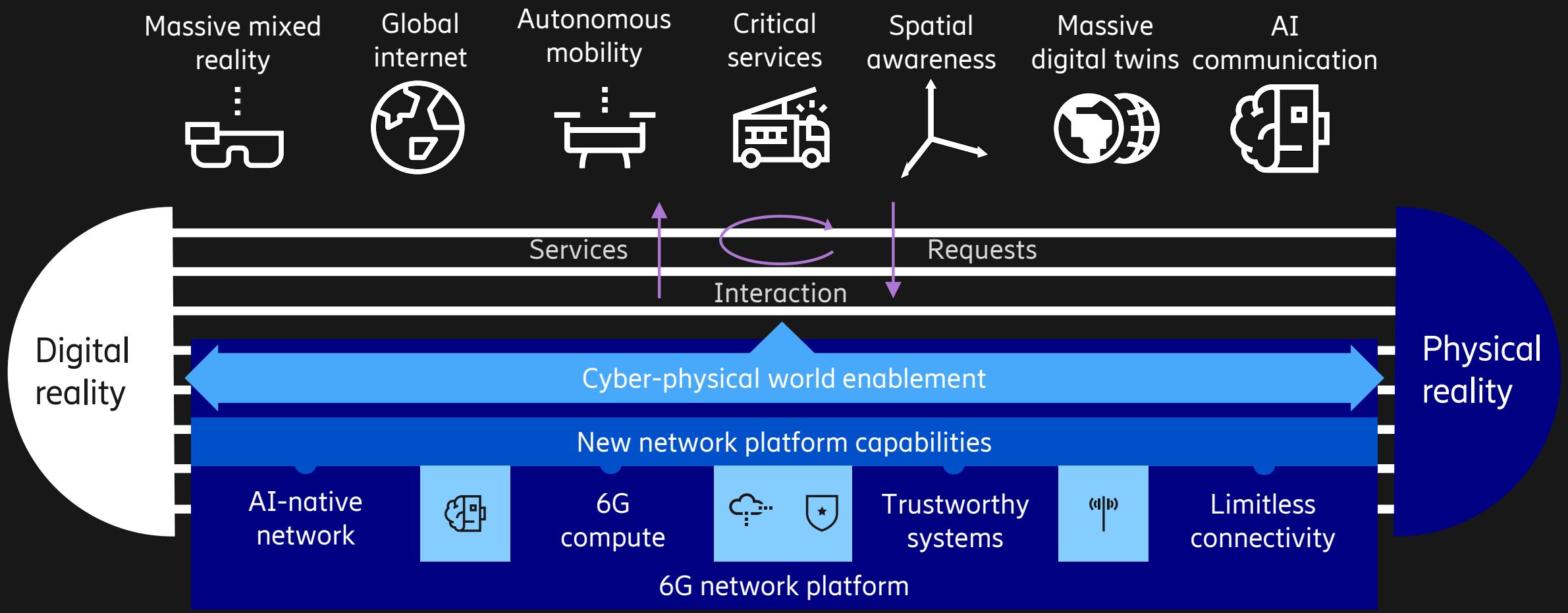


ITU-R M.2160



Co-creating a cyber-physical world

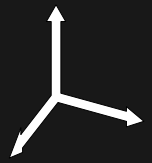
With the 6G network platform



AI-powered communications with smart glasses

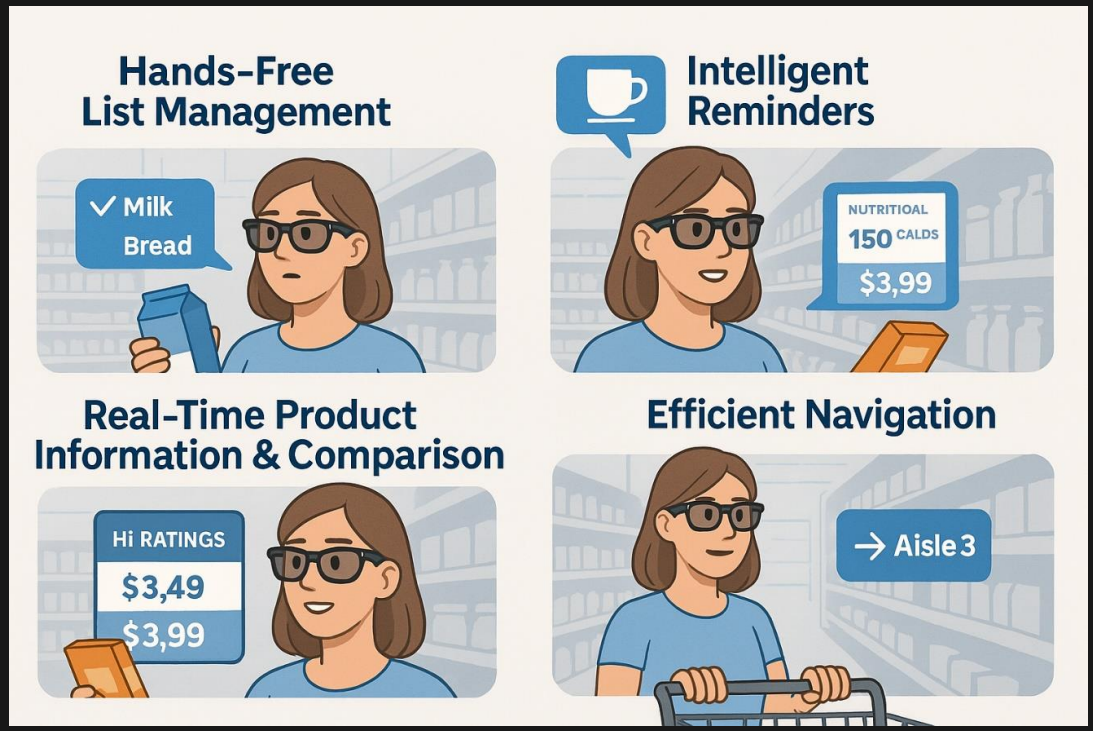
Massive MR

Spatial awareness

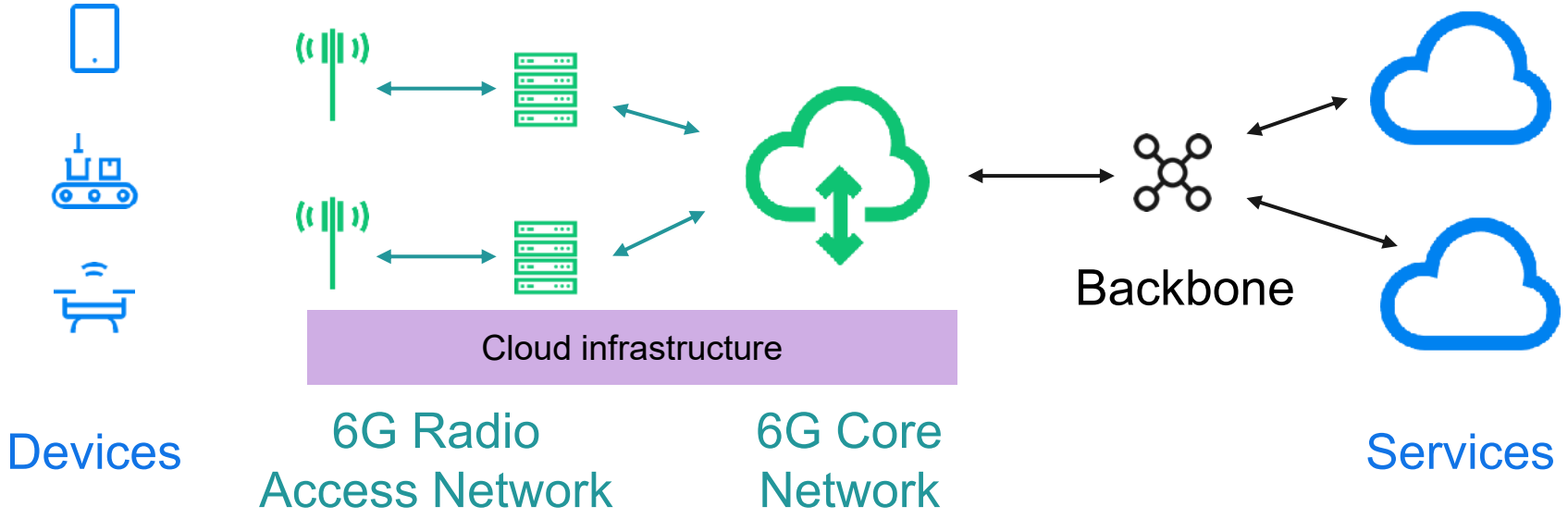


AI communication

Example: AI shopping assistant



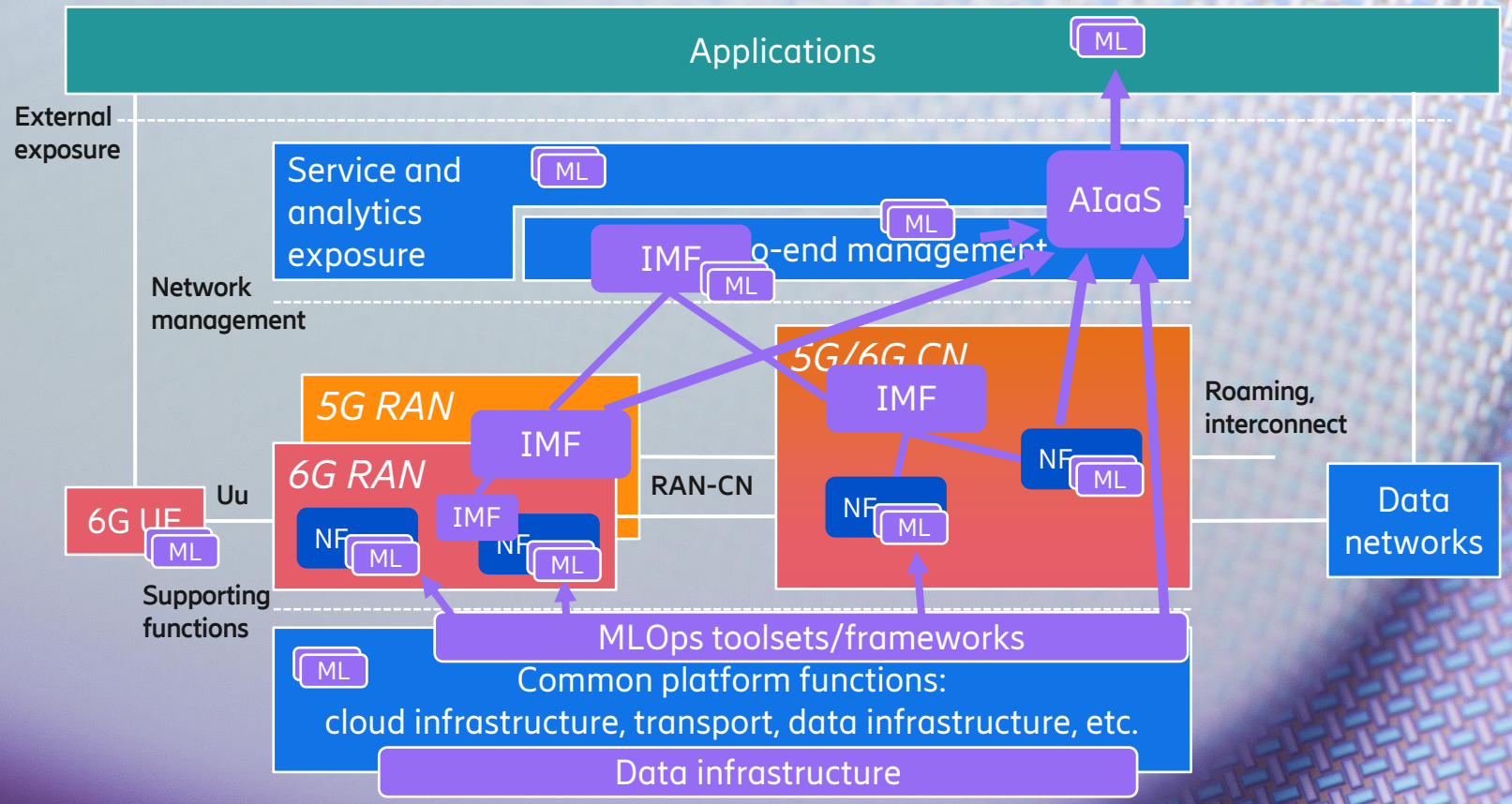
6G continues to evolve Cloud-native: Cloud-based all the way to the base-station!



Today's mobile networks are both a communication and computing platform

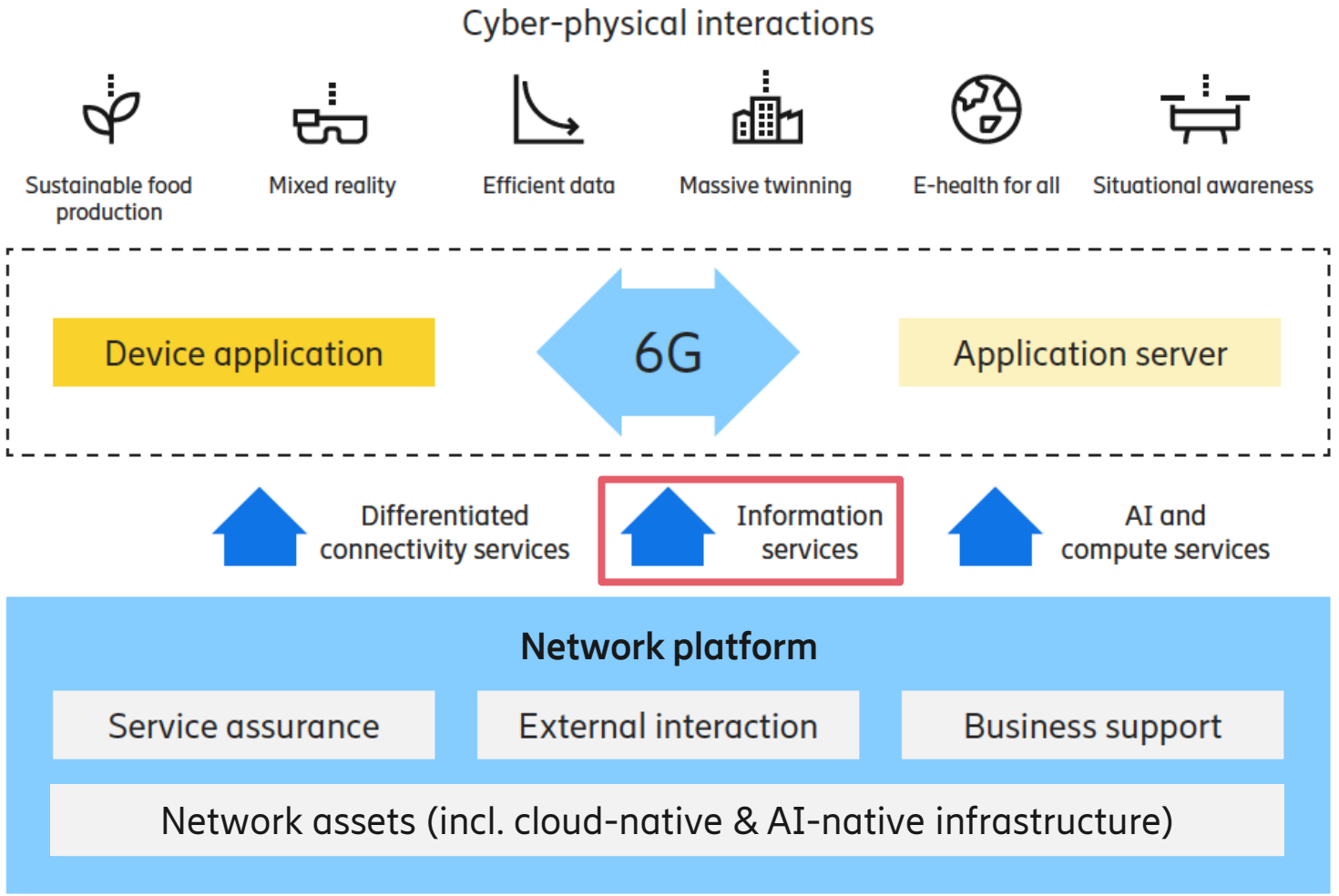
The 6G network will be AI native

Mobile networks will include a distributed data and AI infrastructure



The network platform

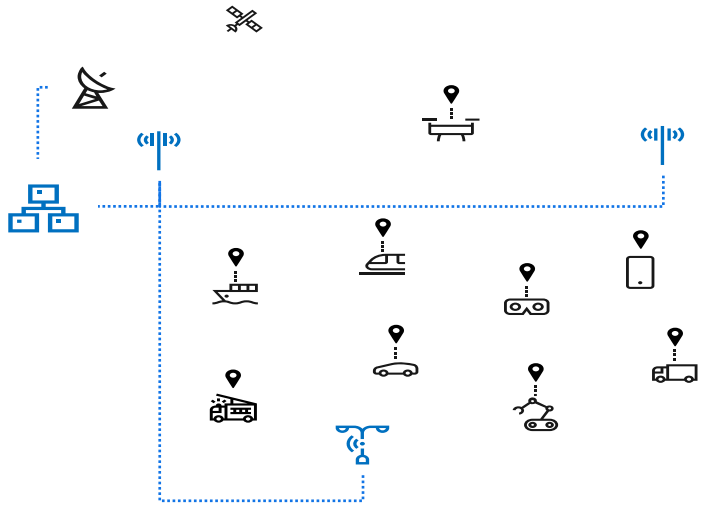
built on 6G technology and services



Network positioning and sensing

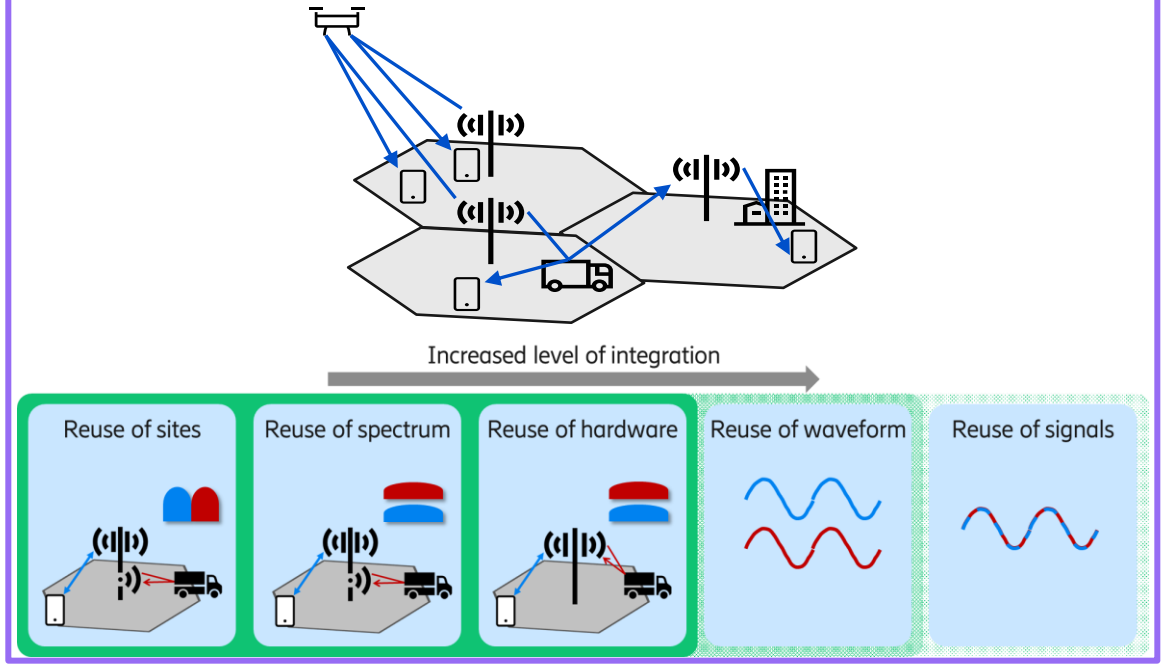
Positioning: an important part of mobile networks

- Regulatory, e.g., emergency call positioning and lawful interception
- Part of the IMT-2030 6G requirements
  - Mission critical
  - Smart factory
  - AR/XR and immersive experiences
  - Digital airspace (incl. drones)



Integrated Sensing and Communication (ISAC)

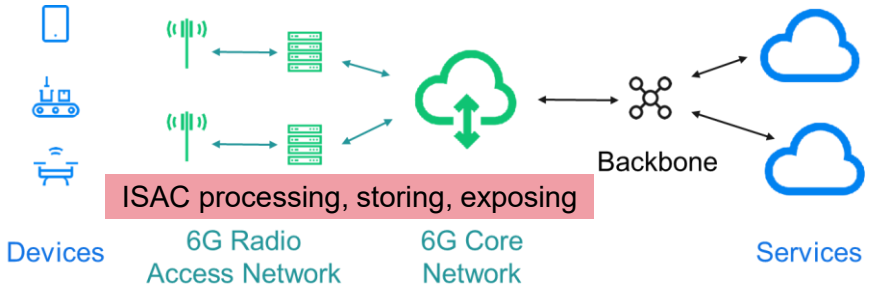
- Observing passive objects, i.e., not connected to the network
- Reuse mobile network infrastructure
- Some key design principles:
  - Limit additional complexity/cost of deployments
  - Signal waveform aligned with "normal" 6G waveforms

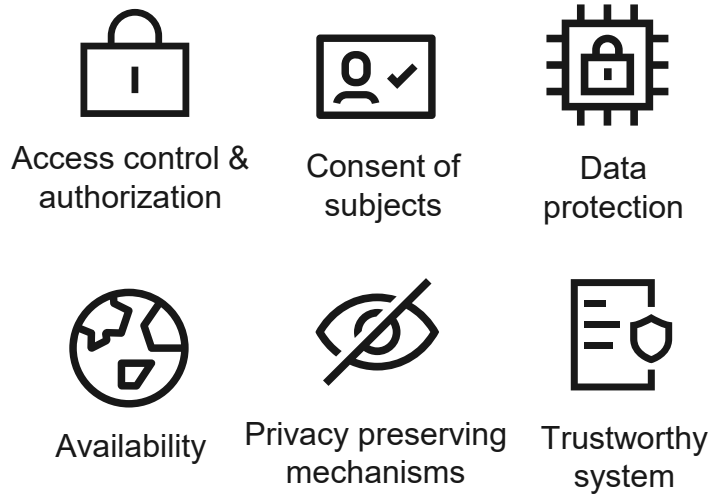


Network positioning and sensing

Some Challenges

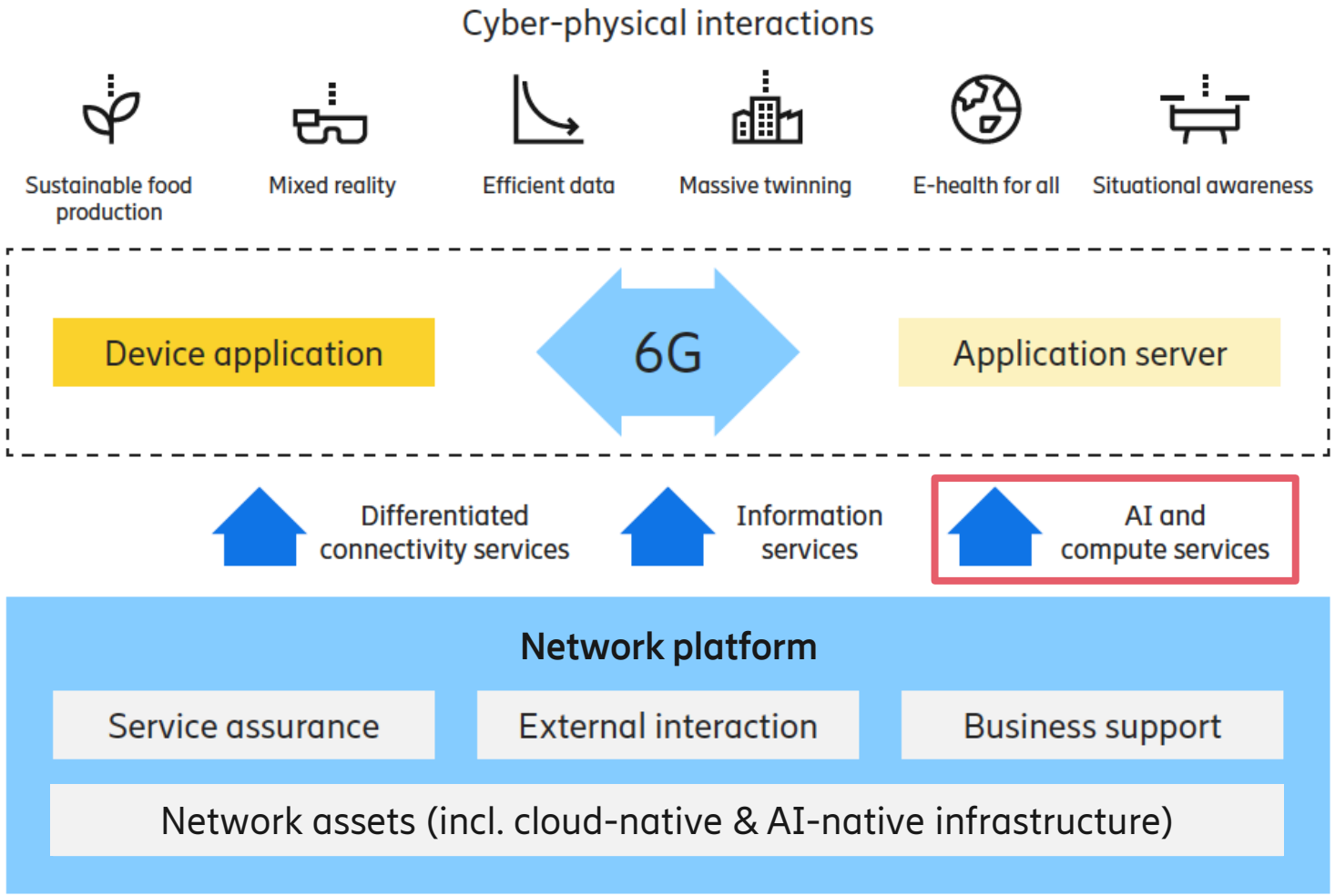
- Sensing data processing –
  how distribute data and processing in the network
  - Significant data flows inside the network (generated by RAN)
  - Processing and AI needs
- Security and privacy
  - Respect for privacy and regulations
  - Details depend on use cases:
    1. Easy cases (e.g., safety, weather monitoring)
    2. Controlled cases (private networks)
    3. General cases (very high-resolution, wide-area)
  - Some basic ingredients for solutions on the right



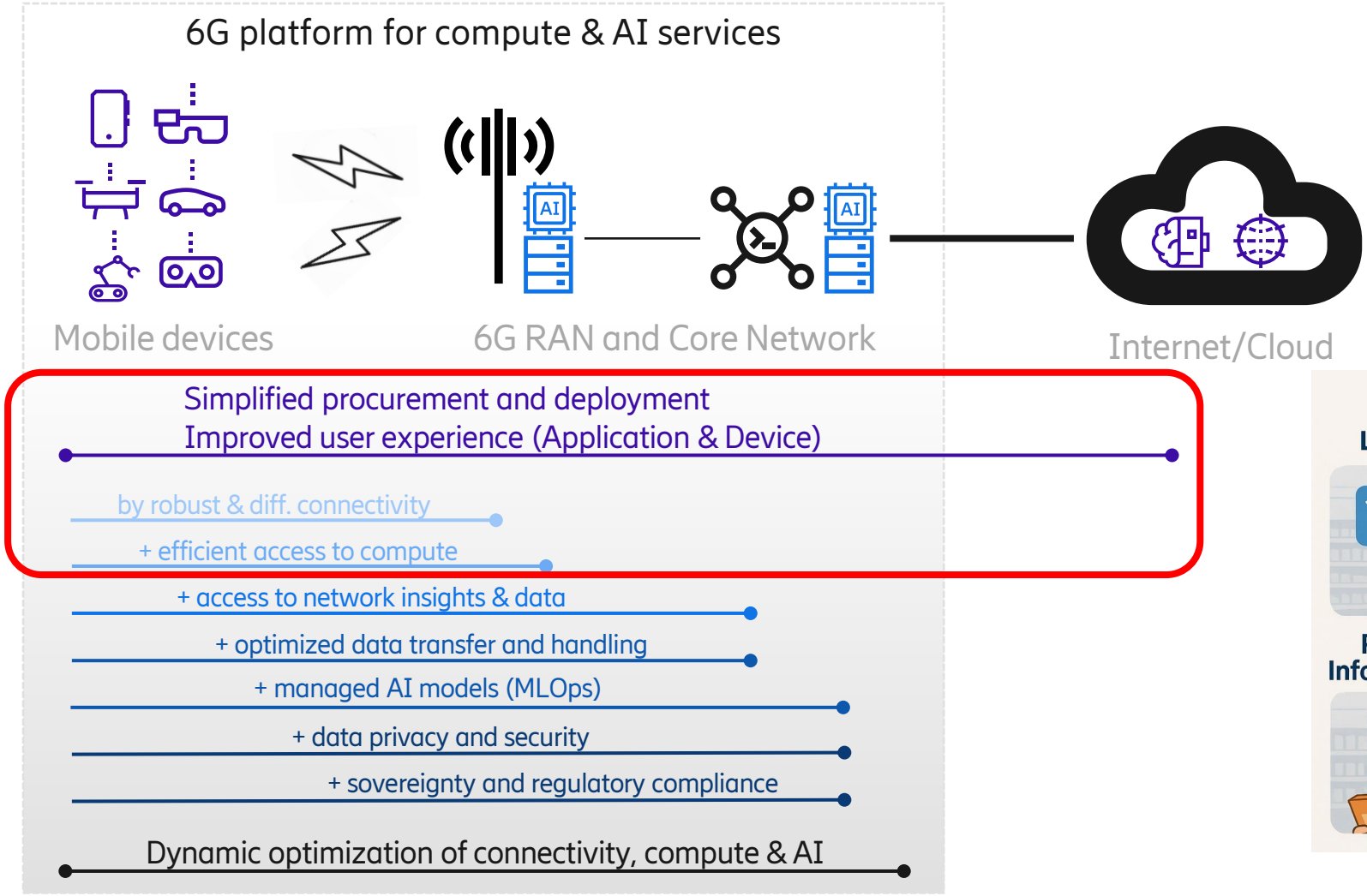


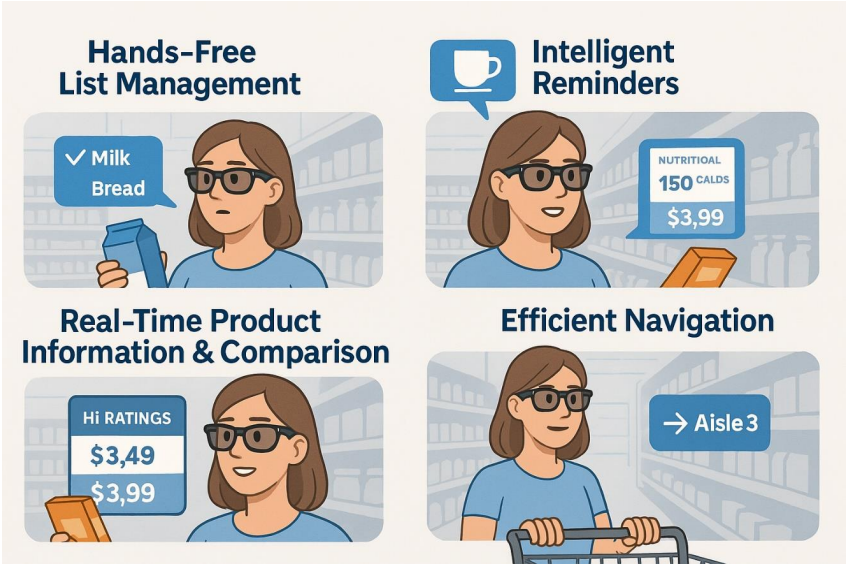
The network platform

built on 6G technology and services



6G AI compute continuum

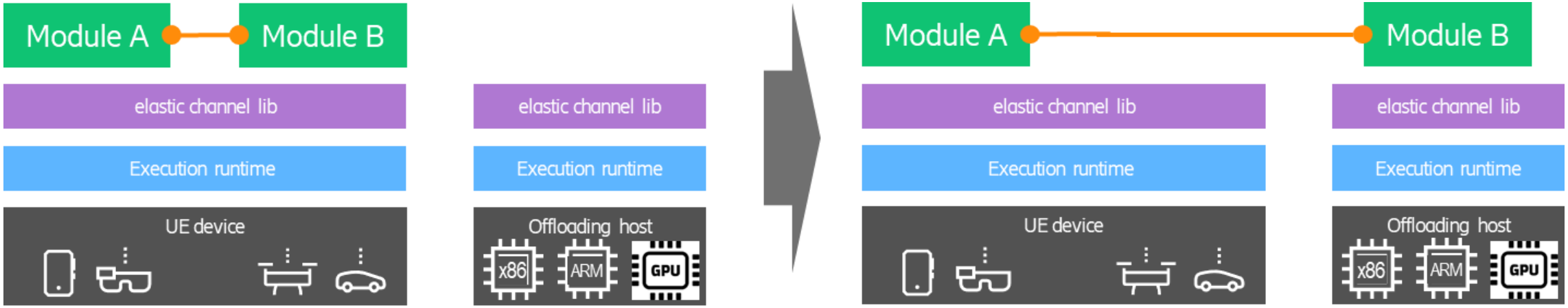




Example use-case:
Smart glasses shopping assistant

Dynamic device offloading: Background and basic idea

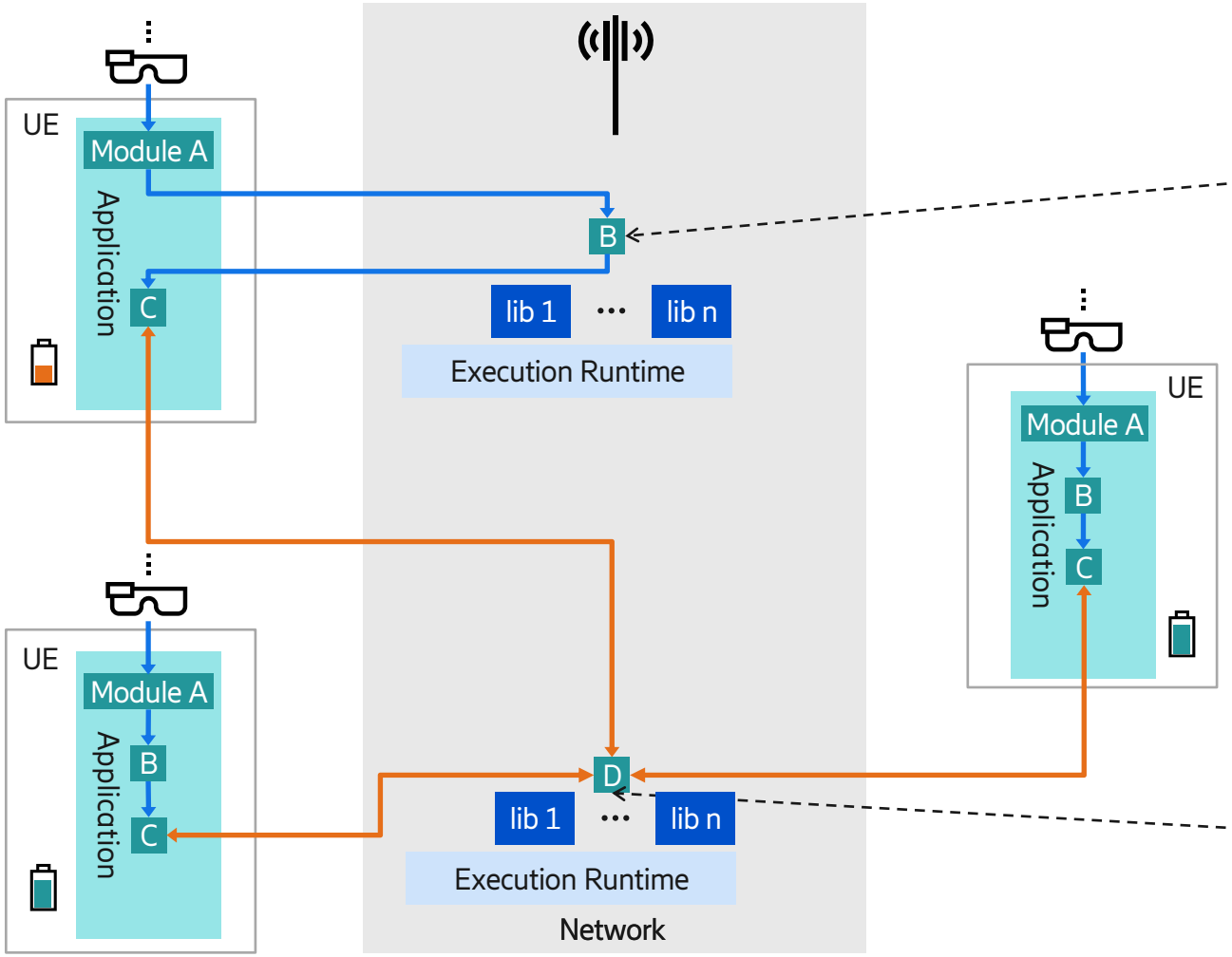
Extending the functionality of mobile devices through dynamic use of remote compute resources



Bring your own code

Deploy anywhere

Adapt dynamically



Computational heavy functions can be offloaded to reduce device heat, battery consumption, or computation times

Demonstrated results of this work:

- Use-cases and device agnostic solution
- Dynamic task deployment during runtime
- Supporting platform heterogeneity
- Minimized development overhead
- Quantified benefits for the mobile device
- Network integration with 3GPP EEL and AKMA

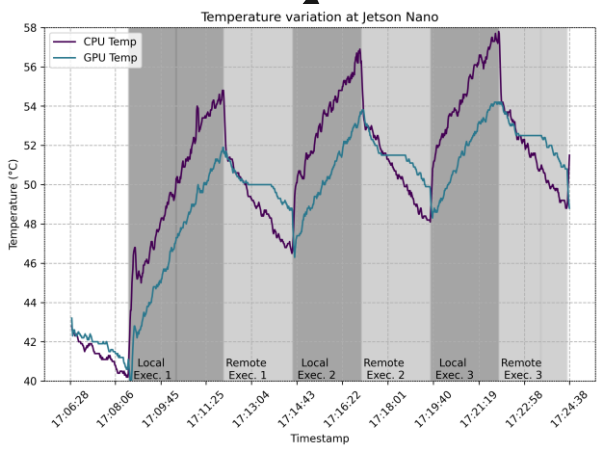
A common coordination task can be offloaded, e.g., to realize collaborative perception or to save overall bandwidth

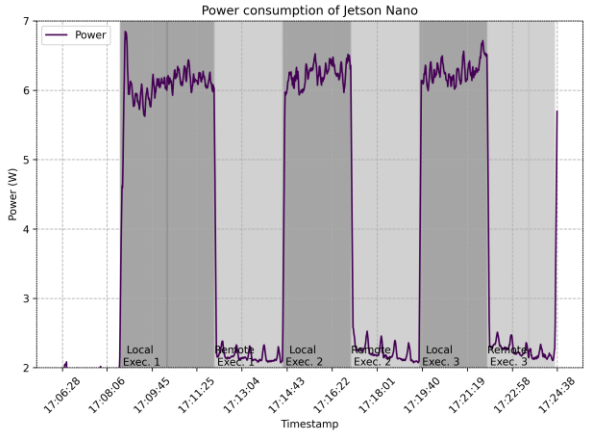
Measurement results

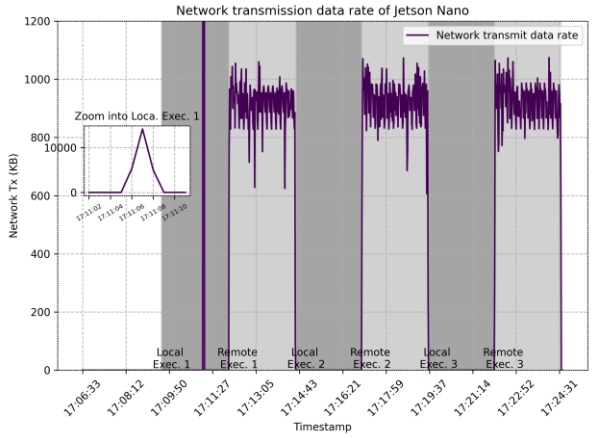
Beneficial
Neutral
Extra Cost

Gains and costs wrt offloading

	CPU Usage	GPU Usage	CPU Temp	GPU Temp	Device Power Consumption	Network TX	Network RX	Computational Response Time
Raspberry Pi 3B (limited HW CAP)								
Jetson Nano (moderate to high HW CAP)								







Summary of results published at CLOSER

Demonstrated use-case scenarios and UE types

taking advantage dynamic device offloading

Search & Rescue with robots or drones

- Unknown location, highly dynamic
- Happens rarely – but when it happens, it needs compute infra

Demonstrated benefits

- Improved QoE, e.g., fast detection times
- Facilitate efficient coordination and collaborative perception
- Prolong robot operation

Demonstrated and published at Closer'24

Environment-aware avatar communication with XR headset

- Requires heavy 3D object detection
- Mobility use-case

Demonstrated benefits

- Reduce HMD heat and prolong battery life
- Improve application QoE
- Possibility to coordinate local users

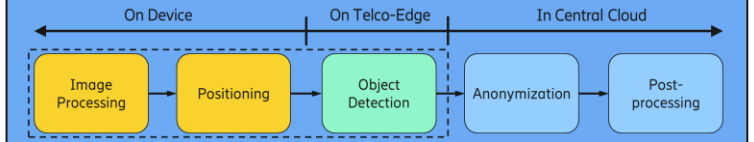
Demonstrated at MWC'24 and IMC'24

Smartphone-based SmartCity app

Demonstrated benefits

- Solving real-life problems encountered out in the field
  - Reduce device heat on sunny days
  - Maximize chance of unplugged devices to survive until re-plugging
- Possiblity to increase detection accuracy in areas of special interest

Demonstrated as part of Celtic-next project ANIARA



Device offloading

Computational offloading:

→ Improved behavior of portable devices

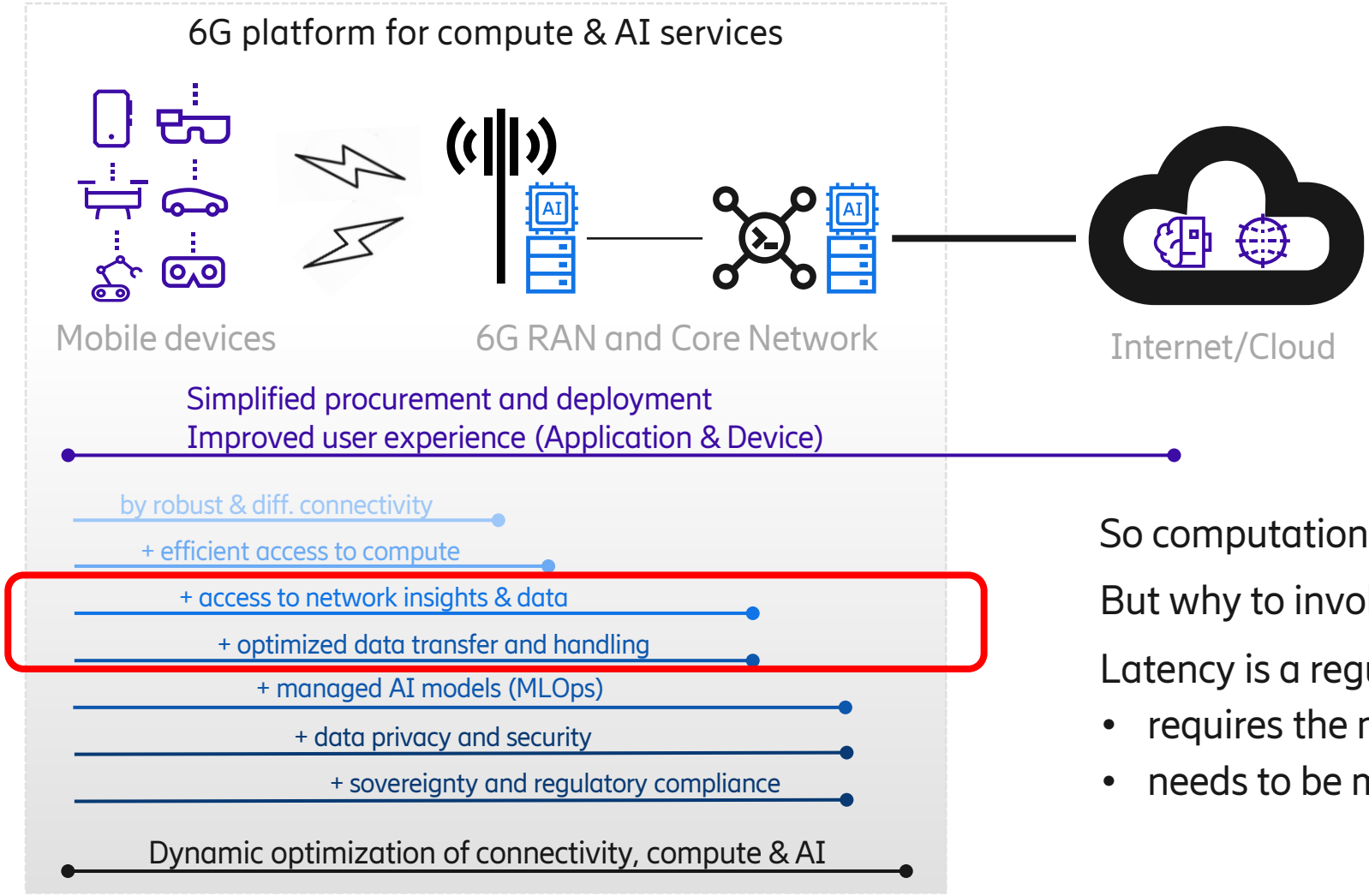
→ Improved applications
on devices with resource constrains

Some open challenges

- When is it beneficial to offload?
  - Predictions based on full situation analysis (conditions of device, site, network, app, user, env., cost,…)
- How can we handle state in a more effective way?
- Improved communication & compute security/isolation
- Wasm/WASI evolution (incl. component model) for this type of use-cases
- Network integration of the signaling path
- Exposure – how to reach the developer communities (APIs, libs, …)
- ….

[Note: our results are indented to provide a proof point in terms of potential benefit of offloading. The results depend on the computational task and the choice of HW on both sides!]

6G AI compute continuum



So computational offloading can be well motivated.

But why to involve a network platform?

Latency is a regular answer! This is true, but

- requires the network architecture to support that
- needs to be motivated by (large scale) use-cases

NW platform as source

Examples of relevant data and insights within a mobile network

Fraud detection and prevention

Call status (ongoing or forwarded), Device status (reachable or roaming), Location verification, or SIM Swap indications as APIs to apps by, e.g., financial institutes or pay-per-use businesses

Connectivity status and prediction

Current and future expectations of connectivity quality at a location, e.g. for UAVs to propose routes with coverage, or apps to predict QoE or adapt application properties

Spatial data

Network positioning and sensing data, fused with other data sources, e.g., to support AR/XR device with improved scene understanding, orientation, navigation, object recognition, etc.

User density

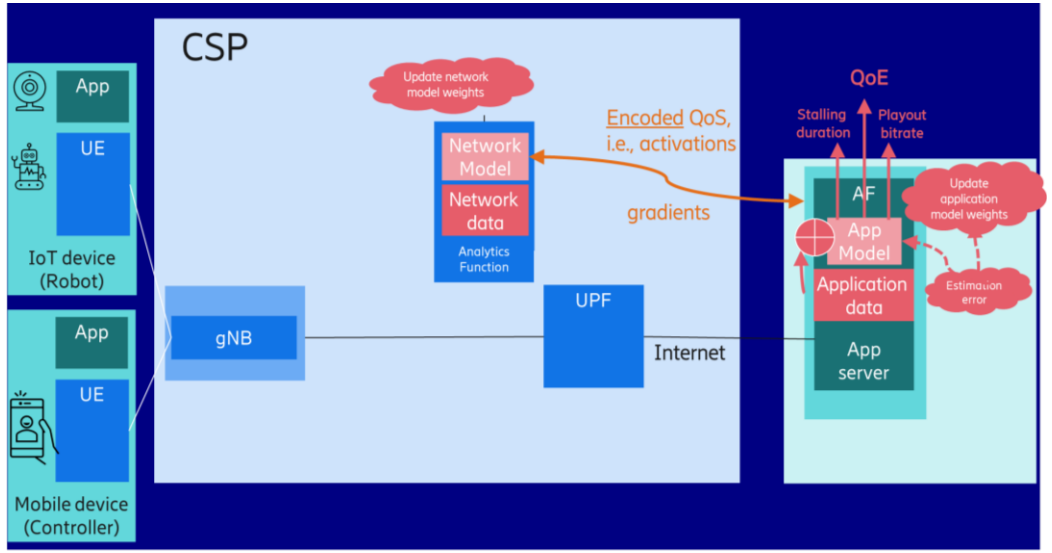
Density of devices (SIM) and detected crowds (sensing-based), e.g., for mobile apps to avoid or seek hotspots, warn of traffic/obstacles, etc.

Use case examples

utilizing network insights

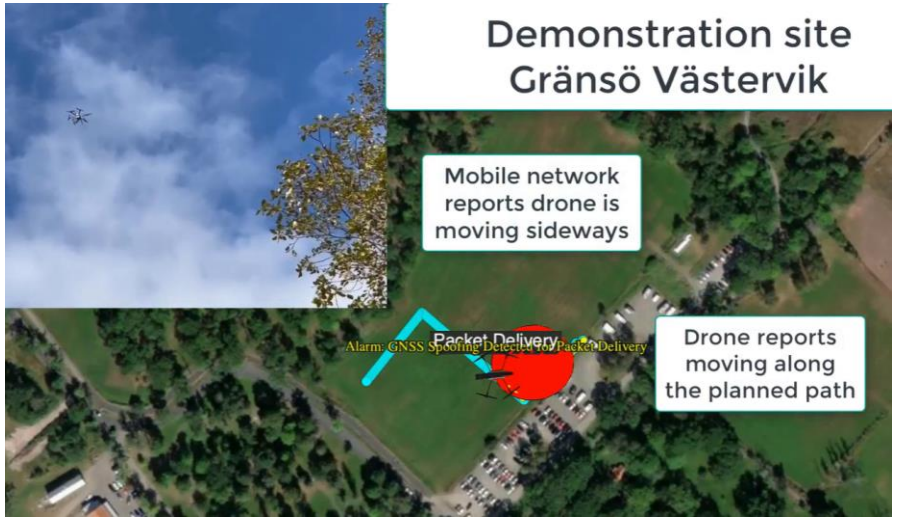
Example: Collaborative QoE estimation for remote controlled Robots

- Improved QoE estimation using vertical federated learning between a network model and an application model
- Sharing activations (inference) and gradients (continuous learning) instead of large, business sensitive data



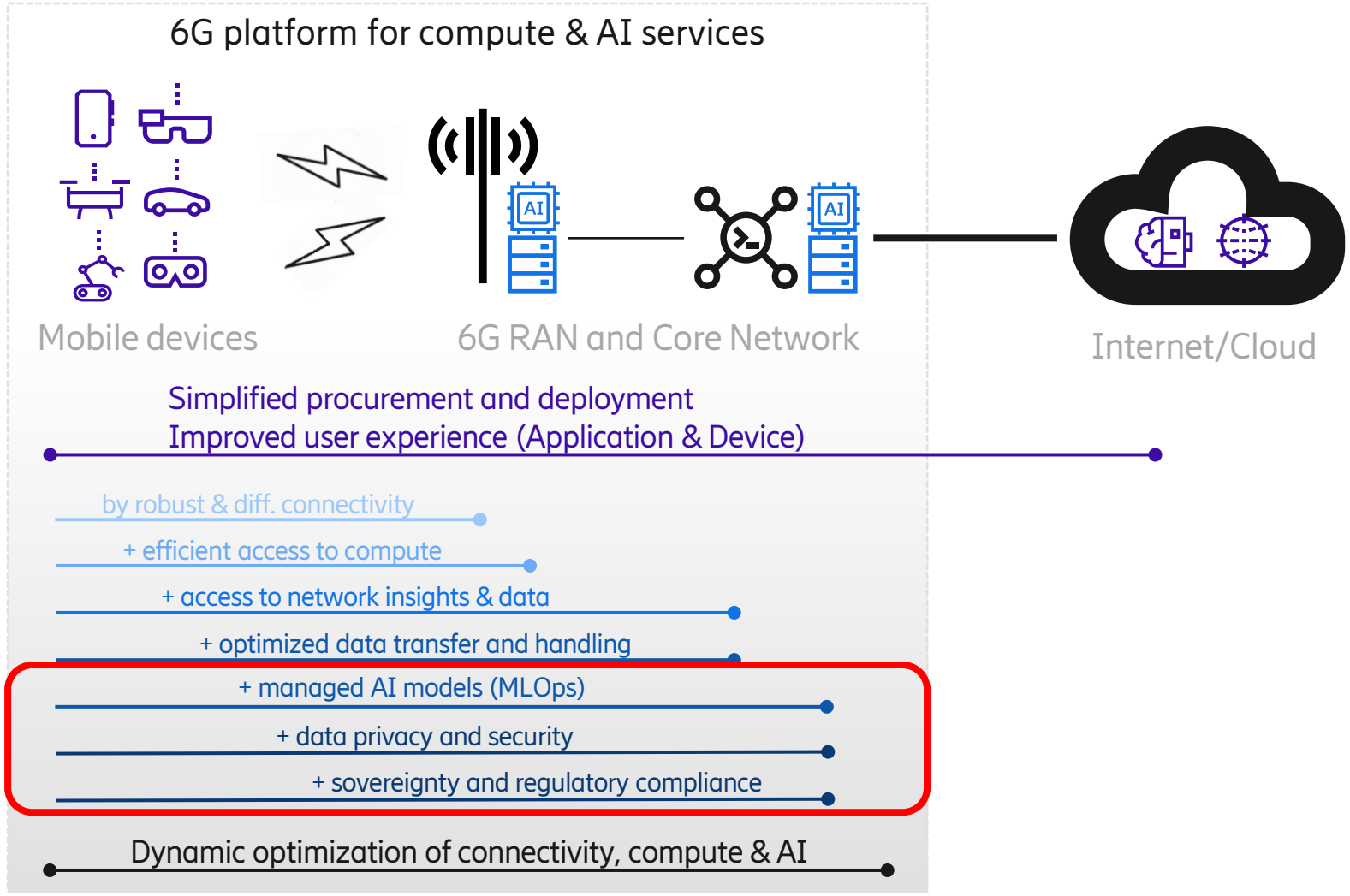
Example: Position Retrieval and Validation (PRV) for digital airspace (UAVs ~ drones)

- Non-line of sight UAVs can report wrong GNSS (~GPS) positions due to spoofing, failures, or deceiving
- UAV traffic management critically depends on real positions
- Trustworthy positioning information by the network to validate report UAV positions



Part of Hexa-X-II D3.5

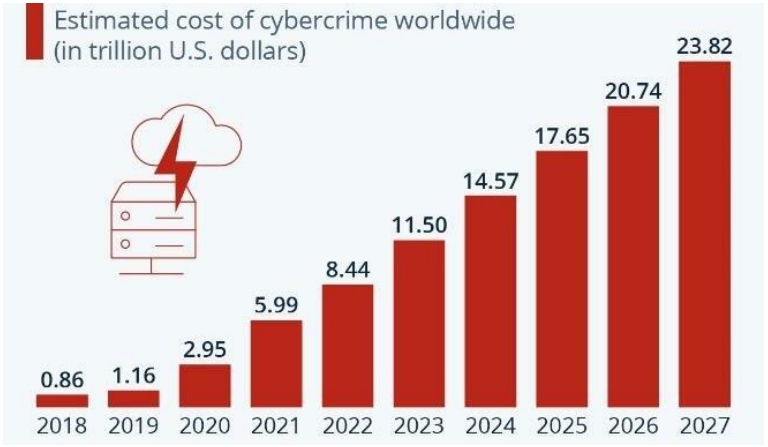
6G AI compute continuum



Trust, privacy, and data sovereignty

as a potential driver for local compute & AI services

- Individuals: Data privacy continues to be an important aspect
- Enterprises and industries: Data privacy and trustworthy infrastructures are business-critical – but face new threats (remote work, GenAI, state-sponsored cyberwar, etc.)
- Regulators: Digital sovereignty is increasingly regulated by regional directives/laws - fueled geo-political developments



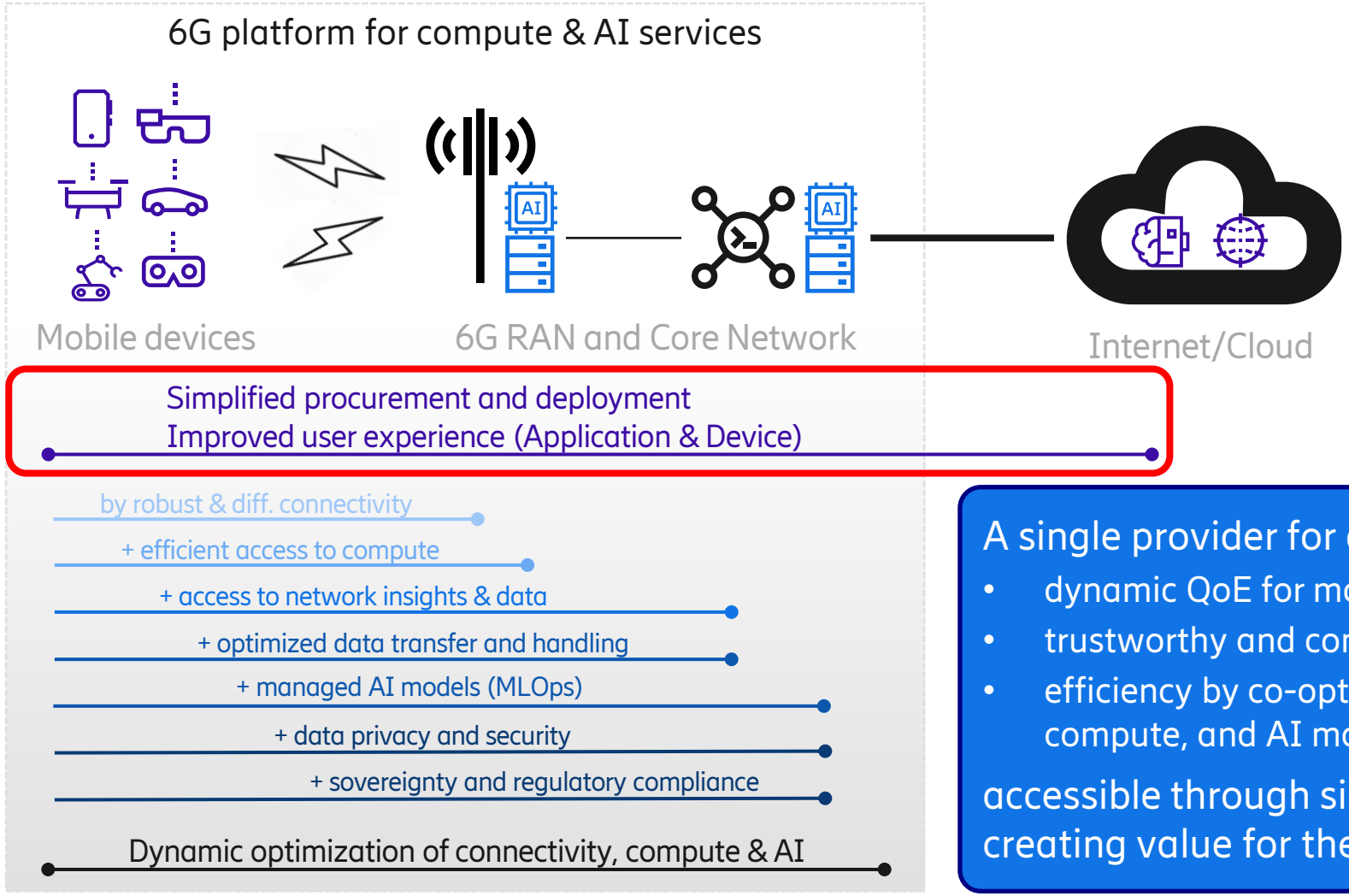
Source: Statista

Some examples:

- **EU: GDPR, EU Data Act (promoting exchange of data)**
- **India: Digital personal data protection act (DPDPA)**
- **Japan: Act on the protection of personal information (APPI)**

Opportunity for Mobile Network operators (MNOs) as a nationally regulated business partners: offering trustworthy bundles of communication, compute, data handing, and AI model management?

6G AI compute continuum



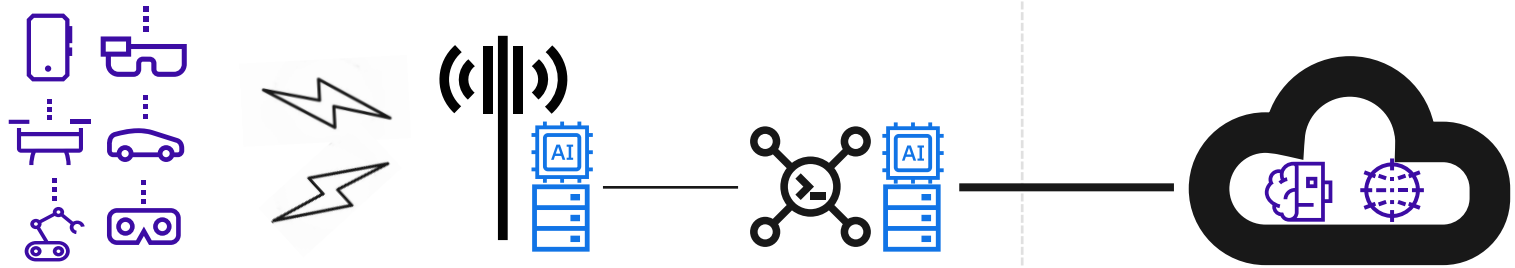
A single provider for a bundle of managed network services

- dynamic QoE for mobile devices & apps
- trustworthy and compliant infrastructure and services
- efficiency by co-optimization of connectivity, compute, and AI model management

accessible through simple APIs,
creating value for the application ecosystem.

Research topics
relevant for a 6G AI compute continuum

6G platform for compute & AI services



Mobile devices

6G RAN and Core Network

Internet/Cloud

Simplified procurement and deployment
Improved user experience (Application & Device)

by robust & diff. connectivity

+ efficient access to compute

+ access to network insights & data

+ optimized data transfer and handling

+ managed AI models (MLOps)

+ data privacy and security

+ sovereignty and regulatory compliance

Dynamic optimization of connectivity, compute & AI

Research topic areas

AI/ML Model and LCM
(model split, deployment,...)

Service Orchestration and Assurance

Exposure and APIs

Compute continuum
(device- infra compute sub-systems and offloading framework)

Complementary Network functions

Architecture and deployment

Energy supply and consumption

Trustworthiness

6G in the AI Compute Continuum

- Telco sector is going towards platform business
  - Starting with connectivity related APIs
  - Adding information, compute, and AI services
- Cloud and compute technology is of increasing relevance for mobile networks
  - Compute for 6G –Cloud- and AI-native network architecture
  - 6G for Compute – offering a wider-platform for 3rd party applications
    - Potential of compute and AI services as part of the AI compute continuum

ERICSSON

CELTIC-NEXT

Eureka Cluster

https://www.ericsson.com/en/edge-computing/network-compute-fabric