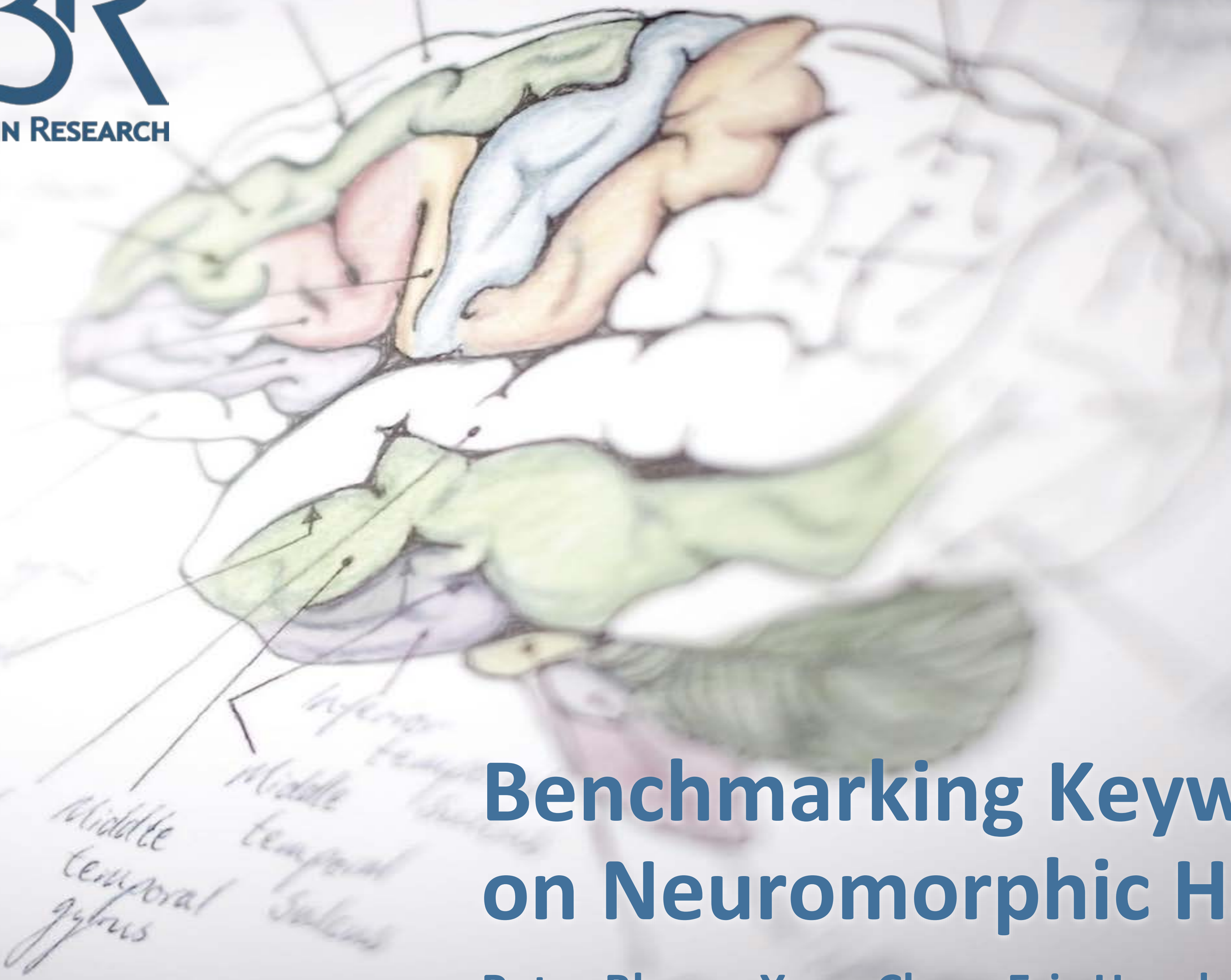
APPLIED BRAIN RESEARCH



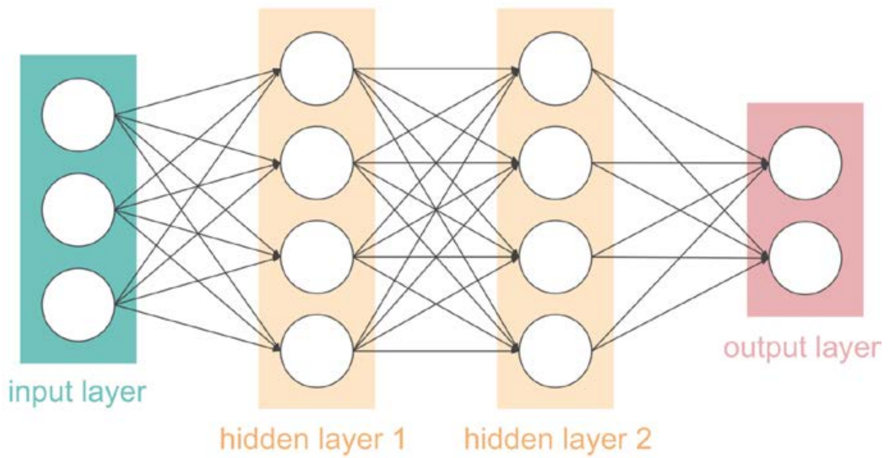
Benchmarking Keyword Spotting on Neuromorphic Hardware

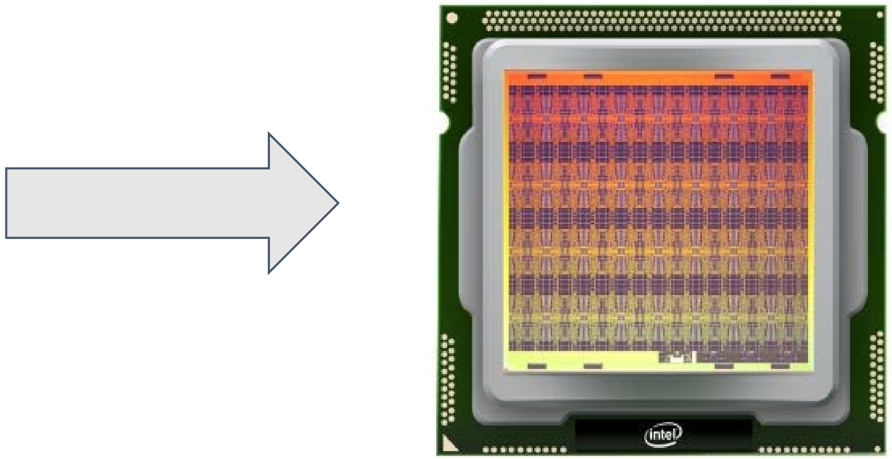
Peter Blouw, Xuan Choo, Eric Hunsberger, & Chris Eliasmith

Background

- **Neuromorphic chips exploit architectural parallelism, event-driven computing**
  - Temporal sparsity improves power efficiency, parallelism improves latency

- **Intel’s Loihi chip is a great tool for analysing the benefits of neuromorphic HW**
  - Software from ABR can be used to run high-level applications on Loihi
  - Specifically, we can convert arbitrary DNNs to functionally comparable SNNs.



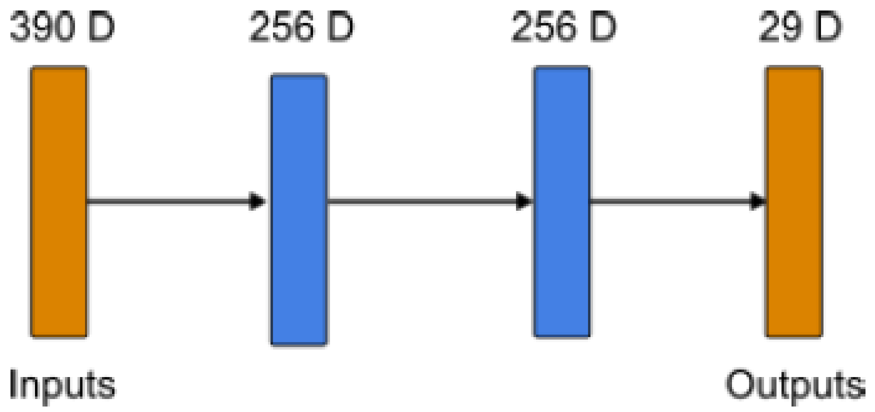


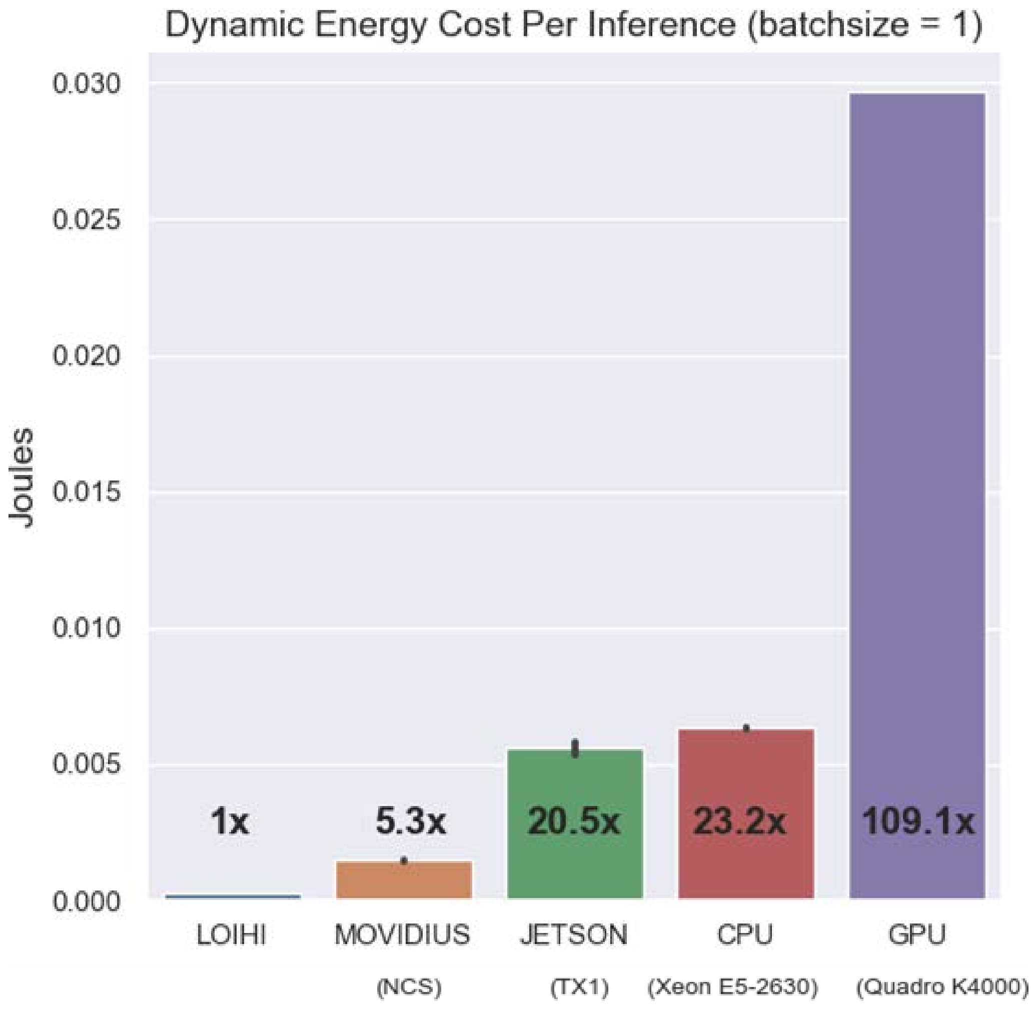
Some Specifics

- **Goal: compare Loihi's speed, efficiency to conventional HW for DNN inferences**
  - Aim is to provide rigorous, quantitative assessment of chip performance
- **Task: keyword spotting (small scale speech recognition)**
  - Use a two-layer feedforward neural network to recognize the phrase "aloha"
- **Metrics: inference speed, dynamic power consumption, energy cost / inference**
  - Measure these on different devices running same network with same data
- **Devices: CPU, GPU, Nvidia Jetson, Movidius NCS, Loihi Research Chip**
  - Note that comparison is between production and research devices

The Highlights

- **Same 2-layer architecture on all HW**
  - Trained in Nengo DL on Loihi
  - Trained in TensorFlow elsewhere (identical params across devices)





The Highlights

Model	True Positive (%)	False Negative (%)	True Negative (%)	False Positive (%)
TensorFlow	92.7	7.3	97.9	2.1
Nengo Loihi	93.8	6.2	97.9	2.1

- **Spiking DNN on Loihi maintains near-equivalent performance characteristics**
  - Training methodology applies to arbitrary DNNs, nothing specific to ASR
- **Dataset collected from 96 speakers using Amazon's Mechanical Turk platform**
  - ~2000 training utterances split 3:1 between positive and negative examples
  - 192 test utterances, 1 positive and 1 negative example per speaker

Speech Recognition Basics

Animation Credit: https://distill.pub/2017/ctc/

Speech Recognition Basics

ABR

1. Get windowed feature inputs (MFCCs) paired with target outputs (characters)

“Aloha”

2. Train rate model (with differentiable LIFs) to match feature inputs to target outputs.

TensorFlow

NengoDL

3. Save the network parameters, swap in spiking LIFs, and port onto Loihi (SNN only)

NengoLoihi

Benchmarking Methodology

1. Estimate idle power consumption on each device (averaged over 15 min)
1. Log power readings at fixed interval during runtime (usually every 200ms)
1. Estimate dynamic power by subtracting idle baseline from logged readings
1. Calculate average logging interval and number of inferences per interval
1. Calculate average energy cost per inference from (3) and (4)

(Energy profiling tools: s-tui, nvidia-smi, power meters for non-CPU/GPU devices)

Benchmarking Methodology

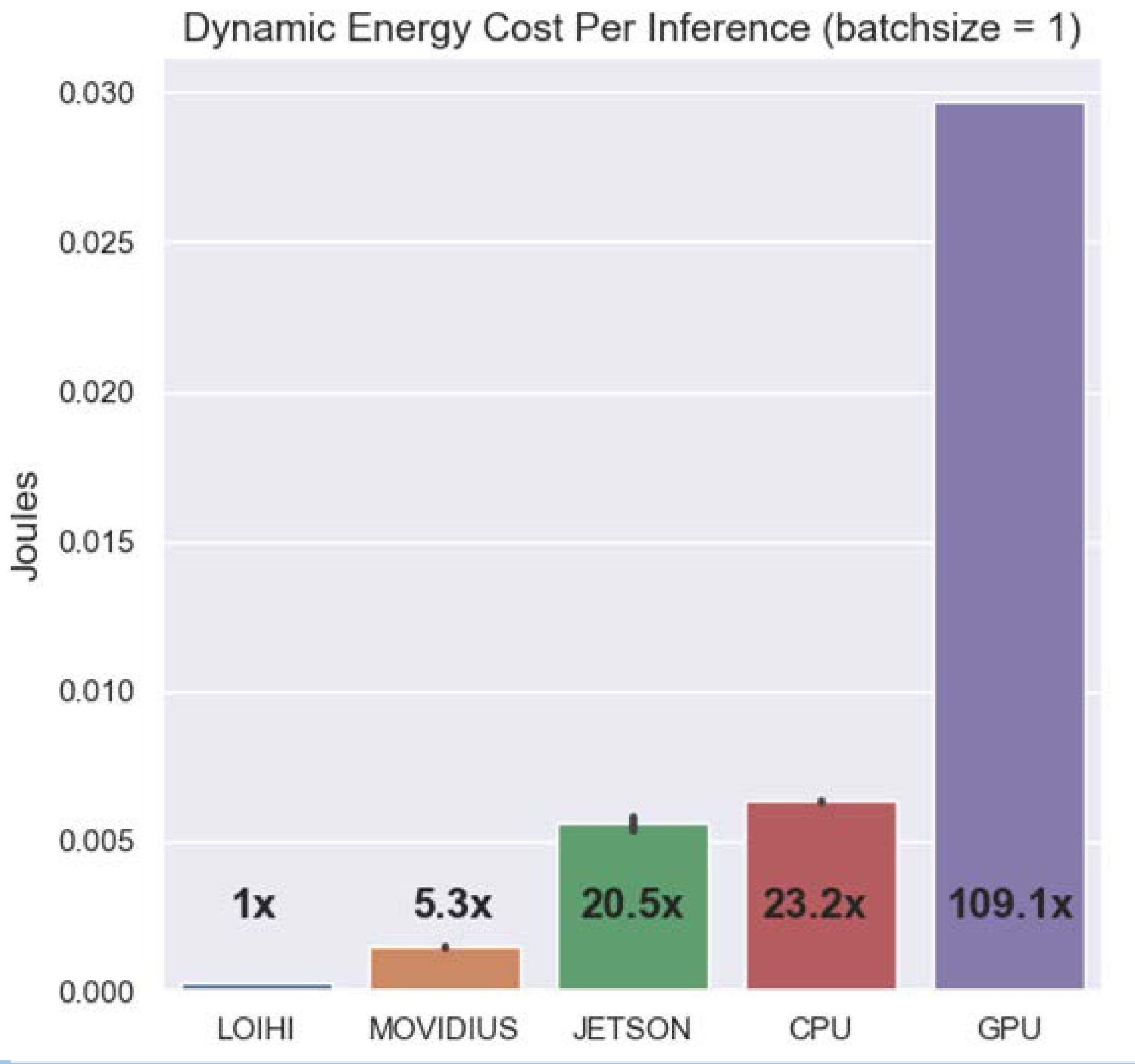
Hardware	Idle (W)	Running (W)	Dynamic (W)	Inf/Sec	Joules/Inf
GPU	14.97	37.83	22.86	770.39	0.0298
CPU	17.01	28.48	11.47	1813.63	0.0063
Jetson	2.64	4.98	2.34	419	0.0056
Movidius	0.210	0.647	0.437	300	0.0015
Loihi	0.029	0.110	0.081	296	0.00027

- **Measurements of power consumption and inference speed tell us everything**
  - Joules per second (W) / inferences per second = joules per inference

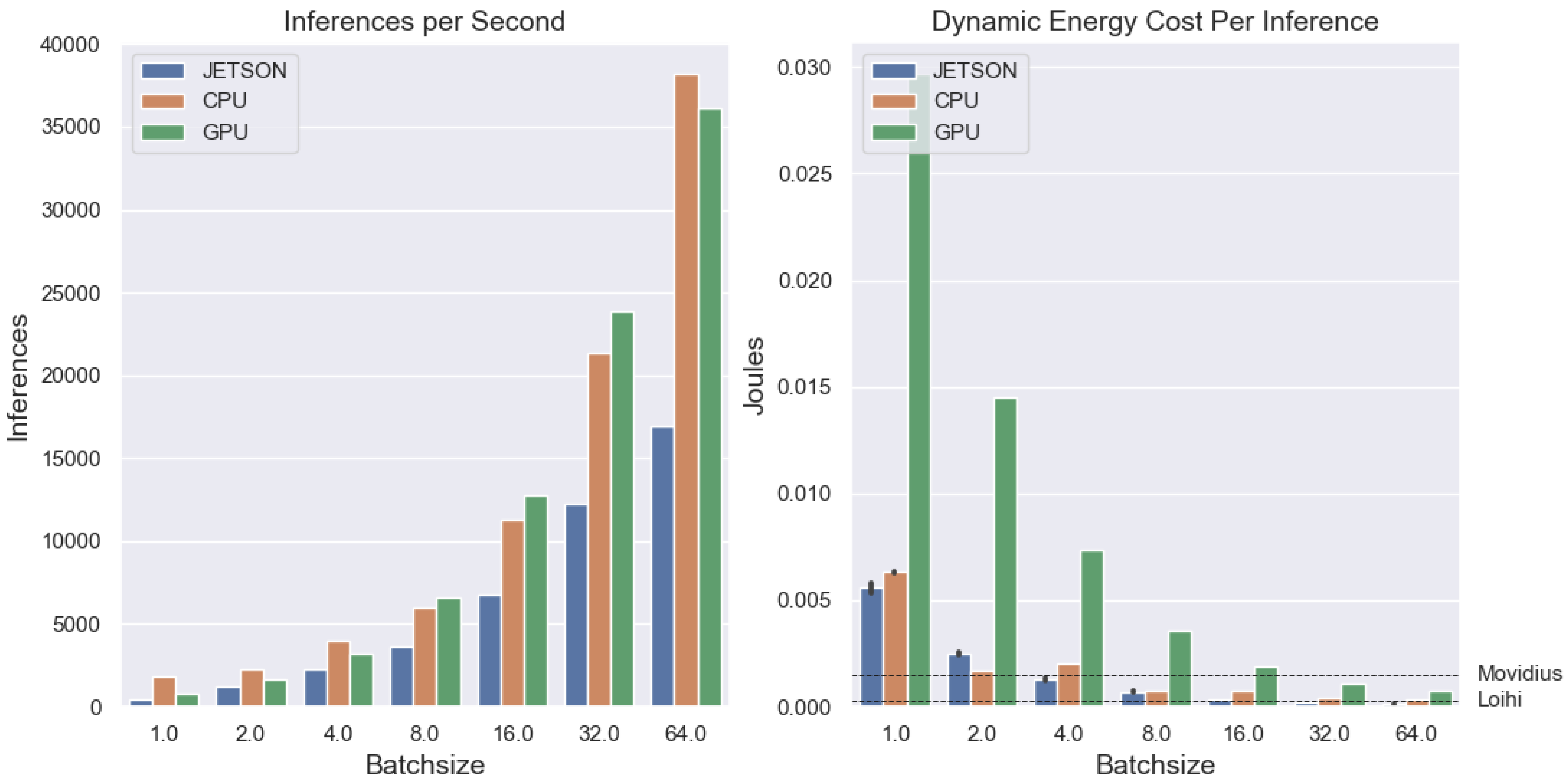
- **Methodology is highly conservative, designed to be generous to other devices**

(Note that batchsize=1)

Results for Online Inference

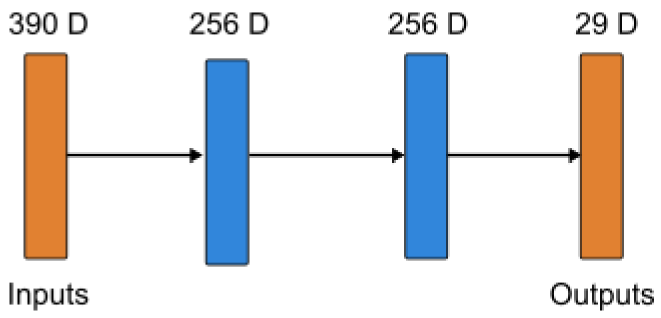


Results for the Effects of Batching

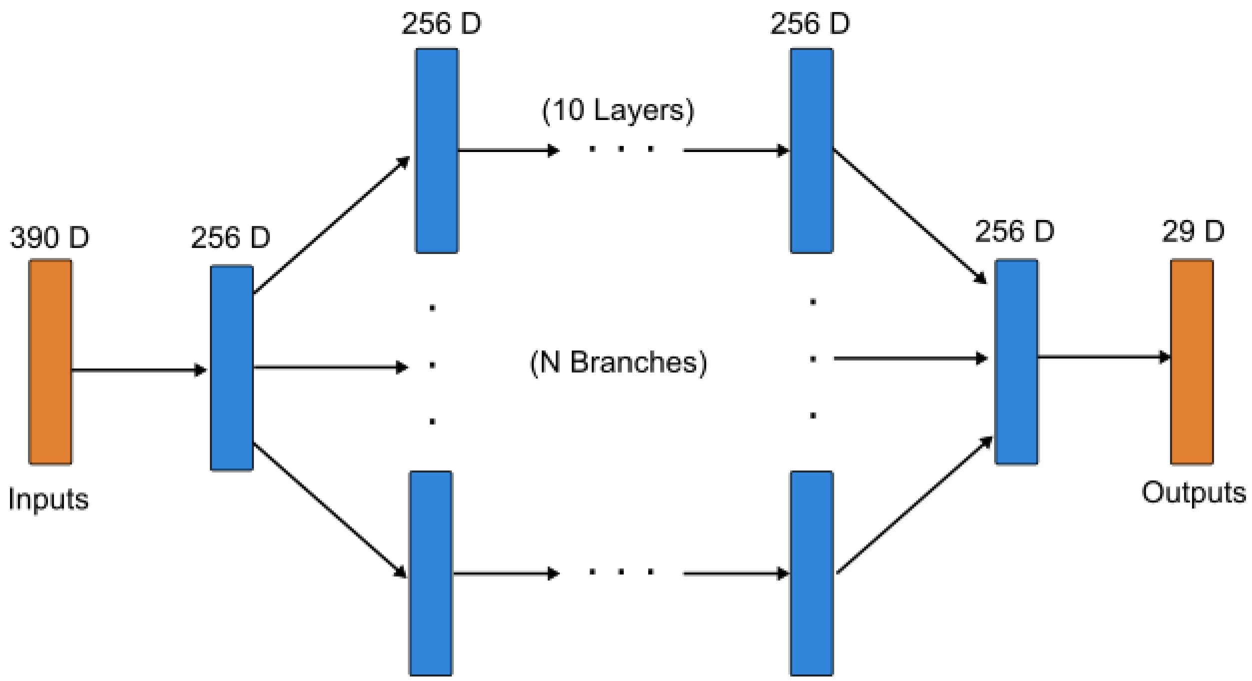


Scaling Network Size with Fixed I/O

Original Network

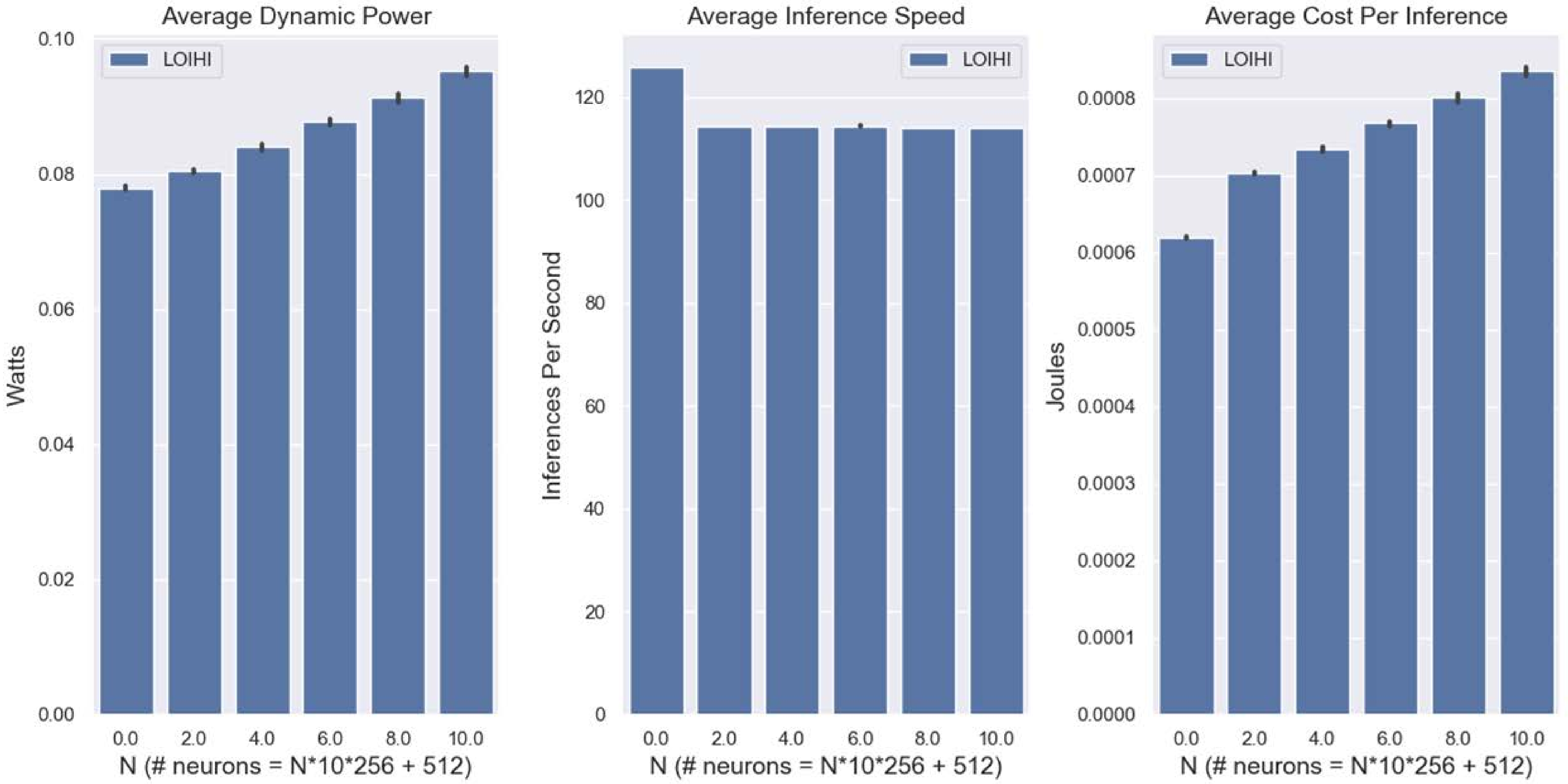


Scaled Network

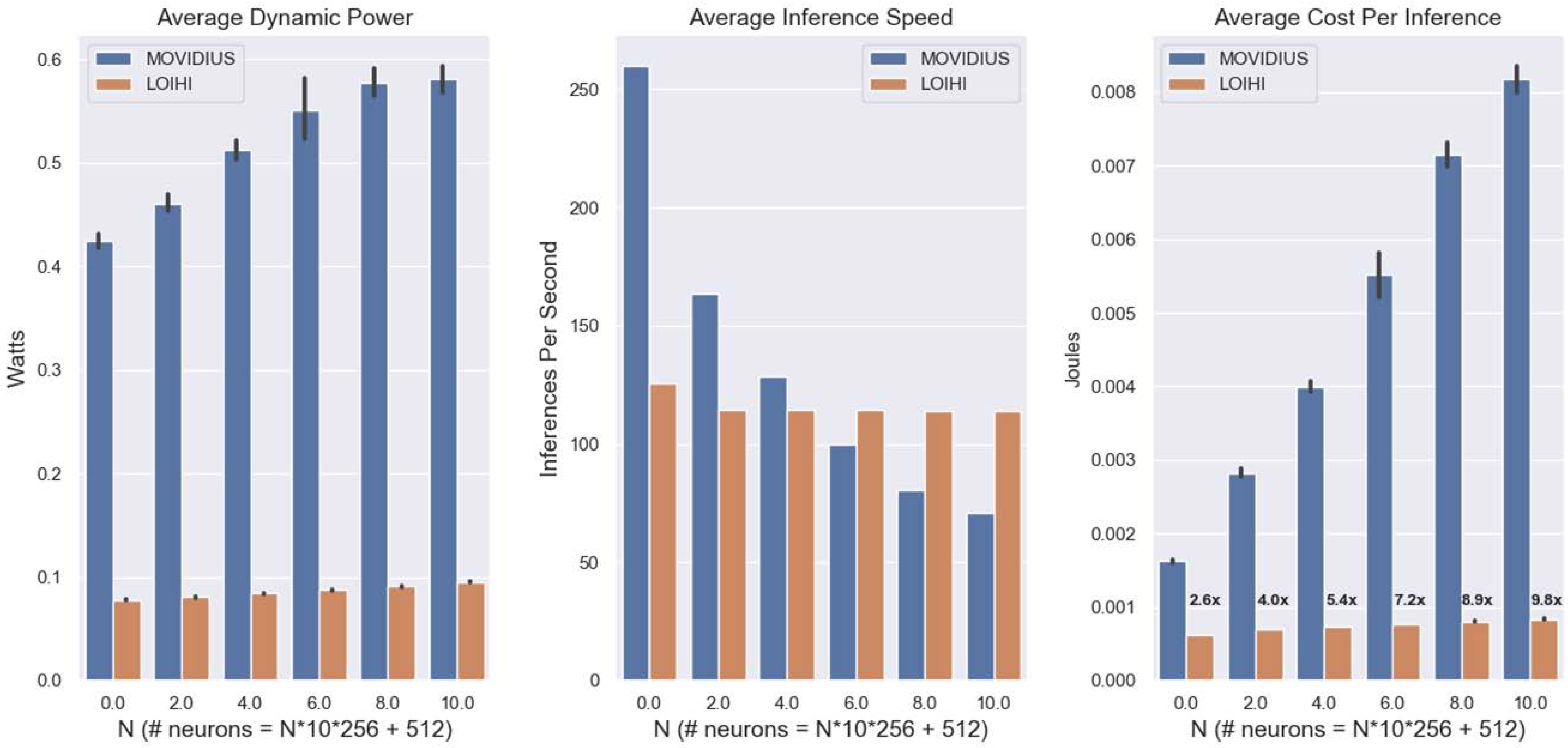


- Helps us isolate relative contributions of I/O and compute to energy consumption
- Scaled network is randomly parameterized with fixed weight distribution

Results for Scaled Networks - Loihi



Results for Scaled Networks



Discussion

- **Inference speed and dynamic power change very slowly w/ scaling on Loihi!**
  - This is architectural parallelism and event-driven computing at work.
- **I/O bottlenecks overall inference speed, so cost/inference can potentially shrink**
  - Doesn’t look like spike trafficking is a main source of energy consumption
- **Important to note that comparison is between research and production devices**
  - Lots of room for further analysis and benchmarking on other applications
- **Methods generalize to arbitrary deep networks, nothing application specific**
  - Plenty of reasons to be optimistic about Loihi as platform for low-power DL

Where to Learn More

NengoDL

https://www.nengo.ai/nengo-dl/

NengoLoihi

https://www.nengo.ai/nengo-loihi/

ABR Collaborators

Dr. Xuan Choo

Dr. Eric Hunsberger

Dr. Chris Eliasmith

APPLIED BRAIN RESEARCH

Thanks for listening!