2026 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) | 979-8-3315-5103-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISPASS69572.2026.00068

CPU Bottlenecks in Rack-Stand Brain-Computer Interfaces: A Phase-Aware Characterisation

Victor Kariofillis*, Iris Uwizeyimana*, Sara Ahmad*, Tanvi Manku*, Raghavendra Pradyumna Pothukuchi†, Abhishek Bhattacharjee‡, Natalie Enright Jerger*
*University of Toronto †University of North Carolina at Chapel Hill ‡Yale University
{viktor.karyofyllis, iris.uwizeyimana, tanvi.manku}@mail.utoronto.ca, sarayyahmad@gmail.com, raghav@cs.unc.edu, abhishek.bhattacharjee@yale.edu, enright@ece.utoronto.ca

***Abstract*— Rack-stand brain-computer interfaces (BCIs) remain the dominant platform for exploratory and clinical BCI research, yet their CPU behaviour is poorly understood [1]–[7]. We characterise four rack-stand BCI pipelines—seizure detection [8], movement intent [9], speech decoding [10], and neural-data compression [11]—using peer-reviewed implementations and redistributable datasets on a commodity x86 CPU. Using offline replay of recorded neural streams, we combine end-to-end runtime across three core-frequency operating points with phase-aware Top-Down execution-time and memory-boundedness breakdowns [12]. The results show that rack-stand BCIs do not form a single bottleneck category, that channel count can change CPU demand by about 3×, and that a single "memory-bound" label is too coarse because different cache and memory levels dominate in different stages.**

I. Introduction

Brain-computer interfaces (BCIs) connect neural activity to computers and machines [1], [2]. Motivated by BCI successes and computational needs, computer architects have proposed processors and systems for implanted and wearable BCIs [3], [4]. Rack-stand BCIs, by contrast, are mobile external stations with acquisition and compute hardware, a display, and a power supply that interface with implanted electrodes [5]–[7]. They sit adjacent to the subject [5], [6].

Rack-stand BCIs are essential during algorithm development, clinical studies, and neurosurgical workflows [5]–[7]. They must remain portable, precluding large server-class hardware [5], [6], yet still process hundreds of Mb/s of neural data and respond within a few hundred milliseconds for clinician-in-the-loop use [5], [6]. Software is often developed in high-level environments with extensive visualisation and GUIs, increasing compute load. Heat dissipation and acoustics matter because the compute system operates near patients and staff in the operating room. For example, in some intraoperative microelectrode-recording workflows, neural activity is converted to audio during electrode placement [13].

We study four representative invasive rack-stand workloads built from peer-reviewed implementations and redistributable datasets: seizure detection [8], movement intent [9], speech decoding [10], and neural-data compression [11]. These workloads span different sensing targets and algorithmic structures. Some stages apply the same computation across signal channels or time windows, whereas others are dominated by spectral analysis, graph search, or compression. Yet end-to-end rack-stand BCI pipelines remain largely uncharacterised at the CPU microarchitectural level, limiting principled tuning and co-design.

II. Workloads and Methodology

We replay pre-recorded datasets on a 10-core Intel E5-2640v4 x86 CPU, pin each workload to a single hardware thread, and instrument phase boundaries in the released code. We collect runtime and instruction count and use Top-Down Performance Analysis to attribute execution slots and memory stalls [12]. Figure 1 summarises end-to-end runtime at three core-frequency operating points: 2.4, 1.8, and 1.2 GHz. In our experiments, lowering core frequency also lowers uncore frequency to the same point. Figure 2 summarises highlighted phases at the default operating point with two stacked bars per phase. The left stacked bar shows the execution-time breakdown. In Top-Down analysis, Backend Bound is split into Core Bound and Memory Bound. The right bar breaks down Memory Bound into L1, L2, L3, DRAM, and Store.

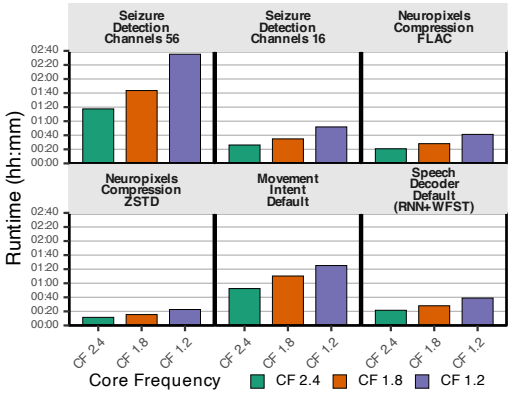

Fig. 1. End-to-end runtime for the six configurations at core frequencies of 2.4, 1.8, and 1.2 GHz.

Seizure Detection processes overlapping 0.5 s intracranial electroencephalography (iEEG) windows from implanted electrodes through spatial encoding, temporal encoding, classification, and postprocessing [8]. A channel is one recorded neural-signal stream, so Figure 1 compares the full 56-channel input with a reduced 16-channel subset. Figure 2 highlights *Spatial*, the dominant phase that converts per-channel voltage samples into channel-specific hypervectors [8]. *Movement Intent* partitions trial windows, estimates the fractal background via irregular-resampling auto-spectral analysis (IRASA), computes spectra, averages across trials, and subtracts the fractal

979-8-3315-5103-2/26/$31.00 ©2026 IEEE
DOI 10.1109/ISPASS69572.2026.00068

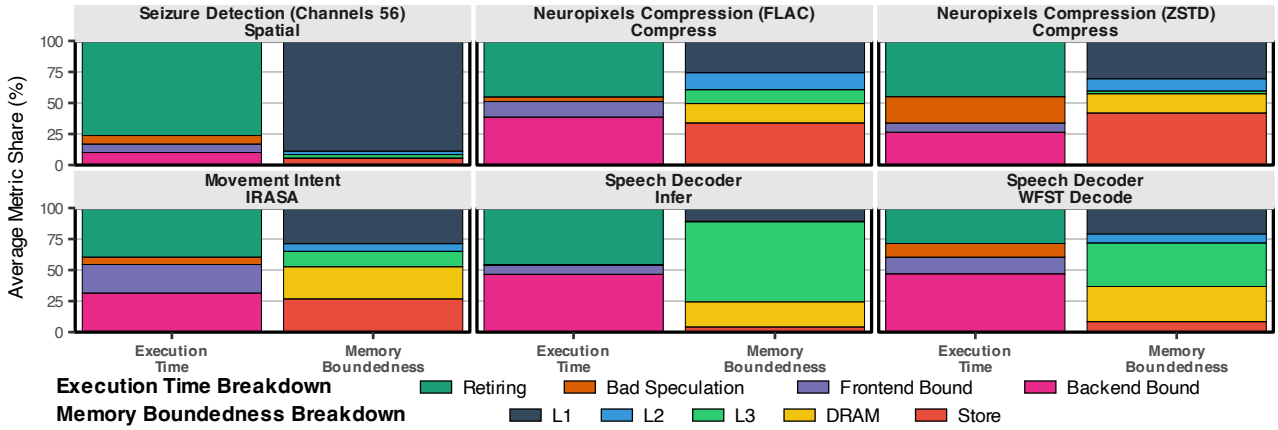

Fig. 2. Phase-aware Top-Down execution-time and memory-boundedness breakdowns for the highlighted workload phases.

component [9]. Figure 2 highlights *IRASA*, the dominant phase in that pipeline [9]. *Speech Decoder* smooths and normalises neural features, runs recurrent-neural-network (RNN) acoustic-model inference, and then performs weighted finite-state-transducer (WFST) beam search [10]. Figure 2 highlights *Infer* and *WFST Decode* to separate inference from graph search. *Neuropixels Compression* divides multichannel voltage traces into fixed-duration chunks and compresses them with either FLAC or ZSTD [11]. Figure 2 highlights *Compress* for both codecs [11].

III. Results

A. Runtime and Channel Count

Observation #1: Channel count is a first-class performance axis.

Figure 1 compares end-to-end runtime for six configurations across three operating points. Lowering the operating point increases runtime in every case, but the clearest difference is between the two Seizure Detection channel counts, not among the operating points. Seizure Detection shows a large gap between 56-channel and 16-channel execution at all three points. At the default operating point, reducing channels from 56 to 16 decreases instruction count from 30 trillion to 9 trillion and runtime from 1 hour 17 minutes to 27 minutes. Channel count is therefore a first-class performance axis for rack-stand BCIs. It materially changes CPU demand and must be explicit when workloads are compared or provisioned. Because many channel-parallel stages apply similar computations independently across channels, channel count is also a useful scaling dimension for future SIMD, multicore, and batching studies. Figure 1 also shows a stable runtime gap within compression, with ZSTD faster than FLAC at all three points.

B. Heterogeneous Bottlenecks Across Pipelines

Observation #2: Rack-stand BCIs do not form a single bottleneck category.

Figure 2 is the main architectural result. The two stacked bars separate execution-slot breakdown from memory-stall composition, and the highlighted phases do not look like one workload class. *Seizure Detection/Spatial* is compute-dominant: the execution-time breakdown is dominated by Retiring, and the Memory Bound portion is negligible. *Movement Intent/IRASA* is more mixed, with substantial Frontend Bound and Backend Bound slots and memory pressure spread across L1, DRAM, and Store. *Speech Decoder* already contains two CPU behaviours within one pipeline. *Infer* is backend-heavy with moderate memory stalls, whereas *WFST Decode* is the most memory-latency-heavy stage, with substantial L3 and DRAM pressure and non-trivial Frontend Bound and Bad Speculation. *Neuropixels Compression/Compress* shows moderate memory pressure with visible L1 and Store components, and FLAC and ZSTD do not share the same stall profile. Rack-stand BCIs therefore should not be treated as a single bottleneck category.

C. Memory-Bound Is Too Coarse

Observation #3: "Memory-bound" is too coarse.

The right-hand stacked bars of Figure 2 show why backend pressure must be qualified. Backend Bound is the sum of Core Bound and Memory Bound, and Memory Bound does not have one universal shape. *Movement Intent* spreads pressure across L1, DRAM, and Store, with little L2/L3 contribution. *WFST Decode* concentrates stalls in L3 and DRAM, exposing read-side miss latency. The compression phases show a different mix, with visible L1 and Store components and smaller L3/DRAM terms, while *Seizure Detection* stays near the floor. Even when phases are backend-pressured to some extent, the dominant cache and memory level changes across stages.

IV. Conclusion

Rack-stand BCIs remain CPU-centric systems that must support diverse pipelines under portability, thermal, and acoustic constraints. Across four representative workloads, the highlighted phases span compute-dominant, mixed, and distinct L1/Store versus L3/DRAM behaviour, while channel count changes CPU demand by about 3×. These results motivate phase-aware characterisation and make channel count explicit in future workload representation and provisioning.

REFERENCES

[1] R. A. Andersen, T. Aflalo, L. Bashford, D. Bjånes, and S. Kellis, "Exploring cognition with brain–machine interfaces," *Annual Review of Psychology*, vol. 73, pp. 131–158, 2022. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-030221-030214

[2] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 513–525, Sep 2016. [Online]. Available: https://doi.org/10.1038/nrneurol.2016.113

[3] I. Karageorgos, K. Sriram, J. Veselý, M. Wu, M. Powell, D. Borton, R. Manohar, and A. Bhattacharjee, "Hardware-software co-design for brain-computer interfaces," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, pp. 391–404.

[4] G. Eichler, L. Piccolboni, D. Giri, and L. P. Carloni, "Mastermind: Many-accelerator soc architecture for real-time brain-computer interfaces," in *2021 IEEE 39th International Conference on Computer Design (ICCD)*, 2021, pp. 101–108.

[5] Alpha Omega, "Neuro Omega," https://www.alphaomega-eng.com/Neuro-Omega-System.

[6] FHC, "Guideline 5," https://www.fh-co.com/product/guideline-5/.

[7] R. Zelmann, A. C. Paulk, I. Basu, A. Sarma, A. Yousefi, B. Crocker, E. Eskandar, Z. Williams, G. R. Cosgrove, D. S. Weisholtz, D. D. Dougherty, W. Truccolo, A. S. Widge, and S. S. Cash, "Closes: A platform for closed-loop intracranial stimulation in humans," *NeuroImage*, vol. 223, p. 117314, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920308004

[8] A. Burrello, L. Cavigelli, K. Schindler, L. Benini, and A. Rahimi, "Laelaps: An energy-efficient seizure detection algorithm from long-term human ieeg recordings without false alarms," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019, pp. 752–757.

[9] A. Stolk, L. Brinkman, M. J. Vansteensel, E. Aarnoutse, F. S. Leijten, C. H. Dijkerman, R. T. Knight, F. P. de Lange, and I. Toni, "Electrocorticographic dissociation of alpha and beta rhythmic activity in the human sensorimotor system," *eLife*, vol. 8, p. e48065, oct 2019. [Online]. Available: https://doi.org/10.7554/eLife.48065

[10] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann, K. V. Shenoy, and J. M. Henderson, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, pp. 1031–1036, 2023.

[11] A. P. Buccino, O. Winter, D. Bryant, D. Feng, K. Svoboda, and J. H. Siegle, "Compression strategies for large-scale electrophysiology data," *Journal of Neural Engineering*, vol. 20, no. 5, p. 056009, sep 2023. [Online]. Available: https://doi.org/10.1088/1741-2552/acf5a4

[12] A. Yasin, "A top-down method for performance analysis and counters architecture," in *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2014, pp. 35–44.

[13] T. Martin, G. Gilmore, C. Haegelen, P. Jannin, and J. S. H. Baxter, "Adapting the listening time for micro-electrode recordings in deep brain stimulation interventions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 8, pp. 1371–1379, 2021.