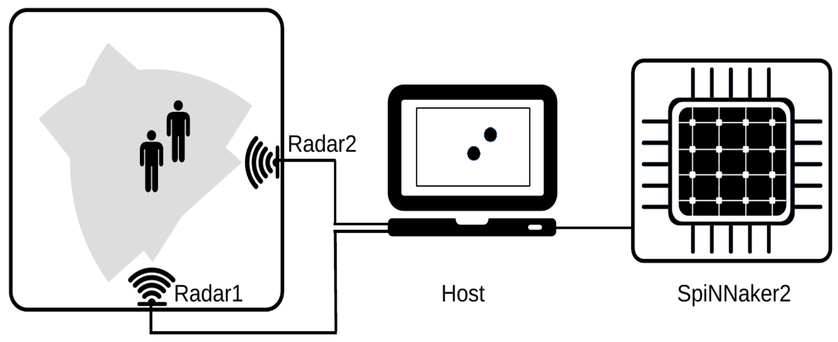
ANDANTE

UC1.1 Setup B: Indoor Positioning Recognition and People Counting

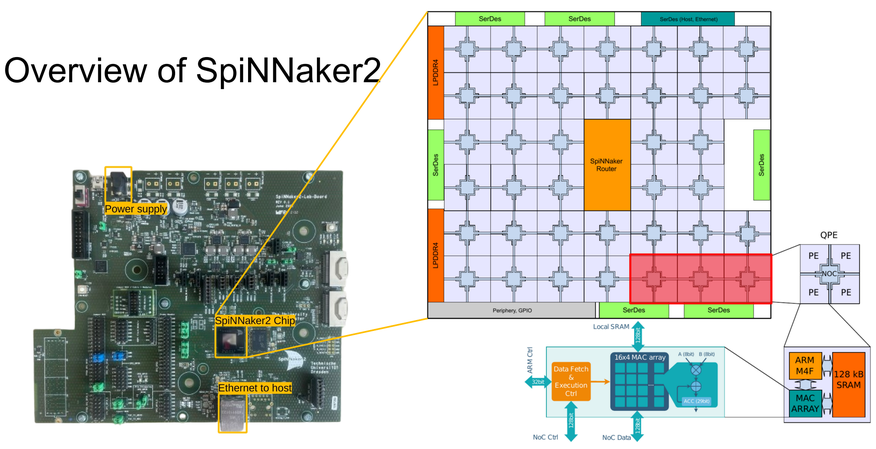
Overview & Setup

- RANet is a DNN model to predict the range-angle map of an indoor scene.
- Address people counting and positioning recognition.
- Objectives:
  - Achieve super-resolution performance beyond conventional radar imaging.
  - Replace radar processing chain with a DNN predictor.
  - Efficient execution in terms of memory, latency and energy.
- Demonstrator setup:
  - Two Infineon 60GHz BGT60TR13C radars.
  - Host machine for data forwarding and results visualization.
  - SpiNNaker2 board for model inference, targets counting and tracking.



Technology

Overview of SpiNNaker2



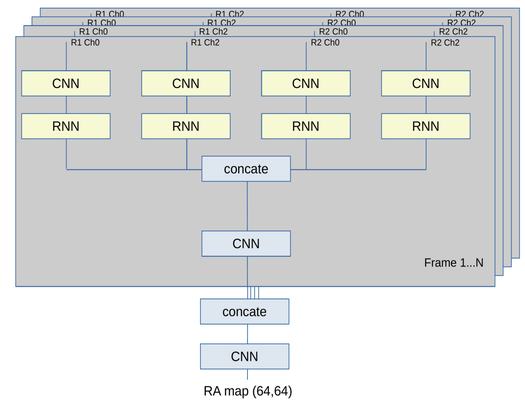
Hardware:

- SpiNNaker2: a 152 PEs digital neuromorphic chip
- Each PE contains 128kB SRAM, one ARM Cortex-M4F, and a ML Accelerator (conv & mm, 2.5 TOPS/W)
- I/O: Ethernet, JTAG, SPI, etc.

Software:

- Use multiple channels and frames of radars
- Combine convolutional and recurrent layers
- Discard dense layers for small model size

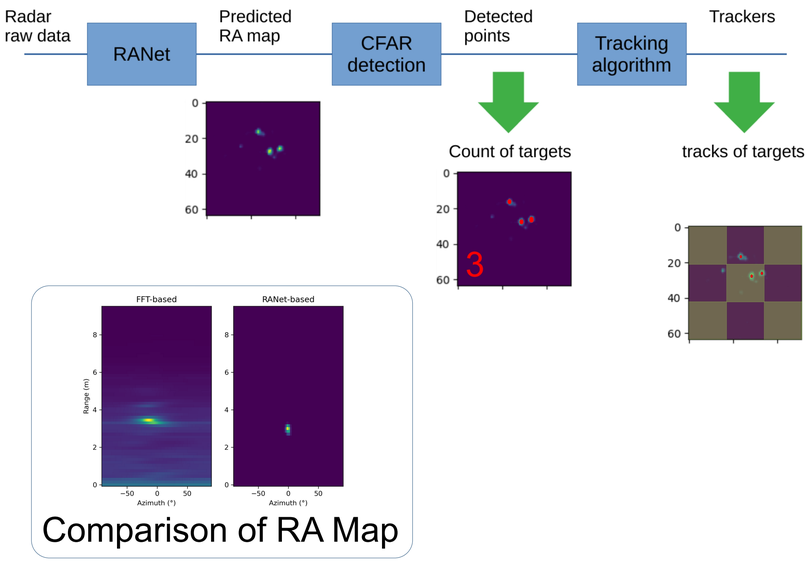
RANet Architecture



Results

- 3-stage pipeline processing
  - Prediction
  - Detection/Counting
  - Tracking
- Higher imaging quality from RANet
- Positioning recognition within a 3X3 a
- 93% counting accuracy
- 97.73% position F1 score
- 80 FPS processing throughput
- Inference with low energy cost

Data flow of UC1.1B



Benchmarking Results

Task	KPI Name	Accuracy/F1	Model Size [kB]	Inference Time [ms]	Inference Energy [mJ]
Counting	FFT baseline (PC)	35.0%	-	51	NA
	RANet (Orin Nano)	93.0%	1200	105	314
	RANet (SpiNNaker2)	93.0%	300	11	0.418
Tracking	FFT baseline (PC)	59.43%	-	58	NA
	RANet (Orin Nano)	97.73%	1200	111	356
	RANet (SpiNNaker2)	97.73%	300	12	0.453

* Benchmarking Results are not final.

Impact

- ML technology can break through the limitations of traditional radar signal processing.
- Conventional radar processing chain can be (partially) replaced by a DNN model.
- GRU quantization useful for many other applications and hardware platforms.

Progress beyond SoA

- Significant performance improvement
- 10X latency improvement (vs. Nvidia Orin Nano)
- 1000X energy efficiency improvement (vs. Nvidia Orin Nano)

Lessons learned

- Quantization of recurrent layers is more challenging than that of convolutional layers.
- A mature end-to-end model deployment tool is still unavailable.
- Efficient AI on edge requires software hardware codesign.

ANDANTE

AI for New Devices And Technologies at the Edge

andante-ai.edu linkedin.com/company/andante-ai mario.diaznava@st.com (coordinator)

"This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement № 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Germany, Netherlands, Portugal, Spain and Switzerland"

Scan Me to visit website