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

Degree of offloading
- No offloading
- Partial offloading
- Full offloading

Architecture
- Cloud offloading
- Edge offloading
- Hybrid offloading (edge and cloud)

Scheduling
- Static offload
- Dynamic offload

Initiator
- Device
- Edge or cloud
- Network

Optimization objectives
- Response Time
- Resource Optimization
- Deployment Cost
- Model Accuracy

Application challenges
- Privacy
- Mobility
- Multi-user

Resources deployment
- Adhoc
- Pre-deploy
Application domains and use cases
Mining: Mine inspection through autonomous vehicles

KPIs:
- Latency
- Uplink bandwidth

Applications to offload:
- hazard detection
- navigation

Heavy machinery close to the device can be used as mobile edge server

Bad network coverage due to limited radio penetration through rock

WiFi or private 4G/5G

Small robots

Heavy machinery

Central server and cloud connection
Automotive: Driver assisted vehicles

KPIs: Safety, Comfort, Convenience

Applications to offload:
- multimedia
- user assistance (e.g., voice recognition)
- direction finding (navigation)
- backup vehicle control

Partial offloading
Cloud offloading
Static offloading
Initiation on device
Service robotics: Service robots in public environments

Applications to offload:
- Localization
- Navigation and motion planning
- Perception
- Obstacle avoidance

KPIs:
- Cost
- Latency
- Reliability

Network coverage can be limited

Infrastructure sensors

Partial/Full offloading
Hybrid offloading
Dynamic offloading
Initiation on device
Manufacturing: Mobile robots in factory facilities

Stable private 5G network. Potential congestion of UE's

Applications to offload:
- Fleet control
- SLAM
- Motion planning
- Perception
- Control

KPIs:
- Latency
- Security
- Reliability

Automated forklifts
Mobile robots

Partial/Full offloading
Edge offloading
Dynamic offloading
Initiation on device
Farming: Autonomous farming machinery

- **Applications to offload:**
  - navigation
  - perception
- **KPIs:**
  - Latency
  - Revenue

- **Bad network coverage**

- **Applications to offload:**
  - navigation
  - perception
  - machinery control
- **KPIs:**
  - Cost
  - Obsolescence
  - Revenue

- **Drones**

- **Edge server**

- **Initiation on device**

- **Partial offloading**

- **Hybrid offloading**

- **Static offloading**
Healthcare: Health wearables and XR-assisted physiotherapy
XR: Remote rendering

Applications to offload:
- 3D simulation
- Rendering
- Gaming
- Training
- Monitoring
- Healthcare

Nearest cloud (local cloud)

Cloud

AR/VR user

KPIs:
- Bandwidth
- Latency
- Privacy

Partial offloading
Cloud offloading
Static offloading
Initiation on device
Commonalities between application areas

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

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

<table>
<thead>
<tr>
<th>Use case</th>
<th>Now</th>
<th>Near future</th>
<th>Distant future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mine inspection through autonomous vehicles</td>
<td>None</td>
<td>Hazard detection and motion control</td>
<td></td>
</tr>
<tr>
<td>Driver assisted vehicles</td>
<td>Multimedia, user assistance, directions and control backups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service robots in public environments</td>
<td>Motion planning and perception</td>
<td>Localization, mapping and tracking</td>
<td>Full navigation stack and Low-level control</td>
</tr>
<tr>
<td>Mobile robots in factory facilities</td>
<td>Fleet control and mapping</td>
<td>Perception and SLAM</td>
<td>Full control stack</td>
</tr>
<tr>
<td>Autonomous farming machinery</td>
<td>Perception</td>
<td>Navigation and machinery control</td>
<td></td>
</tr>
<tr>
<td>Health wearables</td>
<td>Data processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XR physiotherapy</td>
<td>Data processing and rendering</td>
<td></td>
<td>SLAM</td>
</tr>
<tr>
<td>Remote rendering</td>
<td>Rendering</td>
<td></td>
<td>SLAM</td>
</tr>
</tbody>
</table>
# Offloading targets for different application areas

<table>
<thead>
<tr>
<th>Application Domains</th>
<th>Mobile robots</th>
<th>Condition monitoring</th>
<th>XR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Automotive</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immersive interaction</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Defense and railway</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Mining, construction and ports</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Healthcare</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
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.)
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