Computation offloading in Edge and cloud environment

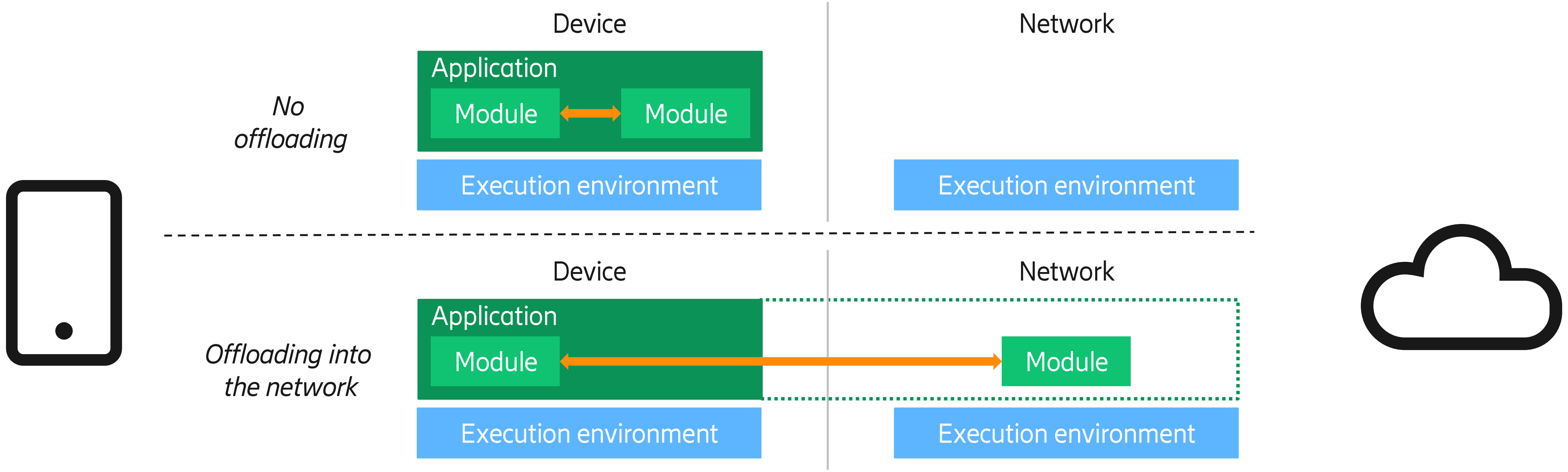
Ali Balador, Ericsson Research, Sweden

September 10, 2024

Compute Offloading

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 (battery, storage, processing capacity and network) to an edge server or cloud server, or from an edge server to a cloud server.
- Task offloading aims to **achieve performance objectives**, such as reducing overall computation time, minimizing network resources usage, maximizing battery life, among others.



Offloading classification

- Offloading
  - 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

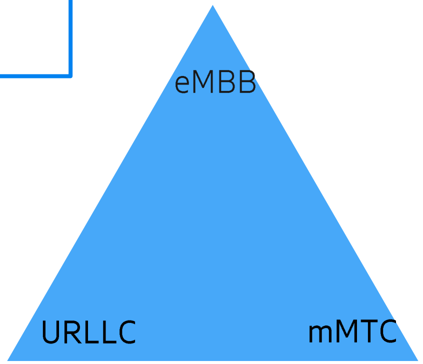
Reference: Offloading for the future: current use cases and scenarios, https://www.ericsson.com/en/blog/2023/8/computational-offloading-for-future-innovations.

6G and compute offloading

5G new services:

5G new communication services provide enhancements

- Mobile BroadBand (eMBB)
- Ultra-Reliable Low Latency Communication (URLLC)
- massive Machine-Type Communication (mMTC)



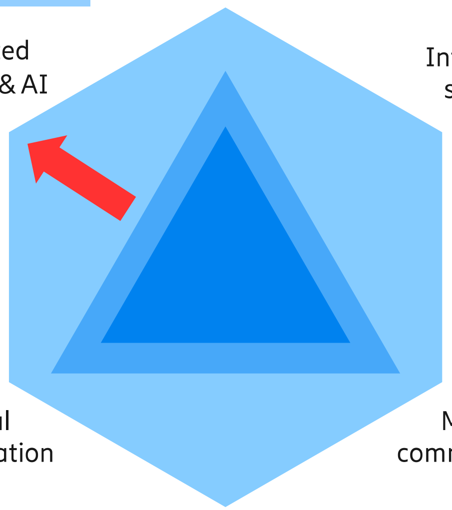
5G

These serve as a backbone of today's digital society.

Services beyond connectivity

Beyond-communication networks

New services on 6G platform



Network sensing

Positioning

AI as a Service

Compute services

Beyond-communication functions are offered by the network platform "as a service"

New services are an important opportunity to increase operator revenue

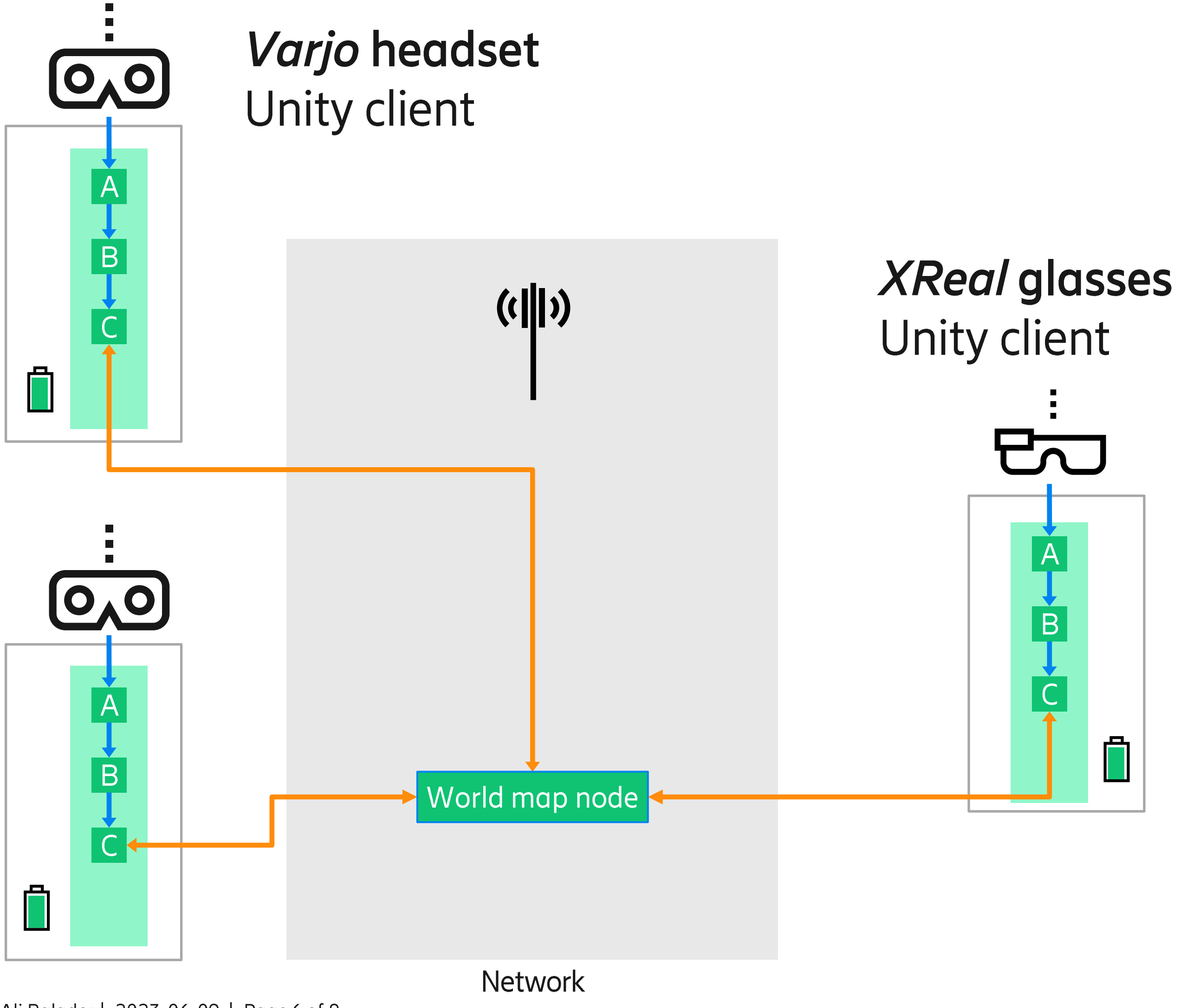
Positioning is a service in today's networks – others are either new or not yet offered by networks

References:

1-ITU Radiocommunication Sector (ITU-R), "Recommendation ITU-R M.2160-0: Framework and overall objectives of the future development of IMT for 2030 and beyond," Nov 2023.
2-ATIS NextG Alliance (NGA), "Roadmap to 6G," Feb. 2022.

Dynamic Device offloading

Case 1



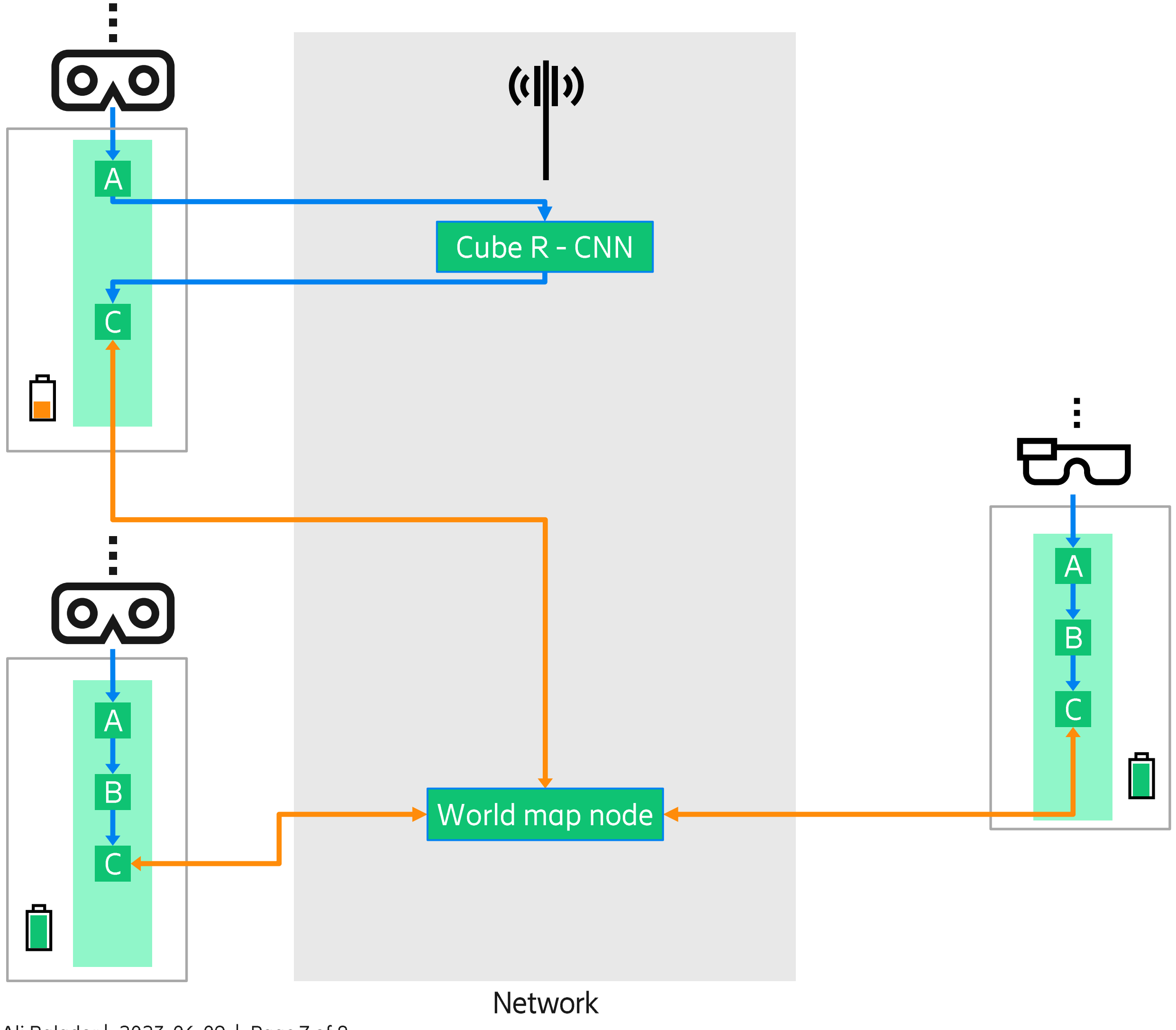
Multiple users wearing XR glasses would like to share a common reality.

Each client performs a series of tasks, which can be computationally taxing or require additional resources.

The communication task can be offloaded into the network to allow multiple devices to coordinate information.

Dynamic Device offloading

Case 1

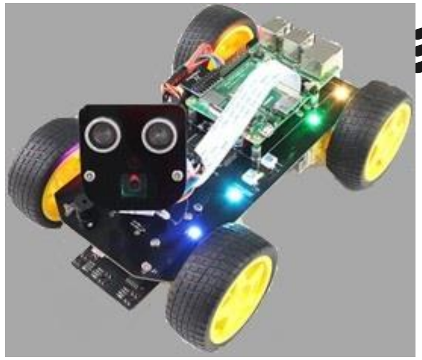


Object detection is a complex task that may run too slowly on a lightweight XR device.

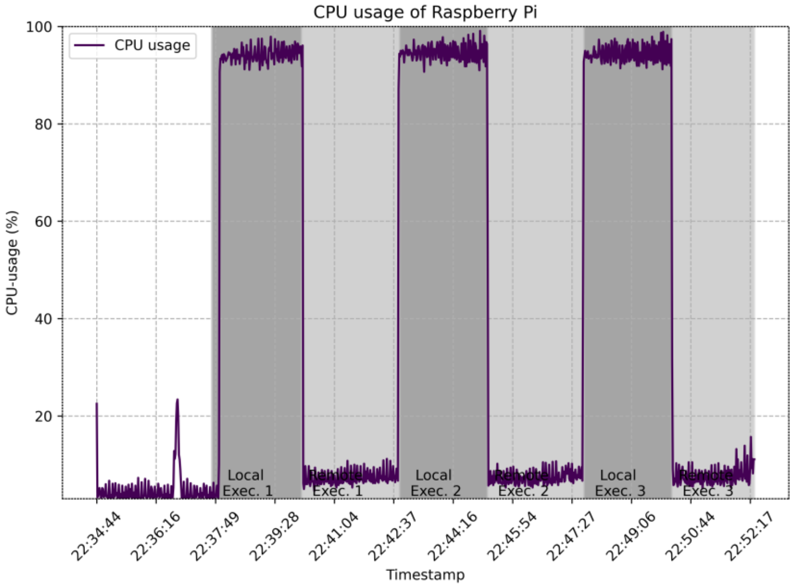
Or the same task may also drain the battery too quickly.

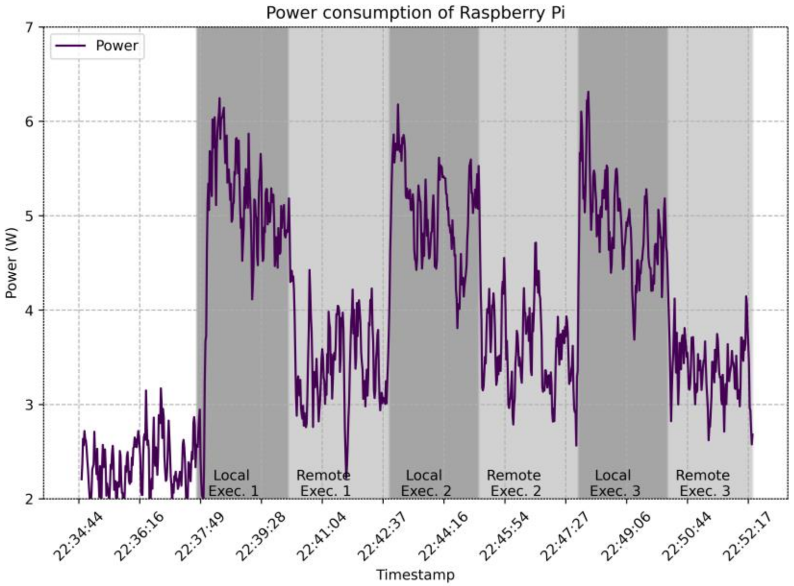
Offloading such complex tasks to the network allows the device to be lighter and the battery to last longer.

The gains and the costs …
Case 2



CPU Usage	CPU Temperature	Device Power Consumption	Network Usage	Performance

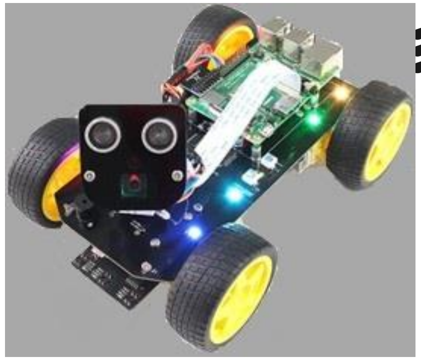




The gains and the costs …

References:

Benefits of Dynamic Computational Offloading for Mobile Devices, Vinay Yadhav, Andrew Williams, Ondrej Smid, Jimmy Kjällman, Raihan Islam, Joacim Halén, Wolfgang John, CLOSER 2024, https://www.scitepress.org/publishedPapers/2024/127198/pdf/index.html.

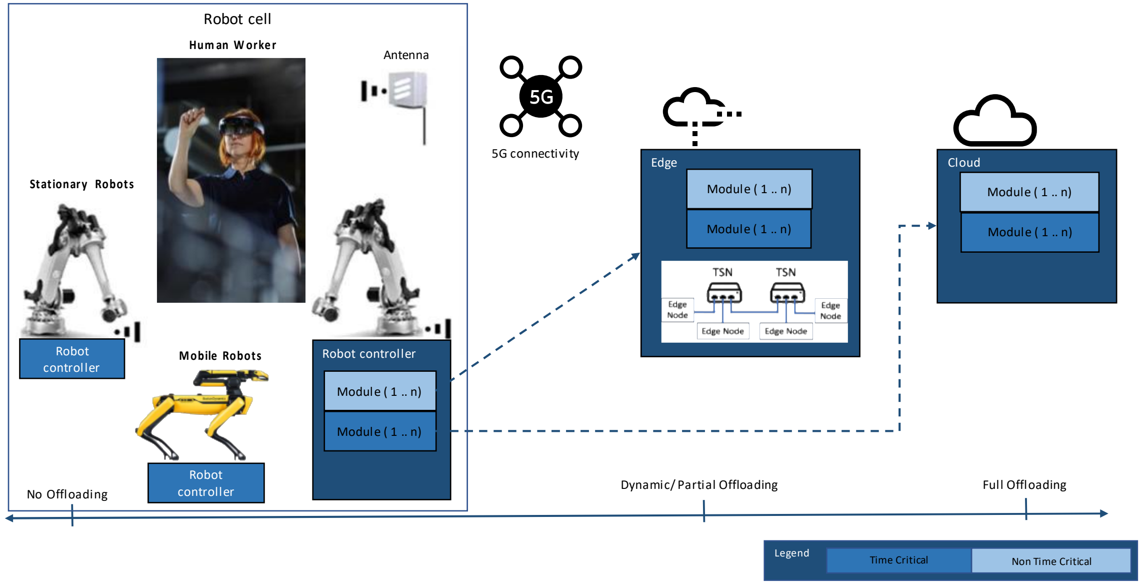


CPU Usage	CPU Temperature	Device Power Consumption	Network Usage	Performance

Power vs Performance		
Offloading Configuration	**Energy** / mWh per frame (on the device)	**Average Response Time** (ms) for object detection
Local Execution	2.02	1250
Remote Execution 5G	0.09	90
Remote Execution LAN	0.06	60

AORTA
Advanced Offloading for Real-Time Applications

- **Project key information:**
  - Start Date: April 2023
  - End Date: September 2026
  - Budget (total): 2 M euros
  - 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 real-time applications).

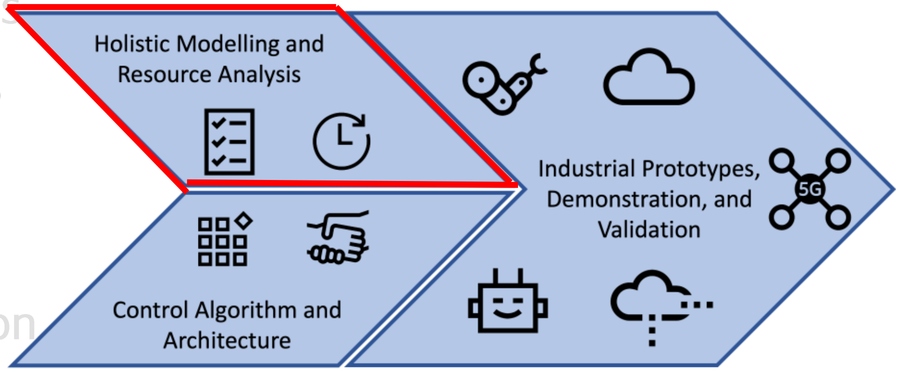


VINNOVA
Sweden's Innovation Agency

ERICSSON COGNIBOTICS Mälardalens universitet LUNDS UNIVERSITET

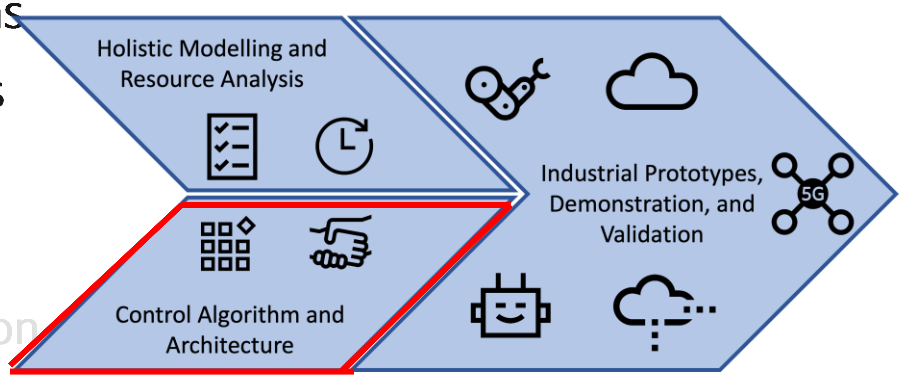
AORTA initial plans

- WP1: Techniques and algorithms for task offloading
  - Develop techniques and algorithms to enable online task offloading in edge-cloud computing architectures considering real-time properties of the tasks.
  - Develop techniques to evaluate the properties of tasks during and after task offloading.
  - Develop techniques to analyze timing properties and resource utilization of systems where tasks can be offloaded to edge-cloud during the system run-time.
- 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



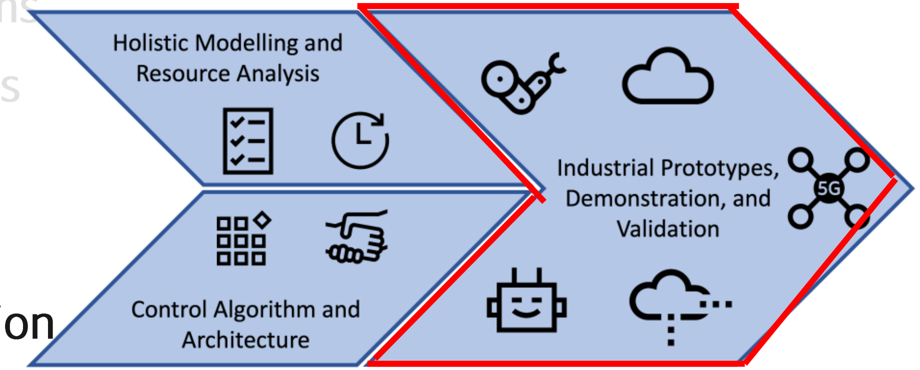
AORTA initial plans

- WP1: Techniques and algorithms for task offloading
  - Develop techniques and algorithms to enable online task offloading in edge-cloud computing architectures considering real-time properties of the tasks.
  - Develop techniques to evaluate the properties of tasks during and after task offloading.
  - Develop techniques to analyze timing properties and resource utilization of systems where tasks can be offloaded to edge-cloud during the system run-time.
- **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



AORTA initial plans

- WP1: Techniques and algorithms for task offloading
  - Develop techniques and algorithms to enable online task offloading in edge-cloud computing architectures considering real-time properties of the tasks.
  - Develop techniques to evaluate the properties of tasks during and after task offloading.
  - Develop techniques to analyze timing properties and resource utilization of systems where tasks can be offloaded to edge-cloud during the system run-time.
- 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

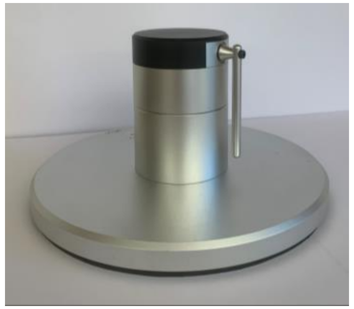


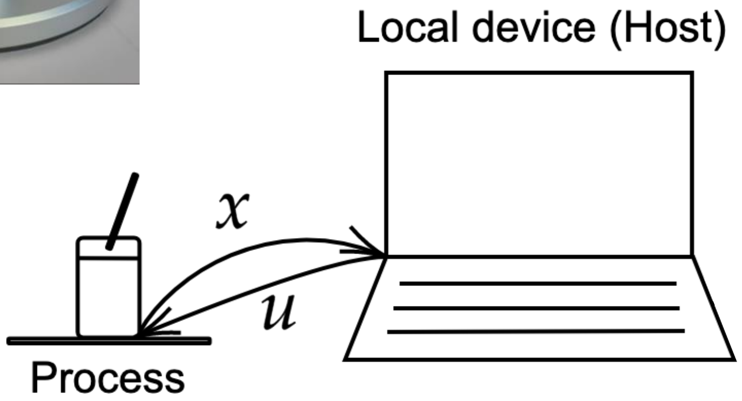
Use cases and objectives

- **UC 1: Offloading MPC controller for furuta pendulum**
  - Offloading using WASM
- **UC 2: Offloading motion (pose) planning for HKM robot**
  - Investigate which module benefits from offloading in this use case
  - How both decision-making component and motion planning works in a real setup
- **UC3: Collaborative offloading scenario**
  - Investigate challenges when different devices need to collaborate

Offloading MPC controller for Furuta Pendulum

Master thesis work in Lund University

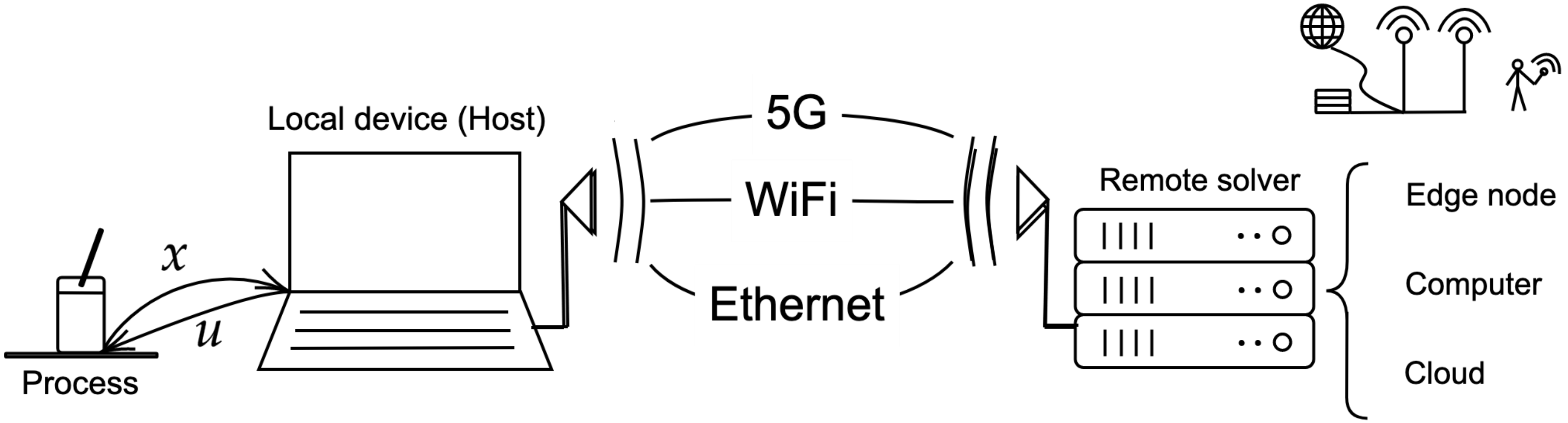




- First priority is stability
- Second priority is offloading

- Could be controlled using different control algorithms
- Here, LQR and MPC
- LQR (Linear–quadratic regulator)
  - Fast execution (one line of C)
  - OK control performance
- MPC (Model predictive control)
  - Quadratic optimization problem solved every sample → resource demanding
  - Better control performance
  - May benefit from edge/cloud execution

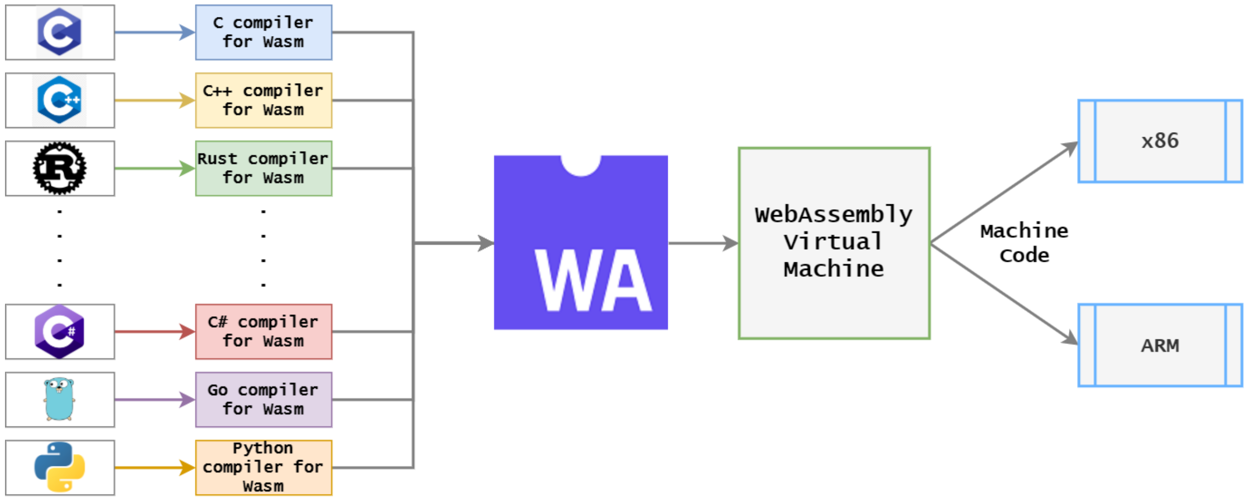
Overview


Local device (Host)
5G
WiFi
Ethernet
Remote solver
Edge node
Computer
Cloud
x
u
Process

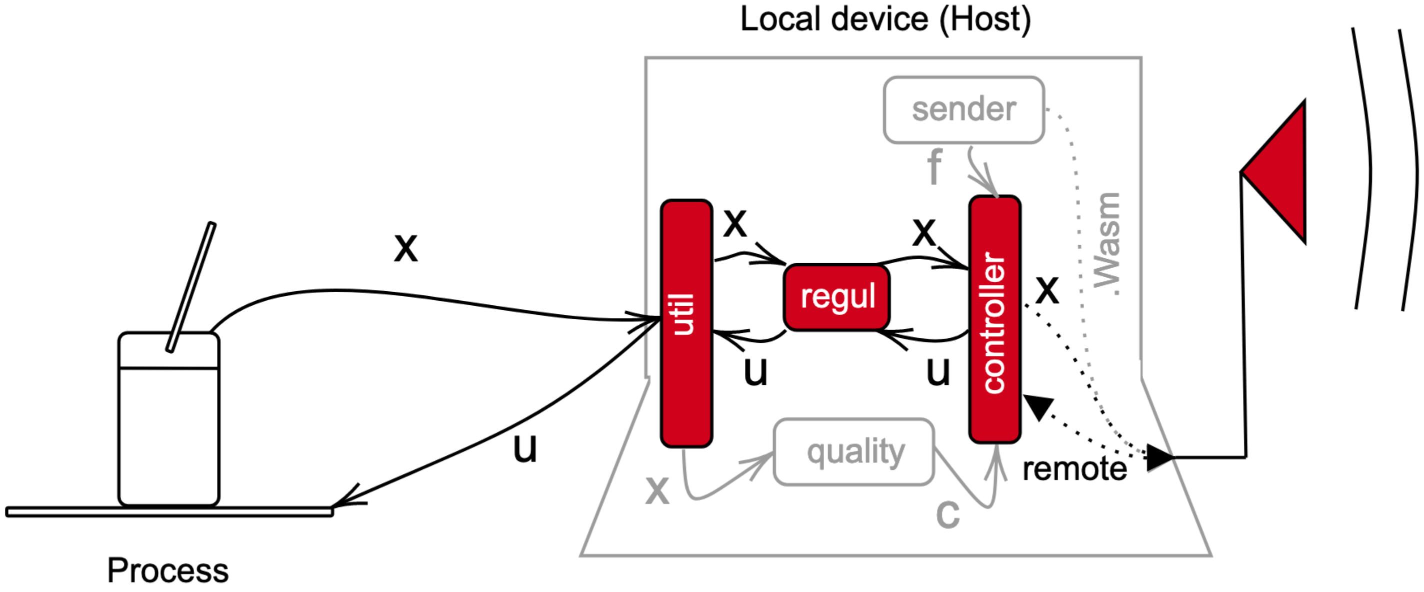
WebAssembly

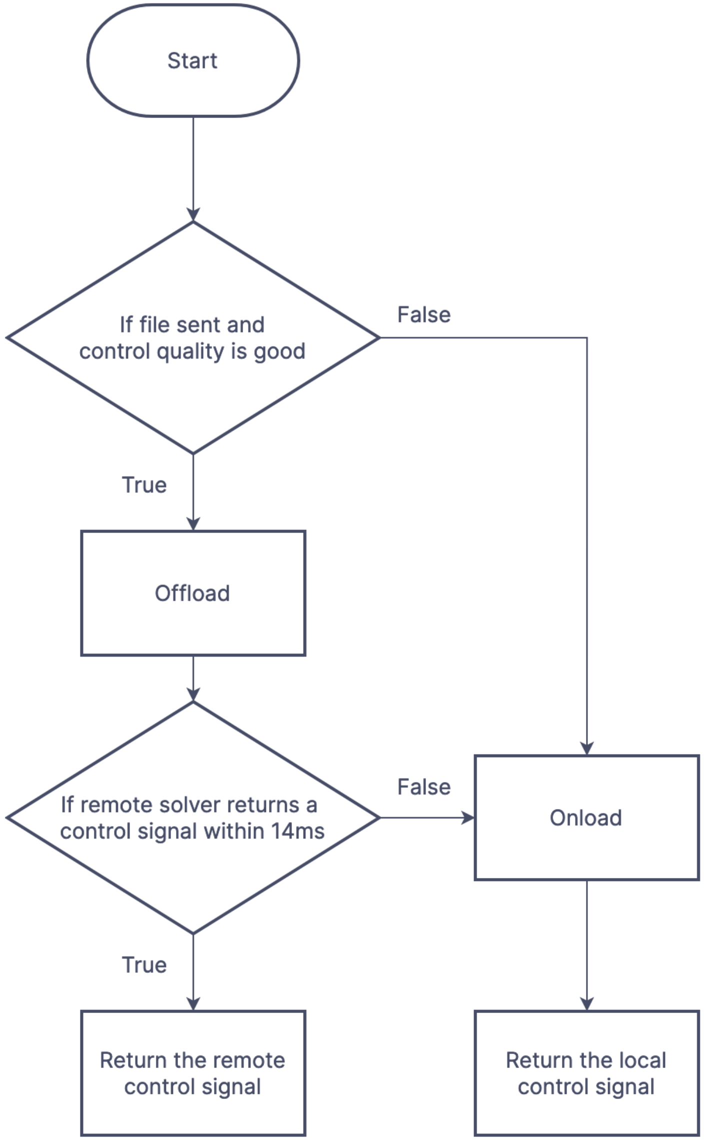
- WebAssembly (abbreviated Wasm) define a **portable**, size- and load-time-efficient **binary format** to serve as a compilation target which can be compiled to execute at native speed by taking advantage of common hardware capabilities available on a wide range of platforms, including mobile and IoT.

- Features
  - Efficient and fast (cold start time)
  - Portable
    - WebAssembly's binary format is designed to be executable efficiently on a variety of operating systems and instruction set architectures, on the Web and off the Web.
  - Secure (Sandboxed)
    - Each WebAssembly module executes within a sandboxed environment separated from the host runtime using fault isolation techniques.



- If the control application and the model parameters are sent and the control quality is good the thread will offload the control task to the remote solver using a UDP socket. Otherwise, the thread will onload the control task.
- If the thread chooses to offload it waits for the response from the remote solver. If the the thread get a response within 14 ms it will use the control signal in its next iteration. If the deadline is missed the thread will use the local controller instead to get a control signal.





Round-Trip Time incl MPC computations

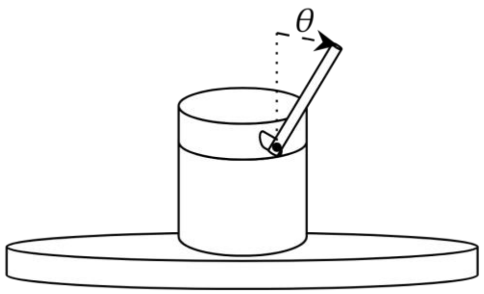
Execution Mode	Controller and Communication	Average RTT (ms)	Worst-Case RTT (ms)
Device	LQR	0.00012	0.032
	MPC C-code	1.3	6.4
	MPC WASM	3.1	10
Device w containers	MPC WASM	4.3	13
Edge using containers	MPC WASM over 5G	11	20
	MPC WASM over Wifi	6.3	15
	MPC WASM over wired Ethernet	5.1	14

Round-Trip Time incl MPC computations

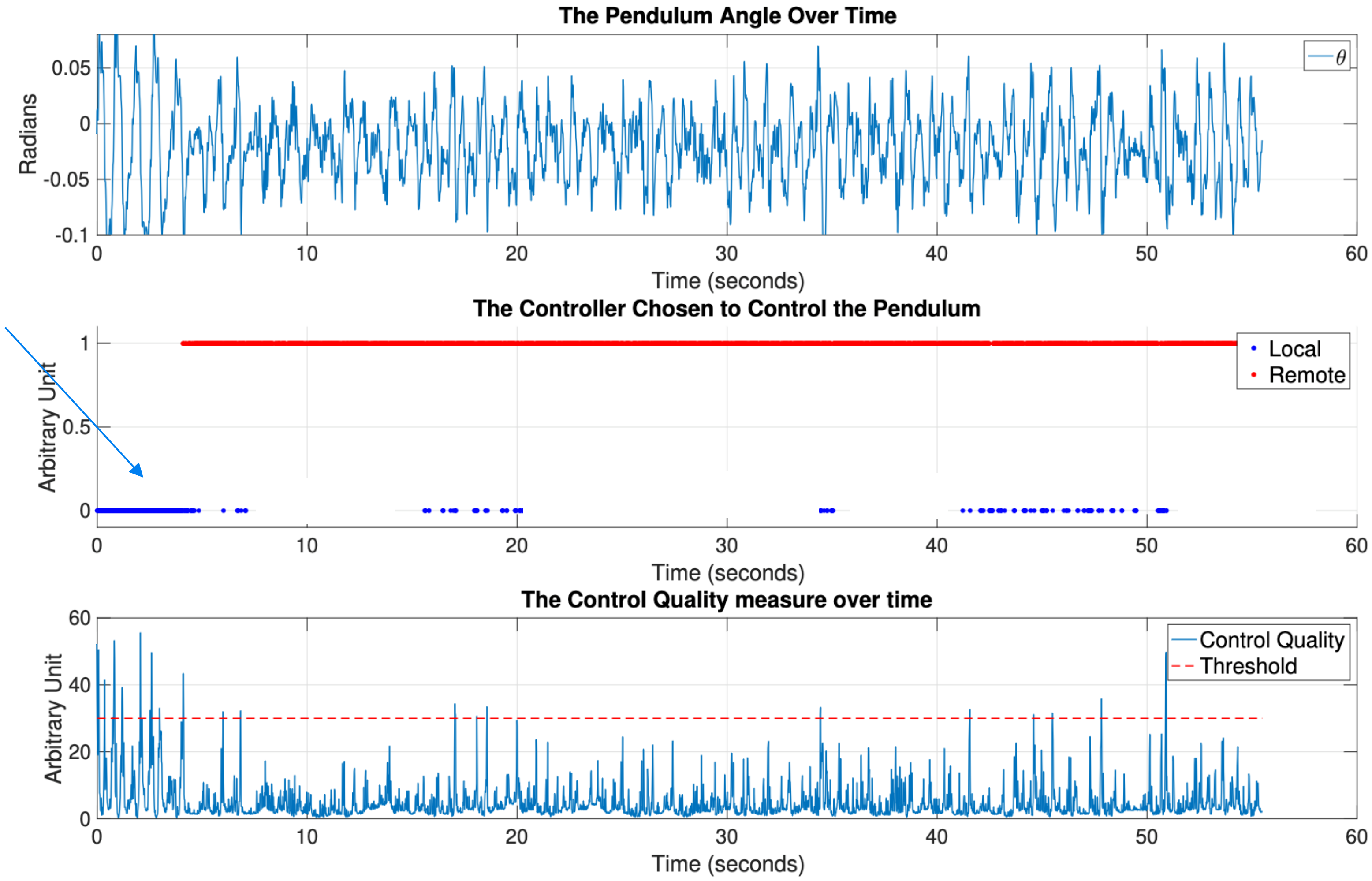
Execution Mode	Controller and Communication	Average RTT (ms)	Worst-Case RTT (ms)
Device	LQR	0.00012	0.032
	MPC C-code	1.3	6.4
	MPC WASM	3.1	10
Device w containers	MPC WASM	4.3	13
Edge using containers	MPC WASM over 5G	11	20
	MPC WASM over Wifi	6.3	15
	MPC WASM over wired Ethernet	5.1	14

Demonstration Data

Reference: Ahmed Al Bayati, Karl-Erik Arzzen, “Dynamic Offloading of Control Algorithms to the Edge using 5G and WebAssembly”, Link: https://www.ecrts.org/wp-content/uploads/2024/07/Dynamic-Offloading-of-Control-Algorithms.pdf



Code offload



Additional references

- Ali Balador, Johan Eker, Raihan Ul Islam, Raquel Mini, Klas Nilsson, Mohammad Ashjaei, Saad Mubeen, Hans Hansson, Karl-Erik Arzen, AORTA: Advanced Offloading for Real-time Applications, RT-Cloud 2023, https://retis.sssup.it/luca/RT-Cloud23/RT-Cloud23_paper_2.pdf.
- Opportunities with dynamic device offloading as a 6G service, Ericsson blogpost, https://www.ericsson.com/en/blog/2023/9/dynamic-device-offloading-as-a-6g-service.
- Benefits of Dynamic Computational Offloading for Mobile Devices, Vinay Yadhav, Andrew Williams, Ondrej Smid, Jimmy Kjällman, Raihan Islam, Joacim Halén, Wolfgang John, CLOSER 2024, https://www.scitepress.org/publishedPapers/2024/127198/pdf/index.html.