Heterogeneous Neural Amplifier Integration for Scalable Extracellular Microelectrodes

Jörg Scholvin, *Member, IEEE*, Justin P. Kinney, Jacob G. Bernstein, Caroline Moore-Kochlacs,
Nancy J. Kopell, Clifton G. Fonstad, *Fellow, IEEE*, Edward S. Boyden, *Member, IEEE*

Abstract— **We here demonstrate multi-chip heterogeneous integration of microfabricated extracellular recording electrodes with neural amplifiers, highlighting a path to scaling electrode channel counts without the need for more complex monolithic integration. We characterize the noise and impedance performance of the heterogeneously integrated neural recording electrodes, and analyze the design parameters that enable the low-voltage neural input signals to co-exist with the high-frequency and high-voltage digital outputs on the same silicon substrate. This heterogeneous integration approach can enable future scaling efforts for microfabricated neural probes, and provides a design path for modular, fast, and independent scaling innovations in recording electrodes and neural amplifiers.**

I. Introduction

For neural recording electrodes, the voltage signals that arise from extracellular activity in the brain are picked up by exposed metal recording sites, which are inserted into the brain. Insulated wires connect these recording sites out of the brain, and allow the signals to be amplified, digitized, and processed (Fig. 1). In this signal chain, the amplifiers need to be in close proximity to the recording sites, to reduce noise and cross-talk. Neural amplifier integrated circuits allow low-noise and low-power amplification, and often also multiplex several signals onto a single wire to reduce the total number of downstream wires required. Digitization may also be included on the same chip, but is not necessary and can take place at a greater distance away, because signals with higher signal strength (and potentially multiplexed, e.g. 10's of ±1V wires instead of 100's of ±1mV wires) are easier to protect from noise or cross-talk.

The neural amplifiers and the multiplexer in the signal path of Fig. 1 can be implemented through different approaches [1], and varying degrees of direct integration [2]. A *passive probe* uses complete separation of the recording electrodes ("probe") and the amplifiers ("headstage"). The probes are often attached to a printed circuit board (PCB) or flexible ribbon cable, and then connected to a separate PCB that contains the amplifiers (e.g. [3], [4]). Passive probes have the benefit of independent development and fabrication of recording devices and amplifiers, with a clear interface abstraction between them (e.g. the cable or connector between the two PCBs). But, this architecture comes at the expense of a physically larger system and the requirement to have as many large-scale interconnections (e.g. between PCBs) as there are neural recording sites.

In contrast, *active probes* monolithically integrate neural amplifiers and multiplexers with the recording electrodes [5], [6], reducing the number of external wires by the multiplexing ratio. The resulting devices are much more compact and avoid the external wiring constraints of passive probes. However, fabrication and design complexity is increased, and it is more costly to create many different designs with recording site placements specific to a scientific question. A different type of active probe integrates a switching matrix onto the shank itself (e.g. [7]), to select a fraction from the shank's sites to record from. Such probes may not contain amplifiers and multiplexers, but can obtain a lower external wiring count through the selection of a subset of recording sites. While only a fraction of the available data can be captured at one time, this solution is suitable to extend available lithography capabilities that otherwise may make it impractical to wire all sites along the shank in passive probes. As feature sizes shrink, this method may become less critical, but can be used to even further extend the total site count.

In this work, we analyze the merits of heterogeneous integration of neural probes with neural amplifiers, where similar functionality to an active probe is achieved through chip-to-chip packaging techniques of individual neural probe and amplifier chips. This approach can combine the benefits of passive probes (independent design and fabrication of probes and amplifiers) with the benefits of active probes (reducing the burden of external wires and connectors through amplification and multiplexing). A related approach was first demonstrated in [8] by placing a neural amplifier side-by-side with a neural probe, and wirebonding the neural inputs between the two chips together, avoiding the relatively large features of the PCB for the neural input wiring. However, the scalability of the approach in [8] is limited, because the connection between probe and amplifier can occur only at the periphery of two chips placed side by side.

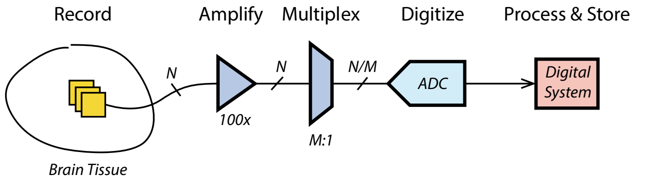


Figure 1. Signal path for extracellular recordings: signals in the brain (10's to 100's of µV) are sensed, conducted out, amplified, and usually multiplexed to reduce the wire count, and allow data transmission over longer distances; this makes better use of the available bandwidth per wire.

This work was supported by the Simons Center for the Social Brain at MIT, the Paul Allen Family Foundation, NIH, NSF, DARPA, IET Harvey Prize, Google, New York Stem Cell Foundation, NSF CBET 1053233.

J. Scholvin, J.P. Kinney, J.G. Bernstein, and E.S. Boyden are with the MIT Media Lab and McGovern Institute, Massachusetts Institute of Technology, Cambridge MA. C.G. Fonstad is with the Department of Electrical Engineering and Computer Science, MIT, C. Moore-Kochlacs and N. Kopell are with Boston University, Boston MA. Correspondence should be addressed to E.S. Boyden (e-mail: esb@media.mit.edu).

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

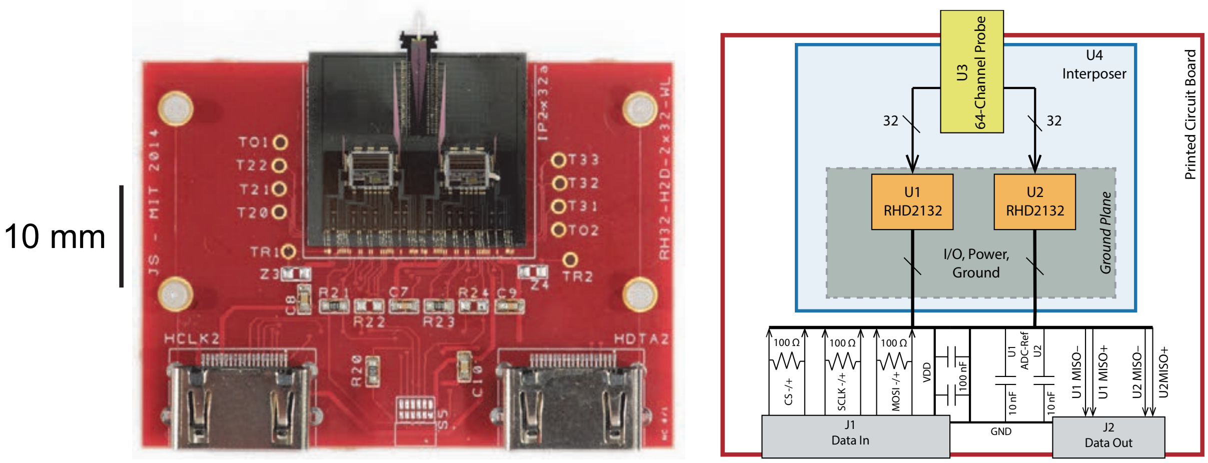


Figure 2. Photograph of a packaged device (left) prior to encapsulation, and simplified schematic (right). The interposer chip (U4) contains a wirebonded 64-channel neural recording probe (U3) we previously fabricated [9], as well as two 32-channel neural amplifiers (Intan Tech. RHD2132, U1 and U2). The data and clock HDMI connectors connect the device to a Willow data acquisition system (LeafLabs, Cambridge MA). The ground plane (darker area on the interposer die) provides isolation between output and input signals. For this investigation, we did not reduce the interposer nor PCB area – significant reductions in form factor and system size are possible. The ground plane is connected to by a large via below the neural amplifier chips (U1 and U2).

In contrast, the more generalized approach we are investigating here is to package neural amplifier chips directly on top of the probe chip. This enables a one- or two-dimensional packaging approach (wirebonding and flip-chip bonding respectively). But for such heterogeneous integration, the neural amplifier outputs need to be routed across the probe chip, raising concerns of cross-talk between the relatively high-voltage, high-frequency output wires and the low-voltage inputs. Adequate isolation between neural inputs and digital signals is therefore a key requirement. We demonstrate for the first time the heterogeneous chip-on-chip integration of neural probes with neural amplifiers. We show that a ground plane implemented on the probe can provide sufficient signal isolation. We quantify the physical separation needed between input and output signals, and characterize the impedance and noise performance of an assembled device.

II. Fabrication and Packaging

The test-chip design used in this work is shown in Fig. 2, with the chip layout shown in Fig. 3. We utilize an existing 64-channel close-packed neural probe previously fabricated (and described in [9]). For this study, we designed and fabricated a suitable silicon interposer chip that uses chip-on-chip packaging of both the neural probe and neural amplifiers. This interposer chip has both neural inputs and multiplexed digital outputs routed across the same substrate (see Fig. 3), with the neural inputs routed over a distance comparable to integrating the neural probe directly with the interposer (Table 1).

The interposer device in Fig. 3 was fabricated at MIT on 150 mm silicon wafers, with cross-sections and a process flow for the dual-metal process shown in Fig. 4. The lower metal layer was used as a ground plane and the upper metal layer was used for signal routing. The contact openings to the upper layer received a 5 µm electroless nickel plating with immersion gold finish (ENIG) to facilitate gold wirebonding. We then attached commercially available neural amplifiers in die form (Intan Technologies, RHD2132, [10]). The ground plane was placed below the neural amplifier dies as well as the signal output wiring, to provide isolation. We did not place a ground plane below the neural inputs, because the cross-talk capacitance between individual input wire pairs is sufficiently small even without a ground plane. After attaching and wirebonding the neural probe, the interposer was wirebonded to a PCB (see Fig. 2), where several passive components are included following the neural amplifier's data sheet. The outputs of the PCB are routed to two HDMI connectors, connecting to the data acquisition system [11].

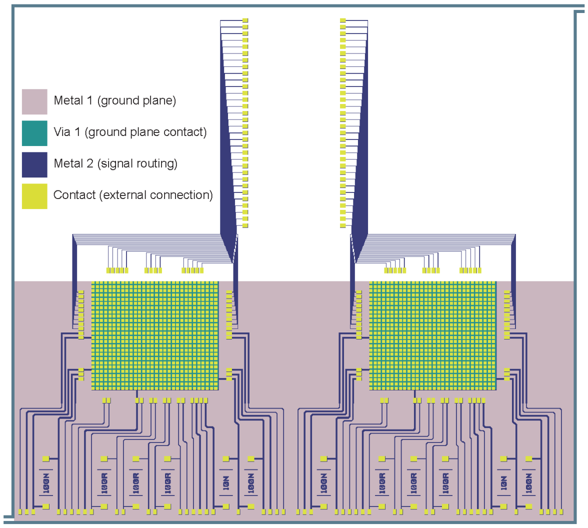
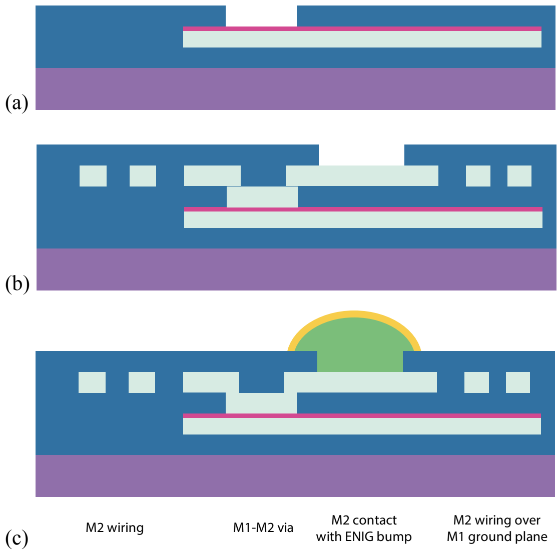


Figure 3. Wiring diagram for the interposer chip (*U4* in Fig. 2). The ground plane (Metal 1) is contacted through a large via (Via 1) located below the neural amplifier ICs (*U1*,*U2* in Fig. 2). Neural inputs are routed on Metal 2, without a ground plane necessary. In contrast, the digital I/O, power, and ground wiring for the neural amplifiers are routed above a Metal 1 ground plane. Wiring geometries are detailed in Table 1 below.

TABLE I
Summary of Wiring Geometries

Location	Signal Type	Wire Width	Wire Length
Interposer	Power, Ground	50 µm	4.1 – 8.4 mm
Interposer	Digital I/O	20 µm	4.0 – 5.8 mm
Interposer	Neural Inputs	6 µm	4.3 – 18 mm
Probe Body	Neural Inputs	2 – 5 µm	2.3 – 11 mm
Probe Shank	Neural Inputs	0.2 – 0.5 µm	3 – 3.5 mm



1. Start with standard Si wafer (150 mm diameter)
2. Deposit 1 µm of PECVD SiO_2
3. Sputter 0.5 µm of Al and 0.05 µm of TiN. Pattern with contact lithography (mask "M1"), and dry etch in Cl-based plasma etcher
4. Deposit 1 µm of PECVD SiO_2. Pattern with contact lithography (mask "Via1"), and dry etch in CF_4/CHF_3-based plasma etcher

-- cross section (a) --

5. Sputter 1 µm of Al. Pattern with contact lithography (mask "M2"), and dry etch in Cl-based plasma etcher
6. Deposit 1 µm of PECVD SiO_2. Pattern with contact lithography (using mask "Contact2"), and dry etch in CF_4/CHF_3-based plasma etcher

-- cross section (b) --

7. Clean, zincate pre-treat, and deposit 5 µm of electroless nickel
8. Die-saw
9. Process individual dies in immersion gold bath

-- cross section (c) –

🟦 PECVD SiO2	🟥 Titanium Nitride	🟨 Immersion Gold
🟪 Silicon	⬜ Aluminum	🟩 Electroless Nickel

Figure 4. Simplified process cross-sections at key steps of the fabrication, carried out on 150 mm wafers at MIT. Drawings are not to scale. The process consists of two metal layers (we chose Al for convenience) insulated by silicon dioxide. Fabrication involves patterning of the lower layer of metal (a), the upper layer (b), and a wafer-scale post-processing (c) to deposit a gold coated nickel bump suitable for wirebonding or flip-chip bonding.

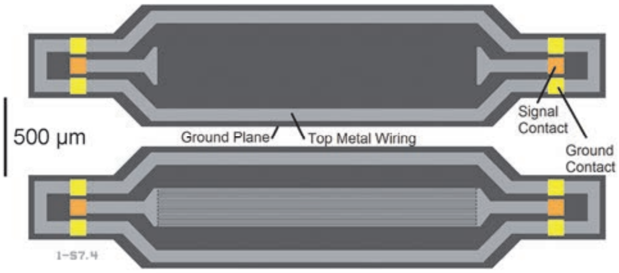


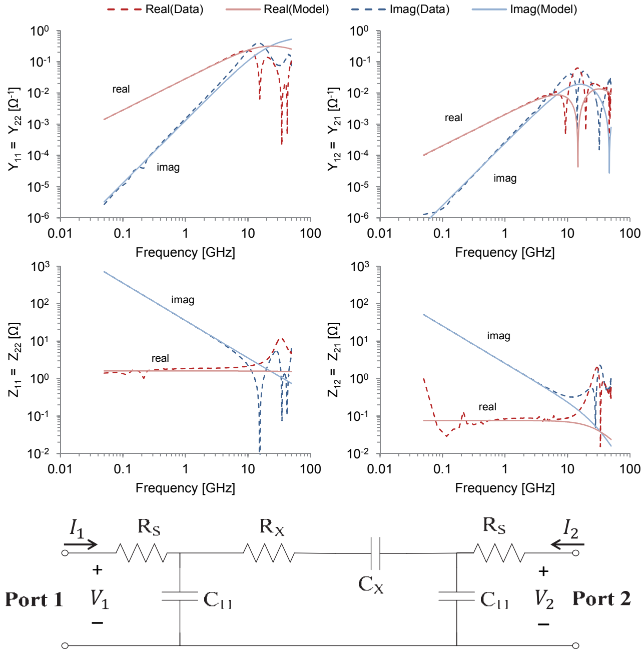
Figure 5. Layout drawing for an interdigitated finger capacitance test structure (*bottom*), also fabricated in the interposer process. An open-circuit de-embedding structure (*top*) is used to subtract the contribution of the test pads [14]. The test structures are designed to be measured using a two-port network analyzer, and vary in line spacing, width, and length.

We used a manual wirebonder (MEI 1204B) to electrically connect the PCB to the interposer, and the interposer to the neural probe and amplifier. When bonding to the interposer aluminum pads, it was easy to crater through the dielectric and cause short-circuits. We therefore post-processed the interposer chips with a 5 µm thick electroless nickel immersion gold (ENIG) finish, to avoid these problems. The entire interposer is then encapsulated in a dark epoxy to protect the amplifiers and interposer from liquids, mechanical damage, and light interference (unlike for the neural probes themselves [9], we did not use a highly doped Si substrate for the interposer, necessitating the choice of dark epoxy over clear one). The tip of the neural probe extends beyond the interposer's top edge (Fig. 2), and is not encapsulated.

We post all design files and images for this paper at http://scalablephysiology.org/probes/.

III. Results

Compared to passive probes, the interposer of Fig. 2 conducts both the neural input signals (e.g. 10 - 500 µV, 10 - 30,000 Hz), and the neural amplifier outputs (e.g. 350 mV, 40 MHz digital, for the RHD2132). The interposer wiring must be able to accommodate both signal types and sufficiently isolate the inputs from the outputs.



$$Z_{11} = \frac{V_1}{I_1}\bigg|_{I_2=0} \quad Z_{12} = \frac{V_1}{I_2}\bigg|_{I_1=0} \quad Y_{11} = \frac{I_1}{V_1}\bigg|_{V_2=0} \quad Y_{12} = \frac{I_1}{V_2}\bigg|_{V_1=0}$$

$$Z_{21} = \frac{V_2}{I_1}\bigg|_{I_2=0} \quad Z_{22} = \frac{V_2}{I_2}\bigg|_{I_1=0} \quad Y_{21} = \frac{I_2}{V_1}\bigg|_{V_2=0} \quad Y_{22} = \frac{I_2}{V_2}\bigg|_{V_1=0}$$

Figure 6. Example of measured and model-fitted two-port Y- and Z-parameters [13] for one of the interdigitated finger test-structures (see Fig. 5). The equivalent circuit (*bottom*) was used to create a model for the data (*top*), and the symmetry of the equivalent circuit implies that Z_{11}=Z_{22}, Z_{12}=Z_{21}, Y_{11}=Y_{22}, Y_{12}=Y_{21}. The model is valid into the GHz range, and allows extraction of coupling (C_X) and ground plane capacitances (C_{11}). R_X and R_S are also extracted, but are sufficiently small (0.1 to 100 Ω range) and thus negligible in the cross-talk estimates.

A. Cross-Talk Characterization and Modeling

To characterize the isolation, we designed a series of radio frequency (RF) test structures that allow us to measure the parasitic capacitances between adjacent metal wires as well as a ground plane. Fabricated on the same wafer as the interposer chips, these test structures contain interdigitated capacitors designed as two-port networks (Fig. 5). We used a vector network analyzer to obtain two-port S-parameter measurements for each test structure [12]. An equivalent circuit model for the structures is shown in Fig. 6. For this topology, converting the S-parameter into an impedance as well as admittance network (Z- and Y-parameters, [12], [13]) allows us to derive a set of relatively simple expressions that relate the model to the two-port Z- and Y-parameters (detailed e.g. in [13], and defined in Fig. 6):

$$C_X = \frac{Imag(Y_{21})}{-\omega} \qquad R_S = Real(Z_{11}) + Real(Z_{21})$$

$$C_{11} = \frac{Imag(Y_{11}) + Imag(Y_{21})}{\omega} \qquad R_X = -Real(Z_{21})\left(1 - \frac{Imag(Y_{11})}{Imag(Y_{21})}\right)^2$$

A typical model-to-data fit is shown in Fig. 6. The model adequately describes the electrical behavior well beyond 1 GHz. Based on the measurements, we characterized the capacitances to ground and between neighboring wires as a function of the wire spacing (Fig. 7). Finite element modeling of capacitance cross-sections matches the measured capacitances, and allows us to extend the results beyond the geometries tested. An upper bound for cross-talk between adjacent wires is the ratio of the cross-to-ground capacitances C_X/C_{11}. The actual cross-talk in a neural probe can be much lower if e.g. the recording site impedance or the neural amplifier's input capacitance provides an additional parallel shunt to ground. The ratio of C_X/C_{11} is independent of the wire length, and thus the shorter the wiring becomes, the more these additional shunt capacitances will dominate and reduce cross-talk significantly further than our upper bound estimate.

Between two neural inputs, a cross-talk of 1% is typically acceptable: even a strong neural signal (e.g. 500 µV) only exerts a weak influence on a neighboring wire at 1% cross-talk, with an interfering signal comparable in magnitude to the recording site noise level (e.g. <10µV). In contrast, the neural amplifier's output signal is much stronger (e.g. 350 mV for the RDH2132). Therefore, isolation from the output to input must be ≪1%, and based on the above example should be two to three orders of magnitude smaller. Fig. 7 shows that cross-talk can be adequately reduced to 10^{-4} when the spacing between two neighboring wires is 100 µm, and extrapolating to a value of 10^{-5} suggests a separation of around 400 µm (approximately twice the pitch between neighboring bondpads). Of course, a tighter spacing would be possible if not only a single ground plane was used, but also additional ground wires in-between the two signals, or an additional ground plane above the signals. However, we recognize that a single ground plane is sufficient to separate the inputs from the outputs. A 400 µm separation is easily achieved because a neural amplifier design will naturally place the two signal types on different sides of its chip periphery, often millimeters apart. To scale the number of recording channels, multiple neural amplifiers can then be used, and a routing scheme suggested in Fig. 8 can ensure that inputs and outputs are always well separated. Based on the results of Fig. 6, more aggressive multiplexing would be no problem, since the frequency capability of the interposers (GHz range) extends well beyond typical multiplexed output frequencies of neural amplifiers (100 MHz range).

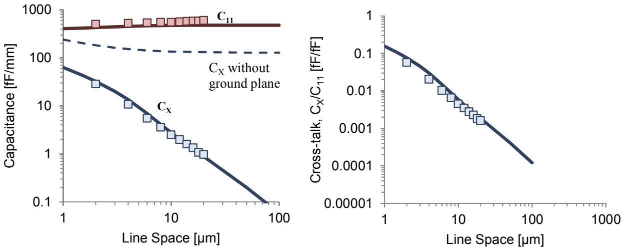


Figure 7. *Left:* Measured capacitance per wire length (squares) for C_{11} and C_X as a function of spacing between signal wires, for 10 µm wide wires. C_{11} is primarily a parallel plate capacitance to the ground plane, making it independent of wire spacing. C_X strongly depends on the spacing, because of the shielding effect of the ground plane. The lines show finite element simulations, and removing the ground plane would dramatically increase C_X as indicated. *Right:* The ratio of C_X/C_{11} shows the isolation between two neighboring wires, for both measured and simulated results. We extrapolate a value of 10^{-5} for C_x/C_{11} around 400 µm spacing.

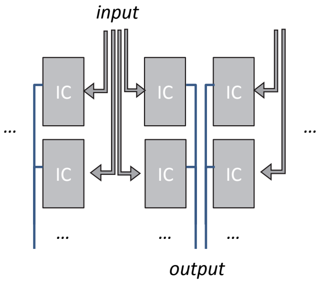


Figure 8. Suggested routing scheme for using a large number of neural amplifiers. Here, sufficient spatial separation of inputs and outputs is readily achieved through chip orientation, which can provide millimeters of signal separation.

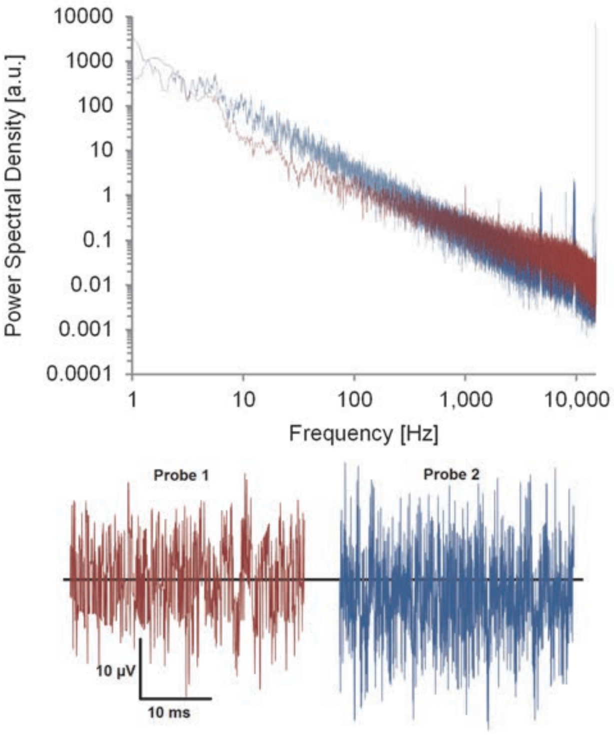


Figure 9. Noise spectrum *(top)* and time series snapshot *(bottom)* with the probe tip immersed into a grounded saline bath. Data shows two assembled interposer probes, acquired on different electrophysiology setups ("rigs").

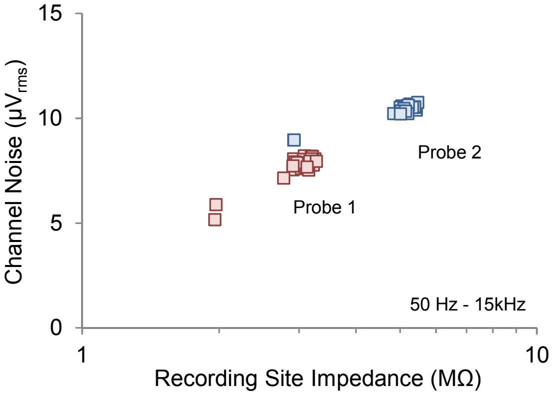


Figure 10. Comparison of the noise (50 Hz - 15 kHz) and recording site impedance for the two probes of Fig. 9. Recording sites damaged by wirebonding on probe 2 were removed from the data. On each probe, one pair of recording sites is short-circuited together by design, and shows up in the data with roughly a 50% lower impedance, as expected.

B. Noise and Impedance Measurements

We packaged two heterogeneous integrated probes and for each submerged its silicon shank in 0.9% saline. A ground reference wire was also placed into the saline, and connected to the PCB ground as well as the neural amplifier reference. Using a 30-second recording snapshot, sampled with 16-bits at 30 kHz using the Willow system (LeafLabs, Cambridge, MA), we calculated the root-mean-square noise voltage (V_{rms}) for each recording site for the frequency range of 50 Hz to 15 kHz. A sample trace and the noise spectrum for a typical recording site are shown in Fig. 9.

In addition, the recording site impedances were measured using the built-in functionality of the Intan neural amplifiers [10]. Fig. 10 shows the relationship between impedance and the V_{rms} noise. For the probes we used, the impedances were moderately high because we did not electroplate the recording sites to obtain lower values, but the V_{rms} at a given impedance is comparable to the levels we observed in [9].

C. Packaging Improvements

We used manual wirebonding, and found that overly aggressive ultrasonic bond-settings can damage both neural amplifiers and the neural probes. During manual wirebonding of the devices in this study, we inadvertently damaged one of the neural amplifier chips (on probe 2), and also cratered through several of the neural probe bondpads (which, unlike the interposer pads, did not receive an ENIG finish). These problems can be avoided, either by calibrating the manual wirebond settings or by using commercial wirebond services. Improved bondpad fabrication of the neural probes can also prevent damage (e.g. use of thicker metal layers or, similar to the interposer process, the use of ENIG plating). Any one of these steps can help prevent device damage during packaging.

IV. Conclusion

In this work, we demonstrated a design to route on the same silicon substrate both low-voltage, low-frequency neural recording inputs and the relatively higher voltage, high-frequency outputs of neural amplifiers. We showed that isolation of inputs from outputs is achieved through a ground plane and only 400 µm of separation is needed between groups of input and output wires; a constraint easily satisfied. We characterized such heterogeneously integrated devices, showing typical values for noise and impedances in saline.

Our results illustrate that it is possible to develop probes and amplifiers independently, yet combine them into a single system using chip-to-chip packaging. This insight can enable a viable solution for scaling up the number of recording sites for microfabricated neural probes, reducing the form factor while uncoupling probe and amplifier design efforts. With this approach, a broad range of neural probes can be fabricated using a single amplifier type, potentially reducing design and fabrication cost compared to fully integrated active probes.

Acknowledgements

We would like to thank Prof. J.A. del Alamo at MIT for the use of a Cascade Summit RF probe station and HP8510C network analyzer in his lab. The devices were fabricated at MIT's Microsystems Technology Laboratories (MTL).

References

[1] T. Jochum et al., "Integrated circuit amplifiers for multi-electrode intracortical recording.," *J. Neural Eng.*, vol. 6, p. 012001, 2009.

[2] P. Ruther and O. Paul, "New approaches for CMOS-based devices for large-scale neural recording," *Curr. Opin. Neurobiol.*, vol. 32, pp. 31–37, 2015.

[3] M. Chamanzar et al., "Ultracompact Optoflex Neural Probes for High- Resolution Electrophysiology and Optogenetic Stimulation," pp. 682–685, 2015.

[4] J. L. Shobe et al., "Brain activity mapping at multiple scales with silicon microprobes containing 1024 electrodes," *J. Neurophysiol.*, p. jn.00464.2015, 2015.

[5] K. Najafi et al., "A High-Yield IC-Compatible Multichannel Recording Array," *Electron Devices IEEE Trans.*, vol. 32, no. 7, pp. 1206–1211, 1985.

[6] C. M. Lopez et al., "An implantable 455-active-electrode 52-channel CMOS neural probe," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 248–261, 2014.

[7] K. Seidl et al., "CMOS-Based High-Density Silicon Microprobe Array for Electronic Depth Control in Neural Recording," *2009 IEEE 22nd Int. Conf. Micro Electro Mech. Syst.*, pp. 232–235, 2009.

[8] J. Du et al., "Multiplexed, high density electrophysiology with nanofabricated neural probes," *PLoS One*, vol. 6, no. 10, 2011.

[9] J. Scholvin et al., "Close-Packed Silicon Microelectrodes for Scalable Spatially Oversampled Neural Recording," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 120–130, 2016.

[10] R. R. Harrison, "RHD2000 Series Datasheet," 2012. [Online]. Available: http://intantech.com/files/Intan_RHD2000_series_datasheet.pdf.

[11] J. P. Kinney et al., "A direct-to-drive neural data acquisition system," *Front. Neural Circuits*, vol. 9, no. September, pp. 1–8, 2015.

[12] Agilent, "S-Parameter Techniques," *Appl. Note 95-1*.

[13] D. M. Pozar, *Microwave Engineering*, 2nd ed. 1998.

[14] Agilent, "De-embedding and Embedding S-Parameter Networks Using a Vector Network Analyzer," *Appl. Note 1364-1*, pp. 1–24.