Avancerad Digitalisering

VINNOVA

Smart and Secure Gateways for a Secure Internet of Things

Mötesplats Avancerad Digitalisering 2023

Joakim Eriksson, RISE
Mohammad Riazati, Imsys
2023-05-26

imsys

IoT BRIDGE

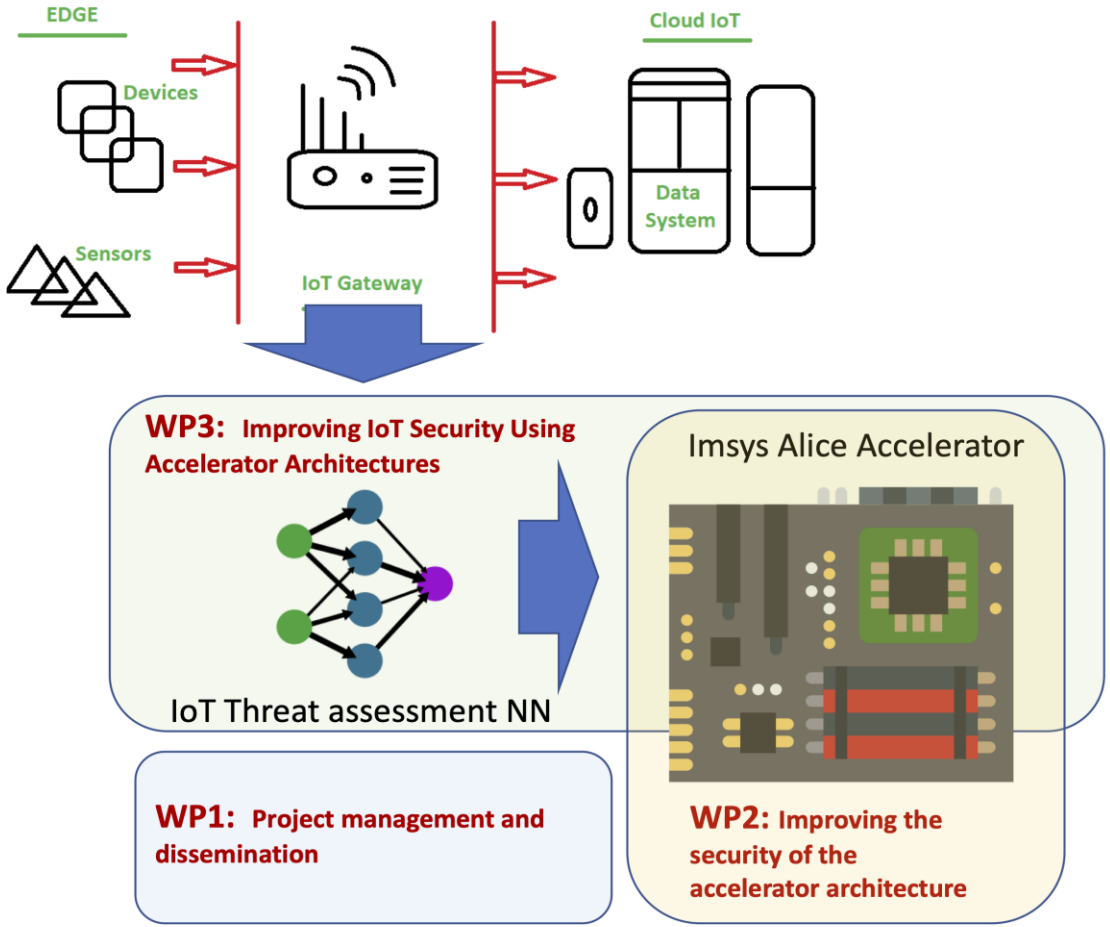
WITRA

RI. SE

UBSALIENSIS S·ACADEMIAE GRATIAE VERITAS NATURAE

Smart and Secure IoT Gatway

- Project Duration 2021-07-01 – 2023-12-31
- Partners
  - **RISE** - project lead and use-case provider: IoT – wireless device fingerprinting
  - **Uppsala University** – research and development of security mechanisms for the AI accelerator
  - **Imsys** – design and implementation of secure AI accelerator
  - **IoT Bridge** – IoT company – use-case provider:
    Bridge Safety IoT application using the AI accelerator
  - **Wittra** – IoT company – use-case provides:
    Secure IoT for asset tracking and asset lock system
- Main contribution
  - Smart and Secure IoT Gateway based on Imsys Alice AI-accelerator
  - Use-cases evaluating Imsys AI accelerator
  - Energy-efficient AI at the extreme edge

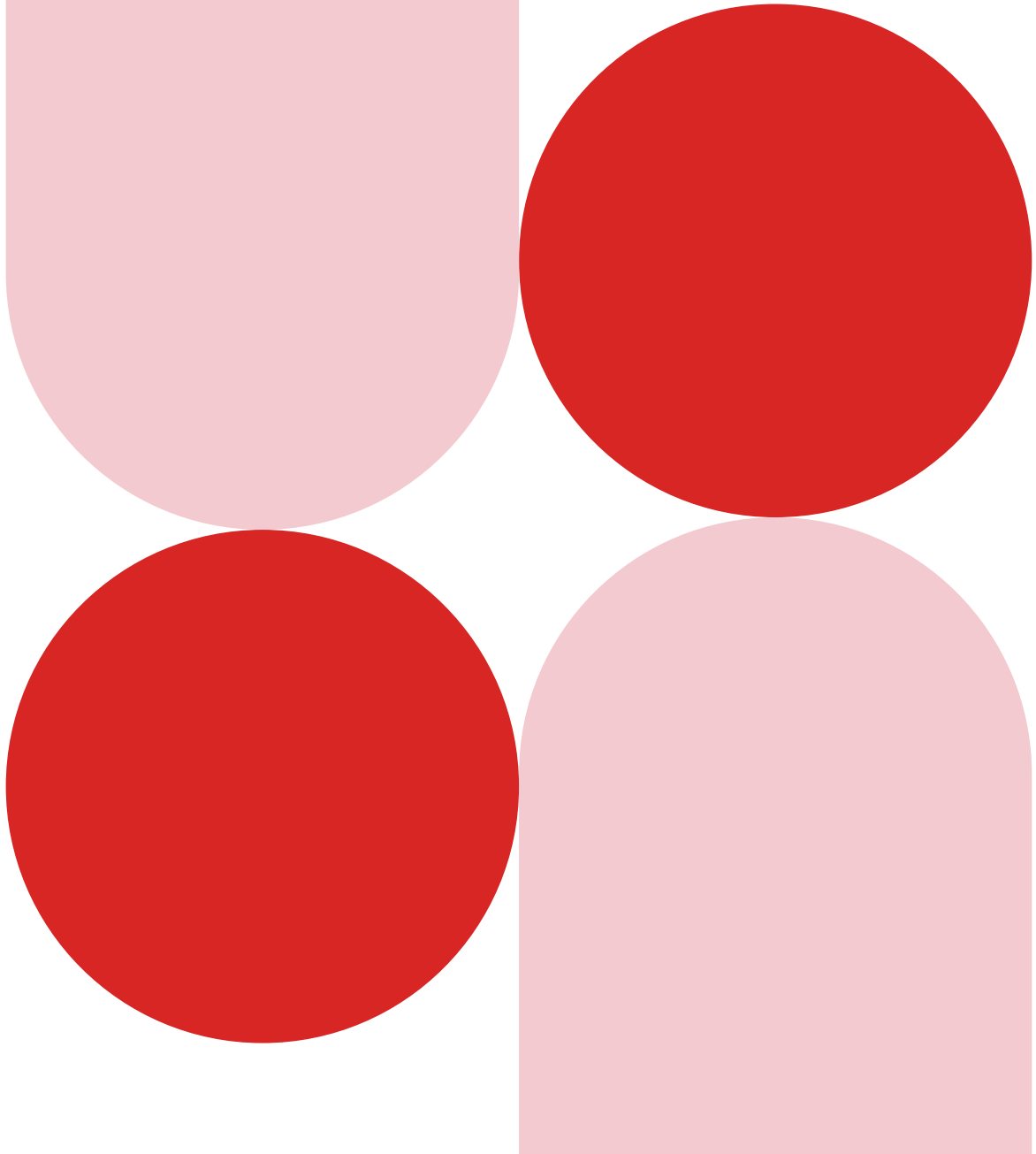


AI Acceleration in SecureGW

Mohammad Riazati

Senior AI Software Engineer

Imsys take on AI acceleration

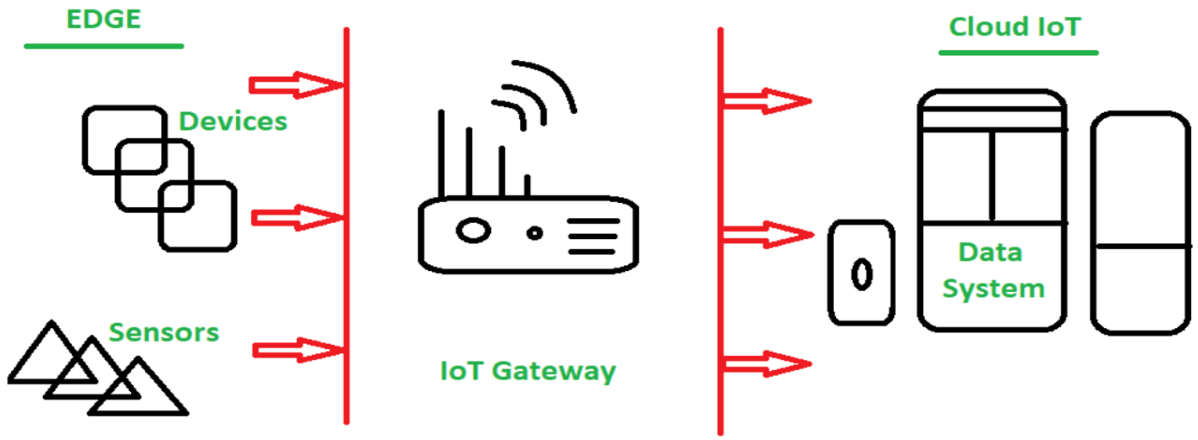


imsys

Why inference at the edge?

- Self-Contained (mission criticality)
- Resource restricted
- Fast response
- Lot of data

AI in smart sensors or aggregation point



Imsys Alice accelerator

High throughput, 10 TOP/S

Energy efficiency, 4 TOP/J

Scalable and programmable

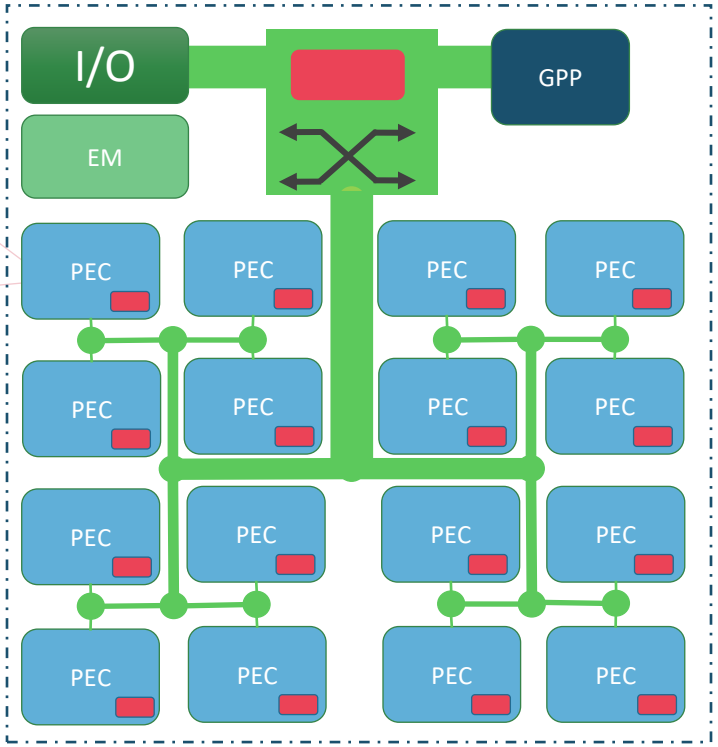
imsys

Alice Accelerator Platform Architecture

- **Many-Processor Architecture**
  - Processing Element Clusters
    - Shared memory
    - 16 Processing Elements each with 8 MAC units.
  - Network on chip
    - High speed transport between I/O and PEC
    - Data exchange between PECs
  - Prepared for tiling and inclusion in a system on chip
  - Tools support to avoid the use of caches.
- **The General-Purpose Processor**
  - Sequences the model execution
    - Using the Deep Neural Network Instruction Set of the PEC
    - Using the NoC data transport and switching capabilities.
  - The GPP can act in cooperation with a system processor in a host system, like the Secure GW, or manage the whole application on a smart sensor.
- **I/O**
  - External memory and highspeed interfaces (PCIe, USB4.0, etc.)
- **Energy manager (EM)**
  - Sleep modes
  - Performance (Supply voltage versus clock speed)
  - External power source (Manage energy bursts for battery operated devices)

imsys

Flexible
Deep
Neural
Network
Instruction
Set
(ISA-A)

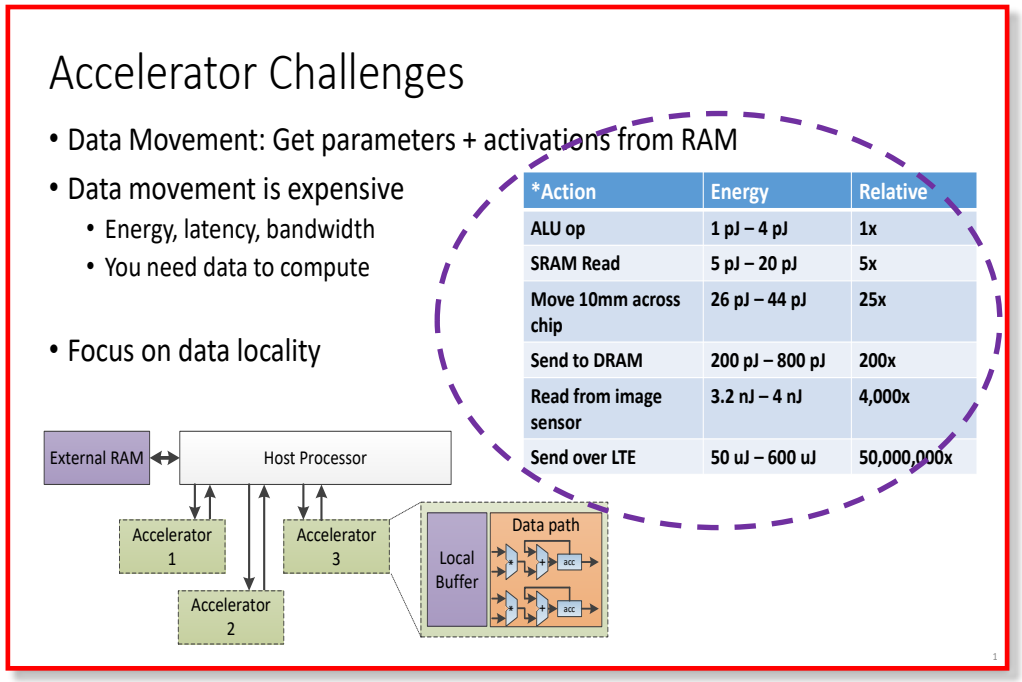


GPP: General-Purpose Processor
EM: Energy Manager
PEC: Processing Element Clusters
ISA-A: Instruction Set Architecture for Accelerators

The accelerator in this project

- Accelerator on Simulator and Emulator made available.
- Security architecture introduced.
- Firmware to execute application models efficiently upgraded.
- Automation tools for model implementation adapted.

The Imsys Accelerator Design for Low energy



DSD 2018 AMDL Keynote, Prof. Dr. Henk Corporaal
Sources of energy consumption challenging the system solution

Don't move data around
- Automated tools for data flow analysis
- Cache-less memory access
- Data reuse
- Processing near memory

Efficient processing
- Lean data types (uint8 most efficient)
- Low power circuit design matching architecture and advanced technology nodes* for system on chip implementation.

Quantized models proved to have same precision to less than a quarter of the energy.

Automated application optimization
- Minimize memory usage & maximize utilization
- Layer fusion, zero pruning, operator fusion, …

imsys

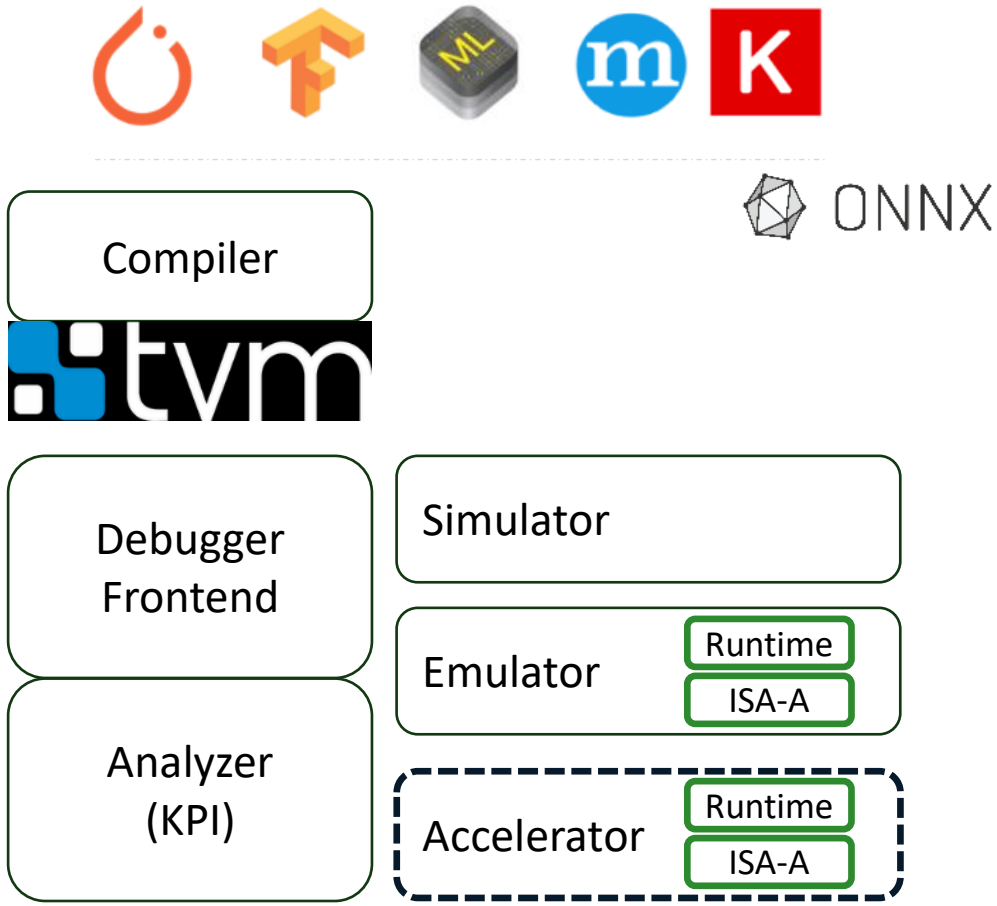
* Synthesized on Global Foundries 22nm FDX process

SecureGW demonstrator platform

imsys

Optimizer, Compiler, and Runtime

- Supports development flow from inference model graph to optimized target object code, which is used by the GPP to execute the model
- Quantization support
  - Training aware
  - Post training
- Customizable optimization
  - pipelining, layer fusing, memory usage …
- Seamlessly integrates with existing AI development frameworks



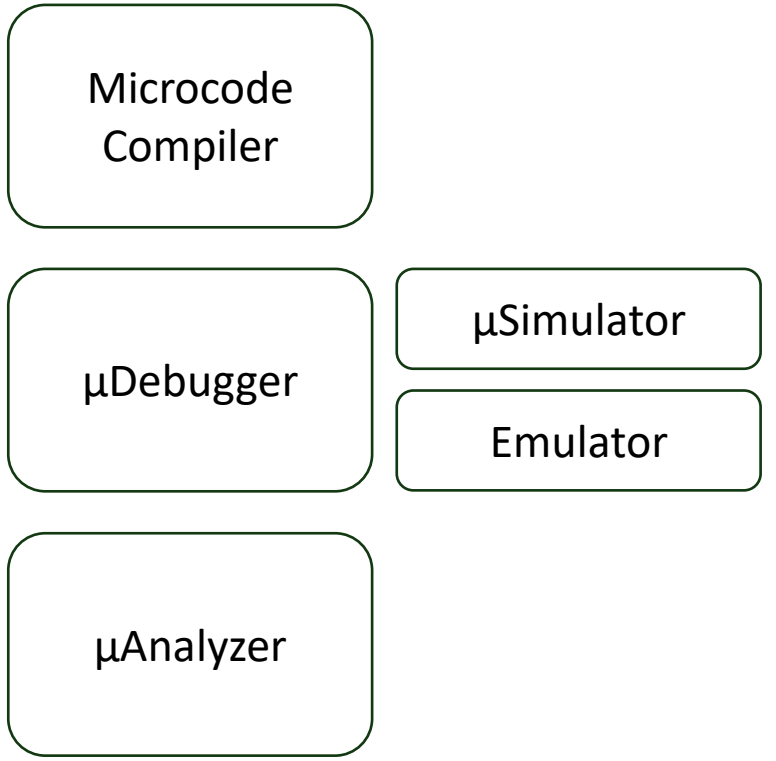
Accelerator on pre-production chip in follow-up project for interested participants

imsys

SDK for extending the DNN Instruction set (ISA-A)

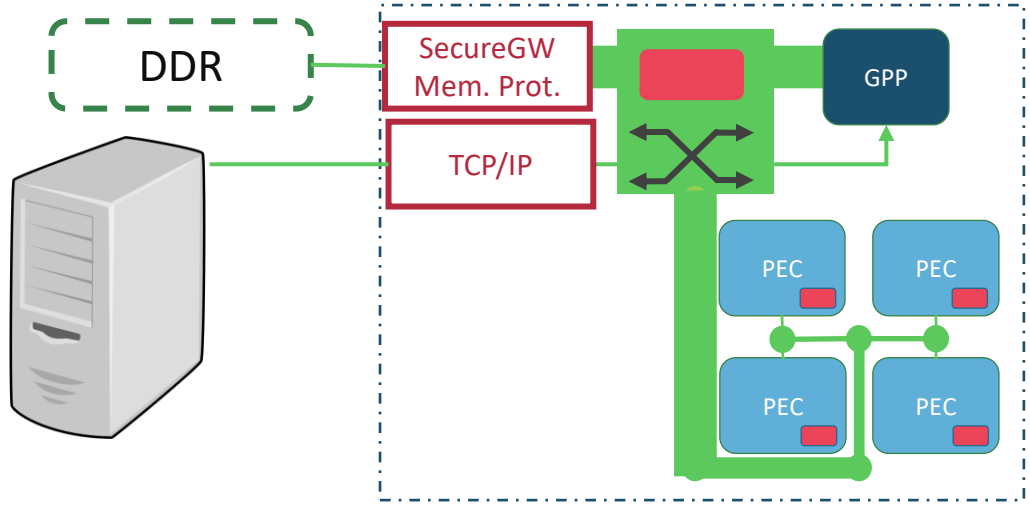
- ISA-A:
  - Instruction Set Architecture for Accelerators
- Library of instructions
  - Extensive instructions for quantized neural network operations
    and other kernel-based operations, e.g., FFT
- Programmable user customization

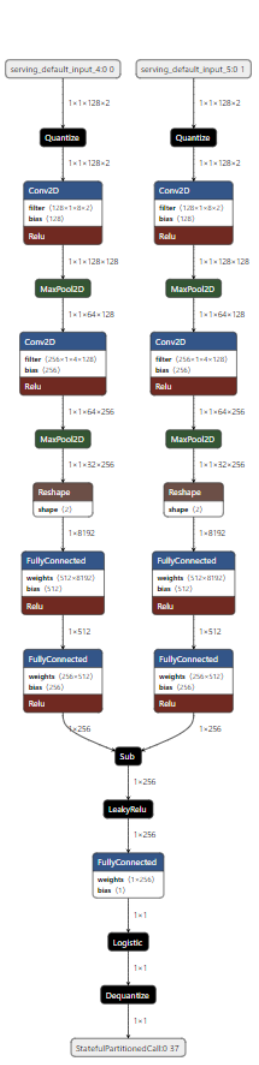
- *The project's three validation use-cases has resulted in new optimizations and kernel library extensions.*

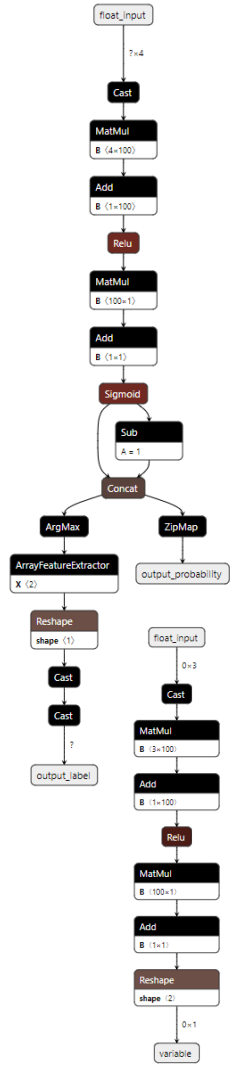


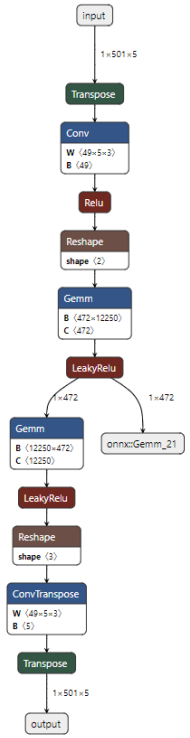
imsys

SecureGW use-case verification







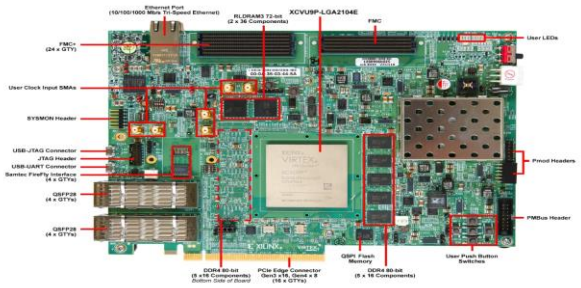


Verification of models

- Simulated in floating point and quantized to 8-bit integers
- Different configurations 1, 4 and 16 clusters
- Different optimizations analyzed

Emulator + Application use-cases on HW

Available on FPGA board

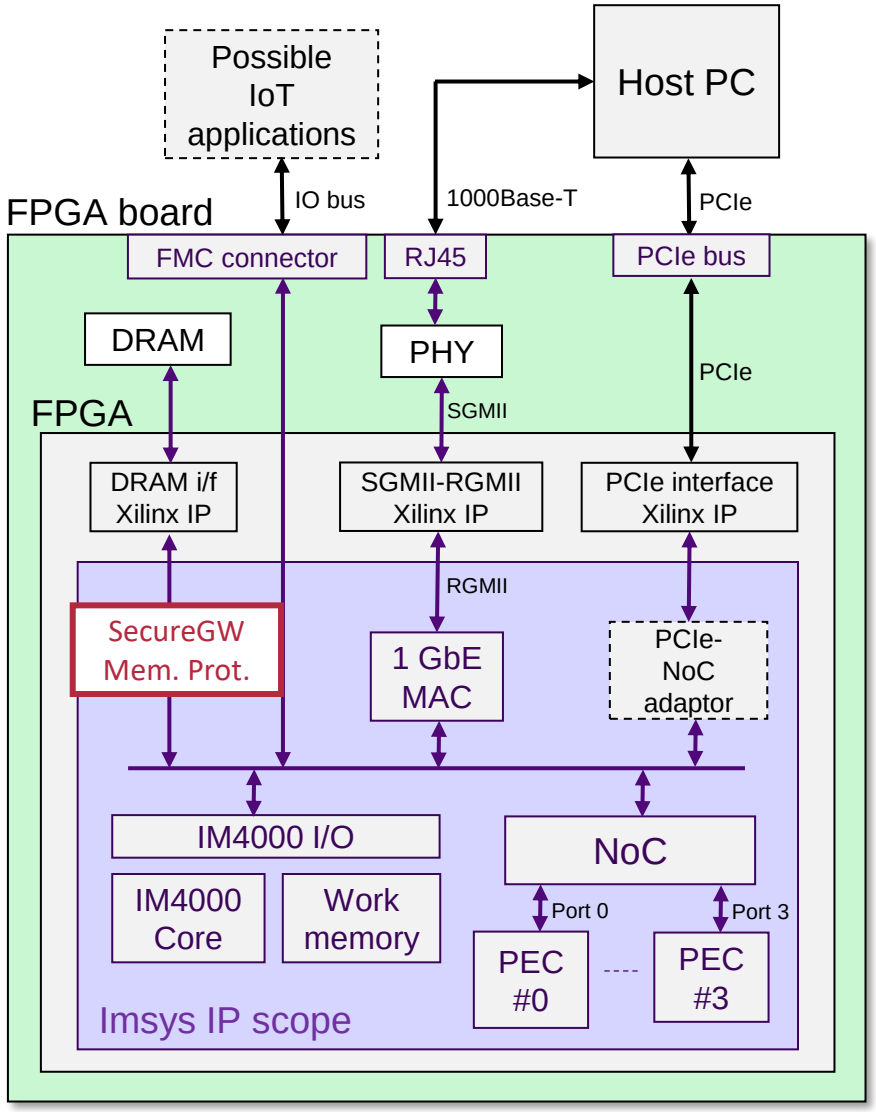


Application
- ✓ Presentation and input on Host PC

Processing elements
- ✓ 4 PEC configuration. (64 PE = ~150 GOP/s)
- ✓ Micro coded DNN operations (ISA-A firmware)
- ✓ Full stack SW on GPP (IM4000)

Security
- ✓ SecureGW memory protection HW for secure and performant high-speed memory accesses

High-speed data transfer
- ✓ DRAM
- ✓ Ethernet

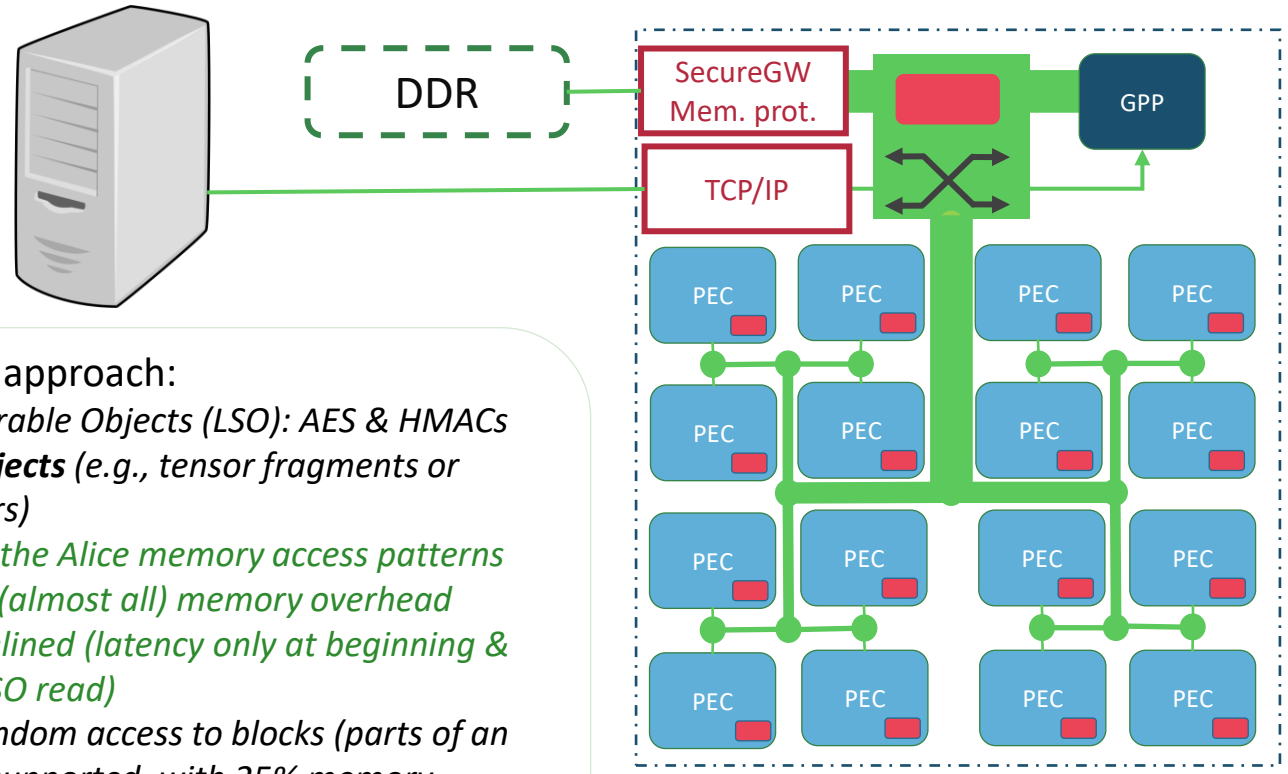


imsys

Security solution demonstrated in SecureGW

Secure Accelerator Memory (DDR):

- *Confidentiality: AES Encryption*
- *Integrity: Hash-based Message Authentication Codes* (SHA3 HMACs)



SecureGW approach:

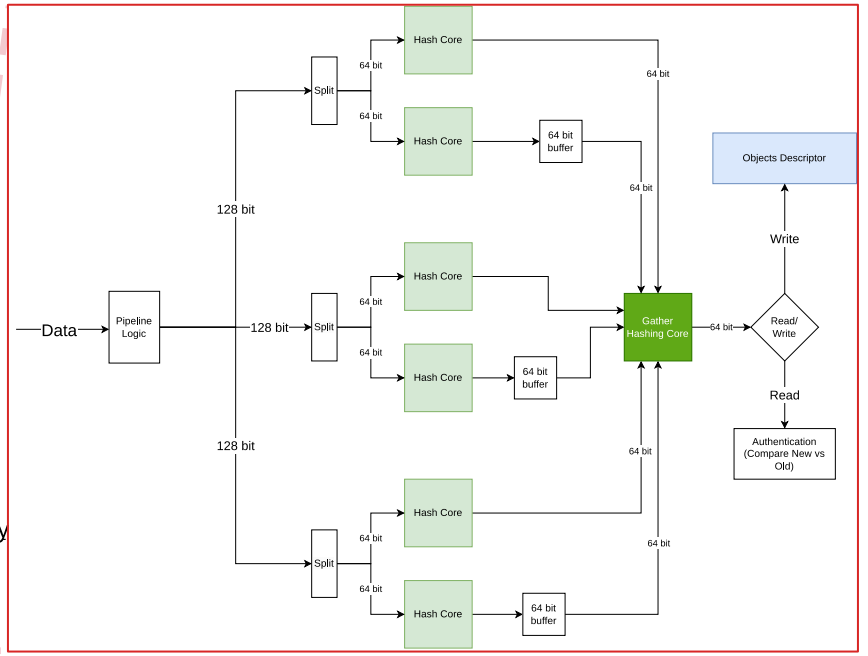
- *Large Securable Objects (LSO): AES & HMACs per **large objects** (e.g., tensor fragments or whole tensors)*
- *Tailored to the Alice memory access patterns*
- *Eliminates (almost all) memory overhead*
- *Is fully pipelined (latency only at beginning & end of the LSO read)*
- *General random access to blocks (parts of an LSO) is also supported, with 25% memory overhead.*

- **Same HW for read and write**
  - on the FPGA demonstrator platform
  - but programmed differently.
- **Confidentiality**
  - ensures even if data is stolen it is secure.
- **Integrity**
  - checks for any changes in the data during transit.
  - We know that what we read is what we wrote.

imsys

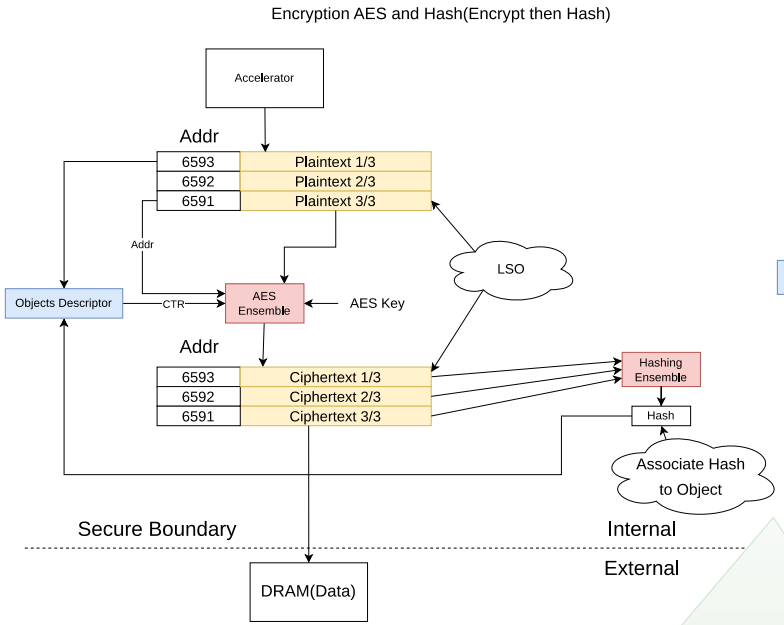
SecureGW Design: Integrity

Pipelined Hash Ensemble Engine
(3 x dual-64bit cores)

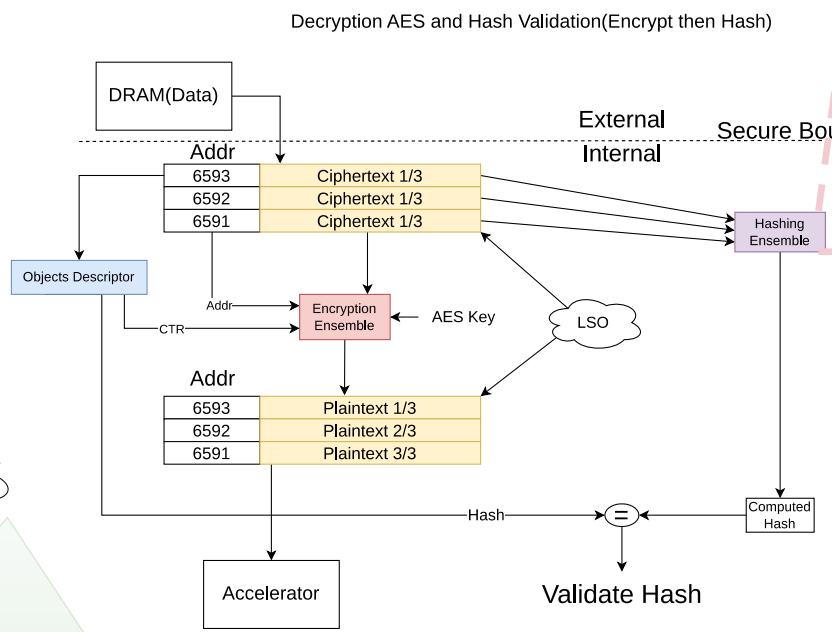


Memory Security in SecureGW is based on LSOs; HMAC generation/validation optimized for this case; Process can be fully pipelined for large objects.

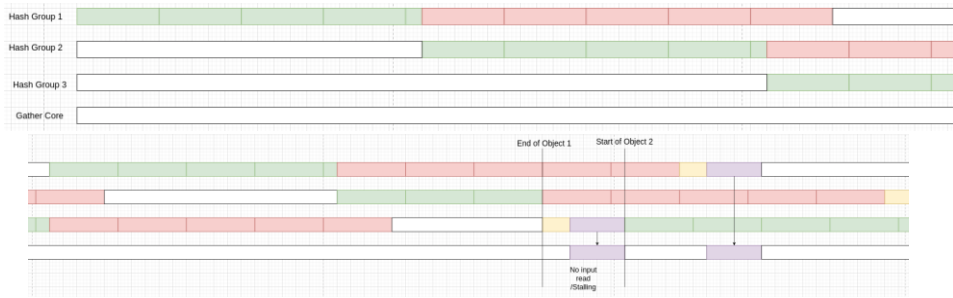
LSO Write (create HMAC)



LSO Read (validate HMAC)

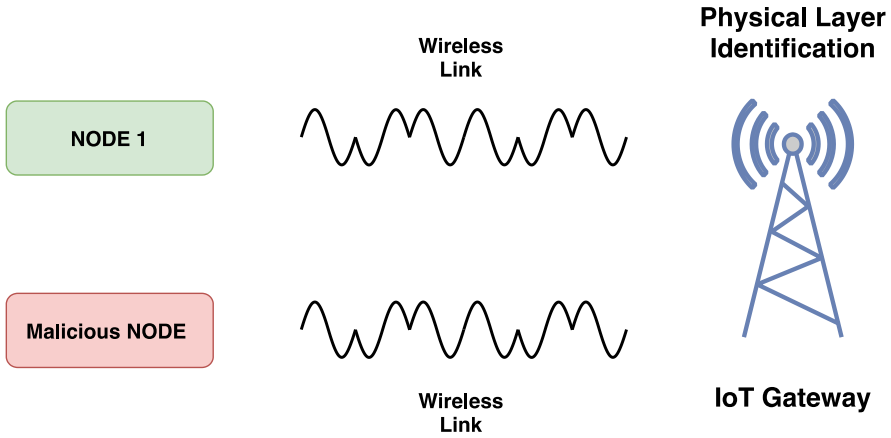


Pipelining diagrams
latency at start and end of a long read)



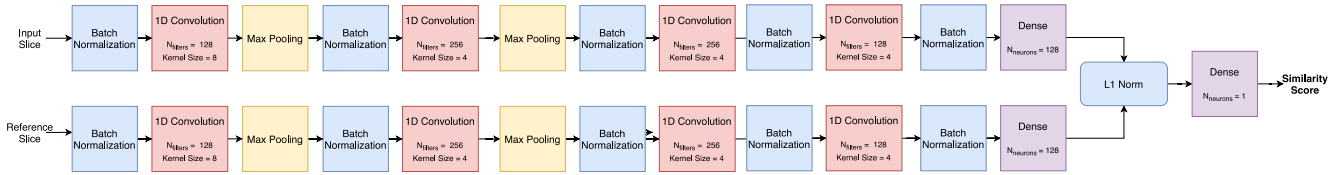
zero memory overhead: LSO Secure Hashes stored inside secure boundary (~ #objects)

IoT authentication application



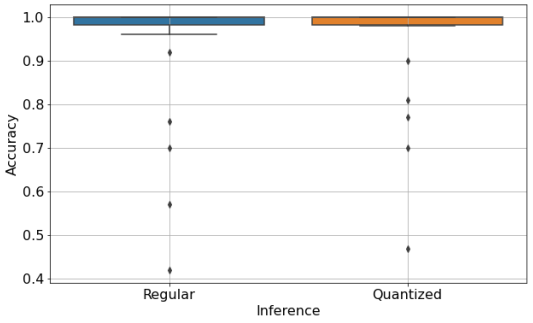
Identifying Rogue Devices

- **Problem:** Complex cryptographic solutions are not suitable for all devices
  - Can be forged to send malicious data
  - Vulnerable to replay attacks
- **Solution:** Identify devices based on their unique signatures
  - Transmitted signal has unique signatures due to hardware imperfections

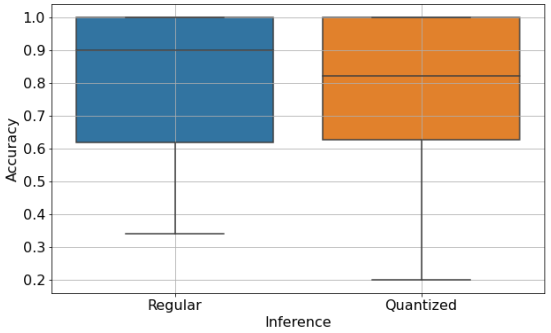


Siamese neural network selected based on its performance*

Similar Chipset



Different Chipset



More Precision degradation with different chipsets

The network was trained with 110 training devices.
Evaluated with rogue device selected from 26 devices (similar chipset) and 28 devices (different chipsets)
The model was quantized to 8-bit to prepare it for efficient acceleration.
The precision degradation was acceptable.

* PLIO: Physical Layer Identification using One-shot Learning | IEEE Conference Publication | IEEE Xplore

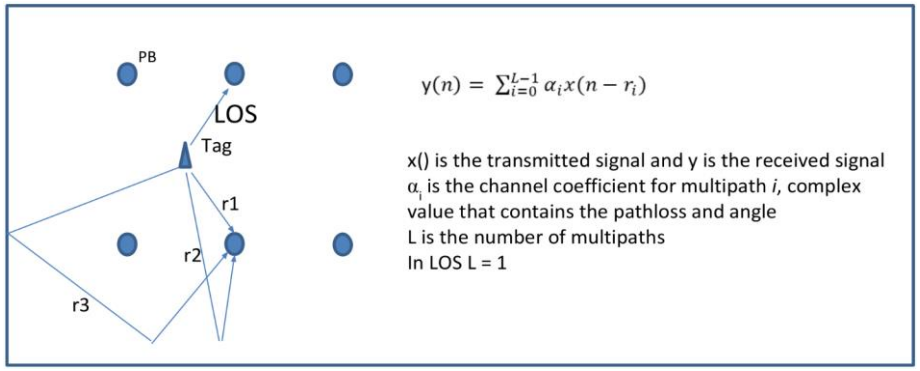
Background:

Wittra Tool Lock is used to track and manage (lock/unlock) equipment for safe handling on construction sites.

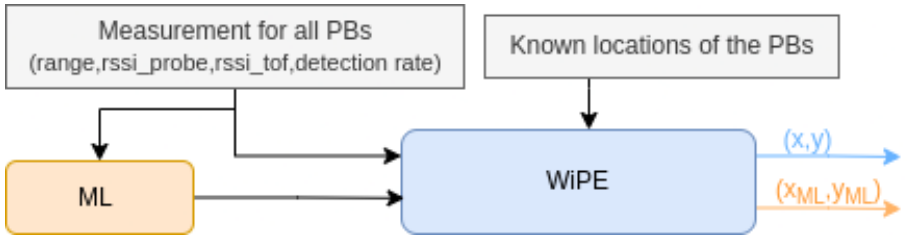
Goal:

Use ML to classify estimates of the distance between a Tag and a Positioning Beacon (PB) to achieve better positioning

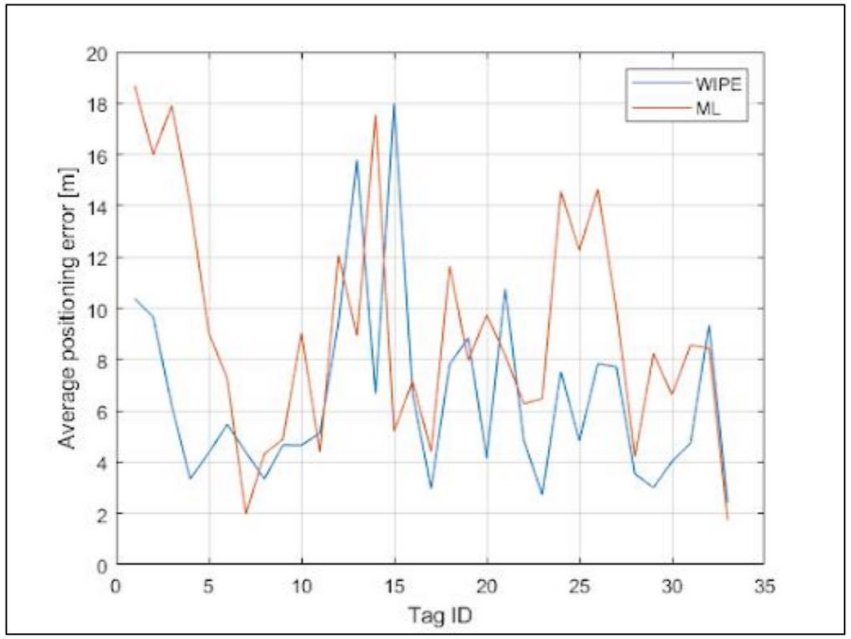
Problem:



Solution:



Result:



Comparison of rage estimates prediction by ML model and existing WIPE system

Conclusion and future work:

- In some cases, ML outperformed WiPE. ML is not far from WiPE classification despite only a small amount of data.
- Improvements to the existing design to include more inputs to better train the ML model

Conclusions …

Next steps, and reflections

- The project ends in 6 months – evaluation of use-cases ongoing
- High value but challenging to work with both hardware (CPU and AI accelerator) and software tools and use-cases simultaneously (had some delivery issues with FPGA manufacturer)
  - Long projects 24+ months is needed!
- We are interested in continuing to work with both the existing use cases and expanding into use cases with a need for energy-efficient and protected AI models in embedded / physical products (not necessarily in an IoT Gateway)

- Interested? – get in touch!
  - Joakim Eriksson, joakim.eriksson@ri.se
  - Dag Helmfrid, dag.helmfrid@imsystech.com or Mohammad Riazati, mohammad.riazati@imsystech.com

imsys IoT BRIDGE WITRA RI. SE