FPGA Accelerated Brain-Computer Interfaces

A Co-design for Future Integrated Systems

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

[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

NSF-2117997

Introduction

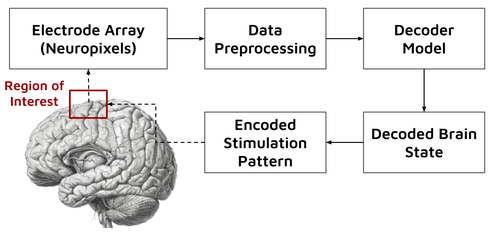
- Brain-Computer Interfaces (BCIs) have greatly advanced neuroscience over the past decade[1]
- For real-time studies, BCIs must operate within a latency constraint to respond to neural activity
  a) Action potentials occur on the order of milliseconds
  b) Data transmission and brain signal decoding impose large latencies
  c) Spatial and temporal resolution of neural recordings are rising with novel technologies, further increasing data processing latency

- **Aim:**

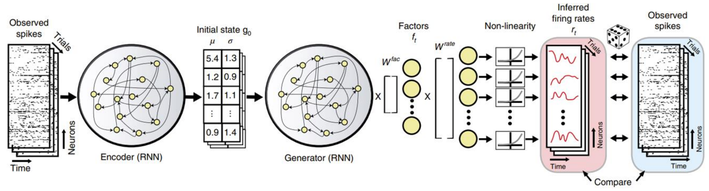
Reduce brain signal processing (i.e. decoding, state estimation, etc.) latency to within constrains using hardware acceleration with FPGAs

Brain State Estimation

- The decoder model is used to transform neural activity to a representation of brain state



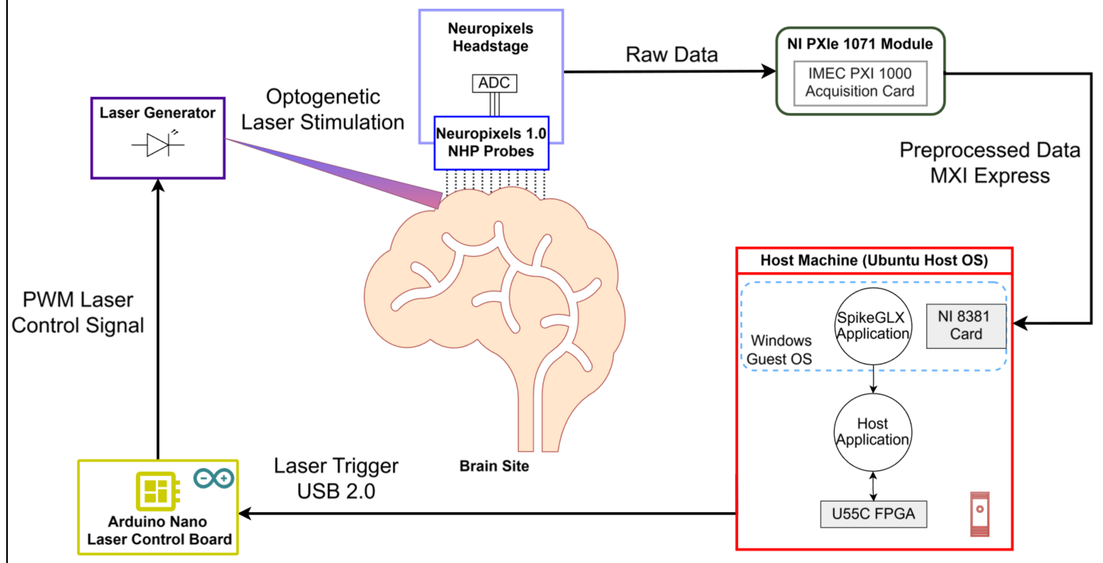
- We deploy LFADS to decode raw neural data and model neural dynamics



LFADS Architecture [3]

System Design

- Hardware accelerated Closed-Loop BCI:



- The system is functionally described by three subsystems:
  1. **Data Acquisition** with Neuropixels probes sampling from 384 channels at 30KHz
  2. **Accelerated Signal Decoding on the FPGA** using a quantized LFADS architecture to produce brain state estimations
     a) Each batch contains 20ms of spiking data across 384 channels to produce a set of latent factors representing brain state
  3. **Optogenetic Stimulation** using a laser targeting light-sensitive neurons local to the brain site under study

System Latency Characterization	
Component	**Latency (ms)**
Headstage & Module Preprocessing	2.512
Module to Host	5.047
Host to FPGA	0.018
Kernel Execution	0.156
FPGA to Host	0.008
Host to Laser Control Board	1.610
Total	**9.351**

The system achieves an action latency of 9.351ms, enabling experiments to respond to neural dynamics in real-time and identify the effects of system perturbations

Results

- FPGA acceleration enables sub-millisecond inference with LFADS, a speedup of 1650x, 118x versus a CPU, GPU

Architecture Comparison		
Architecture	**Inference Latency (ms)**	**Power (W)**
AMD Ryzen 16-Core CPU	257.516 ± 56.693	84.345
Nvidia RTX3090 GPU	18.551 ± 2.995	87.393
Xilinx U55C FPGA	0.156 ± 0.002	27.867

- Hardware resources are not fully utilized, can support additional compute still (i.e larger models, additional postprocessing)

Resource Usage on U55C (8-bit LFADS)	
Resource	**Utilization**
LUTs	13.90%
Flip Flops	6.15%
BRAM	6.03%
DSPs	18.44%

Future Work

- Our co-designed system serves as a silicon testbed for future real-time experiments
a) Extending compatibility to various recording modalities (e.g., EEG)
b) Exploring advanced neural decoder models (e.g., transformer architectures)
c) Investigating neuroplasticity[2]

References

[1] J. Robinson et al., "An application-based taxonomy for brain-computer interfaces," Nat. Biomed. Eng, 2024
[2] A. L. Orsborn et al., "Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control," Neuron, 2014.
[3] C. Pandarinath et al., "Inferring single-trial neural population dynamics using sequential auto-encoders," Nat. Methods, vol. 15, no. 10, pp.805–815, 2018.