Tampere University

FIIF Event with AI for Situational Awareness (AISA) project :

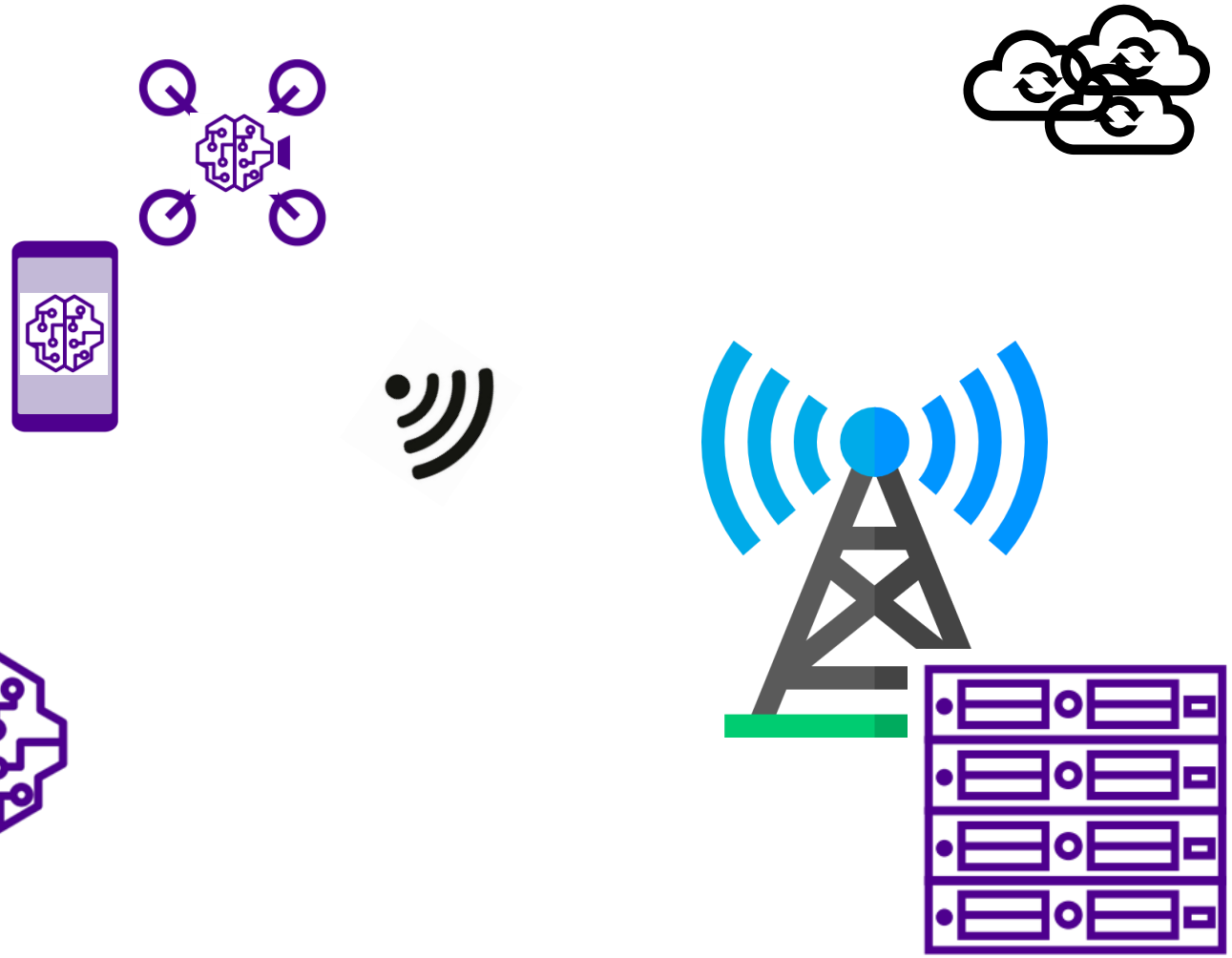
Edge Software Stack for Portable AI Offloading from Battery-Powered Devices

Pekka Jääskeläinen (pekka.jaaskelainen@tuni.fi)
Customized Parallel Computing (CPC) group (https://tuni.fi/cpc)

In Finnish Industrial Internet Forum’s (FIIF) Event with AISA Project: AI for Situational Awareness (Nov 21 2024, Tampere and online)

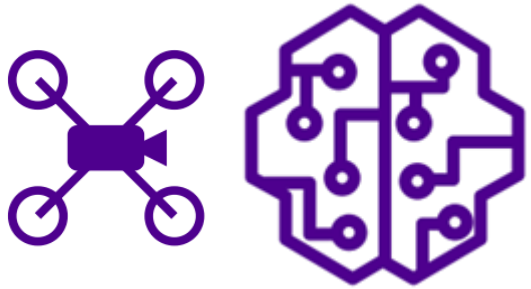
The basic concept of offloading compute to the ”edge of network”

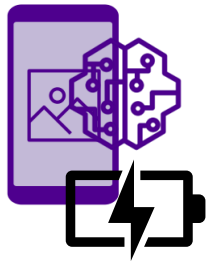
- Instead of performing heavy computation on a “local device”…
- “Offload” it to a remote server that is close to the network access point
- Similar to “cloud computing”, but supports latency-critical real-time tasks

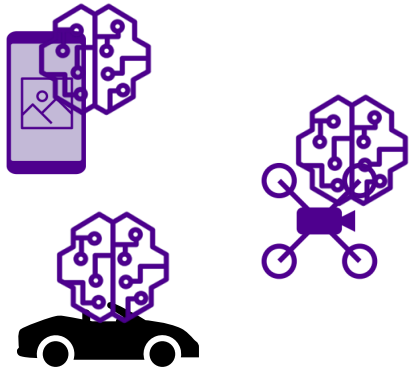


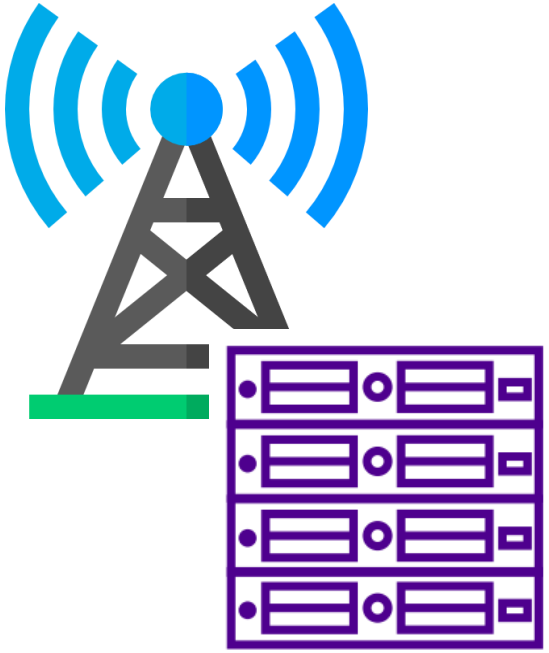
Motivations for edge offloading

- Perform complex processing for simple local devices
  - Mobile devices have limited computational capabilities
  - Server-side can provide 100X the performance
- Local device battery saving
  - Computation consumes energy
  - Spend it instead on the remote server which is connected to the grid
- Improve remote hardware utilization
  - Multiple users on the same edge GPUs minimize the idle time of the expensive hardware









Edge offloading challenges and our software stack's solutions

Key challenges:

- **Low latency responses**
- **Hardware vendor neutral heterogeneous computing**
- **Mobile device roaming**

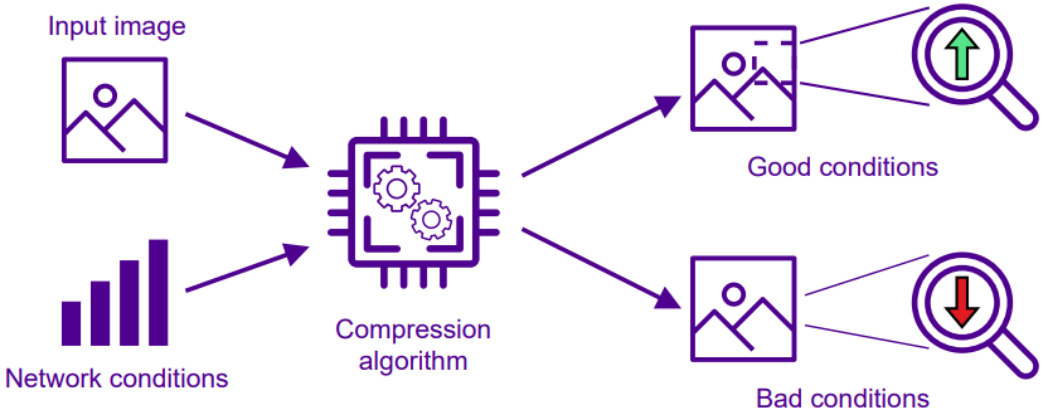
AISA edge offloading framework:

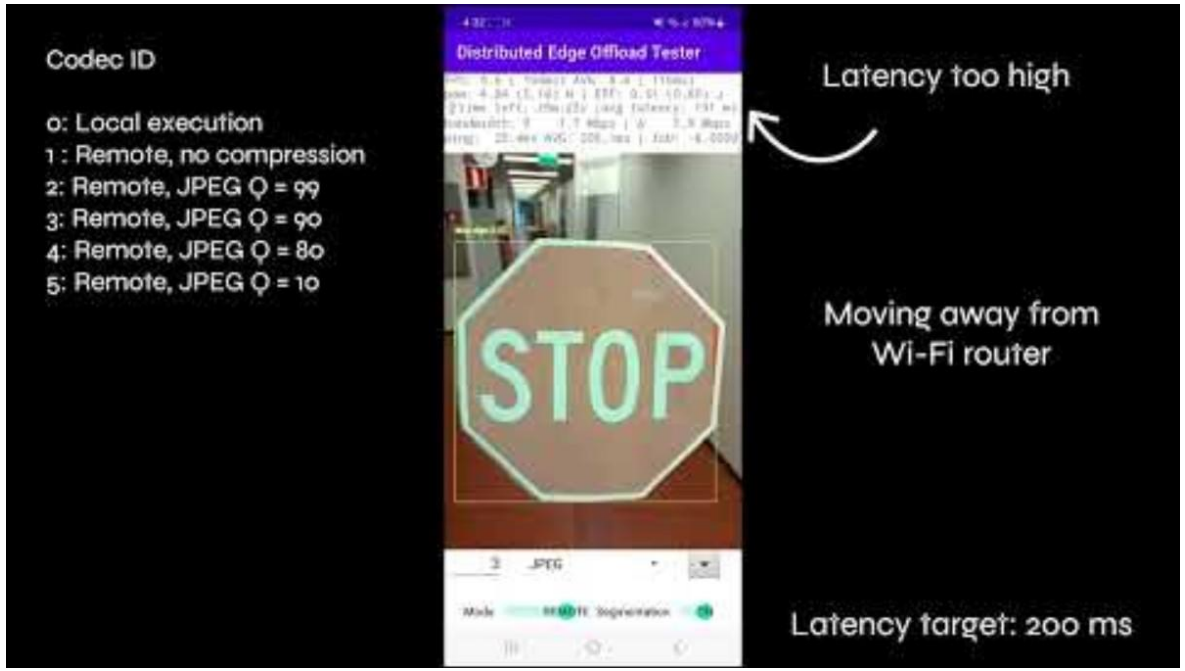
- **Lightweight edge software stack with few layers - all transparent and optimized together**
- **Use efficient networking solutions (e.g. RDMA)**
- **Adaptive lightweight image compression for low latency computer vision**
  - **Compression quality tuned based on the AI results**
- **Cross-vendor OpenCL API in the core of the open source software stack:**
  - **Open standard for heterogeneous diverse computing for CPU, GPU, DSP, FPGA**
- **Edge server-to-server buffer migration**
- **Automated server discovery**

Demo booth: Edge computing demo 1

Automatic image compression based on offloading circumstances

- Automatically increase or decrease compression bitrate (image quality) to maintain target latency
- Decisions driven by monitoring the current latency which depends on the current network conditions
- Compression choices:
  - JPEG
  - HEVC – WIP

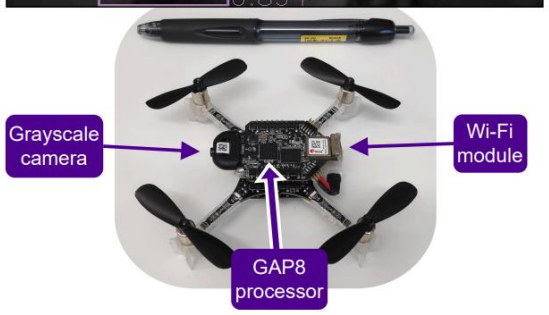




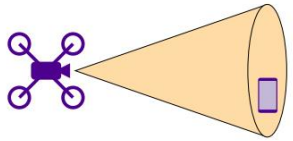
Demo booth: Edge computing demo 2

Application: remote offloading from a lightweight nano-drone

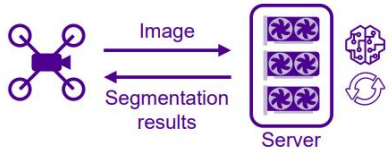
- Nano-drone offloads heavy computing workload while controlling the application logic
- Drone runs on a single RI5CY core (250 MHz) with 512 KB on-chip RAM
- Only the AI inference model is 14 MB



① Capture image of surroundings



② Send image to server for DNN segmentation



③ Adjust position based on segmentation results

Conclusions and Future work

- A software framework for heterogeneous compute edge offloading
  - Open standard-based
  - Resource discovery
  - Automatic image compression adaptation with AI-result quality tuning
- Thanks to AISA, the framework is now stable enough for out-of-the-lab testing

Next:

- Secure edge offloading
  - Encryption, authentication, server-side security via isolation and execution risk levels
- Can we implement a peer-to-peer compute resource marketplace?
  - Reduce idle time of local GPU resources by enabling renting your computers to nearby edge offload users
- Towards supercomputer-level multi-device scalability

Tampere University

Contact information and links

- Demo videos available in https://tinyurl.com/edgeaisw
- The software stack is published in open source
  - https://code.portablecl.org
  - Look for ”the remote driver”
  - Liberal MIT licence
- Meet you at the TAU demo booth!

Customized Parallel Computing group

Pekka Jääskeläinen
Associate Professor

pekka.jaaskelainen@tuni.fi
+358407390750

tuni.fi/cpc

github.com/cpc