RETROSPECTIVE: Hardware-Software Co-Design for Brain-Computer Interfaces

Abhishek Bhattacharjee
abhishek@cs.yale.edu
Yale University

Rajit Manohar
rajit.manohar@yale.edu
Yale University

Karthik Sriram
karthik.sriram@yale.edu
Yale University

1 CONTEXT

In the summer of 2019, we began working on a multi-person/year project that has become a career ambition for us – the design of accelerator-rich computer architectures and chips for brain-computer interfaces. Several graduate students, post-doctoral scholars, and PIs with expertise in computer architecture, hardware accelerators, hardware-software co-design, high-level synthesis, and chip design came together to discuss this common goal.

A common point of conversation among us was our shared research interest in exploring the balance between the performance per watt benefits of specialization with the flexibility benefits of general-purpose execution. Another shared research interest was in neural decoding and how neural computation gives rise to cognition. In response to our interests, we decided to leverage our background in computer architecture and chip design to enable a new processing fabric to decode neural signals and facilitate breakthrough research in the neurosciences, as well as enable cutting-edge treatment of neurological and neuropsychiatric disorders.

2 THE MOTIVATION FOR HALO

The accelerator-rich HALO computer architecture and chip design for brain-computer interfaces, first published and introduced widely at ISCA 2020, was the fruit of our efforts in building a new computer architecture designed to interact with brain signals. Figure 1 illustrates a block diagram of HALO, its key components, and its computational functionality. The illustration builds atop Figure 2 from our ISCA 2020 paper, showcases HALO's hardware accelerators (referred to as Processing Elements or PEs in the ISCA 2020 paper), and illustrates inter-PE organization and communication.

HALO was conceived to balance several competing needs of cutting-edge neural decoding. The most sophisticated treatment options for neurological and neuropsychiatric disorders rely on extracting the highest-fidelity signal data from as many biological neurons as possible. Today, implantable devices embedded under the skull and placed on cortical tissue are necessary to extract the highest fidelity brain data. Implantation requires brain surgery.

While implanted brain-computer interfaces pose surgical risks, they are already used by thousands of individuals worldwide to treat disorders like epilepsy and Parkinson's disease. Clinical studies have also demonstrated that implanted brain-computer interfaces can help paralyzed individuals directly control prostheses or walk with brain signals, can help decode intended speech (and emit it to a computer terminal), can treat depression, anxiety, suicidal ideation, and much more.

By publishing HALO at ISCA 2020, the academic computer architecture community entered the rapidly growing discussion on the best ways to design processors for brain-computer interfaces. This discussion had, until our paper, remained in the realm of a robust industry and startup ecosystem (which raised $531M in 2021 alone). This ecosystem focused on a variety of issues including the design of sensors, wireless radios, and safer surgical/microsurgical techniques. Our goal in building HALO was to meet the needs of *all* implantable brain-computer interface companies by building a processing fabric that addressed the following considerations:

Low power: Temperature increases of even one degree Celsius can damage brain tissue. Consequently, brain implants are generally designed to operate under a few milliwatts of power. We target 15 mW implants. We estimate that the analog components on the brain implant take up 3 mW, leaving 12 mW for the HALO processor.

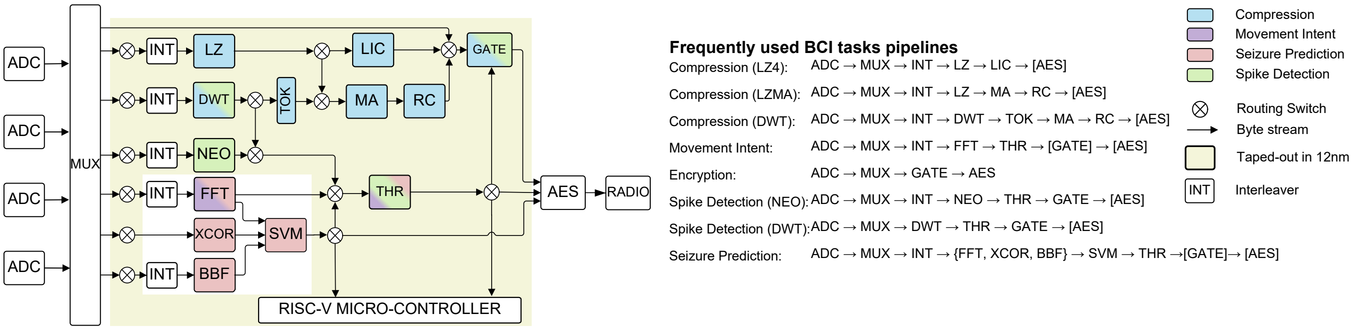
High-bandwidth neural decoding: There are close to a hundred billion neurons with over a trillion connections in the human brain. Quick calculations reveal that reading the activity of all these neurons at desirable sampling rates and fidelity amounts to interfacing rates in the Tbps range. Not all these neurons need to be read for useful brain-computer interactions, but prior to HALO, interfacing rates generally remained in the mere tens of Kbps. It is challenging to achieve these interfacing rates under 12 mW. While higher interfacing rates have occasionally been achieved prior to HALO, they have eschewed the other goals listed below.

Fast neural decoding: Further complicating brain implant design is that many neurological and neuropsychiatric treatments must meet tight real-time constraints. For example, effective treatment of epilepsy requires electrical stimulation of the brain within milliseconds of seizure advent. Meeting real-time constraints with a 12 mW power constraint at high interfacing rates is extremely challenging.

Flexible computational capabilities: A natural response to the power and computational goals outlined above is to specialize processing hardware to extract maximal performance per watt. Indeed, this has historically been the state-of-art in neuroengineering. Processors for brain-computer interfaces have either consisted of power-hungry micro-controllers and FPGAs, or extremely power-efficient but specialized ASICs that treat only one type of disorder in one specific way. Neither approach is ideal. HALO strikes a compromise between both these approaches.

ISCA '23, June 17–21, 2023, Orlando, FL, USA.

ACM ISBN 979-8-4007-0095-8/23/06.
https://doi.org/10.1145/xxx.xxxxx



Frequently used BCI tasks pipelines

Compression (LZ4):	ADC → MUX → INT → LZ → LIC → [AES]
Compression (LZMA):	ADC → MUX → INT → LZ → MA → RC → [AES]
Compression (DWT):	ADC → MUX → INT → DWT → TOK → MA → RC → [AES]
Movement Intent:	ADC → MUX → INT → FFT → THR → [GATE] → [AES]
Encryption:	ADC → MUX → GATE → AES
Spike Detection (NEO):	ADC → MUX → INT → NEO → THR → GATE → [AES]
Spike Detection (DWT):	ADC → MUX → DWT → THR → GATE → [AES]
Seizure Prediction:	ADC → MUX → INT → {FFT, XCOR, BBF} → SVM → THR →[GATE]→ [AES]

Figure 1: HALO consists of low-power hardware PEs and a RISC-V micro-controller. The PEs are configured into pipelines to realize tasks ranging from compression (in blue) to spike detection (in green). PEs taped-out in the latest 12nm technology node are shown within the grey background. Optional PEs (e.g., AES encryption) are shown in square brackets. PEs operating in parallel (e.g., FFT, XCOR, and BBF for seizure prediction) are shown in curly brackets.

3 THE HALO CONCEPT

HALO was the first standardized processing fabric for a wide range of brain-computer interfaces. In order to achieve flexibility, HALO was designed to support a widely-used set of algorithms for signal processing, off-the-shelf compression algorithms, and linear classifiers (specifically, a support vector machine). Standard low power design dictates that we realize one ASIC per application, an approach we refer to as a monolithic ASIC. But, our ISCA 2020 paper showed that monolithic ASICs exceed the permitted power budget, and do not achieve our desired flexibility in hardware design.

To address these problems, HALO realized both flexibility and low power operation. The ISCA 2020 paper showed how we systematically mapped the design space of brain-computer interface applications to identify several target capabilities. These include disease treatment, signal processing and filtering, and secure transmission of neuronal data (e.g., compression and encryption).

The ISCA 2020 paper then refactored the underlying algorithms into distinct components or kernels that realized different phases of the algorithm. The kernels facilitated the design of modular, ultra-low-power hardware PEs. Each PE is clocked at the lowest frequency required to sustain bandwidth while minimizing power, thereby permitting each component of an algorithm to be optimized independently. HALO was then rounded out by including a low-power RISC-V microcontroller to configure PEs into processing pipelines and support computation for which there are no PEs.

A key design principle of HALO was its top-down, modular design. This approach permitted agile design as well as easy integration of newly-developed PEs. We originally synthesized HALO in 28 nm, and have later synthesized and taped out several modules in 12 nm. Figure 2 shows pictures of the HALO tape-outs.

4 WHERE WE GO FROM HERE

We are in the process of testing our first generation of HALO chip tape-outs, which integrate most of the PEs described in the ISCA 2020 paper. In future tape-out runs (expected over the next few years), we envision adding the remaining PEs described in our ISCA 2020 paper, as well as new ones. Our silicon measurements will help us validate our pre-silicon physical synthesis power numbers. Our longer-term goal with HALO is to transition from bench-testing to animal testing, with ex-vivo experimentation with rats among our goals in the next few years. We are interacting with neurosurgeons at the Yale School of Medicine to determine a path and timeline to animal testing.

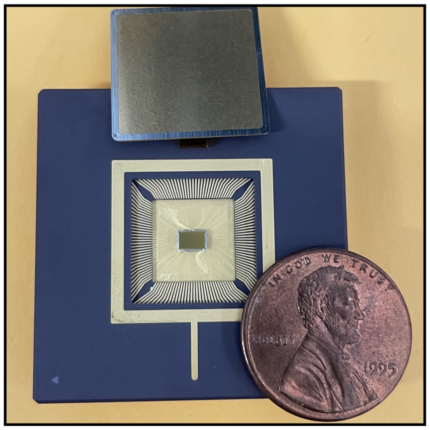


Figure 2: HALO chips from our first tape-out in 12 nm technology node. We have packaged and are in the process of testing these chips.

Our work on HALO has also served to open up more questions on building hardware accelerators. One way to view our HALO work is through the lens of asking, what is the set of canonical mini-accelerators (or in our terminology, PEs) that we need to build (and what is the communication fabric that they should integrate) for a future-proofed processing fabric? Because implantation is ideally a one-time scenario for individuals, given its surgical risks, it is vital that the processing fabric we build is useful for evolving future applications and workloads. An ideal choice of accelerators is one that would stand the test of time in accommodating emerging workloads. When integrating new PEs, we must deliberate on how adding these new PEs change the overall palate of computation that can be supported. Going forward, we envision addressing these questions, which lie at the heart of heterogeneous accelerator-rich computer architecture.