Tampere University

AI for Situational Awareness (AISA):
Situation-Aware Platform for Smart Industry

Edge Computing Software Stack for Reliable Artificial Intelligence Roaming

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

In Finnish Industrial Internet Forum's AI in Industry Networking Seminar
January 27, 2022

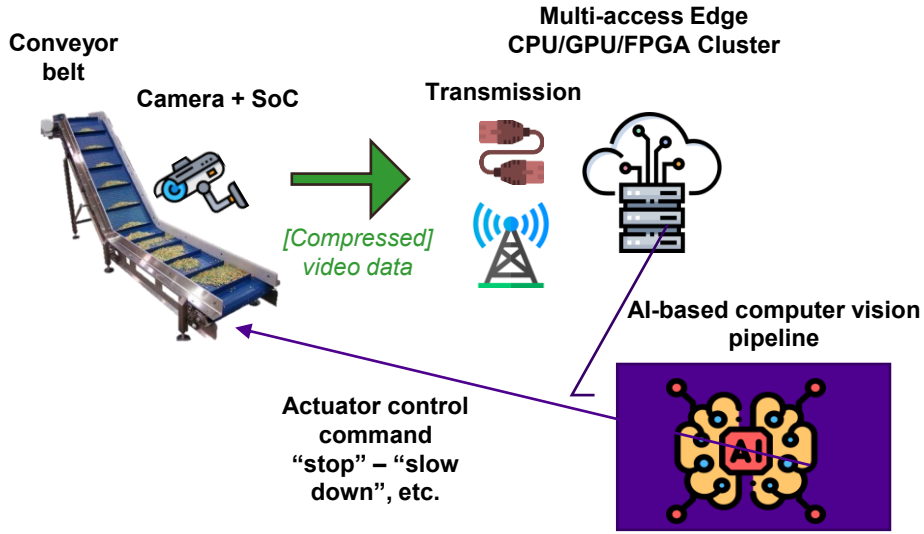
Edge Computing Software Stack for Reliable Artificial Intelligence Roaming - Example application domain:

Smart industry

Examples: A light weight video camera based product quality control loop using AI compute clusters servicing multiple networked cameras or manufacturing robot control.

Technical challenges:

- **Low latency responses**
  - Combined with network distributed compute offload leads to a challenging whole stack optimization problem
- **Reliability**
  - When computing decisions in life-critical environments (people inside the factory)
- **Efficient compute resource sharing**
  - Together with low latency quarantees
  - More of a problem in large scale deployments with hundreds of cameras



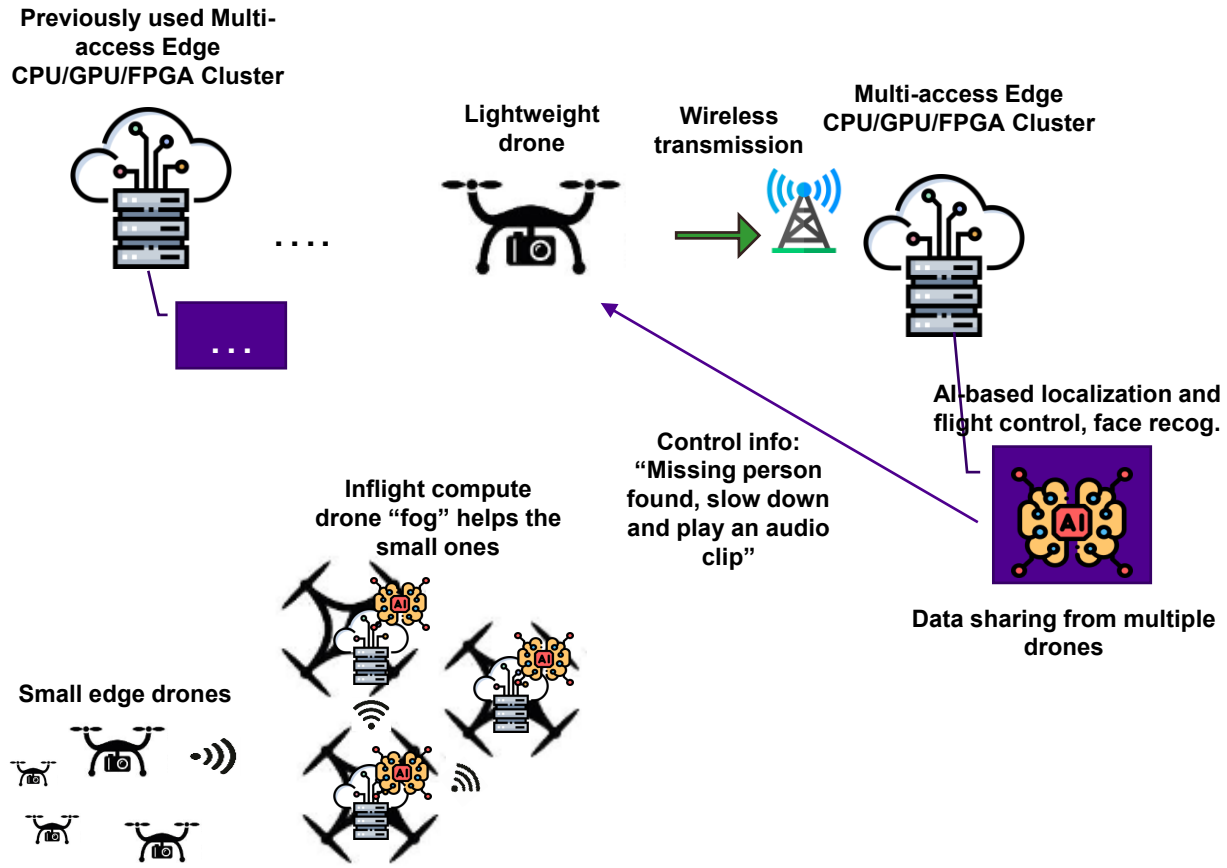
Edge Computing Software Stack for Reliable Artificial Intelligence Roaming - Example application domain:

Autonomous Drones

Examples: Drones for warehouse management.
Swarms of nano-sized drones used e.g. in rescue, police or military operations.
Drones with limited near-sensor compute capabilities.

Additional technical challenges:

- **Compute roaming**
  - Seamless compute continuation when switching from one compute cluster to another (roaming)
- **Diverse heterogeneous** computing
  - Compute servers and the local system have various types of CPUs, GPUs, DSPs and FPGAs from various processor companies
- **Adhoc compute cloud formation**
  - Efficient workload distribution to peers



Edge Computing Software Stack for Reliable Artificial Intelligence Roaming - Example application domain:

Automotive

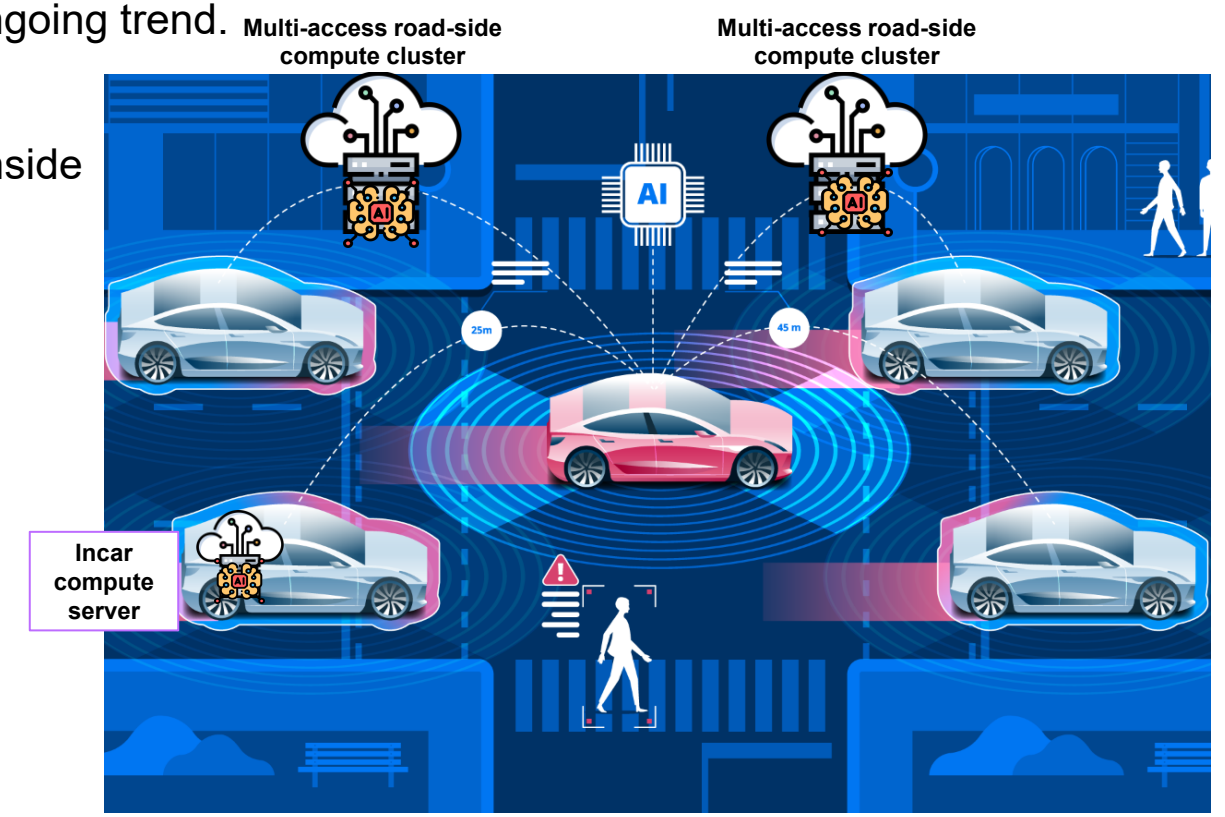
The ongoing compute chip shortage and the green transition has forced the car industry to consider **compute economy, ecology and supply security**. Reducing the use of chips via compute centralization and sharing of compute resources for multiple processing tasks is an ongoing trend.

Examples:

- In-car compute server servicing near-sensor devices inside the car
- Road-side servers servicing multiple cars passing by

Has all of the mentioned technical challenges:

- **Low latency responses**
- **Reliability**
- **Efficient compute resource sharing**
- **Diverse heterogeneous computing**
- **Compute roaming**
- **Adhoc compute cloud formation**



https://www.smartcitiesworld.net/opinions/opinions/driving-autonomous-vehicles-forward-with-intelligent-infrastructure

Tampere University

Edge Computing Software Stack for Reliable Artificial Intelligence Roaming: Customized Parallel Computing research group's research

Challenges:

- **Low latency responses**
- **Reliability**
- **Efficient compute resource sharing**
- **Diverse heterogeneous computing**
- **Compute roaming**
- **Adhoc compute cloud formation**

Technology we are researching:

- **Lightweight software stack with few layers - all transparent and optimized together**
- **Harness efficient networking solutions (e.g. RDMA and SmartNICs)**
- **Lightweight image compression for low latency computer vision (only machines, no humans in the loop)**
- **Automated critical function replication across heterogeneous devices within latency limits**
- **Bare metal server solutions for efficient and fine-grained compute resource sharing**
- **OpenCL API in the core of the open source software stack: Open standard for heterogeneous diverse computing for CPU, GPU, DSP, FPGA**
- **Proactive compute context switching/transfer tech.**
- **Intelligent discovery, probing and connect**

Tampere University

Contact Information

Customized Parallel Computing &
Virtual-reality and Graphics Architectures
groups

Pekka Jääskeläinen
Associate Professor

pekka.jaaskelainen@tuni.fi
+358407390750

tuni.fi/cpc
tuni.fi/vga

github.com/cpc