Significance

Extracellular electrophysiology of millisecond-timescale neural activity is a gold standard method for observing neural population dynamics and exploring the neural code. It plays a fundamental role in research on a variety of basic neuroscientific topics such as development, learning, memory, and cognition, and in probing the mechanisms underlying diseases such as Alzheimer's, epilepsy, Parkinson's, schizophrenia, and depression. The advent of in vivo multielectrode recording has indicated the importance of recording from populations of neurons, in order to understand how neurons within a region coordinate their activity in sequences with complex temporal dynamics (1,2), how multiple sets of neurons coordinate their activity across regions to implement complex processes such as attention (3), how neurons multiplex information so as to encode for multiple kinds of variables at different times (4), and many other complex phenomena in neuroscience. Such technologies are also helpful for exploring the principles of how to create brain-machine interfaces (25).

Increasing the electrode count of neurological recordings allows more neurons to be recorded simultaneously. Not only does this open up new lines of scientific inquiry, but also reduces the number of trials required to achieve results for a given experiment as well as the time and effort required to identify (both online and during offline analysis) and record from experimentally salient units. If the electrode sites are densely packed enough to spatially oversample neurons, then novel, automated, and potentially far more accurate spike sorting techniques can be developed along with greatly increased capability for tracking individual units despite tissue-drift.

To record from larger populations of neurons, next-generation recording systems must scale the number of recording channels and gracefully manage the acquisition, storage, and processing of larger amounts of data, and yet retain a compactness required for freely-moving experiments and ease of use that matches the fast pace of competitive research labs. The magnitude of the challenge is reflected in the absence of commercially available 1000 channel multielectrode arrays for ***freely-moving*** neural recording (**Commercialization Plan, Table 1**). Recently, we validated a partial solution with close-packed silicon probes (5) and a direct-to-disk data acquisition architecture (6) to enable 1000 channel neural recording in head-fixed animals (**Preliminary Data**). 1000+ channel *head-fixed* extracellular recordings have been performed in few academic labs using custom-built systems (7,8,9, **Figure 2**). These systems were composed of hand-made probes, custom headstages, home-grown acquisition systems, and optimized computer setups (hardware and software). Each of these systems were developed independently by the labs that used them. Not all labs possess the engineering skills and resources to devote to these significant technology development tasks. Thus, 1000 channel systems are unavailable to all except current Willow users and a few other labs, and 1000 channel probes for freely-moving neural recording are nonexistent.

Furthermore, pilot studies and customer development interviews at more than 20 Boston-area academic neuroscience labs, as well as over conversations with over 100 visitors to our booths at SFN 2015 and 2016, indicate several remaining barriers to the ease and efficacy of high-channel-count electrophysiology in general, preventing its widespread adoption. High-channel-count recordings produce datasets of unprecedented size, and researchers are unprepared to deal with the logistical challenges of storing, transferring, and analyzing them. The difficulty of managing these datasets bottlenecks open access to high channel count neural data and thus the development of sophisticated analytical tools (such as automated spike sorting) for utilizing and learning from them.

Innovation

We seek to make performing and analyzing freely-moving rodent experiments easy and affordable for all researchers. To accomplish this, we will marry two existing technologies to build an active 1000-channel silicon probe for freely-moving neural recording and stimulation, in combination with a data acquisition system that implements hardware acceleration for easy data analysis online and offline.

New probes are made possible by NeuroProbe, our novel 1024-channel neural recording and stimulation chip (**Preliminary Results**) which will reduce size, weight, and cost each by a factor of 10 compared to commercially available headstages (**Commercialization Plan, Table 1**). This device integrates the functionality of what used to require a dozen or more commercial chips into a single 6 mm x 6 mm silicon chip. Furthermore, the chip allows not just recording but also stimulation on all channels, in addition to other built-in features useful for electroplating and testing such as voltage and current source and sink, not found in commercially available devices.

The market for multichannel electrophysiology probes is crowded. But the field is open for whomever brings to market high-channel count, freely-moving active probes. One notable potential entrant is the Neuropixels project, a collaboration between HHMI, UCL, Allen Institute for Brain Science, and engineers at the European consortium IMEC, a nanoelectronics research center in Belgium, that recently announced an intention to bring to market in Q1 2018 an active probe for freely-moving experiments. Of the 960 electrodes on the probe, only up to 384 can be recorded simultaneously, whereas the NeuroProbe can record from all of 1024 channels simultaneously. Also, the electrodes are arranged in columns with a 40 um pitch such that adjacent columns are offset so the sites form a checkerboard pattern. In initial experiments the observation was made that drift of the probe relative to the brain can cause neurons (perhaps a cell type with a compact electric field) to drift into a region between recording sites and be temporarily lost. To address this type of problem, we have achieved a 11.5um pitch with high-speed electron beam lithography (EBL). Given the constraints on allowable processes, equipment, and facilities demanded by a commercializable cost structure, contact mask lithography enables us to create feature sizes of only 2um, whereas our EBL feature sizes are 200nm. This gives us an order of magnitude increase in the number of wires for a given shank width, whilst remaining cost effective. Increased site density may also be crucial for spatial oversampling of neuron activity to allow automated spike sorting. Finally, in this early stage of development, data is acquired from the IMEC probe using off-the-shelf FPGA boards and open source software, whereas the torrents of data that come from high channel count recordings demand high-throughput commercial-ready acquisition system like the proposed second generation tool.

From our pilot studies in Phase I, we learned that success in electrophysiology experiments often depends on the experimentalist receiving fairly sophisticated feedback from the data collected. For that reason, we will improve not only offline analysis tools, but also improve the state of real-time experimental feedback. Any data acquisition system hoping to provide easy data analysis for high-channel count neural recordings must confront the fact that these recordings generate far more data than can be processed by an average workstation. At a single recording station, if 1000 channels are sampled at 30 kHz at 2 byte/sample, then the data flow from 1 recording station is 60 MB/sec, which equates to 1 TB of neural data every 5 hours. This raw data rate exceeds the typical capacity of shared gigabit network connections to files servers and compute clusters. Thus, during a high-channel count recording, local storage and analysis of data is favored.

To locally store and analyze data from high-channel count recordings, we will assemble and program a custom computer that utilizes a high-speed bus (PCIe) to connect components for data acquisition (FPGA), storage (SSD), memory (RAM), computing (GPU, CPU), and networking (NIC) (10-12). The design leverages considerable technology advancements in data acquisition and processing performed in Phase I of this grant (NIMH Award Number R43MH101943, 2014) and another Phase I grant (NIMH Award Number R43MH109332, 2016) to build an ultra high throughput lightfield microscope, and will allow hardware acceleration of both online analysis of live neural data and rapid offline analysis of static data stored on drives. This design allows full-bandwidth data to be copied from the FPGA to RAM to SSD for safe storage. From RAM the data is accessible to GPUs for massively-parallel computing and to CPUs for performing arbitrary computation. Processed results can be displayed to the user for data visualization, used for closed-loop experimental feedback, saved to SSD, and shared over network. After a recording is complete and before the start of the next recording, raw and processed data can be copied over ethernet network to a compute cluster for exhaustive analysis and archiving or remain in place.

Preliminary Data

Phase I of Willow development focused on high-channel count data acquisition and storage. Willow combines innovations such as close-packed silicon multi-electrodes, a highly scalable data acquisition system, and an extensible graphical user interface (**Figure 1**). With Willow, experimenters can perform single unit electrophysiological recordings at the unprecedented scale of 1020 electrodes per implant in awake head-fixed animals. Our modular, direct-to-drive architecture stores data local to the acquisition hardware. This means that a single Willow data-node can capture over 1000 channels of electrophysiological inputs, and multiple data-nodes can be synchronized for even further scaling. Our close-packed electrode arrays can sample across cortical layers and across multiple cortical regions simultaneously.

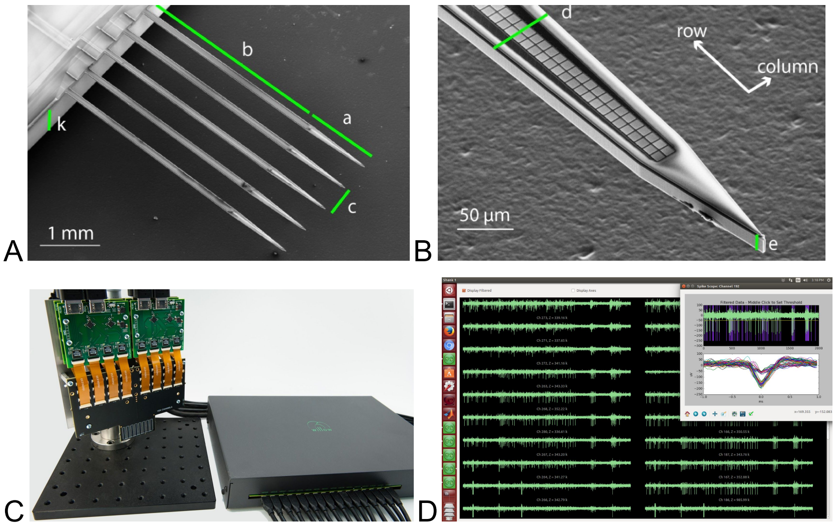

Figure 1. **Willow neural recording system** (A) SEM images of 1020 channel probe showing layout of recording sites. Lengths [a,b,c] are [3,5,0.5] in mm. (B) Each shank contains 204 sites measuring 10 um on a side and separated by 1 µm. Lengths [d,e] are [60,15] in µm.(C) A complete Willow 1000-channel system. (D) WillowGUI visualization software showing channel-mapped plots of channel traces from the 1020-channel recording. This 2x10 subarray of channels shows redundant oversampling of spiking activity. (Inset) A “spike scope” view of one channel from the 1020-channel recording. Top panel shows 2 seconds of band-pass filtered data, with the purple lines indicating detected spikes, whose waveforms are plotted in the bottom panel.

Recently, to validate the combination of close-packed silicon probes (5) and a direct-to-disk data acquisition architecture (6) as a partial solution to scalable neural recording, and to fulfill Specific Aim 1 of our Phase I grant, we performed a 1020 channel neural recording in visual cortex of awake head-fixed mouse (**Figure 2**). In March 2016 at the Boyden lab at MIT, a mouse was prepared for a recording using one of these probes. The specimen (C57BL/6, 8–12 weeks old, male, Jax) was placed under general anesthesia, using 0.5–2% isoflurane in pure oxygen. Five 300 micron diameter craniotomies were opened in the skull, one hole for each of the 5 shanks on the 1020-channel probe. After head-fixing the rodent, the probe was inserted acutely into the visual cortex using computer-controlled stereotaxy. All procedures were in accordance with the National Institute for Laboratory Animal Research Guide for the Care and Use of Laboratory Animals and approved by the MIT Committee on Animal Care.

Once the probe was implanted, a direct-to-disk recording session was initiated using Willow. While still under anesthesia, the specimen was then shown a series of drifting sinusoidal gratings on a display. After one hour of anesthetized recording, the mouse was brought into a waking state by reducing the concentration of isoflurane. Visual stimulation resumed, using grating patterns as well as natural scenes, while the Willow system continued recording 1020 channels from the visual cortex. During the course of the experiment, data was streamed to a workstation computer for real-time visualization, in addition to “snapshots” taken for a more detailed analysis of the data.

After 2 hours and 17 minutes of recording, the total amount of data acquired was nearly one terabyte on disk, and the quality and richness of the data is remarkable - we are still in the process of analyzing this

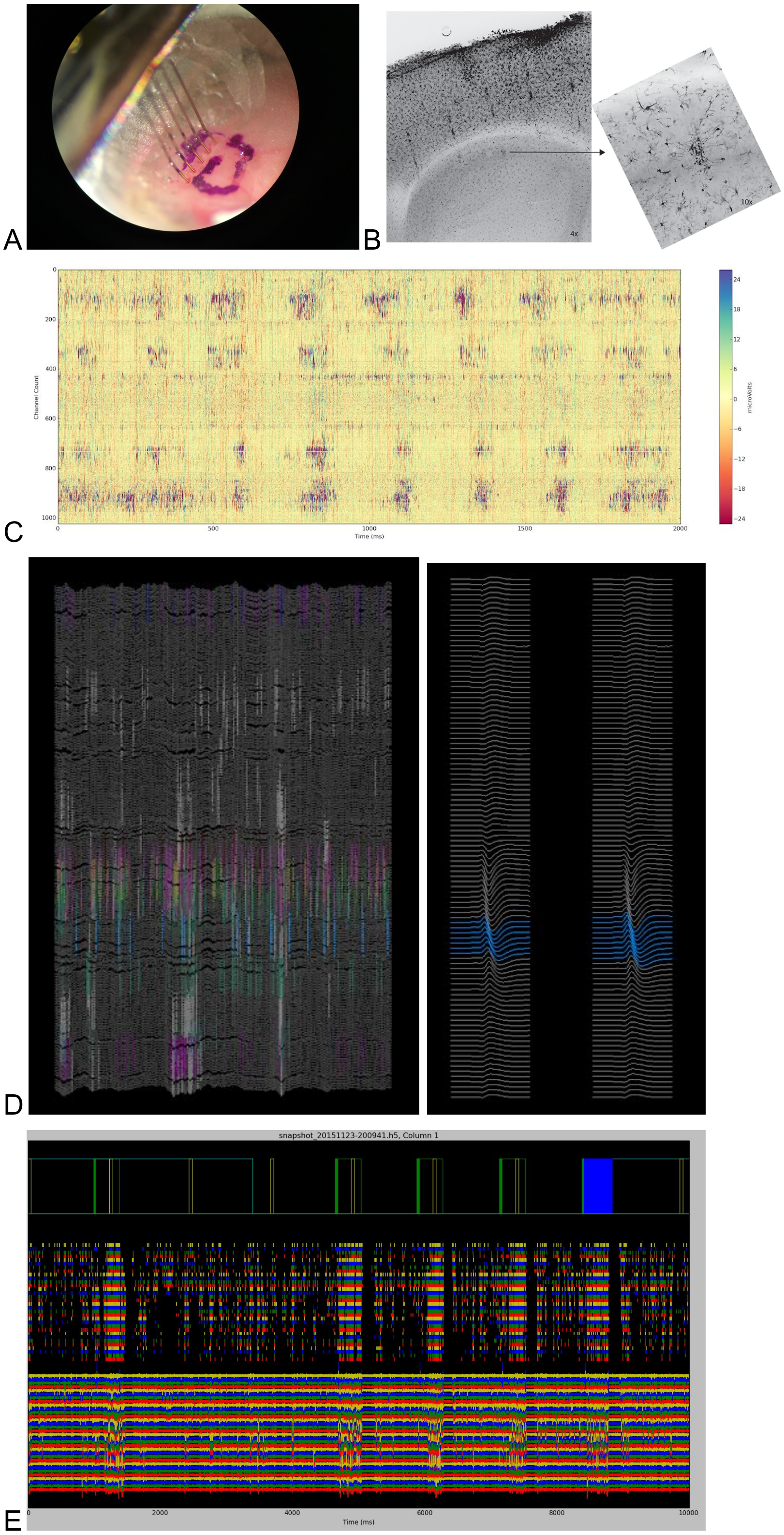

Figure 2. **Terabytes of high-channel count neural recordings** (A) Stereoscope image showing the 1020-channel probe penetrating into brain through equally spaced 300 μm diameter holes in mouse skull. Cresyl violet was added to mark the general location of the craniotomy. (B) Brains were harvested from 2 mice post-experiment for subsequent reconstruction of the location of the patched neuron and dye track from the probe. (C) A “waterfall plot” of 1020 channels of neural data, recorded from the visual cortex of an awake mouse. Channel number is along the y-dimension, while the x-axis represents 2 seconds of time. Variation in time is due to periodic visual stimulus, while modulation in channel count is due to channels being distributed over 5 shanks spaced 500 microns apart. (D) Left: Raster plot of spikes from 1 second of data for all 204 channels belonging to one shank as rendered in Phy (13). Different colors correspond to spikes from different neurons as sorted in KiloSort (14). Right: Average waveform on each electrode for one sorted neuron labelled blue on Left. Sixteen channels with largest waveform amplitude are colored blue. (E) Modulation of neural activity by optostimulation. At the bottom, 10 seconds of neural data (32 channels) is shown; in the middle, a raster plot of detected spikes; at the top, GPIO signals from the various stimulation signals - laser light pulse is shown in green, showing correlation with the spike bursts below. Trial number is encoded in blue. (Data courtesy of the Hasenstaub lab, UCSF)

dataset, but we are excited to report on some preliminary figures here. About 60% of the channels displayed spiking activity, often with multiple units per channel. Spike rates were high, especially in the awake state where we occasionally observed over 100 spikes per second (multi-unit activity) on some channels during bursty periods.

Table 1. Pilot Studies conducted from Nov 2015 to Oct 2016. In the pilot studies, we collaborated with six labs on acute head-fixed neural recordings in mouse, ferret, and both mouse and human stem cell-derived cerebral organoids. N = number of animals or organoids. From these pilot studies, three scientific papers and two grant proposals are in preparation, and we expect the same or better from future collaborations.

Institute	**Lab**	**Days of recording**	**Amount of Data (TB)**	**Maximum channel count**	**Animal model**	**Awake / Anesthetized**	**Brain region**	**N**
UCSF	Hasenstaub	5	1	256	Mouse	Awake	Auditory cortex	5
Brandeis	van Hooser	6	5	1000	Ferret	Anesthetized	Visual cortex	6
MIT	Boyden	2	2	1000	Mouse	Awake	Visual cortex	2
Novartis	NA	2	1	64	Human organoids	NA	NA	3
Harvard	Arlotta	2	3	64	Mouse organoids	NA	NA	6
Columbia	Hussaini	2	3	256	Mouse	Anesthetized	Hippocampus	3

Signal-to-noise was also very high due to the low-impedance afforded by our plating protocol: many spikes were 20-fold larger than the noise floor (e.g. 1mV amplitude spike compared to 47 μV RMS noise in brain and 22 μV RMS in saline above skull). Finally, the spatial oversampling of the probe channels provides a high-resolution view of neural activity which is evident in the channel-mapped plots - this shows great promise for automated spike sorting algorithms like ICA, which we are currently applying in preparation for a journal publication summarizing this experiment.

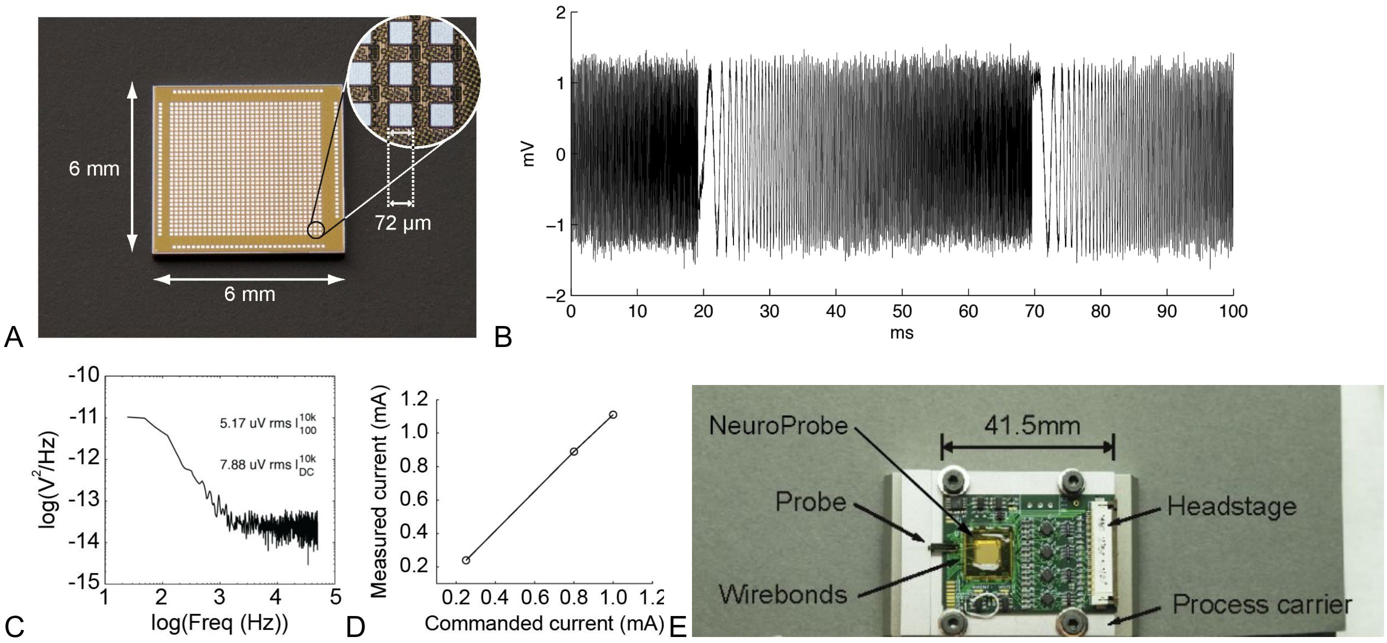

Figure 3. **1024-channel neural recording and stimulation integrated-circuit** (A) Photograph of the 1024-channel neural recording and stimulation IC. The inner 32 x 32 pads are the connection points for the electrodes. The perimeter pads are for power, chip control, and signal outputs. (B) Voltage recording from the IC while feeding an input channel with a 2.5-mV sine wave. The frequency was swept from 10 Hz to 50 kHz over 50 ms duration. The system's response is flat across the entire frequency range. (C) Input-referred power spectral density plot of the recording and analog-to-digital conversion chain. The integrated noise was 5.17 μV rms in the 10 ~ 10k Hz range (for spike recordings), and 7.88 μV rms in the DC to 10 kHz range (for LFP recordings). (D) Per-channel current sink. When enabled, it can sink between 250 μA to approximately 1 mA of current. Note the linearity of the current sink. (E) Prototype 64-channel MIT probe wire bonded to Columbia headstage, with the 1024-channel IC (NeuroProbe).

Additional validation testing for Willow has come from our “pilot study” program (**Table 1**). Over the past 12 months, we have conducted pilot studies in six labs and recorded neural activity with three different animals in three different brain regions both awake and anesthetized. The studies were crucial to assess the functionality and ease of use of the Willow system in actual in-vivo electrophysiology experiments with the involvement of graduate students, postdocs, and PIs. The goal was to lay bare what aspects of the hardware and software functioned well and what could be improved. In addition, data from the pilot experiments has provided material for writing papers and grants. The results from our Phase I grant are encouraging, however much important research involves the study of freely-moving behavior. But like all commercially available headstages, our headstages are too bulky for 1000-channel freely-moving experiments. What is needed is a single silicon chip device composed of an integrated circuit that can record and stimulate from 1024 channels.

Introducing NeuroProbe, a multipurpose neural interface chip (**Figure 3**) designed and tested in the Bioelectronic Systems Lab at Columbia University. NeuroProbe is a multipurpose, low-noise ASIC for neurobiological experiments. It has 1024 input channels implemented as a 32-by-32 pixel array. Signals from these input channels are amplified and multiplexed down to 16 analog outputs. Each input channel may be independently configured to perform one of the following tasks: voltage recording, current recording, voltage source, and current sink. The input channels are capable of delivering voltage-based electrical stimuli. In combination with the on-chip potentiostat amplifier, each input channel is also able to perform cyclic voltammetry measurements (such as used for measuring local concentration of neurotransmitters in brain).

These chips were manufactured by IBM. The diced ICs were shipped to the Columbia Team for flip-chip bonding onto miniature, light-weight printed circuit board for testing (June 2015). Future chips will be manufactured by TSMC (Taiwan Semiconductor Manufacturing Company), using their 180 nm 1.8/3.3 mixed-voltage technology, multi-project wafer (MPW) runs, as part of their CyberShuttle Program.

To accelerate both online analysis of live streaming neural data and rapid offline analysis of pre-recorded data, we will assemble and program a custom computer built from off-the-shelf parts with a PCIe bus topology, including an FPGA for data acquisition, NVMe drives for high-speed storage, and GPU's for massively parallel data processing. In preliminary tests on a benchmark system composed of a Kintex-7 FPGA connected by PCIe bus to NVMe non-volatile storage, the LeafLabs team was able to consistently achieve 1500 MB/s write speeds to a single NVMe drive by employing thermal dissipation, utilizing asynchronous IO, and tuning our write parameters. Adding a second drive saturates the PCIe bus at 2500 MB/s across the bus. These data rates are an order of magnitude are higher than the raw data rate for the 1000 channel NeuroProbe. The FPGA development (Verilog) resulted in two HDL modules: a control module for register IO, and a data module for data streaming. Both modules communicate with the host computer over PCIe using the open-source RIFFA framework (riffa.ucsd.edu). At the lowest level, RIFFA handles the PCIe transactions with the FPGA, which controls register and data IO. To create a GUI, the RIFFA library calls are wrapped into methods for basic transactions with the hardware: initialization, parameter settings, frame requests, and streaming. A daemon which runs on a CPU coordinates communication with a client. The user interface itself is contained in an easy-to-use, graphical thin client which presents controls and feedback to the user by communicating with the daemon.

Approach

Aim 1. Design and implement 1000-channel silicon probes for freely-moving neural recording.

We will utilize previous research in our groups to integrate two existing technologies - 1020 site neural probes and 1024 channel neural recording and stimulation chips - into a compact device capable of 1000 channel freely-moving neural recording and stimulation. The design and implementation will be optimized for high yield and low cost. We will first produce at least 90 re-designed 1020-channel probes to create a sufficient quantity of probes to perform pilot studies, and then continue with a focus on new designs for innovative pilot studies.
Rationale. Like all commercially available headstages, our headstages are too bulky for 1000-channel freely-moving experiments. What is needed is a single silicon chip device composed of an integrated circuit

that can record and stimulate from 1024 channels. To address this need, we will use NeuroProbe, a multipurpose neural interface chip (**Figure 3**) designed and tested in the Bioelectronic Systems Lab at Columbia University. NeuroProbe is a multipurpose, low-noise ASIC for neurobiological experiments.

By connecting directly the NeuroProbe from the Columbia team to a silicon probe from the MIT team, we can create a novel silicon probe that can record and stimulate 1000 closed-packed sites, be compact enough for freely-moving experiments in rodent, and reduce headstage cost by a factor of 10 to $1 per channel.

Method M1.1 Fabricate silicon probes with 1000 close-packed recording sites.

In the Microsystem Technology Laboratory (MTL) at MIT, we have developed a process to fabricate 1020-site close-packed multielectrode arrays on silicon (5,15, **Figures 1,2**). The fabrication process has been partially optimized for high yield and minimized tool time to reduce cost. To leverage existing tools and avoid any new process steps, any new 1020-channel probes will be made with the same MTL process. To minimize probe size, we will simplify the probe design by excising the wire-bond pads and directly connect the multielectrode array to the contact array for interfacing with the chip (a method validated in (15)). We will produce re-designed 1020-channel probes, using the existing MIT process, to create at least 90 probes in Year 1, about the same number of probes as our existing inventory of NeuroProbes (about 90), and a sufficient quantity of probes to perform pilot studies beginning in Year 2 (Aim 2). Probe fabrication will continue in Year 3 with a focus on new designs for innovative pilot studies.

Method M1.2 Assembly of probe and NeuroProbe chip with hermetic packaging.

The new probe design (M1.1) provides an interface to all connections on the NeuroProbe chip. Each NeuroProbe chip will be flip-chip bonded to a silicon probe by automated assembly machine. To reduce cost, the NeuroProbes can be connected to probes still in the wafer before break-out. Probes will be quality-tested to avoid connecting chips to bad probes.

Our existing headstage PCB design (**Figure 3E**) is a full implementation with all 1024 channels streamed out through the connector on 16 output lines. The design is optimised to be compact (about 1/4 of a credit card), light-weight (about 4 g), and dissipate minimal heat. We will modify this board as necessary to address any concerns that may arise during pilot studies (**M2.2**). The 16 output lines carry 64x multiplexed analog signals to be digitized by a high-speed analog-to-digital converter (ADC) chip contained on a new simple interface board to connect a NeuroProbe to a Willow datanode (6). If needed, a design with the ADC on the headstage will be considered.

To complete the fully-assembled probe, the integrated multielectrode array and NeuroProbe is glued and wire-bonded to a fully-populated PCB. As a final step, encapsulation of the NeuroProbe in opaque epoxy prevents light artifacts and provides protection from breakage.

Method M1.3: Refactor Willow datanode to communicate with NeuroProbe

For testing, we can communicate with any NeuroProbe using an existing test system programmed with a custom interface. For neural recording, the Willow datanode firmware will be modified. In Phase I, the FPGA core architecture of Willow was comprised of four main cores - Control, Data Acquisition (DAQ), SATA, and Ethernet - all written from scratch with the exception of the SATA core. For Phase II, the core architecture will be refactored to take advantage of technology development trends in FPGAs, as described in detail in M3.1. As part of Aim 1, the DAQ core will be modified to communicate with NeuroProbe. In the Phase I design, the DAQ core included drivers to communicate over a low-level SPI bus protocol to individual Intan chips, higher-level drivers to initialize the chips and acquire data, and an aggregation layer to combine data from all of the chips and make this aggregate data available to other cores through a RAM interface. For Phase II, the single-chip interface of the 1024-channel NeuroProbe allows the aggregation layer to be dropped, and uses a custom low-level bus protocol as specified in the NeuroProbe Chip Manual (16). The rest of the structure and function of the DAQ core from Phase I is retained in Phase II. Success means that our FPGA-based Willow datanode can communicate with a connected NeuroChip and save data to attached storage drive.

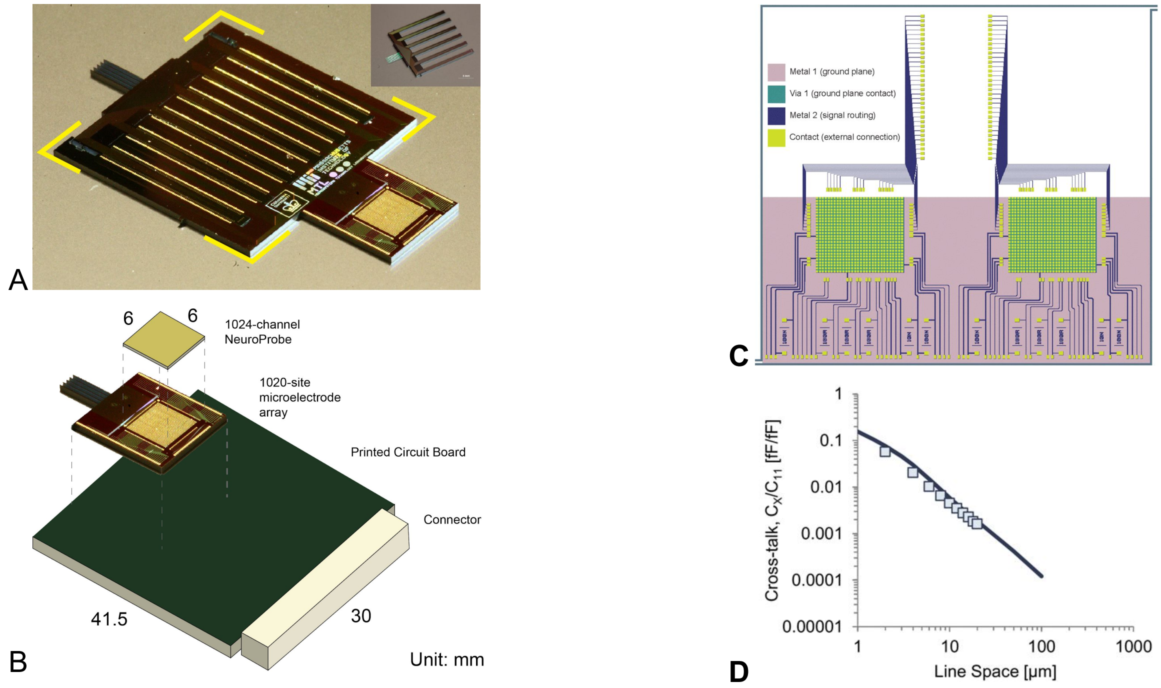

Figure 4. **Progression of probe packaging.** (A) A silicon interposer chip provides a quick way to integrate existing 1K channel silicon probe (inset) with existing 1K channel neural amplifier chip (Figure 3). The principal goal of this design was to demonstrate end-to-end functionality with little regard to compactness. (B) As described in M1.1-2 and illustrated here, the interposer design can be simplified and integrated directly into the silicon probe. The silicon shrinks in size allowing a similar reduction in PCB size. The final size of the assembly will be determined by the connector. See Figure 3E for an example implementation. (C) One way to have mixed analog and digital signals on a single silicon chip is to separate analog and digital signals and include a ground plane beneath the digital side. Adapted from (15). (D) The isolation between two neighboring wires for both measured and simulated results. The design in B will utilize wire spacing of 400 µm or more and experience crosstalk of 1e-5 or less by extrapolation. Adapted from (15).

Method M1.4: System integration and end-to-end test in saline.
For testing, fully-functional systems will be assembled at MIT or LeafLabs. To begin, the electrical noise of the complete, powered-on system will be measured (as an RMS input-referred value from 100-10000 Hz), including a Willow datanode connected by cable to 1020-site multielectrode array with integrated miniature headstage, characterized by submerging the array and reference wire and ground wire in 0.9% saline. Testing, cleaning, and electroplating steps are performed on every probe before first use to assess the quality of the electrode array (i.e. number of good, short, and open recording channels) and to lower impedance of recording sites for improved signal quality. Electrical noise is acceptable if 20 µV (RMS from 100 to 10k Hz) or less, excellent if less than 10 µV, and outstanding for values approaching 5.17 **µ**V (the noise of a NeuroProbe (**Figure 3C**)). Optionally, a low-amplitude signal will be injected into the saline (e.g. 1 mV amplitude, 100 to 1 kHz, sine wave) and recorded for comparison. Success means that the data acquisition side of the Willow system is ready for neural recording.
Potential pitfalls and alternative strategies. If for any reason more 1000-channel silicon probes are needed, a new wafer (about 90 1000-channel probes) can be fabricated in MTL at MIT in less than 4 weeks (**See Letter of Support**). Likewise, if for any reason more NeuroProbes are needed, a CyberShuttle MPW run at TSMC yields about 80 chips for $20k, with subsequent runs costing a small fraction of this cost.

Aim 2 - Perform 1000-channel freely-moving neural recording in rodent.

Using functional 1000-channel silicon probes from Aim 1, our goal in Aim 2 is to perform 1000-channel neural recordings from freely-moving rodent, e.g. rat or mouse. To our knowledge, these recordings, if successful, will be the first of their kind.

Rationale: As basic functionality of the modified system is tested in Aim 1, Aim 2 will test the viability of the system for in vivo neural recording. We will ensure the probes are compact and robust enough for implantation on the head of a rodent (either temporarily or permanently) while still allowing freely-moving behaviour. We will first demonstrate that a new probe and Willow can record from the brain of a head-fixed anesthetized/awake rodent. Then, in collaboration with experts at chronic neural recording with silicon probes, we will perform pilot studies in at least three labs (see **Letters of Support**) to demonstrate 1000-channel freely-moving neural recording in rodent.

Method M2.1 Perform 1000-channel head-fixed neural recording in rodent.

Proceeding to the next step after saline test, the new probes and Willow are used to record neural activity in (any region of) cortex in head-fixed anesthetized rodent at MIT in collaboration with LL. Success for these head-fixed recordings will not depend on the compactness of the probes, only on the electrical performance and signal quality of the probes and system in general. Demonstrating compactness will happen in the next subaim M2.2. The recordings will be performed at Columbia in mouse or rat or both. If needed, neural activity can be modulated by presentation of stimuli, e.g. for visual cortex show drifting gratings and natural scenes.

Any surgical procedures performed at MIT willow follow guidelines laid out by the Committee on Animal Care at MIT, and utilize isoflurane anesthesia (1.5-2.5%) with administration of multiple analgesics. First, stainless steel headplates will be attached to male C57Bl/6 mice of 8-12 weeks of age. After 1-5 weeks of recovery, on the day of recording or one day before, craniotomies will be drilled under isoflurane anesthesia. 200-300 μm diameter circular craniotomies will be drilled at stereotaxically-defined coordinates above the visual cortex (e.g. 2.8 mm A/P, 3.0 mm M/L to bregma), either with a hand drill or with the autodriller robotic system (17). At the start of a recording session, a mouse will be initially anesthetized with isoflurane in an induction chamber and affixed by his head plate to a metal holder, with his body fit snugly inside a 3D-printed tube. For an awake session, a mouse anesthetized as above will be woken up by the cessation of isoflurane delivery. The probe will then be inserted perpendicular to the brain surface by a software-controlled, motor-driven linear stage to its target depth, and allowed to settle for at least 10 minutes. A small computer screen will be placed at a roughly 45 degree angle in the mouse's right visual field. 8 minutes of a visual stimulus can be played of either sinusoidal drifting gratings or a natural scene of reeds blowing in the wind (Chicago motion database) to elicit neural activity. A photodiode will be placed in the lower left of the screen and digitized for synchronization. A typical recording session consists of 1 or more 8 minute presentations in succession. In a successful experiment the activity of populations of neurons is observed and recorded for hours, and the activity is modulated by stimuli and brain state (e.g. anesthesia) which are also captured for correlative analysis.

Method M2.2 Perform 1000-channel freely-moving neural recording in rodent.

To demonstrate 1000-channel freely-moving neural recording in rodent, we will collaborate with three different labs with published expertise at chronic neural recording in rodents using silicon probes to perform multi-day pilot studies (see **Letters of Support** for Pilot Studies). This will determine whether the probes are compact and robust enough for implantation on the head of a rodent (either temporarily or permanently) while still allowing freely-moving behaviour. In Phase I, pilot studies were aimed at acute head-fixed neural recording, with the longest continuous pilot study lasting 5 days. For chronic freely-moving recordings, we expect each pilot study to continue for weeks to months. Accordingly, each pilot study will begin with a couple days of installation and training, followed by us leaving the system with the lab to perform recordings independently for several more weeks. One collaborator in particular (Dan Polley at Mass Eye and Ear) has a lab in close proximity to LeafLabs, convenient for longer duration chronic studies with recordings spread out over months.

Previous pilot studies in Phase I were invaluable in determining ways in which Willow could be improved to better suit the needs of each user. After each pilot study we implemented new hardware and software features in Willow to provide missing functionality. We will repeat a similar iterative process of refinement here by taking the system to multiple labs for pilot studies. Aims 2 and 3 will unfold concurrently and allow rapid iteration cycles to develop the tools that analyze data to enable better optimization of experimental methods (see Timeline as **Table 2**).

Potential pitfalls and alternative strategies.

Freely-moving recordings have been performed in mouse with headplants with a range of size (1900 to 4500 mm^3) and weight (1.8 to 5 g) (28-30). If the headstage is discovered to not be compact enough for freely-moving recordings in mouse, the weight and size can be reduced through simple means such as selecting a smaller connector or smaller components on the PCB or light-weight PCB, without changing the functionality. To shrink the headstage still further, it may be possible to pursue more radical design changes such as hardening the chip configuration to lower the pin count on the connector and cable, or even create a variant of the headstage that only records with no stimulation. Alternatively, freely-moving recordings will be pursued in rat which can carry larger head implants than mouse - size (500 to 12,210 mm^3) and weight (3.1 to 20 g) (28, 31-34).

Aim 3: Implement hardware and software solutions for scalable data analysis.

We will develop Willow into an easy-to-use system for integrated data acquisition, storage, and real-time processing, based on our novel computing architecture called Aspen.

Rationale: One of our aims in Phase I was to develop a simple and intuitive user-interface for Willow. This was achieved with WillowGUI, graphical desktop application for interacting with the recording hardware and visualizing the data (1). WillowGUI was successful in making control of the Willow hardware accessible to neuroscientists in pilot labs and installation sites. But one thing we discovered in our user studies was that the torrent of data produced by high channel count probes introduced challenges downstream of the data acquisition – what do we do with the terabytes of data produced in a typical recording session with Willow?

Our user studies have identified three main classes of data processing that are important in electrophysiology experiments. *Realtime* processing applies to a continuous stream of data, giving the user immediate feedback (within seconds) on the effect of an action. This is typically implemented as a live visualization of streaming data, perhaps with basic processing such as filtering, spike detection and simple sorting, to be observed during the course of the experiment. *Semi-realtime* analysis is performed on selected chunks of data, typically 1 to 10 seconds in length, in an off-line but immediate way (within minutes). This gives users a chance to explore the data interactively and in detail, which is particularly necessary for high channel counts. Finally, *post-hoc* analysis occurs after an experiment (within hours to days), and typically involves the full, methodological analysis of the data for resulting conclusions and publication.

Each of these requirements becomes challenging to the point of disruption with high-channel count systems. To achieve widespread adoption in neuroscience labs, Willow must offer a streamlined solution to data analysis that supports each of these paradigms simply and with minimal overhead.

M3.1: Port the NeuroProbe acquisition core from M1.3 into the Aspen acquisition framework.

Developing the NeuroProbe acquisition core on the Phase I Willow system will allow us to rapidly prototype the communication protocol on a proven 1024-channel acquisition system (M1.3). To polish this into a full-featured system for modern neuroscience experiments, we will port this acquisition core (HDL) to our newly-developed framework called Aspen. Aspen is a unified architecture for acquiring, storing, and processing data from high-throughput experiments. It is based on a PCIe bus topology, and includes an FPGA for data acquisition, NVMe drives for high-speed storage, and GPU's for massively parallel data processing.

Aspen development has already begun in support of Lotus, a high-speed lightfield microscopy system for neuroimaging (see **Preliminary Data**). The modular nature of Aspen's architecture allows us to plug any acquisition front-end into the framework. A single Aspen node is capable of acquiring over 1500 MB/s of data to disk – more than 20x the throughput needed for the 1024-channel Neuroprobe. Beyond being highly capable and easy-to-use, this architecture has the advantage of data locality: once the data is acquired, it is immediately available for visualization and analysis in the context of a robust and high-speed file system.

M3.2: Implement a pipeline for hardware-accelerated data analysis using GPU's

As noted above, the Aspen system makes acquired data immediately accessible for local processing. One way to efficiently process large datasets is using general-purpose GPU (GPGPU) methods (11, 14). We have already explored the off-line (post-hoc) analysis of neural data using GPU acceleration (CUDA) in previous work, with promising results. Here, we will develop a pipeline for realtime and semi-realtime (defined above in **Rationale**) GPGPU processing in parallel with acquisition. This will enable researchers to write custom analysis routines that can be applied to live visualization, and the interactive exploration of high-channel count data over the course of an experiment, using technology that can scale to ultrahigh channel counts.

M3.3: Implement exemplary analysis routines useful for Willow data

Once an analysis pipeline is in place (M3.2), we will supplement it with analysis routines that we have found to be useful in our experiences with high-channel count experiments, and with these probes in particular. In the modern jargon, M3.2 will introduce the platform for data analysis, and M3.3 will develop the “killer apps”. These include visualization, processing, and in particular: spike sorting. Spike sorting of neural data, even at modest channel counts, has been identified as a significant pain point for electrophysiologists. High-density silicon probes offer the promise of automated spike sorting, but for this to be adopted, it will need to fast, accurate and easy-to-use.

We have explored several options for the automated spike sorting of Willow data. One method is Independent Components Analysis (ICA), which has been tested against ground-truth Willow data with very good results (26). Besides being accurate, ICA has the potential for significant hardware acceleration on account of it's heavy use of matrix multiplication (23,24). Another contestant is Kilosort, the latest release in the open-source lineage of KlustaKwik, etc. (14). We have tested KiloSort on data from a 1020-channel head-fixed recording in mouse, using both CPU and GPU hardware for the computation. The results are very promising (**Figure 2**). However, we found it nontrivial to enable GPU-acceleration and configure KiloSort for our data; also, realtime spike sorting is not offered by default. By integrating these state-of-the-art approaches into our analysis pipeline, we can make them accessible to researchers interested in high channel counts for the first time.

Table 2. Timeline. LL = LeafLabs. MIT = Synthetic Neurobiology Lab at MIT. CU = Bioelectronic Systems Lab at Columbia University. Where possible work will be performed in parallel, e.g. make probes, and order PCB and parts, and re-program FPGA, and think about hardware acceleration.

	Year 1	**Year 2**	**Year 3**
M1.1 - Probe fabrication	MIT	MIT	MIT
M1.2 - Probe assembly and packaging	LL, MIT, CU	LL, MIT, CU	LL, MIT, CU
M1.3 - FPGA re-program	LL		
M1.4 - System integration and end-to-end test	LL, MIT, CU	LL, MIT, CU	
M2.1 - Perform 1000-channel head-fixed neural recording in rodent		LL, MIT, CU	
M2.2 - Perform 1000-channel freely-moving neural recording in rodent		LL, Pilots	LL, Pilots
M3.1 - Port NeuroProbe acquisition core into Aspen framework	LL		
M3.2 - Hardware-accelerated data analysis using GPU's	LL	LL	
M3.3 - Implement analysis routines for Willow data		LL	LL

Our goal for Year 1 is to have the first alpha prototype systems installed in the labs of our existing collaborators for pilot testing (See Letters of Support). These systems will have all of the functionality of the final system but will have many bugs, lack polish, and and require significant support effort. Our goal for Year 2 is to have the first paid beta systems installed, though prior to final cost engineering and with considerable subsidies in price and support. Working closely with these early users, the bugs will be fixed and polish added to develop the system to a state where pilot users can perform neural recordings independently (generally, without significant support). Finally, in Year 3 the engineering focus will be to steer the design towards reduced costs and a full implementation of the production plan without sacrificing functionality or ease of use, culminating in the general product launch at a profitable configuration needed to sustain Phase III without additional outside funding.