TWO WATTS IS ALL YOU NEED

Enabling In-Detector Real-Time Machine Learning for Neutrino Telescopes Via Edge Computing

This Work: MJ, Y. Hu, C.A. Argüelles, JCAP 06 (2024) 026
arxiv 2311.04983

Enabling In-Detector Real-Time Machine Learning for Neutrino Telescopes Via Edge Computing

Edge Computing

- *"Edge computing is a distributed computing framework that brings enterprise applications closer to data sources such as IoT devices or local edge servers."*
—IBM

- In other words, it refers to data processing very close, if not on the site of, data acquisition.

- It is efficient, low-latency and scalable.

Example: Traffic Light Control

- Key problems:
  - Latency: observation and decision are time-delayed, but many times traffic flow needs immediate attention
  - Scalability: central facility can only process a number of crossroads, prioritizing over some and neglecting others by choice
    - Power Consumption
    - Data transmission
- Solution:
  - Data processing “on the edge”

The Neutrino Telescope Analogy

- Processing: In the detector, we trigger on local coincidences; in the local lab, we apply simple line fit or regression methods

- Scale:
  - Data transmission: we select the triggered/filtered data and send them to a central facility for further, more complicated reconstruction and treatment
  - Power consumption: we do not require (nor do we have access to) a lot of power on the site, but we have huge supercomputer clusters in a centralized location

The Neutrino Telescope Analogy

- Latency: In the detector, we trigger on local coincidences; in the local lab, we apply simple line fit methods. We are not aware of interesting signals that require sophisticated treatment until we see the data in the centralized facility.

- Scale:

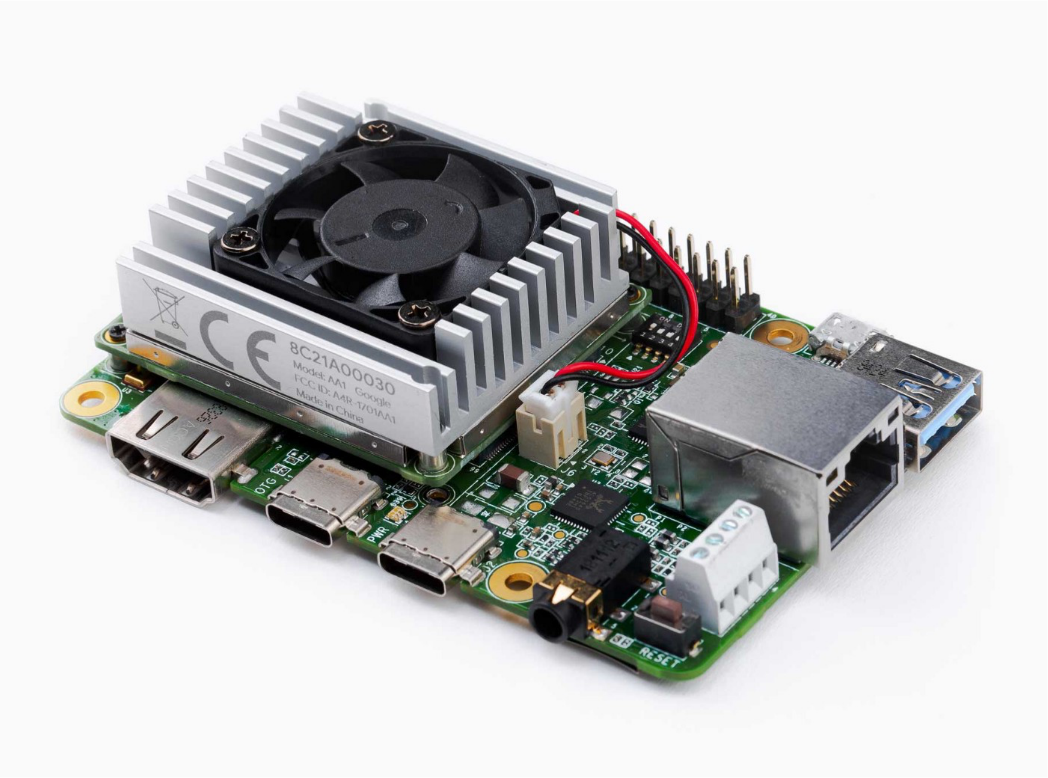
  - Data transmission: we select the triggered/filtered data and send them to a central facility for further, more complicated reconstruction and treatment

  - Power consumption: we do not require (nor do we have access to) a lot of power on the site, but we have.huge supercomputer clusters in a centralized location

  - As we move forward to larger detectors, pressure on both data transmission and power consumption we be further exacerbated, forcing us to postpone (even give up) transporting and processing a larger fraction of data.

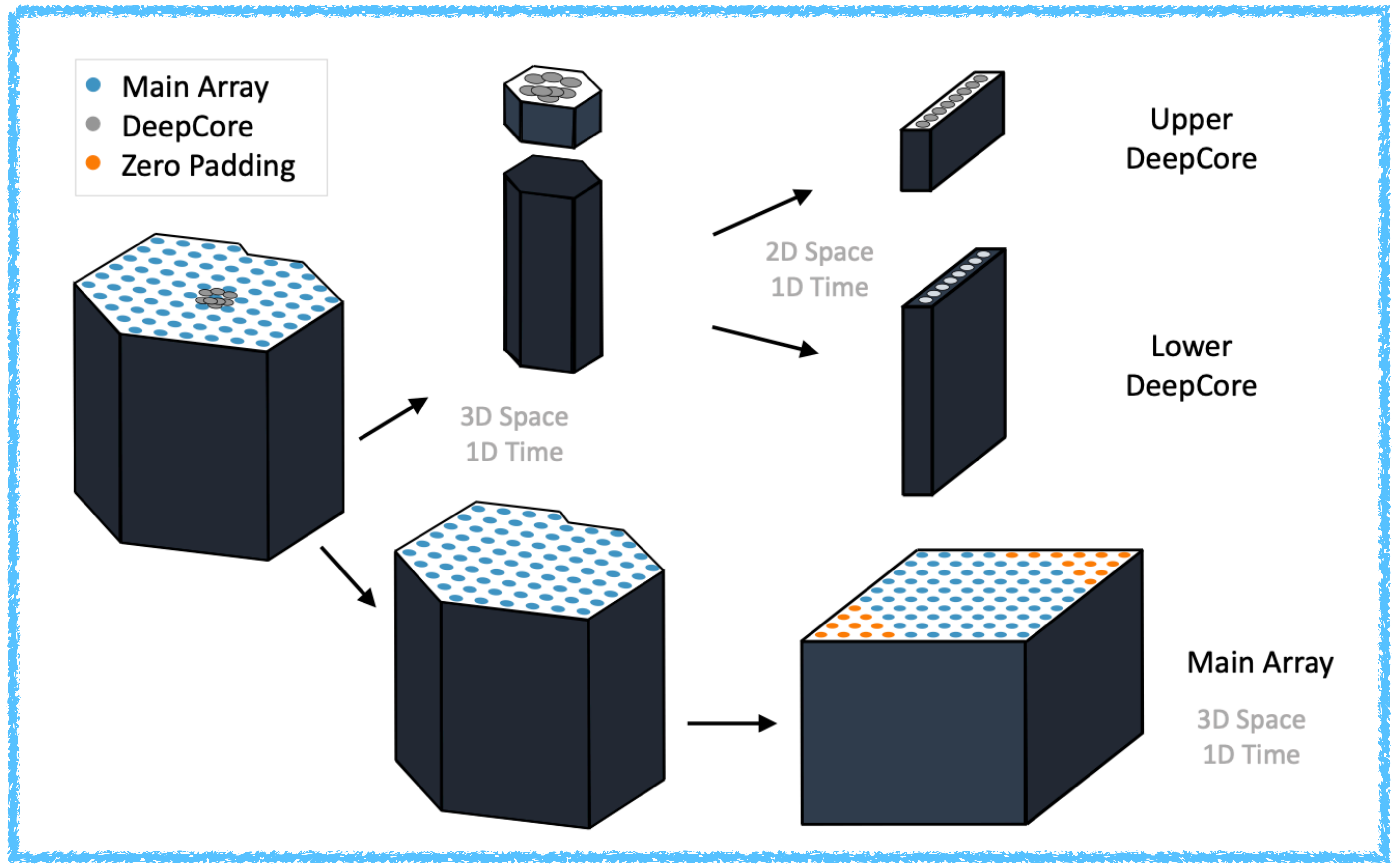
Realization via Edge TPUs

- Features that enable edge computing for us:
  - Low power consumption: 2 watts
  - Versatile utilization and coding: general-purpose computing chip
  - *<u>Specifically engineered for speeding up ML **inference** (enabled by MXUs)</u>*



Difficulties to Overcome

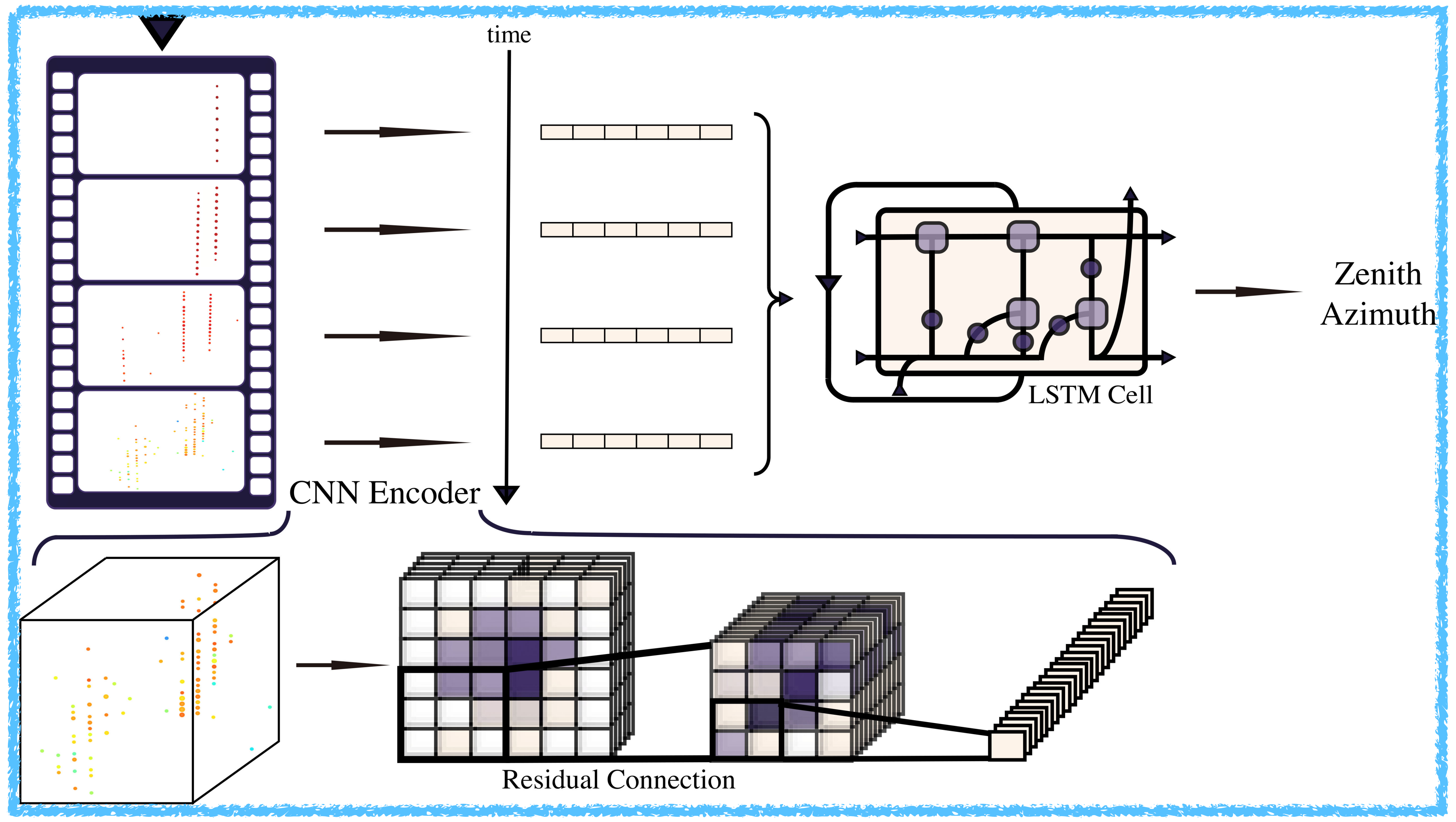
- Types of operations and data format allowed:
    - > 3-dimensional tensors not allowed: convolution limited to 2 dimensional grid data with an extra channel dimension



IceCube Collaboration

Difficulties to Overcome

- Types of operations and data format allowed:
  - 3-dimensional tensors not allowed: convolution limited to 2 dimensional grid data with an extra channel dimension
  - We convert to *Recurrent Neural Network*



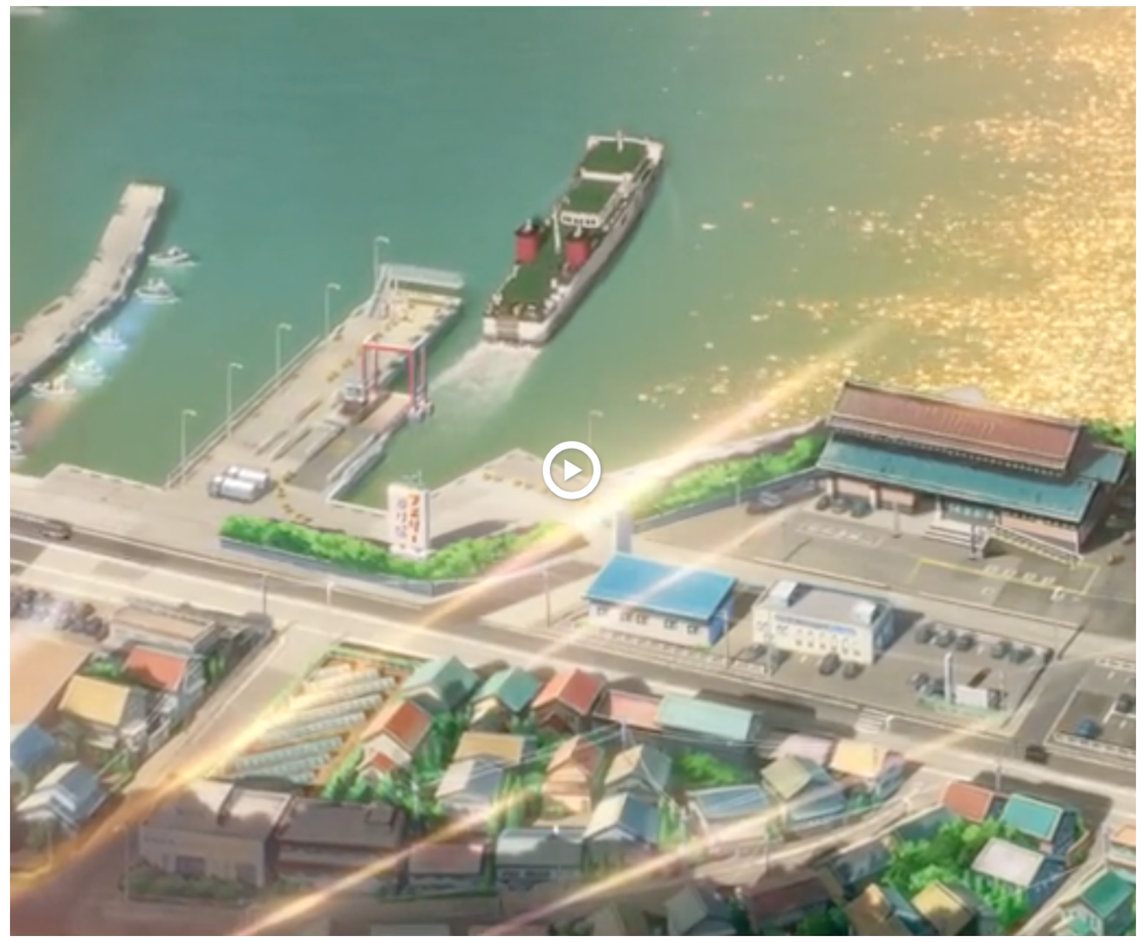
MJ, Y. Hu, C.A. Argüelles

Rethink: Time-series "Speech" problem

A door is opened

Disaster emerges

Door is closed,
Disaster is avoided



They set on a trip to close opened doors

It is a fantasy movie reflecting upon the people, places and associated emotions and histories surrounding natural disasters

Difficulties to Overcome

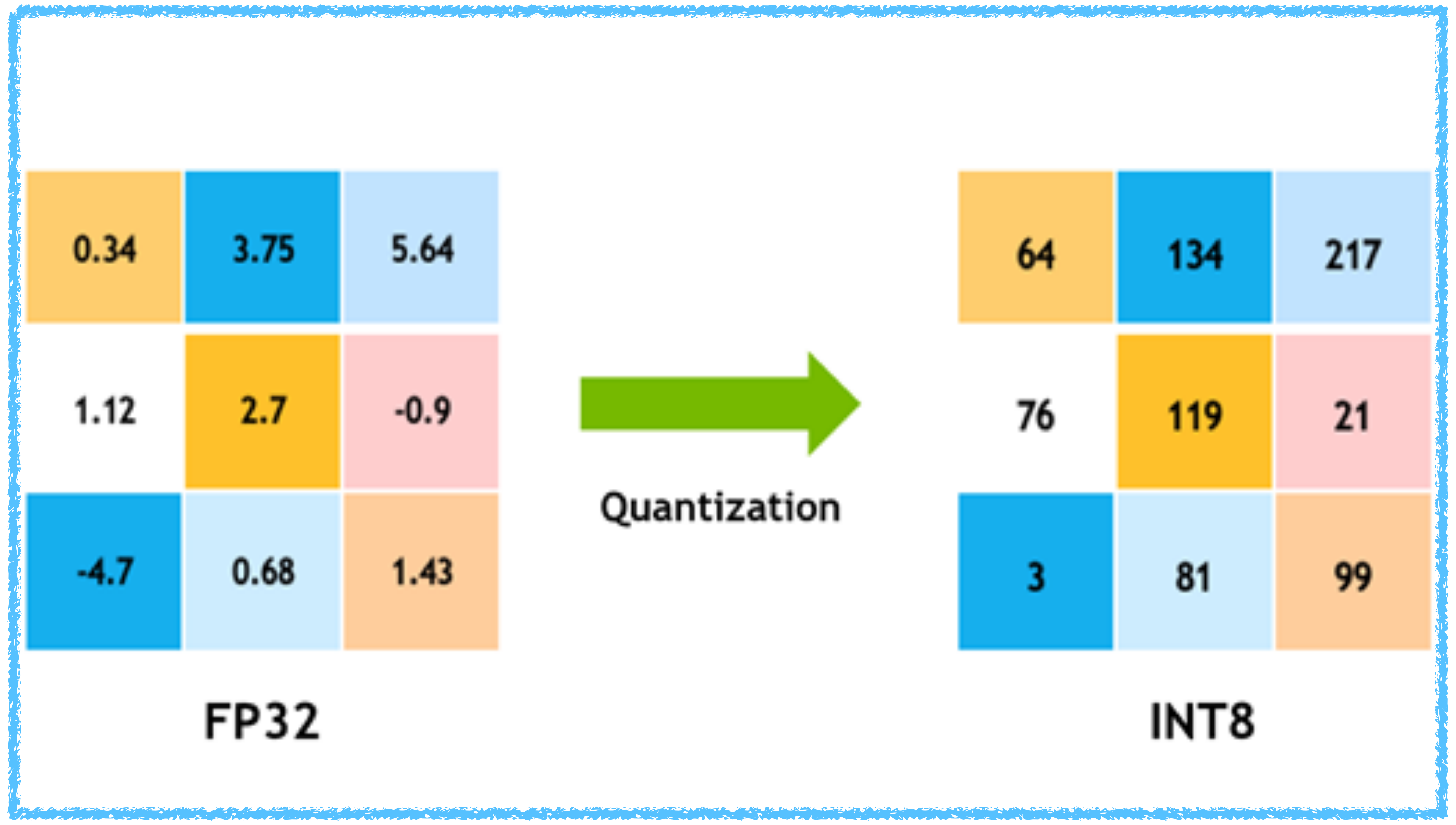
- Types of operations and data format allowed:
  - > 3-dimensional tensors not allowed: convolution limited to 2 dimensional grid data with an extra channel dimension
  - We convert to *Recurrent Neural Network*
- Precision of computation:
  - Only 256 integers are allowed in real-time inference on a TPU.

FP32

Neta Zamora et al.

Difficulties to Overcome

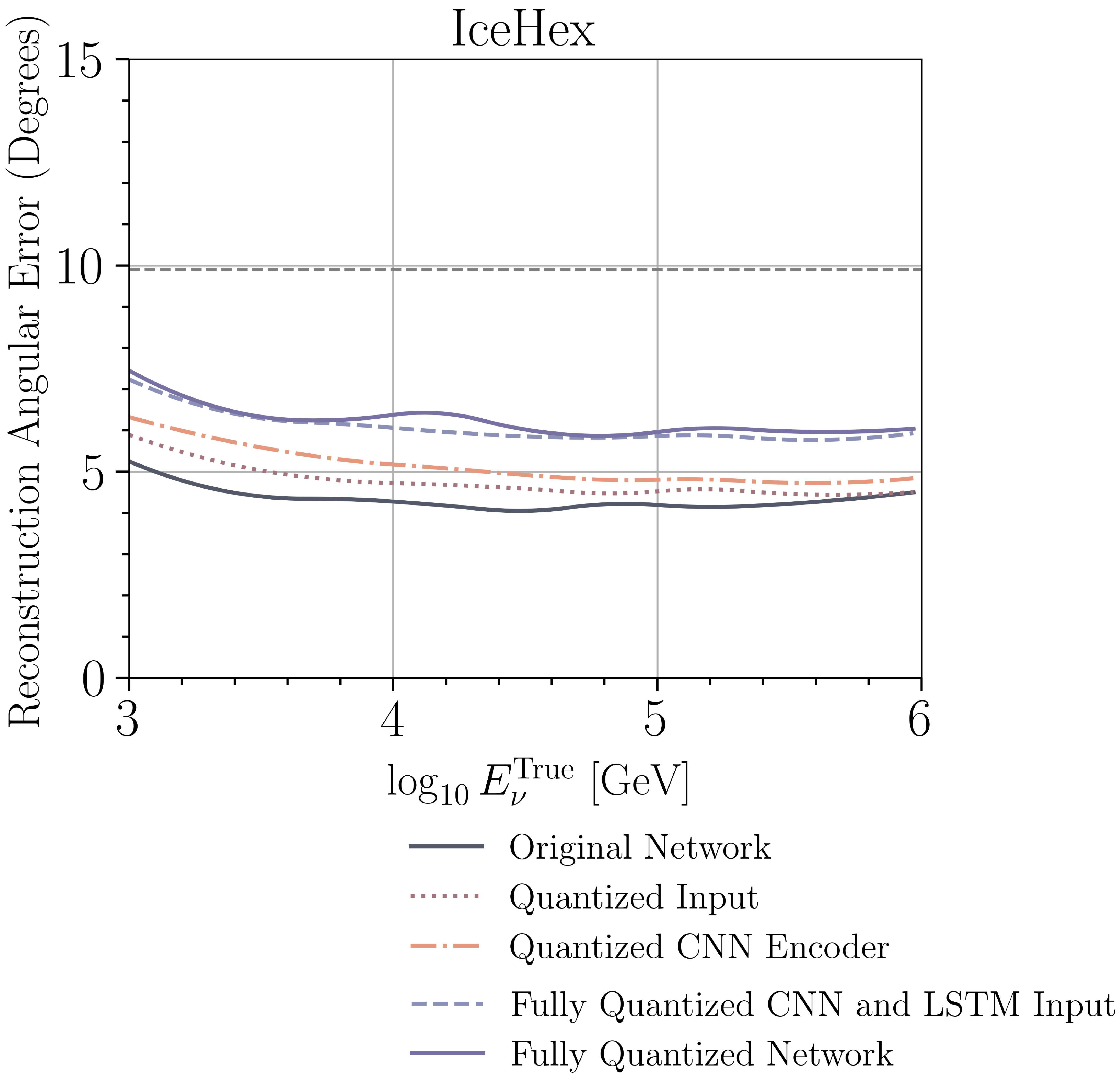
- Types of operations and data format allowed:
  - > 3-dimensional tensors not allowed: convolution limited to 2 dimensional grid data with an extra channel dimension
  - We convert to *Recurrent Neural Network*
- Precision of computation:
  - Only 256 integers are allowed in real-time inference on a TPU.
  - We apply *quantization* to the weights



Neta Zamora et al.

Demonstration on Reconstruction Task

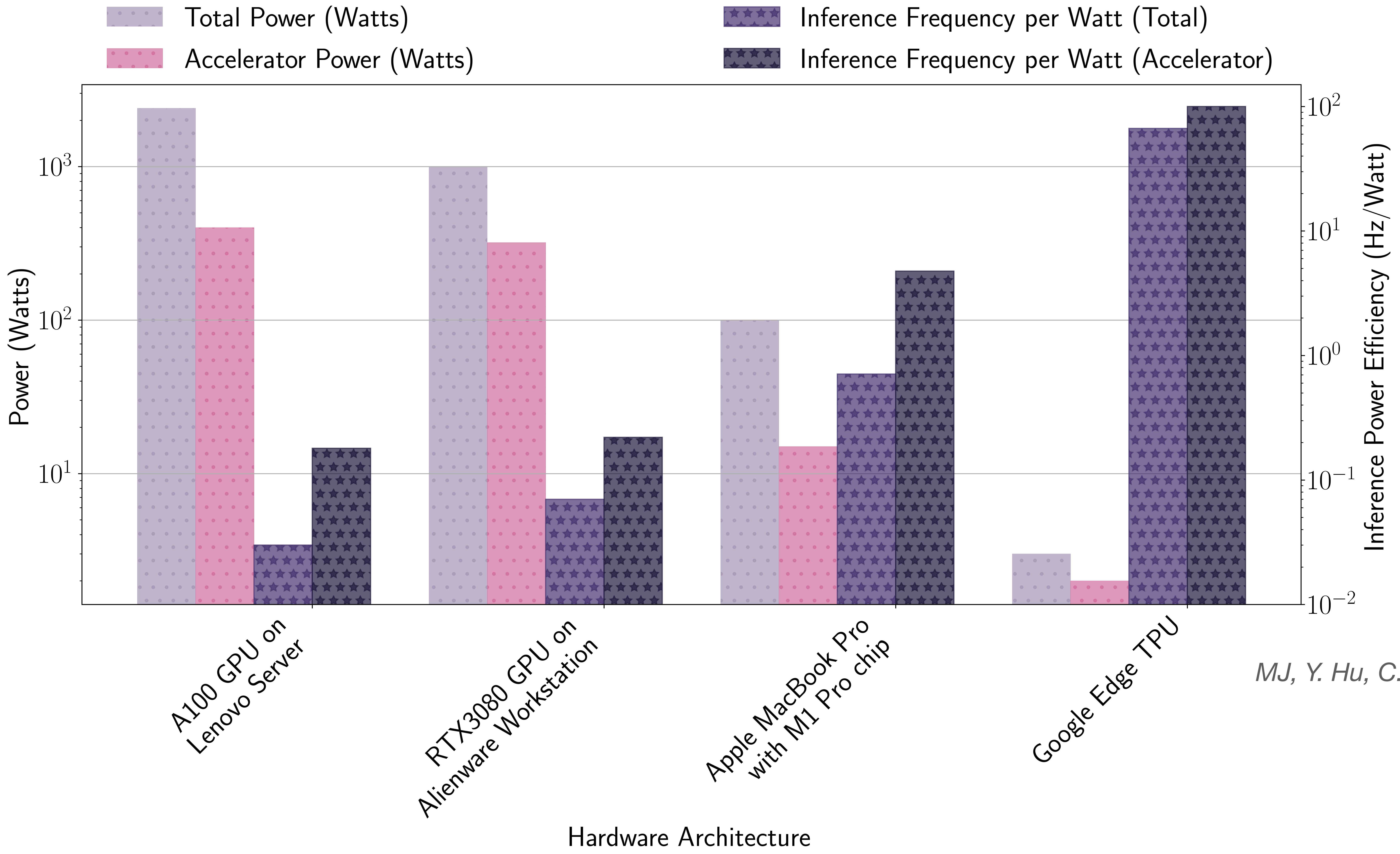
- We demonstrate the feasibility of edge computing by performing an angular reconstruction task on simulated data. The network is capable of recovering a good resolution despite the restriction on data formatting and precision.



MJ, Y. Hu, C.A. Argüelles

Demonstration on Reconstruction Task

- … and at an astonishing power efficiency



MJ, Y. Hu, C.A. Argüelles

Future Prospects

- The reconstruction task is chosen to demonstrate feasibility, but it is not the only task (not even a good task) for edge computing to shine. Here are some future prospects of such a technology:

- ***Real-time data processing***: trace/waveform-based in-detector triggering system

- ***Data compression***: generative model for encoding data to alleviate transmission limitations

- Any large-scale experiment with limited power access and data bandwidth (e.g. satellites)

- …

Thank you!

Hold on…

- These are come caveats and concerns that might be on your mind:
- Q: Why do we need these TPUs in the first place?
  - A: We don't necessarily need them, but they will help a lot
- Q: (ctnd) Why don't we just go with GPUs in the local on-site laboratory?
  - A: It is a great idea. Another paper has explored the option of accelerating reconstruction on GPU, see *F. Yu et al.* Even in this case, a lower-level TPU implementation would make the pipeline even more scalable.
- Q: Do TPUs take full control over how the data is processed from the lowest level?
  - A: No, it is possible to implement a "seatbelt" that circumvents the TPU treatment of trace/pulse data. (Ironically) that can be implemented too on the TPU thanks to its coding versatility.
- Q: You have shown TPUs work on simulated hit-level data, but you are advocating for TPUs to be employed on DOMs, how do you know a network on trace-level would also work post-quantization?
  - A: Unfortunately, we do need further investigation and algorithm development before claiming this technology to be ready. We still got a long way to go…