Silicon Neural Probes for Stimulation of Neurons and the Excitation and Detection of Proteins in the Brain

Thesis by

Trevor Michael Fowler

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2019
(Defended August 27, 2018)

To Mom.

Acknowledgments

I would like to first and foremost express my sincere gratitude to my advisor, Prof. Michael Roukes, for all of his guidance over the course of my Ph.D. study and related research. His motivation, advice, and instruction were crucial for the completion of this work.

Besides my advisor, I would like to thank the rest of my committee, Prof. Andrei Faraon, Prof. Changhuei Yang, Prof. Henry Lester, and Dr. Laurent Moreaux for their encouragement and insightful questions.

My sincere thanks to Dr. Jessica Arlett, Dr. Warren Fon, and Dr. Laurent Moreaux, the senior scientists who helped guide me through this work. Without their collaboration, completing the research in this thesis would have been an insurmountable task. I would also like to thank Dr. Wesley Sacher and Dr. Eran Segev for their guidance and collaboration in the development of novel photonic devices. Thanks to Andrei Faraon for his collaboration in developing the optical devices described in this thesis. Great thanks to Dr. Scott Lewis for his collaboration in developing novel electron beam resists.

Thanks to Derrick Chi and the KNI cleanroom staff (Dr. Guy Derose, Melissa Melendes, Nils Asplund, Steven Martinez, Dr. Matthew Hunt, Nathan Lee, and Alex Wertheim), who spent innumerable hours training me in the cleanroom.

I would also like to thank my mother for her infinite love and encouragement during my graduate studies. Finally, I would like to thank all my friends, especially John, Mike, Anya, Andrew, Steve, Mark, Scott, Idrees, Mark II, Yunia, Max and Ann-Lauriene for keeping me sane during this crazy process.

Published Content and Contributions

Eran Segev, Jacob Reimer, Laurent C. Moreaux, Trevor M. Fowler, Derrick Chi, Wesley D. Sacher, Maisie Lo, Karl Deisseroth, Andreas S. Tolias, Andrei Faraon, Michael L. Roukes, *Patterned Photostimulation via Visible-Wavelength Photonic Probes for Deep Brain Optogenetics.* Neurophotonics 4(1), 011002 (2017), doi: 10.1117/1.NPh.4.1.011002.

TMF contributed to the fabrication, design and testing of the AWG devices.

Segev, E., Moreaux, L., Fowler, T., Faraon, A., Roukes, M. *Implantable, highly collimated light-emitters for biological applications.* US Patent Application 15/295,991, filed October 2016.

TMF contributed the development of the optical emitters.

Roukes, M., Fowler, T., Arlett, J. *Highly multiplexed optogenetic neural stimulation using integrated optical technologies.* US Patent Application 20140142664, filed on November 2013.

TMF contributed to the conceptualization and application of this patent, including the development of AWG optical demultiplexers on neural probes and the optical multiplexing scheme described in the patent application.

Abstract

This thesis describes the development of a number of novel microfabricated neural probes for a variety of specific neuroscience applications. These devices rely on single mode waveguides and grating couplers constructed from silicon nitride thin films, which allows the use of planar lightwave circuits to create advanced device geometries and functions. These probes utilize array waveguide gratings to select an individual emitter from a large array of emitters using the wavelength of incoming light, allowing for spatial multiplexing of optical stimulation. These devices were tested in the laboratory and in living tissue to verify their efficacy. This technology was then modified to create steerable beam forming for stimulation of neurons using optical phase arrays. This technology was also tested for use in fluoresence lifetime imaging microscopy and the first application of pulsed light through the photonic circuits. Finally, this technology was again modified to create laminar illumination patterns for light sheet fluorescence microscopy applications. These devices were further improved by adding embedded microfluidics to the probes. The process of creating embedded microfluidic channels by the dig and seal method is described in detail, including modifications to the procedure that were added to address potential pitfalls in the fabrication process. Next, two projects which combine microfluidics with the optical devices described in the previous chapter are detailed. One project involves combining the use of optical emitters with microfluidic injections containing caged neurotransmitters to stimulate neurons is described. The other project involves microfluidic sampling of the extracellular space for neuropeptides which are detected using ring resonator biosensors. The sensitivity of these biosensors was analyzed in detail, determining both the physical limit of detection and the effect of biological noise due to non-specific binding on the sensors.

Contents

Chapter 1

Introduction

1.1 Motivation

Fully understanding neural computation is one of the great biological unknowns. Although the properties of neurons have been studied extensively since the advent of the electronic amplifier, we have still only scratched the surface of understanding how the brain works as a whole.

Neural computation occurs at a number of scales. The most well understood scale is what occurs at the membrane level. Neural membranes are themselves voltage sensors, in which sodium channels are triggered by a rising membrane potential above the resting potential of the membrane. Once the potential reaches a threshold, these channels open, allowing sodium to flow into the neuron. The membrane potential of the neuron increases rapidly until potassium channels are opened, returning the membrane potential to below threshold. The action potential provides the simplest method for accessing neural information by tracking electrical spikes. Traditional electrophysiological techniques (such as patch clamps and extracellular electrodes) measure the variations in the electrical potential surrounding the neuron, which in turn provides a proxy for neural information flow in the brain.

If we wish to communicate back with the brain electrically, however, it is much more difficult than simply placing electrodes in the vicinity of a neuron[12]. To target single neurons, electrodes must come in contact with the neuron, often using patch clamp electrodes. This, unfortunately, is not scalable if access to a larger neural circuit is desired. Thus, optical methods have been developed to stimulate action potentials in neurons. Light can easily be focused and manipulated over large areas using technologies such as spatial light modulation microscopes, allowing scientists to stimulate multiple neurons over large brain regions[92]. The limitation of optical technologies is the ability for light to penetrate tissue without scattering, which can be limited to a few hundred microns using single photon techniques and approximately 1mm using multi-photon techniques.

Optical methods can also be used to measure neural activity using calcium sensitive fluorescent proteins[3]. These methods are powerful because they allow for the measurement of neural activity through non-contact means, and can leverage the technology already developed for the microscopy community to create powerful measurement tools. These tools, however, are still limited by the scattering limit of light in tissue. Since these techniques often require the use of blue light, scattering becomes the critical factor in understanding the measurement depth of these techniques. Bringing the measurement technology into the brain is a potential solution to solve the scattering problem.

Neural information isn't completely contained within the action potential. Synaptic transmission is a critical and essential step in neural communication as well. To interface with neural circuits, it is desirable to utilize neurotransmitters directly to induce or inhibit neural spiking, as the neurotransmitter used will have an effect on the signal induced in the downstream neuron[13]. Caged neurotransmitters have been developed, where a neurotransmitter is passivated chemically and is only activated by exposure to light at a specific wavelength. Thus, it is possible to leverage optical technology to control synaptic transmission in neurons as well.

In what follows we describe a number of new silicon probe based technologies for directly modulating and measuring neural activity. Utilizing silicon as a substrate allows one to co-opt the great development that has been made in semiconductor fabrication at the micro- and nanoscale. These processes will allow for the creation of novel devices, enabling a new frontier for our ability to interface and understand neurons in their native environment. Silicon based neural probes for extracellular recording have already made significant impacts on the way neuroscientists conduct research in living animals and even humans, which is described in the following section. New technologies which utilize light, fluidics, and electrochemistry are on the horizon for the measurement and modulation of the chemical and electrical environment of neurons in situ.

1.2 Review of microfabricated probe-based sensing

Although the electrical nature of the neural tissue was understood from the experiments of Galvani in the late 18th century, the first single-unit recordings of neural activity were elucidated by Renshaw et al. in 1940 by measuring the electrical activity of hippocampal pyramidal cells using a glass pipette microelectrode[89]. In these experiments, Renshaw et al. began with a glass capillary which they heated and pulled in such a way to preserve the lumen of the capillary while reducing in size the tip of the tube. Inserted into the lumen of the capillary was a silver-silver chloride electrode submerged in a potassium chloride electrolyte. Using this device, along with newly developed analog amplifiers, Renshaw et al were able to record some of the first transients from single neurons. Unfortunately, the amplifier being used had insufficient bandwidth to observe single action potentials from neurons. Technical developments over the next ten years allowed for the measurement of single unit spiking in the brain of living animals, resulting in the seminal work of Hodgkin and Huxley in 1950, who developed a model of the time dependent dynamics of a single neural action potential using the squid giant axon as a model system[46].

As time went on and the techniques associated with single unit recording matured (such as the development of the patch clamp, better amplification schemes, etc.), the desire to record action potentials from multiple units simultaneously became very desirable, the concept being that by measuring multiple neurons simultaneously would allow scientists to measure multiple interrelated neurons simultaneously to elucidate the behavior of neural circuits. First attempts at extracellular recording from multiple sites was through the use of tetrodes, or twisted quadrupoles of nichrome wire which can be used to triangulate the location of separate neurons based on differential delays between the wires[130]. Tetrodes are fabricated by first twisting four insulated nichrome wires together (first using a magnetic stirplate). The insulation was then melded together by melting using a torch, and then the tip was cut away, exposing fresh nichrome wires. Finally, the tetrode is inserted in a cannula for implantation. Using triangulation and finite element analysis, it was theorized to be possible to separate signals from separate neurons within the detection radius of the tetrode[37, 18]. The use of tetrodes has been a driving force in multiunit recording, allowing for some of the best multi-unit recordings available.

Tetrodes, however, have a major downside in that even if the size of the sensitive area of the tetrode is small, the device themselves are quite large. Since it is difficult to pack multiple tetrodes together to cover a large volume of neural activity, new technologies were required to measure larger areas of the brain with large scale coverage. Two technological platforms were developed to address this issue, one developed at the University of Utah (the “Utah” multi-electrode array), and one at the University of Michigan (the “Michigan” probe).

The Utah probe was first developed by Richard Norman, the fabrication protocol first published in [14] in 1991. The Utah probe (initially intended for stimulation), is fabricated from a silicon wafer using semiconductor fabrication techniques. Briefly, conductive p-type pillars are created within the n-type silicon wafer through a thermomigration process, where deposited aluminum is drawn through the wafer by a thermal gradient, leaving the wafer doped with p-type dopant aluminum as it passes eutectically through the wafer. Next, a dicing saw was used to define the long pilars which would be used to implant into the brain. These columns, however, would still be square and unsuitable for implantation, so an isotropic etch was developed using 5% hydrofluoric and 95% nitric acid was used to round the columns, and finally sharpen the tips of the probes.

Since their first fabrication in the early 1990’s, the Utah probe has continued to be refined over the last three decades. Glass and polyimide was added to the probes to improve biocompatibility[52]. Recently, a version with individual optical fibers has been fabricated for optogenetic stimulation and electrical detection [126]. However, the most significant consideration about the Utah array is that it is the only currently FDA approved neural implant for human use. This has resulted in a large body of literature describing the use of these implants in primates and humans for clinical applications[95, 103]. Utah type MEAs have also been shown to be better at preventing postimplantation immunogenicity[84].

For all of the merits of the Utah array, there are two major problems to the technology moving forward; the Utah array still doesn't allow for high density recording along the length of each shank and, due to the fabrication techniques used for the array, the Utah array doesn't allow for the integration of non-electrical technologies such as microfluidics. The alternative, introduced by K.D. Wise at the University of Michigan is the so-called Michigan Probe.

The original iteration of the Michigan probe was developed by Wise during his doctoral studies at Stanford University[132]. The first probes were fabricated monolithically from a thick silicon wafer. The implantable shank was patterned from the top using standard photolithograhy, and then an isotropic backside etch was performed to create a tapered implantable probe. Unfortunately, due to the isotropic backside etch, the implantable portion of the shank was very short.

As a solution to that problem, an alternative process was developed in the 1980's in Wise's lab in Michigan[75]. Instead of anisotropically etching the backside, Najafi et al. first used ion implantation to alter the doping profile under the location of the probe. After patterning and deposition of the conductive lines (tantalum or polysilicon) and an insulating overcoat to insulate the wires, the wafer was thinned from the backside to 15-25μm and finally the probes were released by a ethylene diamine-pyrocatechol etch which will only etch the region left un-doped after ion implantation. This process produced the first thin, deeply implantable probes using planar semiconductor fabrication techniques. This also implies that any of the techniques currently available to semiconductor fabrication can also be utilized on these shanks. For example, the Michigan style probes have been converted for electrochemical applications.

With the advent of silicon-on-insulator (SOI) technology, a new pathway for the fabrication of Michigan style probes was developed at Caltech[90]. These probes incorporate a combination of passive electrical recording probes (with 1024 active, available recording sites) with separated, active electronics to minimize the amount of heat deposited to the tissue. SOI wafers consist of a three layer stack of materials. On top, there is a 400nm-50μm layer of single crystal silicon called the device layer. In the semiconductor world, this layer is used to fabricate transistors for integrated circuits. Underneath this layer is an insulating layer made of pure SiO_2. The bottom layer consists of a thick single crystal silicon wafer, often called the wafer "handle." This layer is intended to be a mechanical base for the upper layers. SOI wafers significantly simplify the fabrication process of neural probes from a silicon substrate. Using an SOI wafer, the device layer can be used as the thin, implantable shank. The insulator layer acts a simple etch stop while defining the shape of the shank. Finally, the handle wafer can be etched from the back to produce a more stable base for the probe, allowing for simplified handling of the probe as well as a spacer for stacked probes.

The latest attempts at silicon based neuroelectrophysiological recording probes have used CMOS technology to leverage the scalability and technological maturity of the semiconductor fabrication industry. Recently the Neuropixel project, managed by Harris et. al at Janelia Farms Research campus in collaboration with IMEC, has been released to the neuroscience community[53]. These probes incorporate analog CMOS switches to reduce the total active sites on the shank (960 sites) to the 384 possible output channels on the base of the probe. On the base of the probe are electronics which are able to multiplex, filter, amplify and digitize these signals, thereby reducing the number of necessary channels coming from the probe to the setup headstage.

An additional application area for silicon based neuroelectrophysiological probes is to detect neurotransmitter concentrations in the brain. Since electrical information encoded in spiking only contains stimulatory information, it is critical to understand which neurotransmitters are active in individual brain circuits at different times to determine the role of excitatory and inhibitory information flow during neural computation. Electrochemical analysis of neurotransmitters has been a goal of the neuroscience community since the 1970's, when simple *in vitro* electrodes were first being pioneered for biological samples[1]. Since then, electroanalytical techniques have evolved significantly with the advent of advanced techniques such as fast scan cyclic voltametry and novel coatings including enzymes to to detect specific analytes within the extracellular space[118]. Selectivity of the probes is the major concern, as there are many electroactive compounds which mimic common neurotransmitters (e.g., ascorbic acid and DOPAC interfering with dopamine sensing), causing many scientists to spend significant effort engineering coatings for these probes to reject interfering chemicals from the surface of the electrode (e.g., nafion, tyrosinase coatings). Sensitivity of these probes is often near the concentration limit of these compounds in the brain, often in the mid to low nanomolar range, increasing the difficulty of their application.

The first demonstration of *in vivo* neutransmitter detection using modified electrophysiological probes was by the Kipke group in 2009[51]. Johnson et al. began with a Michigan-style probe which had electrphyisological electrodes instead coated with liquid Nafion, an ion exchange membrane, and the solvent was allowed to evaporate. The Nafion membrane is intended to filter incoming molecules based on charge, making it preferentially selective towards dopamine. Experiments were executed using single potential amperometry at +500mV to optimize for the peak oxidation potential of dopamine. All transients were measured against a silver-silver chloride reference electrode. In this paper, Johnson et al. showed that it was possible to detect electrochemical transients in the striatum (attributed to dopamine) during the stimulation of the medial forebrain bundle.

Similarly, a neural probe system for measuring choline concentration is described by Frey et al.[32]. In this paper, the authors describe a 250μm by 250μm by 8mm silicon based probe with platinum microelectrodes coated with choline oxidase to specifically detect choline in the extracellular space. The choline oxidase produces hydrogen peroxide when in the presence of choline, which can then be reduced at 0.7V to produce a signal proportional to the concentration of choline. These probes were successfully able to detect 10μM concentrations of choline in a laboratory setting while successfully rejecting signals from dopamine when introduced to the probe.

Glutamate, the primary excitatory neurotransmitter in the brain, has also been a target for electrochemical detection on implantable probes[127]. These probes were developed to detect glutamate on platinum electrodes also using an enzyme (glutamate oxidase) to convert the neurotrasmitter concentration to hydrogen peroxide, which may then be detected using constant potential amperometry. These probes were fabricated from silicon wafers, producing probes of 150μm thickness, 120μm width, and length between 2 and 9mm, with four individual sensor electrodes with dimensions 40μm in width and 100μm in length. These devices used a more complicated functionalization to ensure that the signal was free of interfering compounds. Each electrode was coated first with a layer of polypyrrole intended to block signal from dopamine leaking into the sensor. Nafion was then coated on top of the polypyrrole to reduce ascorbic acid interference. Finally, glutamate oxidase was coated onto the nafion membrane, completing the functionalization stack. These probes were successfully able to detect 10μM concentrations of glutamate without responding significantly to ascorbic acid or dopamine. Further advancements have been made by combining detection of glutamate and dopamine on the same probe[119].

The alternative to using electrochemical analysis techniques is microdialysis, with the long standing field reviewed in Chefer et al[17]. This technique utilizes a microfluidic cannula which can be inserted into living tissue, often with a silica tip. Surrounding the tip is a semi-permeable membrane which allows small molecules, such as neurotransmitters or smaller neuropeptides, to diffuse across the membrane while preventing contamination of the tissue by larger macromolecules or composites such as viruses. These devices are set up to perfuse a fluid (such as artificial cerebrospinal fluid) through the cannula, allowing molecules to be dissolved in the passing fluid. The fluid is then collected and analyzed using any desired biomolecular detection technique (e.g., mass spectrometry, ELISA, etc.). These cannulas are considered minimally invasive and are used in a variety of animals to measure local biomolecule concentrations in all regions of the brain.

The major advantages and disadvantages of this technique are summarized in [83]. One of the most obvious advantages is the generic nature of the sampling of the extracellular space. The perfusate can be sampled for numerous compounds, and each target can have a customized analysis procedure to improve detection. This provides a superior option to the electrochemical probes discussed above, which must use complex chemistry customized to identify individual analytes from one another. The probes are small (less than 1mm in diameter) and are considered minimally invasive and suitable even for use in humans. The drawbacks to microdialysis include slow measurement rates, dilution of target compounds, neural damage, and contamination of tissue. Perfusion rates for microdialysis experiments are typically in the low μL per minute range, requiring tens of minutes to obtain a sample large enough for external detection. Furthermore, there is a trade-off between the size of the exposed membrane and the amount of analyte captured by the probe. Larger probes capture more analyte, but the uncertainty in the location of the signal also is increased. By requiring a diffusion gradient to move molecules across the membrane, the analyte will always be diluted compared to the concentration in neural tissue. Furthermore, this requires careful calibration to ensure that the concentration of the extracellular space is properly reflected in the measured data.

No examples of neural probes with microdialysis membranes has been published, although it was originally suggested in [112]. These probes utilize microfluidic channels created first by depositing a series of lines of platinum on top of the silicon substrate. The silicon beneath these platinum wires is then etched away using XeF_2 gas while the surrounding area was masked using photoresist. Membranes are made using a phase inversion process in cellulose acetate or polyethersulfone (PES). These films were cast in a solvent (such as DMSO) and spin coated onto the wafers over the microfluidic channels. The platinum wires were sufficient to prevent the film from entering the microfluidic channels, instead sitting on top of the platinum wires. The wafer was then placed in a water bath, in which the cast solvent diffuses out of the membrane and the water diffuses into the membrane rapidly. This process is described in detail in [66]. This solvent exchange process produces a porous membrane through which small molecules can diffuse. The degree of porosity is controlled by the concentrations of multiple solvents, the type of solvent used, and the mass percent of the polymer used to create the membrane. These chips were shown to isolate large proteins while allowing smaller molecules to pass through, which is the intent of the microdialysis membrane. This also shows that it is possible to create microfabricated microdialysis membranes for use on neural probes.

1.3 Probe based stimulation of neurons

1.3.1 Optoprobes for optogenetic stimulation of neural tissue

There has been a revolution in experimental techniques for neural stimulation upon the discovery of channelrhodopsins [10, 137, 139]. These proteins were isolated from microalgae which use channelrhodopsins to orientate themselves towards sources of light. These proteins are light activated ion channels which become conductive when exposed to light. After being genetically isolated and either transfected or knocked-in to animals, they were then used to create depolarizing (and later hyperpolarizing in the case of halorhodopsins) currents across neural membranes, thereby manipulating the activity of said neuron. This tool has proven to be extremely powerful, allowing targeted stimulation of single neurons or groups of neurons in living tissue. The use of light allows for spatiotemporal activation of neurons, and the genetic component allows these genes to be coupled to known promoters so that these proteins can be transcribed only in a specified type of neuron.

Channelrhodopsins were activated in early experiments through the direct implantation of optical fibers into mice or through projection of blue light through microscope objectives onto whole brains or neural slices. Naturally the development of silicon based electrically active neural probes with optical excitation was quickly realized. First demonstrations of optically active neural probes were simple yet effective. These devices simply used commercially available Michigan style probes produced by NeuroNexus and glued an optical fiber to the top of the probe[93]. This was the first demonstration of combining optical and electrical recordings in the same shank, and was effective in demonstrating the principle of combining optogenetic stimulation with other recording technologies. The major disadvantage of this technique is that the beam is highly divergent in the tissue, essentially activating a large region of the brain to the same stimulus. This technology is also not scalable for improved spatiotemporal stimulation of neural circuits.

One of the first implementations of on-shank waveguides described by Zorzos et al. in 2010[142]. The devices described in this reference are based on silicon oxynitride waveguides deposited on a silicon substrate. Probes were first deposited with 3μm of silicon oxide as the bottom cladding for the waveguides, after which 9μm of silicon oxynitride was deposited. Wafers were patterned with photoresist and the silicon oxynitride layer was then etched using an undisclosed RIE process. Waveguides were then coated with an additional 3μm of PECVD silicon oxide. Probes were then coated with aluminum to act as a mirror surface which will aid in the containment of light in these large waveguides, especially when the waveguide is bent. Probes were defined using a DRIE etch process from 6 inch wafers, resulting in extremely thick shanks. 12 waveguides per probe were realized using this technology. These probes showed high optical throughput from both laser and LED sources due to the very large size of the multimode optical waveguides. Due to the large cross section, coupling efficiency was high but beams produced from the waveguide were highly divergent in tissue. Furthermore, due to the multimode design of the waveguides emission was only possible from the edges of the shank; emission out of the plane of the shank is impossible with such large multimode waveguides. This idea was expanded to include metal electrodes by Wu et al. in 2013[133].

Continuing with the theme of multimode waveguide stimulation, Im et al. demonstrated a neural probe with polymer waveguides in 2011[48]. These probes utilized a SOI based technique to create silicon neural probes with significantly thinner shanks, similar to [90]. These devices started by patterning the top surface of the wafer with oxide as an insulating layer. Wafers were then patterned with metal for electrophysiological recordings, and were capped with an additional layer of PECVD silicon oxide. Next, SU-8 (a clear, photocurable polymer) was patterned on top of the wafer and waveguides were patterned photolithographically on the surface of the wafer. These waveguides were also quite large, with a square cross-section of 15μm on a side. Probes were then etched from the top side using DRIE, and released from the backside using DRIE up until the BOX was reached. The BOX layer was then removed. These probes were innovative in that they showed the first applications of photonic circuits on shanks, with patterned optical mixers and splitters on the shank for improved multiplexing. However, these devices were limited by the multimode nature of the waveguides and the size of the waveguides, ultimately limiting the multiplexability of the waveguides on shank. This method for creating optical waveguides was further improved in [108] using the Michigan probe fabrication technique and multiple optical splitters to create a large excitation volume using a single optical input.

Recent interest has been moving towards placing light emitting diodes directly on neural probe shanks. One of the first examples of this technology is demonstrated by Stark et al. in 2012[111]. These probes also utilize the commercial probes sold by NeuroNexus. These probes are similar to those described in [93], using optical fibers glued onto the shank of the probe. The probes described in [111] differ in that they have cemented a LED to the base of the probe coupled directly to the optical fibers. This allows for superior temporal multiplexing. The next logical step along this path of investigation is to fabricate LEDs directly on shank. This was accomplished by Wu et al. in 2015[134]. InGaN microLEDs were fabricated onto a silicon wafer (typically these LEDs are fabricated on sapphire substrates). Making the shift to producing LEDs directly on the shank has a significant advantage when considering both spatial and temporal multiplexing of optical signals. Probes were produced with 12 optical emitters and 32 electrophysiological measuring lines, and were used to stimulate and record from hippocampal pyramidal cells in CA1. MicroLEDs can be patterned anywhere on the shank and are limited in density only by the size of the traces leading to the LEDs. They produce significant optical fluence (up to 1mW per square mm at the surface), sufficient for activation of channelrhodopsin at a significant distances from the probe. The major downside to this technique is that the microLEDs produce waste heat which is deposited into the tissue and that the optical beam produced by the LEDs is still highly divergent. Although the physiological effect of local heating on neurons is unknown, it is thought that heating neural tissue to 39.5^{o}C can have effect spiking rates of neurons.

1.3.2 Microfluidic probes for drug delivery

Although a very powerful tool, optogenetic actuators have their limitations. Primarily, they require genetic modification of tissue. Although techniques for genetic modification have advanced greatly in the last few years, requiring the modification of the genome of the animal requires significant effort and can have unintended effects on the tissue being modified. Furthermore, optogenetic tools utilize ion channels to actuate (or inhibit) neurons, which only encodes spiking information. Synaptic transmission, however, utilizes a variety of chemical reporters which can transmit intent in a much more information rich manner. Thus, utilizing chemicals which modulate neural activity directly, such as pharmacological agents, is desirable for neuroscientists to better understand information transfer between neurons as well as natural neuromodulation of brain circuits *in situ*. To introduce pharmacological agents directly into the brain, injections of neuromodulators are required. Although using a standard syringe and needle is the most common method for introducing pharmacological agents, many attempts to make microfluidic neural probes have been made.

One of the first examples of microfluidic neural probes was demonstrated by Cheung et al. in 2003[20]. In this paper, the authors developed microfluidic channels in SOI neural probes. The wafer stack started with a top layer of 500nm silicon nitride on top of the 25μm device layer. Holes were etched into the nitride film with widths of either 1μm where the channel will be sealed or 3μm where the channel will be open. The wafer was then placed in a silicon wet etch to produce a channel under the silicon nitride layer. Holes were then sealed by depositing 2μm of PECVD silicon oxide, which would leave the 3μm holes open to the environment while sealing the 1μm holes. Electrodes were patterned on the surface using a liftoff process. Probes were defined using photolithography and deep reactive ion etching (DRIE) processing using the oxide as an etch stop. These probes demonstrated combined electrical detection in the visual cortex with microfluidic channels on individual shanks. Drawbacks of this technique are the large size of the microfluidic channels (which reduces multiplexing ability) and the thin cover created for the microfluidic channel (risking mechanical failure during implantation and handling).

One of the most complex and robust methods for creating microfluidic probes was developed in Seidl et al. in 2010[97]. The intent of these probes was to combine drug delivery through microfluidics with standard electrophysiology electrodes on a single probe. The technique used to create the microfluidic channels is simply bonding two silicon substrates together through wafer bonding. These probes have a unique inlet design in that they developed a system in which inlets are available for interface either in the probe plane or through the bottom substrate of the probes. 300μm thick wafers were prepared with a 2μm layer of PECVD oxide, which is patterned using reactive ion etching to act as a mask for future DRIE steps. Microfluidic channels are then defined in the silicon substrate using DRIE in a two layer process, one defining a deep etch for inlets and the other defining a 50μm deep microfluidic channel for fluid flow. Oxide is then removed with buffered oxide etch and a 100μm silicon wafer is bonded to the top of the substrate wafer to seal the channels using direct wafer bonding at 1050^{o}C. An oxide layer of 500nm is left between the two wafers as an etch stop for later steps. The surface is then patterned with an insulating sandwich of PECVD deposited SiO_2/SiN/SiO_2 to insulate electrodes from the conductive silicon below. Electrodes are then patterned on top of the insulating layer, and is electrically passivated using a second SiO_2/SiN/SiO_2 sandwich. Wafers are then etched from the backside to define the thickness of the shanks, leaving behind 150μm of the bottom wafer in addition to the 100μm top wafer. Wafers are finally etched from both the front and backside to release the probes. The major benefit of this fabrication protocol is that it leaves the top surface of the probes intact and undisturbed for devices on the top of the wafer. It also allows for efficient coupling of fluidic inlets to the devices. The downside is that the probes are extremely thick ($\approx 250\mu$m), which puts tissue at risk of being damaged. A three-dimensional version of this probe was released by the same research group in 2011[109]. Interestingly, these showed a floating design (utilizing a flexible PDMS microfluidic cable) to help minimize tissue damage and were highly parallelized, including sixteen shanks and four separate microfluidic inputs. Similar to this methodology, Lee et al. demonstrated a protocol using an oxide top wafer instead of the silicon for a similar purpose[58].

Following this, a parylene based microfluidic device was developed in John et al. in 2011[50]. These probes used oxide coated single crystal silicon wafers. First, metal wires were patterned on top of the wafers to create metallic interconnects and pads for electrophysiological recordings. The wafer was then coated with parylene C, a biocompatible polymer used commonly in medical devices. A trench was etched in the parylene, exposing a thin section of silicon. Silicon was then etched isotropically using XeF_2 gas to create a channel without attacking the parylene layer. The trench was then sealed with a second coating of parylene. The wafer was then etched using DRIE from the front side to define the depth of the probe, and probes were then released by backside DRIE etching. These probes also used parylene interconnects to conduct fluid and electrical signals through a flexible cable which is often desired for chronic implants. Probes had final dimensions of 100μm by 100μm by 2.8mm. These probes were implanted in live rats with the intention of inducing epileptiform activity in the brain when injecting artificial cerebrospinal fluid (ACSF) containing 4-aminopyridine. The authors successfully demonstrated abnormal spiking activity upon injection. These probes showed promising interface technology with the flexible cables; however, they also suffer from extremely large cross-sectional dimensions which are likely to induce tissue damage upon insertion.

The final microfluidic neural probe in the literature was developed by Pongrcz et al. in 2013[85]. Their devices utilized a dig and seal method for developing microfluidic channels, similar to the protocol described in this thesis. The substrate for this process is an 1μm silicon oxide coated silicon wafer. A 2μm wide, straight wall trench was etched through the oxide layer and into the silicon approximately 20μm. The wafer was then oxidized to passivate the sidewalls of the trench for the next steps. Aluminum was deposited onto the top of the wafer, protecting the corners of the trench during etching. SiO_2 was removed using a plasma etch, exposing the bottom of the trench. An isotropic silicon etch was then used to create a large channel while the passivated trench remains small. After removing the aluminum and SiO_2, the trench is filled with PECVD amorphous silicon. Finally, the wafers were metalized and probes were released using a DRIE process similar to the above examples. Probe dimensions ranged from 15-70mm long, 200-380μm thick, and 200-380μm wide. These probes were used to inject bicuculline (a light-sensitive competitive GABA agonist) into a number of brain regions, inducing spiking measured by the probes. The advantage of these probes is the robustness of the microfluidic channel design and the control over the size of the channels, which range from 5-30μm.

1.4 Overview of Thesis

This thesis describes the development of a number of novel microfabricated neural probes for a variety of specific neuroscience applications. Optical neural probes for optogenetic stimulation were first developed, with the goal of allowing for arbitrary spatiotemporal stimulation of multiple neural circuits using both static and steerable emitters. This technology was then adapted for neural imaging applications, specifically for the stimulation of GCaMP6, a calcium sensitive fluorescent protein. Next, microfluidics were added to the probes to sample the extracellular space for peptides and allow for injections of pharmacological agents. One project utilizes photonic ring resonator biosensors to detect proteins extracted from the brain. The other project combines optical emitters with microfluidics to engage in neurotransmitter uncaging in deep brain regions.

Chapter 2 focuses on the sensitivity of detection of optical ring resonator biosensors, which are used in the probes described in Chapter 5. First, a physics based analysis of the expected sensitivity of these devices is undertaken. This determines the limit of detection for mass binding to a ring resonator biosensor in ideal conditions. Next, non-specific binding is taken into account, in which the binding of unintended proteins to the ring resonator compete with the target analyte for space on the surface of the sensor. This analysis shows that sensitivities of approximately 10nM is expected for label free assays, and sensitivities of approximately 1pM is expected for sandwich assays. These values fit what is shown in the literature very well. Finally, the effect of the microfluidic flow rate on the binding rate of the analyte is analyzed, showing that the devices created should be within the reaction limited regime for reasonable flow rates and pressures in the microfluidic channels.

Chapter 3 discusses the details of device fabrication. All aspects of the fabrication process are described, including photolithography, electron beam lithography, wet and dry etch processes, thermal oxidation of silicon, and deposition of thin films. In addition, the development of a novel electron beam resist technology is described. Principals of operation and critical dimension controls are discussed. Theory of etch processing, lithography, and deposition are discussed in the context of practical applications of these technologies in Appendix A. Deposition of conformal thin films are discussed as well as how they relate to sealing microfluidic channels. Finally, fabrication protocols for probe fabrication and optical device fabrication are detailed.

Chapter 4 describes the development of optics-only neural probes. These devices rely on single mode waveguides and grating couplers constructed from silicon nitride thin films, which allows the use of planar lightwave circuits to create advanced device geometries and functions. These probes utilize array waveguide gratings to select an individual emitter from a large array of emitters using the wavelength of incoming light, allowing for spatial multiplexing of optical stimulation. These devices were tested in the laboratory and in living tissue to verify their efficacy. This technology was then modified to create steerable beam forming for stimulation of neurons using optical phase arrays. This technology was also tested for use in fluoresence lifetime imaging microscopy and the first application of pulsed light through the photonic circuits. Finally, this technology was again modified to create laminar illumination patterns for light sheet fluorescence microscopy applications.

Chapter 5 describes the development of microfluidic neural probes. The process of creating embedded microfluidic channels by the dig and seal method is described in detail, including modifications to the procedure that were added to address potential pitfalls in the fabrication process. Next, two projects which combine microfluidics with the optical devices described in the previous chapter are detailed. One project involves combining the use of optical emitters with microfluidic injections containing caged neurotransmitters to stimulate neurons is described. The other project involves microfluidic sampling of the extracellular space for neuropeptides which are detected using ring resonator biosensors.

Chapter 6 concludes the thesis, with a discussion of the projects as a whole and the future direction of these projects.

Chapter 2

Theory and Design of Optical Ring Resonator Biosensors for Neural Applications

In this chapter an estimate for the limit of detection of ring resonators due to physical and biological noise will is made. To begin, the responsivity of the effective index of refraction of a rib waveguide to a change in refractive index of the surrounding environment (in this case protein dissolved in water) will be calculated from a simple model of a waveguide. Next, due to the concentrating effect of the antibodies on the surface of this type of sensor, the magnitude of the shift in effective index in the binding region will be estimated using the evanescent field profile of the waveguide. A discussion of the physical noise of the sensor is included which will determine the physical limit of mass detection for a ring resonator biosensor. Following this, a model for the biological noise of a generic affinity biosensor is proposed to explain the discrepancy between the sensitivity described by the physical model and the actual limit of detection of affinity biosensors in actual experiments (i.e., detection of proteins in blood serum or cell lysate). The final portion of this chapter discusses factors relevant to biosensing on probes as well as other modalities described in this thesis, including the physics of flow in microfluidic channels, and the design of optical emitters for neurotransmitter uncaging experiments.

2.1 Waveguide ring resonators

Although there are many biosensing modalities currently used, optical systems are advantageous in that they have a number of advantages over other physical measurement systems. Optical systems operate well in water (unlike mechanical systems which are heavily damped in liquid environments) and are unaffected by the ionic strength of the surrounding solution (as are nanowire biosensors). Among optical sensors, there is a divide between integrated solutions and fluorescence solutions. Due to the constraints of the form factor necessary to accomplish fast biosensing experiments directly on a neural probe, an integrated solution is required. Ring waveguide sensors were chosen over interferometric devices due to a smaller footprint and the superiority of resonance based measurement systems. In this section, the design of waveguide ring resonators is described, and the sensitivity of such a sensor using the chosen materials described previously is made. In addition, practical considerations in the design of ring resonators is discussed.

2.1.1 Waveguide design in silicon nitride films

When designing waveguides for optical ring resonators, the first step in the design is to determine the dimensions of the waveguide. Although it generally applies to circular waveguides, a simple guideline to determine the maximum size of the waveguide is determined by calculating the normalized frequency pararameter, V. V can be found from the following equation:

$$V = 2\pi \frac{a}{\lambda} \sqrt{n_{core}^2 - n_{cladding}^2}, \tag{2.1}$$

where a is the radius of the fiber, λ is the wavelength, and n_{core} and $n_{cladding}$ are the core and cladding refractive indexes, respectively. Although this is calculated for cylindrical waveguides, it provides an upper bound for the size of the waveguides used to maintain single mode operation. V must be less than the first root of the Bessel function, J_0, to only allow single mode propagation. Therefore, rearranging equation 2.1 the core size can be approximated.

$$V = 2.4048 > 2\pi \frac{a}{\lambda} \sqrt{n_{core}^2 - n_{cladding}^2} \tag{2.2}$$

$$a < 2.4048 \frac{\lambda}{2\pi \sqrt{n_{core}^2 - n_{cladding}^2}}. \tag{2.3}$$

Given the material used to create the waveguides is silicon nitride with an index of refraction of 2.0, a central wavelength of the laser of 673nm, and an outer cladding of water (index 1.33), we find

$$a < 2.4048 \frac{673\text{nm}}{2\pi \sqrt{2^2 - 1.33^2}} = 172.5\text{nm}. \tag{2.4}$$

Thus, the width of the waveguides using must be less than 345nm. To maximize quality factors and sensitivity, the largest waveguide width while maintaining single mode operation should be used. For the following calculations a square cross section waveguide of width and height of 300nm will be used as a starting point because the cross sectional area of the square waveguides matches the area of a 345nm circular waveguide.

The critical parameter when describing the properties of optical waveguides is the propagation coefficient of the waveguide, which also defines the effective refractive index of the waveguide (i.e., the speed light travels in the waveguide). This can be solved for using the method described by

Marcatili[68]. Mercatili's method solves Maxwell's equations for a rectangular waveguide and surrounding cladding, while simplifying the problem by ignoring the regions near the corners of the waveguides, as shown in Figure 2.1.1. In the specific case described here, region 1 is the silicon nitride waveguide and regions 2 and 3 are the cladding (water), which are identical because $a = d = 300$nm. The primary assumption is that the mode is well guided, which is a reasonable assumption in this case considering the large index difference between core and cladding and the assumptions used to calculate the waveguide size in equation 2.2.

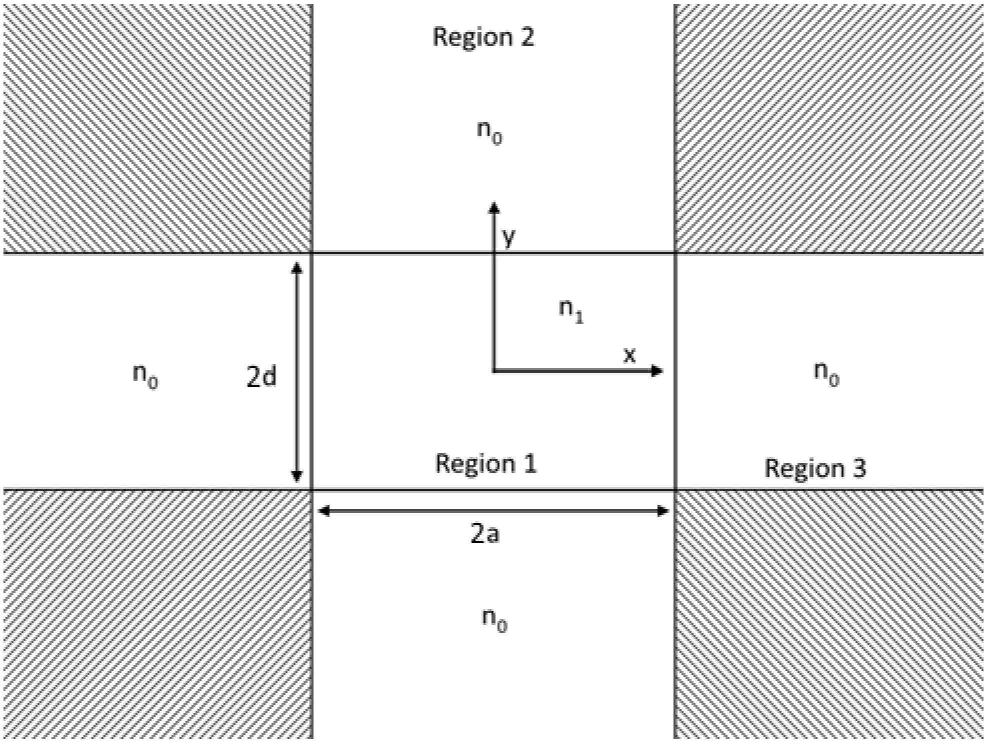

Figure 2.1: Cross sectional digram defining the boundary used in Marcatili's method.

For simplicity, we will assume that the bottom cladding is water as well, with an index of 1.33, instead of silicon oxide which has an index of refraction of 1.45. This will result in a small error, but is still sufficient for basic design of the waveguides. Starting with Maxwell's equations, where

$$\nabla \times E = -\mu_0 \frac{\partial H}{\partial t} \tag{2.5}$$

$$\nabla \times H = \epsilon_0 n^2 \frac{\partial E}{\partial t}. \tag{2.6}$$

Solving these equations, we will solve for the E_{00}^x mode with E along the x-axis, thus assuming $H_x = 0$. Assuming plane wave propagation,

$$E = E(x,y)\exp[j(\omega t - \beta z)] \tag{2.7}$$

$$H = H(x,y)\exp[j(\omega t - \beta z)]. \tag{2.8}$$

Applying the above equations to equations 2.5 and 2.6, and assuming $H_x = 0$, the above simplifies to

$$\frac{\partial^2 H_y}{\partial x^2} + \frac{\partial^2 H_y}{\partial y^2} + (k^2 n^2 - \beta^2) H_y = 0, \tag{2.9}$$

where

$$E_x = \frac{\omega \mu_0}{\beta} H_y + \frac{1}{\omega \epsilon_0 n^2 \beta} \frac{\partial^2 H_y}{\partial x^2} \tag{2.10}$$

$$E_y = \frac{1}{\omega \epsilon_0 n^2 \beta} \frac{\partial^2 H_y}{\partial x \partial y} \tag{2.11}$$

$$E_z = \frac{-j}{\omega \epsilon_0 n^2} \frac{\partial H_y}{\partial x} \tag{2.12}$$

$$H_z = \frac{-j}{\beta} \frac{\partial H_y}{\partial y}. \tag{2.13}$$

Equation 2.9 is the Helmholtz equation with eigenvalues β. Thus, the following solution may be assumed for the envelope of H_y:

$$H_y = \begin{cases} Acos(k_x x)cos(k_y y) & \text{(region 1)} \\ Acos(k_x x)cos(k_y a)e^{-\gamma_y (y-a)} & \text{(region 2)} \\ Acos(k_x a)cos(k_y y)e^{-\gamma_x (x-a)} & \text{(region 3).} \end{cases} \tag{2.14}$$

This solution to the Helmholtz equation describes the envelop function of the light propagating in the waveguide, with the majority of the mode concentrated within the waveguide and an exponentially decaying evanescent field outside of the waveguide. The transverse wavenumbers that result from this solution are

$$\begin{cases} -k_x^2 - k_y^2 + k^2 n_1^2 - \beta^2 = 0 & \text{(region 1)} \\ \gamma_x^2 - k_y^2 + k^2 n_0^2 - \beta^2 = 0 & \text{(region 2)} \\ -k_x^2 + \gamma_y^2 + k^2 n_0^2 - \beta^2 = 0 & \text{(region 3).} \end{cases} \tag{2.15}$$

Applying the boundary conditions where $E_z \propto (1/n^2)\partial H_y/\partial x$ is continuous at $x = a$ and $H_z \propto \partial H_y/\partial y$, the following transcendental equations must be satisfied:

$$k_x a = \tan^{-1}\left(\frac{n_1^2 \gamma_x}{n_0^2 k_x}\right) \tag{2.16}$$

$$k_y a = \tan^{-1}\left(\frac{\gamma_y}{k_y}\right), \tag{2.17}$$

where γ_x and γ_y can be found from equation 2.15. These equations must be solved numerically, with the ultimate goal of solving for the propagation coefficient, β. The solution to the above transcendental equations for β is shown in Figure 2.1.1 for the waveguide parameters $\lambda \in (500, 1000)$nm, $a = 300$nm, and $n_0 = 2$, $n_1 = 1.33$. At the center wavelength of the laser used, 673nm, the propagation constant is $\beta = 15.1761$. The effective refractive index can be calculated as,

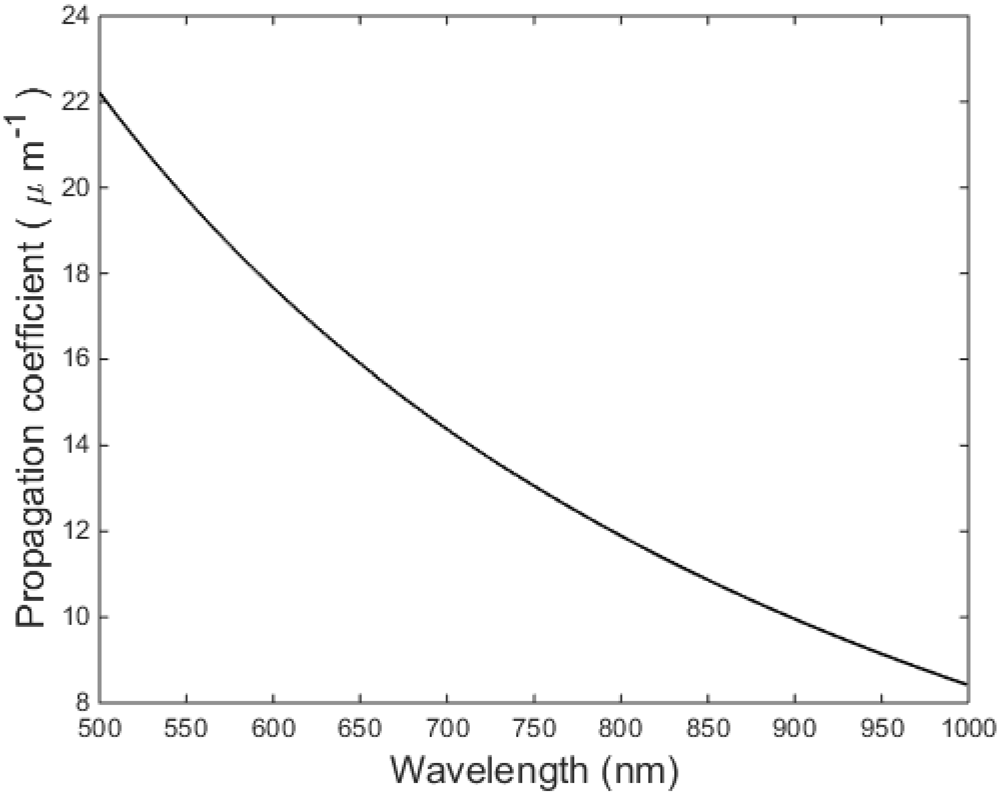

Figure 2.2: Dispersion relationship of a 300nm square silicon nitride waveguide in water.

$$n_{\text{eff}} = \frac{\beta}{k} = \frac{\beta\lambda}{2\pi}. \tag{2.18}$$

Therefore, for the above waveguide the effective refractive index is $n_{\text{eff}} = 1.63$. Approximating the waveguide propagation coefficient is an important first step, but these calculations will also prove useful in the following sections.

2.1.2 Resonance Condition of ring resonators

Ring resonators are simple resonant structures where light circulates in a circular waveguide cavity, as shown in Figure 2.1.2. Due to interference between incoming light from the bus waveguide and light making a full round trip of the ring, resonance will only occur when there is constructive

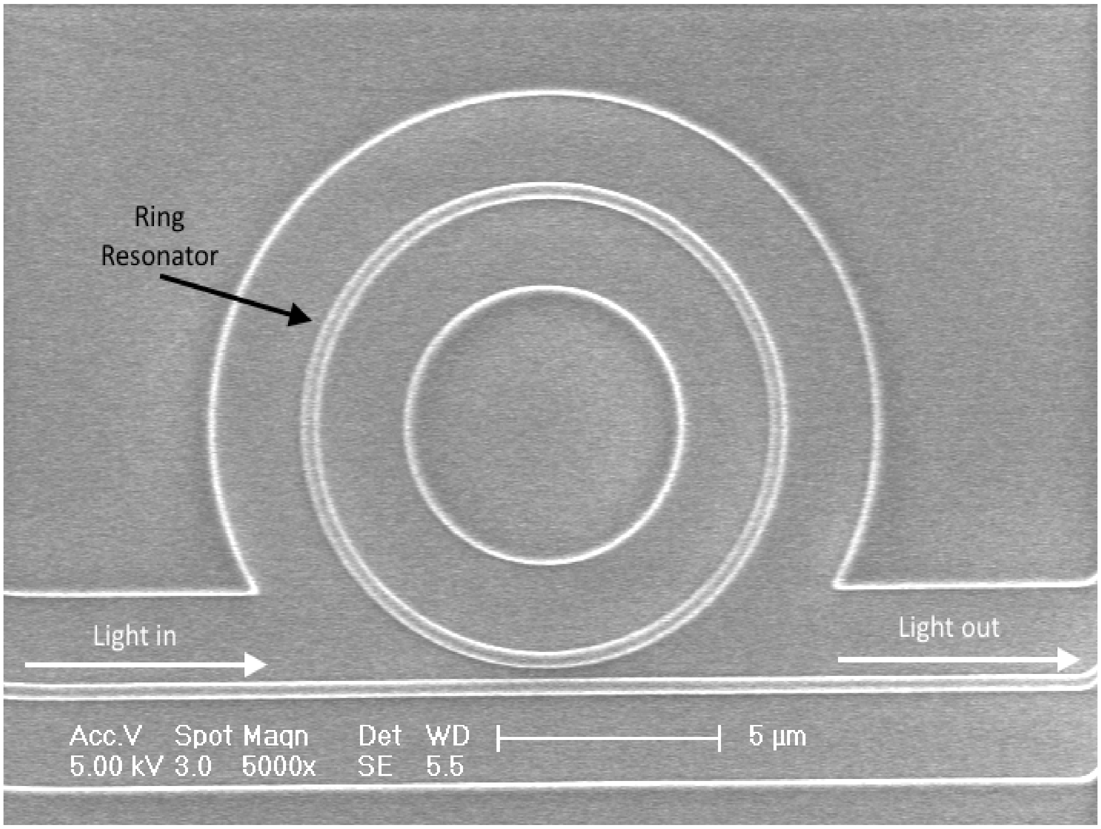

Figure 2.3: Scanning electron microgram of a planar ring resonator structure.

interference between these two waves. If there is significant destructive interference, the circulating light will be quenched by the incoming light, and thus the intensity in the ring will be insignificant. Given that light gains phase while propagating through the waveguide equal to $2\pi n_{\text{eff}} d/\lambda$, where d is the distance traveled along the waveguide, and the round trip phase must match the incoming phase (i.e., the shift is an integer multiple of the incoming phase); therefore

$$2\pi r n_{\text{eff}} = m\lambda_m, \tag{2.19}$$

where r is the radius of the resonant cavity, λ_m is the resonant wavelength, and $m \in \mathbb{Z} > 0$ is mode number. This can be taken a step further to define the distance between the resonance wavelength λ_n and the next resonance, λ_{n+1}. This distance is called the free spectral range, and can be calculated as

$$\Delta\lambda = -\frac{2\pi}{2\pi R}\left(\frac{\partial\lambda}{\partial\beta}\right) = -\frac{1}{R}\frac{\partial\lambda}{\partial\beta} \approx \frac{\lambda^2}{2\pi R n_{\text{eff}}}. \tag{2.20}$$

Thus, to calculate the free spectral range accurately, the slope of the dispersion relationship must be calculated. This can be done numerically given the result of the dispersion relationship solved for in section 2.1.1. Around the central wavelength of 673nm, this value is $\partial\lambda/\partial\beta = -3.26 \times 10^{-2}\mu\text{m}^2$.

2.1.3 Coupling light into ring resonators

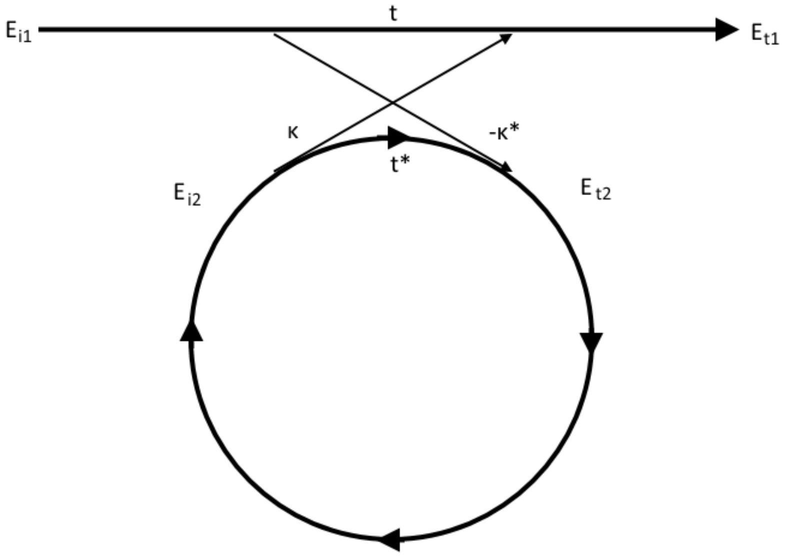

Figure 2.4: Diagram describing the coupling coefficients used in the coupling model by Yariv.

Coupling light into ring resonators is typically accomplished by fabricating a waveguide (i.e., a bus waveguide) very close to the ring, allowing the evanescent field from the bus to ‘leak’ into the ring, thus exciting a mode within the ring. This configuration can be seen in Figure 2.1.2. Although closed form calculations are possible to understand the coupling gap (this is the purview of coupled mode theory), it is difficult to accurately make conclusions based on this technique without eventually using complicated simulations, and thus is beyond the scope of this thesis.

A simplified mathematical model used to describe the properties of light within and the output properties of ring resonator filters was first described by Yariv in [135, 136] and described in a more modern context in [87]. This model utilizes practical parameters, described diagrammatically in Figure 2.1.3, which can be directly measured from ring resonator systems and used to understand and design ring resonator systems. The energy states of the system are described by four parameters: E_{i1} is the input light electric field (herein assumed to be 1 for simplicity), E_{t1} is the output of the ring resonator system, E_{i2} is the electric field circulating in the ring before the coupler, and E_{t2} is the electric field in the ring after the coupler. There are also two parameters which describe the coupling, t is the fraction of energy which is passed through the coupler and κ is the fraction of energy which crosses the coupler to the ring. Since the couplings to and from the ring are reciprocal, therefore

$$|\kappa^2| + |t^2| = 1. \tag{2.21}$$

Since the ring is lossy, the relationship between the energy in the ring immediately after the coupling region (E_{t2}) and the energy after a round trip are related by the following relationship:

$$E_{i2} = \alpha \exp[j\theta] E_{t2}, \tag{2.22}$$

where α is the round trip loss of the ring resonator and θ is the phase shift from a single round trip of the ring. Determining α will be discussed in the following sections. The phase shift can be calculated geometrically from the dimensions of the ring and the properties of the waveguide, where the wavelength of the light in the ring is kn_{eff}, therefore the round trip phase shift is

$$\theta = kn_{\text{eff}}(2\pi R) = \frac{2\pi}{\lambda} n_{\text{eff}}(2\pi R) = 4\pi^2 n_{\text{eff}} \frac{R}{\lambda}. \tag{2.23}$$

Therefore, the equations describing the relationship diagrammed in Figure 2.1.3 can be written as

$$E_{t1} = tE_{i1} + \kappa E_{i2} = t + \kappa E_{i2} \tag{2.24}$$

$$E_{t2} = -k^* E_{i1} + t^* E_{i2} = -k^* + t^* E + i2. \tag{2.25}$$

Along with equation 2.22, This system of equations can be simplified to

$$E_{t1} = \frac{-\alpha + t\exp(-j\theta)}{-\alpha t^* + \exp(-j\theta)} \tag{2.26}$$

$$E_{i2} = \frac{-\alpha\kappa^*}{-\alpha t^* + \exp(-j\theta)} \tag{2.27}$$

$$E_{t2} = \frac{-\kappa^*}{1 - \alpha t^* \exp(j\theta)}. \tag{2.28}$$

The output intensity of the ring resonator system is simply the square of E_{t1}:

$$P_{t1} = \frac{\alpha^2 + |t|^2 - 2\alpha|t|\cos(\theta + \phi_t)}{1 + \alpha^2|t|^2 - 2\alpha|t|\cos(\theta + \phi_t)}, \tag{2.29}$$

where ϕ_t is the phase shift induced by the coupler. On resonance ($\theta + \phi_t = 2\pi m$), the output of the ring resonator system will be

$$P_{t1} = |E_{t1}|^2 = \frac{(\alpha - |t|)^2}{(1 - \alpha|t|)^2}. \tag{2.30}$$

Thus, if α, the loss in the ring resonator, equals $|t|$, the amount of light coupled into the ring from the bus, then the output power is zero. This is the critical coupling condition, where all of the energy from the incoming light is concentrated in the resonator, and occurs because the light leaving the ring resonator through the coupler is equal in magnitude and 180 degrees out of phase from the incoming light. The optical power stored in the ring itself can similarly be calculated as

$$P_{i2} = \frac{\alpha^2(1 - |t|^2)}{1 + \alpha^2|t|^2 - 2\alpha|t|\cos(\theta + \phi_t)}. \tag{2.31}$$

On resonance,

$$P_{i2} = |E_{i2}|^2 = \frac{\alpha^2(1-|t|^2)}{(1-\alpha|t|)^2}. \tag{2.32}$$

2.1.4 Determination of quality factors in ring resonators

The primary figure of merit used in describing resonators is the quality factor of the resonator. The quality factor is defined as

$$Q = 2\pi\frac{\text{Energy stored}}{\text{Energy dissipated per cycle}} = \frac{\lambda}{2\delta\lambda}, \tag{2.33}$$

where λ is the resonant wavelength and $2\delta\lambda$ is the width of the resonance peak at half maximum intensity. The Quality factor can be calculate using this definition, and equations 2.31 and 2.32, first by finding the phase shift θ at which the ring power will be half of the maximum (assuming the coupler phase shift is negligible),

$$\frac{1}{2}\frac{\alpha^2(1-|t|^2)}{(1-\alpha|t|)^2} = \frac{\alpha^2(1-|t|^2)}{1+\alpha^2|t|^2-2\alpha|t|\cos(\theta)}. \tag{2.34}$$

Since the numerators of both terms are the same, only the denominators need to be compared. Thus,

$$2(1-\alpha|t|)^2 = 1+\alpha^2|t|^2-2\alpha|t|\cos(\theta). \tag{2.35}$$

Solving for $\cos(\theta)$,

$$\cos(\theta) = \frac{(\alpha|t|)^2-4(\alpha|t|)+1}{2(\alpha|t|)}. \tag{2.36}$$

Applying the small angle formula and solving further for θ,

$$\theta = \left(2+\frac{(\alpha|t|)^2-4(\alpha|t|)+1}{\alpha|t|}\right)^{1/2}. \tag{2.37}$$

Applying equation 2.23, the the full width half maximum of the resonance can be found

$$2\delta\lambda = \frac{\lambda}{4\pi^2 R n_{\text{eff}}}2\theta = \frac{\lambda^2}{2\pi^2 R n_{\text{eff}}}\left(2+\frac{(\alpha|t|)^2-4(\alpha|t|)+1}{\alpha|t|}\right)^{1/2}. \tag{2.38}$$

The quality factor, Q, follows

$$Q = \frac{\lambda}{2\delta\lambda} = \frac{2\pi^2 R n_{\text{eff}}}{\lambda}\left(2+\frac{(\alpha|t|)^2-4(\alpha|t|)+1}{\alpha|t|}\right)^{-1/2}. \tag{2.39}$$

From this equation, the major controllable factors which affect the quality factor of the resonator are α and R. The relationship between these two parameters and other factors which affect *alpha* are discussed in the following sections.

2.1.4.1 Bending loss

When guided modes change direction (as in a ring resonator), there are losses due to mode conversion and radiation of the evanescent field tail along the lagging edge of the curve. An early closed form theory was proposed by Marcatili and Miller in [67]. This method proposes an exponential relationship between bending radius and loss. Although this has been shown to be mostly true, this method is still inconsistent with experimental results by at least 60%. Thus, to better understand quality factors versus bending radius, simulations must be used. In this case, MEEP[80] was used to simulate an ideal circular waveguide (in this case 300nm by 300nm silicon nitride waveguides with water cladding on 3 sides, and silicon oxide cladding on the bottom). The simulation code can be found in appendix ZZZ, with the results shown in Figure 2.1.4.1.

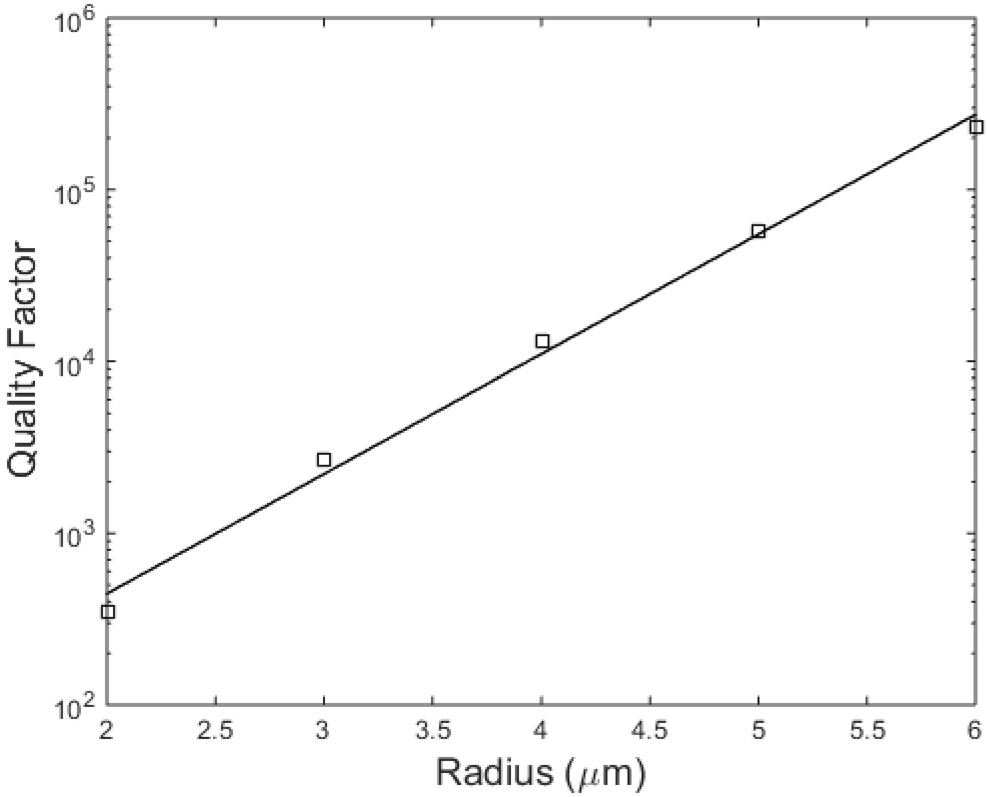

Figure 2.5: Simulated quality factor versus bending radius.

The results of the simulation do show an approximately exponential relationship between Q and R, and provide maximum possible quality factors for a given ring radius. Note that these are 'unloaded' quality factors (i.e., not coupled to a bus waveguide), loaded Q's will always be lower due to coupling losses. To ensure that the quality factors of the devices are not limited by the bending radius of the device, the ring radius needs to be greater than or equal to approximately 5.5μm, which would place the $Q > 1 \times 10^5$.

2.1.5 Determining coupling parameters from data

Returning to the model described in section 2.1.4, the values of α and $|t|$ can be determined using data. The technique is described by [REF], and relies on two easy to measure parameters, the finesse of the resonance, $\mathscr{F}$, and the extinction coefficient of the resonance, $\mathscr{E}$. These parameters are defined as,

$$\mathscr{F} = \frac{\text{Free Spectral Range}}{ResonanceFWHM} = \frac{\Delta\lambda}{2\delta\lambda} \tag{2.40}$$

$$\mathscr{E} = \frac{\text{Max Transmission}}{\text{Resonance Transmission}} = \frac{P_{t1}(\theta = \pi)}{P_{t1}(\theta = 0)}. \tag{2.41}$$

These values can be expressed in terms of α and $|t|$ as

$$\cos(\pi/\mathscr{F}) = \frac{2\alpha|t|}{1 + \alpha^2|t|^2} \tag{2.42}$$

$$\mathscr{E} = \left(\frac{(\alpha + |t|)(1 - \alpha|t|])}{(\alpha - |t|)(1 + \alpha|t|)}\right)^2. \tag{2.43}$$

Something worth noticing is that in equation 2.39, all the terms are expressed as the product of α and $|t|$. This leads to an ambiguity in that although it will be possible to solve for these parameters, it will be impossible to determine which of the results correspond to α or $|t|$. Thus, the experiment used to determine these parameters will require a single parameter to be changed to change the coupling coefficient, $|t|$, to determine which parameter corresponds to the coupling term. Examples of usable parameters to sweep are the resonant wavelength or the coupling gap, both of which shouldn't change α significantly.

The above equations can be solved for α and $|t|$ using the following parameters:

$$A = \frac{\cos(\pi\mathscr{F})}{1 + \sin(\pi\mathscr{F})} \tag{2.44}$$

$$B = 1 - \left(\frac{1 - \cos(\pi\mathscr{F})}{1 + \cos(\pi\mathscr{F})}\right)\frac{1}{\mathscr{E}}. \tag{2.45}$$

These parameters can be used to solve the quadratic resulting from system of equations as

$$(\alpha, |t|) = \left(\frac{A}{B}\right)^{1/2} \pm \left(\frac{A}{B} - A\right)^{1/2}. \tag{2.46}$$

This relationship will be useful in the understanding of the impact of etching to the unloaded quality factor of real resonators in Chapter 4.

2.2 Waveguide mass responsivity

To understand the responsivity of the sensor, the primary concern is the how the effective refractive index of the ring varies with the refractive index of the surrounding solution in bulk. However, only understanding this relationship is insufficient as only a small region of the bulk solution will be concentrated with the target molecule, as the capture agents are coated on the surface of the resonator. Thus, the relationship between the evanescent field profile and sensitivity of the ring must be determined. The relationship between protein concentration and refractive index is also important to this calculation and must be determined. Finally, the responsivity can be calculated based on the resonance condition of the ring, and the detectable shift in the resonance by the measurement apparatus. This section will detail all of these steps, ultimately determining a order-of-magnitude estimate for the mass sensitivity of a ring resonator biosensor.

2.2.1 Waveguide effective index responsivity to bulk refractive index shift

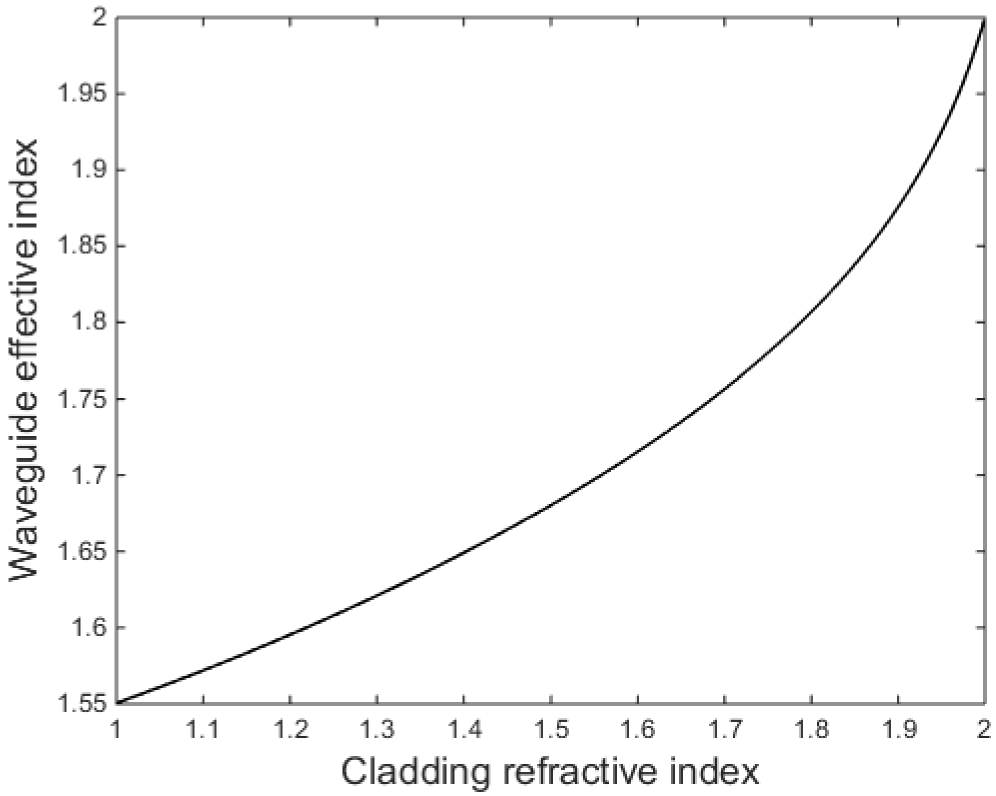

Figure 2.6: Waveguide effective index vs. bulk cladding refractive index at 669nm.

Using the physics discussed in 2.1.1 for the dispersion relationship for the waveguides we can instead solve for the effective refractive index for varying cladding refractive index numerically. The result of this analysis is shown in Figure 2.2.1. As is expected, the effective index of the increases with the cladding index. However when considering the sensitivity of microring resonator devices, the most important factor is the responsivity, R, which in this case is simply

$$R = \frac{\partial n_{\text{eff}}}{\partial n_{\text{cl}}}. \tag{2.47}$$

Using the data in Figure 2.2.1 the derivative can be taken numerically, as is shown in Figure 2.2.1. Thus, the responsivity of silicon nitride waveguides to bulk changes in the cladding index can be read off of this chart. Around 669nm and n_{cl} the responsivity is $R = .24\frac{\text{RIU}}{\text{RIU}}$.

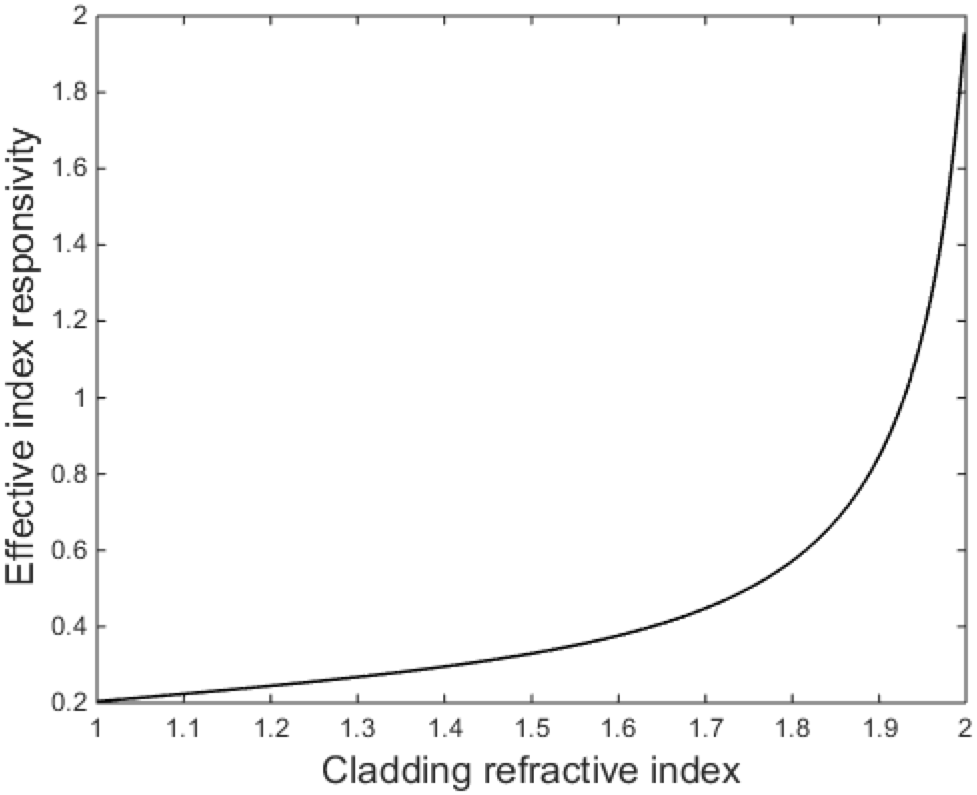

Figure 2.7: Partial derivative of waveguide effective index against bulk cladding index.

2.2.2 Waveguide effective index responsivity to surface refractive index shift

The responsivity discussed in the previous section is an unrealistic representation of the system. In reality, the surface binding doesn't diffusely occur over the volume defined by the evanescent field of the resonator but in a concentrated band near the surface of the device. Furthermore, binding will only occur on three of four sides of the waveguide. The sensitivity of the waveguide

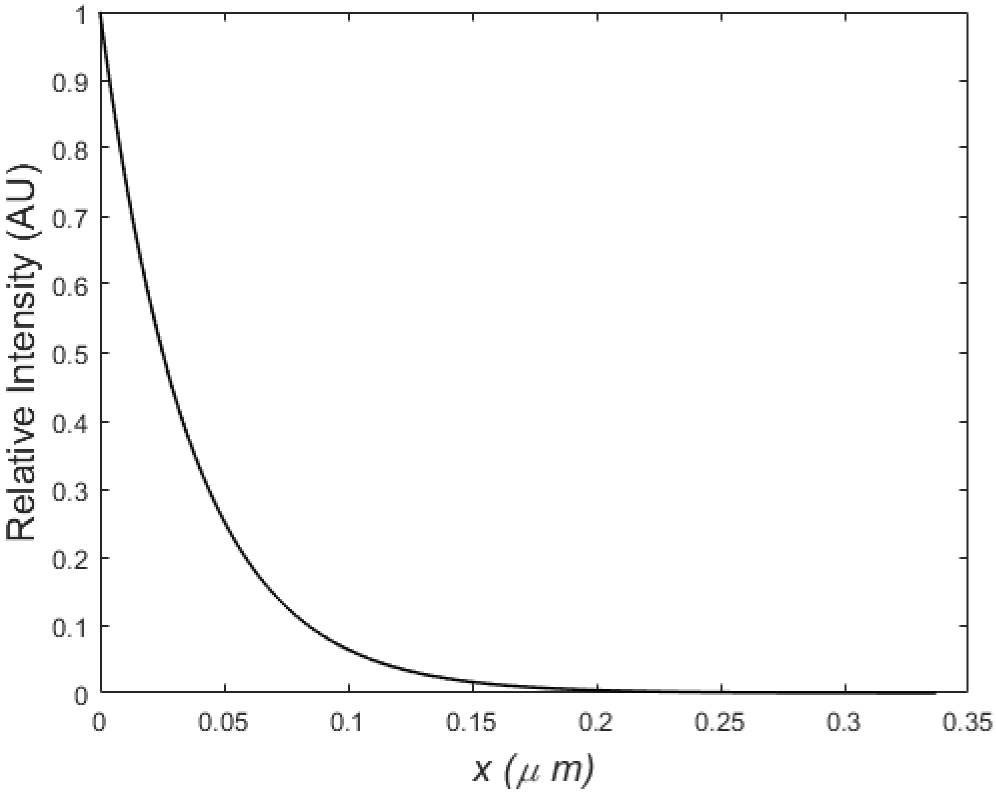

Figure 2.8: Relative intensity of light as a function of distance from the edge of the waveguide.

to external binding of sequential protein monolayers appears to be proportional to the intensity of the evanescent field, as shown in [64]. Thus, by defining a small region of potential binding, the responsivity calculated in the previous section can be corrected for the localization of protein binding to the surface.

The first assumption which must be approximate length of the cross-linking chemistry between the surface of the waveguide and the capture agents. A typical chemical protocol would be first to vapor deposit a monolayer of an sulfhydryl-functionalized silane on the surface of the waveguide. This layer adds 2-3 atoms of thickness, or about 0.5nm of distance. Next, a heterobifunctional crosslinker is added with a short spacer arm with a length of approximately 1nm is used to couple the sulfhydryl-functionalized surface with free amines on the capture agent. Finally, the length of the capture agents vary depending on their type (antibodies $\approx$ 10 nm, antibody fragments $\approx$ 5nm). Thus, it is reasonable to assume that the binding occurs will occur approximately 6 to 12nm from the surface. Given this volume, the concentration of bound analyte can be approximated in a shell 6 to 12 nm from the surface.

Next, the field profile of the evanescent field must be determined. Solving the Helmholtz equation for the field outside of the waveguide, the intensity profile surrounding the waveguide will be a decaying exponential function,

$$I(x) \propto H^2(x) \propto \left(cos(k_x a)e^{-\gamma_x(x-a)}\right)^2, \tag{2.48}$$

where in this case, γ_x and $k_x a$ can be solved from region two in the system of equations 2.14. For 669nm light, $\gamma_x \approx 13.75\mu m^{-1}$ and $kx_a \approx 7.87\mu m^{-1}$. The resulting field profile is shown in Figure 2.2.2. From this, to determine how much of the field interacts with the analytes, the intensity of the light must be integrated in the 6 to 12nm shell and compared to the total field intensity. Thus, the fraction of power in the shell can be calculated by

$$\frac{I_{a+6nm \to a+12nm}}{I_{a\to\infty}} = \frac{\int_{a+6nm}^{a+12nm} e^{-\gamma_x(x-a)}dx}{\int_a^\infty e^{-\gamma_x(x-a)}dx} = 0.13. \tag{2.49}$$

In other words, 13 percent of the intensity is concentrated in the shell surrounding the waveguide between 6 and 12nm. Therefore, the responsivity calculated in section 2.2 must be modified by the above figure. Furthermore, since binding will only occur on three sides of the device, the responsivity must also be reduced by approximately 75%. Therefore, the responsivity to changes in refractive index in the region where binding is expected to occur is

$$R_{bound} = R_{bulk} \times \frac{I_{a+6nm \to a+12nm}}{I_{a\to\infty}} = 0.24 \times .13 \times \frac{3}{4} = .023\frac{\text{RIU}}{\text{RIU}}. \tag{2.50}$$

2.2.3 Protein concentration and refractive index

Determining the incremental change in index of refraction of a protein solution versus concentration is a complex problem which is well beyond the scope of this thesis. However, estimates were made by Zhao et al. in reference [141]. In this work amino acid refractive index data was combined with volumetric estimates of protein density via a database of protein structures to determine a estimated distribution of the effective refractive indexes of all proteins. The resulting distribution of dn/dc for human proteins was found to be $0.1899ml/g$ with a standard deviation of $0.0030ml/g$. Given that this distribution is relatively compact, using the mean value as an estimate for the incremental change in index of refraction with changing protein concentration should be a good estimate for any given protein.

2.2.4 Responsivity of resonance wavelength to surface binding

In the previous sections the responsivity of the refractive index of a waveguide with respect to the concentration of bound protein. The final step in determining the responsivity of an optical ring resonator biosensor is to incorporate the resonance condition into the responsivity calculation. Recall from the discussion in section 2.1.2, resonance occurs when the light coupled into the resonator interferes constructively with the light making a trip around the ring. Thus, the resulting resonance condition dictates that

$$\lambda_m = \frac{2\pi r n_{\text{eff}}}{m}, \tag{2.51}$$

where m is the resonance order, λ_m is the resonance wavelength of order m, n_{eff} is the effective refractive index, and r is the radius of the ring resonator. Given a specific ring resonator, r and m will remain constant, and therefore any shift in λ_m is due to a change in n_{eff}. Therefore, the wavelength responsivity of a ring resonator to a changing amount of bound protein can be determined via differentiation, where

$$\frac{\partial \lambda}{\partial n_{cl}} = \frac{2\pi r}{m} R_{bound}. \tag{2.52}$$

Given the above responsivity, the physical limit of detection for such a sensor can be determined as

$$\text{LOD} = \left(\frac{\partial \lambda}{\partial n_{cl}}\right)^{-1} \cdot \delta\lambda_{min}, \tag{2.53}$$

where LOD is the limit of detection, and $\delta\lambda_{min}$ is the smallest change in wavelength measurable by the experimental setup.

As a demonstration, it is possible to calculate the limit of detection for a common protein (in this case TNF-α) in a pure solution. Assuming that $\delta\lambda_{min}$ is the width of a resonance (e.g., 16.75pm for a resonator with Q = 40,000 at a central wavelength of 667nm), the limit of detection is calculated as

$$\text{LOD} = \left(\frac{2\pi r}{m} R_{\text{bound}}\right)^{-1} \cdot \delta\lambda_{min} = \left(\frac{2\pi 5.6625\mu\text{m}}{88}.031\right)^{-1} \cdot 16.5\text{nm} = 1.3 \times 10^{-3}\text{RIU}. \tag{2.54}$$

Next the limit of detection for mass must be calculated using the relationship protein concentration and index of refraction. Given that this relationship is .19 RIU per (mg per mL), and the volume of binding around the ring resonator is approximately 4.6×10^{-13}mL, the mass LOD can be calculated as,

$$\text{LOD}_{\text{mass}} = \text{LOD} \times \left(.19\frac{\text{RIU} \cdot \text{mL}}{\text{mg}}\right)^{-1} \times 4.6 \times 10^{-13}\text{mL} = 1.9\text{MDa}. \tag{2.55}$$

Thus, it is possible to detect the binding of 1.9MDa to the ring resonator surface. This corresponds to 100 molecules binding to the surface of the resonator, or a solution concentration of 10fM at equilibrium. This is well beyond the limit of detection of real sensors[6], which often show sensitivities in the nanomolar range. Thus, there must be another process which is dominating the sensitivity of these devices in experimental solutions.

2.3 Sources of biological noise

In addition to the physical noise described in the previous section, there will also be noise induced by the biological environment of the sensor. This biological noise is the result of any proteins in the sample which are not the analyte binding non-specifically to the sensor surface. The claims of the physical noise floor of many biosensors is many orders of magnitude lower than the observed noise floor for making actual measurements of proteins in solutions of high protein concentration. Although this result is widely known, there isn't a comprehensive theory of the biological noise floor for affinity based biosensors in the literature. In this section, a set of "rules of thumb" for the minimum detectable protein concentration in a solution with high protein content compared to the concentration of the analyte is discussed. This theory applies to equilibrium and initial rate measurements of label free assays, and equilibrium measurements of sandwich assays.

2.3.1 Pure solutions

To begin, the binding of molecules of interest (herein called the substrate, "S") to a protein with which the substrate has an affinity (an antibody or a receptor, "R") can most simply be described through a simple first order reaction system. In this reaction system, the concentration of the substrate ($[S]$), the concentration of the receptor ($[R]$) and the concentration of the complex of the substrate and the receptor ($[RS]$) is related as follows,

$$R + S \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} RS \tag{2.56}$$

where k_{on} is the rate of the forward reaction (units of $Lmol^{-1}s^{-1}$) and k_{off} is the rate of the reverse reaction (units of s^{-1}). Thus, the differential equation describing the concentration of the receptor-substrate complex is

$$\frac{d[RS]}{dt} = k_{on}[S][R] - k_{off}[RS]. \tag{2.57}$$

Next, an assumption about the relative concentrations of the receptor and the substrate must be made to simplify the above differential equation. Simply, the total concentration of the receptor in relation to the total amount of receptor, R_{max} is

$$[R] = R_{max} - [RS]. \tag{2.58}$$

Substituting the above relationship into the differential equation, the differential equation can be simplified to a first order ODE,

$$\begin{aligned}\frac{d[RS]}{dt} &= k_{on}[S](R_{max} - [RS]) - k_{off}[RS] &(2.59)\\ &= k_{on}[S]R_{max} - (k_{on}[S] + k_{off})[RS]. &(2.60)\end{aligned}$$

Solving this linear first order ordinary differential equation, the solution is quite simply expressed as a percentage of bound receptors,

$$\frac{[RS]}{R_{max}} = \frac{k_{on}[S]}{k_{on}[S] + k_{off}} \left(1 - \exp\left[-(k_{on}[S] + k_{off})t\right]\right). \tag{2.61}$$

As $t \to \infty$, the system approaches an equilibrium of

$$\frac{[RS]}{R_{max}} = \frac{k_{on}[S]}{k_{on}[S] + k_{off}} = \frac{K_a[S]}{1 + K_a[S]} \text{ s.t. } K_a = \frac{k_{on}}{k_{off}}, \tag{2.62}$$

where K_a is the association equilibrium constant between the receptor and the substrate. Thus, given the concentration of S and the association constant associated with the reaction, the percentage of the receptors bound to the substrate can be calculated in a pure solution using the above equation. This situation, however, is unrealistic for any real biosensing experiment since there will always be other proteins in the solution which can also bind to the surface, albeit in a non-specific manner.

2.3.2 Mixed solutions and non-specific binding

Taking the next step in making the binding reactions more realistic non-specific binding of proteins must be considered. A simplified model of such a system is represented as

$$\begin{aligned} RI + S \underset{k^I_{off}}{\overset{k^I_{on}}{\rightleftharpoons}} R + S + I \underset{k^S_{off}}{\overset{k^S_{on}}{\rightleftharpoons}} RS + I &&(2.63)\\ R_{max} = [RI] + [RS] + [R], &&(2.64)\end{aligned}$$

where I is the inhibiting compound, k^I_{on} and k^I_{off} are the on and off rates for the inhibitor-receptor interaction, and k^S_{on} and k^S_{off} are the on and off rates for the substrate-receptor interaction. This simplified model makes a number of assumptions, however the utility of this simple model will be shown in the following sections. These assumptions include:

- although the exact nature of the inhibitor binding to the surface is unknown, there will still be steric hindrance of the receptor-substrate interaction,

- the receptor-inhibitor interaction has a relatively well defined equilibrium coefficient with the surface,

- all reactions follow first order kinetics,
- binding of substrate to substrate, substrate to inhibitor and inhibitor to inhibitor are small compared to substrate to receptor and inhibitor to receptor.

Based on these assumptions, a model of biosensing in dirty solutions for both label free detection and sandwich assays is developed which has good agreement with experimental data in the literature.

2.3.2.1 Equilibrium analysis

Starting with the model shown in 2.63, the first step is to write the system of differential equations for this system:

$$\frac{d[R]}{dt} = 0 = -k_{on}^{I}[R][I] + k_{off}^{I}[RI] - k_{on}^{S}[R][S] + k_{off}^{S}[RS] \tag{2.65}$$

$$\frac{d[RI]}{dt} = 0 = k_{on}^{I}[R][I] - k_{off}^{I}[RI] \tag{2.66}$$

$$\frac{d[RS]}{dt} = 0 = k_{on}^{S}[R][S] - k_{off}^{S}[RS]. \tag{2.67}$$

Each of these differential equations are zero, by definition, because the analysis is being done for the equilibrium state. The next operation should be done on both equations 2.66 and 2.67, but the operation will only be demonstrated for one of them.

Rearranging 2.67,

$$[R] = \frac{k_{off}^{S}[RS]}{k_{on}^{S}[S]} = \frac{[RS]}{K^{S}[S]}, \tag{2.68}$$

where K^S is the association constant for the receptor-substrate reaction. Substituting the above relationship for $[R]$ into 2.66,

$$0 = \frac{K_{on}^{I}[I][RS]}{K^{S}[S]} - k_{off}^{I}[RI] \tag{2.69}$$

$$[RI] = \left(\frac{k_{on}^{I}}{k_{off}^{I}}\right)\frac{[I][RS]}{K^{S}[S]} = \frac{K^{I}[I][RS]}{K^{S}[S]}, \tag{2.70}$$

where K^I is the association constant for the receptor-inhibitor reaction. Substituting the above and equation 2.68 into equation 2.64 and simplifying,

$$R_{max} = [RS]\left(\frac{K^{I}[I]}{K^{S}[S]} + 1 + \frac{1}{K^{S}[S]}\right) \tag{2.71}$$

$$= [RS]\left(\frac{K^{I}[I]}{K^{S}[S]} + \frac{K^{S}[S]}{K^{S}[S]} + \frac{1}{K^{S}[S]}\right) \tag{2.72}$$

$$\frac{[RS]}{R_{max}} = \frac{K^{S}[S]}{1 + K^{I}[I] + K^{S}[S]}. \tag{2.73}$$

Note that $[RS]/R_{max}$ is the percentage of the substrate bound to the receptor at equilibrium.

Likewise for $[RI]$,

$$\frac{[RI]}{R_{max}} = \frac{K^I[I]}{1 + K^I[I] + K^S[S]}. \tag{2.74}$$

When doing an experiment in a dirty solution with a high concentration of proteins able to non-specifically bind to the biosensor surface, the measurement will only be accurate when the concentration of the target protein is high enough to guarantee that it can out compete the other, non-specific interactions. Thus, the minimum amount of binding can be expressed as a ratio of the specific to non-specific binding,

$$\frac{\frac{[RS]}{R_{max}}}{\frac{[RI]}{R_{max}}} = \frac{\frac{K^S[S]}{1+K^I[I]+K^S[S]}.}{\frac{K^I[I]}{1+K^I[I]+K^S[S]}} = \frac{K^S[S]}{K^I[I]} > \alpha, \tag{2.75}$$

where α is the desired minimum ratio of specific to non-specific binding. If α is 5, then the safety margin between the specific and non-specific binding is a factor of 5. Rearranging this equation, the relationship which defines the minimum biochemical specificity for a label-free biosensor at equilibrium,

$$[S] > \alpha \frac{K^I}{K^S}[I]. \tag{2.76}$$

This is the critical result of the analysis, providing a rule of thumb by which the minimum sensitivity of a label free biosensor. Given the affinity coefficients or the specific and non-specific binding to the surface and the concentration of the non-target proteins in the solution, it is possible to estimate the minimum target protein concentration that can be measured reliably.

A realistic example of this would be the detection of TNF-α in blood serum. The relevant values from the literature,

$$\begin{aligned} K^S &= 5.25 \times 10^{10}\text{M}^{-1} \\ K^I &\approx 1 \times 10^{5}\text{M}^{-1} \\ [I] &= 600 \times 10^{-6}\text{M}. \end{aligned}$$

Thus, requiring the threshold of specific binding to be 10 times greater than non-specific binding to have confidence in the experiment,

$$[S] > 10\frac{1 \times 10^5\text{M}^{-1}}{5.25 \times 10^{10}\text{M}^{-1}}\left(600 \times 10^{-6}\text{M}\right) \tag{2.77}$$

$$[S] > 11\text{nM}. \tag{2.78}$$

Measurement results in the low nM concentration range for a label free assay agree well with the literature when measuring in blood serum.

2.3.2.2 Initial reaction rates

Another method by which label free biosensor measurements are made is to measure the initial rate of the reaction instead of watching the time response until equilibrium is reached on the surface. This is an attractive method because it allows measurements to be taken much more quickly than an equilibrium measurement. As will be demonstrated, depending on the kinetics of the reactions this method can be better or worse than the equilibrium method described previously.

The analysis begins in the same way as the previous section with the differential equations governing the reaction system; however instead of the rates of change being zero, the concentrations $[RI] = [RS] = 0$, and $[R] = R_{max}$. Thus,

$$\frac{d[R]}{dt} = -k^I_{on}[R][I] + k^I_{off}[RI] - k^S_{on}[R][S] + k^S_{off}[RS] = -k^I_{on}[I]R_{max} - k^S_{on}[S]R_{max} \tag{2.79}$$

$$\frac{d[RI]}{dt} = k^I_{on}[R][I] - k^I_{off}[RI] = k^I_{on}[I]R_{max} \tag{2.80}$$

$$\frac{d[RS]}{dt} = k^S_{on}[R][S] - k^S_{off}[RS] = k^S_{on}[S]R_{max}. \tag{2.81}$$

Similar to the previous section, the goal is to determine when the initial rate of the desired reaction (the specific interaction) is faster than the initial rate of the undesired reaction (the non-specific binding reaction). Thus, the ratio of the initial rates of the binding of the substrate to the receptor vs. the inhibitor to the receptor is expressed as,

$$\frac{\frac{d[RS]}{dt}}{\frac{d[RI]}{dt}} = \frac{k^S_{on}[S]_0R_{max}}{k^I_{on}[I]_0R_{max}} > \alpha \tag{2.82}$$

$$= \frac{k^s_{on}[S]_0}{k^I_{on}[I]_0} > \alpha \tag{2.83}$$

$$[S]_0 > \alpha\frac{k^I_{on}}{k^S_{on}}[I]_0. \tag{2.84}$$

The similarity between the above conclusion and equation 2.76 is obvious, simply exchanging the equilibrium coefficients for the on rates for the specific and non-specific interactions. There is, however, no data in the literature for the on rates for abundant proteins in most measurements (such as serum albumin), and it is difficult to make these measurements. However, this theoretical framework can provide some insight for when to use the initial rate method versus the equilibrium method. When measuring the kinetic coefficients of an antibody-antigen or receptor-substrate interaction, by looking at the on rates an educated guess about the best method can be made based on the balance between the on and off rates of the interaction. Proteins with higher than average on rates will benefit from the initial rate method, whereas proteins with higher than average off rates prefer the equilibrium method. This distinction can be improved with improved methods for measuring the on and off rates of low affinity reactions.

2.3.3 Sandwich assays

To improve the sensitivity of a protein measurement, typically a sandwich assay is used. Sandwich assays utilize a secondary antibody as a to improve the sensitivity of the measurement by adding another affinity reaction step to the measurement. Typically the secondary antibody is bound to a reporter such as a fluorophore or enzyme. The most common form of the sandwich assay is the enzyme linked immunosorbent assay, or ELISA. An ELISA assay is performed by first coating a well with the primary antibody of the desired antigen. Next the solution containing the antigen (e.g., cell lysate) is added to the well and allowed to react. The well is then washed with clean buffer, and finally the secondary antibody solution is added to the well and allowed to react for a time. The well is washed again, and the reporter is read from the well to determine the concentration of the antigen in the solution. By adding the secondary antibody step a second affinity step is used, thereby increasing the sensitivity of the assay.

The simplest way of understanding the sandwich assay is by considering the primary antibody incubation as a concentration step. Since there is a higher affinity for the target than to interfering compounds in the solution, thereby making the effective concentration of the analyte significantly higher on the surface relative to the non-specific inhibitors. Next, a pure solution of the secondary antibody is added. There will still be non-specific binding between the interfering proteins on the surface and the secondary antibodies, however it will proportionally be less due to the concentration by the primary antibodies on the surface of the chip. This provides a significant increase in the desired over the undesired binding, thereby increasing the affinity of the overall reaction.

Considering this in a mathematical sense, the specific binding of the of the secondary antibody to the substrate, $[SAb]$ can be modeled as binding in a pure solution at equilibrium. Using equation 2.62,

$$[SAb] = \frac{K_a[Ab]}{1 + K_a[Ab]}[RS]_{eq}, \tag{2.85}$$

where $[Ab]$ is the antibody concentration and K_a is the affinity coefficient between the secondary antibody and the substrate. Likewise, the non-specific interaction between the secondary antibody and the surface can likewise be modeled as

$$[IAb] = \frac{K^I[Ab]}{1 + K^I[Ab]}[RI]_{eq}. \tag{2.86}$$

For simplicity for this calculation, it is assumed that the K^I for the primary and secondary antibodies are approximately the same.

Similar to the previous analysis, the critical factor determining the soundness of the measurement is ensuring that the ratio of the specific to non-specific interactions is greater than 1. Therefore,

$$\frac{[SAb]}{[IAb]} = \frac{\frac{K_a[Ab]}{1+K_a[Ab]}}{\frac{K^I[Ab]}{1+K^I[Ab]}} \left(\frac{[RS]_{eq}}{[RI]_{eq}}\right) > \alpha. \tag{2.87}$$

Note that the portion in the parenthesis can be replaced with the result of equation 2.75. Thus,

$$\frac{[SAb]}{[IAb]} = \frac{\frac{K_a[Ab]}{1+K_a[Ab]}}{\frac{K^I[Ab]}{1+K^I[Ab]}} \left(\frac{K^S[S]}{K^I[I]}\right) > \alpha, \tag{2.88}$$

again where α is the safety factor for the binding ratio. Simplifying this equation and solving for $[S]$, the minimum concentration detectable,

$$[S] > \alpha \frac{(K^I)^2[I](1+K_a[Ab])}{K_a K^S(1+K^I[Ab])}. \tag{2.89}$$

Extending the example discussed previously, the detection of TNF-α can be extended to a sandwich assay by adding a secondary antibody step to the reaction. The values used are

$$\begin{aligned} K^S &= 5.25 \times 10^{10}\text{M}^{-1} \\ K^I &\approx 1 \times 10^5\text{M}^{-1} \\ K_a &= 1 \times 10^{11}\text{M}^{-1} \\ [I] &= 600 \times 10^{-6}\text{M} \\ [Ab] &= 1 \times 10^{-9}\text{M}. \end{aligned}$$

Therefore, given a safety factor, α, of 10, the sensitivity of the sandwich assay is

$$[S] > 10\frac{(1\times 10^5\text{M}^{-1})^2 600\times 10^{-6}\text{M}(1+1\times 10^{11}\text{M}^{-1}\cdot 1\times 10^{-9}\text{M})}{5.25\times 10^{10}\text{M}^{-1}\cdot 1\times 10^{11}\text{M}^{-1}(1+1\times 10^5\text{M}^{-1}\cdot 1\times 10^{-9}\text{M})} \approx 1pM. \tag{2.90}$$

Again, this is a good approximation of what is demonstrably possible in the literature. Occasionally it is possibly to make measurements below this threshold in real solutions, but usually they are in more favorable conditions than blood serum.

It is also interesting to note that the result obtained from equation 2.89 shows that the sensitivity will increase as the secondary antibody concentration is reduced. This makes sense, since at high concentrations the specific binding of the secondary antibody to the analyte will saturate, and thus by increasing the concentration of the antibody only non-specific binding will occur. This effect is seen experimentally very often in poorly designed ELISA assays. If the antibody concentration is too high, the noise in the measurement will saturate the signal. On the other hand the concentration of the secondary antibody cannot be too low otherwise the experiment will take too long to accomplish.

In the extreme case, if the secondary antibody concentration is so low that the binding rate of the secondary antibody is comparable to the off rate of the primary antibody and the substrate, a measurement will never be accurate due to the desorption of the analyte from the surface. Therefore, a compromise must always be made between the incubation time and the sensitivity of the assay.

2.4 Microfluidics, flow and timescales

A major portion of the technology described in this thesis is reliant on piping fluid to and from the brain. Understanding the dynamics of these channels will be crucial in a theoretical sense will have an impact on the design of the probes as well as the interface between the probe and macroscopic fluidic elements. Understanding expected flow rates is crucial for the perfusion of caged neurotransmitters and drugs into the brain. Furthermore, when comparing assays which are built into a microfluidic framework to those done on the bench top, the presence of flow over the sensor can help or hinder the ability of the reaction to occur. For example, slow flow rates limit the amount of material reaching the sensor compared to a fast flow rate, which will slow the rate that the reaction reaches equilibrium. In this section, the theory required to design biosensing and perfusion experiments will be discussed.

2.4.1 Poiseuille flow through microfluidics

The first thing to consider when designing flow channels is to determine the flow regime in which the dynamics of the fluid will reside. To determine whether the flow in the channel is laminar or turbulent, the Reynolds number is used:

$$Re = \frac{\rho v L}{\mu} = \frac{Q \rho D_H}{\mu A}, \tag{2.91}$$

where ρ is the density of the fluid, v is the average velocity of the flow, L is the characteristic length of the channel, μ is the dynamic viscosity of the flow, Q is the flow rate of the fluid in the channel, and D_H is the hydraulic diameter of the channel. Above on the left is the general form of the Reynolds number and on the right is the form typically used in microfluidic applications. Typically the flow in the channel is considered laminar when $Re < 2000$. For a typical buried silicon microfluidic channel as described in this thesis, the diameter of the channels are typically 3-4μm. The flow rate can be approximated using the volume of a typical channel, the longest of which is about 1.5cm and given a diameter of 3μm results in a volume around 100pL. Thus, if the channel is cleared once per minute, $Q = 100$pL per minute. The viscosity of cerebrospinal fluid is typically close to water, which is 1×10^{-3} Pa·s. Therefore, the Reynolds number for a typical microfluidic channel is

$$Re = \frac{(100\text{pL per min}) \cdot (1000\text{Kg per m}^3) \cdot (3\mu\text{m})}{(1 \times 10^{-3}\text{Pa} \cdot \text{s}) \cdot \pi \cdot (1.5\mu\text{m})^2} = 7.1 \times 10^{-4}] \ll 2000. \tag{2.92}$$

For all intents and purposes, all microfluidic channels exhibit laminar flow.

Given that the flow will be laminar, a simple solution to the Navier-Stokes equation can be used. This solution, the HagenPoiseuille equation, provides a simplified resistance model for laminar flow through a circular tube. The governing equation is

$$\Delta P = \frac{8\mu L Q}{\pi r^4}, \tag{2.93}$$

where ΔP is the pressure differential over the channel, μ is the dynamic viscosity of the fluid, L is the length of the channel, Q is the volumetric flow rate, and r is the radius of the channel. This provides a simple way to design channels to ensure that the desired flow rates are within a nominal desired range. Due to process variation, the flow rate will have to be measured empirically before beginning an experiment where the volume of a bolus or a steady, known perfusion rate is desired.

It is important to determine the volumetric flow rate of a typical channel for the next section, and thus an estimate of the flow rate will be made for a microfluidic loop given a shank length of 5 mm and a radius of 3.5μm (this is the worst case scenario). The pressure differential used will be 1PSI, which is safe for most microfluidic devices. Using these values,

$$Q = \frac{\pi r^4}{8\mu L}\Delta P = \frac{\pi \cdot (1.75\mu\text{m})^4}{8 \cdot (1.4 \times 10^{-3}\text{Pa} \cdot \text{s}) \cdot (15\text{mm})} 1PSI \approx 72.5\text{pL per minute}. \tag{2.94}$$

This is a reasonable value from a functional perspective, resulting in the channel from the exchange area to the sensors to be cleared once every minute.

2.4.2 Convective and diffusive influences on biosensor reactions

Interaction between convection, diffusion and reaction is a common theme in biomass transfer. In reference [110], Squires et al. uses standard mass transfer theory and applies it to biosensing applications. For this section, I will summarize the results relevant to the sensors in this thesis and apply the theory to the practical parameters of the ring resonator biosensors described above. The goal of this section is to determine the timescale for the detection reaction when placed in a microfluidic context.

2.4.2.1 Definitions

A generic sensing area with dimensions of length, L, and W_s are within a microfluidic channel of cross sectional dimensions height, H, and width, W_c. The flow through the channel is laminar (as described above), with velocity, u, and volumetric flowrate, Q. The dimensionless sensor scale is defined as $\lambda = L/H$.

2.4.2.2 Reaction and Diffusion: The boundary layer

Starting with the simplest case where there is no flow in the channel, a diffusive flux will arise from S, the substrate, binding to the surface. As molecules bind to the surface and since by design $k_{on}[S]_0 > k_{off}$, they are removed from the surface, thereby lowering $[S]$ near the surface. This will create a concentration gradient, causing more of S to diffuse toward the surface of the sensor. The concentration gradient created by this process is called the boundary layer, the size of which is

$$\delta \approx \sqrt{Dt}, \tag{2.95}$$

where δ is the size of the boundary layer, D is the diffusion constant of the analyte in water, and t is time. The boundary layer will grow with time in a stationary system, collecting any molecules that diffuse toward the sensor. In a flow system, the boundary layer will reach a steady state size as the solution above the boundary layer is replenished. As will be described later, the size of the boundary layer is a useful feature to understand when compared to the size of the channel under flow conditions, as it helps determine the limitation of the rate of reaction, whether that limitation be diffusive or reactive.

Another important concept is the time in which the boundary layer grows to be the height of the channel. Dimensionless time $\tilde{t}$ is defined as

$$\tilde{t} \equiv Dt/H, \text{ s.t. } \tau_D = H^2/D. \tag{2.96}$$

τ_D is the dimensionless diffusion timescale.

2.4.2.3 Convection, Diffusion and Reactions: the Damköhler number

When considering the combined effects of convection, diffusion and surface reactions, any calculation will be concerned with the relative flux of molecules due to diffusive/convective processes vs the flux of molecules binding with the sensor surface. In a typical experiment, convection and diffusion will continually expose the sensor surface to a fresh solution containing the analyte, while the reactive flux will remove the analyte from the solution above. While the sensor surface is not saturated, the system will quickly reach a steady state where the size of the boundary layer neither grows nor shrinks. Given the assumption of a steady state, it is clear that the diffusion caused by the concentration gradient in the boundary layer must be equal to the flux of the analyte binding to the surface. In the case of laminar flow through a flow cell over the sensor, the diffusive flux is

$$J_D \approx D([S]_0 - [S]_S)W_S\mathscr{F}, \tag{2.97}$$

where J_D is the diffusive flux through the boundary layer, $[S]_S$ is the surface concentration of the concentration of the substrate at the sensor surface (therefore $([S]_0 - [S]_S)$ is the concentration gradient), W_S is the sensor width, and $\mathscr{F}$ is the dimensionless flux function. $\mathscr{F}$ is dependent on the properties of the flow cell and will be discussed in the next section in detail. The reactive flux of the sensor is

$$J_R \approx k_{on}[S]_S b_m L W_S, \tag{2.98}$$

where b_m is the density of receptors on the surface, and $L \cdot W_S$ is the area of the sensor. Given the steady state assumption,

$$J_D \approx D([S]_0 - [S]_S)W_S\mathscr{F} = k_{on}[S]_S b_m L W_S \approx J_R. \tag{2.99}$$

The interesting value to solve for in this situation is the ratio of the surface concentration $[S]_S$ to the bulk solution concentration $[S]_0$. In this context, it is possible to determine the dilution due to the boundary layer. Thus, solving for the ratio,

$$\frac{[S]_S}{[S]_0} = \left(1 - \frac{k_{on} b_m L}{D\mathscr{F}}\right)^{-1}. \tag{2.100}$$

The factor in the parenthesis is the Damköhler number,

$$Da \equiv \frac{k_{on} b_m L}{D\mathscr{F}}. \tag{2.101}$$

This dimensionless constant is useful as it defines the transition between the system being transport limited and reaction limited. When $Da \gg 1$ the system is transport limited whereas when $Da \ll 1$ the system is reaction limited. Furthermore, when the system is in the reaction limited regime, the rate of the reaction will approach the time constant for the reaction. However, when the system is in the transport limited regime, the reaction time is more complicated.

In the transport limited regime ($Da \gg 1$), the convective and diffusive flux becomes the dominant term,

$$j_D \approx \frac{D[S]_0\mathscr{F}}{L}. \tag{2.102}$$

In this case, since the boundary layer is very large, $[S]_S$ is small compared to $[S]_0$ and can be ignored. As the flow continues over the sensor, nearly all of the analyte will be captured by the surface. Thus, the sensor should approach equilibrium when the amount of molecules that have crossed the sensor in the flow channel is equal to the number of the analyte bound to the sensor at equilibrium. Thus,

$$\tau_{TL} = frac[RS]_{eq} j_D, \tag{2.103}$$

where tau_{TL} is the transport limited equilibration time constant of the system. Recall that $[RS]_{eq}$ is defined in equation 2.62, and therefore the transport limited time constant is

$$\begin{aligned} \tau_{TL} \approx \frac{[RS]_{eq}}{j_D} &= \left(\frac{b_m k_{on}[S]_0}{k_{off} + k_{on}[S]_0} \right) \cdot \left(\frac{L}{D[S]_0 \mathscr{F}} \right) && (2.104) \\ &= \left(\frac{k_{on} b_m L}{D \mathscr{F}} \right) \left(\frac{1}{k_{off} + k_{on}[S]_0} \right) && (2.105) \\ \tau_{TL} &\approx Da \tau_R, && (2.106) \end{aligned}$$

where b_m and R_{max} are interchangeable, only one is used in a concentration context and the other in a number context. Thus, the real rate of the reaction in the transport limited case is simply the product of the Damköhler number and the reaction only time constant. The only factor yet to be addressed is how the flux function $\mathscr{F}$ is affected by the geometry of the channel and flow rate, which will be discussed in the next section.

2.4.2.4 Convection and Diffusion: Péclet Numbers

When considering the flux of molecules reaching the surface of the sensor, it is important to understand which factor is the dominant factor in the delivery of the analyte to the surface, diffusion or convection. For example, if the flow rate is slower than diffusion, all of the analyte will likely be collected by the sensor. However if the flow rate is fast, not all of the analyte will be able to make it to the surface. Thus the balance of these two terms is important to understand the amount of analyte exposed to the sensor.

Thus a dimensionless comparison of the diffusive and convective time scales is required to determine which term dominates. The size of the boundary layer grows toward the height of the channel as

$$\tau_D = \frac{H^2}{D}. \tag{2.107}$$

In comparison, the convective timescale compared to the height of the channel is

$$\tau_C = \frac{\text{Channel Cross Section} \cdot \text{Critical Dimension}}{\text{Volumetric Flow Rate}} = \frac{W_c H^2}{Q}. \tag{2.108}$$

Taking the ratio of these two factors results in the Péclet number with respect to the height of the channel,

$$Pe_H = \frac{\tau_D}{\tau_C} = \frac{Q}{D W_c}. \tag{2.109}$$

If $Pe_H \ll 1$, mass transfer is dominated by diffusion and the percentage of analyte collected from the solution is high. In this case, the relationship between the flux and the flow is simple:

$$\mathscr{F} \approx Pe_H. \tag{2.110}$$

Since all of the analyte will be collected as it flows by, the dimensionless flux collapses directly to Pe_H.

If $Pe_H \gg 1$, the mass transfer is dominated by convection and the percentage of analyte collected from the solution is low and the properties of the flow must be further considered. In a channel with laminar flow, the velocity profile of the flow along the z-axis is parabolic,

$$u = \frac{6Q}{w_c H^3} z\left(H - z\right). \tag{2.111}$$

However in the case where the boundary layer is small relative to the height of the channel, the shear flow above the sensor must be considered. Linearizing around $z \approx 0$, the flow velocity becomes

$$u_s = \left(\frac{6Q}{H^2 W_c}\right) z. \tag{2.112}$$

Thus, the effective time scale of the shear flow becomes

$$\tau_C = \frac{6Q}{H^2 W_c}. \tag{2.113}$$

Substituting this into the equation for the Péclet number,

$$Pe_s = \frac{\tau_D}{\tau_C} = \frac{L^2/D}{H^2 W_c/(6Q)} \tag{2.114}$$

$$= 6\left(\frac{L^2}{H^2}\right)\left(\frac{Q}{DW_C}\right) \tag{2.115}$$

$$Pe_s = 6\lambda Pe_H, \tag{2.116}$$

where Pe_s is the shear Péclet number. When $Pe_H \gg 1$, Pe_s is used to determine $\mathscr{F}$. Using computer simulations described in [110], the dimensionless flux is

$$\mathscr{F}(Pe_H \gg 1 \ \&\ Pe_S \gg 1) = .81 Pe_s^{1/3} + .71 Pe_s^{-1/6} - .2 Pe_s^{-1/3} + \ldots \tag{2.117}$$

$$\mathscr{F}(Pe_H \gg 1 \ \&\ Pe_S \ll 1) = \pi \left(\ln(4/Pe_s^{1/2}) + 1.06\right)^{-1}. \tag{2.118}$$

Thus, using these factors the dimensionless flux for the specific case can be found using the two versions of the Péclet number.

2.4.2.5 Convection, diffusion and reaction: calculation example

To demonstrate the analysis of a typical system, we will use a sample system similar to the measurement system described later in this thesis. The critical parameters are

$$\begin{aligned}
W_c &= h = 100\mu m \\
L_s &= 25\mu m \\
Q &= 150pLmin^{-1}(\text{typicall flow rate at 1PSI}) \\
D &= 3 \times 10^{-11}m^2s^-1(\text{TNF-}\alpha) \\
b_m &= 1 \times 10^{12}\#cm^{-2} \\
k_{on} &= 1 \times 10^5 M^{-1}s^{-1} = 1.66 \times 10^{-19}L\#^{-1}s^{-1}(\text{TNF-}\alpha).
\end{aligned}$$

First, calculate the Péclet number with respect to the channel height,

$$Pe_H = \frac{Q}{DW_c} = \frac{150pLmin^{-1}}{(3 \times 10^{-11}m^2s^-1) \cdot (100\mu m)} = 0.83. \qquad (2.119)$$

Given this result, the flow rate is slow enough that nearly all the analyte that passes by should be collected, but not slow enough that it's guaranteed. In this case it is a good idea to calculate the sheer rate Péclet number to make sure it isn't large. Thus,

$$Pe_s = 6\lambda^2 Pe_H = 6\left(\frac{25\mu m}{100\mu m}\right)^2 0.83 = 0.31. \qquad (2.120)$$

Thus, the worst case scenario is that $\mathscr{F} = Pe_H$. Given an estimate for $\mathscr{F}$, the Damköhler number can be calculated,

$$Da = \frac{k_{on}b_mL}{D\mathscr{F}} = \frac{(1.66 \times 10^{-19}L\#^{-1}s^{-1}) \cdot (1 \times 10^{12}\#cm^{-2}) \cdot (25\mu m)}{(3 \times 10^{-11}m^2s^-1) \cdot 0.83} = 1.67, \qquad (2.121)$$

and thus the time scale for the reaction is 1.67 × τ_R. By speeding up the flow to 2 or 3 PSI, the system reaction should closely approximate τ_R. For more examples, see [110].

Chapter 3

Fabrication

3.1 Photolithography

Photolithography is a primary method for wafer patterning that is critical for the creation of MEMS devices. Photolithography utilizes external energy to create a chemical change in a polymer. In this case, photolithography uses a photographic film which is sensitive to ultraviolet light to create a pattern on the wafer. In this section, the considerations and recipes for the photolithographic process are discussed. For a discussion of the theory of photoresists, see appendix A.

3.1.1 Photoresist Processing

In this section, the practical elements of photoresist processing will be discussed. Primary among those, is the equipment that is used to process these resists. When using novolac-DNQ based resists, there is a missmatch in the surface properties of the wafer and the polarity of the resist, meaning that the resist on an untreated wafer (which is typically polar due to the growth of native oxide on the wafer surface) can peel off the wafer if mishandled. To improve adhesion, before coating the wafer with photoresist the wafer is exposed to hexamethyldisilazane (HMDS), which reacts with the native oxide and forms a non-polar surface.

After exposure to HMDS, photoresist is applied via spin coating using a Laurel 150mm spin coating system (Laurell Technologies, North Wales, PA). The thickness of the photoresist coating is controlled by the angular velocity of the spinning chuck, which generally has the relationship of

$$\text{Thickness} \propto \frac{1}{\omega}, \tag{3.1}$$

where ω is the angular velocity of the spin chuck, typically measured in RPMs. Resist is dispensed on the wafer such that it covers one-half to two-thirds of the surface of the wafer. Most spin recipes start with a slow (typically 500RPM) spin rate to distribute the resist over the surface of the wafer. After the spread step, the high velocity step will reduce the film thickness to the desired thickness, typically running for 1 minute. After the spin step, the wafer is baked to drive off excess cast solvent in the film.

Exposure of resists was done using contact lithography on a Karl Suss MA-6 mask aligner. This method of lithography utilizes a chrome mask layer on a soda lime glass plate and brings the plate in direct contact with the wafer to transfer the pattern. The advantage of contact lithography is that there is no limitation on the size of the exposure region and it is simple and inexpensive to produce masks. The limitation of this method is that since the light impinging on the mask is not focused, diffractive effects create a limitation in feature size at around 1-2μm. Alignment is accomplished using a pair of microscopes which visualize alignment marks on each side of the wafer. Micrometers are used to position the wafer underneath the mask aligner. Once in position, the wafer and mask are brought into contact using a stepper motor. To improve contact uniformity the chamber is first evacuated so that air cannot be trapped between the wafer and mask. After contact is achieved, the mask aligner exposes the wafer through the mask plate at either 405nm or 365nm using a filtered mercury lamp.

Only puddle development is currently available for wafer processing. The wafer is submerged in a bath of resist developer solution (typically a commercial solution of .26N tetramethyl ammonium hydroxide with additives to improve development) such as MF-319 or CD-26. The wafer is then transferred to a water bath after a set amount of time and allowed to soak for an additional minute to remove any trace developer from the wafer surface. The wafer is then dried using a stream of high pressure N_2 gas. Specific recipes for the resists used in the processing of wafers follows.

3.1.2 S1813

The S1800 series of resists by Shipley is a standard series of novolac-DNQ resists that is optimized for liftoff and wet etch processing. S1813 specifically spins to 1.3μm at a spin speed of 4000 RPM.

1. Dehydration bake for 2 minutes at 180^{o}C,
2. Expose to HMDS vapor for 10 minutes
3. Spin coat S1813, 2500 RPM for 30s (appx. 1.75μm thick)
4. Bake on hotplate for 3 minutes at 110^{o}C
5. Expose wafer 200mJ/cm^2 with 405nm light
6. Develop in CD-26 for 1 minute

3.1.3 AZ-9245

1. Dehydration bake for 2 minutes at 180^{o}C
2. Expose to HMDS vapor for 10 minutes
3. Spin coat AZ-9245, 4000 RPM for 30s (appx. 4.2μm thick)
4. Bake on hotplate for 2 minutes at 110^{o}C
5. Expose wafer 625mJ/cm^2 with 405nm light
6. Develop in 1:4 AZ 400K developer:H_2O for 4 minutes

3.1.4 Liftoff Process

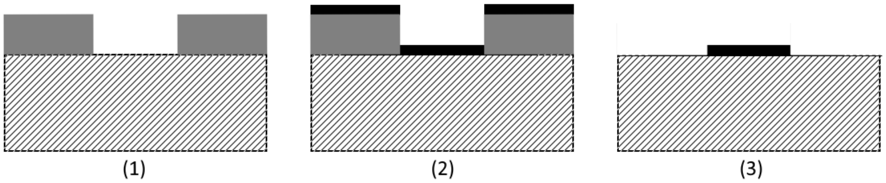

Figure 3.1: Liftoff process. 1, patterned photoresist on substrate. 2, deposited hard mask. 3, photoresist removed by acetone soak, leaving behind patterned hard mask.

A liftoff process is a photolithographic process which allows for the patterning of a thin film without the use of etching. This process is typically used to deposit hard masks for deep etch processes. This process is shown in figure 3.1. Photoresist is first applied to the wafer surface, with a mask with a negative version of the desired final pattern (i.e., photoresist will be removed in the places we want the hard mask to remain on the wafer). The hard mask is then deposited over the patterned wafer. Finally, the wafer is placed in a bath of acetone which dissolves the resist. This will undercut any of the hard mask material that is deposited on top of the resist, freeing it and allowing it to be released into the acetone bath. Thus, only the hard mask that was in contact with the wafer itself will remain. This technique can be used to pattern hard mask materials like aluminum oxide or metal films.

3.2 Electron Beam Lithography

Similarly to photlithography, electron beam lithography (EBL) also uses a polymer film to pattern wafers. However, instead of utilizing ultraviolet light shining through a mask-plate to expose the photographic film, electron beam lithography uses a serial approach where an electron beam is traced over the polymer film to create a pattern. Although significantly slower, electron beam lithography has the advantage of being limited to the diffraction limit of an electron. Since it is much easier to create and control high energy electrons, electron beam lithography makes it easy to create nanometer scale features in resists. Electron beam lithography has the additional advantage of being reprogrammable allowing for rapid prototyping of new patterns without having to create expensive mask plates each time a new pattern is required. In this section, the practicalities of EBL will be discussed, culminating in the development of a novel EBL resist by the author and collaborators.

3.3 Electron beam lithography resists

3.3.0.1 PMMA and SML Resist

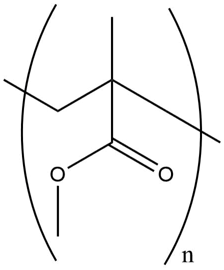

Figure 3.2: PMMA polymer.

The most commonly used resist in electron beam lithography is Poly(methyl methacrylate) (PMMA), as shown in figure 3.2. This polymer is used in a variety of industrial applications, and is most often seen in its solid form as acrylic glass. However, this polymer can be cast in anisole (at a mass percent between 2% and 11%), allowing it to be spun onto wafers. Since these resist utilize a main-chain polymer scission process, the molecular weight (and thereby the length of the chains) is a critical parameter in the properties of the resist. PMMA can be purchased in 495kDa and 950kDa variants.

PMMA is a popular resist due to its low cost, however it is limited in its application predominantly due to poor etch resistance. A variant of PMMA is available for purchase called SML (EM resist, Manchester, UK.). Originally developed to create mask plates directly using electron beam lithography, SML is a composite resist[91]. SML utilizes PMMA to provide the lithographic properties of the deposited film, but is impregnated with a number of dyes which, when exposed to an electron beam, are broken down into smaller, non-absorbing materials. The intention of this resist was to be absorbing where the electron beam is absent during writing, but clear where the electron beam exposed the resist. Although this application is not used in developing the devices described in this thesis, the addition of these additive dyes were at high enough concentration to effect the etch resistance of the SML resist. Thus, SML acts as a version of PMMA with a stronger etch resistance.

The dyes added to PMMA to create SML are Sudan black, sudan orange, trans-Stilbene, and 4-phenylazophenol, shown in figure 3.3. It is quite obvious that the structure of all of these compounds

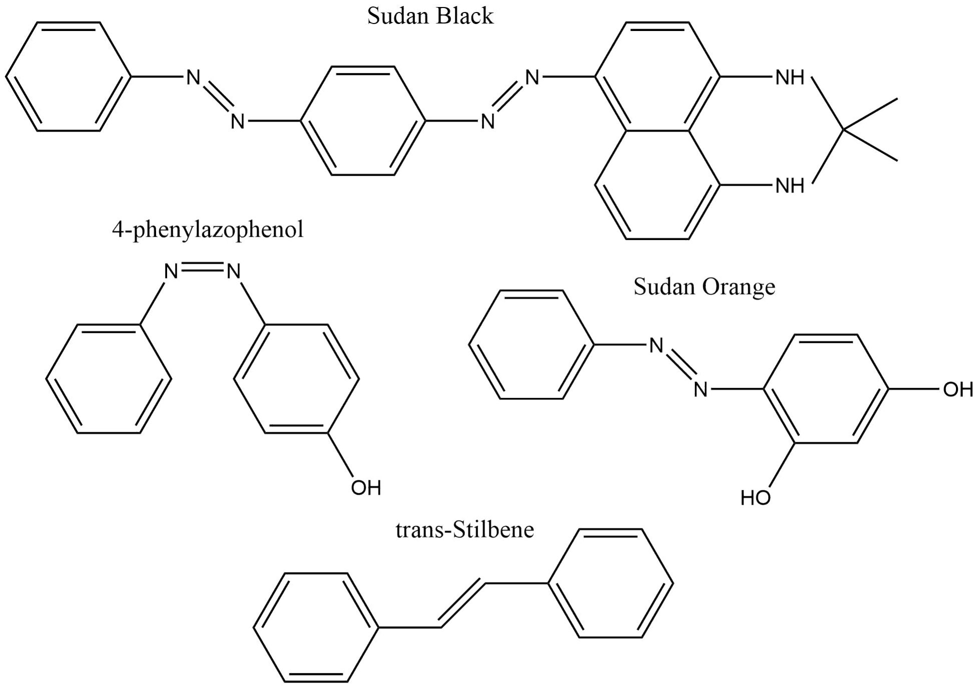

Figure 3.3: Dye additives to PMMA to create the SML series of resists.

have many features in common, including conjugated pi-bond networks and many benzene rings. Due to the stability of the benzine rings (as is shown in novalac resins used for photolithography), these compounds improve the etch resistance of the film, which is highly desirable, although it does not quite match up to the etch resistance of ZEP, discussed in the next section. The unfortunate trade-off when using SML is that the dyes also act to absorb electrons during the exposure process, resulting in incredibly slow writes. Doses required for the thickest films (e.g, SML-2000) can exceed 3000μC per square cm.

For the processes described in this chapter, SML-2000 was used to create structures resistant to long, harsh silicon dioxide etch processes. The resulting film was typically greater than 3 microns thick, but allowed for pseudo-bosch etching for durations greater than 20 minutes while creating feature sizes well below what is able to be done with contact photolithography. Wafers are first cleaned, dried, and baked at 180^{o}C for at least 2 minutes to remove any residual water or solvents which might be on the surface of the wafer. Wafers are then coated with a puddle of SML-2000, spread around the surface of the wafer using a swab, and spun at 1500 RPM for 45 seconds to evenly distribute the resist over the surface. The back of the wafer is cleaned with a wipe soaked in Acetone to remove any residual resist, and is then baked for 3 minutes at 180^{o}C. Exposure parameters for the resist using a 100kV electron beam are as follows. Typical dose for the film is 3500μm per square centimeter. Features produced using SML-2000 were usually quite large for electron beam lithography (300-600μm), and thus large beams could be used. A 100nA beam with a beam step size of 50μm was used to write the resist. Patterns were developed similarly to PMMA in 3:1 IPA:MIBK developer. Wafers were developed in a bath for 1 minute, rinsed in isopropyl alcohol and dried with compressed nitrogen gas.

3.3.0.2 ZEP Resist

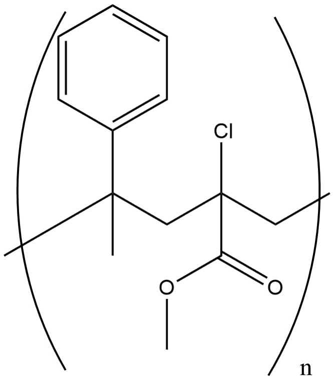

Figure 3.4: ZEP polymer.

ZEP-520a(Zeon Corporation, Tokyo JP) resist is a proprietary polymer resist (dissolved to 11% by weigth in anisole) that is extremely popular in academic nanofabrication facilities. Structurally similar to PMMA (see figure 3.4), it adds an additional 2 carbons to the backbone, one of which includes a benzene ring, which improves the etch resistance of the resist significantly. In addition, a CH_3 side chain is replaced with a Cl group, which acts as a secondary electron generator, improving the write speed of the resist by a factor of 3. This resist is preferred for many applications due to its high etch selectivity and improved write speed over PMMA.

This resist is used for the production of high quality optical devices in this thesis. Typical processing is as follows. Wafers are first cleaned, dried, and baked at 180^{o}C for at least 2 minutes to remove any residual water or solvents which might be on the surface of the wafer. Wafers are then coated with a puddle of ZEP-520a and spun at 4500 RPM for 30 seconds, evenly distributing the resist over the surface. The back of the wafer is cleaned with a wipe soaked in Acetone to remove any residual resist, and is then baked for 2 minutes at 180^{o}C. Exposure parameters for the resist using a 100kV electron beam are as follows. Typical dose for the film is 280μm per square centimeter. For fine features such as ring resonators and other optical devices where edge roughness is a critical, a beam current of 300pA is used with a beam step of 2.5nm. The 300pA beam has a spot size of 5nm, and allowing for this overlap between beam shots produces a smoother result. Features where sidewall roughness was not critical (e.g., alignment marks), a beam with a 100nA current and a step size of 50nm was typically used to increase write speed. Wafers were puddle developed in a proprietary developer, ZED-N50 (Zeon Corporation, Tokyo, JP), consisting of amyl acetate and proprietary additives, for 1 minute. Wafers were then rinsed in a bath of isopropyl alcohol and dried using compressed nitrogen gas.

Reflow Processing To improve sidewall roughness, it is possible to reflow the resist. Reflow processing involves baking the resist after development to slightly liquefy the resist. Liquefaction of the resist surface allows high frequency roughness to be reduced by surface energy minimzation of the resist in liquid form. Essentially, small variations in the surface will be removed by the surface tension of the resist in liquid form. This is a standard procedure as described in the literature by [72]. The result of this paper showed that the critical temperature for ZEP reflow is between 145-155^{o}C. Thus, this is the range for testing used when developing this process. Ring resonators were fabricated using the standard electron beam lithography procedure using ZEP resist described above. Rings were fabricated from 325nm wide by 300nm tall waveguides with a ring diameter of 11μm, and the quality factor of the resonance was used to optimize the reflow procedure using 672nm light. It was found that baking the ring resonators on a hotplate at 152^{o}C for 1 minute produced extremely high quality factor resonators (g.t. 100,000 with air cladding) while not significantly degrading the quality of grating couplers.

3.3.0.3 Ma-N Resist

Although positive tone electron beam resists are the norm in academic settings, there are some drawbacks which necessitate using negative tone resists. For example, positive tone resists are often slow to write, both because of the required doses to induce main-chain scission, but also due to geometric constraints. For example, when producing waveguide structures by subtractive fabrication processes (e.g., etching) using a positive tone resist, the pattern which is defined is actually an outline of the desired waveguide, not the waveguide itself. To prevent coupling of light into the surrounding material, the trenches that define the border of the waveguide must often be much larger than the waveguide itself, creating a situation where the pattern is much larger than the desired structure. Positive tone resists often take significantly longer than negative tone resists to write comparable features. Thus from a practical standpoint negative tone resists are desirable when long, narrow structures are desired (such as waveguides).

To improve write times on devices with long waveguide structures where some waveguide loss is tolerable, the use of positive tone resists were used. Specifically Ma-N 2403 (Microchemicals GMBH, Ulm, Germany), a deep-UV resist which is also electron activated [28]. This resist is a chemically amplified novolac resist which crosslinks when exposed to an electron beam. Unfortunately the chemical amplification process creates roughness in the features, thus creating inferior waveguide structures to ZEP-520a.

Ma-N 2403 is a highly hydrophobic material and thus wafer preparation requires multiple steps to prepare the surface for consistent coating. This is especially true with SiO_2 and Si_3N_4 films, which tend to have a thin oxide layer at their surface which is highly polar. Wafers are first cleaned in a solution of a proprietary, acid-peroxide based cleaning solution called Nanostrip (KMG Chemicals). This solution is similar to typical "pirahna" etches consisting of a mixture of sulfuric acid and hydrogen peroxide, but Nanostrip is stabilized for long term storage and safety. The wafer is first subjected to a 5 minute bath in Nanostrip at room temperature, and is then rinsed with deionized water and dried with compressed nitrogen gas. The wafer is then dipped in buffered oxide etch for 10 seconds to remove any native oxide on the surface of the wafer, and is again rinsed with deionized water and dried with compressed nitrogen gas. Immediately after drying, the wafer is placed on a 180^oC hotplate for at least 2 minutes to drive away any additional water on the surface of the wafer. After baking, the wafer is then placed in a container where HDMS is applied in the vapor phase, converting the typically polar surface of the wafer into a non-polar surface chemistry.

Next, Ma-N 2403 is applied to the surface of the wafer using a pipette, is generously spread across the wafer using the pipette, and is spun at 3500 RPM for 30 seconds. The resist is then baked for 1 minute at 90^oC. Wafers are placed in the EBPG-5000+ electron beam lithography tool, and patterns are exposed with a dose of $250\mu C$ per cm^2. Fine features were exposed using a 500pA beam with a beam step resolution of 2.5nm. Coarse features (such as alignment marks) were exposed using a 100nA beam and a 50nm beam step resolution.

Features are developed in a bath of Ma-D 520 proprietary developer (0.2N tetramethylammonium hydroxide plus additives) for 1 minute, and rinsed with deionized water and dried with compressed nitrogen gas. An issue that is common with negative tone resists is that, since the majority of the film is dissolved away by the developer, there is a risk of re-deposition of developed resist onto the surface of the wafer, creating a thin film of organic residue on the surface of the wafer. This can create surface roughness during etching, which is highly undesirable. To mitigate this issue, wafers are developed upside down to mitigate surface redeposition of the development products. To keep the wafer suspended, a wafer cassette spring was used at the bottom of the development vessel to raise the surface of the wafer off of the bottom of the vessel during development.

Reflow Processing Similar to the process used for ZEP-520a, reflow was also used for the optical devices fabricated using Ma-N 2403. In this case it was determined that a superior result was achieved when the wafer was baked in an oven for 5 minutes. The optimal temperature of the oven was found to be 145^oC.

3.3.1 Advanced positive tone resist

Although the SML series of resist allows for extremely long etch processing, there is a fundamental limitation that arises when dealing with thick resist films. The primary concern (although partially allayed by the use of a 100kV electron beam) is that there is a limit on the aspect ratio that can be achieved with a thick resist. Due to the structural constraints of the resist as well as proximity effects, it is impossible to successfully create and etch features in resists at extremely high aspect ratios. For example, it is impossible to create 100nm features in a film of SML-2000. Thus, if longer etches of smaller features are desired, the resist must become more robust to etching, allowing for thinner films to be used.

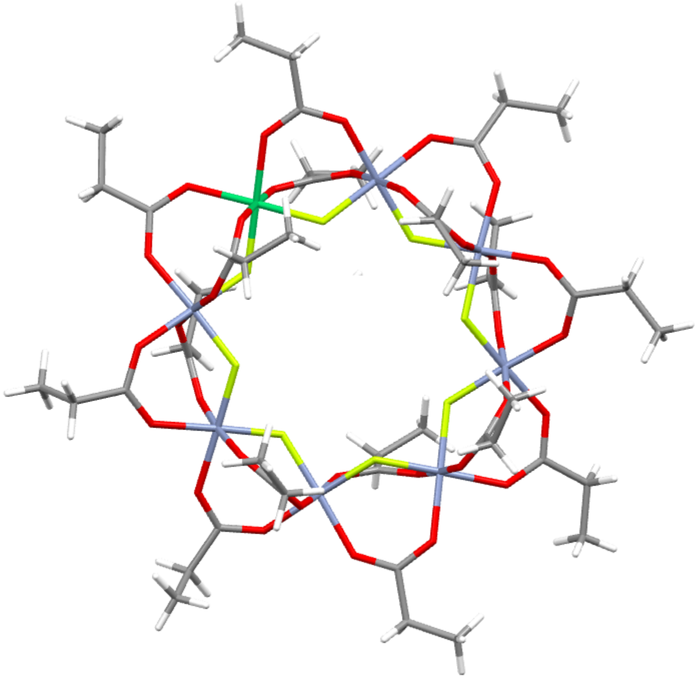

Figure 3.5: Cr_7Ni organometallic ring structure, color coding: Light blue = Chromium, Green = Nickel, White = Hydrogen, Yellow = Fluorine, Grey = Carbon, Red = Oxygen.

Similar to the SML series of resists, it is possible to improve the etch resistance of a resist by impregnating the polymer with other compounds with high etch resistance. This process has been attempted with somewhat unsuccessful results previously using alumina nanoparticles and other materials[61]. However, new negative tone macromolecular resist are currently being developed at the University of Manchester by Scott Lewis, Richard Winpenny, and Steven Yeates in collaboration with Guy Derose at Caltech. These resists consist of an organometallic ring structure of nickel and chromium with attached organic moieties to form a non-polar soluble spin on metal oxide film (unpublished work). An example of one of these ring structures is shown in figure 3.5. This compound, when exposed to an electron beam, breaks apart into its constituent metal oxides. Since the oxides are then separated from the organic moieties which lend the property of non-polar solubility, the resist becomes insoluble when exposed to the electron beam. This effect produces the negative tone behavior of the resist. These resists have shown supreme etch resistance in experiments at the Kavli

Nanoscience Institute at Caltech, demonstrating etch selectivities of etching of over 100:1 versus silicon.

For the processes described in this thesis, a positive tone version of this resist is highly desirable, which provides high etch resistance as well as smaller features than those that can be produced in thick SML films. To accomplish this, a composite resist was formulated by combining an already existing positive tone resist (which would lend the positive tone behavior to the composite resist) with a lower concentration of the macromolecular structure shown in figure 3.5 (which would provide superior etch resistance to the film). Although this film is unlikely to match the etch resistance of the negative tone version, it is still possible to improve the etch resistance of the film significantly.

3.3.1.1 Formulations of composite resist

Composite resists were formulated using both PMMA and ZEP as the base polymer. The Cr_7Ni ring structure needed modification for solubility in anisole (the cast solvent for both ZEP and PMMA). The organic side chains were replaced with propionate groups to improve solubility. Through empirical testing, it was found that a mass percentage of the Cr_7Ni ring of 30% provided good film coating. When a mass percent of 50% was attempted, the film cracked severely.

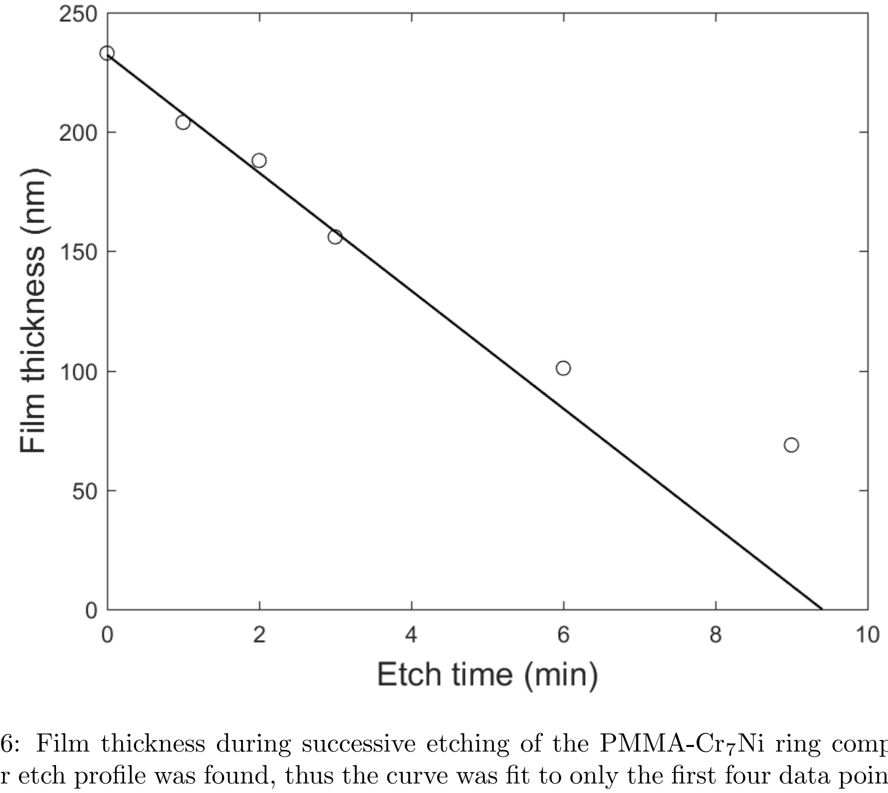

Figure 3.6: Film thickness during successive etching of the PMMA-Cr_7Ni ring composite film. A non-linear etch profile was found, thus the curve was fit to only the first four data points to estimate the etch rate. The etch rate was found to be 24.7 nm per minute (R^2 = .986).

PMMA based composite formulation PMMA-Cr_7Ni ring composites were formulated using PMMA-950-A4 (950,000Da molecular weight, 4% solids by weight). 30mg of the Cr_7Ni ring (provided by Scott Lewis, University of Manchester, Manchester UK.) was measured and dissolved in 300mg anisole and allowed to dissolve fully. The final solution weighed 330mg. This mixture was then filtered through a 450nm pore size syringe filter, resulting in 190mg of filtered solution. The filtered solution was added to 1g of PMMA-950-A4 and rolled until the ring was evenly distributed within the solution. This resulted in a solution with a ratio of Ring:(PMMA+Ring) of 30%.

The resulting PMMA-Cr_7Ni ring composite resist was spun at a rate of 2500RPM for 30 seconds and baked for 2 minutes at 180^oC. The resist was exposed in single pass lines of a 300pA beam with a dose array ranging from 100-10000 μC per square cm using the Raith EBPG 5000+ in the KNI cleanroom. Samples were developed for 30 seconds in a 3:1 solution of IPA:MIBK, following the development procedure of PMMA. Optimal single line doses were found to be 6000μC per square cm.

Samples were etched using the silicon pseudoBosch etch described in section 3.4.2.1 below. This etch as a bulk silicon etch rate of 250nm per minute. The resulting etch rate of the composite resist film (as measured by scanning electron microscopy) is shown in figure 3.6. Interestingly, a nonlinear trend is observed in the etch rate data, indicating that the resist becomes more etch resistant as time goes on. The current theory is that the plasma also generates secondary electrons, which continually exposes the resist during the etching process. The negative tone version of this resist will continue to become more etch resistant with exposure, and thus it is possible that the plasma is continuing the cross-linking process within the resist as the etch is proceeding. This hypothesis is currently being tested.

Comparing the etch rate of this film (24.7nm per minute) to that of silicon (240nm per minute), the etch selectivity over silicon was found to be 9.71:1. This is extremely impressive, especially considering that what is widely considered to be the best polymer resist (ZEP) has a selectivity over silicon of 3.5:1. Resulting patterns of writing and etching the PMMA-Cr_7Ni ring composite resist are shown in figure 3.7.

ZEP-520a based composite formulation To attempt to further improve the etch resistance, the composite was formed with ZEP as the base polymer instead of PMMA. ZEP-Cr_7Ni ring composites were formulated using ZEP-520a (11% solids by weight). 47mg of the Cr_7Ni ring (provided by Scott Lewis, University of Manchester, Manchester UK.) was measured and dissolved in 606mg anisole and allowed to dissolve fully. The final solution weighed 653mg. This mixture was then filtered through a 450nm pore size syringe filter, resulting in 418mg of filtered solution. The filtered solution was added to 1.007g of PMMA-950-A4 and rolled until the ring was evenly distributed within the solution. This resulted in a solution with a ratio of Ring:(ZEP+Ring) of 27%.

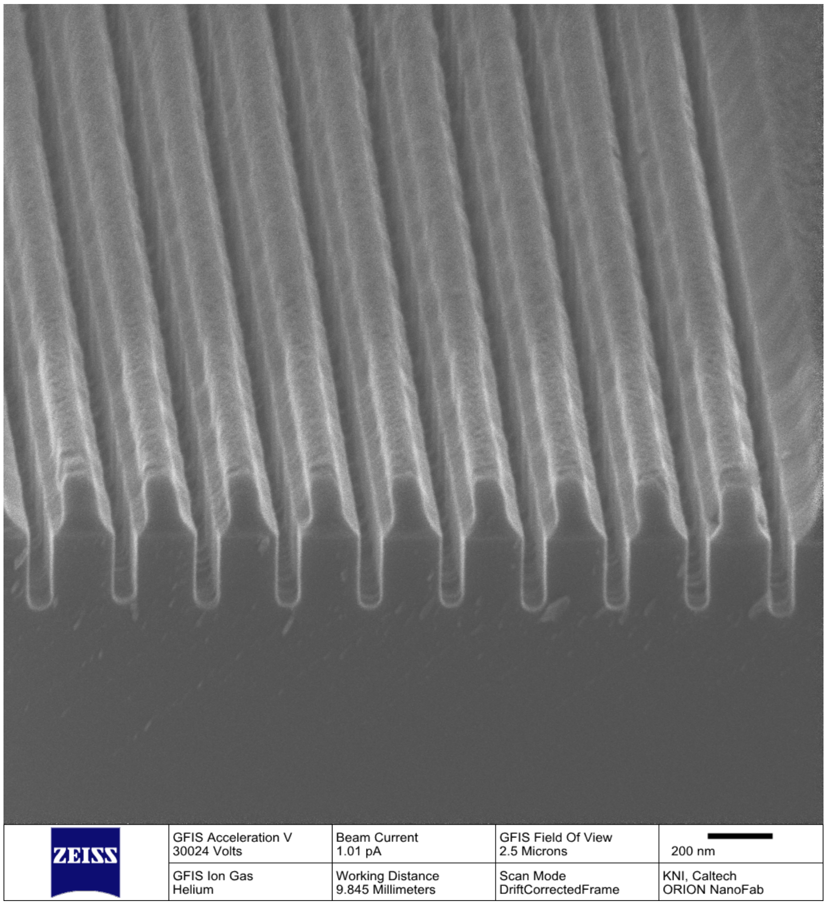

Figure 3.7: Electron micrographs of PMMA-Cr_7Ni ring composite resist after etching.

The resulting ZEP-Cr_7Ni ring composite resist was spun at a rate of 4000RPM for 30 seconds and baked for 2 minutes at 180^oC. The resist was exposed in single pass lines of a 300pA beam with a dose array ranging from 100-10000 μC per square cm using the Raith EBPG 5000+ in the KNI cleanroom. Samples were developed for 1 minute in ZED-N50 developer, following the development procedure of ZEP-520a. Optimal single line doses were found to be 1200μC per square cm. It is interesting to note that the dose for the ZEP version of the resist is much faster than the PMMA version. This is due to the chlorine in the ZEP polymer acting as a secondary electron generator, which speeds up the write time of the resist. Write times of these resists can be further improved by adding additional secondary electron generators such as diallylamine.

Samples were etched using the silicon pseudoBosch etch described in section 3.4.2.1 below. This etch as a bulk silicon etch rate of 250nm per minute. The resulting etch rate of the composite resist film (as measured by scanning electron microscopy) is shown in figure 3.8. Comparing the etch rate of this film (17.4nm per minute) to that of silicon (240nm per minute), the etch selectivity over silicon was found to be 13.8:1. This is even more impressive, increasing the etch resistance of ZEP by 4 times. Resulting patterns of writing and etching the PMMA-Cr_7Ni ring composite resist are

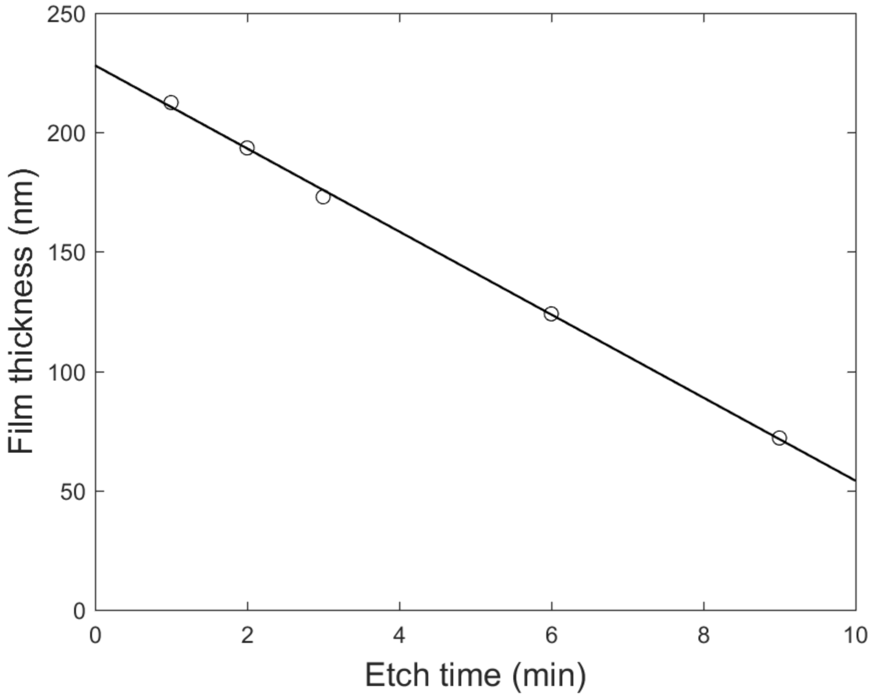

Figure 3.8: Film thickness during successive etching of the PMMA-Cr_7Ni ring composite film. The etch rate was found to be 17.4 nm per minute (R^2 = .999).

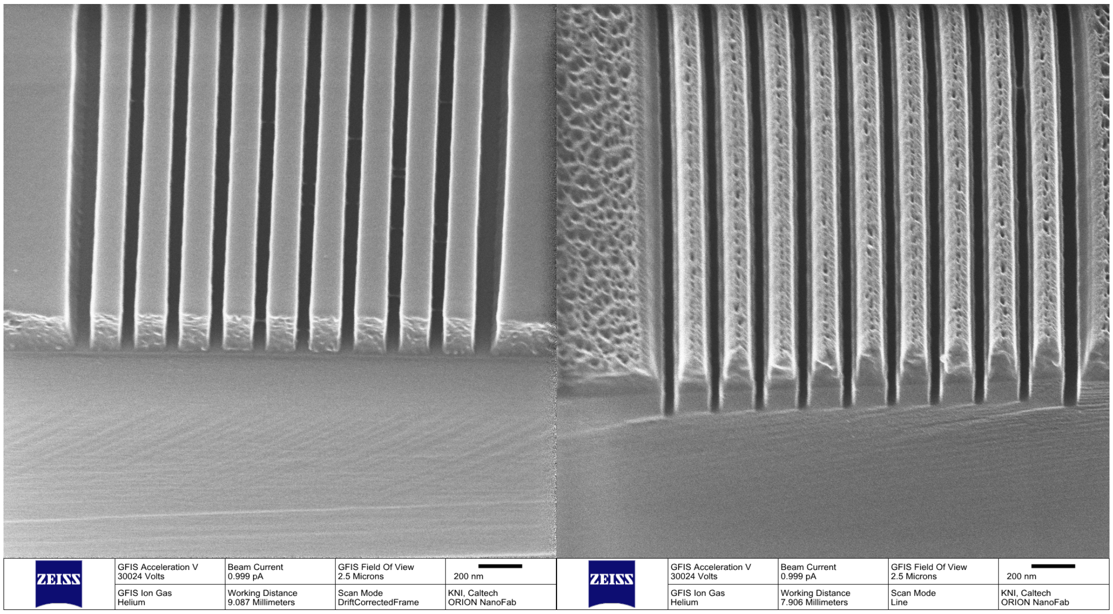
GFIS Acceleration V 30024 Volts
Beam Current 0.999 pA
GFIS Field Of View 2.5 Microns
200 nm
GFIS Ion Gas Helium
Working Distance 9.087 Millimeters
Scan Mode DriftCorrectedFrame
KNI, Caltech ORION NanoFab
GFIS Acceleration V 30024 Volts
Beam Current 0.999 pA
GFIS Field Of View 2.5 Microns
200 nm
GFIS Ion Gas Helium
Working Distance 7.906 Millimeters
Scan Mode Line
KNI, Caltech ORION NanoFab

Figure 3.9: Electron micrographs of ZEP-Cr_7Ni ring composite resist before etching (left) and after etching (right).

shown in figure 3.9.

3.4 Pattern Transfer: Plasma Etching

Heretofore the discussion has focused on creating patterns on a given substrate. However, the purpose for this has not been discussed. The next step in any fabrication process after a resist has been patterned and developed is called pattern transfer, wherein the pattern on the resist is copied to the substrate in some way (via etching, deposition, impression, etc.). This section is concerned with pattern transfer via reactive ion etching in plasmas. This is a common technique which utilizes the combination of chemically activated species created by a plasma (ions and free radicals) with the physical properties of accelerated ions to create highly anisotropic etch profiles. A theoretical discussion of plasma processing can be found in appendix A. This is advantageous in many process steps, as will be shown by example in the following chapters.

3.4.1 Etch Process: Bosch

The Bosch Process is specifically designed for etching high aspect ratio features in silicon. This etch is often used for MEMS devices as well as through wafer vias. This is accomplished by alternating between two etch recipes, one that etches silicon quasi-isotropically (using SF_6 gas), and one that deposits a film of Teflon-like polymer on the features being etched. The Teflon-like polymer derived from C_4F_8 gas is used to preferentially protect the sidewalls of the structure being etched, while allowing etching downwards. This effect is caused by the differences etch rates between the sidewalls and the bottom of the channel. Since the etch process occurs at low pressures, the flux of ions in the etch chamber will be very directional, thus the flux of ions striking the bottom of the channel will be orders of magnitude greater than the flux of ions hitting the sidewalls. Therefore, due to ion-assisted etching, the etch rate of the polymer will be high at the bottom of the channel, but very slow along the sidewalls. This leaves the sidewalls protected while allowing downward etching at the bottom of the channel. Thus, by alternating sidewall protection and etching steps, it is possible to etch down quite far while maintaining sidewall integrity. The alternating of the two recipes does have one undesired effect in that the sidewalls will have ridged features, as shown in figure A.9. The parameters for this recipe are shown in the table below, utilizing an Oxford Instruments 380 ICP-RIE etcher (Oxford Instruments, Yatton, Bristol, UK). The etch rate of this etch is 1.02 microns per minute.

	Strike	**Deposition**	**Etch**
Pressure:	5 mTorr	23 mTorr	23 mTorr
Temperature:	20^oC	20^oC	20^oC
ICP Power:	1500 W	1500 W	1500 W
Forward Power:	250 W	10 W	30 W
DC Bias:	-	88 V	128 V
Time:	3 s	9 s	14 s
Gasses:	SF_6 = 150 sccm	C_4F_8 = 140 sccm	SF_6 = 150 sccm

3.4.2 Etch Process: Pseudo-Bosch

The so called "pseudo-Bosch" process is an etch process which, similarly to the Bosch process described above, utilizes SF_6 and C_4F_8 gasses during the etch process. The differentiator between this etch and the Bosch process is that in pseudo-Bosch, the gasses are both utilized concurrently in the plasma. Thus, there is a subtle interplay between the etching by the SF_6 plasma products and the sidewall protection by the C_4F_8 plasma products. This results in the producting of high aspect ratio feature definition (similar to the Bosch process), while leaving smooth sidewalls. Smooth sidewalls are advantageous when etching devices like optical waveguides where sidewall roughness induces performance degradation of the device. Furthermore, pseudo-Bosch can etch a wide variety of substrates such as SiO_2 and Si_3N_4, whereas the Bosch process is usually only effective when etching silicon.

There are a few downsides to using the pseudo-Bosch etch. First, the etch rate is much slower (typically 1/4 to 1/10th the etch rate of Bosch) and the maximum feature depth is much shallower than the Bosch process (approximately 10 microns vs 100's of microns for Bosch). Secondly, the sidewalls produced by the pseudo-Bosch etch will typically be slanted to some degree, although this angle can be quite low (even vertical) when the etch is well optimized. The angle of the sidewall slant, however, can be tuned and is often advantageous, as will be discussed later. Multiple pseudo-Bosch recipes are employed in the fabrication of devices described in this thesis. These recipes are described below and utilized a Oxford Instruments 380 ICP-RIE etcher (Oxford Instruments, Yatton, Bristol, UK).

3.4.2.1 Pseudo-Bosch processes for Silicon

The recipe and parameters for the straight sidewall silicon pseudo-Bosch recipe is detailed below. Typically the etch rate for this process in silicon is 250nm per minute.

Pressure:	10 mTorr
Temperature:	15^oC
ICP Power:	1200W (strike: 1500W)
RIE Power:	23W (strike: 50W)
DC Bias:	70V
Gasses:	SF_6 = 42 sccm
	C_4F_8 = 55 sccm

Etch Process: Tapered Pseudo-Bosch Purposely tuning the sidewall angle of an etch process is a valuable tool in process design. Although often straight sidewall profiles are desired, this is often not the case when considering trench filling. As will be discussed later, during deposition there typically is preferential growth at the top of high aspect ratio trenches. Since one of the goals of this research is to fill narrow, high aspect ratio trenches, it was desirable to create an etch with an angled sidewall profile such that the filling of the trench will start from the narrowest point at the bottom of the channel, and fill from the bottom. This will ensure that the channel fills evenly and completely. The recipe below accomplishes this by increasing the concentration of C_4F_8 in the plasma chamber, thereby increasing the ratio of passivation to etching. It is possible to accomplish this by increasing the ICP power as well, however the following recipe was found to work best.

Pressure:	10 mTorr (Strike: 5mTorr)
Temperature:	15^oC
ICP Power:	1200W (strike: 1500W)
RIE Power:	23W (strike: 50W)
DC Bias:	80V
Gasses:	SF_6 = 42 sccm
	C_4F_8 = 70 sccm

3.4.2.2 Pseudo-Bosch process for Silicon Nitride films

The following etch recipes were designed to minimize sidewall roughness when etching into silicon nitride films. Since optical devices were created from LPCVD silicon nitride, it is crucial to minimize any sidewall roughness, as roughness will translate into propagation losses in the silicon nitride waveguides. Different etches were designed for different resists due to the positive vs. negative tone property of each resist. The amount of exposed silicon nitride and resist influences the properties of the etch gas, and thereby requires the development of different recipes.

Pseudo-Bosch process for Si_3N_4-ZEP resist

Pressure:	10 mTorr
Temperature:	15^{o}C
ICP Power:	1300W (Strike: 75W)
RIE Power:	23W (Strike: 1750W)
DC Bias:	75V
Gasses:	SF_6 = 37 sccm
	C_4F_8 = 53 sccm

Pseudo-Bosch process for Si$_3$N$_4$-Ma-N resist

Pressure:	10 mTorr
Temperature:	15^{o}C
ICP Power:	1200W (strike: 1500W)
RIE Power:	15W (strike: 50W)
DC Bias:	
Gasses:	SF_6 = 12 sccm
	C_4F_8 = 20 sccm

3.4.2.3 Pseudo-Bosch process for silicon dioxide films

The silicon dioxide pseudo-Bosch recipe is described below.

Pressure:	10 mTorr
Temperature:	15^{o}C
ICP Power:	1200W (strike: 1500W)
RIE Power:	23W (strike: 50W)
DC Bias:	70V
Gasses:	SF_6 = 42 sccm
	C_4F_8 = 55 sccm

Etch Process: Tapered Pseudo-Bosch Pseudo-bosch can also be used to etch silicon dioxide films, albeit at a very slow rate (appx. 20 nm per minute).

Pressure:	10 mTorr
Temperature:	15^{o}C
ICP Power:	1200W (strike: 1500W)
RIE Power:	23W (strike: 50W)
DC Bias:	90V
Gasses:	SF_6 = 42 sccm
	C_4F_8 = 70 sccm

3.4.3 Isotropic silicon etch

The following etch was designed to create a nearly isotropic etch of silicon for creating buried microfluidic channels. This recipe achieves an isotropic profile due to a higher pressure in the chamber (inducing further gas scattering) and a reduced forward power, thereby reducing ion assisted etching and creating a more chemical etch profile. An etch rate of 1.15 microns per minute was found using an oxide carrier wafer.

Pressure:	35 mTorr
Temperature:	15^{o}C
ICP Power:	2500 W
RIE Power:	10 W
DC Bias:	50V
Gas:	$SF_6 = 100$ sccm

3.4.4 Fast silicon dioxide etch

Anisotropically etching silicon dioxide films using pseudo-Bosch is a difficult task due to the extremely slow etch rates. The processes described in this thesis often require etching through multiple microns of silicon dioxide and thus necessitates a faster etch process. The following etch was designed to be highly anisotropic due to a high forward power (inducing high rates of ion assisted etching) and the choice of gas (C_4F_8 can be used to passivate the sidewalls of the trench as well as etch in this case). The etch rate was found to be 250nm per minute.

Pressure:	8 mTorr
Temperature:	20^{o}C
ICP Power:	2200W
RIE Power:	150W
DC Bias:	175V
Gasses:	$O_2 = 5$ sccm
	$C_4F_8 = 70$ sccm

3.4.5 Parylene Etching

This is the only etch utilizing a purely capacitive etch system (Plasmatherm SLR-720 RIE, Plasmatherm, St. Petersburg, FL). Oxygen is a strong etchant of any polymer film, as carbon is highly reactive with oxygen to create CO_2. This etch was used to remove parylene from inlets and outlets of microfluidic devices, as well as clear parylene away from optical structures.

Pressure:	140 mTorr
RIE Power:	80 W
DC Bias:	70V
Gasses:	O_2 = 30 sccm
	CF_4 = 7 sccm

3.5 Pattern Transfer: Wet Etching

3.5.1 Buffered Oxide Etch

Hydrofluoric acid is a weak acid which, in solution, will etch silica glass compounds (including pure SiO_2). Typically pure HF is dangerous to handle, is difficult to mask, and has a very high etch rate of silicon (typically greater than 2μm per minute)[129]. Thus, to moderate the etch rate of the solution and to improve photoresist masking, a 5:1 buffered HF solution with NH_4F is instead used, often called buffered oxide etch. The reaction between the silicon dioxide surface and the buffered solution is often cited as

$$SiO_2 + 4HF + 2NH_4F \rightarrow (NH_4)_2SiF_6 + 2H_2O. \tag{3.2}$$

This solution has a nominal etch rate of 100nm per minute of thermal oxide, and a slightly higher etch rate for PECVD oxides around 150-200nm per minute.

3.5.2 Aluminum Oxide Etch

Aluminum oxide, either sputtered or evaporated, can be removed utilizing a simple solution known colloquially as RCA-1. Typically used for cleaning organic compounds from wafers, it also acts to etch alumina films. RCA-1 is composed of a mixture of 5 parts de-ionized water, 1 part ammonia (29% dissolved in water), and 1 part hydrogen peroxide (30% dissolved in water). This etch is often used to remove alumina after use as a hard mask, and has an etch rate of approximately 10nm per minute.

3.6 Deposition of Thin Films

The previous section described the etching of a number of materials and thin films common to semiconductor processing. However, the means by which these materials are created has yet to be addressed. The substrate used for all of these processes is a bare silicon wafer, which has had a number of films produced on top of it. This series of films is often called a “stack,” and is the precursor for the devices being fabricated. In this section, the relevant practice of thin film deposition will be discussed, with a focus on depositing conformal films as is relevant to trench filling for the process used in this thesis. Theoretical analysis of thin film deposition with an emphasis on conformal processing can be found in appendix A.

3.6.1 LPCVD silicon nitride deposition

Low-pressure chemical vapor deposition is a common technique for creating stoichiometric or near-stoichiometric thin films of various insulators, including silicon nitride. In the case of LPCVD, high temperatures are required to overcome the activation energy barrier to convert gaseous precursors into a solid on the surface of a wafer. Reactors are often defined by the pressure of the gasses (in the case of LPCVD, the low pressure indicates that the samples are under vacuum) and whether the reactor has hot walls (in which better wafer uniformity is traded for allowing material to deposit on the walls of the reactor) and cold walls (where the reactor walls remain clean but film uniformity is sacrificed). For optical quality thin films as those desired for the devices in this thesis, hot wall reactors are required to maintain good wafer uniformity. Typically deposited between 700^oC and 800^oC, the reaction for LPCVD nitride is

$$3SiCl_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2. \tag{3.3}$$

This reaction produces a film with little impurities, and optical properties identical to single crystal silicon nitride. Due to the need for a high quality film and the capabilities of the Kavli Nanoscience Cleanroom at Caltech, thin films were prepared off site at Rogue Valley Microdevices in Medford Oregon.

3.6.2 PECVD oxide deposition

Plasma enhanced chemical vapor deposition (PECVD) encompasses a large variety of thin film deposition processes, but with the common theme that a plasma is used to supply the dissociative energy to the precursor gas, allowing for lower temperature deposition. In semiconductor processing, low temperature deposition is critical for multiple reasons, including minimization of dopant diffusion in semiconductor structures and stress minimization in MEMS devices.

Choosing PECVD oxide deposition was a necessity for optical neural probe devices to minimize the stress between the different layers in the probe stack. Furthermore, PECVD silicon dioxide produces a less dense film due to a higher hydrogen content (from the precursor) when compared to hot deposited SiO_2, making it easier to remove from optical devices using buffered hydrofluoric acid.

The construction of a PECVD reactor is typically a capacitively coupled plasma reactor, however the gas inlet to the chamber is distributed over a larger area. The gas distribution nozzles are typically distributed over the breadth of the reactor to create a more uniform deposition profile. Within the chamber, the wafer sits on an adjustable heated table to control the deposition rate of the film (and thereby control the film's properties). The precursor gas is flowed in through the nozzle array, and the plasma applied to the wafer. The major difference between an etching system and a deposition system is that the plasma products have very low vapor pressures at the given wafer temperature, thereby allowing a film to grow on the surface (as opposed to the etching case where a high vapor pressure is desired to remove the etch products from the surface).

3.6.2.1 Precursors

The choice of silicon precursors is critical when understanding the properties of the deposited film. There are two common types of precursors, inorganic and organic precursors, examples of which are described below.

Silane Silane (SiH_4) is the most common precursor for silicon, silicon dioxide and silicon nitride deposition. However, when depositing a SiO_2 film, an oxidizer must be added to the reaction mixture. Typically, this can be any oxidizer with a presence of oxygen, however N_2O is the most commonly used oxidizer with silane since directly mixing oxygen gas with silane tends to form particulates very quickly.

$$SiH_4 + 2N_2O \rightarrow SiO_2 + 2H_2 + 2N_2. \tag{3.4}$$

The main disadvantage of using silane as a source for silicon atoms is that it proves to have a very high sticking coefficient, resulting in poor step coverage. Since the goal of many of these projects is to create microfluidic channels via trench filling, silane is not an ideal precursor for creating silicon neural probes with microfluidics.

TEOS The alternative to inorganic precursors are organic precursors, such as tetraethyl oxysilicate (TEOS). TEOS, with a chemical formula of $Si(CH_3CH_2O)$, has been shown to have superior properties to Silane-SiO$_2$. These properties include denser films with less residual hydrogen (although water will form in TEOS films and can be removed by annealing), lower-stress films, and superior trench filling properties[98]. TEOS is typically oxidized using oxygen gas resulting in the following chemical equation:

$$Si(OCH_2H_5)_4 + 6O_2 \rightarrow SiO_2 + 10H_2O + 8CO_2. \tag{3.5}$$

3.6.2.2 Step Coverage and Trench Filling

The key application for the use of deposited SiO_2 was intended to be the sealing of microfluidic channels. There are multiple considerations that are critical to step coverage and trench filling.

First, the geometry of the trench will greatly affect the filling of the trench. The higher the aspect ratio of the trench, the more difficult it will be to fill the trench without having a void in the center. This can be explained by line-source model, wherein a random point within the trench (if not the center point) will be exposed to a non-even flux from the left and right due to differential shadowing from the walls on either side of that point. This differential flux results in cusping or "breadloafing" of the top of the trench compared to the bottom, and is roughly described using the equation,

$$\Gamma_s \propto 1 - \cos(\theta_s), \tag{3.6}$$

where Γ_s is the deposition rate, and θ_s is the angle subtended on one point on the sidewall of the trench shadowed by the corner of the other side of the trench [33]. Thus, all things being equal, the directionality of the plasma deposition will cause a non-uniformity in the deposition of the oxide along the sidewalls of the trench. Although this is a fundamental problem with the geometry, if the geometry can be changed slightly it is possible to encourage better filling. To accomplish this, trenches were etched with a slanted sidewall. This increases the angle seen by the sidewall in the above equation, and encourages filling from the bottom of the trench instead of the top. An etch with tapered sidewalls was developed to improve the trench filling properties of the deposition process.

The second point to consider, which can compensate the effect of the directionality of the incoming flux, is the sticking coefficient of the precursor molecules. Discussed in detail above theoretically, the practical coefficients for deposition were not discussed above. Silane, the limiting factor in the deposition of silane-SiO_2, has a sticking coefficient in the range of 0.35 at typical PECVD flow rates and pressures [33]. In comparison, TEOS based SiO_2 deposition typically exhibits a lower sticking coefficient in the range of 10^{-3} range. The result of this difference in sticking coefficients when comparing PECVD silicon oxide deposition from TEOS and Silane is shown in figure 3.10. Unfortunately, once we had established a functioning process for the trench filling, our vendor for TEOS oxide deposition (Noel Technologies, Campbell CA.) discontinued their TEOS-SiO_2 process and we were forced to start using an alternative technology for trench filling, parylene.

3.6.3 Parylene deposition

Parylene is a polymeric thin film that is deposited via pyrolysis of a dimer precursor. Parylene was originally utilized as a moisture barrier for electronic circuits predominantly due to the fact that it deposits on a room-temperature substrate and that parylene creates a highly conformal layer over high aspect ratio features. Later it was discovered that parylene is highly biocompatible[77], making it an ideal compound for coating and creating biological implants. For example, parylene has been used to make biocompatible microfluidic devices[71], drug delivery[62], and even implantable neural probes[56], among many other bioMEMS applications.

Parylene starts as a cyclical dimer, each monomer consisting of a p-xylene aromatic ring. Different precursors are available with different functional groups on the xylene ring. The most common parylene precursor used for biological applications is parylene-C, which has a single chlorine group substitution on the aromatic ring. The dimer is placed in an oven which is heated to 200^oC, above the sublimation temperature of the dimer. The sublimated vapor is passed into a 650^oC pyrolysis furnace at pressures less than 1 Torr, where the stressed bonds between the two monomers break and the parylene monomer goes through a bond rearrangement forming p-xylylene. The vapor is finally passed over the sample at a temperature of less than 30^oC, where the monomer deposits on the surface, diffuses and reacts with other adsorbed monomers, forming a film of poly-para-xylylene. Parylene has a sticking coefficient on the order of 10^{-3}, similar to TEOS-silane[31], leading to a highly conformal coating.

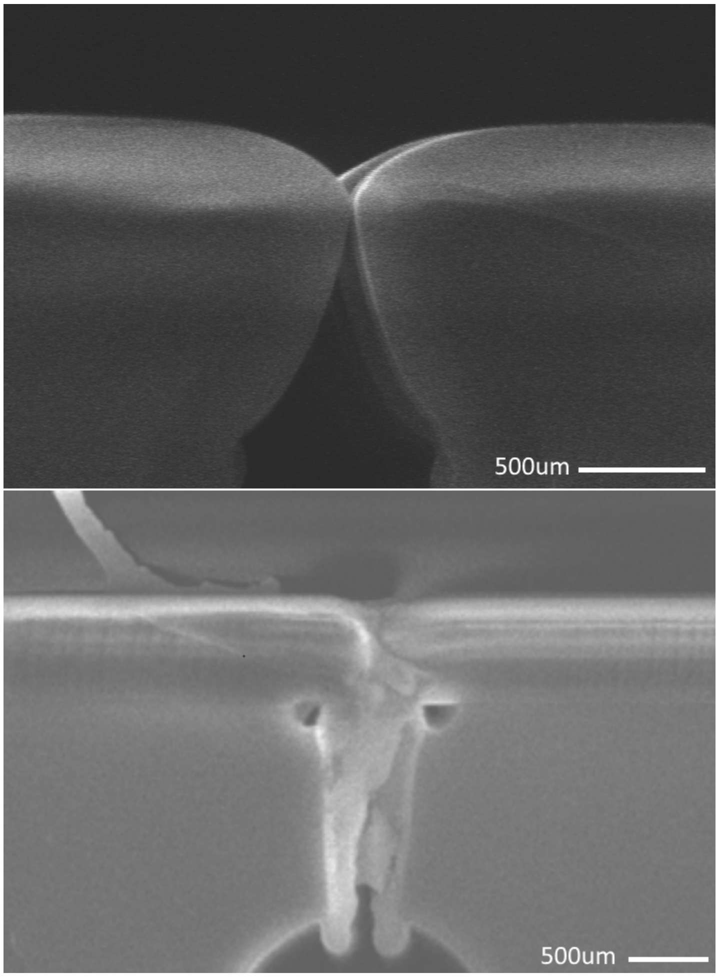

Figure 3.10: Top, silane deposited SiO_2 cusping during trench filling. Bottom, TEOS deposited SiO_2 of trench with same geometry, properly filled.

3.6.4 Electron Beam Evaporation

Evaporation of compounds is a very common technique in thin film coatings. All solid (or liquid) compounds, when heated to a sufficient temperature, will release component atoms to the vapor phase. Under controlled conditions and in an evacuated chamber, the vapor will travel in a straight line from the source outward from the sample forming a plume (semicircular in cross-section) from the evaporation source. There are many methods for evaporation of compounds. The container holding the source material can be heated directly as in simple thermal evaporation (although this technique requires the container to have a much higher vapor pressure and melting point than the source material). However, the most common method for evaporation of materials is electron beam evaporation.

Electron beam evaporation is a technique where a electron emission source such as a tungsten filament is heated until the electrons in the filament overcome the work function of tungsten and are emitted from the filament. These electrons, being charged particles, are then attracted towards a crucible containing the source material using an electric field. The electrons strike the source material, and induce heating of the material through Joule heating in the source material. As the source material is heated to temperatures in excess of 1000^{o}C, the source material will begin to evaporate. Since the electron beam hits the surface of the source material and the crucible is cooled from the backside, there is minimal contamination from the crucible due to crucible heating. Also due to this configuration, it is possible to evaporate materials with high melting points such as ceramics as the temperature of the surface of the source material will always be much higher than the temperature of the crucible. Evaporation rates are controllable within an approximate range of .1 - 100nm per minute. The sample to be coated is placed within the path of the plume, typically between 30 and 100cm away from the source material. This allows for the evaporated material to cool partially before being deposited on the surface of the sample. Samples are typically mounted on a semi-spherical holder so that all the samples are the same distance from the evaporating sample, thereby ensuring uniformity of the deposition. Furthermore, the samples are usually rotated to improve uniformity.

The only electron beam evaporation process used in the fabrication of neural probes is for the deposition of an Al_2O_3 (Alumina) hard mask for deep etch processing. Due to long etch times and material choices, there are certain processing steps which a photoresist mask will not hold up for the duration of the etch. This includes the deep etch process to define the base of the neural probes and etching through hard materials such as silicon dioxide. Alumina provides etch resistances over silicon in the range of 100:1[44]. Alumina has a density of 3970 kg per m^2, a 10^{-4} vapor pressure at 1550^{o}C, and a melting point of 2072^{o}C[57].

Alumina was deposited on the wafers using a liftoff process as described above. After S1813 photoresist had been coated and patterned on the surface, the wafer was loaded into a Temescal

FC-1800 electron beam evaporator and pumped to a pressure of at least 5.00×10^{-6} torr. After reaching pressure, the electron gun was engaged and ramped up to a power of 8% over a period of 1 minute, after which the crucible was exposed to the beam at 8% power for an additional minute. This step is called the “soak”, and begins the heating of the alumina slowly. The power is ramped again, to 13% over the period of a minute, and again the alumina is exposed to this higher intensity electron beam for another minute. After this pre-deposit step, the shutter is moved away from the crucible and the wafer is exposed to the vapor from the crucible. The throw of the TES FC-1800 is about a half a meter. The deposition rate and thickness is controlled using a feedback controller connected to a quartz crystal monitor, and is set to a constant deposition rate of 1Å per minute. This typically results in a maximum power output of the electron beam of no more than 14.5%. The film is typically deposited to a thickness of 350-450nm depending on the process being run. Since the sample often gets very hot, which can lead to issues in the liftoff process if the photoresist undergoes thermal changes, the process is usually done in steps of 75nm deposition with 5 minute breaks between each deposition step. This mitigates any issues of heating of the sample.

3.7 Fabrication Protocols

3.7.1 Neural Probe Platform

The neural probe platform, on which all of the following devices are fabricated, is a critical component of the fabrication of all of the devices described in this thesis. This procedure was originally described in [90], and was modified for the purposes of photonic and microfluidic neural probes.

The substrate used for the fabrication of neural probes is a silicon-on-insulator (SOI) wafer. These wafers consist of a thick base layer of silicon 300μm thick called the handle, a 15μm thick layer of silicon on top called the device layer, and sandwiched between the two silicon layers is the buried oxide, a 1-2μm thick layer of silicon dioxide. This material stack sets the geometry of the probes being fabricated, as the device layer will form the implantable portion of the probe, often called the shank of the probe. The handle will only remain at the base of the probe as a structural support for handling the probes.

The SOI wafers were first oxidized to produce a silicon dioxide film of thickness of 1.5 or 2μm on both the front and back surfaces of the wafer. This was done by an outside vendor, Rogue Valley Microdevices (Medford, OR). This layer is used as an insulating layer for any structures which are to be fabricated on the top surface of the wafer. For example, silicon is highly absorbing of visible wavelength light, and thus any optical devices on the wafer must be isolated from the silicon device layer using a transparent film, such as SiO_2.

After oxidizing the wafer surface, any necessary layers were added to the top of the wafer, such as a layer of Si_3N_4 for optical devices. At this stage, fabrication of any structures on the top surface would be fabricated before continuing with the probe fabrication process, such as optical devices and microfluidic channels.

Next, wafers were coated with S1813 photoresist per the recipe described in section 3.1.2. The mask designed for this step is the negative image of the desired pattern as is necessary for a liftoff process. The resist is exposed for 8 seconds on an Karl Suss MA-6 mask aligner under vacuum contact to ensure high resolution image transfer. Following development of the resist, the wafer was placed in the TES FC-1800 electron beam evaporator to apply an alumina hard mask for etching. Alumina was deposited at a final power of 14.5% at a rate of 1Å per minute to a final desired thickness of 300nm. Actual thickness of deposition was typically greater than the desired thickness, typically 350-360nm. The wafer was left overnight in an acetone bath, after which the surface was gently brushed with a swab to remove any undesired Al_2O_3 remaining after the liftoff process.

Following the liftoff process, the wafer will now be etched from the front side. This etch defines the device layer of the SOI wafer, and thus is the basis for creating the shank structures for implantation. The front side layer stack will consist of SiO_2 on Silicon on SiO_2 at this stage of the fabrication process. The first stage is to etch through the top SiO_2 layer using either the C_4F_8/O_2 based silicon dioxide etch described in section 3.4.4 or to use the pseudo-Bosch recipe optimized for silicon dioxide etching described in section 3.4.2.3. The former has the benefit of having a much faster etch rate (appx. 250nm per minute) and produces a superior facet for edge coupled devices, while the latter has a slower etch rate (appx. 35nm per minute) but etches the hard mask much slower, allowing for a thinner layer of Al_2O_3 to be used. Using either recipe, the wafer is placed in the DRIE enabled Oxford ICP 380 tool in the KNI cleanroom and is etched for the requisite time to remove the top silicon oxide layer. The etch depth is monitored visually (color change from a colored film to silver is a good indicator that all of the oxide is removed) and by a profilometer, which uses a stylus to measure the relative thickness of the etched region versus the un-etched hard mask.

After removing the top silicon oxide, the silicon device layer is etched via the Bosch process described in section 3.4.1. The etch rate for the Bosch etch is typically 1μm per cycle. Since Bosch etching etches silicon dioxide at an extremely slow rate, the buried oxide layer is used as a hard etch stop, and a slight over etch using the Bosch process is tolerated. Thus, for a 15μm device layer 20 cycles are used to ensure that all of the silicon is removed before etching the buried oxide layer. This also ensures that there is not a slight overhang in the bottom silicon due to the scalloped shape of the sidewalls as a result of the Bosch process. Finally, the burried oxide layer is etched using the same SiO_2 etch as the top oxide, running either 35nm per minute for the pseudo-Bosch process or 250nm per minute for the C_4F_8/O_2 based silicon dioxide etch. The etch is intended to penetrate the entirety of the buried oxide layer to reduce overhanging oxide at the bottom of this layer. It is not uncommon for there to be overhanging due to sidewall non-uniformity at the base of the etched trench due to the depth of the trench itself. As any trench becomes deeper, there will always be a reduced etch rate at the corner created between the base of the substrate and the sidewall due to diffusion of the etch species. Since the handle layer of the SOI wafer will eventually be removed anyway, it is possible (and desirable) to over-etch the BOX layer into the underlying silicon. This provides better sidewall uniformity and a more reproducible result for the devices.

After defining the shank and base structure of the probes on the front side of the wafer, the aluminum oxide hard mask must be removed. This is accomplished by a wet-etch process in a solution of 1 part ammonium hydroxide (27% by mass in water), one part hydrogen peroxide (30% by mass in water) and 5 parts water, also known colloquially as RCA-1. This solution is heated to 60-70^{o}C, after which the wafer is submerged in this solution for 30 minutes. Once the aluminum oxide hard mask has been removed, the wafer is rinsed in deionized water and dried by pressurized N_2.

To define the base of the probe from the handle layer of the SOI wafer, the back side of the wafer must be patterned. The backside mask includes a duplicate of the top-side trench; however instead of patterning shanks the backside mask simply has an open trench. This will allow the creation of cantilevered, 15+μm shanks from the device layer. Feature definition is accomplished using the Karl Suss MA-6 mask aligner in the KNI cleanroom. These tools are equipped with underside facing cameras which allow for the alignment of masks to features on the opposite side of the wafer. Wafers were coated with S1813 photoresist per the recipe described in section 3.1.2 on the unpatterned side of the wafer. The mask plate is loaded into the tool and using the backside alignment microscope, the alignment marks on the mask are found and lined up on the MA-6 monitoring screen. At this point it is critical not to move the mask plate or the backside microscope, as the tool takes a still image of the alignment marks which will be used as an overlay once the wafer has been placed in tool. After taking the still image, the wafer then can be loaded into the tool such that the wafer is patterned side down. Often a glass wafer is placed between the chuck and the front side of the wafer to protect the features already on the frontside of the wafer. The backside microscope is then used to locate the alignment mark features on the front side of the wafer corresponding to the mask being used. Finally, the stage micrometers are used to align the still image of the mask alignment marks to the corresponding alignment marks on the front of the wafer. The wafer is then imaged, in the case of S1813 it is imaged for 8s at 25mW per cm^2 (dose of 200mJ per cm^2). The wafer is then developed in CD-26 for one minute.

Following backside exposure, the frontside of the wafer is protected with a plastic film was placed in the TES FC-1800 electron beam evaporator to apply an alumina hard mask for etching to the backside of the wafer. Alumina was deposited at a final power of 14.5% at a rate of 1Å per minute to a final desired thickness of 300nm. Actual thickness of deposition was typically greater than the desired thickness, typically 350-360nm. The wafer was left overnight in an acetone bath, after which the surface was gently brushed with a swab to remove any undesired Al_2O_3 remaining after the liftoff process.

Finally, the backside of the wafer may now be etched to produce released probes. The frontside of the wafer is first protected by spinning a thick layer of PMMA A11 on to the front side and baking for 2 minutes at 180^{o}C. The wafer is then mounted on the carrier using a generous amount of Santovac grease to ensure good thermal contact between the device wafer and the carrier wafer. The wafer is then etched first using the pseudoBosch recipe described in section 3.4.2.3. This etch will remove the top Si_3N_4 layer (if necessary) as well as the underlying SiO_2 layer. Etching is monitored using a profilometer, which will give an approximate etch depth as the etch proceeds. Given an etch rate of 35nm per minute, this etch will require approximately 43 minutes to penetrate the 1500nm SiO_2 film used on wafers designed for 472 and 405nm light, and will take approximately 58 minutes to penetrate the 2000nm films used by wafers designed for 670nm light. This is typically done in 10 minute increments to minimize wafer heating (and thereby etch rate drift) during the etch process, and is slowed to shorter increments when nearing the point of etch-through to the next layer.

After removing the top silicon oxide, the silicon handle layer is etched via the Bosch process described in section 3.4.1. The etch rate for the Bosch etch is typically 1μm per cycle. Thus, for a 300μm handle layer approximately 300 cycles are used to ensure that all of the silicon is removed. In reality, it often takes 20-50 more cycles than expected as the etch rate slows as the wafer is thinned. This could be due to changes in thermal conductance as the wafer is thinned, but the mechanism is unknown. Typically, the etch is done procedurally starting with 2x 100 cycle rounds of etching, followed by an 80 cycle round, followed by 10 cycle rounds of etching until the silicon behind the shanks has been completely removed. Unfortunately, due to the depth of the trenches, it is impossible to use the profilometer to monitor etch depth, thus inspection under microscope is the only method to determine when the etch is completed.

After completing the etch, the wafer is released from the carrier wafer by overnight soaking in an acetone bath. Acetone is mild enough not to damage any of the optical structures or the parylene used to cover the microfluidics. After soaking overnight, the wafer is submerged in isopropyl alcohol to remove any leftover acetone (which if improperly removed can leave an organic film on the wafer), and gently dried using compressed nitrogen gas.

3.7.2 Optical Devices

Two protocols were developed to create optical devices in silicon nitride. One technique was developed using ZEP-520a e-beam resist, another using Ma-N 2403. Two recipes were developed to address different needs for the optical devices. ZEP-520a is a positive tone resist, thus requiring longer write times due to the requirement to write the region around waveguides instead of the waveguides themselves. ZEP-520a, however, has superior etch resistance and line edge roughness, which results in superior smoothness of optical waveguide structures. Ma-N 2403 is a negative tone resist which allows for much faster writing, but has weaker etch resistance and line edge roughness, and thus is reserved for larger structures which take a long time to expose.

Process involving ZEP-520a Wafers are first prepared by baking for 2 minutes at 180^oC to remove any water vapor on the silicon nitride surface. ZEP-520a is applied to the wafer surface using a plastic pipette, and is spun at 4500 rpm for 30 seconds. The wafer is then baked at 180^oC for 2 minutes. Wafers are placed in the EBPG-5000+ electron beam lithography tool, and patterns are exposed with a dose of 280μC per cm^2. Fine features were exposed using a 300pA beam with a beam step resolution of 2.5nm. Coarse features (such as alignment marks) were exposed using a 100nA beam and a 50nm beam step resolution. Fracturing of these features used a bulk-and-sleeve method to improve side wall roughness, especially along the edges of ring resonators. Fracture optimization is critical for creating smooth curved structures.

Features were developed in a bath of ZED-N50 proprietary developer (amyl acetate plus additives). To further improve sidewall roughness before etching, the wafer is baked after development to reflow the resist. The temperature used in the reflow process is critical, as the goal is to slightly liquefy the resist to smooth out any high frequency roughness in the resist while not damaging the overall structures defined in the resist. The optimal reflow temperature used for ZEP-520a at this thickness was found to be 152^oC, and the reflow process is done for 1 minute on the hotplate.

After reflow, the sample is etched in the Oxford Instruments 380 ICP plasma etcher designated for clean samples only, including silicon and silicon nitride etching. After preconditioning the chamber with the pseudoBosch recipe described in section 3.4.2.2, the sample is placed on the carrier wafer and is processed using this recipe optimized for sidewall smoothness using a positive tone resist. The etch rate of this process is 75nm per minute, thus for devices utilizing a 200nm Si_3N_4 film (designed for 405 and 480nm light), the total etch time is 2 minutes and 40 seconds. For devices utilizing a 300nm Si_3N_4 film (designed for 678nm light), the total etch time is 4 minutes. After etching, the sample is exposed to a gentle oxygen plasma to remove any additional resist.

Process involving Ma-N 2403 A wafer is first subjected to a 5 minute bath in Nanostrip at room temperature, and is then rinsed with deionized water and dried with compressed nitrogen gas. The wafer is then dipped in buffered oxide etch for 10 seconds to remove any native oxide on the surface of the wafer, and is again rinsed with deionized water and dried with compressed nitrogen gas. Immediately after drying, the wafer is placed on a 180^oC hotplate for at least 2 minutes to drive away any additional water on the surface of the wafer. After baking, the wafer is then placed in a container where HDMS is applied in the vapor phase.

Next, Ma-N 2403 is applied to the surface of the wafer using a pipette, is generously spread across the wafer using the pipette, and is spun at 3500 RPM for 30 seconds. The resist is then baked for 1 minute at 90^oC. Wafers are placed in the EBPG-5000+ electron beam lithography tool, and patterns are exposed with a dose of 250μC per cm^2. Fine features are exposed using a 500pA beam with a beam step resolution of 2.5nm. Coarse features (such as alignment marks) are exposed using a 100nA beam and a 50nm beam step resolution.

Features are developed in a bath of Ma-D 520 proprietary developer for 1 minute, and rinsed with deionized water and dried with compressed nitrogen gas. Wafers are developed upside down to mitigate surface redeposition of the development products. To further improve sidewall roughness before etching, the wafer is baked after development to reflow the resist. Wafers are reflowed in a laboratory oven at 145^oC.

After reflow, the sample is etched in the Oxford Instruments 380 ICP plasma etcher designated for clean samples only, including silicon and silicon nitride etching. After preconditioning the chamber with the pseudoBosch recipe described in section 3.4.2.2, the sample is placed on the carrier wafer and is processed using this recipe optimized for sidewall smoothness using a positive tone resist. The etch rate of this process is 40nm per minute, thus for devices utilizing a 200nm Si_3N_4 film (designed for 405 and 480nm light), the total etch time is 5 minutes. The etch rate for this recipe is likely slower due to higher plasma loading caused by the properties of the negative tone resist used for these devices.

Cladding optical devices Following etching, samples are prepared for cladding with PECVD silicon oxide. This requires a rigorous cleaning procedure to remove any impurities from the wafer surface. The first step is bathing the wafer in a proprietary solvent, Remover PG, which is designed to remove residual resists from wafer surfaces by swelling and dissolving any hydrophobic materials on the wafer. Following the solvent clean, the wafer is rinsed in acetone and isopropanol and finally is dried by compressed nitrogen gas. Following the solvent clean, the wafer was submerged in a solution of 5 parts deionized water, 1 part 27% hydrogen peroxide, 1 part 30% ammonium hydroxide, known colloquially as RCA-1. This will further remove any organic residue from the wafers. After submerging the wafer in this solution (heated to at least 60^oC for 10 minutes), the wafer is then rinsed in clean water. Following RCA-1 cleaning, the wafer is dipped in buffered oxide etch for 10 seconds to remove any native oxide which is created upon exposure to hydrogen peroxide. Finally, the wafer is exposed to a solution of 6 parts deionized water, 1 part 27% hydrogen peroxide, and 1 part 37% hydrochloric acid, at 60^oC for 10 minutes. This clean, so called RCA-2, removes any metallic impurities from the surface. The wafer is rinsed with deionized water and is dried with compressed nitrogen gas. After cleaning, wafers were then placed in the Oxford Systems 100 Plasma Enhanced CVD chamber available in the KNI. The SiH_4-Ar precursor was used to produce an amorphous SiO_2 film on the surface of the wafer at a rate of 82.6nm per minute. Typically devices were clad for 18 minutes 10 seconds, resulting in a 1.5μm film on the surface of the wafer.

3.7.3 Silicon Microfluidics

The silicon microfluidics developed utilize a dig-and-seal method for fabrication. The alternative option for developing microfluidic probes was to create microfluidic channels via wafer bonding, where a channel is defined in the silicon substrate by etching, and a SiO_2 wafer is bonded to the silicon wafer via anodic bonding. This method, however, has many problems. First, thinning the top SiO_2 requires the use of precision chemical-mechanical polishing techniques which are unavailable in the KNI cleanroom at Caltech. Secondly, due to the temperature difference during the bonding (typically 550^oC), unknown stresses will be created at the interface between the silicon and SiO_2 wafers. Finally, the fabrication protocol to combine the microfluidic devices with the functional optical devices are incompatible. Thus, a dig-and-seal approach was utilized as it allows for co-integration of microfluidics with optical waveguide structures on the same probe.

Fabrication of microfluidic channels requires a 250nm silicon oxide layer on top of silicon. This silicon dioxide layer will eventually be used as an etch mask. SML2000 is spin coated on to the wafer surface at a velocity of 1500RPM for 45 seconds. This produces a film in thickness in excess of 3μm, sufficient to withstand pseudoBosch etching of approximately 20 minutes in duration. The wafer is then baked for 2 minutes at 180^oC to remove excess solvent from the film. The wafer is loaded into a Vistec EBPG 5000+ EBL tool, and the SML2000 film is exposed using a 100nA beam with a dose of $4000\mu C$ per square cm. The pattern consists of simple trenches of width 500nm which will define the path of the microfluidic channles; the resolution of features is set to 50nm per shot. After exposure, wafers are developed in a solution of 3:1 isopropyl alcohol:methyl isobutyl ketone for 1 minute.

The wafer is mounted to a 6 inch silicon carrier wafer with Santovac oil as a thermal contact solution and is inserted into a Oxford Instruments ICP 380 etching tool for pattern transfer. The etch described in section 3.4.2 for tapered sidewalls is used. The tapering in the sidewall cross-section is intended to improve deposition uniformity later in the process. To penetrate the silicon oxide layer, the etch is allowed to proceed for 11 minutes. Next, a short burst of the isotropic silicon etch described in section 3.4.3 for 30 seconds is applied to the wafer to prevent over-etching in later steps of the process. Finally, the tapered sidewall etch is continued for an additional 6 minutes to create the trench in the silicon layer.

After etching, the wafer is thoroughly cleaned using Cyantek Nanostrip (a stabilized solution of sulfuric acid and hydrogen peroxide) to remove any organic residue from the wafer. The wafer was left in the Nanostrip solution overnight. Nanostrip leaves a thin layer of dirty oxide on the surface of the wafer, so following the Nanostrip soak the wafer is dipped in buffered oxide etch for 10 seconds. Finally the wafer is cleaned in an oxygen plasma briefly to ensure that the surface of the wafer is clear of organic residue. The wafer was then oxidized in the Tystar Tytan furnace under dry oxygen conditions at 1000^oC for 45 minutes, creating a thin, conformal oxide film.

The wafer is then mounted on a 6 inch oxide coated silicon carrier wafer Santovac oil as a thermal contact solution. Oxide wafers are used in this step because they provide a more consistent etch profile when comparing with the test chips used to determine the etch process parameters that day. Again, the etch described in section 3.4.2 for tapered sidewalls is used to penetrate the bottom of the passivated channel while leaving the sidewall oxide in tact. This takes 1 minute and 45 seconds, typically. Finally, the isotropic silicon etch described in section 3.4.3 is applied to the wafer for 1 minute and 30 seconds, producing a round channel at the bottom of the trench with diameter approximately 3.5μm. Wafers are then removed from the carrier and cleaned thoroughly with acetone, isopropyl alcohol and dried with compressed nitrogen gas.

Channels are finally filled using parylene C, as described in section 3.6.3 above. Before coating, an adhesion promoter of γ-methacryloxypropyltrimethoxy silane was applied to the surface in a bath after dilution to 0.5% in a 50:50 solution of isopropyl alcohol and water. The silane was left overnight to bond to the wafer surface, after which the wafer was removed and allowed to air dry for 30 minutes. The wafer was then rinsed with isopropyl alcohol for 5 minutes, dried using compressed nitrogen gas, and baked at 115^{o}C for 30 minutes. Next, parylene is applied to the surface using a Paratech parylene coating system via sublimation and pyrolysis of the parylene-C precursor. The film was deposited to 1μm.

To open the inlets and outlets of the channels, SPR-220-7 was coated to a thickness of 6.5μm (4000 RPM) on the parylene coated wafer, and was exposed using vacuum contact lithography (Karl Suss MA-6) to a dose of 400μC per cm^2 at 365nm. The wafer was then bath developed in CD-26 for 60s and rinsed with water. This exposed the inlets and outlets of the microfluidic channels while leaving the remainder of the probes protected from etching. Etching was done in a capacitive RIE etcher (Plasmatherm SLR-720 RIE). Oxygen was flowed at 30sccm versus CF4 at 7sccm. Chamber pressure was set to 140mTorr and the forward power of the chamber was set to 80W. The parylene was etched for 15 minutes, which was the maximum possible etch resistance for the photoresist film. This allowed the inlets to be completely removed of any parylene, leaving an open channel. After the etch step, it was found that the channels were open and permitting flow.

Chapter 4

Optical stimulation and excitation of proteins using neural probes

4.1 Motivation

Optical stimulation of neurons is a transformative technology. Since the inception of modern neuroscience, the development of new technology has played a central role in the growth of the field. Beginning with the development of early instrumentation amplifiers for electrophysiology following the Second World War, recording and stimulating neurons has been at the core of modern electrophysiological instrumentation. Traditionally, measurements were taken using simple pipette-based electrodes but, as the field progressed, so did the desire to measure networks of neurons, especially in vivo. Eventually the microelectromechanical systems (MEMS) field provided an alternative and reproducible path to integrated circuit fabrication technology and production of high density arrays of extracellular neural probes. Notable achievements have been made by the Universities of Michigan and Utah[14, 75]. Although extracellular recording does not allow for the same richness of information provided by intracellular recordings (i.e., only spiking can be effectively measured using extracellular electrodes), they are still the most prevalent method for measurement of neural networks due to their scalability. Accordingly, these probes are in common use today in research labs for acute and chronic neural recording. Extracellular probes are able to independently record spikes from hundreds of neurons in the brain over periods of days or even months[124].

The recording of neural spiking *in vivo* is a well-validated methodology, leading to increased interest in the integration of novel/enhanced functionalities, one of the most compelling of which is *in situ* stimulation of neurons. Specifically, the goal is to perturb individual or small groups of nodes in working neural networks. This effort specifically targets this goal for which the potential impact is substantial and far reaching. Specifically, this technology will facilitate the ability to stimulate neural networks not only through external sensory perturbations in chronic studies, but instead to stimulate small neural networks and observe behavioral changes in the brain as a whole as well as the external behavior of the test subject[137, 139].

Due to the extracellular nature of implanted neural probes, it is impossible to effectively modulate neural activity in small, localized groups of specific cells by electrical means. Thus, the goal of the nascent field of optogenetics is to utilize optical means, taking advantage of its ability to interact with specifically-modified neurons. Over the past decade there has been a great interest in stimulating neurons using light-activated ion channels[137, 138]. Light-activated channels have the capability of not perturbing damaging neurons, while allowing for light-activated modulation of neuron membrane potentials. Furthermore, light-activated channels can be delivered to individual neuron types, allowing for specific activation of targeted neurons. A growing class of proteins has been used to modulate neural signals; however the most effective stimulatory protein to date is channelrhodopsin-2[137, 139, 10]. ChR-2 is a light stimulated cation channel which was discovered in the green algae C. reinhardtii by Nagel and Hegemann at the Max Plank Institute for Biophysics[104]. This species of algae is positively phototaxic, making use of the membrane spanning ChR-2 to sense the direction of impingent light. This protein has been inserted into rat hippocampal cells by Karl Deisseroth, Ed Boyden and colleagues at Stanford, using lentiviral vectors. Their work demonstrated convincingly that ChR-2 can be used to stimulate neurons optically[10]. The significance of ChR-2 modulated neural activity, when compared to other modalities, is that it is able to act at rates that are nearly commensurate with those of neural spiking (i.e., on the order of 10ms). Other photoactivated channels currently operate only at slower rates, on the order of seconds, given their biochemical properties. Thus, ChR-2 is the first protein to allow for physiologically relevant modulation of neural membrane potentials. Since then, there has been a concerted effort to create genetically-modified light-activated ion channels for research purposes; this has recently had much success. There are now tens of different proteins available for optogenetic excitation or inhibition of neurons, each optimized for a range of specific applications[137].

Biotechnology is outpacing probe technology. The burgeoning biotechnology behind the production of light activated channels has attained significant momentum. By contrast, progress on neural probes that can enable realization of the full potential of this optogenetic biotechnology lags far behind. Since the spatial organization of neural structures begets a neural networks functionality, it becomes critical for experimentalists to be able to spatially, as well as temporally, activate neurons arbitrarily. The current state-of-the-art for neural stimulation and detection was developed by the Diesseroth group[140]. In this method, an array is created from individual fiber tapers, the 6x6 grid of which are to be implanted into the brain of a rat or mouse. Each taper is coated with aluminum, which can be used to measure local spiking as well as local field potentials in the brain. Although this is the first multiplexed optical probe technology that combines multiplexed stimulation and spike detection in the same device, the major issue with this technique is a practical one, as the "optrode" array requires a 1:1 point of stimulation-to-fiber optic ratio. As the desired channel number increases these devices will become more cumbersome due to the consummate number of optical fibers needed. Furthermore, this technique does not lend itself to more advanced integration for the detection of neural activity allowed by a traditional neural probe may allow due to the bulky nature of the individual optrode. For example, it could be difficult to measure cells within a column of tissue because it is impossible to stack optrodes directly on top of one another, something that is trivial with a Michigan-style neural probe. The Boyden group demonstrated a neural probe with 12 optical waveguides fabricated on a simple silicon oxynitride shank[142]. Although this is a significant step forward toward integrating probes with directly-fabricated optical elements, it still requires the use of an individual source for each point of stimulation. The logical extension of this work and approach would appear to indicate that, to make this technology viable, many sources need to be integrated directly onto high-density probe arrays. This method is likely not scalable, however; placing LEDs or lasers on the base of the shank can create significant local heating. It is well understood that small changes in the temperature of order 0.1^{o}C of neural tissue can strongly bias spiking behavior. Furthermore, such an approach would take up significant "real estate" on the probe base, and would necessitate edge coupling of the emitted light directly into a waveguide this is a highly non-trivial task.

At the time the project described in this section was begun, the only commercially available technology utilized the crudest approach: an optical fiber that was simply glued to a standard neural probe[93]. Although this is sufficient for preliminary experimentation, it lacks any ability for spatial multiplexing and, especially, robust production for delivery to the neurophysiological research community.

For these reasons the author views it as highly advantageous, instead, to multiplex many channels of light outside of the neural probe, then to deliver it through a single optical fiber, and perform compact demultiplexing locally, right at the base of the shank. By using a single optical fiber to the neural probe which can control at least 32 or more independent points of stimulation, this technology becomes scalable in that a few optical fibers can then potentially stimulate arbitrarily large regions of the brain. Furthermore, it is important to look forward to integrating spatiotemporally-controllable optical stimulation with high-density electrical recordings to create a powerful experimental system for neuroscientists. The integrated optical technologies described below are compatible with the fabrication of state-of-the-art electrophysiological neural probes currently being produced. Here, a plan to is described for developing a system for the spatiotemporal modulation of optical neural stimulation using integrated optical wavelength-division demultiplexing.

Integrated optics and microfabrication techniques can allow for large scale spatial multiplexing of optical signals. A great deal of time and money has been invested in developing integrated optoelectronic technology, mostly driven by the telecommunications industry. Fiber-optic technology has been quite popular for some time due to its capability of accomodating the high data rates and low losses sustained by optical communications, and thus significant thought has gone into integrating fiber optics with silicon-based transmitters and receivers. Essentially, integrated optics is the technology enabling the manipulation of light on chips using microfabricated structures. There are direct parallels to microelectronics integrated circuit technology, in which electrons are manipulated. Currently there are a multitude of well-validated methods for coupling light on and off of silicon chips[2, 23], creating optical waveguides (the optical equivalent of electrical wires)[43], realizing multiplexers and demultiplexers[70], and optical resonators[70], to name only a few amongst the variety of applications and structures. Although much of this technology is designed for use in the telecommunications band (around 1550 nm), these technologies can be readily adapted for different wavelengths (necessary for optogenetics) by using alternative materials to silicon.

4.2 Application: Multiplexed Optogenetic Stimulation

Planar lightwave circuits (PLCs) can be used for on-shank wavelength-division demultiplexing. The basic unit of integrated optical structures is the waveguide. Similar to electrical wires, waveguides act as pipes to conduct light from one place to another. Created from non-absorbing dielectric materials, optical waveguides are most often produced by etching a thin film into long, rectangular slabs in which light can be confined using the principle of total internal reflection. This requires that the waveguide material has an index of refraction greater (ideally much greater) than that of the surrounding cladding. Air is commonly used as cladding, with an index of 1, so most optically clear dielectric materials can guide light effectively.

One of the most ubiquitous applications of PLCs in the telecommunications industry is wavelength-division multiplexing (WDM). WDM is an optical technique which divides a region of the electromagnetic spectrum into a number of channels which can be used to address a single or multiple devices assigned to that spectral segment[11]. Signals are multiplexed by assigning each signal to a channel where light in the corresponding spectral segment, modulating the amplitude of each channel, and combining all of the channels into a single optical fiber. At the terminus of the optical fiber another device is designed to break the signal into spectral components. The WDM strategy is advantageous in this case because channelrhodopsin-2 has a broad action spectrum, thus allowing for many wavelengths (spanning 450 to 500 nm at over 80% efficacy) to stimulate neurons. Furthermore, most WDM strategies are passive, reducing the complexity of the optical devices on the neural probe itself. The authors will utilize WDM to define a number of channels, each of which correspond to a specific point of stimulation, and combine those signals into a single fiber optic which will conduct that information to a neural probe, where the signals will be demultiplexed on the probe.

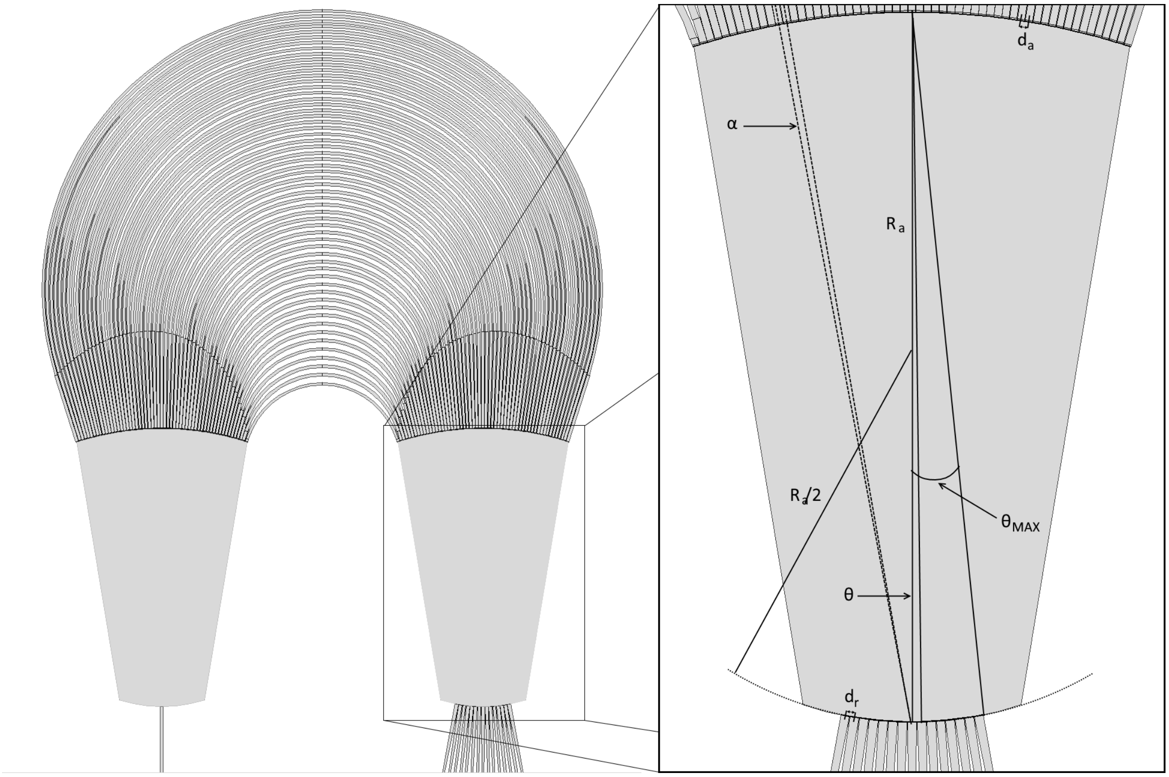

Figure 4.1: Array waveguide grating.

The technique chosen to accomplish the on-probe demultiplexing is the array waveguide grating (AWG). AWGs have excellent performance and are robust, simple to fabricate and used extensively in the telecommunications industry[105]. AWG technology is an integrated dense wavelength-division scheme which utilizes optical waveguides as delay lines to act as a grating and direct light to a number of receiver waveguides (similar conceptually to a phased array in radio communications). An example of an AWG can be seen in figure 4.1. The AWG device consists of three regions: two slab waveguides which are used for shaping of the input beam and divergence on the receiving end, and an array of waveguides, with each waveguide slightly longer than the previous. Light begins in the waveguide on the left and is emitted into the first multi-mode region of the AWG. The purpose of this region is to selectively illuminate the array of waveguides which form the grating such that they begin with the same phase. The light then propagates through a series of delays created in each subsequent waveguide, which act in a similar fashion to a diffraction grating. However, unlike a grating where the path length difference is induced by illuminating the facets of a grating at an angle, instead waveguides of slightly different lengths are used to create the delays. The end result is the same as in a traditional grating, where the phase of the light emerging from the grating produces an interference-induced wavelength dependent focus on the image plane of the spectrometer. When a secondary array of waveguides is placed at the focal plane of the AWG spectrometer the device acts as a wavelength-dependent demultiplexer. Thus by shaping (i.e., optically multiplexing) the spectrum delivered to the neural probe, individual waveguides (and thereby individual points of stimulation) can be selectively illuminated on the shank, providing spatial control of stimulation.

AWGs have been demonstrated in the near infrared telecommunication band (around 1550 nm) with hundreds of channels, and thousands when using cascaded AWG systems[116]. Visible wavelength AWGs were designed by the authors using the method described in [106] to assess the feasibility of producing high density AWGs practical for neural applications at around 485nm. Given a channel density of 1nm, the instrument imposed fabrication limit of 50nm minimum spacing between waveguides, 200nm x 360nm waveguides (determined via simulation), maximum cross talk of -3dB, and a 3dB roll off it was determined that AWGs up to at least 32 x 1nm channels are feasible at a central wavelength of 480nm with a bandwidth of 32nm, potentially more by using different geometries than the one chosen or by cascading multiple AWGs. It was also found that a typical AWG designed in this wavelength range has a small footprint, on the order of 300μm x 300μm, which will easily fit on the base of a neural probe.

Finally, the terminus of each waveguide can be implemented with a variety of structures depending on the desired application. The simplest terminus would be a blunt end. A blunt end projects light preferentially down along the waveguides central axis and is then diffracted outwards. The degree of diffraction depends on the size of the waveguide. Alternatively, the waveguide can be narrowed down to a point, allowing for an approximately spherical evanescent emission pattern from the tip of the waveguide. Facets can also be etched into the waveguide to direct light out away from the shank, or to the sides of the shank. Finally, grating structures can be used to diffract light terminating in the waveguide in a particular direction. The work described in this section was published in reference [96] and accomplished in collaboration primarily with Dr. Eran Segev, Roukes Group, Caltech.

4.2.1 Implementation of Array Waveguide Gratings for Visible Spectrum Light

Of primary concern to this project is the propagation of light contained in optical waveguides. Recalling equation 2.1, an estimate for the desired geometry of the waveguide at a given frequency can be calculated. Given a core index of refraction of 2 (silicon nitride) and a cladding refractive index of 1.45 (PECVD oxide), single mode propagation requires that the radius of a waveguide be less than

$$a = 2.4048 \frac{473\text{nm}}{2\pi\sqrt{2^2 - 1.45^2}} = 131\text{nm}. \tag{4.1}$$

Therefore, to include a margin of error, waveguides designed for single mode propagation will be designed to have a rectangular cross section with height 200nm and width 240nm when leaving the taper of the grating coupler. However, once single mode propagation is established, it is possible to widen the waveguide to improve losses. Thus, for the AWG, the waveguide widths were widened to 600nm. Given this geometry, the effective index can be calculated similarly to the procedure described in section 2.1.1. Given a central wavelength of 473nm (corresponding to the maximum of the channelrhodopsin action spectrum), the effective index of the waveguide is approximately 1.82 for the TE mode favored by the grating couplers.

Design of the AWG was based on the theoretical analysis in reference [106] and is reproduced in brief below using relevant parameters. All relevant measures used to draw the AWG are shown on the right in figure 4.1. Using the procedure described in this reference, the design begins at the emission end of the demultiplexer and is built towards the input. The first step in designing the device is to determine the receiver size and spacing at the far end of the AWG structure. This can be defined either by the lithographic limit of the tool (if the user attempts to pack together as many receiver waveguides as possible), or by defining the maximum desired crosstalk between waveguides. In this case, since the application of biological stimulation requires less strict limits on crosstalk, it was decided to determine these parameters on the lithographic limits imposed by the EBPG. Furthermore, as power throughput is critical in this application to stimulate channelrhodopsin, it is logical to use a sockets that are large with small gaps to improve power throughput. Therefore, a socket width (w) of 710 nm was chosen with a spacing (d_r) of 760nm from waveguide to waveguide. These taper to single mode waveguides of width 360nm to ensure high power throughput.

Next, the geometry of the free propagation regions (FPRs) must be defined. The FPRs are typically identical on both ends of the AWG, however they serve different purposes. The FPR on the input side of the AWG allows the light from the input waveguide to spread to illuminate the arrayed waveguides. On the output end of the AWG, the FPR allows free space for the focusing of the grating onto the receiver waveguides. The first step in defining this is to determine the intensity difference between the central waveguide and an outer waveguide. Since waveguides produce a Gaussian beam, it is impossible for the waveguide array to be illuminated evenly. Thus, the maximum allowable drop off in side channels must be chosen. Here it is set to 3dB. Using this parameter, it is possible to determine the angular extent of the AWG side of the FPR. Using the far field Gaussian beam approximation,

$$L_u = -10\log\left(e^{-2\theta_{max}^2/\theta_0^2}\right) \approx 8.7\left(\frac{\theta_{max}}{\theta_0}\right)^2, \tag{4.2}$$

where L_u is the relative intensity at the edge of the FPR in dB, θ_0 is the width of the Gaussian far-field, and θ_{max} is the angular extent of the FPR curve. Solving the above for $L_u = 3$dB, we find that $\left(\frac{\theta_{max}}{\theta_0}\right) = 0.59$. Next, to solve for θ_{max} the angle for the Gaussian far field must be solved for.

To solve for θ_0, first the Gaussian beam waist must be calculated. This is approximated using the following relationship given $V = 6.5$ from equation 2.1 for the 600nm waveguide, and by applying this value to the following approximation:

$$w_e \approx w_{wg}\left(0.5 + \frac{1}{V - 0.6}\right) = 710\text{nm}\left(0.5 + \frac{1}{5.9}\right) = 475\text{nm}. \tag{4.3}$$

Therefore, using the Gaussian approximation,

$$\theta_0 = \frac{\lambda}{n_{\text{FPR}}}\frac{1}{w_e\sqrt{2\pi}}, \tag{4.4}$$

where n_{FPR} is the index of the free propagation region (1.84). Thus, for the parameters given $\theta_0 = .22$radians and $\theta_{max} = .125$ radians. Therefore, the angular extent of the FPR on the side of the arrayed waveguides must be .125radians to ensure that the loss remains below 3dB. Since this simply traces out a circle, the radius for the curve can simply be found using the definition of the arclength of a circle in radians given that the arclength, s_{max}, must be long enough to contain half of the receiver waveguides. The radius R_a of the output aperture of the arrayed waveguides is then calculated as

$$R_a = \frac{s_{max}}{\theta_{max}} = \frac{d_r(N+1)/2}{\theta_{max}}. \tag{4.5}$$

Using the above parameters and assuming that the desired number of output channels (N) is 15, the resulting radius of curvature for the AWG waveguides to ensure proper focusing and a loss of no more than 3dB is $R \approx 56.25\mu$m. To find the radius of the receiver end of the FPR, we can simply design the FPR region as a Rowland-type fixture. In this case, the radius of curvature for the receiver aperture should be one half of the output array aperture. The distance between the two curved apertures is R_a.

Next the length increment of the waveguides in the array must be calculated. First, the dispersion needed to separate the channels into discrete spectral segments with a defined size must be calculated. Given that receiving waveguides are 760 nm apart, and a channel spacing of 1 nanometer (1.34×10^{12}Hz at 473nm) is desired, the dispersion can be found by the ratio of these two values,

$$D = \frac{d_r}{\Delta f_{ch}} = \frac{760\text{nm}}{1.34 \times 10^{12}\text{Hz}} = .568\mu\text{m per THz}. \tag{4.6}$$

The dispersion also can be calculated using the phase matching conditions of the arrayed waveguides,

$$D = R_a\frac{d\theta}{df} = \frac{1}{f_c}\frac{\widetilde{n_g}}{n_{\text{FPR}}}\frac{\Delta L}{\Delta\alpha}, \tag{4.7}$$

where f_c is the central frequency of the AWG corresponding to 473nm, $\widetilde{n_g}$ is the group velocity in the waveguide (calculated by simulation to be approximately 1.93), ΔL is the path length difference subsequent waveguides in the waveguide array, and $\Delta\alpha$ is the angular difference between each subsequent waveguide in the array ($\Delta\alpha = d_a/R_a$). Substituting for the dispersion D and solving for ΔL,

$$\Delta L = Df_c \frac{n_{\mathrm{FPR}}}{\widetilde{n_g}} \Delta\alpha = (.568\mu\mathrm{m\ per\ THz})\,(633\mathrm{THz}) \left(\frac{1.84}{1.93}\right) \left(\frac{.67\mu\mathrm{m}}{56\mu\mathrm{m}}\right) = 4.08\mu\mathrm{m}. \tag{4.8}$$

Therefore, when building the array waveguides the path length difference between subsequent waveguides must be 4.08 microns to ensure that center-to-center wavelength difference between channels is 1nm.

Finally, the last parameter determines the throughput of the AWG. To ensure that the loss at the central waveguide is less than 3db, the aperture for the arrayed waveguides must create an angle of 1.5 times $theta_0$, as shown in figure 3 of reference [106]. Thus, the angular extent of the arrayed waveguides is $\theta_a = .328$ radians.

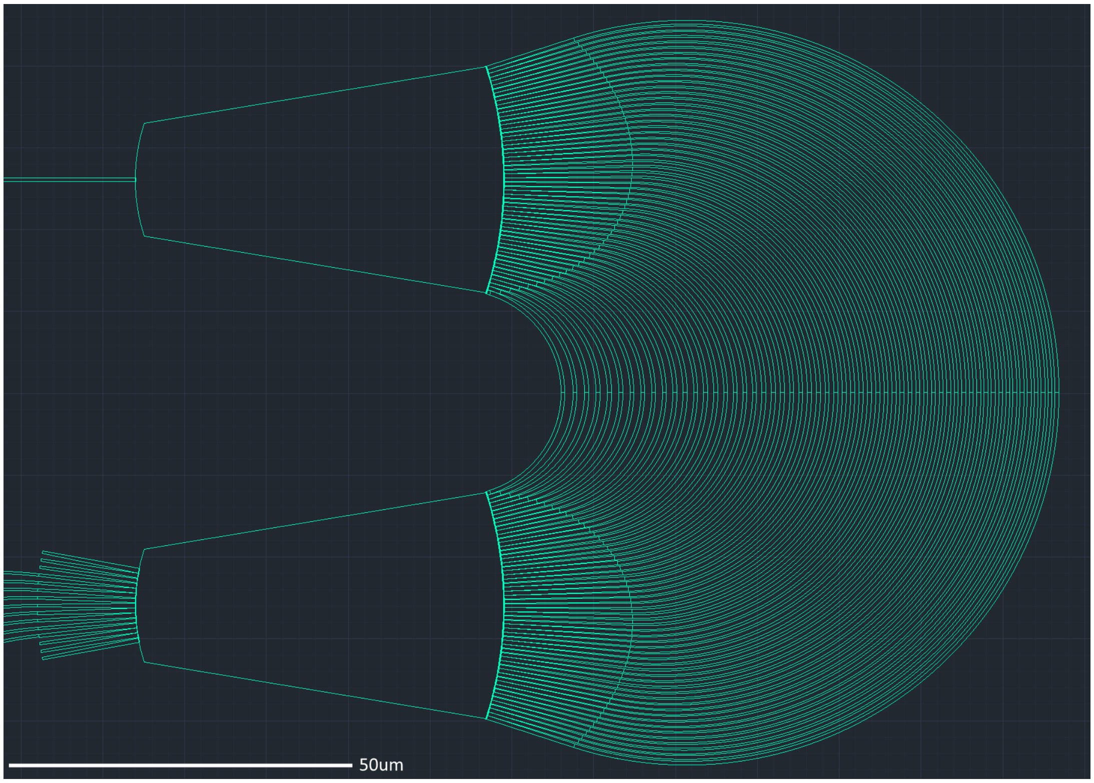

Figure 4.2: Array waveguide grating.

Given the above calculations, the AWG may now be drawn in CAD and produced. Due to the complexity of the waveguide structures, they were composed using automated Matlab code, which produces the DXF file. The resulting AWG structure is shown in figure 4.2. For simplicity and functionality, even though the AWG was designed for 15 outputs, only 9 were used in the final design. The outputs of the waveguide were tapered from the 710 nm receiver socket down to 340nm for routing on the shank. This ensures higher packing densities of waveguides on the shank of the probe.

4.2.2 Fabrication of Optically Integrated Neural Probes

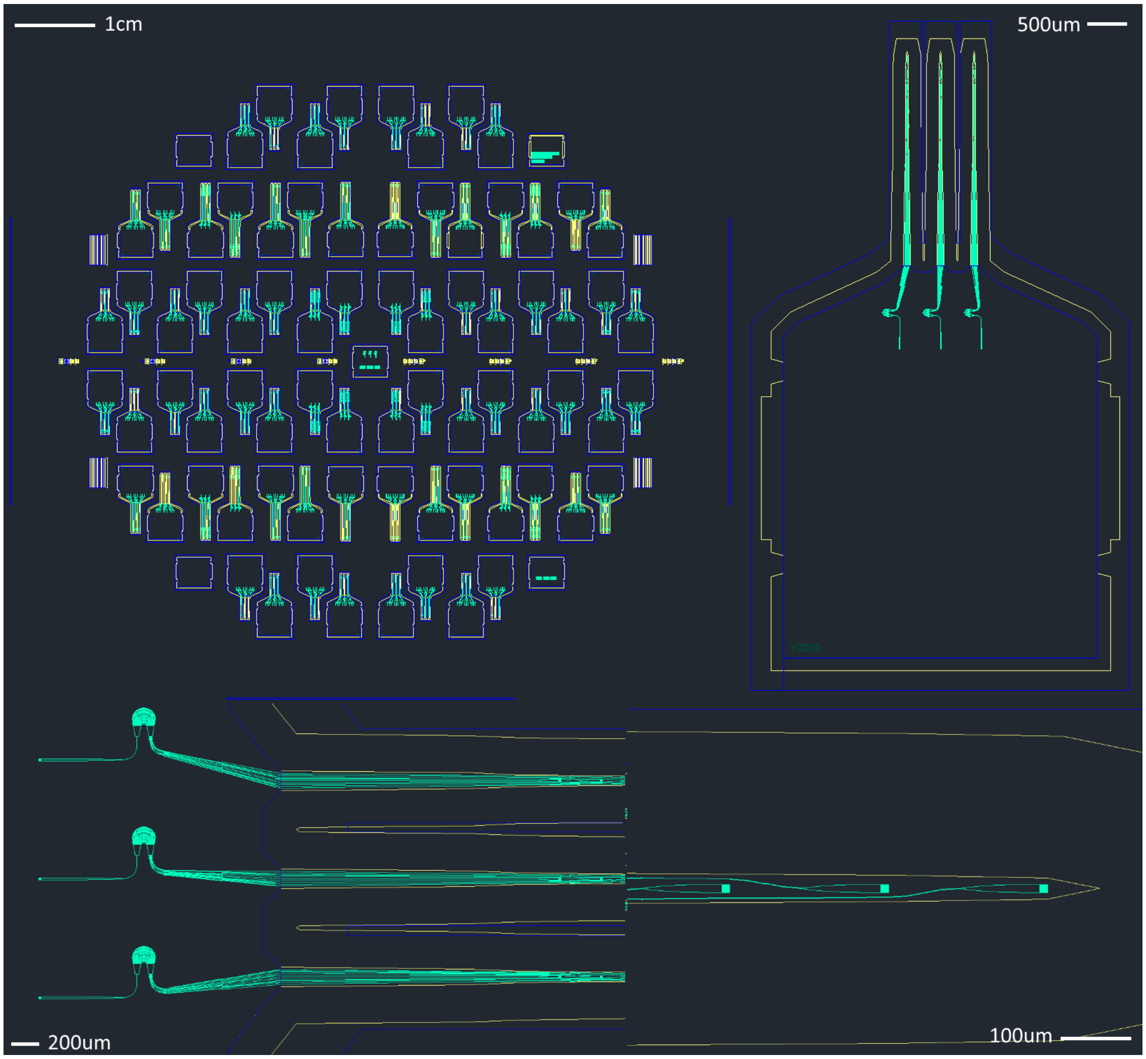

Figure 4.3: Composite image showing mask design for AWG-based probes for optogenetic stimulation. Layers: green = optical layer, yellow = frontside mask, blue = backside mask.

The mask set used for the fabrication of the AWG based neuro-optical probes consists of three total patterns, one patterned using electron beam lithography and two using contact photolithography. An overview of the patterns is shown in figure 4.3.

The layer defined by electron beam lithography contains the AWG structures described in the previous section, as well as the necessary routing of waveguides and grating couplers for coupling light into and out of the waveguides. AWGs are designed using the parameters in the section above, with critical parameters $R_a = 56.25\mu$m for the FPR, and $\Delta L = 4.08\mu$m.

Input and output grating couplers were designed using a combination of simulation and empirical testing. Optimal parameters for an output angle of 6^o were found to be 293nm with a duty cycle of 40% (resulting in silicon nitride slabs of 117.2nm). Grating couplers had a socket length of 15μm and a width of 10μm to accommodate the geometry of a single mode fiber and to ensure that there was enough length for the majority of the light exiting the fiber taper to be emitted from the grating. Grating couplers were designed with a taper length of 400nm at the input and 100nm at the output. Gratings were distributed evenly down the center of the shanks with a pitch of 200μm to accommodate the tapers for the grating couplers.

Routing waveguides for the AWG were larger (600nm) to minimize cross talk between channels and minimize losses in the AWG structure. Routing waveguides were made smaller 380nm with the plan of creating higher density probes in the future which requires smaller waveguide structures for increased packing efficiency onto shanks.

The structure of the probes themselves have a number of important features. Because of the use of SOI wafers, released shanks had a uniform thickness of 18μm and widths of 90μm at the base of the shank. This is tapered down to 20μm at the tips to ensure that there is minimal disruption of the neural tissue during implantation. Furthermore, the shank ends were sharpened to tips with a radius of curvature less than 1μm to ensure that penetrating the neural tissue would be smooth and problem-free. Initial probes are 3mm long, ensuring sufficiently deep penetration into neural structures is possible in small animals such as mice. This length allows for patterned optogenetic stimulation in deep structures such as the mouse hippocampus. As can be seen in figure 4.3, it is obvious that the majority of the base of the probe is unused. This “dead space” is used when probes are being packaged with external fibers, as is described later in this section. It is also possible to see that the edges of the base of the probe have 4 tabs of 200μm width. These are used to keep the probe in place after the final backside release etch is completed. In this way, the probes do not float out of position when the wafer is released from the carrier wafer. These tabs can be broken very simply using a pair of tweezers when it is time to use the probe.

The specifics of the fabrication protocol can be found in Chapter 3, however a brief summary of the major steps of the fabrication protocol follows. Prime 100mm SOI wafers with $< 1, 0, 0 >$ orientation, device layer thickness of 15μm, buried oxide thickness of 2μm and handle thickness of 300μm were first prepared by wet thermal oxidation to a thickness of 1.5μm (Rogue Valley Microdevices, Medford, OR). LPCVD silicon nitride was then coated to a thickness of 200nm (Rouge Valley Microdevices, Medford, OR). Wafers were prepared and coated with Ma-N 2403 per section 3.3.0.3. The wafer were patterned in a Raith EBPG 5000+ with a 500pA beam and a dose of 250μC per square cm and developed for 1 minute in Ma-D 525. The wafer was then etched using the Ma-N specific pseudoBosch recipe described in section 3.4.2.

The wafer was then cleaned thoroughly using solvents, RCA-1, buffered oxide etch, and RCA-2

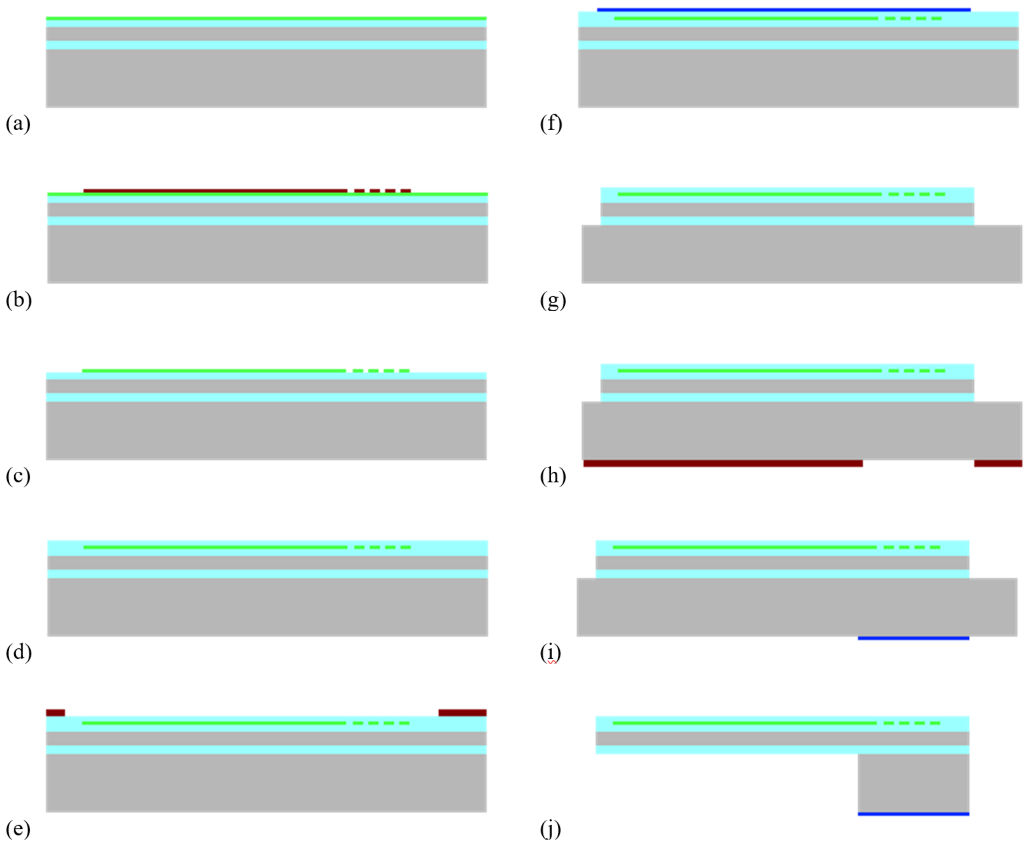

Figure 4.4: Fabrication process. (a) Initial wafer stack consisting of an SOI wafer (silicon in grey, oxide in aqua) coated with thermal oxide (aqua) and LPCVD silicon nitride (green). (b) Wafer is coated with Ma-N 2403 and is patterned using electron beam lithography. (c) Pattern is transferred to silicon nitride wafer using pseudoBosch etch. (d) Wafer is coated with PECVD silicon oxide to clad optical layer. (e) Wafer is patterned photolithographically using S1813 resist. (f) Alumina (blue) is coated on surface using electron beam deposition and is patterned by liftoff process. (g) Pattern is transferred to wafer topside through top oxide, device layer, and buried oxide. (h) Backside of wafer is patterned photolithographically using S1813 resist. (i) Alumina (blue) is coated on back surface using electron beam deposition and is patterned by liftoff process. (j) Pattern is transferred through wafer backside to box layer.

before coating with PECVD oxide. 1.5μm of PECVD oxide was coated onto the frontside of the wafer using the Oxford Systems 100 Plasma Enhanced CVD using Silane (in argon) and N$_2$0 as the oxidizer, as described in section 3.6.2.

Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per cm^2 using mask 1 (negative tone, yellow in figure 4.3). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited on the wafer using the TES FC-1800 electron beam evaporator to a thickness of approximately 350nm, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is lightly brushed with a swab to remove any excess alumina on the surface, and is rinsed in isopropyl alcohol before drying with compressed nitrogen.

After depositing the alumina hard mask, the wafer is etched in a Oxford Instruments 380 ICP etcher with DRIE pod. The wafer is secured to a 6-inch carrier wafer using a thermal contact solution (Fomblin, Solvay, Bruxelles Belgium) and is placed in the etch chamber. Wafers were first etched using the pseudoBosch recipe designed for SiO_2 removal described in section 3.4.2. Due to the relatively slow etch rate (35nm per minute), approximately 80 minutes of etching were required to penetrate the oxide layer. The etch depth was monitored using a profilometer. After penetrating the top oxide layer, the wafer was exposed to the Bosch etch described in section 3.4.1 for 20 cycles to etch through the device layer of the SOI wafer. Finally, the oxide etch is resumed to penetrate the buried oxide layer, an additional 50 minutes. Wafers are placed in an acetone bath overnight to remove the thermal contact layer from the surface, releasing the wafer from the carrier.

Next, backside of the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 2 (negative tone un-mirrored mask, blue in figure 4.3). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited on the wafer using the TES FC-1800 electron beam evaporator to a thickness of approximately 350nm, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is lightly brushed with a swab to remove any excess alumina on the surface, and is rinsed in isopropyl alcohol before drying with compressed nitrogen.

Similarly to the front side etch, after depositing the alumina hard mask, the wafer is etched in a Oxford Instruments 380 ICP etcher with DRIE pod. To protect the frontside structures on the wafer, the wafer is first coated with a thick layer of PMMA A11. The wafer is secured to a 6-inch carrier wafer using a thermal contact solution (Fomblin, Solvay, Bruxelles Belgium) and placed in the etch chamber. Wafers were first etched using the pseudoBosch recipe designed for SiO_2 removal described in section 3.4.2. Due to the relatively slow etch rate (35nm per minute), approximately

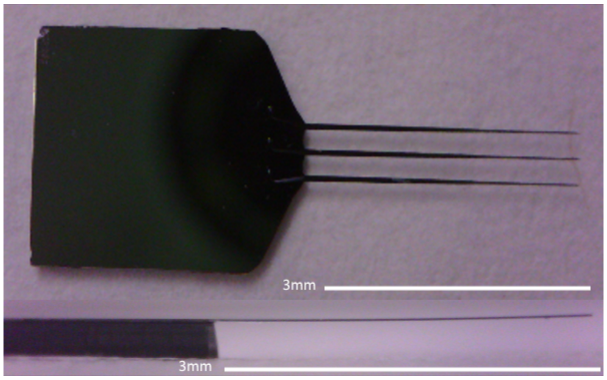

Figure 4.5: Photographs of a released probe.

50 minutes of etching were required to penetrate the oxide layer. The etch depth was monitored using a profilometer. After penetrating the top oxide layer, the wafer was exposed to the Bosch etch described in section 3.4.1 for 100 cycles at a time until 300 cycles have been completed. The wafer is then etched at a rate of 10 or 20 cycles at a time until all of the back silicon has been removed by microscopic inspection. This typically takes 340-360 cycles to remove all 300μm of the handle wafer. The wafer (with the carrier) is then placed in an acetone bath overnight to release the wafer from the carrier. The following morning, the wafer is carefully removed, rinsed with isopropyl alcohol, and gently dried with compressed nitrogen gas.

The resulting released probe can be seen in figure 4.5. Although in this image it is difficult to see, the probes are patterned with the optical layer showing upward. The side view of the shank shows the height of the implantable probe at approximately 18μm. The base of the probe has dimensions 4.4mm by 5.4mm. Micrographs of the probe showing the optical layer is shown in figure 4.6. These show the AWG structures fabricated onto the top surface of the probe, as well as the output gratings used to project light into the tissue.

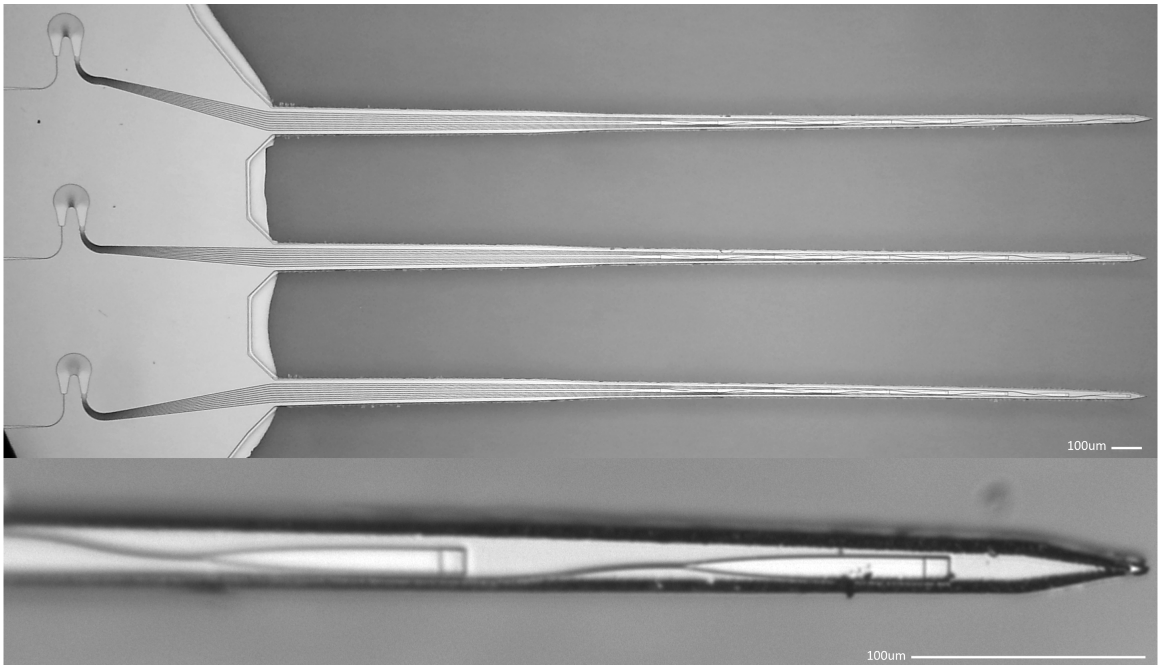

Figure 4.6: Micrographs of the released probes showing the length of the probe shanks (top), and a close up of the emitter pixels (bottom). Modified from [96].

4.2.3 Packaging of probes

The major complication in packaging probes such as these is that, for implantation, it is highly desirable to ensure that the optical fiber used to bring light to the probe is in-line with the axis of the shank. This creates a more easily implantable probe and ensures that the microscope objective

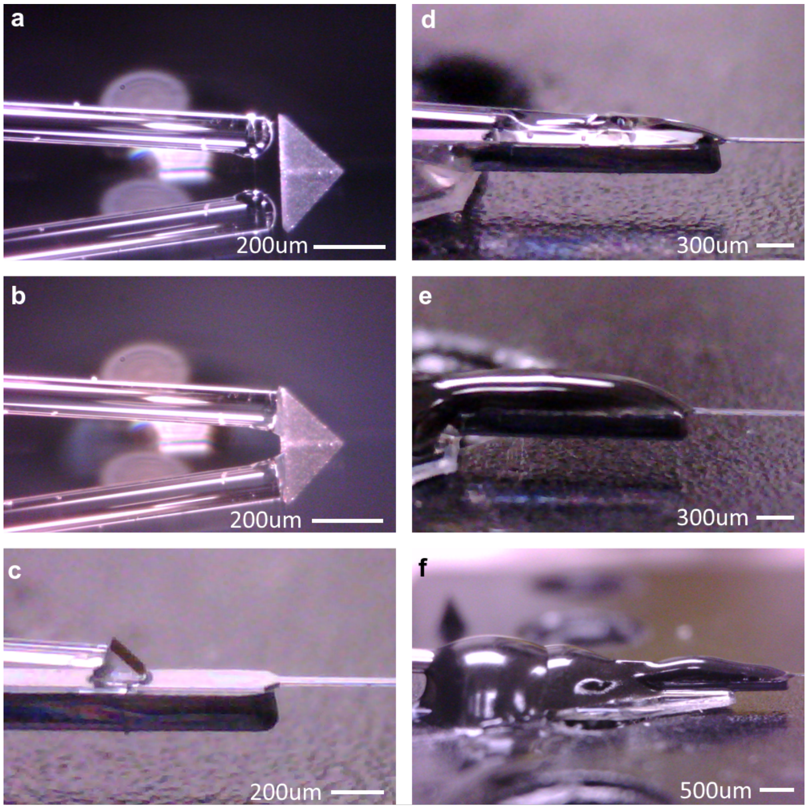

Figure 4.7: Microprism attachment and probe coupling procedure. Reproduced from [96].

used for experimental monitoring of the behavior of neurons does not interfere with the location of the implanted probe. Two solutions were explored for changing the angle of light exiting the end facet of an optical fiber. The first solution attempted was to use an angle-polished fiber, where the end facet is cut at an angle using a laser and then polished to a flat surface. Optionally, the end facet can be coated with aluminum to improve reflectivity. This allows the beam to propagate down, through the cladding of the fiber at a controlled angle. This solution, however, was found wanting: The facets obtained were rough and propagation through the cladding caused problems even when using index-matching liquids.

As an alternative solution to this problem, the use of microprisms was explored. These prisms, only 80μm on a side, allow for light to be redirected via total internal reflection of the beam at a 45^o angle. This was found to be a much superior solution for the re-direction of light into the grating coupler. To prepare the fiber, first the fiber was cleaved and cleaned as to produce a good surface for light propagation. The fiber was then mounted in a thin stainless steel tube and mounted to a 3-axis micromanipulator stage. The facet of the fiber is then dipped in UV-activated epoxy (figure 4.7.a). The fiber is positioned such that it touches one face of the prism, as is shown in figure 4.7.b. The UV epoxy was then cured using a handheld UV light source typically used for dental procedures. Next, a thin layer of glue is carefully applied to the exposed base of the prism by touching the base of the prism to a bead of UV epoxy. The fiber with attached prism was then moved over a released probe. Light is coupled through the fiber into the chip, and the intensity of the output gratings is optimized by moving the grating with prism until it is in place (figure 4.7.c). To maximize coupling, the fiber is tilted to 6^o with respect to the flat surface of the probe. Once in position, the UV epoxy on the bottom of the prism is cured using UV light. Once cured, an additional layer of UV epoxy is applied to the surface of the probe and fiber and cured to provide further stability of the fiber (figure 4.7.d). Finally, a layer of black medical epoxy is applied over the surface of the probe. This is done to absorb any stray light that might become decoupled from the waveguide or is otherwise lost during the coupling process, as grating couplers are not 100% efficient. The final result is shown in figure 4.7.e.

To complete the packaging procedure, L-shaped holders were produced using 3D printing to offset the micromaniplator axis from the probe axis. These were attached to stainless steel rods which, in turn, connected the 3D printed part to the micromanipulator stage. Probes were secured with epoxy to the 3D printed holder using medical grade epoxy. The fiber was also attached directly to the holder using medical grade epoxy to minimize tension and bending of the fiber.

4.2.4 Results

4.2.4.1 Insertion and propagation loss of waveguides

Transmission properties of the optical waveguide structures were determined using a switch-back arrangement on supplementary test chips. These structures and measurements are intended to differentiate the insertion loss (the loss from coupling into and out of the optical waveguide circuit) from the loss during propagation to the waveguide. This is accomplished by creating a series of waveguides with a serpentine structure with varying straight lengths of the straight waveguide structure. Curves are maintained in each chip to ensure that any losses due to waveguide curvature are not also inserted into the propagation loss calculation. Six identical waveguides at six lengths were measured. Chips were measured using a solid state 473nm laser coupled into a single-mode fiber. The fiber was brought into proximity of the chip using a 3-axis micromanipulator stage. Output from the chip was captured using a larger multimode fiber (0.22 NA) to ensure that all of the outgoing light was captured. The multimode fiber output was coupled to a standard bolometric power meter, and then compared with the intensity of the input light. The resulting propagation loss curve is shown in figure 4.8. Waveguide loss was measured to be 13dB per cm, which is comparable to, but more lossy than industry-standard waveguides at this wavelength (10dB loss per cm).

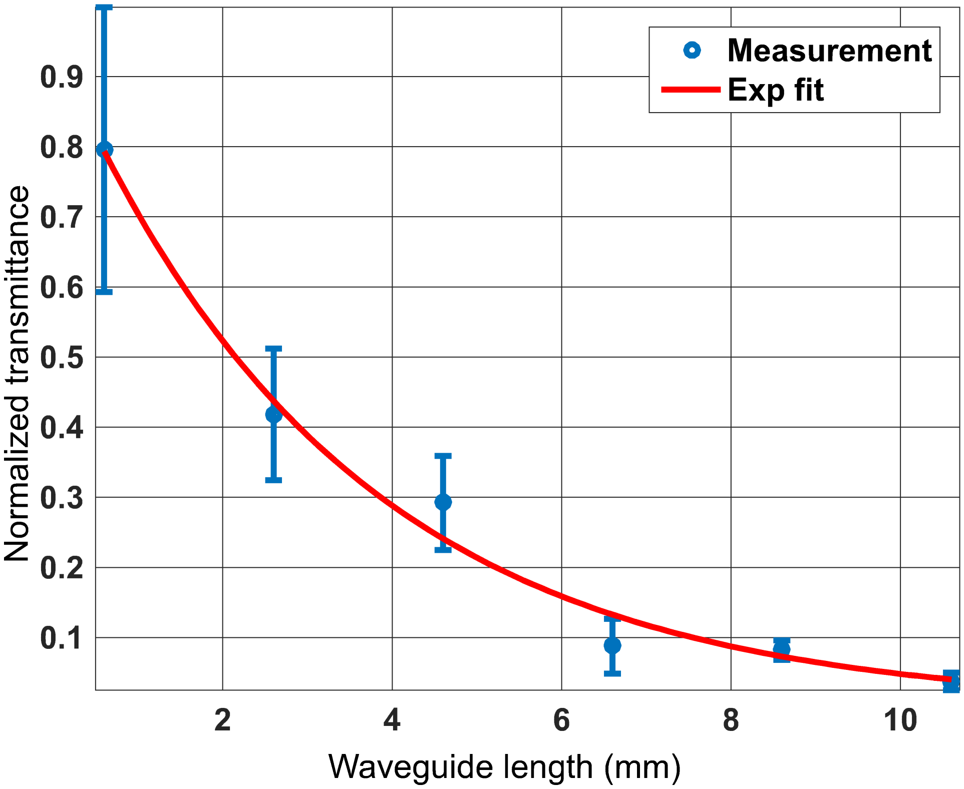

Figure 4.8: Propagation loss is Si_3N_4 waveguides at 473nm fabricated at Caltech in the Kavli Nanoscience Center cleanroom. Modified from [96].

4.2.4.2 Emission profile of emitter pixels

The next critical factor in the quantitative understanding of these devices is the emission profile of the beam emerging from the output of the probe. To accomplish this, an apparatus was designed to obtain micrographs of the beam profile of the emitted light using Fluorescein (a fluorescent dye) as a contrast agent. A cuvette consisting of two microscope coverslips was created to act as a container for the fluorescent solution. This is shown in figure 4.9, bottom. The probe was inserted between the coverslips on the open end of the cuvette using a micromanipulator such that the probe was perpendicular to the image plane of the microscope objective. A fluorescein solution (10μM, pH> 9.5) was prepared in deionized water and was injected into the cuvette. The laser was then engaged and the beam profile was imaged using the microscope, as shown in figure 4.9. A highly collimated beam was observed with an emission angle of approximately 6^o. The intensity data is summarized in figure 4.10. The theoretical beam profile showed a slowly diverging beam in the far field; the beam diameter starts at a waist of less than 5μm, and expands to approximately 10μm at 250μm from the probe. Comparing this to the sample in Fluorescein, the beam waist at its narrowest is approximately 6μm and expands to approximately 21μm at 250μm from the probe. Finally, when

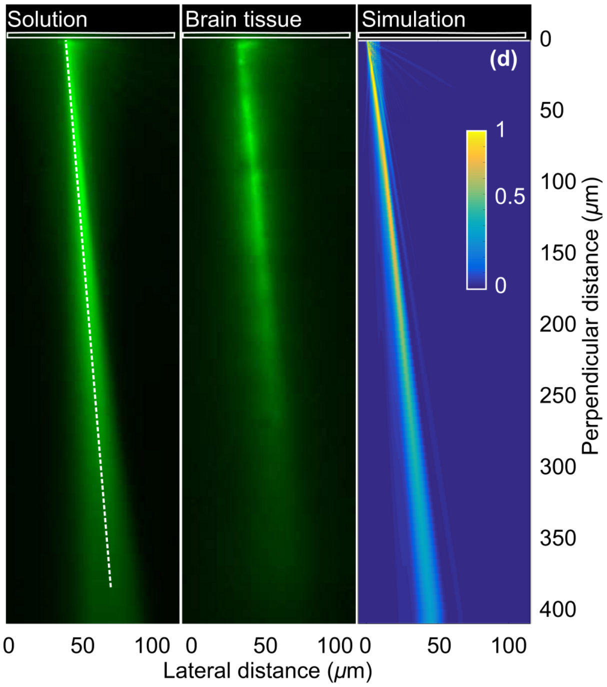

Figure 4.9: Measured beam profile of e-pixel in fluorescein solution (left), brain tissue (center), and simulated (right). Below, photograph of the probe sandwiched between two cover slips to take beam profile images above. Modified from [96].

the probe is placed in the neural tissue impregnated with Fluorescein, the beam is found to be more divergent, but still quite collimated several hundred microns away, due to the scattering of the neural tissue. This also shows that the beam remains relatively narrow in scattering media; the beam waist in neural tissue at its smallest is approximately 15μm and expands to 30μm at 250μm from the probe. This data shows that the minimum beam pitch for these devices can be reduced to 50μm to ensure total coverage of the length of the shank while minimizing overlap between beams.

4.2.5 AWG performance

Array waveguide grating performance was evaluated using a simplified experimental setup using a tunable laser source and a filter to produce the desired wavelength of light. The experimental setup is shown in figure 4.11. The tunable source used was a tunable titanium-sapphire laser (Chameleon Vision, Coherent Inc., Santa Clara, CA. USA) tuned to 946nm. It pumps a second-harmonic-generating BBO crystal, which produces a coherent beam at half the wavelength of the pump laser (473nm). Since the efficiency of the BBO crystal is not 100%, the beam is then filtered using a dichroic mirror to reject any light at the pump wavelength and allow the upconverted 473nm beam to pass. Due to the pulsed nature of this light, the bandwidth of the beam is actually larger than required to measure a single channel of the AWG. Thus, the beam was then sent through an optical band pass filter on a rotating stage. By rotating the bandpass filter it is possible to tune the

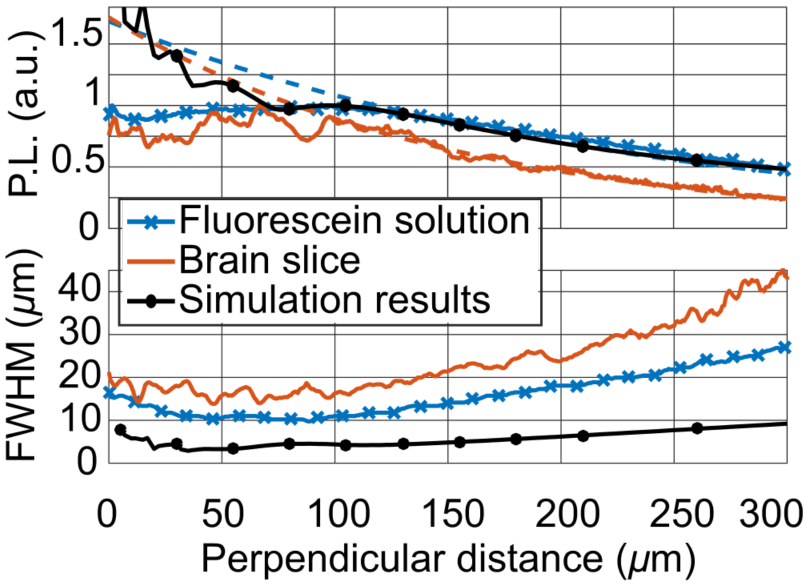

Figure 4.10: Quantitative representation of data in figure 4.9. Modified from [96].

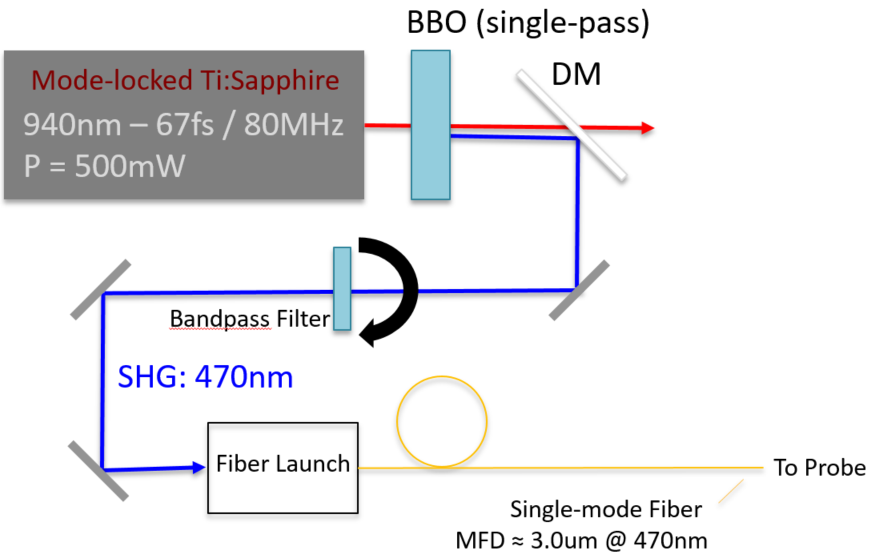

Figure 4.11: Experimental setup used to test spectral selectivity of AWG demultiplexed probes. Pulsed light from a Ti:Sapphire laser tuned to 940nm was frequency doubled and sent through a bandpass filter on a rotational stage. By rotating the bandpass filter, it was possible to tune the spectrum within that produced by the BBO crystal. The light was then fiber coupled and sent to the probe.

pass band of the filter to allow a narrow, controllable band of light through the filter to stimulate individual e-pixels. This light is then coupled into an optical fiber which was sent to the test chip containing an AWG with a single input and multiple outputs (one for each optical channel). The light was coupled into the AWG through the input grating coupler by bringing the optical fiber's cleaved facet into proximity with the grating coupler at an angle of 6^o from the normal. Light was coupled off of the waveguide using a multimode fiber, and sent to a spectrometer to analyze the output signal.

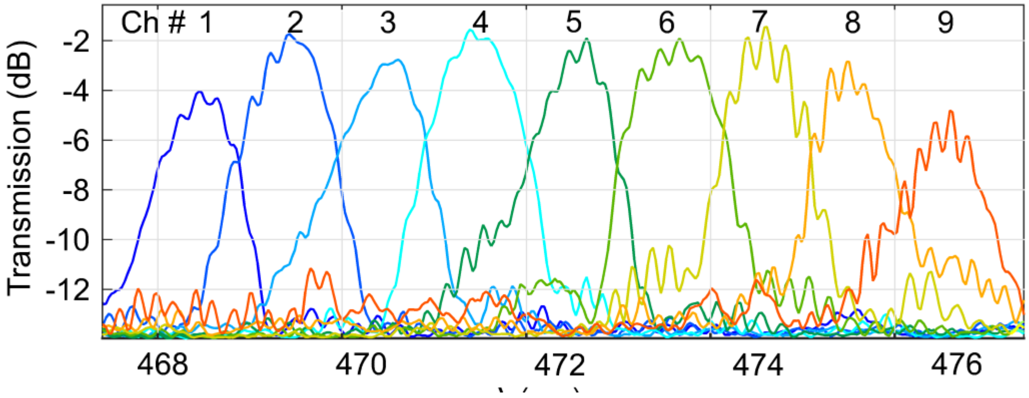

Figure 4.12: AWG performance, including the output spectrum of each e-pixel measured using a spectrometer. Broadband illumination was used to illuminate all channels simultaneously.

The resulting AWG performance curves are shown in figure 4.12. A bandwidth of approximately 1nm was achieved successfully. Optical throughput was normalized to a single fiber on the same chip. Central wavelength loss through the AWG was minimized to less than 3dB. Furthermore, losses between the center channel and outer channels were less than 3dB as well, per the design of the AWG in section 4.2.1. Due to the design decision to pack receiver waveguides as close together as possible, crosstalk was relatively high in this prototype design. However, by designing the source of the light for this system to have a bandwidth narrower than 1nm, this can be avoided. Future designs can reduce this channel crosstalk by increasing d_r. Emission by multiple waveguides simultaneously is shown in figure 4.13.

4.2.5.1 Validation in living tissue

The efficacy of these probes for stimulating neural tissue was tested in two separate experiments by Dr. Eran Segev, and is described herein as validation of these probes in neural tissue. The first experiment involved co-localizing an electrode onto the shank of the probe with an active e-pixel. This experiment was done in the lab of Karl Deisseroth at Stanford University in collaboration with the Roukes group at Caltech. To accomplish this, an insulated tungsten wire electrode was glued to a shank 100μm above an active e-pixel, as shown in figure 4.14.a. This setup would allow for the electrical validation of optogenetic stimulation by the optical probe.

Transgenic Thy1:18-ChR2-EYFP male mice were used for this experiment. They were housed in a vivarium on the Stanford University campus 3 to 5 to a cage on a reverse 12 hour light-dark cycle with an *ad libitum* feeding schedule. Experimental protocols were approved by the Stanford

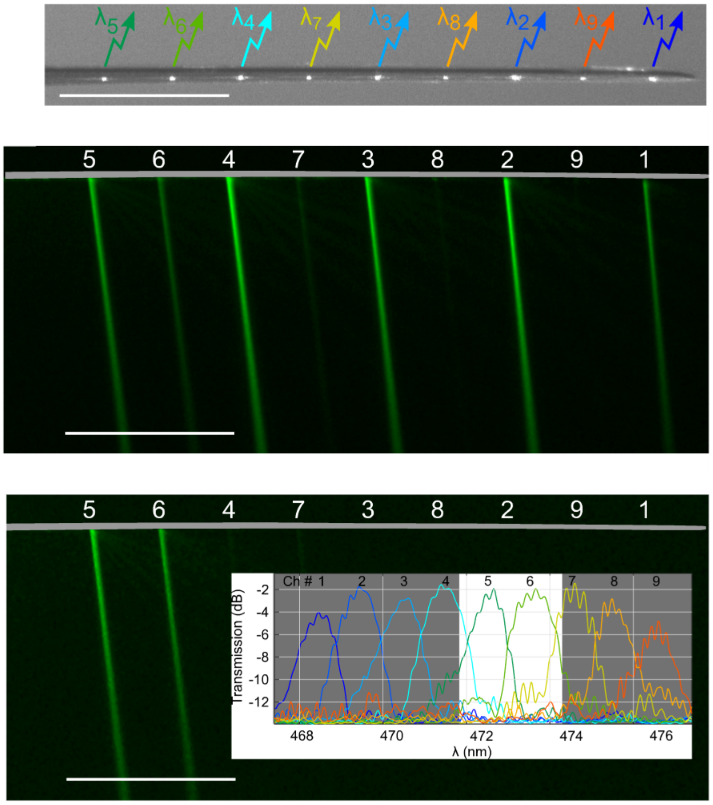

Figure 4.13: Emission from probe wafer using setup described in figure 4.11 under two angles of the bandpass filter. Bars are 500μm.

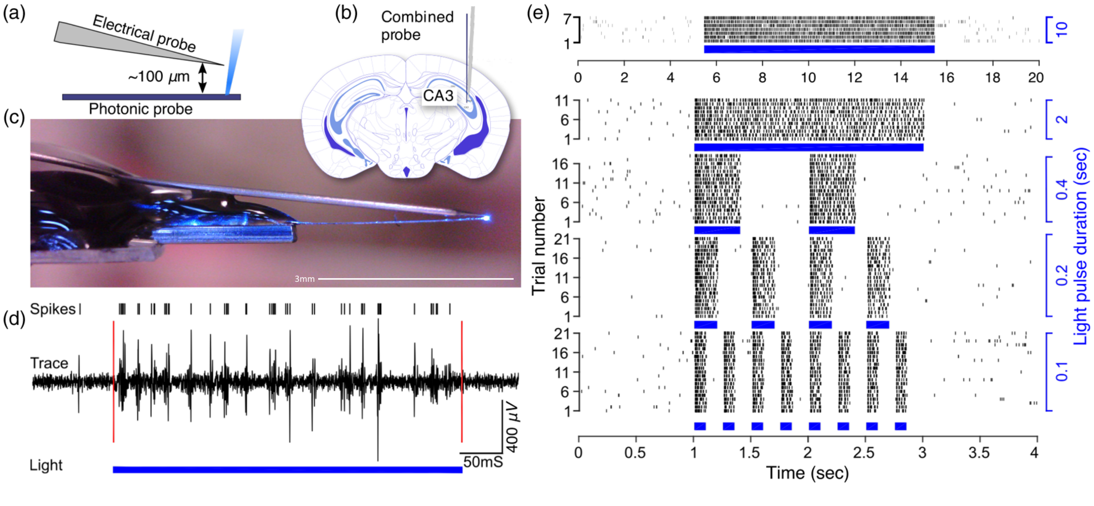

Figure 4.14: Validation of e-pixel performance in living tissue using a tungsten electrode. (a) Geometry of the optical probe and tungsten electrode, (b) location of implantation of the probe into CA3 of the hippocampus, (c) photograph of the probe with attached electrode, (d) spike trace from the electrode during optical stimulation, and (e) raster plot of spiking resultant from multiple stimulation frequencies using the optical probe. Reproduced from [96].

University IACUC. Animals were anesthetized with isofluorane gas (1-4%) and anesthesia levels were monitored by response to physical stimulus. Once anesthetized, mice had their heads shaved and were immobilized in a Kopf Instruments (Tujunga, CA. USA) stereotaxic apparatus. The exposed skin on mouse was treated with betadine and rinsed with 70% ethanol. A craniotomy was performed to remove the scull, exposing a region of diameter 0.5-1mm of the brain. The composite probe was inserted into the brain using the stereotaxic apparatus at a location of X = +2.75mm, Y = -2.54mm, Z = -2.60mm, intended to implant the tip of the probe into the CA3 layer of the hippocampus. Clampex software was used for the electrical recording and to conrol the illumination frequency of the optical probe.

The resulting behavior of pyramidal cells to electrical stimulation are shown in figure 4.14.c, 4.14.d and 4.14.e. The first figure shows the result of stimulation on the electrode. Spikes are clearly visible during the period of stimulation. This data is often simplified as a raster, as shown in figures 4.14.d and 4.14.e, where a threshold is used to identify neural spiking. Each dark pixel on the raster corresponds to a single neural spike. Using the raster plots, it is obvious that during optical stimulation (shown as a dark blue bar), spiking frequency is greatly increased. This correlates directly with the length of the pulse. Therefore, the probe is successfully stimulating ChR2 positive neurons in the tissue.

The second experiment utilized a different method for measuring neural stimulation. In this case, a two-photon scanning confocal microscope was used to measure the intensity of fluorescence from a calcium sensative protein, GCaMP6. During stimulation of neurons, a critical part of the generation of action potentials and synaptic transmission is the influx of calcium ions from the exterior of the cell and from inside the endoplasmic reticulum of the cell into the cell body, or soma. GCaMP6 a hybrid protein, composed of a fluorescent protein connected to a portion of calmodulin, which binds to calcium ions. When calcium is present, the shape of the chromophore protein changes, which increases its fluorescence efficiency. Therefore, we can measure spiking by the proxy of calcium influx into the neurons. This technique, however, does not measure spiking directly, we only see a peristimulus effect which evolves over a slower timescale than neuron's action potential. These experiments were done in collaboration between Andreas Tolias' research group at the Baylor College of Medicine and the Roukes group at Caltech. All experiments were approved by the IUCAC at BCM.

Experiments were performed on IP-Cre/ChR2-tdTomato mice (C57Bl/6 background) which were injected with 1μL of AAV1.Syn.GCamp6s.WPRE.SV40 virus mixed 1:1 with AAV1-CamKIIa-ChR2(E123T/T159C)-mCherry 3-5 weeks before the experiment. Injections were preformed stereotactically through a burr hole at a 60^o angle targeting the visual and extra striate cortex $\approx$ 300μm below the surface of the cortex.

Animals were anesthetized with isofluorane gas (3%). Once anesthetized, mice had their heads

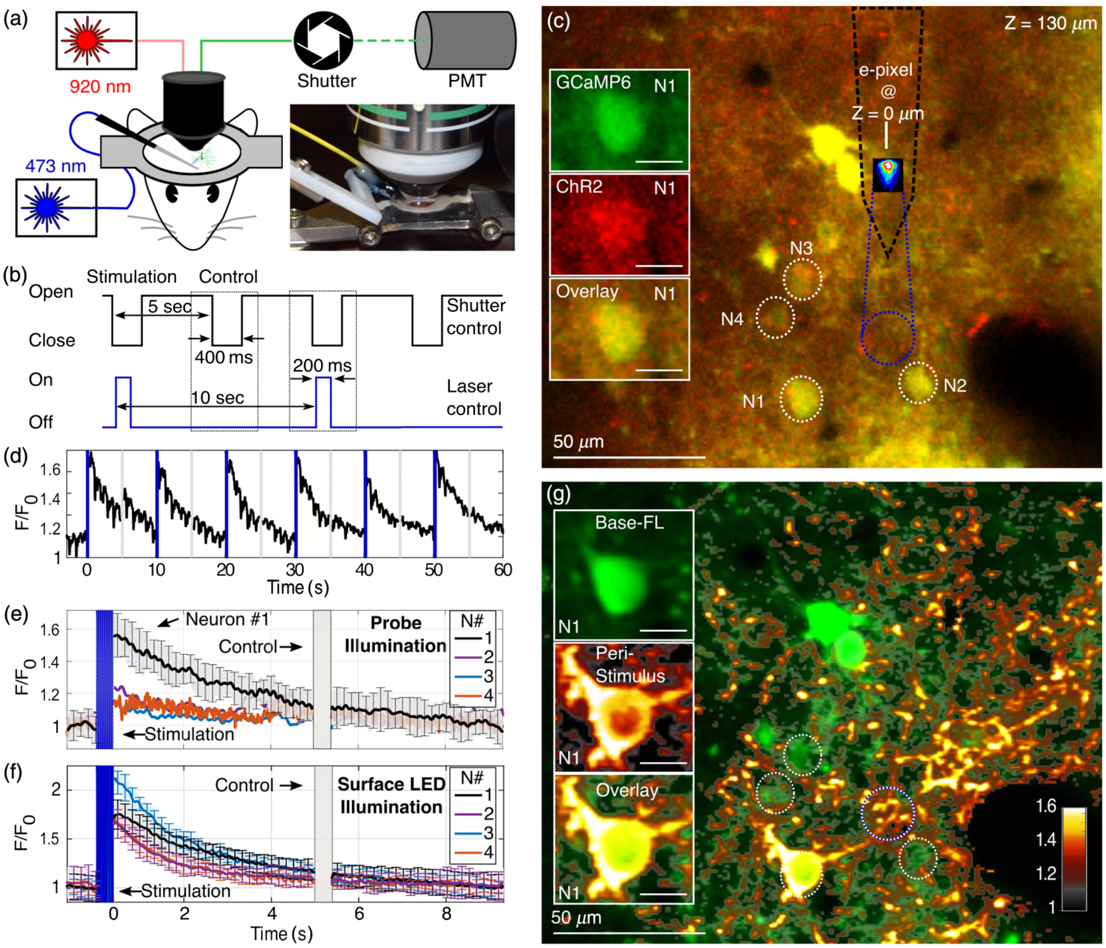

Figure 4.15: Validation of e-pixel performance using optical reporters measured by two-photon microscopy. (a) Experimental setup including mouse headfix and orientation of inserted probe, (b) light excitation sequence with respect to imaging frequency, (c) visualization of the expression of optical actuator (ChR2) and optical sensor (GCaMP6) with probe overlay, (d) calcium transients measured in neuron 1, (e) peristimulus histograms averaged over 19 stimulation cycles for neurons 1-4, (f) peristimulus histograms averaged over 15 stimulation cycles using wide field blue illumination, (g) migrograph of peristumulus properties of a single stimulation event by the optoprobe. Reproduced from [96].

shaved and were immobilized in a stereotaxic apparatus (Kopf Instruments, Tujunga, CA). Mice were injected with 5 to 10 mg per kg ketoprofen subcutaneously at the start of the surgery. Bupivicane was injected subcutaneously, and after 10-20 minutes a 1 square cm area of scalp was removed from the skull and the boundaries were sealed with surgical glue. A headbar was then attached with dental cement (Dentsply Grip Cement). Using a surgical drill and HP 12 burr a craniotomy was preformed, opening a 3mm diameter hole exposing the cortex. The craniotomy was washed with artificial cerebrospinal fluid. The mouse was then positioned beneath the 2-photon microscope, where the optical probe was fixed to a micromanipulator stage. The probe was advanced through the dura mater and tracked by the reflection of the probe in the microscope, and insertion was continued until the probe was positioned within the injection site of the AAV1 viruses.

Next, light pulses from the optical probe were produced at a rate of 0.2-1Hz, and the focal plane was scanned above the probe location until cells were found which seemed to be synchronized with the pulsing of the laser were found. Once the image plane was established, the experiment was conducted. In this case, two phases of measurement were repeated to ensure that the activation of the neurons was actually due to optical stimulation by the e-pixel. In the first phase, the e-pixel was activated while the shutter to the photomultiplier tube was closed (to ensure that the PMT was not oversaturated), stimulating the neurons. The laser was then deactivated and the shutter opened. During this phase, the intensity of GCaMP6 fluorescence was measured. After this measurement was completed, the shutter was closed again and reopened while the laser was off. This was done as a control to ensure that the sound of the shutter did not directly stimulate the neurons being observed.

The results of this experiment are shown in figure 4.15.d-f. A direct correlation is evident between the stimulation of the neurons by exposure to 473nm light, as shown in the peristimulus plot in figure 4.15.d. By measuring the beam profile from the e-pixel, it was subsequently determined that the beam stimulated the dendritic processes of neuron 1 (N1 in figure 4.15.c), while leaving neurons 2-4 unstimulated. This was confirmed by flood exposure by the microscope objective of the field that neurons 2-4 were also stimulatable. This further reinforces that these probes can be used to effectively stimulate individual neurons *in situ*, and that they provide beam diameters at the cellular scale able to target individual neurons.

4.2.6 Limitations and Future Work

The major limitation of this work doesn't arise from on-chip demultiplexing, as described above, but instead from complications with multiplexing the signal before being sent to the probe. Signal generation requires separately-controllable wavelength sources, some of which are impractically expensive and some of which are unable to deliver enough power to the probe to induce a conformational change in channelrhodopsin.

The simplest (and most expensive) solution would be to drive the AWG neural probes with an array of lasers. Although AWG-based systems are driven in this manner in industry at telecom wavelengths, due to the low volume of production of lasers around 473nm is difficult to use a die-picking method to develop an array of lasers with exactly 1nm spacing. Furthermore, due to lack of strong commercial interest in this wavelength range, tunable lasers at this wavelength simply do not exist. Therefore, laser technology does not provide an avenue for creating the illumination source for this probe.

As an alternative, diffractive optics and spatial light modulation feasible approach, since sources could be developed for such a system. This is the approach proposed by this author to solve the problem. A monochromator could be built, but instead of filtering with a slit, multiple wavelengths could be selected simultaneously by a spatial light modulator such as a digital micromirror display (DMD). The selected wavelengths could be collimated into a fiber and sent to the optical fiber, and then the chip. This technology requires a spatially- and temporally-coherent light source that is sufficiently broadband for this approach. Current technologies which can satisfy this demand are superluminescent LEDs (SLEDs) and supercontinuum lasers. Supercontinuum sources are currently available commercially, however they typically have spectral power densities of at most 1mW per nm. Although this is sufficient on its own to stimulate neurons, when considering coupling losses in photonic neural probes could still be insufficient to create a workable solution. Promising research is being conducted in the area of creating blue SLEDs, however they are not yet commercially available and thus are not an option at this point in time [55, 54].

4.3 Phased Array for Multi-Site Optogenetic Stimulation

4.3.1 Motivation

The work described in the previous section describes a novel method to stimulate neurons in a densely packed fashion. However there is a limitation of this arrangement of e-pixels; this setup requires that the neurons be in an arrangement which guarantees that the beam impinges on a desired neuron when the probe is implanted. In reality, the distribution of neurons is unknown and irregular and it is difficult or impossible to determine neurons of interest before implantation. Furthermore, exact placement of the probe apriori to stimulate a predetermined individual neuron would be nearly impossible. Although there are many solutions to this problem, the superior solution when dealing with implantable probes would be to steer the beam to activate any desired neurons in front of the probe's surface. To accomplish this, however, is not a trivial task. Mechanical reflection systems are a possible solution, however it is undesirable to fabricate MEMS devices on an implantable probe. Furthermore, mechanical devices would likely be much larger than the probe itself. Acousto-optical elements are also a possible option for beam steering, but it is unclear how steering with these materials could be accomplished easily using a planar circuit. However, there are non-mechanical methods available for beam stearing as well. In this section, a phased array system is described which can use wavelength to encode the angle of emission for an e-pixel. This method is conceptually similar to the AWG system described above, however it utilizes delay lines to induce free-space diffraction of a beam.

Phased arrays have many common characteristics to the AWG structures described in the previous section. Phased arrays are also PHASAR lightwave circuits which utilize delay lines to create a phase difference between a plurality of optical pathways. However instead of being emitted into a free propagation region within the planar circuit, the delayed light is immediately emitted from a series of grating couplers. Thus, the interference of the beams occurs in free space and interferes to redirect the beam based on the wavelength supplied to the input of the phased array. The simplest way to understand the steering ability of a phased array is to utilize the Huygens principle. Imagine that each grating is a point emitter with a tunable phase delay. If all the emitters exhibit no phase delay, a plane wave will be produced with a Poynting vector perpendicular to the plane of the phased array. However, if each emitter has a small, linear phase delay difference from its neighbors, a plane wave with an angular diffraction offset will be produced. The angle is determined by the difference in the phase of each subsequent emitter. Furthermore, focusing is achievable using phase offsets which are in opposite directions on either side of the central waveguide, similar to the wavefront bending induced when light propagates through a lens.

Phased arrays have been implemented in 1-D and 2-D configurations previously [27, 114, 122], however these implementations only function at infrared wavelengths. These devices were used for beam steering, typically utilizing a tunable laser source to sweep the wavelength or utilize heating on the chip to change phase differences while using a monochromatic source. In this section, a similar system is implemented for use with visible wavelength light. First prototypes were designed and fabricated around a 673nm central wavelength to test the ability to do beam steering in the visible, since a process for high quality optical device fabrication had been developed around this wavelength and a source was available at this wavelength. Further implementations were then developed for beam steering to enable use of these devices for shorter wavelengths for optogenetic stimulation.

4.3.2 Implementation of Phased Array Technology

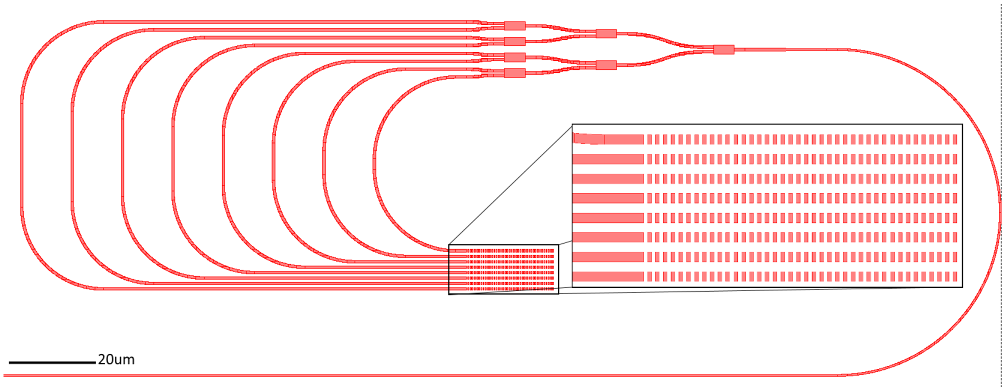

Figure 4.16: CAD layout for individual phased array circuit, central wavelength = 673nm. Designed by W. Sacher, Roukes Group, Caltech.

In the phased array configuration, the length of delay lines establish the dispersion of the beam. This, of course, is set during fabrication similarly to the design of the AWG above. Typically for beam steering applications, the delay between subsequent lines is constant, creating a uniform wavefront from the surface of the gratings in a direction defined by the difference in phase between each phase delay line, although focusing configurations are also possible.

When considering the fabrication of such a device, it is possible to borrow from microwave engineering to implement phased array designs. The mathematics behind the design of the phased array was described in detail in reference [121], with the relevant mathematics reproduced below. The analysis of the phased array is typically taken from the perspective that an individual emitter within the phased array has a geometrically defined far-field beam profile, and this is multiplied by a factor, T, which includes the summation of all the beam phases. Therefore, the analysis and design of the phased array is predominantly focused on T, the array factor. The array factor for a uniform array in 1-D is written as

$$T = \sum_{n=0}^{N-1} A_n e^{-j\beta_n} e^{jks_n} = e^{-j\beta_n} e^{jk_0(n\Lambda_y \sin\phi)}, \tag{4.9}$$

where n is the numerical index of the delay line, N is the number of delay lines, A_n is the intensity of each emitter (assumed to be 1), β_n is the phase delay in each waveguide, Λ_y is the spacing between subsequent output gratings, and ϕ is the angle of measurement from the normal vector of the gratings. Therefore, it is important to determine at which angle, ϕ, T is at a maximum. Noting that $\beta_n = n_{\text{eff}}(2\pi/\lambda)n\Delta L$, and by assuming that each emitter is identical except for its phase, the summation can be rewritten as

$$T = e^{j(k_0\Lambda_y \sin\phi - k_0 n_{\text{eff}}\Delta L)(N-1)/2} \left(\frac{\sin\left(N\frac{k_0\Lambda_y \sin\phi - k_0 n_{\text{eff}}\Delta L}{2}\right)}{\sin\left(\frac{k_0\Lambda_y \sin\phi - k_0 n_{\text{eff}}\Delta L}{2}\right)} \right). \tag{4.10}$$

Thus, the angle of emission for a grating is calculated by maximizing T, which occurs when the term inside the sin functions is $(q \times \pi)$, where q is a constant integer. Therefore, substituting for k_0 we obtain

$$\sin\phi = q\frac{\lambda}{\Lambda_y} + \frac{n_{\text{eff}}\Delta L}{\Lambda_y}. \tag{4.11}$$

This illustrates that, given a spacing of gratings, the phase delays in each delay line, and the wavelength, the emission angle can be calculated.

Phased arrays were designed using the following parameters. For the given waveguide dimensions (w = 500nm, l = 300nm), the effective index of the TE mode was found to be 1.8108 and the group index of the waveguides was found to be 2.2071. The gratings have a period of 400nm and a width of 200nm. Phased array elements were spaced 1 micron apart from each other. The resulting ΔL that optimized the parameters for the desired angle per degree was found to be $\Delta L = 21\mu$m, with a q of 69. The free spectral range of the device therefore was found to be

$$\text{FSR} = \frac{(673\text{nm})^2}{2.2071 \times 21\mu\text{m}} = 9.77\text{nm}. \tag{4.12}$$

In addition, the distance in angular space between the each maximum of the diffraction in the phased array direction can be calculated. From theory, the distance between maxima is calculated as

$$\sin(\theta) = \frac{\lambda}{\Lambda_y} = .673, \tag{4.13}$$

where the arcsin of .673 is 42.3 degrees.

4.3.3 Fabrication of Phase Array

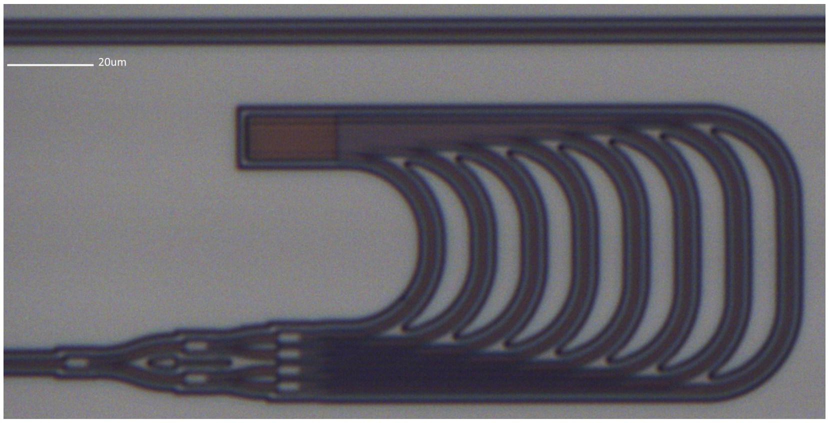

Figure 4.17: Micrograph of resulting phased array structure after fabrication.

Fabrication of the phased array designed for 673nm light started with chips with a layer structure starting with a 300nm top layer of LPCVD silicon nitride on top, followed by a 2μm thermal silicon dioxide cladding layer, and a 550μm silicon handle. Chips were diced from 4 inch wafers, and were prepared using a solvent clean followed by a 2 minute dehydration bake at 180^{o}C. Chips were spin coated with a layer of ZEP-520a at 4500RPM for 30 seconds, and were baked for 2 minutes at 180^{o}C. Chips were then exposed to electron beam lithography in a Raith EBPG 5000+ EBL tool, patterns were fractured with a resolution of 2.5nm, and a 300pA beam was used to write the structures. A 2μm trench around the structures was created during fracturing due to the positive tone of the resist. Resist was developed in ZED-N50 developer for 1 minute, and chips were then rinsed with isopropyl alcohol. Chips were then reflow baked at 152^{o}C for 1 minute on a hotplate to reduce edge roughness of the resist.

Next, chips were etched in a Oxford Instruments ICP 380 etching tool using the recipe in section 3.4.2 intended for ZEP resist etching. The rate of etching for this process is approximately 75nm per minute, thus the chip was etched for 4 minutes. After etching, the chip was exposed to an oxygen plasma for 40 seconds to remove any ZEP from the surface. Chips were then cleaned in hot PG remover for 5 minutes, rinsed with acetone and isopropyl alcohol and were finally dried using compressed nitrogen gas.

4.3.4 Results

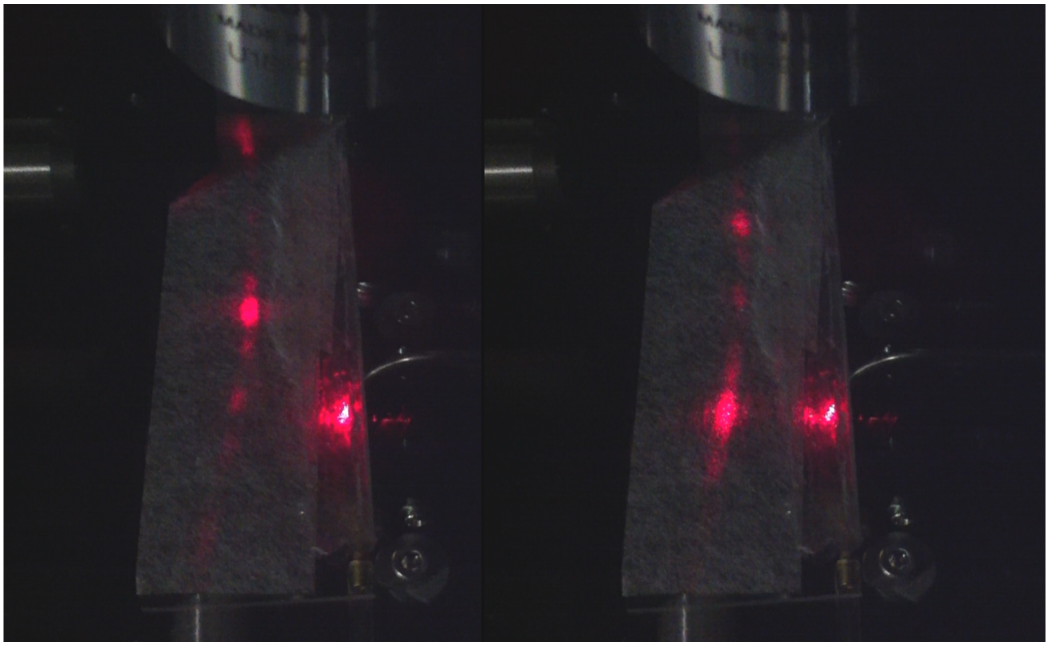

Figure 4.18: Wavelength induced phased array beam scanning using tunable laser source. The beam was projected onto a Kimwipe which scattered the light, showing the location of the beam maximum.

Qualitative testing of the phased array was accomplished using a very simple experimental setup. Light from a fiber-coupled tunable laser source (Newport Velocity TLB-6309, 6mW, tunable from 668-678nm, Newport Corp.) was coupled through a simple grating coupler into the silicon nitride waveguides patterned onto test chips. The light was projected off the chip from the output grating couplers onto a scattering sheet (a Kimwipe), and video was taken as the wavelength of the source was scanned from 668-678nm. The results of this experiment are shown in figure 4.18. The beam was scannable over an output range of approximately $\theta = \pm 20^o$.

The extent of the scanning ability and the resulting beam profile of the was next determined using a a more sophisticated Fourier imaging system, intended to measure the angular deflection of the beam by transforming the resulting beam into k-space. This was accomplished by coupling the light from the phased array into a long working distance infinity-corrected microscope objective (Mitutoyo 20x, NA = 0.42). Since the objective is infinity corrected, a second lens (f = 60mm, achromatic) was added to the setup to act similarly to the tube lens in a microscope, and also to transform the light from the objective into the Fourier domain. Therefore, the beam would be deflected based on the angle it entered the objective. Light passing through the second lens was then allowed to illuminate a linear CCD sensor (FLIR Grasshopper3 CCD). Therefore, based on the locations of the pixels illuminated, the angle of the incoming beam could be determined, based on the deflection from the center of the optical path. The position of the second lens was determined by ensuring that the entire back aperture of the objective was present in the field of view, and that it filled the majority of the frame.

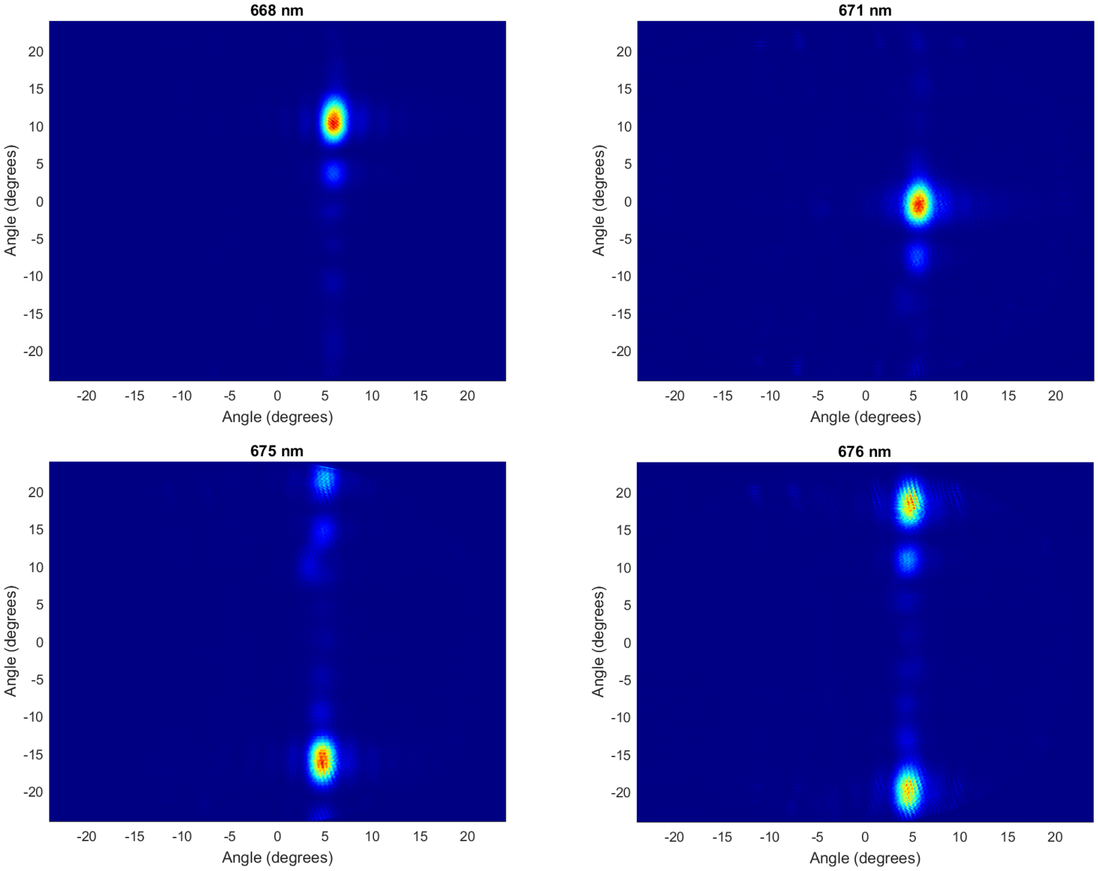

Figure 4.19: Fourier image of phased array beam steering at four wavelengths.

Angular calibration was accomplished by illuminating the entire back aperture of the microscope objective by focusing on the grating and turning up the gain on the camera. Given the known numerical aperture of the objective, imaging the extent of the back aperture provides a concrete fiducial which defines the maximum extent of the angle admitted by the objective. Thus, by measuring the location of the back aperture, it is possible to directly calibrate the angle-per-pixel linear transform simply by counting the extent of the back aperture in pixels, and dividing the angular extent of the objective by the number of pixels measured. This is allowed because the angular projection onto the Fourier plane is very linear with these objectives.

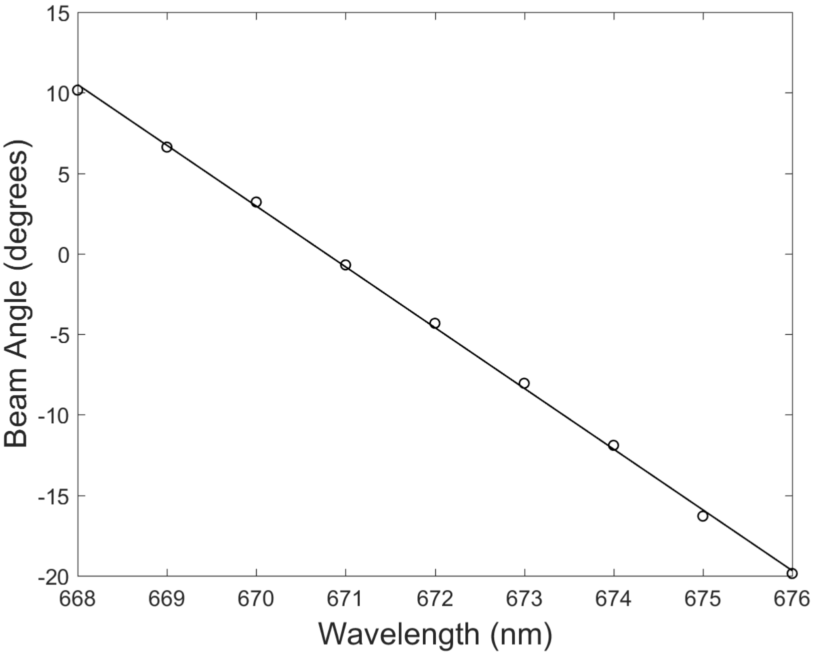

Figure 4.20: Angle versus wavelength relationship of beam. The resulting dispersion relationship shows a 3.77 degree deflection per nm.

The resulting angular deflection of the beam is shown in figure 4.19. The full-width-at-half-maximum of the beam is 2.2 degrees in the grating direction and 4.7 degrees in the array direction. The angular extent of the beam in the array direction can be improved by adding more phased array elements, which will be done in the future iterations of these devices. The angular extent between diffraction orders is shown in figure 4.19 in the lower right image. Diffraction orders are separated by 42 degrees, allowing for a large tuning range without interference from multiple diffraction orders. The resulting beam deflection versus wavelength is shown in figure 4.20. This also approximately fits the theoretical prediction of 42.3 degrees. The resulting profile is very linear, with an angular rate of deflection of 3.77 degrees per nm.

4.3.5 Limitations and Future Work

There are two important next steps in the development of these devices. First, it will be ideal to continue to scale these devices down to shorter wavelengths which correspond with the excitation maximum of the most popular opsin proteins in the range for 470-500nm. This shouldn't be too difficult from a fabrication standpoint as we have already demonstrated the ability to fabricate optical devices at these wavelengths. The primary issue to overcome is with the source of the light. The presence of tunable laser sources at this wavelength is likely required for these applications; however they are currently not produced commercially. The most attractive alternative is the use of a supercontinuum laser which produces light over the entire visible spectrum, however at a lower power per nm. These supercontinuum lasers can be tied to a computer-tuned optical cavity resonator which will filter the incoming broad spectrum into a narrow, 1.3 nm wide band. These can be tuned quickly, but they still suffer from lower power than a traditional laser.

Multiple improvements can be made to the next generation of phased arrays. A primary issue with the design described above is that the beam profile has significant energy located in the side-lobes of the beam. This is due to the windowing chosen by the use of MMI couplers. MMI couplers split the power evenly between each phased array element, which results in a rectangular window in k-space. The Fourier transform of this square window produces a sinc function in the beam space, which has significant power in the side-lobes, as seen in figure 4.19. By replacing the MMI tree with a star coupler, a Gaussian profile is produced instead of the rectangular window, which will produce a beam with significantly smaller side-lobes.

A second issue is the anisotropy in the beam profile. Due to the number of elements in the grating versus the number of elements in the phased array, there is a mismatch in the geometry of the beam causing an elliptical beam profile. Although it may be tolerable in some applications, the ability for different neurons to be addressed is defined by the width of the beam profile. By narrowing the beam profile, it follows that more neurons can separately be stimulated by the beam. This issue is solved simply by increasing the number of phased array elements incorporated into the phased array. By increasing the number of elements to 12 or 16, we believe that the beam profile resulting from the phased array will show a circular or nearly circular beam profile.

The final issue observed with these devices is that the power was found to be quite variable over the wavelengths tested (668-678nm). This is likely due to a Fabry-Perot resonance in the lengths of waveguide that make up the phased array. Although reflections cannot fully be removed from an optical system such as the one described above, there are certain design elements which can be used to minimize this behavior. Changing from the MMI tree to a star coupler may help in this regard. Further design consideration must be made when designing the next generation of these devices.

4.4 Spatial Excitation of Optical Waveguides

4.4.0.1 Edge Couplers and Illumination Optics

The optogenetic probes described above utilized grating couplers to bring light onto the planar lightwave circuits which compose the optical layer of the probe. Grating couplers were a convenient starting point because they are highly efficient and easy to design. However, grating couplers have two major practical drawbacks. The first drawback is specifically related to packaging of probes, as it is impossible to package probes such that the fiber is coming off the surface at an angle near 90 degrees. This necessitated the use of the microprism, as described in section 4.2.3 above. Packaging probes in this way is labor intensive and slow, showing that lateral coupling using microprisms is undesirable for probes which require higher input densities.

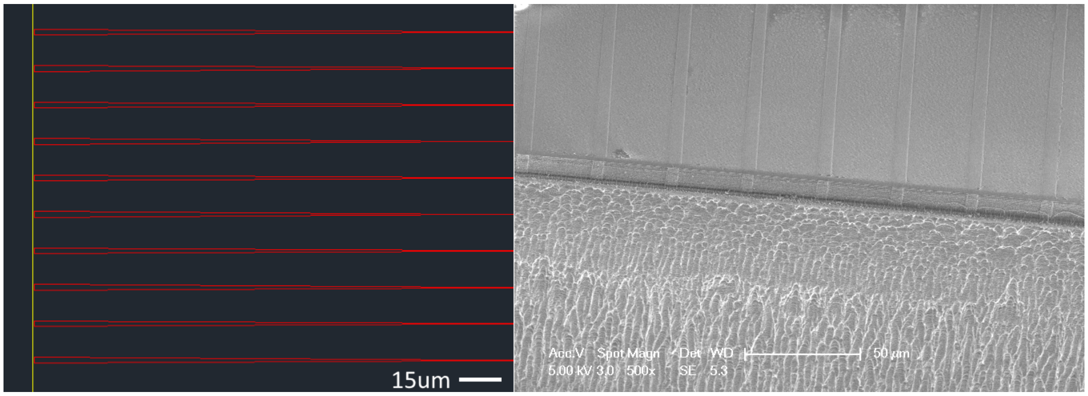

Figure 4.21: Edge couplers. Left, CAD drawing of edge coupler tapers. Right, scanning electron micrograph of edge couplers.

The second major drawback to grating couplers is coupling a large number of optical fibers to a single probe. The packaging method using fibers directly to gratings is bulky and unwieldy, requiring individual gluing of fibers for each socket. There are technologies available for coupling multiple fibers to a chip simultaneously (such as v-groove fiber holders), however these are incomparable with the probe designs described previously. There is just not a good way to laterally couple multiple fibers to multiple grating couplers while maintaining an in-plane mechanical arrangement for probe implantation.

To remedy these problems, a paradigm shift to using butt-coupled waveguides was made. Butt couplers have the advantage of allowing for dense coupling of multiple fibers all within the plane of the probe. This is critical for implantability of probes in tight spaces, such as when coupled to microscopes. Furthermore, butt couplers are regularly spaced (see figure 4.21), allowing for simultaneous coupling of multiple fibers of known spacing quickly and easily. Finally, butt coupling is superior in that it is wavelength- and polarization- independent, issues which can cause problems with grating couplers especially when using multiple wavelengths in the same waveguides. Light is coupled from wide sockets (5.2μm) matching the core diameter of the input fiber, and are adiabatically tapered[34] to a single mode waveguide of width 600nm. The major disadvantage of edge couplers is that they are not as efficient as grating couplers, losing as much as 10dB per coupler.

To easily address multiple waveguides of a probe using a single laser, a MEMS mirror was used for optical switching between channels. Laser light is directed to different cores of a fiber bundle with thousands of micron scale optical cores via a scanning system. The scanning system consists of a MEMS mirror for $\approx$10ms switching between cores, and relay optics for coupling to the fiber bundle. The fiber bundle is edge-coupled to the probe chip using on-chip adiabatic edge couplers, enabling broadband optical coupling over the visible spectrum. Light coupled from the fiber bundle

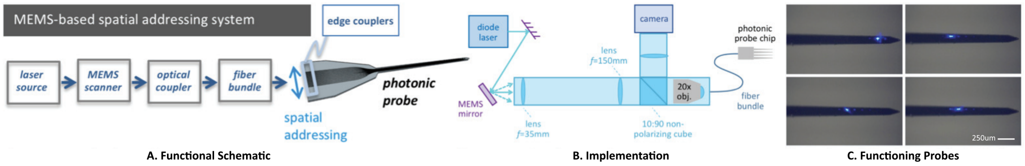

Figure 4.22: Input coupling methodology. (a) Using this system, a laser is directly coupled into a fiber bundle, where a 1-to-1 mapping of the fibers-to-probe inputs is created. Thus by scanning the laser to the correct fiber core, it is possible to illuminate that circuit. (b) schematic of the scanning system, (c) illumination of individual cores in the fiber bundle with corresponding output from an optical probe.

to the probe chip is routed by the on-chip waveguides to the E-pixels and emitted. The scanning system was designed and constructed by W. Sacher and L. Moreaux in the Roukes Group, Caltech, in collaboration with Prof. Joyce Poon's group at the University of Toronto.

4.5 Application: Excitation of GCaMP-6

4.5.1 Genetically-Encoded Indicators

Current trends in neuroscience see optical methods as the future for parallel, multiscale recording of neural activity. Although the driving force in 20th century neuroscience was electrical recording, these methods are currently reaching their maximal potential for recording densities. The major issue with high density electrical recording is that as the desired number of neurons under concurrent measurement is scaled up, it requires larger and larger implants into neural tissue, which is disruptive and often destructive to the neural tissue. Optical technologies have the benefit of being less perturbative and damaging while allowing for massive parallelization under the right circumstances. Therefore, modern approaches to neuroscience are currently focused on the production of new optical techniques for the detection of neural activity.

Since the isolation and subsequent cloning of GFP, genetically-coded indicators using fluorescence has been one of the most important technological advances in biological experimentation, even leading to the Nobel Prize in 2008[100, 86]. Although originally limited to tracking proteins by creating fusion proteins, fluorescent proteins have been developed into a number of interesting and useful applications as sensors. The first fluorescent sensors were based on fluorescence resonance energy transfer (FRET), where binding of a molecule to a coupled pair of fluorophores would alter the wavelength of fluorescence by transferring energy from one fluorophore to the other. Other methods of sensing are also possible, primarily where binding of a molecular species to the protein changes the electronic state of the fluorophore, inducing a change in the fluorescence intensity.

The primary physical targets of genetically-encoded indicators for neuroscience applications is to measure either calcium concentration as a proxy for neuronal spiking or as voltage indicators that directly measure membrane voltages in cells[63]. A variety of calcium sensors have been developed, starting with the discovery of the calcium sensitive aequorin from jellyfish[100]. The first series of genetically-encoded calcium indicators which were man-made were the Chameleons, which coupled GFP to a calcium binding protein, calmodulin[74]. This protein uses FRET to modulate the signal from the GFP molecule; when calmodulin binds to calcium it undergoes a conformational change, which will bring a CFP molecule in proximity to a YFP molecule, inducing FRET. This showed a sensitivity on the scale of μM, with a range between 0-1mM.

As time went on, incremental improvements were made to the Chameleon sensors until a new paradigm entered the field. The new competitor was the GCaMP series of calcium indicators. Originally developed by Junichi Nikai[76], these proteins are a fusion protein consisting of a single, circularized GFP molecule fused with the M13 chain of myosin light chain kinase and calmodulin. Due to the circularization of the GFP molecule, a new terminus is located within the core of the protein, to which the calmodulin and M13 are attached. When calcium is bound to calmodulin, the GFP remains in its normal state which fluoresces strongly. However, when calcium is not bound to the calmodulin, the GFP molecule is stressed, quenching the fluorescent activity of the molecule. This creates a calcium indicator with significantly better signal-to-noise performance compared to the chameleon indicators. The structure of GCaMP has been continually improved by enhancing the stability of the molecule and through improved fluorescence of the GFP molecule. Currently GCaMP is on its 6th iteration (GCaMP6).

Although calcium imaging was revolutionary to the field, it has some major drawbacks. The primary drawback of calcium imaging is that it doesn't directly report the activity of the cell, only reporting a proxy of activity in intracellular calcium concentration. As a consequence, calcium transients are much slower than the individual action potentials and thus calcium imaging has a limitation in that it has a much slower time constant than the actual activity of the neuron. Thus, the development of genetically-encoded voltage dyes were motivated. Since the mechanism of transduction of an action potential is voltage, it has clear advantages over calcium concentration as a parameter to measure.

The first genetically encoded voltage sensor, FlaSh, was developed by Micah Siegel and Ehud Isacoff in 1997[101]. This protein was a fusion between GFP and a potassium channel. Since then, a variety of different protein fusions have been attempted, however no option has yet to satisfy all of the demands necessary for robust functional imaging of membrane voltages[63]. The main sensors used by these fusion proteins are either opsins or ion channels, both of which respond to transmembrane voltage changes. These can be coupled to FRET pairs or directly to fluorophores. To some, establishing a robust and bright genetically encoded voltage indicator is the holy grail of functional neural imaging, however the perfect combination of proteins has yet to be discovered and research is still ongoing in this area.

4.5.2 Modern Microscopy

With the current development of all these interesting, novel genetically encoded indicators, the focus moved to determine the best methods to image these probes. By and large the enabling technology for high quality optical interrogation of tissues was the development of the scanning confocal microscope by Marvin Minsky in 1955[73]. This technique uses a standard epifluorescence microscope along with specialized illumination optics to create high spatial accuracy images of three-dimensional specimens with low noise from fluorescence in the outside the focal plane of the microscope. Confocal microscopes combine a trick of optics with a pinhole to ensure that only light from the focal plane of the microscope is collected. This is accomplished by adding a pinhole to the optical path, where the fluorescent light is focused through the objective and into a secondary tube lens, which then focuses the light through the pinhole before reaching the detector (typically a photomultiplier tube). As light is collected from the objective lens, only the light originating from the focal point of the objective will be perfectly collimated when leaving the back aperture of the objective. Therefore, by re-focusing the light coming from the objective and placing a pinhole at the focus, only the perfectly collimated light originating from the focal point of the objective will pass through the pinhole. Thus, any out of focus light from outside of the focal point of the objective will be rejected by the pinhole filter, thereby ensuring that the light is coming from the focus of the objective.

The confocal technique is able to produce high quality images in 3D structures by scanning the focal point over the image plane of interest. This is accomplished using a pair of mirror galvanometers which can rapidly scan the focal point optically over the surface without having to move the sample. Due to the confocal nature of this type of microscopy, images are created by scanning the focal point of the optical plane over the entire optical field, taking data from one pixel at a time. Images from multiple focal planes can be taken from a sample by scanning the sample stage up or down in the z-plane of the imaging system. Thus, 3D images of complected samples can be formed at a resolution nearing the diffraction limited focal point of the objective.

Although confocal microscopy produces high quality 3D images at or near the diffraction limit, the technique suffers from a number of drawbacks. One drawback is the depth of imaging. As light is projected into any tissue, as a consequence of the heterogeneity of any tissue light will become scattered. Although the majority of light is scattered ballistically, enough scattering will cause the focal point to become larger and less intense. To solve this problem, two-photon microscopy was developed in the lab of Watt Webb at Cornell University[26]. As a consequence of cellular geometry and the mathematics behind scattering, the scattering process is wavelength dependent. As the wavelength of light becomes longer, loss of beam energy and direction is reduced while propagating through tissue. Therefore, using two photon excitation fluorophores can be excited by light at twice the wavelength of their normal absorption, increasing the depth of imaging significantly compared to confocal microscopy. This technique has revolutionized functional imaging by allowing the measurement of neurons below the superficial layers allowed by confocal microscopy alone.

However, this technique is still limited by the signal-to-noise ratio for each pixel as depth increases. Although two-photon imaging improves delivery of energy to the targeted fluorophores, the light returning from the fluorophores to the microscope are still in the visible regime and subject to scattering and absorption at shorter wavelengths. Thus, as as the focus moves deeper into tissue, longer integration times are required to overcome the noise in the detector. This causes integration times to increase as the focal depth increases, and sets the imaging speed of the microscope system.

Currently, the primary drawback is the speed of imaging required for multicellular calcium imaging studies. The current state-of-the-art for confocal microscopy is the RAMP microscope [88, 21] which utilizes acousto-optic modulation to deflect the beam in lieu of using a galvanometer system and two-photon imaging to improve the possible depth of imaging. Although the use of acousto-optical modulation can improve speeds significantly, they max out at a rate of 50,000 samples per second. The newest techniques in this field of microscopy utilize random access to take only a single sample per neuron and a sample rate for the neurons was 30Hz (which is necessary to measure the decay of the GCaMP6 signal), only 1600 neurons would be addressable. Although this is a leap forward beyond what is currently possible for live cell imaging, it reaches a maximum potential that is still far away from measuring full neural circuits. To achieve the next generation of live cell GCaMP6 imaging, other techniques beyond confocal microscopy must be explored.

Furthermore, the use of probe-based illumination and optical detection has the possibility to greatly impact the imaging space by accessing geometries not allowed by conventional microscopy. Unlike the rigid coordinates imposed by a typical microscope, probes can be inserted with a vast array of orientations will enable the imaging of deep and/or highly constrained brain regions. Furthermore, novel illumination patterns will enable novel imaging applications for neuroscience. Below, two imaging schemes which are currently in development in the Roukes group are described, including laminar illumination for rapid imaging of tissue as well as probe-based fluorescence microscopy.

4.5.3 Implantable Laminar Illumination of Tissue

Light sheet fluorescence microscopy (herein LSFM) is a new technique under development. This technique utilizes structured illumination to minimize fluorescence signal out of the image plane (similar to confocal microscopy) while eliminating the need to scan a single point along the image plane to create an image (similar to standard epifluorescence microscopy). This will, in turn, increase the imaging speed hundreds to thousands of times by using a CCD detector instead of a single point detector like a photomultiplier tube.

Originally called orthogonal plane fluorescence optical sectioning by Voie et al.[125], light sheet microscopy utilizes a thin plane of illumination (the light sheet) projected from the side of the sample to create a single plane of illumination in the sample. The illumination was originally created by a cylindrical lens projecting the flat beam into the side of a sample, and using a microscope to image the sample from the top. Multiple setups have since been developed, but the basic principle is the same[38]. Since the plane of propagation of the light sheet is coincident with the image plane of the microscope, there is no out of plane excitation of the fluorophores in the sample. This significantly improves signal to noise ratio and provides high quality sectioning of biological samples.

Many methods of light sheet illumination are currently under development. The first method of light sheet was to scan a Gaussian beam across the image plane of the imaging objective. However, due to diffraction of the beam, it was found to be a problematic implementation due to the width of the beam. As an improvement of using a galvo-scanned Gaussian beam, Bessel beam light sheet utilizes a galvo-scanned Bessel beam to create planar illumination of the image plane from the side of the sample [82]. The Bessel beam is a specific type of beam which is said to be "diffractionless," meaning that it has low divergence as it propagates through space. This is critical for light sheet imaging, as it is desired to provide a uniform sheet of illumination as light propagates through the sample. The main disadvantage of this type of beam is that it keeps much of its energy in side lobes, and thus image processing is often necessary to remove out-of-plane artifacts from the resulting images. This also induces increased phototoxicity outside of the image plane of the microscope. Continuing this pathway of successive improvement, a lattice based approach was developed and described in [19]. Similar to Bessel beams, a 2D optical lattice has limited diffraction over long distances. Optical latices are essentially interference patterns of complex beams formed using a spatial light modulator. This technique is also advantageous in that it doesn't require mechanical scanning of the beam. These can be designed to have low divergence, and are utilized in the current state of the art light sheet microscopes.

One major drawback associated with LSFM comes from the geometric considerations when illuminating from the side of a sample. Since the beam must propagate through the sample from the side, this technique is currently limited to rather small samples which can be easily suspended, such as zebrafish embryos. These are often fixed samples suspended in a clear gel which can be easily manipulated, which severely limits the application space of this technology. Furthermore, light sheet broadening due to scattering is also an issue. The light sheet loses the ability to remain in the image plane if it is projected through tissue for too long, a similar problem to standard confocal microscopy.

To address these issues, we are currently developing an implantable optical system for light sheet fluorescent microscopy. The concept for this system was developed in the Roukes group by Laurent Moreaux and Wesley Sacher. This system will utilize an photonic neural probe, similar to the one discussed above, to create laminar illumination by broadening the output of the grating couplers in one direction and spacing them so that the beam profile emerging from consecutive grating couplers overlap, thereby creating an even sheet of illumination. The main advantage of this technique is that it is directly implantable, thereby circumventing the disadvantage of LSFM described in the preceding paragraph. Implantable light sheet technology will allow for the use of light sheet microscopy on arbitrarily large samples.

The eventual goal of this project is to create a rapid method for parallel measurement of calcium reporters (such as GCaMP6) in vivo. To accomplish this, however, it is not as simple as implanting a light sheet probe and imaging with a microscope. Due to the compactness of the microscope objective in proximity to the probe, it would be difficult or impossible to take images using this setup. Thus, a gradient index (GRIN) of refraction lens will be used to improve the mechanical constraints of the microscope-light sheet probe system. GRIN lenses utilize a gradient of refractive index in their glass to create a focusing effect (instead of using curvature as in standard lenses). This allows for compact, cylindrical lenses to be created for imaging systems. In this case, a GRIN lens will be used to extend the focal length of a microscope objective to create more space between the objective and the probe. Rapid focusing between layers will also be necessary for high data rate applications. This will be accomplished by coupling the microscope objective to a liquid lens, allowing for rapid tuning of the focal depth of the microscope.

Overall this system will be revolutionary for rapid, parallel imaging of fluorescence in tissues. The illumination scheme allows for low-noise capture of deep signals in parallel, and should have depth resolution comparable to that of two-photon microscopy. By allowing a single sheet to be measured simultaneously, the limitation of the rate of capture of data is not limited to single-pixel-at-a-time but by the necessary integration time of an optical imager such as a CCD or scientific CMOS sensor. Overall, this should allow for high data rate monitoring of thousands of neurons in a relatively large sample volume (500μm by 500μm by 320μm).

4.5.4 Implementation of Waveguide Routing for Light Sheet Probes

4.5.4.1 Optical Circuitry

The major complication in the design is the routing of light to multiple output gratings using single layer photonic lightwave circuit. Since the desired applications for this system require multiple layers of illumination and multiple gratings per layer, it is necessary that there be crossings between waveguides to simultaneously route light to each layer of gratings. This will be done using in-plane waveguide crossings, as described in [9, 94]. This technique utilizes tapered crossings in a four-pointed star pattern, as shown in figure 4.23 on the left. Each waveguide enters the crossing and the light is tapered out, before reaching a straight, multi-mode section of the waveguide. The length and width of the straight region is designed such that there is an interference maximum in the center of the crossing. In this way, a minimal amount of light is interacting with the corners from the orthogonal waveguide in the crossing, allowing minimized losses into the orthogonal waveguide.

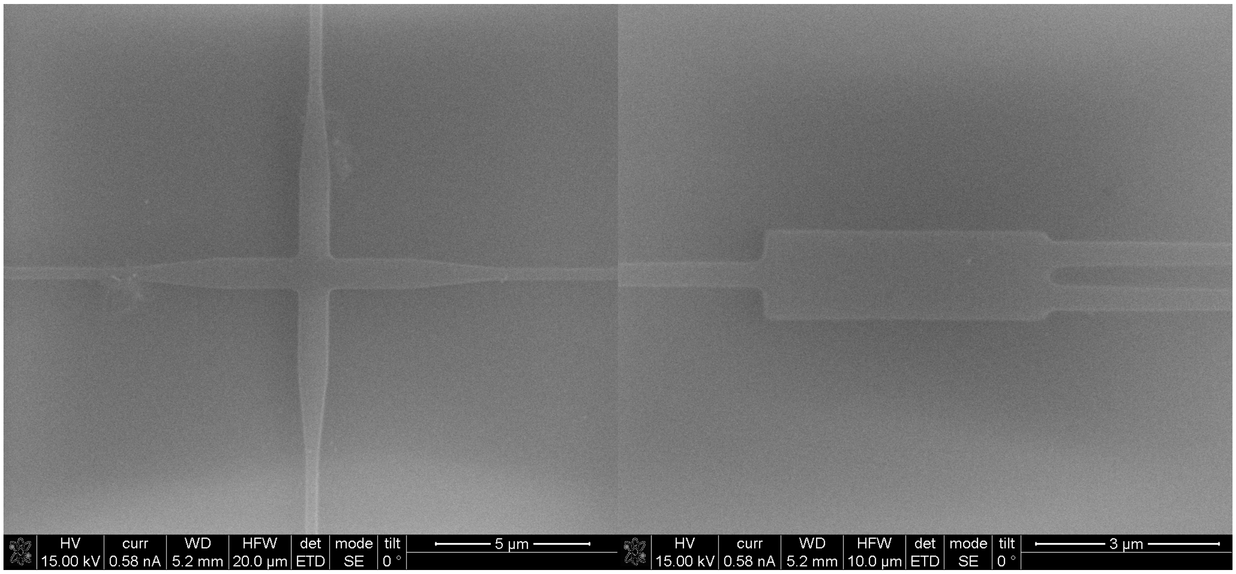
HV curr WD HFW det mode tilt
15.00 kV 0.58 nA 5.2 mm 20.0 μm ETD SE 0 °
5 μm
HV curr WD HFW det mode tilt
15.00 kV 0.58 nA 5.2 mm 10.0 μm ETD SE 0 °
3 μm

Figure 4.23: Critical optical structures in routing design. On the left an inplane crossing is shown. On the right, a multimode interference splitter is shown.

Light is then tapered back into a routing waveguide, and sent to the next optical structure. The specific geometry of the crossings, as shown in figure 4.23, were designed and optimized using finite differences methods.

The other primary structure utilized is the multimode interference splitter[47, 120]. Since it is desired to have uniform, even and simultaneous stimulation of all gratings on a single layer of the light sheet probe, power splitting is necessary to ensure even and uniform illumination. Multimode interference power splitters allow for multimode propagation of light in a widened waveguide, where reflections off of the sidewalls are allowed to interfere with the central beam creating multiple maxima and minima within the multimode region of the waveguide. By selecting the width and the length of the mutlimode region, it is possible to ensure that the maxima of the propagating light correspond to output sockets of waveguides, thereby efficiently coupling the light into an arbitrary number of output waveguides. This structure for a 1 × 2 splitter is shown in figure 4.23. The geometry of this splitter was designed and optimized using finite differences methods.

4.5.5 Fabrication of Optically-Integrated Neural Probes

4.5.5.1 Mask Design

The critical consideration when designing the optical layer of the mask was the trade-off between minimizing the number of crossings between waveguides while maintaining uniform illumination of the gratings as to produce an even light sheet. When attempting to minimize routing of the waveguides (8 gratings per sheet, 5 sheets), the minimum number of crossings necessary to route

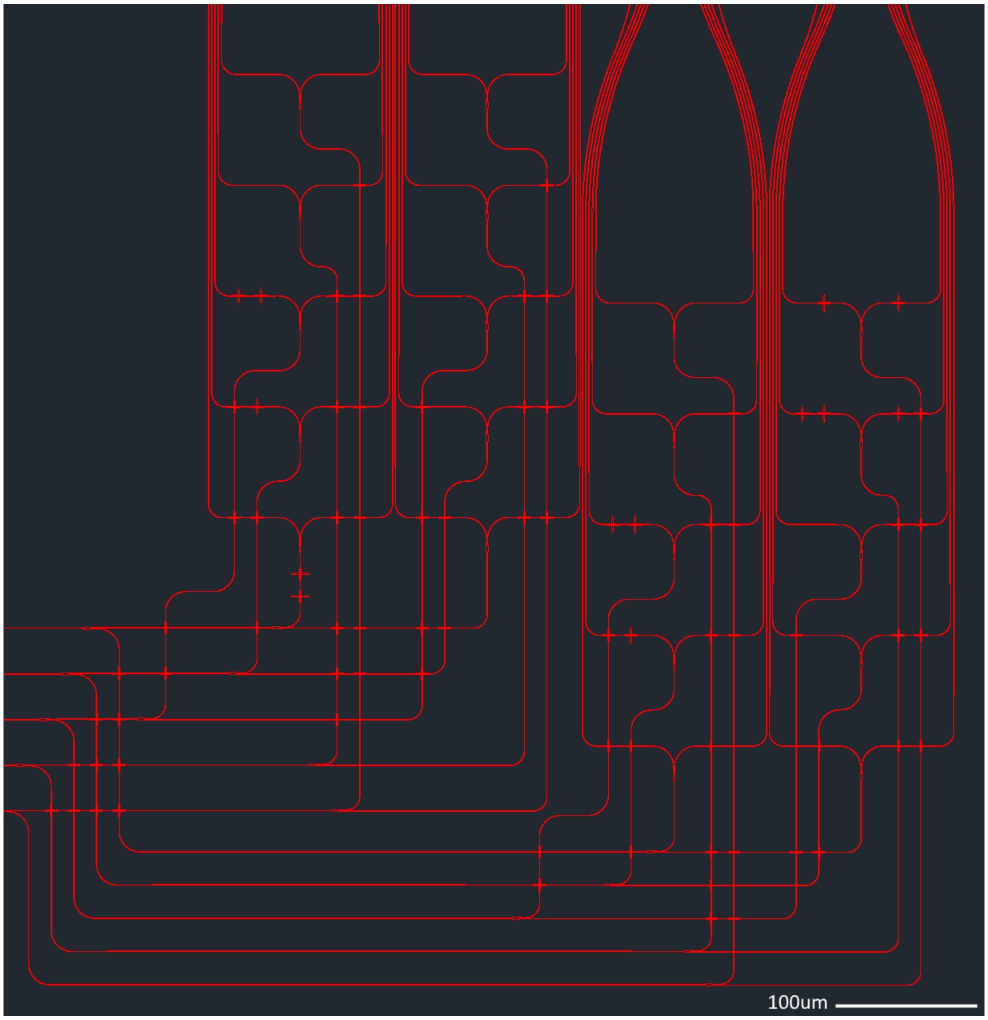

Figure 4.24: Waveguide routing including MMI splitters and in-plane waveguide crossings. Light enters the waveguides from the left, and is subsequently split down to address the 8 gratings per light sheet layer. Light leaves the router towards the top of the figure and is sent down the length of the shanks to the emitter gratings. Designed by Wesley Sacher, Roukes Group, Caltech.

all of the waveguides is 7 crossings per waveguide. However not every waveguide needs this many crossings. Thus, to ensure even illumination, so-called "dummy" crossings were added where only one waveguide passes through the crossing. This ensures that the losses accumulated over the length of the waveguide are equal when reaching each grating coupler at the end of transmission. The light reaching each grating will also need to go through 4 MMI power splitters. These photonic circuits were designed by Wesley Sacher, Roukes Group, Caltech.

Gratings for the light sheet probe are designed differently than for the optogenetics probe described above. The grating couplers do not have a nearly symmetrical, square socket but an elongated, narrow socket. The narrowness of the grating ensures that light will be scattered along a broad cross-section laterally, as the socket width is inversely proportional to the emission angle of the light from the grating. Likewise, the grating is made extremely long to ensure that the emission profile in the other direction is narrow. Combining these two design parameters, the resulting emission from the grating each grating coupler will have the shape of a thin, laterally divergent beam. The beams from each grating on a single layer will propagate and overlap with each other, creating a thin, even sheet of illumination in the tissue.

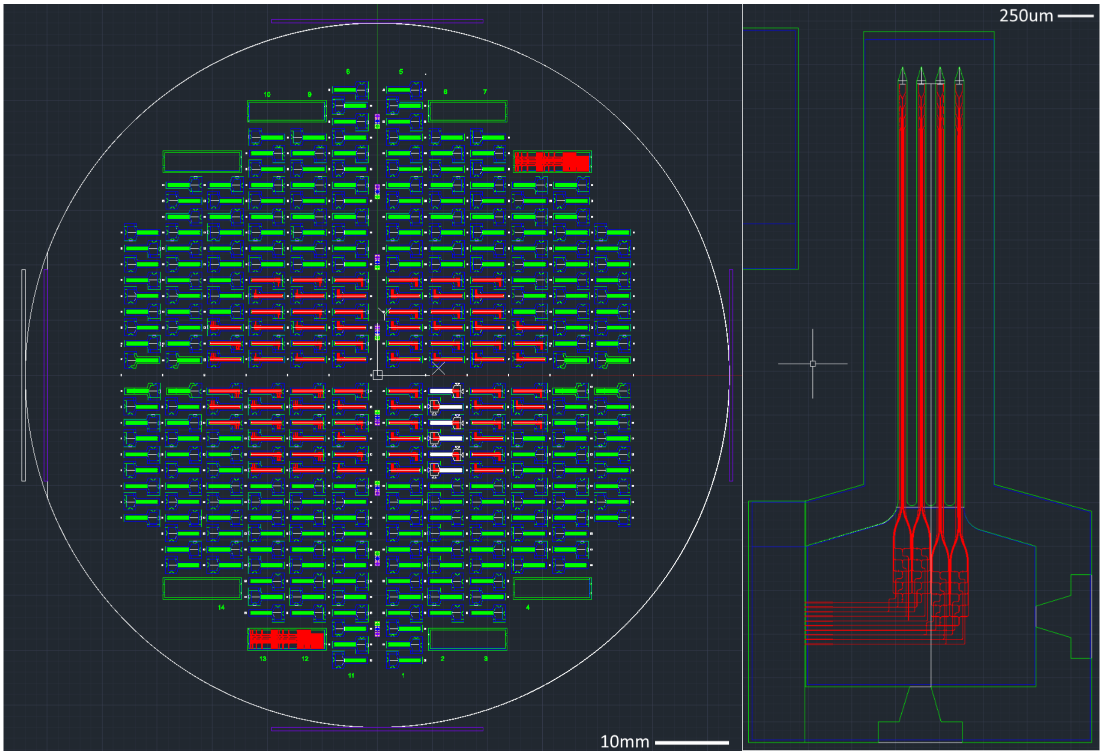

Figure 4.25: Mask design for lightsheet probes. Red = optical waveguide layer, green = top side photolithography mask, blue = backside photolithography mask, white = reference features.

Wafer scale mask design is shown in figure 4.25, as designed by the author. Probes were designed with two geometries, either with four shanks (2 gratings per layer per shank) or with 8 shanks (1 grating per layer per shank). The 8-shank design is intended to be superior in the sense that it displaces less tissue when implanted, but it was feared that the shanks might be too fragile, so for preliminary tests the 4 shank design was preferred. To ease hindrance during implantation, probes were designed such that the optical input would come from the left side of the probe instead of the back end of the probe. This ensures that the fiber bundle will not interfere with the microscope objective during implantation and measurement. 14 designs with variable configurations of the light sheet (1 or 2 inputs per layer, different numbers of crossings, different arrangements of gratings, etc.) were designed and placed on the same optical layer for testing. Shanks have lengths of 3.1mm and widths of 65μm for the 4 shank design. Base dimensions were 1.65mm×1mm, with a 230μm taper. A total of 324 probes are available per wafer.

4.5.5.2 Fabrication protocol

See section 4.2.2 above for fabrication details.

4.5.6 Packaging of Optical System

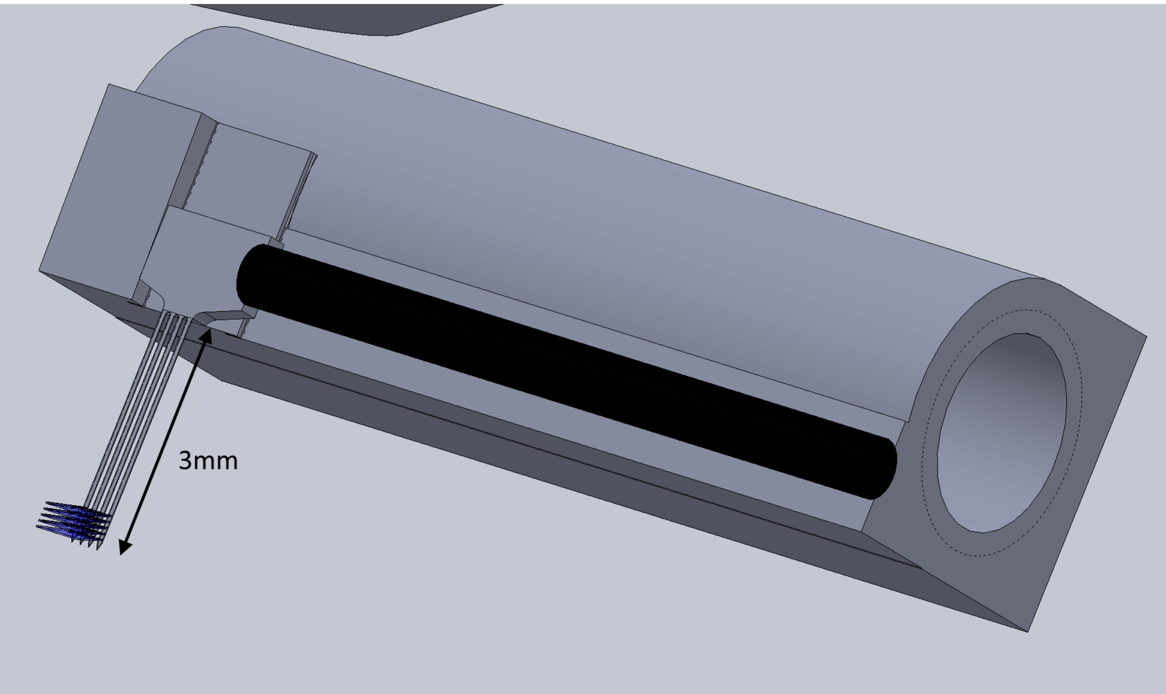

Figure 4.26: Probe holder modeled in SolidWorks including light sheet probe and fiber bundle (black).

To begin testing of the light sheet probes, a simple packaging setup was devised to secure the probe and fiber bundle to each other along the axis of a threaded rod. A 3D printed part was created to serve this purpose, serving as a structural element for both the probe and the fiber bundle. The model of the resulting part is shown in figure 4.26. All 3D printed parts were created using a Stratasys PolyJet 3D printer using Vero materials by Studio Fathom (Oakland, CA, USA). A flattened facet was created near the end of a cylindrical feature with dimension slightly smaller than the base of the probe. Along the right side of the probe is a wall which the probe can be squared to to ensure good alignment. The probe is glued to the facet using Loctite 5-minute epoxy (Henkel Corp., Scottsdale, AZ, USA). After securing the probe to the holder, the fiber bundle is brought into contact with the probe, and UV-curing optical epoxy (Norland NOA 146H, CRANBURY, NJ, USA) is applied to the facet and probe and is cured using UV light. After curing, black medical epoxy is applied to the length of the fiber bundle and the top of the light sheet probe to secure all optical parts and to ensure no stray light leaks out of the optical structures. A set screw is then applied to the hole in the cylinder, and a stainless steel rod is attached to the other end of the set screw. The

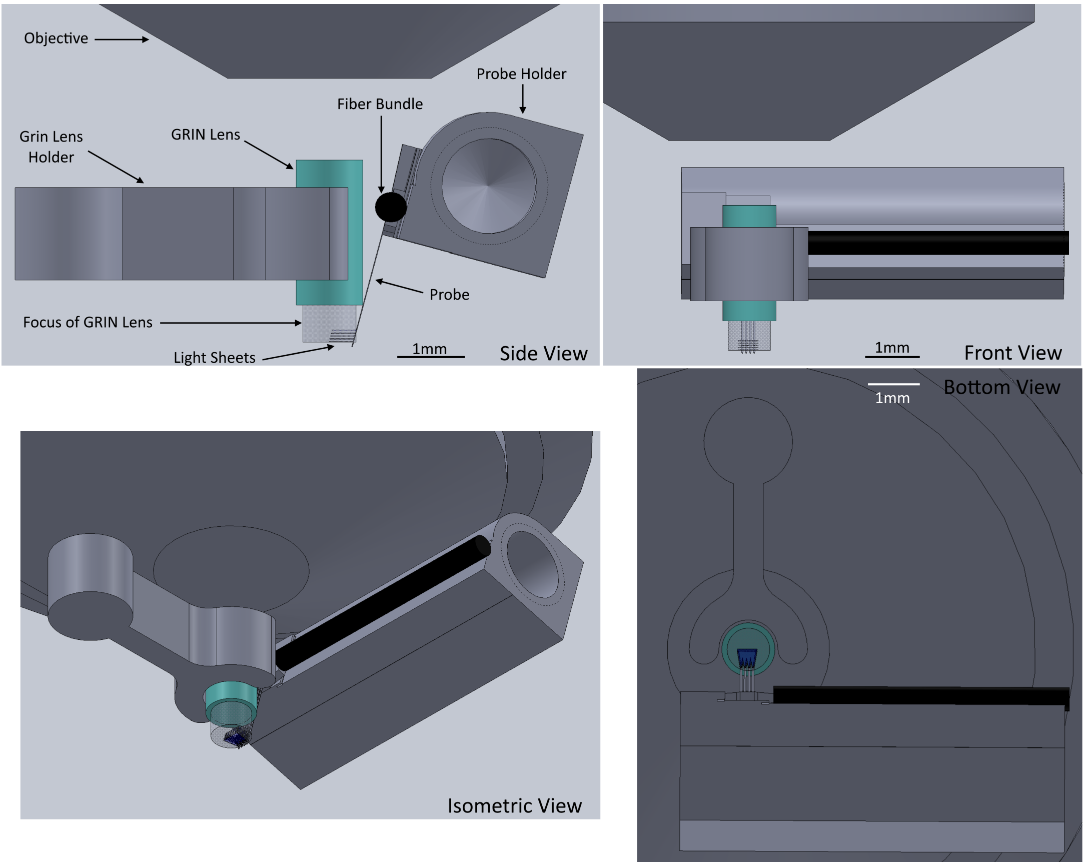

Figure 4.27: 3D model of complete packaging assembly including probe holder with probe and fiber bundle, microscope objective with accurate distance from GRIN lens, GRIN Lens holder with attached GRIN LENS.

stainless steel rod may then be attached to a micro manipulator for testing and implantation.

The complete optical arrangment for the testing setup is shown in figure 4.27, including the probe holder, GRIN lens in 3D printed GRIN lens holder and microscope objective. Using this model, it is possible to predict any possible mechanical interference designing the 3D printed parts. The GRIN lens holder is simply a C-shaped clamp allowing a grin lens to be glued into, and a cylindrical part which fits into a proprietary clamp attached to a micromanipulator. Two models of GRIN lenses are currently being tested: the Thorlabs G2P10 (1 mm, L = 3.4 mm, WD = 0.25 mm (Water), NA = 0.5) and the Edmund Optics 64-520 (1.8 mm, L = 3.93 mm, WD = 0.23 mm (Water), NA = 0.55). The G2P10 has the advantage of having an anti-reflective coating, but is smaller and more difficult to work with than the 64-520. Overall this setup allows for the flexibility to test the entire system while working out any mechanical constraints in the system for the next generation of optical system mounting for the light sheet probe.

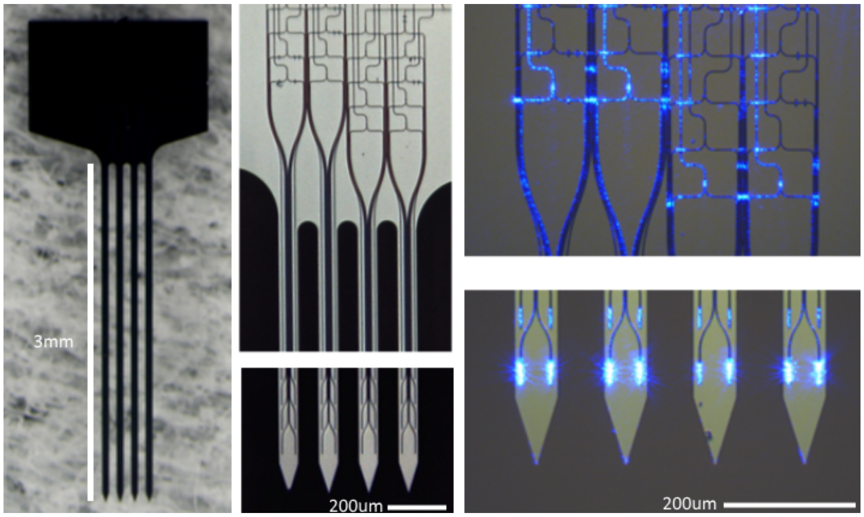

Figure 4.28: Released light sheet probes, including probe profile (left), micrograph of optical circuitry (center), and circuitry conducting blue light (right).

4.5.7 Results

Results from the first successful batch of laminar illumination probes can be seen in figure 4.28. On the left, a fully released probe is seen. Micrographs of the photonic circuitry are shown in the center and the right. When light is coupled into the photonic circuit, there is non-negligible scattering from the optical waveguides, an issue that is currently being resolved. As a result, some layers exhibit uneven illumination, as shown in figure 4.28 on the bottom right, as it is obvious that one grating is dimmer than the other gratings. Further development of the etching of the optical layer is required to minimize losses within the photonic circuitry, however the current set of probes is sufficient for initial testing of the properties of the devices.

4.5.8 Limitations and Future Work

This project is in its infancy and thus there is considerable future work that must be done. Primarily, the imaging resolution of the light sheet probe-optics combination must be thoroughly tested to ensure that the quality of images produced are sufficient for experimental needs. There is some risk involved in this project due to out-of-plane scattering of the light sheet which may broaden the light sheet to stimulate out of plane flurophores, degrading the quality of the image. Scattering is also a concern for light emitted from fluorophores as they exit the tissue and are collected by the GRIN lens. Thus, the maximum depth of implantation must be determined that can produce high quality images using the aforementioned optical system.

The other major concern is the power of light required to create the light sheet. Since the beam is focused in one direction, a significant amount of power will be continually pumped into the sheet.

This is concerning because there is some risk of damaging tissues, especially when utilizing the short wavelengths necessary to stimulate fluorescence. Studies of tissue death around the grating coupler emitters will need to be done to assess phototoxicity of the light emitted from the probe.

4.5.9 Excitation of GCaMP6 using ePixels

An additional drawback to two-photon microscopy is that, although superior to standard confocal microscopy, it is still limited in the penetration depth of the optical field. Two-photon microscopy may allow a researcher to image deep brain structures in small animals such as rodents; however, as this technology is translated to larger animals with more complex neural structures, there will be a limit of the penetration depth of the light and thereby the utility of this technique. One solution is to increase the number of photons which interact within the tissue, thereby further increasing the wavelength of the photons from the laser source and increasing the penetration depth. Three-photon microscopy has been demonstrated and is currently becoming more popular within the neuroscience community, however increasing to four-photon systems becomes increasingly more difficult and can't be repeated ad-infinitum given the laser power levels required. Thus, other techniques must be developed to improve depth resolution of microscopic tools for neuroscience research.

One method to improve the depth resolution of optical microscopy techniques is to bring the light to the deep neural structures instead of projecting it through long lengths of scattering tissue. This technique would use neural probes to deliver light directly to the brain structure of interest. A stimulation and detection scheme is currently being developed as a collaboration between the Roukes group and the Shepard group at Columbia University in NY, which is described below. This tehcnique utilizes implantable emitters and sensors to create a micro-scale imaging system within the tissue.

Conventional microscopy techniques utilize color filters such as dichroic mirrors or dielectric stacks to manipulate and separate light of different wavelengths within the optical path. However it is currently difficult if not impossible to create such films on micro- and nanoscale devices which are small, robust, and angular insensitive. Thus, other methods for identifying fluorescence must be used. One such mechanism is fluoresence lifetime imaging microscopy (FLIM), where the decay rate of light from a group of fluorophores is used to identify the type of fluorophore instead of the color of emission. The typical decay of fluorescence from a fluorophore is modeled as

$$I(t) = I_0 e^{-t/\tau}, \tag{4.14}$$

where I_0 is the initial intensity of the fluorescence after the cessation of stimulation, t is time, and τ is the lifetime of the fluorophore, which is specific for each fluorophore. Therefore, since the time scale of the fluorescence decay is specific to a given fluorophore, it is a means of generating contrast and identifying florophores from their inherent properties. Thus the identity of a florophore can be determined by directly measuring the rate of decay of fluorescence instead of measuring the color emitted by a fluorophore.

Measuring the decay of the fluorescence signal requires the use of highly sensitive optical detectors since the number of photons received by the detector from a single pulse of excitation (as opposed to the continuous pumping of an epifluorescence microscope) is extremely small. In microscopic applications, this is typically a photomultiplier tube in a photon-counting configuration. There is, however, a semiconductor alternative to PMTs which allow for photon counting in large arrays. These devices are avalanche photodiodes, which utilize the avalanche breakdown effect in a semiconductor substrate to convert the interaction of photons with semiconductor to induce a large current within the substrate, which is then detected as a spike in current through the substrate. When in single photon counting mode, these devices are typically called single photon avalanche photodiodes (SPADs). Given the sensitivity of these devices, they can be used to detect and count single photon counts from biological samples.

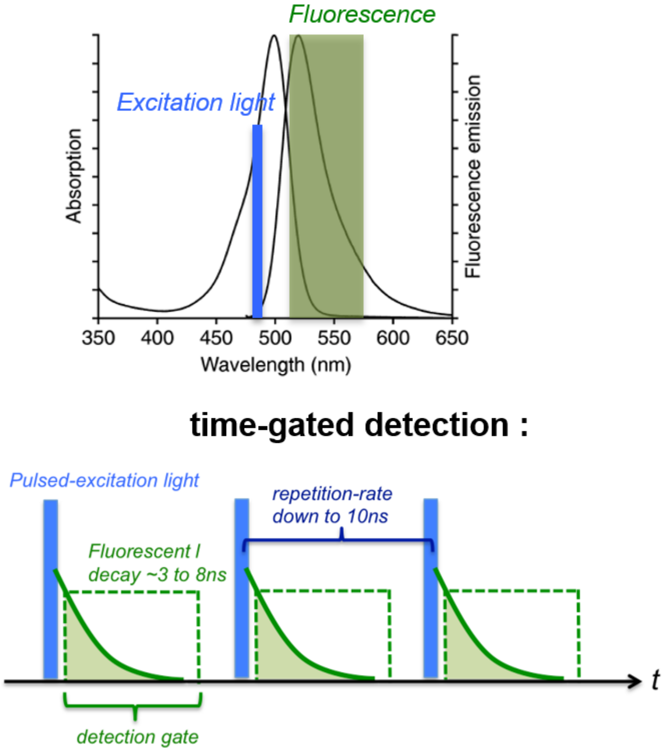

Figure 4.29: Stimulation pulse and SPAD integration timing diagram. Blue pulses of stimulation at a frequency of 80MHz begin each cycle which energizes the fluorescence intensity of the fluorophores (shown in green). Each pulse is then integrated for a given amount of time to reconstruct the exponential decay of the fluorescence signal.

A typical measurement of the fluorescence decay of a fluorophore is accomplished through photon counting with a specific delay after the initial pulse of stimulation of the florophore. Since this is, in essence, integration of the signal, the integrated decay will also be an exponential function with a constant scaling factor. Thus, the rate of decay of the integrated signal is identical to the decay of the actual fluorescence. The detection scheme is shown in figure 4.29. Each pulse and fluorescence response is identical, however the gating of the integration (shown in green above the light intensity curve) is progressively shortened. By plotting the delay versus the integrated fluorescence intensity, it is possible to extract the decay constant from the data, and thereby extract the identity of the fluorophore being measured. The intensity of the fluorescence is simply proportional to the integrated value of the fluorescence with the shortest decay.

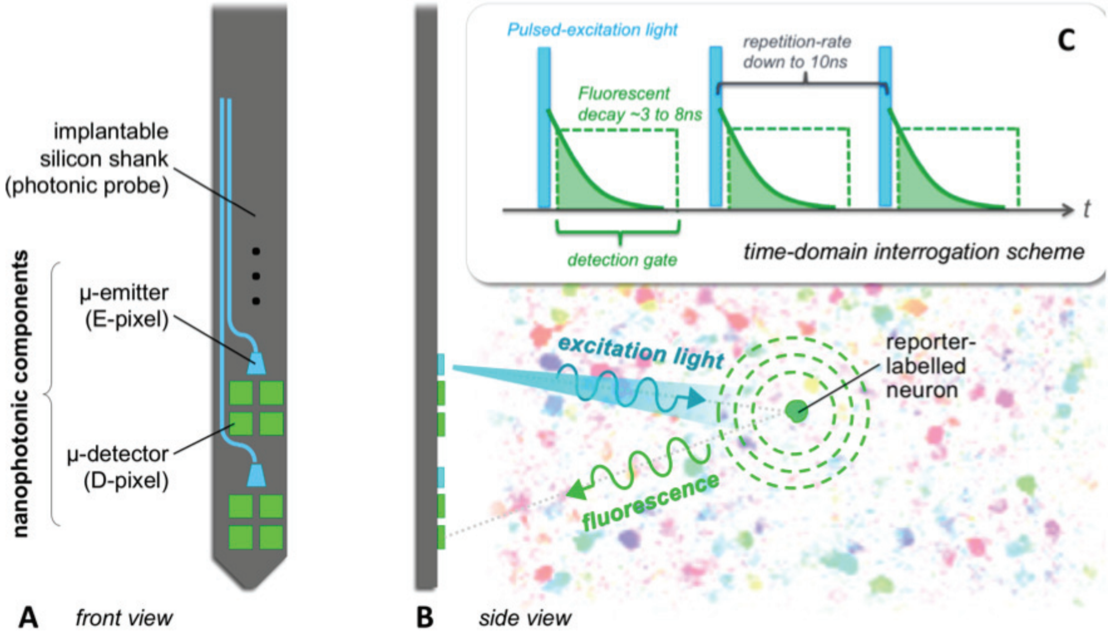

Figure 4.30: Stimulation and detection of neurons using photonic circuit based emitter pixels (e-pixels) and SPAD based detector pixels. Picosecond pulses of light are emitted from e-pixels, creating an angular excitation field where the location in brain is known by the emission field of a single e-pixel. The pulse of light stimulates fluorophores in the neurons while quickly extinguishing itself, and the fluorescence of cell bodies is collected by the d-pixels elsewhere on the same probe.

A simplified detection scheme is shown in figure 4.30. In this detection scheme, a probe containing both SPAD detectors and optical emitters (like those described previously in this chapter) would be implanted into neural tissue. The grating emitters will be used to locally and directionally stimulate neurons in the space between the probes. Elsewhere on the probe, CMOS electronics fabricated at a foundry with an array of SPADs to be used as detectors. These detector pixels will be used to collect the light from a stimulated neuron and be integrated using the scheme from the paragraph above to determine the source and intensity of the fluorescence signal. Using these devices together, it is possible to measure the intensity of GCaMP6 neurons and determine the activity of the neurons at any given point in time.

4.5.9.1 Picosecond pulses through photonic circuits

Understanding the properties of picosecond and shorter pulses through photonic circuits is critical for timing the gating of the SPADs for fluorescence lifetime imaging in the configuration described above. Furthermore, it is unknown what the properties of light emitted from a grating coupler will look like using a pulsed source. Beam profile broadening is a significant risk of using a broadband pulse instead of a narrow spectrum continuous wave laser to excite the grating coupler. These parameters are critical for executing the experiment described above.

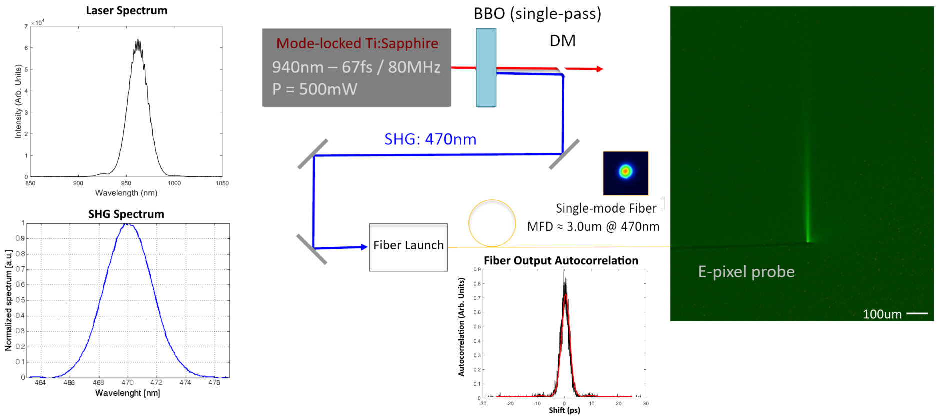

Figure 4.31: Experimental setup for producing picosecond pulses in emitter pixels utilizing Chameleon Vision S Ti:Sapphire laser, BBO frequency doubler, and associated optics. Spectra from the laser and second harmonic generator shown, as well as the autocorrelation signal from the fiber which is then coupled into the photonic probe.

The pulsed light was created using a Coherent Chameleon Vision S Ti:Sapphire laser. This laser utilizes a 18W Verdi pump laser with wavelength 532nm to excite the Ti:Sapphire crystal within the Chameleon laser system. The Chameleon laser is tuneable between 660nm-1060nm, with a peak power of approximately 3.5 W at 800nm. Pulse widths typically range from 60-100fs depending on the frequency used, and have a pulse repetition frequency of 80MHz. The spectrum for this laser is shown in figure 4.31 at the typical operating wavelength of 960nm. The bandwidth of this pulse is approximately 30nm, and results in pulses with duration of approximately 100fs.

Light from the Ti:Sapphire laser is then coupled into a APE HarmoniXX second harmonic generator (SHG). Contained within the device is a nonlinear medium (typically a lithium niobate or lithium triborate crystal) which, when high intensity light propagates through this medium, the photons generate a nonlinear polarization of the crystal at twice the frequency of the incoming light. This polarization relaxes, causing the emission of light at twice the frequency (half the wavelength) of the incident light. This efficiency of this conversion is proportional to the square of the intensity of the incoming light, and therefore these devices are most efficient with high intensity light like that produced by a pulsed laser. Peak efficiencies for SHG are typically 40% for pulsed sources. The resulting spectrum coming from the second harmonic generator is shown in figure 4.31. The resulting bandwidth of the optical signal has been significantly reduced to approximately 4nm. This is due to the nonlinear conversion of the energy in the SHG crystal lengthening the pulse.

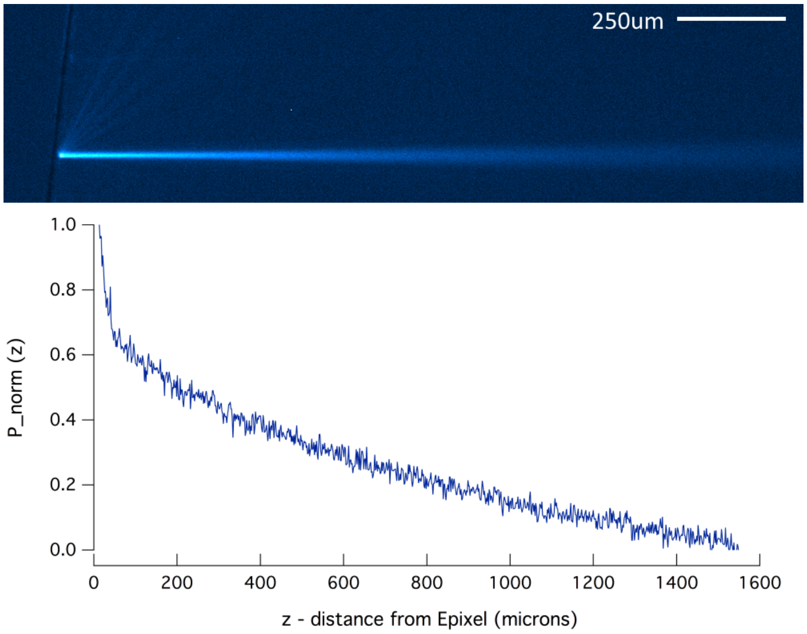

Figure 4.32: Beam profile of e-pixel using pulsed SHG generated signal in fluorescein solution (top), and beam profile intensity versus distance from the shank (bottom).

Frequency doubled light is then coupled from free space into an optical fiber using a standard fiber optic launch system, where an objective lens is used to focus the light in free space down to a small point, which is coincident with a 460HP optical fiber from Thorlabs. This fiber has a core diameter of 3.5μm and an NA of 0.13, allowing only for single-mode propagation at 480nm. Light was coupled into a 2 meter length of this fiber, which was then sent into an autocorrelator to determine the pulse width of the frequency doubled pulse. The resulting autocorrelation signal is shown in figure 4.31, showing a 3.5ps pulse width. This shows significant peak broadening from non-linear dispersion in the fiber.

The resulting beam profile is shown in figure 4.32. Due to the relatively broad bandwidth of the frequency doubled light from the BBO crystal and the non-linear propagation through the optical fiber (as well as the on-probe waveguide), it was unknown what the emission profile from the optical probe would be. As shown in the figure, the beam has a well collimated profile, similar to the profile emitted using a monochromatic source from similar waveguide structures (although slightly wider in cross-section). This is encouraging, as it will allow for targeted illumination of neurons expressing GCaMP6 for imaging purposes. Unfortunately it is currently impossible to measure the temporal pulse width of the signal from the probe, as the power emitted from the probe is too weak to be measured by the autocorrelator.

Chapter 5

Microfluidic Neural Probes

In this chapter, the development of silicon microfluidics and the integration of this technology within neural probes is described. Microfluidics on probes provides a point of access for chemical information to be sent and received by the probe to the brain. Microfluidics can be used to sample the chemical composition of the extracellular space, or microfluidics can be used to send chemicals (e.g., pharmacological agents such as muscimol) to the brain to effect behavior[8]. Two major projects were developed around the combination of microfluidics with the optical techniques described in the previous chapter. The first project centers around the use of caged neurotransmitters, where the compounds are continually injected into neural tissue while optical emitters are used to induce photolysis of the caged compounds, thereby releasing the neurotransmitter into the neural tissue in its active form. The second project details the use of microfluidics to obtain protein samples of the extracellular milieu in deep brain structures, and conduct it to the base of a neural probe for *in situ* analysis by ring resonator biosensors (as described in Chapter 2). These two projects highlight the utility of embedded microfluidics in neural probes and combine microfluidics with other technologies to accomplish novel experiments in deep brain regions.

5.1 Motivation

5.1.1 Uncaging of Neurotransmitters

In the previous chapter, the importance of neural stimulation in the elucidation of neural circuit behavior and functionality was discussed in depth. However, in spite of the powerful optogenetic techniques currently in use, there is a drawback to this technique[49]. Namely, although optogenetic techniques are able to stimulate neurons, it does this in a markedly non-biomimetic fashion. There is a vast amount of information lost by circumventing synaptic (in)activation and simply just stimulating the neurons themselves. The dendridic tree is understood to be an area of significant calculation by the neuron, with the constant push-and-pull of excitatory and inhibitory stimulus being summed together to determine the spiking characteristics of the neuron itself. Therefore it is very desirable to have a mechanism for direct stimulation of synapses to manipulate neural circuits in situ.

Caged neurotransmitters have emerged as an important experimental tool for synaptic stimulation. A caged compound is typically a chemical that has been passivated by a labile chemical group which can be subsequently photo-cleaved to activate the functional component. In the case of caged neurotransmitters, this typically means that a neurotransmitter has been chemically modified such that it's active site is bound to a photolabile group, which when exposed to light will leave and expose the active site of the neurotransmitter to the environment. Since enzymes that might break down or bind to the neurotransmitter are sterically specific, the caged version of the compound is much more stable in biological environments as compared to the uncaged form. Since there are numerous housekeeping enzymes that break down free neurotransmitters in the extracellular matrix of brain tissue, the use of caged compounds for infusion into tissue is critical. It is impossible to simply inject neurotransmitters into neural tissue and expect a quantitative, sustained response from the neurons[4].

A variety of caged neurotransmitters were first developed in 1990 by Wilcox et al. at Cornell University[128]. The first experimental use of caged glutamate, the primary excitatory neurotransmitter in the central nervous system, was demonstrated in brain slices by Callaway and Katz in 1993[13]. These researchers used an ultraviolet argon laser source to uncage glutamate near specific neurons in slices of ferret visual cortex while measuring the response using whole cell recording. With this approach, they were able to successfully stimulate neurons using this technique more than 30 times without seeing any fall-off in responsivity of the neuron.

The development of this technique has proven to be extremely important in the study of synaptic and dendridic computation, and has spawned a large body for scientific research using these compounds [16, 5, 35]. However, it is important to note that the vast majority of these efforts rely on tissue slices in their studies. It turns out, the main drawback of this technology stems from the wavelength of light necessary to release the photolabile group attached to the neurotransmitter. Ultraviolet light has a number of undesirable properties when interacting with neural tissue that makes it non-ideal for experimentation. Primarily, short wavelengths of light are scattered strongly, thereby limiting the delivery of UV light to structures at superficial depths. Furthermore, photons of UV light have large energy densities, which can easily induce phototoxicity, especially with an intense, focused beam in tissue. Finally, single-photon excitation permits uncaging to occur all along the entire beam profile, which is not ideal. One remedy for this was to attempt two-photon uncaging, where a longer wavelength could be used for uncaging[25], reducing scattering loss and improving localization through the two-photon spatiotemporal focusing. New caged molecules were developed with high two-photon absorption cross-sections[39], allowing for efficient and safe uncaging experiments deeper in the brain. This technique was subsequently expanded to the production of photoreleasable ions, which can also be used to modulate neural activity[143]. Improvements to the chemistry of photolabile groups have been continuously evolving since 1990. Recent advances include the development of the very popular 4-methoxy-7-nitroindolinyl caged compounds[15]. The latest, commercially-available caged neurotransmitters are based on ruthenium-bipyridine chemistries, as described in [29]. These compounds have superior properties with respect to uncaging speed, have superior two-photon absorption, and allow for uncaging at visible wavelengths (around 470nm).

Although two-photon excitation has been revolutionary in allowing for localized dendridic stimulation[81], the ability to carry out uncaging experiments in living animals has been limited to superficial experimentation with maximum depths of 200μm[22, 78]. To broaden the applicability of these techniques to enable scientists to utilize not only optogenetic techniques but neurotransmitter based techniques, we conceived of an optically-enabled microfluidic probe, as described here. Our premise is that these probes offer the benefit of localized neurtransmitter based stimulation directly in neural tissue, allowing access to deep brain regions which have . This probe incorporates the optogenetic e-pixel technology, described in the previous chapter, with microfluidic channels to allow for bolus and/or perfusion based supply of caged neurotransmitters localized to the region surrounding the emitter pixels. This can enable direct stimulation of synapses in deep brain regions.

5.1.2 Biosensing in Neural Tissue

To develop next generation brain-machine interfaces, we must have better understanding of the brain's immune response to chronic neural implants. To enable this research, we will develop an innovative biophotonic, nanodialysis probe for real-time monitoring of the brain's evolving adaptation to implanted devices.

Currently, the primary mechanism for studying the brain's immune response is through immunohistochemistry on brain slices from sacrificed animals[99, 115, 131]. Alternative approaches have been developed in order to perform real-time *in vivo* measurements. However, the majority of these approaches do not provide a direct measure of extracellular protein concentration, e.g. optical imaging methods of mapping receptor sites[102]. Positron emission tomography (PET) has been use to attempt *in vivo* imaging of microglia activation[36], but has not proven successful to date. For example Bartels et al. conclude: *"In current practice, [11C]-PK11195 seems an unsuitable tracer for accurate or reliable quantification of neuroinflammation."*[7] This has led to increased interest in microdialysis sampling for a reliable *in vivo* technique for quantification of inflammatory molecules in the brain. A key challenge to microdialysis sampling for chemokines and cyotokines is the low concentrations. Basal concentrations of picograms/mL are typical, rising to nanograms/mL with stimulation[45, 123]. The most successful use of microdialysis for detection of immune response molecules has been in humans[41, 40, 42, 69], where the ability to use relatively large microdialysis membranes (typically 10mm in length) enables the volume of fluid required for analysis to be collected at relatively low flow rates (typically 0.3μL per min). There is significant interest in performing real time, in vivo detection of cytokines and chemokines in the brains of rodents and other small animals[45, 123]. However, the smaller brain size presents a unique challenge when it comes to collecting a sufficient volume of fluid for analysis. Specifically, greater flow rates must be used to collect comparable volumes of fluid on the same time scale. As a result of these high flow rates, the molecular concentration in the perfusate does not have time to equilibrate with concentrations in the brain, reducing the detection efficiency. Multiple researchers within the neuroscience community have recently highlighted these unique challenges, e.g. Vasicek *et al.* (2013): *"While cytokine collections using microdialysis catheters have been performed in humans and multiplexed measurements have been made, there are challenging differences between the current state of the art with respect to human cytokine measurements vs those for use in basic research"*[123]. Herbaugh and Stenken (2011) make the same point: *"Microdialysis has been used to collect chemokines and cytokines in the brain but these studies have been predominantly in humans since the length of the membranes is typically 10mm vs. the 1-4mm range used for rodents."*[45] These studies point to a compelling need for an approach capable of achieving a comparable limit of detection (LOD) on smaller volumes. This would enable reduced rates of perfusion, increasing the concentration of target molecules in the perfusate.

Nanophotonic micro-ring resonators have a proven track record for the detection of cytokines and chemokines with a limit of detection comparable to that which can be achieved on a more traditional immunoassay. Sensitivity has been demonstrated through such applications as the detection of the cytokine interleukin-2 (IL-2) secreted by Jurkat cells at concentrations down to 100pg/mL[64]. A key motivation for the development of this technology has been its natural integration within microfluidics to enable high sensitivity detection from small volumes of fluid.

5.2 Development of Silicon Microfluidics

5.2.1 Development of Fabrication protocol

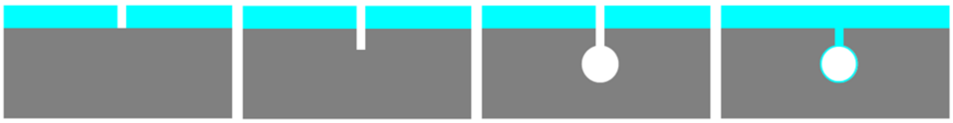

Figure 5.1: Dig and seal process flow, from left to right. First, a trench is produced in the top oxide layer, followed by continuing that trench into the underlying silicon. The sidewalls of the trench are passivated, and the bottom of the trench is removed. An isotropic etch is then applied to the trench to produce the body of the microfluidic channel. Finally, an isotropic deposition process is used to fill the neck of the channel while leaving the body of the channel free for fluid flow.

An overview of the “dig-and-seal” process we have employed is shown in figure 5.1. The process begins with a 250nm film of SiO_2 atop the silicon substrate. The SiO_2 layer is used as an etch mask when defining the microfluidic channel later in the process. The first step of the process utilizes electron beam lithography to pattern an SML-2000 film coated on the surface of the wafer. SML-2000 was chosen due to it’s thickness and etch resistance, which allows for a long etch process required to penetrate the SiO_2 film. SML-2000 was spun at 1500RPM and baked for 2 minutes at 180^oC to prepare the film. A line is defined in the SML-2000 film using the Vistec EBPG-5000+ elecron beam lithography tool, 500nm wide, which will be used to etch a trench in the silicon substrate. Due to the width of the line, a high beam current (typically 100nA) and a low resolution (50nm between shots) are used in the EBPG-5000+ to speed up the processing of the wafer. Due to the thickness of the resist, a high dose is required (3500μC per square cm).

5.2.1.1 Trenches via Bosch process

The first attempts to define the trenches for the dig-and-seal process utilized the Bosch process, as it produces well defined deep trenches easily in silicon. In the first attempts, the SiO_2 layer was absent, as it was theorized that it would be possible to utilize the polymer layer deposited as part of the Bosch process to protect the sidewalls and top surface while etching a microfluidic channel at the bottom of the trench.

In this procedure, after the pattern was defined in the SML-2000 resist, the Bosch recipe shown in section 3.4.1 was used to etch a trench. The ideal depth for the trench was unknown at the outset, so for testing purposes the trench was etched between 4-8 cycles. After etching, the resulting trench was exposed to the deposition cycle for an additional amount of time, typically between 30 seconds and 2 minutes. Normally, this should be sufficient to protect the sidewalls and the bottom of the trench equally, but the etch rate at the bottom of the trench should be much higher than the sidewalls due to a lack of ion-assisted etching on the sidewalls of the trench. This is because the sidewalls should be shielded from the ions structurally, while the bottom of the trench should not be shielded at all. Early results, however, showed that there was undesired side-wall etching, which, in turn, resulted in a shallow microfluidic channel. This is shown in figure 5.2. Shallow microfluidic channels are not desirable as the channels must be sturdy and unlikely to break since they will be eventually integrated onto neural probe shanks, which can bend during insertion. Thus, the plan to use deposited polymer based on C_4F_8 as side wall protection was abandoned.

To remedy this issue, thermally grown SiO_2 was chosen as an alternative for side wall protection. SiO_2 is an ideal candidate because instead of using a deposition process, it is possible to directly oxidize the sidewalls of the trench. This guarantees that the sidewalls will be coated equally with the passivation layer. After etching the trench, the silicon was thoroughly cleaned using Nanostrip to remove any organic residue from the wafer. The wafer was left in the Nanostrip solution overnight.

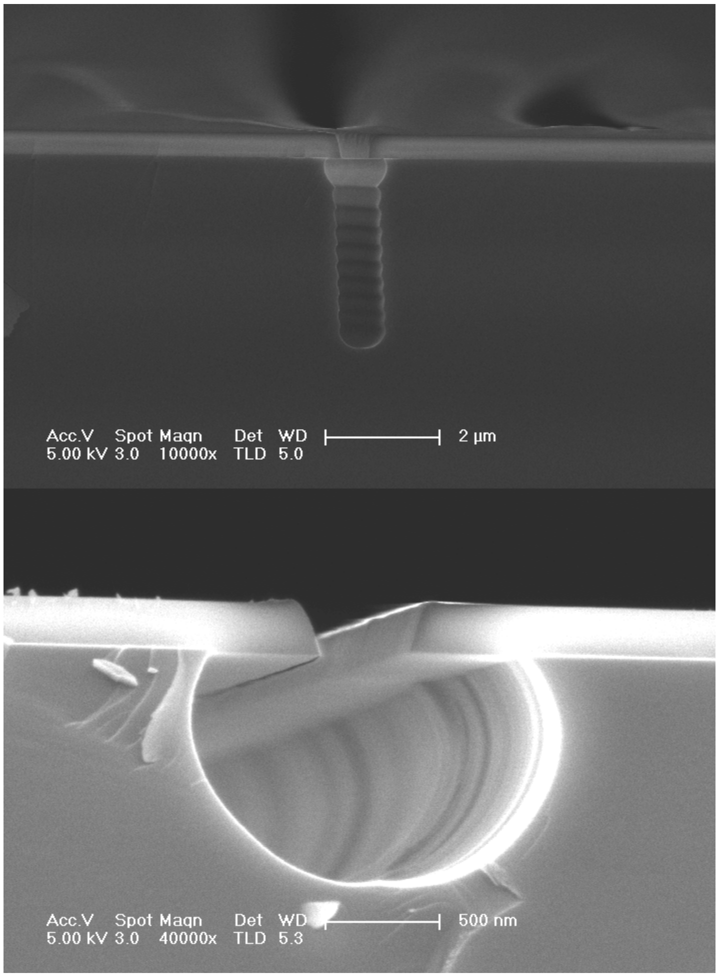

Figure 5.2: Top, trench etched using Bosch process. Bottom, anisotropic etch of trench after sidewall protection via Bosch polymer deposition. Unfortunately the trench has been etched away as well, leaving only a thin cover over the channel.

Nanostrip leaves a thin layer of dirty oxide on the surface of the wafer, so following the Nanostrip soak the wafer is dipped in buffered oxide etch for 10 seconds. Finally the wafer is cleaned in an oxygen plasma briefly to ensure that the surface of the wafer is clear of organic residue.

Next, the wafer is placed in a Tystar Tytan furnace for oxidation. The wafer is oxidized using dry O_2 gas at 1000^oC for 45 minutes, followed by a 30 minute annealing step in N_2 gas at 1000^oC. The resulting oxide film is shown in figure 5.3. After oxidation, the silicon is placed back in the etcher, and the oxide is etched using the recipe listed in section 3.4.4. After etching for 25s, the oxide in along the bottom of the trench was removed while maintaining the sidewall oxide. Finally, an isotropic silicon etch, described above in section 3.4.3 was used to create the body of the microfluidic channel. The result is shown in figure 5.4. The resulting channels are typically 2.5-4 microns wide.

Following the completion of the microfluidic channel etch process, the final step in the process is to deposit material conformally such that the neck of the channel is filled, while leaving the body

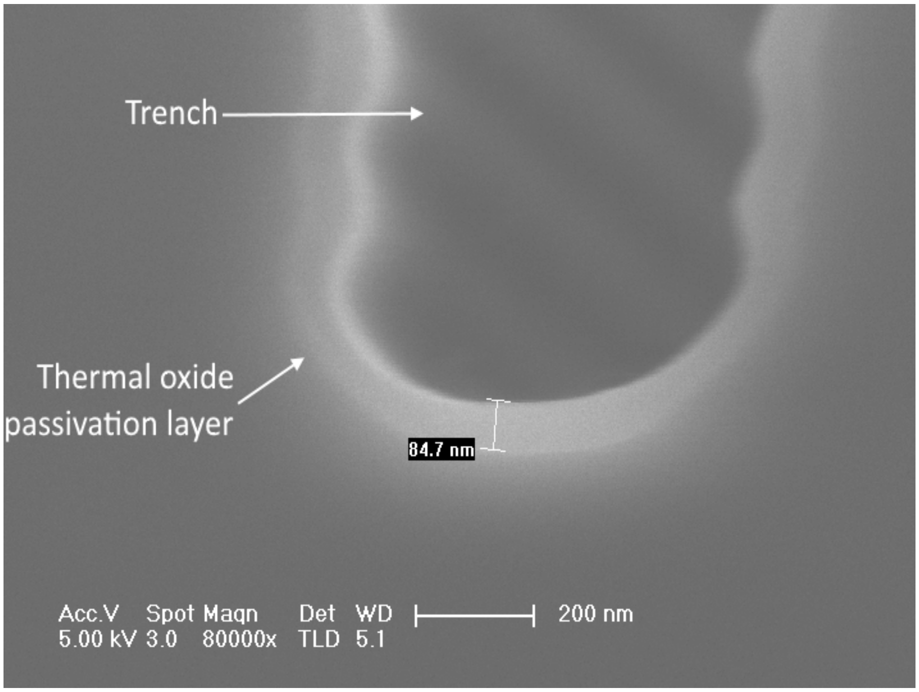

Figure 5.3: Trench after dry thermal oxidation for 45 minutes. This produces a stronger passivation layer for the trench sidewalls.

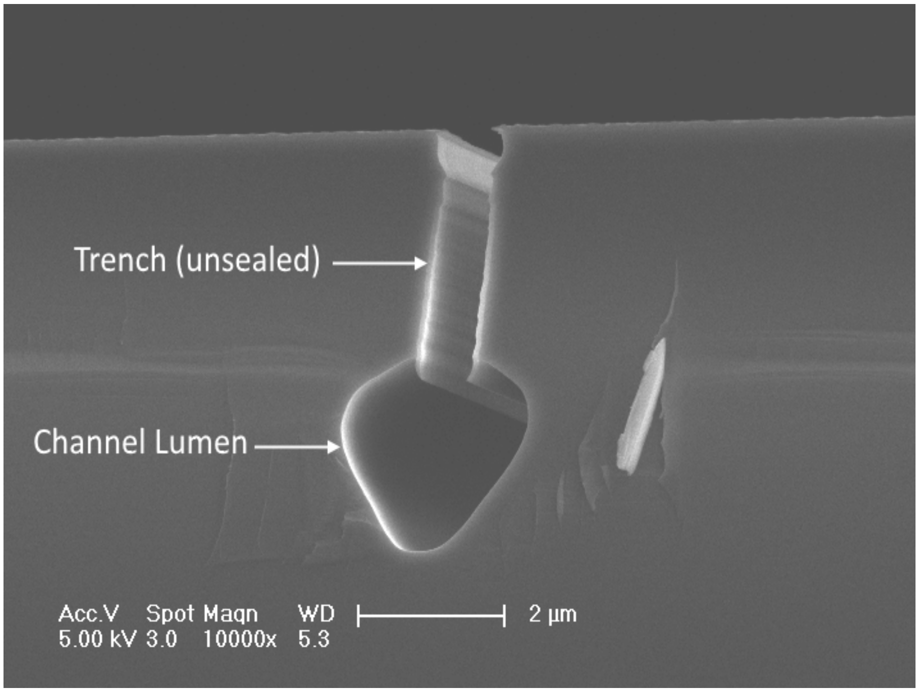

Figure 5.4: Completed microfluidic channel etch using thermal oxide sidewall passivation. This leaves the trench in tact, allowing for a strong structural layer above the microfluidic channel.

of the channel open to flow. Two separate techniques were attempted for the filling of the channel. The first test was using PECVD oxide using the Oxford Systems 100 Plasma Enhanced CVD chamber available in the KNI. This deposits a Silane-based SiO_2 film, which is unfortunately not very conformal. The resulting deposition left the entire neck of the microfluidic channel completely empty, only sealing the channel with a very thin layer of oxide at the top. This is shown in figure 5.5. The second choice was to deposit parylene to seal the channel, which has a much lower sticking coefficient and thus should provide superior coating conformality. Before coating, an adhesion promoter of γ-methacryloxypropyltrimethoxy silane was applied to the surface in a bath after dilution to 0.5% in a 50:50 solution of isopropyl alcohol and water. The silane was left overnight to bond to the wafer surface, after which the wafer was removed and allowed to air dry for 30 minutes. The wafer was then rinsed with isopropyl alcohol for 5 minutes, dried using compressed nitrogen gas, and baked at 115^{o}C for 30 minutes. Next, parylene is applied to the surface using a Paratech parylene coating system via sublimation and pyrolysis of the parylene-C precursor. The film was deposited to achieve a 1.5μm thickness. The resulting film was superior with respect to conformality, but due to the shape of the neck of the microfluidic channel it suffered from a similar problem to the SiO_2 deposition, where there was still incomplete filling of the neck and a thin seal at the top of the trench, leaving a weak seal of the channel, which is undesirable, as seen in figure 5.5 on the left. Therefore modifications to the fabrication procedure were necessary to ensure proper filling of the neck of the channel.

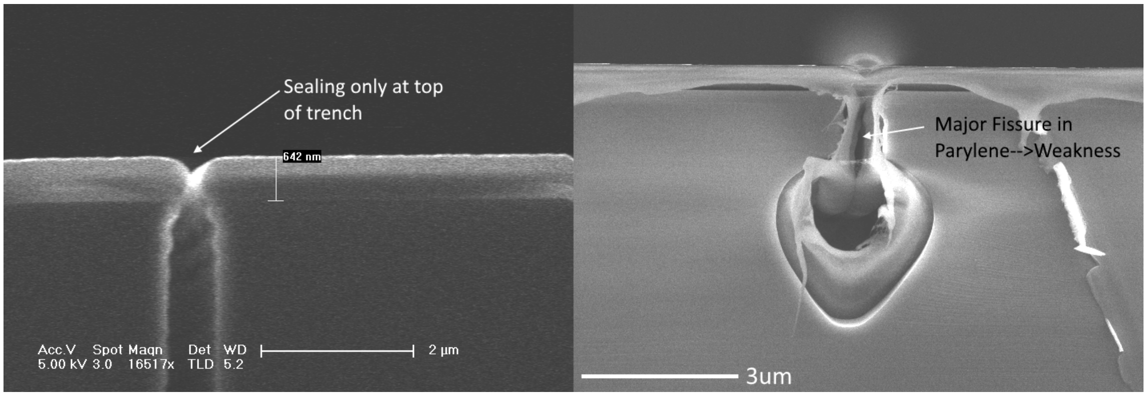
Sealing only at top of trench
642 nm
Acc.V Spot Magn Det WD
5.00 kV 3.0 16517x TLD 5.2
2 µm
Major Fissure in Parylene-->Weakness
3um

Figure 5.5: Sealing of microfluidic channels by PECVD (left) and by Parylene deposition (right). Both processes show incomplete filling of the neck of the channel. This is due to the geometry of the trench having straight sidewalls and scallops.

5.2.1.2 Trenches via pseudo-Bosch process

To solve the trench filling problems described above, two changes to the process were made. First, the etch used to define the trench was changed from Bosch to pseudo-Bosch. The key advantage is that this will allow the sidewall angle of the trench etch to be slanted, which was expected to improve trench filling due to the discussion in section 3.6.2.2. The second improvement was to switch from a silane-based PECVD oxide to a TEOS-Based PECVD oxide. Oxides formed from a TEOS precursor have a much lower sticking coefficient and precursors have a higher diffusion coefficient on silicon surfaces, and thus the film should be more conformal.

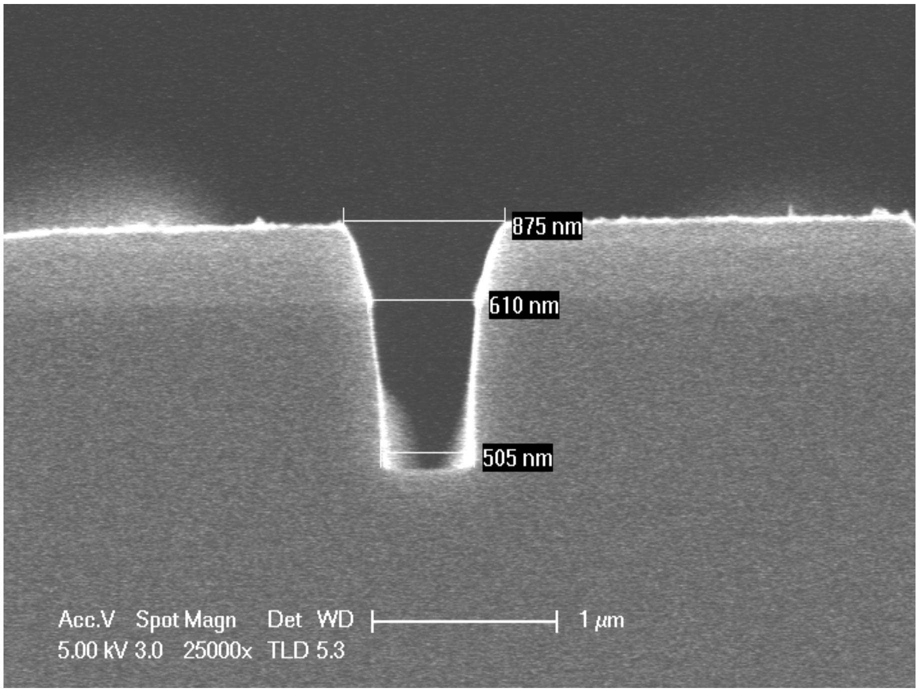

Figure 5.6: Result of slanted sidewall pseudo-Bosch etch. This should remedy the filling problem by inducing trench filling from the bottom of the trench upwards.

Thus, the first step in updating the above process was to develop a pseudo-Bosch recipe which produces slanted sidewalls. Two major parameters can be used to alter the sidewall profile, gas composition (increase C_4F_8 flow rate or decrease SF_6 flowrate) and forward power (counter-intuitively, increasing forward power will increase sidewall protection). Both of these options were explored, however it was found that altering the C_4F_8 flow rate produced the best results. Comparing the recipes described previously, the only difference in the recipe for straight and slanted sidewalls is that the flowrate was increased from 55 to 70 sccm to increase the sidewall angle. The result is shown in figure 5.6. An added benefit of using pseudo-Bosch is that the same etch can be used to etch through the top oxide as well as the silicon.

Although these two changes should remedy the trench filling process, another issue arose with the modifications to the process. After completing the process (described in detail below), a consistent problem with the etch process was observed, as shown in figure 5.7 on the left. At the interface between the top layer of SiO_2 and the silicon below, there are additional voids that form during the final isotropic silicon etch at the end of the process. Upon further investigation, it was found that the SiO_2 layer formed in the oxidation process had two unexpected properties. First, the oxidized layer protrudes out from the boundary defined by the extent of the top SiO_2, as is shown in figure

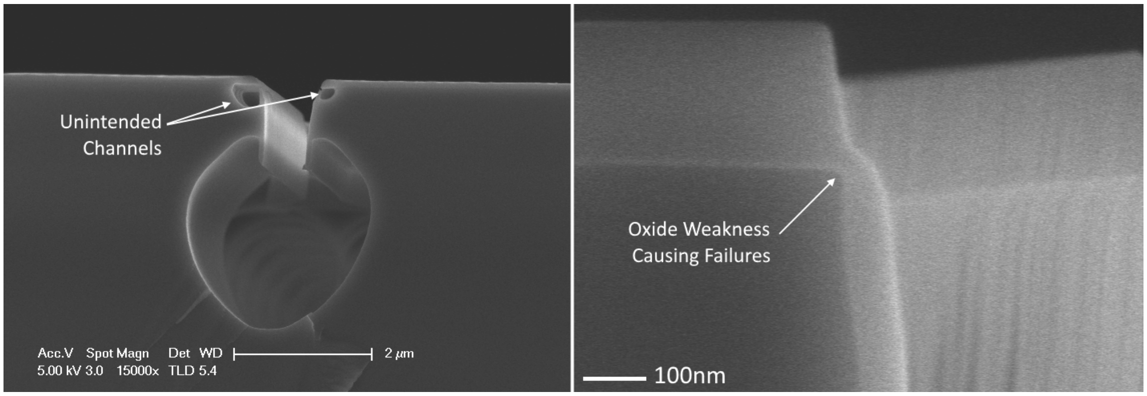

Figure 5.7: Left, Result of pseudo-Bosch based microfluidic fabrication with failure at the SiO_2-Silicon interface. These additional voids reduce the structural integrity of the channel and allow for another, undesired path for fluid to flow in the channel. Right, SEM showing the thin oxide film at the interface between the top oxide and the bulk silicon.

5.7 on the right. This overhang allows for ions to bombard the overhanging oxide, increasing the etch rate at that point by ion-assisted etching. Second, the thickness of the oxide film at the corner interface of the top SiO_2 and the underlying silicon is thinner than the rest of the film produced by oxidation. This is described by the Deal-Grove model discussed in appendix A; the diffusive flux is dependent on the thickness of the SiO_2 layer oxygen is diffusing through.

To solve this problem, an additional step was added to the etching procedure. Instead of maintaining the same pseudo-Bosch etch straight through the top oxide and into the underlying silicon, an intermediate step was added to the etching protocol. After penetrating the top SiO_2 layer using the slanted sidewall pseudo-Bosch recipe, an isotropic silicon etch was used to induce an indent in the sidewall profile. The silicon dioxide is nearly immune to the isotropic etch, so only the underlying silicon would be etched isotropically. This creates an indent in the side wall profile of the silicon, similar to the scallops seen in the Bosch etch described in the previous section. After the indent step, the tapered sidewall pseudo-Bosch was continued to finish the etching of the trench. The resulting structure is shown in figure 5.8. This structure resolves the issue of the protruding corner of the oxidized surface being etched quickly by ion-assisted etching by protecting it under the overhang created by the top SiO_2. The final etch protocol follows.

Similar to the pseudo-Bosch procedure above, the modified process begins with the definition of a 500nm wide line in SML-2000 resist spun at 1500 RPM for 45 seconds and baked for 2 minutes on a hotplate at 180^oC. The spatial resolution of the EBPG was set to 50nm, and the dose was $3500\mu C$ per square cm. After developing the SML-2000 resist in a solution of 3:1 Isopropyl Alcohol:Methyl Isobutyl ketone for 1 minute, the wafer is placed in a Oxford Instruments 380 ICP plasma etcher. Wafers were etched for 11 minutes using the tapered pseudo-Bosch etch to penetrate the 250nm SiO_2 layer on top of the silicon device layer. Next, the etch recipe is switched to the isotropic silicon

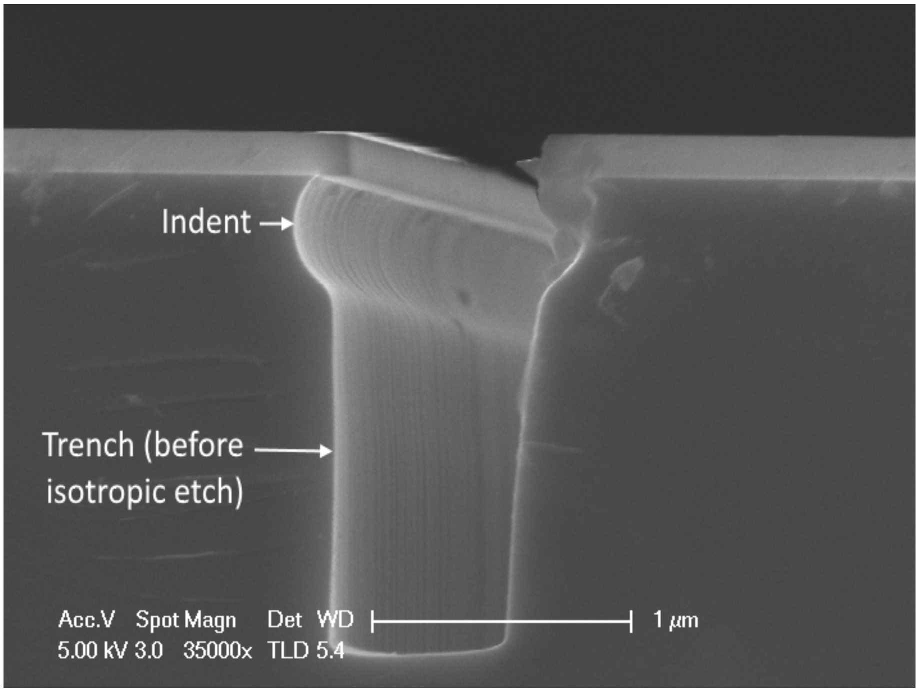

Figure 5.8: Microfluidic trench etched with indent procedure to prevent etching of additional voids at the SiO_2-silicon interface.

etch for 32 seconds to create the indent in the silicon sidewall. Finally, the recipe is switched back to the tapered pseudo-Bosch etch for an additional 6 minutes to complete etching the trench structure into the silicon layer. The resulting trench is approximately 2 microns deep into the silicon layer.

Next, the wafer is placed in a Tystar Tytan furnace for oxidation. The wafer is oxidized using O_2 gas at 1000^oC for 45 minutes, followed by a 30 minute annealing step in N_2 gas at 1000^oC. After oxidation, the wafer is placed back in the etcher on a silica carrier wafer, and the oxide is etched using the tapered sidewall pseudo-Bosch etch. Using this etch also helps reduce the probability of producing voids at the SiO_2-silicon interface. After etching for 1 minute and 45 seconds, the oxide along the bottom of the trench was removed while maintaining the sidewall oxide. Finally, an isotropic silicon etch, described above in section 3.4.3, was used to create the body of the microfluidic channel, etching for 1 minute and 30 seconds. The result is shown in figure 5.9. The resulting channel is approximately 4 microns wide.

Following the completion of the microfluidic channel etch process, the final step in the process is to deposit material conformally such that the neck of the channel is filled, while leaving the body of the channel open. TEOS based oxide was attempted using an off-site facility. TEOS-PECVD oxide was deposited by Noel Technologies (Campbell, CA) to a thickness of 950nm. The result was very successful, as shown in figure 5.10 on the left. Unfortunately, after testing the process and verifying the correct thickness for trench filling, the supplier discontinued the TEOS-PECVD deposition process. Thus, as an alternative, parylene was used to seal the microfluidic channel.

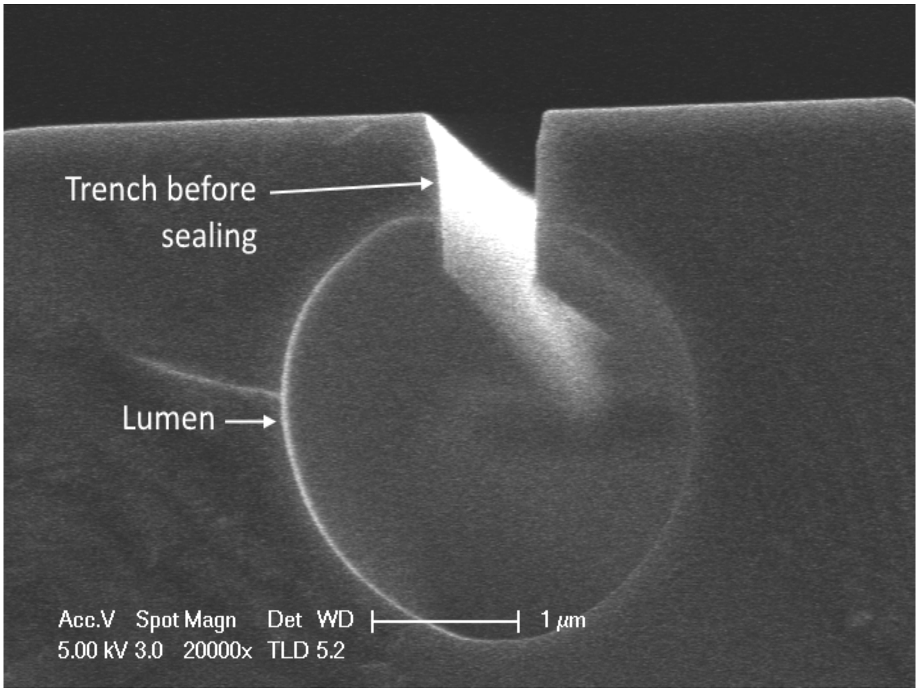

Figure 5.9: Completed microfluidic channel etching, before trench filling.

Before coating, an adhesion promoter of γ-methacryloxypropyltrimethoxy silane was applied to the surface in a bath after dilution to 0.5% in a 50:50 solution of isopropyl alcohol and water. The silane was left overnight to bond to the wafer surface, after which the wafer was removed and allowed to air dry for 30 minutes. The wafer was then rinsed with isopropyl alcohol for 5 minutes, dried using compressed nitrogen gas, and baked at 115^{o}C for 30 minutes. Next, parylene is applied to the surface using a Specialty Coating Systems parylene coating system via sublimation and pyrolysis of the parylene-C precursor. The film was deposited to a thickness of 1μm. The result was also very successful, as shown in figure 5.10 on the right. Due to its availability, parylene was chosen for the fabrication of microfluidic channels for the final wafers.

5.2.1.3 Solving inlet-outlet issues

After the fabrication process was completed, fluid flow was attempted through the microfluidic channels. Simple 3D printed devices were fitted with micro-O-rings to create a seal between the probe inlet and the 3D printed microfluidics. The packaging was simplified to create the shortest channel possible in the 3D printed parts to minimize any issues with the plastic parts, therefore any issues with fluid flow would be isolated to the probes themselves. First attempts to flow fluid through the dig and seal microfluidics failed. Even at high pressures there was no indication that fluid entered the channels.

Multiple theories of the potential issue with these channels were theorized. The first step was

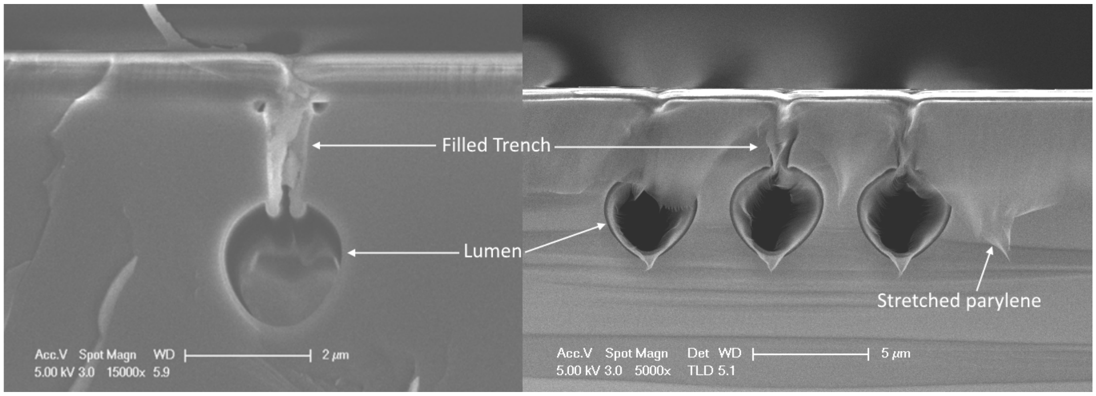

Figure 5.10: Left, microfluidic channel filled by TEOS-PECVD SiO_2. Right, microfluidic channel filled by Parylene-C.

to determine if there was PMMA clogging the inlets and outlets of the channels. The front side of the wafer was protected with PMMA during the backside etch, so it was possible that some of that PMMA wasn't completely dissolved from the inlets. To alleviate this, probes were placed in PGMEA to dissolve any remaining resist for a period of 5 days. Probes were observed under the microscope and there was an apparent change in the coloration of the channel trenches, indicating that this may be helping to clear out the channels. However, upon testing, no flow was observed in these channels. Due to the visual change in the channels, a soak in PGMEA was added to the protocol even if it didn't directly solve the flow issues.

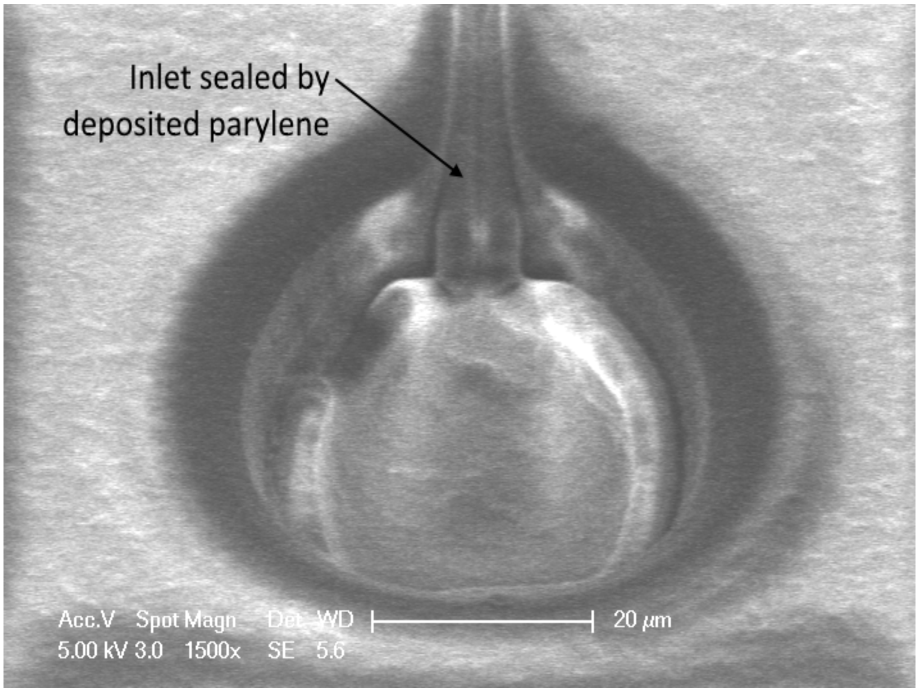

Figure 5.11: SEM micrograph of channel inlet after coating with parylene. The lumen of the channel has completely been covered with parylene.

Next, SEM images of the inlets were then obtained, as shown in figure 5.11. Looking at this image, the first set of probes had a much thicker film than desired (appx. 2μm thick), and it was hypothesized that the inlets may be clogged by excess deposition near the inlets. The physical characteristics of the deposition process are complex, and it was assumed that the channels would remain open after deposition since they were designed to have a circular profile and a significantly larger diameter than the trenches which were filled to create the sealed microfluidic channels. This assumption, however, was incorrect, as seen in figure 5.11.

To solve this problem, an additional process step was added to the fabrication protocol, in which the parylene was etched away from the inlets. The etch for parylene is primarily an oxygen chemistry, so thick photoresist was required to ensure that the parylene intended to seal the length of the channel was sufficiently protected. SPR-220-7 was coated to a thickness of 6.5μm (4000 RPM) on the parylene coated wafer, and was exposed using vacuum contact lithography (Karl Suss MA-6) to a dose of 400μC per cm^2 at 365nm. The wafer was then bath developed in CD-26 for 60s and rinsed with water. This exposed the inlets and outlets of the microfluidic channels while leaving the remainder of the probes protected from etching. Etching was done in a capacitive RIE etcher (Plasmatherm SLR-720 RIE). Oxygen was flowed at 30sccm versus CF4 at 7sccm. Chamber pressure was set to 140mTorr and the forward power of the chamber was set to 80W. The parylene was etched for 15 minutes, which was the maximum possible etch resistance for the photoresist film. This allowed the inlets to be completely removed of any parylene, leaving an open channel. After the etch step, it was found that the channels were open and permitting flow.

5.3 Application 1: Combining Optical Waveguide Biosensors and Microfluidics for Protein Measurements *In Vivo*

This section describes the development of a number of components necessary for the development of neural probes for the detection of neuropeptides *in vivo*. The intent of these probes is to act as a fully-integrated microdialysis probe. By integrating the detection with the microfluidics for sampling the extracellular space, it will be possible to make the detection process faster and more accurate. Furthermore, by controlling the size of the dialysis region, it is possible to pinpoint the location of origination of the target molecules compared to microdialysis. Undertaking this project fully is outside the scope of this thesis,F as this project requires the time and dedication of multiple individuals over a long period of time. Individual components necessary for this probe are described, including the development of ring resonator biosensors and microfluidic devices.

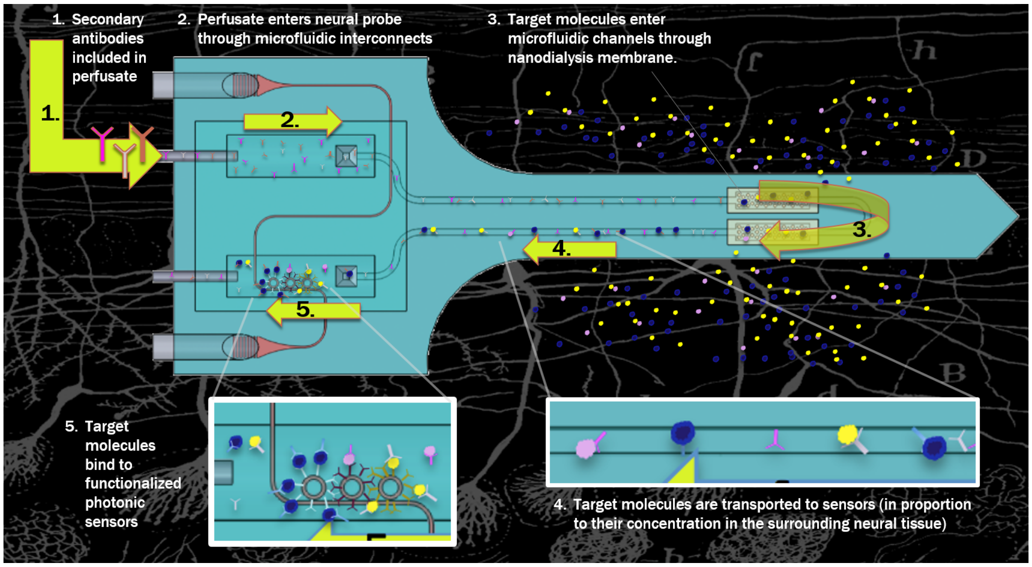
1. Secondary antibodies included in perfusate
2. Perfusate enters neural probe through microfluidic interconnects
3. Target molecules enter microfluidic channels through nanodialysis membrane.
1.
2.
3.
4.
5.
5. Target molecules bind to functionalized photonic sensors
4. Target molecules are transported to sensors (in proportion to their concentration in the surrounding neural tissue)

Figure 5.12: Information flow in a nanodialysis experiment using a nanodialysis probe.

5.3.1 Mechanism For Detection

The flow of an experiment using a nanodialysis probe is shown in figure 5.12. Probes will first be prepared by functionalization of the ring resonator biosensors on the base of the probe by coating the surface with primary antibodies against the target of interest. To ensure that each ring has its own coating, a microarray spotter is used to localize the chemistry over each ring. To cancel out any drift in the resonances of the rings, one ring is not functionalized and may be coated with silicon oxide to ensure there is no binding to the surface. The base of the probe is then sealed to create a microfluidic channel over the ring resonators (which are exposed on the surface of the base of the probe), and to act as a seal between the inlet/outlet of the embedded microfluidics and the pumping apparatus. 3D printed microfluidics were designed for this purpose, and are described in detail later in this section.

After preparation, the probe is next inserted into the brain region of interest. The subsequent measurement protocol is as follows: the microfluidic channels will begin filled with sterile artificial cerebrospinal fluid (ACSF). After allowing the implant to settle and the animal to heal, fluid flow will be initiated to begin bringing analytes from the extracellular space into the probe. Secondary antibodies may be diluted into the ACSF within the probe to improve capture of analytes, as shown in figure 5.12. This fluid will continue through the microfluidic loop where it will reach the nanodialysis region. This region consists of a number of openings forming a grid of holes in a silicon membrane. The silicon membrane acts as a support for a dialysis membrane (typically made from polymers such as cellulose acetate), which will enforce the size selectivity for analytes capable of diffusing into the microfluidic channel. The size exclusion can be tuned when fabricating the membrane. Finally, the fluid (containing the analytes of interest) is returned to the base of the probe through the remainder of the microfluidic loop, and is flowed over the ring resonator biosensors, upon binding of the analytes to the ring resoators, the resonant wavelength are shifted due to the presence of analyte on their surface. This permits detecting the concentration of said analyte in the extracellular space within the brain. Since these rings are coated with specific capter molecules achieve selectivity, it is possible to target a variety of proteins. Additional microfluidic channels may be added to efficiently flush the sample and remove bound analytes to the rings, thereby refreshing their ability to measure the analyte of interest.

5.3.2 Fabrication of Protein-Sensing Probes

5.3.2.1 Mask Design

Masks were designed using a similar template to the uncaging probes described in the previous section, as shown in figure 5.13. In fact, the photolithography masks used to define the top side and back side etches were identical to reduce fabrication cost, as they can be used with both sets of probes. The design of the sensor probes was simplified by the fact that it is unnecessary to place both optical and microfluidic devices onto the shank at the same time. Therefore the probe was only designed with microfluidics in mind. Furthermore, this simplified the stress analysis for the probes, as all of the back oxide can be removed to minimize shank bending. Probe base dimensions were 6.5mm by 7mm. Shanks were 4.75mm long, 60μm wide at the tip and 200μm wide at the base. Due to a more complicated microfluidic design, this probe was not designed around the constraints of o-ring dimensions. Instead, 3D printed gaskets are employed to define the microfluidic geometry, discussed below.

Optical device design was based on the analysis discussed in Chapter 2. Ring resonators of diameter 11μm were used to ensure that losses were not dominated by waveguide curvature, while maximizing the free spectral range of the resonance. Grating couplers were designed empirically, resuling in a periodicity of 420nm and a slab width of 231nm (55% duty cycle), resulting in a maximum acceptance of light at 7^o. The socket dimension of the gratings were 10μm×10μm and were tapered down to a waveguide width of 325nm over 500μm. three ring resonators were multiplexed onto each bus waveguide, with each ring designed to be 20nm larger than the previous to ensure equal spacing within the FSR of the resonance. This result is described later in this chapter.

The other critical design feature is the microfluidic interface between the tissue and the probe. The intent of this region, shown in the lower right corner of figure 5.13, is to provide a region for diffusive exchange between the extracellular space around the implant and the microfluidics on the probe itself. The intent is that there should be a barrier between the two spaces where diffusion of analytes into the microfluidics occurs, while preventing large particles being exchanged between the

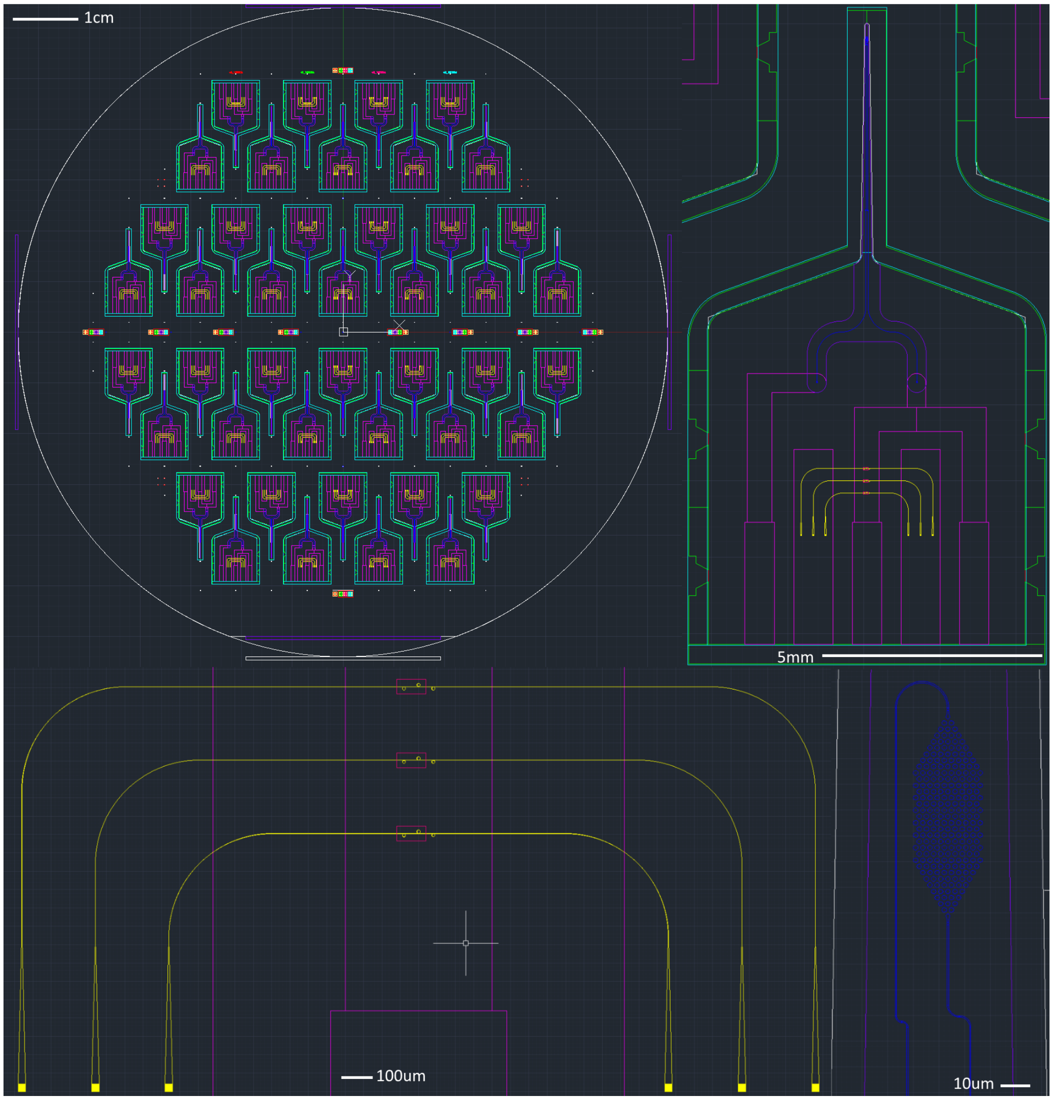

Figure 5.13: Mask design for biosensing neural probes. Due to the need for a microfluidic interface, the probe base dimensions were made larger than previous probes. Layers: optical devices = yellow, microfluidics = blue, top side photolithography mask = green, backside photolithography mask = aqua, parylene = purple, windows for optical clading removal = red.

two volumes. To accomplish this, a phase inversion based membrane will be coated onto the tip of the probe. To support this membrane, a silicon membrane was designed using an array of holes, 2μm in diameter with 3.5μm spacing between holes. The benefit of this arrangement is that the holes are large enough to not be sealed by the parylene during the deposition process, allowing a passageway for particles between the exterior of the probe and the interior. The resulting cross-section of the exchange region is shown in figure 5.14.

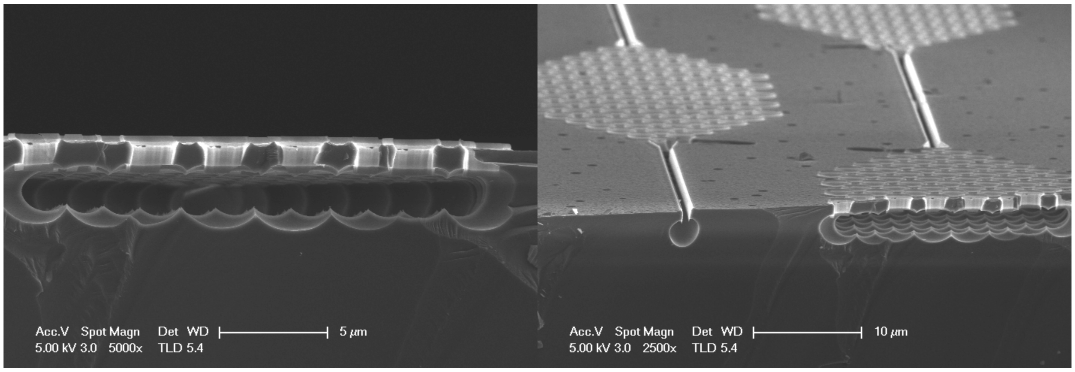
Acc.V Spot Magn Det WD 5 μm
5.00 kV 3.0 5000x TLD 5.4
Acc.V Spot Magn Det WD 10 μm
5.00 kV 3.0 2500x TLD 5.4

Figure 5.14: Fluid exchange region between embedded microfluidics and extracellular space. Left, looking at the nanodialysis region edge on, and right looking at the top of the nanodialysis region on angle.

5.3.2.2 Fabrication Protocol

The specifics of the fabrication protocol can be found in Chapter 3, however a brief summary of the major steps of the fabrication protocol follows, and summarized in figure 5.15. Prime 100mm SOI wafers with $< 1, 0, 0 >$ orientation, device layer thickness of 25μm, buried oxide thickness of 2μm and handle thickness of 300μm were first prepared by wet thermal oxidation to a thickness of 1.5μm by Rogue Valley Microdevices (Medford, OR.). LPCVD silicon nitride was then coated to a thickness of 200nm also by Rogue Valley Microdevices. Wafers were prepared and coated with ZEP-520a per section 3.3.0.2. The wafer was patterned in a Raith EBPG 5000+ with a 300pA beam and a dose of 280μC per square cm and developed for 1 minute in ZED-N50. The wafer was then etched using the ZEP specific pseudoBosch recipe in section 3.4.2.

The wafer was then cleaned thoroughly using solvents, RCA-1, buffered oxide etch, and RCA-2 before coating with PECVD oxide. 1.5μm of PECVD oxide was coated onto the frontside of the wafer using the Oxford Systems 100 Plasma Enhanced CVD using Silane (in argon) and N_20 as the oxidizer, as described in section 3.6.2.

Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 1 (negative tone, green in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. The wafer is then etched with buffered oxide etch at a rate of approximately 250nm per minute for PECVD oxide, and 100nm per minute for thermal oxide. The wafer was monitored during the etch to ensure that a final oxide thickness of 250nm remained in the regions surrounding the probes and where the microfluidics would be etched. Resist was removed after etching in an acetone bath for 15 minutes.

Next the wafer was baked for 2 minutes at 180^{o}C, and spin coated with SML-2000 at 1500 RPM per the recipe in section 3.3.0.1. The wafer was patterned in a Raith EBPG 5000+ with a 100nA

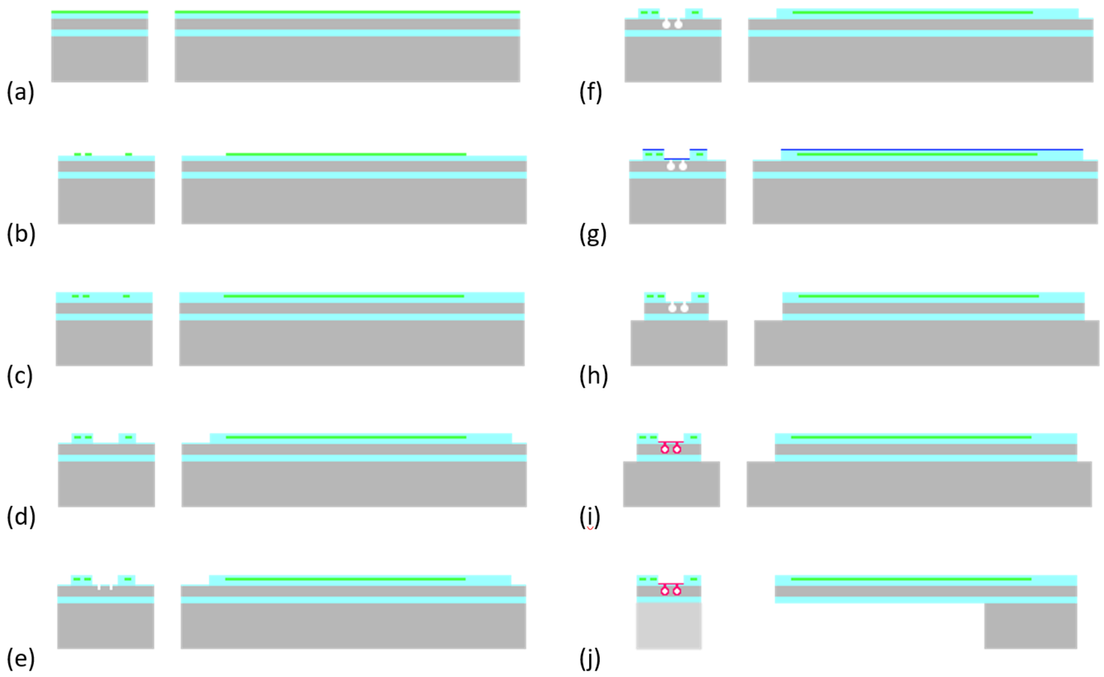

Figure 5.15: Fabrication of uncaging probes. (a) Initial stack of SOI wafer with 1.5μm top silicon oxide and 200nm top silicon nitride, (b) optical layer patterned by EBPG, pseudoBosch etch, (c) PECVD silicon oxide cladding is applied, (d) top oxide is thinned for microfluidics fabrication using BOE, (e), trenches etched for micrfluidics, (f) microfluidic channels are defined with isotropic silicon etch, (g) deposition of alumina for top side hard mask, (h) top side etching via psuedoBosch and Bosch process, (i) deposition of parylene to seal channels, (j) backside etch to release probes.

beam and a dose of 4000μC per square cm and developed for 1 minute in 3:1 IPA:MIBK. Wafers were etched according to the protocol defined in section 5.2.1.2 using the Oxford Instruments 380 ICP etcher. The first step was a 11 minute pseudoBosch etch using the tapered etch described in section 3.4.2.3, followed by a 30s isotropic etch (section 3.4.3), and finished with another 6 minute tapered pseudobosch etch. The wafer was left in the Nanostrip solution overnight, followed by a 10 second dip in BOE. Finally the wafer was cleaned in an oxygen plasma briefly to ensure that the surface of the wafer is clear of organic residue. The wafer was then oxidized under dry conditions at 1000^{o}C for 45 minutes to passivate the sidewalls of the trenches.

The wafer was returned to the Oxford Instruments 380 ICP etcher on an oxide carrier wafer to complete the formation of the microfluidic channels. To penetrate the oxide selectively on the bottom of the wafer, the tapered pseudobosch etch was used for 1 minute 45 seconds. After penetrating the oxide, the isotropic etch described in section 3.4.3 was used to hollow out the body of the channel for 1 minute 30 seconds. The wafer was then cleaned thoroughly in heated PG Remover, acetone, and isopropyl alcohol.

Next, the optical ring resonators must be exposed from the cladding oxide. The wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h-

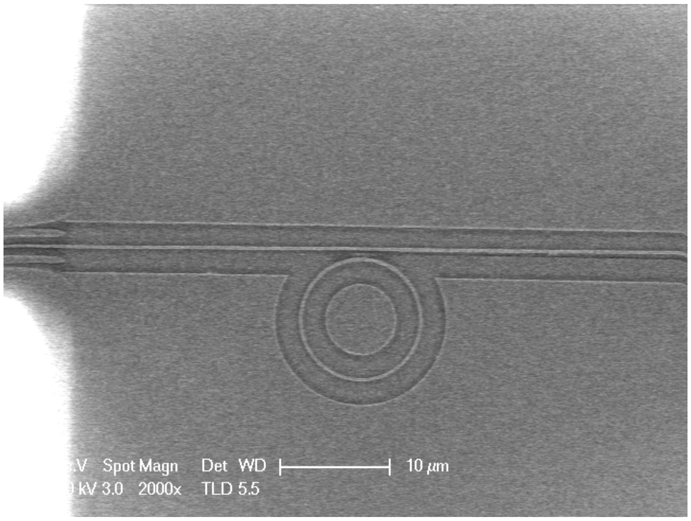

Figure 5.16: Ring resonator after etching away the top oxide cladding the optical waveguides.

and i-line lithography tool for 8 seconds at 25mW per square cm using mask 1 (positive tone, red in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. The wafer is then etched with buffered oxide etch at a rate of approximately 250nm per minute for PECVD oxide. The wafer was etched for 6 minutes in one-minute increments to track the etch rate of the PECVD oxide and ensure that the top oxide was removed while preventing etching into the thermal oxide underneath the ring resonators. Resist was removed after etching in an acetone bath for 15 minutes. The result of this step of the process is shown in figure 5.16

Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 1 (negative tone, green in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited on the wafer using the TES FC-1800 electron beam evaporator to a thickness of approximately 350nm, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is lightly brushed with a swab to remove any excess alumina on the surface, and rinsed in isopropyl alcohol before drying with compressed nitrogen.

After depositing the alumina hard mask, the wafer is etched in a Oxford Instruments 380 ICP etcher with DRIE pod. The wafer is secured to a 6 inch carrier wafer using a thermal contact solution and is placed in the etch chamber. Wafers were first etched using the pseudoBosch recipe designed for SiO_2 removal described in section 3.4.2. At this point the oxide layer is thinned to 100nm, and thus only 5 minutes of etching is required to expose the underlying silicon. After penetrating the top oxide layer, the wafer was exposed to the Bosch etch described in section 3.4.1 for 30 cycles to etch through the device layer of the SOI wafer. Finally, the oxide etch is resumed for an additional 50 minutes to penetrate the buried oxide layer. Wafers are placed in an acetone bath overnight to remove the thermal contact layer from the surface, releasing the wafer from the carrier.

After completing the front-side etch of the probes, parylene must be subsequently deposited on the front-side before etching the back of the wafer. Parylene is deposited as described in section 3.6.3. Briefly, after coating the surface of the wafer with the adhesion promoter, parylene-C is applied to the wafer via pyrolysis at 650^{o}C, resulting in a conformal film on the surface of the wafer. Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 2 (negative tone, purple in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. The wafer is then etched in an oxygen-C_4F_8 polymer using a Plasmatherm SLR-720 reactive ion etcher. Parylene requires 7 minutes and 30 seconds to be removed using this etch. Finally, the remaining resist was removed by soaking in acetone.

Next, backside of the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 2 (negative tone un-mirrored mask, aqua in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited on the wafer using the TES FC-1800 electron beam evaporator to a thickness of approximately 350nm, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is lightly brushed with a swab to remove any excess alumina on the surface, and rinsed in isopropyl alcohol before drying with compressed nitrogen.

Micrographs of the released biosensor probes are shown in figure 5.17.

5.3.2.3 Probe Packaging

Packaging of microfluidic probes was accomplished using 3D-printed microfluidic parts. All 3D printed parts were created using a Stratasys PolyJet 3D printer using Vero materials by Studio Fathom (Oakland, CA, USA). The constraints on this design are much stricter than for the previous devices since there needs to be a microfluidic channel over the surface of the device to allow the ring resonator biosensors to be exposed to the biomolecules from the neural tissue. To accomplish this, a flexible 3D printed material was used instead of using o-rings. This material, so called TangoBlack, was developed to be a 3D printable soft rubber with a hardness of 60-62 on the Shore A scale. A deformable gasket was designed from this material to create this microfluidic channel, as is shown in figure 5.18 on the right. Both materials were printed onto the same part. The gasket protrudes from below the hard plastic portion of the part to ensure that a good seal is made, and the gasket has a curved profile, somewhat like an o-ring, to ensure proper sealing of the channel.

The other component that must be considered is coupling optical fibers to the probe. In this design, there are two possible inlets for optical fibers, one on top of the device (if a fiber is to be glued perpendicularly to the surface of the probe), and the other groove on the bottom of the top

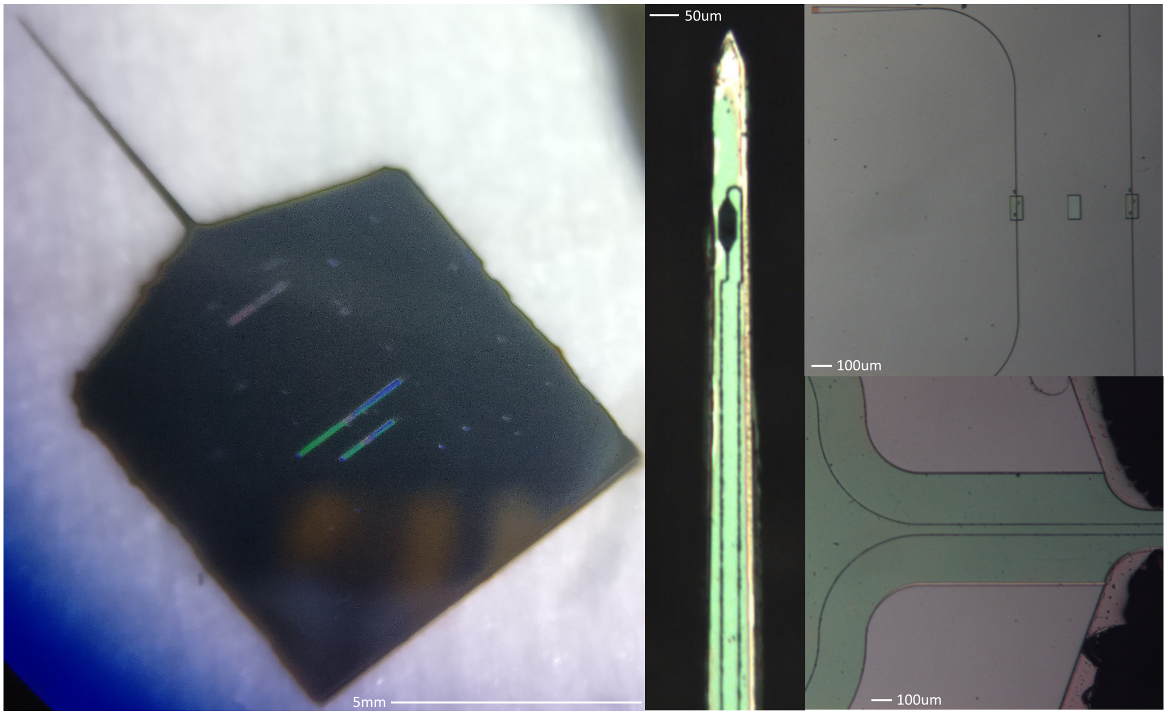

Figure 5.17: Released microfluidic probe micrographs. Left, photograph of a released probe. Center, micrograph of the shank tip. Upper right, ring resonator biosensors. Lower right, microfluidic routing after sealing with parylene near the base of the shank.

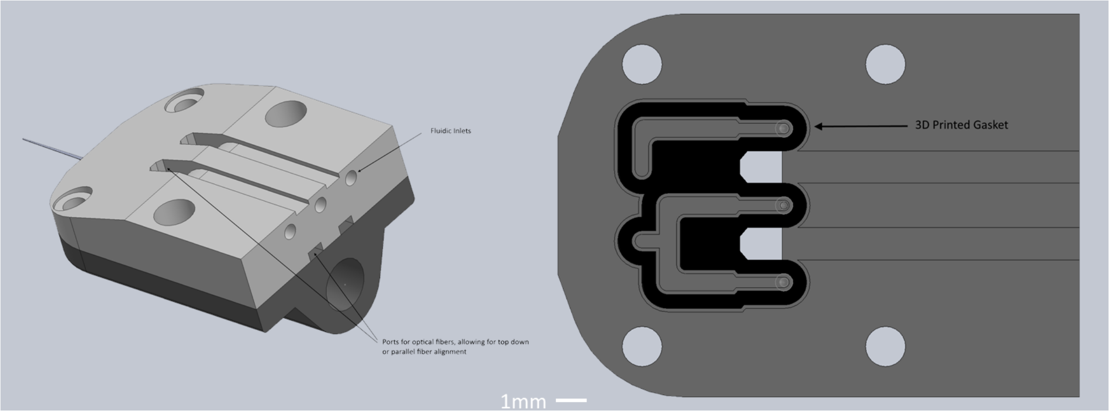
Fluidic Inlets
Ports for optical fibers, allowing for top down or parallel fiber alignment
3D Printed Gasket
1mm

Figure 5.18: Renderings of 3D printed packaging for biosensor probes. The packaged and assembled probe is shown on the left. The gasket used to seal the microfluidics is shown on the right.

piece if the fiber is glued perpendicularly to the probe, as described in Chapter 4. These grooves can be seen in figure 5.18.

The bottom section of the 3D printed part was intended to precisely fit the dimensions of the probe, thereby registering the location of the probe to the top 3D-printed part, ensuring that the o-rings and microfluidic glands are registered to the ports on the probe. The top and bottom portion are held together using 1mm metric screws, with the bottom section of the 3D printed part being

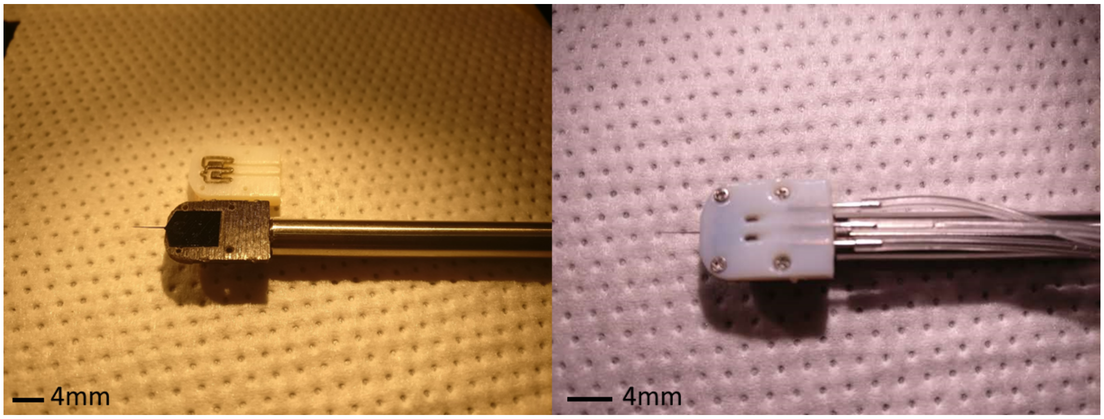

Figure 5.19: Photographs of a packaged probe. Left, before sealing and right, after sealing.

threaded. A probe is shown in figure 5.19 after packaging.

5.3.3 Results

5.3.3.1 Optical Ring Resonator Biosensors

Resonances were fit to a Lorentzian function, defined as

$$I(\lambda) = I_{\mathrm{av}} - \frac{(I_{\mathrm{av}} - I_{\mathrm{min}})(\gamma/2)^2}{(\lambda - \lambda_{\mathrm{min}})^2 + (\gamma/2)^2}, \tag{5.1}$$

where I is the intensity of the optical signal, γ is the loss factor. The quality factor of the resonance can then be calculated as

$$Q = \frac{\sqrt{\lambda_{\mathrm{min}} - 2(\gamma/2)^2}}{\gamma}. \tag{5.2}$$

Multiple parameters in the fabrication of the ring resonator devices were optimized to maximize quality factors of resulting devices. The primary factors to consider are the diameter of the ring resonator (discussed in Chapter 2), the etch process itself (discussed in detail in Chapter 3), the ZEP resist reflow temperature (discussed in Chapter 3), and the gap between the bus waveguide and the resonator. The trade-off incurred when setting the distance between the ring and the bus waveguide is that the extinction of the resonance will reach a maximum (critical coupling), and will then fall off as the gap continues to increase. However, since the bus waveguide also acts as a scatterer in proximity to the ring, the quality factor continues to improve as the gap is increased. Therefore, there is no clear cut distance that is "best" for the operation of the ring resonator, but typically it is good to be close to critical coupling.

Measurements were obtained using the setup shown in figure 5.20. This experimental setup was designed to allow for rapid detection of multiple ring resonators for a biosensing experiment. The major components of the setup include a Velocity TLB-6700 series laser with mode hop free

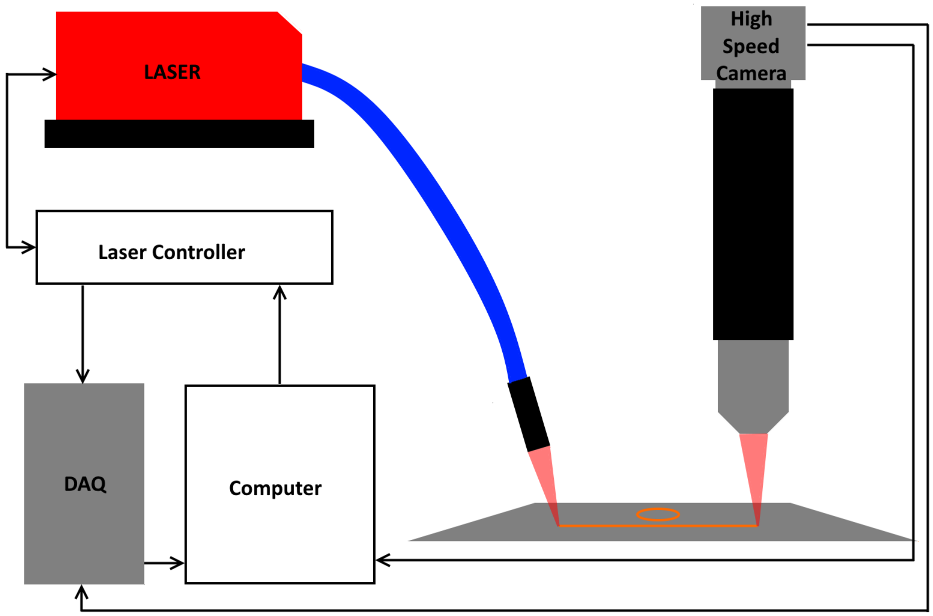

Figure 5.20: Ring resonator measurement setup with information flow between components. A computer controls the flow of information, accepting wavelength data from the laser controller (which, in turn, controls the laser) via a DAQ and accepts data from a camera, which captures the intensity information from the ring resonator output.

tunability from 668-678 nm (Newport Corp., Irvine CA.). This laser source can be scanned up to 5nm per second. This was controlled by a laser controller, which was connected to a personal computer for control of the laser. The laser controller also produces a signal corresponding to the current wavelength of the laser. This was connected to a data acquisition board to measure the current wavelength during the sweep (National Instruments, Austin TX). The DAQ converted the analog signal from the laser to a digital signal which was used by the computer to correlate the intensity of the device with the current wavelength in the scan. Detection of the light intensity was done using a high speed Firefly3 USB 3 camera (FLIR, Richmond, BC, Canada), allowing for reduced frame rates of over 500FPS. The camera could detect multiple ring resonators at a time, allowing for high-throughput measurements. The camera was connected directly to the computer where the data could be extracted. Furthermore, the camera sent a trigger signal to the DAQ to acquire laser wavelength data with each frame, so that the camera and the wavelength of the laser were correlated with each other. Due to the noise of the power supply in the laser controller box, a 21-point averaging filter was implemented on this data. Since the data is known to be linear, there is little worry of filtering artifacts in the data. It should also be noted that this setup was primarily designed to measure the central wavelength of the resonance and, due to the camera used, there may be some distortion of the shape of the peaks recorded. Quality factors may actually be higher than they appear in the data due to saturation of the detector.

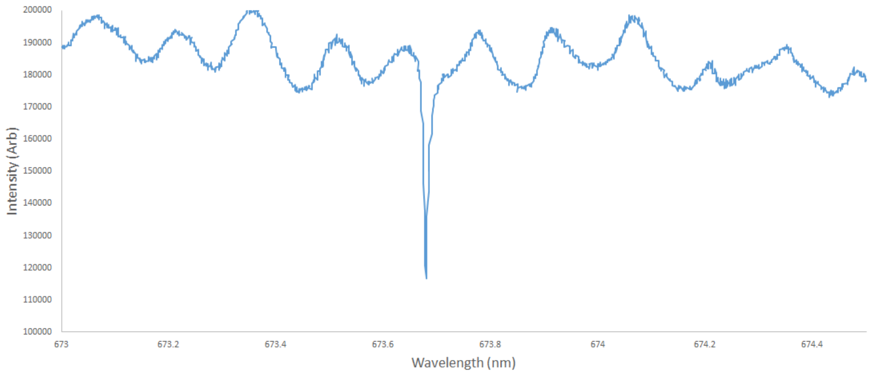

Figure 5.21: Typical resonance curve for an air clad ring resonator biosensor. $Q \approx 59500$.

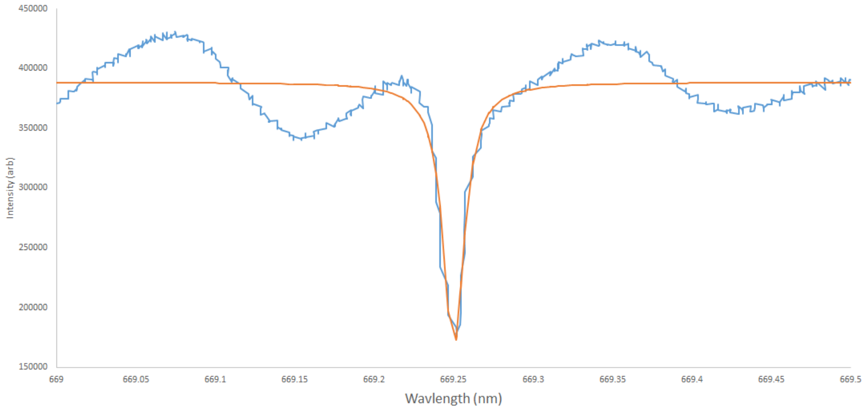

Figure 5.22: Typical resonance curve for a water clad ring resonator biosensor, including best fit line. $Q \approx 42000$.

A typical resonance from a device with air cladding is shown in figure 5.21. The resonance has significantly higher extinction compared to the background Fabry-Perot osculations that are shown in the background, this device showing an extinction ratio of approximatly 45%. Although the typical resonance quality factor is in the range of 50,000-75,000, resonators up to quality factors of 200,000 have been observed using this process. The medium of the ring has a strong influence on the quality factor of the device, with water routinely halving in the quality factor of the same device. This is likely due to small bubbles interacting with the surface of the resonator, causing scattering to occur on the surface of the device. A typical resonance in water is shown in figure 5.22. The typical full width half maximum for these devices in water is typically 16pm, although fitting can further improve the ability to resolve the shift in the resonance wavelength of theses devices.

The trend shown by increasing the gap between the ring and the bus waveguide is shown in figure

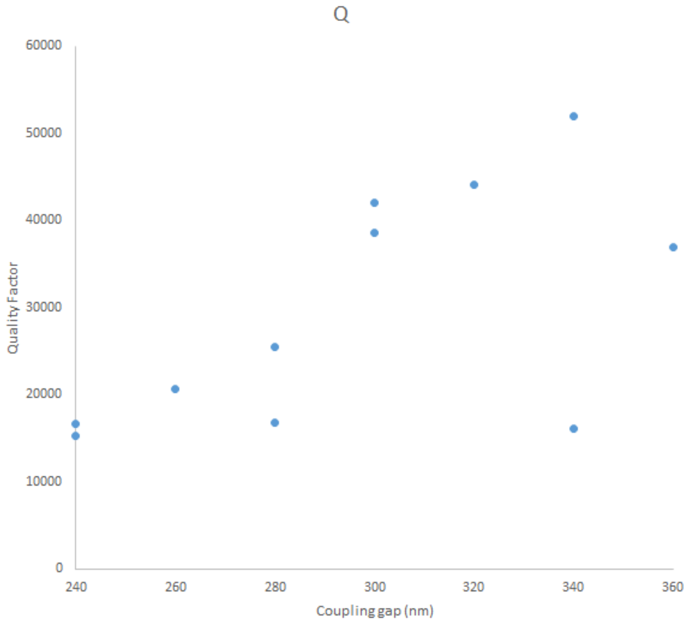

Figure 5.23: Typical quality factor of resonators in water as the gap between the ring and the bus waveguide is varied. These devices were immersed within water.

5.23. This data shows that there is a linear trend between the gap between the bus waveguide and the ring resonator and the quality factor of the ring. This is expected because the bus waveguide, in essence, acts as a source of loss for the ring since light is coupled away from the ring and back into the bus waveguide as the light circulates. Therefore, the further away the ring, the higher the quality factor of the resonance. Using this data, the optimal gap between bus waveguide and resonator was determined to be 340nm. This is a good optimum for typically-achieved quality factors (typically 40,000-50,000), which maintains a strong enough extinction to keep the signal below the Fabry-Perot oscillations in the bus waveguide.

Another important question to consider is the number of ring resonators that can be coupled to a single bus waveguide. The major problem with simply placing multiple ring resonators on a single bus is that the central wavelength of the resonator must be separated from the other rings in wavelength space. Most applications of this technology utilize heaters and the thermo-optic effect to separate the resonances from each other. However, this is undesirable as it adds significant complexity to the fabrication process. Therefore, it was hypothesized that, due to the extremely high resolution of the EBPG used to pattern the ring resonators, it might be possible to utilize slight changes in the writing to induce a difference in the central wavelength of each resonator. The first attempt was to try increasing the diameter of each ring on the waveguide by a small amount. The result of this experiment is shown in figure 5.24. Resonances were successfully multiplexed onto the same bus waveguide by increasing the diameter of each subsequent resonator by 20nm,

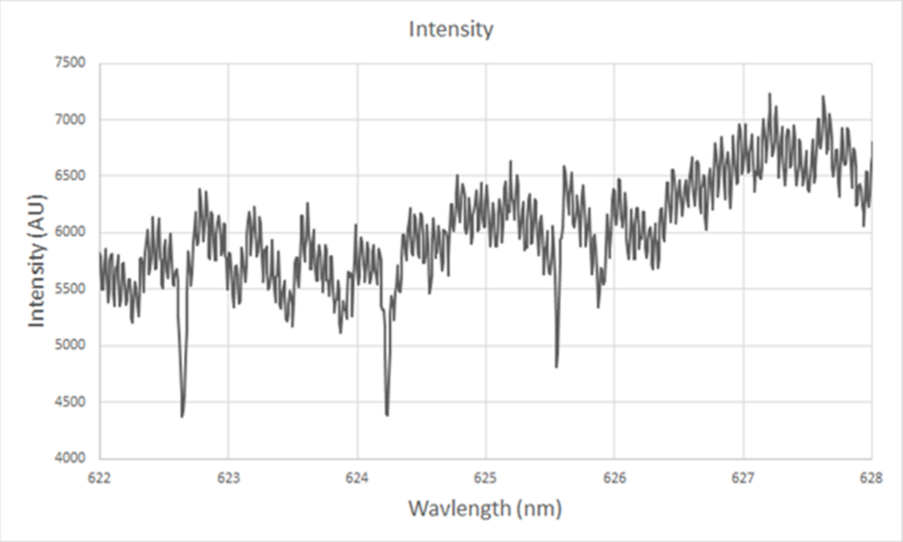

Figure 5.24: Rings multiplexed onto a single bus waveguide, spacing of approximately 1.75nm.

allowing for 3 separate resonators to be measured on the same bus waveguide. This is critical for minimizing the number of connections from the base of the probe to the sensors, allowing for a single bus (and therefore a single input and output fiber) to accommodate two measurement rings and one non-functionalized reference ring.

5.3.3.2 Functionalization of Silicon Nitride Surfaces

Functionalization of the ring resonators utilizes a three-component process. The first component converts the essentially non-reactive silicon nitride surface to a reactive but controllable surface. The second step adds a molecule which is optimized for binding proteins. The final step binds the antibody to the surface of the resonator. The functionalization protocol loosely follows the description in [30].

The chemistry of the functionalization of silicon nitride begins with the surface of a typical silicon nitride film. Silicon is easily oxidized by ambient oxygen in the atmosphere, and is typically hydrophilic due to free -OH groups on the surface. This is advantageous as we can utilize an -OH binding chemistry to convert the surface to a more reactive surface. This is typically done with silane chemistry. Silanes, which have a central silicon surrounded by oxygen groups coupled to an organic chain, will create new Si-O-Si bonds by dehydration of the ether between the silicon and carbon groups. Silanes, such as (3-Mercaptopropyl)trimethoxysilane (MPTMS), have one functional group (in this case an sulfhydryl group) along with the remaining oxygens surrounding the central silicon atom. When exposed to a surface with exposed Si-OH groups, the silane will react to the surface and create a monolayer with the functional group exposed. Thus, by exposing the surface to APTES the surface chemistry moves from a hydroxide to an amine functionalized terminus. To accomplish this, silicon nitride was exposed to 2% MPTMS in anhydrous ethanol. Note that the solvent must be anhydrous, otherwise the silane will undergo hydrolysis and not react with the surface. Surfaces were then washed with anhydrous ethanol and baked at 110^oC to remove any remaining byproducts.

Next, a crosslinker is used to conjugate the surface chemistry to a protein reactive group. These crosslinkers are typically linear molecules, with a specific functional group at each end. Since antibodies have exposed amines which are useful for external chemistry, a crosslinker which is reactive with both the sulfhydryls from silanization of the surface and the free amines on the antibodies is desired. There is a commercially available crosslinker, N-ϵ-maleimidocaproyl-oxysulfosuccinimide ester (Sulfo-EMCS), which has this property. On one end is a maleimide group which is reactive with sulfhydryl groups. On the other end is a Sulfo-NHS ester which is highly reactive with amines. These two functional groups are linked by a 5 carbon backbone. Thus, to functionalize the surface a solution of sulfo-EMCS was disolved to 1mM in a pH 9 sodium tetraborate buffer and exposed to the surface for 2 hours at room temperature. A clean buffer solution was then used to wash the surface. Finally, to complete the process, antibodies or antibody fragments were prepared by dissolution in phosphate buffered saline at a concentration of 0.1mM. These proteins were exposed to the surface for an additional 2 hours.

5.3.4 Future Work

Multiple components of the biosensing neural probe have been demonstrated here, including probe fabrication protocols, development of ring resonator biosensors, and packaging schemes for these devices. Work still needs to be done with the fabrication protocol, as there is a problem when co-integrating microfluidics with optics which causes clogging of the microfluidic channels during fabrication. This problem is described in detail later in the chapter. Furthermore, the development of dialysis membrane technology which can be applied to the shank of the probe must be developed and optimized. Although a plan for developing these membranes is in place, this is outside of the scope of this thesis and should be considered future development work. Finally, once all of the components are co-integrated onto a single probe, calibration of the probe is necessary to determine the ultimate limit of detection of the entire device, including diffusion through the dialysis membrane, capture by antibodies, and detection by ring resonator biosensors.

5.4 Application 2: Integrating Optical Emitters and Microfluidics for Uncaging Experiments

This section describes the development of probes which integrate microfluidics for the injection of caged neurotransmitters with optical devices used to photocleave said caged neurotransmitters, as described in the motivation section above. These probes incorporate straight lengths of microfluidics extending from the base of the probe to the tip with open ends intended for bolus injections of the caged neurotransmitters. Also traversing the length of the shank are optical waveguides terminated with emitter-pixels, with the intent of utilizing the emitters to uncage the neurotransmitters *in vivo*. This technology will enable the stimulation of neurons chemically in deep brain regions.

5.4.1 Fabrication of Uncaging Probes

5.4.1.1 Mask Design

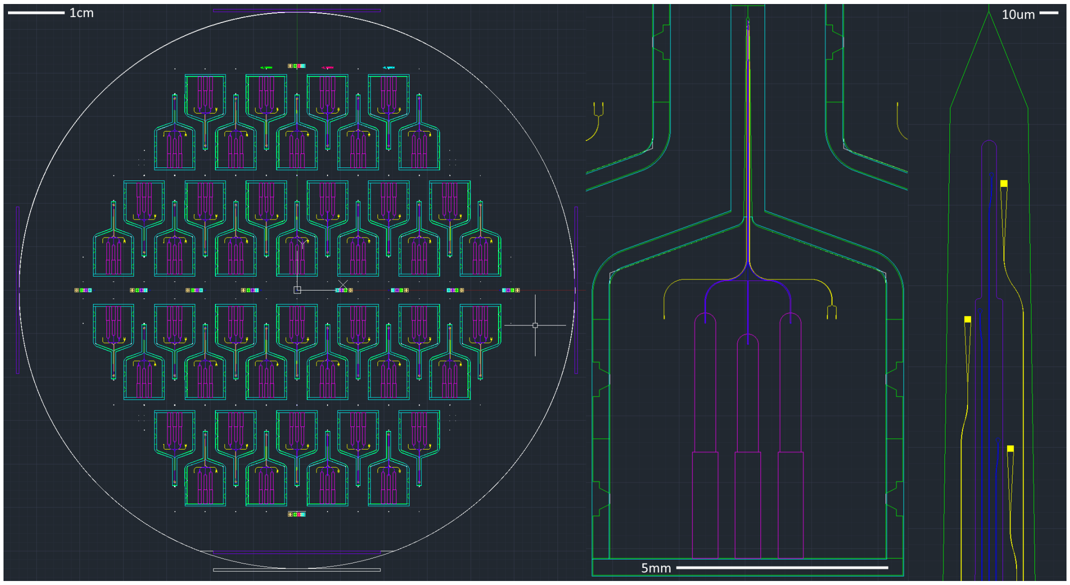

Figure 5.25: Mask design for uncaging neural probes. Due to the need for a microfluidic interface, the probe base dimensions were made larger than previous probes. Layers: optical devices = yellow, microfluidics = blue, top side photolithography mask = green, backside photolithography mask = aqua, parylene = purple.

The finalized mask design for the first generation of neural probes for uncaging is shown in figure 5.25. Waveguide dimensions of 200nm were used to ensure single mode propagation. Optimal grating couplers were determined by testing (data below), with a period of 240nm, and a slab hight of 132nm (55% duty cycle). Grating sockets were 10 × 10μm for input gratings, and 5 × 5μm for output gratings to save space on the shank. Trench width for the microfluidics was 500nm, as determined in the previous section. Ports were fabricated with a circular diameter of 5μm. Probe base dimensions were 6.5mm by 7mm. Shanks were 4.75mm long, 60μm wide at the tip and 200μm wide at the base.

The major consideration for this design architecture is the interface between the microfluidic channels on the probe and the packaging. The base of the probe was made much larger to ensure adequate room for the microfluidic interface. The method determined to ensure an adequate seal is obtained between the probe and external microfluidics is to use o-rings to create a watertight seal. Specialized micro-scale o-rings were purchased, with dimensions of inner diameter 710μm and cross section 410μm were purchased from Apple Rubber (Lancaster, NY. USA). The dimensions and locations of the microfluidic ports were designed around these constraints. Due to the method used for fabrication of the microfluidics, port design was quite simple. Since the seal of the dig-and-seal process is width-dependent, input and output ports were simply fabricated by expanding the trench dimension to be large enough such that the conformal coating of parylene would be unable to seal the trench in that area. Thus, circular regions at the ends of channels with larger diameters naturally act as open ports for fluid flow.

Another major concern with the design was deciding whether to place the mesa for the optical devices on the outside of the shank or in the center of the shank. To ensure structural support, it was decided that the microfluidics should be fabricated in a trench at the center of the shank, as shown in figure 5.25. This was intended to create even stress balance over the length of the shank, especially near the edges of the shank. This design decision included a trade-off between microfluidic channel length and optical device length. Since the microfluidic channels were placed in the center of the shank, they also had a shorter length. It was determined based on the loss coefficient of waveguides at 473 nm that longer waveguides shouldn't induce too much trouble as the propagation losses were low in those designs. It was later discovered, however, that there was much more loss at 405nm than expected. This is likely due to surface roughness of waveguide sidewalls becoming more important as wavelength is reduced. Future designs will include the waveguides in the center of the shank to reduce the length of the waveguides, thereby minimizing losses due to propagation, as described in the following section.

5.4.1.2 Fabrication Protocol

The specifics of the fabrication protocol can be found in Chapter 3, however a brief summary of the major steps of the fabrication protocol follows. Prime 100mm SOI wafers with $< 1, 0, 0 >$ orientation, device layer thickness of 15μm, buried oxide thickness of 2μm and handle thickness of 300μm were first prepared by wet thermal oxidation to a thickness of 1.5μm by Rogue Valley Microdevices (Medford, OR.). LPCVD silicon nitride was then coated to a thickness of 200nm also by Rogue Valley Microdevices. Wafers were prepared and coated with ZEP-520a per section 3.3.0.2. The wafer was patterned in a Raith EBPG 5000+ with a 300pA beam and a dose of 280μC per square cm and developed for 1 minute in ZED-N50. The wafer was then etched using the ZEP specific pseudoBosch recipe in section 3.4.2.

The wafer was then cleaned thoroughly using solvents, RCA-1, buffered oxide etch, and RCA-2 before coating with PECVD oxide. 1.5μm of PECVD oxide was coated onto the frontside of the wafer using the Oxford Systems 100 Plasma Enhanced CVD using Silane (in argon) and N_20 as the oxidizer, as described in section 3.6.2.

Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 1 (negative tone, green in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. The wafer is then etched with buffered oxide etch at a rate of approximately 250nm per minute for PECVD oxide, and 100nm per minute for thermal oxide. The wafer was monitored during the etch to ensure that a final oxide thickness of 250nm was remaining in the regions surrounding the probes and where the microfluidics would be etched. Resist was removed after etching in an acetone bath for 15 minutes.

Next the wafer was baked for 2 minutes at 180^{o}C, and spin coated with SML-2000 at 1500 RPM per the recipe in section 3.3.0.1. The wafer was patterned in a Raith EBPG 5000+ with a 100nA beam and a dose of 4000μC per square cm and developed for 1 minute in 3:1 IPA:MIBK. Wafers were etched according to the protocol defined in section 5.2.1.2 using the Oxford Instruments 380 ICP etcher. The first step was a 11 minute pseudoBosch etch using the tapered etch described in section 3.4.2.3, followed by a 30s isotropic etch (section 3.4.3), and finished with another 6 minute tapered pseudobosch etch. The wafer was left in the Nanostrip solution overnight, followed by a 10 second dip in BOE. Finally the wafer was cleaned in an oxygen plasma briefly to ensure that the surface of the wafer is clear of organic residue. The wafer was then oxidized under dry conditions at 1000^{o}C for 45 minutes to passivate the sidewalls of the trenches.

The wafer was returned to the Oxford Instruments 380 ICP etcher on an oxide carrier wafer to complete the formation of the microfluidic channels. To penetrate the oxide selectively on the bottom of the wafer, the tapered pseudobosch etch was used for 1 minute 45 seconds. After penetrating the oxide, the isotropic etch described in section 3.4.3 was used to hollow out the body of the channel for 1 minute and 30 seconds. The wafer was then cleaned thoroughly in heated PG Remover, acetone and isopropyl alcohol.

Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 1 (negative tone, green in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited on the wafer using the TES FC-1800 electron beam evaporator to a thickness of approximately 350nm, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is lightly brushed with a swab to remove any excess alumina on the surface, and is rinsed in isopropyl alcohol before drying with compressed nitrogen.

After depositing the alumina hard mask, the wafer is etched in a Oxford Instruments 380 ICP etcher with DRIE pod. The wafer is secured to a 6-inch carrier wafer using a thermal contact solution and is placed in the etch chamber. Wafers were first etched using the pseudoBosch recipe designed for SiO_2 removal described in section 3.4.2. At this point the oxide layer is thinned to 100nm, and thus only 5 minutes of etching is required to expose the underlying silicon. After penetrating the top oxide layer, the wafer was exposed to the Bosch etch described in section 3.4.1 for 20 cycles to etch through the device layer of the SOI wafer. Finally, the oxide etch is resumed to penetrate the buried oxide layer, an additional 50 minutes. Wafers are placed in an acetone bath overnight to remove the thermal contact layer from the surface, releasing the wafer from the carrier.

After completing the front side etch of the probes, parylene must be deposited on the frontside before etching the back of the wafer. Parylene is deposited as described in section 3.6.3. Briefly, after coating the surface of the wafer with the adhesion promoter, parylene-C is applied to the wafer via pyrolysis at 650^oC, resulting in a conformal film on the surface of the wafer. Next, the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 2 (negative tone, purple in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. The wafer is then etched in an oxygen-C_4F_8 polymer using a Plasmatherm SLR-720 reactive ion etcher. Parylene requires 7 minutes and 30 seconds to be removed using this etch. Finally, the remaining resist was removed by soaking in acetone.

Next, backside of the wafer is coated with S1813 per section 3.1.2. The wafer is exposed using contact lithography in a Karl Suss MA-6 h- and i-line lithography tool for 8 seconds at 25mW per square cm using mask 2 (negative tone un-mirrored mask, aqua in figure 5.25). The wafer is then developed in CD-26 developer for 1 minute, and rinsed with deionized water. Alumina is next deposited to a thickness of approximately 350nm on the wafer using the TES FC-1800 electron beam evaporator, per section 3.6.4. After deposition, the wafer is left overnight in an acetone bath to complete the liftoff procedure. The wafer is then lightly brushed with a swab to remove any excess alumina on the surface, and is rinsed in isopropyl alcohol before drying with compressed nitrogen.

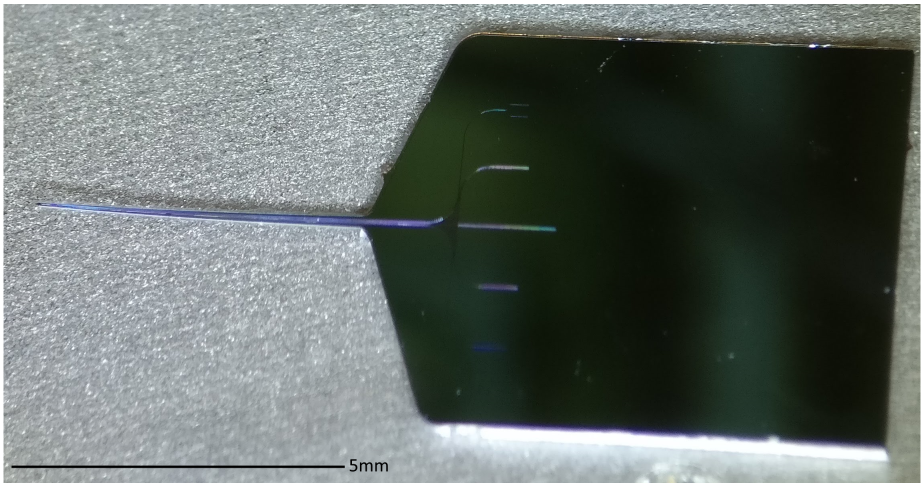

Figure 5.26: Micrograph of released uncaging probe.

After depositing the alumina hard mask, the wafer is etched in a Oxford Instruments 380 ICP etcher with DRIE pod. To protect the frontside structures on the wafer, the wafer is first coated with a thick layer of PMMA A11. The wafer is secured to a 6-inch carrier wafer using a thermal contact solution (Fomblin, Solvay, Bruxelles Belgium) and is placed in the etch chamber. Wafers were first etched using the pseudoBosch recipe designed for SiO_2 removal described in section 3.4.2. Due to the relatively slow etch rate (35nm per minute), approximately 50 minutes of etching were required to penetrate the oxide layer. The etch depth was monitored using a profilometer. After penetrating the top oxide layer, the wafer was exposed to the Bosch etch described in section 3.4.1 for 100 cycles at a time (to allow for wafer cooling) until a total of 300 cycles have been completed. The wafer is then etched at a rate of 10 or 20 cycles at a time until all of the back silicon has been removed by microscopic inspection. This typically takes 340-360 cycles to remove all 300μm of the handle wafer. The wafer (with the carrier) is then placed in an acetone bath overnight to release the wafer from the carrier. The following morning, the wafer is carefully removed, rinsed with isopropyl alcohol, and gently dried with compressed nitrogen gas.

5.4.1.3 Packaging of Probes

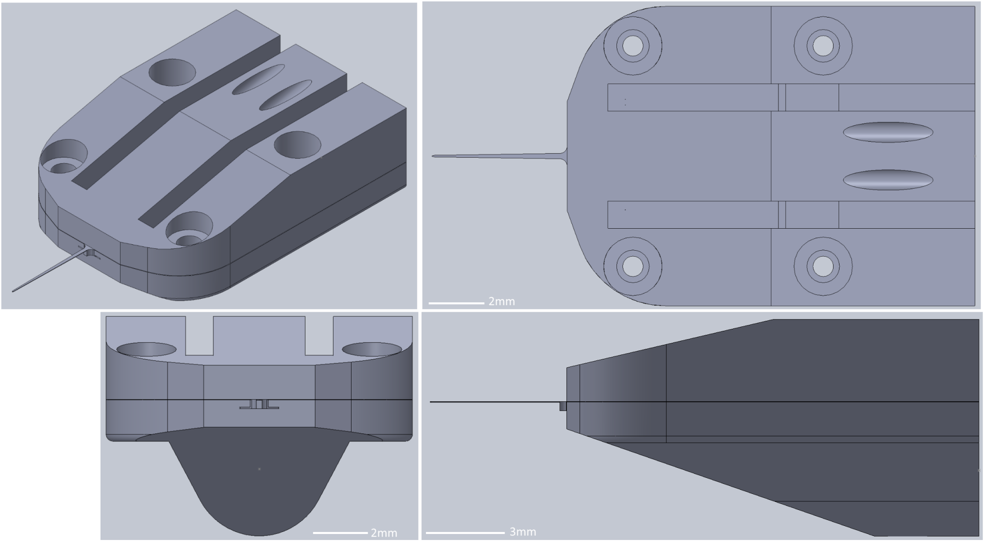

Figure 5.27: CAD design for 3D printed packaging intended for uncaging probes.

Packaging of microfluidic probes was accomplished using 3D printed microfluidic parts. All 3D printed parts were created using a Stratasys PolyJet 3D printer using Vero materials by Studio Fathom (Oakland, CA, USA). The major limitation when designing 3D printed microfluidics is ensuring that the microfluidic lumen is clear of the structural material used to create voids in the structure by the 3D printer. This necessitates using straight lines to create the ports for the microfluidics, hence the ports are positioned on top of the 3D printed part, as shown in figure

5.27. Thus, the tubing inserted into the 3D printed part will make a straight line directly to the input ports on the probe itself. Glands for o-rings were added to the design, intended for o-rings with dimension of inner diameter 710μm and cross section 410μm that were purchased from Apple Rubber (Lancaster, NY. USA). These o-rings were chosen as a compromise between the minimum feature size for voids on the 3D printer and the minimum cross-section of the o-ring itself. Since the gland must be smaller than the inner diameter of the o-ring, a hole diameter of 550μm was chosen to allow the 3D printer to resolve the hole in the center of the gland. This, however, still required post-processing, where a pin was used to ensure that the hole was completely clear of plastic material.

The bottom section of the 3D printed part was intended to precisely fit the dimensions of the probe, thereby registering the location of the probe to the top 3D-printed part, ensuring that the o-rings and microfluidic glands are registered to the ports on the probe. The top and bottom portion are held together using 1mm×4mm metric screws, with the bottom section of the 3D printed part being threaded. This ensures that there is sufficient pressure between the o-rings and the probe to prevent leakage from the probe.

5.4.2 Results

5.4.2.1 Optical Properties of Waveguides at 405nm

Grating Couplers An array of grating couplers with varying pitches (220nm - 300nm) and duty cycles (30% - 70%) were constructed using the EBPG-5000+ using the ZEP process. The pattern consisted of simple single-mode (200nm wide) waveguides connected to tapers at either end of the 2mm waveguide. The tapers fanned out to 5μm, which were direcly coupled to the grating couplers of varying dimensions. Upon the wafer was deposited a "photonics stack" comprised of a 200nm silicon nitride layer on top of a 2μm layer of silicon oxide on top of silicon. The pattern was written with a 300pA beam in ZEP-520A, 2.5nm step size, 280μC per cm^2 dose. After writing, samples were baked for 1 minute at 152^{o}C to reflow the resist. Patterns were etched using the standard nitride pBosch etch described in Chapter 3. The chip was finally clad using silane-based PECVD oxide to a thickness of 1μm.

The measurement apparatus consisted of two micromanipulators, each carrying an individual fiber within stainless steel hypodermic tubing. The input stage also had a goniometer to control the angle of incidence of the optical fiber on the input side of the setup. Light originated from a fiber coupled semiconductor laser (Coherent CUBE, Santa Clara, CA) at 405nm. This was patched into a polarization-maintaining PM-S405-XP fiber patch cable (Thorlabs, Newton, NJ) which was cleaved at one end. This fiber was used to direct the light into the grating couplers. On the other side was a cleaved multimode fiber to collect all light at the output fiber grating (Thorlabs, Newton, NJ). This fiber was coupled to a optical power meter (Newport Corp., Irvine CA) where the output power was measured.

Grating couplers were tested at a nominal angle of 7^o from the normal. The resulting data is shown in figure 5.28. There are two islands of optimal coupling, one with a 240nm pitch and another with a 260nm pitch. Since the coupling efficiency was highest and the duty cycle was closest to 50% (which is easier to fabricate), it was decided that the best option was to use the 240nm pitch and 45% duty cycle moving forward.

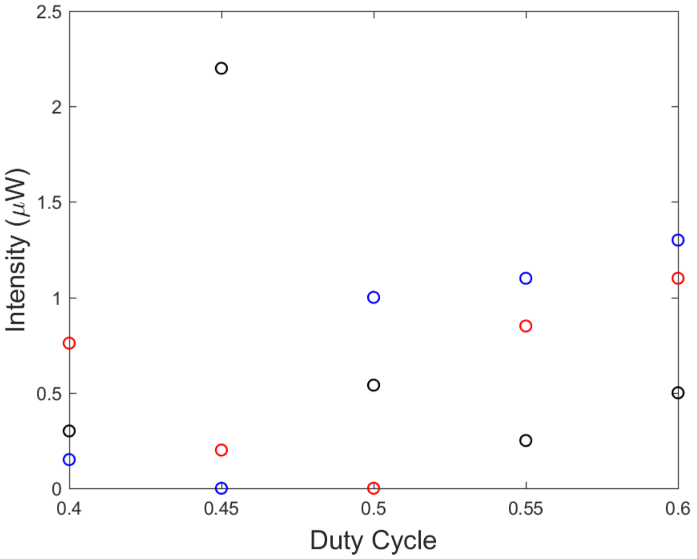

Figure 5.28: Grating coupler throughput data at 405nm using frequency doubled Ti-Sapphire laser. Black curve = 240nm pitch, blue curve = 260nm pitch, red curve = 280nm pitch. Input power = 630μW.

Propagation Losses Measurement of propagation losses was conducted using test chips, where an array of waveguides were patterned with lengths of 2mm, 4mm and 6mm to measure internal losses due to the propagation of light through the silicon nitride waveguides. Waveguides were tapered to 200nm width briefly to ensure single-mode propagation, and were then tapered out to 600nm to improve mode confinement. Each fiber was connected with gratings on each end with the parameters described in the previous section. The silicon wafer had a photonics stack of 200nm silicon nitride on top of a 2μm layer of silicon oxide on top of silicon. The pattern was written with a 300pA beam, 2.5nm step size, 280μC per cm^2 dose. After writing, samples were baked for 1 minute at 152^{o}C to reflow the resist. Patterns were etched using the standard nitride pBosch etch described in Chapter 3. The chip was finally clad using silane based PECVD oxide to a thickness of 1μm.

The measurement apparatus was identical to the one described in the previous section, but for two important changes. The laser source used for these experiments was a frequency doubled Ti-

Sapphire laser tuned to 405nm. Second, a polarization controller was added to the input fiber to ensure that the gratings were receiving light at the optimal polarization angle.

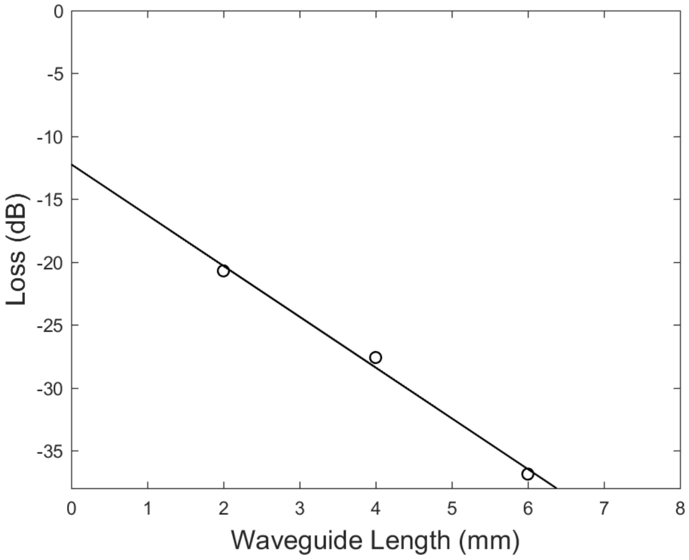

Figure 5.29: Measured waveguide propagation losses at 405nm.

The resulting loss curves are shown in figure 5.29. This figure shows that the losses incurred by common factors (mostly the grating couplers) is approximately 12dB. Although this is an acceptable loss, it could be further improved by optimizing grating parameters or the incident angle of the input fiber. The troubling data is the slope of the line, showing a loss of over 4dB per mm. Thus, for a 5mm waveguide a source of 100mW is necessary to create an output beam of 1mW for uncaging, and this does not include grating coupler losses. Multiple attempts were made to improve the waveguide losses, however no good solution was found.

When this project was started, the longest wavelength usable for uncaging neurotransmitters was 405nm. Since then, caged compounds have been developed which function at 470nm[29]. This now allows the use of the waveguide structures operating at longer wavelengths, as discussed in the previous chapter on optogenetics. Loss data for those waveguides is detailed in Chapter 4.

5.4.2.2 Microfluidic injections of Dye

To test the microfluidic architecture designed and described above, simplified probes were fabricated which only included the microfluidic fabrication while omitting the optical devices. This was a test of the efficacy of the microfluidic fabrication protocol and ensure that the microfluidic devices (including packaging and interfacing to the probe microfluidics from 3D-printed parts) were able to permit fluid flow.

Testing of the microfluidic channels utilized a set of two simple 3D printed parts to bring fluid to

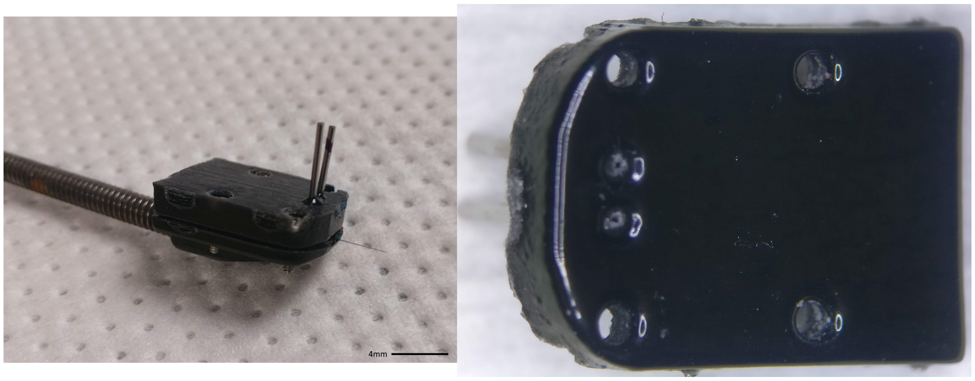

Figure 5.30: Simplified packaging for microfluidics-only probes. The complete packaging assembly is shown on the left, and the bottom surface of the top section (which forms the seal with the probe) is shown on the right.

the chip and create a seal between the stainless steel hypodermic tubing used to deliver fluid and the probe itself. 3D printed parts were designed in-house and printed using Stratysis Polyjet 3D printers (Studio Fathom, Oakland CA). Seals were created using a micro-O-ring (Apple Rubber, Lancaster NY) with inner diameter 710μm and cross section 410μm. The grooves for the o-rings is shown in figure 5.30 on the right. O-rings are seated within the grooves, and the assembly is flipped onto the probe seated in the base part (shown in figure 5.30 on the left). The pieces are screwed together using M1 screws (1mm×4mm), and tightened if needed using nuts. The screws were sufficient as guides to ensure that the location of the inlets on the probes was within the extent of the o-rings.

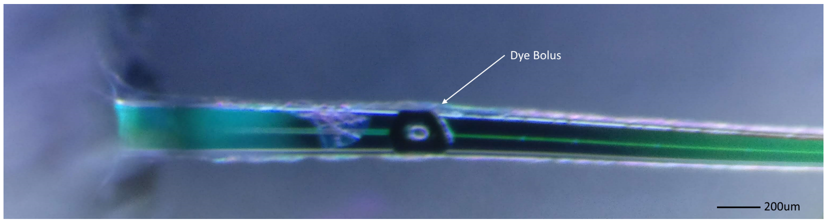

Figure 5.31: Blue dye escaping a microfluidic outlet on probe shank.

Fluid was pressurized into the channels through a syringe using a syringe pump (Harvard Apparatus, Holliston, MA). To expedite the filling process, a high flow rate was chosen, typically 10μL per minute. Blue food dye was loaded into the channels to demonstrate flow, as the food dye is highly light-absorbing, making it easy to visualize under a microscope. When fluid successfully travels through the channel, a bead of liquid is formed at the inlet, as is shown in figure 5.31. In this image, blue food dye is shown accumulating and covering a portion of the shank.

Next, the probe was then submerged in a water bath to visualize the flow coming from the outlet and to simulate the bolus injection of fluid into a liquid environment. Upon activating the syringe pump at a flow rate of 10μL per minute, a stream of dye was emitted from the microfluidic

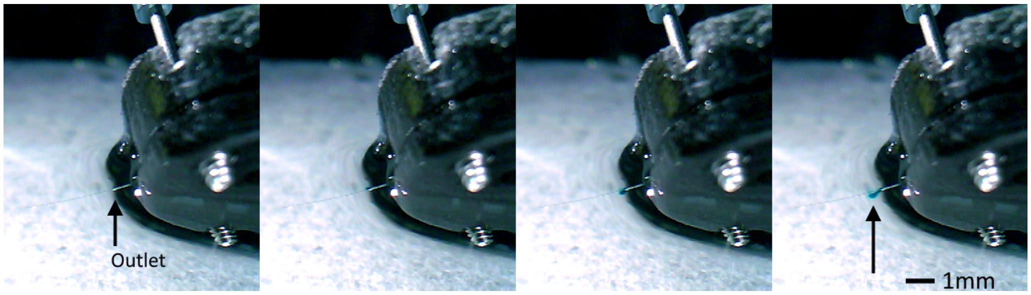

Figure 5.32: Time series of dye injected into water bath from microfluidic probe using a syringe pump (10μL per minute). Frames are 5 seconds apart.

outlet. This is shown as an image sequence in figure 5.32, with each frame being approximately 5 seconds apart. Expected channel performance under a head of 1PSI should produce 10nL per minute of flow for a 4mm long channel. Although this may seem low, this actually clears the channel (volume 0.2nL) once every second, resulting in a flow velocity of approximately 4mm/sec. With the successful injection of fluid from a microfluidic probe, the next step is to combine the microfluidic technology with other technologies to create a functional, multi-application neural probe for scientific experimentation.

5.5 Flow issues in combined optics + microfluidics probes

After fabricating multiple probe wafers, it was found that there was an extremely high failure rate in the microfluidics of these probes. Although this could be due to the fact that these probe designs necessitate longer microfluidic channels which will result in higher failure rates, it seems unlikely that all of the microfluidic channels tested would fail. Thus, it seems that something in the fabrication protocol is to blame for the failures in the microfluidic channels when adding the optical devices to the devices.

There are two major differences in the fabrication protocol when adding the optical devices. The first major difference is that the optical devices must sit on a $1.5-2\mu$m silicon oxide mesa to isolate the evanescent field from the underlying silicon. The microfluidic stack, however, requires that the top oxide be 225nm thick. Therefore, a wet etch using BOE was used to thin down the top oxide in the region where the microfluidics would be patterned. Using a wet etch has the advantage of a consistent and fast etch rate (100nm per minute) using buffered oxide etch, however it is isotropic, so, depending on the design of the photomask, there is the possibility that an uneven substrate will be formed. For these probes, the width of the valleys between mesas is between 21.5μm or 11.5μm, depending on the location on the probe. All of the devices in the 11.5μm region failed due to curvature in the etched substrate, however the interference color of the substrate in the 21.5μm region appeared functional on visual inspection under a microscope. However, it is possible that this did not correlate well with test chips and may have caused trenches to form improperly.

The second issue with adding the optics to the fabrication protocol is the fact that during the second etch process which forms the bottom of the channel the optical features must be protected from the etch process. This necessitates coating the mesas with photoresist to ensure they are not etched during processing of the microfluidic channels. Although this should be a simple task, it is possible that photoresist may remain in the channel during the development process due to the geometry of the trenches. This is especially a concern given that the photoresist will be coated more thickly within the valleys created by the mesas where the microfluidic channels reside. If, after development, small amounts of photoresist are left over, the areas coated with the remaining resist will not be etched, and this will create a blockage in the channel. This has been observed at times on test wafers when they were imaged, however it only takes one small region of photoresist remaining over the entire length of the channel to create a blockage and reduce flow to zero.

There are a number of possible solutions to these problems. The first solution involves eliminating the mesa structures altogether. This would require etching through 1μm of PECVD oxide and 1.5μm of thermal oxide to get down to the silicon below. Although this is possible, it would require moving to a hard mask, such as aluminum oxide, to ensure that there was enough etch resistance to penetrate these layers by the etch. This would likely preclude the use of a pseudoBosch etch due to low etch rates. However, this problematic alternative etches do not allow for the control of side wall angle, which is critical for channel filling.

An alternative plan would be to continue etching the mesas as before, but use a hard mask instead of a photoresist mask to protect the mesas. The benefit of this is that it would allow the photoresist to completely be removed by a long soak in solvent, while still protecting the mesas sufficiently. By the end of the soak the trenches should be completely clear of material. There is still some risk of the hard mask material settling into the trenches and sticking, however utilizing sonication may be sufficient to ensure that the channels are clean of debris. An alternative solution to ensure that no material is left in the trenches during the protection step would be to expose the patterned wafer to an oxygen plasma to remove any residue. This may be the simplest next step to solving this problem.

Chapter 6

Conclusion

This thesis describes the development of a number of silicon based neural probe technologies for the optical stimulation and excitation of proteins and chemicals in the brain. Furthermore, it details the development of microfluidics for the delivery of pharmacological agents and the detection of neuropeptides deep within the brain. The concluding summary of the thesis follows.

Chapter 2 was concerned primarily with theoretical considerations of optical ring resonator biosensors and their sensitivity of protein detection. It was shown that, using a physics based approach, the sensitivity of ring resonator biosensors is phenomenally high (tens of femtomolar in concentration), and much higher than any realistic measurements have shown. This is confirmed by estimates of physical biosensor sensitivities shown in the literature, which have never met expectations in real biological solutions. Next, a chemical approach to non-specific binding was undertaken to help understand why these biosensors do not perform to this level in real solutions (such as cell lysate or blood plasma). By accepting that there will be competitive binding between the target analyte and other proteins in solution, it is possible to understand that the concentration of the analyte must be high enough to out-compete the other proteins for space on the sensor surface. This approach gave estimates of 10nM for label-free detection and 1pM for sandwich assays, which fits the ultimate limits of detection described in the literature quite well. The next step in this research would be to use a gold standard tool (such as a Biacore surface plasmon resonance biosensor) to better understand the non-specific binding processes on antibody coated surfaces. This could lead to a verification of this theory with data by altering non-specific vs. specific binding in a variety of combinations, allowing for the theory to be rigorously tested.

Chapter 3 was concerned with the fabrication of the devices described in this thesis. The practicum and design of processes for photolithography, electron beam lithography, etching and deposition of thin films for this project was described in detail. The process flow for each type of device or process was also described. In addition to the practical concerns of fabrication, a novel electron beam lithography resist was developed in collaboration with Dr. Scott Lewis. This resist was developed to confer a positive tonality to a very effective negative tone resist currently being developed by Dr. Lewis. By combining metallic macromolecules with standard positive tone resists (PMMA and ZEP), it was possible to improve the etch resistance of these resists by a factor of 4. This could allow for the production of extremely high aspect ratio features, or allow electron beam lithography to be used with long etch processes. Further research into these resists is possible, including the addition of secondary electron generators (such as diallylamine or mercury chloride) to speed up resist writing, and to further explore the etch resistance and resolution of this resist.

Chapter 4 concerned the development of optical neural probes for a variety of applications. The concept of spectrally addressing point emitters on neural probes was introduced, including a plan to use array waveguide gratings as a means of spectral demultiplexing signals from the bench top to the probe. This was a challenging task because these devices had yet to be demonstrated in the visible regime. These probes were implemented and data showing spectral addressing of emitter pixels was demonstrated. These probes were tested for optogenetic stimulation in living animals and showed the ability to stimulate single, targeted neurons within the hippocampus of mice. Spectral information could also be used to steer light using a phased array. This was similar in design to the AWG, however light was focused out of plane of the probe and allowed to propagate into tissue. These devices were demonstrated at 670nm and showed good very good steering ability in the 10nm of bandwidth provided by the laser, tuning over 42 degrees.

Due to limitations of optical sources currently available, a switch to spatial addressing of emitters was made. This utilizes a MEMS mirror system to steer a beam impinging on a fiber bundle, which in turn is coupled to a number of edge couplers along the side of a probe. This technique removes a number of limitations of grating couplers, such as low bandwidth and the need to couple fibers from the top of the probe.

Novel imaging probes were also discussed in Chapter 4. The fundamentals of a probe based light sheet fluorescent microscope were established, utilizing optical neural probes with long, narrow grating designs to create lamina of illumination within tissue. These probes were fabricated and preliminary results were demonstrated. In addition, picosecond pulses for a probe-based fluorescent lifetime imaging system were developed, and the resulting spectra of pico-second pulses through these optical devices were shown. Future research directions coming from this research are numerous. The continuation of the laminar illumination pathway and light sheet microscopy project is clear, in that there needs to be significant testing of the probes moving forward. The divergence of the beam both in solution and in tissue must be determined to understand how effective this technique will be for illuminating live tissue. Polarization control must be implemented as well to ensure that the sheets are single mode. The imaging system must also continue development to realize a full system for *in vivo* imaging. Furthermore, laminar illumination is possible with the phased array devices by providing a broadband source of illumination, and should be tested for efficacy.

Chapter 5 concerned the development of microfluidics on silicon neural probes. A dig-and-seal approach was decided upon to create small microfluidic channels (1-5μm) which were buried deep within the silicon substrate to minimize risk of probe fracture. Most comparable microfluidic probes had larger lumens and much thicker shanks to their probes. The development of the fabrication protocol began with Bosch based trenches, however it was found that the polymer passivation was insufficient for maintaining sidewall integrity during a long isotropic etch to create the channel. The process was changed to include thermal oxidation for sidewall passivation. Although this process was successful in creating an appropriate cross-section, the trenches created by the Bosch process did not seal well. To remedy this, a change to a pseudo-Bosch etch was made because it allows for the sidewall angle of the etch to be adjusted. An etch with an appreciable sidewall slant was created to accommodate the trench filling process, inducing filling from the bottom of the trench upwards. This ensures that the trench is completely filled, creating a strong seal for the channel below. Two techniques were attempted for trench filling, and parylene was chosen to fill the channels. Fluid flow was demonstrated through microfluidic channels using a dye solution.

Next, two projects were conceived to utilize on-shank microfluidics and the optical devices described in Chapter 4. The goal of the first project was to create probes which could be used for neural stimulation using caged neurotransmitters. This involved the patterning of emitter-pixels on a shank with microfluidic ports. Caged neurotransmitters would be injected into the subject, and e-pixels could then be used to stimulate the caged compounds, causing them to be activated. The second project intended to create neural probes for rapid detection of neuropeptides in the brain via nanodialysis. These probes would utilize microfluidics to sample the extracellular space within the brain, capturing proteins of interest. Optical ring resonator biosensors on the base of the probe would be used to make measurements of protein concetrations within the dialysate. Although both of these projects showed promising results, there was an issue when combining the microfluidic fabrication with the constraints imposed by the optics fabrication process. This was likely due to the fabrication of the microfludic channels within the valleys created by the mesas which contained the optical devices. Modifying the fabrication process of the microfluidic channels is a necessary next step in developing these devices. This could include utilization of the thick oxide layer to act as the neck of the channel in the dig-and-seal approach, fabricating the microfluidics before adding the optical layers, or modifying the optics to better accommodate the microfluidic fabrication.

Bibliography

[1] Adams, R. N. Probing brain chemistry with electroanalytical techniques. *Analytical Chemistry 48*, 14 (1976), 1126A–1138A.

[2] Adar, R., Henry, C. H., Kazarinov, R. F., and Kistler, R. C. Optical waveguide comprising bragg grating coupling means, Mar. 16 1993. US Patent 5,195,161.

[3] Akerboom, J., Chen, T.-W., Wardill, T. J., Tian, L., Marvin, J. S., Mutlu, S., Calderón, N. C., Esposti, F., Borghuis, B. G., Sun, X. R., et al. Optimization of a gcamp calcium indicator for neural activity imaging. *Journal of Neuroscience 32*, 40 (2012), 13819–13840.

[4] Amara, S. G., and Pacholczyk, T. Sodium-dependent neurotransmitter reuptake systems. *Current opinion in neurobiology 1*, 1 (1991), 84–90.

[5] Araya, R., Jiang, J., Eisenthal, K. B., and Yuste, R. The spine neck filters membrane potentials. *Proceedings of the National Academy of Sciences 103*, 47 (2006), 17961–17966.

[6] Arlett, J., Myers, E., and Roukes, M. Comparative advantages of mechanical biosensors. *Nature nanotechnology 6*, 4 (2011), 203.

[7] Bartels, A., Willemsen, A., Doorduin, J., De Vries, E., Dierckx, R., and Leenders, K. [11c]-pk11195 pet: quantification of neuroinflammation and a monitor of anti-inflammatory treatment in parkinson's disease? *Parkinsonism & related disorders 16*, 1 (2010), 57–59.

[8] Beaumont, K., Chilton, W., Yamamura, H., and Enna, S. Muscimol binding in rat brain: association with synaptic gaba receptors. *Brain Research 148*, 1 (1978), 153–162.

[9] Bogaerts, W., Dumon, P., Van Thourhout, D., and Baets, R. Low-loss, low-cross-talk crossings for silicon-on-insulator nanophotonic waveguides. *Optics letters 32*, 19 (2007), 2801–2803.

[10] Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., and Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience 8*, 9 (2005), 1263.

[11] Brackett, C. A. Dense wavelength division multiplexing networks: Principles and applications. *IEEE Journal on Selected Areas in Communications 8*, 6 (1990), 948–964.

[12] Branner, A., and Normann, R. A. A multielectrode array for intrafascicular recording and stimulation in sciatic nerve of cats. *Brain research bulletin 51*, 4 (2000), 293–306.

[13] Callaway, E. M., and Katz, L. C. Photostimulation using caged glutamate reveals functional circuitry in living brain slices. *Proceedings of the National Academy of Sciences 90*, 16 (1993), 7661–7665.

[14] Campbell, P. K., Jones, K. E., Huber, R. J., Horch, K. W., and Normann, R. A. A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array. *IEEE Transactions on Biomedical Engineering 38*, 8 (1991), 758–768.

[15] Canepari, M., Nelson, L., Papageorgiou, G., Corrie, J., and Ogden, D. Photochemical and pharmacological evaluation of 7-nitroindolinyl-and 4-methoxy-7-nitroindolinyl-amino acids as novel, fast caged neurotransmitters. *Journal of neuroscience methods 112*, 1 (2001), 29–42.

[16] Carter, A. G., and Sabatini, B. L. State-dependent calcium signaling in dendritic spines of striatal medium spiny neurons. *Neuron 44*, 3 (2004), 483–493.

[17] Chefer, V. I., Thompson, A. C., Zapata, A., and Shippenberg, T. S. Overview of brain microdialysis. *Current protocols in neuroscience 47*, 1 (2009), 7–1.

[18] Chelaru, M. I., and Jog, M. S. Spike source localization with tetrodes. *Journal of neuroscience methods 142*, 2 (2005), 305–315.

[19] Chen, B.-C., Legant, W. R., Wang, K., Shao, L., Milkie, D. E., Davidson, M. W., Janetopoulos, C., Wu, X. S., Hammer, J. A., Liu, Z., et al. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science 346*, 6208 (2014), 1257998.

[20] Cheung, K. C., Djupsund, K., Dan, Y., and Lee, L. P. Implantable multichannel electrode array based on soi technology. *Journal of Microelectromechanical Systems 12*, 2 (2003), 179–184.

[21] Cotton, R. J., Froudarakis, E., Storer, P., Saggau, P., and Tolias, A. S. Three-dimensional mapping of microcircuit correlation structure. *Frontiers in neural circuits 7* (2013), 151.

[22] Crowe, S. E., Kantevari, S., and Ellis-Davies, G. C. Photochemically initiated intracellular astrocytic calcium waves in living mice using two-photon uncaging of ip3. *ACS chemical neuroscience 1*, 8 (2010), 575–585.

[23] Dakss, M., Kuhn, L., Heidrich, P., and Scott, B. Grating coupler for efficient excitation of optical guided waves in thin films. *Applied physics letters 16*, 12 (1970), 523–525.

[24] Deal, B. E., and Grove, A. General relationship for the thermal oxidation of silicon. *Journal of Applied Physics 36*, 12 (1965), 3770–3778.

[25] Denk, W. Two-photon scanning photochemical microscopy: mapping ligand-gated ion channel distributions. *Proceedings of the National Academy of Sciences 91*, 14 (1994), 6629–6633.

[26] Denk, W., Strickler, J. H., and Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science 248*, 4951 (1990), 73–76.

[27] Doylend, J. K., Heck, M., Bovington, J. T., Peters, J. D., Coldren, L., and Bowers, J. Two-dimensional free-space beam steering with an optical phased array on silicon-on-insulator. *Optics express 19*, 22 (2011), 21595–21604.

[28] Elsner, H., Meyer, H., Voigt, A., and Grützner, G. Evaluation of ma-na 2400 series duv photoresist for electron beam exposure. *Microelectronic engineering 46*, 1-4 (1999), 389–392.

[29] Fino, E., Araya, R., Peterka, D. S., Salierno, M., Etchenique, R., and Yuste, R. Rubi-glutamate: two-photon and visible-light photoactivation of neurons and dendritic spines. *Frontiers in neural circuits 3* (2009), 2.

[30] Fixe, F., Chu, V., Prazeres, D., and Conde, J. An on-chip thin film photodetector for the quantification of dna probes and targets in microarrays. *Nucleic acids research 32*, 9 (2004), e70–e70.

[31] Fortin, J. B., and Lu, T.-M. *Chemical vapor deposition polymerization: the growth and properties of parylene thin films.* Springer Science & Business Media, 2003.

[32] Frey, O., Van Der Wal, P., Spieth, S., Brett, O., Seidl, K., Paul, O., Ruther, P., Zengerle, R., and De Rooij, N. Biosensor microprobes with integrated microfluidic channels for bi-directional neurochemical interaction. *Journal of Neural Engineering 8*, 6 (2011), 066001.

[33] Fridman, A. *Plasma chemistry.* Cambridge university press, 2008.

[34] Fu, Y., Ye, T., Tang, W., and Chu, T. Efficient adiabatic silicon-on-insulator waveguide taper. *Photonics Research 2*, 3 (2014), A41–A44.

[35] Gasparini, S., and Magee, J. C. State-dependent dendritic computation in hippocampal ca1 pyramidal neurons. *Journal of Neuroscience 26*, 7 (2006), 2088–2100.

[36] Gerhard, A., Pavese, N., Hotton, G., Turkheimer, F., Es, M., Hammers, A., Eggert, K., Oertel, W., Banati, R. B., and Brooks, D. J. In vivo imaging of microglial activation with [11c](r)-pk11195 pet in idiopathic parkinson's disease. *Neurobiology of disease 21*, 2 (2006), 404–412.

[37] Gray, C. M., Maldonado, P. E., Wilson, M., and McNaughton, B. Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex. *Journal of neuroscience methods 63*, 1 (1995), 43–54.

[38] Greger, K., Swoger, J., and Stelzer, E. Basic building units and properties of a fluorescence single plane illumination microscope. *Review of Scientific Instruments 78*, 2 (2007), 023705.

[39] Gug, S., Charon, S., Specht, A., Alarcon, K., Ogden, D., Zietz, B., Léonard, J., Haacke, S., Bolze, F., Nicoud, J.-F., et al. Photolabile glutamate protecting group with high one-and two-photon uncaging efficiencies. *ChemBioChem 9*, 8 (2008), 1303–1307.

[40] Helmy, A., Antoniades, C. A., Guilfoyle, M. R., Carpenter, K. L., and Hutchinson, P. J. Principal component analysis of the cytokine and chemokine response to human traumatic brain injury. *PloS one 7*, 6 (2012), e39677.

[41] Helmy, A., Carpenter, K. L., Menon, D. K., Pickard, J. D., and Hutchinson, P. J. The cytokine response to human traumatic brain injury: temporal profiles and evidence for cerebral parenchymal production. *Journal of Cerebral Blood Flow & Metabolism 31*, 2 (2011), 658–670.

[42] Helmy, A., Carpenter, K. L., Skepper, J. N., Kirkpatrick, P. J., Pickard, J. D., and Hutchinson, P. J. Microdialysis of cytokines: methodological considerations, scanning electron microscopy, and determination of relative recovery. *Journal of neurotrauma 26*, 4 (2009), 549–561.

[43] Henry, C. H., Blonder, G., and Kazarinov, R. Glass waveguides on silicon for hybrid optical packaging. *Journal of lightwave technology 7*, 10 (1989), 1530–1539.

[44] Henry, M., Walavalkar, S., Homyk, A., and Scherer, A. Alumina etch masks for fabrication of high-aspect-ratio silicon micropillars and nanopillars. *Nanotechnology 20*, 25 (2009), 255305.

[45] Herbaugh, A. W., and Stenken, J. A. Antibody-enhanced microdialysis collection of ccl2 from rat brain. *Journal of neuroscience methods 202*, 2 (2011), 124–127.

[46] Hodgkin, A. L., and Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology 117*, 4 (1952), 500–544.

[47] Hosseini, A., Kwong, D. N., Zhang, Y., Subbaraman, H., Xu, X., and Chen, R. T. 1× n multimode interference beam splitter design techniques for on-chip optical interconnections. *IEEE Journal of Selected Topics in Quantum Electronics 17*, 3 (2011), 510–515.

[48] Im, M., Cho, I.-J., Wu, F., Wise, K. D., and Yoon, E. Neural probes integrated with optical mixer/splitter waveguides and multiple stimulation sites. In *Micro Electro Mechanical Systems (MEMS), 2011 IEEE 24th International Conference on* (2011), IEEE, pp. 1051–1054.

[49] Jerome, J., and Heck, D. H. The age of enlightenment: evolving opportunities in brain research through optical manipulation of neuronal activity. *Frontiers in systems neuroscience 5* (2011), 95.

[50] John, J., Li, Y., Zhang, J., Loeb, J. A., and Xu, Y. Microfabrication of 3d neural probes with combined electrical and chemical interfaces. *Journal of Micromechanics and Microengineering 21*, 10 (2011), 105011.

[51] Johnson, M. D., Franklin, R. K., Gibson, M. D., Brown, R. B., and Kipke, D. R. Implantable microelectrode arrays for simultaneous electrophysiological and neurochemical recordings. *Journal of neuroscience methods 174*, 1 (2008), 62–70.

[52] Jones, K. E., Campbell, P. K., and Normann, R. A. A glass/silicon composite intracortical electrode array. *Annals of biomedical engineering 20*, 4 (1992), 423–437.

[53] Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydin, Ç., et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature 551*, 7679 (2017), 232.

[54] Kafar, A., Stańczyk, S., Grzanka, S., Czernecki, R., Leszczyński, M., Suski, T., and Perlin, P. Cavity suppression in nitride based superluminescent diodes. *Journal of Applied Physics 111*, 8 (2012), 083106.

[55] Kopp, F., Lermer, T., Eichler, C., and Strauss, U. Cyan superluminescent light-emitting diode based on ingan quantum wells. *Applied Physics Express 5*, 8 (2012), 082105.

[56] Kuo, J. T., Kim, B. J., Hara, S. A., Lee, C. D., Gutierrez, C. A., Hoang, T. Q., and Meng, E. Novel flexible parylene neural probe with 3d sheath structure for enhancing tissue integration. *Lab on a Chip 13*, 4 (2013), 554–561.

[57] Kurt J. Lesker Co. Aluminum oxide (al2o3) pieces evaporation materials, 2018.

[58] Lee, H. J., Son, Y., Kim, J., Lee, C. J., Yoon, E.-S., and Cho, I.-J. A multichannel neural probe with embedded microfluidic channels for simultaneous in vivo neural recording and drug delivery. *Lab on a Chip 15*, 6 (2015), 1590–1597.

[59] Levinson, H., et al. Spie handbook of microlithography, micromachining and microfabrication. *The Society for Photo-optical Instrumentation Engineers, United States of America* (1997).

[60] Lewis, S. Resists: From tip to chip. Presented at the KNI Lecture Series, California Institute of Technology.

[61] Lewis, S., Haynes, V., Wheeler-Jones, R., Sly, J., Perks, R. M., and Piccirillo, L. Surface characterization of poly (methylmethacrylate) based nanocomposite thin films containing al2o3 and tio2 nanoparticles. *Thin solid films 518*, 10 (2010), 2683–2687.

[62] Li, P.-Y., Shih, J., Lo, R., Saati, S., Agrawal, R., Humayun, M. S., Tai, Y.-C., and Meng, E. An electrochemical intraocular drug delivery device. *Sensors and Actuators A: Physical 143*, 1 (2008), 41–48.

[63] Lin, M. Z., and Schnitzer, M. J. Genetically encoded indicators of neuronal activity. *Nature neuroscience 19*, 9 (2016), 1142.

[64] Luchansky, M. S., Washburn, A. L., Martin, T. A., Iqbal, M., Gunn, L. C., and Bailey, R. C. Characterization of the evanescent field profile and bound mass sensitivity of a label-free silicon photonic microring resonator biosensing platform. *Biosensors and Bioelectronics 26*, 4 (2010), 1283–1291.

[65] Mack, C. *Fundamental principles of optical lithography: the science of microfabrication.* John Wiley & Sons, 2008.

[66] Mahendran, R., Malaisamy, R., and Mohan, D. R. Cellulose acetate and polyethersulfone blend ultrafiltration membranes. part i: Preparation and characterizations. *Polymers for Advanced Technologies 15*, 3 (2004), 149–157.

[67] Marcatili, E., and Miller, S. Improved relations describing directional control in electromagnetic wave guidance. *Bell System Technical Journal 48*, 7 (1969), 2161–2188.

[68] Marcatili, E. A. Dielectric rectangular waveguide and directional coupler for integrated optics. *Bell System Technical Journal 48*, 7 (1969), 2071–2102.

[69] Mellergård, P., Åneman, O., Sjögren, F., Pettersson, P., and Hillman, J. Changes in extracellular concentrations of some cytokines, chemokines, and neurotrophic factors after insertion of intracerebral microdialysis catheters in neurosurgical patients. *Neurosurgery 62*, 1 (2008), 151–158.

[70] Melloni, A., Costa, R., Monguzzi, P., and Martinelli, M. Ring-resonator filters in silicon oxynitride technology for dense wavelength-division multiplexing systems. *Optics letters 28*, 17 (2003), 1567–1569.

[71] Meng, E., and Tai, Y.-C. Parylene etching techniques for microfluidics and biomems. In *Micro Electro Mechanical Systems, 2005. MEMS 2005. 18th IEEE International Conference on* (2005), IEEE, pp. 568–571.

[72] Meyer, D., Downey, B., Bass, R., Katzer, D., and Binari, S. 40 nm t-gate process development using zep reflow. CS MANTECH Conference.

[73] Minsky, M. Memoir on inventing the confocal scanning microscope. *Scanning 10*, 4 (1988), 128–138.

[74] Miyawaki, A., Llopis, J., Heim, R., McCaffery, J. M., Adams, J. A., Ikura, M., and Tsien, R. Y. Fluorescent indicators for ca 2+ based on green fluorescent proteins and calmodulin. *Nature 388*, 6645 (1997), 882.

[75] Najafi, K., Wise, K., and Mochizuki, T. A high-yield ic-compatible multichannel recording array. *IEEE Transactions on Electron Devices 32*, 7 (1985), 1206–1211.

[76] Nakai, J., Ohkura, M., and Imoto, K. A high signal-to-noise ca 2+ probe composed of a single green fluorescent protein. *Nature biotechnology 19*, 2 (2001), 137.

[77] Nguyen, Q. P. Characterization of biocompatible parylene-c coating for biomems applications. Master's thesis, Louisiana State University, 2011.

[78] Noguchi, J., Nagaoka, A., Watanabe, S., Ellis-Davies, G. C., Kitamura, K., Kano, M., Matsuzaki, M., and Kasai, H. In vivo two-photon uncaging of glutamate revealing the structure–function relationships of dendritic spines in the neocortex of adult mice. *The Journal of physiology 589*, 10 (2011), 2447–2457.

[79] Nojiri, K. *Dry Etching Technology for Semiconductors.* Springer, 2012.

[80] Oskooi, A. F., Roundy, D., Ibanescu, M., Bermel, P., Joannopoulos, J. D., and Johnson, S. G. Meep: A flexible free-software package for electromagnetic simulations by the fdtd method. *Computer Physics Communications 181*, 3 (2010), 687–702.

[81] Pettit, D. L., Wang, S. S.-H., Gee, K. R., and Augustine, G. J. Chemical two-photon uncaging: a novel approach to mapping glutamate receptors. *Neuron 19*, 3 (1997), 465–471.

[82] Planchon, T. A., Gao, L., Milkie, D. E., Davidson, M. W., Galbraith, J. A., Galbraith, C. G., and Betzig, E. Rapid three-dimensional isotropic imaging of living cells using bessel beam plane illumination. *Nature methods 8*, 5 (2011), 417.

[83] Plock, N., and Kloft, C. Microdialysistheoretical background and recent implementation in applied life-sciences. *European Journal of Pharmaceutical Sciences 25*, 1 (2005), 1–24.

[84] Polikov, V. S., Tresco, P. A., and Reichert, W. M. Response of brain tissue to chronically implanted neural electrodes. *Journal of neuroscience methods 148*, 1 (2005), 1–18.

[85] Pongrácz, A., Fekete, Z., Márton, G., Bérces, Z., Ulbert, I., and Fürjes, P. Deep-brain silicon multielectrodes for simultaneous in vivo neural recording and drug delivery. *Sensors and Actuators B: Chemical 189* (2013), 97–105.

[86] Prasher, D. C., Eckenrode, V. K., Ward, W. W., Prendergast, F. G., and Cormier, M. J. Primary structure of the aequorea victoria green-fluorescent protein. *Gene 111*, 2 (1992), 229–233.

[87] Rabus, D. G. *Integrated ring resonators.* Springer, 2007.

[88] Reddy, G. D., and Saggau, P. Fast three-dimensional laser scanning scheme using acousto-optic deflectors. *Journal of biomedical optics 10*, 6 (2005), 064038.

[89] Renshaw, B., Forbes, A., and Morison, B. Activity of isocortex and hippocampus: electrical studies with micro-electrodes. *Journal of Neurophysiology 3*, 1 (1940), 74–105.

[90] Rios, G., Lubenov, E. V., Chi, D., Roukes, M. L., and Siapas, A. G. Nanofabricated neural probes for dense 3-d recordings of brain activity. *Nano letters 16*, 11 (2016), 6857–6862.

[91] Robinson, A., and Lawson, R. *Materials and Processes for Next Generation Lithography*, vol. 11. Elsevier, 2016.

[92] Ronzitti, E., Ventalon, C., Canepari, M., Forget, B. C., Papagiakoumou, E., and Emiliani, V. Recent advances in patterned photostimulation for optogenetics. *Journal of Optics 19*, 11 (2017), 113001.

[93] Royer, S., Zemelman, B. V., Barbic, M., Losonczy, A., Buzsáki, G., and Magee, J. C. Multi-array silicon probes with integrated optical fibers: light-assisted perturbation and recording of local neural circuits in the behaving animal. *European Journal of Neuroscience 31*, 12 (2010), 2279–2291.

[94] Sanchis, P., Villalba, P., Cuesta, F., Håkansson, A., Griol, A., Galán, J. V., Brimont, A., and Martí, J. Highly efficient crossing structure for silicon-on-insulator waveguides. *Optics letters 34*, 18 (2009), 2760–2762.

[95] Santhanam, G., Ryu, S. I., Byron, M. Y., Afshar, A., and Shenoy, K. V. A high-performance brain–computer interface. *nature 442*, 7099 (2006), 195–198.

[96] Segev, E., Reimer, J., Moreaux, L. C., Fowler, T. M., Chi, D., Sacher, W. D., Lo, M., Deisseroth, K., Tolias, A. S., Faraon, A., et al. Patterned photostimulation via visible-wavelength photonic probes for deep brain optogenetics. *Neurophotonics 4*, 1 (2016), 011002.

[97] Seidl, K., Spieth, S., Herwik, S., Steigert, J., Zengerle, R., Paul, O., and Ruther, P. In-plane silicon probes for simultaneous neural recording and drug delivery. *Journal of Micromechanics and Microengineering 20*, 10 (2010), 105006.

[98] Seshan, K. *Handbook of thin film deposition.* William Andrew, 2012.

[99] Seymour, J. P., and Kipke, D. R. Neural probe design for reduced tissue encapsulation in cns. *Biomaterials 28*, 25 (2007), 3594–3607.

[100] Shimomura, O., Johnson, F. H., and Saiga, Y. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea. *Journal of Cellular Physiology 59*, 3 (1962), 223–239.

[101] Siegel, M. S., and Isacoff, E. Y. A genetically encoded optical probe of membrane voltage. *Neuron 19*, 4 (1997), 735–741.

[102] Signore, A., Mather, S., Piaggio, G., Malviya, G., and Dierckx, R. Molecular imaging of inflammation/infection: nuclear medicine and optical imaging agents and methods. *Chemical reviews 110*, 5 (2010), 3112–3145.

[103] Simeral, J., Kim, S.-P., Black, M., Donoghue, J., and Hochberg, L. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *Journal of neural engineering 8*, 2 (2011), 025027.

[104] Sineshchekov, O. A., Jung, K.-H., and Spudich, J. L. Two rhodopsins mediate phototaxis to low-and high-intensity light in chlamydomonas reinhardtii. *Proceedings of the National Academy of Sciences 99*, 13 (2002), 8689–8694.

[105] Smit, M. K. Progress in awg design and technology. In *Fibres and Optical Passive Components, 2005. Proceedings of 2005 IEEE/LEOS Workshop on* (2005), IEEE, pp. 26–31.

[106] Smit, M. K., and Van Dam, C. Phasar-based wdm-devices: Principles, design and applications. *IEEE Journal of selected topics in quantum electronics 2*, 2 (1996), 236–250.

[107] Smith, D. L. *Thin-Film Deposition: Principles and Practice.* McGraw-Hill, 1995.

[108] Son, Y., Lee, H. J., Kim, J., Shin, H., Choi, N., Lee, C. J., Yoon, E.-S., Yoon, E., Wise, K. D., Kim, T. G., et al. In vivo optical modulation of neural signals using monolithically integrated two-dimensional neural probe arrays. *Scientific reports 5* (2015), 15466.

[109] Spieth, S., Brett, O., Seidl, K., Aarts, A., Erismis, M., Herwik, S., Trenkle, F., Tätzner, S., Auber, J., Daub, M., et al. A floating 3d silicon microprobe array for neural drug delivery compatible with electrical recording. *Journal of Micromechanics and Microengineering 21*, 12 (2011), 125001.

[110] Squires, T. M., Messinger, R. J., and Manalis, S. R. Making it stick: convection, reaction and diffusion in surface-based biosensors. *Nature biotechnology 26*, 4 (2008), 417–426.

[111] Stark, E., Koos, T., and Buzsáki, G. Diode probes for spatiotemporal optical control of multiple neurons in freely moving animals. *Journal of neurophysiology 108*, 1 (2012), 349–363.

[112] Subrebost, G. L. Silicon-based microdialysis chip with integrated fraction collection and biofouling control. *Subrebost Ph. D. Thesis, The Robotics Institute, Carnegie Mellon University* (2005).

[113] Sugawara, M. *Plasma Etching: Funtamentals and Applications.* Oxford Press, 1998.

[114] Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S., and Watts, M. R. Large-scale nanophotonic phased array. *Nature 493*, 7431 (2013), 195.

[115] Szarowski, D., Andersen, M., Retterer, S., Spence, A., Isaacson, M., Craighead, H., Turner, J., and Shain, W. Brain responses to micro-machined silicon devices. *Brain research 983*, 1-2 (2003), 23–35.

[116] Takada, K., Abe, M., Shibata, T., and Okamoto, K. 10-ghz-spaced 1010-channel tandem awg filter consisting of one primary and ten secondary awgs. *IEEE Photonics Technology Letters 13*, 6 (2001), 577–578.

[117] Thornton, J., and McGuire, G. Semiconductor materials and process technology handbook. *Noyes 329* (1988).

[118] Troyer, K. P., Heien, M. L., Venton, B. J., and Wightman, R. M. Neurochemistry and electroanalytical probes. *Current opinion in chemical biology 6*, 5 (2002), 696–703.

[119] Tseng, T. T.-C., and Monbouquette, H. G. Implantable microprobe with arrayed microsensors for combined amperometric monitoring of the neurotransmitters, glutamate and dopamine. *Journal of Electroanalytical Chemistry 682* (2012), 141–146.

[120] Ulrich, R., and Kamiya, T. Resolution of self-images in planar optical waveguides. *JOSA 68*, 5 (1978), 583–592.

[121] Van Acoleyen, K. *Nanophotonic beamsteering elements using silicon technology for wireless optical applications.* PhD thesis, Ghent University, 2012.

[122] Van Acoleyen, K., Bogaerts, W., and Baets, R. Two-dimensional dispersive off-chip beam scanner fabricated on silicon-on-insulator. *IEEE photonics technology letters 23*, 17 (2011), 1270–1272.

[123] Vasicek, T. W., Jackson, M. R., Poseno, T. M., and Stenken, J. A. In vivo microdialysis sampling of cytokines from rat hippocampus: comparison of cannula implantation procedures. *ACS chemical neuroscience 4*, 5 (2013), 737–746.

[124] Vetter, R. J., Williams, J. C., Hetke, J. F., Nunamaker, E. A., and Kipke, D. R. Chronic neural recording using silicon-substrate microelectrode arrays implanted in cerebral cortex. *IEEE transactions on biomedical engineering 51*, 6 (2004), 896–904.

[125] Voie, A., Burns, D., and Spelman, F. Orthogonal-plane fluorescence optical sectioning: Three-dimensional imaging of macroscopic biological specimens. *Journal of microscopy 170*, 3 (1993), 229–236.

[126] Wang, J., Wagner, F., Borton, D. A., Zhang, J., Ozden, I., Burwell, R. D., Nurmikko, A. V., van Wagenen, R., Diester, I., and Deisseroth, K. Integrated device for combined optical neuromodulation and electrical recording for chronic in vivo applications. *Journal of neural engineering 9*, 1 (2011), 016001.

[127] Wassum, K. M., Tolosa, V. M., Wang, J., Walker, E., Monbouquette, H. G., and Maidment, N. T. Silicon wafer-based platinum microelectrode array biosensor for near real-time measurement of glutamate in vivo. *Sensors 8*, 8 (2008), 5023–5036.

[128] Wilcox, M., Viola, R. W., Johnson, K. W., Billington, A. P., Carpenter, B. K., McCray, J. A., Guzikowski, A. P., and Hess, G. P. Synthesis of photolabile precursors of amino acid neurotransmitters. *The Journal of Organic Chemistry 55*, 5 (1990), 1585–1589.

[129] Williams, K. R., and Muller, R. S. Etch rates for micromachining processing. *Journal of Microelectromechanical systems 5*, 4 (1996), 256–269.

[130] Wilson, M. A., and McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science 261*, 5124 (1993), 1055–1059.

[131] Winslow, B. D., Christensen, M. B., Yang, W.-K., Solzbacher, F., and Tresco, P. A. A comparison of the tissue response to chronically implanted parylene-c-coated and uncoated planar silicon microelectrode arrays in rat cortex. *Biomaterials 31*, 35 (2010), 9163–9172.

[132] Wise, K. D., Angell, J. B., and Starr, A. An integrated-circuit approach to extracellular microelectrodes. *IEEE Transactions on Biomedical Engineering*, 3 (1970), 238–247.

[133] Wu, F., Stark, E., Im, M., Cho, I.-J., Yoon, E.-S., Buzsáki, G., Wise, K. D., and Yoon, E. An implantable neural probe with monolithically integrated dielectric waveguide and recording electrodes for optogenetics applications. *Journal of neural engineering 10*, 5 (2013), 056012.

[134] Wu, F., Stark, E., Ku, P.-C., Wise, K. D., Buzsáki, G., and Yoon, E. Monolithically integrated μleds on silicon neural probes for high-resolution optogenetic studies in behaving animals. *Neuron 88*, 6 (2015), 1136–1148.

[135] Yariv, A. Universal relations for coupling of optical power between microresonators and dielectric waveguides. *Electronics letters 36*, 4 (2000), 321–322.

[136] Yariv, A. Critical coupling and its control in optical waveguide-ring resonator systems. *IEEE Photonics Technology Letters 14*, 4 (2002), 483–485.

[137] Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M., and Deisseroth, K. Optogenetics in neural systems. *Neuron 71*, 1 (2011), 9–34.

[138] Zhang, F., Prigge, M., Beyrière, F., Tsunoda, S. P., Mattis, J., Yizhar, O., Hegemann, P., and Deisseroth, K. Red-shifted optogenetic excitation: a tool for fast neural control derived from volvox carteri. *Nature neuroscience 11*, 6 (2008), 631.

[139] Zhang, F., Wang, L.-P., Boyden, E. S., and Deisseroth, K. Channelrhodopsin-2 and optical control of excitable cells. *Nature methods 3*, 10 (2006), 785.

[140] Zhang, J., Laiwalla, F., Kim, J. A., Urabe, H., Van Wagenen, R., Song, Y.-K., Connors, B. W., Zhang, F., Deisseroth, K., and Nurmikko, A. V. Integrated device for optical stimulation and spatiotemporal electrical recording of neural activity in light-sensitized brain tissue. *Journal of neural engineering 6*, 5 (2009), 055007.

[141] Zhao, H., Brown, P. H., and Schuck, P. On the distribution of protein refractive index increments. *Biophysical journal 100*, 9 (2011), 2309–2317.

[142] Zorzos, A. N., Boyden, E. S., and Fonstad, C. G. Multiwaveguide implantable probe for light delivery to sets of distributed brain targets. *Optics letters 35*, 24 (2010), 4133–4135.

[143] Zucker, R. Effects of photolabile calcium chelators on fluorescent calcium indicators. *Cell calcium 13*, 1 (1992), 29–40.

Appendices

Appendix A

Fabrication theory and practice

A.1 Photolithography

A.1.1 Photoresist Chemistry

The photographic films, or photoresists, that are most common in MEMS processing fall into the category of non-chemically amplified photoresists. These resists consist of three major components: a phenol-formaldehyde resin called Novolac, a diazonaphthoquinone (DNQ) dissolution inhibitor, and a casting solvent. Each of these components plays an integral role in the behavior and properties of the resist. The first component, the solvent, allows the solid polymer components to be cast on a wafer via spin coating, and then removed by a gentle baking process to create a solid film on the surface of the wafer. Most photoresists are dissolved using propylene glycol monomethyl ether acetate, or PGMEA. The viscosity (and thereby the spin-thickness curve of the resist) is controlled by varying the polymer-to-solvent ratio of the resist.

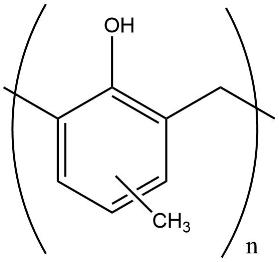

Figure A.1: Novolac polymer.

The novolac resin is a phenol formaldehyde based polymer derived from a mixture of meta- and para-cresol reacted with formaldehyde to form a long-chain polymer with a molecular weight around 20kDa. The novolac resin makes up the bulk of the dry mass of the resist, typically around 80% by mass of the solids in the resist. This compound was chosen in part for photoresists because it is relatively clear in the near UV range used to expose these resists. Structurally, this polymer will have strong etch resistance due to the significant concentration of benzene in the polymer. This is exhibited in the superior etch resistance of ZEP (with an additional styrene group on the main chain) compared to PMMA, as described below.

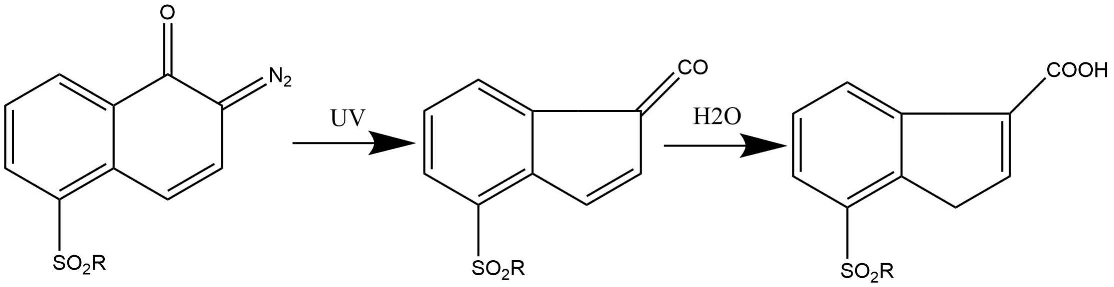

Figure A.2: Photoactivation of DNQ[65].

Although the novolac resin makes up the majority of the solids in the resist, the most critical component is the diazonaphthoquinone dissolution inhibitor. DNQ is typically attached to a ballast R-group (see figure A.2), which improves the solubility of the DNQ in the novolac resin and can be used to improve optical properties of the resist. This R-group is often itself a short chain novolac polymer. DNQ has an interesting property with respect to novolac, in that when present it inhibits the dissolution of the polymer in aqueous basic solvents such as TMAH or potassium borate. The reason for this property is unknown, but has been theorized to have to do with hydrogen bonding within the polymer film. When exposed to light, DNQ loses the azide to nitrogen gas and forms an indene intermediate[65]. Water is then consumed and the compound forms a carboxylic acid, resulting in the final stable state of the compound.

Undergoing this reaction, DNQ undergoes two interesting physical changes. First, the compound bleaches, thereby reducing the absorbance in the UV. This is critical during the exposure process, since as the resist is exposed from top to bottom, the resist that is exposed becomes clear to UV. This is advantageous as it provides a more even exposure of the resist. The other important change that DNQ undergoes after exposure is that it no longer inhibits the dissolution of the novolac resin in basic solution. In fact, after being exposed to UV light, the product of the DNQ-UV reaction increases the dissolution rate of the novolac polymer. Thus, where the resist is exposed to UV, the rate of dissolution increases significantly. Therefore, by soaking the polymer film in a basic developer, the exposed area will be removed while leaving behind the un-exposed photoresist. This is the mechanism of action for these resists.

The key metric for describing the “quality” of a resist is the resists contrast ratio. The contrast ratio may be calculated from a dose exposure experiment, where the resist is exposed to increasing doses of light energy and the resulting thickness of resist after developing is measured. An example

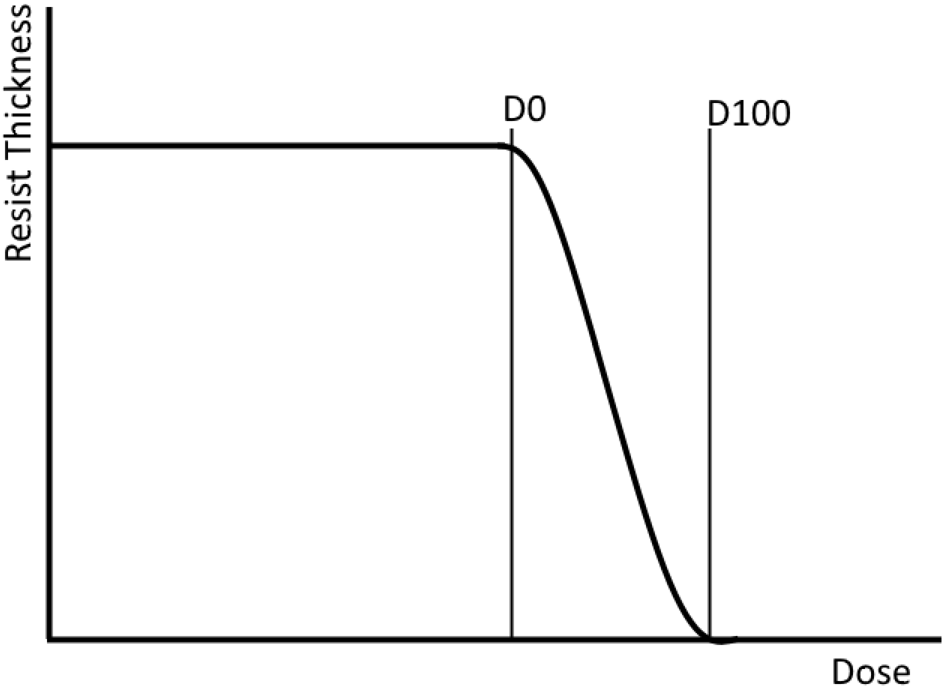

Figure A.3: Dose-Thickness curve for a positive tone resist.

of such a plot is shown in figure A.3. The contrast ratio is calculated as

$$\gamma = \frac{1}{\log(D_{100}/D_0)}, \tag{A.1}$$

where γ is the contrast, D_{100} is the dose where there is 100% clearance of the resist, and D_0 is the does at which the resist begins to develop away. Typically the contrast ratio for a high quality novolac-DNQ resist is on the range of 1-2, although chemically amplified resists can have contrast up to 7.

A.2 Electron Beam Lithography

A.2.1 Theory

A.2.1.1 Lithography tool structure

The overall function of the electron beam column is described in detail in [59] and summarized below. Electron beam lithography requires the generation of a tightly focused electron beam which is painted across a surface to accomplish lithographic patterning. The most common beam type (which is created by both tools used in this thesis) is a Gaussian profile beam. Column arrangement is nearly identical to that of a scanning electron microscope, however the tool is designed for higher energy electrons (which reduce proximity effect, as described below). The column is evacuated to high vacuum to minimize scattering of the beam within the column. To ensure tight focusing, the electrons are produced from a Mller-type, sharply pointed emitter. The sharpness of the emitter guarantees small beam spot sizes at the end of the focusing column. The emitter is exposed to a high electric field inducing field emission in the small chip. Then, the electrons are accelerated under a 100kV electric field to produce a very high energy electron beam.

Upon leaving the accelerator, the electron beam, similar in many ways to an optical microscope, will undergo manipulation by a number of electromagnetic lenses. However, unlike with light, these lenses are created using coils of wire inside of an iron ferrule, which are used to focus the magnetic field lines. The first lens is used to align the beam so that it propagates down the center of the column. Following these coils, the electron beam passes through a series of two lenses intended to act as a condenser. These two lenses are adjusted to determine the illumination of the objective lens, which essentially sets the beam current propagating through the column. After propagating through the condenser, electrons are exposed to the beam blanking electrodes which steers the beam into a blanking body when the beam needs to be turned off. Following blanking, the beam propagates through the final aperture, typically 300μm, which only allows properly focused electrons to propagate, while blocking any out of focus electrons. This is very similar to the setup used in a confocal microscope. Finally, the beam is sent through the objective lens which focuses the beam onto the substrate. Other lenses are utilized here as well to efficiently scan the beam over the surface of the sample without having to move the sample stage.

Due to the geometry of the Gaussian beam and the column, there is a direct relationship between the beam current and the spot size of the Gaussian beam at the end of the column. Thus, if a tight focus is required for high resolution writing, the beam current must also be very small (100-500pA beam currents produce 3-5nm beam diameters). Whereas for coarse writing, it is acceptable to use a higher beam current (50+nA beams produce 50nm spot sizes). Unfortunately this relationship limits write times by the pattern generator as high resolution writing requires low beam currents. Write speed is determined by the simple equation,

$$t = \frac{DA}{I}, \tag{A.2}$$

where t is the dwell time of the beam at a particular point, D is the dose, A is the spot size area, and I is the beam current. This is often also expressed as beam frequency, which is simply the reciprocal of t, in Hz. Therefore, the only way to increase write speed practically is to increase the beam current, as A is defined by the beam current and D is a property of the resist.

A.2.1.2 Physical mechanism of lithography

Similar to that of photolithography, electron beam lithography resists require sufficient external energy delivered to the resist to create a physical change in the polymer. Typically the easiest way to do this is to increase the solubility of the polymer in a specific solvent. Although, as discussed previously, novolac polymers utilize a photosensitive compound (DNQ) which, when exposed to light, changes structure which changes the solubility of the resist. Unfortunately, there isn't as elegant of an alternative under exposure to electronic radiation. Thus electron activated polymers require the direct scission of their main chain to increase solubility. The process by which this occurs is simple. By starting with a very long polymer (typically 4000-10000 monomers long), polymeric chains will easily become entangled when cast onto a wafer. Thus, solubility of the film will be low due to steric hindrance. If a region of that film, however, is broken down in some way to much smaller subunits, the solubility in that area will be much higher than the un-exposed regions. Therefore, by using an electron beam to break apart the backbone of the polymer, the solubility in that area will be increased and will provide a mechanism for development of the pattern.

Electron beams are able to create energies much larger than the carbon-carbon bond energy (3.6eV). Typically this occurs in PMMA (see figure 3.2) when the tertiary carbon on the main chain is exposed to radiation (in this case, high energy electrons). The bond between the tertiary carbon and the carbon bonded to the carbonyl becomes unstable, breaking apart and creating a radical at the tertiary carbon. During rapid bond rearrangement, the main chain of the polymer is broken, inducing the main chain scission event described previously.

Propagating electrons interact with their medium using a number of mechanisms, primarily through scattering processes (described in detail in [60] and summarized below). Elastic scattering is described by Rutherford scattering, as electrons are charged particles. The Rutherford cross section for an electron in a lithography tool is described as

$$\sigma_{\text{elastic}} = 5.21 \times 10^{-21} \left(\frac{Z^2}{E^2}\right) \left(\frac{4\pi}{\alpha(1+\alpha)}\right) \left(\frac{E+511}{E+1024}\right)^2, \tag{A.3}$$

where E is the electron energy in keV, Z is the atomic number of the material, α is the screening factor, where

$$\alpha = 3.4 \times 10^{-3} \frac{Z^{.67}}{E}. \tag{A.4}$$

It is also useful to calculate the mean free path of an electron in a material,

$$\lambda = \frac{A}{N_a \rho \sigma_{\text{elastic}}}. \tag{A.5}$$

At high energies, elastic scattering dominates other scattering processes. This is a desirable effect when doing lithography as elastic scattering will typically only deviate the path of an electron insignificantly, which will in turn maintain a straight pathway through the material.

On the other hand, inelastic scattering is also common, especially at low beam energies. Inelastic scattering produces additional electrons by removing them from the chemical structure that the electron beam is propagating through. These electrons are called secondary electrons, and have some interesting properties. The generation of secondary electrons is described by the following equation:

$$\frac{d\sigma}{d\Omega} = \left(\frac{\pi e^4}{E^2}\right)\left(\frac{1}{\Omega} + \frac{1}{(1-\Omega)^2}\right), \tag{A.6}$$

where ΩE is the energy of the secondary electron produced (i.e., Ω is the share of energy taken from the primary electron and given to the secondary electron). The emission angle of the primary electron (α) and the secondary electron (γ) are defined by conservation of momentum as

$$sin^2(\alpha) \quad = \tfrac{2\Omega}{2+t-t\Omega}, \tag{A.7}$$

$$sin^2(\gamma) \quad = \tfrac{2(1-\Omega)}{2+t\Omega}, \tag{A.8}$$

where t is the kinetic energy of the electron. By consequence of the energy of the primary electrons being produced by the tool (100keV), secondary electrons are emitted on average at an 80^o angle from the beam axis. This creates a cascade of lateral scattering events, as secondary electrons have less energy than primary electrons, and thus are more likely to create more secondary electrons by inelastic scattering. This generates what is called the proximity effect, where the size of written features is larger than the beam diameter due to lateral scattering of secondary electrons. This is minimized by using high beam energies, which ensures elastic scattering dominates over inelastic scattering.

Using these two equations, plus the energy transfer equation (stopping power of the material), it is possible to predict the behavior of materials and resists when exposed to an electron beam. This is typically done using Monte Carlo statistical simulations. Furthermore, these simulations are helpful in predicting and reducing proximity effects when creating structures with small gaps between features.

A.3 Pattern Transfer: Plasma Etching

A.3.1 Theory

The study of plasma physics is interesting in its own right as a separate state of matter. The technical definition of a plasma is that a plasma is a quasineutral gas of charged and neutral particles which exhibits collective behavior; however this definition belies the functional aspects of the plasma and the chemical properties of a gas where electrons are separated from a majority of the rest of the matter in the gas. There are many different types of plasmas (thermal, arc, flames, etc.), but for semiconductor processing we are entirely concerned with the class of glow discharge plasmas. These plasmas are created when an electrical current is passed through a rarefied gas, wherein electrons are separated from their nuclei and are conducted within the electric field established in the plasma chamber. This is similar to the idea of electrons moving in a semiconducting material, and in fact the electrical properties of a plasma are very similar to that of a diode.

As was stated in the previous paragraph, a plasma must act as a quasineutral entity and exhibit collective behavior. Thus, the physical properties of the charged gas cloud can be defined in physical terms. The first critical parameter to understand when considering plasmas is the Debye screening length, defined in the case of a gas as

$$\lambda_D = \frac{\epsilon_0 k_B / q_e^2}{n_e/T_e + \sum_{ij} j^2 n_{ij}/Ti} \approx \frac{\epsilon_0 k_B T_e}{n_e q_e^2} \tag{A.9}$$

where ϵ_0 is the permittivity of free space, k_B is Boltzmann's constant, q_e is the charge of an electron, n_e is the number density of free electrons in the gas, T_e is the "temperature" of the electrons in the gas, n_{ij} is the number density of the i'th ionic species with charge j, and T_{ij} is the "temperature" of that species in the gas. The temperature of an electron or an ion is a measure of the energy of an electron or ion, and is calculated as $T = (2/3)E/k_B$. It turns out that typically in plasmas the energy of electrons is much higher than that of the ions because electrons and ions are accelerated with the same force in the electric field of a glow discharge plasma, however the ions are heavier and thus gain less energy. Thus the electronic term dominates and we get the approximation on the far right of the equation. For all intents and purposes this approximation is sufficient for all rarefied plasmas.

The Debye length describes the ability of charges to be surrounded (or screened) by charges of the opposite type due to electrostatic attraction. This is a common occurrence around electrodes in electrochemical experiments where a positively charged electrode will be surrounded by negatively charged ions and vice versa. In the case of a charged gas the same thing will occur, where a positive ion might be screened by electrons in the gas. This phenomenon is called screening (or alternatively dampening), where an electric field is observed to be much smaller than it should be due to the surrounding mobile electronic charges. The consequence of this screening property is that when viewed from far enough away, the charge of the central ion of interest will appear to lose its apparent charge due to the charges of opposite polarity surrounding it. The screening "length" is the distance away from the charged object where the apparent charge is a certain amount smaller than the actual charge (in this case a single exponential fold).

Recall that from the definition of the plasma, the gas must be "quasineutral." Therefore, from the screening length, we know that if a single particle is observed far enough away, the particle will appear neutral. When we consider the entire ensemble of particles in the gas, we can define the first rule of definition for a plasma, where if the debye length is significantly smaller than the size of the plasma (L), then the entire plasma must appear neutral. Mathematically, this is written as

$$\lambda_D << L. \tag{A.10}$$

To address the second part of the definition of a plasma, wherein the plasma must exhibit collective behavior, the number of particles which interact with each other must be understood. Imagine a scenario where an ionized gas is so sparse that particles are not significantly interacting with each other (i.e., the distance between particles is much larger than the Debye length, λ_D). In this case if an external electric field is applied to the ionized gas, each particle will act independently to move within the electric field. In other words, the each particle will experience a force proportional to the electric field by itself. This behavior, is independent and thus the gas will not exhibit the collective behavior described in the definition of a plasma. Thus, this thought experiment informs that the density of the plasma must be such that particles are interacting more with their local environment than the external field. This distance is measured by the Debye length, and thus we want the density of the majority carrier (electrons in this case) in the center of the plasma to be great enough such that multiple electrons and ions are interacting with each other and have effectively screened out the external influence of the electric field. That's not to say a plasma won't move in an electric field, but it should be moving as an ensemble, not as individual particles. This criterion is defined as the plasma parameter (Λ),

$$\Lambda = V_{Debye} \times n_e = \frac{4}{3}\pi\lambda_D^3 n_e >> 1, \tag{A.11}$$

where V_{Debye} is the sphere defined by the Debye length, and n_e is the number density of electrons in the plasma. This calculates the number of charged particles within a single Debye length of another particle. This must be much greater than one to provide the collective behavior described in the definition of the plasma. In other words, if Λ is large, there are many charged particles within the Debye sphere of a single chosen particle, and they are close enough that they feel the influence of their surrounding particles predominantly instead of the influence of a surrounding electric field.

These two factors ensure that the definition of the plasma is satisfied. However, there is another factor that is commonly considered with a plasma as well, and that is that the dynamics of the plasma are dominated by electrostatic forces as opposed to the gas kinetics of the plasma. The plasma frequency is a measure of the timescale over which the plasma responds to a change in the electrostatic environment. It is often also defined as the maximum frequency that can pass through a plasma without attenuation. The plasma frequency for electrons (the fastest responder to electromagnetic energy) is defined as

$$\omega_p = \left(\frac{n_e e^2}{\epsilon_0 m}\right)^{1/2} \tag{A.12}$$

Thus, if ω_p is faster than the electron-neutral collision frequency, the plasma should be dominated by the applied electromagnetic forces to the plasma as a whole and not the collisional kinetics of the gas itself.

A.3.1.1 Plasma Chemistry

The development (or “strike”) of a plasma is often established in a rarefied gas with an applied electric field. Electrons are separated from their host atoms (creating positive ions), and if the velocity of electrons is high enough collisions between neutrals and electrons occur. This creates predominately negative ions and free radicals, the latter of which are very important in etch processing. If the electric field is great enough, an avalanche process can begin, in which when electrons ionize neutral particles more electrons are created, each of which can collide with another neutral, and so on until the plasma is created.

Due to the ionization of gas molecules and the collisions that occur within a plasma, the chemical composition of plasmas is rich and varied. The major constituents of a plasma include the neutral gas, electrons, positive and negative ions, and free radicals. Each of these species have a distinct effect on the properties of a plasma and the to what happens when an object (such as a wafer) is placed within the plasma itself. The general composition of plasmas can be divided into two broad categories,

- Cold Plasmas (e.g., glow discharge plasmas): These plasmas have a low degree of ionization and are usually created in a high voltage environment. They have low current densities where most of the energy is carried by electrons.

- Hot Plasmas (e.g., arc plasmas): These plasmas have an extremely high degree of ionization and are usually created in a high current environment. Most of the energy in a ”hot” plasma are carried by positive and negative ions.

The plasmas used in semiconductor processing are typically cold plasmas (as they are less damaging to the substrate), and thus herein the discussion herein will assume a glow discharge plasma.

Within a glow discharge plasma, the predominant energy carrying particles are electrons due to their relatively small mass (thus gaining more velocity within an electric field). Although electrons are the primary energy carriers, it is possible for electrons to create high energy ions via elastic scattering. However, the most important interactions are the nonlinear interactions between electrons and the other species in the plasma. The predominant chemical reactions are as follows:

- Excitation: $A + e^- \rightarrow A^* + e^- \rightarrow A + e^- + h\nu$
- Ionization: $A + e^- \rightarrow A^+ + 2e^-$
- Dissociation: $AB + e^- \rightarrow A + B + e^-$
- Electron Attachment $A + e^- \rightarrow A^-$

The first two reactions provide the major properties of a glow discharge plasma. Excitation, or an electron-neutral collision which induces an excited state in the neutral molecule, is responsible for the glow that is created by the plasma. This interaction also creates neutrals which are more reactive than they would be in their ground state.

The second reaction described above, ionization, is the key reaction in creating and maintaining the glow discharge plasma. In this reaction, a high energy electron impacts the neutral molecule such that a secondary electron is produced from the atom, leaving a positive ion behind. Since electrons are accelerated in the electric field, both of the remaining electrons gain energy and will have the ability to further ionize other neutral particles. This process, called an electron avalanche, induces the “breakdown” in the gas from a generally neutral collection of molecules into a plasma. Thus, the plasma within an electric field is self-regenerating and generally is stable once started.

The third reaction, dissociation, is very important when considering etching reactions. In this reaction, an electron collides with a molecule and splits that molecule into two constituent molecules. Since the charge of the electron isn't transferred to the new molecules, A and B, they can either be positively charged, negatively charged, or uncharged, so long as the sum of the charges is equal to AB. This reaction is extremely important when considering the chemical reactions involved in the etching process as dissociation creates free radicals. Free radicals are highly reactive molecules due to an un-paired electron existing in its valence band. These free radicals are often one of the major reactive species during the etching process, as well as inducing free radical polymerization of passivation compounds used by the Bosch process (discussed in detail below).

The final reaction, although an important part of any plasma process, is of minimal importance for semiconductor processing due to the creation of heavy weight ions and also the negative charge of these ions.

The resulting free radicals and ions are what drive the etching process. By choosing the appropriate gas, it is possible to utilize the free radicals and ions to react with the surface of the substrate (such as silicon) such that bonds are broken from the bulk substrate and are freed from the surface. When determining a gas chemistry to use, it is important to not only consider that the chemicals in the gas will react with the surface (this is common due to the highly energetic state of the molecules in the plasma), but also that these species will easily evaporate from the surface once freed from covalent bonds.

For example, when deciding what gas to use to etch silicon, it is critical to check the vapor pressure of the molecules created by silicon and the gas reactants. Both fluorine and chlorine are common etch chemistries for many materials, but when etching silicon there is a large difference in the properties of the resulting molecules. $SiCl_4$ reaches a vapor pressure of 10 torr at -34.7^oC. While this enables etching to occur, having such a high vapor temperature is not ideal as it will take some time for the $SiCl_4$ to leave the surface. In contrast, fluorine chemistry results in a product of SiF_4 which has a 10 torr vapor pressure at -130.4^oC. This is far superior to chlorine and thus fluorine is the preferred chemistry for silicon etching. It is important to understand the vapor pressures of the products of etching reactions to understand which gasses will produce an appreciable etch rate at a reasonable temperature. A table of such values for multiple semiconductor and metal compounds is found in chapter 2 of reference [79].

The interplay between etching via free radicals and positive ions is a difficult one to dissect. It is know that free radicals are the primary source of isotropic etching, however modern etch recipes show etch rates much higher than what would be predicted from free radical etching alone. The current theory behind the effect of ions in the etching process is the concept of ion assisted etching, wherein ions bombard the surface of the material being etched where unmasked, and thereby heating the surface. As discussed in the last paragraph, temperature is a key determinant of etch rate and therefore if the unmasked area is being locally heated, the etch rate should increase only where there is ionic bombardment. This effect leads to anisotropic etch recipes that don't require sidewall passivation. Typically, however, these recipes have high forward power and will be tough on polymer resists, requiring the use of hard masks such as aluminum oxide to withstand the ionic bombardment and sputtering that occurs during etches with high forward power. More information on ion assisted etching can be found in references [79] and [113].

A.3.1.2 Capacitively Coupled Plasmas

Capacitively coupled plasmas are the basis of most commercial etching system. Even if the plasma is driven by another source (such as an inductor or RF energy, as described in the following section), there is always a capacitive component to the plasma which brings the reactive species in the plasma to the wafer. A schematic of such an etch chamber is shown in figure A.4. Such a device is created by filling an evacuated chamber with a rarefied gas and applying a large, alternating electric field between the two plates of a capacitor spanning the width of the chamber.

The RF source is coupled to the chamber through a capacitor (as shown in the figure), as it is desirable to bring ions to the surface of the substrate. When considering the electronic properties of a plasma, one must always consider the difference in the mobilities of electrons and ions. Electrons also have the special ability to be emitted from and absorbed by the metallic capacitor plates. Thus, due to the large difference in mobility of the electrons, if a capacitor is added to the substrate side

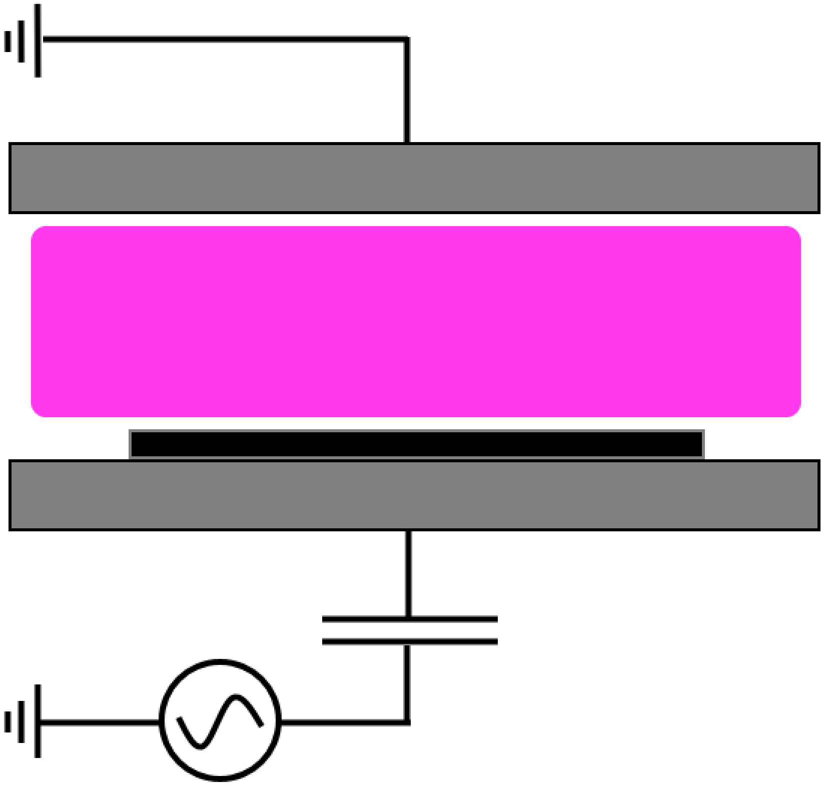

Figure A.4: Capacitively coupled plasma. The plasma is created between two parallel plates, with the wafer (black) placed on top of the plate attached to the external capacitor.

of the etch chamber the high speed electrons will quickly accumulate in the capacitor. Thus, the substrate plate will be negatively charged, which is desirable to attract positively charged ions to the substrate for reactions between the ions and the substrate to occur. The so-called DC bias is created in the reactive ion etcher. The potential of the plasma is shown in figure A.5, showing the large negative potential near the electrode with the connected capacitor. This negative potential attracts positively charged ions to this electrode, which are critical for the etching process. This is shown in figure A.5.

Another important feature of a capacitively coupled plasma is the sheath regions near each of the etcher plates. Due to the varying field in the plasma (when the RF frequency is slower than the plasma frequency) and the slower effect of the field on ions, a sheath is formed between the body of the plasma and electrodes. This is the region in which electrons are accelerated by the electric field but don't have enough energy to excite the electronic state of the gas molecules, and thus the sheath is the region in which there is no glow to the plasma. Returning to figure A.5, the sheaths are the regions nearest the electrodes in which the potential is gradually increasing or decreasing.

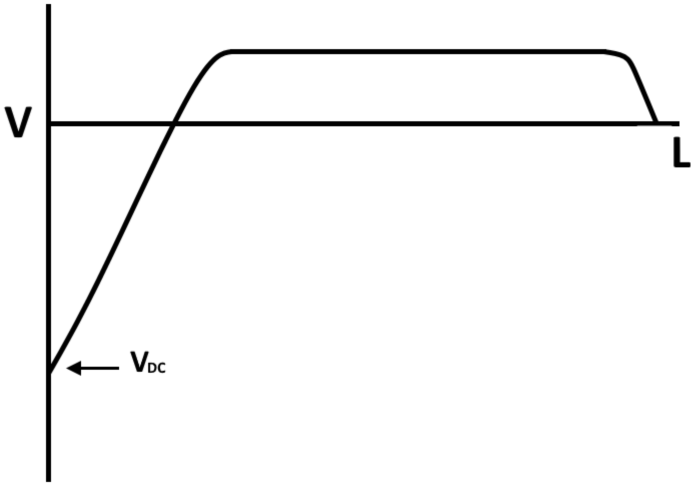

Figure A.5: Plasma potential of a DC coupled plasma. The voltage of the plasma is along the vertical axis. The horizontal axis is the length along the chamber, with $x = 0$ being the substrate plate.

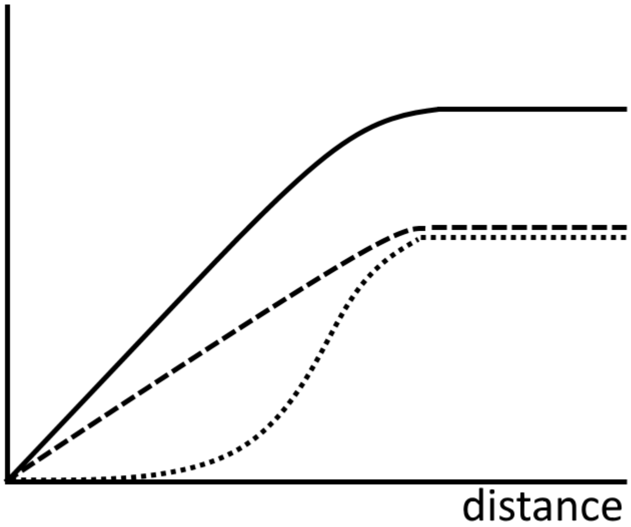

Figure A.6: Plasma potential of a DC coupled plasma zoomed in on the sheath region. The plasma potential is shown as a solid line. The concentration of positive ions is shown as a dashed line, and the concentration of electron is shown as a dashed line

The glow region of the plasma is the neutral region (designated by a flat potential).

The voltage drop between electrodes is proportional to the area of the electrodes. This is an important feature to understand as the electrode on which a wafer is placed is typically smaller than the counter electrode on the opposite side of the chamber due to design constraints. The relationship between voltage and electrode area is as follows:

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1}\right)^4. \tag{A.13}$$

This equation is also critical for determining the magnitude of the DC bias for a plasma, as the sample holder for a typical etcher is typically the size of the wafer and is smaller than the size of the top electrode.

A.3.1.3 Modern Etcher Construction

Modern etcher designs are still currently based around a capacitvely coupled plasma which is used to bring ions to the surface of the material to be etched. However they typically use a different technology to create the plasma itself. By decoupling the plasma generation from the capacitively coupled plasma, it is possible to independently control the plasma density (i.e., the chemistry of the plasma itself) and the degree by which the surface is bombarded by ions. This decoupling is not possible in a strictly capacitively coupled plasma, and is advantageous because it allows for a more versatile and controllable plasma to be created and applied to the substrate being etched.

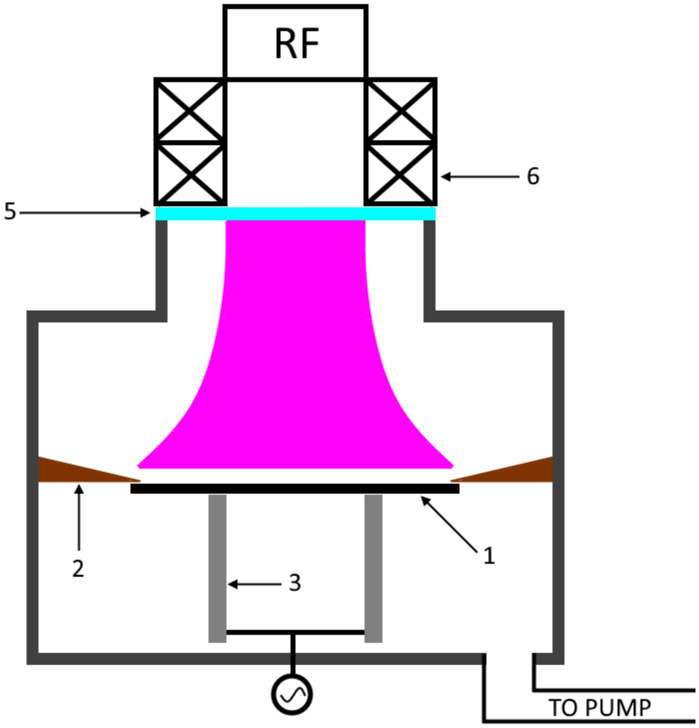

Figure A.7: Electron-cyclotron resonance etcher schematic. 1. Wafer, 2. Quartz wafer clamp, 3. Table, 4. Microwave Source, 5. Quartz Window, 6. Wire solenoid.

One of the most common tools developed for etching silicon is the electron-cyclotron resonance (ECR) etcher, shown in figure A.7. These etchers utilize a pair of air-core electromagents to create a stationary magnetic field within the top region of the etch chamber. Due to the Lorentz force, an electron moving in this field will feel a force perpendicular to its direction of motion, creating a cyclical motion of electrons within the field. To create an electric field which will start the electrons moving within this magnetic field, the etch chamber is exposed to microwaves generated by a magnetron at the top of the chamber. This creates an electric field which is perpendicular to the magnetic field which has been established in the chamber. Thus, electrons will begin to circulate and become accelerated throughout the chamber in the stationary magnetic field. At an optimized point in the magnetic field for the given microwave frequency (where the microwave frequency matches the cyclotron resonance of the electrons), resonance will occur. Energy is transfered to the gas in the form of electron-neutral collisions and thus will generate a plasma.

The advantages of an ECR plasma are that it can sustain a stable plasma at very low pressures (something necessary for ion assisted etch processes), controllable ion energy by setting the RF bias, and a controllable ion-radical influx ratio. These etchers were popular within the semiconductor industry in the 1990's, but fell out of favor due to two major issues. First, the matching network for the ECR chamber would have to be hand optimized by a technician to ensure that the plasma was operating properly. Matching networks were so complex that it was impossible to optimize automatically. The second issue has to do with the consequence of bad network matching. If the matching network wasn't properly optimized, it is possible that the plasma would not absorb all of the energy from the microwaves being emitted by the magnetron. Thus, it was possible that the RF would hit the wafer, and cause it to heat up, thereby changing the properties of the etch and potentially damaging the substrate. Thus, ECR was abandoned by the industry for a safer option, the Inductively Coupled Plasma (ICP) etcher.

The currently preferred technology for etchers is the ICP etcher, shown in figure A.8. The ICP Etcher acts similarly to the competing electron cyclotron resonance (ECR) etcher in that it accelerates electrons using a magnetic field, however the ICP etcher is much simpler in design. A coil is either placed on top of the etch chamber or surrounding the top of the etch chamber and a high current is applied to the coil. The coil will create a magnetic field within the top of the etch chamber, wherein any moving electrons will feel a Lorentz force perpendicular to the magnetic field, circulating the electrons and adding energy to them. As in the ECR plasma, electron-neutral collisions will create the plasma via ionization. The main advantages to the ICP etcher are similar to that of the ECR etcher in that the ICP etcher can easily create plasmas in low pressure environments, control the plasma density independently of the ion flux to the substrate, and results in high etch rates. However ICP is deemed superior due to its simpler design which allows for automatic network matching of the plasma to the etcher's high power electronics, making these tools more user friendly

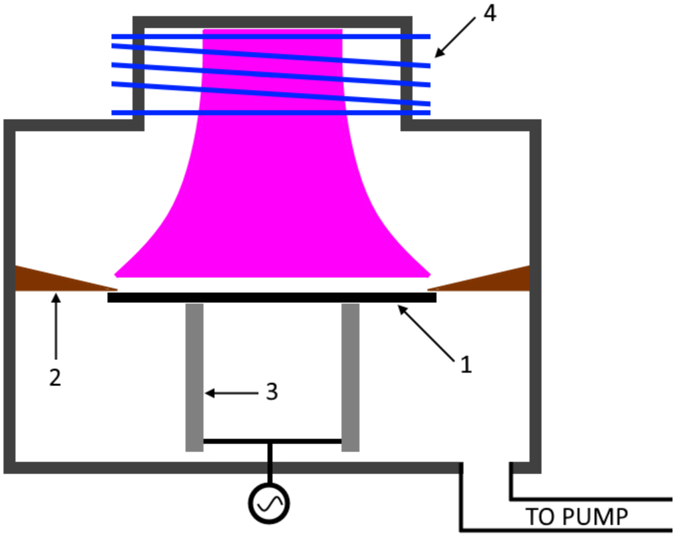

Figure A.8: Inductively coupled plasma etcher schematic. 1. Wafer, 2. Quartz wafer clamp, 3. Table, 4. Inductor coil.

and repeatable in industrial settings. The etchers used in the research detailed in this thesis were all done on Oxford Plasmatherm 380 ICP etchers.

A.3.2 Physical Parameters Affecting Etch Properties

In this section the critical parameters used for tuning etch properties will be discussed. These factors typically manipulate one of the physical properties of the plasma described in the previous sections, such as the properties of the sheath, the plasma density, or the plasma chemistry. It is important to note that these parameters, since they have a physical effect on the plasma, cannot be tuned to their extremes without causing the plasma to fail to strike, and thus tuning plasmas for etch processes requires the use of multiple factors simultaneously.

A.3.2.1 Pressure

The first factor which can be used to tune the properties of a plasma is the pressure of the gas in the chamber. Typically pressure is used to tune the anisotropy of an etch. The reasoning goes that when considering the relationship between mean free path (λ) and pressure, that $\lambda \propto P^{-1}$. Therefore the higher the pressure, the shorter the path length an ion must take before it's original direction is lost to scattering. It is logical that by increasing the pressure in the chamber, the etch will become more isotropic. Inversely, if the pressure is decreased, the etch will become more anisotropic.

Furthermore, as the pressure decreases, due to the minimization of collisions in the sheath region where ions are accelerated at the target, the energy at which ions strike the surface will increase.

This induces a more "mechanical" etch at lower pressures, where sputtering by the ions of the surface material occurs. If the pressure is increased, the only mechanism by which etching may occur is by a chemical process due to the low ion energies.

To give an idea of the range of pressures, a rule of thumb is that a "high" pressure plasma is around 1 torr, and a "low" pressure plasma has a chamber pressure of around 1 mtorr. Typical etching pressures are between these two values, as most modern etch processes use an ion assisted etch process (described in detail in the next section), which typically occurs optimally around 10-50 mtorr.

Changing the pressure also has a significant effect on the properties of the electrode sheaths, as described in [113]. As pressure increases, the resistance of the sheath decreases. This results in a higher voltage drop between the plasma and the electrode, although it isn't enough to overcome the collisions that occur in the higher pressure gas to create higher energy ions. Furthermore, the size of the sheath increases with decreasing pressure, causing the particles to be accelerated over a longer distance in low pressure plasmas.

A.3.2.2 Temperature

Reiterating a concept discussed in the sections above, temperature is a major determinant of the etch rate and anisotropy of an etch recipe. As temperature increases, the byproducts of the reaction more readily desorb from the surface and the rate of the etch will increase. Interestingly, the temperature can also be used to tune the degree of anisotropy in an etch when done in the ion-assisted regime. Since the ions will locally increase the temperature in unmasked regions of the substrate, it is possible to tune the temperature so that there is minimal desorption of reaction species on the unmasked regions (i.e., the sidewalls of the etched pattern), while allowing the ionic bombardment of the substrate to heat up those areas exposed to the ion flux. Therefore, the anisotropy can be tuned by reducing the substrate temperature while increasing the forward power of the etch. This is commonly done in cryogenic silicon etching. Silicon fluoride evaporates readily even at relatively low temperatures (SiF_4 has a vapor pressure of 10 torr at -130.4 centigrade); therefore by cooling the substrate to cryogenic temperatures (around -150 centigrade), it is possible to create very anisotropic etch profiles using the ions impinging on the wafer as a heater for the unmasked material.

A.3.2.3 Forward Power

Forward power is another factor closely related to the processes described above. The forward power defines the amount of energy in the capacitvely coupled plasma, and due to the accumulation of electrons in the RF source capacitor also is the primary determinant of the DC bias in an etcher. The DC bias is the measure of the sheath voltage drop which ions are accelerated through, so increasing the RF power (and thereby increasing the DC bias) will increase the energy of ions accelerated towards the substrate. For a given chamber pressure, increasing the DC Bias will typically increase etch rate through surface heating and ion assisted etching processes. In cases of high forward power, it is often advantageous to simultaneously reduce the substrate temperature to promote anisotropy, if desired.

A.3.2.4 Plasma Density

Plasma density is a critical factor in plasma etching. In fact, the reason ICP and ECR etchers have gained so much popularity in industry is due to their ability to independently tune the density of the plasma using the ICP and microwave power delivered to the plasma, respectively. The major effect on a plasma's physical properties is the properties of the ion sheath. The size of the ion sheath is defined as

$$d_{is} = \frac{2}{3}\left(\frac{2e}{m_i}\right)^{1/4}\left(\frac{\epsilon_0}{i_i}\right)^{1/2}(V_p - V_{dc})^{3/4}. \tag{A.14}$$

Therefore by increasing the ion current density, i_i, the sheath will shrink. Since the ion current density is proportional to the amount of ICP power, the sheath properties can also be tuned by sadjusting the ICP power. The major benefit of a shortened ion sheath is that ions are less likely to be scattered as they are accelerated towards the sample, creating a more isotropic etch profile. Chemically a higher density plasma will create more reactive species and will thereby increase the etch rate in most cases.

A.3.2.5 Sidewall Protection

Sidewall protection is a method used in multiple etch processes described below to improve anisotropy over longer etch depths and thereby allow for deeper etch processes. The first etching technique to use sidewall protection by plasma deposition was the Bosch process, patented by Robert Bosch GmbH in 1993. This process uses two plasmas to create deep trenches in silicon wafers (and is often called deep reactive ion etching, or DRIE, for this reason). The first plasma used in this process is a simple SF_6 plasma, which is the standard quasi-isotropic etch that is commonly used for silicon etching. This creates, in theory, a nearly cylindrical etch profile when applied to silicon. Realistically, however, this etch preferentially etches the bottom of a trench due to ion assisted etching and sputtering in a typical ICP etcher, creating an elliptical cross-sectional etch.

The second plasma utilizes C_4F_8 as its carrier gas. This molecule is quite interesting, being a four membered carbon ring with fluorines making up the side chains of the molecule. Carbon-fluorine bonds are quite stable, and thus won't be etched easily in a plasma by chemical means. The other interesting feature of this molecule is that when a single bond is broken, it maintains it a large molecular weight due to its circular shape. These two factors, plus the high concentration of radicals in the plasma, allow this compound to polymerize, thereby uniformly and conformally coating the trench created by the SF_6 plasma. This Teflon-like polymer has the property of etch resistance in the SF_6 plasma. When the SF_6 plasma properties are tuned properly, the ion assisted etching process allows for the bottom of the trench, where the polymer is exposed to the ion flux from the plasma, to be etched while leaving the sidewalls unetched by the neutral species in the plasma. By alternating these two plasmas, it is possible to alternate between etching down a small amount and protecting the newly created sidewalls, thereby allowing for very deep structures to be etched into silicon.

An example of a Bosch process etch is shown in figure A.9. The ridges formed from the alternating etch and polymer deposition are quite obvious in this picture. The major advantage of this etch is that it is possible to create straight sidewalls through hundreds of microns of silicon. The major disadvantage is that the sidewalls will always have this scalloped effect, which may not be desirable. Another etch that uses sidewall protection and creates straight sidewalls is described in section 3.4.2.

A.3.2.6 Gas Composition

Gas composition is a critical factor in multi-gas etch processes. The roles of individual gases in relation to the plasma chemistry must be understood to determine the effect on the etch process. As an example, a commonly used etch is the so called pseudo-Bosch process. This process utilizes a mixture of SF_6 and C_4F_8 simultaneously to create a Bosch-like deep etch where there is polymeric protection of the sidewalls while the SF_6 etches the trench downwards. However, instead of cycling

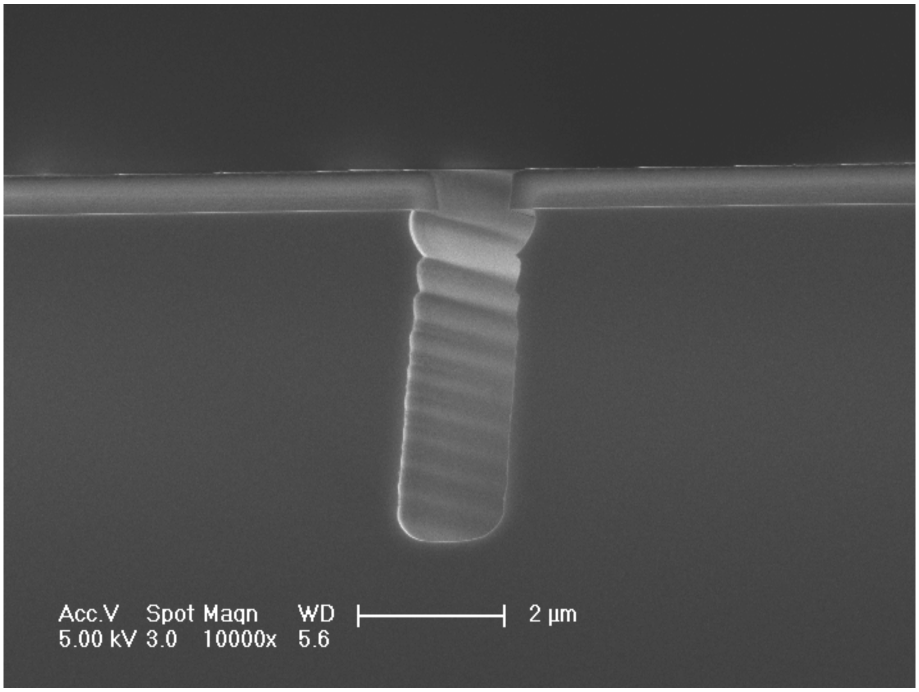

Figure A.9: Trench etched via Bosch process.

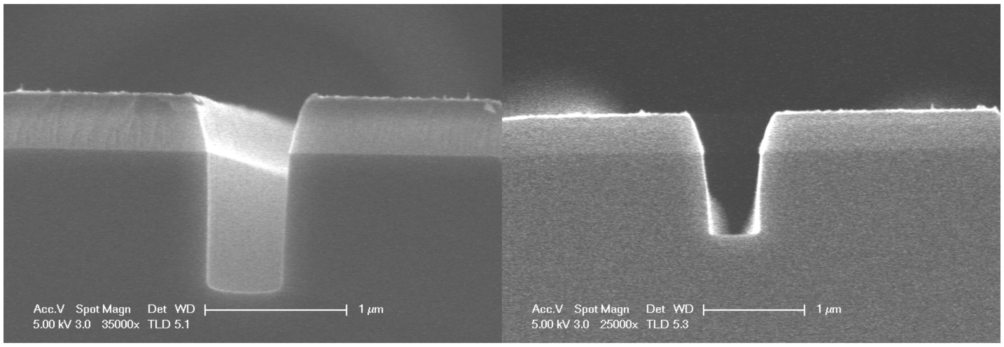

Figure A.10: Pseudo-Bosch process with normal C_4F_8 concentration (left) and increased C_4F_8 concentration (right).

one gas after the other, the pseudo-Bosch process utilizes both gasses simultaneously and are tuned such that the sidewalls are as close to vertical as possible. However, by de-tuning this protection, it is possible to define the sidewall angle as desired. By increasing the C_4F_8 concentration in the plasma, it is possible to reduce the sidewall angle and create slanted sidewalls. Decreasing the C_4F_8 concentration too far can do the opposite, creating sidewalls with an angle greater than 90 degrees. An example of C_4F_8 concentration tuning is shown in figure A.10.

A.3.2.7 Plasma Loading and ADER

The topics of this section have less to do with the plasma themselves, but with how the substrate effects the properties of the etch process. Plasma loading is the concept that the material used to mask the wafer will have an effect on the plasma itself. As the polymeric material is etched away, it is released into the plasma and can effect the etch rate of the substrate in that particular plasma. In general, adding polymer byproducts will slow the etch rate considerably. When using a fully coated wafer with small features, that effect can be in the range of 15-30%. Conversely, it is possible to speed up an etch process by using a carrier wafer coated with a low etch rate material. This is a common effect when using a SiO_2 wafer as a carrier instead of a standard Si wafer. In the author's research, to be able to directly compare etch rates between test chips and full wafer scale etching, a four-inch wafer was coated with photoresist, broken, and used to surround a number of test chips to ensure that the etch rate would be comparable to the etch rate of a full wafer, which was the end goal of the project. This setup is shown in figure A.11.

Area dependent etch rate (ADER) is another effect which must be considered when dealing with etch rates. Due to the physical effects of ion penetration, charging, etc. of the plasma when etching a sample, it is common that the etch rate of small features is dependent on the size of the feature itself. Therefore, small features have lower etch rates than features greater than 2 microns (in the author's experience, this is the point where the bulk etch rate is the same as the feature etch rate). An example of this effect is shown in figure A.12, where a series of devices were etched together with

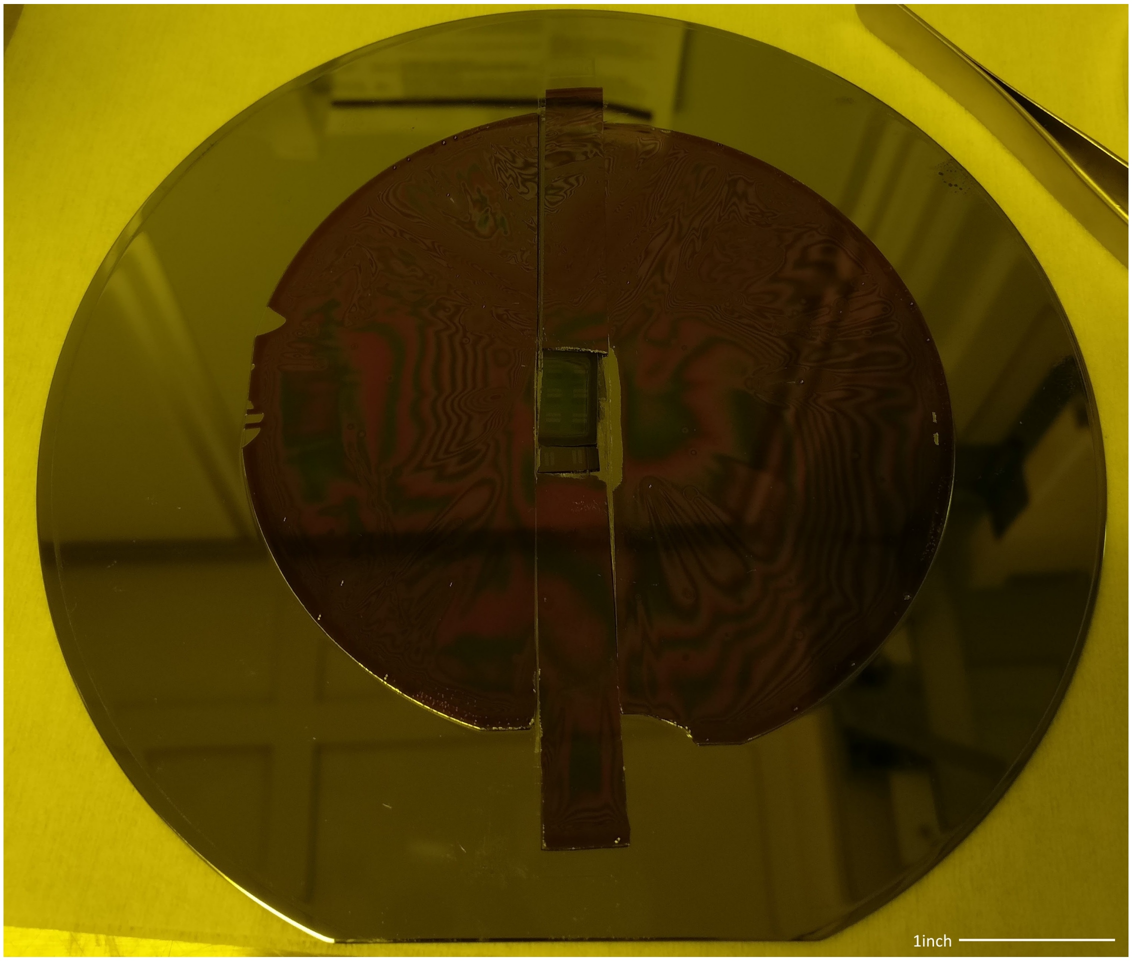

Figure A.11: Carrier wafer with photoresist coated wafer in addition to test samples in center. This technique is used to simulate the properties of the etch when applied to a wafer-scale process as opposed to chips alone.

a decreasing trench width. It is evident that the devices with smaller features are etched less deeply into the substrate.

A.4 Deposition of Thin Films

A.4.1 Theory

A.4.1.1 Flux to the Surface

The first factor to consider when dealing with deposition and predicting deposition rates is the amount of vapor or gas that impinges on the surface of a target. To consider this, it is necessary to understand that the probability of a gas molecule impinging on the surface in the molecular flow regime is approximately one half (given an infinite substrate). Thus, the flux at the surface will

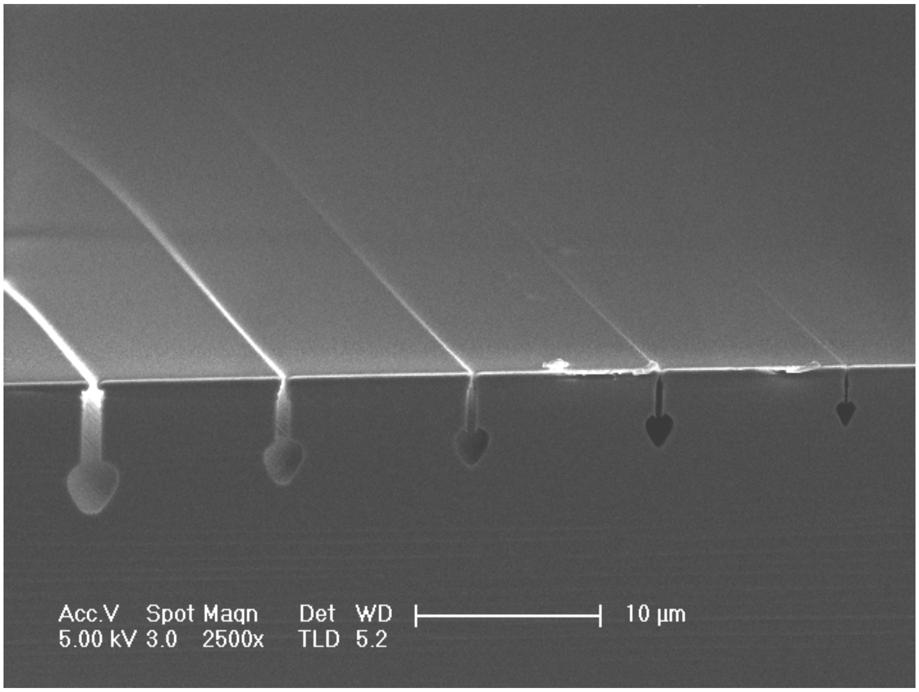

Figure A.12: Example of area dependent etch rate effect.

be approximately one-half of the number density (n) times the velocity of the particle towards the surface. Furthermore, when considering the average velocity in the direction of the surface, when integrating the solid angle flux on the surface the distribution of velocities turns out to reduce the flux by another factor of two, as shown by [107]. Thus, given the average velocity of the gas $\bar{c}$, the relationship for the flux is

$$Ji = \frac{1}{4} n\bar{c}. \tag{A.15}$$

Next, to find the average velocity of a gas molecule, the thermodynamics of the average particle must be considered. Thermal energy is described by a Boltzmann distribution with three degrees of freedom of translation. The mean velocity of the particles can be found by integrating over the Boltzmann distribution, resulting in the following relationship:

$$\bar{c} = \sqrt{\frac{8RT}{\pi M}}, \tag{A.16}$$

where R is the gas constant, T is the temperature of the gas, and M is the molar mass of the molecule.

Next, it is possible to convert the above equation, using the ideal gas law, into the molecular impingement flux in terms of mean velocity. By combining the above equations and using some basic arithmetic, it is possible to write the flux impinging on a surface as a function of the properties of the gas (pressure, temperature, and composition) as

$$Ji = \frac{N_A p}{\sqrt{2\pi MRT}}, \tag{A.17}$$

where N_A is Avogadro's number. Ji is called Knudsen's number and is critical for determining the typical rate of deposition for a thin film in a hot-wall gas reactor.

A.4.1.2 Adhesion

The next critical factor when considering the properties of deposited thin films is the adhesion of the precursor to a substrate. The amount of reactant that remains "stuck" to the surface is critical to determining the deposition rate and to understanding the profile of the deposition process. The derivation of the sticking coefficient is described in detail in [107] and is summarized below.

When considering the deposition of the precursor onto the surface, the first assumption that must be made is that the adhesion process has already reached a steady state. This should occur rapidly in almost all cases. Therefore, using simple first order kinetics, the rate of the reaction is simply

$$R_k = k_k n_s = k_k n_{so}\theta, \tag{A.18}$$

where R_k is the rate of the k^{th} per unit surface area, n_s is the surface concentration of the reactant, n_{so} is the monolayer surface concentration, and θ is the fractional surface coverage of the reactant. In other words, the difference between n_s and n_{so} is that n_s equals n_{so} when there is a full monolayer of reactant on the surface and otherwise n_s is the concentration of the fraction of covered surface.

Given the steady state approximation, a simple mass balance may be used to determine the amount of physisorbed precursor on the surface,

$$J_i\delta(1-\theta) = R_r + R_d = (k_r + k_d)n_{so}\theta, \tag{A.19}$$

where δ is the trapping probability (the fraction of the incoming flux which end up physisorbed to the surface), $(1-\theta)$ is the fraction of the surface not occupied by a precursor molecule, R_r is the rate of the reaction forming the film, R_d is the desorption rate, and k_r and k_d are the reaction and desorption rate constants, respectively. Therefore at steady state, the rate of adsorption on the right is equal to the rate of disappearance of the precursor on the surface (whether due to reaction or desorption). This expression is often rearranged to determine the fractional occupancy of reaction sites on the target surface, where

$$\theta = \frac{J_i\delta/n_{so}}{J_i\delta/n_{so} + k_r + k_d}. \tag{A.20}$$

Although θ alone provides valuable information as to the adsorption properties of the reactant on the surface, it is often more popular to create a dimensionless version of this factor which is relevant to the reaction rate on the surface R_r and incoming flux of molecules from the gas phase. This value is called the sticking coefficient S_c, and is defined as

$$S_c = \frac{R_r}{J_i} = \frac{\delta k_r}{J_i \delta / n_{so} + k_r + k_d}. \tag{A.21}$$

The sticking coefficient is an absolutely critical factor when considering the conformality of thin films deposited in high aspect ratio features. There are two possible mechanisms for molecules to find their way to shadowed features such as sidewalls or into trenches, either by bouncing multiple times off of surfaces or by adhering to the surface and diffusing long distances before reacting with the surface. Since it is important to have somewhat rapid growth of thin films when processing for semiconductor applications, conformality by surface diffusion is usually too slow to allow for rapid processing of wafers and thus has been neglected by industrial development. Therefore the primary physical mechanism for conformal film growth is to allow multiple adsorption/desorption events before a molecule reacts with the surface. This mechanism, therefore, requires the sticking coefficient to be significantly lower than unity.

A.4.1.3 Surface Diffusion

The other critical factor determining the conformality of a thin film is the ability for a precursor molecule to diffuse along the surface before reacting with either the surface itself or a growing nucleus of atoms. The fundamental diffusion length is simply calculated using the following formula:

$$\Lambda = a\sqrt{k_s t}, \tag{A.22}$$

where a is the distance between potential chemisorption sites on the surface (i.e., the distance between low energy states on the surface), k_s is the diffusion rate, and t is time. This, however, is not very enlightening as this equation doesn't account for any of the physical parameters within the deposition chamber. Thus, it is possible to develop an equation that describes k_s as a Boltzmann distributed parameter,

$$k_s = \left(\frac{k_B T}{h}\right) e^{-E_s/RT}, \tag{A.23}$$

where k_B is Boltzmann's constant, T is temperature, h is the Planck constant, E_s is the surface energy and R is the gas constant. This relationship determines the rate of hopping between sites for adsorbed molecules, however it doesn't determine the length of time that a molecule may diffuse before either desorbing or binding to a growing nucleus.

Next, the time of residence on the surface must be determined. This, however, is tricky, as there are two different regimes that are common. There is the regime in which the incoming flux is so fast (or the surface temperature so low) that particles are able to stack one on top of another, thereby burying underlying molecules. The second regime is when the flux is slow (or the surface temperature is high), allowing molecules to desorb before being buried.

The timescale in the burial regime is quite simple. The time for a molecule to diffuse before being buried is approximately,

$$t = \frac{n_o}{J_i}, \tag{A.24}$$

where n_o is the number density of adsorption sites, and J_i is the flux of deposition. Therefore, substituting equation A.23 into A.22,

$$\Lambda = a\sqrt{\frac{k_B T n_o}{h J_i}}e^{-E_s/2RT}. \tag{A.25}$$

The key result from this equation (describing the surface diffusion length for high deposition rates) is that as temperature increases, the diffusion length of the molecule also increases and as deposition flux increases, diffusion decreases.

In contrast, in the desorption regime the rate of desorption is the critical factor. Again, we will assume that this factor is Boltzmann distributed, where

$$t = \frac{1}{k_c} = \frac{1}{A_{oc}}e^{E_c/RT}, \tag{A.26}$$

where k_c is the rate of chemisorption, A_oc is the pre-exponential coefficient for the chemisorption exponential, and E_c is the energy of chemisorption. Substituting for t using the above equation into equation A.22,

$$\Lambda = a\sqrt{\frac{k_B T}{h A_{oc}}}e^{(E_c - E_s)/2RT}. \tag{A.27}$$

In this case, the direction of the relationship between temperature and surface diffusion depends on the difference $(E_c - E_s)$. Since we are in the regime where desorption is dominant, this number is always positive, so there will be an increase in surface diffusion with decreasing temperature.

There is an interesting contrast between the two different regimes, where in the burial regime, increasing temperature will increase the ability for molecules on the surface to diffuse, whereas in the desorption regime the opposite is true. This leads to an optimal temperature between these two regimes where the ability for molecules to diffuse is maximized. Finding this temperature is critical for producing films with high conformality.

A.4.2 Growth of Thermal Oxide

The growth of optical quality silicon oxide films via thermal oxidation is a very common process in semiconductor fabrication. SiO_2 is extremely stable and thermodynamically favorable, as exhibited by the prevalence of SiO_2 in sand and other minerals. Bare silicon will immediately oxidize in an oxygen or water vapor environment, however the oxidation of the surface is self limiting. Once a few layers of SiO_2 are formed on the surface of a wafer, it becomes exponentially more difficult for oxygen to penetrate the oxide film to oxidize the silicon below.

To overcome this barrier, like most other thin film processes, the temperature must be increased far above room temperature to overcome a chemical barrier, in this case diffusion through the oxide film. The model used to describe this process is thanks to Bruce Deal and Andrew Grove, and will be described in brief below [24].

This model assumes that there are three different phases for the oxidant to flow through, and therefore three fluxes to consider. The first flux, $\mathcal{J}_1$, is the flux of the oxidant from the bulk flow into the silicon oxide film. The second flux, $\mathcal{J}_2$, is the diffusion of the oxidant in the film from the surface to the interface between the silicon and the silicon oxide. The final flux, $\mathcal{J}_3$, is the reaction rate of the oxidant with the silicon after passing the interface. These fluxes can be written as

$$\begin{aligned} \mathcal{J}_1 &= h(C^* - C_0) && \text{(A.28)} \\ \mathcal{J}_2 &= D_{eff}\frac{dC}{dx} \approx D_{eff}\frac{C_0 - C_i}{x_0} && \text{(A.29)} \\ \mathcal{J}_3 &= kC_i, && \text{(A.30)} \end{aligned}$$

where C^* is the bulk flow concentration of oxidant, C_0 is the concentration of oxidant in the oxide film at the interface between the film and the gas phase, C_i is the concentration of oxidant at the interface between the oxide and the silicon, h is the gas phase transfer coefficient, x_0 is the current thickness of the oxide film, and k is the reaction rate coefficient for the reaction between the oxidant and silicon.

Assuming this process is at steady state, $\mathcal{J}_1 = \mathcal{J}_2 = \mathcal{J}_3$. The above equations also assume that the flux of oxidant in the film is approximately linear through the oxide film, an assumption which produces a very good fit to experimental data. Given the above three equations and the steady state approximation, it is possible to eliminate the two hidden parameters, C_0 and C_i, and solve for the flux over the density of oxidant molecules in the crystal film,

$$\frac{\mathcal{J}}{N} = \frac{dx_0}{dt} = \frac{kC*/N}{1 + k/h + kx_0/D_{eff}}. \tag{A.31}$$

It is important to note that the factor $\mathcal{J}/N$, or the flux over the number density of the reaction, is equivalent to the growth rate of the film, dx_0/dt. This differential equation can be solved easily by separation of variables, resulting in the following solution (after applying the initial condition of

$x_0 = x_i$ at $t = 0$):

$$x_0^2 + 2D_{eff}\left(\frac{1}{k} + \frac{1}{h}\right)x_0 = 2D_{eff}\frac{C^*}{N}t + x_i^2 + 2D_{eff}\left(\frac{1}{k} + \frac{1}{h}\right)x_i \tag{A.32}$$

This is often simplified by combining the coefficients into terms A and B, such that,

$$A = 2D_{eff}\left(\frac{1}{k} + \frac{1}{h}\right) \tag{A.33}$$

$$B = 2D_{eff}\frac{C^*}{N}. \tag{A.34}$$

Equation A.32 is a quadratic in x_0 and can be solved by the quadratic equation, but intuitively this shows that with time, the film grows slower as the square of time for each time step. Thus, as the film gets thicker, the rate of growth decreases and therefore is slower. Also, the increase in growth with time looks like the inverse of the square function, or the film grows as the square root of time.

When applying the model, the terms A and B must be fit to data to determine the growth times for arbitrary thicknesses. These have been tabulated in the literature[117], and are dependent on the crystal orientation of the silicon, the temperature used to grow the film, and the choice of oxidant (for example, using steam as an oxidant is significantly faster than simply using oxygen gas).