Heat-Assisted Magnetic Recording: Fundamental Limits to Inverse Electromagnetic Design

Samarth Bhargava

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-106
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-106.html

May 14, 2015

Heat-Assisted Magnetic Recording: Fundamental Limits to Inverse Electromagnetic Design

By
Samarth Bhargava

A dissertation submitted in partial satisfaction of the
requirement for the degree of
Doctor of Philosophy
in
Engineering – Electrical Engineering and Computer Sciences
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Eli Yablonovitch, Chair
Professor Ming Wu
Professor David Bogy

Spring 2015

Heat-Assisted Magnetic Recording: Fundamental Limits to Inverse Electromagnetic Design

Samarth Bhargava

To my loving family and friends.

Abstract

Heat-Assisted Magnetic Recording: Fundamental Limits to Inverse Electromagnetic Design

by

Samarth Bhargava

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Eli Yablonovitch, Chair

In this dissertation, we address the burgeoning fields of diffractive optics, metals-optics and plasmonics, and computational inverse problems in the engineering design of electromagnetic structures. We focus on the application of the optical nano-focusing system that will enable Heat-Assisted Magnetic Recording (HAMR), a higher density magnetic recording technology that will fulfill the exploding worldwide demand of digital data storage. The heart of HAMR is a system that focuses light to a nano- sub-diffraction-limit spot with an extremely high power density via an optical antenna. We approach this engineering problem by first discussing the fundamental limits of nano-focusing and the material limits for metal-optics and plasmonics. Then, we use efficient gradient-based optimization algorithms to computationally design shapes of 3D nanostructures that outperform human designs on the basis of mass-market product requirements.

In 2014, the world manufactured ~1 zettabyte (ZB), ie. 1 Billion terabytes (TBs), of data storage devices, including ~560 million magnetic hard disk drives (HDDs) [1]. Global demand of storage will likely increase by 10x in the next 5-10 years, and manufacturing capacity cannot keep up with demand alone. We discuss the state-of-art HDD and why industry invented Heat-Assisted Magnetic Recording (HAMR) [2][3] to overcome the data density limitations. HAMR leverages the temperature sensitivity of magnets, in which the coercivity suddenly and non-linearly falls at the Curie temperature. Data recording to high-density hard disks can be achieved by locally heating one bit of information while co-applying a magnetic field.

The heating can be achieved by focusing 100 μW of light to a ~30nm diameter spot on the hard disk. This is an enormous light intensity, roughly ~100,000,000x the intensity of sunlight on the earth's surface! This power density is ~1,000x the output of gold-coated tapered optical fibers used in Near-field Scanning Optical Microscopes (NSOM), which is the incumbent technology allowing the focus of light to the nano-scale. Even in these lower power NSOM probe tips, optical self-heating and deformation of the nano- gold tips are significant reliability and performance bottlenecks [4][5]. Hence, the design and manufacture of the higher power optical nano-focusing system for HAMR must overcome great engineering challenges in optical and thermal performance.

There has been much debate about alternative materials for metal-optics and plasmonics to cure the current plague of optical loss and thermal reliability in this burgeoning field. We clear the air. For an application like HAMR, where intense self-heating occurs, refractory metals and metals nitrides with high melting points but low optical and thermal conductivities are inferior to noble metals. This conclusion is contradictory to several claims and may be counter-intuitive to some, but the analysis is simple, evident and relevant to any engineer

working on metal-optics and plasmonics. Indeed, the best metals for DC and RF electronics are also the best at optical frequencies.

We also argue that the geometric design of electromagnetic structures (especially sub-wavelength devices) is too cumbersome for human designers, because the wave nature of light necessitates that this inverse problem be non-convex and non-linear. When the computation for one forward simulation is extremely demanding (hours on a high-performance computing cluster), typical designers constrain themselves to only 2 or 3 degrees of freedom. We attack the inverse electromagnetic design problem using gradient-based optimization after leveraging the adjoint-method to efficiently calculate the gradient (ie. the sensitivity) of an objective function with respect to thousands to millions of parameters. This approach results in creative computational designs of electromagnetic structures that human designers could not have conceived yet yield better optical performance.

After gaining key insights from the fundamental limits and building our Inverse Electromagnetic Design software, we finally attempt to solve the challenges in enabling HAMR and the future supply of digital data storage hardware. In 2014, the hard disk industry spent ~$200 million dollars in R&D but poor optical and thermal performance of the metallic nano-transducer continues to prevent commercial HAMR product. Via our design process, we successfully computationally-generated designs for the nano-focusing system that meets specifications for higher data density, lower adjacent track interference, lower laser power requirements and, most notably, lower self-heating of the crucial metallic nano-antenna. We believe that computational design will be a crucial component in commercial HAMR as well as many other commercially significant applications of micro- and nano- optics.

If successful in commercializing HAMR, the hard disk industry may sell 1 billion HDDs per year by 2025, with an average of 6 semiconductor diode lasers and 6 optical chips per drive. The key players will become the largest manufacturers of integrated optical chips and nano-antennas in the world. This industry will perform millions of single-mode laser alignments per day. All academic and industrial players in micro- and nano- optics should excitingly watch what Seagate, Western Digital, HGST and TDK accomplish in the next 5-10 years.

Contents

List of Figures

List of Tables

Acknowledgements

I begin by thanking my advisor, Prof. Eli Yablonovitch, who has led me through the journey of graduate school at UC Berkeley. I am very grateful for his demystification of physics and many hard lessons learned. I also thank Prof. Ming Wu, Prof. David Bogy and Prof. Connie Chang-Hasnain for their tough questions and for being on my qualifying exam and dissertation committees.

I continued to graduate school only after first being thrown into the water via my undergraduate research experience with Prof. David Greve at Carnegie Mellon University. I also had the good fortune of mentorship from Prof. James Bain, who was my first electromagnetics teacher at CMU, and he helped me brainstorm technical ideas and personal career paths during a dozen conferences throughout my PhD.

I would not have started my work in Heat-Assisted Magnetic Recording and Inverse Electromagnetic Design without the mentorship of Matteo Staffaroni and Owen Miller. Matteo's breadth of knowledge kept me motivated to learn, and his realism kept me honest. Matteo's help far exceeded the expectations of an alumnus and friend. Owen started the inverse design project in our group. He taught me to be methodical in the chaos of research, and the many months of pair-coding during my first summer at Berkeley was a phenomenal learning experience.

I will cherish all of the insightful scientific discussions and great friendships in the Yablonovitch research group and the entire Cory Hall optoelectronics office. I worked with Vidya Ganapati practically everyday during my PhD, and I owe much my learning and research progress to our discussions and debates. I collaborated with Chris Keraly on new features and applications of the nano-optics optimization software. I learned much about optical antennas from Kevin Messer, and I enjoyed working with Gregg Scranton and Patrick Xiao on optimizing new and larger electromagnetic problems. I had many fantastic discussions about physics, engineering and manufacturing with the research teams at Seagate, HGST and Western Digital. I fortunately had financial support from the NSF Graduate Research Fellowship Program, Advanced Storage Technology Consortium (ASTC) and Western Digital.

All of my fantastic friends in Berkeley and the Bay Area have made the past 5 years a ton of fun. Thanks to my lovely house for treating me like family. Thanks to the dinner club for the delicious homemade meals and for keeping me grounded when things got crazy. Thanks to all of my hiking and backpacking friends for the great adventures. Thanks to the wonderful coffee shops of Berkeley, especially Philz and the original Peet's, where I did all of my best studying.

I am very lucky to have such an awesome family. My grandmother, aunts, uncles and cousins have helped me in every way imaginable before and during this work. Most of all, I would like to thank my mother, father and brother for always believing in me and pushing me to aim high.

1 Introduction

1.1 Worldwide Data Storage Demand

With the proliferation of personal computing products around the world, humans have access to an abundance of information in the form of text, music, images and video. In the modern ecosystem, this data is increasingly stored in the cloud and digital data storage is considered a public utility with the Internet as its instantaneous worldwide delivery agent. In 2014, the world manufactured ~1 zettabyte (ZB), ie. 1 Billion terabytes (TBs), of data storage devices. The bulk of this storage was supplied by magnetic Hard Disk Drives (HDDs), which has been the staple of digital storage for the last 50 years. The HDD manufacturers roughly sold 560 million drives in 2014 [1]. Solid-State Drives (SSDs) have an increasingly dominant role as the lightweight, high-speed and energy-efficient storage device for consumer electronics, including laptops and smartphones, as well as front end servers in datacenters. It is clear that SSDs have greatly benefited society as an enabling storage device for personal and wearable computing, and SSDs will have a strong and permanent future in the computing economy. In revenue and number of units, SSD production is very impressive. In units of ZBs/year, SSD production is dwarfed by HDD production. For the foreseeable future, the bulk of data storage hardware as a function of ZBs/year will be supplied by HDDs, which can be predicted based on the lack of Silicon nano-fabrication facilities to supply storage devices at the scale of ZBs/year.

There are three key reasons that worldwide data storage demand will grow in the future: a) more users of computing devices in the developing world, b) increased daily usage of existing users and c) larger memory sizes of media files. With billions of users publishing photos and videos of higher resolutions and higher dimensions of space (ie. 3D), there is no foreseeable upper limit to the data storage to be demanded by the world. It is hard to precisely predict the storage demand in 5 years or 10 years, because of the variability in the exponent in this nonlinear trend. Nevertheless, the trend of global demand will look like Figure 1. Assuming moderate increases in manufacturing capacity, there is a large gap between society's future demand and the hardware supply. The most inexpensive way for society to meet exploding global demand is to further scale the data areal density on HDDs.

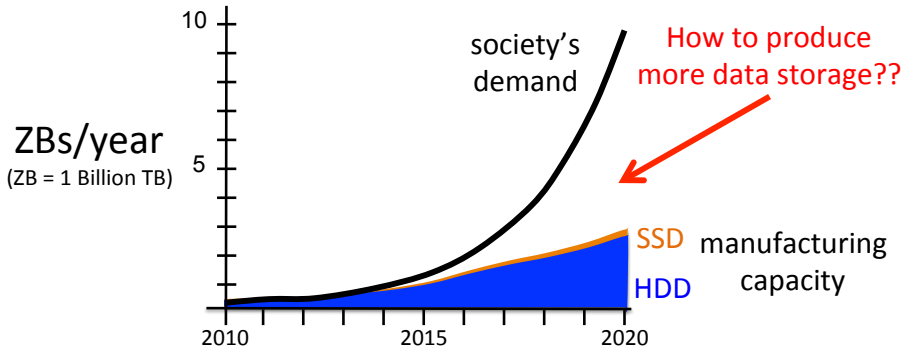

Figure 1: Global supply and demand curves of digital data storage versus time, projecting into the future.

1.2 Optical Chips and Plasmonics in Billions of Storage Devices

Historically, data density in HDDs has increased by 40% per year, faster than Moore's Law of transistors on computing chips. However, the density growth rate for HDD manufactures started to stall in ~2012, as the industry reached the limits of Perpendicular Magnetic Recording (PMR), the technology that led the previous decade of density growth. Typically, areal density is improved by reducing the area of the magnetic grains on the hard disk. To counter the increased rate of spontaneous magnetic moment flipping due to thermal fluctuations, smaller grain size is accompanied by an increase in coercivity of the magnetic material through novel metallurgy. The data recording process involves applying a sufficiently high magnetic field via an electromagnet to forcefully orient magnetic moments on the disk to encode the digital data. Further scaling of PMR became limited by the inability to produce a stronger magnetic field from the electromagnet, because the necessary writing field was higher than that of the magnetic flux saturation in the typically CoFe-based electromagnet. The hard disk industry invented another magnetic recording method to overcome this limitation, dubbed Heat-Assisted Magnetic Recording (HAMR) [2][3].

HAMR leverages the temperature sensitivity of magnets, in which the coercivity suddenly and non-linearly falls at the Curie temperature. Hence, if the disk is at an elevated temperature, then one can rewrite data with a lower magnetic field, which is below the saturation limits of realizable electromagnets. The heating needs to be applied locally to only one bit of information at a time while co-applying a magnetic field. The energy to provide the heating must be delivered using the reading/writing head across an ~2nm air gap to the hard disk. The most efficient energy transfer to ~30 nm spot on the disk occurs with optical frequencies of electromagnetic excitation. For a typical recording rate and hard disk medium, HAMR requires 100 μW of light absorbed in a ~30nm diameter spot on the FePt-based storage media. This is an enormous light intensity, roughly ~100,000,000x the intensity of sunlight on the earth's surface! This power density is ~1,000x the throughput of gold-coated tapered optical fibers used in Near-field Scanning Optical Microscopes (NSOM), which is the incumbent technology allowing the focus of light to the nano-scale. Typical optical transmission efficiency of NSOM probe tips is less than 0.001% [6], while the minimum optical coupling efficiency required for commercial HAMR to be powered by inexpensive 10 mW diode lasers is 1%. Even though the power requirements for NSOM probe tips are many orders of magnitude less than that required for HAMR, optical self-heating and deformation of the nano- gold tips are significant reliability and performance bottlenecks [4][5]. Hence, the design and manufacture of the optical nano-focusing system for HAMR must overcome great engineering challenges in terms of optical and thermal performance.

1.3 Inverse Electromagnetic Design

A key hypothesis in this dissertation is that electromagnetic designs are sub-optimal if designed to be intuitive shapes. Although intuitive shapes like circles with few degrees of freedom may be simple to understand and optimize, the complicated wave-nature of light compels us to ask whether an unintuitive shape may be better. Since one cannot analytically solve or inversely solve Maxwell's equations, electromagnetic designers and researchers are typically restricted to ponder and dream about a magical shape that may cure their system challenges. A more realizable approach is to solve the inverse problem by computationally optimizing 3D electromagnetic structures with thousands to millions of geometric degrees of freedom. Many degrees of freedom are required for the optimization to be 'creative' and allow convergence to structures not fed by an engineer's intuition. To inversely solve for the nano-optics system for HAMR, we are confined to full wave optics simulations (not geometric or Fourier optics) in which we discretize space and solve the coupled differential equations in the entire 3D volume. A typical HAMR model may take 2 hours on a high-performance computing cluster. Because we cannot afford to sample the enormous parameter space, we developed a gradient-based optimization algorithm for wave-optics, dubbed Inverse Electromagnetic Design.

Figure 2 shows an early optimization of a 2D gold nano-antenna illuminated with a Gaussian beam of wavelength 830 nm at an incoming angle of 45° to the left. The objective function was electric field intensity 10 nm below the antenna tip. The antenna boundary iteratively converged to geometry in the right-most frame that resembles a person sitting in a chair. This highly unintuitive structure would never have been designed by intuition or analytic calculation, yet it offers superior performance. This result gave hope and birth to the work in this dissertation. Although a crude first step, we eventually applied this optimization technique to full 3D dielectric and metallic nano-structures within holistic computational models of a commercial HAMR system in collaboration with researchers and engineers of major HDD manufacturers.

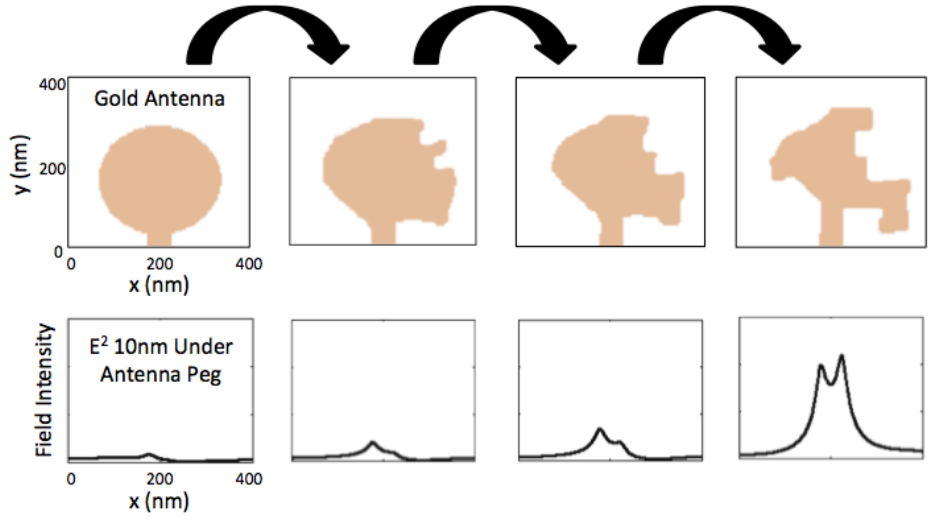

Figure 2: An Iterative and Creative Inverse Design of an Optical Antenna's shape. The antenna in the rightmost frame is not likely to be designed without computation and achieves significantly stronger field localization.

2 Heat-Assisted Magnetic Recording

2.1 Hard Disk Drive State-Of-Art

In 2014, the Hard Disk Drive (HDD) manufacturers produced ~560 million HDDs with total storage capacity of ~1 ZB, or ~1 Billion TBs [1]. A typical HDD contains 2-6 platters or glass disks with 2 surfaces for storage each. Amongst many layers deposited on the glass disks, the storage medium itself is a ~10 nm film of FePt grains of ~5 nm diameter separated by oxide as shown in Figure 3. The result is trillions of individual magnetic domains fully encompassing the entire surface. Data is encoded in the magnetization vector, which in today's Perpendicular Magnetic Recording (PMR) paradigm is perpendicular to the disk's surface. The nano-grains are not patterned using expensive nano-lithography tools used in Silicon processing, but rather through a very inexpensive sputtering process. As such, the FePt grains are not patterned into a regular array of bits that can be accessed individually but rather contain a variation in shape and size. For a high signal-to-noise ratio (SNR), a bit of information is represented by a group of 15-20 grains. Digital bits are assigned as the transition (1) or lack there of (0) between opposite direction magnetization vectors as shown in Figure 4. In 2014, commercial data density was ~600 Gb/in^2 and each bit had dimensions of ~22x50 nm^2. The typical noise in this encoding is the jitter in the arrival of transition edges, because the random sputtered media does not contain rigid bit edges. Because of variation in grain shape, size and coercivity and a finite magnetic field gradient at the edges of the magnetic field spot generated by the electromagnet, the edge between regions of opposite magnetization is not exact. At a larger scale, data is represented as elliptical tracks of width ~50 nm. There is no electrical contact to each bit of information and, thus, a mechanical system to align a chip, designated the 'head', containing a reading sensor and writing transducer, is positioned to a particular track. The disk itself rotates at common rates of 5400, 7200 or 15000 rpm with surface velocities up to 20 m/s. The head is mounted on a wing and the head is suspended in a closed-loop at a flying height above the disk of 2-4 nm. The ratio of velocity to height is the same as a Boeing 747 flying at its full speed of 600 mph precisely 50 nm above the ground as shown in Figure 5. The mechanical system is very robust and leverages pressure feedback in the air gap between the head and disk to maintain fly height. Track alignment leverages a closed-loop with the read sensor actively measuring SNR to prevent mechanical deviation from a skinny ring of data.

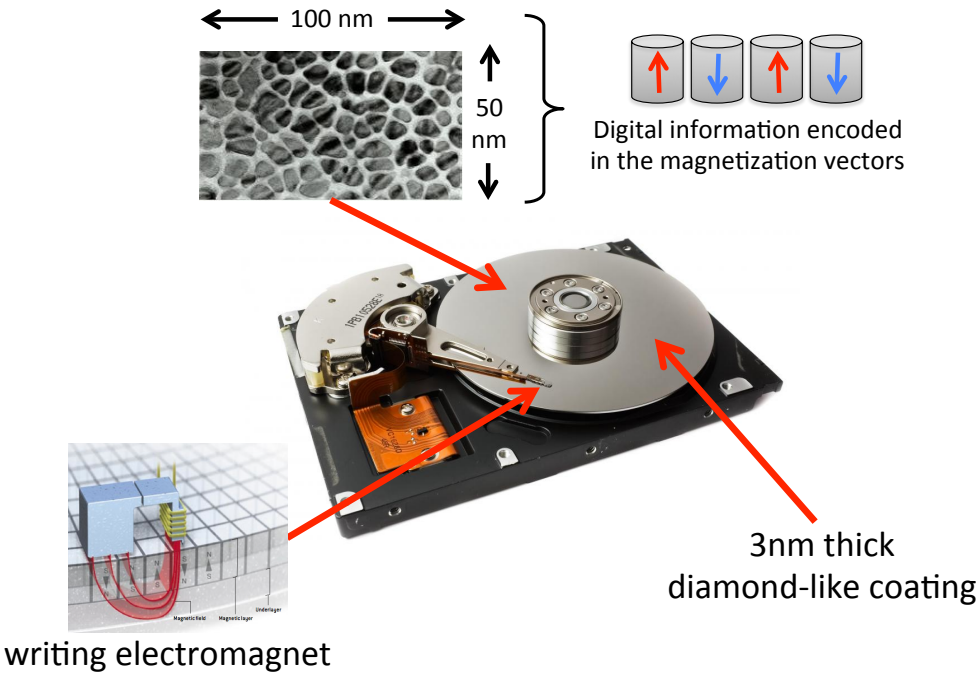

Figure 3: The HDD's surface contains trillions of individual magnetic domains separated by oxide. Data is manipulated using an electromagnet mounted on a mechanical arm that flies above the disk. The surface is protected from abrasion using a diamond-like coating and polymer lubricant.

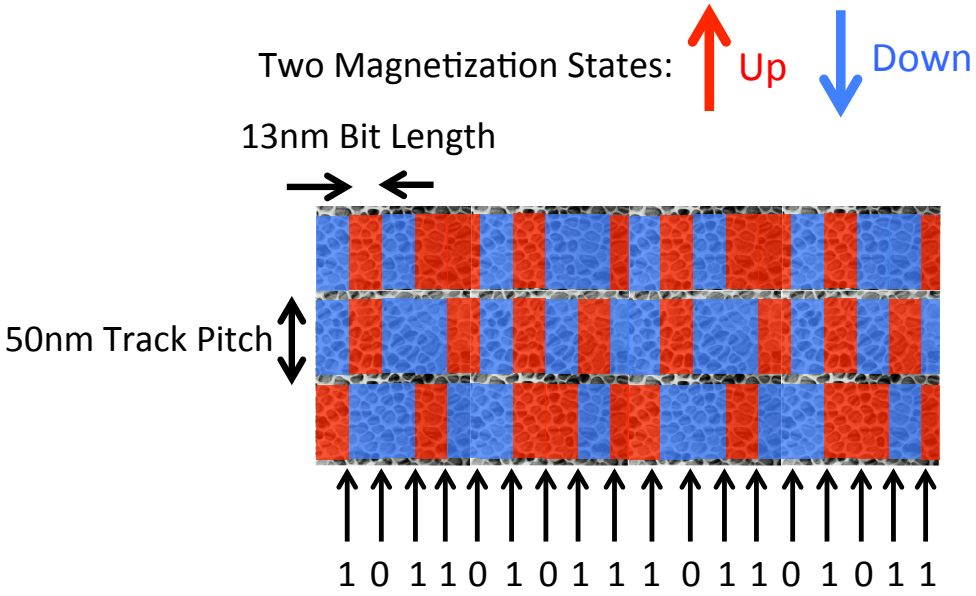

Figure 4: The HDD data encoding scheme of transitions (1) or lack there of (0) of magnetization states. Many grains are use to represent each bit. Data density shown here is 1Tb/in^2 where as the 2014 commercial data density was ~600Gb/in^2.

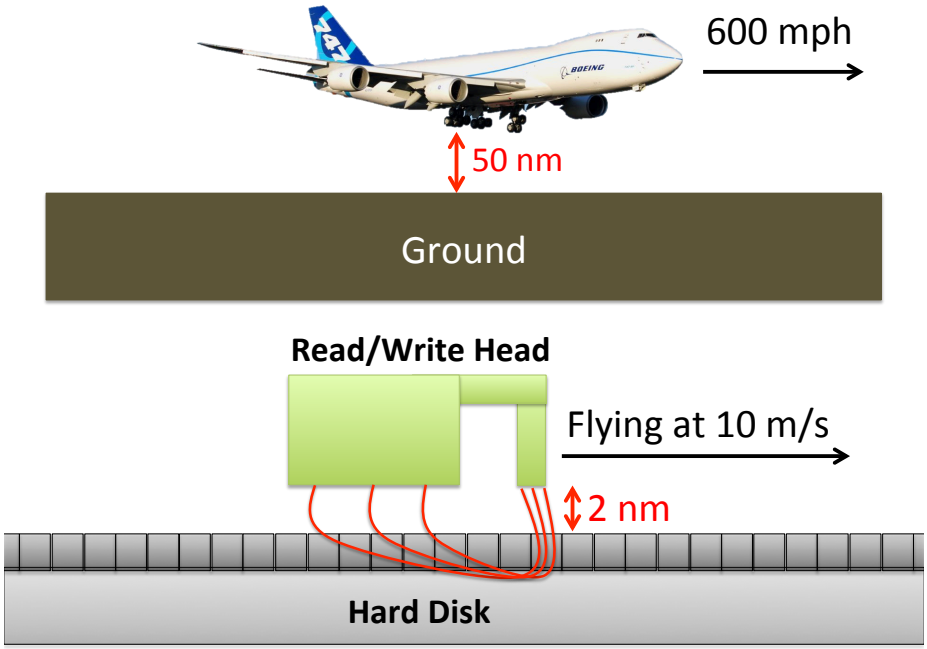

Figure 5: A modern read/write head typically flies faster than 10 m/s at a stable height of 2 nm above a hard disk. The velocity to height ratio is comparable to Boeing 747 flying a full speed 50 nm above the ground.

2.2 Areal Density Scaling

Historically, data areal density has grown by 40% per year, faster than the Moore's Law for transistor computing. Every scaling step requires engineering feats in metallurgy, mechanics, controls, magnetics, electronics, signal processing and manufacturing. The scaling strategy for data recording involves a few key steps as shown in Figure 6. First, we must shrink the size of the magnetic grains to reduce the bit dimension while maintaining a high grain-to-bit ratio for high SNR encoding. Each magnetic domain contains magnetic energy, $K_u V$, and thermal energy, $k_b T$. K_u is the magnetocrystalline anisotropy energy in J/m^3, V is the volume and T is the temperature of the magnetic domain. To prevent thermal noise from randomly causing bit flips during a data storage lifetime of 5-10 years, the storage media must have $K_u V / k_b T \gg 50$ [7]. Every decrease in volume of the magnetic grain is accompanied with an inversely proportional increase in anisotropy energy. An increase in K_u is equivalent to an increase in the magnetic field, H_k, required to switch the magnetization of the grain. Correspondingly, higher K_u requires a stringent condition on the recording transducer to apply a larger magnetic field to align the grains according to the desired digital encoding. Thus, in order to scale density, we must not only scale the hard disk technology but also the recording head technology. However, we cannot generate any arbitrarily high magnetic field by increasing the coil current and decreasing the size of the magnetic yolk tip. Normally, we model magnetic flux B as linearly proportional to the magnetic field H that can be generated by current loops around the magnetic yolk. However, when all the spins in the magnetic material (typically CoFe) align, there is a saturation in the magnetization at $\sim 2.5T/\mu_0$ [7] as depicted in Figure 7. Traditional PMR has lost scaling momentum due to this hard material limit, and this is the motivation behind HAMR as a replacement recording strategy.

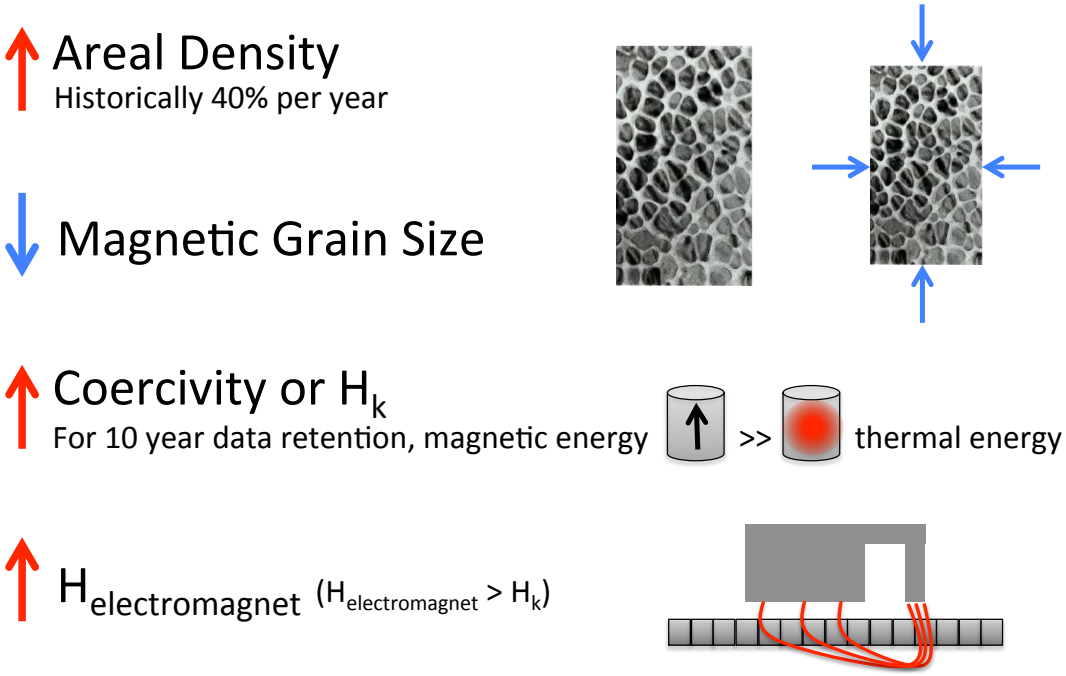

Figure 6: Strategy to scale HDD areal density and writing transducer.

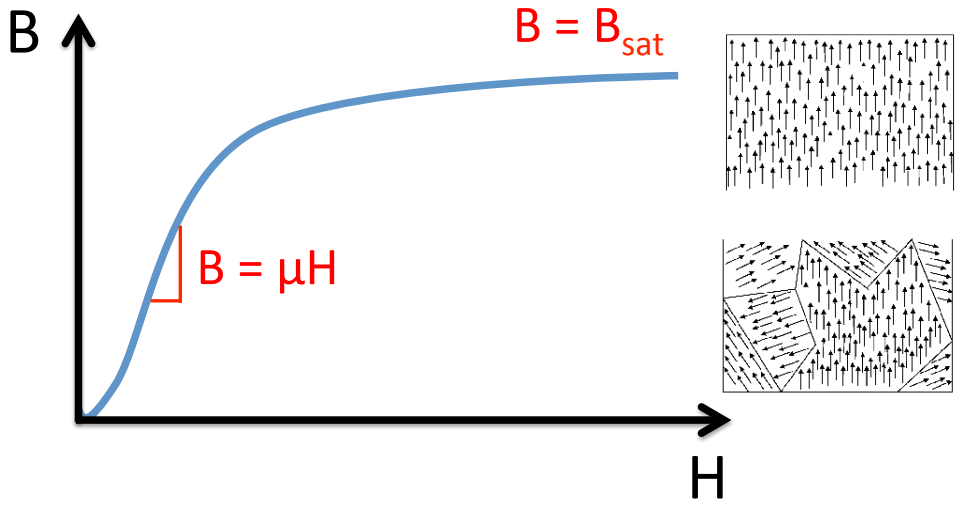

Figure 7: Magnetic flux in a write head eventually saturates and does not linearly increase with magnetic field.

HAMR leverages the temperature sensitivity of magnets to allow for long-term data reliability as well as high-speed write-ability. The mean switching field in the magnetic media falls extremely non-linearly at the Curie temperature as shown in Figure 8. If the media is stored at room temperature where the coercivity is high, then the data may be retained for many years. If a single bit of information is heated to its Curie temperature, then the switching field is very small. The magnetic field from a traditional electromagnet would then be capable of forcefully aligning the magnetization vectors in the granular media according to the encoded data.

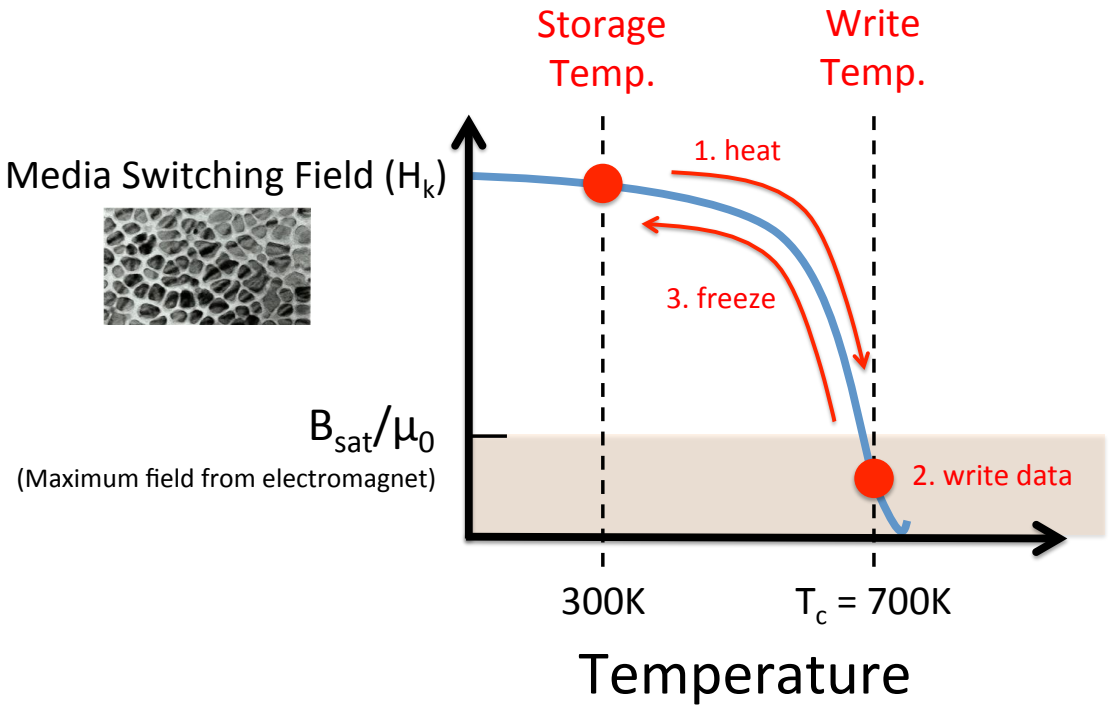

Figure 8: The temperature sensitivity of magnets allows for high coercivity at room temperature and low coercivity at the Curie temperature, where a weaker magnetic field is capable of switching the grain magnetization.

2.3 HAMR System Specifications

First, we must discuss the nature of the heating that must be applied for HAMR to allow for increased areal density while maintaining other important system functionalities. Some of the following arguments could be robustly modeled using micro-magnetic simulations of the magnetization vectors in a statistical distribution of grains perturbed by a thermal spot and write field. Nevertheless, through simple analysis, we can achieve some rough system specifications to achieve high-SNR high-density storage.

Assuming that the magneto-crystalline anisotropy is only along the perpendicular axis as shown in Figure 9, then the magnetic grain's energy as a function of magnetization angular orientation, θ, is modeled by (2. 1) [7]. H_k is the mean switching field of the magnetic grain. H_{eff} is the magnetic field witnessed by the magnetic grain according to (2. 2). $H_{appl}(\phi)$ is the applied magnetic field at an angle ϕ with respect to the perpendicular anisotropy axis. Both θ and ϕ range between 0° and 180° which reflect the two opposite magnetization directions. When H_{eff} is zero, then the energy landscape is proportional to $\sin^2(\theta)$ and two energy minima exist at θ equals 0° and 180°. When H_{eff} is nonzero (and ϕ is not equal to 90°) then one minimum rises in energy while the opposite lowers in energy. The energy difference between the peak energy barrier and the higher of the two minima lowers with increasing H_{eff} and increasing $|\phi - 90°|$.

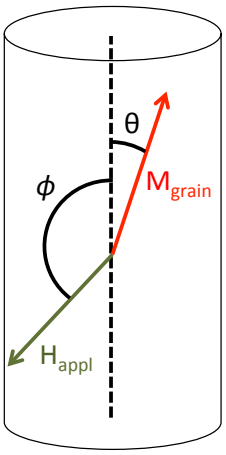

Figure 9: Schematic of an anisotropic magnetic grain with magnetization at angle θ under an applied magnetic field at angle ϕ with respect to the perpendicular axis.

$$E(\theta,\phi) = K_u V\left[\sin^2(\theta) - 2\frac{H_{eff}}{H_k}\cos(\phi-\theta)\right] \quad (2.\ 1)$$

$$H_{eff} = H_{appl} + H_d + H_{ex} \quad (2.\ 2)$$

The energy barrier to switch a magnetization state to the opposite state for $\phi = 0°$ can be approximated by (2. 3) [7][8]. The statistical time constant of a particle overcoming this energy barrier is based on an attempt frequency (on the order of 10^9 – 10^{12} Hz [7]) and the Boltzmann distribution in (2. 4). Clearly, even when H_{eff} equals zero, the time constant is nonzero and approaches zero as the temperature is increased. Hence, the storage metallurgy must be chosen such than $E_B(T_{ambient}) >> k_b T_{ambient}$.

$$E_B(\mathrm{T}) = K_u V\left[1 - \frac{|H_{eff}|}{H_k(T)}\right]^2 \quad (2.\ 3)$$

$$\tau = \frac{1}{f_0}\, e^{\frac{E_B(T)}{k_b T}} \quad (2.\ 4)$$

HAMR is attractive for recording to high coercivity media, because the switching time constant τ can be greatly reduced with temporary heating and not only by increasing H_{eff}. In traditional PMR, linear data density is a function of the write field gradient with respect to downtrack position. In HAMR, the effective write field gradient is dependent on the steepness of $H_k(T)$ and the temperature gradient with respect to downtrack position as shown in

$$\frac{dH_{write}}{dx} \sim \frac{dH_k}{dT} \times \frac{dT}{dx} \quad (2.\ 5)$$

The Curie temperature varies with the material systems. It was observed that $SmCo_5$ and CoPt have a high T_C of 840–1000 K, while $L1_0$ FePt has a lower T_C of 750 K [8]. Hence, FePt-

based granular media has become the standard storage medium for HAMR. Although the time constant in (2. 4) is very non-linear near the Curie temperature, there are numerous reasons why the applied heating must be very local as shown in Figure 10. A typical HDD specification is to ensure data reliability of adjacent tracks after 10,000-100,000 writes of nearby tracks. This makes a stringent case that the heating must be very local to ensure the probability of erasing information elsewhere is minimal. Also, the linear density of data is dependent on the thermal gradient (K/nm), which must be ~15 K/nm to achieve 1 Tb/in^2 areal density for ~50 nm track width. For the reliability of other layers in the hard disk, especially the polymer lubricant, the peak temperature in the thermal spot should be less than ~900 K. Note, T_C is not a constant among all magnetic grains but rather has a distribution depending on the shape and volume of the grains. Considering a T_C distribution width of 50 K, the thermal gradient would need to be 15 K/nm within that whole range for a high SNR. Such a high thermal gradient would be challenging to generate if the media was preheated rather than sitting at ambient temperature, typically 45°C in computing environment. Depending on radial position, the disk may be spinning at 30 m/s, which offers an interaction time between the head and a 15 nm long bit of 500 ps. Within that interaction time, the bit must experience the heating, writing and cooling processes.

The power required to heat a 50×15×10 nm^3 volume of FePt by 400°C in ~200 ps is approximately 45 μW according to (2. 6), where C is the volumetric heat capacity. Of course, the area of one bit in the 10 nm FePt layer is not the only material being heated, because heat dissipates into the remaining volume of the disk. The power dissipated in the disk can be approximated as spherical heat conduction from a hemisphere of diameter 30 nm out to an infinite distance away via a solid angle of 2π through a homogenous glass medium of thermal conductivity ~1 W/mK as shown in (2. 7). The total optical power that is needed for a HAMR system is the sum of these two powers and is on the order of 100 μW.

$$P_{hotspot} \approx \frac{\Delta T \times C \times V}{t} = \frac{400K \times \left(3\times10^6 \frac{\mathrm{J}}{\mathrm{Km^3}}\right) \times (50\times15\times10\mathrm{nm^3})}{200ps} = 45\mu W \qquad (2.\ 6)$$

$$P_{dissipated} = \Delta\mathrm{T}\times\mathrm{K}\times\Omega\times\mathrm{d} = 400K \times 1\frac{\mathrm{W}}{\mathrm{mK}} \times 2\pi \times 30\mathrm{nm} = 75\mu\mathrm{W} \qquad (2.\ 7)$$

In order to record information, a CoFe writing electromagnet must apply a magnetic field on the disk within ~10nm of the optical hotspot. As such, the electromagnet must be near the optical nano-focusing system, and it will likely self-heat from optical absorption or heat via dissipation from nearby hot elements. In the same way that the coercivity of the granular media is reduced with increasing temperature, the magnetization of the electromagnetic yolk is also reduced with increasing temperature. The typical specification is that the electromagnet must remain under 180°C to avoid significant deterioration of the magnetic flux output from the write head.

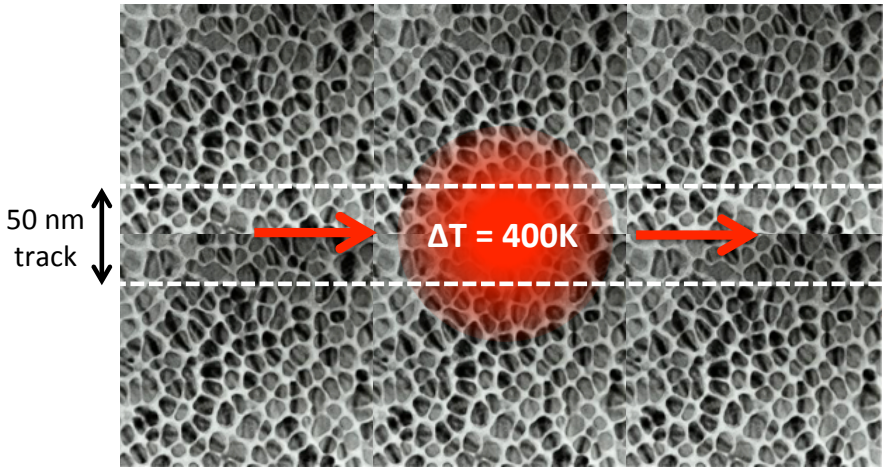

Figure 10: The thermal hotspot on the HAMR media must tightly localized to ~30 nm spot to ensure abrupt bit transitions in the downtrack direction as well as low track erasure even after potentially 10,000 rewrites of an adjacent track.

2.4 Near-Field Transducer (NFT)

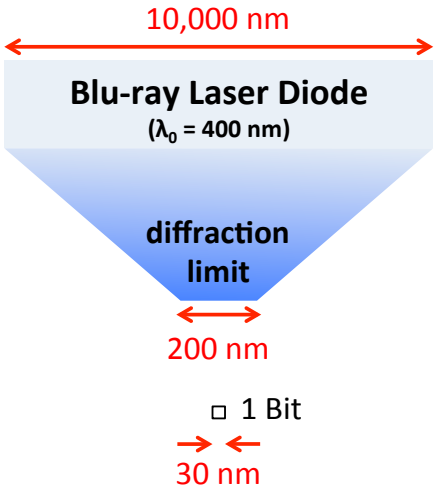

Figure 11: The shortest-wavelength light source (among those with potential supply in billions of units), a 400 nm Blu-ray laser diode, offers a 200 nm minimum diameter spot in the far-field in air with an infinite aperture. To reach a 30 nm spot, we would need a medium with refractive index of 7, which does not exist in nature.

2.4.1 Optical Near-Field Transducer

The challenge in electromagnetic design in HAMR is to focus ~100 μW of energy into a ~30 nm spot on the disk. Traditionally, focusing of light is performed in the far-field with lenses. The shortest-wavelength light source (among those with potential supply in billions of units), a 400 nm Blu-ray laser diode, offers a 200 nm minimum diameter spot in the far-field in air with an infinite aperture as shown in Figure 11. To reach a 30 nm spot in the far-field, we would require an index of refraction of 7, which does not exist in nature. A similar

constraint is found with guided modes in dielectric waveguides, which also yield a spot size much larger than the required 30 nm. Thus, the strategy is to couple energy through the non-radiative near-field interactions via a sharp metal tip. The HDD industry has dubbed this device as a Near-Field Transducer (NFT). In the simplest form, a sharp tip constructed with conductive material has a key advantage over its dielectric counterpart as shown in Figure 12. When an electric field is applied to a conductor, it imparts a force unto free charges which may gather near the surface of the sharp tip. In this volume, the charge density may be enormous and is limited only by a finite skin depth, which has many varied definitions according to structure and material properties (some examples are described well by M. Staffaroni [9]). If the applied electric field and the resultant charge population are static, then the electric field and the spot size decays inversely proportional to r^2 away from the tip. If the charge population is oscillating, then the resultant radial electric field is that of Hertzian dipole (or collection there of) expressed in (2. 8).

$$E_r = \frac{Z_0 I \delta l}{2\pi}\left(\frac{1}{r^2} - \frac{i}{kr^3}\right)e^{-ikr}\cos(\theta) \tag{2. 8}$$

The nano-focusing action can occur if the tip is sized at ~30 nm and the load (the hard disk) is several nanometers away. Considering that electromagnetic absorption scales with the electromagnetic frequency ω according to (2. 9), clearly we must operate at nonzero ω (ie. not DC) to have absorption and joule heating in the load.

$$P_{absorbed} = \frac{1}{2}\omega\epsilon_0\epsilon''|E^2| \tag{2. 9}$$

The air gap between the head and disk defines a distance of ~2 nm between the metal tip transducer to the disk. We can model the air gap as a capacitor with an impedance of $1/i\omega C$, and this impedance clearly decreases with increasing ω. From these simple expressions of the impedance of the air gap and the resistance of the load, it is clear that optical frequencies (100's THz) are desirable to excite the sharp metal transducer. The missing piece of information here is the impedance of the metal transducer at these frequencies, which is drastically different than the impedance of conductors at RF frequencies (refer to M. Staffaroni [9]) for circuit theory models of metal transmission lines at optical frequencies. We will discuss the metal transducer's impedance, which has a real component (ie. lossy and absorbing), in Section 3.3 to derive the temperature ratio between the hotspot in the recording media and the metal transducer. The temperature ratio can determine the optimal optical frequency and material choice for the metal transducer.

2.4.2 Probe Tips for Near-Field Scanning Optical Microscopes

Previous work achieved the focus of light to a 10 nm spot size with gold-coated tapered optical fibers, used in Near-Field Scanning Optical Microscopes (NSOMs). Although they are very valuable in spectroscopy, tapered fibers are remarkably inefficient, and typical optical transmission efficiency to a sub-100 nm spot is on the order of 10^{-5} to 10^{-7} [6]. Moreover, the maximum sustainable optical power at the output of a tapered fiber probe is limited by the significant self-heating via optical absorption experienced in the gold coating and nano-scale gold aperture, which causes structural deformation of the probe tip and thermal instabilities of nearby devices [4][5]. Transmission efficiency and self-heating are thus significant bottlenecks towards higher optical power, higher signal-to-noise ratio and higher scan rate in

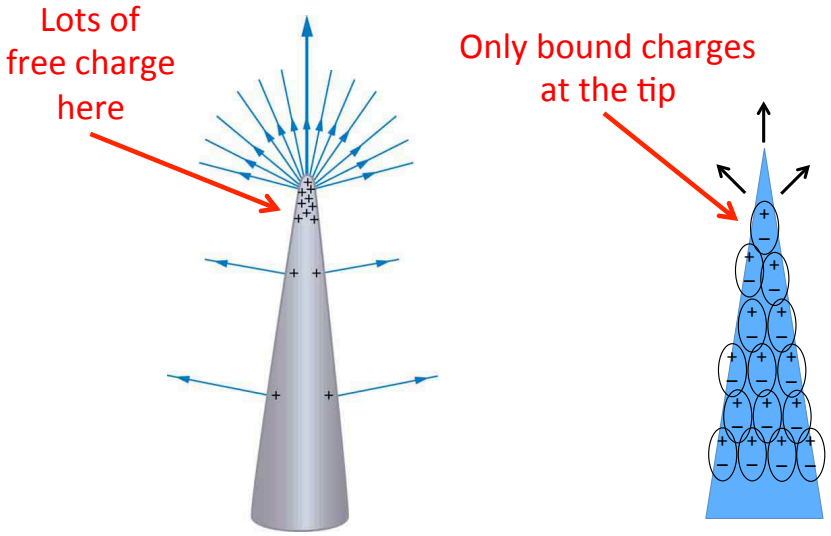

Figure 12: Sharp tips of conductive (left) and dielectric (right) materials have very different responses under electromagnetic excitation. In a conductive material, free charge from a large volume collects at the tip to generate enormous electric field intensity near the tip. In a dielectric material, only the small amount of polarized charged exists at the tip, because the polarization density is uniform throughout the dielectric and the polarization density only linearly increases with the applied electric field.

sub-wavelength optical probes. Metals are very lossy at optical frequencies, and it is not surprising that much of the input laser light is lost while propagating along a ~100 μm metal transmission line (the gold-coated taper). Choosing the optimal metal for optical loss and heat dissipation will be discussed in Chapter 3.

2.4.3 A More Efficient NFT = The Optical Antenna

In terms of metal transducer structure, there is another method to excite a charge resonance in the metal transducer tip. Rather than a 'wired' configuration of exciting a wave that propagates down a long metal transmission line, one can use a 'wireless' configuration of propagating light solely through a confined relatively lossless dielectric mode that is efficiently received by a resonant nano-antenna. In a typical RF antenna, incoming radiation that oscillates spatially according to a radio wavelength of centimeters or meters excites an oscillating current in a conductor. The oscillation is most efficient when the current oscillation is resonant with the radiation oscillation, which occurs when the dimension (of a perfectly conducting rod antenna) is an odd multiple of $\lambda/2$. For RF electronics, the load is attached where the current is the highest (the center point of the current oscillation) and ultimately the incoming radiation will have focused to a tiny wire or transistor that is significantly smaller than the original free space wavelength. Similarly, an optical antenna if sized appropriately for an optical wavelength can capture incoming light radiation to cause a current oscillation within the conducting antenna. To produce near-field excitation, rather than the position of peak current as a location of interest, we can place the sharp metal tip at a position of peak charge density as shown in Figure 13 and Figure 14. Thus, the optical antenna is a device that contains a charge resonance excited by external far-field radiation, and a strong near-field is generated by a narrow tip at a resonant node of the antenna. Because the antenna is commonly sized to be sub-wavelength, the antenna is inherently a more optically efficient NFT compared to a lossy metal transmission line of 100-1000 wavelengths in length. Simply, the light spends less time in the lossy metal via an antenna than a transmission line.

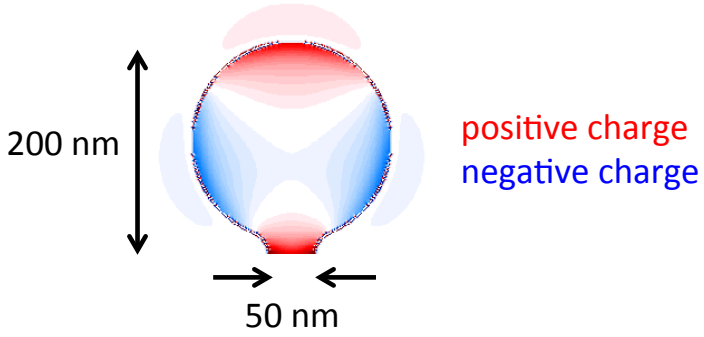

Figure 13: The charge distribution in a quadrupole-sized nano-antenna illuminated at 830 nm wavelength. This charge distribution oscillates with the optical frequency, and its resonance is similar to that of an RF antenna. This optical antenna is also called a Near-Field Transducer, because the charge density inside the antenna tip generates a huge near-field light intensity.

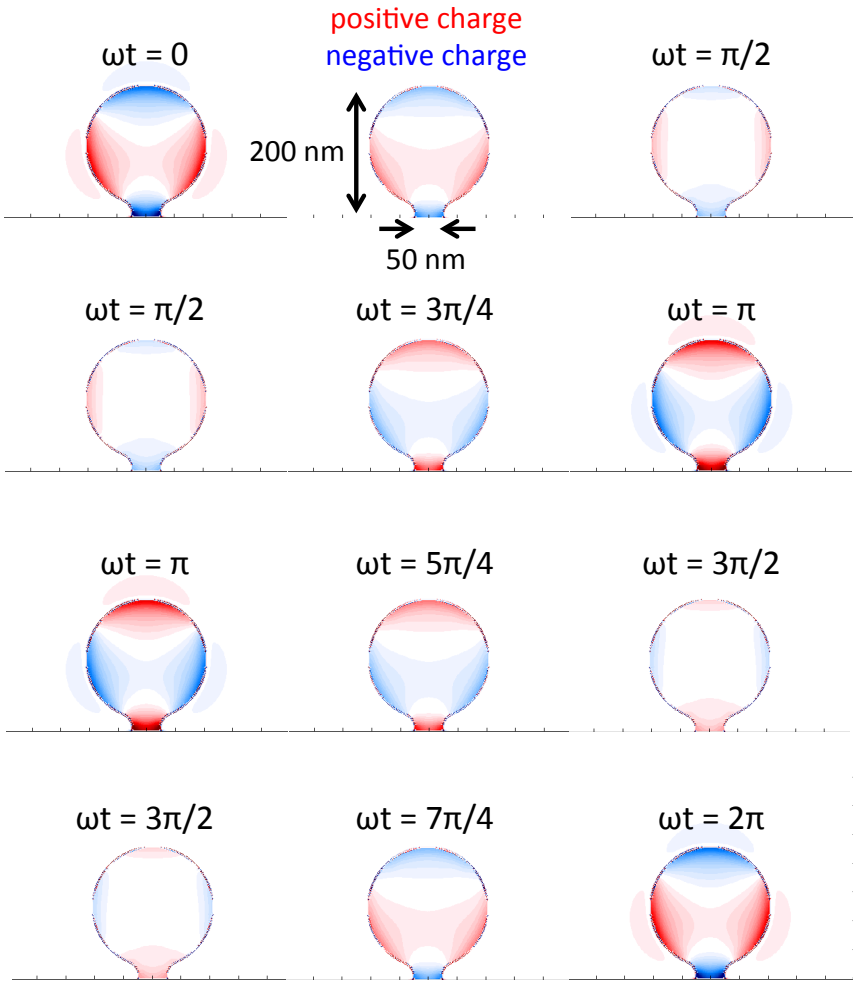

Figure 14: A 'movie' of the charge distribution in a gold optical antenna versus time. We plot here one full oscillation of the optical frequency ($\omega t: 0 \rightarrow 2\pi$). Charge oscillates back and forth between the various poles in the resonant antenna in a similar fashion to an RF antenna. The nano-focusing action occurs when a sharp tip of metal is placed at a node of the resonance. There is only one resonance mode (at a fixed wavelength) that couples charge into the sharp tip. Hence, this is a single-mode device.

2.5 Industry Optical Designs for HAMR

Seagate invented and first published a HAMR optical system using an optical NFT in 2009 [2], and a mock schematic of their system is shown in Figure 15. The red beam indicates 830nm light from a semiconductor diode laser incoming perpendicular to a grating pattern on a TE-mode optical waveguide. The waveguide is a large multimode Ta_2O_5 slab patterned as a parabola with gold coatings on the left/right sides to act as mirrors. Hence, the structure acts as a planar parabolic mirror, which Seagate dubbed as a Parabolic Solid Immersion Mirror (PSIM). The PSIM focuses the incoming light at the bottom toward the Near-Field Transducer. The evanescent light from the waveguide is coupled to the gold NFT, which produces an intense near-field sub-diffraction-limit focusing on to the hard disk, only a few nanometers away. Out of plane from the waveguide is a CoFe magnetic yolk (which is actuated via current coils not shown here), whose focused tip is 10-30nm above the NFT and large return pole is below the waveguide. The electromagnet and optical system are made with wafer-level processing on the read/write head, which is flown above the hard disk at a distance of a few nanometers. To be clear, the author is not a proponent of the multimode waveguide or lollipop antenna for HAMR. We will often compare results against this design, because it a well-accepted benchmark system. It is the most published design from the hard disk industry, and every HAMR optical designer has modeled it before.

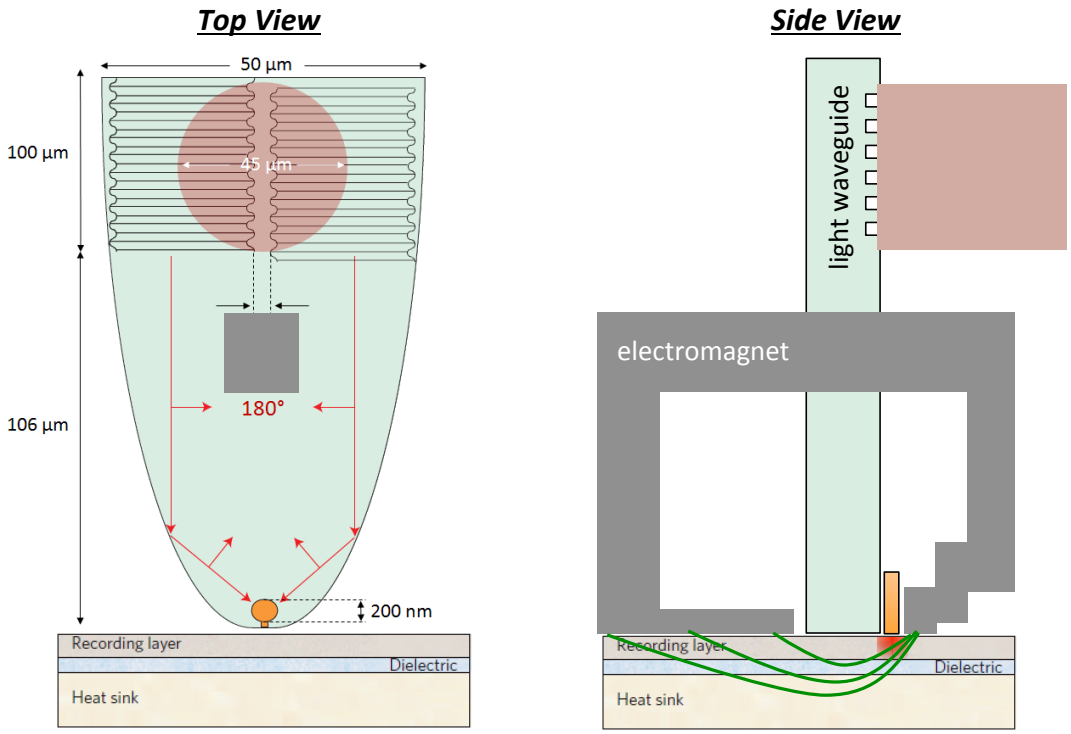

Figure 15: A not-to-scale schematic of the HAMR recording head and hard disk media, resembling the first published HAMR system design (published by Seagate [2]). The red beam indicates 830nm light from a semiconductor diode laser incoming perpendicular to a grating pattern on an optical waveguide. The waveguide is a large multimode slab patterned as a parabola with gold coatings on the left/right sides to act as mirrors. The waveguide acts as a parabolic condenser, which focuses the incoming light at the bottom toward the Near-Field Transducer. The evanescent light from the waveguide is coupled to the gold NFT, which produces strong near-field sub-diffraction-limit hotspot in the hard disk, only a few nanometers away. Out of plane from the waveguide is the writing electromagnet, whose focused tip is 10-30nm above the NFT and large return pole is below the waveguide.

Figure 16 shows a close up cross-sectional view of the NFT and hard disk, and TABLE 1 shows quantitatively the thicknesses and materials in the multi-layered hard disk, often called the 'media stack'. The NFT is integrated on a chip (the head) along with other optical, magnetic and electrical components. This head is flown above disk at a distance of a few nanometers. The air-bearing surface (ABS) of the head is coated in a sub-2nm thick diamond-like carbon (DLC) layer as a durable protective overcoat. Within the media stack, the most crucial layer is a 10nm FePt granular film, which is the recording layer. FePt is typically grown on a MgO underlayer. Underneath the underlayer is typically a metal heatsink, which electromagnetically interacts with the NFT. Moreover, the heatsink conducts heat away to quickly cool the FePt spot after it has been heated and data has been written to it. The disk and substrate itself is typically a sub-mm thick glass disk. The surface of the disk above the FePt, like the head, has a sub-2nm DLC protective overcoat. On the disk above the DLC is a polymer lubricant coating, which serves as a non-stick coating that self-heals after the head impacts the disk. Such impacts, called 'touchdowns', are not accidental and are actually part of the head flying process. In order to precisely fly 2nm above the disk, the head is lowered via multi-axis thermal expanders until a touchdown event is sensed via touch/vibrational sensors at all corners of ABS. Then, the current through the heat expanders is precisely lowered to trigger a thermal contraction of 2nm or the desired fly height.

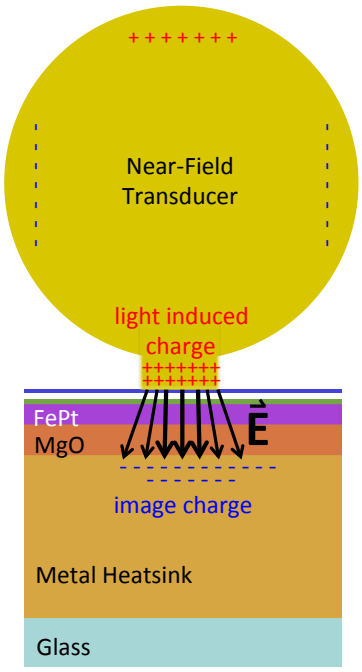

Figure 16: A close-up schematic of the interface between the NFT and hard disk. The NFT is typically a gold thin film with a sharp tip protruding toward the disk. The NFT is integrated on a chip along with other optical, magnetic and electrical components. Inside the hard disk media, there is a 10nm FePt film, which is the recording layer. FePt is typically grown on MgO. Underneath FePt's underlayer is typically a metal heatsink, which electromagnetically interacts with the NFT. Moreover, the heatsink conducts heat away to quickly cool the FePt spot after it has been heated and data has been written to it.

TABLE 1: TYPICAL STRUCTURE OF NFT AND HARD DISK MEDIA

Structure	Material	Dimension
NFT Body	Gold	Large 3D Shape
NFT Tip	Gold	Rectangle of length 30nm
Head Overcoat	Diamond-Like Carbon	1.5nm Thick Film
Air Gap	Ambient Air	2.5nm Thick Gap
Lubricant	Polymer	1.5nm Thick Film
Media Overcoat	Diamond-Like Carbon	1.5nm Thick Film
Storage Layer	Granular FePt	10nm Thick Film
Underlayer	MgO	15nm Thick Film
Heatsink	Gold	80nm Thick Film
Substrate	Glass	300µm Disk

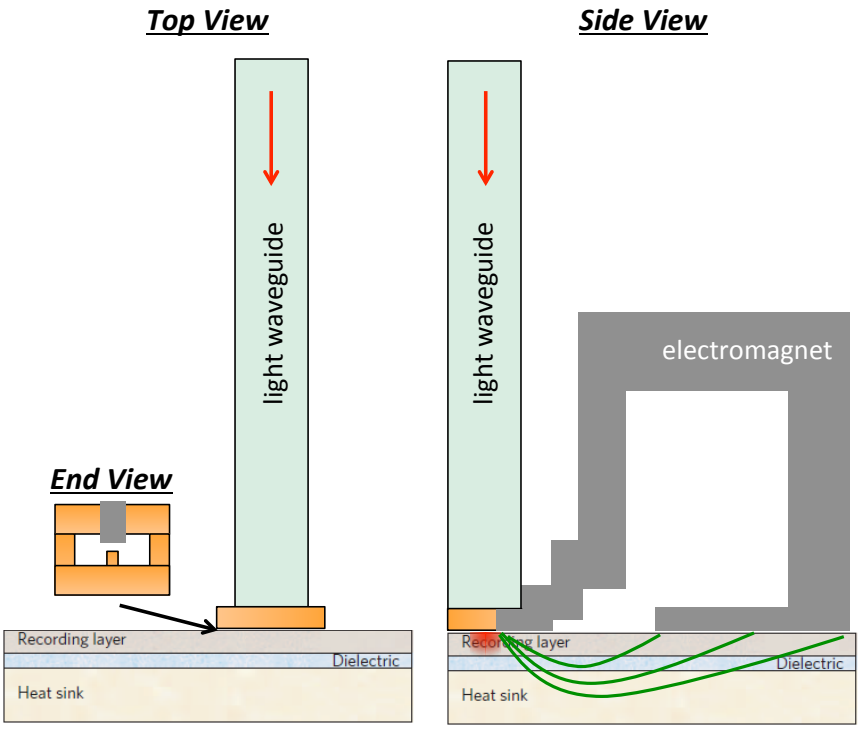

Figure 17: A not-to-scale schematic of the HAMR light delivery system published by HGST in 2010 [3]. Light may be butt-coupled from a diode laser into a single-mode TM waveguide. The NFT is an inverted E-antenna that is placed in the cross-section of the optical waveguide. The electromagnetic write-pole slightly intersects the E-antenna such that the magnetic-thermal offset is 10-30nm.

Figure 17 shows another important industry optical system for HAMR that was published by HGST in 2010 [3], which includes a single-mode rectangular waveguide (the standard for all Silicon Photonics and optical computing chips) coupled to an E-aperture antenna. HGST used a TM-mode waveguide where the electric field is perpendicular to the wafer plane. The 'ridge' of the E-aperture is where the charge density is highest in the NFT and where the near-field coupling to the hard disk occurs. A major difference in NFT manufacturing in the HGST design is that the E-shape of the NFT is not in the top-down direction of the wafer, meaning that it must be made in numerous deposition, lithography and etching steps. The black lines in the end-view diagram in Figure 17 roughly depict the gold structures must be created in separate top-down steps.

2.6 Single-Mode Optical Nano-Focusing System

An important light delivery system specification that is very standard in optical communication and optical computing is the requirement of a single-mode system. The NFT that produces a nano-spot is inherently single-mode, meaning that only one resonance mode of the NFT produces the desired hotspot. This requires that every element in the optical system exciting the NFT also be single-mode, which includes the laser and waveguide. The PSIM is fundamentally flawed, because it is multi-mode. After 100μm of propagation and crosstalk in a multi-mode waveguide, the output mode cannot be guaranteed. This is especially true considering that the waveguide is rough and bumpy, and the incoming laser alignment onto the grating will have some positional and angular error. If the light is randomized in perhaps 100 optical modes, then the total system efficiency could never be more than 1%. HGST had the correct notion of a single-mode system in their 2010 paper [3].

The earliest work that we accomplished in HAMR was to design simple planar NFTs, resembling the lollipop NFT by Seagate, that coupled to single-mode waveguides. The author's designs for simple single-mode light delivery systems for HAMR are shown in Figure 18. The incoming rectangular mode may either be TE or TM. The NFT is a planar film of gold that sits on top of the waveguide. For the TE mode, an asymmetry about the center axis of the waveguide is required in the system, which can be accomplished via an asymmetric NFT shape. For the TM mode, symmetry about the axis of the waveguide is required. In this dissertation, we will also introduce the *fat* version of these NFTs, in which the entire resonant body of the NFT is attached to bulk gold in order to keep the NFT and NFT tip at a low operating temperature. The need for the low temperature will be discussed in Chapter 3 and the designs to achieve low temperature operation will be discussed Chapter 5.

In another work [10], we showed computational optimizations of the planar NFTs for the single-mode waveguide designs using the Inverse Electromagnetic Design software that will be discussed in Chapter 4. If the goal is to have the simplest shape possible, then the author recommends the rectangular shapes shown here amended by the aggressive heatsink discussed in Chapter 5.

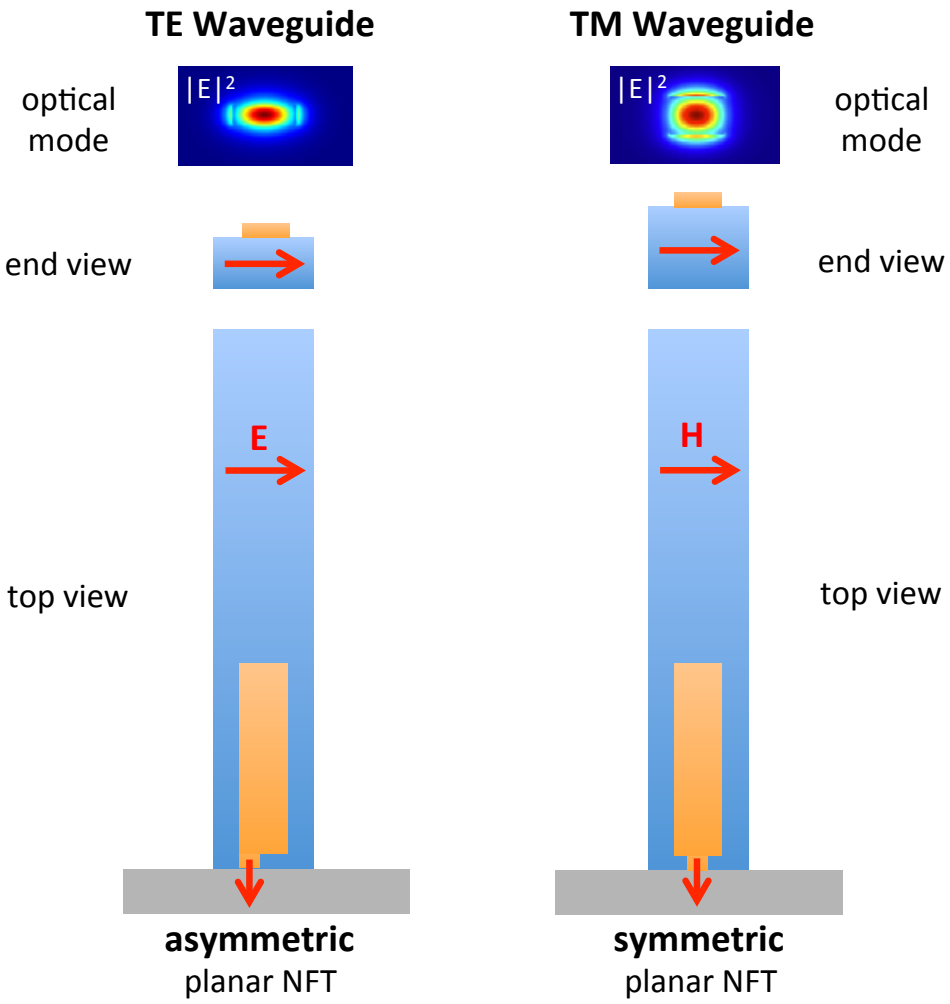

Figure 18: The author's designs for simple single-mode light delivery systems for HAMR. The incoming rectangular mode may either be TE or TM. The NFT is planar film of gold sits on top of the waveguide. For the TE mode, an asymmetry about the center axis of the waveguide is required in the system, which can be accomplished via an asymmetric NFT shape. For the TM mode, symmetry about the axis of the waveguide is required.

3 Fundamental Limits

3.1 Energy and Heat Transfer Methods for Nano-Focusing

3.1.1 Optical Nano-Focusing and Heat Dissipation

Within an optical nano-focusing system like that used for HAMR, energy is transferred at the nano-scale via electromagnetic and thermal physics. We must first get an understanding of all the different mechanisms of energy transfer before being able to suggest informed design choices for the various optical and thermal structures needed to implement a high power nano-focusing system. As calculated by (2. 6) and (2. 7), the desired energy that must be deposited in the hotspot to achieve a 400°C temperature rise in a ~30 nm hotspot within 100 ps is ~100 μW, which is thus ~10^7 W/cm^2 of power density. In HAMR, the primary source of energy is light from a diode laser that is coupled to an on-chip waveguide in the read/write head, which then couples light to the disk via the Near-Field Transducer (NFT). To achieve proper heating of the disk, we must set the laser output power such that the electromagnetic energy transferred from head to disk at the laser's wavelength will also be ~10^7 W/cm^2. Assuming a modest peak temperature rise of 100°C in the NFT tip due to electromagnetic absorption and a large NFT thermal conductivity like 300 W/mK, the heat conduction through the NFT is also ~10^7 W/cm^2 via a similar expression as (2. 7).

3.1.2 Heat Transfer via the Air Gap

Next, there are various methods of energy transfer between the NFT and disk that are not electromagnetic energy at the laser wavelength. The obvious one is heat conducted through the diffusion of hot air particles between opposite sides of a several-nanometer wide gap. The net power transfer can be modeled by (3. 1) and (3. 2). n is the volumetric density of particles, v_{th} is the mean thermal velocity, $k_B T$ is the energy carried by each particle in the air gap, m is the mean mass per particle, and P_{atm} is the air pressure in the gap. For a pressure of 20 atm, which occurs when the head is flying at 10 m/s or faster at a height of several nanometers, the power density of heat transfer through the air is ~10^5 W/cm^2.

$$P_{air} = n \times v_{th} \times k_B T \tag{3. 1}$$

$$P_{air} = \left(\frac{P_{atm}}{k_B T}\right) \times \left(\frac{k_B T}{m}\right)^{\frac{1}{2}} \times k_B T = P_{atm} \left(\frac{k_B T}{m}\right)^{\frac{1}{2}} \tag{3. 2}$$

3.1.3 Nano-scale Enhanced Blackbody Radiation

Next, there are two energy phenomena that are only significant in the nanoscale and are important to discuss even though they are often ignored. The former is plasmon-enhanced blackbody radiation. Blackbody-radiation is typically modeled assuming the density of optical states in free space, and the total radiated power is small for temperatures achieved in HAMR like 700 K (a very low temperature compared to a hot tungsten filament or the sun). However, the optical density of states may be much higher in a nanometric gap between metals, which is the typical phenomenon exploited in plasmonics and metal-optics. The radiated power is given by (3. 3), where n is the number of optical modes and ω is the frequency of the radiated mode. For the limit of $\hbar\omega \ll k_B T$ and a density of states of 2 (as per a 1D transmission line), then we reach the expression for typical Johnson noise in (3. 4). Rather than a wire, if we approximate the head/disk interface as two metals separated by distance d, then the optical density of states is inversely proportional to d^2 as shown in (3. 5). After accounting for the various overcoats on the head and disk, the distance between the NFT and FePt recording layer is effectively ~7 nm. The net power transfer between the opposite sides of the air gap, integrating over 10^{13} Hz, is thus ~10^5 W/cm^2. A more rigorous calculation was performed by Mulet et al. [11] to achieve the expression in (3. 6), where ϵ_1 and ϵ_2 are the respective permittivity of the two objects that are separated. Mulet et al. calculated the same scaling trend of $1/d^2$, and quantitatively also reaches a power density through enhanced blackbody radiation of ~10^5 W/cm^2. Note that this value may be a significant overestimate of a HAMR system, because the density of states is higher for an infinite metal slab structure compared to an NFT tip of diameter ~30 nm.

$$P_{blackbody} = \int n \times \left(\frac{\hbar\omega}{e^{\frac{\hbar\omega}{k_B T}} - 1} \right) d\omega \tag{3. 3}$$

$$P_{1D} \approx 2 k_B T \Delta f \tag{3. 4}$$

$$P_{nanogap} \approx \frac{k_B T \Delta f}{{d_{eff}}^2} \tag{3. 5}$$

$$P_{Mulet.} = \int \frac{1}{\pi d^2} \frac{\epsilon_1 \epsilon_2}{|1+\epsilon_1|^2 |1+\epsilon_2|^2} k_B \left(\frac{\hbar\omega}{k_B T} \right)^2 \frac{\hbar\omega}{\left(e^{\frac{\hbar\omega}{k_B T}} - 1 \right)^2} d\omega \tag{3. 6}$$

3.1.4 Nano-scale Phonon Tunneling

The latter mechanism of nanoscale energy transfer between the head and disk is phonon tunneling, which is a model for whether two objects that are separated by several nanometers are effectively in direct contact or not. Through van der Waal's forces, phonons (or vibrational modes) in one object can couple to another object despite a physical separation. This mechanism is only significant across very short distances, because the van der Waal potential energy function decays as $\sim 1/d^2$ and the overlap integral between the potential energy functions of two objects decays even faster. This effect was examined by Budaev and Bogy [12]. For a HAMR head/disk interface, they calculated the heat transport coefficient across a 2 nm

air gap with 0.25 nm thick carbon overcoats to be ~10^2 W/Km^2 at 0K temperature differential. Unfortunately, no literature before this dissertation has calculated the heat transport coefficient at a ~300K temperature difference between head and disk. If the heat transport coefficient is constant with respect to ΔT, then the phonon tunneling effect in a typical HAMR system yields a net energy flux of ~1 W/cm^2. But, if the heat transport coefficient increases steeply and non-linearly with ΔT, then the phonon tunneling effect may be very significant. This is important future work, because building a HAMR system with low NFT self-heating is hopeless if the heat transfer via phonon tunneling dominates.

A summary of the energy transfer mechanisms relevant to optical nano-focusing is shown in Figure 19. The most dominant mechanisms are the optical near-field coupling at the laser wavelength and thermal conduction through the transducer and load.

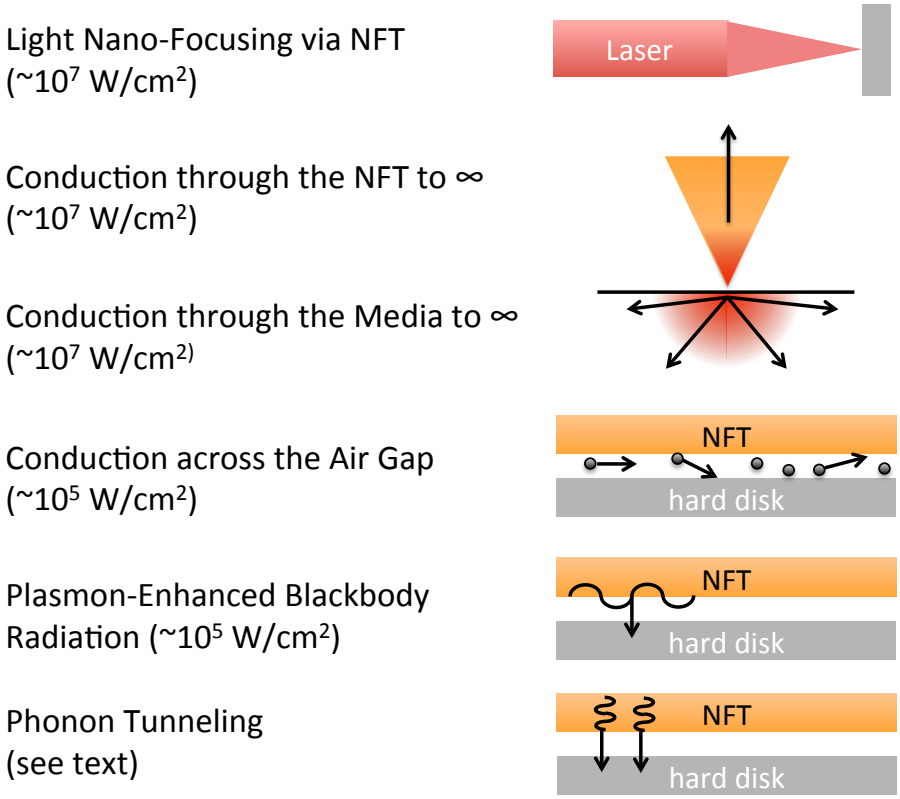

Figure 19: A summary of the various energy transfer methods in the HAMR NFT, air gap and hard disk. The most dominant energy transfers are via near-field light coupling from NFT to disk, heat conduction through the high-conductivity NFT and heat-conduction into the large hard disk substrate.

3.2 Comparison of Metals for Plasmonics and Metal-Optics

3.2.1 Trade-Offs between Optical, Thermal and Mechanical Properties

To engineer any physical device, a natural first question to ask is with which material should the device be constructed. In DC and RF electronics, silver and gold are the most electrically conductive albeit expensive. Copper is less expensive and still offers a very high conductivity. Aluminum is then used for applications like power lines, where the sheer volume begs for an even more economic material. Copper and aluminum are also often used as

heatsinks to dissipate heat from electronics like CPUs and power converters, because of their high thermal conductivities. In incandescent filament light bulbs, tungsten (or wolfram) is used for its high melting point to ensure durability and slow evaporation at the high temperatures (~2500 K) needed to achieve reasonable blackbody emission at visible wavelengths. In plasmonics and metal-optics, gold has been the standard material of choice due to its high conductivity and low corrosiveness.

In plasmonics and metal-optics, massive electromagnetic absorption has always been a plague, the highly inefficient NSOM probe tip being a ubiquitous victim. For HAMR, we desire high optical performance (to deliver 100 μW to a ~30 nm spot on the hard disk), low self-heating (to ensure low deterioration of the writing electromagnet's permeability) and high durability (nanoscale tip must not deform or corrode over many years of use). The noble metals, being the most electrically and thermally conductive, are also the softest metals with the lower melting points compared to metal nitrides and refractory metals. It may not be obvious where in the trade-off curve all of these metals lay, and which will meet all of HAMR's specifications. Generally, HAMR is the most energy intense nano-focusing system that has ever been made, so this application definitely pushes the limits on material robustness.

Sections 3.3 and 3.4 will explore the self-heating that the metallic NFT experiences under illumination and the self-diffusion of the NFT at the appropriate operating temperature. We require that the NFT be less than ~180°C to keep the nearby writing electromagnet from overheating and its permeability deteriorating. We also depend on an exact unmoving shape of the NFT's nano-tip to ensure that the local heating of the disk is of a consistent pattern, such that the reading sensors and signal processing can reliably decode the data written to the disk. So, the NFT tip must not deform or corrode, which are material-dependent properties that are exponentially accelerated at higher temperatures.

3.2.2 Alternative Plasmonic Materials

It should be noted that there has been much discussion (and debate) about refractory metals and metal nitrides as an alternative to noble metals for plasmonics and metal-optics [13][14][15][16][17][18][19]. These alternative materials including W, Mo, TiN and ZrN indeed have higher melting points and display stiffer mechanical properties, while the current king of plasmonics, Au, is relatively soft. Hence, it may naively seem that Au would be inferior for a high-power application like HAMR, where the NFT's nano-sharp tip needs to survive high temperatures without deformation. So, is it not obvious that a material like TiN would be superior? TiN's mechanical properties are sufficiently robust to be used as a drill bit coating, and its optical properties mimic gold to the naked eye. TiN is golden colored and often called Ti-Gold after all. Indeed, there is no doubt that a nanostructure of TiN will be more stable than a nanostructure of Gold at a fixed temperature. However, in most plasmonics and metal-optics applications, including HAMR, the nanostructure in question is not at a fixed temperature but rather is illuminated until a fixed optical output is achieved. Despite the higher melting point, TiN has ~10x higher optical loss and ~10x lower thermal conductivity. Accordingly, TiN is grossly inferior to gold for plasmonic elements that must achieve a particular optical output as it will experience ~100x higher operating temperatures than its gold counterpart for a fixed optical functionality. This will be discussed more rigorously in Section 3.3. The conclusion that the noble metals far outperform metal nitrides in application to HAMR NFTs was also reached through computational modeling performed by Xu et al. [20].

3.2.3 Johnson & Christy's Measurements of Metal Nano-films

A popular misconception propagated by many articles on alternative plasmonic materials is that nano-scale Au and bulk Au have drastically different optical properties. Thus, many

articles make comparisons not to the well accepted measurements of Au from Johnson & Christy [21], but rather to a fictitious version of Au that is 3x more absorbing [13][14][15][16][17][18][19]. The blunder that several articles made was to assume that Johnson & Christy merely measured the optical properties of bulk gold and to make the arbitrary assumption that nano-films of Au will observe roughly 3x higher imaginary permittivity. The truth is that Johnson & Christy performed measurements on 18.5 – 50 nm metal thin-films, which is on-par or thinner than common plasmonic and metal-optic devices (HAMR NFT's are typically 40-100 nm thick). Johnson & Christy specifically discusses the phenomenon of the degradation of optical conductivity with decreasing film thickness, and observed that thin-films offer the same optical properties as bulk material down to 20-30 nm thicknesses.

"The optical properties of evaporated thin films have been found to be the same as for bulk materials, provided the thickness of the films is greater than about 200-300 Å." - Johnson & Christy, 1972 [21]

"As mentioned, the inferred dielectric constants- are independent of film thickness only above a certain critical thickness, which for gold is about 250 A. In order to check this limit, we made a thinner gold film of I86-Å thickness. The initial optical measurements on this film as evaporated failed to converge to any values of n and k in the visible or ultraviolet. The observed reflection and transmission contours did not intersect in the n-k plane, presumably because the film was not homogeneous and continuous. After annealing this film for 12 h at 150°C, the structure changed enough to give convergent solutions for n and k. The values for ϵ_2 (and n) were well outside of the error estimates for thicker films. Although the annealing apparently improved the uniformity of the 186-Å film, inferred values of the dielectric constants are not representative of bulk gold. Thus bulk values for ϵ_1 and ϵ_2 can only be obtained from films whose thickness is about 300-Å or more." - Johnson & Christy, 1972 [21]

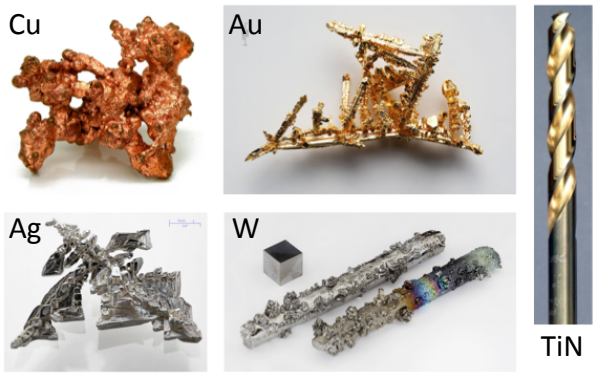

Figure 20: Various metals used for with varying electronic, thermal, optical and mechanical properties.

3.3 Self-Heating Limits for Optical Nano-Focusing

3.3.1 Figure of Merit = Load/Transducer Temperature Ratio

Previous work in optical nano-focusing was achieved through gold-coated tapered optical fibers used in NSOMs. The maximum sustainable optical power at the output of a tapered fiber probe is limited by the significant self-heating via optical absorption experienced in the gold coating and nano-scale gold aperture, which causes structural deformation of the probe tip and thermal instabilities of nearby devices [4][5]. Thus, rather than optical transmission efficiency, a more important metric for NSOM probe tips may be the ratio of light intensity in the sample versus the temperature rise in the metal tip. In HAMR, it is not the light intensity in the hard disk that is crucial, but rather the temperature rise in the hard disk which must be ~400°C. In order to reach the desired media temperature, the NFT will inevitably self-heat

and possibly fail. Hence, for HAMR, the temperature rise ratio between the media and NFT is an important Figure of Merit.

As discussed in Section 3.1, the dominant heat transfer mechanisms in the transducer's nano-tip and load's nano-hotspot are thermal conduction via the metallic transducer and via the load's substrate. To derive the temperature ratio between the hard disk media and NFT, one may approximate the system with a simple model of spherical heat conduction from heat sources due to optical absorption in the medium and the tip of the NFT. As shown in Fig. 1, we approximate the metallic NFT as a cone and the multi-layered media stack as a homogenous hemi-sphere. It is further assumed that all the significant optical absorption is in the NFT tip and in the media hotspot.

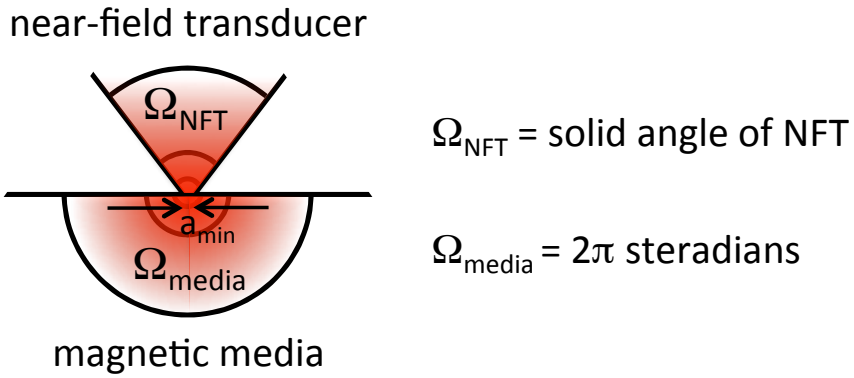

Figure 21: Model of spherical heat conduction in a hemispherical media and conical NFT.

Relative to ambient temperature infinitely far away, the temperature rise in the NFT tip and media hotspot are described by

$$\Delta T_{NFT} = \frac{P_{NFT}}{\Omega_{NFT} \times K_{NFT} \times a_{min}} \tag{3. 7}$$

$$\Delta T_{media} = \frac{P_{media}}{\Omega_{media} \times K_{media} \times a_{min}} \tag{3. 8}$$

where P is the heat generated in the NFT tip or media, Ω is the solid angle, K is the thermal conductivity, and a_{min} is the minimum diameter of the NFT tip. It is assumed that the diameter of the hotspot in the media is also a_{min}. Next, we relate the heat generated to the optical absorption in these respective regions by

$$P_{NFT} = \frac{1}{2}\omega\, \varepsilon_0\, \varepsilon''_{NFT}\, |\boldsymbol{E_{NFT}}|^2 V_{tip} \tag{3. 9}$$

$$P_{media} = \frac{1}{2}\omega\, \varepsilon_0\, \varepsilon''_{media}\, |\boldsymbol{E_{media}}|^2 V_{hotspot} \tag{3. 10}$$

where ω is the frequency of the excitation laser light, ε_0 is the free-space permittivity, ϵ'' is the imaginary part of the permittivity, $|E|^2$ is the light intensity in the NFT tip or media, and

V is the volume of heated region in the NFT tip and media hotspot. We can estimate the ratio of light intensity in the NFT tip to media according to electromagnetic boundary conditions at the NFT-air-media interface. Assuming the electric field is perpendicular to the interface, which is the case for the Hertzian dipole mode in the NFT tip as described by (2. 8), the light intensities in the NFT and media are simply proportional as shown in (3. 11). Assuming that all electromagnetic absorption in the media occurs in the FePt granular layer, then $V_{hotspot}$ is the area of the hotspot multiplied by the grain thickness t_{grain}, typically ~10 nm. At the NFT tip, the electromagnetic field is confined according to a skin depth δ_{NFT} (please refer to M. Staffaroni [9] for different types of skin depths) of the electric field penetrating and decaying into the metal NFT. Hence, V_{tip} is the area of the NFT tip multiplied by the skin depth (or some multiple of it). Assuming that the area of the NFT equals the area of the hotspot, then these respective volumes are related by (3. 12).

$$|\epsilon_{NFT}\boldsymbol{E_{NFT}}|^2 \approx |\epsilon_{media}\boldsymbol{E_{media}}|^2 \tag{3. 11}$$

$$\frac{V_{NFT}}{\delta_{NFT}} = \frac{V_{hotspot}}{t_{grain}} \tag{3. 12}$$

By combining equations (3. 7)(3. 12) we derive the following dimensionless ratio for media/NFT temperature.

$$\frac{\Delta T_{media}}{\Delta T_{NFT}} \approx \frac{|\epsilon_{NFT}|^2}{|\epsilon_{media}|^2} \times \frac{\varepsilon''_{media}}{\varepsilon''_{NFT}} \times \frac{K_{NFT}}{K_{media}} \times \frac{\Omega_{NFT}}{\Omega_{media}} \times \frac{t_{grain}}{\delta_{NFT}} \tag{3. 13}$$

For a low temperature transducer, this ratio must be as high as possible. Clearly, there are significant factors that are not accounted for in this expression, such as the anisotropic thermal conductivity of HAMR granular media and its under-layers, or the exact structural design of the NFT. Nevertheless, this expression correctly emphasizes some key design requirements for low temperature NFT operation.

1) The media must have minimum heatsinking.

2) NFT metallurgy must be optimized for $K_{NFT} \times \frac{|\epsilon_{NFT}|^2}{\varepsilon''_{NFT}} \times \frac{1}{\delta_{NFT}}$.

3) NFT structural design must have the largest solid angle of heat conduction at the NFT tip.

3.3.2 Temperature Ratio versus NFT Material and Excitation Wavelength

Figure 22 shows the optical properties of noble metals, refractory metals and metal nitrides from measured data from Johnson & Christy [21], Rakic [22], and Boltasseva [13]. Thermal conductivities assumed here are 300W/mK (Au), 430W/mK (Ag), 400W/mK (Cu), 240W/mK (Al), 170W/mK (W), 140W/mK (Mo), 20W/mK (TiN) [23][24] and 20W/mK (ZrN) [23][25]. In plasmonic applications like the optical NFT in HAMR, we desire specific system functionality related to the output optical field of the metallic transducer. Specifically for HAMR, we must excite the NFT with sufficient laser power to accomplish intense nano-focusing in the media to achieve a media temperature rise of ~400°C. The first major limit to material choice for the NFT is the significant self-heating that the nano-tip will experience in order to localize ~100 μW into a ~30 nm spot, which represents a power density 100,000,000 more intense than

sunlight on the earth's surface. The self-heating is not only a function of the joule heating density (from optical absorption) but also a function of how quickly heat dissipates (a function of thermal conductivity and structure). Figure 23 shows the absolute NFT temperature under operating conditions for HAMR, according to (3. 13), for different NFT materials and different wavelengths of operation. The operating condition is defined as sufficient laser illumination onto the NFT to achieve a temperature in the nano-hotspot of 700 K from an ambient temperature of 300 K. Note that the y-axis in Figure 23 is shown on a log scale, and the respective melting points are shown to the right. Other metals and metal nitrides were considered too, but were not sufficiently interesting to discuss here. The noble metals are dramatically colder, which was expected due their higher electrical and thermal conductivity. Both the refractory metals and the metal nitrides show drastically higher NFT operating temperatures. The author recommends that this analysis be treated as lower limits of the NFT operating temperatures, as many considerations were ignored like absorption in the NFT other than within the tip as well as temperature-based deterioration of the optical and thermal conductivity. The conclusions from our analytic results correlate well to those via computational modeling of NFTs of different metals performed by Xu et al. [20].

Note, although the comparison in this figure is to the respective melting points, the NFT melting point far exceeds the actual temperature limits in an optical nano-focusing system. For HAMR, ~10 nm away from the NFT tip is the tip of a CoFe electromagnet, whose permeability deteriorates above ~180°C. By this limit, the only metals worth pursuing are noble metals and Mo (if operated ~2 μm wavelength). This completely rules out metal nitrides for HAMR unless future metallurgists grow better alloys. Moreover, temperature aggravates self-diffusion and mechanical deformation of the NFT tip, which will be discussed in Section 3.4.

To emphasize the dependence of NFT geometry toward self-heating, Figure 24 shows the equivalent of Figure 23 but for a very narrow solid angle (or cone angle) of the NFT. The same analysis for a skinny NFT, which is a norm among thin film optical antenna designs, yields that most NFT materials have too high of an operating temperature. In this skinny configuration, only Au and Ag are low enough in temperature for consideration in HAMR, especially since the wavelength of interest for HAMR is ~830 nm. The un-physically high operating temperature for the metal nitrides implies that it is impossible to focus the required power to the nano- hotspot using those materials.

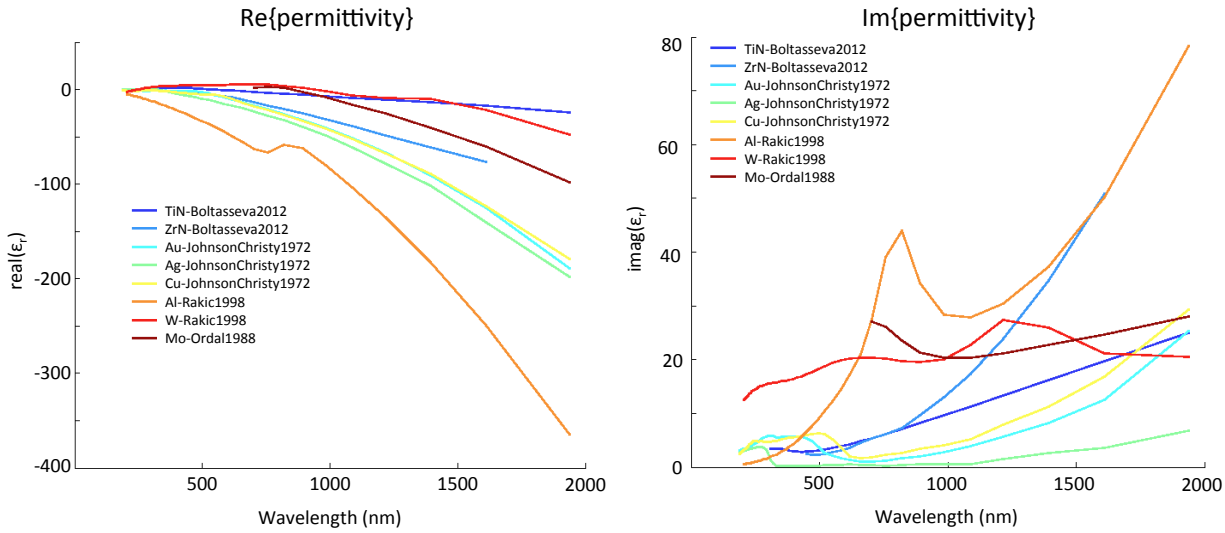

Figure 22: The real and imaginary parts of metal permittivities at optical frequencies, which is relevant to plasmonic and metal-optic applications [21][13].

TABLE 2: THERMAL CONDUCTIVITIES: METALS FOR NEAR-FIELD TRANSDUCERS

Material	Thermal Conductivity (W/mK)
Silver	430
Copper	400
Gold	300
Aluminum	240
Tungsten	170
Molybdenum	140
TiN [23][24]	20
ZrN [23][25]	20

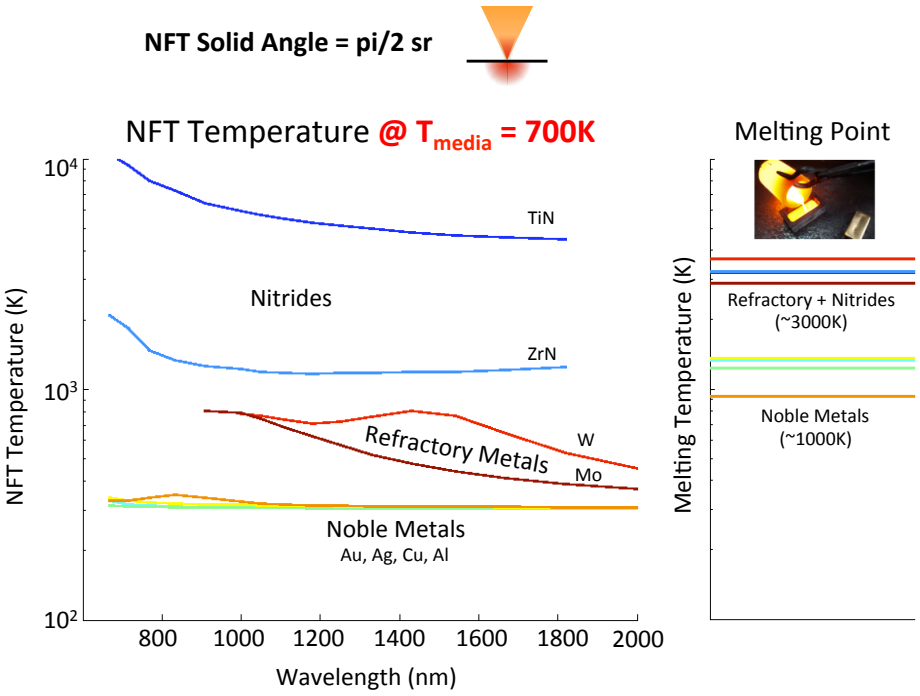

Figure 23: The temperature of a Near-Field Transducer's nano-tip induced by the self-heating under optical illumination to achieve a fixed desired optical output, which is defined here as a peak media hotspot temperature of 700 K. Ambient temperature is assumed to be 300 K. Calculation is determined by (3. 13). The NFT is assumed to be very wide with a large structural solid angle of 0.5π.

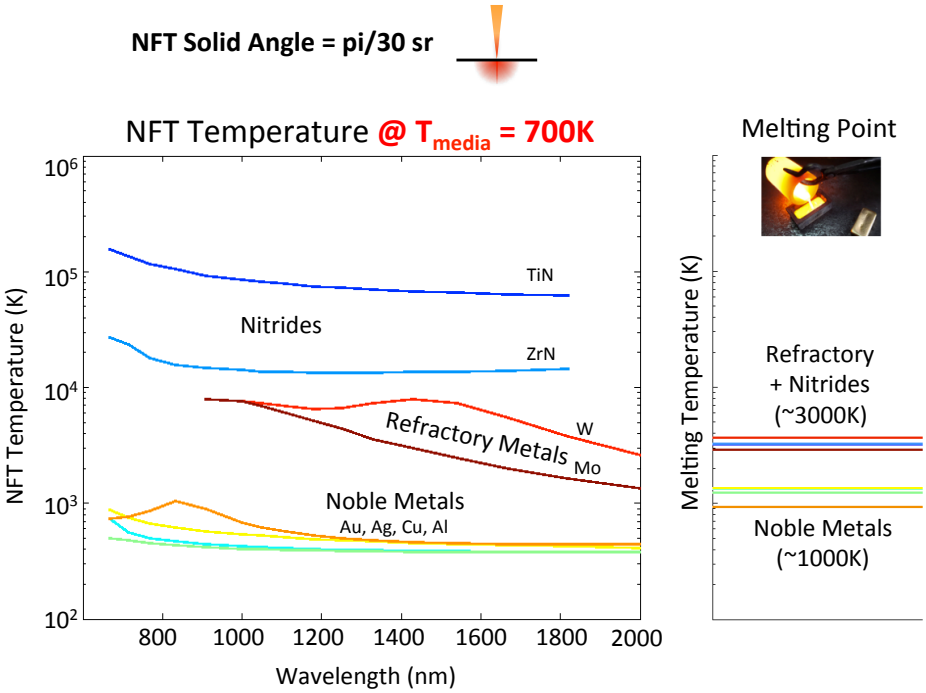

Figure 24: The temperature of a Near-Field Transducer's nano-tip induced by the self-heating under optical illumination to achieve a fixed desired optical output, which is defined here as a peak media hotspot temperature of 700 K. Ambient temperature is assumed to be 300 K. Calculation is determined by (3. 13). The NFT is assumed to be very skinny with a narrow structural solid angle of π/30.

3.4 Self-Diffusivity and Deformation of Nano-Transducers

Mechanical deformation, aggravated by the high operating temperature of an optical nano-focusing transducer, is a crucial failure mechanism as depicted in Figure 25. In HAMR, we rely on the tip of the NFT to remain intact over many years of operation. We typically view diffusion of atoms or molecules as a phenomenon within the liquid and gas phases, but atoms actually diffuse distances much greater than the atomic spacing within solids as well. Because the bit size is ~30 nm and the fly height is ~2-4 nm, there is very little tolerance for atomic movement which may distort the NFT peg shape. If the metal in the NFT tip recedes into the body of the NFT or starts to round, then the effective fly height of the NFT and the shape of heated spot on the disk can quickly fall out of specification. Hence, the rate at which atoms diffuse within the NFT metal may be a useful metric to gauge the rigidity of various metals for nano-focusing operation.

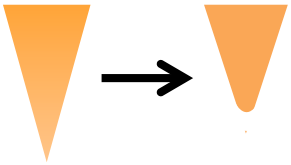

Figure 25: Nature abhors a sharp tip. A crucial failure mechanism of an NFT is the rounding of the nano-tip over time, which is accelerated at elevated temperatures.

The self-diffusivity (units cm^2/s) of atoms in a solid can be modeled by (3. 14), and like many physical phenomenon, it scales as constant multiplied by a Boltzmann distribution with a particular activation energy. A summary of self-diffusion activation energies is given in TABLE 3. The activation energies summarized here for refractory and noble metals are specifically of surface self-diffusion, which may be of specific interest to the deformation of nano-tip surfaces. The surface self-diffusion activation energies for TiN and ZrN were not found in literature, but rather the volume self-diffusion energies are reported here. Much higher activation energies are observed for volume diffusion compared to surface diffusion. So, the analysis here is perhaps overly optimistic for the metal nitrides. The important observation to conclude from this table of activation energies is that the most conductive metals in nature are also the softest metals. It is unfortunate for applications like optical nano-focusing that rigid materials like TiN and W (used for resistance to abrasion and high temperatures) are not great materials choices for optics, as discussed in the Section 3.3.

$$\mathrm{D} = \frac{l^2}{t} = D_0 e^{-\frac{E_A}{k_B T}} \tag{3. 14}$$

Figure 26 and Figure 27 show the self-diffusivity of various metallic NFT's under HAMR's operating conditions, defined here as illumination with sufficient laser light to heat a nano-spot in the media to 700 K from an ambient temperature of 300 K. For reference, the volume self-diffusivity of ZrN at ambient temperature is $\sim 10^{-30}$ cm^2/s. Accordingly, all of the self-diffusivities calculated here are extremely high due to the softness of most metals (especially the noble metals) which is then aggravated by the high operating temperatures induced by the transducer's self-heating. For the case of a large NFT with a wide solid angle (Figure 26), ZrN, Mo and W offer the possibility of lower diffusivity despite the much higher operating temperatures. However, all of these diffusivities pale in comparison to what may be needed, 10^{-20} cm^2/s or below. For the case of a skinny NFT with a narrow solid angle (Figure 27), the

trend within the metals is different. For skinny NFT's, the noble metals are the most stable despite their low activation energies. The metal nitrides have much higher diffusivities despite the higher activation energy, because their operating temperatures are an order of magnitude higher than that of the noble metals.

TABLE 3: SELF-DIFFUSION ACTIVATION ENERGIES: METALS FOR NEAR-FIELD TRANSDUCERS

Material	Self-Diffusion Activation Energy (eV)	Source
TiN	2.0	Hultman, 2000 [26]
ZrN	2.0	Hultman, 2000 [26]
Molybdenum	0.9	Antczak & Ehrlich, 2014 [27]
Tungsten	0.8	Antczak & Ehrlich, 2014 [27]
Aluminum	0.4	Antczak & Ehrlich, 2014 [27]
Gold	0.4	Antczak & Ehrlich, 2014 [27]
Silver	0.3	Antczak & Ehrlich, 2014 [27]
Copper	0.25	Antczak & Ehrlich, 2014 [27]

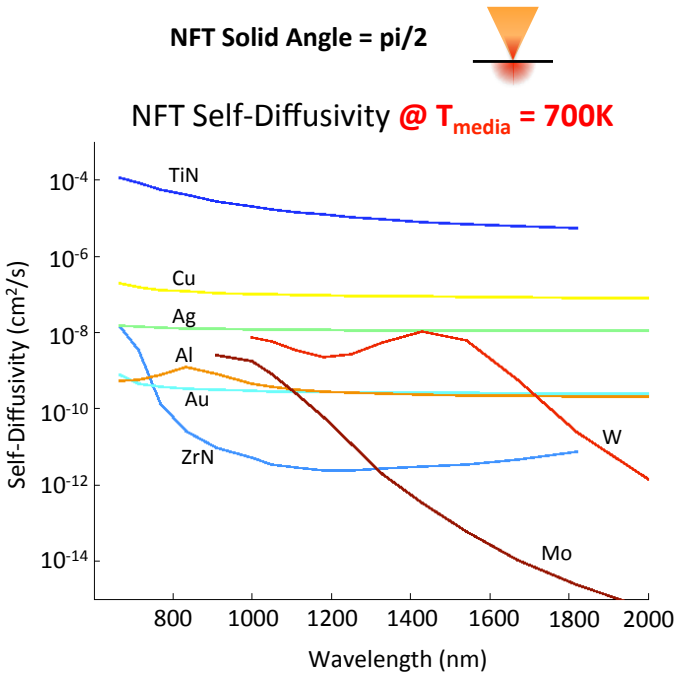

Figure 26: The self-diffusivity of a Near-Field Transducer's nano-tip accelerated by the self-heating under optical illumination to achieve a fixed desired functionality, which is defined here as a peak media hotspot temperature of 700 K. Ambient temperature is assumed to be 300 K. Calculation is determined by (3. 13) and (3. 14). The NFT is assumed to be very wide with a large structural solid angle of 0.5π.

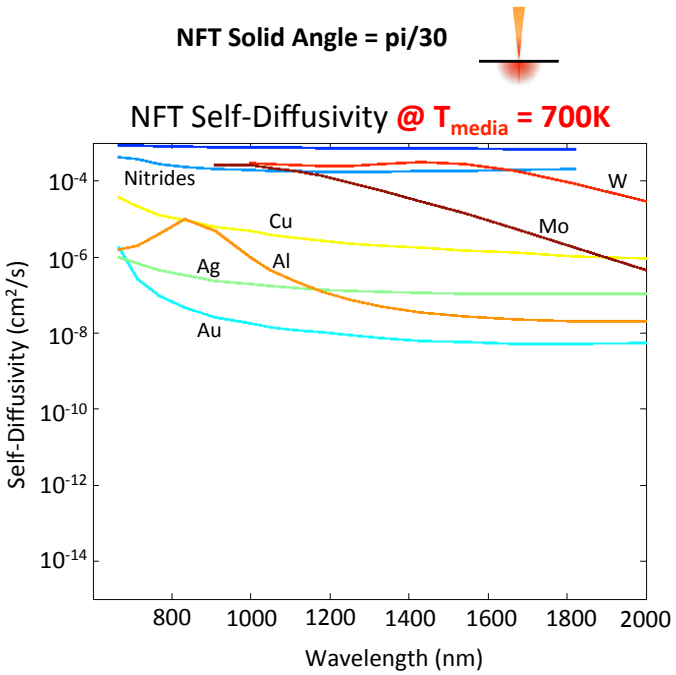

Figure 27: The self-diffusivity of a Near-Field Transducer's nano-tip accelerated by the self-heating under optical illumination to achieve a fixed desired optical output, which is defined here as a peak media hotspot temperature of 700 K. Ambient temperature is assumed to be 300 K. Calculation is determined by (3. 13) and (3. 14). The NFT is assumed to be very skinny with a narrow structural solid angle of π/30.

3.5 Self-Assembly of NFT Tip via Surface Tension Forces

Nature abhors sharp tips. This is intuitive for liquid droplets, which tend to a spherical shape to reduce surface energy (an effect that we can visibly see in the macro-scale on the timescale of seconds). Even when below its melting point, atoms within a solid are constantly self-diffusing within itself. For macro-scale objects, the effect is so minimal that we tend to approximate solid objects as being stable objects without internal kinetics. However, if the diffusion length of an atom over an hour is on the order of micrometers, than nano-scale objects cannot be modeled as kinetically stable shapes. Defining a nano-geometry via lithography and chemical etching does not ensure that the object will remain in the desired shape over a time scale of hours, days or years. Over such timescales, the solid nano-particle will tend to an equilibrium shape similar to a droplet.

Figure 28 shows the equilibrium shape of gold on glass surrounded by air. The arrows denote the 3 surface energy forces that pull on the tri-material vertex. It is the balance of these 3 vectors that determines the wetting angle of the gold. If the glass is rigid, then the wetting angle of the gold particle is ~130°C [28]. Although it seems like a burden that a desired nano-particle shape will unintentionally and inevitably morph into a droplet-type structure, the wetting and equilibrium shape of solid metals can also be thought of as a robust self-assembly process. Syms et al. used the wetting of Pb-Sn solder on a fixed substrate and hinged micro-mirrors to achieve the self-assembly of micro-mirrors tilted in 3D [29]. Syms et al. was able to use the volume of solder to apply different intensities of surface tension forces unto the mirror in a predictable process to achieve various desired tilted mirrors. Syms et al. performs a detailed analytic treatment of the equilibrium geometries induced by surface energy forces. In

this section, we will show qualitatively how the surface energies can used to benefit the construction of the nano-focusing transducer rather than the mechanism of failure.

The phenomenon of capillary action is based on these surface energy forces as shown in Figure 29. Generally, a material with a wetting angle less than 90°C will seep through a rectangular tube and materials with a wetting angle greater than 90°C will be pushed out. The general trend in HAMR development has been to find an interfacial layer that gold will wet more strongly. A common material for this is Cr, but the drawback is that these adhesion layers are optically absorbing and will self-heat dramatically, based on Section 3.3. For plasmonic and metal-optics devices, all of the field intensity is on the surface, so the placement of lossy Cr at the surface is detrimental. Another method to engineer the wetting of gold within the NFT peg (the tube) is to make the peg conical rather than rectangular. Rather than the strict condition that $\theta_W < 90°$, we have a condition with a free degree of freedom that $\theta_W < 90° + \alpha$ where α is the half cone angle. As long as we make the cone sufficiently wide, then the material will be seeped in rather than pushed out. The advantage of this mechanism is emphasized in Figure 30. A typical failure mode in HAMR is when the gold in the NFT peg recedes in to the main cavity of the NFT. This is expected because of the wetting of gold on glass and that most HAMR NFT pegs are rectangular in shape. If the peg is conical, then the gold material is actively pushed toward the air-bearing surface from surface tension forces. In this way, the peg will retain its shape and will self-level even if some deformation, thermal expansion or contraction occurs. Thus, a conical NFT peg self-assembles and self-heals. Despite the intense internal kinetics of the gold atoms as predicted by Section 3.4, the boundary of the gold in the most crucial region (the peg and air-bearing surface) will be fixed in place.

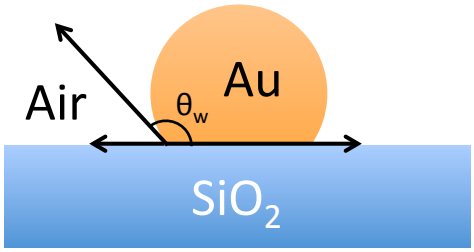

Figure 28: The equilibrium shape of gold in air on glass has a wetting angle of ~130°C.

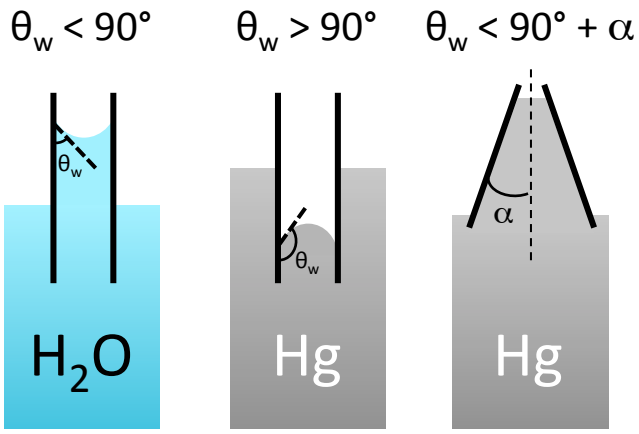

Figure 29: Materials with acute wetting angles will be pulled inward and materials with an obtuse wetting angle will be pushed outward. An angled tube may retain materials of large wetting angle.

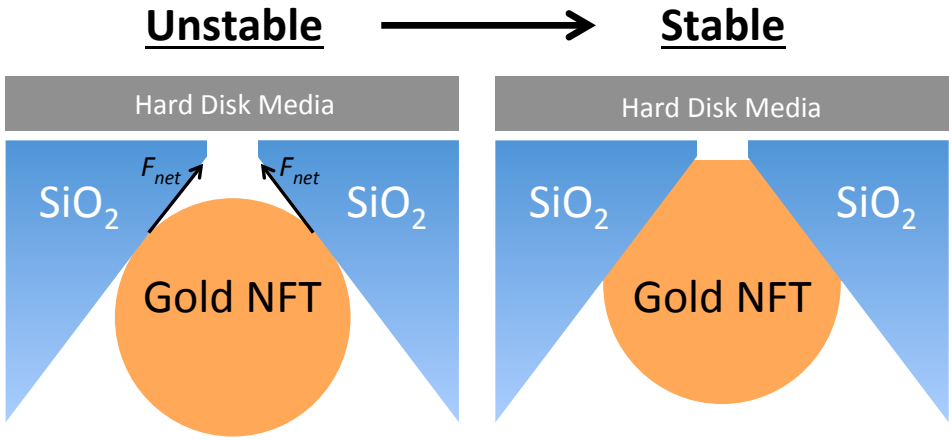

Figure 30: Because of wetting and surface tension forces, a sphere far away from the air-bearing surface (ABS) is not the equilibrium shape of a nano- gold NFT. If the glass cavity that the gold is encapsulated is sufficiently angled, then the equilibrium shape is that of a droplet pulled toward the ABS.

3.6 Thermal Gradient in Hard Disk Media

Since the goal of HAMR is increase data capacity per unit of hard disk drive, understanding the specifications for increasing the data density is very important. Recalling the relationship from (2. 5), the linear recording density is proportional to the downtrack thermal gradient. Specifically, the crucial metric is the temperature fall per nanometer at the Curie point contour of the hotspot near the magnetic write-pole as labeled in Figure 31. This implies that the peak temperature must actually be above the Curie point in order to achieve a high gradient. In the example shown here, the peak temperature is 800K and Curie point is 750K. In real media, the Curie point is not a single value but rather a distribution, so we actually require a particular downtrack gradient for a small temperature range around the Curie point. The crosstrack width and crosstrack gradient determines the width of the written tracks of data and level of noise written to nearby tracks. Hence, the crosstrack gradient determines the track density and the downtrack gradient determines the linear density. Typically, the optimal bit shape for SNR given by a particular read/write head and media has an aspect ratio between 1 and 10 with the track width much larger than the bit length. A typical lollipop antenna on the reference media described earlier offers only 9K/nm downtrack thermal gradient, whereas we need a gradient >15K/nm to achieve the 1Tb/in^2 and higher data densities promised by HAMR.

The main reason that a sharp thermal gradient exists in the hard disk media under HAMR operation is, of course, that light is being focused to a small ~30nm spot on the disk. Since the disk is kept at an ambient temperature around 45°C, the intense heating within the small spot cause a large temperature rise at the peak of the optical hotspot and a sharp temperature gradient toward ambient as the heat diffuses away. For layered hard disk media, it is important to note that the optical absorption is mainly within the FePt grains within a 10nm thin layer. The MgO underlayer and various overcoats are weakly absorbing at 830nm optical excitation. Because the storage layer is granular with oxide separating the many grains, there is a strong anisotropy of heat conduction in the in-plane and perpendicular directions. Typically, the effective thermal conductivity in the plane of the grains is ~0.5W/mK while the conductivity perpendicular into the MgO underlayer is ~5W/mK. This anisotropy is of great benefit to HAMR, because the thermal spot will broaden less, which increases the thermal gradient. However, this is not a variable to be engineered. The only way to increase the thermal

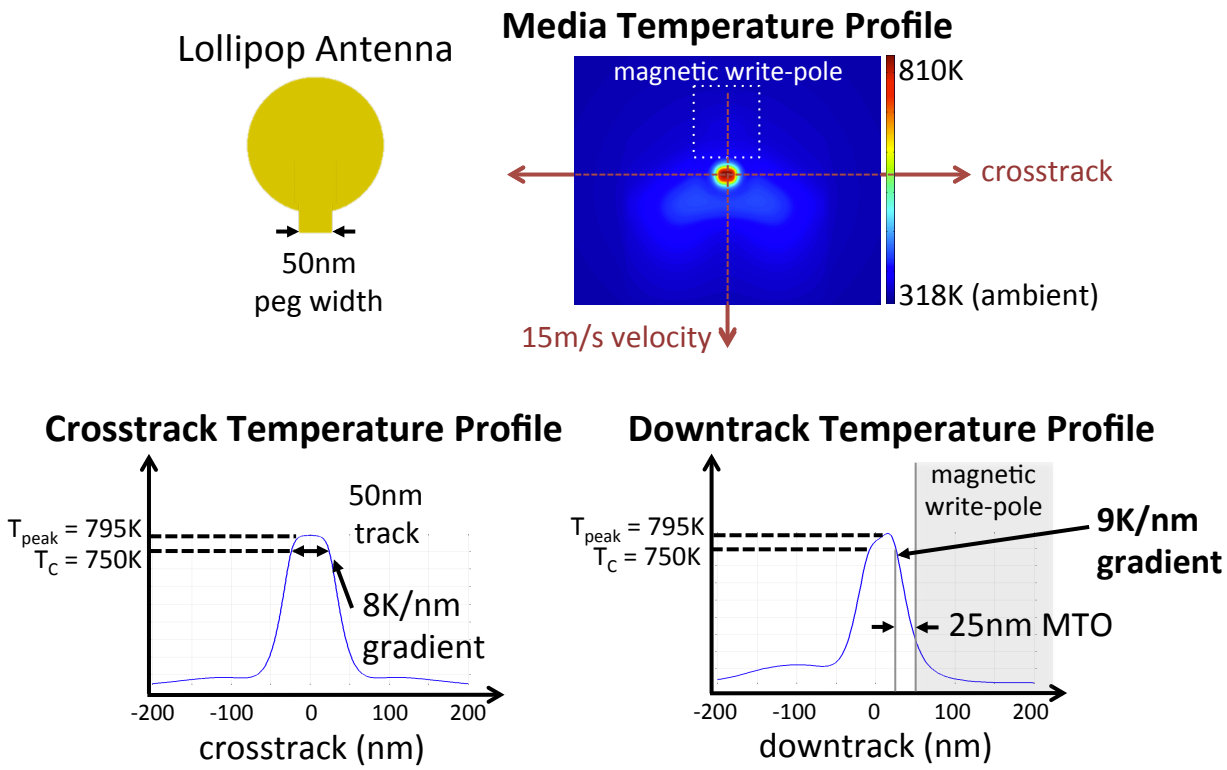

Figure 31: For a Lollipop antenna excited by a PSIM structure, the temperature in the middle cross-section of the FePt recording layer in the hard disk media is shown in the top-right. Linear plots in the crosstrack and downtrack directions are shown on bottom. The thermal gradients are calculated at the contour of the Curie point of ~750K.

conduction anisotropy of the FePt layer is to have larger oxide spacing between the FePt grains, but this would be at the expense of lower grain density (which is undesirable as we need higher density data storage). The main degree of freedom in the media is the structure of the layers above and below the FePt.

3.6.1 Size of Optical Hotspot

To solve this design problem, let us first discuss the optical hotspot shape. Referring back to Figure 16, the intense electric field in the FePt is due to the high charge density induced in the sharp NFT tip via optical excitation and the image charge density induced in the metal heatsink (like a typical RF antenna on a ground plane). The charge density in the NFT tip can be higher if the input optical power delivered to the tip is increased, the NFT tip is made sharper, or if the NFT is made of a more conductive material. The optical coupling efficiency of the NFT can indeed be improved through better NFT design (will be discussed in Chapters 5-7). The latter two variables have already been exhausted by existing NFT designs that consist of gold tips as narrow as 30nm. Alternatively, the charge density in the media heatsink can be increased if the total distance between NFT and heatsink is reduced or if the heatsink is made of more conductive material, as shown in Figure 32 and Figure 33. The distance between the NFT and heatsink can be modulated by reducing the thickness of the MgO underlayer. Both of these effects increase the image charge density, thereby making the electric field more intense and more confined. The combined effect is a higher optical gradient in the FePt.

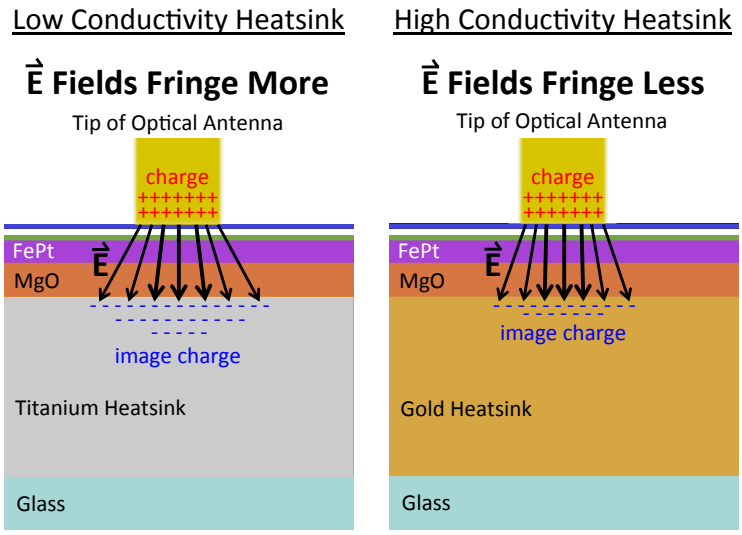

Figure 32: The effect on electric field in the optical hotspot from the conductivity of the media heatsink.

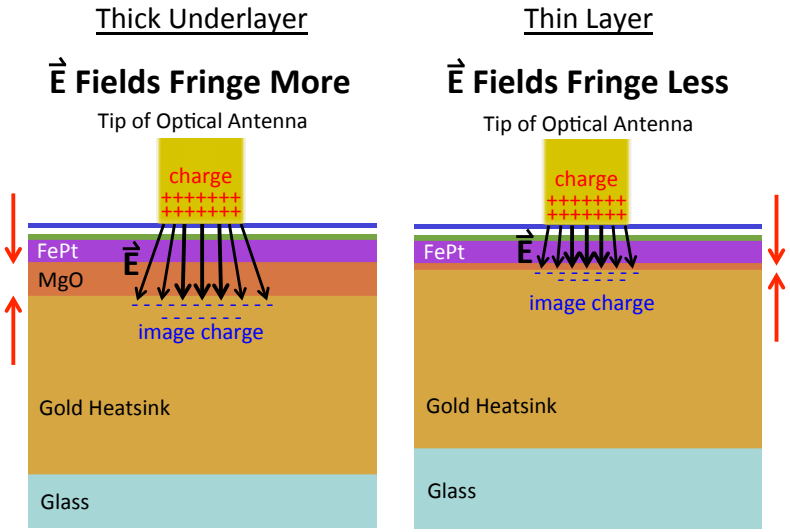

Figure 33: The effect on electric field in the optical hotspot from the underlayer thickness.

3.6.2 Broadening of Thermal Hotspot

Next, let us discuss the thermal properties of the media stack that can be engineered. We already discussed that the thermal conduction anisotropy within the FePt granular layer cannot be higher without sacrificing areal density. Limiting the broadening of the thermal spot can be achieved with the same strategy of a high conductivity heatsink and thin underlayer as we just discussed. We can decrease the thermal blooming of the hotspot and increase the sharpness of the thermal gradient by reducing the underlayer thickness as shown in Figure 34. With the heatsink closer, heat in the FePt dissipates more quickly into the hard disk, which effectively increases the thermal conduction anisotropy of the FePt.

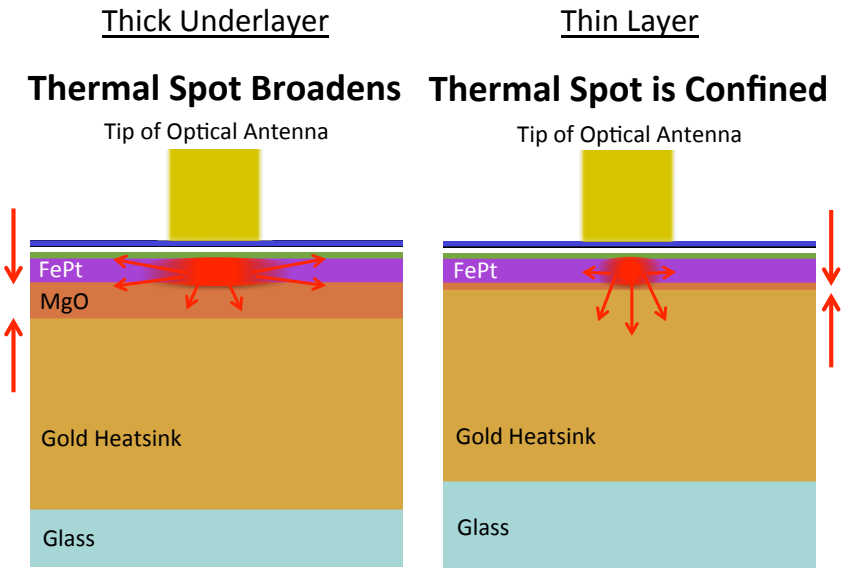

Figure 34: The effect on thermal hotspot broadening from the underlayer thickness.

The last comment pinpoints the crucial trade-off in HAMR. Sharper thermal gradient can be achieved at the expense of cooling the media faster, which means that the media/NFT temperature ratio will decrease. With traditional thin NFTs like the lollipop antenna, which nominally operates at a high temperature, experiments with media containing optically and thermally conductive heatsinks would show a faster failure time of NFTs. Because the failure rate in most HAMR experiments has been orders of magnitudes below the desirable 5 year lifetime, a myth in the hard disk industry developed that we cannot tolerate high conductivity heatsinks in the media.

However, there is no alternative method to achieve higher thermal gradients and higher areal density via HAMR. Instead, we must look back at Section 3.3 and decide that we must compensate and increase the media/NFT temperature ratio via an aggressive heatsink on the NFT tip. An aggressive heatsink contains s large solid angle of heat conduction from the NFT tip. The media must look like that in the right frame of Figure 34, in which there is a heatsink very close to the FePt (perhaps 5nm below) with high electrical and thermal conductivities (perhaps gold). The NFT must not look like the traditional skinny lollipop antenna but rather the fat NFT proposed in this dissertation, shown in Figure 35. The need for large solid angle of heat conduction was the third key conclusion from Section 3.3.1, and the implementation of the fat NFT will be discussed in Chapters 5-7.

To summarize, to achieve high thermal gradient and low NFT temperature:

1) A media heatsink with high electrical and thermal conductivities must be close to the FePt storage layer.

2) The NFT must be fat with a large solid angle of heat conduction from the NFT tip.

3) The NFT and excitation waveguide must have a high optical coupling efficiency.

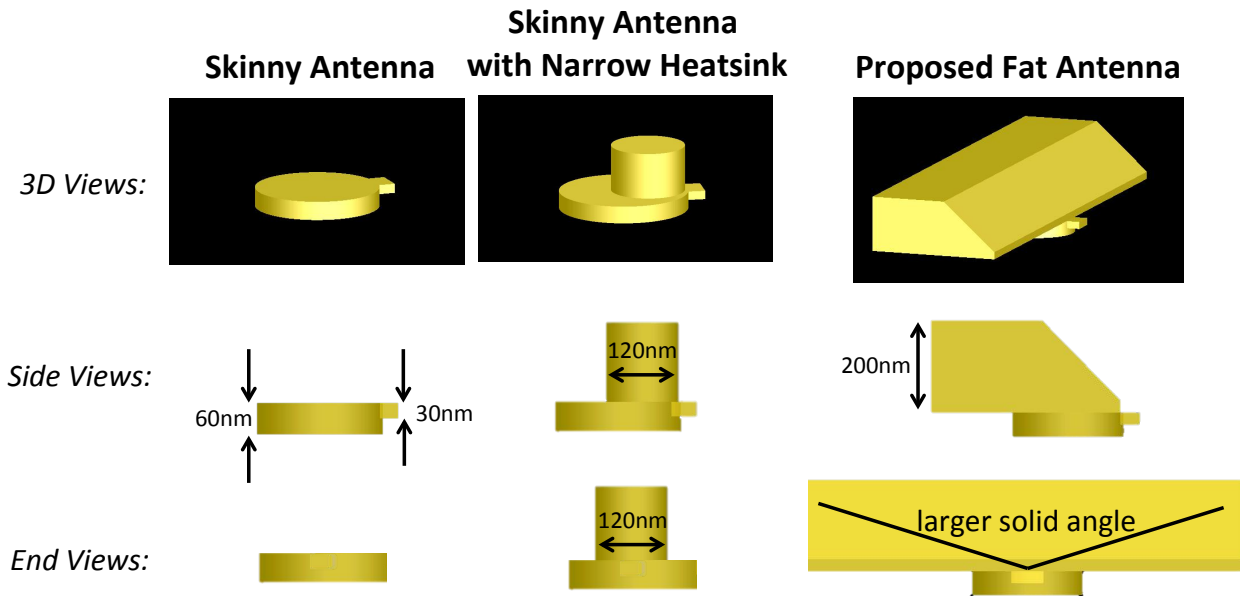

Figure 35: On the left are typical skinny NFTs used in HAMR. The NFT is typically a thin film of metal with a narrow heatsink touching only the 'deadspot' of the resonant body. We propose a fat NFT on the right, in which the entire NFT body directly touches bulk gold and there is huge solid angle of heat conduction from the NFT tip.

3.7 Scaling Strategy for HAMR

In the typical scaling strategy for hard disk drives discussed in Section 2.2, there must be advancements to manufacturing and metallurgy to enable simultaneous improvements to the media (to increase areal density) and to the head (to read/write to the smaller bits). For the future HAMR HDDs, there are new scaling trends needed to ensure that HAMR can provide a decade of areal density growth. To improve areal density, we need higher thermal downtrack and crosstrack gradients in the FePt storage layer of the disk. If we already have an optimized optical nano-focusing system, then there are two ways of further increasing the gradient: a) increase the peak intensity of the disk hotspot via reducing the NFT peg dimensions, and b) increasing the vertical heat conduction out of the FePt to reduce lateral thermal blooming. Both of these methods will reduce the crucial temperature ratio, $\Delta T_{media}/\Delta T_{NFT}$, which in turn reduces the reliability and robustness of the NFT. Thus, NFT reliability will incessantly be a barrier to overcome in every future scaled HAMR product, not just a problem for the generation 1 product. Thus, we must also have a method to keep lowering ΔT_{NFT} and improving the NFT peg reliability. As discussed in Sections 3.3 – 3.5, increasing the NFT solid angle reduces the NFT temperature and further leverages surface energy forces to keep the gold boundary in place. Thus, it may be possible to follow a new scaling trend in which we aim to constantly improve thermal gradient and improve the NFT cone angle to obtain higher areal density while striving to maintain or increase $\Delta T_{media}/\Delta T_{NFT}$. There are incremental improvements to manufacturing and metallurgy processes that can enable this process as described in Figure 37.

In the proposed fat NFT shown in Figure 35, we increased the solid angle of the NFT body. In order to increase the media-NFT temperature ratio even farther and to leverage surface tensions forces, then we must also make the NFT tip conical as well as shown in Figure 36. In all the designs presented in this dissertation (and most of the published and patented material

on HAMR), the NFT tip is a rectangular extruded rod with essentially zero solid angle. The bottleneck preventing a conical NFT tip is the nano-precision lapping process that is used to define the air-bearing surface of the head chip. ~1 million heads are fabricated per wafer in an enormous array. The wafer is diced into chiplets containing the head. Each head has one edge that is to become to air-bearing surface. Exactly on that edge, the tip of the NFT, magnetic write pole and GMR reader must be exposed. In order to expose those elements with nano-precision, industry uses a nano-grinding process with a feedback loop of measuring resistance through alignment markers as depicted in Figure 38. This lapping may have a typical variation of ~10 nm. For a half cone angle of 40°, a 10 nm variation in peg length leads to ~17 nm variation in peg width which is grossly unacceptable. A reasonable peg width variation may be ~5 nm, which yields a max NFT cone angle of 14° given a 10 nm variation in peg length. If the lapping tolerance was improved to 5 nm, then the NFT cone angle could be increased to ~27°. Other processing improvements in deposition and etching techniques may also be required to increase the vertical tapering of the NFT peg to make the solid angle wider in the vertical direction as well.

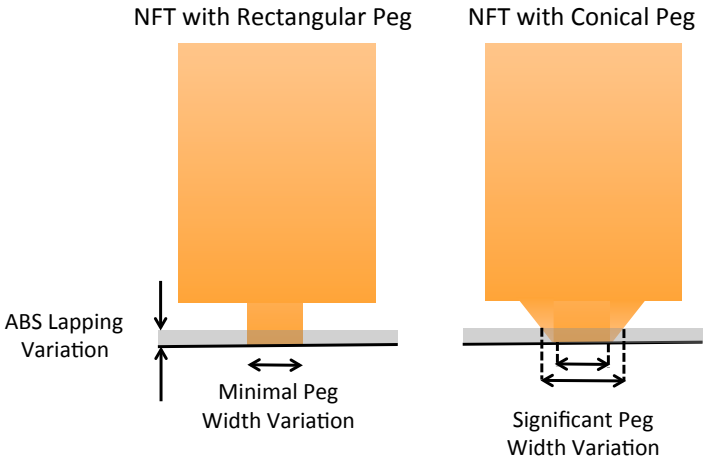

Figure 36: In addition to a fatter NFT body, we also desire the NFT peg itself to be conical rather than rectangular. A conical NFT peg may improve reliability but may add significant variation to the peg width due to variation in ABS position from the lapping process.

Second, the bottleneck preventing higher thermal gradients is very similar to the current challenges and material science accomplishments in scaling the media. Higher thermal gradients can be achieved with a high conductivity heatsink with a very thin MgO underlayer to the FePt granular layer. Depositing good quality FePt grains on thin underlayers on never-before-used substrates like MgO on a gold media heatsink is certainly a huge challenge. Nevertheless, the scaling strategy on the media side is the incrementally decrease the MgO thickness (as well as continued R&D into alternative media heatsinks with higher thermal and optical conductivities).

Before scaling of HAMR can be accomplished, there are numerous manufacturing challenges that must be addressed to commercialize a Generation 1 product: good quality FePt deposition on a thin underlayer on a high conductivity metal, single-mode alignment of the laser, precise nano-patterning of the NFT tip, high precision nano-grinding of the ABS, and deposition of high-temperature-robust overcoats on the head and media. To meet the global storage demand, the hard disk industry must overcome these challenges and produce >6 million heads and >3 million disks per day.

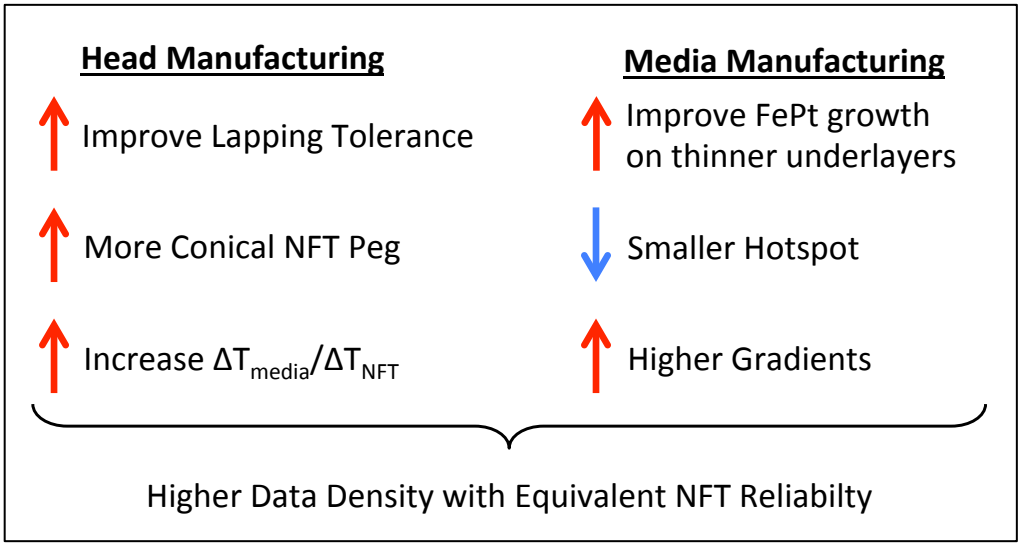

Figure 37: Scaling strategy for HAMR to increase areal density while maintaining the reliability and low operating temperature of the nano- metallic NFT.

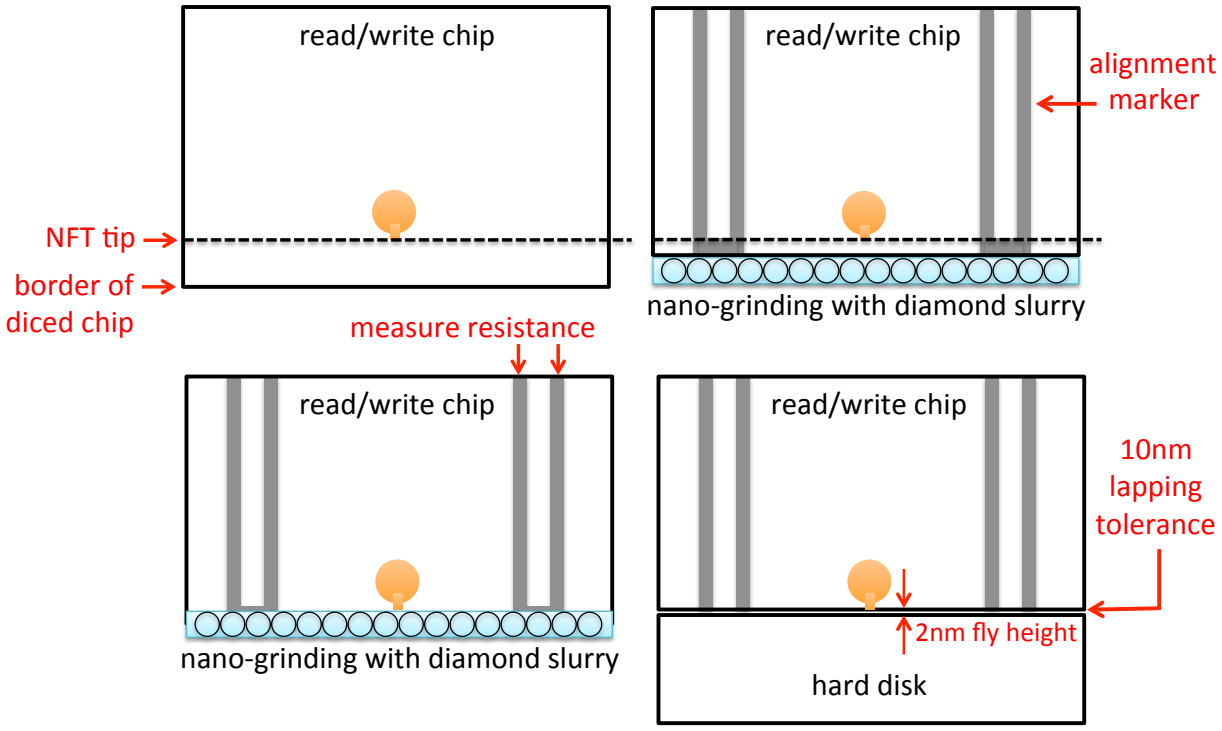

Figure 38: The lapping process to define the air-bearing surface of the head. Nano-grinding with feedback by measuring resistance through an alignment marker loop allows for exposing a surface within a chip with precision of ~10 nm. Precision is determined by slurry consistency and lithography precision.

4 Inverse Electromagnetic Design

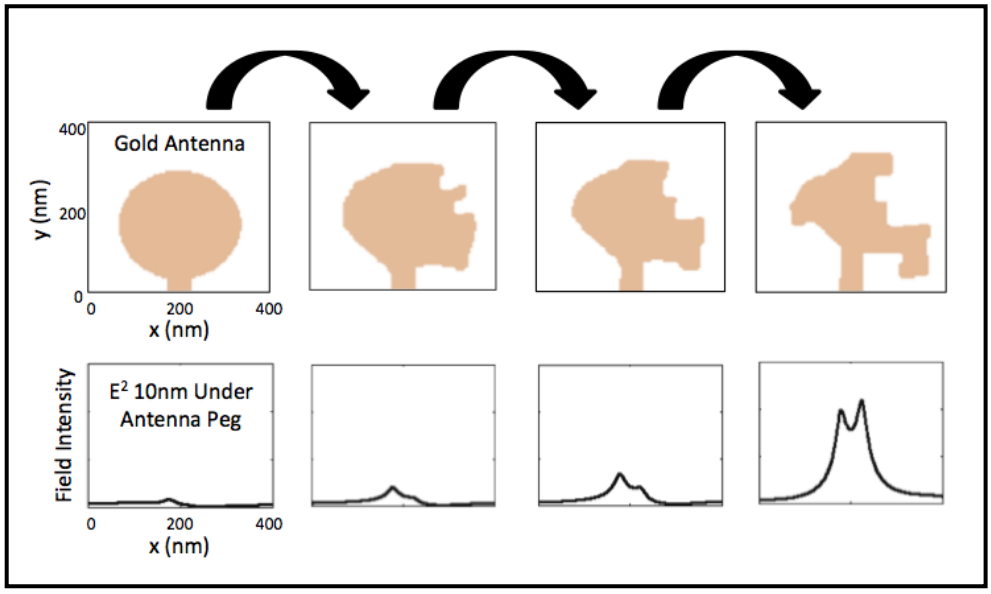

Figure 39: An Iterative and Creative Inverse Design of an Optical Antenna's shape. The antenna in the rightmost frame is not likely to be designed without computation and achieves significantly stronger field localization.

Although intuitive shapes like circles with few degrees of freedom may be simple to understand and optimize, the complicated wave-nature of light compels us to ask whether an unintuitive shape may be better. Since one cannot analytically solve or inversely solve Maxwell's equations, electromagnetic designers and researchers are typically restricted to ponder and dream about a magical shape that may cure their system challenges. A more realizable approach is to solve the inverse problem by computationally optimizing 3D electromagnetic structures with thousands to millions of geometric degrees of freedom. Many degrees of freedom are required for the optimization to be 'creative' and allow convergence to structures not fed by an engineer's intuition. To inversely solve for the nano-optics system for HAMR, we are confined to full wave optics simulations (not geometric or Fourier optics) in which we discretize space and solve the coupled differential equations in the entire 3D volume. A typical HAMR model may take 2 hours on a high-performance computing cluster. Because we cannot afford to sample the enormous parameter space, we developed a gradient-based

optimization algorithm for wave-optics, dubbed Inverse Electromagnetic Design. This computational algorithm is capable of automatically designing 3D optimal geometries of dielectric or metal objects governed by Maxwell's equations. Applications include designing antenna shapes to efficiently deliver optical energy to sub-wavelength spots [30], designing textures for optimal light-trapping in sub-wavelength thick solar cells [31], designing efficient couplers between waveguides and devices in integrated photonics [32], and designing lithography masks [33].

Inverse Electromagnetic Design is based on two concepts: a) parameter-free 'freeform' boundary representation, and b) gradient-based optimization via the adjoint method to efficiently optimize the freeform shape. In the context of HAMR, important electromagnetic figures of merit include the optical absorption in the media hotspot, the ratio of absorption in the media versus the NFT, and the ratio of absorption in the hotspot versus secondary unwanted hotspots in the media. The gradient is the derivative of the chosen Figure of Merit with respect to all of the geometric parameters, which may be the shape boundaries of the NFT and waveguide structures in the HAMR write-head. The gradient allows us to use a fast deterministic optimization algorithm like steepest descent., in which we iteratively approach a local optimum. In contrast, heuristic methods like genetic algorithms and particle-swarm optimizations rely on random trials, which can be useful for problems where the simulation of the physics is very fast. However, it is too cumbersome for applications in which a single 3D simulation of Maxwell's equations is only feasible on high-performance computing resources.

The simplest but most inefficient method to calculating the gradient is finite-difference, in which we modulate every parameter by a small perturbation. This requires at least N simulations to calculate N independent derivatives, where N is the number of parameters in the system. Calculating the gradient with finite-difference for a complex shape with 1000 geometric parameters where each 3D simulation of Maxwell's Equations takes one hour on a High-Performance Computing Cluster is completely unfeasible. Because both gradient and heuristic optimization methods are impractical, current designers in RF and Optics often limit themselves to simple structures like circles and rectangles. Often, designers optimize in a brute force fashion via parameter sweeps on geometries with few degrees of freedom like the C-aperture antenna shown in Figure 40. Adding more degrees of freedom may allow for improved structures but the parameter space increases exponentially. The goal of the proposed Inverse Design method is to break this trade off and greatly reduce the computation required to optimize non-parametric geometries (when N tends to infinity). With our ability to manufacture devices with sub-wavelength features and new applications of metal optics today, we no longer should restrict ourselves to simple structures.

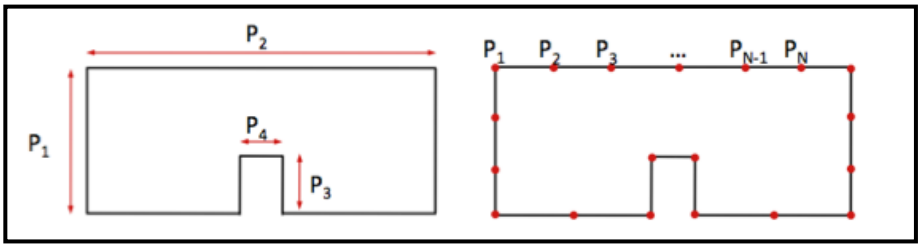

Figure 40: A C-Aperture Antenna can be represented by 4 (left) or N (right) geometric parameters. If confined to a design methodology of parametric sweeps, then more parameters offer more degrees of freedom at the expense of exponentially more intensive computation.

The proposed Inverse Electromagnetic Design algorithm is described as Adjoint-Based Gradient Descent. Gradient Descent may refer to any hill climbing method including

Newton's Method, which uses the second derivative in addition to the first. To find a local maximum or minimum, one simply needs to calculate an instantaneous derivative of a Figure of Merit (FOM) with respect to a Variable and iteratively increment or decrement the Variable until the derivative is zero. A completely freeform shape can be described by having N Variables distributed around the shape's perimeter, where N tends to infinity for an increasingly continuous representation of the shape's boundary, like that in the right frame of Figure 40. Adjoint-Based refers to using reciprocity in electromagnetics (or more generally the dual method in linear algebra) to calculate the derivative of the FOM with respect to all N Variables in an efficient manner. For example, to characterize an antenna, there are two symmetric simulations of Maxwell's Equations. The first involves the antenna being illuminated from the far-field and the localized electric field is calculated. The second involves the antenna being excited in the near-field and the far-field radiation is calculated. The significant realization is that only these 2 simulations of Maxwell's equations are required to calculate the gradient along the entire shape's boundary, regardless of the number of parameters. Figure 41 shows one iteration of the iterative gradient-based method. Positive (red) and negative (blue) gradients indicate where the boundary should be pushed outward or inward in order to navigate toward a local optimum in the shape parameter space. Converging with thousands or millions of parameters has challenges as well which was addressed in this work via quasi-Newton gradient descent in which we also approximate the second derivative.

An early result of freeform inverse design in electromagnetics is shown in Figure 39. This figure shows an optimization of a 2D gold nano-antenna illuminated with a Gaussian beam of wavelength 830 nm at an incoming angle of 45° to the left. The objective function was electric field intensity 10 nm below the antenna tip. The antenna boundary iteratively converged to geometry in the right-most frame that resembles a person sitting in a chair. This highly unintuitive structure would never have been designed by intuition or analytic calculation, yet it offers significantly better performance.

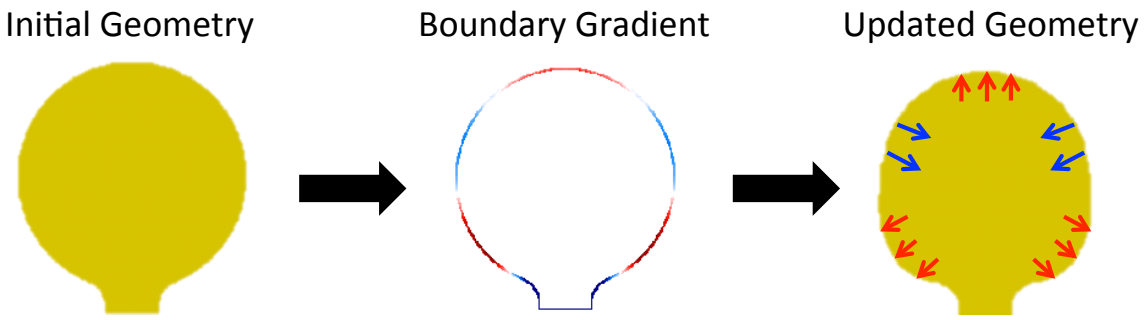

Figure 41: The gradient along the boundary of an object allows for iterative deterministic optimization of freeform shapes. The arrows indicate an iterative change to the boundary.

The dual or adjoint method is a commonplace technique in linear systems. In electromagnetics, a common application is to efficiently compute the sensitivity of an objective function with respect to parameters [34][35][36]. The sensitivity of the scattering from metallic objects was specifically studied by Dadash et. al. [37] and they derived similar expressions to those shown in Section 4.8. This technique has also been used as an image reconstruction method in quasi-static electromagnetic probing under the ground or ocean [38]. Using the adjoint method specifically for engineering design has been performed for a long time in mechanical engineering especially in fluid dynamics [39].

4.1 Dual Method in Linear Algebra

To avoid the implication that the mathematical methods used in this dissertation are unique to Maxwell's equations, it is worth discussing the general dual method in linear algebra. This dissertation applies this very general property of linear systems to electromagnetics.

Consider the simple problem of calculating the objective function F in (4. 1). The inherent inefficiency in this system is that F is a vector, but we had to solve for the larger matrix $\boldsymbol{X}$ in the process. Through simple manipulation in (4. 2), we can group the calculations differently. After a substitution of variables with a dummy vector $\boldsymbol{s}$, we arrive at the dual problem in (4. 3). Rather than $\boldsymbol{X}$, we must first solve for the vector $\boldsymbol{s}$ in order to calculate F. If $\boldsymbol{s}$ is dimensionally smaller than $\boldsymbol{X}$, then the dual method is computationally more efficient. If s and $\boldsymbol{X}$ have the same dimension, then the dual method is not clearly better or worse. In either case, the fact that every linear problem has a dual problem is a key phenomenon.

$$F = \boldsymbol{g}^T \boldsymbol{X} \;\; such\ that \;\; \boldsymbol{A}\boldsymbol{X} = \boldsymbol{B} \tag{4. 1}$$

$$F = \boldsymbol{g}^T(\boldsymbol{A}^{-1}\boldsymbol{B}) \; = (\boldsymbol{g}^T\boldsymbol{A}^{-1})\boldsymbol{B} = ((\boldsymbol{A}^{-1})^T\boldsymbol{g})^T\boldsymbol{B} = ((\boldsymbol{A}^T)^{-1}\boldsymbol{g})^T\boldsymbol{B} \tag{4. 2}$$

$$F = \boldsymbol{s}^T \boldsymbol{B} \;\; such\ that \;\; \boldsymbol{A}^T\boldsymbol{s} = \boldsymbol{g} \tag{4. 3}$$

4.2 Dual Method to Efficiently Calculate the Gradient

In a similar fashion, it is commonplace to use the dual method to calculate the gradient with respect to input parameters in a linear system with dramatically less computation compared to finite-difference. Consider the system in (4. 4) where F is the objective function. We wish to calculate the derivative of F with respect to parameter ξ_i on which system matrix $\boldsymbol{A}$ is dependent. If we differentiate both expressions with respect to ξ_i, then we arrive at (4. 5). The system problem is dependent on x, so we first must solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ to calculate x. Next, if we have many parameters ($i \in [1, N]$), then we would need to solve the system, $\boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial \xi_i} = \left(\frac{\partial \boldsymbol{b}}{\partial \xi_i} - \frac{\partial \boldsymbol{A}}{\partial \xi_i}\boldsymbol{x}\right)$, N times to obtain N independent derivatives, $\frac{\partial \boldsymbol{x}}{\partial \xi_i}$. So, in total, we require $N + 1$ solves to calculate the gradient. Similarly, using finite-difference to calculate the gradient would involve 1 solve to calculate F and N solves to calculate N $\delta F's$ for N independent $\delta \xi_i's$.

$$F = f(\boldsymbol{x}) \;\; such\ that \;\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{4. 4}$$

$$\frac{\partial F}{\partial \xi_i} = \frac{\partial f}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \xi_i} \;\; such\ that \;\; \boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial \xi_i} = \left(\frac{\partial \boldsymbol{b}}{\partial \xi_i} - \frac{\partial \boldsymbol{A}}{\partial \xi_i}\boldsymbol{x}\right) \tag{4. 5}$$

Recognizing that (4. 5) is like (4. 1), we can instead convert to the dual problem in (4. 6). Here, we would only need to solve the system, $\boldsymbol{A}^T\boldsymbol{s} = \frac{\partial f}{\partial \boldsymbol{x}}$, once since it is independent of ξ_i. We still need to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ once to calculate $\boldsymbol{x}$. So, in total, we require 2 solves to calculate

the gradient using the dual method, regardless of the value of N. This is a huge computational saving, reducing the total number of solves from scaling linearly with N to just a constant.

$$\frac{\partial F}{\partial \xi_i} = \boldsymbol{s}^T \left(\frac{\partial \boldsymbol{b}}{\partial \xi_i} - \frac{\partial \boldsymbol{A}}{\partial \xi_i} \boldsymbol{x} \right) \; such\ that \; \boldsymbol{A}^T \boldsymbol{s} = \frac{\partial f}{\partial \boldsymbol{x}} \tag{4. 6}$$

4.3 Dual Method on Linear Operators

Sections 4.1 and 4.2 show the dual method applied to linear algebra problems. To apply this technique to optimization in electromagnetics, we are more interested in the dual method applied to linear operators, which encompasses the various differential operators present in Maxwell's equations or any set of Partial Differential Equations (PDEs). Once again, this section is not a contribution of this dissertation but rather a review of well-known mathematics. Let us consider the objective function and system equation in (4. 7). Here, $\boldsymbol{A}$ is a linear operator, not a matrix of numbers. Very similar to the previous analysis, there exists a dual problem expressed in (4. 8), where $\boldsymbol{A}^*$ is called the adjoint of the operator $\hat{A}$. The star notation here does NOT indicate conjugate transpose. Formally, $\boldsymbol{A}^*$ is defined by (4. 9) for an arbitrary $\boldsymbol{x}$ and $\boldsymbol{y}$. There is no single expression to calculate the adjoint of any arbitrary operator. To construct the dual problem, we have to determine the adjoint of the specific operator in question. If we can determine the adjoint operator, then we can enjoy the same huge computational benefits that were discussed in Sections 4.1 and 4.2. In the following sections, we will derive the adjoint operator for Maxwell's equations.

$$F = \boldsymbol{g}^T \boldsymbol{x} \; such\ that \; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{4. 7}$$

$$F = \boldsymbol{s}^T \boldsymbol{b} \; such\ that \; \boldsymbol{A}^* \boldsymbol{s} = \boldsymbol{g} \tag{4. 8}$$

$$\boldsymbol{A}\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{A}^* \boldsymbol{y} \tag{4. 9}$$

Although we do not wish to introduce more vocabulary, it is worth noting that there are special operators that are self-adjoint, meaning that the adjoint operator is identically the original operator, shown in (4. 10). The self-adjoint property is not necessary to use the dual method as framed above. The benefit of a self-adjoint operator is that the dual problem will use identical computation as a normal forward solution of the original operator.

$$\boldsymbol{A}\ is\ self\ adjoint\ iff\ \boldsymbol{A} = \boldsymbol{A}^* \tag{4. 10}$$

4.4 Reciprocity in Electromagnetics

To apply the adjoint method to electromagnetics, we need to derive the adjoint of the system operator for Maxwell's equations. In electromagnetics, the properties of the adjoint operator are typically discussed as the phenomenon of reciprocity [38]. So, we will first discuss reciprocity in typical fashion. This work was first coherently written by John Henry Poynting in 1883 [40] and Hendrik Lorentz in 1896 [41]. Their derivations essentially apply two basic properties of differential operators shown in (4. 11) and (4. 12) to Maxwell's equations.

$$\nabla \cdot (\boldsymbol{A}\times\boldsymbol{B}) = \boldsymbol{B} \cdot (\nabla\times\boldsymbol{A}) - \boldsymbol{A} \cdot (\nabla\times\boldsymbol{B}) \tag{4. 11}$$

$$\iiint_V (\nabla \cdot \boldsymbol{A})dV = \oiint_S \boldsymbol{A} \cdot \hat{\boldsymbol{n}}dS \tag{4. 12}$$

4.4.1 Poynting's Theorem

Poynting started with the closed surface integral of the normal component of $(\boldsymbol{E}\times\boldsymbol{H})$, which is now called the Poynting vector, as shown in (4. 13). After invoking (4. 11) and (4. 12), we arrive at Poynting's theorem in (4. 14), where the left term is the net energy leaving a closed surface and the right terms are the net energy produced by the electric and magnetic currents, $\boldsymbol{J}$ and $\boldsymbol{J_m}$, within the enclosed volume. The fact that energy is conserved is of course obvious, but Poynting's theorem is a great description of the various forms of energy production or loss in an electromagnetic system. Energy production happens when the currents, $\boldsymbol{J}$ and $\boldsymbol{J_m}$, are out of phase with the respective fields, $\boldsymbol{E}$ and $\boldsymbol{H}$. Energy absorption occurs when the currents are in phase with the fields. It is also common to consider an energy-neutral volume where no energy is escaping the closed surface, in which the left term is zero and we have the relationship in (4. 15). This is also just a statement of electromagnetic energy conservation.

$$\oiint_S (\boldsymbol{E}\times\boldsymbol{H}) \cdot \hat{\boldsymbol{n}}dS = \iiint_V \big(\nabla \cdot (\boldsymbol{E}\times\boldsymbol{H})\big)dV = \iiint_V \big(\boldsymbol{H} \cdot (\nabla\times\boldsymbol{E}) - \boldsymbol{E} \cdot (\nabla\times\boldsymbol{H})\big)dV \tag{4. 13}$$

$$\oiint_S (\boldsymbol{E}\times\boldsymbol{H}) \cdot \hat{\boldsymbol{n}}dS = \iiint_V (-\boldsymbol{H} \cdot \boldsymbol{J_m} - \boldsymbol{E} \cdot \boldsymbol{J})dV \tag{4. 14}$$

$$0 = \iiint_V (\boldsymbol{H} \cdot \boldsymbol{J_m} + \boldsymbol{E} \cdot \boldsymbol{J})dV \tag{4. 15}$$

4.4.2 General Reciprocity in Electromagnetics

Following in Poynting's footsteps, Lorentz discussed a reciprocity relationship between electric currents and electric fields. Using his methods, we will derive all 3 important reciprocity relationships in electromagnetics. Let us consider two different systems denoted by subscripts 1 and 2. Each system contains the same geometric structures and materials, but have different electric and magnetic currents that produce different electric and magnetic fields. Using the same steps as Poynting, we can show the relationships in (4. 16) and (4. 17). By the subtraction of these two equations, we derive (4. 18). If only far-field waves propagate through this closed surface, then the subtraction term on the left is zero and we derive (4. 19). (4. 20) is merely a rearrangement of the equation, which is preferred by the author. In his 1896 paper, Lorentz published (4. 20) with $\boldsymbol{J_{m1}}$ and $\boldsymbol{J_{m2}}$ equals zero.

$$\oiint_S (\boldsymbol{E_1}\times\boldsymbol{H_2}) \cdot \hat{\boldsymbol{n}}dS = \iiint_V (-\boldsymbol{H_2} \cdot \boldsymbol{J_{m1}} - \boldsymbol{E_1} \cdot \boldsymbol{J_2})dV \tag{4. 16}$$

$$\oiint_S (\boldsymbol{E_2}\times\boldsymbol{H_1}) \cdot \hat{\boldsymbol{n}}dS = \iiint_V (-\boldsymbol{H_1} \cdot \boldsymbol{J_{m2}} - \boldsymbol{E}_2 \cdot \boldsymbol{J_1})dV \tag{4. 17}$$

$$\oiint_S (\boldsymbol{E_1}\times\boldsymbol{H_2} - \boldsymbol{E_2}\times\boldsymbol{H_1}) \cdot \hat{\boldsymbol{n}} dS = \iiint_V (\boldsymbol{H_1}\cdot\boldsymbol{J_{m2}} - \boldsymbol{H_2}\cdot\boldsymbol{J_{m1}} + \boldsymbol{E_2}\cdot\boldsymbol{J_1} - \boldsymbol{E_1}\cdot\boldsymbol{J_2}) dV \quad (4.\ 18)$$

$$0 = \iiint_V (\boldsymbol{H_1}\cdot\boldsymbol{J_{m2}} - \boldsymbol{H_2}\cdot\boldsymbol{J_{m1}} + \boldsymbol{E_2}\cdot\boldsymbol{J_1} - \boldsymbol{E_1}\cdot\boldsymbol{J_2}) dV \quad (4.\ 19)$$

$$\iiint_V (\boldsymbol{E_2}\cdot\boldsymbol{J_1} - \boldsymbol{H_2}\cdot\boldsymbol{J_{m1}}) dV = \iiint_V (\boldsymbol{E_1}\cdot\boldsymbol{J_2} - \boldsymbol{H_1}\cdot\boldsymbol{J_{m2}}) dV \quad (4.\ 20)$$

4.4.3 Reciprocity of an Electric Dipole to Electric Field

If we consider system 1 to only contain an electric current source at $\boldsymbol{x'}$ and system 2 to only contain an electric current source at $\boldsymbol{x}$, then we derive the reciprocal relationship shown in (4. 21). This relationship is also written here in the form of a Green's function, which is simply a transfer function.

$$\boldsymbol{E_2}(\boldsymbol{x'})\cdot\boldsymbol{J_1}(\boldsymbol{x'}) = \boldsymbol{E_1}(\boldsymbol{x})\cdot\boldsymbol{J_2}(\boldsymbol{x}) \leftrightarrow \boldsymbol{G^{EJ}}(\boldsymbol{x},\boldsymbol{x'}) = \boldsymbol{G^{EJ}}(\boldsymbol{x'},\boldsymbol{x}) \quad (4.\ 21)$$

4.4.4 Reciprocity of an Electric Dipole to Magnetic Field

If we consider system 1 to only contain an electric current source at $\boldsymbol{x'}$ and system 2 to only contain a magnetic current source at $\boldsymbol{x}$, then we derive the reciprocal relationship shown in (4. 22). This relationship is once again also written here in the form of Green's functions. This relationship also covers the case of magnetic dipole to electric field, so we will not re-write that as a separate case.

$$\boldsymbol{E_2}(\boldsymbol{x'})\cdot\boldsymbol{J_1}(\boldsymbol{x'}) = -\boldsymbol{H_1}(\boldsymbol{x})\cdot\boldsymbol{J_{m2}}(\boldsymbol{x}) \leftrightarrow \boldsymbol{G^{HJ}}(\boldsymbol{x},\boldsymbol{x'}) = -\boldsymbol{G^{EM}}(\boldsymbol{x'},\boldsymbol{x}) \quad (4.\ 22)$$

4.4.5 Reciprocity of an Magnetic Dipole to Magnetic Field

If we consider system 1 to only contain a magnetic current source at $\boldsymbol{x'}$ and system 2 to only contain a magnetic current source at $\boldsymbol{x}$, then we derive the reciprocal relationship shown in (4. 23). This relationship is once again also written here in the form of Green's functions.

$$-\boldsymbol{H_2}(\boldsymbol{x'})\cdot\boldsymbol{J_{m1}}(\boldsymbol{x'}) = -\boldsymbol{H_1}(\boldsymbol{x})\cdot\boldsymbol{J_{m2}}(\boldsymbol{x}) \leftrightarrow \boldsymbol{G^{HM}}(\boldsymbol{x},\boldsymbol{x'}) = \boldsymbol{G^{HM}}(\boldsymbol{x'},\boldsymbol{x}) \quad (4.\ 23)$$

4.5 The Adjoint Operator for Maxwell's Equations

Steady-state solutions to Maxwell's equations for a given frequency of oscillation can be written in the form of (4. 24). The electric fields, $\boldsymbol{E}(\vec{\boldsymbol{r}})$, and magnetic fields, $\boldsymbol{H}(\vec{\boldsymbol{r}})$ are 3D vectors that vary in space and are complex-valued (ie. oscillates in time according to frequency ω). These fields are the results of electric and magnetic sources $\boldsymbol{J}(\vec{\boldsymbol{r}})$ and $\boldsymbol{J_m}(\vec{\boldsymbol{r}})$, which are also complex-valued 3D vectors as a function of 3D space. These sources produce the resultant fields according to the operator that is shown that represents Maxwell's equations. ϵ is the

complex electrical permittivity as a function of space, and μ is the magnetic permeability as a function of space. These material property functions fully describe all of the structures inside the 3D domain. In the interest of readability, we will no longer explicitly write $(\vec{r})$ and assume that the reader understands that all of these quantities are functions of space.

$$\begin{bmatrix} -i\omega\epsilon(\vec{r}) & \nabla(\vec{r})\times \\ \nabla(\vec{r})\times & i\omega\mu(\vec{r}) \end{bmatrix} \begin{bmatrix} \boldsymbol{E}(\vec{r}) \\ \boldsymbol{H}(\vec{r}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}(\vec{r}) \\ -\boldsymbol{J_m}(\vec{r}) \end{bmatrix} \tag{4. 24}$$

If we write the general reciprocity statement from (4. 20) as dot products of vectors, then we have (4. 25). If we substitute the left term of (4. 24) for the current sources, then we obtain (4. 26). This equation is in the same form as the definition of an adjoint matrix from (4. 9). Thus, we have found the adjoint of the Maxwell's equations operator, and it happens to be self-adjoint as noted in (4. 27).

$$\begin{bmatrix} \boldsymbol{J}_1 \\ -\boldsymbol{J_{m_1}} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{E}_2 \\ \boldsymbol{H}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_2 \\ -\boldsymbol{J_{m_2}} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{E}_1 \\ \boldsymbol{H}_1 \end{bmatrix} \tag{4. 25}$$

$$\begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \boldsymbol{E}_1 \\ \boldsymbol{H}_1 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{E}_2 \\ \boldsymbol{H}_2 \end{bmatrix} = \begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \boldsymbol{E}_2 \\ \boldsymbol{H}_2 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{E}_1 \\ \boldsymbol{H}_1 \end{bmatrix} \tag{4. 26}$$

$$\begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix}^* = \begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \tag{4. 27}$$

4.6 Adjoint-Based Gradient Calculation in Electromagnetics

Consider the objective function F that is a function of electric fields and magnetic fields as written in (4. 28), which are subject to the partial differential equations in (4. 29).

$$F = f(\boldsymbol{E}, \boldsymbol{H}) \tag{4. 28}$$

$$\begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J} \\ -\boldsymbol{J_m} \end{bmatrix} \ (Forward\ Simulation) \tag{4. 29}$$

The total derivative of F with respect to parameter ξ_i is shown in (4. 30), and differentiating (4. 29) leads to (4. 31). We notice that (4. 31) has the same form as the original system equation. Hence, we can re-write (4. 31) as (4. 32), where $\frac{\partial \boldsymbol{E}}{\partial \xi_i}$ and $\frac{\partial \boldsymbol{H}}{\partial \xi_i}$ are the fields produced by the effective sources, $\boldsymbol{J}_{eff,i}$ and $\boldsymbol{J}_{\boldsymbol{m}_{eff,i}}$. These effective sources are defined in (4. 33). And, we can re-write (4. 30) in terms of these new field quantities. In this form, in order to calculate the gradient of F, we must solve (4. 32) N times if we have N parameters ($i \in [1, N]$).

$$\frac{\partial F}{\partial \xi_i} = \frac{\partial f}{\partial \boldsymbol{E}} \cdot \frac{\partial \boldsymbol{E}}{\partial \xi_i} + \frac{\partial f}{\partial \boldsymbol{H}} \cdot \frac{\partial \boldsymbol{H}}{\partial \xi_i} \tag{4. 30}$$

$$\begin{bmatrix} -i\omega\dfrac{\partial \epsilon}{\partial \xi_i} & 0 \\ 0 & i\omega\dfrac{\partial \mu}{\partial \xi_i} \end{bmatrix} \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{bmatrix} + \begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \dfrac{\partial \boldsymbol{E}}{\partial \xi_i} \\ \dfrac{\partial \boldsymbol{H}}{\partial \xi_i} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \boldsymbol{J}}{\partial \xi_i} \\ -\dfrac{\partial \boldsymbol{J_m}}{\partial \xi_i} \end{bmatrix} \tag{4. 31}$$

$$\begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \dfrac{\partial \boldsymbol{E}}{\partial \xi_i} \\ \dfrac{\partial \boldsymbol{H}}{\partial \xi_i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_{eff,i} \\ -\boldsymbol{J}_{\boldsymbol{m}_{eff,i}} \end{bmatrix} \tag{4. 32}$$

$$\begin{bmatrix} \boldsymbol{J}_{eff,i} \\ -\boldsymbol{J}_{\boldsymbol{m}_{eff,i}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \boldsymbol{J}}{\partial \xi_i} + i\omega \dfrac{\partial \epsilon}{\partial \xi_i} \boldsymbol{E} \\ -\left(\dfrac{\partial \boldsymbol{J}_{\boldsymbol{m}}}{\partial \xi_i} + i\omega \dfrac{\partial \mu}{\partial \xi_i} \boldsymbol{H} \right) \end{bmatrix} \tag{4. 33}$$

$$\frac{\partial F}{\partial \xi_i} = \frac{\partial f}{\partial \boldsymbol{E}} \cdot \frac{\partial \boldsymbol{E}}{\partial \xi_i} + \frac{\partial f}{\partial \boldsymbol{H}} \cdot \frac{\partial \boldsymbol{H}}{\partial \xi_i} \tag{4. 34}$$

Instead, we can construct the dual problem as written in (4. 35) and (4. 36). $\boldsymbol{E}_{\boldsymbol{A}}$ and $\boldsymbol{H}_{\boldsymbol{A}}$ are the dummy variables equivalent to s in Section 4.1. We will refer to $\boldsymbol{E}_{\boldsymbol{A}}$ and $\boldsymbol{H}_{\boldsymbol{A}}$ **as** the 'Adjoint' fields, which are the resultant fields produced by sources $\boldsymbol{J}_{\boldsymbol{A}}$ and $\boldsymbol{J}_{\boldsymbol{m}_{\boldsymbol{A}}}$ as defined in (4. 37). $\boldsymbol{E}_{\boldsymbol{A}}$ and $\boldsymbol{H}_{\boldsymbol{A}}$ are independent of ξ_i and are only dependent on the objective function, $f\left(\boldsymbol{E}(\vec{r}), \boldsymbol{H}(\vec{r})\right)$. Thus, using the dual method, only 2 simulations are needed: the 'Forward' simulation in (4. 29) to determine $\boldsymbol{J}_{eff,i}$ and $\boldsymbol{J}_{\boldsymbol{m}_{eff,i}}$ according to (4. 33), and the 'Adjoint' simulation in (4. 36) to determine $\boldsymbol{E}_{\boldsymbol{A}}$ and $\boldsymbol{H}_{\boldsymbol{A}}$. Because Maxwell's equations are self-adjoint, the forward and adjoint problems are steady-state solutions of electromagnetic fields produced by oscillating electromagnetic sources. Any Maxwell solver including FDTD, FDFD, FIM and FEM that are used for full wave electromagnetics simulations can be used for both of the forward and adjoint solves needed here. Moreover, the observation that the Maxwell's equations operator is self-adjoint is not the root of the computational savings but only the root of the convenience that the forward and adjoint problems can be solved with the same solver. Since this analysis was performed on Maxwell's equations, this analysis can of course be used in Fourier Optics (an approximation to efficiently model far-field light propagation with diffraction) as well as Ray Optics (an approximation to efficiently model non-diffracting far-field light propagation). In this dissertation, we were specifically interested in applying this method to designing 3D nanostructures that can only be characterized by fully modeling Maxwell's equations (Full Wave Optics).

$$\frac{\partial F}{\partial \xi_i} = \boldsymbol{E}_{\boldsymbol{A}} \cdot \boldsymbol{J}_{eff,i} - \boldsymbol{H}_{\boldsymbol{A}} \cdot \boldsymbol{J}_{\boldsymbol{m}_{eff,i}} \tag{4. 35}$$

$$\begin{bmatrix} -i\omega\epsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \boldsymbol{E}_{\boldsymbol{A}} \\ \boldsymbol{H}_{\boldsymbol{A}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_{\boldsymbol{A}} \\ -\boldsymbol{J}_{\boldsymbol{m}_{\boldsymbol{A}}} \end{bmatrix} \ (Adjoint\ Simulation) \tag{4. 36}$$

$$\begin{bmatrix} \boldsymbol{J}_{\boldsymbol{A}} \\ -\boldsymbol{J}_{\boldsymbol{m}_{\boldsymbol{A}}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial \boldsymbol{E}} \\ \dfrac{\partial f}{\partial \boldsymbol{H}} \end{bmatrix} \tag{4. 37}$$

4.7 Gradient Calculation in 3D Wave Optics (intuitive analysis)

Let us consider the application of optimizing the shape of a gold nano-antenna as shown in Figure 42, in which the antenna is illuminated by light. This antenna produces a large near-field light intensity near the antenna's tip from the charge resonance and lightning rod effect shown previously in Figure 13. The objective function noted in (4. 38) is the electric field intensity at the point x_0 near the tip of the antenna. For application to optical nano-focusing, x_0 may be where the load, sample or hard disk media is placed. The simulation of this system is equivalent to solving the forward problem in (4. 29).

$$F = \frac{1}{2}|E(x_0)|^2 \tag{4. 38}$$

Let us consider that the degrees of freedom in the structure to be optimized are the positions of the boundary of gold object. The clumsy way to calculate the gradient, the derivative of F with respect to all of the boundary points, is using finite-difference. As depicted in Figure 43, finite-difference would involve running N separate simulations to model each of N possible boundary pertubations. By taking the difference of F from these N simulations and the original F from Figure 42, we can approximate the gradient of F. If each simulation takes 1 hour on a computing cluster, then calculating the gradient is very expensive and time-consuming.

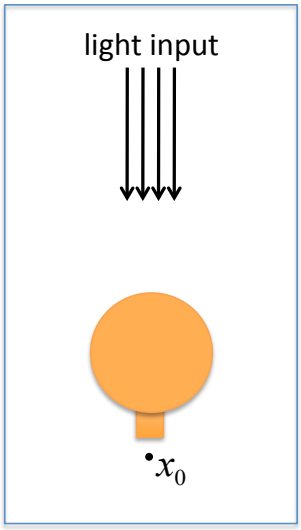

Figure 42: A gold nano-antenna modeled by full wave optics with a light input and an objective function of electric field intensity at x_0.

The first key trick to simplify the size of this optimization problem is to approximate the effects of a perturbation to the shape's geometry. Clearly, as we perturb the geometry, the fields everywhere in space are perturbed as well as shown in Figure 44. The external source induces a polarization in the perturbation (a localized change in permittivity) that oscillates at the electromagnetic frequency. If the perturbation is small enough, then it will only support the dipole resonance mode and will thus act as a **dipole scatterer** as shown in Figure 45. Hence, we can approximate the effect of a perturbation at any location x' with a point current source

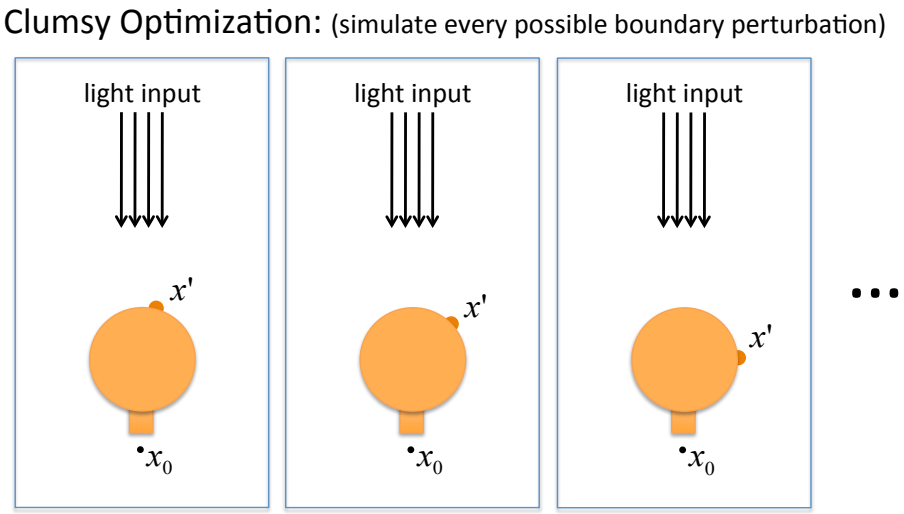

Figure 43: The clumsy method of gradient calculation. An inefficient way to calculate the gradient of a Figure of Merit is finite-difference. In this brute-force method, we can model every possible boundary change separately to measure the change in Figure of Merit with respect to each perturbation independently. For N parameters, this requires N+1 simulations.

J_1 at x' and solve Maxwell's Equations to determine the electric field distribution E_1 everywhere in the volume. To evaluate ΔF for a perturbation at x', one must simply observe $E_1(x_0)$, the electric field produced by the respective current source at x_0. This simulation neither requires modeling the excitation light source nor any structural changes within the 3D domain as shown in Figure 46. These simulations are equivalent to solving (4. 32), where the source term, $\boldsymbol{J}_{eff,i}$ and $\boldsymbol{J}_{\boldsymbol{m}_{eff,i}}$, is the dipole scatterer that is discussed here. The quantitative expression for the dipole moment for a specific geometric perturbation is the evaluation of (4. 33) and will be discussed in Section 4.8. After performing this first key trick, one still needs to perform many simulations as shown in Figure 46 to calculate ΔF due to each dipole scatterer independently.

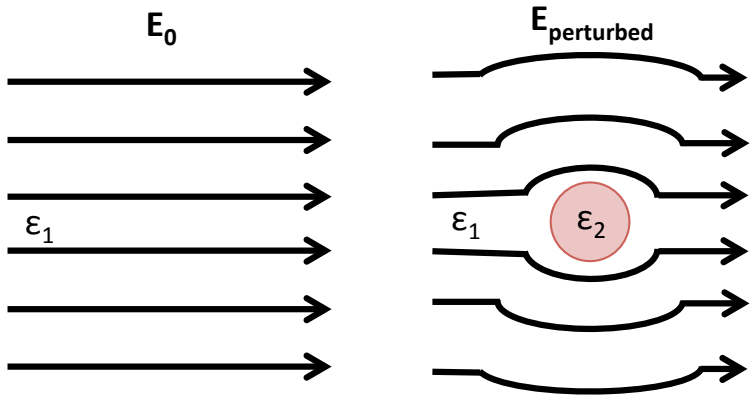

Figure 44: (Left) shows the electric field within a sea of material 1. (Right) shows the electric field when a perturbation in the form of a small sphere of material 2 is added.

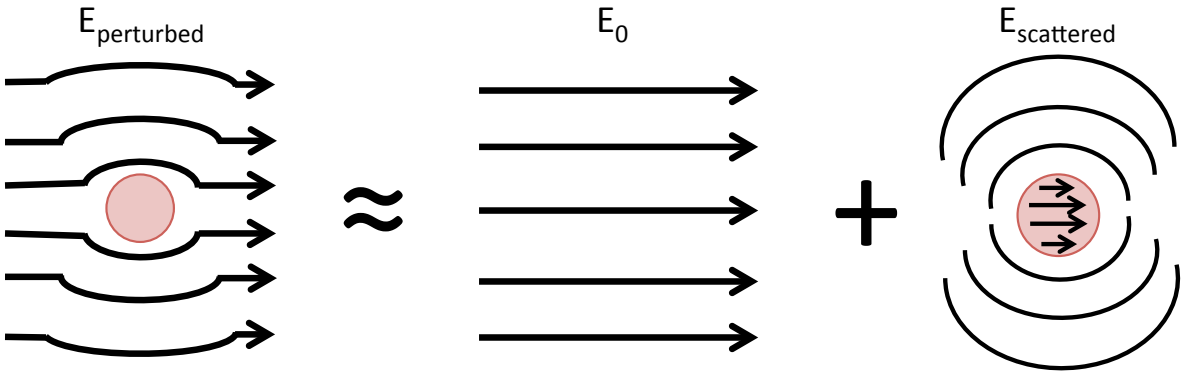

Figure 45: In the presence of electric field, a polarization is induced inside of a spherical perturbation of a different material. This polarization oscillates at the excitation frequency and can be modeled as a current source. We can approximate the perturbed fields as the coherent summation of the original electric field and a scattered field from a current source in the location of the perturbation.

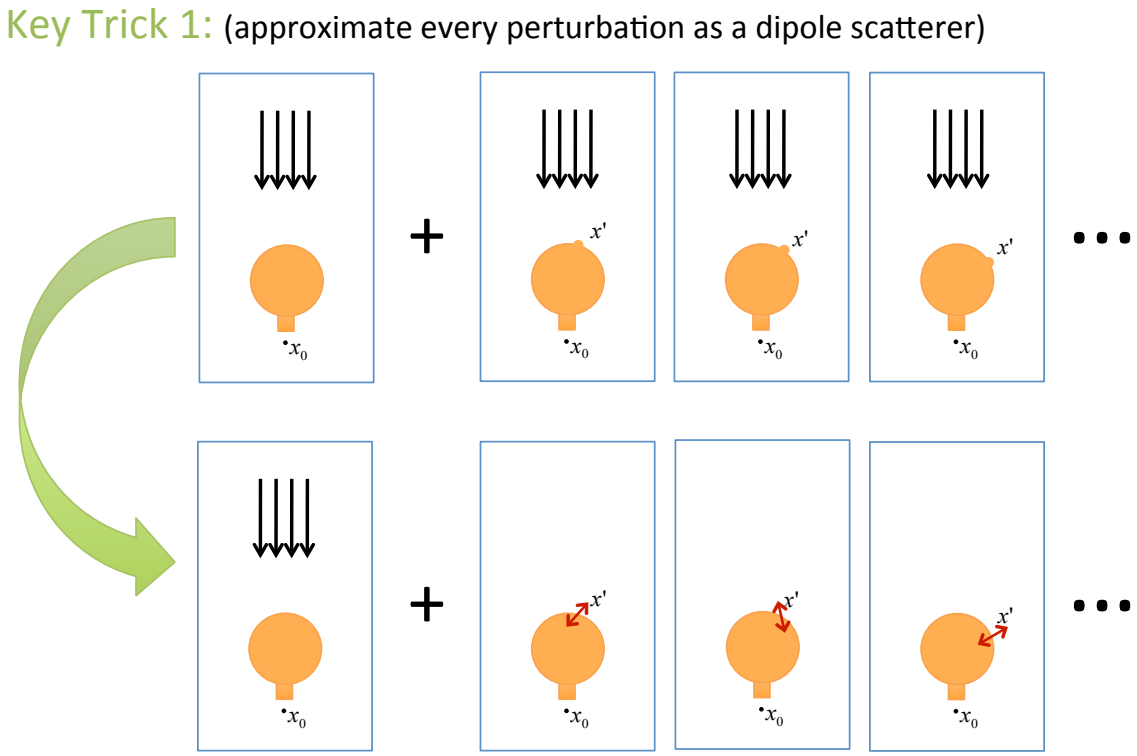

Figure 46: Model every possible perturbation (addition or removal of material) as a **dipole scatterer**, whose dipole moment is proportional to the electric field induced in the perturbation and oscillates at the excitation frequency.

As shown in Figure 49, the second key trick uses **Reciprocity** to greatly simplify the calculation of the electric and magnetic field contributions to location x_0 from every possible perturbation or dipole scatterer. In the left frame of Figure 47, consider the scenario of a dipole source at x' and observing the radiated electric field at x_0. If we relocate the dipole source to x_0 and the observation point to x', we will observe the exact same electric field. This is easy to justify in ray optics if we realize that all path lengths between two points are

identical in either direction. This is generally true in wave optics, because Maxwell's equations are linear and symmetric. And, this holds for any arbitrary set of electromagnetic structures in a 3D domain as long as all material properties are linear (complex values to model gain or loss is okay).

The common form of Lorentz reciprocity [41] relates two current sources $\boldsymbol{J_1}$ and $\boldsymbol{J_2}$ that independently produce electric field distributions $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$, respectively, in a volume of arbitrary materials as written in (4. 39).

$$\iiint \boldsymbol{J_1} \cdot \boldsymbol{E_2}\ dV = \iiint \boldsymbol{J_2} \cdot \boldsymbol{E_1}\ dV \tag{4. 39}$$

The induced dipole moment in a geometrical perturbation at x' is equivalent to a point current source $\boldsymbol{J_1}$ at x', which produces the electric field distribution $\boldsymbol{E_1}$ in the entire simulation volume. Also, consider the same volume instead excited with a point current source $\boldsymbol{J_2}$ at the observation point x_0 which produces the electric field distribution $\boldsymbol{E_2}$. The previous Reciprocity relationship becomes (4. 40).

$$\boldsymbol{J_1}(x') \cdot \boldsymbol{E_2}(x') = \boldsymbol{J_2}(x_0) \cdot \boldsymbol{E_1}(x_0) \tag{4. 40}$$

Rather than solving for the electric fields E_1 everywhere in the volume V produced by a current source at x' to evaluate $\boldsymbol{E_1}(x_0)$, we could alternatively solve for the fields $\boldsymbol{E_2}$ from a current source at x_0. Using the relationship above, we can calculate $\boldsymbol{E_1}(x_0)$ by instead observing $\boldsymbol{E_2}(x')$ and knowing the currents $\boldsymbol{J_1}$ and $\boldsymbol{J_2}$. The advantage is that from one simulation with a current source J_2 at x_0, we know $\boldsymbol{E_2}(x')$ for every x' in the simulation volume as shown in Figure 48. Hence, from this one simulation one can calculate $\boldsymbol{E_1}(x_0)$ and therefore evaluate ΔF for all perturbations at any x' in the volume. Now, the problem has been greatly simplified as we can calculate the gradient everywhere in space from one simulation where we excite the geometry with current sources at the same locations as where the objective function is evaluated as shown in Figure 49. Solving the bottom frame in Figure 49 is equivalent to solving the dual or adjoint problem in (4. 36).

In literature, Lorentz reciprocity is also commonly referred to as reciprocity of Green's Functions. Essentially, every time one models a perturbation at x' as a dipole scatterer or current source, one solves Maxwell's Equations to determine the function $G_{x'\rightarrow x_0}(J)$, that relates a current J at x' to an electric field at x_0. The reciprocity of Green's Functions is essentially a simplified case of Lorentz Reciprocity as well. $G_{x'\rightarrow x_0}(J)$ and $G_{x_0\rightarrow x'}(J)$ are equivalent for an arbitrary volume of geometries and materials, and one can perform either simulation to determine the other.

Figure 50 shows why there is an efficiency to be gained in solving for the gradient of many geometric parameters in electromagnetics. The inefficient method from Figure 43 requires solving Maxwell's equations everywhere in the volume for every possible perturbation at all x', even though the objective function only depends on the field at x_0. Essentially, in each of these simulations we are calculating and throwing away field data that is not of interest. However, in this reciprocal or adjoint simulation, every field data point that is calculated is used to evaluate the gradient at those respective x' locations. Although the x' locations are drawn only on the boundary in this figure, the adjoint method calculates the gradient at every location in the volume.

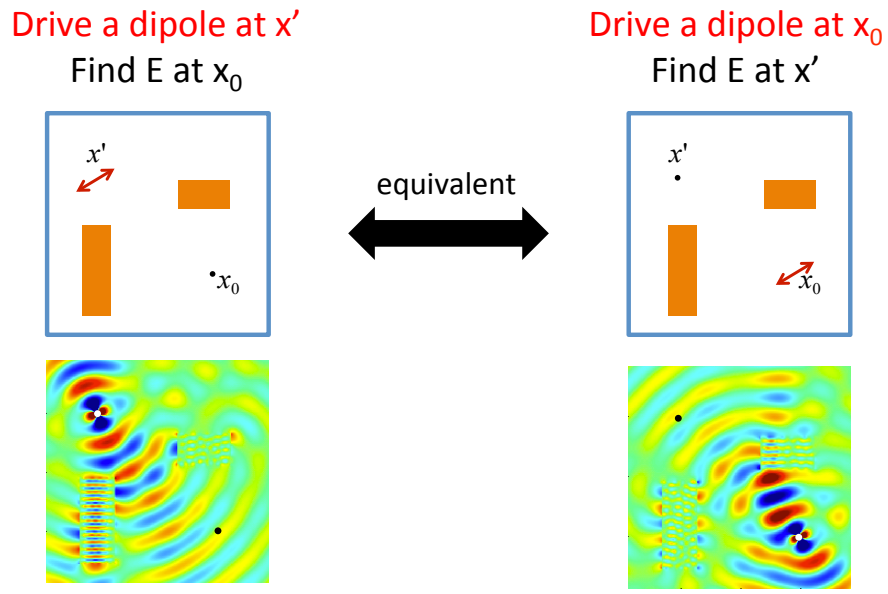

Figure 47: Because Maxwell's equations are symmetric, there are symmetries in the solutions to Maxwell's equations. Pictured here is Rayleigh-Carson Reciprocity (a derivative of Lorentz Reciprocity), which states that one can swap the source and observation point and witness the same electric field at the opposite location. Note that there are no geometric symmetries, and the resultant electric fields are also not symmetric. Yet, the electric fields at x_0 in the left frame is exactly identical to that at x' in the right frame.

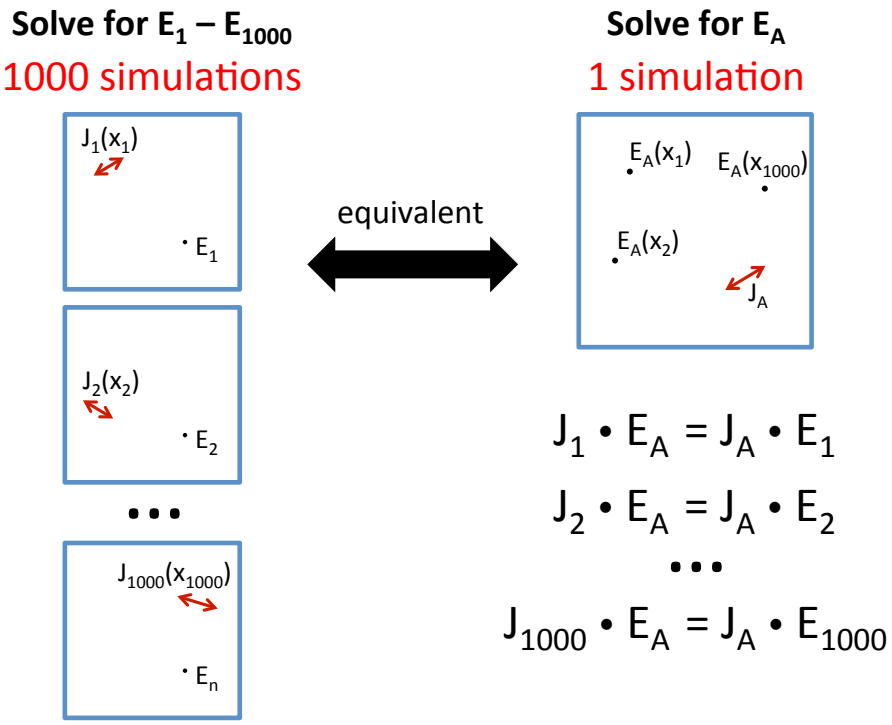

Figure 48: Using reciprocity (ie. the dual method applied to Maxwell's equations), we can parallelize the calculation of the electric field from N independent sources into just one simulation.

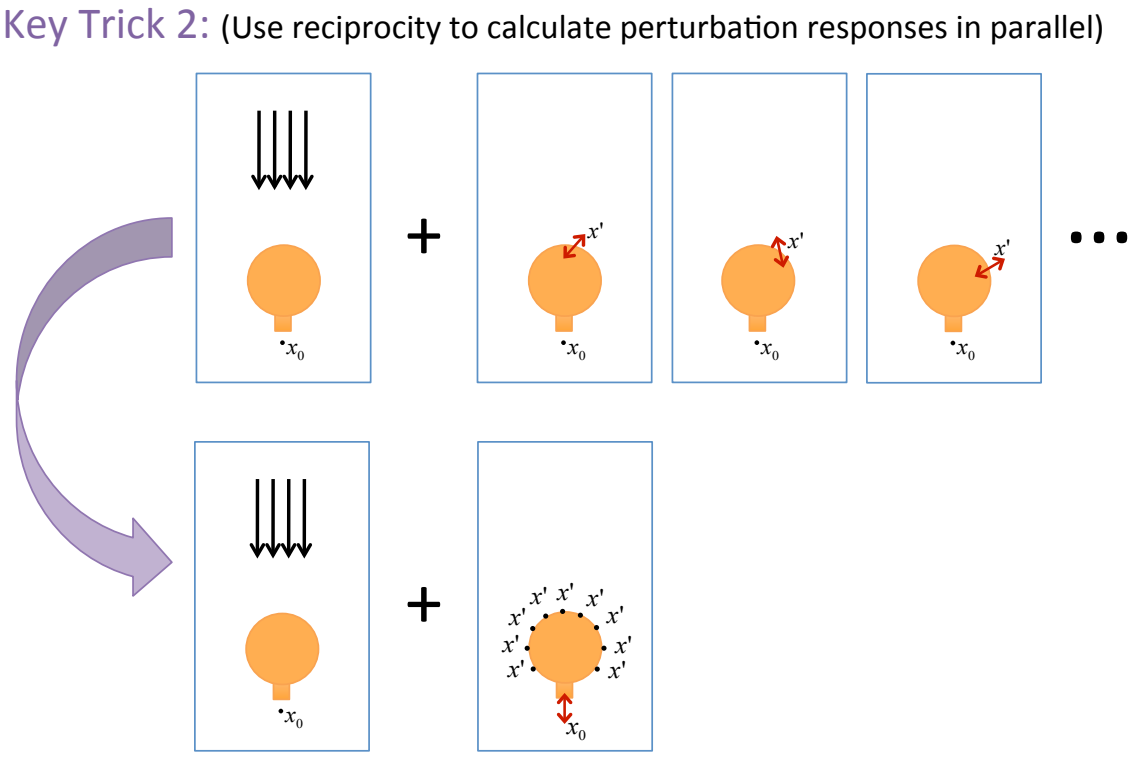

Figure 49: Reciprocity parallelizes the calculation of the Green's Function between x' and x_0.

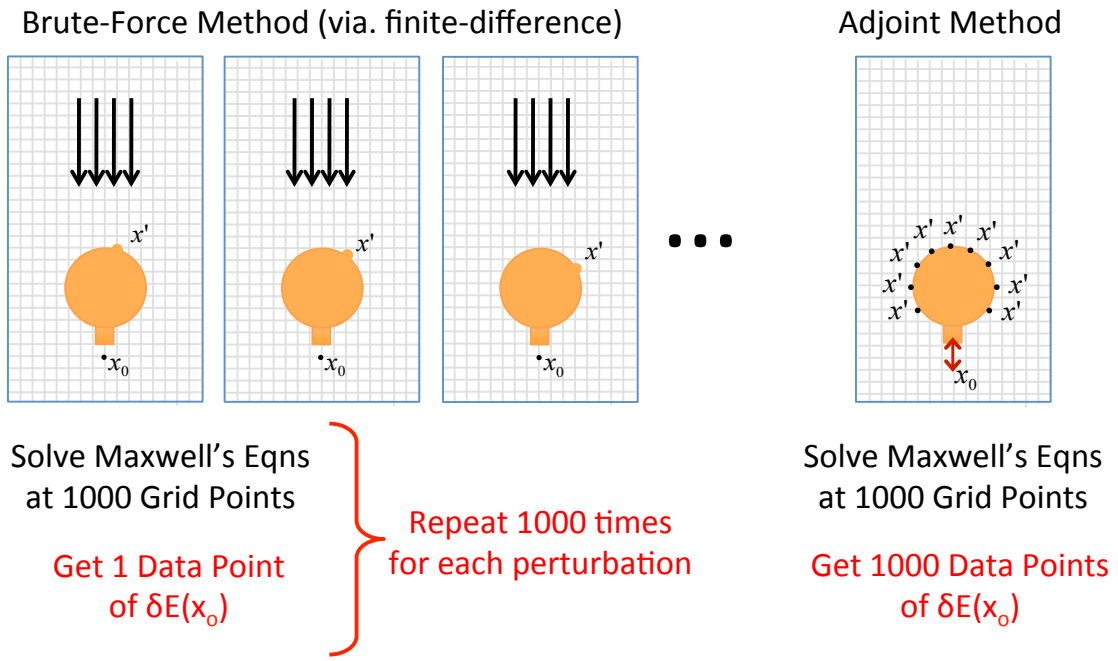

Figure 50: An information theory viewpoint of the dual method applied to calculating the gradient of an electromagnetic objective function. The brute-force method requires solving for the electric field everywhere in the domain even though the objective function is only evaluated at one point. Instead, we can avoid intensive and wasteful computation via the dual method (ie. adjoint method).

4.8 Gradient Calculation in 3D Wave Optics (derivation)

First, let's denote the objective function as the integral of an arbitrary function of electric and magnetic field at locations $\boldsymbol{x}$ within a particular volume V_F.

$$F = \iiint_{V_f} f\big(\boldsymbol{E}(\boldsymbol{x}), \boldsymbol{H}(\boldsymbol{x})\big)\, d^3x \tag{4. 41}$$

The electromagnetic fields in this region are a function of electromagnetic sources and of geometric structures. For a geometry optimization, we must model the electromagnetic effects of a small perturbation to the geometry. We will consider two possible structural perturbations. First, a *sparse* perturbation is the inclusion of an isolated small sphere of permittivity ϵ_2 displacing a volume within a sea of permittivity ϵ_1 as shown in Figure 51a. Second, a *boundary* perturbation at the interface between two objects of permittivity ϵ_1 and ϵ_2 is the inclusion of a bump of ϵ_2 replacing a volume of ϵ_1 as shown in Figure 51b. For either perturbation type, if the perturbation is electrically small, the electric field in this perturbed volume of ϵ_2 is the same as the original electric fields in the displaced volume of ϵ_1, only differing by a different set of boundary conditions. For the *sparse* perturbation, applying boundary conditions around the perturbed sphere leads to equation (4. 42), relating the electric field in the sphere to the original electric field in the sea of ϵ_1 by the Clausius-Mossotti factor [42]. Similarly, for the *boundary* perturbation, we arrive at (4. 43), which are the familiar boundary conditions of a flat interface, where $\parallel$ and $\perp$ denote the parallel and perpendicular vector components of the electric field. In both (4. 42) and (4. 43), $\boldsymbol{x}'$ denotes the location of the perturbed volume.

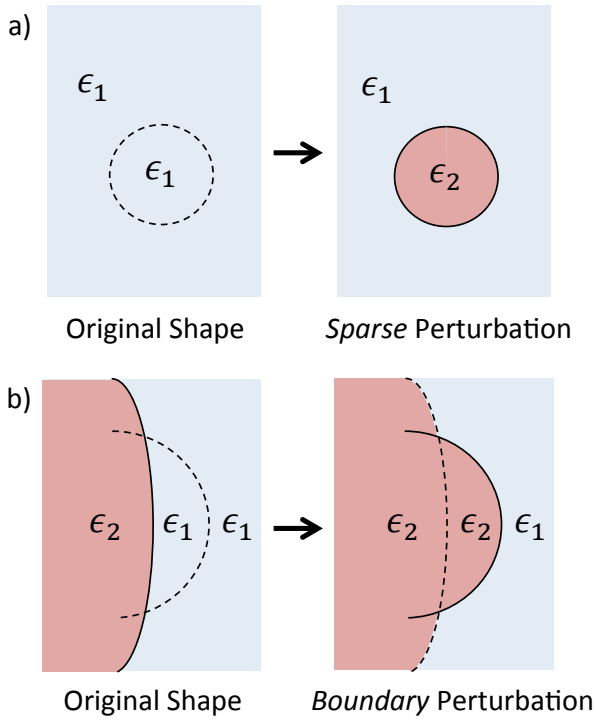

Figure 51: (a) A *sparse* perturbation is the inclusion of an isolated small sphere displacing a material of different permittivity. (b) A *boundary* perturbation is the inclusion of a locally flat bump at the interface between materials of different permittivity.

$$\boldsymbol{E_{perturbed}}(\boldsymbol{x'}) \approx \frac{3}{\epsilon_2/\epsilon_1 + 2}\boldsymbol{E_{orig}}(\boldsymbol{x'}) \tag{4. 42}$$

$$\boldsymbol{E_{perturbed}}(\boldsymbol{x'}) \approx \boldsymbol{E_{orig\parallel}}(\boldsymbol{x'}) + \frac{\epsilon_1}{\epsilon_2}\,\boldsymbol{E_{orig\perp}}(\boldsymbol{x'}) \tag{4. 43}$$

The electromagnetic effects of these perturbations are effectively modeled by a change in dipole moment density, $\boldsymbol{J_{pert}}$, in the perturbed volume as described in (4. 44) and (4. 45) for the *sparse* and *boundary* perturbations. These equations are an approximation in evaluating (4. 33). Johnson et. al. studied the perturbation theory of dielectric boundary shifts with more detail and recommends a few fitting parameters in (4. 45) to account for non-flat boundary shifts [43]. Following in Johnson's footsteps, we calculated the fitting parameters for metal-dielectric boundaries rather than dielectric-dielectric boundaries, but ultimately found that the correction did not offer major advantages in the optimizations and applications explored in this work. Thus, the simpler expression is shown here.

$$\frac{d\boldsymbol{J_{pert}}}{dV_{sparse}}(\boldsymbol{x'}) \approx (\epsilon_2 - \epsilon_1)\frac{3}{\epsilon_2/\epsilon_1 + 2}\boldsymbol{E_{orig}}(\boldsymbol{x'}) \tag{4. 44}$$

$$\frac{d\boldsymbol{J_{pert}}}{dV_{bnd}}(\boldsymbol{x'}) \approx (\epsilon_2 - \epsilon_1)\left(\boldsymbol{E_{orig\parallel}}(\boldsymbol{x'}) + \frac{\epsilon_1}{\epsilon_2}\,\boldsymbol{E_{orig\perp}}(\boldsymbol{x'})\right) \tag{4. 45}$$

This change in dipole moment causes a change in fields elsewhere in space. Of interest, the electric field is perturbed according to (4. 46), where $\boldsymbol{G^{EJ}}(\boldsymbol{x}, \boldsymbol{x'})$ is the electromagnetic Green's function, which is simply a transfer function relating a unit current source at the perturbation location $\boldsymbol{x'}$ to the electric field induced at location $\boldsymbol{x}$. Similarly, the magnetic field is perturbed according to (4. 47), where $\boldsymbol{G^{HJ}}(\boldsymbol{x}, \boldsymbol{x'})$ is a transfer function relating a unit current source at the perturbation location $\boldsymbol{x'}$ to the magnetic field induced at location $\boldsymbol{x}$. In application to the complex optical systems used for HAMR, this Green's function can only be evaluated by a full 3D Maxwell simulation with a current source at $\boldsymbol{x'}$ and observing the numerically-calculated electromagnetic fields at $\boldsymbol{x}$.

$$\boldsymbol{E_{perturbed}}(\boldsymbol{x}) \approx \boldsymbol{E_{orig}}(\boldsymbol{x}) + \left(\boldsymbol{G^{EJ}}(\boldsymbol{x}, \boldsymbol{x'})\,\boldsymbol{J_{pert}}(\boldsymbol{x'})\right) \tag{4. 46}$$

$$\boldsymbol{H_{perturbed}}(\boldsymbol{x}) \approx \boldsymbol{H_{orig}}(\boldsymbol{x}) + \left(\boldsymbol{G^{HJ}}(\boldsymbol{x}, \boldsymbol{x'})\,\boldsymbol{J_{pert}}(\boldsymbol{x'})\right) \tag{4. 47}$$

By differentiating (4. 41) and using the chain rule, we arrive at the expression (4. 48) for the gradient, which is the derivative of the FOM with respect to a volumetric change in permittivity. The $2Re\{\quad\}$ is a result of carefully taking the total derivative with respect to the complex valued functions $\boldsymbol{E}$, $\boldsymbol{H}$ and $\boldsymbol{J}$, which is not shown in detail here for brevity. By

replacing $\boldsymbol{E}(\boldsymbol{x})$ in (4. 41) with the approximate perturbed fields given by (4. 46) and (4. 47), we arrive at (4. 49).

$$\frac{\partial FOM}{\partial V_{pert}(\boldsymbol{x}')} = 2Re\left\{\iiint_{V_f} \frac{df}{d\boldsymbol{E}}(\boldsymbol{x}) \cdot \frac{\partial \boldsymbol{E}(\boldsymbol{x})}{\partial \boldsymbol{J_{pert}}(\boldsymbol{x}')} \cdot \frac{d\boldsymbol{J_{pert}}}{dV_{pert}}(\boldsymbol{x}') + \frac{df}{d\boldsymbol{H}}(\boldsymbol{x}) \cdot \frac{\partial \boldsymbol{H}(\boldsymbol{x})}{\partial \boldsymbol{J_{pert}}(\boldsymbol{x}')} \cdot \frac{d\boldsymbol{J_{pert}}}{dV_{pert}}(\boldsymbol{x}')\, d^3x\right\} \tag{4. 48}$$

$$\frac{\partial FOM}{\partial V_{pert}(\boldsymbol{x}')} = 2Re\left\{\iiint_{V_f} \frac{df}{d\boldsymbol{E}}(\boldsymbol{x}) \cdot \left(\boldsymbol{G^{EJ}}(\boldsymbol{x},\boldsymbol{x}')\frac{d\boldsymbol{J_{pert}}}{dV_{pert}}(\boldsymbol{x}')\right) + \frac{df}{d\boldsymbol{H}}(\boldsymbol{x}) \cdot \left(\boldsymbol{G^{HJ}}(\boldsymbol{x},\boldsymbol{x}')\frac{d\boldsymbol{J_{pert}}}{dV_{pert}}(\boldsymbol{x}')\right) d^3x\right\} \tag{4. 49}$$

Note that when using (4. 49), if we desire the unique value of the gradient at N possible perturbations at locations $\boldsymbol{x}'$, the terms $\boldsymbol{G^{EJ}}(\boldsymbol{x},\boldsymbol{x}')$ and $\boldsymbol{G^{HJ}}(\boldsymbol{x},\boldsymbol{x}')$ must be evaluated via N individual Maxwell simulations of a current source equal to the dipole moment of each possible perturbation at each $\boldsymbol{x}'$, respectively. This is computationally expensive. Instead, we leverage reciprocity in electromagnetics, which we proved in Section 4.4, and summarize the necessary relationships here in (4. 50) and (4. 51). $\boldsymbol{G^{EM}}(\boldsymbol{x},\boldsymbol{x}')$ is a transfer function relating a unit magnetic current source at the perturbation location $\boldsymbol{x}'$ to the electric field induced at location $\boldsymbol{x}$.

$$\boldsymbol{G^{EJ}}(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{G^{EJ}}(\boldsymbol{x}',\boldsymbol{x}) \tag{4. 50}$$

$$\boldsymbol{G^{HJ}}(\boldsymbol{x},\boldsymbol{x}') = -\boldsymbol{G^{EM}}(\boldsymbol{x}',\boldsymbol{x}) \tag{4. 51}$$

After substituting (4. 50) and (4. 51) into (4. 49), we obtain a fundamentally different expression for calculating the gradient that is computationally inexpensive, shown in (4. 52). According to equations (4. 44) and (4. 45), the first term depends on the electric fields from a single *Forward* simulation of the original source and the unperturbed geometry. The latter term comprises of the electric fields from a single *Adjoint* simulation where the source is the superposition of electric and magnetic sources of amplitude $\frac{df}{d\boldsymbol{E}}(\boldsymbol{x})$ and $\frac{df}{d\boldsymbol{H}}(\boldsymbol{x})$, respectively. Hence, only 2 Maxwell simulations are required to obtain the gradient at all potential perturbation positions $\boldsymbol{x}'$.

$$\frac{\partial FOM}{\partial V_{pert}(\boldsymbol{x}')} = 2Re\left\{\frac{d\boldsymbol{J_{pert}}}{dV_{pert}}(\boldsymbol{x}') \cdot \left[\iiint_{V_F} \left(\boldsymbol{G^{EJ}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{E}(\boldsymbol{x})}\right) + \left(\boldsymbol{G^{EM}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{H}(\boldsymbol{x})}\right) d^3x\right]\right\} \tag{4. 52}$$

By substituting (4. 44) and (4. 45) in (4. 52), we arrive at the gradient formulas shown in (4. 53) and (4. 54) for the *sparse* and *boundary* perturbations that are explicitly used in this dissertation.

$$\frac{\partial FOM}{\partial V_{bnd}(\boldsymbol{x}')} = 2Re\left\{\left[(\epsilon_2 - \epsilon_1)\left(\boldsymbol{E_{orig\parallel}}(\boldsymbol{x}') + \frac{\epsilon_1}{\epsilon_2}\boldsymbol{E_{orig\perp}}(\boldsymbol{x}')\right)\right] \cdot \left[\iiint_{V_f}\left(\boldsymbol{G^{EJ}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{E}(\boldsymbol{x})}\right) + \left(\boldsymbol{G^{EM}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{H}(\boldsymbol{x})}\right)d^3x\right]\right\} \tag{4. 53}$$

$$\frac{\partial FOM}{\partial V_{sparse}(\boldsymbol{x}')} = 2Re\left\{\left[\frac{3(\epsilon_2 - \epsilon_1)}{\epsilon_2/\epsilon_1 + 2}\boldsymbol{E_{orig}}(\boldsymbol{x}')\right] \cdot \left[\iiint_{V_f}\left(\boldsymbol{G^{EJ}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{E}(\boldsymbol{x})}\right) + \left(\boldsymbol{G^{EM}}(\boldsymbol{x}',\boldsymbol{x})\frac{df(\boldsymbol{x})}{d\boldsymbol{H}(\boldsymbol{x})}\right)d^3x\right]\right\} \tag{4. 54}$$

4.9 Gradient-Based Freeform Optimization

By using this gradient calculation, one can easily implement an iterative optimization using steepest descent, where every iterative geometry update is in the direction of the gradient. In this work, we also used finite-difference between consecutive iterations to approximate the second derivative $\frac{\partial^2 FOM}{\partial V_{bnd}(\boldsymbol{x}')^2}$, specifically the diagonal of the Hessian, and implemented a quasi-Newton update method. This was crucial to converge to an optimum within the enormous parameter space occupied by the thousands to millions of degrees of freedom that we usually represent. Moreover, the freeform nature of the boundary optimization was implemented by representing the boundaries on a binary bitmap, where 1s and 0s represent the material inside and outside the various boundaries. Every pixel along the boundary was treated as a separate degree of freedom, and the boundary could expand outwards or contract inwards on a per-pixel basis. Hence, the optimized boundaries were allowed to completely diverge from the initial shape fed to the optimization algorithm. We did not use the Level Set Method, which uses a signed-distance function (SDF) as the representation of a contour. An SDF allows for nearly continuous changes in a contour's position, which is often a desirable trait, but we needed a boundary representation that was stair-cased on a uniform discrete grid to match the geometry representation used in FDTD solvers. The SDF caused major problems when a contour moved by less than a mesh spacing of the simulation model. Convergence is superior when the geometry model represented by the optimization software is the same as the representation used by the simulator. We implemented radius of curvature and minimum dimensions constraint on the binary bitmap geometry representation using mathematical morphology, which is a basic application of convolutions with various filters. We extended some of the mathematical morphology techniques to properly weigh the gradient along the boundary according to the curvature constraints to maintain stable convergence when geometric constraints were enforced. This freeform geometric representation combined with gradient-based optimization allows for creative objective-first design of 3D electromagnetic structures, which we call *Inverse Electromagnetic Design.*

4.10 Example: Volume Hologram for Solar Spectral Splitting

4.10.1 High-Efficiency Solar Modules via Multiple Bandgaps Cells

Solar energy is clearly a desirable pathway toward renewable energy not dependent on fossil fuels like coal and petroleum. The strategy toward 50% efficient modules requires light to be absorbed by cells of different electronic bandgaps to absorb more light and to reduce energy loss due to thermalization of carriers. The most popular strategy, commercialized by companies like SpectroLab (a subsidiary of Boeing), is to use multi-junction cells where several cells with different bandgaps are grown on top of each other. Currently, this is a very expensive product that is reserved for niche applications like solar modules in outer space. A key limit to multi-junction cell efficiency stems from the maximum number of bandgaps that are lattice-matched and stackable. Also, multi-junction cell efficiency is limited by current-matching losses in the junctions when connected in series, especially since the solar spectrum is not constant over a day or year. Perhaps, a better module would consist of solar cells of different bandgaps placed laterally next to each other. This would involve simple single junction cells perhaps diced into strips and laid flat. Of course, this module would require splitting the incoming light with respect to wavelength (ie. color) onto the respective cells to maximize efficiency. In this example, we applied the wave optics Inverse Design software to design a periodic holographic structure that could split solar light onto periodic solar cell strips as shown in Figure 52.

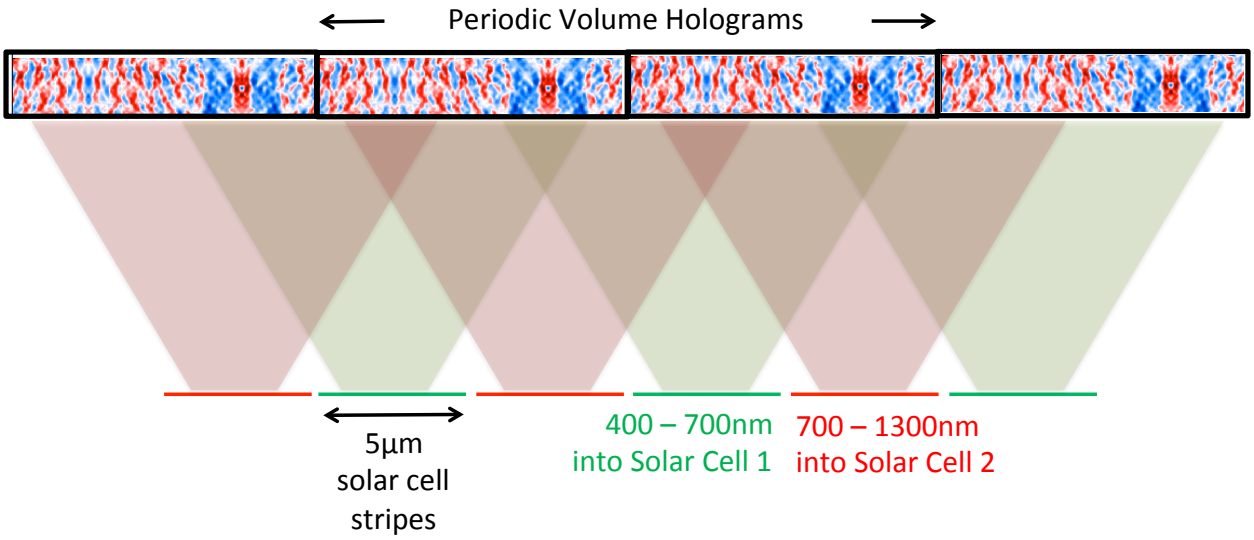

Figure 52: A schematic of computer-generated volume holograms to the split the solar spectrum onto different bandgap solar cells.

4.10.2 Computer-Generated Volume Hologram for Broad-Bandwidth Performance

In this design problem, we wish to split colors of light spatially onto different locations. One challenge is that we need the hologram to be functional over an entire continuous bandwidth of wavelengths rather than one fixed wavelength (typical holography often uses lasers as light sources). To achieve continuous bandwidth performance, we used the minimax technique (or perhaps its opposite), in which we iteratively maximized the worst performing wavelength. We defined our objective function to be (4. 56). The partial derivative of the objective function with respect to **E** and **H** are (4. 58) are (4. 59), which were used as the electromagnetic source currents in the adjoint simulation.

$$maximize\ F = f(\boldsymbol{E}, \boldsymbol{H}, \lambda) \tag{4. 55}$$

$$f = \min_{\lambda}\left[\int_{S_1} \alpha(\lambda)\cdot\left(\boldsymbol{E}(\lambda)\times\boldsymbol{H}(\lambda)\right)\cdot\hat{\boldsymbol{n}}dS + \int_{S_2}(1-\alpha(\lambda))\cdot(\boldsymbol{E}(\lambda)\times\boldsymbol{H}(\lambda))\cdot\hat{\boldsymbol{n}}dS\right] \tag{4. 56}$$

$$\alpha(\lambda) = \begin{cases} 1, & 400nm < \lambda < 700nm \\ 0, & 700nm < \lambda < 1300nm \end{cases} \tag{4. 57}$$

$$\frac{df}{d\boldsymbol{E}} \approx \alpha(\lambda_{min})\cdot\left(\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{y}}\\ \hat{\boldsymbol{z}}\end{bmatrix}\times\boldsymbol{H}(\lambda_{min})\right)\cdot\boldsymbol{S_1} + (1-\alpha(\lambda_{min}))\cdot\left(\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{y}}\\ \hat{\boldsymbol{z}}\end{bmatrix}\times\boldsymbol{H}(x_{S_2},\lambda_{min})\right)\cdot\boldsymbol{S_2} \tag{4. 58}$$

$$\frac{df}{d\boldsymbol{H}} \approx \alpha(\lambda_{min})\cdot\left(\boldsymbol{E}(\lambda_{min})\times\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{y}}\\ \hat{\boldsymbol{z}}\end{bmatrix}\right)\cdot\boldsymbol{S_1} + (1-\alpha(\lambda_{min}))\cdot\left(\boldsymbol{E}(\lambda_{min})\times\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{y}}\\ \hat{\boldsymbol{z}}\end{bmatrix}\right)\cdot\boldsymbol{S_2} \tag{4. 59}$$

Figure 53 shows the results from the iterative gradient-based optimizations of a 10μm thick volume hologram with a horizontal periodicity of 10μm. There was a 100nm gap between the bottom of the 10μm thick volume hologram and a periodic linear array of solar cells of 2 different bandgaps. This was a 2D model and simulation, with a plane wave of broad-spectrum light propagating through the hologram from the top onto the solar cells. The simulation was performed with Lumerical FDTD at 28 discrete wavelengths within the 400-1300nm range. This sampling of wavelengths was found to be sufficient in the minimax optimization to approximate the entire color spectrum. The geometry representation was a 500×500 array of 20nm×20nm pixels, each with an index of refraction continuously varying between 1.8 and 2.2. Because this geometry is not a binary set of materials with a boundary, we used the *sparse* gradient calculation analysis discussed earlier. The initial structure was homogenous material with an index of refraction of 2.0 with no pre-existing holographic features. These 250,000 pixels were successfully optimized in 145 iterations. The last iteration's design provided near-perfect splitting with a step function response between the two color bands. Figure 54 shows the light intensity through a vertical cross-section of the volume hologram for different incoming wavelengths.

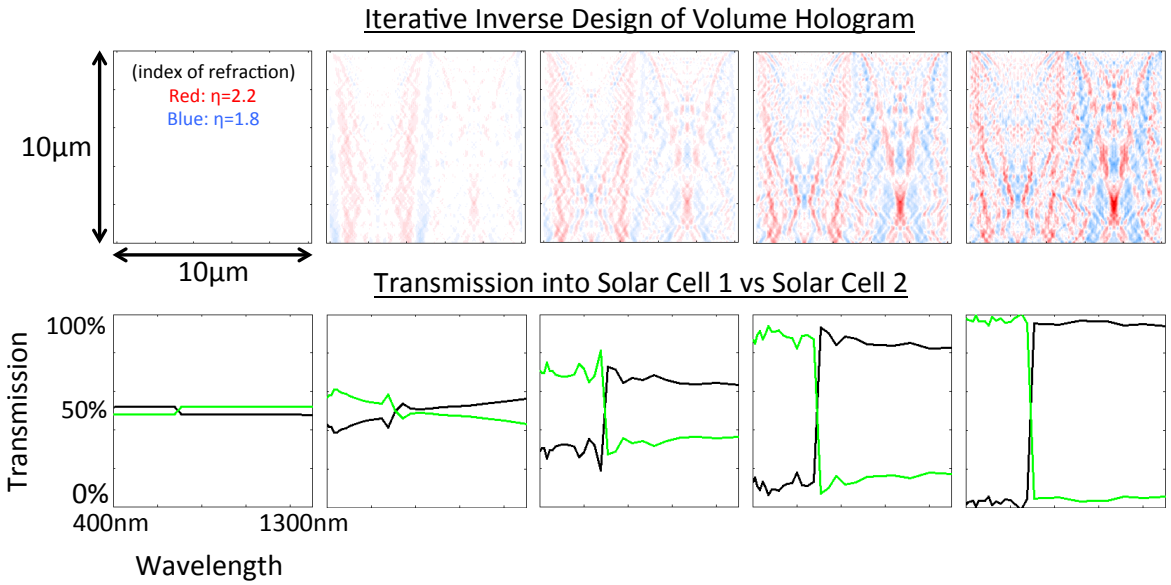

Figure 53: The iterative optimization of a volume hologram for solar spectral splitting.

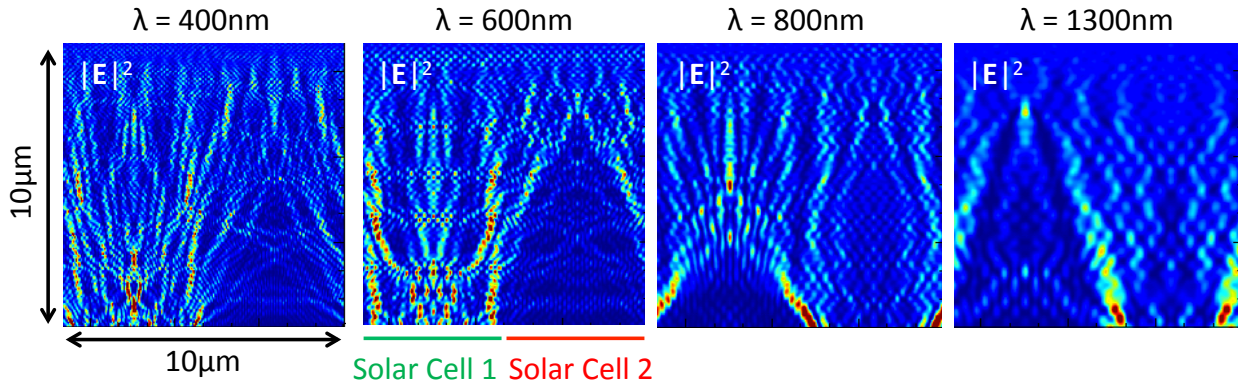

Figure 54: The light intensity at different wavelengths in a vertical cross-section of the optimized volume hologram. The hologram correctly splits 400-700nm light to the left solar cell and 700-1300nm light to the right solar cell.

4.11 Example: Optical Antenna for an Ultrafast LED

4.11.1 Antenna-Enhanced Spontaneous Emission

It was first discussed by Edward Purcell that electromagnetic emission could be suppressed or enhanced by encapsulating an optical emitter inside a resonant cavity [44]. This effect has seen much interest in SERS spectroscopy and more recently in enhancing emission from semiconductors for ultrafast LEDs [45][46][47] with the worthy goal of making the spontaneous emission rate in semiconductors faster than stimulated emission. Stimulated emission is ubiquitous in laser emitters but comes at the price of energy inefficiency via very high currents to reach high modulation bandwidths. Rather than discussing the optical density of states or the Purcell effect, the antenna-enhanced radiation from a semiconductor has also been well explained via traditional circuit and antenna theory [46][48][49]. If the emitter is a III-V semiconductor (as opposed to dye molecules which are orders of magnitude slower), then a rate enhancement of ~200x is required for spontaneous emission to beat stimulated emission.

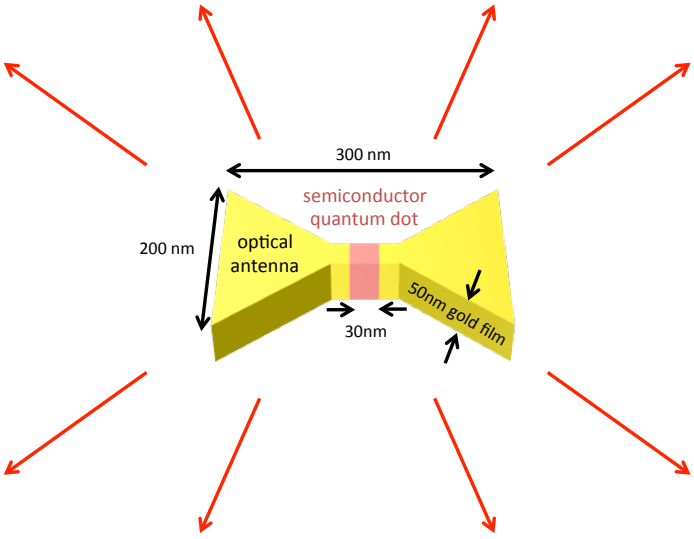

Figure 55: An optical antenna consisting of 50nm thin film of gold with a 2D contour can enhance the spontaneous emission of a semiconductor quantum dot.

4.11.2 Optimized Optical Antenna Shapes for Enhanced Emission

In this example, the objective function used in our Inverse Design software was the Poynting vector integrated over a closed surface fully encompassing the quantum-dot and antenna structure (ie. a transmission box) as noted in (4. 61). The light emission from the quantum dot was modeled as an infinitesimally small current source (single wavelength of 830 nm) within a cube of InP with the current polarization parallel to the antenna orientation. The objective function is not a measure of the light emitted by the quantum dot but rather the light emitted by the whole structure after the loss in the metal is taken account for. In the structures presented here, the energy absorbed by the antenna was roughly double the energy emitted by the whole structure. Even with the high absorption, the enhanced light emission can be huge. In this optimization, only the boundary between the gold antenna and surrounding air was optimized. The gold was constrained to have constant thickness of 50 nm, radius of curvature of 5 nm, minimum feature size of 10 nm, and fixed gap width of 30 nm. Figure 56 shows the iterative shape optimization of the antenna. The initial structure was the bowtie structure in Figure 55, and the final structure resembles the cross-section of a fictional Star Wars plane. Figure 57 shows the convergence plot of this boundary optimization and the radiation enhancement versus wavelength. The final structure seems unintuitive but actually matches the optical antenna circuit analysis performed by Eggleston and Messer [46], in that capacitance over the antenna gap is minimized and the fringing capacitance between the far edges of the antenna is maximized. The exact shape of the optimized antenna would not have been designed by any traditional design methods but rather only from an efficient freeform optimization tool like our wave optics Inverse Electromagnetic Design software.

$$maximize\ F = f(\boldsymbol{E}, \boldsymbol{H}) \tag{4. 60}$$

$$f = \oiint_S (\boldsymbol{E} \times \boldsymbol{H}) \cdot \hat{\boldsymbol{n}} dS \tag{4. 61}$$

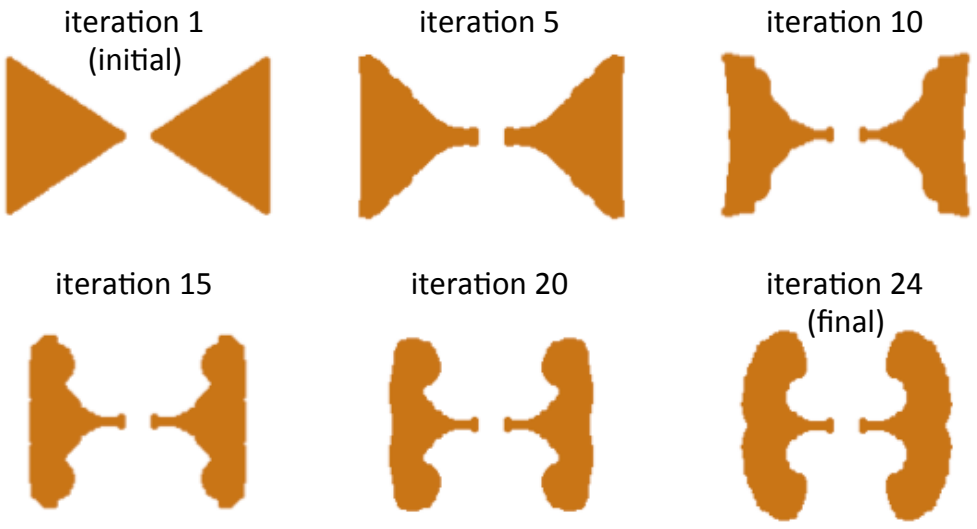

Figure 56: The iterative optimization of an optical antenna of a uniform thickness 50nm gold thin film coupled to an InP quantum dot (not shown) at the center. The initial structure was a bowtie of 200nm height, and the optimized unintuitive structure resembles the cross-section of a fictional Star Wars fighter plane.

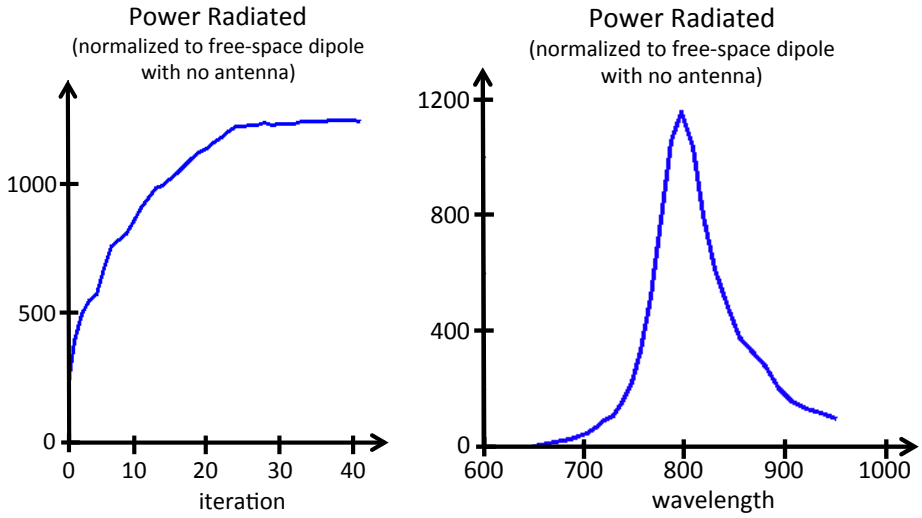

Figure 57: The optimization convergence plot showing the objective function versus iteration (left). The objective function plotted versus wavelength of operation showing the significant enhancement of radiation at the antenna resonance near 800 nm.

5 A Fat Antenna with Reduced Self-Heating

To ensure thermal stability of data over a 10 year lifetime in hard disks of beyond 1Tb/in^2 areal density, the magneto-crystalline anisotropy of the magnetic granular media must be increased while scaling the magnetic grains to smaller dimensions [7]. Recording information to such a medium is a monumental challenge. The current state-of-art employs writing electromagnets that are already limited by magnetic field saturation of permeable metals, placing an upper bound on the strength of magnetic field that can be applied during the recording process. Heat-Assisted Magnetic Recording (HAMR) promises writing to highly anisotropic media, by temporarily heating the area of a single datum to its Curie temperature while simultaneously applying a magnetic field from a conventional electromagnet [8][50]. In practice, a metallic optical antenna or near-field transducer (NFT) focuses light onto the highly absorbing magnetic recording layer in the media and locally heats a 30nm×30nm spot on the medium to near 700K [3][51]. Since the metal comprising the NFT, typically gold, is itself absorbing at optical frequencies, the NFT also heats by several hundreds of degrees [2]. This NFT self-heating is a significant cause of failure in HAMR systems and limits the lifetime of today's prototype HAMR write-heads to be orders of magnitude less than the desired 10 year lifespan [42]. Hence, an important Figure of Merit for reliability in a HAMR write-head is the temperature ratio between the media hotspot and NFT.

The difficulty in designing a low temperature HAMR write-head is two-fold: (1) the fundamental limits on the NFT temperature are not well known; (2) designing the light delivery system that produces nano-scale heating requires understanding the complex electromagnetic interactions of the illuminating waveguide, metallic NFT, magnetic write pole and multi-layered hard disk medium. On the first note, we derived in Section 3.3 a simple analytic model for the ratio of temperature rise in the NFT to the temperature rise in the media hotspot. This provides important limits and constraints on the structural design of the NFT that must be satisfied for low temperature operation. On the second note, because of the wave nature of light, the optimal shapes of electromagnetic structures are often unexpected [52]. Traditional design approaches, based on intuition or highly constrained optimization such as parameter sweeps, are inadequate. Instead, we used the Inverse Electromagnetic Design software, which provides fast optimization of 3D electromagnetic structures with thousands of degrees of freedom. With such a large parameter space, an optimization can search for unexpected shapes of the NFT or feeding waveguide that offer superior performance for HAMR. The drawback of such an optimization is computational expense. For applications like HAMR, a single 3D Maxwell simulation of nano-scale metallic structures and a multi-layered medium may take hours on a modern high-performance computing cluster. Since heuristic

algorithms like particle swarm and genetic algorithms rely on random trials, they are too computationally burdensome for practical engineering design for optics in the nano-scale. We wish to have a computational design process that takes 1 day and not 1000 hours. In contrast, the Inverse Electromagnetic Design software performs gradient-based optimization using the adjoint method, which results in fast deterministic optimization that is computationally inexpensive [[10], [32], [53]].

In this chapter, our strategy towards achieving a low temperature HAMR write-head is to make major NFT design choices based on a simple analytic expression for the NFT/media temperature ratio. Then, we use our Inverse Electromagnetic Design software to find unexpected shapes of the waveguide feeding the proposed NFT design to provide the desired optical performance.

5.1 Media/NFT Temperature Ratio

We derived in Section 3.3 the following temperature ratio between the hard disk media and the nano-focusing NFT. For HAMR, we wish the rise in temperature in the media to be ~400°C but need the temperature rise in NFT tip to significantly less. As shown in (5. 1), this temperature ratio is a function of many electromagnetic and thermal material properties as well as the solid angle of the NFT and media.

$$\frac{\Delta T_{media}}{\Delta T_{NFT}} \approx \frac{|\epsilon_{NFT}|^2}{|\epsilon_{media}|^2} \times \frac{\varepsilon''_{media}}{\varepsilon''_{NFT}} \times \frac{K_{NFT}}{K_{media}} \times \frac{\Omega_{NFT}}{\Omega_{media}} \times \frac{t_{grain}}{\delta_{NFT}} \tag{5. 1}$$

For a low temperature transducer, this ratio must be as high as possible. Clearly, there are significant factors that are not accounted for in this expression, such as the anisotropic thermal conductivity of HAMR granular media and its under-layers, or the exact structural design of the NFT. Nevertheless, this expression correctly emphasizes some key design requirements for low temperature NFT operation.

1) The media must have minimum heatsinking.

2) NFT metallurgy must be optimized for $K_{NFT} \times \frac{|\epsilon_{NFT}|^2}{\varepsilon''_{NFT}} \times \frac{1}{\delta_{NFT}}$.

3) NFT structural design must have the largest solid angle of heat conduction at the NFT tip.

5.2 A Fat NFT

For low temperature operation, a crucial NFT design choice is to have the largest solid angle of heat conduction from the tip of the NFT. The left of Figure 58 shows a simplified HAMR system consisting of a magnetic write pole, media stack, gold lollipop NFT, and incident light in a slab waveguide similar to that of Seagate's parabolic solid-immersion mirror (PSIM) [2]. Through 3D FDTD modeling, we observe that this system produces a sharply confined hotspot in the media's recording layer, whose dimensions are defined by the cross-sectional dimensions of the NFT tip at the air-bearing surface (ABS). However, the lollipop NFT is skinny and has very little solid angle of heat conduction. Such NFTs experience a rise in temperature by hundreds of degrees, which is a significant cause of failure in HAMR systems. Therefore, we need a fat NFT that has a large solid angle of heat conduction.

We propose a *Fat* NFT, shown on the right of Figure 58, consisting of a thin-film gold pattern embossed on a bulk chunk of gold. The only part of the thin-film pattern that does not touch the bulk gold is the NFT tip, because the magnetic write pole tip was fixed at a 30nm offset from the NFT tip at the ABS. This new Fat NFT is NOT the familiar lollipop NFT. We illuminated the Fat NFT with the same PSIM-like waveguide mode, and the observed light intensity in the media is shown in the bottom right of Figure 58. The PSIM-like waveguide mode is a poor mode-match to the Fat NFT and the system delivers a poorly confined hotspot. The side lobes in the light intensity pattern are unacceptable, because high temperatures in the media outside of the hotspot would unintentionally erase information. A typical data storage specification is to allow for 100,000 writes to a particular track without erasing data on nearby tracks. To meet this specification, the peak light intensity in the hotspot versus the peak intensity elsewhere in the media should be at least 5x, and achieving this light intensity ratio with the proposed Fat NFT was the goal in this chapter.

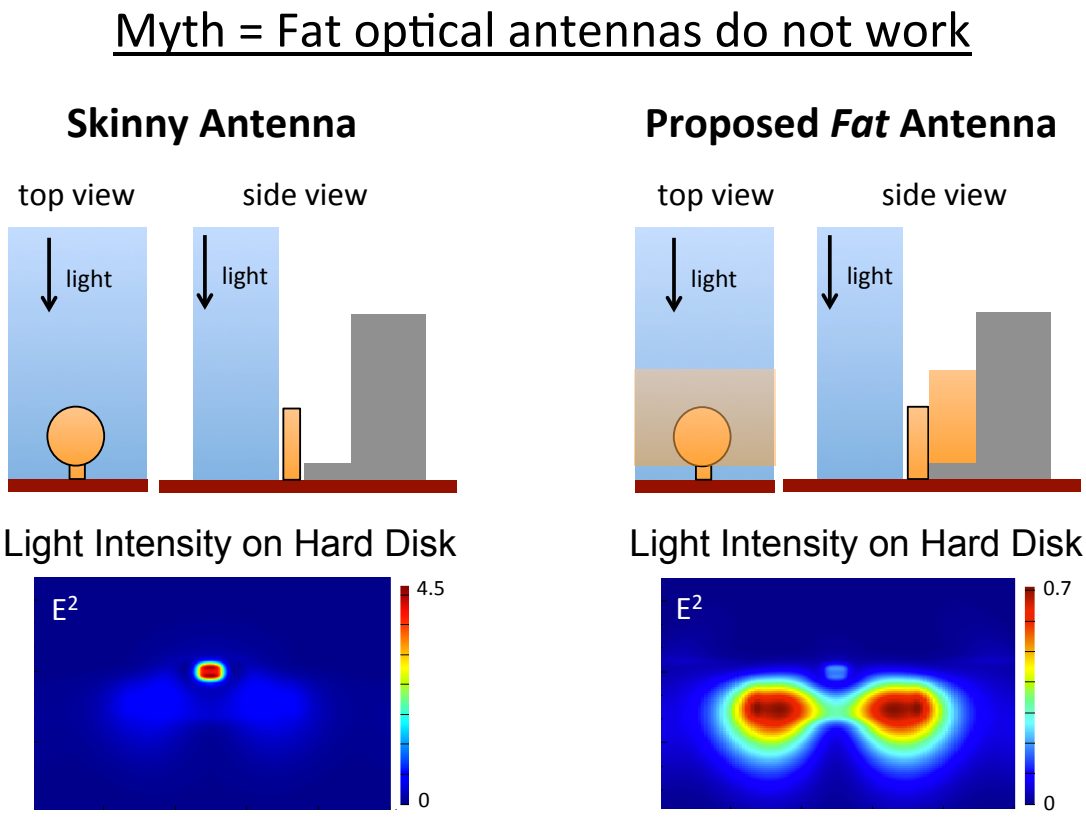

Figure 58: A HAMR optical system composing of a Ta_2O_5 waveguide (blue), gold NFT (yellow), CoFe write pole (grey) and magnetic media (red). On left, a typical skinny lollipop NFT produces a confined hotspot in the storage layer. On right, the proposed Fat NFT has different electromagnetic behavior and is a poor mode match to a PSIM-like waveguide mode.

5.3 Maxwell Simulation Methods

The ability to perform accurate 3D electromagnetic simulations is imperative to computational optimization. Simulation results in this chapter use a commercial finite-difference time-domain Maxwell solver, Lumerical FDTD, in which a pulse of light is injected

into the waveguide of the HAMR system and propagated in the time domain towards the NFT, write pole and media until pulse energy has decayed beyond our desired precision. A detailed mesh convergence test was performed to ensure minimal computational error due to discretization. The most crucial and computationally demanding mesh requirements were 1nm cubic Yee cells in the metallic NFT and 0.5nm Yee cell thicknesses in the media to resolve the various nanometer-thin layers of the media stack. We used an in-house high-performance computing cluster consisting 336 cores and 668GBs RAM over 26 nodes. By parallelizing the solver through a Message-Passing Interface, Open MPI [54], and 40Gb/s Infiniband interconnects, our in-house cluster can simulate FDTD models of 500 million Yee cell nodes. Typically, we ran simulations on 64 to 128 cores at a time, with which we could run iterative optimizations of HAMR structures in roughly one day's time.

Figure 59 shows 3D views of the HAMR structure that was modeled, and TABLE 4 contains structural and optical properties at the operation laser wavelength of 830nm for the numerous write-head and media components. These properties were chosen to closely mimic designs from industry publications and patent literature. TABLE 5 shows thermal properties that were assumed for a thermal finite-element model performed in COMSOL Multiphysics to predict the media and NFT temperatures.

TABLE 4: STRUCTURAL AND OPTICAL PROPERTIES IN PROPOSED HAMR SYSTEM

Device	Dimensions	n	k
Au NFT	125 nm radius, 60 nm thick	0.16	5.08
Au NFT Peg	50 nm wide at ABS, 30 nm thick	0.16	5.08
Ta_2O_5 Waveguide	100 nm thick	2.1	-
SiO_2 Cladding	-	1.4	-
CoFe Writepole	120 wide at ABS	3	4
Head Overcoat	2.5 nm thick	1.6	-
Air Gap	2.5 nm thick	1.0	-
Media Overcoat	2.5 nm thick	1.2	-
FePt Recording Layer	10 nm thick	2.9	1.5
MgO Interlayer	15 nm thick	1.7	-
Au Media Heatsink	80 nm thick	0.26	5.28
Media Substrate	infinite	1.5	-

TABLE 5: THERMAL PROPERTIES IN PROPOSED HAMR SYSTEM

Material	Specific Heat (J/m^3K)	Thermal Conductivity (W/mK)
Au	$3{\cdot}10^6$	100
Ta_2O_5	$2{\cdot}10^6$	2
SiO_2	$2{\cdot}10^6$	1
CoFe	$3.5{\cdot}10^6$	20
FePt - Lateral	$3{\cdot}10^6$	5
FePt - Vertical	$3{\cdot}10^6$	50
MgO	$2{\cdot}10^6$	3

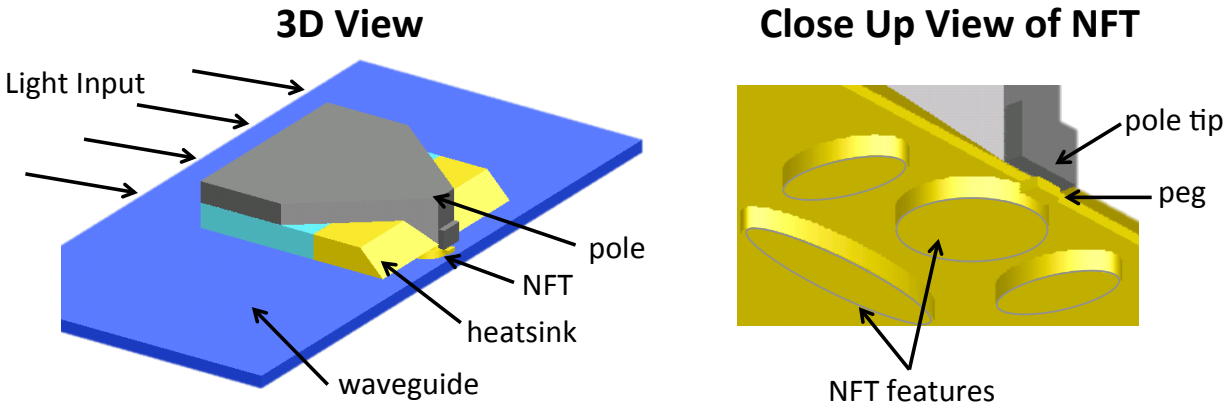

Figure 59: 3D views of the proposed HAMR light delivery structure. The Fat NFT is a thin-film gold pattern embossed on a bulk chunk of gold illuminated by the Seagate parabolic condenser or PSIM.

5.4 Inverse Design of Fat Antenna

The first strategy to improve the optical performance of the Fat NFT was to optimize the boundaries of the thin-film pattern that is embossed onto the bulk chunk of gold. A 3D perspective view of the proposed HAMR system is shown in Figure 59. The incident light enters a slab Ta_2O_5 waveguide and evanescently couples to the Fat NFT. The waveguide that is shown is the bottom-most portion of the Seagate parabolic condenser or PSIM that was described in Figure 15. The NFT consists of a gold thin-film pattern directly touching a bulk chunk of gold with a gold peg protruding towards the ABS. A CoFe magnetic write pole sits on top of the NFT, and the write pole tip is 30nm above the top surface of the NFT peg. Not shown in these 3D views is the magnetic media stack, described in TABLE 4, which is adjacent to the right side of the waveguide, NFT peg and write pole tip.

We used the Inverse Electromagnetic Design software to computationally generate the optimal shapes in the thin-film pattern of the Fat NFT. The objective function in the optimization was the light intensity ratio between the hotspot and unwanted side lobes in the

media. The geometry that was optimized was a 2D binary bitmap of 350,000 pixels, where each pixel represented a 3D voxel of either SiO_2 or Au of dimensions 1nm×1nm×60nm occupying a total volume of 700nm×500nm×60nm, which was the region of the etched thin-film gold layer of the Fat NFT. Using mathematical morphology (MM), additional constraints on the binary bitmap were employed to enforce that the minimum feature size was greater than 50nm and the radius of curvature of any boundary was at least 25nm. The initial structure was 4 ellipses as shown in Iteration 0 of Figure 60, where the shaded region indicates where this thin-film layer touches the thick gold layer. The software calculated the *boundary* gradient and was constrained to only make iterative changes to the boundaries of the 4 shapes. Figure 60 shows the iterative optimization of the NFT thin-film pattern over 60 iterations, representing a total of only 120 simulations to optimize 350,000 degrees of freedom. Figure 61 shows a plot of the FOM versus iteration, showing smooth stable convergence towards a locally optimal design. Figure 62 shows a comparison of the optical coupling performance of an un-optimized and optimized Fat NFT design. The left case looks like the traditional lollipop antenna that is directly touching a thick film of gold. This is the same design shown on the right of Figure 58, and its poor performance led to the popular belief that optical antennas must be skinny thin-films by themselves. However, the optimized Fat NFT performs remarkably well, and a light intensity cross-section 5nm into the storage FePt layer of the hard disk medium shows a confined hotspot without significant background radiation. The big advantage of the Fat NFT, which will be discussed later in this chapter, is that the NFT tip will undergo less self-heating. Because of the thick film layer, the Fat NFT has a large solid angle of heat conduction from the NFT tip.

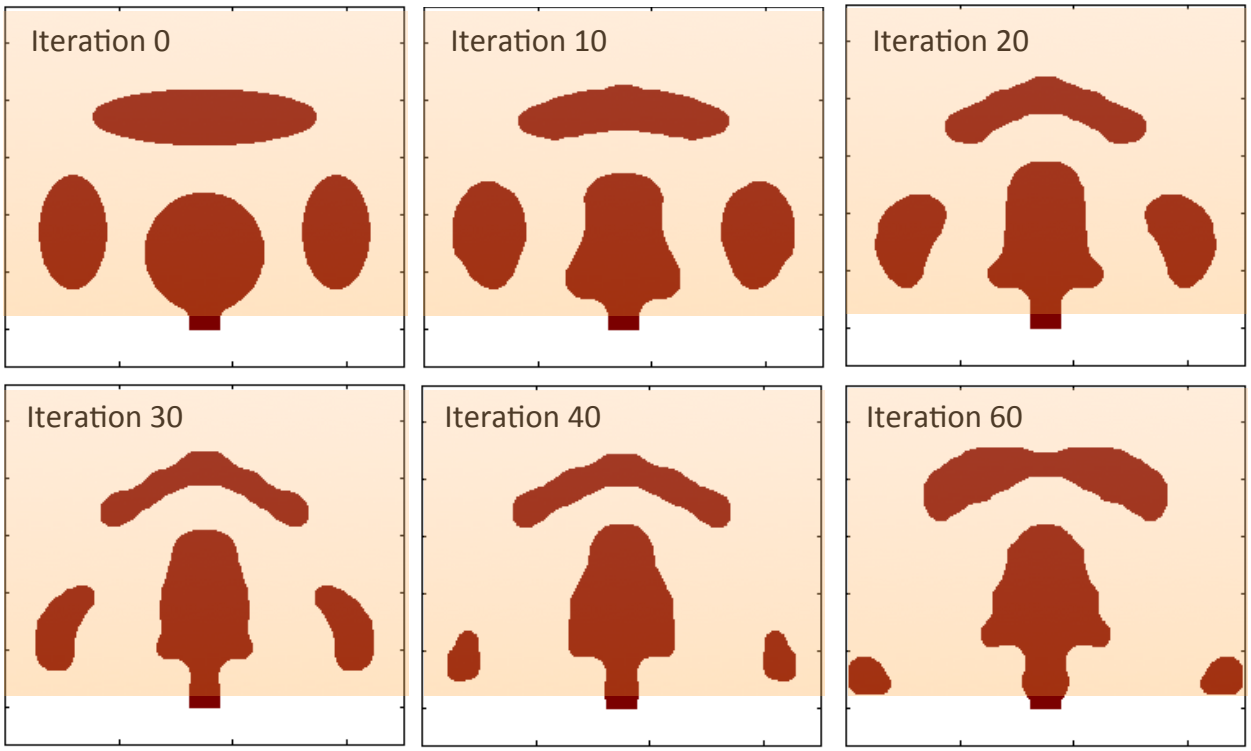

Figure 60: Top view and iterative optimizations of the thin-film pattern in the proposed Fat NFT. Red indicates Au, white indicates SiO_2 and the shaded regions indicates where the thin-film NFT pattern touches the thick film portion of the Fat NFT.

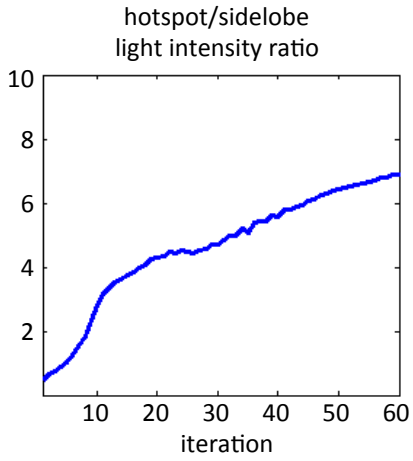

Figure 61: Convergence plot of the optimized FOM versus iteration. The FOM was the peak light intensity in the media hotspot divided by the peak light intensity in unwanted side lobes.

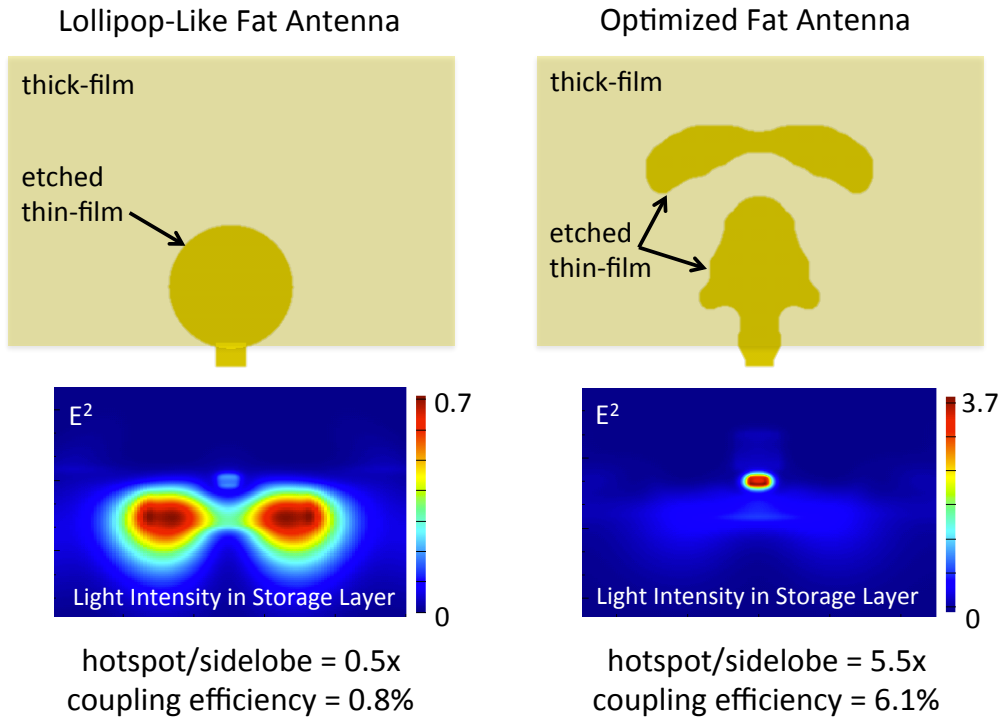

Figure 62: Comparison of the un-optimized and optimized Fat NFT. The un-optimized case shows poor optical performance, which is why the myth developed that fat antennas do not work. But, after running the Inverse Design software, it is clear that a properly designed Fat NFT performs very well.

5.5 Inverse Design of Low-Index Mode-Matching Grating

The previous section optimized the Fat NFT to better match the incoming light mode. This section aims to change the light mode in the waveguide to better match the NFT without any geometric design changes to the NFT. The physical strategy to improve the mode match was to

insert an array of holes of low index material etched into the high index slab waveguide. A 3D perspective view of the proposed HAMR system is shown in Figure 63. The incident light enters a slab Ta_2O_5 waveguide and evanescently couples to the Fat NFT. In this section, we let the thin-film pattern of the Fat NFT be the un-optimized disk design shown on the left of Figure 62. Instead, the waveguide is patterned with holes of SiO_2 to reshape the incident mode to better couple to the Fat NFT. A CoFe magnetic write pole sits on top of the NFT, and the write pole tip is 30nm above the top surface of the NFT peg. Not shown in these 3D views is the magnetic media stack, described in TABLE 4, which is adjacent to the right side of the waveguide, NFT peg and write pole tip.

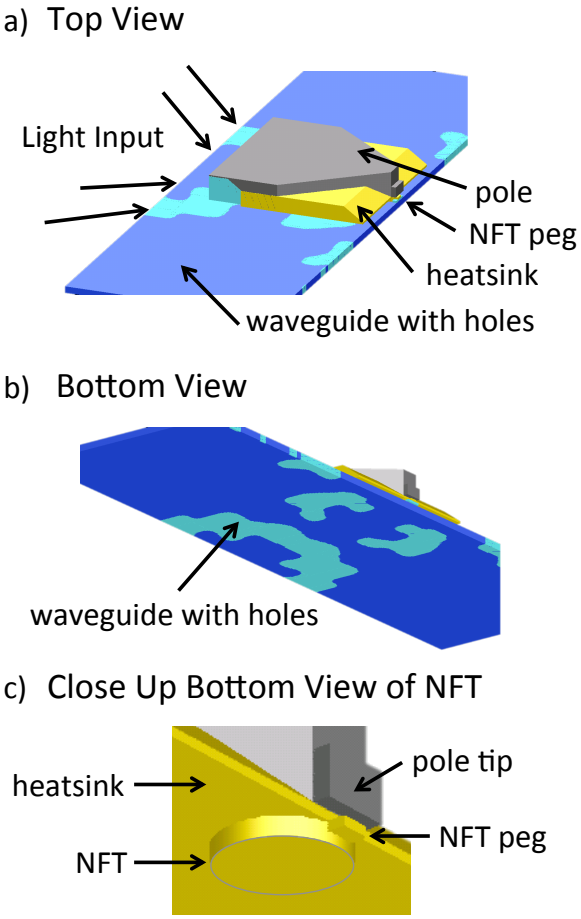

Figure 63: 3D views of the proposed HAMR light delivery structure. The Fat NFT is a thin-film disk embossed on a bulk chunk of gold, and the slab waveguide contains a pattern of low index material.

We used the Inverse Electromagnetic Design software to computationally generate the optimal waveguide pattern. The objective function in the optimization was the light intensity ratio between the hotspot and unwanted side lobes in the media. The geometry that was optimized was a 2D binary bitmap of 75,000 pixels, where each pixel represented a 3D voxel of either SiO_2 or Ta_2O_5 of dimensions 8nm×8nm×100nm occupying a total volume of 4μm×1.2μm×0.1μm, which was the region of the waveguide core under the NFT and adjacent to the ABS. Using mathematical morphology (MM), additional constraints on the binary bitmap were employed to enforce that the minimum diameter of a SiO_2 hole was greater than 128nm and the radius of curvature of any boundary was at least 64nm. Figure 64 shows the

iterative optimization of the holey waveguide pattern over 15 iterations, representing a total of only 30 simulations to optimize 75,000 degrees of freedom. In the first iteration, the software was configured to use the *sparse* gradient and added many new SiO_2 holes into the Ta_2O_5 waveguide core. In the latter iterations, the software calculated the *boundary* gradient and was constrained to make boundary changes only. Figure 65 shows a plot of the FOM versus iteration, showing smooth stable convergence towards a locally optimal design. Note, that between iterations 10 and 15, the geometry kept changing with little change to the FOM, suggesting that the optimal solution is robust to small variations in the boundaries of the waveguide pattern. And, although not specifically studied here, this also suggests that the optimal solution is robust to small variations in material permittivity.

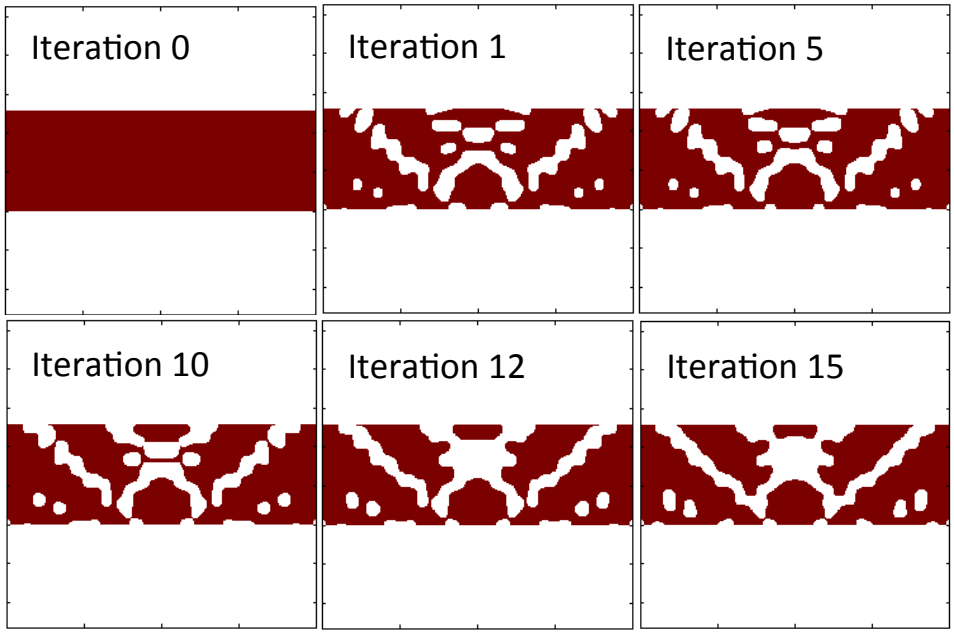

Figure 64: Top view and iterative evolution of a Ta_2O_5 slab waveguide core (red) patterned with SiO_2 holes (white). This computer-generated pattern offers more absorption in the hotspot and reduced unintentional erasure of adjacent tracks.

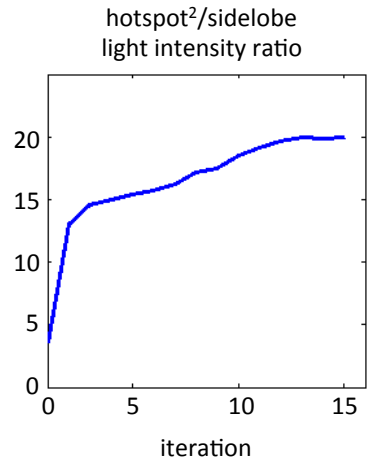

Figure 65: Convergence plot of the optimized FOM versus iteration. The FOM was the square of the peak light intensity in the media hotspot divided by the peak light intensity in the unwanted sidelobes.

For a fair comparison, we modeled a typical heatsink structure for a lollipop NFT, shown in Figure 66, that consists of a 120nm diameter gold cylinder connecting the center of the NFT to a bulk chunk of gold of the same dimensions used in the Fat NFT design. The narrowness of the cylindrical heatsink limits the heat conduction out of the NFT peg. The same geometries of the waveguide, cladding, write pole and media used in the Fat NFT model were used in the lollipop NFT model. We imported the optical absorption profile as a volumetric heat source in a thermal FEM model, in which we observed a peak temperature rise in the NFT peg of 450°C above ambient when injecting enough light into waveguide to achieve a desired 400°C temperature rise in the media hotspot.

The optimized waveguide pattern coupled to the proposed Fat NFT is shown in Figure 67. Using identical simulation models, we observed that the new proposed structure produces nearly identical optical properties of the media hotspot. Specifically, the hotspot shape is well defined by the NFT peg dimensions, the power absorbed in the hotspot normalized to power injected into waveguide is ~6%, and the light intensity ratio between the hotspot and undesired sidelobes is greater than 5. More importantly, we observed that the proposed Fat NFT has a peak temperature rise in the NFT peg of only 230°C above ambient when injecting enough light into the waveguide to achieve the same 400°C temperature rise in the media. This represents a ~50% lower temperature rise (220°C) compared to a lollipop NFT with a typical cylindrical heatsink. This is expected, because the typical cylindrical heatsink offers little solid angle of heat conduction. A ~50% reduction in NFT temperature could result in exponential improvements to HAMR write-head lifetimes.

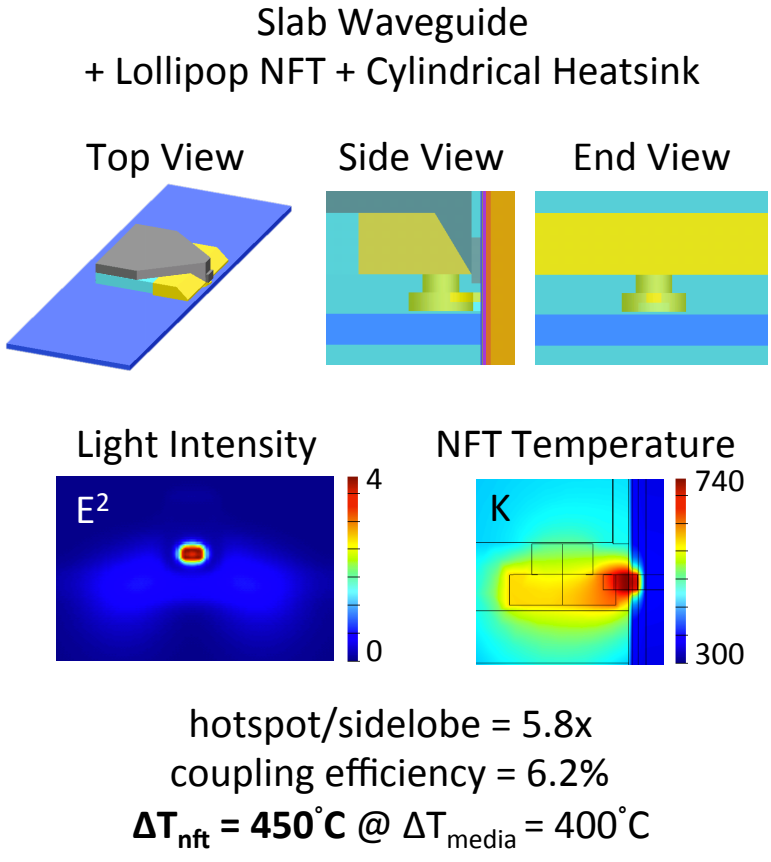

Figure 66: (top) Structural model of a slab waveguide, lollipop NFT, narrow cylindrical heatsink and write pole. (mid) Simulated light intensity in the media and side-view temperature profile of the NFT, heatsink and media. This design suffers from severe self-heating.

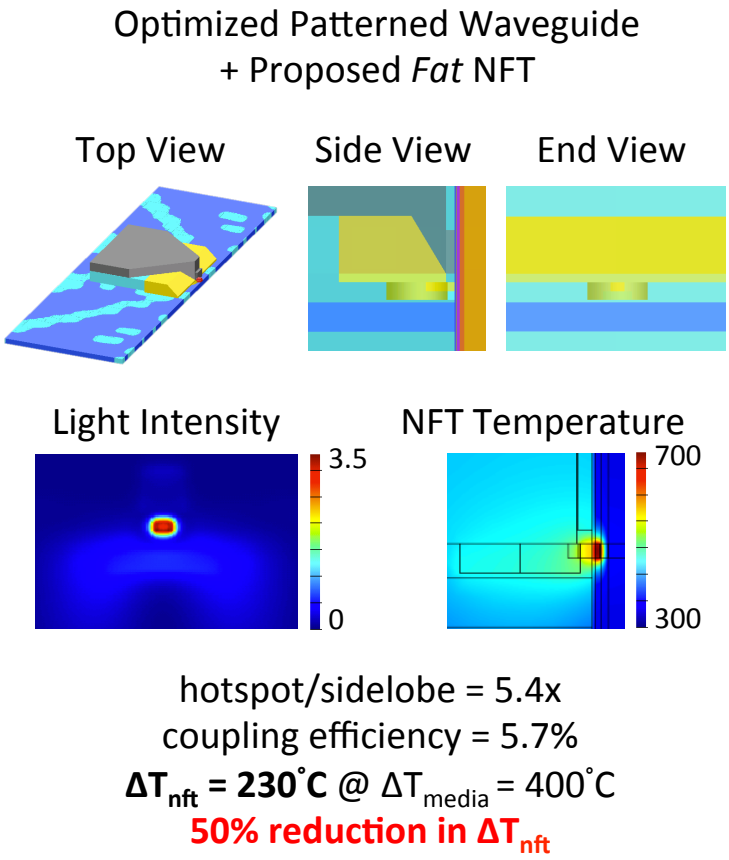

Figure 67: (top) Structural model of the proposed HAMR structure consisting of a patterned waveguide, Fat NFT, and write pole. (mid) Simulated light intensity in the media and side-view temperature profile of the NFT, heatsink and media. This design achieves desirable optical properties and significantly reduced self-heating.

5.6 Conclusions

The simple thermal analysis from Section 3.3 mandates that we have a *Fat* NFT with a large solid angle of heat conduction. The Fat NFT proposed in this chapter has different electromagnetic properties than the familiar skinny lollipop NFT. Thus, a different incident waveguide mode is required to properly excite the proposed NFT. With the power of Inverse Electromagnetic Design, we computationally generated complex waveguide patterns that mode-matched to the Fat NFT. The combined system of a patterned waveguide and Fat NFT produced the desired optical properties for HAMR nano-scale light delivery as well as a greatly reduced operation temperature. Reliability of structural and electronic devices often varies with the exponential of temperature. Hence, the new structures proposed in this chapter may offer orders of magnitude improvements to reliability by reducing the NFT self-heating by ~50% (220°C) compared to typical industry designs. We expect that computationally generated electromagnetic structures will have an important role in commercial HAMR technology.

6 Multi-Objective Design

Previous work achieved the focus of light to a 10 nm spot size with gold-coated tapered optical fibers, used in Near-Field Scanning Optical Microscopes (NSOMs). Although they are very valuable in spectroscopy, tapered fibers are remarkably inefficient, and typical optical transmission efficiency to a sub-100 nm spot is on the order of 10^{-5} to 10^{-7} [6]. Moreover, the maximum sustainable optical power at the output of a tapered fiber probe is limited by the significant self-heating via optical absorption experienced in the gold coating and nano-scale gold aperture, which causes structural deformation of the probe tip and thermal instabilities of nearby devices [4], [5]. Transmission efficiency and self-heating are thus significant bottlenecks towards higher optical power, higher signal-to-noise ratio and higher scan rate in sub-wavelength optical probes.

Recently, interest in sub-wavelength optical focusing has been renewed by the data storage industry toward commercializing Heat Assisted Magnetic Recording (HAMR). This data-recording scheme relies on focusing optical energy to locally heat the area of a single bit, several hundred square nanometers on a magnetic hard disk [8], [50]. In order to write data with a scan rate of 10 m/s, the optical system must heat the media by 400°C in less than 1ns, which amounts to ~1mW delivered to a 30nm spot [2]. Hence, for practical diode laser powers, the system must achieve an energy coupling efficiency of at least 1% or 10^5 times the transmission of tapered fiber. To achieve this, a diode laser is coupled to an on-chip optical waveguide that illuminates a metallic optical antenna or near-field transducer (NFT). The hotspot dimensions on the hard disk media are defined by the narrow tip of the metallic NFT that is ~30nm×30nm in cross-section and abuts the air-bearing surface (ABS) of the write-head. Like NSOMs, nano-scale metallic NFTs suffer from self-heating and structural deformation. Typical industry prototypes of HAMR light delivery systems experience self-heating of several hundreds of degrees in the NFT [23]–[25], [28]. This may severely limit the lifetime of prototype devices to orders of magnitudes below the required lifetime of many years for a commercial product. Therefore, designing more efficient and thermally stable sub-wavelength optical focusing systems, although a historically challenging problem, is imperative for commercial HAMR technology.

The optical focusing in a HAMR write-head involves the complex electromagnetic interactions between the illuminating waveguide, metallic NFT, magnetic write-pole and multi-layered hard disk media. Traditional geometric design approaches, based on intuition or highly constrained optimization such as parameter sweeps, are inadequate. We propose Inverse Electromagnetic Design software, which provides fast optimization of 3D electromagnetic structures with 100,000's of degrees of freedom. With such a large parameter space, an optimization can search for unexpected shapes of the NFT or waveguide that offer superior performance for HAMR. The drawback of such an optimization is computational expense. For applications like HAMR, a single 3D Maxwell simulation at optical frequencies of nano-scale

metallic structures and multi-layered media is computationally demanding even on modern high-performance computing clusters. Since heuristic algorithms like particle swarm and genetic algorithms rely on random trials, they are too computationally burdensome for practical engineering design for optics in the nano-scale. In contrast, the Inverse Electromagnetic Design software performs gradient-based optimization using the adjoint method, which results in fast deterministic optimization that is computationally inexpensive [10], [30], [32], [53].

Moreover, the optical spot size, optical throughput, and self-heating from optical absorption in the light focusing system are coupled electromagnetic phenomena. Hence, it is a weak strategy to treat the list of optical specifications as independent problems to be solved separately. In this chapter, we propose multi-objective optimization, where we actively maximize numerous objective functions simultaneously. This functionality is a natural extension to the Inverse Electromagnetic Design software, because we can calculate the gradient of numerous objective functions rather than the gradient of one objective function alone. We developed multi-objective Inverse Design and computationally generated unexpected geometric designs for the metallic and dielectric structures in a HAMR light delivery system that simultaneously achieves the desired optical focusing and low self-heating.

6.1 Figures of Merit for HAMR

6.1.1 Optical coupling efficiency

For practical and economic reasons, a commercial HAMR product must use a very inexpensive diode laser as the optical source. Hard disk drives are commoditized products with steep competition, and the addition of a sub-wavelength optical focusing system must be a marginal increase in cost. A practical upper limit on the diode laser power may be 10mW, and roughly 100μW must be absorbed in a 30nm spot on the media. Hence, an important Figure of Merit (FOM) for the optical design is the optical coupling efficiency, defined as

$$FOM_1 = \frac{\int_{V_{hotspot}} \frac{1}{2}\omega\epsilon_0\epsilon''_{media}\left|\vec{E}(x)\right|^2 dx}{P_{in}} \tag{6. 1}$$

where ω is the frequency of the excitation laser light, ε_0 is the free-space permittivity, ϵ'' is the imaginary part of the permittivity, $|E|^2$ is the light intensity in the media, and P_{in} is the input optical power. Relative to laser output power, this efficiency must be at least 2%. Figure 68 shows the qualitative difference between an unacceptable and a desired light intensity profile in the media.

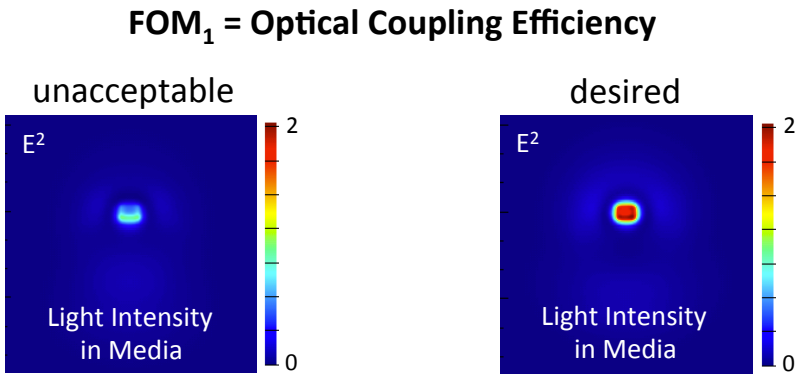

Figure 68: Unacceptable versus desired optical profile on the hard disk media. The coupling efficiency from laser to media must be at least 1%, given the maximum output power of inexpensive laser diodes to be used in a commercial HAMR hard disk drive.

6.1.2 Media/peg ratio

Similar to NSOM probe tips, HAMR systems rely on a conductive, but unfortunately soft, material like gold to provide the sub-wavelength focusing. The optical absorption in the metal causes it to self-heat, which is a fundamental root of failure mechanisms in HAMR systems and a crucial bottleneck that has yet to be solved. Hence, another FOM to be optimized is the temperature ratio between the media and the NFT tip or peg. In this chapter, we optimized the media/peg temperature ratio by optimizing a related electromagnetic function, which is the absorption ratio between the media and peg, shown in (6. 2). Figure 69 shows example cross-sectional temperature profiles through the NFT, NFT peg and hard disk media. The NFT temperature must be low enough to avoid structural deformation, to prevent heating and subsequent lowering of the nearby magnetic write pole's permeability, and to ensure chemical stability of the surrounding materials for many years.

$$FOM_2 = \frac{\int_{V_{hotspot}} \epsilon''_{media} \left|\vec{E}(x)\right|^2 dx}{\int_{V_{peg}} \epsilon''_{NFT} \left|\vec{E}(x)\right|^2 dx} \tag{6. 2}$$

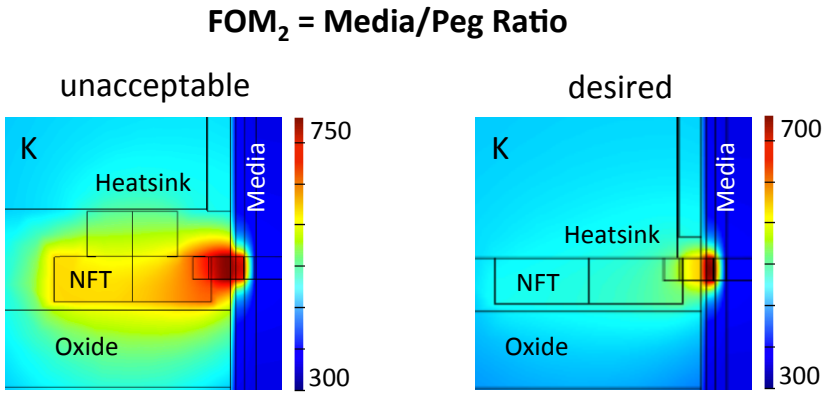

Figure 69: Unacceptable versus desired media/NFT temperature ratio for improved reliability and lifetime of HAMR systems.

6.1.3 Hotspot/sidelobe ratio

For high SNR recording, it is not only sufficient to have the desired temperature profile on the recorded track. In addition, the optical system must guarantee that adjacent tracks experience low enough temperatures to suppress unintentional recording or erasing of data. The FOM to describe this effect is the hotspot/sidelobe temperature ratio. Once again, in this chapter, we attempt to optimize a related electromagnetic FOM that is the hotspot/sidelobe absorption ratio given in (6. 3). Figure 70 depicts an unacceptable hotspot/sidelobe ratio in which adjacent tracks would experience high temperatures, thus interfering with nearby data on the disk, versus the desired profile.

$$FOM_3 = \frac{\max\limits_{x \in V_{hotspot}} \left|\vec{E}(x)\right|^2}{\max\limits_{x \notin V_{hotspot},\ x \in V_{media}} \left|\vec{E}(x)\right|^2} \tag{6. 3}$$

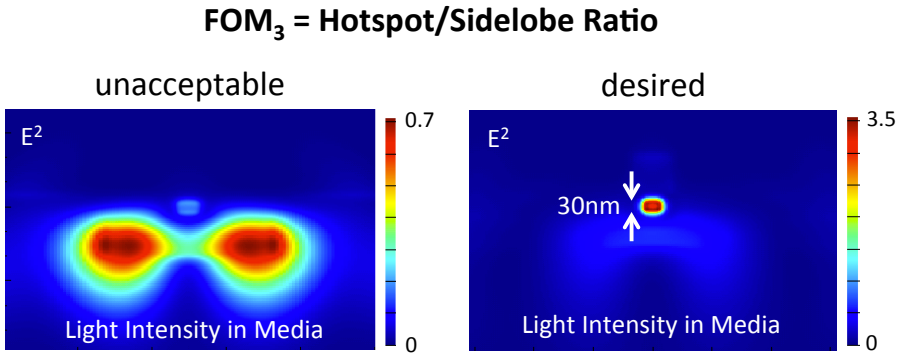

Figure 70: Unacceptable versus desired hotpot/sidelobe light intensity ratio in the media to suppress adjacent track interference.

6.2 Multi-objective optimization

Multi-objective optimization is a simple extension of single-objective optimization. Since we cannot easily visualize N-dimensional vectors, we will depict the gradient with respect to only 2 geometric parameters. If we are maximizing FOM_1 given the gradient $\vec{\nabla} FOM_1$, then there exists a hemisphere of possible geometry updates $\Delta\vec{x}$ such that $\Delta\vec{x} \cdot \vec{\nabla} FOM_1 > 0$. Given the gradient for two different FOMs, as shown in Figure 71a, there exists a cone of possible geometry updates $\Delta\vec{x}$ such that both FOMs increase in value. In a similar scenario, we sometimes want to optimize FOM_1 and apply a constraint that FOM_2 is preserved at a particular value. Then, the optimal iterative geometry update is $\Delta\vec{x}$ shown in Figure 71b, where $\Delta\vec{x}$ has a positive projection onto $\vec{\nabla} FOM_1$ and has zero projection onto $\vec{\nabla} FOM_2$. This latter case is a typical constrained optimization that can be formulated using the method of Lagrange multipliers. For electromagnetic problems that are highly non-linear, Sequential Quadratic Programming (SQP) is a superior iterative optimization method to the linear programming methods shown in Figure 71. SQP involves approximating the second derivative and performing a quadratic programming routine to solve for the optimal $\Delta\vec{x}$. SQP was not used for the results in this chapter, but was implemented in a later distribution of our Inverse Electromagnetic Design software and indeed offered superior convergence.

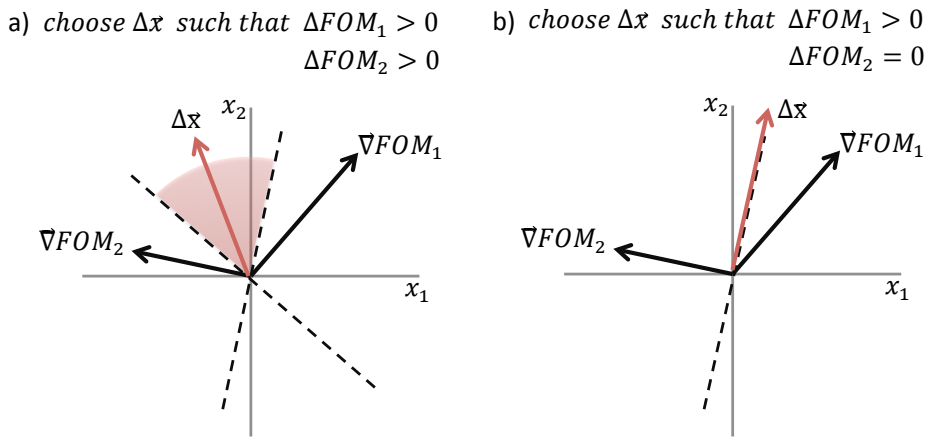

Figure 71: Plots of allowable iterative geometry updates $\Delta\vec{x}$ given gradient vectors of two different objective functions. a) The optimization goal is improve both objectives. b) The optimization goal is to improve one objective without sacrificing the other.

6.3 Mock industry system

For comparison, we modeled a typical HAMR light delivery system shown in Figure 72, consisting of a magnetic write pole, multi-layered media stack, lollipop NFT, cylindrical heatsink and incident light injected into a slab waveguide similar to that of the parabolic solid-immersion mirror (PSIM) [2]. The NFT is a 60nm-thick 210nm-diameter cylinder with a rectangular peg that is 25nm in length and 30nm×50nm in cross-section. A 120nm diameter cylinder acted as a heatsink and was attached to the center of the NFT body. Since there is very little electric field at the center of the NFT, the combined heatsink-NFT structure does not overtly deviate in electromagnetic behavior from that of a typical thin film lollipop NFT. The narrowness of the cylindrical heatsink limits the heat conduction out of the NFT peg and is major bottleneck towards increasing the media/NFT temperature ratio [30].

6.4 Proposed system

The proposed structure shown in Figure 72 has many similar elements to the mock industry structure. There are 2 key differences. First, the NFT is not a thin film device connected to a narrow heatsink, but rather it is a thin film pattern of gold embossed on a bulk chunk of gold. This *fat* NFT has a large solid-angle of heat conduction from the NFT peg. The proposed *fat* NFT is not a lollipop NFT, because the electromagnetic behavior is very different. The bulk gold layer in the *fat* NFT not only acts as an aggressive heatsink but also significantly affects the NFT's electromagnetic behavior. The aggressive heatsink is part of the antenna. Hence, a typical PSIM may mode-match to the lollipop NFT but does not mode-match to the *fat* NFT [30]. Nonetheless, the *fat* NFT can indeed focus light into a hotspot on the media with dimensions defined by the NFT peg as long as the incoming illumination mode-matches to it. This leads us to the second key difference. In the proposed system, the waveguide core is patterned with holes of higher-index material. Here, the core is Ta_2O_5, the cladding is SiO_2, and the patterned holes in the waveguide are poly-Silicon. The light is thus not only

interacting with the metallic NFT but also with a grating structure built into the waveguide. By optimizing the shapes of both the NFT and the waveguide grating, we found that there were enough degrees of freedom to improve all of the FOMs simultaneously.

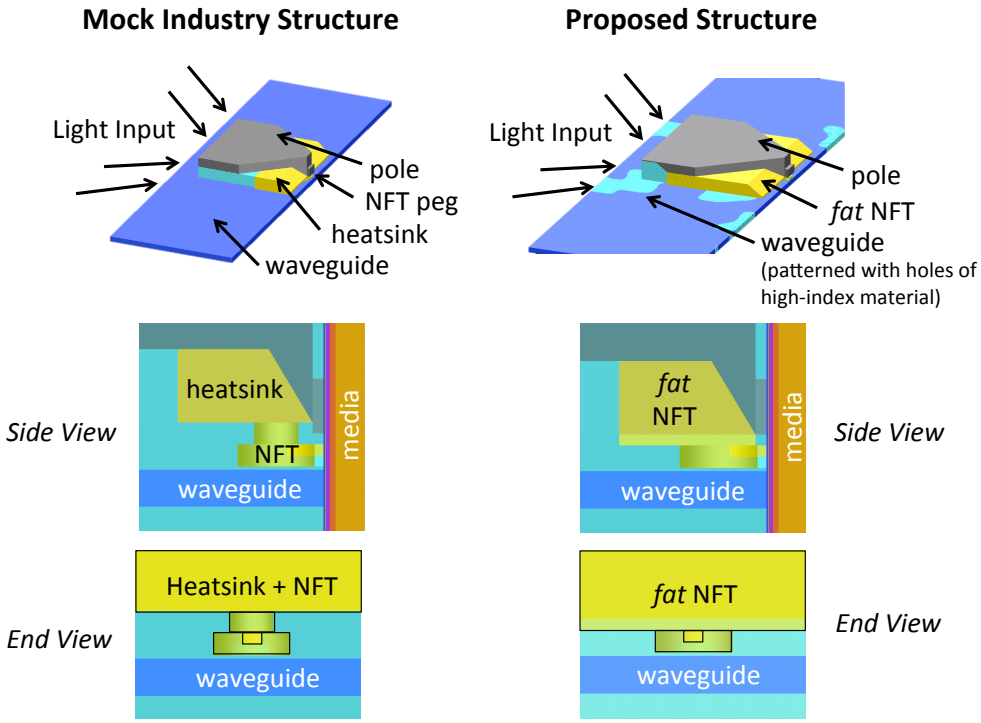

Figure 72: Mock industry structure, on left, consists of a lollipop NFT and cylindrical heatsink. The proposed structure, on right, consists of a *fat* NFT and a high-index grating embedded in the waveguide.

6.5 Computational models

Simulation results in this chapter use a commercial finite-difference time-domain Maxwell solver, Lumerical FDTD, in which a pulse of light is injected into the waveguide of the HAMR system and propagated in the time domain towards the NFT, write pole and media until the pulse energy has decayed beyond our desired precision. A detailed mesh convergence test was performed to ensure minimal computational error due to discretization. The most crucial and computationally demanding mesh requirements were 1nm cubic Yee cells in the metallic NFT and 0.5nm Yee cell thicknesses in the media to resolve the various nanometer-thin layers of the media stack. We used an in-house high-performance computing cluster consisting 336 cores and 668GBs RAM over 26 nodes. By parallelizing the solver through a Message-Passing Interface, Open MPI [54], and 40Gb/s Infiniband interconnects, our in-house cluster can simulate FDTD models of half a million Yee cell nodes. Typically, we ran simulations on 64 to 128 cores at a time, with which we could run a single-objective optimization of a HAMR optical system in ~1 day.

TABLE 6 contains structural and optical properties at the operation laser wavelength of 830nm for the numerous write-head and media components. These properties were chosen to closely mimic designs from industry publications and patent literature. TABLE 7 shows

thermal properties that were assumed for a thermal finite-element model performed in COMSOL Multiphysics to predict the media and NFT temperatures caused by volumetric heat sources from the optical absorption in the media hotspot, NFT peg and NFT body.

TABLE 6: STRUCTURAL AND OPTICAL PROPERTIES IN PROPOSED HAMR SYSTEM

Device	Dimensions	n	k
Au NFT	60 nm thick	0.16	5.08
Au NFT Peg	50 nm wide at ABS, 30 nm thick	0.16	5.08
Ta_2O_5 Waveguide	100 nm thick	2.1	-
Si Holes in Waveguide	100 nm thick	3.66	0.0045
SiO_2 Cladding	-	1.4	-
CoFe Writepole	120 wide at ABS	3	4
Head Overcoat	2.5 nm thick	1.6	-
Air Gap	2.5 nm thick	1.0	-
Media Overcoat	2.5 nm thick	1.2	-
FePt Recording Layer	10 nm thick	2.9	1.5
MgO Interlayer	15 nm thick	1.7	-
Au Media Heatsink	80 nm thick	0.26	5.28
Glass Substrate	infinite	1.5	

TABLE 7: THERMAL PROPERTIES IN PROPOSED HAMR SYSTEM

Material	Specific Heat (J/m^3K)	Thermal Conductivity (W/mK)
Au	$3 \cdot 10^6$	100
Ta_2O_5	$2 \cdot 10^6$	2
SiO_2	$2 \cdot 10^6$	1
CoFe	$3.5 \cdot 10^6$	20
FePt - Lateral	$3 \cdot 10^6$	5
FePt - Vertical	$3 \cdot 10^6$	50
MgO	$2 \cdot 10^6$	3

6.6 Inverse Design of High-Index Grating and NFT

We applied our Inverse Electromagnetic Design software to the proposed sub-wavelength optical focusing system to improve all of the HAMR FOMs in Section 6.1 simultaneously. The degrees of freedom in the optimization were the boundaries of a silicon-Ta_2O_5 grating pattern in the waveguide core and the thin film pattern of gold in the *fat* NFT. Specifically, the waveguide grating geometry was represented as a 666x216 binary bitmap, where each pixel represents a 6nm×6nm×100nm voxel of Si or Ta_2O_5 within a 4000nm×1300nm×100nm region of the feeding Ta_2O_5 waveguide that is underneath the NFT and abuts the ABS. Similarly, the geometry of the thin film portion of the NFT was represented as a 400x350 binary bitmap, where each pixel represents a 1nm×1nm×60nm voxel of Au or SiO_2 within a 400nm×350nm×60nm region. In total, there were ~300,000 degrees of freedom. The silicon geometry was constrained to have a minimum dimension of 96nm and a radius of curvature of 48nm. The gold NFT geometry was constrained to have a minimum dimension of 30nm and a radius of curvature of 15nm. During each iteration, we calculated the gradient of the latter two FOMs listed in Section 6.1: media/peg absorption ratio and hotspot/sidelobe absorption ratio. For a multi-objective optimization of 2 FOMs, 1 *forward* simulation and 2 *adjoint* simulations were needed per iteration. Each 3D FDTD Maxwell simulation took ~1 hour when parallelized on 64 cores in our in-house computing cluster. Hence, we ran ~30 iterations of the optimization method in ~3.75 days.

Figure 73 shows convergence plots versus iteration for 4 different FOMs. The optical coupling efficiency and the media/NFT ratio were not optimized but were evaluated every iteration for extra information. In these calculations, the media hotspot absorption was taken to be the integral of optical absorption within a 100nm×100nm×10nm volume containing the hotspot in the FePt layer of the hard disk media. Optical coupling efficiency was normalized to the power injected into the waveguide and does not include coupling losses between the laser and the waveguide. Figure 74 shows the top view of the initial and optimized geometries of the NFT and high-index waveguide grating. The top of Figure 75 shows the light intensity 5nm into the FePt layer on a logarithmic scale and normalized to the peak intensity. This shows a significant increase in the hotspot/sidelobe ratio of ~16x in the proposed design versus ~6x in the mock industry design. The bottom of Figure 75 shows the temperature profile through a vertical cross-section through the NFT and media in the proposed and mock industry designs. The NFT in the mock industry design heated by 450°C above ambient when enough light is injected to reach a 400°C temperature rise in the media. Notably, the proposed system only heated by 170°C above ambient when the media rises by 400°C. This amounts to a ~60% reduction in temperature rise, which perhaps could lengthen the lifetime of HAMR write-heads by several orders of magnitude. TABLE 8 contains a detailed list of numerous FOMs comparing the proposed design with the mock industry design, which were studied using identical electromagnetic and thermal models.

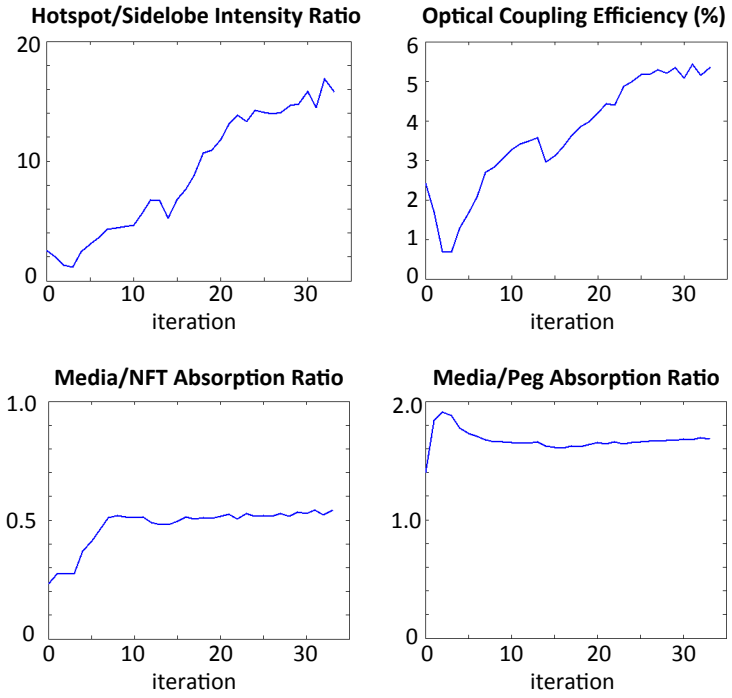

Figure 73: Convergence plots of 4 FOMs versus optimization iteration in a multi-objective optimization. Every FOM was either improved or preserved. No FOM was sacrificed in order to improve another.

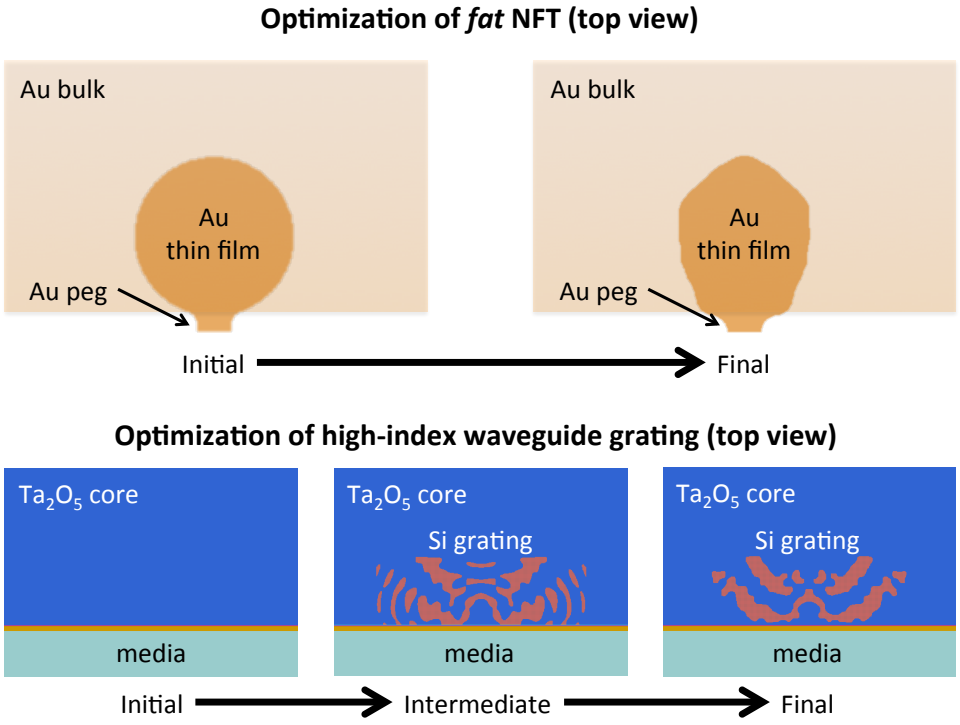

Figure 74: Initial and final shapes (top view) of the proposed fat NFT and high-index waveguide grating. These 3D planar geometries were represented by 2D binary bitmaps containing ~300,000 degrees of freedom.

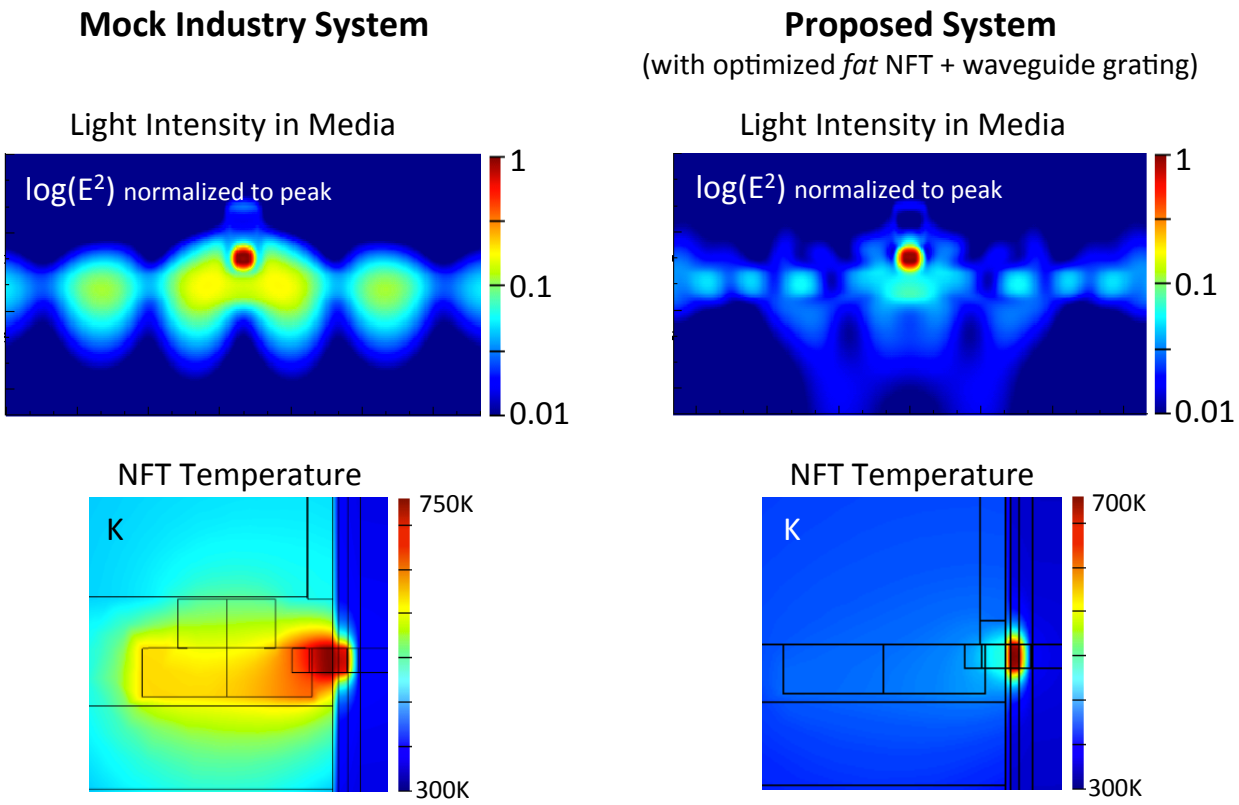

Figure 75: The top shows the light intensity on a logarithmic scale normalized to the peak in a cross-section 5nm into the FePt layer of the hard disk. The bottom shows the temperature profile in a vertical cross-section through the NFT and media. The proposed system experiences a significant reduction in adjacent track interference and NFT self-heating compared to the mock industry system.

TABLE 8: SIMULATED OPTICAL AND THERMAL BEHAVIOR

	Industry System	Proposed System
Optical Coupling Efficiency	6.2%	5.4%
Hotspot/Sidelobe Intensity Ratio	5.8x	15.8x
Media/NFT Absorption Ratio	0.28x	0.54x
Media/Peg Absorption Ratio	0.96x	1.7x
Media/NFT Temperature Rise Ratio	0.9x	2.4x
NFT Temperature @ ΔT_{media}=400°C	450°C	170°C

6.7 Conclusions

Inverse Electromagnetic Design is a powerful optimization tool for nano-optics design. In this chapter, the software optimized ~300,000 degrees of freedom in the 3D shapes of a gold NFT and silicon-Ta_2O_5 waveguide grating in a practical sub-wavelength optical focusing system for HAMR. Before this work, it was unknown whether the various optical specifications for HAMR could be met simultaneously, but we successfully determined that indeed multiple electromagnetic objective functions could be optimized together. We computationally generated the design of a HAMR light delivery system that has a sufficient optical coupling efficiency, an improved suppression of adjacent track interference, and a greatly reduced NFT temperature rise. Reliability of structural and electronic devices often varies with the exponential of temperature. Hence, the new structures proposed in this chapter with a 60% reduction in NFT temperature rise (280°C) may improve the NFT reliability and HAMR write-head lifetime by orders of magnitude. The optimized shapes that are presented here would be very challenging to generate via other design methods.

7 Higher Areal Density

Throughout this dissertation, we constantly evaluate what the important metric functions for HAMR really are. So far, the emphasis has been on obtaining the Curie point temperature (~450°C rise) in the media without stray light that obviously erases information and without the NFT operating at too high of temperature. But, have we actually achieved higher areal recording density? If not, then we have merely designed a new hard disk drive that is more expensive and complex than today's Perpendicular Magnetic Recording (PMR) drive.

Recalling the relationship from (2. 5), the linear recording density is proportional to the downtrack thermal gradient. Specifically, the crucial metric is the temperature fall per nanometer at the Curie point contour of the hotspot near the magnetic write-pole as labeled in Figure 31. This implies that the peak temperature must actually be above the Curie point in order to achieve a high gradient. In the example shown here, the peak temperature is 800K and Curie point is 750K. In real media, the Curie point is not a single value but rather a distribution, so we actually require a particular downtrack gradient for a small temperature range around the Curie point. The crosstrack width and crosstrack gradient determines the width of the written tracks of data and level of noise written to nearby tracks. Hence, the crosstrack gradient determines the track density and the downtrack gradient determines the linear density. Typically, the optimal bit shape to achieve the highest SNR for a particular read/write head and media has an aspect ratio between 1 and 10 with the track width much larger than the bit length. A typical lollipop antenna on the reference media described earlier offers only 9K/nm downtrack thermal gradient as shown in Figure 31, whereas we need a gradient >15K/nm to achieve the 1Tb/in^2 and higher data densities promised by HAMR.

In experiment, the simple way to increase the thermal gradients is to increase the laser input power. Increasing the input power will increase the peak temperature in the medium and consequently increase the gradient. However, increasing the input power also means that the hotspot contour at the Curie point will increase in diameter, meaning that the track density will be lower. Moreover, increasing the input power will increase the operating temperature of the NFT and thereby ruin the NFT's reliability. The goal of this chapter is to discover whether we can increase the thermal gradient while neither sacrificing track density nor NFT operating temperature.

7.1 Optimizing the Hard Disk Media for Thermal Gradient

Sharper thermal gradient can be achieved at the expense of cooling the media faster, which means that the media/NFT temperature ratio will decrease. This is the crucial trade-off that is currently halting the hard disk industry from commercial product. With traditional thin NFTs like the lollipop antenna, which nominally operates at a high temperature, experiments with media containing optically and thermally conductive heatsinks would show a faster failure

time of NFTs. Because the failure rate in most HAMR experiments has been orders of magnitudes below the desirable 5 year lifetime, a myth in the hard disk industry developed that we cannot tolerate high conductivity heatsinks in the media.

However, there is no alternative method to achieve higher thermal gradients and higher areal density via HAMR. Instead, we must look back at Section 3.3 and decide that we must compensate and increase the media/NFT temperature ratio via an aggressive heatsink on the NFT tip. An aggressive heatsink contains s large solid angle of heat conduction from the NFT tip. The media must look like that in the right frame of Figure 34, in which there is a heatsink very close to the FePt (perhaps 5nm below) with high electrical and thermal conductivities (perhaps gold). The NFT must not look like the traditional skinny lollipop antenna but rather the fat NFT proposed in this dissertation, shown in Figure 35.

To summarize, to achieve high thermal gradient and low NFT temperature:

1) A media heatsink with high electrical and thermal conductivities must be close to the FePt storage layer.

2) The NFT must be fat with a large solid angle of heat conduction from the NFT tip.

3) The NFT and excitation waveguide must have a high optical coupling efficiency.

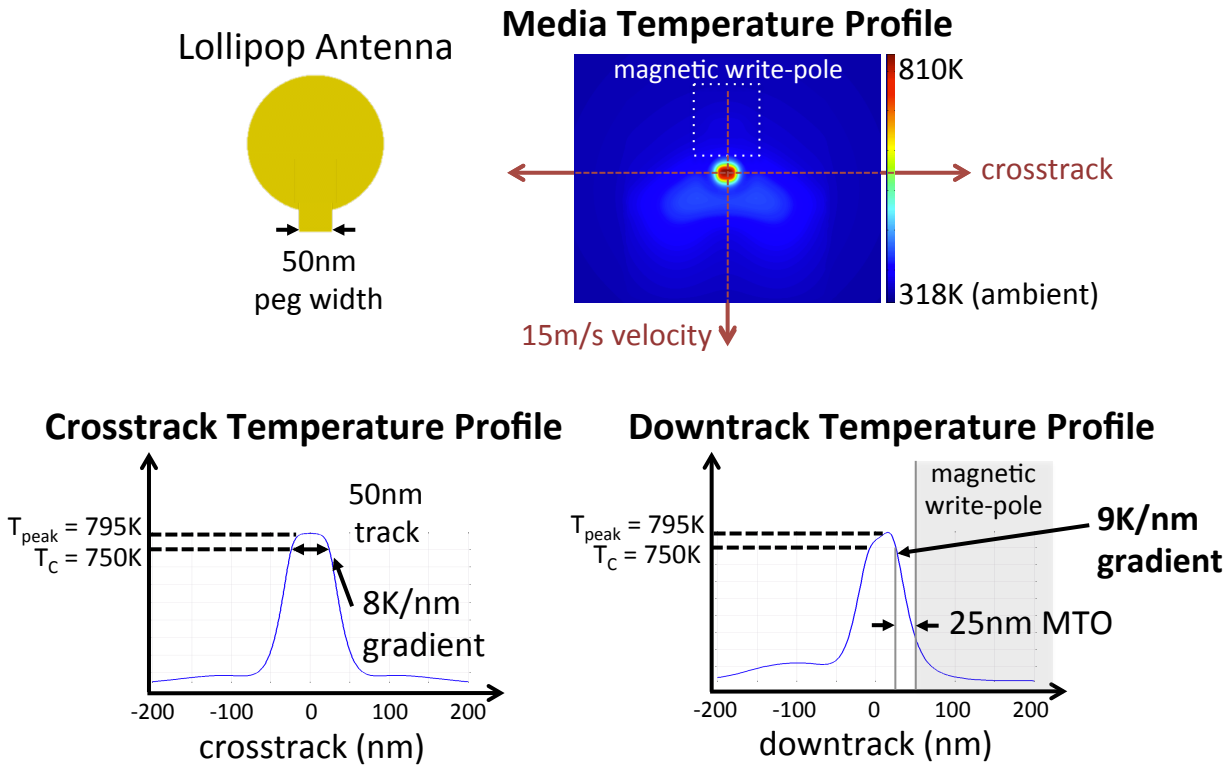

Figure 76: For a Lollipop antenna excited by a PSIM structure, the temperature in the middle cross-section of the FePt recording layer in the hard disk media is shown in the top-right. Linear plots in the crosstrack and downtrack directions are shown on bottom. The thermal gradients are calculated at the contour of the Curie point of ~750K.

7.2 Single-Mode Nano-Focusing with a Fat NFT

As discussed in Section 2.6, we need single-mode illumination of the single-mode NFT. The author does not recommend the designs previously shown in this dissertation with PSIM illumination modes. Results with the PSIM were shown, because it is the most standard HAMR system model for our industry partners to compare against. In this chapter, we will show computationally-optimized designs for a Fat NFT coupled to a TM single-mode rectangular waveguide. We will compare its results with two other systems.

In this chapter, there are many variables that we are modifying, so it is worth discussing in detail the three systems shown in Figure 77 that will be discussed here. System 1 is the canonical Seagate lollipop antenna with narrow heatsink that was compared against in Chapters 5 and 6. System 2 is the system that was optimized in Chapter 6 (optimized fat NFT with TE PSIM patterned with high-index material). System 3 is the new proposed system of TM single-mode rectangular waveguide coupled with the fat NFT. System 1 was coupled with the typical media stack that has been studied previously in this dissertation. Systems 2 and 3 used a modified stack, where the MgO underlayer had a thickness of 1nm. Typically, making the underlayer thinner would cause the NFT operating temperature to rise, but instead we combat the NFT temperature rise with the aggressive heatsinking inherent to the fat NFT. A detailed list of the media stack is provided in TABLE 9.

TABLE 10 shows the thermal properties that were assumed in an FEM model for heat conduction. The media velocity with respect to the head was taken to be 15 m/s. Since it was not discussed earlier, media velocity smears the thermal profile on the disk and lowers the thermal gradient (but is of course a necessary consideration). In attempt to better match industry simulation models, the thermal properties assumed in this chapter are very different than the previous chapters. Mainly, we assumed a higher thermal conductivity gold (in the NFT body, NFT tip and media heatsink) and lower thermal conductivities in the FePt. We will also added thermal barriers between metals and dielectrics in the simulation model.. The author believes that this model offers a gross underestimate of the NFT temperature rise, because a nano-rod of metal smaller than the mean-free-path of electrons in bulk material should have dramatically lower thermal conductivity. Taking into account the degradation of thermal conductivity in the NFT peg would not only predict higher NFT operating temperatures than described in this chapter, but also it would predict a larger relative reduction in temperature for the fat NFT systems.

7.3 Inverse Design of a Fat NFT for TM-Mode Excitation

In this chapter, we applied the Inverse Electromagnetic Design software to computationally optimize a gold fat NFT structure coupled to the single-mode waveguide in System 3. The NFT is composed of 3 structures of gold, the NFT tip, a thin-film structure and thick-film structure. The thick-film structure in the fat NFT does not merely act heatsink but rather it also affects the electromagnetics of the NFT. However, in this chapter, we only optimized the thin-film body of the NFT, so we expect further optimization is possible of the fat NFT. We enforced the thin-film structure to have constant thickness of 60nm, and the initial structure was a rectangle rather than a circle. The NFT tip was constrained to be a rectangular prism with a width of 50nm at the ABS, thickness of 20nm and length 30nm. The entire thin-film structure in the NFT was in direct contact with the thick-film gold. We enforced a radius of curvature of the gold boundary of 25nm and minimum dimension of 50nm.

TABLE 9: STRUCTURAL AND OPTICAL PROPERTIES IN PROPOSED HAMR SYSTEM

Device	Dimensions	n	k
Au NFT	60 nm thick	0.16	5.08
Au NFT Peg	50 nm wide at ABS	0.16	5.08
Au NFT Peg (Systems 1 & 2)	30 nm thick at ABS	0.16	5.08
Au NFT Peg (System 3)	20 nm thick at ABS	0.16	5.08
Ta_2O_5 Waveguide	100 nm thick	2.1	-
Si Holes in Waveguide (System 2)	100 nm thick	3.66	0.0045
SiO_2 Cladding	-	1.4	-
CoFe Writepole	120 wide at ABS	3	4
Head Overcoat	2.5 nm thick	1.6	-
Air Gap	2.5 nm thick	1.0	-
Media Overcoat	2.5 nm thick	1.2	-
FePt Recording Layer	10 nm thick	2.9	1.5
MgO Interlayer (System 1)	15 nm thick	1.7	-
MgO Interlayer (System 2 & 3)	1 nm thick	1.7	-
Au Media Heatsink	80 nm thick	0.26	5.28
Glass Substrate	infinite	1.5	-

Figure 78 shows the optimized NFT shape for System 3 and various HAMR metric functions to compare the various systems. Both Systems 2 and 3 show an improved media/NFT temperature ratio despite being coupled to media stacks with more aggressive heatsinking. System 3 showed ~10% optical coupling efficiency, 3x higher than the lollipop/PSIM system and 2x higher than the optimized fat NFT with PSIM illumination. System 3 showed a significant increase in optical absorption in the NFT tip. Although it was not studied here, we expect that making the peg slightly conical or perhaps 30nm thick rather than 20nm could reduce the NFT tip absorption. If the NFT tip absorption could be reduced in System 3, then perhaps it could offer as great of a temperature reduction as System 2 offers.

Figure 79 shows the thermal simulations of the three systems in HAMR operation, defined here as injecting enough input laser power to reach a track width at 750K of 50nm assuming 318K ambient temperature. We assumed a media velocity of 15m/s. Despite a hard disk that quickly wicks heat away, Systems 2 and 3 with much higher optical efficiencies require much less input laser power. The next row of plots shows the temperature profile in a vertical cross-section through the NFT, write-pole and media stack. As predicted by the temperature ratios

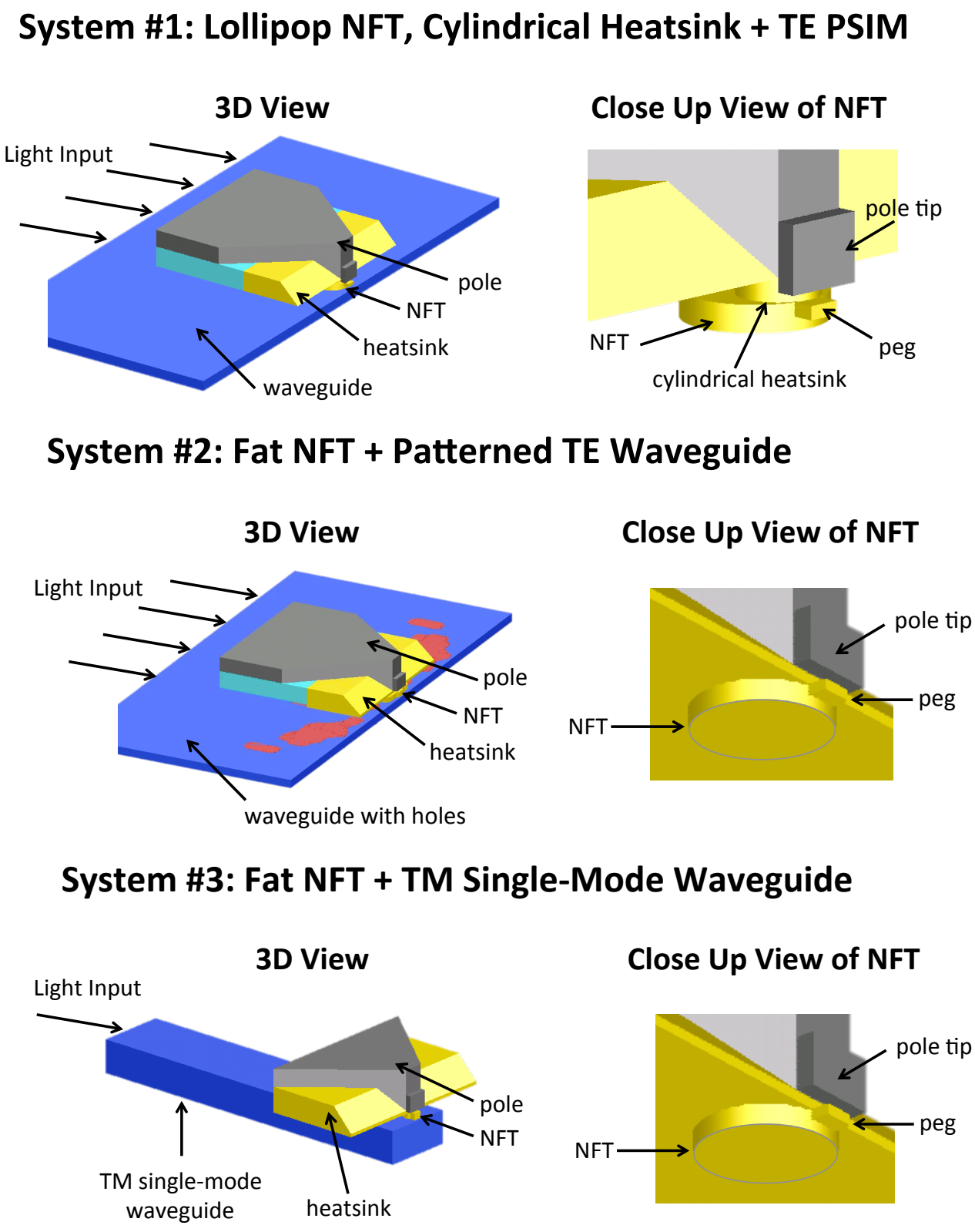

Figure 77: The 3 optical nano-focusing systems for HAMR that are compared in this chapter. System 3 is the only design that is recommended by the author, because it couples a fat NFT to a single-mode waveguide.

TABLE 10: THERMAL PROPERTIES IN PROPOSED HAMR SYSTEM

Material	Specific Heat (J/m^3K)	Thermal Conductivity (W/mK)
Au	$3 \cdot 10^6$	200
Ta_2O_5	$2 \cdot 10^6$	2
SiO_2	$2 \cdot 10^6$	1
CoFe	$3.5 \cdot 10^6$	20
FePt - Lateral	$3 \cdot 10^6$	0.5
FePt - Vertical	$3 \cdot 10^6$	5
MgO	$2 \cdot 10^6$	5
Thermal barrier between metals and dielectrics	Thin Approximation	1nm barrier with 0.5W/mK
Media Velocity = 15 m/s		

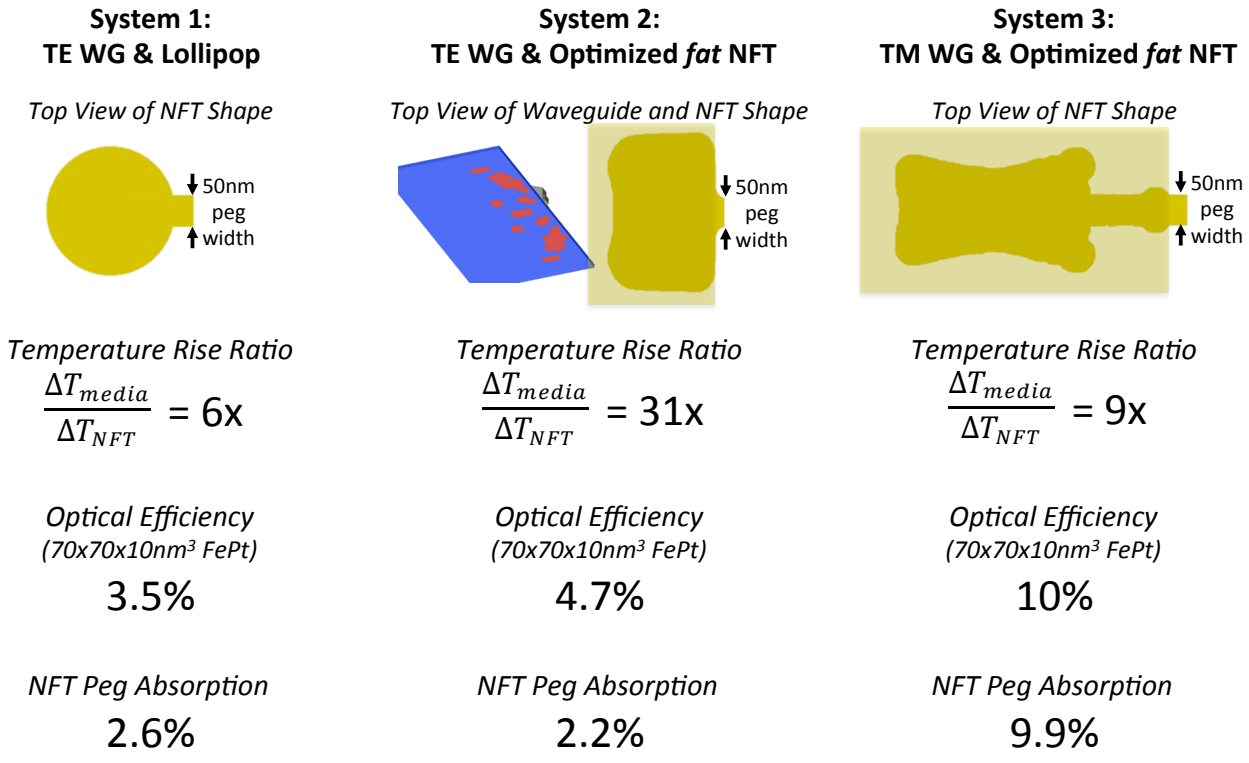

Figure 78: A comparison of HAMR merit functions between the 3 systems compared in this chapter. System 3 is the computationally generated shape of a fat NFT coupled to a single-mode waveguide that is the center-point of this chapter. System 1 is the canonical lollipop antenna with narrow heatsink. System 2 is the optimized design from Chapter 6.

shown in Figure 78, System 2 and 3 have a lower peak temperature of the NFT tip under practical optical illumination. The next row shows the temperature profile 5nm into the FePt layer of the media. Because of the much greater optical nano-focusing efficiency of System 3, the hotspot is qualitatively sharper and the background heating of the media is greatly reduced. The final two rows show quantitatively the thermal hotspot shape and gradient in the downtrack and crosstrack directions for the 3 systems. System 1 merely achieves 8K/nm crosstrack gradient and 9K/nm downtrack gradient. System 2 achieves 16K/nm crosstrack gradient and 21K/nm downtrack gradient. System 3 achieves 23K/nm crosstrack gradient and 26K/nm downtrack gradient. System 2 and 3 offer downtrack thermal gradients not only matching the HAMR specifications for 1Tb/in^2 recording, but these gradients actually drastically exceed the requirement of 15K/nm downtrack gradient. Note, that the grand achievement here is not merely a higher downtrack thermal gradient, but rather that it was achieved while lowering the NFT tip temperature, lowering the injected laser power, maintaining a narrow 50nm track width, and lowering background media heating.

System 3's 26K/nm is a much higher downtrack gradient than is necessary for HAMR, which is a good problem to have. We can be less stringent on design and illumination of System 3 lower the NFT tip temperature even further while still achieving 15K/nm downtrack gradient. Figure 80 shows the simplest change to System 3, which is to illuminate at 2.0mW rather than 2.5mW. This causes the NFT tip temperature to further reduce by 20%, the track width reduces from 50nm to 38nm, and the downtrack gradient is still above 15K/nm. In case it is not clear, smaller track width means higher density recording and is favorable. Although not studied here, we could also increase the MgO underlayer thickness from 1nm to 5nm, widen the NFT tip width from 50nm to 60nm, thicken the NFT tip from 20nm to 30nm. We expect that this scaled-down version of System 3 will still achieve 15K/nm gradient and a 50nm track width with much reduced NFT tip temperature.

7.4 Conclusions

In this chapter, we successfully demonstrated that high linear density recording could be achieved with low injected laser power and low operating temperature NFTs. The only HAMR optical nano-focusing system that we recommend is one with a fat NFT illuminated by a single-mode waveguide like that in System 3. Specifically the System 3 specifications shown here are more stringent than required, and the design can be scaled back.

The general requirements for a HAMR system that achieves high density recording with high NFT reliability is described by the following: The NFT must have a large of solid angle of heat conduction from the peg. As depicted in the fat NFT proposed in this dissertation, perhaps the entire volume between the NFT tip and magnetic write-pole could be filled with gold such that the NFT body is as conical as possible. It was not designed and modeled here explicitly, but we further advise that the NFT tip itself be conical rather rectangular. A conical NFT tip may be more optically efficient and will improve the media/NFT temperature ratio. We recommend that the media stack contain a heatsink that has high electrical and thermal conductivity (perhaps gold), at a very close distance to the FePt granular storage layer (perhaps MgO underlayer thickness of 5nm). Less conductive heatsinks farther away form the FePt will likely not achieve high thermal gradients simultaneously with a low NFT operating temperature. The excitation optical mode must be a single-mode waveguide to most efficiently couple light to the inherently single-mode NFT. The single-mode nature of the entire nano-focusing system requires that that packaging of the laser be performed in closed-loop or with

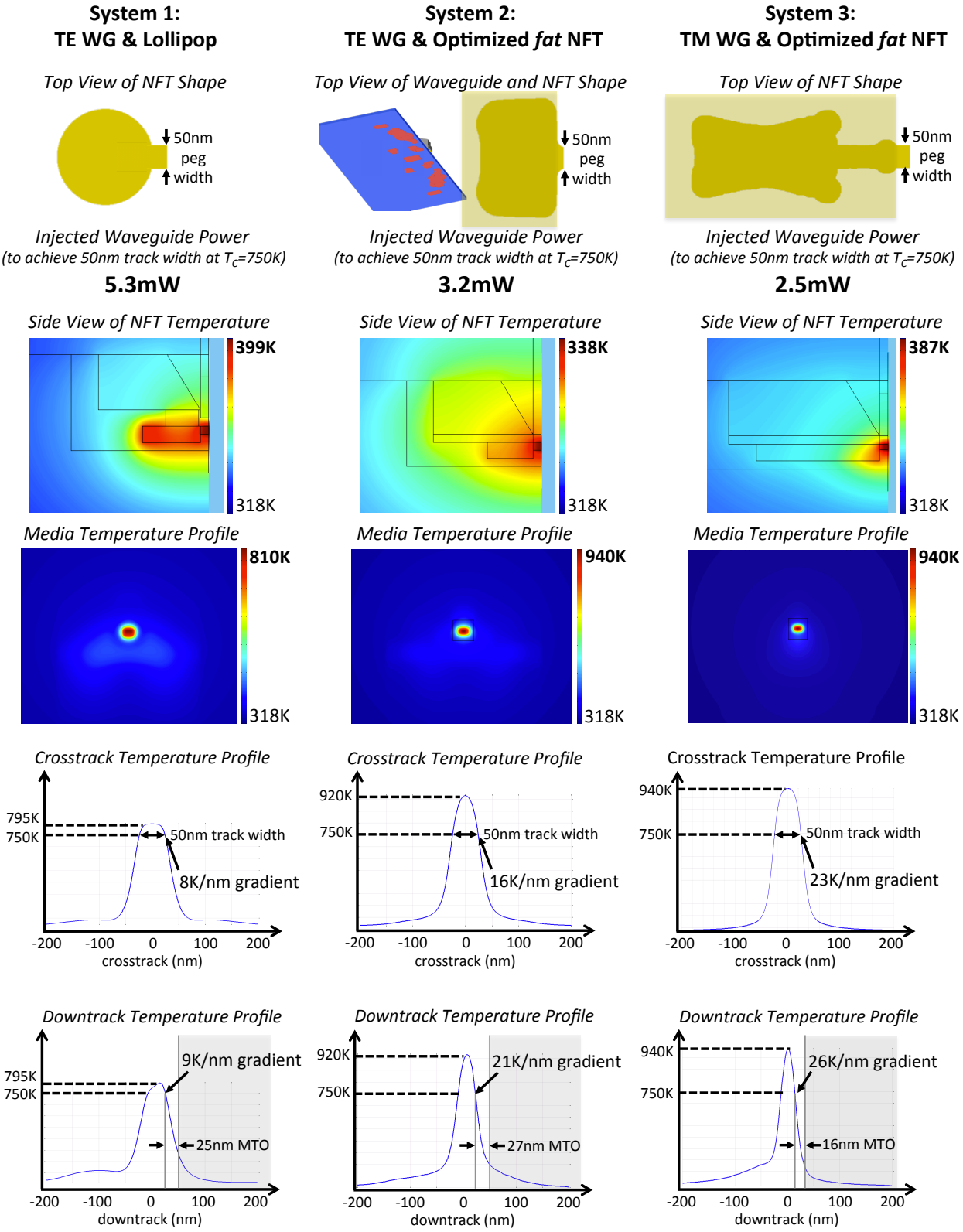

Figure 79: Thermal simulation results showing the enormous thermal gradients delivered by Systems 2 and 3, which also require less injected waveguide power and offer lower NFT tip operating temperature.,

active alignment. Active alignment requires that there is a second waveguide on the head that taps a small percentage of light through evanescent coupling from the main excitation waveguide. The second waveguide may terminate at the edge of the chip such that the light output may be measured and be maximized to actively align the laser to the main waveguide.

There are numerous manufacturing challenges that must be overcome to achieve HAMR: good quality FePt deposition on a thin underlayer on a high conductivity metal, single-mode alignment of the laser, precise nano-patterning of the NFT tip, high precision nano-grinding of the ABS, and deposition of high-temperature-robust overcoats on the head and media. To meet the global storage demand, the hard disk industry must overcome these challenges and produce >6 million heads and >3 million disks per day.

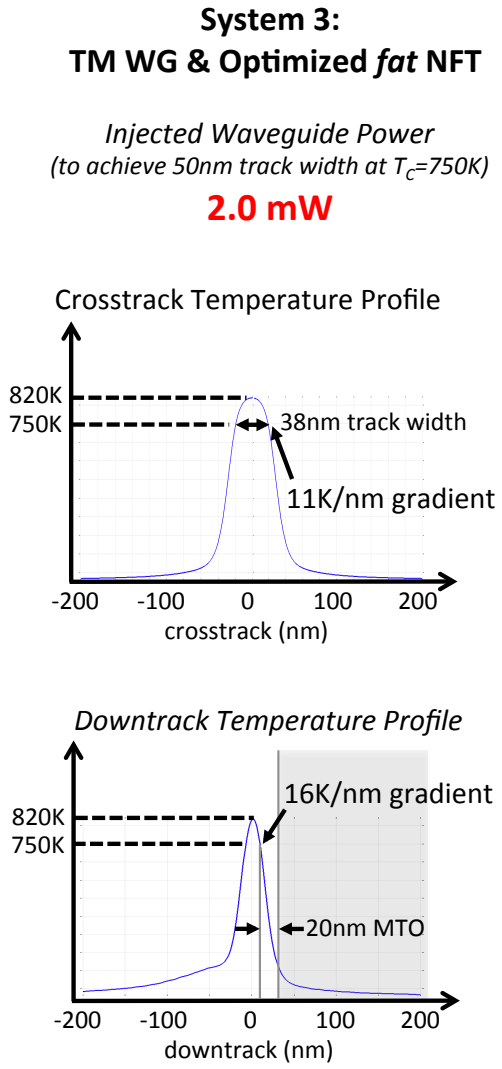

Figure 80: Thermal profiles of System 3 when illuminated with less injected laser power. The track width reduces to ~40nm, the NFT tip temperature is reduced by 20% as compared to the illumination used in Figure 79, and downtrack thermal gradient is still above the desired 15K/nm.

8 Bibliography

[1] "Global shipments of hard disk drives (HDD) from 4th quarter 2010 to 4th quarter 2014 (in millions)." statista.com, Feb-2015.

[2] W. A. Challener, C. Peng, A. V. Itagi, D. Karns, W. Peng, Y. Peng, X. Yang, X. Zhu, N. J. Gokemeijer, Y.-T. Hsia, G. Ju, R. E. Rottmayer, M. A. Seigler, and E. C. Gage, "Heat-assisted magnetic recording by a near-field transducer with efficient optical energy transfer," *Nat. Photonics*, vol. 3, no. 4, pp. 220–224, Apr. 2009.

[3] B. C. Stipe, T. C. Strand, C. C. Poon, H. Balamane, T. D. Boone, J. A. Katine, J.-L. Li, V. Rawat, H. Nemoto, A. Hirotsune, O. Hellwig, R. Ruiz, E. Dobisz, D. S. Kercher, N. Robertson, T. R. Albrecht, and B. D. Terris, "Magnetic recording at 1.5 Pb m^2 using an integrated plasmonic antenna," *Nat. Photonics*, vol. 4, no. 7, pp. 484–488, Jul. 2010.

[4] B. I. Yakobson, A. LaRosa, H. D. Hallen, and M. A. Paesler, "Thermal/optical effects in NSOM probes," *Ultramicroscopy*, vol. 61, no. 1–4, pp. 179–185, Dec. 1995.

[5] A. H. L. Rosa, B. I. Yakobson, and H. D. Hallen, "Origins and effects of thermal processes on near-field optical probes," *Appl. Phys. Lett.*, vol. 67, no. 18, pp. 2597–2599, Oct. 1995.

[6] Y. Wang, W. Srituravanich, C. Sun, and X. Zhang, "Plasmonic Nearfield Scanning Probe with High Transmission," *Nano Lett.*, vol. 8, no. 9, pp. 3041–3045, Sep. 2008.

[7] T. W. McDaniel, "Ultimate limits to thermally assisted magnetic recording," *J. Phys. Condens. Matter*, vol. 17, no. 7, p. R315, Feb. 2005.

[8] M. H. Kryder, E. C. Gage, T. W. McDaniel, W. A. Challener, R. E. Rottmayer, G. Ju, Y.-T. Hsia, and M. F. Erden, "Heat Assisted Magnetic Recording," *Proc. IEEE*, vol. 96, no. 11, pp. 1810–1835, Nov. 2008.

[9] M. Staffaroni, "Circuit Analysis in Metal-Optics, Theory and Applications," Technical Report, University of California, Berkeley, 2011.

[10] S. Bhargava, "Inverse Electromagnetic Design of Optical Antennas for Heat-Assisted Magnetic Recording," University of California, Berkeley, 2012.

[11] J.-P. Mulet, K. Joulain, R. Carminati, and J.-J. Greffet, "Enhanced Radiative Heat Transfer at Nanometric Distances," *Microscale Thermophys. Eng.*, vol. 6, no. 3, pp. 209–222, 2002.

[12] B. V. Budaev and D. B. Bogy, "Heat transport by phonon tunneling across layered structures used in heat assisted magnetic recording," *J. Appl. Phys.*, vol. 117, no. 10, p. 104512, Mar. 2015.

[13] A. Boltasseva, G. V. Naik, J. L. Schroeder, X. Ni, A. V. Kildishev, and T. D. Sands, "Titanium nitride as a plasmonic material for visible and near-infrared wavelengths," *Opt. Mater. Express*, vol. 2, no. 4, pp. 478–489, Apr. 2012.

[14] A. Boltasseva and H. Atwater, "Low-Loss Plasmonic Metamaterials," *SCIENCE*, Jan. 2011.

[15] G. V. Naik, J. Kim, and A. Boltasseva, "Oxides and nitrides as alternative plasmonic materials in the optical range [Invited]," *Opt. Mater. Express*, vol. 1, no. 6, p. 1090, Oct. 2011.

[16] A. Boltasseva, "Empowering plasmonics and metamaterials technology with new material platforms," *MRS Bull.*, vol. 39, no. 05, pp. 461–468, 2014.

[17] U. Guler, A. Boltasseva, and V. M. Shalaev, "Refractory Plasmonics," *Science*, vol. 344, no. 6181, pp. 263–264, Apr. 2014.

[18] U. Guler, A. Kildishev, A. Boltasseva, and V. Shalaev, "FD 178: Plasmonics on the slope of enlightenment: the role of transition metal nitrides," *Faraday Discuss.*, Nov. 2014.

[19] P. r. West, S. Ishii, G. v. Naik, N. k. Emani, V. m. Shalaev, and A. Boltasseva, "Searching for better plasmonic materials," *Laser Photonics Rev.*, vol. 4, no. 6, pp. 795–808, Nov. 2010.

[20] B. X. Xu, Z. H. Cen, J. F. Hu, and J. W. H. Tsai, "Alternative material study for heat assisted magnetic recording transducer application," *J. Appl. Phys.*, vol. 117, no. 17, p. 17C112, May 2015.

[21] P. B. Johnson and R. W. Christy, "Optical Constants of the Noble Metals," *Phys. Rev. B*, vol. 6, no. 12, pp. 4370–4379, Dec. 1972.

[22] A. D. Rakic, A. B. Djuri?ic, J. M. Elazar, and M. L. Majewski, "Optical Properties of Metallic Films for Vertical-Cavity Optoelectronic Devices," *Appl. Opt.*, vol. 37, no. 22, pp. 5271–5283, Aug. 1998.

[23] H. O. Pierson, *Handbook of refractory carbides and nitrides: properties, characteristics, processing, and applications.* Park Ridge, N.J: Noyes Publications, 1996.

[24] R. E. Taylor and J. Morreale, "Thermal Conductivity of Titanium Carbide, Zirconium Carbide, and Titanium Nitride at High Temperatures," *J. Am. Ceram. Soc.*, vol. 47, no. 2, pp. 69–73, Feb. 1964.

[25] J. Adachi, K. Kurosaki, M. Uno, and S. Yamanaka, "Thermal and electrical properties of zirconium nitride," *J. Alloys Compd.*, vol. 399, no. 1–2, pp. 242–244, Aug. 2005.

[26] L. Hultman, "Thermal stability of nitride thin films," *Vacuum*, vol. 57, no. 1, pp. 1–30, Apr. 2000.

[27] G. Antczak and G. Ehrlich, *Surface Diffusion: Metals, Metal Atoms, and Clusters.* Cambridge: Cambridge University Press, 2010.

[28] E. Ricci and R. Novakovic, "Wetting and surface tension measurements on gold alloys," *Gold Bull.*, vol. 34, no. 2, pp. 41–49, Jun. 2001.

[29] R. R. A. Syms, E. M. Yeatman, V. M. Bright, and G. M. Whitesides, "Surface tension-powered self-assembly of microstructures - the state-of-the-art," *J. Microelectromechanical Syst.*, vol. 12, no. 4, pp. 387–417, Aug. 2003.

[30] S. Bhargava and E. Yablonovitch, "Lowering HAMR Near Field Transducer Temperature via Inverse Electromagnetic Design," *IEEE Trans. Magn.*, 2015.

[31] V. Ganapati, O. D. Miller, and E. Yablonovitch, "Light Trapping Textures Designed by Electromagnetic Optimization for Subwavelength Thick Solar Cells," *IEEE J. Photovolt.*, vol. 4, no. 1, pp. 175–182, Jan. 2014.

[32] C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, "Adjoint shape optimization applied to electromagnetic design," *Opt. Express*, vol. 21, no. 18, pp. 21693–21701, Sep. 2013.

[33] G. Scranton, S. Bhargava, V. Ganapati, and E. Yablonovitch, "Single spherical mirror optic for extreme ultraviolet lithography enabled by inverse lithography technology," *Opt. Express*, vol. 22, no. 21, p. 25027, Oct. 2014.

[34] P. R. McGillivray, D. W. Oldenburg, R. G. Ellis, and T. M. Habashy, "Calculation of sensitivities for the frequency-domain electromagnetic problem," *Geophys. J. Int.*, vol. 116, no. 1, pp. 1–4, Jan. 1994.

[35] A. Hördt, "Calculation of electromagnetic sensitivities in the time domain," *Geophys. J. Int.*, vol. 133, no. 3, pp. 713–720, Jun. 1998.

[36] Z. Martinec and J. Vel��msk��, "The adjoint sensitivity method of global electromagnetic induction for CHAMP magnetic data," *Geophys. J. Int.*, vol. 179, no. 3, pp. 1372–1396, Dec. 2009.

[37] M. S. Dadash, N. K. Nikolova, and J. W. Bandler, "Analytical Adjoint Sensitivity Formula for the Scattering Parameters of Metallic Structures," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 9, pp. 2713–2722, Sep. 2012.

[38] J. Chen, D. W. Oldenburg, and E. Haber, "Reciprocity in electromagnetics: Application to marine magnetometric resistivity," no. Draft 7, Jul. 2003.

[39] M. B. Giles and N. A. Pierce, "An Introduction to the Adjoint Approach to Design," *Flow Turbul. Combust.*, vol. 65, pp. 393–415, 2005.

[40] J. Poynting, "On the Transfer of Energy in the Electromagnetic Field," *Philos. Trans. R. Soc. Lond.*, vol. 175, pp. 343–361, 1884.

[41] H. Lorentz, "The theorem of Poynting concerning the energy in the electromagnetic field and two general propositions converning the propagation of light," *HA Lorentz Collect. Pap.*, vol. 3, pp. 1–11, 1896.

[42] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. Wiley, 2004.

[43] S. G. Johnson, M. Ibanescu, M. A. Skorobogatiy, O. Weisberg, J. D. Joannopoulos, and Y. Fink, "Perturbation theory for Maxwell's equations with shifting material boundaries," *Phys. Rev. E*, vol. 65, no. 6, p. 066611, Jun. 2002.

[44] E. Purcell, "Spontaneous emission probabilities at radio frequencies," *Phys Rev 69*, pp. 11–12, 1946.

[45] N. Kumar, "Spontaneous Emission Rate Enhancement Using Optical Antennas," University of California, Berkeley, 2013.

[46] M. S. Eggleston, K. Messer, L. Zhang, E. Yablonovitch, and M. C. Wu, "Optical antenna enhanced spontaneous emission," *Proc. Natl. Acad. Sci.*, vol. 112, no. 6, pp. 1704–1709, Feb. 2015.

[47] K. Messer, M. Eggleston, M. Wu, and E. Yablonovitch, "Enhanced Spontaneous Emission Rate of InP using an Optical Antenna," 2014, p. STu1M.3.

[48] M. Eggleston, K. Messer, E. Yablonovitch, and M. Wu, "Circuit Theory of Optical Antenna Shedding Light on Fundamental Limit of Rate Enhancement," 2014, p. FM2K.4.

[49] M. Staffaroni, J. Conway, S. Vedantam, J. Tang, and E. Yablonovitch, "Circuit analysis in metal-optics," *Photonics Nanostructures - Fundam. Appl.*, vol. 10, no. 1, pp. 166–176, Jan. 2012.

[50] R. E. Rottmayer, S. Batra, D. Buechel, W. A. Challener, J. Hohlfeld, Y. Kubota, L. Li, B. Lu, C. Mihalcea, K. Mountfield, K. Pelhos, C. Peng, T. Rausch, M. A. Seigler, D. Weller, and X. Yang, "Heat-Assisted Magnetic Recording," *IEEE Trans. Magn.*, vol. 42, no. 10, pp. 2417–2421, Oct. 2006.

[51] M. A. Seigler, W. . Challener, E. Gage, N. Gokemeijer, G. Ju, B. Lu, K. Pelhos, C. Peng, R. E. Rottmayer, X. Yang, H. Zhou, and T. Rausch, "Integrated Heat Assisted Magnetic Recording Head: Design and Recording Demonstration," *IEEE Trans. Magn.*, vol. 44, no. 1, pp. 119–124, Jan. 2008.

[52] S. Bhargava and E. Yablonovitch, "Inverse Design of Optical Antennas for Sub-Wavelength Energy Delivery," in *CLEO: 2014*, 2014, p. FW3K.6.

[53] O. Miller, "Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design," University of California, Berkeley, 2012.

[54] E. Gabriel, G. E. Fagg, G. Bosilca, T. Angskun, J. J. Dongarra, J. M. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. H. Castain, D. J. Daniel, R. L. Graham, and T. S. Woodall, "Open MPI: Goals, concept, and design of a next generation MPI implementation," in *In Proceedings, 11th European PVM/MPI Users' Group Meeting*, 2004, pp. 97–104.