UC San Diego

Offload Rethinking by Cloud Assistance for Efficient Environmental Sound Recognition on LPWANs

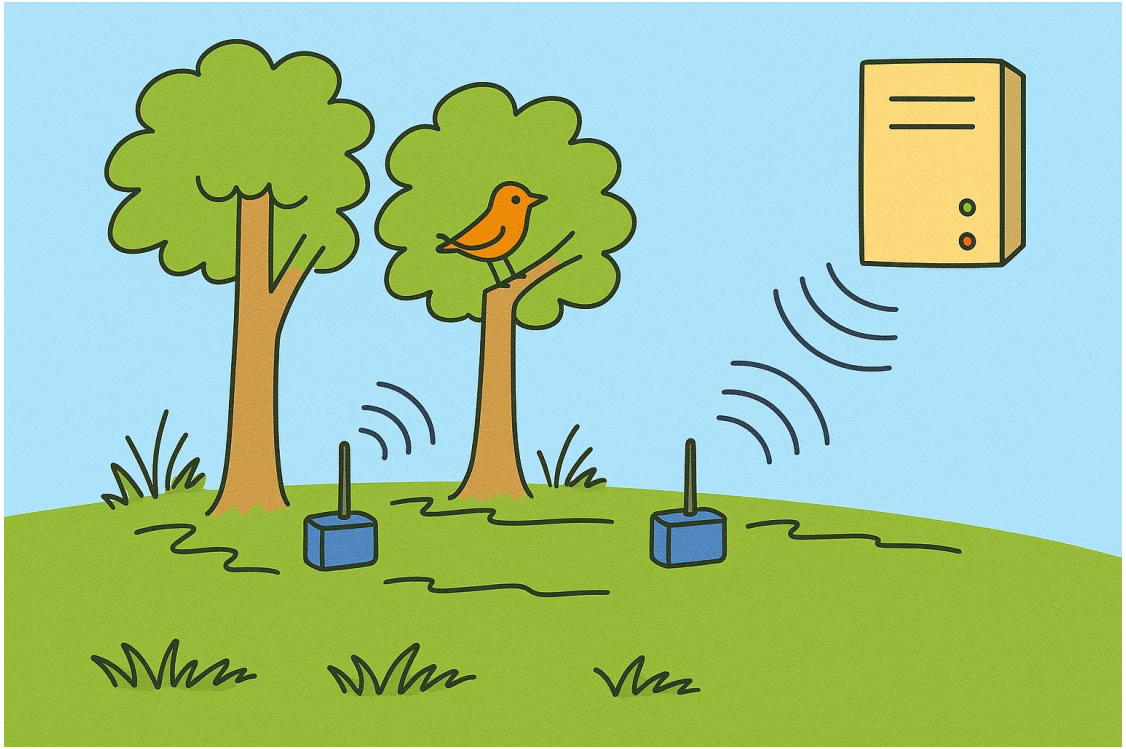
Le Zhang*, Quanling Zhao*, Run Wang, Shirley Bian, Onat Gungor, Flavio Ponzina, Tajana Rosing
University of California San Diego (UCSD)

(*These authors contributed equally to this research)

System Energy Efficiency Lab

seelab.ucsd.edu

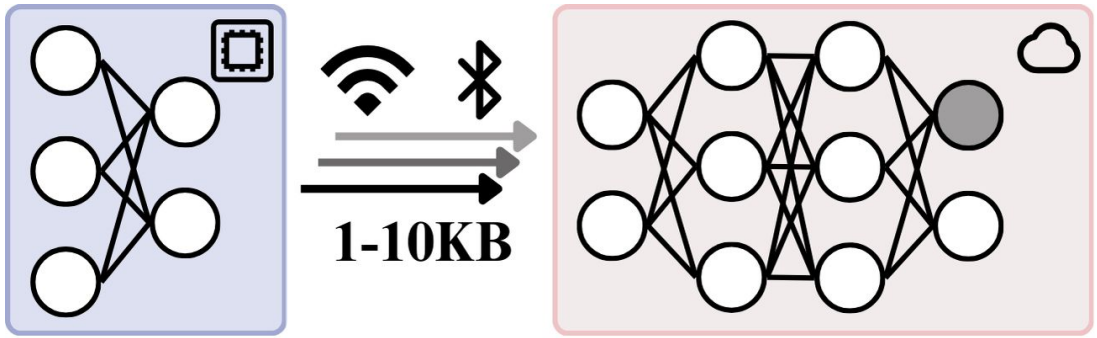
Background: Batteryless intelligent acoustic applications



System Requirements:

1. Wide-area application
2. Resource efficiency
3. Reliable service
4. Accurate prediction

Baseline: Cloud offloading [IPSN '24]



(b) Cloud Offloading

Ahn J. et al. "Split Learning-based Sound Event Detection in Energy-Constrained Sensor Devices." IPSN '24

Baseline: Cloud offloading [IPSN '24]

Pros:

+ **High accuracy** by leveraging server

+ **LPWANs** to support **long-range wide-area** communication

Cons:

~~**Short range radio** (WiFi, BLE)~~

- **Resource inefficient: Large payload and multi-round communication**

- **Not reliable and dependent on server**

?

Prior Works	Edge Optimize	Wireless	Payload Size	Comm-adapt	ULP	Reliable
DeepCOD [58]	Cloud offloading	Wi-Fi,LTE	0.1-5KB	✓	✗	✗
FLEET [17]	Early exits	BLE	2-9KB	✓	✗	✗
SEDAC [3]	Cloud offloading	BLE	N/A	✗	✓	✗
CACTUS [42]	Micro-classifier	N/A	11.5-43.9MB	✗	✗	✗
LimitNet [15]	Cloud offloading	LPWANs	0.3-3.2KB	✓	✗	✗
ORCA	**Cloud assistance**	**LoRa**	**0-0.1KB**	✓	✓	✓

Reliability

Wide-area scenarios

Resource efficiency

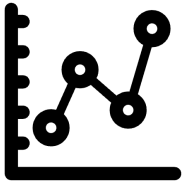
[58] Yao S. et al. "Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency." SenSys '20
[17] Huang J. et al. "Re-thinking computation offload for efficient inference on IoT devices with duty-cycled radios." MobiCom '23
[3] Mohammad Mehdi R. etal. "CACTUS: Dynamically Switchable Context-aware micro-Classifiers for Efficient IoT Inference." MobiSys '24
[42] Hojjat A. et al. "LimitNet: Progressive, Content-Aware Image Offloading for Extremely Weak Devices & Networks." MobiSys '24
[15] Ahn J. et al. "Split Learning-based Sound Event Detection in Energy-Constrained Sensor Devices." IPSN '24

Challenges for offloading on LPWANs

High energy cost & Low data rate

- 40x for uplink, 6x power for downlink
- 0.3–5.5 kbps (125 kHz LoRa channel)
- Audio has large payload (44.1kHz)

Minimize payload size & audio-adapted offloading



Unstable wireless channel & Dynamic energy cost

- Environmental factors change channel condition
- Reliable comm. requires dynamic energy cost
- Budgets are fixed

Adaptation to dynamic energy costs

Unreliable offloading

- Frequent packet loss
- Costly retransmission

Support cloud-independent inference

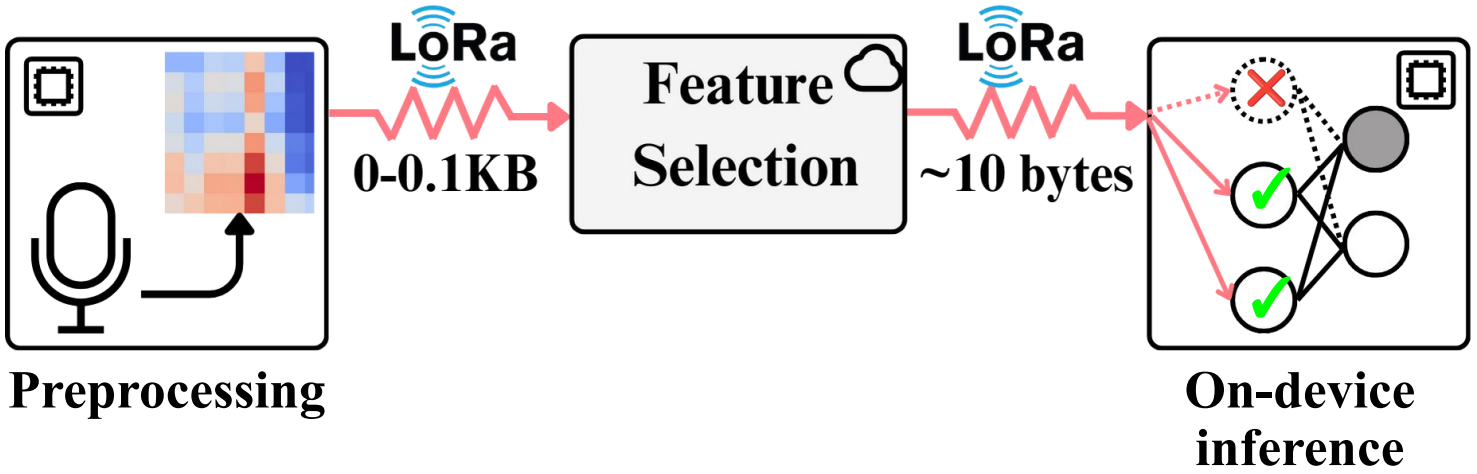
Research Question

Besides generating inference results, what tasks can a server perform under weak connectivity and limited resources?

Cloud Assistance System Design

Step 1: *Low-resolution compression* for *cloud feature selection*

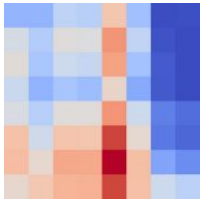
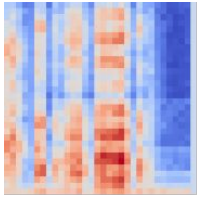
- Feature selection is a resource-intensive process, should be done by server
- Doesn't require high resolution → **smaller uplink payloads (<100 bytes)**
- Feature importance requires **even smaller downlink payloads (~10 bytes)**



Cloud Assistance System Design

Step 1: *Low-resolution compression* for *cloud feature selection*

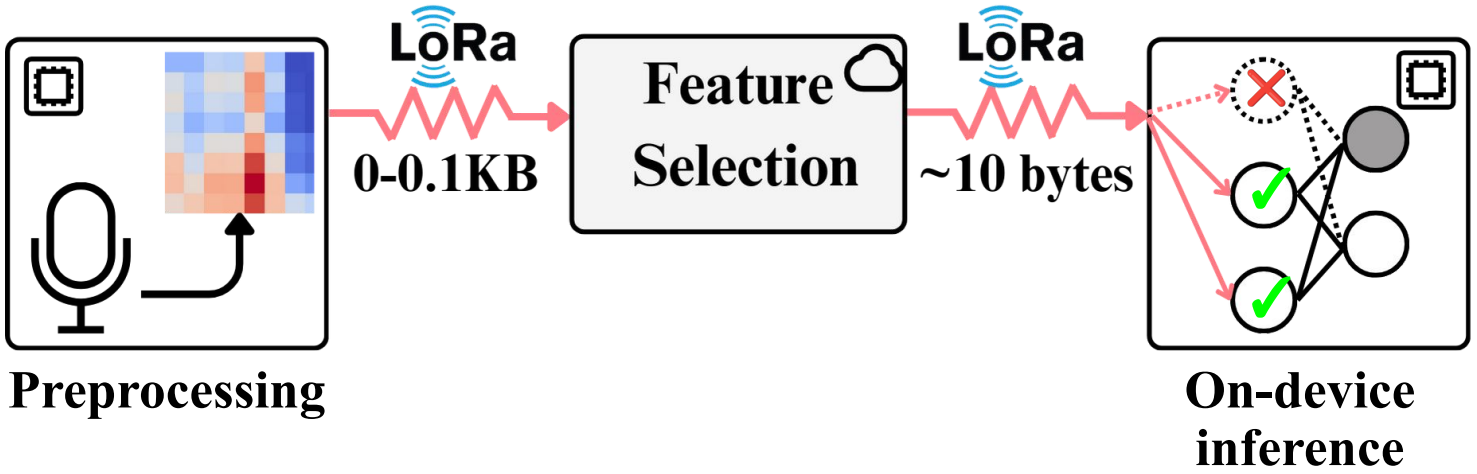
- Feature selection is a resource-intensive process, should be done by server
- Doesn't require high resolution → **smaller uplink payloads (<100 bytes)**
- Feature importance requires **even smaller downlink payloads (~10 bytes)**

Resolution	Classification	Feature selection	Small payloads
Low-resolution 	✗	✓	✓
High-resolution 	✓	✓	✗

Cloud Assistance System Design

Step 1: *Low-resolution compression* for *cloud feature selection*

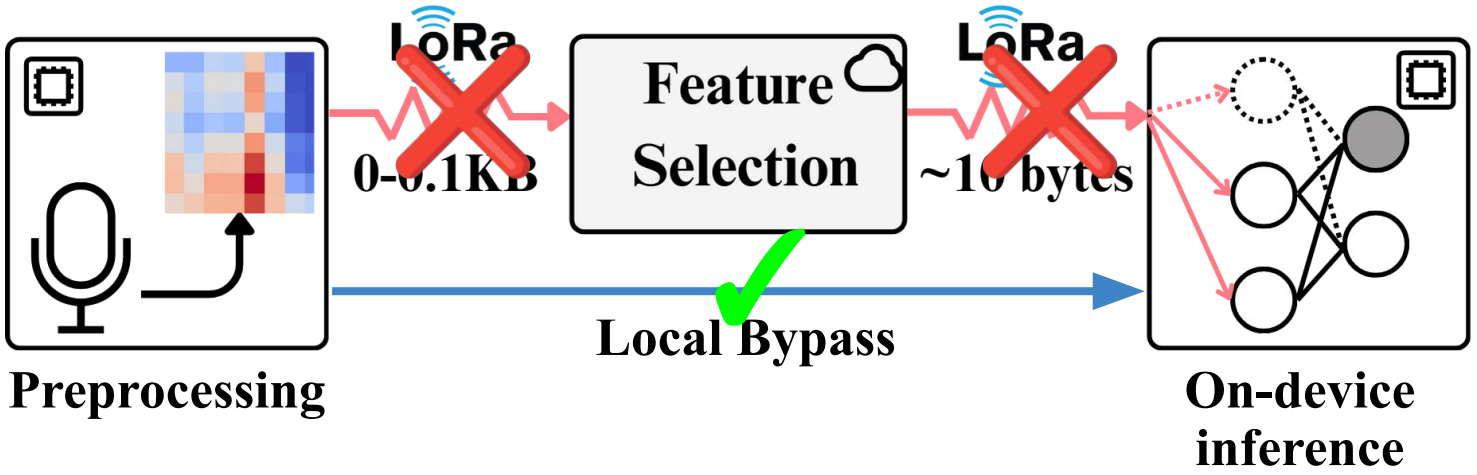
- Feature selection is a resource-intensive process, should be done by server
- Doesn't require high resolution → **smaller uplink payloads (<100 bytes)**
- Feature importance requires **even smaller downlink payloads (~10 bytes)**



Cloud Assistance System Design

Step 2: Local bypass for *reliable cloud-independent inference*

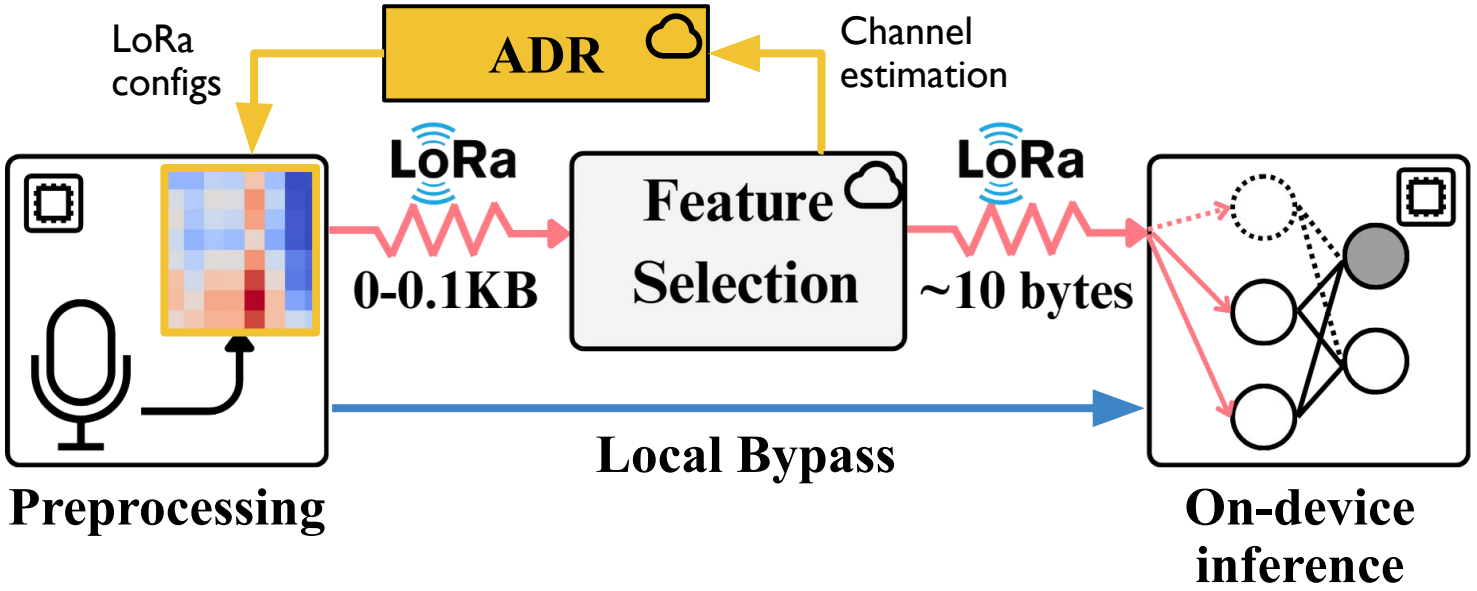
- Keep the inference pipeline on-device
- When packet losses occur, perform **on-device inference**
- **No retransmission required** → Communication and energy efficiency



Cloud Assistance System Design

Step 3: Optimization for *Dynamic Energy Cost* under Fixed Budget

- Wireless channel is unstable → energy cost for reliable transmission is also variable
- Use LoRa configs from Adaptive Data Rate (ADR) to estimate the cost of transmission



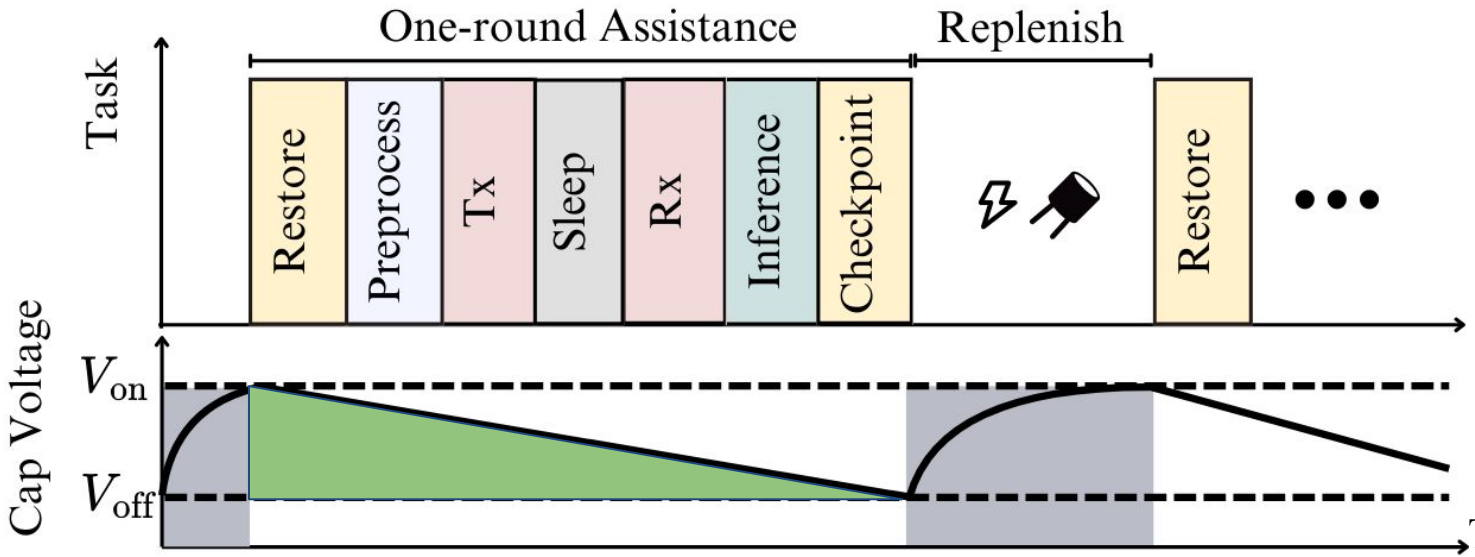
Step 3: Optimization for Dynamic Energy Cost

Energy constraints

Fixed budgets

$$E_{\mathrm{Tx}}(\mathrm{x}) + E_{\mathrm{Pre}} + E_{\mathrm{sleep}} + E_{\mathrm{Rx}} + E_{\mathrm{inf}} \leq E_{\mathrm{cap}}$$

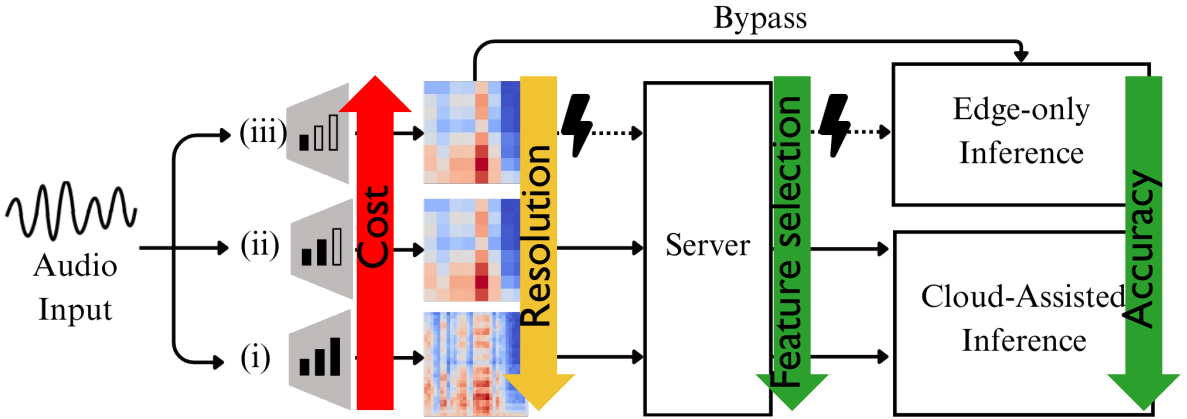
Variable energy cost for uplink



Step 3: Optimization for Dynamic Energy Cost

Intuition: when cost is high, use smaller payload/ lower resolution trade accuracy

↑ Energy cost per byte by configs from ADR $E_{Tx}(x)$ Payload / Resolution ↓ $\leq$ constant budget



Step 3: Optimization for Dynamic Energy Cost

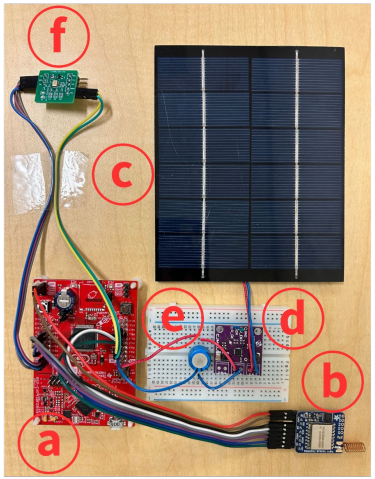
Optimize uplink resolution for accuracy given costs and energy constraints

Accuracy Cost Res Energy constraint

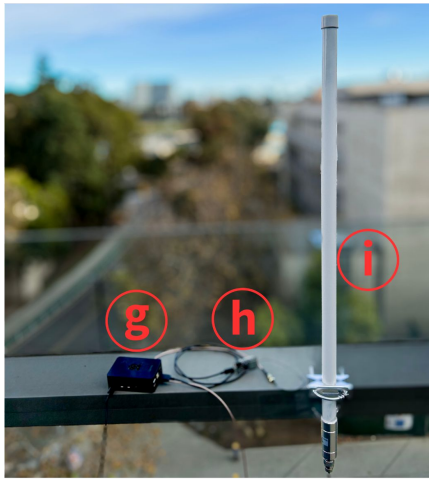
$$\max_{x} \boxed{a^T x} \quad \text{s.t.} \quad \boxed{E_{\text{Tx}}}\boxed{(x)} + \boxed{E_{\text{Pre}} + E_{\text{sleep}} + E_{\text{Rx}} + E_{\text{inf}} \leq E_{\text{cap}}}$$

$$\mathbf{1}^T x = 1, \; x_i = \{0, 1\}$$

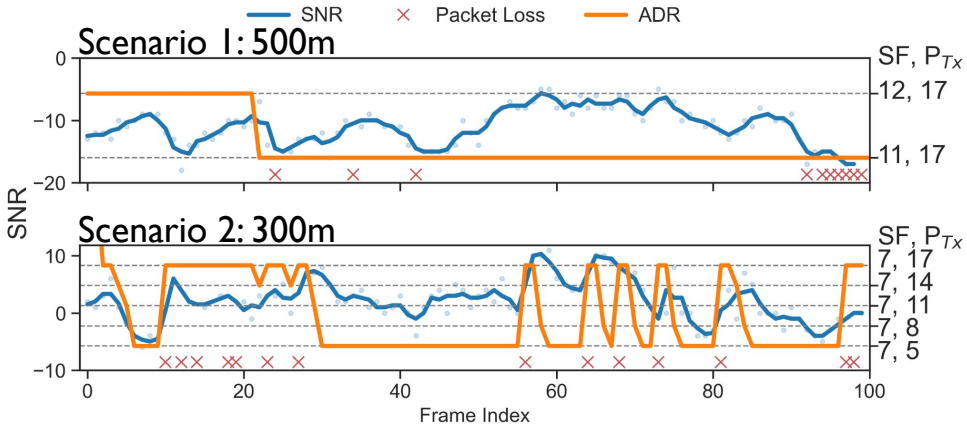
Testbed Implementation



Edge device



Server



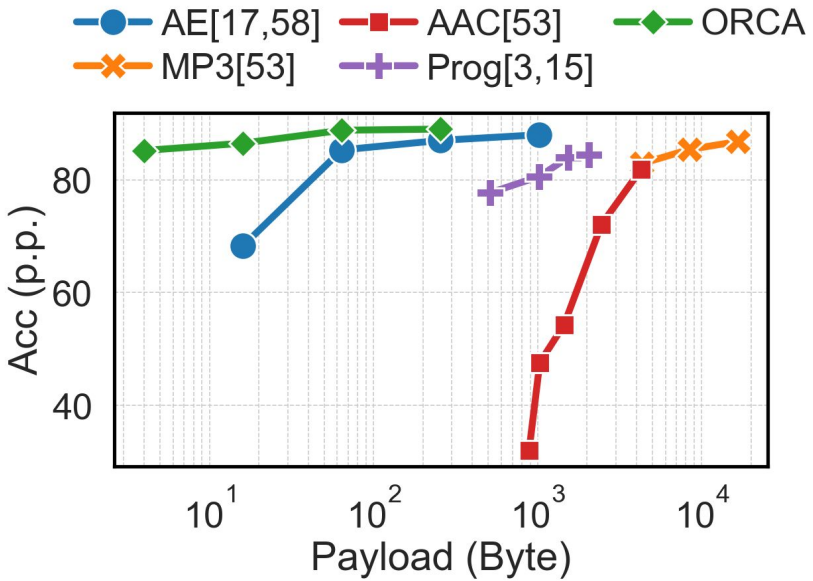
Traces

System Evaluations

Datasets: ESC50, US8K, DESED

Baselines:

- Autoencoder [SenSys '20, MobiCom '23]
- Audio compression: MP3 and AAC
- Progressive offloading with time-domain attention [MobiSys '24, IPSN '24]
- On-device inference [IWMUT '19]

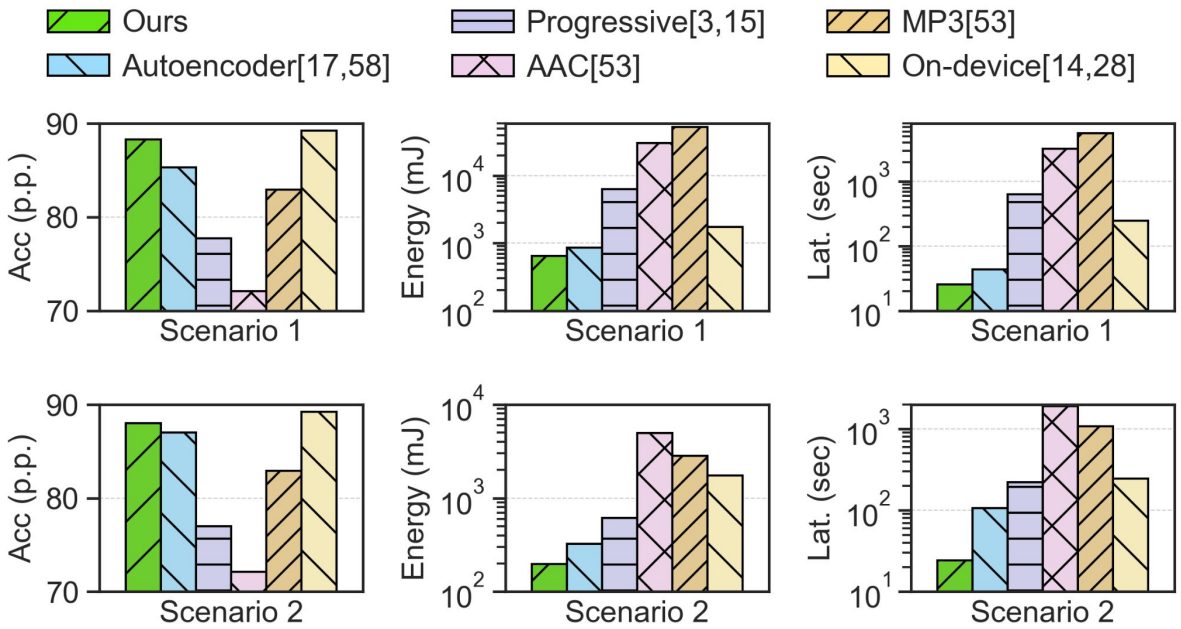


Payload: **4x-8x** payload savings with **5–20 p.p.** accuracy advantages

[58] Yao S. et al. "Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency." SenSys '24
[17] Huang J. et al. "Re-thinking computation offload for efficient inference on IoT devices with duty-cycled radios." MobiCom '23
[15] Hojjat A. et al. "LimitNet: Progressive, Content-Aware Image Offloading for Extremely Weak Devices & Networks." MobiSys '24
[3] Ahn J. et al. "Split Learning-based Sound Event Detection in Energy-Constrained Sensor Devices." IPSN '24
[28] Lee, S., et al. Intermittent learning: On-device machine learning on intermittently powered system. IMWUT '19

System Evaluations

Energy & latency: **80x** Energy savings, **220x** Latency reductions with comparable accuracy



[58] Yao S. et al. "Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency." SenSys '24
[17] Huang J. et al. "Re-thinking computation offload for efficient inference on IoT devices with duty-cycled radios." MobiCom '23
[15] Hojjat A. et al. "LimitNet: Progressive, Content-Aware Image Offloading for Extremely Weak Devices & Networks." MobiSys '24
[3] Ahn J. et al. "Split Learning-based Sound Event Detection in Energy-Constrained Sensor Devices." IPSN '24
[28] Lee, S., et al. Intermittent learning: On-device machine learning on intermittently powered system. IMWUT '19

Conclusion

We propose a novel cloud-assistance ML framework on LPWANs

- Save **4x-8x** payloads through **cloud assistance framework**
- Improve **5-20 p.p.** accuracy using **cloud feature selection and on-device inference**
- Enable **local bypassing** to reduce the cloud dependency and improve reliability
- Reduce energy cost by **80x** and latency by **220x** **under dynamic wireless channel**

Read our paper here!

Thank you for listening!

Q&A