Toward a Networked Implantable Neural Interface: Architecture, Real-Time Constraints, and Open Problems

Martim Gaspar Eric Wu Raghavendra Pradyumna Pothukuchi

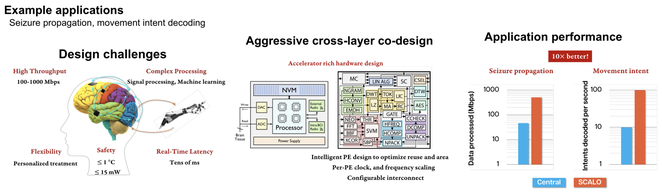
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

With contributions from Erik Jessen and a team of 30+ undergraduate researchers. Supported in part by NSF Major Research Instrumentation (MRI) Grant #2510152.

Prior Work and the Gap

In today's brain-computer interfaces (BCIs), A percutaneous cable connects the implant to a rack of computers that runs the decoder.

- **Tethered hardware restricts the patient.**
  Movement is constrained and coverage is limited to a few brain regions.
- **Decoding lives off-body.**
  Algorithms all run on external hardware. None of it is implantable.
- **No path to multi-region coverage.**
  Existing systems do not scale to several recording sites at once, which the diseases of interest demand.



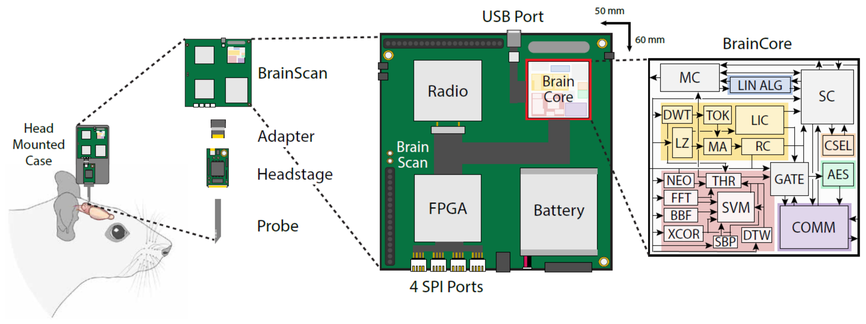
First Pipeline: Streaming Motor Decoder

- **End-to-end streaming Kalman decoder.**
  From raw 30 kHz neural samples to a hand position and velocity estimate every 50 ms.
- **Fully fixed-point datapath.**
  No floating-point on chip. All numerics in fixed-point format (e.g., 14 fractional bits in the high-pass, 40 in the Kalman accumulators).
- **One physical front end, any channel count.**
  Per-channel blocks are time-multiplexed. Scaling from a Utah array (96 channels) to Neuropixels (1536 channels) costs memory, not compute.
- **Auto-generated from MATLAB.**
  Hardware code (the register-transfer level, or RTL, description) is generated by HDL Coder from the MATLAB reference. Simulation, generated hardware, and silicon are bit-exact.

Vision: Untethered and Networked BCIs

- **BCIs today lack processing.**
  More can be done to enhance processing power and efficiency.
- **The brain is networked.**
  Parkinson's disease, epilepsy, and major depression are network diseases. Brain-wide activity drives behavior, function, and dysfunction.
- **Match the topology.**
  One chip per recording site, all running the same processing stack.

Hardware Prototype: First Step Toward the Network



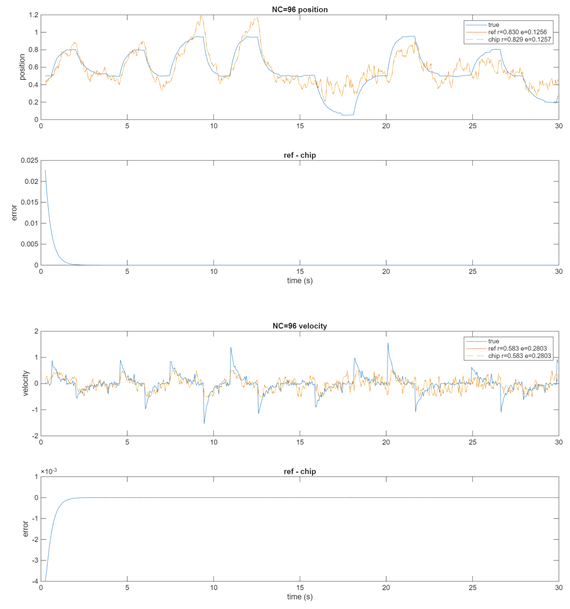
- **Single-node prototype, networked stack.**
  The same processing stack will run at every site once we go networked.
- **FPGA-prototyped on-chip processing.**
  The BrainCore digital pipeline runs on an FPGA today. Application-specific silicon is the target.

Open Questions

- **Spike features: how rich?**
  Threshold crossings, waveforms, on-chip sorting for speech and seizure pipelines?
- **Kalman engines: shared or per-pipeline?**
  One time-multiplexed engine per chip, or one per pipeline for better worst-case latency?
- **Network protocol: what crosses the wire?**
  Spike events, decoder state, or both? At what rate, with what timing guarantees?

Validation: Bit-Exact and on Target

We compare the simulated chip output against double-precision MATLAB reference, on synthetic spike trains generated from the Vaskov motor task.



- **Fixed-point matches the double-precision reference.**
  After about a 3-second startup transient, residuals fall below 10^{-3} in position and 10^{-4} in velocity.
- **Same correlation against ground truth.**
  Position $\rho = 0.830$ (reference) vs. 0.829 (chip). Velocity $\rho = 0.583$ for both.
- **One hardware design, two channel counts.**
  Validated unmodified at 96 channels (Utah array) and at 1536 channels (Neuropixels).

Status

- MATLAB and HDL simulation, $N_c \in \{96,\ 1536\}$: **done**.
- FPGA bring-up on Opal Kelly XEM8320-AU25P: **in progress**.
- End-to-end test with recorded neural data: **next**.

Roadmap

Today — Motor decoder bit-exact in MATLAB and hardware simulation. FPGA bring-up running on the XEM8320 board.

Summer 2026 — Recorded neural data on the FPGA, plus a second pipeline behind the same front end (seizure detection or LFP-band biomarkers).

December 2026 — Multi-pipeline ASIC tape-out.

Late 2029 — Networked, implantable, closed-loop: several chips, several brain regions, one hub deciding when and where to stimulate.