TECoSA Seminar
Computation offloading in Edge and cloud environment: Survey, taxonomy, applications and open challenges

Ali Balador, Ericsson Research

Outline

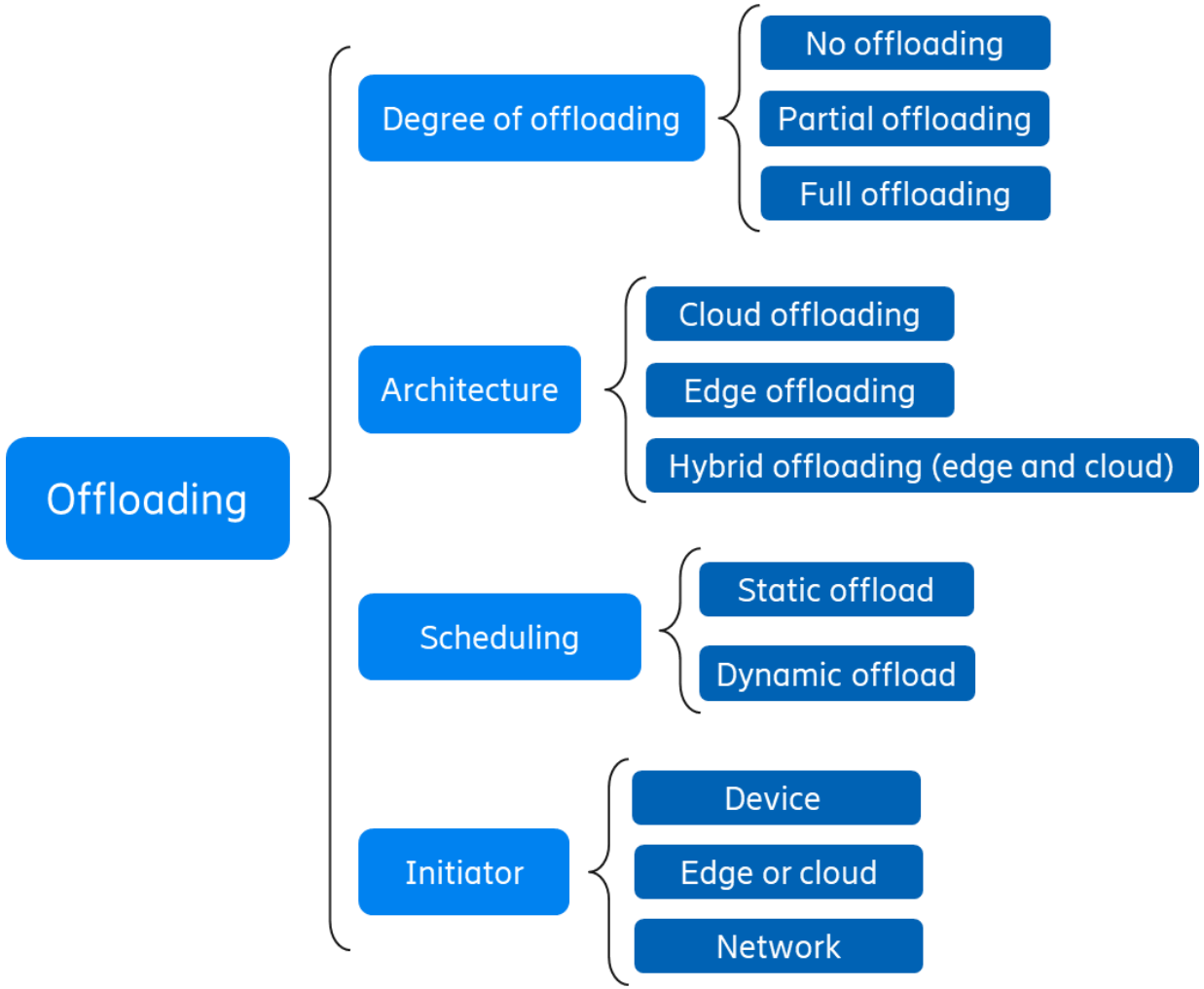
- Definition and classification of offloading
- Application domains and use cases
- Commonalities between application areas
- Mapping use cases to the offloading classification
- Timeline for offloading of applications
- Offloading targets for different application areas
- Overview of our Vinnova funded project AORTA

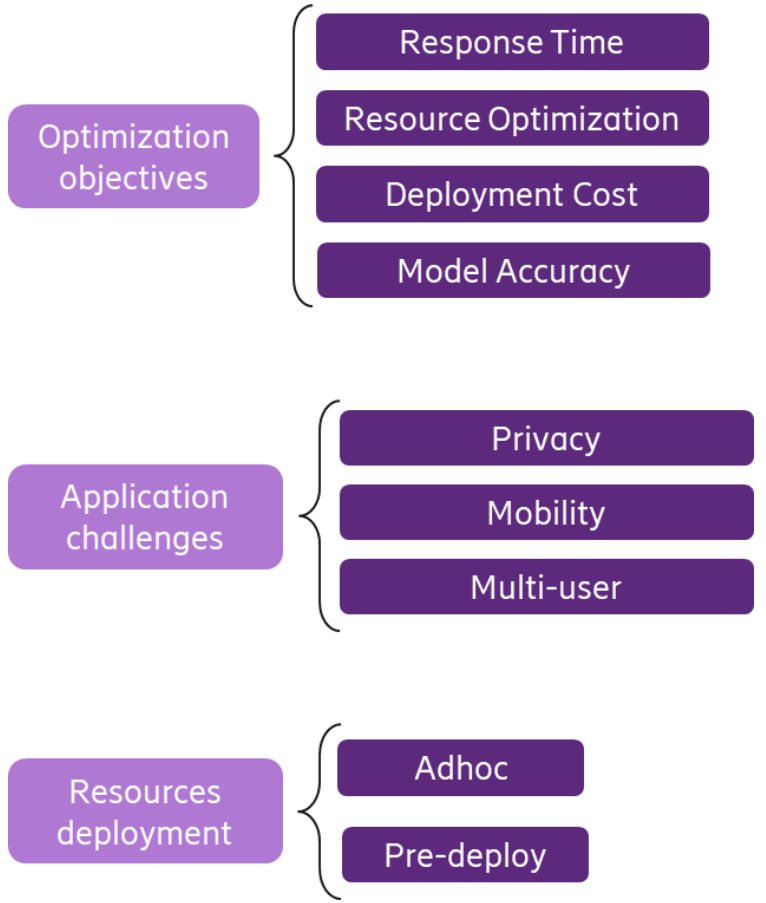
Offloading definition

A mechanism to move the processing or computation from one device to another with more suitable capabilities. Its main characteristics are:

- The main goal is to move **resource-intensive tasks** from a device with limited resources to an edge server or cloud server, or from an edge server to a cloud server.
- We consider **resources** to be battery, storage, processing capacity and network.
- **"More suitable capabilities"** can refer to higher processing capacity, better network access, higher network quality or greater storage capacity.
- Offloading allows the scaling of resources across the computing ecosystem.
- Task offloading aims to **achieve performance objectives**, such as reducing overall computation time, minimizing network resources usage, maximizing battery life, maximizing the revenue for the network operators, among others.

Offloading classification





Application domains and use cases

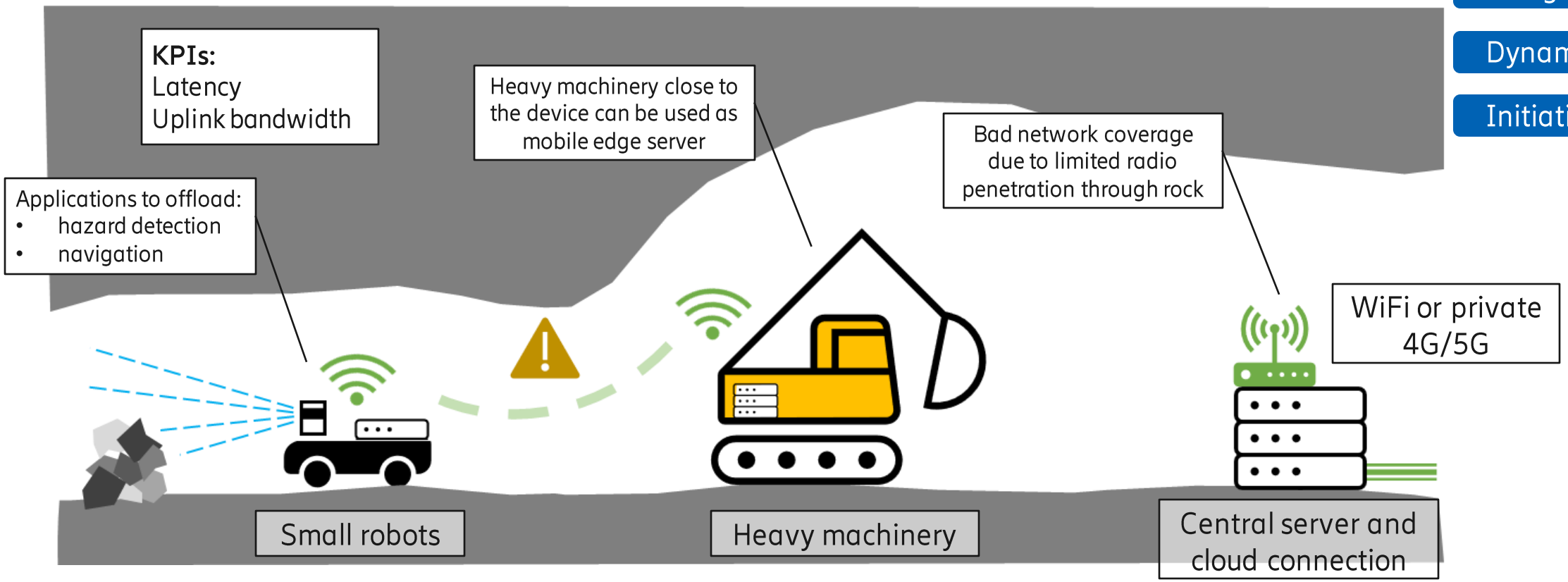
Mining: Mine inspection through autonomous vehicles

Partial/full offloading

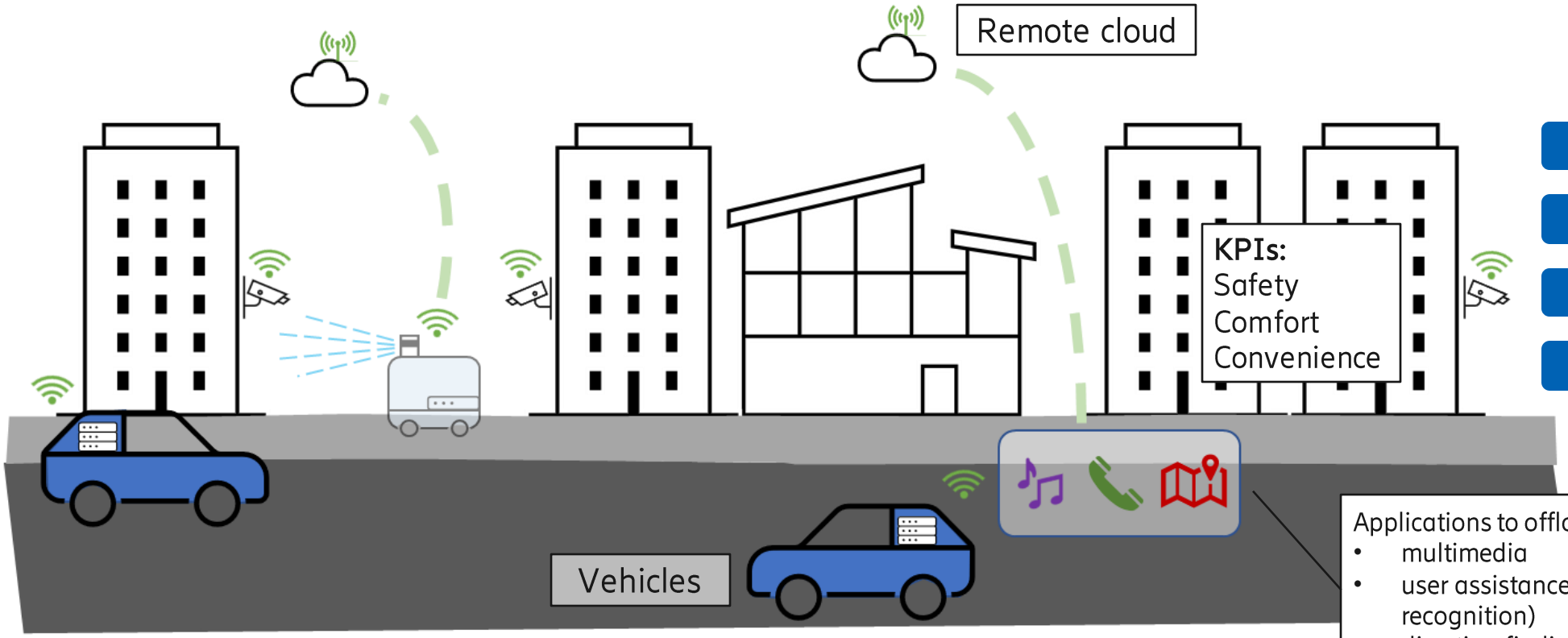
Edge offloading

Dynamic offloading

Initiation on device

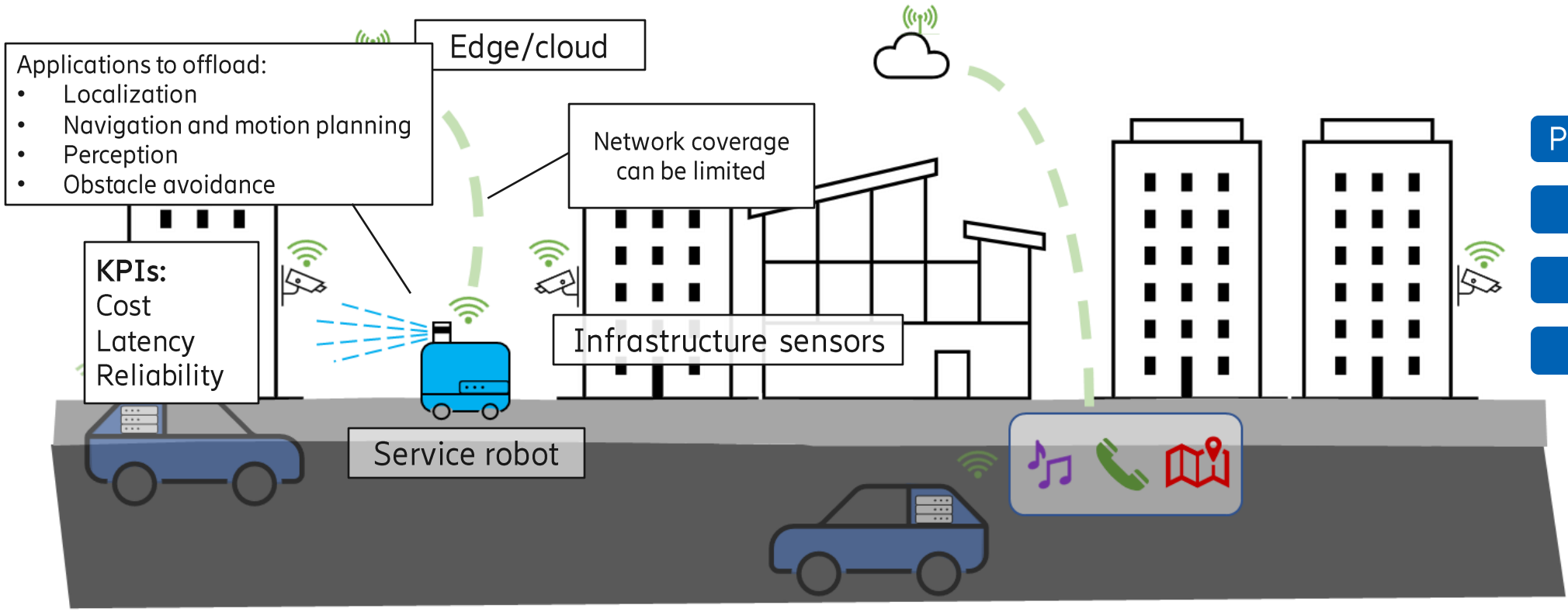


Automotive: Driver assisted vehicles



- Partial offloading
- Cloud offloading
- Static offloading
- Initiation on device

Service robotics: Service robots in public environments



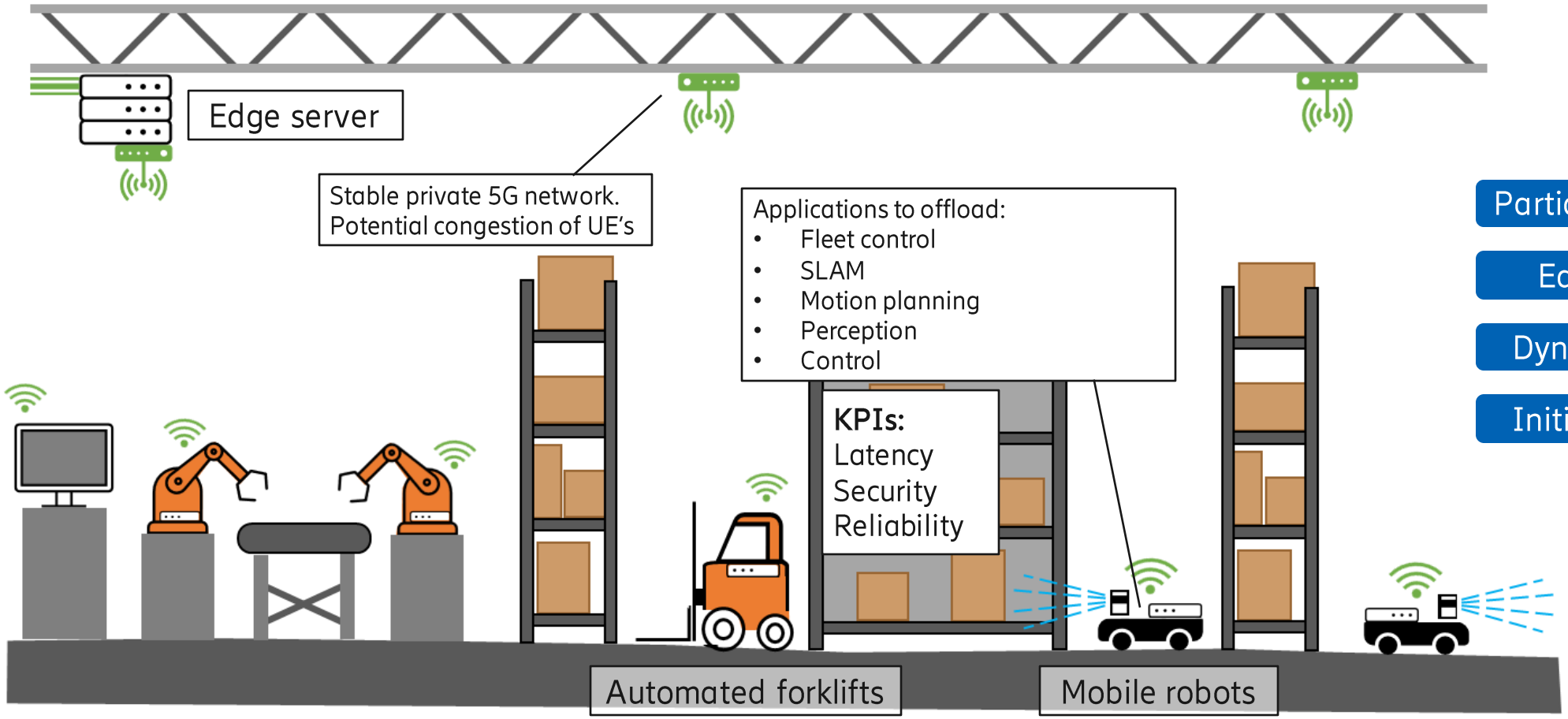
Partial/Full offloading

Hybrid offloading

Dynamic offloading

Initiation on device

Manufacturing: Mobile robots in factory facilities



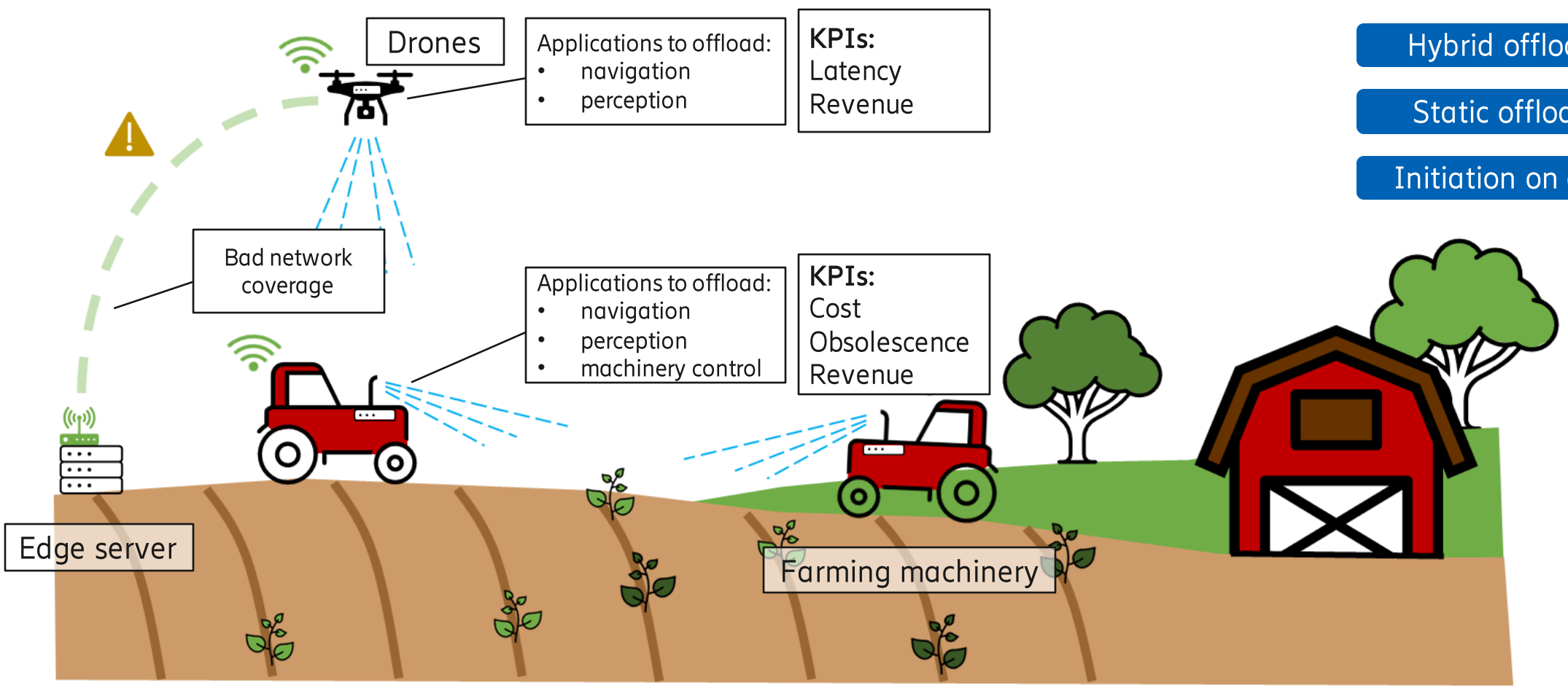
Partial/Full offloading

Edge offloading

Dynamic offloading

Initiation on device

Farming: Autonomous farming machinery



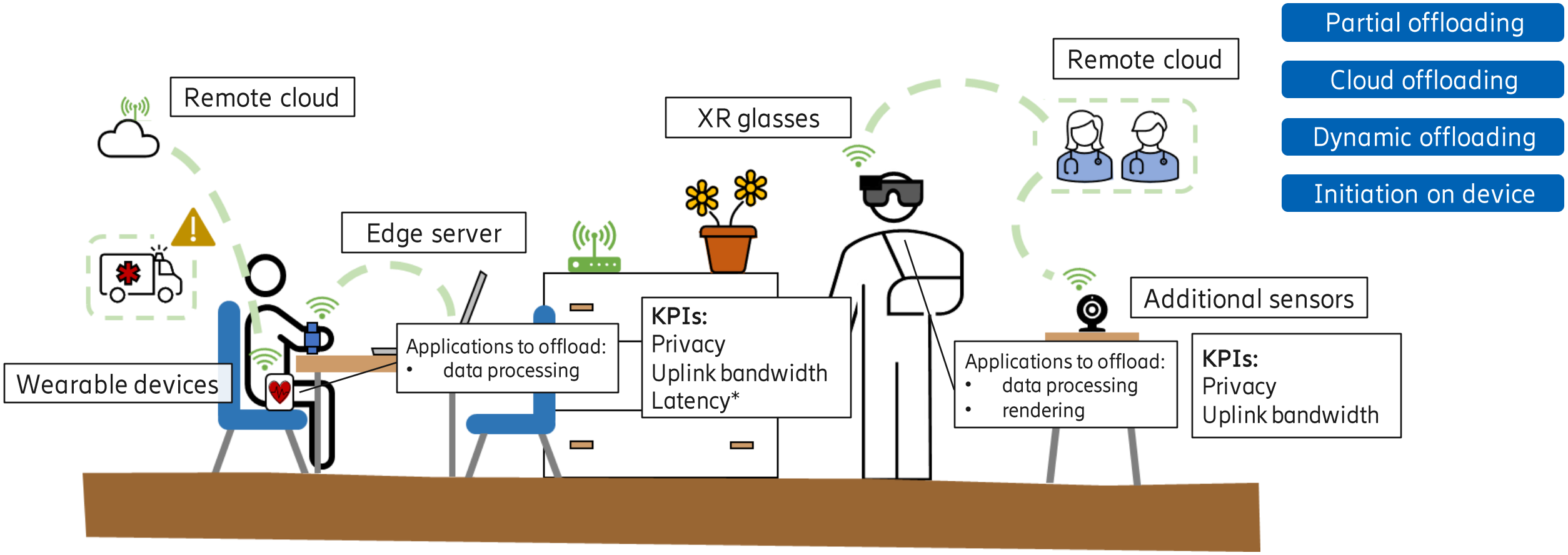
Partial offloading

Hybrid offloading

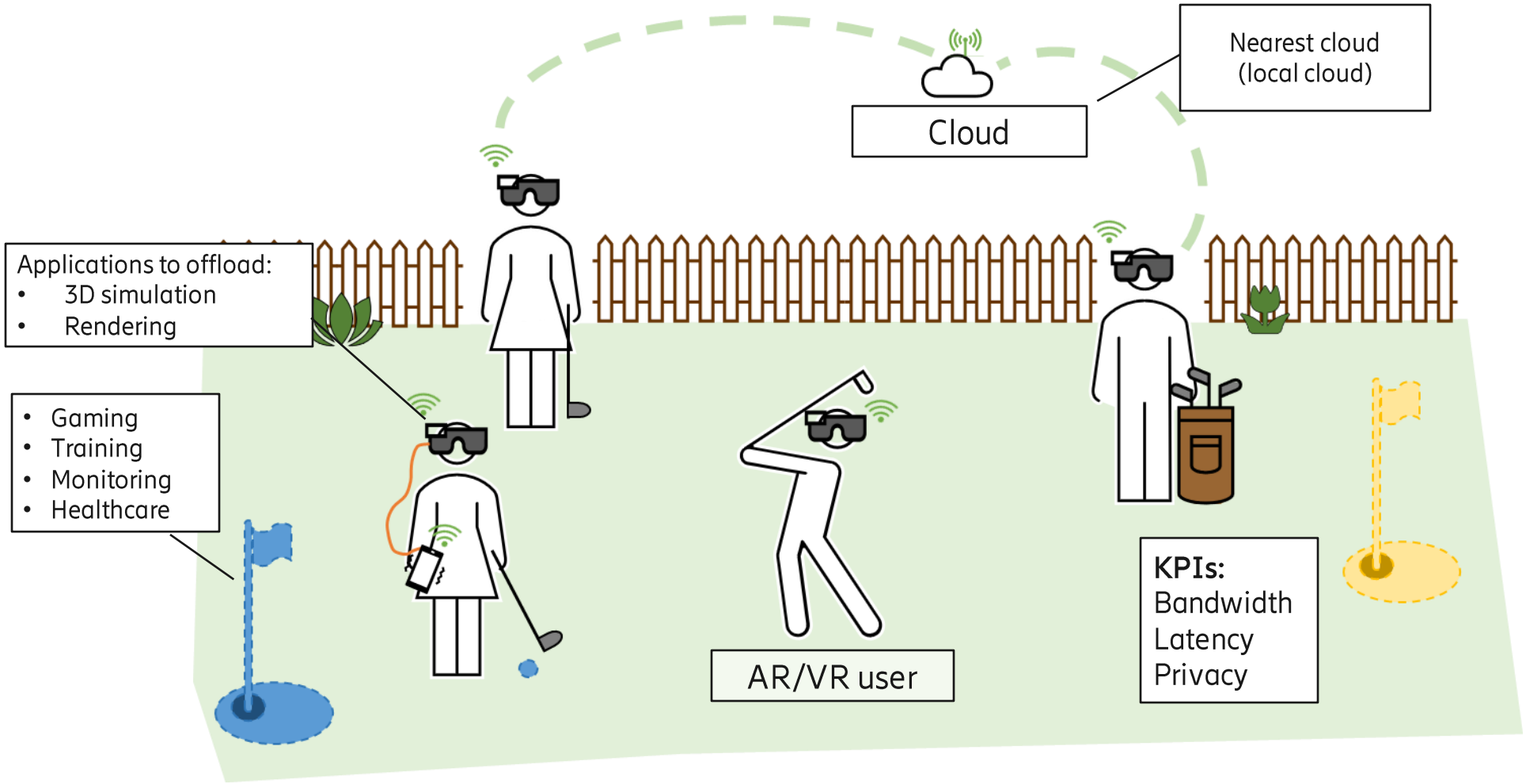
Static offloading

Initiation on device

Healthcare: Health wearables and XR-assisted physiotherapy



XR: Remote rendering



- Partial offloading
- Cloud offloading
- Static offloading
- Initiation on device

Commonalities between application areas

Trade-off between safety and performance

Privacy preservation

- Offloading requires techniques to secure environment and personal data.
- Data encryption increases on board processing and transmitted bandwidth

Federated Learning to preserve privacy

"Thin client"

- General future aim to use offloading to reduce to the minimum on board processing and sensors.
- Reasons: cost, battery life, hardware obsolescence, theft, etc.
- Currently, there are mainly security, deployment and network limitations.

Safety regulations

- Offloading requires additional safety considerations to be taken into account.
- While humans are involved keep control of large vehicles locally.
- Smaller robots need to offload control but keeping a local safety backup or emergency stop and limiting speeds.

Mapping use cases to the offloading classification

Use Case	Degree of offloading	Architecture	Deployment	Scheduling
Mine inspection through autonomous vehicles	Partial/full	Edge	Pre-deploy	Dynamic
Mobile robots in factory facilities	Partial/full	Edge	Pre-deploy	Dynamic
Service robots in public environments	Partial/full	Hybrid	Ad-hoc	Dynamic
Remote rendering, e.g., XR	Partial	Cloud	Ad-hoc	Static
Drones/Robots in disaster site inspection or search & rescue	Partial	Edge	Ad-hoc	Dynamic
XR physiotherapy	Partial	Cloud	Pre-deploy	Dynamic
Health wearables	Partial	Hybrid	Pre-deploy	Dynamic
Autonomous farming machinery	Partial	Hybrid	Pre-deploy	Static
Driver-assistance services for passenger vehicles	Partial	Cloud	Pre-deploy	Static

Timeline for offloading of applications

Use case	Now	Near future	Distant future
Mine inspection through autonomous vehicles	None	Hazard detection and motion control	
Driver assisted vehicles	Multimedia, user assistance, directions and control backups		
Service robots in public environments	Motion planning and perception	Localization, mapping and tracking	Full navigation stack and Low-level control
Mobile robots in factory facilities	Fleet control and mapping	Perception and SLAM	Full control stack
Autonomous farming machinery	Perception	Navigation and machinery control	
Health wearables	Data processing		
XR physiotherapy	Data processing and rendering		SLAM
Remote rendering	Rendering		SLAM

Offloading targets for different application areas

Application Domains	Mobile robots	Condition monitoring	XR
Manufacturing	X	X	X
Automotive	X		
Immersive interaction			X
Defense and railway		X	
Mining, construction and ports	X	X	
Agriculture	X	X	X
Healthcare		X	X

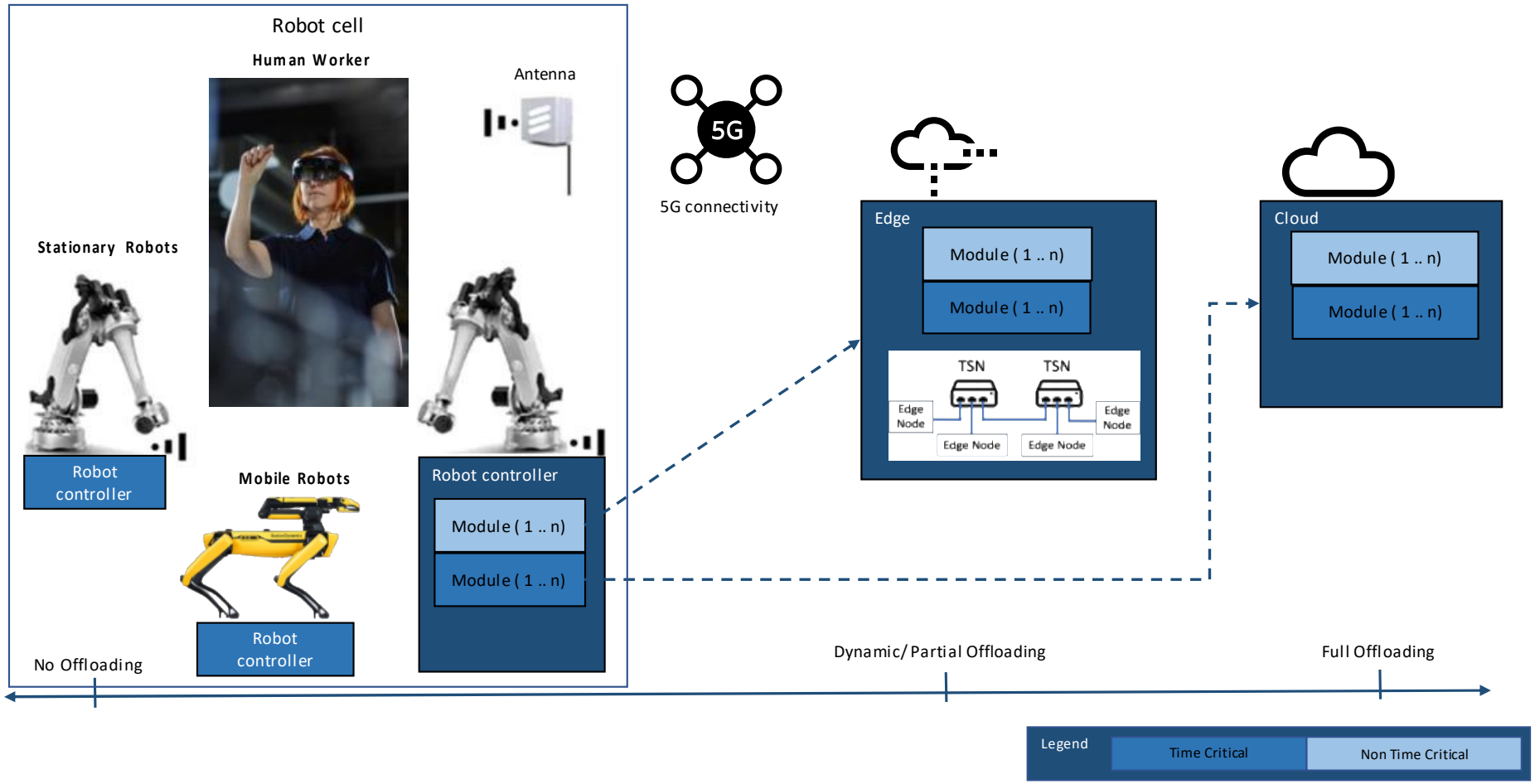
Overview of our Vinnova funded project AORTA

AORTA
(Advanced Offloading for Real-Time Applications)

- **Project key information:**
  - Start Date: January 2023
  - End Date: January 2026
  - Budget (total): 20 MSEK
  - Partners: Mälardalen Uni, Lund Uni, Cognibotics, Ericsson

- **Project ambition:**
  - Support advanced robotics and manufacturing applications in utilizing non-local services in a predictable fashion (ensure deterministic performance and support timing predictability of RT applications).

AORTA (Cont.)


Robot cell
Human Worker
Antenna
Stationary Robots
Robot controller
Mobile Robots
Robot controller
Robot controller
Module (1 .. n)
Module (1 .. n)
5G
5G connectivity
Edge
Module (1 .. n)
Module (1 .. n)
TSN
TSN
Edge Node
Edge Node
Edge Node
Edge Node
Cloud
Module (1 .. n)
Module (1 .. n)
No Offloading
Dynamic/ Partial Offloading
Full Offloading
Legend
Time Critical
Non Time Critical

Project implementation

- **WP1: Holistic modelling and resource analysis:** This WP will develop new techniques and a framework for holistic modelling, on-the-fly adaptation and offloading, and resource verification of real-time applications that utilize edge-cloud continuum and predictable networks like TSN and 5G.
  - T1.1: Develop holistic modelling and resource analysis techniques
  - T1.2: Develop on-the-fly adaptation and offloading framework for edge-cloud based real-time applications
- **WP2: Control Algorithm and Architecture:** This work package will develop the application part of the framework developed in WP1, providing the foundation for an ecosystem for real-time flexible mission-critical wireless automation components that use the edge and cloud for offloading.
  - T2.1: Dynamic and distributed edge and cloud-aware control systems
  - T2.2: Resource management for safety-critical collaborative robotics
- WP3: Industrial prototypes, demonstration, and validation:
  - T3.1: Use-case development and drafting of a virtual demonstrator
  - T3.2: Tailoring real-time computing to edge-cloud controller migration
  - T3.3: Develop and evaluate an integrated demonstrator prototype

Take-aways

- Some application domains as automotive are very restricted by safety regulations and offloading is losing importance
- Some others as manufacturing are very controlled and full offloading is a smaller challenge
- Applications with mobility and dynamicity seems more challenging and need to have dynamic offloading solutions in place.
- Collaborative applications or use cases seems to be one of the most promising cases for (dynamic) offloading as it can save communication bandwidth
- Interesting lines for future research:
  - Environments with limited/low quality network or network congested by devices
  - Offloading for multiple devices
  - High-mobility offloading