Closed-loop Brain Stimulation in NHPs

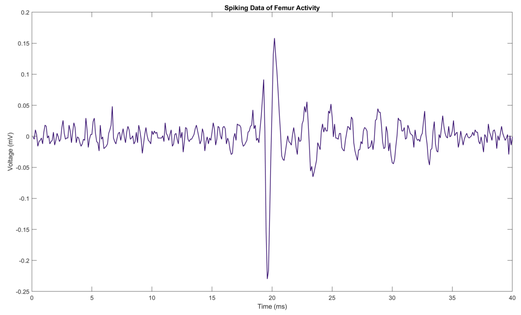
Using Hardware Accelerated Machine Learning

Rajeev B. Botadra[1], ChiJui Chen[5], Leo Scholl[1], Trung Le[1], Hao Fang[1], Ryan Canfield[2], Amy Orsborn[1], Eli Shlizerman[3], Shih-Chieh Hsu[4], Scott Hauck[1]

NSF-2117997

Stimulating the Brian

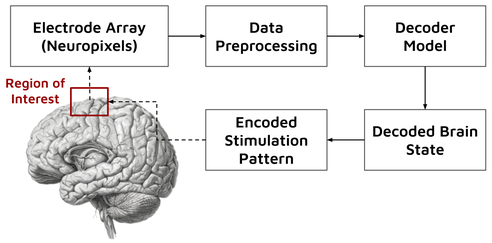
- Electric stimuli can incite/inhibit neuron activity
- Occurs over an order of milliseconds



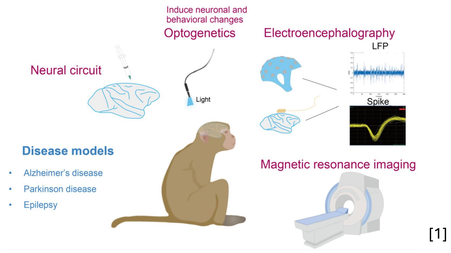
- Need a system that operates within latency constraint to effectively target brain regions

Motivations

- Creating closed-loop Brain-Computer Interfaces (BCIs) for Prosthetics



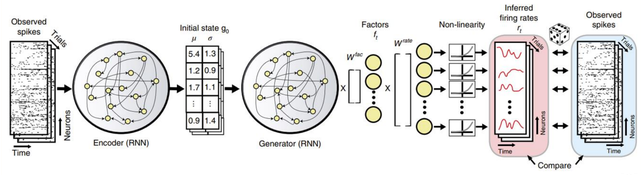
- Studying neuron circuits and their contributions to behavior for Neurotherapy
  - NHPs are ideal candidates of study for their physiological similarities to humans



[1]

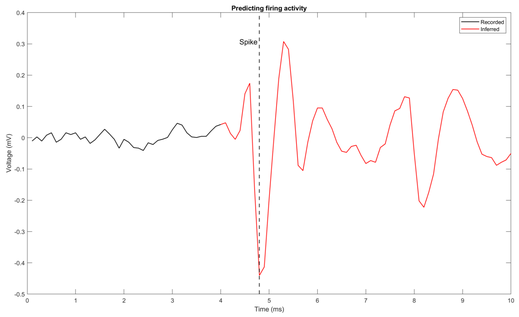
Decoding Neural Activity

- LFADS – a Variation Autoencoder – models firing activity in brain regions

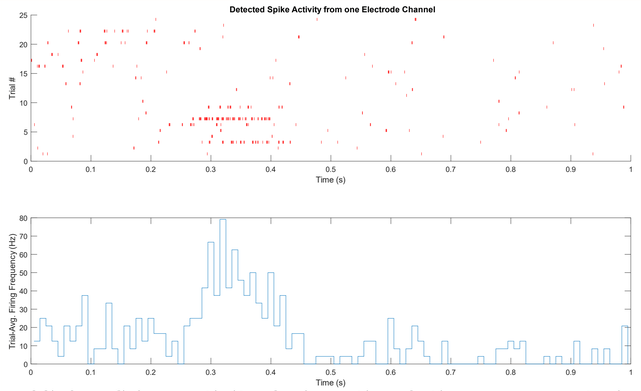


[2]

- Decoded brain states inform the stimulation pattern to inhibit neurons and observe the response



Latency Constraint & FPGAs



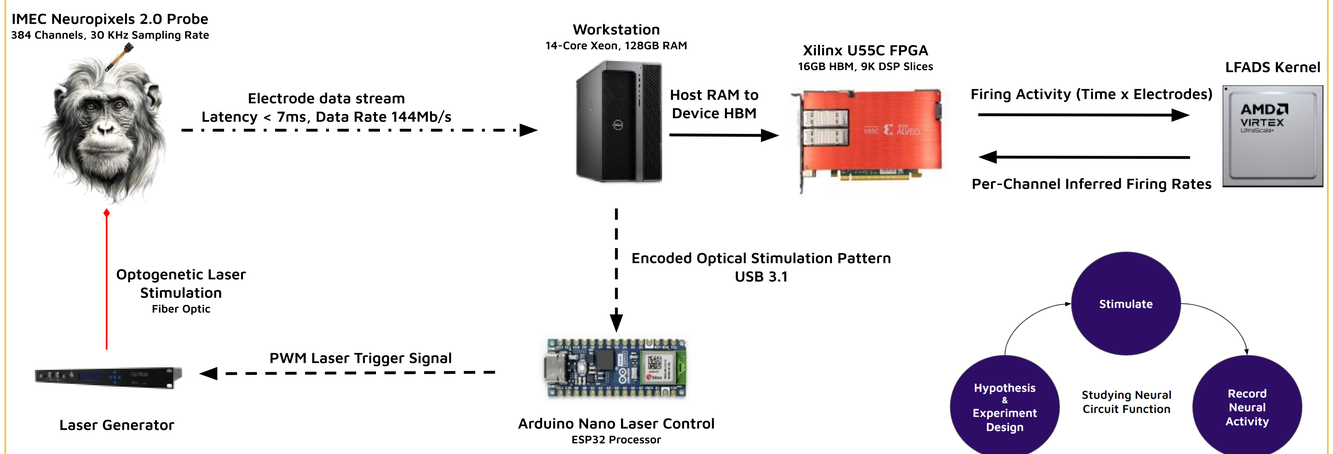
- Higher firing activity during stimulation can constrain inference latencies to <20ms (1/F)

	Model No.	Inference Latency (Batch Size of 1)
CPU	Intel Core i7	30.24 ms
GPU	Nvidia RTX A4000	28.80 ms
FPGA	Xilinx Alevo U55C	0.65 ms

- Modern CPUs and GPUs fail to meet this constraints on sequential architectures, so we look to FPGAs

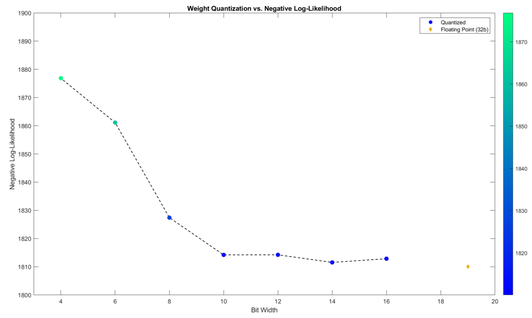
System Design

- The existing experimental apparatus was modified to integrate the FPGA and build an online closed-loop system



- Raw field potentials are preprocessed into spikes before decoding
- LFADS kernel deploys a fine-tuned and quantized architecture

Decoder Accuracy



- The pre-trained LFADS model weights were quantized with QAT

LFADS Performance: Hand-Reach Out Task	
Bit Width	Test Set R^2 Score
32 bit Floating Point	0.91303
8 bit	0.80195

- LFADS Kernel resource utilization on U55C

Resource Usage on U55C (8-bit LFADS)	
Resource	Utilization
LUTs	14.51%
Registers	9.28%
BRAM	14.53%
DSPs	24.27%

Future Works

- Test and deploy additional model architectures (NDT, MRAE, modified LFADS)
- Leverage closed-loop capability in experiment design
  - Investigating motion pre-planning through the hand-reach-out task

References

[1] Nan Qiao, et al. "Update on Nonhuman Primate Models of Brain Disease and Related Research Tools." MDPI Biomedicines, 2023.

[2] Xiaohan Liu, et al. "Sleep Spindles as a Driver of Low Latency, Low Power ML in HLS4ML & TinyML." University of Washington, 2023.

[1]Electrical & Computer Engineering, University of Washington
[2]Bioengineering, University of Washington
[3]Applied Mathematics, University of Washington
[4]Physics, University of Washington
[5]Electrical Engineering, National Yang Ming Chiao Tung University