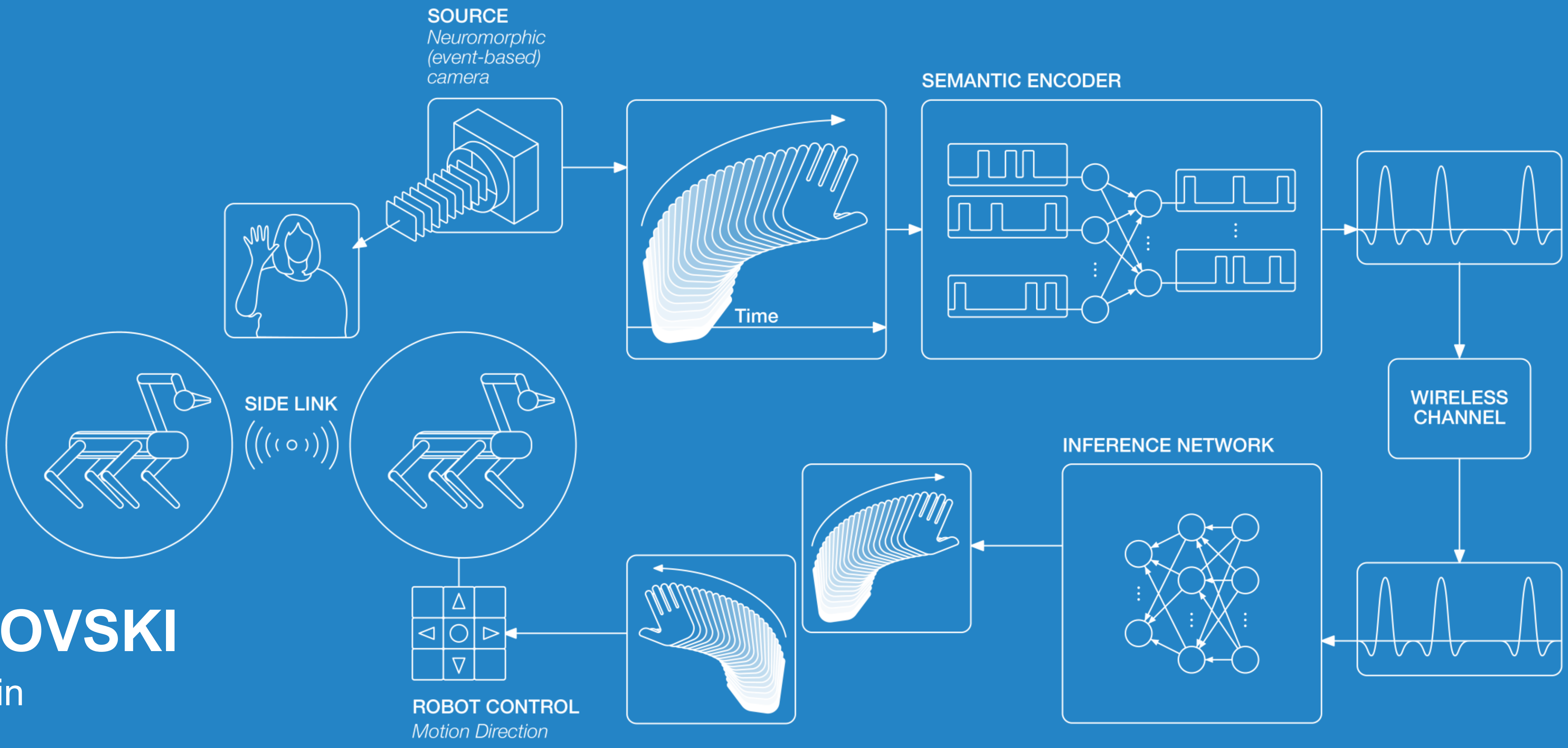
AI FOR RAN OR RAN FOR AI?

6G-RIC
Research and
Innovation Cluster



ZORAN UTKOVSKI
Fraunhofer HHI Berlin
WN Department

Collaborators

Mehdi Heshmati, Yuzhen Ke, Viktor Lorentz, Pengtao Xie, Omar Abdelhaleem, Johannes Dommel, Osvaldo Simeone
and
Slawomir Stanczak

+

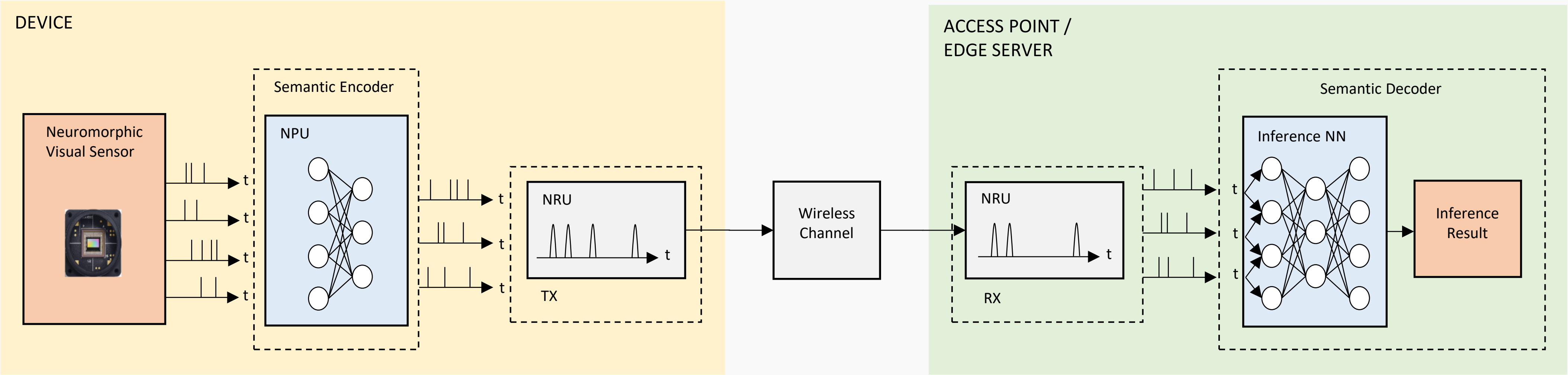
Collaborators from 6G-XCEL

(COLLABORATIVE) EDGE LEARNING AND INFERENCE

COLLABORATIVE EDGE INFERENCE

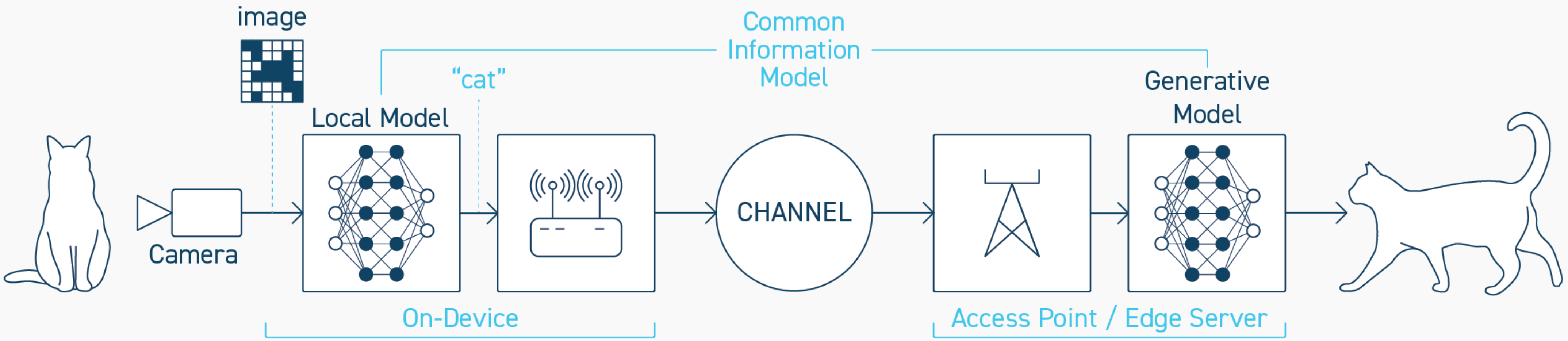
Device-Edge Co-Inference: device and edge server cooperate to perform a certain task

- **Efficiency**: Transmit only **task-relevant information** → reduced complexity, communication overhead, energy consumption
- **Trade-off**: Complexity vs. performance (**complexity is related to energy consumption**)
- Includes **split inference** and **semantic offloading** as special cases



CLOSELY RELATED TO "SEMANTIC COMMUNICATION"

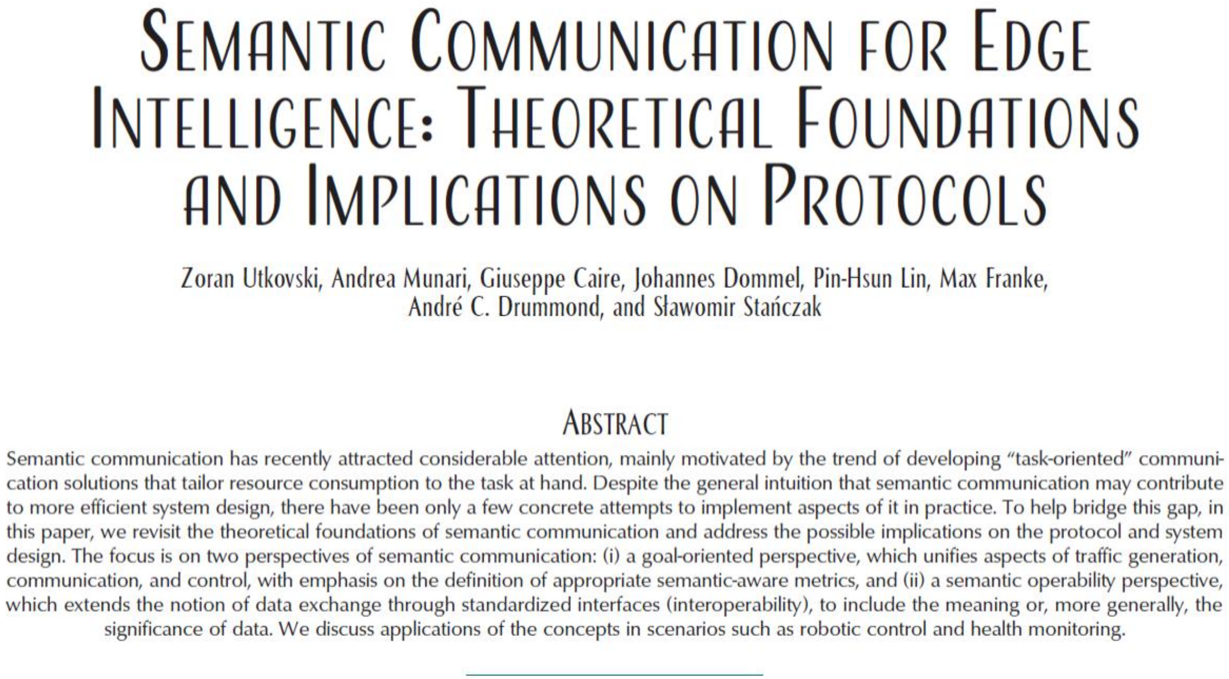
Reduction in communication overhead and energy consumption by extracting/transmitting only "**task-relevant information**"



ANALOGY TO JOINT SOURCE AND CHANNEL CODING

SEMANTIC COMMUNICATION FOR EDGE INTELLIGENCE

IEEE IoT Magazine: Special Issue on Task-Oriented Communication for IoT

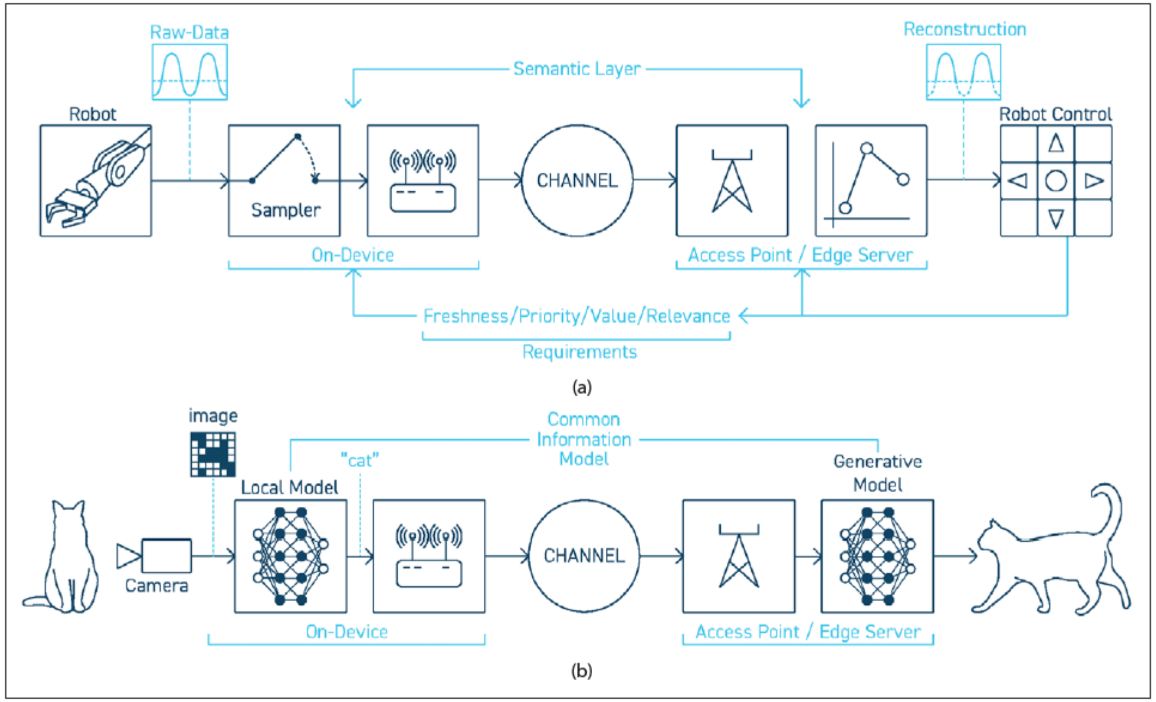


Semantic Communication for Edge Intelligence: Theoretical Foundations and Implications on Protocols

Zoran Utkovski, Andrea Munari, Giuseppe Caire, Johannes Dommel, Pin-Hsun Lin, Max Franke,
André C. Drummond, and Sławomir Stańczak

Abstract

Semantic communication has recently attracted considerable attention, mainly motivated by the trend of developing "task-oriented" communication solutions that tailor resource consumption to the task at hand. Despite the general intuition that semantic communication may contribute to more efficient system design, there have been only a few concrete attempts to implement aspects of it in practice. To help bridge this gap, in this paper, we revisit the theoretical foundations of semantic communication and address the possible implications on the protocol and system design. The focus is on two perspectives of semantic communication: (i) a goal-oriented perspective, which unifies aspects of traffic generation, communication, and control, with emphasis on the definition of appropriate semantic-aware metrics, and (ii) a semantic operability perspective, which extends the notion of data exchange through standardized interfaces (interoperability), to include the meaning or, more generally, the significance of data. We discuss applications of the concepts in scenarios such as robotic control and health monitoring.



EXAMPLE 1:

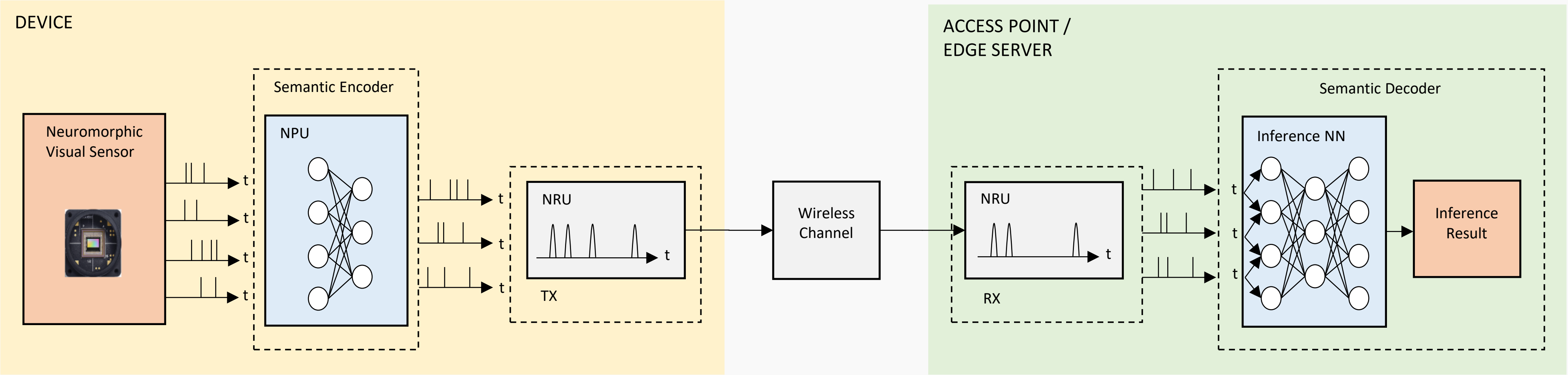
(NEUROMORPHIC) EDGE INFERENCE FOR HUMAN-ROBOT INTERACTION

Neuromorphic Wireless Device-Edge Co-inference via the Directed Information Bottleneck
Y. Ke, Z. Utkovski, M. Heshmati, O. Simeone, J. Dommel, and S. Stanczak, ACM ICONS 2024

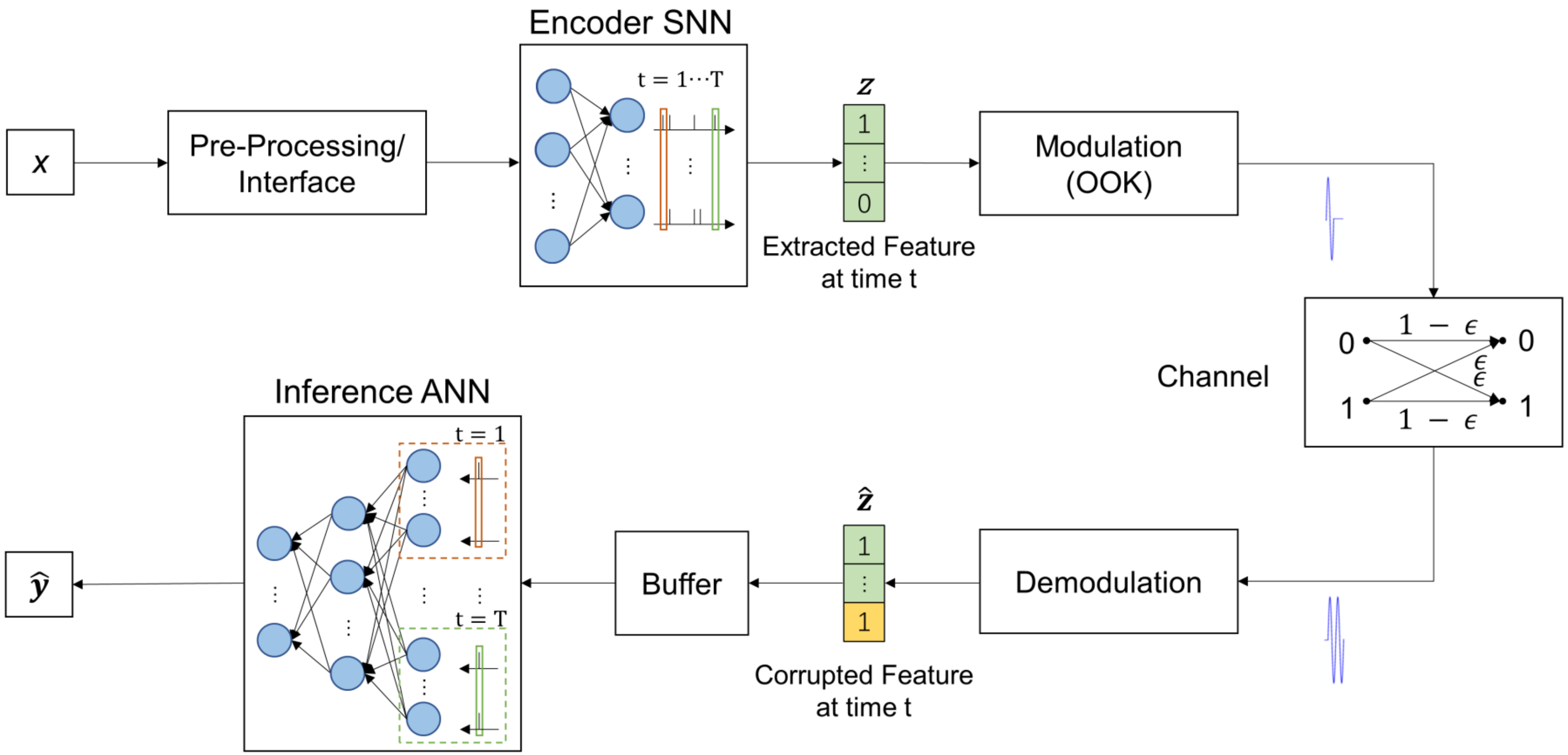
(SEMANTIC-AWARE) NEUROMORPHIC WIRELESS EDGE INFERENCE

Spiking Neural Networks (SNNs) for Edge Intelligence

- network of dynamic spiking neurons
- communicate and process information with the timings of spikes
- mimic the operation of biological neurons
- energy-efficient (**~pJ per spike**)
- low latency & event-based processing



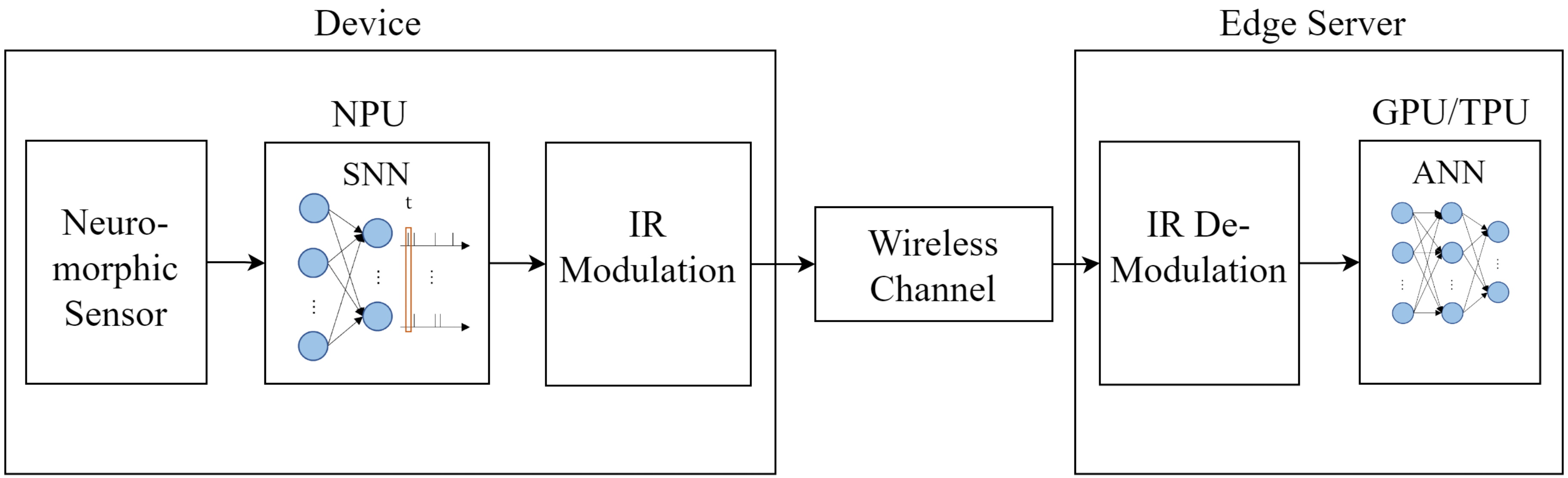
DEVICE-EDGE CO-INFERENCE VIA THE DIRECTED INFORMATION BOTTLENECK



$$\mathcal{L}_{DIB}(\phi) = -I(\hat{\boldsymbol{Z}} \rightarrow \boldsymbol{Y}) + \beta \cdot I(\boldsymbol{X} \rightarrow \hat{\boldsymbol{Z}})$$

ANALOGY TO REPRESENTATION LEARNING

TRANSMITTER AND WIRELESS CHANNEL PARAMETERIZATION



Transmitter

$$p_{\phi}\left(\boldsymbol{z} \,\|\, \boldsymbol{x}\right) = \prod_{t=1}^{T} p\left(\boldsymbol{z}_t | \boldsymbol{z}^{t-1}, \boldsymbol{x}^t\right)$$

$$= \prod_{t=1}^{T} \prod_{i \in \mathcal{V}} p_{\phi}\left(z_{i,t} | u_{i,t}\right)$$

Channel

$$p_{\phi}\left(\hat{\boldsymbol{z}} \,\|\, \boldsymbol{x}\right) = \sum_{\boldsymbol{z}} p_{\phi}\left(\boldsymbol{z} \,\|\, \boldsymbol{x}\right) p_{\mathrm{BSC}}\left(\hat{\boldsymbol{z}} | \boldsymbol{z}\right)$$

SEMANTIC VARIATIONAL DIRECTED INFORMATION BOTTLENECK (S-VDIB)

Design Criterion

$$\mathcal{L}_{DIB}(\phi) = -I(\hat{\boldsymbol{Z}} \to \boldsymbol{Y}) + \beta \cdot I(\boldsymbol{X} \to \hat{\boldsymbol{Z}})$$

Inference Model

$$\begin{aligned} I(\hat{\boldsymbol{Z}} \to \boldsymbol{Y}) &= H\left(\boldsymbol{Y}\right) - H(\boldsymbol{Y} \parallel \hat{\boldsymbol{Z}}) \\ &= H\left(\boldsymbol{Y}\right) + E_{p(\boldsymbol{y}, \hat{\boldsymbol{z}})}\left[\log p\left(\boldsymbol{y} \parallel \hat{\boldsymbol{z}}\right)\right] \end{aligned}$$

$$q_{\boldsymbol{\theta}}\left(\boldsymbol{y} \parallel \hat{\boldsymbol{z}}\right) = \prod_{t=1}^{T} q_{\boldsymbol{\theta}}\left(\boldsymbol{y}_t | \boldsymbol{y}^{t-1}, \hat{\boldsymbol{z}}^t\right)$$

Encoder Model

$$I(\boldsymbol{X} \to \hat{\boldsymbol{Z}}) = E_{p(\hat{\boldsymbol{z}}, \boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\phi}}\left(\hat{\boldsymbol{z}} \parallel \boldsymbol{x}\right)}{p\left(\hat{\boldsymbol{z}}\right)}\right]$$

$$q(\hat{\boldsymbol{z}}) = \prod_{t=1}^{T} q\left(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}^{t-1}\right)$$

Neuromorphic Wireless Device-Edge Co-inference via the Directed Information Bottleneck
Y. Ke, Z. Utkovski, M. Heshmati, O. Simeone, J. Dommel, and S. Stanczak, ACM ICONS 2024

SEMANTIC VARIATIONAL DIRECTED INFORMATION BOTTLENECK (S-VDIB)

Semantic Variational Directed Information Bottleneck (S-VDIB)

$$\mathcal{L}_{VDIB}\left(\boldsymbol{\phi},\boldsymbol{\theta}\right)=$$

$$E_{p(\boldsymbol{y},\boldsymbol{x})}\left\{E_{p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}||\boldsymbol{x})}\left[-\log q_{\boldsymbol{\theta}}\left(\boldsymbol{y}||\hat{\boldsymbol{z}}\right)+\beta\log\frac{p_{\boldsymbol{\phi}}\left(\hat{\boldsymbol{z}}||\boldsymbol{x}\right)}{q(\hat{\boldsymbol{z}})}\right]\right\}.$$

ANN – Gradient estimation via (standard) Backpropagation

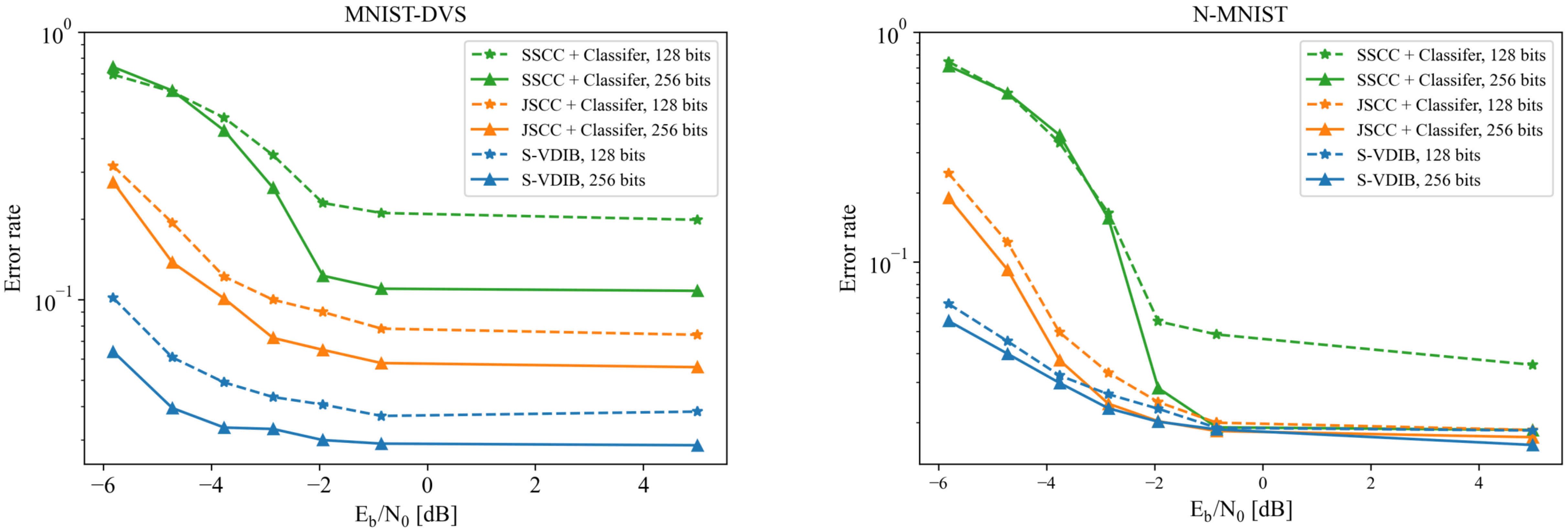
$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{VDIB}}\left(\boldsymbol{\phi},\boldsymbol{\theta}\right)=\mathrm{E}_{p(\boldsymbol{x},\boldsymbol{y})}\mathrm{E}_{p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}||\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}}\ell_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}},\boldsymbol{y})\right]$$

SNN – Gradient estimation via Score Function Estimator

$$\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\mathrm{VDIB}}\left(\boldsymbol{\phi},\boldsymbol{\theta}\right)=$$

$$\mathrm{E}_{p(\boldsymbol{x},\boldsymbol{y})}\mathrm{E}_{p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}||\boldsymbol{x})}\left[\Big(\ell_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}},\boldsymbol{y})+\beta\ell_{\boldsymbol{\phi}}\left(\hat{\boldsymbol{z}},\boldsymbol{x}\right)\Big)\nabla_{\boldsymbol{\phi}}\log p_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}}||\boldsymbol{x})\right]$$

SIMULATION RESULTS



- Test error rate as a function of the SNR per bit, Eb/N0, for the two standard datasets MNIST-DVS (left) and N-MNIST (right).
- Comparison perfomed with Separate Source Channel Coding + Classifier and with Joint Source and Channel Coding + Classifier

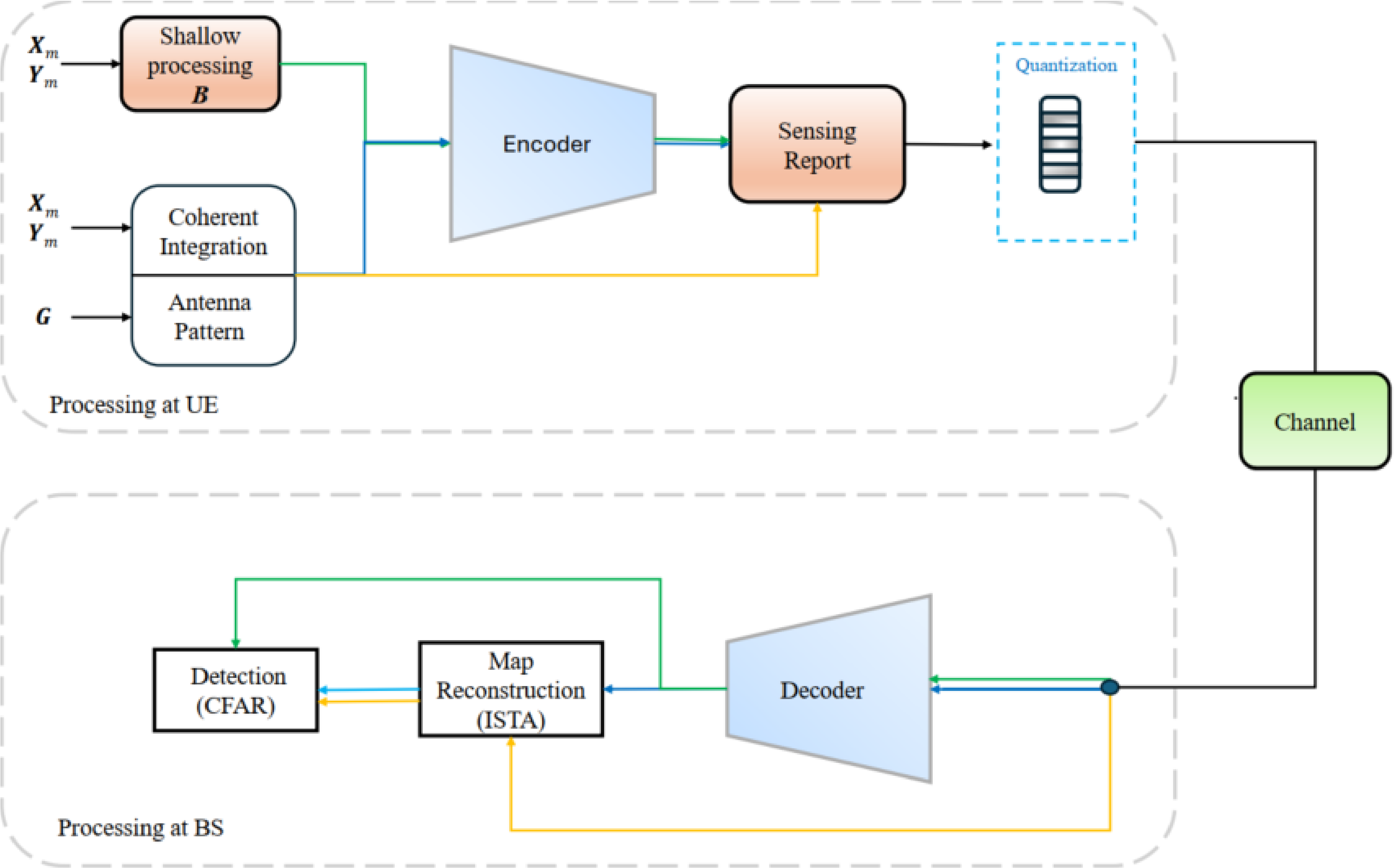
PoC NEUROMORPHIC WIRELESS DEVICE-EDGE CO-INFERENCE

Best Demo Award IEEE ICMLCN 2024

EXAMPLE 2:

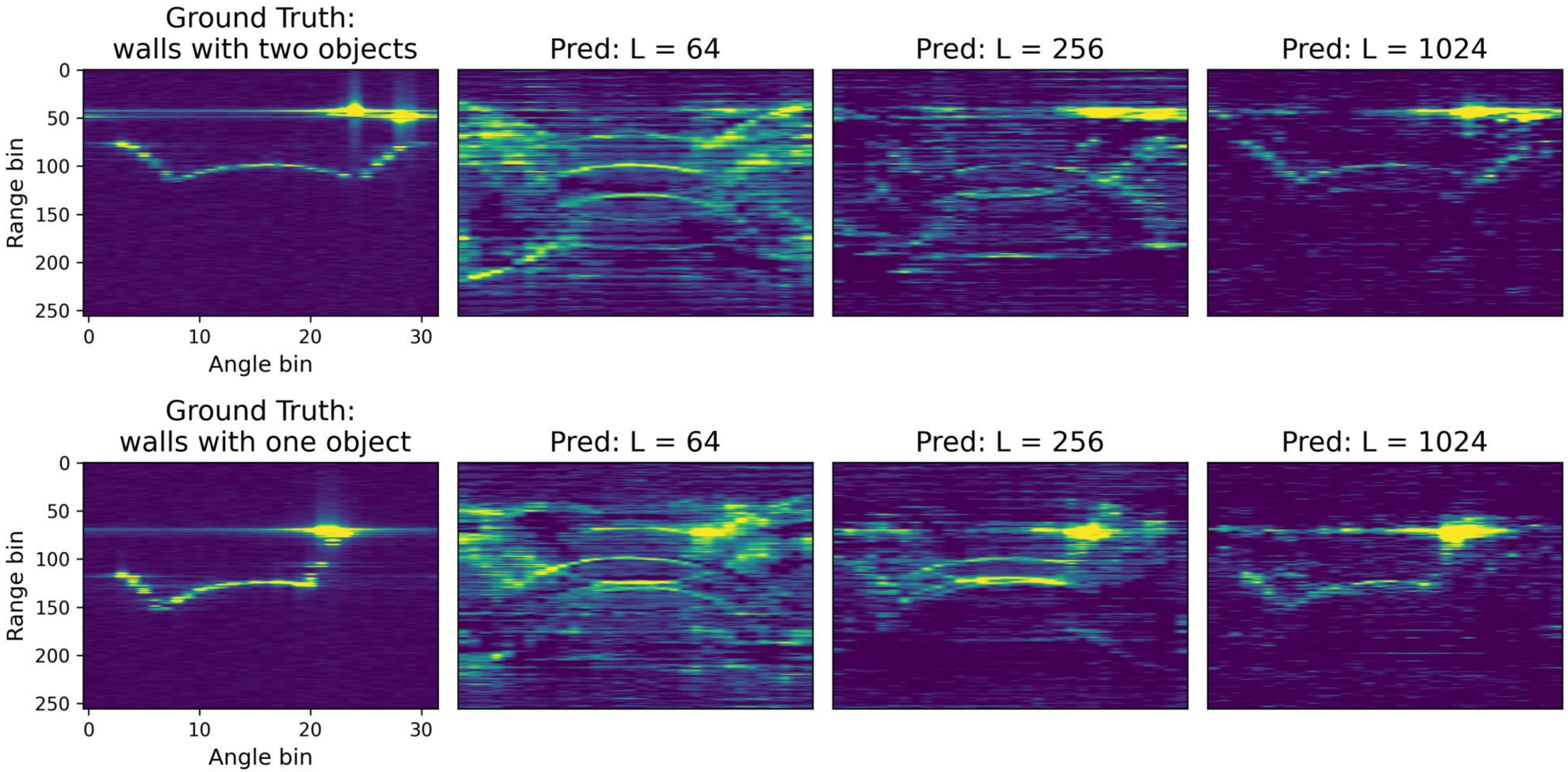
COLLABORATIVE INFERENCE FOR INTEGRATED SENSING AND COMMUNICATION

COLLABORATIVE INFERENCE FOR RADIO SENSING



A. Fazli, Z. Utkovski, E. Tohidi, S. Stanczak, “A Framework for ISAC-related reporting,” in preparation, 2025.

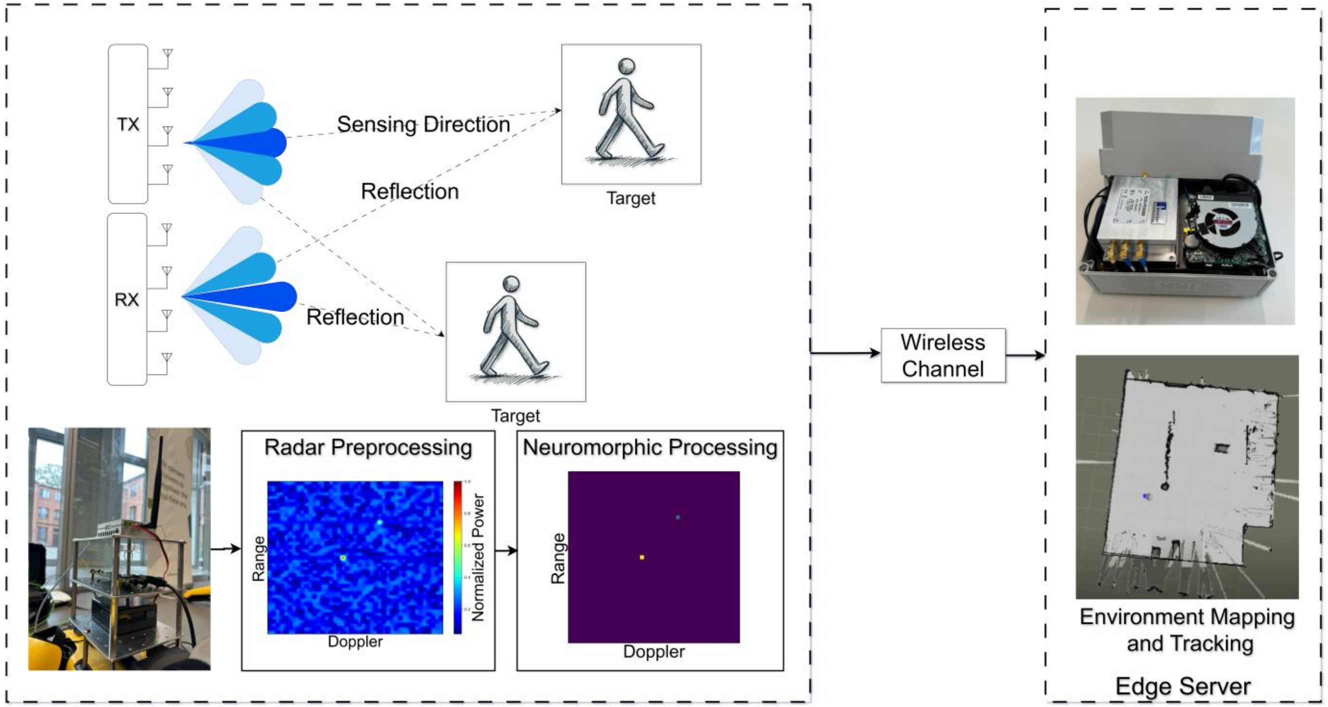
COLLABORATIVE INFERENCE FOR RADIO SENSING



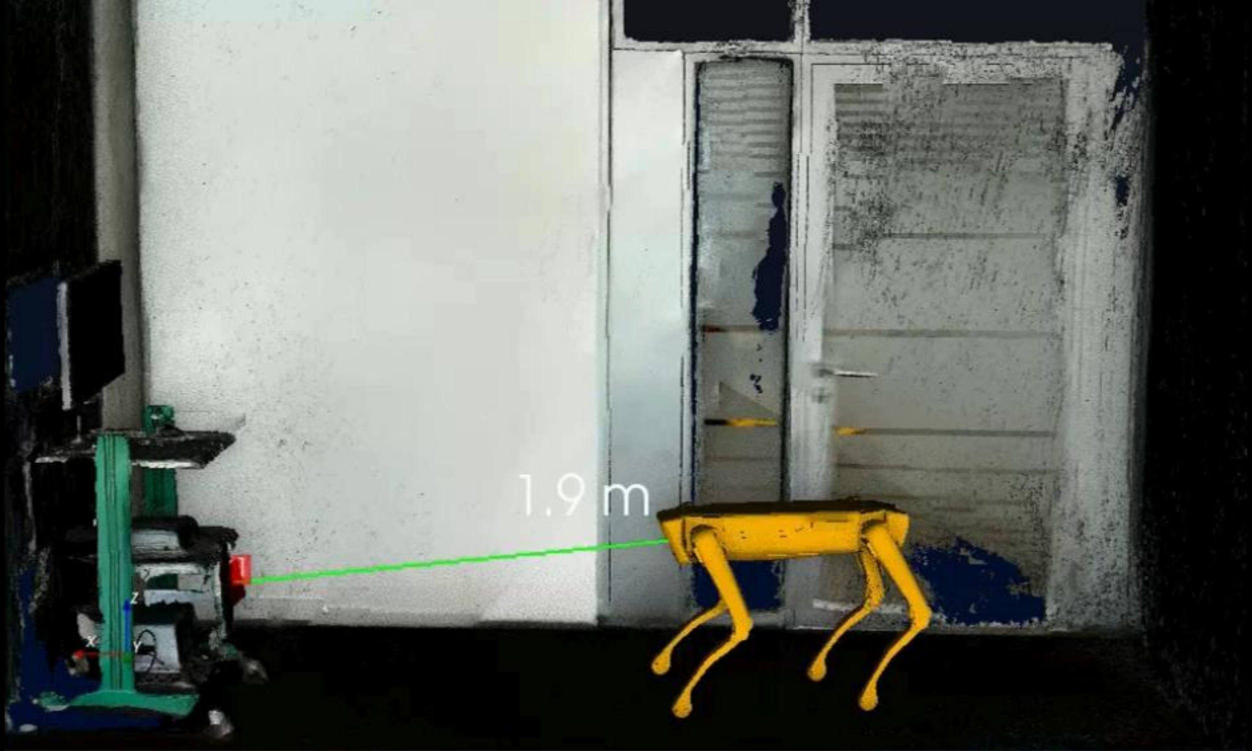
A. Fazli, Z. Utkovski, E. Tohidi, S. Stanczak, “A Framework for ISAC-related reporting,” in preparation, 2025.

(NEUROMORPHIC) COLLABORATIVE INFERENCE FOR RADIO SENSING

- Neuromorphic device-edge co-inference for radio sensing in networked robotics
- Device (robot) generates range-angle maps and uses Spiking Neural Networks (SNNs) with the Spiking-Locally Competitive Algorithm (S-LCA) for sparse representation
- Edge server performs scatterer clustering, extended target detection, and tracking



M. Heshmati, Z. Utkovski, K. Turbic, Y. Ke, S. Wittig, R. Askar, M. Peter and S. Stanczak, “Split-Inference Architecture for Device-Centric Radio Sensing in a Networked Robotics Scenario”, accepted at IEEE International Conference on Machine Learning for Communication and Networking (IEEE ICMLCN), 2025.

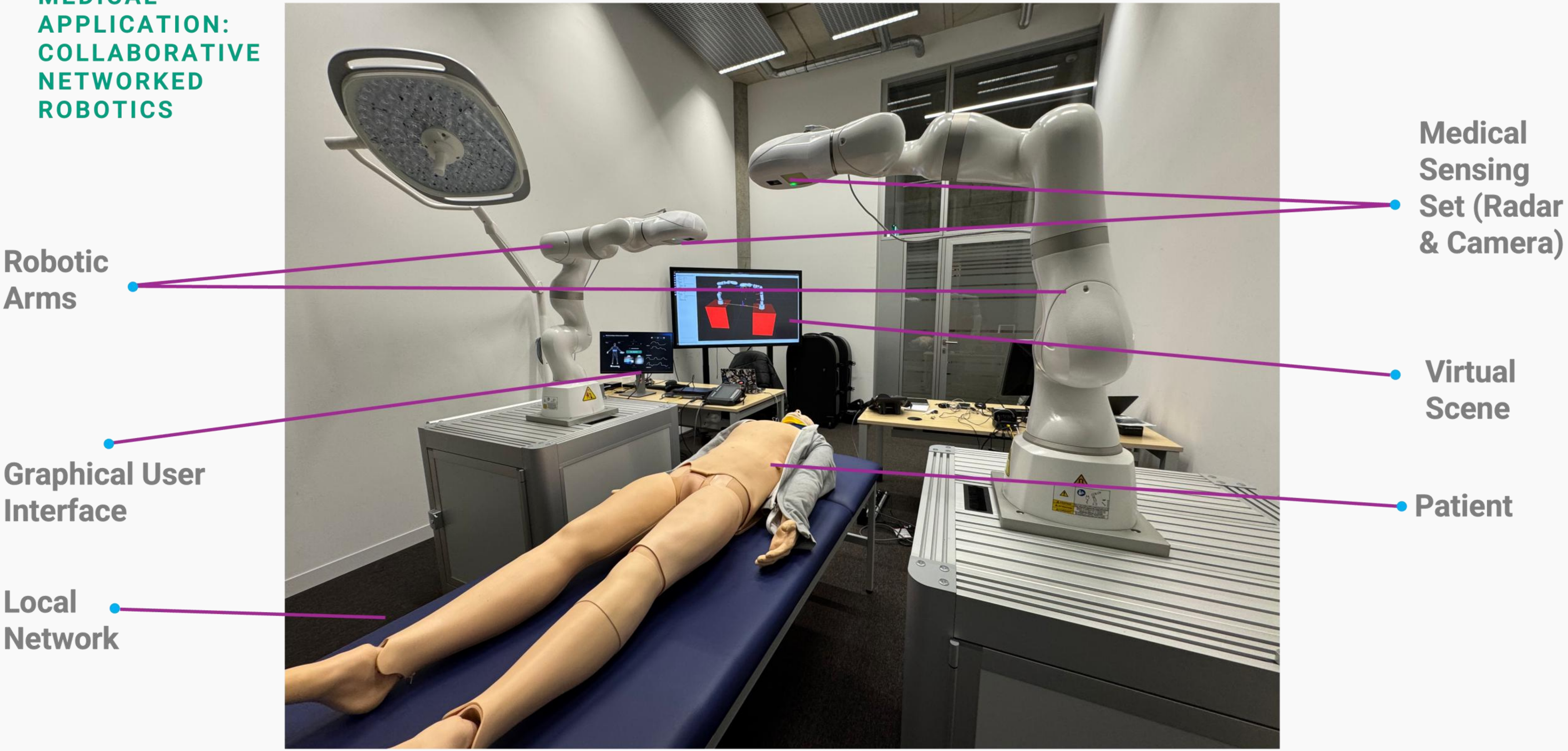

1.9 m
z
x
y

D-Band Sensing and Communication
BostonDynamics

EXAMPLE 3:

COLLABORATIVE INFERENCE FOR MEDICAL ROBOTICS

MEDICAL APPLICATION: COLLABORATIVE NETWORKED ROBOTICS



Medical Sensing Set (Radar & Camera)

Robotic Arms

Virtual Scene

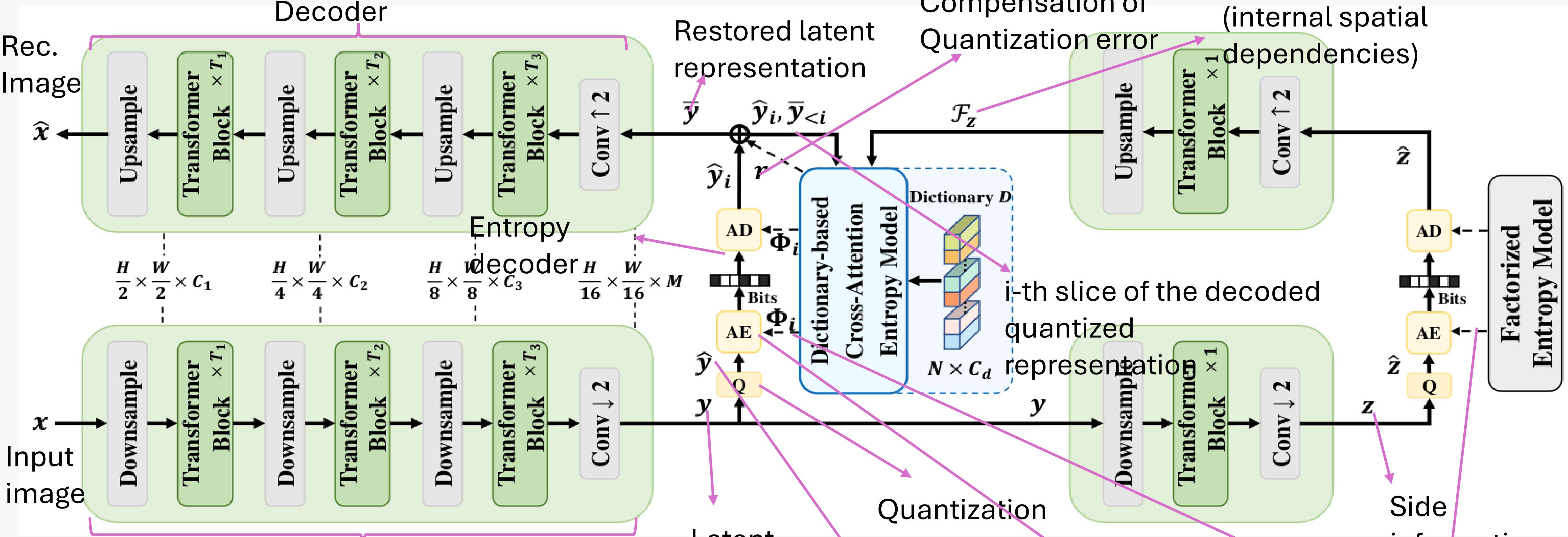
Graphical User Interface

Patient

Local Network

COLLABORATIVE INFERENCE FOR MEDICAL ROBOTICS

Learned Image Compression with Dictionary-based Entropy Model(DCAE)



$\hat{x}$
Upsample
Transformer Block $\times T_1$
Upsample
Transformer Block $\times T_2$
Upsample
Transformer Block $\times T_3$
Conv ↑ 2
$\bar{y}$
$\hat{y}_i, \bar{y}_{<i}$
$\hat{y}_i$
r
AD
Φ_i
Bits
AE
Φ_i
$\hat{y}$
Q
y
Dictionary-based Cross-Attention Entropy Model
Dictionary D
$N \times C_d$
$\mathcal{F}_z$
Upsample
Transformer Block $\times 1$
Conv ↑ 2
$\hat{z}$
AD
Bits
AE
$\hat{z}$
Q
z
Factorized Entropy Model
$\frac{H}{2} \times \frac{W}{2} \times C_1$
$\frac{H}{4} \times \frac{W}{4} \times C_2$
$\frac{H}{8} \times \frac{W}{8} \times C_3$
$\frac{H}{16} \times \frac{W}{16} \times M$
x
Downsample
Transformer Block $\times T_1$
Downsample
Transformer Block $\times T_2$
Downsample
Transformer Block $\times T_3$
Conv ↓ 2
y
Downsample
Transformer Block $\times 1$
Conv ↓ 2

Decoder

Rec. Image

Restored latent representation

Compensation of Quantization error

Hyper-prior Feature (internal spatial dependencies)

Entropy decoder

i-th slice of the decoded quantized representation

Input image

Encoder

Latent representation

Quantized representation

Quantization

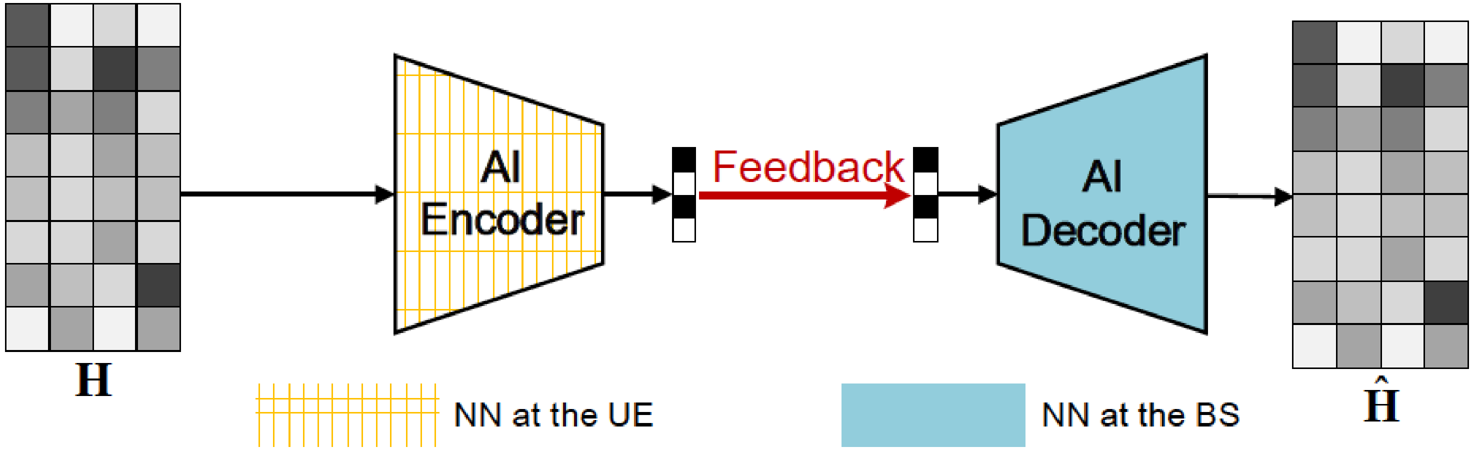
Entropy encoder

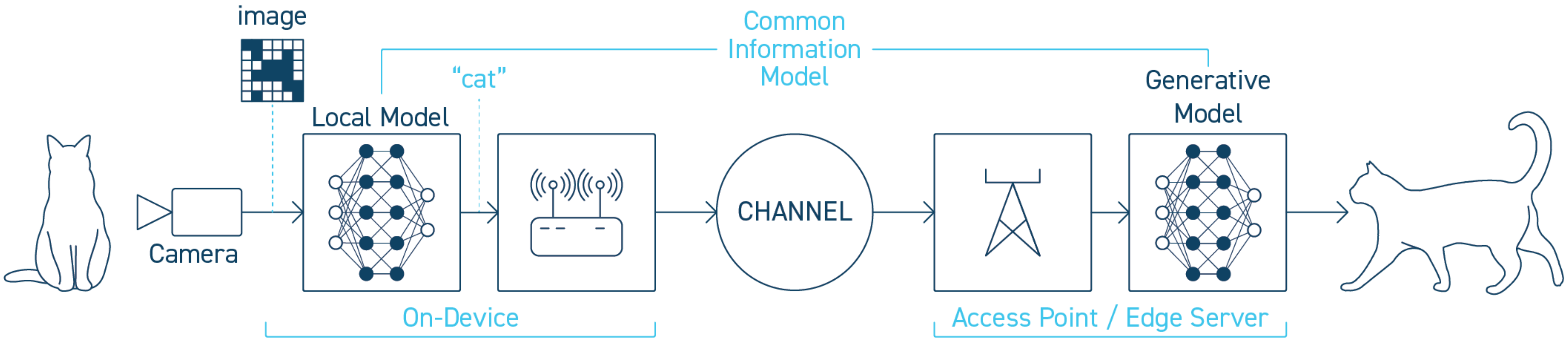
Side information

Distribution para meter of y/z

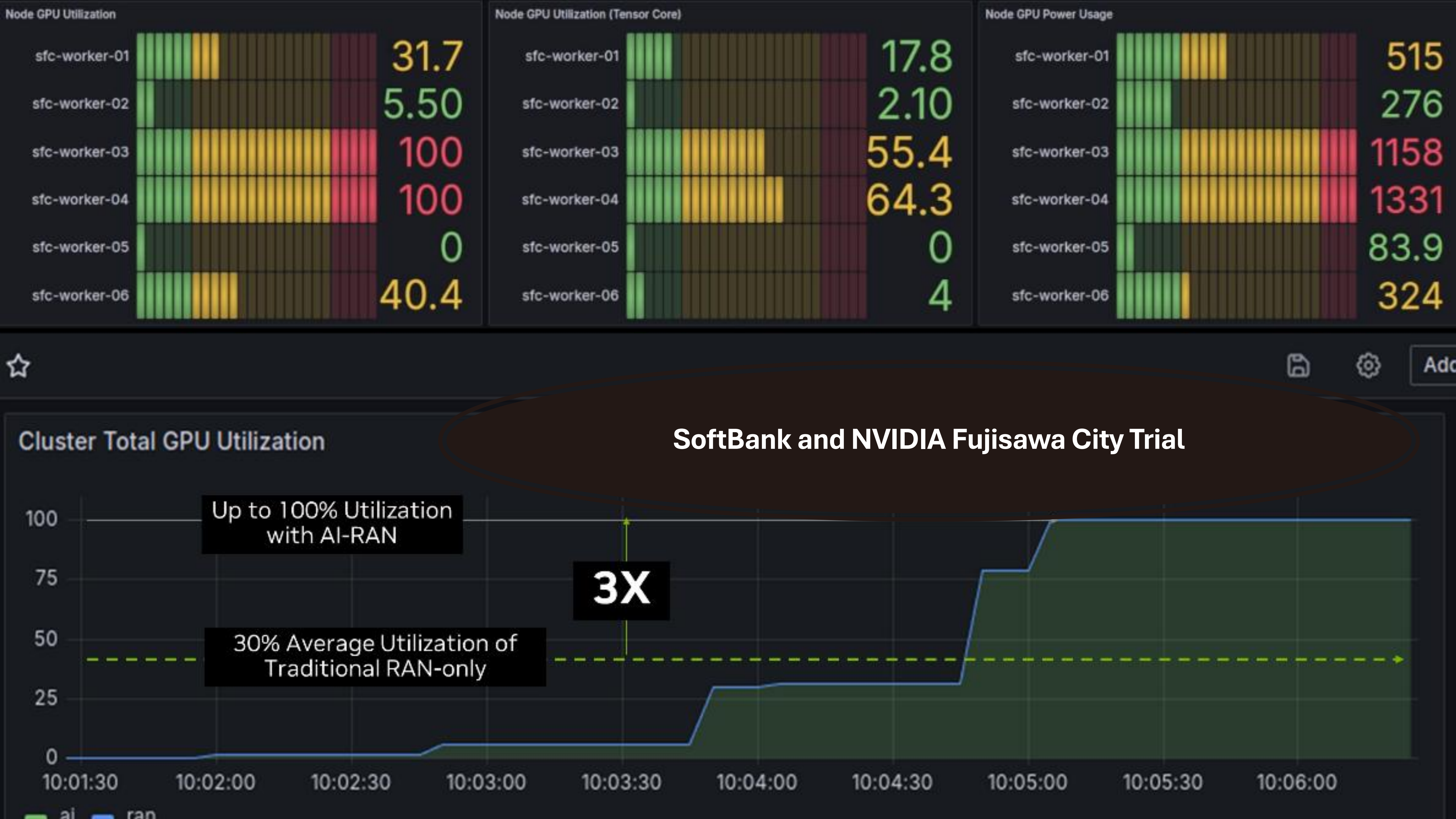
ARCHITECTURAL IMPLICATIONS AND IMPLICATIONS ON PROTOCOLS

SEMANTIC COMMUNICATION AS AN EXTENSION TO THE AI-NATIVE RADIO INTERFACE FRAMEWORK IN 3GPP





SoftBank and NVIDIA Fujisawa City Trial



Node GPU Utilization

sfc-worker-01	31.7
sfc-worker-02	5.50
sfc-worker-03	100
sfc-worker-04	100
sfc-worker-05	0
sfc-worker-06	40.4

Node GPU Utilization (Tensor Core)

sfc-worker-01	17.8
sfc-worker-02	2.10
sfc-worker-03	55.4
sfc-worker-04	64.3
sfc-worker-05	0
sfc-worker-06	4

Node GPU Power Usage

sfc-worker-01	515
sfc-worker-02	276
sfc-worker-03	1158
sfc-worker-04	1331
sfc-worker-05	83.9
sfc-worker-06	324

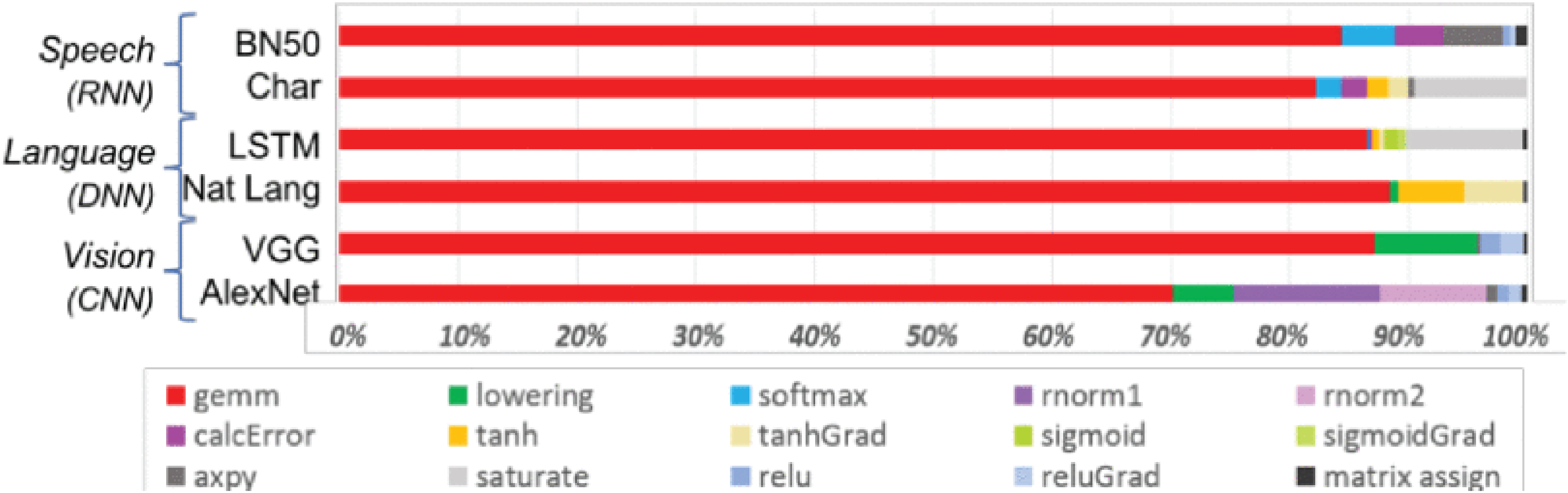
Cluster Total GPU Utilization

Up to 100% Utilization with AI-RAN

3X

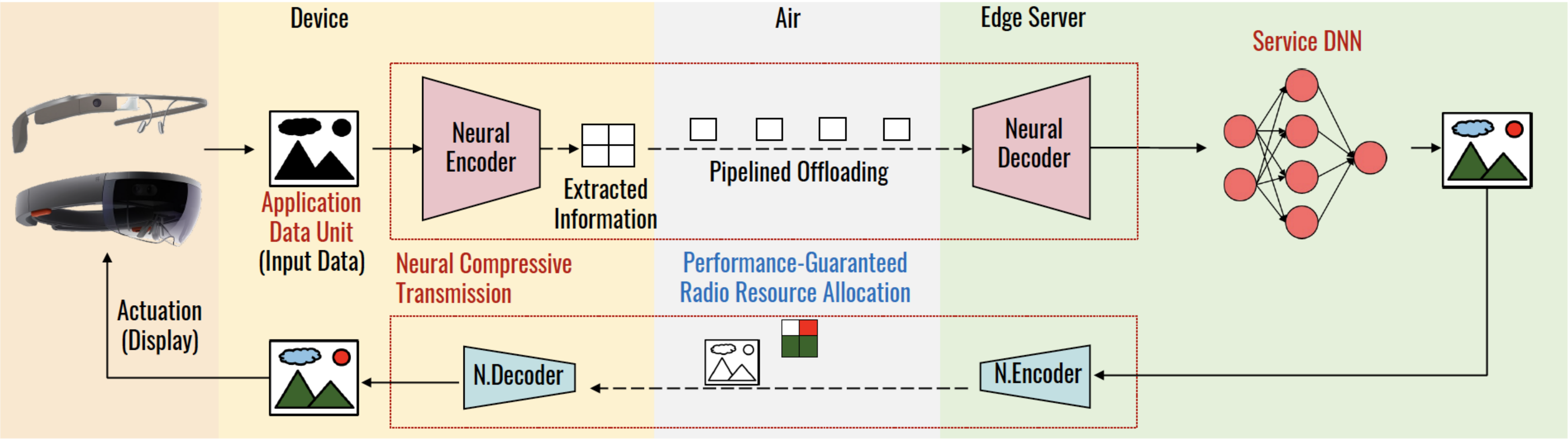
30% Average Utilization of Traditional RAN-only

COMPLEXITY OF AI/ML AND (6G) SIGNAL PROCESSING IN RAN IS DOMINATED BY VECTOR-MATRIX MULTIPLICATION



Source: S. Shukla et al. “A Scalable Multi-TeraOPS Core for AI Training and Inference,” IEEE Solid-State Circuits Letters.

CELLULAR 2.0: ENABLING PERFORMANCE-GUARANTEED NETWORKED COMPUTING



J. Kim, B. Shim, and K. Lee, "Towards Enabling Performance-Guaranteed Networking in Next-Generation Cellular Networks," IEEE Communications Magazine, vol. 61, no. 1, pp.32-37, Jan. 2023

CELLULAR NETWORK SHOULD BE NO MORE A BLACKBOX TO APPLICATIONS!

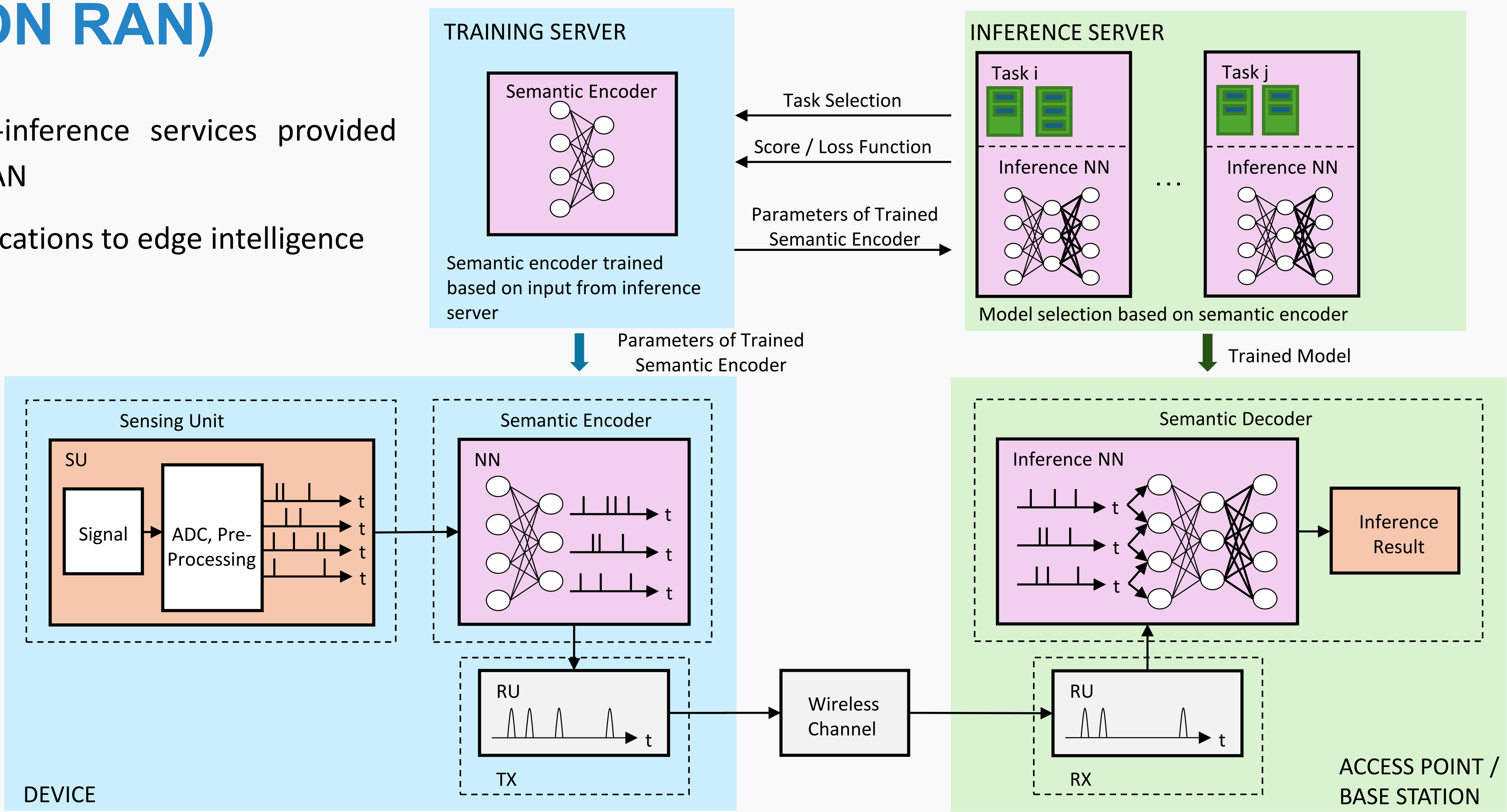
KEY PERFORMANCE METRICS FOR EDGE LEARNING & INFERENCE (from 6G-XCEL)

KPMs for AI/ML capabilities in relation to edge learning and inference	Description
Training complexity	Number of real-valued operations needed for training an AI model until convergence (assuming fixed input data distribution).
Inference complexity	Number of real-valued operations needed for pre-, post-processing, and inference in an AI model. Can also be characterized through the number of real-valued model parameters.
AI/ML- related communication overhead	Overhead incurred for assistance information, data collection, model delivery/transfer, and other required signalling.
Model generalization capability	A model's ability to perform under unseen scenarios/data distributions.
AI/ML performance	Metrics to access the performance of AI/ML models such as, e.g., - accuracy, recall, false-positives rate and precision for classification tasks - MSE, RMSE for regression tasks.
Inference latency	Time delay between the data input and the inference output (including over-the-air transmission in split-inference scenarios).

6G-XCEL Deliverable 2.2: Comprehensive analysis of benefits and implications of the AI/ML-based network control deployment (DTAG, HHI)

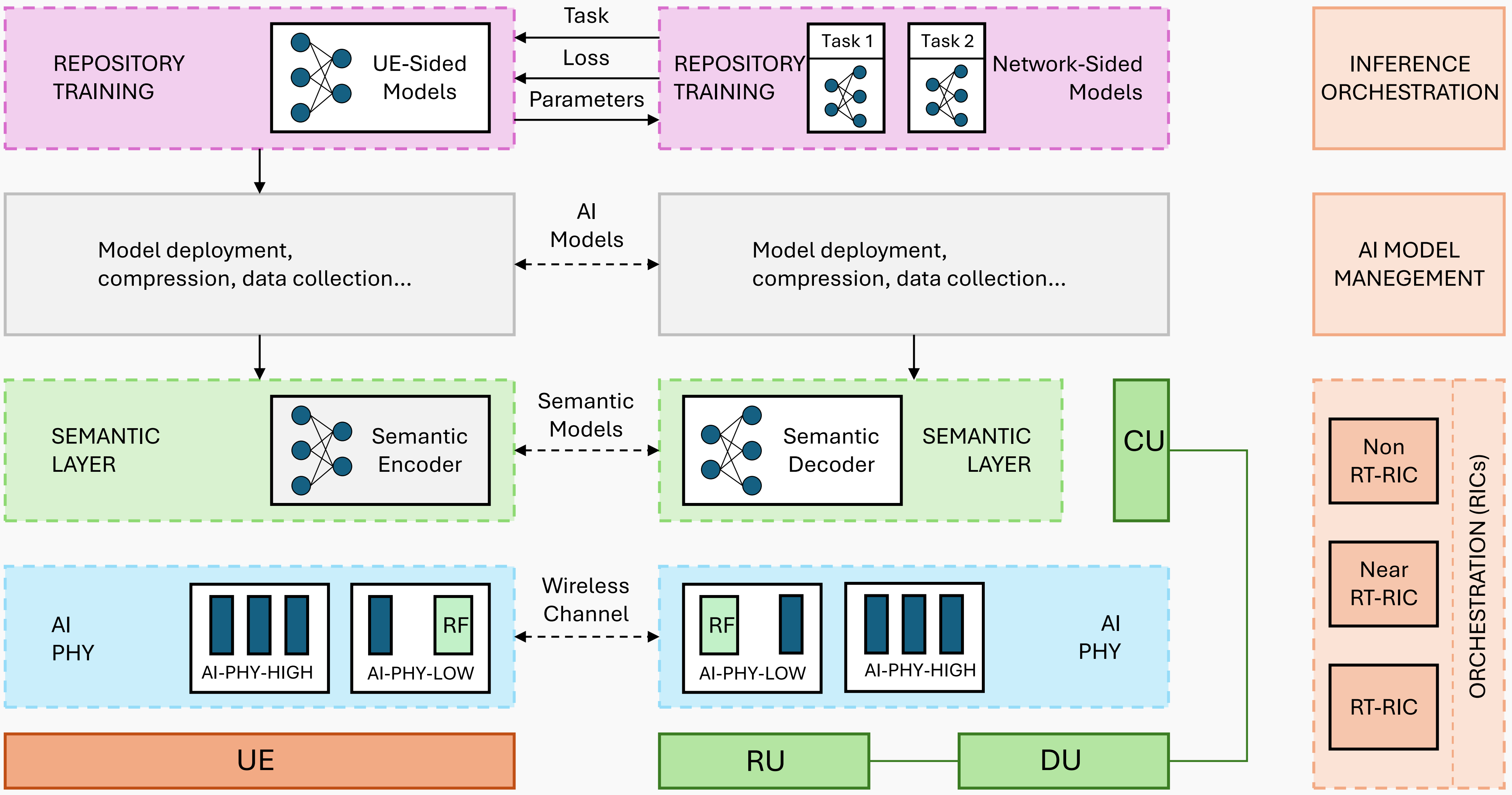
FUNCTIONAL ARCHITECTURE FOR EDGE INFERENCE (AI ON RAN)

- Edge-inference services provided by RAN
- Applications to edge intelligence



UNDER DISCUSSION IN 6G-XCEL

“CELLULAR 2.0: INFERENCE AS A SERVICE”?



MAPPING TO 3GPP: STARTING POINT – AI / ML FOR MEDIA

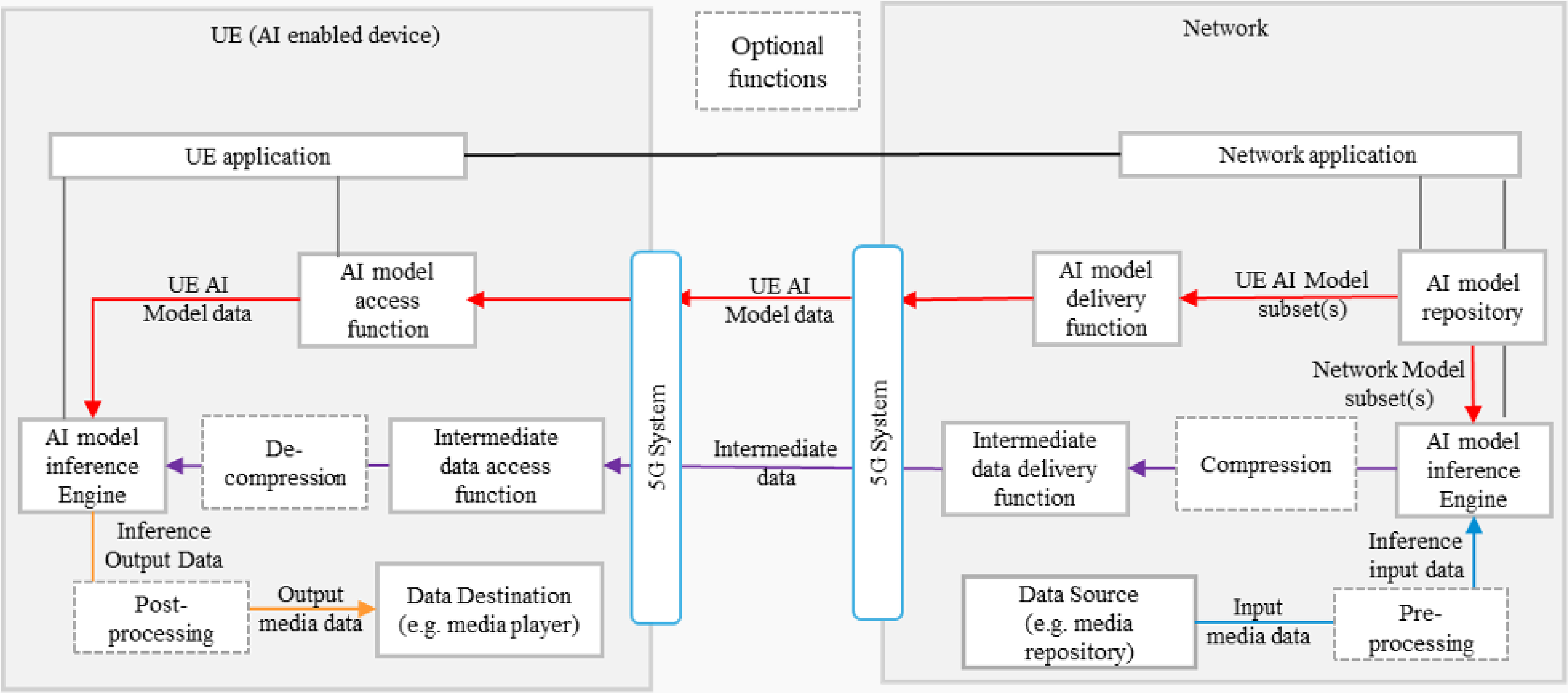


Figure: An example of a basic architecture for split inferences between the UE and the network, as described in 3GPP TR 26.927. The media data source originates from the UE, the first part of the inference is performed in the UE, the second part in the network.

AI FOR RAN OR RAN FOR AI?

CONVERGENCE OF COMMUNICATION AND COMPUTATION (AND SENSING AND CONTROL)

6G-RIC

Research and
Innovation Cluster

https://6g-ric.de