Circuit Analysis in Metal-Optics, Theory and Applications

Matteo Staffaroni

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2011-43
http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-43.html

May 9, 2011

Copyright © 2011, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Circuit Analysis in Metal-Optics, Theory and Applications

By

Matteo Staffaroni

A dissertation submitted in partial satisfaction of the

requirements 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 Xiang Zhang
Professor Ming C. Wu

Spring 2011

The Dissertation of Matteo Staffaroni, titled Circuit Analysis in Metal-Optics, Theory and Applications, is approved:

Chair ____ Date _______________

____ Date _______________

____ Date _______________

University of California, Berkeley

Abstract

Circuit Analysis in Metal-Optics, Theory and Applications

by

Matteo Staffaroni

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Eli Yablonovitch, Chair

In the first part of the dissertation we provide electrical circuit descriptions for bulk plasmons, single-surface plasmons, and parallel-plate plasmons. Simple circuits can reproduce the exact frequency versus wave-vector dispersion relations for all these cases, with reasonable accuracy. The circuit paradigm directly provides a characteristic wave impedance that is rarely discussed in the context of plasmonics. Owing to the presence of kinetic inductance, a plasmonic transmission line can support very large characteristic impedances on the order of kilo-Ohms. The ability to adjust the plasmonic wave impedance allows voltage transformer action at optical frequencies, through tapered metallic structures. This transformer action can be used to engineer efficient delivery of optical power to the nanoscale, or as an impedance matching tool toward molecular light emitters.

In the second part of the dissertation we discuss at length the application of plasmonic impedance matching to the problem of heat assisted magnetic recording (HAMR) where an optical antenna is used to concentrate optical power to nanoscale dimensions on the surface of a magnetic hard-disk drive.

Table of Contents

Introduction

Ambition is a state of permanent dissatisfaction with the present.

– Emanuel Derman

Circuits with distributed inductive and capacitive elements can capture much of the physics in Maxwell's Equations. A circuit model provides powerful insights, and can reveal physics that might otherwise be concealed within an exact analytical solution, or in a brute-force numerical solution. While lumped element circuit approaches are common in the microwave and RF regime, they have played a limited role in optics. At optical frequencies, in addition to capacitance, and Faraday inductance, there is also kinetic inductance arising from the inertia of the electrons in a metal. Kinetic inductance dominates over Faraday inductance at blue frequencies, or when there are characteristic dimensions smaller than the collisionless skin depth. The dominance of kinetic inductance defines the plasmonic regime. In general, all three circuit components, must be included, capacitance, Faraday inductance, and kinetic inductance. In this dissertation we detail the role of kinetic inductance in metal optics and utilize it in circuit models for various configurations not yet considered from this respect in the literature. Along the way we highlight several insights unique to the circuit approach to metal optics as outlined below.

In Chapter 1 we model guided wave propagation in metal optics as a one-dimensional distributed element circuit. Simple circuit models recover key results pertaining to dispersion curves for *(i)* guided surface plasmon waves on a single metal surface, and *(ii)* for parallel plate plasmonic waveguides. The circuit approach provides insights into metal optics that are lost in more rigorous formal or numerical treatments. Namely, wave impedance emerges, and is recognized of equal significance to plasmon dispersion. In Chapter 2 we look at some examples of systems that are well approximated by our circuit model and illustrate how the dominance of kinetic inductance over other circuit elements is synonymous with operation in the plasmonic regime. Owing to the dominance of kinetic inductance, a plasmonic transmission line can have impedance greater than the impedance of free space. The ability to adjust the plasmon wave impedance allows voltage transformer action at optical frequencies, through tapered metallic structures. This transformer action is the subject of Chapter 3 where we discuss how optical voltage transformers can be used to engineer efficient delivery of optical power to the nanoscale, or as impedance matching tools toward molecular light emitters. In Chapter 4 we apply the circuit model to antenna radiators and predict how nanoscale optical antennas can lead to spontaneous emission rate enhancement of five orders of magnitude compared to free space, resulting in the phenomenon of spontaneous hyper-emission. In Chapter 5 we apply the circuit model to the problem of optical power delivery in heat assisted magnetic recording (HAMR) where an optical antenna is required to efficiently concentrate optical power to a nanometer-size spot on the surface of a hard-disk drive as part of a thermally assisted magnetic recording scheme. The HAMR concept is briefly introduced and the circuit approach to the light delivery problem is presented as an intuitive means of qualitatively narrowing the optical antenna design space. In Chapter 6 we conclude by summarizing the results of exhaustive modeling work addressing the quantitative aspects of optical antenna design for HAMR which were carried out as part of extensive three year collaborations with Western Digital Corporation in Fremont CA and the Information Storage Industry Consortium (INSIC).

1. Circuit Analysis in Metal Optics

Lumped optical circuit models are available for metallic nano-spheres[1-2], split-ring resonators[3-4], and nanorods[5], but the simplest case of the electrical circuit for a surface plasmon wave propagating along a flat metallic surface has not been presented. In addition to capacitance, and Faraday inductance, there is also kinetic inductance arising from the inertia of the electrons in a metal. Kinetic inductance dominates over Faraday inductance at blue frequencies, or when there are characteristic dimensions smaller than the collisionless skin depth[6]. The dominance of kinetic inductance defines the plasmonic regime. In general, all three circuit components, must be included, capacitance, Faraday inductance, and kinetic inductance. In this chapter we model guided wave propagation in metal optics as a 1-dimensional distributed element circuit. We use the waveguide geometry to derive expressions for the circuit elements, L' and C', comprising its equivalent transmission line circuit. The dispersion relation and characteristic transmission line impedance are then obtained through the standard transmission line equations $\omega^2 = k^2/L'C'$, and $Z = (L'/C')^{1/2}$, respectively. The simple circuit model presented here is very general and recovers dispersion curves for both: *(i)* guided surface plasmon waves on a single metal surface, and *(ii)* for parallel plate optical waveguides. It is readily apparent from the circuit picture that *(i)* is a special case of *(ii)* in the limit of very large plate spacing. The circuit approach also provides new insights into metal optics that would be lost in more rigorous formal or numerical treatments and these insights are the subject of subsequent chapters.

The format of this chapter is as follows: In section 1.1 we introduce the concept of kinetic inductance by illustrating how it naturally arises from modeling the bulk plasma resonance in a metal as a simple LC circuit. In section 1.2 we illustrate the link between metal dielectric constant and conductivity, which we use to arrive at a general expression for the kinetic inductance associated with an arbitrary current distribution. In sections 1.3-4 we use the expression for kinetic inductance from section 1.2 along with conventional RF concepts of capacitance and Faraday inductance to obtain equivalent transmission line circuits for surface plasmon waves on a single metal surface and in parallel plate waveguides. In the process we make use of expressions derived in appendices D-E for the capacitance and Faraday inductance associated with a surface plasmon wave on a flat surface. The transmission line circuits are then used to obtain dispersion relations in good agreement with exact solutions of the Maxwell equations. Wave impedance also emerges from the circuit model, and in section 1.5 we discuss how this impedance is found to diverge at the nanoscale due to the geometric dependence of kinetic inductance of the reciprocal of a waveguide's width. We conclude with some remarks regarding possible uses for the diverging characteristic impedance of plasmonic waveguides.

1.1 - Bulk Plasma Resonance Condition

The bulk plasma resonance condition may be derived from Newton's second law, $F = ma$. Consider a rectangular metal slab of length z, and cross-sectional area A. Displacing the electron cloud about the background ionic lattice by an infinitesimal distance dz along the length dimension creates an electric field V/z. The force on each electron is then $F = q\,V/z = ma = m(dv/dt)$. It is wise at this point to insert the quantity nqA, in both numerator and denominator

$$F = m\frac{dv}{dt} = m\frac{nqA}{nqA}\frac{dv}{dt}\,, \tag{1.1}$$

where n is the number density of electrons in the metal, and q is the electron charge. Recognizing that $nqAv$ is the electric charge that passes a single point in one second, it represents the flowing electric current, $nqAv = I$. Equation (1.1) can then be re-written as $q\,V/z = (m/nqA)(dI/dt)$ which resembles the voltage response of an inductor: $V = L_k(dI/dt)$, where the inductance is $L_k = (m/nq^2)(z/A)$, the kinetic inductance due to inertia of the electrons in the metal. Likewise the charge separation in the metal slab leads to a capacitance: $C = \varepsilon_o(A/z)$. When the charge cloud is released, it oscillates about the ionic lattice like a mass on a spring, or simply as an LC circuit. The corresponding plasma oscillation frequency is given by

$$\omega^2 = \frac{1}{LC} = \frac{nq^2}{\varepsilon_o m}\frac{z}{A}\frac{A}{z} = \frac{nq^2}{\varepsilon_o m} \equiv \omega_p^2, \tag{1.2}$$

where ω_p is recognized as the conventional bulk plasma frequency of the metal. Thus bulk plasmons can be represented by a kinetic inductance $L_k = (m/nq^2)(z/A)$. Kinetic inductance is a well-known concept in superconductivity[7], where the ordinary resistance is zero, and it is only inductance that impedes current flow.

1.2 - Equivalence Between Metal Dielectric Constant and Conductivity

At the root of the circuit description of metal optics is a recognition of the equivalent parameterization between optical dielectric constant of a metal, $\varepsilon_m(\omega)$, and the less frequently discussed complex frequency dependent optical conductivity $\sigma(\omega) \equiv j\omega\varepsilon_o(\varepsilon_m - 1)$. This can be seen directly from Ampere's Law, which can treat the metal as a dielectric

$$\nabla \times H - \partial D/\partial t = \nabla \times H - j\omega\varepsilon_o\varepsilon_m E = 0, \tag{1.3}$$

with relative dielectric constant ε_m. Alternately the metallic response can be regarded as producing only currents J and charges ρ with otherwise negligible dielectric response, in which case Ampere's Law becomes

$$\nabla \times H = J + j\omega\varepsilon_o E = \sigma E + j\omega\varepsilon_o E. \tag{1.4}$$

The equivalence of Eqns. (3,4) is ensured when the complex conductivity is defined as $\sigma(\omega) \equiv j\omega\varepsilon_o(\varepsilon_m - 1)$. The complex resistivity $\rho(\omega) \equiv 1/\sigma(\omega)$ can be rationalized into real and imaginary parts as

$$\rho(\omega) \equiv \frac{1}{\sigma(\omega)} = \frac{1}{\varepsilon_o\omega}\frac{j(1-\varepsilon_m^{'}) + \varepsilon_m^{''}}{(1-\varepsilon_m^{'})^2 + (\varepsilon_m^{''})^2}.$$

Then a metallic wire will have an impedance that can be derived from the complex resistivity:

$$\mathrm{Z} = \frac{1}{\sigma(\omega)}\frac{length}{area} = \frac{1}{j\omega\varepsilon_o(\varepsilon_m - 1)} \times \frac{length}{area}. \qquad (1.5)$$

Substituting in the metal optical relative dielectric constant, $\varepsilon_m = 1 - nq^2/\varepsilon_o m\omega^2$, that neglects collisions, the impedance becomes $Z= j\omega \times (m/nq^2) \times (length/area) \equiv j\omega L_k$. Thus the kinetic inductance arises from the inertia due to mass, m, of $L_k = (m/nq^2) \times (length/area)$, as in the previous section on bulk plasmons. Equivalently expressed in terms of relative dielectric constant $L_k = (1/\omega^2\varepsilon_o(1-\varepsilon_m)) \times (length/area)$. Including collisions, there is an additional resistance term $Z= R + j\omega L_k$, where for collision time τ, $R = (m/nq^2\tau) \times (length/area)$, which is the usual expression for the dissipative resistance of an electron gas.

1.3 - Circuit Theory for Surface Plasmons

A hallmark of metal-optics is that the interface between a metal and free space can support surface plasmon modes[8-11] with the exact dispersion relation (derived in Appendix A):

$$k = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m}{\varepsilon_m + 1}} \qquad (1.6)$$

where $k = 2\pi/\lambda_k$ is the wave vector of the mode, and λ_k is the corresponding wavelength along the surface. Eqn. (1.6) is plotted as the solid blue curve in Fig.1. But the physics of surface plasmons can be captured in a distributed circuit transmission line model that is quite conventional, except that it includes L_k in series with the more conventional Faraday inductance, L_F.

In developing a transmission line circuit model, there are several challenges to overcome: Is it indeed possible to have a transmission line when there is only a single conductor? Transmission line theory usually applies when there are two conductors, one to transmit current, and one to return current. We want to treat the single metal surface of Fig.2 as a conductor, but there is no return conductor to complete the circuit! Effectively the return currents must flow at infinity, to permit a single metal plate to act as a transmission line. The charge distribution $\sigma(x)$ in Fig.2(a) creates an oscillatory voltage $V(x)$ which can be calculated from electrostatics. With voltage and charge calculated, a capacitance per unit length C' in the x-propagation direction can be determined. A detailed electrostatic calculation, in Appendix D, shows that $C' = 2\varepsilon_o kW$, where W is the width of the metal surface in the y-direction, transverse to k, the wave propagation direction.

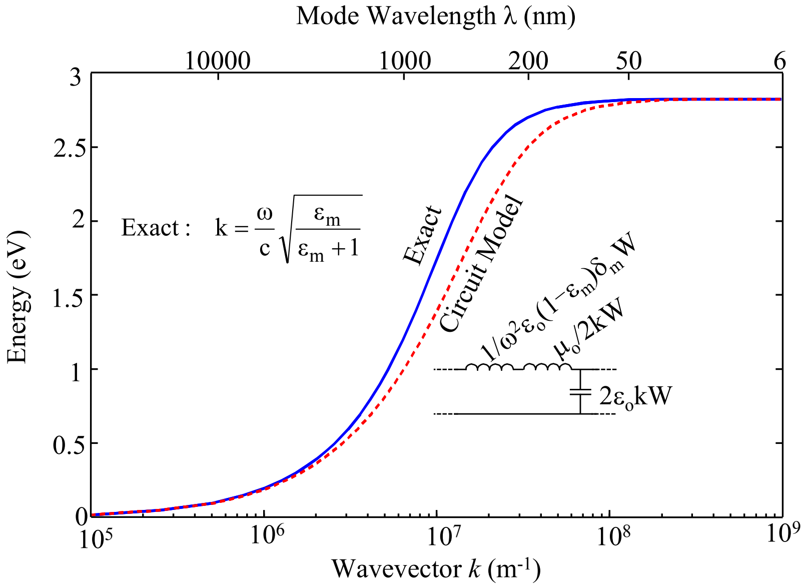


Figure 1: The dispersion relation of surface plasmons for a Drude metal with $\hbar\,\omega_p = 4eV$. The solid blue curve is the exact dispersion, while the red, dashed line is the transmission line circuit model consisting of capacitance C', kinetic inductance L'_k, and ordinary Faraday inductance L'_F, where the prime $'$ represents–per unit length of transmission line.

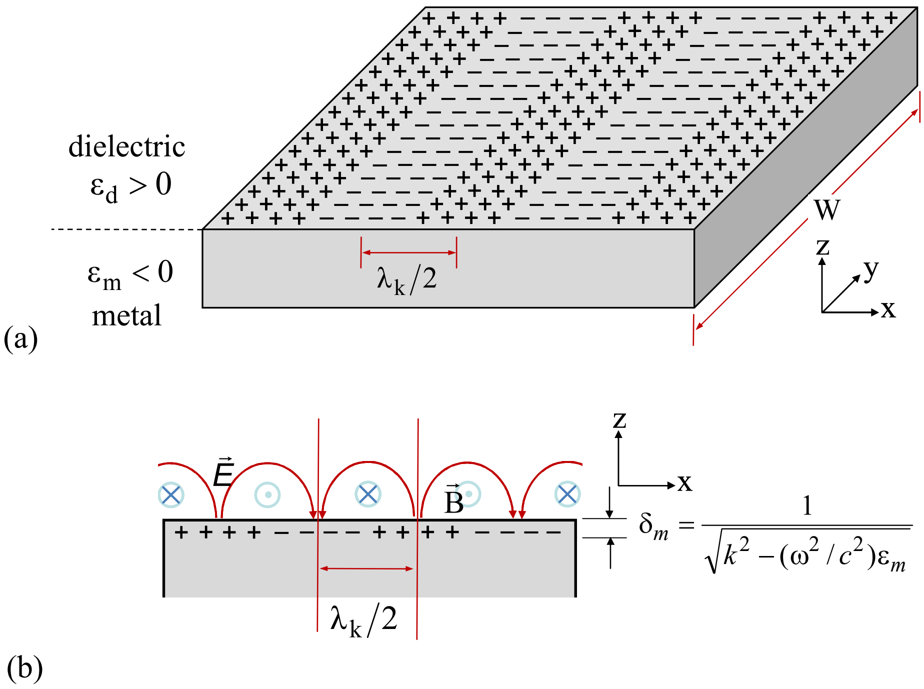


Figure 2: (a) The distribution of charges associated with a surface electromagnetic wave on a metal. (b) The electric and magnetic fields, and the associated skin depth in which metallic current flows.

As the surface charge shown in Fig. 2(a) oscillates in time, sinusoidal surface currents must flow in space. The surface currents produce a magnetic field $B(x,z)$ above the surface, spatially sinusoidal in the x-direction. The effective Faraday inductance L_F can be calculated[13] from $\int Bdxdz = L_F I$, where I is the surface current over the full metal width W, and the magnetic flux is obtained by integrating $\int Bdxdz$ above the metal surface in the $+z$-direction, and in the x-direction. Expressed as inductance/per unit length L' in the x-direction, the magnetic flux is $\int Bdxdz = \int L'_F I dx$. Equating the integrands yields $L'_F I = \int Bdz$. Thus a calculation of $\int Bdz/I$, allows us to determine L'_F, the Faraday inductance per unit length. In Appendix E, it is shown that $L'_F = \mu_o/2kW$.

There remains to calculate the contribution from the kinetic inductance, $L_k = (1/\omega^2\varepsilon_o(1-\varepsilon_m)) \times (length/area)$ which is generally in series with L_F. Per unit length $L'_k = (1/\omega^2\varepsilon_o(1-\varepsilon_m)) \times (1/area)$. It remains to calculate the area of the conduction path, which is the skin depth of the metal slab × the width, $A = \delta_m W$. There is no one skin depth that is appropriate to all situations. Table I presents three different forms of skin depth:

Skin Depth Type	(a) Collisional $\omega\tau < 1$	(b) Collisionless $\omega\tau > 1$	(c) Surface Wave Skin Depth
Skin Depth δ_m	$\sqrt{\frac{2\rho'(\omega)}{\omega\mu_o}} = \frac{c\sqrt{2\varepsilon'_m(\omega)}}{\omega\lvert 1-\varepsilon_m\rvert}$	$\frac{c}{\omega\sqrt{1-\varepsilon'_m}} = \frac{c}{\omega_p}$	$\frac{1}{\sqrt{k^2-\varepsilon_m(\omega^2/c^2)}}$

Table I: (a) Collisional skin depth is appropriate to microwaves, and is found in most Electromagnetics books[14]. (b) Collisionless skin depth[6] pertains to normal incidence plane waves, above the collision frequency, ~10THz. (c) If there is a wave propagation k, parallel to the surface, the surface wave skin depth applies[8]. Generally ε_m is predominantly negative, and ε'_m & ε''_m, and ρ' & ρ'', represent the real and imaginary parts respectively.

For the surface wave propagation that we are considering, the appropriate skin depth δ_m is given as case (c) of Table I. Thus $L'_k = 1/\omega^2\varepsilon_o(1-\varepsilon_m)\,\delta_m W$. All three component values, C', L'_F, and L'_k are now known. Since they are defined per unit length, they contribute toward a distributed transmission line. The two inductors L'_F, and L'_k act in series $L'_F + L'_k$, as shown in the inset of Fig.1. The properties of such a series L–parallel C transmission lines are well worked out[15]. There are two important properties of transmission lines: (a) the dispersion, ω versus k, is given by $\omega^2 = k^2/L'C'$, and (b) the wave impedance is given by $Z = \sqrt{L'/C'}$. The general dispersion is given by $\omega^2 = k^2/(L'_F + L'_k)C'$ which can be rewritten

$$\frac{1}{\omega^2} = \frac{1}{k^2}\left(\frac{\mu_o}{2kW} + \frac{1}{\omega^2\varepsilon_o(1-\varepsilon_m)\delta_m W}\right)2\varepsilon_o kW = \frac{1}{k^2}\left(\mu_o\varepsilon_o + \frac{2k}{\omega^2(1-\varepsilon_m)\delta_m}\right). \quad (1.7)$$

Equation (1.7) is plotted as the "circuit model" in Fig.1.

In the limit $k < \omega_p/c$, the kinetic inductance term is negligible compared to the Faraday inductance. Eqn. (1.7) simplifies to $\omega = kc$, the “light line”, in good agreement with the exact solution in Fig.1. In the limit $k > \omega_p/c$, the kinetic inductance term dominates the Faraday inductance. Moreover, the skin depth $\delta_m = 1/\sqrt{k^2 - \varepsilon_m(\omega^2/c^2)}$, becomes $\delta_m \approx 1/k$. Then Eqn. (1.7) becomes:

$$\frac{1}{\omega^2} = \frac{1}{k^2}\left(\frac{2k^2}{\omega^2(1-\varepsilon_m)}\right) = \frac{2}{\omega^2(1-\varepsilon_m)}\,, \tag{1.8}$$

which further reduces to $1 - \varepsilon_m = 2$, or in other words $\varepsilon_m = -1$, which can also be written $\omega_p = 1/\sqrt{2}$, all of which expressions correspond exactly to the surface plasmon condition, in good agreement with the exact dispersion in Fig.1. In the intermediate regime $k \sim \omega_p/c$, the circuit model deviates from the exact dispersion in Fig.1 by about 15%. The circuit model is distributed, consisting repeating circuit blocks in one dimension, as shown in the insets of Fig.1. A more realistic model would be distributed in two dimensions respecting the two-dimensional character of our problem. A two-dimensional distributed circuit would better describe our situation, but such an avenue would add many more circuit components while doing little for intuitive understanding. Thus we retain our one dimensionally distributed model in spite of the slight disagreement with the exact solution.

The wave impedance is of equal importance to the dispersion, and in the circuit model may be written:

$$Z = \sqrt{\frac{L'}{C'}} = \sqrt{\frac{\dfrac{\mu_o}{2kW} + \dfrac{1}{\omega^2\varepsilon_o(1-\varepsilon_m)\delta_m W}}{2\varepsilon_o kW}} = \frac{1}{W}\sqrt{\frac{\dfrac{\mu_o}{2k} + \dfrac{1}{\omega^2\varepsilon_o(1-\varepsilon_m)\delta_m}}{2\varepsilon_o k}}. \tag{1.9}$$

The final expression in Eqn. (1.9) shows explicitly that the wave impedance $Z \propto 1/W$ becomes very large as the conducting plate becomes narrower. Once again we treat the limits $k < \omega_p/c$ and $k > \omega_p/c$. For small wave vectors near the light line, $k < \omega_p/c$, the Faraday inductance dominates:

$$Z = \sqrt{\frac{L'}{C'}} = \frac{1}{W}\sqrt{\frac{\mu_o}{4k^2\varepsilon_o}} = \frac{1}{2kW}\sqrt{\frac{\mu_o}{\varepsilon_o}} = \frac{1}{2kW} \times 377\Omega. \tag{1.10}$$

Since $W > 1/k$ to maintain the one dimensionality of the problem, the impedance in case $k < \omega_p/c$ cannot exceed the impedance of free space, $377\Omega = \sqrt{\mu_o/\varepsilon_o}$. In general, for free space transverse magnetic (TM) transmission lines without kinetic inductance, 377Ω is an upper limit to the achievable impedance[15].

In the opposite limit, $k > \omega_p/c$, the kinetic inductance dominates in Eqn. (1.9), and the wave impedance becomes

$$Z = \sqrt{\frac{L'}{C'}} = \frac{1}{W}\sqrt{\frac{1}{2\omega^2\varepsilon_o^2(1-\varepsilon_m)k\delta_m}} = \frac{1}{W}\sqrt{\frac{\mu_o}{\varepsilon_o}\frac{c^2}{2\omega_p^2 k\delta_m}} = \frac{1}{W}\frac{\lambda_p}{2\pi}\sqrt{\frac{\mu_o}{2\varepsilon_o}}, \tag{1.11}$$

where c/ω_p was replaced by λ_p , the vacuum wavelength at the plasma frequency, and $\delta_m \approx 1/k$ in the deep plasmonic regime where kinetic inductance dominates, resulting in the simple form on the right side of Eqn. (1.11). When width W is less than the skin depth $\lambda_p/2\pi$, this creates the possibility of a wave impedance $Z > 377\Omega$, which should be regarded as a unique feature of the plasmonic regime.

1.4 - Circuit Theory for the Plasmonic Parallel Plate Waveguide

We now transfer our attention to parallel plate waveguides at optical frequencies. There exists an exact solution[8-9,11] of Maxwell's equations for the parallel-plate waveguide, for general complex dielectric constant ε_m:

$$e^{-K_i d} = \left(\frac{K_i\varepsilon_m + K_m\varepsilon_i}{K_i\varepsilon_m - K_m\varepsilon_i}\right) \tag{1.12}$$

with $K_m = \sqrt{k^2 - \varepsilon_m(\omega/c)^2} = 1/\delta_m$, $K_i = \sqrt{k^2 - (\omega/c)^2}$, and where $k = 2\pi/\lambda_k$ is the actual wave vector of the mode, and λ_k is the corresponding mode wavelength, ω is the optical frequency, c is the speed of light in free space, and d is the plate spacing. The skin depth $K_m = 1/\delta_m$ in the metal is the same as for the single plate waveguide described in Table I. In transmission line theory, parallel plate waveguides operating in the microwave regime are modeled as a distributed-element repeating circuit of series inductors and parallel capacitors[13,15]. The voltage and current waveforms supported by this type of reactive transmission line circuit follow the general dispersion relation given by $\omega^2 = k^2/L'C'$, and wave impedance by $Z = \sqrt{L'/C'}$. Once again, we need to define L'_F, C', and L'_k. For the parallel-plate geometry the kinetic inductance L'_k is the same as for a single plate, but multiplied by 2 to account for the series inductance of the first plate and the return current plate, $L'_k = 2/\omega^2\varepsilon_o(1-\varepsilon_m)\,\delta_m W.$ The parallel plate inductance/unit length $L'_{cF} = \mu_o d/W$, and capacitance/unit length $C'_c = \varepsilon_o W/d$ are easy to derive, and well documented[13,15]. The cross plate capacitance C'_c does not tell the whole story. We learned, when analyzing the single plate case that intra-plate capacitance, now labeled as $C'_i = 2\varepsilon_o kW$ is also present. Likewise the intra-plate inductance $L'_{iF} = \mu_o/2kW$ must also be present.

Some corrections must now be introduced: (*i*) Since the intra-plate inductance L'_{iF} appears in series on both the upper plate and the lower plate, the correct value must be multiplied by 2 ×: $L'_{iF} = \mu_o/kW$. Likewise the intra-plate capacitances on the upper and lower plates appear as reactive impedances in series. Thus the true intra-plate capacitance must be cut in half: $C'_i = \varepsilon_o kW$. (*ii*) Further corrections are needed on the cross-plate inductance and capacitance/unit length. When the plate spacing is larger than the modal wavelength, only a fraction of the electric field lines reach from plate to plate, while the rest contribute to intra-plate capacitance. The presence of a spatially oscillating positive and negative charge on one plate implies that a distant plate sees net cancellation, a weak field that falls off exponentially as $exp(-kd)$. Since the corresponding charge is smaller, for the same plate voltage, the cross-plate capacitance is smaller by the same factor, diminished to $C_c' = \varepsilon_o(W/d)exp(-kd)$. This exponentially decaying term is similar to the screening of electric field through a periodic perforated screen. The spatially oscillating surface charge provides the periodicity. Combining the intra-plate capacitance, C_i', and the cross-plate capacitance, C_c':

$$C' = C_i' + C_c' = \frac{\varepsilon_o W}{d}\left(kd + e^{-kd}\right). \tag{1.13}$$

In the limit of small plate spacing $exp(-kd) \rightarrow 1 - kd$, and Eqn. (1.13) reduces to that of a parallel plate capacitor, $C' \rightarrow \varepsilon_o(W/d)$. In the opposite limit of large plate spacing, $exp(-kd)$ $\rightarrow 0$ and $C' \rightarrow \varepsilon_o kW$, half the intra-plate capacitance, owing to the fact that the two widely separated plates are effectively in series. Likewise the cross-plate inductance L_{cF}' is increased when the plates are widely spaced, since little cross-plate displacement current flows in spite of a high voltage on the plate; $L_{cF}' = \mu_o(d/W)exp(kd)$. This must be combined with the intra-plate inductance that we already know, $L_{iF}' = \mu_o/kW$. Since the current flowing in the metal plate flows either cross-plate or intra-plate the two inductance contributions, L_{cF}' and L_{iF}' must be in parallel:

$$L_F' = \left(\frac{1}{L_{iF}'} + \frac{1}{L_{cF}'}\right)^{-1} = \left(\frac{Wkd}{\mu_o d} + \frac{W}{\mu_o d e^{kd}}\right)^{-1} = \frac{\mu_o d}{W\left(kd + e^{-kd}\right)}. \tag{1.14}$$

In the limit of small plate spacing $exp(-kd) \rightarrow 1 - kd$, and Eqn. (1.14) reduces to that of a parallel plate inductor, $L_F' \rightarrow \mu_o(d/W)$. In the opposite limit of large plate spacing, $(-kd) \rightarrow 0$, and $L_F' \rightarrow \mu_o/kW$, twice the intra-plate inductance, owing to the fact that the two distant plates are in series.

The agreement between the exact dispersion, Eqn. (1.12), and the circuit model, inset of Fig.3, is perfect in the limits of high and low wave-vector k, but there is some discrepancy at the knee of the dispersion. This may indicate of a more distributed interaction between the Faraday and kinetic inductance than was captured by our simple circuit model. As the plate spacing is increased, the fields at each metal plate gradually decouple from each other until the limit of a single surface wave guided by a single metal plate governed by Eqn. (1.6) and shown in Fig.2 (see also Appendix C).

The distributed incremental equivalent circuit for a small section of plasmonic parallel plate waveguide is illustrated in Fig.4(a) and in the inset to Fig.3, and includes both cross-plate and intra-plate circuit components. The single plate case is in Fig.4(b) and the inset of Fig.1, and includes only intra-plate circuit components. Figs.4(a)&(b) also sketch the electric field lines and surface charges for a half-wavelength $\lambda_k/2$ segment of the line, at the wave-vector k.

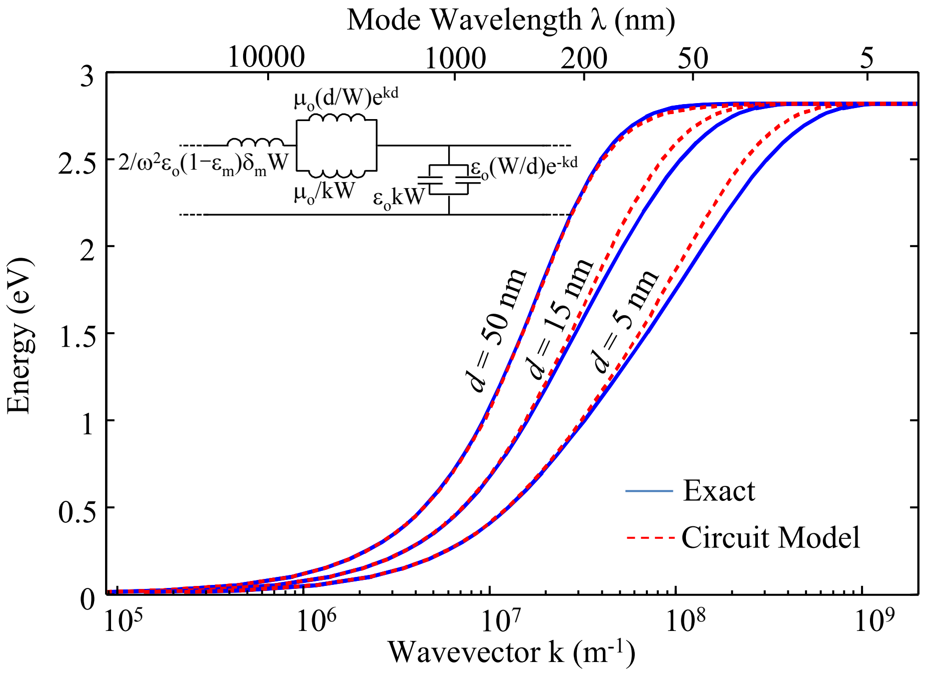


Figure 3: Semi-logarithmic plot of the dispersion relation for a parallel plate waveguide at optical frequencies. The dielectric constant ε_m is that of a free-electron metal with a plasma frequency corresponding to $\hbar\,\omega_p = 4eV$. The parameter d is the plate spacing.

Fig.4(c) operates at low enough frequency so that there are no plasmonic effects, *i.e.* negligible kinetic inductance, yet there remains an interesting competition between cross-plate and intra-plate fields. This case of intermediate wave-vector $1/d < k < \omega_p/c$ of Fig.4(c) is non-plasmonic, yet it doesn't seem to appear in microwave textbooks. For comparison the common case treated in textbooks is given in Fig.4(d).

As in any transmission line, the dispersion of Fig.4(c) is given by $\omega^2 = k^2/L'C'$. Substituting in the capacitance C' of Eqn. (1.13), and the inductance L'_F of Eqn. (1.14), all cross-plate and intra-plate terms cancel, leading to the simplest possible dispersion $\omega = kc$ along the light line. This dispersion in Fig.4(c) is the same as an ordinary transmission line Fig.4(d). The difference between Figs.4(c)&(d) lies in the wave impedance $Z = \sqrt{L'/C'}$. For the ordinary transmission line, Fig.4(d), the wave impedance is controlled by the aspect ratio: $Z = (d/W)\sqrt{\mu_o/\varepsilon_o}$. For the widely spaced transmission plates of Fig.4(c), the role of the spacing d is replaced by the reciprocal wave vector $Z = (1/kW)\sqrt{\mu_o/\varepsilon_o}$. Thus the widely spaced parallel plate waveguide has a wave-impedance even lower than a conventional waveguide, in the non-plasmonic regime, with kinetic inductance absent. In either Fig.4(c) or (d), the wave impedance is always $\ll 377\Omega$, the impedance $\sqrt{\mu_o/\varepsilon_o}$ of free space.

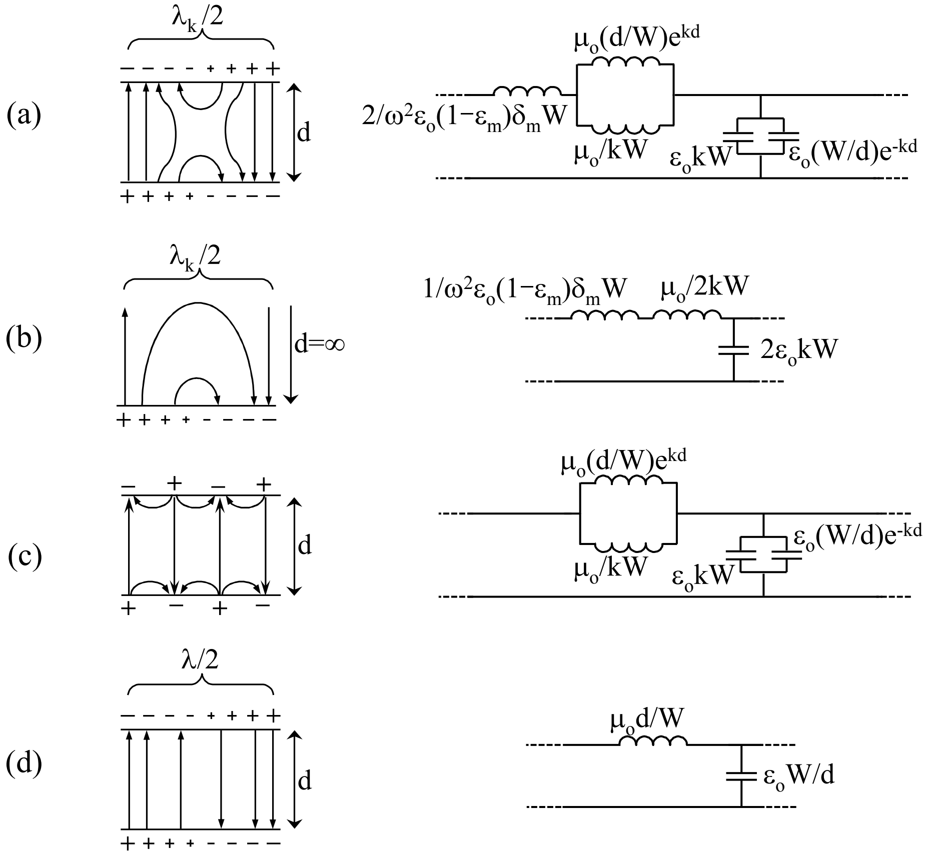


Figure 4: The electric field distribution, surface charges, and distributed equivalent circuits for (a) a half-wavelength $\lambda_k/2$ section of plasmonic parallel plate waveguide, of plate spacing d, and width W. (b) A half-wavelength $\lambda_k/2$of a single-surface plasmonic plate of width W. (c) A two-wavelength section of parallel plate waveguide, at intermediate k along the light line, but having widely spaced plates: $1/d < k < \omega_p/c$. (d) A conventional RF parallel plate waveguide with $k < \omega_p/c$, and $kd < 1$, as is usually covered in Electromagnetics texts[13,15]. Note that when losses are present the fields will additionally be bowed in the direction of propagation[8].

1.5 - Wave Impedance for the Plasmonic Parallel Plate Waveguide

In Equation (1.11) we have already given the surprising wave impedance of a single plasmonic plate, $= (\lambda_p/2\pi W)\sqrt{\mu_o/2\varepsilon_o}$, where $(\lambda_p/2\pi) \sim 25\, nm$ is the collisionless skin depth in the metal. Uniquely for a TM waveguide, the wave impedance of a single narrow plate can become larger than 377Ω, the impedance of free space, owing to the additional contribution by kinetic inductance, L_k. Similar effects occur for the plasmonic parallel plate waveguide. The plasmonic parallel plate wave impedance is $= \sqrt{(L'_k + L'_F)/C'}$, where $L'_k = 2/\omega^2\varepsilon_o(1-\varepsilon_m)\,\delta_m W$, and C' & L'_F are given by Eqns. (13)&(14) respectively. We are particularly interested in the regime where L'_k makes a significant contribution. When $k > \omega_p/c$, and $k > 1/d$, the intra-plate impedances dominate. The expression for Z simplifies to $\left(\lambda_p/2\pi W\right)\sqrt{2\,\mu_o/\varepsilon_o}$. This is twice the single plate impedance, as expected. This provides an opportunity for increasing the wave impedance above 377Ω. When $k > \omega_p/c$, but k is still less than $1/d$, the following formula for wave impedance emerges:

$$Z = \sqrt{\frac{\mu_o}{\varepsilon_o}}\sqrt{\frac{d^2}{W^2} + \frac{2\lambda_p^2 kd}{(2\pi W)^2}}\,, \tag{1.15}$$

which can be high, but not as high as $\left(\lambda_p/2\pi W\right)\sqrt{2\,\mu_o/\varepsilon_o}$. In either instance, $kd > 1$ or $kd < 1$, it is possible to achieve a wave impedance above 377Ω. As we will discuss further in Chapter 3, the ability to taper the dimension W to a narrower waveguide, provides in effect a transformer action at optical frequencies, for both single plate and parallel plate waveguides. The availability of an optical voltage transformer suggests that efficient optical power delivery to the nanoscale is within reach. Before exploring the consequences of such a device, we pause and in Chapter 2 take a closer look at how kinetic inductance can be understood as the defining feature of the plasmonic regime, ultimately being responsible for giving the surface plasmon dispersion relation its characteristic shape.

2. The Role of Kinetic Inductance in Metal Optics

In Chapter 1 we used a simple LC circuit model to recover the surface plasmon dispersion relation. Here we take a closer look at this circuit model and show how it clearly reveals the role of kinetic inductance in shaping the dispersion and other key properties of surface plasmons. In developing a circuit model for surface plasmons on a flat metallic surface, the main challenge to overcome is the fact that only one conductor is present, and that the return currents must flow at infinity instead of on a conventional return conductor. To arrive at a circuit model for the flat metal plate one begins by assuming a sinusoidal charge distribution $\sigma(x)$ as illustrated in Fig. 5. The charge distribution $\sigma(x)$ results in an oscillatory voltage $V(x)$ which can be calculated from electrostatics. With voltage and charge calculated, a capacitance per unit length C' in the x-propagation direction can be determined. A detailed electrostatic calculation in Appendix D shows that $C' = 2\varepsilon_o kW$, where W is the width of the metal surface in the y-direction, transverse to the wave propagation direction, and k is the wave-vector of the surface wave. As the surface charge shown in Fig. 5 oscillates in time, sinusoidal surface currents must flow in space. The surface currents produce a magnetic field $B(x,z)$ above the surface, spatially sinusoidal in the x-direction. The Faraday inductance L_F can be calculated from $\int Bdxdz = L_F I$, where I is the surface current over the full metal width W, and the magnetic flux is obtained by integrating $\int Bdxdz$ above the metal surface in the +z-direction, and in the x-direction. Expressed as inductance per unit length L'_F, the magnetic flux is $\int Bdxdz = \int L'_F Idx$. Equating the integrands yields $L'_F I = \int Bdz$. Thus a calculation of $(\int Bdz)/I$ allows us to determine L'_F, the Faraday inductance per unit length. Carrying out the calculation as in Appendix E one finds $L'_F = \mu_o/2kW$. There remains to calculate the contribution from kinetic inductance, which per unit length evaluates to $L'_k = 1/\left(\omega^2\varepsilon_o(1-\varepsilon_m)\right) \times (1/area)$, where the area of the conduction path is the skin depth in the metal × the width = $\delta_m W$. We thus have $L'_k = 1/\left(\omega^2\varepsilon_o(1-\varepsilon_m)\right) \times (\delta_m W)$.

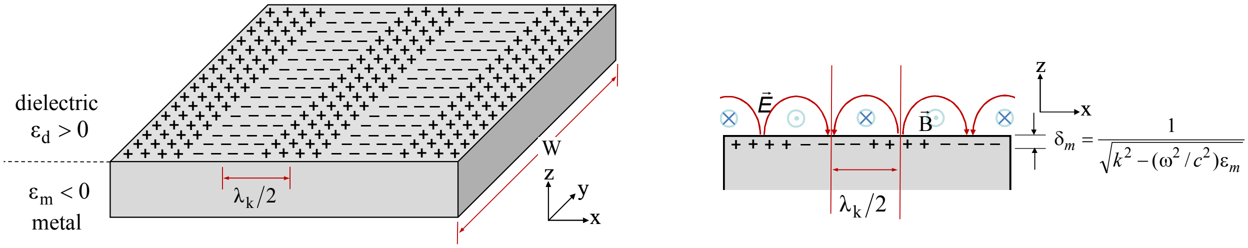


Figure 5: *(left panel)* The distribution of charges associated with a surface electromagnetic wave on a metal plate. *(right panel)* The electric and magnetic fields, and associated ski depth in which the current flows.

Since , , and are defined per unit length, they can be thought of as contributing to a distributed transmission line of series inductors and parallel capacitors. The properties of such transmission lines are well worked out[15], namely the dispersion relation for the line will be given by . The circuit dispersion relation with is plotted as the dashed curve in Fig. 6(a) along with the exact dispersion for a surface plasmon on a flat metal plate[9]. Good agreement is found between the two curves. It is constructive to consider the behavior of

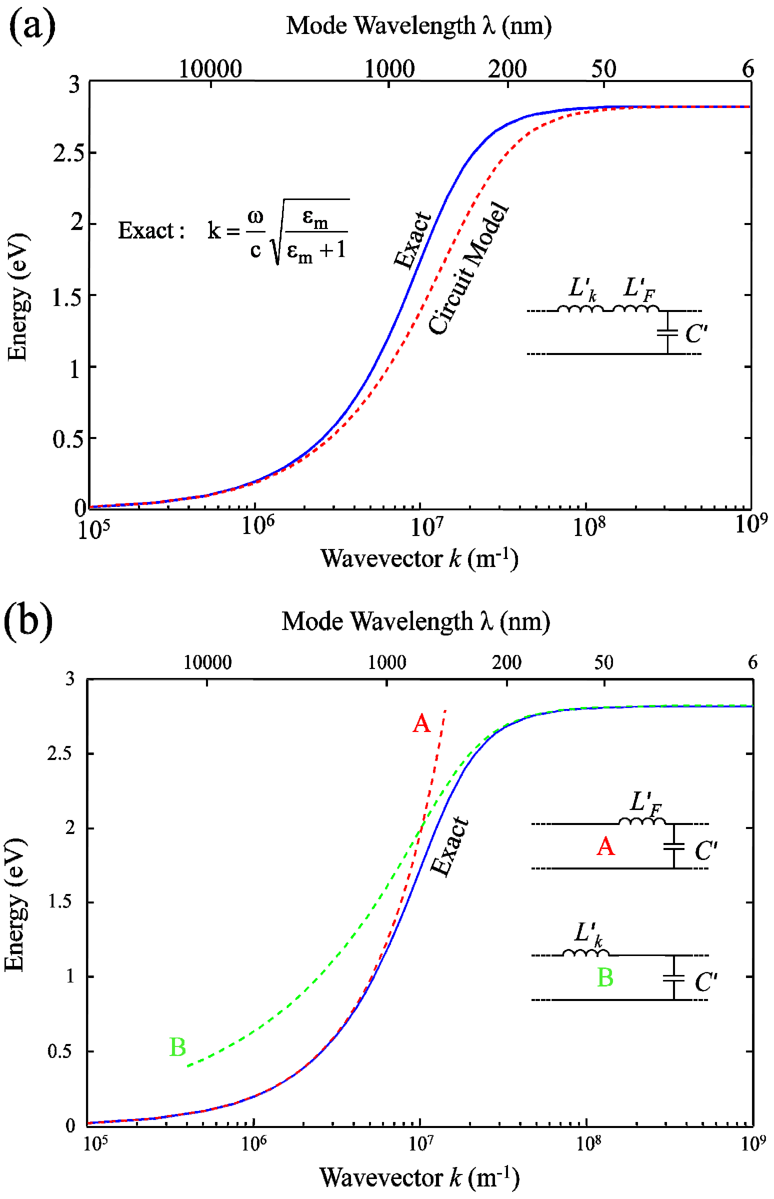


Figure 6: (a) Complete circuit model approximation (dashed line) to the exact dispersion of a surface plasmon (solid line). (b) Low- (A, dashed line) and high- (B, dashed line) circuit element approximations to the exact dispersion of a surface plasmon (solid line).

the dispersion relation from the circuit model in the limits of small and large wave-vector k. The small wave-vector limit is defined by $k < \omega_p/c$, where ω_p is the plasma frequency of the metal comprising the plate, and c is the speed of light in free space. In this regime, comparing the expressions for L'_F and L'_k reveals that the kinetic inductance is negligible compared to the Faraday inductance, so that the former may be dropped from the circuit model. Without kinetic inductance, the dispersion relation from the circuit model reduces to the light line $\omega = kc$, plotted as curve A in Fig. 6(b). On the other hand, the large wave-vector limit is defined by $k > \omega_p/c$. In this regime, comparing the expressions for L'_F and L'_k reveals that the Faraday inductance is negligible compared to the kinetic inductance, so that the former may be dropped from the circuit model. Thus kinetic inductance dominates Faraday inductance in the large wave-vector regime. Without Faraday inductance, the dispersion relation from the circuit model reduces to curve B in Fig. 6(b). We thus see that the asymptotic behavior of the plasmonic dispersion curve is a manifestation of the dominance of kinetic inductance in the large wave-vector limit. Moreover, in the limit of very large k, the plasmon skin depth $\delta_m = 1/(k^2 - \varepsilon_m \, \omega^2/c^2)$ becomes $\delta_m \approx 1/k$ and the circuit model dispersion relation reduces to $1 - \varepsilon_m = 2$, or in other words $\varepsilon_m = -1$, which can also be written $\omega_p = 1/\sqrt{2}$, all of which expressions correspond exactly to the surface plasmon resonance condition. In addition to fixing the surface plasmon resonance condition, the kinetic inductance also plays an important role in shaping the line impedance Z of plasmonic waveguides since $Z = \sqrt{L'/C'}$ and thus in the limit of large wave-vectors this reduces to $Z = \sqrt{L'_k/C'}$, resulting in an impedance that diverges as the reciprocal of the metal plate's width. One can easily show that that the circuit model discussed above for a plasmon wave on a single metal plate is the limiting case of the circuit model representation of a parallel plate plasmonic waveguide where the plate spacing is taken to be infinite. In the parallel plate configuration the kinetic inductance is still found to pin the dispersion relation at large wave-vectors, and once again results in a line impedance that diverges as the reciprocal of the plate width. This is in sharp contrast to conventional microwave and RF transverse magnetic (TM) modes in parallel plate waveguides having line impedances that fundamentally cannot exceed the impedance of free space[16]. The ability of a plasmonic parallel plate waveguide to attain very large impedance TM modes is thus directly related to the dominance of kinetic inductance in the plasmonic regime. In fact, according to the circuit picture of metal-optics, the plasmonic regime may be thought of as the realm where kinetic inductance is the dominant circuit element. The concept of diverging line impedances in plasmonic parallel plate waveguides is expanded upon in Chapter 3 where it is used to devise an optical voltage transformer.

3. The Optical Voltage Transformer

In this Chapter we show that transformer action at optical frequencies arises naturally from a circuit analysis of plasmonic parallel plate waveguides. The plasmonic transmission line impedance in a parallel plate configuration is found to diverge at the nanoscale resulting in a rise in voltage and a decrease in current as the waveguide narrows to a sharp tip. A decreased current, as a result of transformer action, is accompanied by diminished I^2R resistive losses, which are a major problem in metal optics.

It is well known that while electromagnetic fields cannot penetrate far into good conductors, they can be guided long distances by them. In the microwave regime distributed objects can have energy stored in electric fields through capacitance, and magnetic fields through inductance; the interplay between these can give rise to an energy flow as described by transmission line theory[17]. The voltage and current waveforms in a transmission line are related through $V = I \cdot Z$, where Z is the characteristic impedance of the line. An ideal free-space parallel plate transmission line in the microwave regime can be modeled as a distributed element reactive circuit of series inductors and parallel capacitors with characteristic impedance

$$Z = \frac{d}{W}\sqrt{\frac{\mu_o}{\varepsilon_o}}, \tag{3.1}$$

where W is the plate width and d is the plate separation. If d and W are scaled together, as in a tapered transmission line with square cross-section, the impedance of the line remains unchanged. When this geometry is scaled to very small dimension the resistive losses associated with the metallic plates are no longer negligible and actually become prohibitive since resistance scales as $1/W$. In order to take the resistive loss into account, the reactive circuit model has to be modified to include a resistance. This may be done by modeling each infinitesimal unit of length of the line as a voltage divider between characteristic impedance Z of the line and resistance R associated with and incremental unit of distance dx as shown in Fig.7.

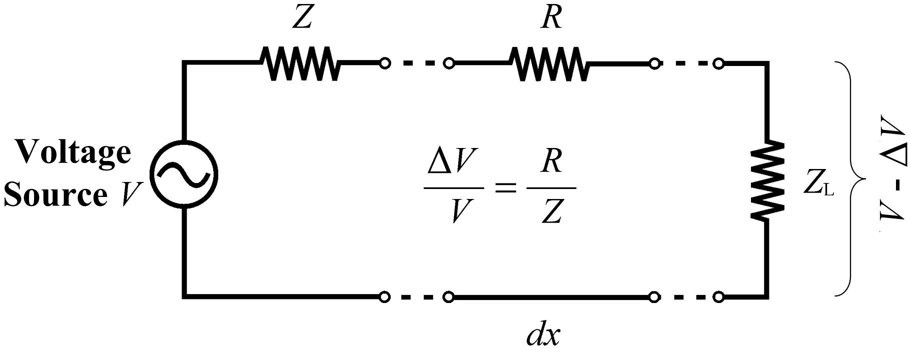


Figure 7: Voltage divider circuit model for an infinitesimal incremental length of transmission line with loss.

The voltage drop across a length of line dx is

$$\Delta V = \frac{V}{Z} R\,, \tag{3.2}$$

where V/Z corresponds to the current carried by the line. Loss can be minimized by ensuring that the transmission line impedance Z is large compared to the resistance R. In the coming sections we show that recasting the standard solution for the propagating surface plasmon in a parallel plate waveguide as a circuit problem reveals the availability of transmission line impedances that diverge as $W \to 0$. A tapered plasmonic parallel plate waveguide then behaves as an optical voltage transformer which can be used to efficiently deliver optical power to the nanoscale.

3.1 - Diverging Transmission Line Impedance

The dispersion relation for a guided surface plasmon wave in a parallel plate metallic waveguide is well known and is conventionally arrived at analytically through use of the Maxwell equations in a boundary value problem[9]. In the previous chapters we have shown that the dispersion relation can also be recovered from a circuit model where in addition to capacitance and Faraday inductance one also includes the kinetic inductance associated with the electrons in the metal. It is found that the transmission line circuit for the plasmonic parallel plate waveguide is equivalent to that of a conventional parallel plate waveguide but the conventional capacitance and inductance per unit length $C' = \varepsilon_o W/d$ and $L' = \mu_o d/W$, are replaced by

$$C' = \frac{\varepsilon_o W}{d}\left(kd + e^{-kd}\right) \tag{3.3}$$

and

$$L' = \frac{\mu_o d}{W\left(kd + e^{-kd}\right)} + \frac{2}{\omega^2 \varepsilon_o (1 - \varepsilon_m)\delta_m W} \tag{3.4}$$

where k is the wave vector of the plasmon mode, ε_m is the dielectric constant of the metal at frequency ω, and $\delta_m = 1/\sqrt{k^2 - (\omega/c)^2}$ is the skin depth of the fields into the metal. Whereas the transmission line impedance $Z = \sqrt{L'/C'}$ of the conventional parallel plate waveguide given in Eqn. (3.1) is constant when d and W are scaled together, the impedance Z_p of the plasmonic line from Eqns. (3.3,3.4) behaves differently:

$$Z_p = \frac{1}{W}\sqrt{\frac{\mu_o d^2}{\varepsilon_o \left(kd + e^{-kd}\right)^2} + \frac{2\left(kd + e^{-kd}\right)}{\omega^2 \varepsilon_o^2 (1 - \varepsilon_m)\delta_m}}\,. \tag{3.5}$$

It is instructive to consider Z_p in the limit $kd \ll 1$ where the dispersion relation is close to the light line, and $kd \gg 1$ where the dispersion relation is flat near the surface plasmon resonance. When $kd \ll 1$, the exponential $e^{-kd} \rightarrow 1 - kd$ and $k \rightarrow \omega/c$ so that we recover the conventional parallel plate impedance given in Eqn. (3.1). When $kd \gg 1$, the exponential $e^{-kd} \rightarrow 0$ and $\delta_m \rightarrow 1/k$ so that the impedance becomes

$$Z_p = \frac{1}{W}\sqrt{\frac{\mu_o}{\varepsilon_o k^2} + \frac{2k^2 d}{\omega^2 \varepsilon_o^2 (1 - \varepsilon_m)}}\,, \tag{3.6}$$

which diverges as $W \rightarrow 0$. The exact plasmonic parallel plate transmission line impedance from Eqn. (3.5) is plotted in Fig. 8 for gold plates at $\lambda_o = 830$ nm with values of ε_m of taken from the literature[18]. When the plate width is fixed at 50 nm and the plate spacing is reduced below this value, the impedance drops as it would for a conventional parallel plate waveguide. When the plate width and plate spacing are scaled together the impedance diverges as $W, d \rightarrow 0$ with a value of roughly 4000 Ω for $W, d = 1$ nm.

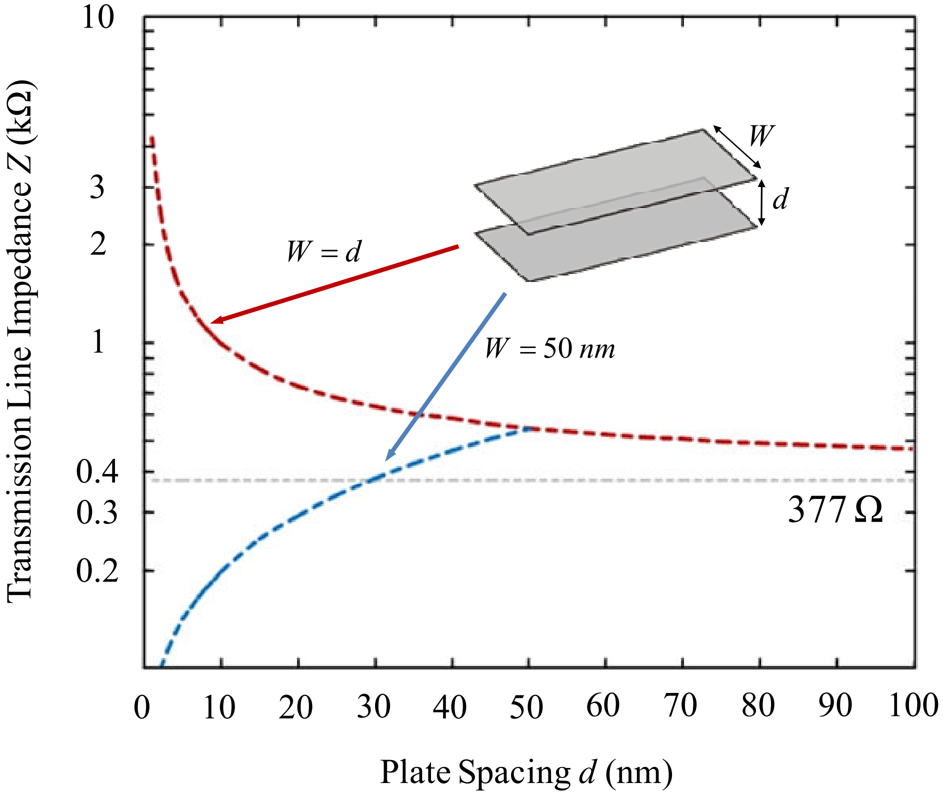


Figure 8: Transmission line impedance as a function of plate spacing d for a plasmonic parallel plate waveguide with fixed plate width $W = 50$ nm (blue line) and with square cross-section $W = d$ (red line). The impedance $Z = \sqrt{\mu_o/\varepsilon_o} \sim 377\Omega$ of a conventional parallel plate waveguide is also shown for comparison.

3.2 - Transformer Action

The ability to taper W to a narrower waveguide provides in effect a transformer action at optical frequencies. Traveling along a tapered waveguide toward a sharp tip the optical ac voltage increases, and the optical ac current decreases. A decreased current, as a result of transformer action, is accompanied by diminished I^2R resistive losses, which are a major problem in metal optics. In Fig. 9 we show two possible tapered plasmonic waveguide configurations that would result in transformer action. We compare the transformer action predicted by our circuit model to results obtained from numerical solution of the Maxwell equations for the configuration shown in the right panel of Fig.9 and in the top inset to Fig. 10. The case of tapering only the plate spacing while keeping the plate width constant as shown in the lower inset of Fig.10 is also considered.

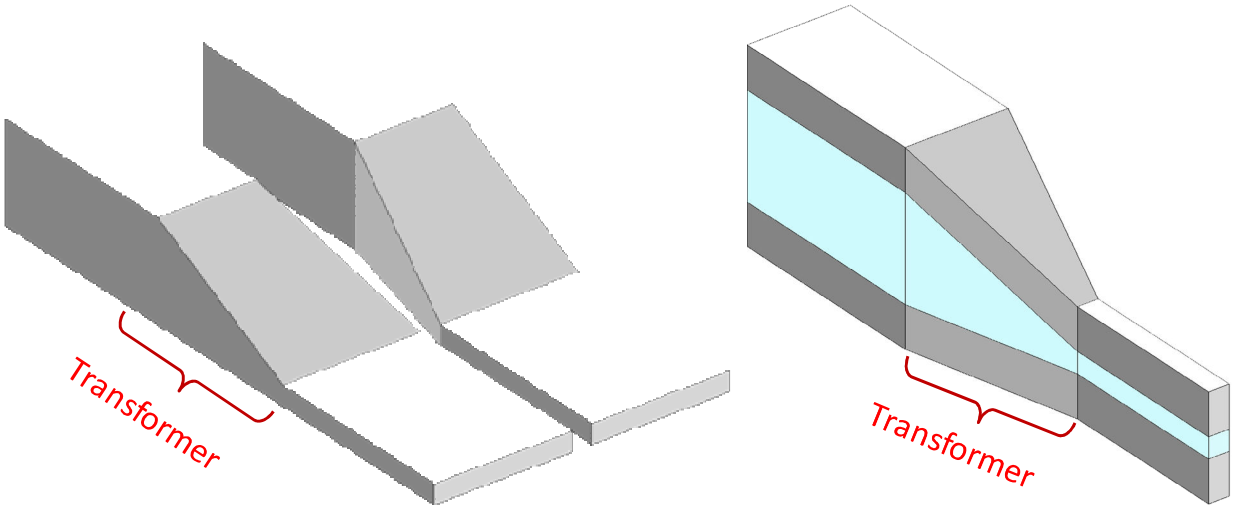


Figure 9: Two possible realizations of an optical voltage transformer. The blue color in the right panel indicates a dielectric sandwiched between the metal plates.

We again work at $\lambda_o = 830$ and use gold as the metal for the plates and free-space as the dielectric. In the numerical simulations we specify the electric field distribution at the tail end of the taper and solve for the distribution at the snout. In Fig.10 we plot using markers the ratio of the average electric field E_{out} at the snout to the average electric field E_{in} at the tail. The electric field enhancement when tapering both W and d is an order of magnitude greater than that obtained when tapering only d. In the latter case it is also found that in going from $d = 50 \rightarrow 1$ nm the electric field enhancement is only 20, as opposed to the expected enhancement of $d_{in}/d_{out} = 50$ one obtains from geometric reasoning. To obtain the electric field enhancement predicted by the circuit model we first convert E_{in} and E_{out} into voltages by using $E_{in} = V_{in}/d_{in}$

and $E_{out} = V_{out}/d_{out}$. Assuming the taper has negligible reflection losses (this will be the case[19] for an taper angle of ~ 20) conservation of power requires that $V_{out}/V_{in} = \sqrt{Z_{out}/Z_{in}}$, where Z_{in} and Z_{out} must be obtained from Eqn. (3.5) and pertain to the plasmonic parallel plate waveguides at the tail end and snout end of the taper, respectively. To covert back to an electric field ratio we once again make use of the relation $E = V/d$ and obtain $E_{out}/E_{in} = (d_{out}/d_{in}) \cdot \sqrt{Z_{out}/Z_{in}}$. The electric field enhancement predicted by the circuit model is shown as the dashed lines in Fig.10 and it is in excellent agreement with numerical results. We see that the 50 fold enhancement predicted by geometric reasoning for the constant width device has to be corrected by a factor $\sqrt{Z_{out}/Z_{in}}$ to account for the impedance mismatch between the waveguide at the tail end and the one at the snout end. According to Eqn. (3.5) and Fig. 8, for this device $Z_{out} < Z_{in}$ so that the correction factor evaluates to less than unity and the effective electric field enhancement is reduced to 20.

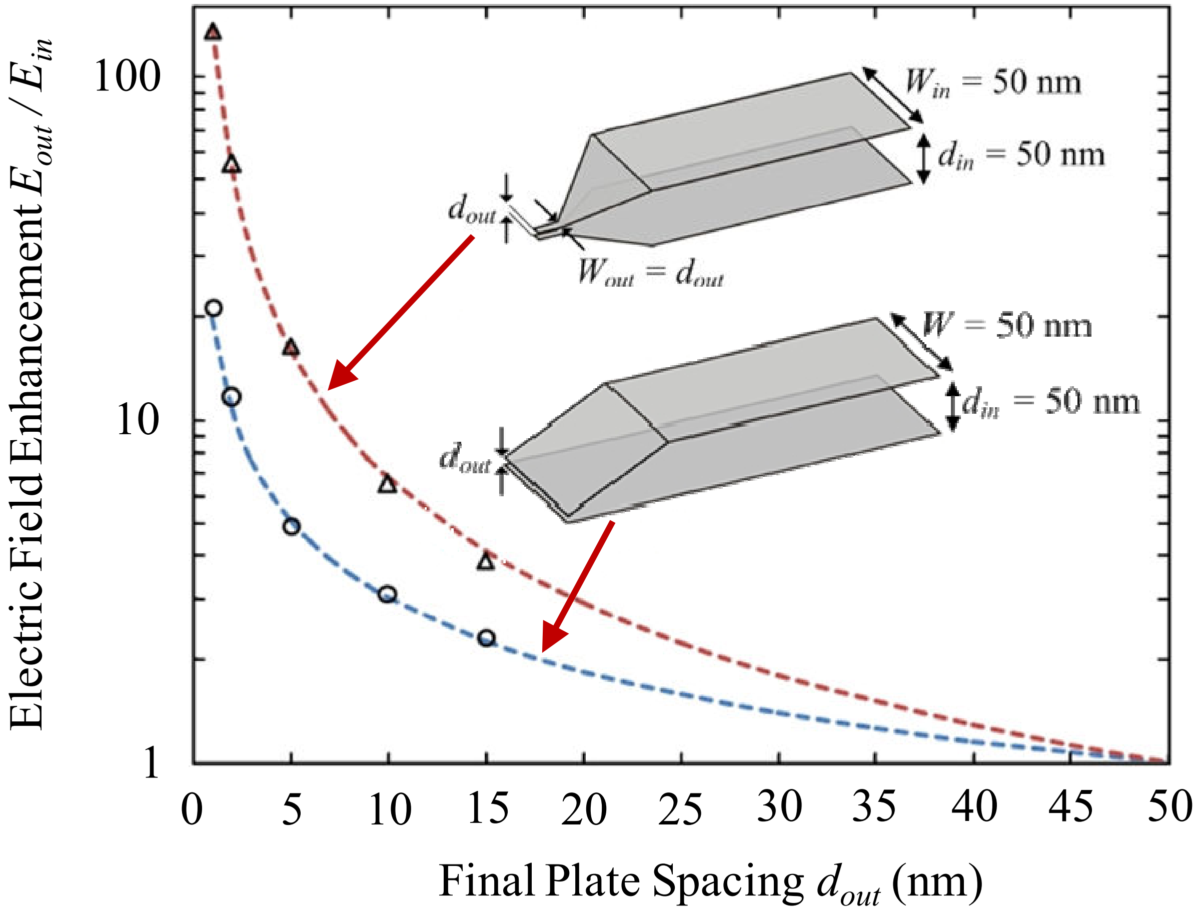


Figure 10: Comparison of transformer action predicted by circuit theory (dashed lines) to results obtained from full-wave solutions of the Maxwell equations (markers) for two different tapered waveguide geometries.

When the imaginary part of ε_m is included in the numerical simulations to account for the resistive loss in the metal, it is found that for short non-adiabatic tapers with taper angles 20-30 the power loss in traversing the taper is less than 3dB when staring with $W, d = 50$ nm and terminating with waveguide dimensions as small as $W, d = 1$ nm.

3.3 - Limits of Transformer Action

As we saw in Chapter 2, the origin of the diverging plasmonic parallel plate waveguide impedance can be traced back to the kinetic inductance associated with the inertia of the electrons in the metal. An expression for the kinetic inductance/per unit length is given by the second term in Eqn. (3.4). The first term corresponds to Faraday inductance. As one moves along the surface plasmon dispersion relation (Fig. 11) from the origin to the surface plasmon resonance asymptote, the Faraday inductance remains the same but the kinetic inductance grows larger, overtaking Faraday inductance roughly at the knee of the dispersion. Past the knee of the dispersion the skin depth $\delta_m \rightarrow 1/k$ and Eqn. (3.4) shows that kinetic inductance $L'_k \propto k/W$. For a given plate width W working at a higher wave vector k results in greater impedance and thus allows for more transformer action. Referring to Fig. 11, the amount of transformer action available from a given taper is proportional to the horizontal distance between the dispersion line corresponding to the plate spacing at the wide end (50 nm in our example) and the dispersion line corresponding to the plate spacing at the narrow end (15 or 5 nm in our example) at any given frequency. The greatest amount of transformer action is then obtained by tapering to the

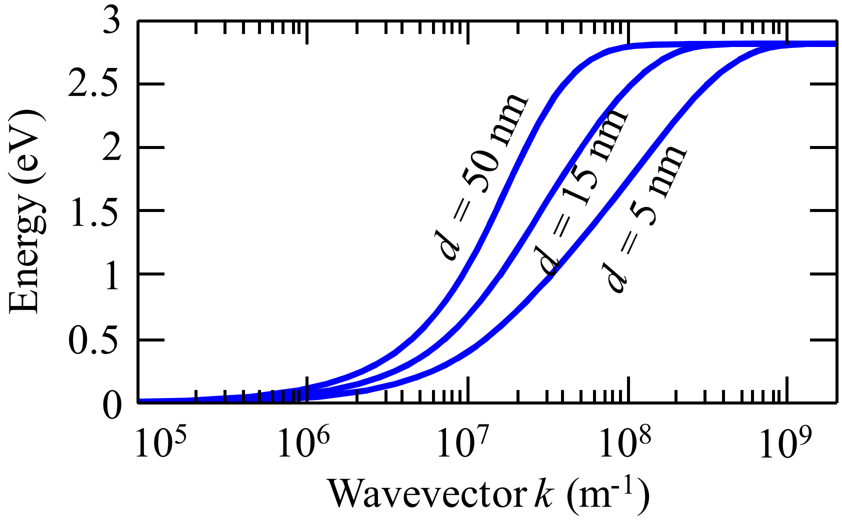


Figure 11: Semi-logarithmic plot of plasmonic parallel plate waveguide dispersion relation for three different values of plate spacing assuming gold plates.

smallest possible dimensions while working at the highest possible frequency. The operating frequency however cannot be arbitrarily large since the dispersion relation is constrained by the surface plasmon resonance of the metal comprising the waveguide plates. In gold and silver, intraband transitions and free electron contributions fix the surface plasmon resonance at around 2.6 eV and 3.5 eV, respectively[20]. Working at these frequencies, a plasmonic parallel plate waveguide tapering from $W, d = 50$ nm to $W, d = 1$ nm would result in a voltage boost of approximately 5x for gold, and 6x for silver. Starting with a wider taper does not improve the transformer action since beyond $W, d = 50$ the impedance of the line asymptotically approaches 377Ω and is rather flat as seen in Fig. 8. Although not shown, this would manifest itself in Fig. 11 as having the dispersion curves for $d > 50$ nm lay almost on top of the one for $d = 50$ nm. Attempting to achieve greater transformer action by tapering to waveguide dimensions below $W, d = 1$ nm is also ineffective due to non-local electron conduction losses which become dominant at these length scales and thus provide a fundamental limitation to the plasmonic propagation[19].

3.4 - Outlook

In Chapters 1-3, through both intuitive and rigorous means, we have shown that the transmission line impedance of the plasmonic parallel plate waveguide can be made very large. The impedance diverges reciprocally with the width of the waveguide. By tapering the waveguide the impedance can be made several times larger than the impedance of free space at the tip. This is the source of several effects which are unique to plasmonic devices. Namely there is the ability to efficiently couple light to the nanoscale by means of a tapered plasmonic waveguide. The resistance of the channel increases as more current is focused into a smaller area. From an initial evaluation, then, this would seem to preclude any kind of efficient focusing of the surface plasmon waves. Fortunately, the transmission line impedance increases along with the resistance, keeping the overall loss independent of the focusing. Plasmonic losses can therefore be kept manageable for a short taper. The availability of large impedances through tapered plasmonic waveguides suggests that efficient optical power delivery to the nanoscale is within reach[21]. Tapered plasmonic waveguide geometries could also be used as replacement for near-field scanning optical microscope (NSOM) probes[22-23] and as a heating element for heat assisted magnetic recording[24] (HAMR). When combined with optical antennas, which are the subject of Chapter 4, tapered plasmonic waveguides could be used as impedance matching tools[25-26] to mediate molecules at high impedances or match optical antennas to plasmonic transmission lines in plasmonic circuits[27-29]. This antenna matching could result in spontaneous emission enhancement[30] and contribute toward[31-34] surface enhanced Raman scattering (SERS). Later, in Chapters 5 and 6 we will consider in detail the application of optical voltage transformers as heating elements for HAMR.

4. Circuit Analysis of Optical Antennas

We begin this chapter on optical antennas by deriving and reviewing several fundamental properties of antennas, and in particular those of electrically small antennas. In section 4.1 we derive an approximate analytical expression originally due to Edward[35] for the scaling law of radiation resistance with antenna dimensions. In section 4.2 we draw an analogy between antennas and RLC circuits which when combined with Edward's scaling law allows us to recover the Wheeler limit[36] for the quality factor Q of electrically small antennas. In the section 4.3 we review the relationship between antenna Q and radiation capture efficiency and present a general proof originally due to Dubost[37] for the invariance of the effective area of an antenna at resonance. We conclude in section 4.4 by combining the well-known circuit model for an antenna with the circuit model for a cavity resonator and obtain fundamental limits pertaining to antenna-enhanced spontaneous emission rates.

4.1 - Radiation Resistance

In order to radiate an antenna has to be driven by an AC current, and ultimately an antenna's radiation pattern is merely the superposition of the radiation fields from the sea of free electrons in the antenna which oscillate in response to a driving current. The total energy flux per cycle due to an oscillating charge is proportional to the average acceleration of the charge squared. Namely it can be shown[38] that the time-averaged radiated power is

$$P = \frac{2}{3} e^2 \frac{\langle a^2 \rangle}{c^3} \ , \tag{4.1}$$

where a is the instantaneous acceleration of the charge, and $e^2 = q^2/4\pi\epsilon_o$, where q denotes the electron charge. In MKS units we thus have for N oscillating charges

$$P_N = \frac{2}{3} \frac{(Nq)^2}{4\pi\epsilon_o} e^2 \frac{\langle a^2 \rangle}{c^3} = \frac{1}{6\pi\epsilon_o c^3} (Nq)^2 \langle a^2 \rangle \ . \tag{4.2}$$

Then, noting that $\langle a^2 \rangle = \langle (\partial v/\partial t)^2 \rangle = \omega^2 \langle v^2 \rangle = \omega^2 v_{RMS}^2$, we have

$$P_N = \frac{\omega^2}{6\pi\epsilon_o c^3} (Nqv_{RMS})^2 \ . \tag{4.3}$$

In the last step we assumed an oscillation frequency (*i.e.* $v = v_0 \cos(\omega t)$). If we now multiply and divide by the length l of the antenna squared, we obtain

$$P_N = \frac{\omega^2}{6\pi\epsilon_o c^3} \left(l \cdot \frac{Nqv_{RMS}}{l} \right)^2 \ . \tag{4.4}$$

Noting that the term Nqv_{RMS}/l has units of charge per unit time and thus corresponds to a root mean square current I_{RMS}, we rewrite the last expression as

$$P_N = \frac{\omega^2}{6\pi\epsilon_o c^3}\left(l \cdot \frac{Nqv_{RMS}}{l}\right)^2 = \frac{\omega^2}{6\pi\epsilon_o c^3}(l \cdot I_{RMS})^2 = \frac{\omega^2 l^2}{6\pi\epsilon_o c^3} I_{RMS}^2 \ , \qquad (4.5)$$

which we recognize as the classic formula for power dissipated in a resistor, $P = IV = I^2R$, where now $R = \omega^2 l^2/6\pi\epsilon_0 c^3$ is the radiation resistance of the antenna.

$$R_{rad} = \frac{\omega^2 l^2}{6\pi\epsilon_o c^3} = \frac{1}{6\pi}\left(\frac{\omega l}{c}\right)^2 \frac{1}{\epsilon_o c}$$

but

$$c = 1/\sqrt{\epsilon_o\mu_o} \qquad \therefore\ 1/\epsilon_o c = \sqrt{\epsilon_o\mu_o/\epsilon_o^2} = \sqrt{\mu_o/\epsilon_o} \approx 377\ \Omega$$

therefore

$$\begin{aligned} R_{rad} &= \frac{1}{6\pi}\left(\frac{\omega l}{c}\right)^2 377\ \Omega \\ &= \frac{1}{6\pi}\left(\frac{2\pi f l}{c}\right)^2 377\ \Omega \\ &= \frac{4\pi^2}{6\pi}\left(\frac{fl}{c}\right)^2 377\ \Omega \\ &= \frac{2\pi}{3}\left(\frac{l \cdot c/\lambda}{c}\right)^2 377\ \Omega \\ &= \frac{2\pi}{3}\left(\frac{l}{\lambda}\right)^2 377\ \Omega\ . \qquad (4.6) \end{aligned}$$

Thus for a half-wave dipole antenna with $l = \lambda/2$, we expect

$$R_{rad} = \frac{2\pi}{3}\left(\frac{\lambda/2}{\lambda}\right)^2 377\ \Omega = \frac{2\pi}{3} \cdot \frac{1}{4} \cdot 377\ \Omega \approx 200\,\Omega\,,$$

which is a bit off the actual value of $v = v_0\cos(\omega t)$. Replacing the I_{RMS} with $I_0/2$ where we are now assuming that $I = I_0\cos(\omega t)$, the radiation resistance is reduced by a factor of 4:

$$R_{rad} = \frac{\pi}{6}\left(\frac{l}{\lambda}\right)^2 \cdot 377\ \Omega \qquad (4.7)$$

and for a half-wave dipole antenna we obtain $R_{rad} \approx 50\ \Omega$, which is a lot closer to the actual value of 73 Ω than the previous expression. Equation (4.7) is sometimes found in the literature[36] as $R_{rad} = 20\pi^2(l/\lambda)^2$.

4.2 - Antenna *Q* and Wheeler Limit

Equation (4.7) suggests that subwavelength antennas have a smaller radiation resistance than their larger counterparts. A consequence of this is that subwavelength antennas have larger Q's and thus perform poorly as radiators. To understand why this is the case, picture an antenna as an *RLC* resonator where here *R* represents the radiation resistance and the Ohmic resistance of the antenna is taken to be negligible. The *RLC* circuit resonates at some frequency ω with a quality factor

$$Q = \frac{L\omega}{R}.$$

Noting that for an RLC circuit it holds that $\omega_0 = 1/\sqrt{LC}$, we have

$$L = \frac{1}{\omega^2 C} \qquad \therefore Q = \frac{1}{\omega RC}\ , \tag{4.8}$$

and it thus follows that since $C \propto l\epsilon_0$ (where l represents the length of the antenna) and $R \propto (l/\lambda)^2$, smaller antennas will have larger quality factors.

This result is due to Wheeler[36] and it is usually expressed as

$$Q = \frac{3}{4\pi^2}\left(\frac{\lambda}{l}\right)^3 \tag{4.9}$$

which is known as the Wheeler limit. This expression gives the lowest bound on the Q of a small antenna of volume l^3 operating at a wavelength λ. We can recover Wheeler's limit using a simple circuit model by combining equations (4.7) and (4.8):

$$Q = \frac{1}{\omega RC} = \frac{1}{\omega R \cdot l\epsilon_o} = \frac{1}{\omega \cdot \frac{4\pi}{6}\left(\frac{l}{\lambda}\right)^2 \sqrt{\frac{\mu_o}{\epsilon_o}} \cdot l\epsilon_o} = \frac{6\left(\frac{\lambda^2}{l^3}\right)}{4\pi\omega\sqrt{\mu_o\epsilon_o}}$$

$$\therefore\ \ Q = \frac{3\left(\frac{\lambda^2}{l^3}\right)c}{4\pi f\pi} = \frac{3}{4\pi^2}\left(\frac{\lambda}{l}\right)^3,$$

which is the Wheeler limit.

4.3 - Antenna Capture Cross-section

In the previous section we saw that from the Wheeler limit, electrically small antennas have higher Q's than their larger counterparts. Since the quality factor is defined as the ratio of total energy stored in a harmonic system to the energy lost in one cycle of oscillation, it would seem that compared to larger antennas, electrically small antennas have a harder time reradiating energy. By reciprocity (for a lossless antenna) this also means that an electrically small antenna captures less power than a larger antenna, and is thus less effective. This is usually the case, unless the antenna is operating exactly on resonance, in which case the capture cross-section becomes independent of antenna dimensions[37]. This rather unintuitive result is summarized graphically in Fig. 12.

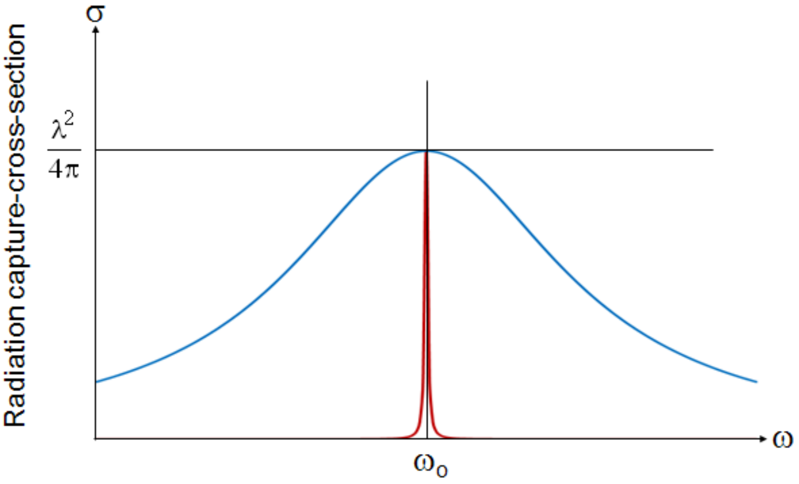


Figure 12: Invariance of capture cross-section for electrically small (red curve) and large (blue curve) antennas with the same central frequency ω_o.

An electrically large antenna (blue line) has a much lower Q than an electrically small antenna (red line) that resonates at the same center frequency ω_o. At the center frequency however, both antennas have the same capture cross-section $\lambda^2/4\pi$. The invariance of capture cross-section with antenna dimensions is the reason why cell phones can get away with antennas a few centimeters long even though they operate around 1 GHz (λ = 30 cm), and car antennas can pick up FM and AM broadcast even though they use λ = 3-300 m carrier wavelengths. To get a qualitative feel for what is happening, we once again draw on the *RLC* resonator analogy. Ignoring Ohmic losses, an antenna can be thought of as an *RLC* resonator: *R* is the radiation resistance of the antenna, *C* is the capacitance associated with the electric field distribution in the near-field of the antenna, and *L* is the inductance associated with the magnetic field distribution in the near-field of the antenna and may include a kinetic inductance contribution. When the *RLC* circuit (*i.e.* the antenna) is driven by an AC signal, energy is shuffled from the near-field inductance (currents on the surface of the antenna) to the near-field capacitance (charge distribution along the length of the antenna) and energy is lost as the current flows through the resistance R (radiation resistance from accelerated and decelerated charges on the surface of the antenna). When the antenna is connected to a load, whether it be a generator in transmission mode, or some circuit waiting for a voltage signal in receiving mode, it can be described by an impedance $Z = R + j\omega L - j/\omega C$. To keep the analysis as simple as possible, and without loss of generality, we assume receiving mode and that the load circuit has a purely resistive impedance which is matched to the antenna so that $Z_{load} = R_{load} = R$. It follows that in general, when radiation is incident on the antenna it will result in a voltage drop across the reactive part of the antenna impedance $j\omega L - j/\omega C$ and a voltage drop across the resistive part $R + R_{load} = 2R$. For electrically small antennas R is small compared to the reactive part, so that most of the energy remains trapped in the inductance and capacitance, resulting in a high Q. In larger antennas R is comparatively larger and thus sustains a larger fraction of the total voltage drop, resulting in a smaller Q and thus a better antenna. At resonance, however, we have that $\omega = 1/\sqrt{LC}$ and the impedance of the antenna becomes $Z = R + j\omega L - j/\omega C = R + j\sqrt{L/C} - j\sqrt{L/C} = R$. It follows that since the reactive part of the impedance cancels completely, all of the voltage drop occurs across the resistance, and power transfer to the load is no longer a function of antenna dimensions (i.e. efficiency is no longer determined by the ratio $|R|/|\omega L - 1/\omega C|$ as is the case off-resonance). This is a great intuitive explanation, but just like the picture on the previous page, it assumes that on resonance the power captured by an antenna is independent of its dimensions; something that we have not proven yet, and which we will take up next. In textbooks on antenna theory, the invariance of antenna capture cross-section is usually derived anecdotally by obtaining cross-section expressions for large and small antennas of different kinds, and showing that aside for a factor accounting for directivity, the answer always comes out to $\lambda^2/4\pi$. This approach is quite unsatisfactory since although recovering the correct answer, it does not provide any physical intuition as to why things are the way they are. To this end, we present below a general derivation solely from statistical thermodynamics (due to Dubost[37]) for the invariance of the cross-section of an antenna at resonance. The capture cross-section or effective area of an antenna is defined as

$$A_{eff} = P_a/N \quad (m^2)$$

where N is the power per unit surface area incident on the antenna and P_a is the power supplied by the antenna to its load assuming perfect impedance matching. In general the effective area is a function of antenna directivity, and we write $A_{eff} = A_{eff}(\theta, \varphi)$. It follows that an impedance matched antenna enclosed in a chamber whose walls are kept at some temperature T, will absorb an amount of power

$$dP_a = \frac{1}{2}\int A_{eff}(\theta, \varphi) E_\lambda d\lambda d\Omega \quad (W)$$

from the part of the black body radiation spectrum with wavelengths lying between λ and $\lambda + d\lambda$, which is taken to coincide with the antenna resonance. Here E_λ denotes the black body radiation energy density in the wavelength range λ to $\lambda + d\lambda$

$$E_\lambda = \frac{2hc^2}{\lambda^5} \frac{d\lambda}{e^{hv/kT} - 1} \quad (J/m^3)$$

and the integral is taken over the solid angle subtended by the antenna and over the wavelength interval λ to $\lambda + d\lambda$. The factor of 1/2 is included because in general the radiation inside the cavity will be isotropic and unpolarized. If $A_{eff}(\theta, \varphi)$ is constant over the spectral range of interest, then the integral is only over the solid angle and we have

$$dP_a = \frac{1E_\lambda d\lambda}{2} \int A_{eff}(\theta, \varphi) d\Omega$$

$$= \frac{hc^2}{\lambda^5} \frac{d\lambda}{e^{hv/kT} - 1} \int A_{eff}(\theta, \varphi) d\Omega \, .$$

At this point it is convenient to define a dummy function

$$r(\theta, \varphi) = A_{eff}(\theta, \varphi)/A_{max}$$

where A_{max} is the maximum effective antenna area. This allows us to express the integral as

$$dP_a = \frac{hc^2}{\lambda^5} \frac{d\lambda}{e^{hv/kT} - 1} A_{max} \int r(\theta, \varphi) d\Omega \, .$$

Provided that the antenna is in equilibrium with its surroundings it will also have an absolute temperature T and because at equilibrium it must radiate energy at the same rate at which it absorbs it, it follows that over the spectral range v to $v + dv$ (corresponding to λ to $\lambda + d\lambda$) the absorbed power will be

$$dP_a = hv \frac{dv}{e^{hv/kT} - 1} \, .$$

Note that $v = c/\lambda$ and so $dv/d\lambda = -c/\lambda^2 \;\; \Rightarrow dv = -cd\lambda/\lambda^2$. It follows that

$$dP_a = hv \frac{dv}{e^{hv/kT} - 1} = \frac{hc^2}{\lambda^5} \frac{d\lambda}{e^{hv/kT} - 1} A_{max} \int r(\theta, \varphi) d\Omega$$

$$\Rightarrow hv \; dv = \frac{hc^2}{\lambda^5} d\lambda \, A_{max} \int r(\theta, \varphi) d\Omega$$

$$\Rightarrow \frac{hc^2}{\lambda^3} d\lambda = \frac{hc^2}{\lambda^5} d\lambda \, A_{max} \int r(\theta, \varphi) d\Omega$$

$$\Rightarrow \lambda^2 = A_{max} \int r(\theta, \varphi) d\Omega \,.$$

In particular, for an isotropic antenna $r(\theta, \varphi) = 1$ and the integral collapses to give the classic result $A_{max} = \lambda^2/4\pi$, independent of antenna dimensions, thus concluding our analysis.

4.4 - Optical Antennas and Spontaneous Hyper-Emission

In this section we combine the well-known circuit model for an antenna introduced in Chapter 4.1 with the circuit model for a cavity resonator and obtain fundamental limits pertaining to antenna-enhanced spontaneous emission rates. Optical antennas amount to metallic structures that are resonant at optical frequencies, and as such they may be modeled as simple series RLC circuits driven by a voltage source. On resonance, the reactive components of the circuit cancel out and one is ideally left with the radiation resistance R_{rad} in series with the voltage source. For a general antenna of length l, equating the power dissipated in this circuit, $P = I^2 R_{rad}$, to the power radiated by a collection of oscillating charges from the Larmor formula and rearranging terms, one obtains[35] a simple expression for the radiation resistance $R_{rad} = (4\pi/6)(l/\lambda)^2 \cdot 377\Omega$, where λ is the working wavelength. It is also well know[36] that a series RLC circuit has a quality factor $Q = 1/\omega RC$, where ω is the resonance frequency of the circuit, and $c = 1/\sqrt{\varepsilon_o \mu_o}$ is the speed of light in free space. Taking $C = l\varepsilon_o$ and $R = R_{rad}$, we obtain the expression

$$Q = (3/4\pi^2)(\lambda/l)^3, \tag{4.10}$$

which gives the lowest bound on Q for a small antenna of volume l^3 operating at a wavelength λ. Eqn. (4.10) was first derived[36] in a lengthier and more formal fashion in 1947and is known as the Wheeler limit. Another well-known function relating the quality factor of a resonator to its volume is the Purcell factor[39] describing the spontaneous emission rate enhancement in a cavity resonator compared to free space. Namely, at a given wavelength, the spontaneous emission rate in a cavity is found proportional to Q/V_{cav}, where V_{cav} is the cavity volume. We can recover this result by once again modeling the resonator cavity as an RLC circuit. We can imagine the cavity comprising a parallel plate capacitor with $C = \varepsilon_0 A/d$ having the plates shunted by a small wire with some associated resistance R and inductance L. A dipole radiator qx located between the capacitor plates (with x perpendicular to the plates) will induce a charge magnitude qx/d on each plate, and give rise to a voltage $V = (qx/d)/C$ across the capacitor. This voltage dissipates a power $V^2/R = \omega Q\, (qx)^2/d^2 C$ through the resistance R, where we used $Q = 1/\omega RC$. Substituting for C and letting $A \cdot d = V_{cav}$, one readily recovers the Purcell effect with power lost by a radiator inside a cavity being proportional to Q/V_{cav}. We now combine the circuit model for an antenna with the circuit model for a cavity resonator discussed above, and obtain a fundamental upper limit for the spontaneous emission enhancement rate of an optical antenna.

We first consider an antenna by itself as in Fig. 13(a), and then the same antenna coupled to a resonator cavity as in Fig. 13(b).

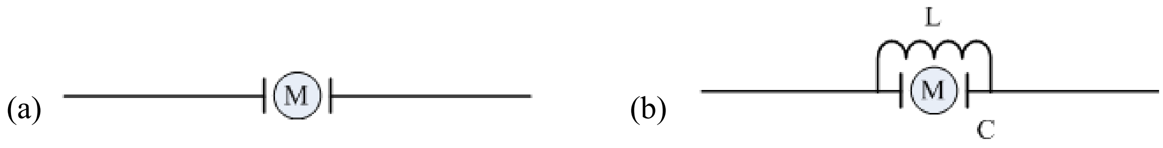


Figure 13: (a) A radiating molecule at the feed gap of an antenna. (b) A radiating molecule in a resonant cavity, coupled to the feed gap of an antenna.

According to the circuit picture[40], on resonance, the power radiated by the antenna is $P = I^2 R$. If we assume the current is arising from the cyclic motion of charge q in an excited molecule trying to return to its ground state by emitting a photon, then we may let $I = (q\omega/2\pi)$and obtain $P = R(q\omega/2\pi)^2$. Taking $R = \sqrt{\mu_o/\varepsilon_o} \approx 377\Omega$ (*i.e.*, assuming the best case scenario of an antenna perfectly matched to free space) and solving for the spontaneous emission rate $\tau_{sp}^{-1} = P/\hbar\omega$ we find

$$\frac{1}{\tau_{sp}} = \frac{\omega}{\pi}\left(\frac{q^2}{4\pi^2\varepsilon_o\hbar c}\right) = \frac{\omega}{\pi}\alpha\,, \tag{4.11}$$

where$\alpha \approx 1/137$ is the fine structure constant.Coupling the antenna to an *LC* resonator housing the driving molecule sharpens up the resonance and allows for even faster spontaneous emission rates via the Purcell effect. Namely, invoking Fermi's golden rule on resonance with a single-mode Lorentzian density of states and designing the resonator optical lifetimeτ_p to equal the spontaneous emission lifetime[5]:$\tau_p = 1/\Delta\omega = \tau_{sp}$ gives

$$\frac{1}{\tau_{sp}^2} = \frac{4}{\hbar^2} q^2 x^2 E_{zp}^2\,,$$

whereqx is the dipole moment of the molecule in the cavity, and $\varepsilon_o E_{zp}^2 V_{cav} = \hbar\omega/2$ is the energy associated with the zero-point field. Rearranging terms we obtain

$$\frac{1}{\tau_{sp}} = \omega\sqrt{\frac{q^2}{4\pi^2\varepsilon_o\hbar c}}\sqrt{\frac{4x^2\lambda}{V_{cav}}} = \omega\sqrt{\alpha}\sqrt{\frac{4x^2\lambda}{V_{cav}}}\,, \tag{4.12}$$

where$\alpha \approx 1/137$ is the fine structure constant and the last term under square root is a volume ratio. It follows that if $V_{cav}/4x^2\lambda < 1/\sqrt{137}$ then the spontaneous emission rate will be much greater than that of a molecule at the feed gap of an antenna without a coupled resonator. For a driving molecule with $x = 0.5$nm dipole at λ_0 = 820 nm and an antenna cavity volume of $20 \times 80 \times 150$ nm^3, Eqn. (4.12) predicts seven orders of magnitude spontaneous emission enhancement compared to the molecule in free space. Furthermore, by reducing the cavity volume, the spontaneous emission may be enhanced great enough to be on the order of the radiation frequency itself. A survey[30,41-42]of the art reveals the very large antenna-assisted spontaneous emission enhancement predicted by the simple circuit model has not yet been achieved, with the largest enhancement observed[42] to date at only 28. This is in part due to poor choices in antenna design and difficulty in fabricating designs that promise to perform well as these usually require very small feed gaps beyond the reach of current processing techniques. Additional obstacles to achieving the fundamental limit set by Eqn. (4.12) include polarization misalignment between the driving molecule and the preferential axis of the antenna, non-optimized spatial alignment of the molecule to the field maximum, resonance misalignment, and the finite losses associated with the metal comprising the antenna. In smaller optical antennas, where the dimensions are comparable to the skin depth of the optical fields, the metal losses become significant (Q_{loss}~ 10's). It is then important to design the radiation-Q of the antenna to equal that of the loss-Q in order to achieve quantum efficiencies close to 50%. Overcoming these obstacles and attaining spontaneous emission rates towards the driving optical frequency would have profound implications. Spontaneous emission rates approaching the limit set by Eqn. (4.12) would allow direct modulation speeds of nano-LED's beyond 100 Gbits/sec for interconnects, far exceeding the modulation speeds of lasers relying on stimulated emission[43]. Additionally, many molecules that do not fluoresce will radiate efficiently when placed near properly designed antenna structures, allowing a whole new class of bio-sensors and improving Surface-Enhanced Raman Scattering (SERS) single molecule sensors[31,32] which would then also find a rational scientific basis in antenna theory.

5. HAMR Optical Power Delivery as a Circuit Problem

Heat Assisted Magnetic Recording (HAMR) is a scheme for encoding information in magnetic media with very small track widths so as to enable very high areal densities. The HAMR concept is exhaustively detailed in Chapter 6 and for the purpose of the present chapter it will suffice to know that HAMR involves using an optical antenna to capacitively couple energy to a metallic (magnetic) recording medium across a 5-10 nm air gap. As the problem has been around for some time, several ideas have been put forth for such optical antennas, however none has come out ahead as a clearly superior device. In this chapter we leverage the lumped element approach to plasmonics developed in Chapters 1-3 to narrow the design space of optical antennas for HAMR. In Chapter 6 we push beyond the limits of the circuit model and examine the HAMR problem by numerically solving the Maxwell equations.

5.1 - Optical Antennas and Antenna Loads in HAMR

While ultimately one will have to rely upon the numerical solution of the Maxwell equations when designing an antenna for HAMR, it is instructive at first to put aside numerical techniques and consider the problem from a more fundamental perspective through circuit theory. The left panel in Fig. 14 schematically illustrates a monopole antenna above the target recording layer as a simple approach to HAMR. The electric field lines associated with the antenna resonance are shown qualitatively in green; when in the recording layer or in the antenna these lines also correspond to real current flow. Note that because of our choice of design, there are return currents immediately below the antenna, in the region we want to heat (highlighted in red), but that there are also return currents in the recording layer far away from the antenna. These latter currents are to be considered parasitic, since they will result in heating outside the target region directly below the antenna. This system can be modeled using simple circuit elements as illustrated in the right panel of Fig. 14. The antenna is reduced to a voltage source in series with

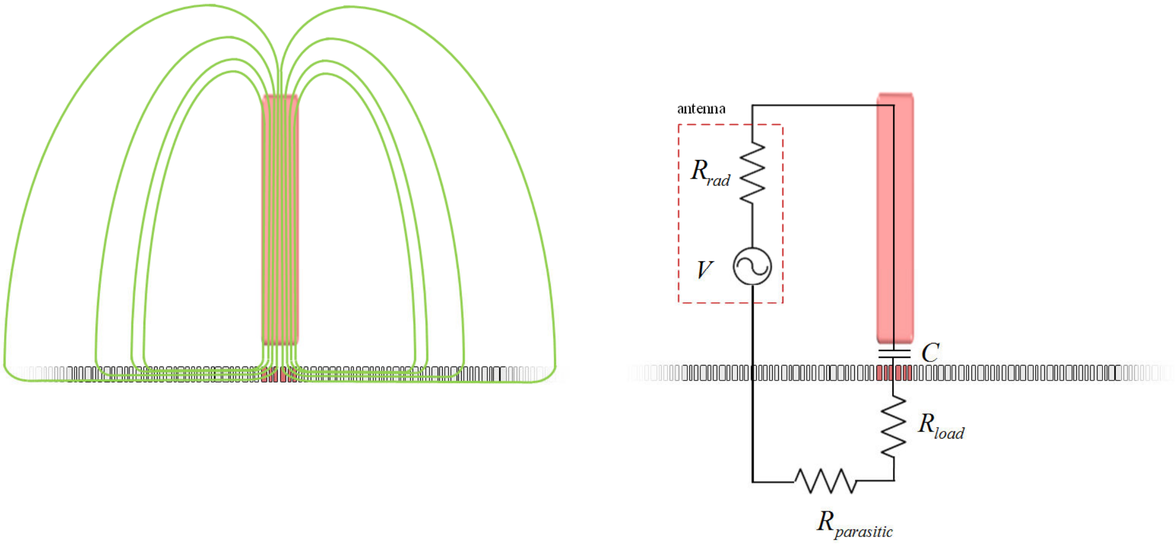


Figure 14: *(Left panel)* a monopole antenna for HAMR in the electromagnetic field picture. *(Right panel)* a monopole antenna for HAMR in the equivalent circuit picture.

a radiation resistance R_{rad}, the gap between the antenna and the recording layer becomes a capacitance C, and the recording layer very roughly reduces to a load resistance R_{load}, corresponding to resistive losses in the region immediately below the antenna, and a parasitic resistance $R_{parasitic}$, corresponding to resistive losses due to the rest of the return currents. The power dissipated at the load resistance may then be expressed as

$$W_{load} = |I|^2 R_{load} = \frac{|V|^2 R_{load}}{\left|R_{load} + R_{rad} + R_{parasitic} - j/\omega C\right|^2}\,. \tag{5.1}$$

First of all, note the reactive term in the denominator will in general be very large since the capacitance will be very small for very small antennas required to heat very small regions of the recording layer. To compensate then one should try to make the angular frequency of the radiation being used as large as possible, to prevent the ωC product in Eq. (5.1) from becoming too small. Also note that among the terms in the denominator, $R_{parasitic}$ is not a physical requirement, but rather a result of our particular choice of antenna which is lowering our efficiency. By simply switching to a different type of antenna, like the one shown in Fig. 15, we can eliminate the parasitic resistance term and improve our throughput. If we consider the antenna in Fig. 15 as a heating element for 1 Tb/in^2 HAMR, then the antenna will have to heat a

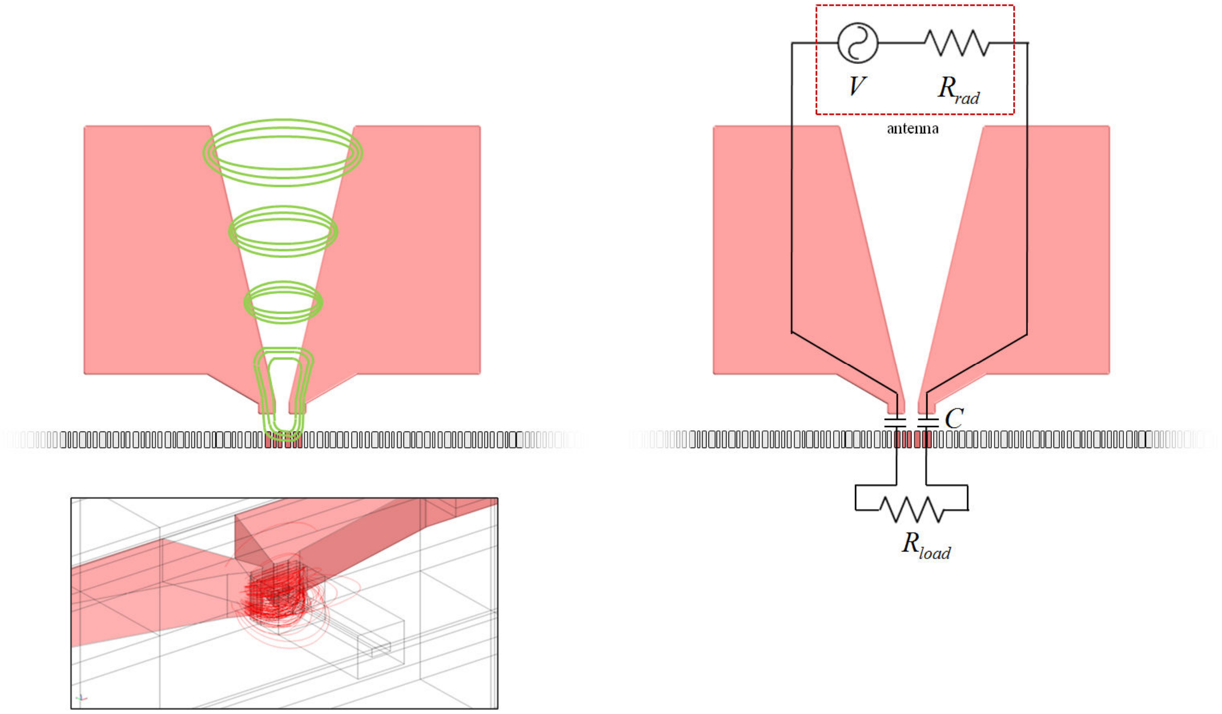


Figure 15: Analogue of Fig. 14 but for an antenna that does not suffer from parasitic resistance at the load. The lower inset in the left panel is a numerical simulation result showing that the current at the load is mostly confined in the region immediately across from the antenna, as desired.

12.5 nm x 50 nm x 5 nm region of the recording layer. If we assume the recording layer comprises Cobalt, then we can model the target region of recording layer as a load impedance by using the results from Chapter 1. Because the antenna is not in contact with the recording layer, we also have to include a reactance associated with capacitive coupling to the load. Working with 830 nm light and choosing antenna dimensions appropriate for 1 Tb/in^2 we obtain a load impedance $Z_{load} = 1035 + j772 - j995\,\Omega = 1035 - j223\,\Omega$. The real term corresponds to Ohmic losses in the load region. The positive imaginary term corresponds to a reactance associated with the kinetic inductance of the electrons in the metal comprising the load. The negative imaginary term corresponds to a reactance associated with capacitive coupling between the antenna and the load. Note that while positive and negative reactive terms partially cancel, there still remains a relatively large reactive component to the impedance, which for this particular antenna will prevent efficient power delivery to the load. Note also that the real part of the load impedance is rather large, on the order of kilo-Ohms, whereas the radiation resistance of a typical good antenna will be around 50-70 Ohms. The mismatch between the radiation resistance of good antennas and the load impedance will result in very poor efficiency, below 5%, in delivering power to the load. What is needed is a matching network that transforms the low antenna impedance to the very high impedance of the load to facilitate power transfer. This is the subject of the following section.

5.2 - Optical Antenna Matching Networks for HAMR

The load matching problem discussed in Chapter 5.1 presents a great opportunity for the optical voltage transformer introduced in Chapter 3. The 1 kΩ resistance presented by the load is well within the reach of tapered plasmonic parallel plate waveguide transformers like those shown in Fig. 9. The circuit approach suggests employing a tapered plasmonic parallel plate waveguide to couple power to HAMR loads with reasonable efficiency. This subject is taken up in Chapter 6 where the tapered plasmonic waveguide approach is compared through rigorous numerical solution of the Maxwell equations to other contending optical antenna designs for HAMR and it is found to be the most efficient even without any optimization.

6. Heat Assisted Magnetic Recording

Most electrical engineers are aware of Moore's law; the exponential growth in the number of transistors industry has been able to squeeze on a silicon die ever since Intel's 4004 microprocessor debuted in 1971 with 2300 of them. Few however are aware of the equally impressive exponential growth in the aerial density of magnetic storage media since IBM first introduced the commercial hard disk drive in 1956. While the semiconductor industry is expected to run into trouble keeping up with Moore's law around 2030 as most device parameters will simultaneously reach fundamental physical limits[17], the magnetic storage industry is already in trouble as the magnetic domains used to store information on hard disks have become so small that they can barely maintain their magnetization in the face of random thermal energy fluctuations[44]. To maintain the historical aerial density growth rate the magnetic industry is expected to begin using a new recording technique known as heat assisted magnetic recording (HAMR). This technique relies on the availability of near field transducers that can efficiently focus energy to the nanoscale. As the problem has been known for some time, several ideas have been put forth for such transducers, however none has come out ahead as a clearly superior device. Here we survey the state of the art and then leverage the lumped element approach to plasmonics developed in the first part of the dissertation to arrive at an optimal design for a plasmonic near field transducer.

The work presented in the remainder of this dissertation is the result of a close collaboration with Western Digital Inc. and the Information Storage Industry Consortium (INSIC) who funded our research. HAMR is expected to go into high volume manufacturing by 2015 so only the work that took place during the pre-competitive phase of the research will be disclosed here.

6.1 - Introduction

The first magnetic hard disk drive, released by IBM in 1956, had an areal density of 250 bytes/in^2 and a capacity of 4.4 MB. The drive consisted of a stack of fifty 24" discs and was housed in an enclosure the size of an automobile. Today's hard drives can pack two terabyte of storage in a form factor about half the size of a videocassette and can be bought off a shelf for less than US $100. By comparison, IBM would lease its original hard drive for US $35,000 a year, and this figure is not even adjusted for inflation. In order to sustain this kind of progress the magnetic storage industry has had to find ways of continually increasing areal recording density while maintaining adequate media signal-to-noise ratio. The most straightforward way of increasing areal density is to decrease the mean media grain volume. Decreasing grain volume, however, jeopardizes data integrity because as grain size is reduced, the recorded magnetization configuration becomes increasingly susceptible to random thermal fluctuations. The stability metric for media with characteristic magnetization switching volume V is given by the ratio of the magnetic energy stored in the volume to the thermal energy of the surroundings [44]:

$$\text{Stability} \propto \frac{\text{Stored magnetic energy}}{\text{Thermal energy}} \propto \frac{\text{Anysotropy} \times \text{Volume}}{k_B T} = \frac{K \times V}{k_B T} \; .$$

In order to assure stability against thermal agitation, the magnetocrystalline anisotropy energy per unit volume K must be increased proportionally as the grain volume V is reduced. However, K cannot be elevated arbitrarily since the output magnetic field of the inductive heads required to record a magnetic medium must scale commensurately with the medium mean switching field, which in turn scales in proportion to K. Because the output field of a write head reaches a hard material limit corresponding to the highest saturation magnetization values found in magnetic solids, the value of the recording medium's K at the instant of recording is strictly limited by the capability of recording heads.

Luckily, for most materials of interest in magnetic recording, K possesses a temperature dependence which can be used to advantage. The anisotropic energy of ferromagnetic materials vanishes as the material temperature is increased toward the Curie temperature, T_C. Magnetic recording media materials, usually alloys containing Fe, Co, Pt, and/or Ni have Curie temperatures that are at least a few hundred Kelvin above room temperature. Supposing that for the recording medium *K(T)* falls monotonically from some high value at ambient temperature to zero at T_C one can conceive that the medium could be brought from a condition of extreme stability at room temperature to very little or no stability at an elevated temperature. In such a situation, if the medium can be locally heated momentarily during recording, the coercivity and the thermal stability issues would be solved simultaneously. Such a process is known as *Heat Assisted Magnetic Recording* (*HAMR)*, and it is presently recognized as the leading candidate for extending the historical growth of magnetic recording past the limitations of thermal stability at room temperature. Achieving the areal densities projected for HAMR will require a heating element capable of focusing energy to extremely small scales, on the order of a hundred square nanometers. Although this level of confinement cannot be achieved through conventional optics, it is well within the reach of plasmonics. Because currently there are no means of reliably lasing surface plasmon modes at room temperature, generation of surface plasmons requires external optical excitation. It becomes clear then that the heating element must double as an optical antenna, converting high frequency free-space radiation into localized high frequency currents. As we saw in Chapter 5, generally the antenna dimensions required to work with optical excitations will result in antenna impedances that are orders of magnitude smaller than those found at the nanoscale. It follows that in order to efficiently deliver the energy captured by the antenna to a nanoscopic load one needs an impedance matching circuit with considerable dynamic range. Although the required transformer action is untenable in the RF and microwave regimes, at optical frequencies such a transformer can be fashioned out of a tapered plasmonic waveguide. Arguably then we have available to us all the fundamental components required to efficiently capture and deliver energy to the nanoscale. All that remains to be done is to integrate these components in the most effective way so as to arrive at an optimal design for a plasmonic heating element – this will be our goal in this chapter. Our metric for viability will be how closely any one scheme comes to attaining a magnetic recording areal density that can surpass the projected limit for the currently employed magnetic recording technology (shingled perpendicular magnetic recording) which is expected to fail beyond 1.5 $\mathrm{Tb/in^2}$.

6.2 - The State of the Art

Optical antenna designs are a dime a dozen, but if one is to operate as a near field transducer (NFT) in a HAMR drive, it has to be rather special. Requirements for an NFT include, but are not restricted to: *(i)* achieving a cross-track full-width at half-maximum (FWHM) optical spot of 30 nm or less (measured half-way into the recording layer), *(ii)* being able to operate with a magnetic-thermal offset (MTO) of 30 or less (meaning that the tip of the magnetic write pole has to be within 30 nm of the NFT, and although it is only the tip, it is quite large with an interface on the order of 100×100 nm^2), *(iii)* be as efficient as possible when working with a continuous wave (CW) laser diode at a wavelength of around 830 nm (minimum 3% NFT-to-recording media power conversion efficiency is required, and obviously the NFT cannot melt during operation), *(iv)* the NFT geometry must be tolerant to fabrication error (all fabrication is lithographic for high volume manufacturing), especially lapping error since the air bearing surface (ABS) of the slider is defined through a lapping process with an electronic lapping guide sigma of around 10 nm. Compatibility with the standard process flow of the slider limits the NFT material to gold, and the surrounding optical media mainly to alumina (Al_2O_3), tantalum oxide (Ta_2O_5), and silica (SiO_2).

The current leading approach[45,46] to realizing HAMR was developed by Seagate Technologies Inc. and it is sketched in Fig.16. Power delivery to the recording media is achieved in three separate stages: *(i)* light from a diode laser is focused to roughly a 50 μm spot and coupled to a wide dielectric transverse electric (TE) waveguide consisting Al_2O_3/ Ta_2O_5 with a core thickness of 120-150 nm by means of a linear grating. The vertical confinement of the light is on the order of the core thickness. *(ii)* The dielectric waveguide is etched in the shape of a parabola and the sides are coated with gold so the light coupled through the grating is focused to a diffraction limited spot at the focal plane of the parabola. This is known as a planar solid immersion mirror[47-49] (PSIM) and for light at a wavelength of 830 nm can provide a lateral confinement of roughly 200-300 nm. *(iii)* A gold disk of radius ~200 nm and roughly 20 nm thick is placed at the focal plane 10-20 nm above the core. It is buried inside the waveguide cladding. The disk has a peg roughly 20 nm wide and 20 nm long that protrudes toward the media and terminates at the ABS. This so-called lollipop structure is an optical antenna and capacitively couples optical power to the region of the recording media located roughly 5 nm directly below the peg. The optical spot size in the media has about the same shape as the cross-section of the peg at the ABS plane, so that for a 20 nm wide peg, the cross-track FWHM can ideally be around 20 nm.

The tip of the magnetic write pole is not shown, but would be located directly above the disk of the NFT and could possibly contact both the NFT disk and peg. A close-up of the lollipop NFT is shown in Fig.17. The lollipop NFT is designed to operate as a quadrupole antenna with the disk diameter and surrounding dielectric fixing the resonance condition. The peg only serves the purpose of locally enhancing the electric field intensity beneath it when the media is present, and it largely does not affect the resonance of the disk. As suggested by the (left) diagram in Fig.17, it is desirable for the purpose of enhancing the electric field inside the media, to have a ground plane underneath the recording layer. As we will see later on, this is however difficult to achieve in practice.

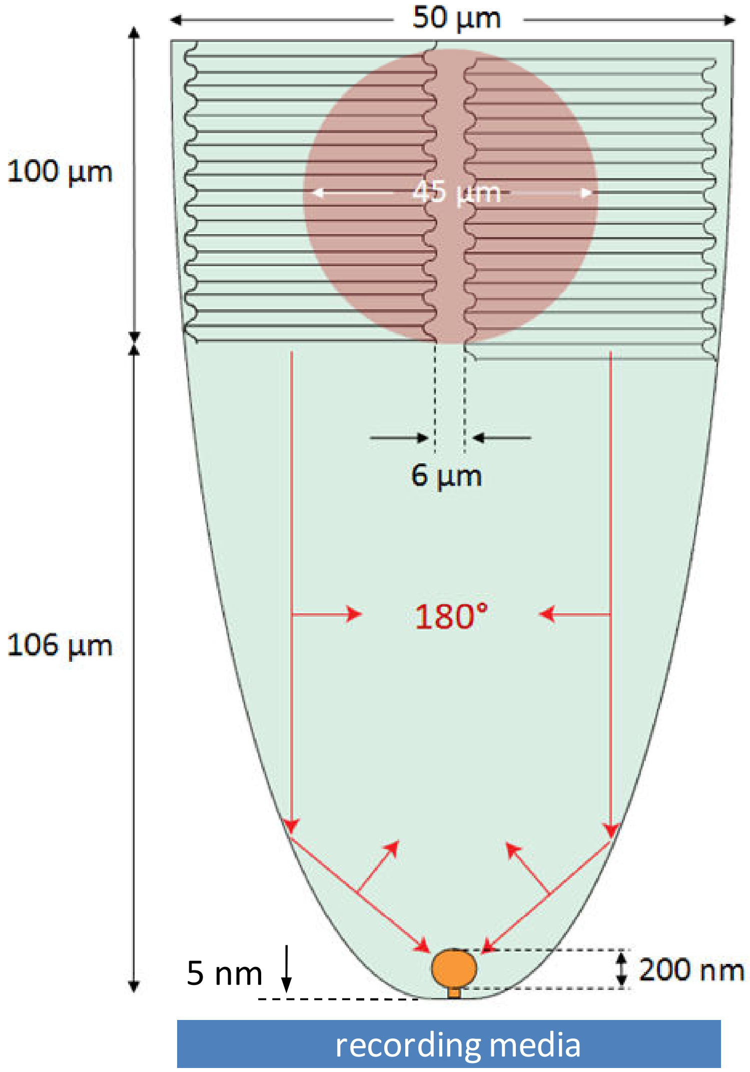


Figure 16: Schematic of the lollipop NFT design illuminated by a PSIM as proposed by Seagate Technologies Inc.

The second most promising approach[50] for implementing HAMR is being championed by Hitachi Global Storage Inc. and comprises butt-coupling light from a laser diode into a large cross-section (~600×300 nm^2) dielectric (Al_2O_3/ Ta_2O_5) waveguide which terminates on a gold film at the ABS with a conventional[51,52] C-aperture etched into it. The thickness of the gold film has to be adjusted to maximize throughput for a given working wavelength, and whereas in the case of the lollipop NFT the optical cross-track FWHM was fixed by the width of the peg, for the C-aperture it is fixed by the width of the metal ridge extending into the opening. In this scenario the magnetic write pole tip is located opposite the metal ridge, replacing a portion of the gold film. Obviously the MTO will be limited by the spacing between the top of the ridge and the inner boundary of the aperture. Variations[53,54] on these designs have been proposed but have yet to catch on with industry.

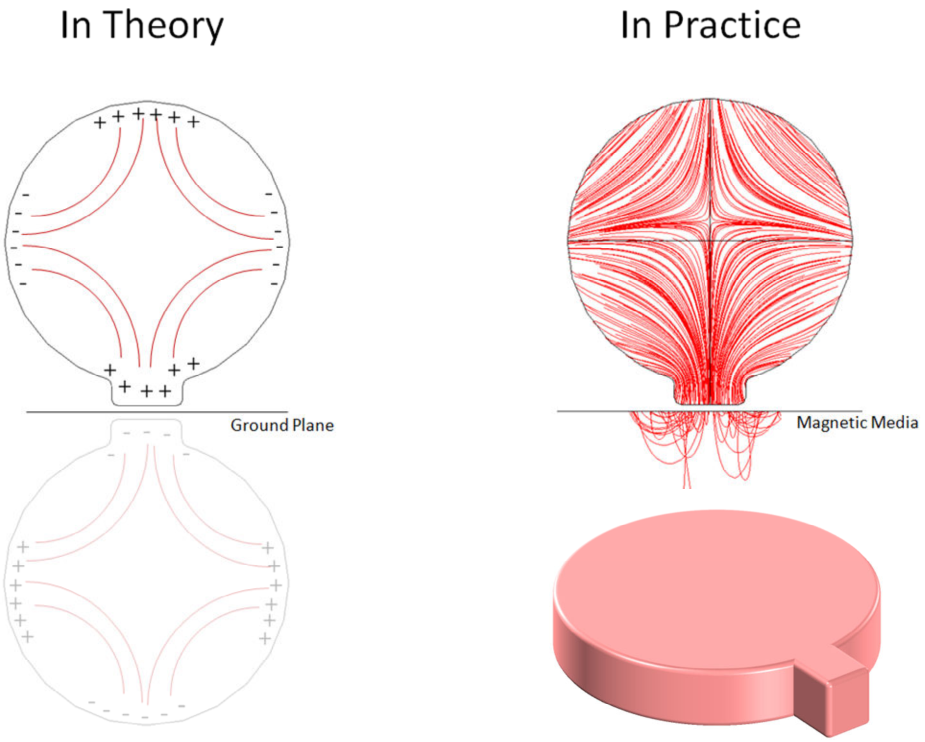


Figure 17: Perspective view of the lollipop NFT along with quadrupole current distributions in theory (left) and in practice, as obtained numerically (right).

6.3 - Impedance Matching

From the circuit perspective outlined in the first part of the dissertation, HAMR boils down to an impedance matching problem. We have shown before[55] that there are two main obstacles to an efficient implementation of NFT-media coupling. The first obstacle is the unavoidable air-gap between the NFT and the recording media. This forces the NFT to be capacitively coupled to the media, and due to the small size of the capacitive coupling region, the reactance associated with this capacitance is overwhelming and dominates the voltage divider between itself and the resistance associated with Joule heating in the media. The second obstacle is the inherent mismatch between the impedance associated with the heating target region on the media (which tends to be quite large due to the small dimensions required for HAMR) and the radiation resistance of the optical antenna. It is reasonable to take the impedance of a good antenna to be around 50Ω. When we factor in the Ohmic losses in the antenna and the disproportionately large capacitive reactance and impedance of the heating target in the media we find that the efficiency with which power can be coupled from a conventional NFT (such as the lollipop antenna or the C-aperture) to a small region of recording medium (on the order of 50×50×8 nm^3) is only a few percent. We previously suggested[55] the two aforementioned problems can be mitigated by using an optical voltage transformer in a tapered waveguide structure in place of more conventional NFTs. Next we will present modeling results that provide a quantitative comparison between the performance of our impedance matched NFT design to that of more common ones mentioned earlier.

6.4 - Modeling Challenges

The optics of heat assisted magnetic recording present a formidable modeling challenge. This is mainly due to the large dynamic range required of the modeling, which has to span from hundreds of microns to capture the grating and PSIM focusing structure, to the sub-nanometer resolution required to resolve the air gap between NFT and media, and the complex structure of the media itself. Additionally, the media stack comprises multiple layers each with thickness on the order of nanometers, interfacing with several square microns of the ABS which must be modeled to capture the effects of side-track erasure due to stray light. Resolving the skin depth of the light in the metal of the NFT, write pole, mirror coatings, and metallic media layer is also a troublesome task due to their large cumulative surface area. A combination of finite-difference time-domain (FDTD) and finite element method (FEM) approaches are required to overcome these challenges and their roles are discussed in Appendix G (while expositions[56,57] of FDTD techniques for electromagnetics are widely available, the same cannot be said for FEM[58], so we have included a brief and very readable introduction to the subject in Appendices H-J for the interested reader).

Irrespective of NFT choice, to properly capture the physics of the problem one aspect is of paramount importance: the illumination must be properly modeled in order to correctly determine the efficiency of the NFT-media coupling. This efficiency is conventionally defined as the ratio of the power lost in a 100×100×*(thickness of recording layer)* nm^3 immediately across from the NFT, which is given by evaluating the integral (see Appendix F)

$$\frac{1}{2}\int \omega\varepsilon_o\varepsilon_m''|E|^2 dV \,,$$

to the total power that reaches the back plane of the NFT from the PSIM or the illuminating waveguide. When comparing efficiencies across different NFTs or media stacks, it is important to use comparable meshes or grids so that the fidelity of the numerical integration is appropriate. Incidentally, since the efficiencies are on the order of a few percent, the meshes and grids also have to always be fine enough so that the numerical error is at least an order of magnitude smaller.

6.5 - Modeling the Seagate Device

At the heart of the Seagate approach to HAMR is the lollipop NFT. This is simply an optical quadrupole antenna with the target current resonance pattern shown in Fig.17. While dipole resonators respond to the local strength of electric field, quadrupole resonators respond to the local gradient of the electric field, which makes it difficult to illuminate the lollipop antenna with a conventional optical waveguide. The purpose of the PSIM is to create an electric field gradient at the focal plane of the parabola where the NFT is located. Furthermore, as any antenna, the lollipop has a particular radiation pattern which is illustrated (top view) as the transparent colored plot in Fig.18. The shape of the PSIM and the position of the NFT within it have to be optimized so that the illumination from the PSIM (shown as the blue curve in Fig.18) matches the radiation pattern of the antenna as closely as possible so as to optimize energy coupling. Note

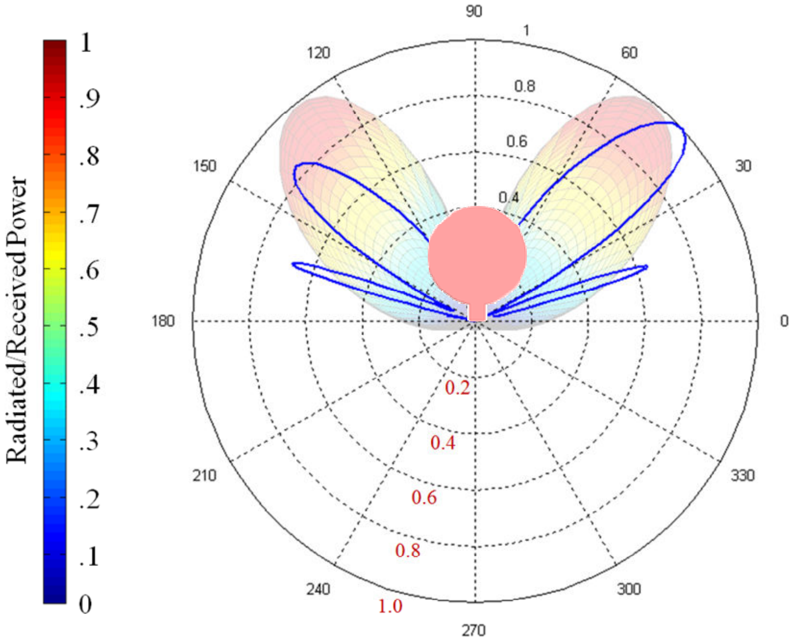


Figure 18: Top view of the radiation pattern (transparent colored) of a lollipop antenna fed at the base of the peg (an infinite recording medium plane is assumed in the lower half-space), and the illumination from the PSIM (blue curve). The two should be as close as possible in shape to optimize antenna performance.

that part of matching the radiation pattern of the antenna requires impinging light on it with the correct polarization, and this is the reason for the split-grating in Fig.10. The two arms of the grating are shifted vertically so that the left and right light paths are offset by 180° to guarantee the proper light polarization conditions at the antenna location. The linear grating can be easily designed[59] for a given incident Gaussian beam spot size and required incidence angle tolerance with a theoretical efficiency of around 50% assuming the aid of a bottom reflector. A gap of width equal to the opening of the PSIM at the ABS is placed between the two grating arms since if a grating were present there, the coupled light would not interact with the PSIM walls and partake in focusing but would rather just propagate forward and result in poor confinement at the PSIM focus. While it would intuitively seem that offsetting the two gratings by half a period would result in the desired 180° phase shift, a simple analytical analysis[60] will reveal otherwise. Ultimately however the interaction of the light with the offset gratings cannot be exactly described through a simple analytical model and we must resort to numerical techniques. The effects of grating offset on phase shift are illustrated in Fig.19 for a nominal choice of grating, revealing that neither intuition nor the simple analytical model are correct. An alternative means of achieving the desired phase shift is to use a delay line on one of the arms. The delay line would simply consist of a long grating tooth located a few microns downstream from the grating, as illustrated in Fig.20. To conclude the discussion on gratings we refer to Fig.21 which shows the impact of grating layout on the overall efficiency of the lollipop NFT. When modeling a grating numerically with FDTD it is important to guarantee that the discretization grid captures the periodicity of the grating exactly otherwise the results will be meaningless.

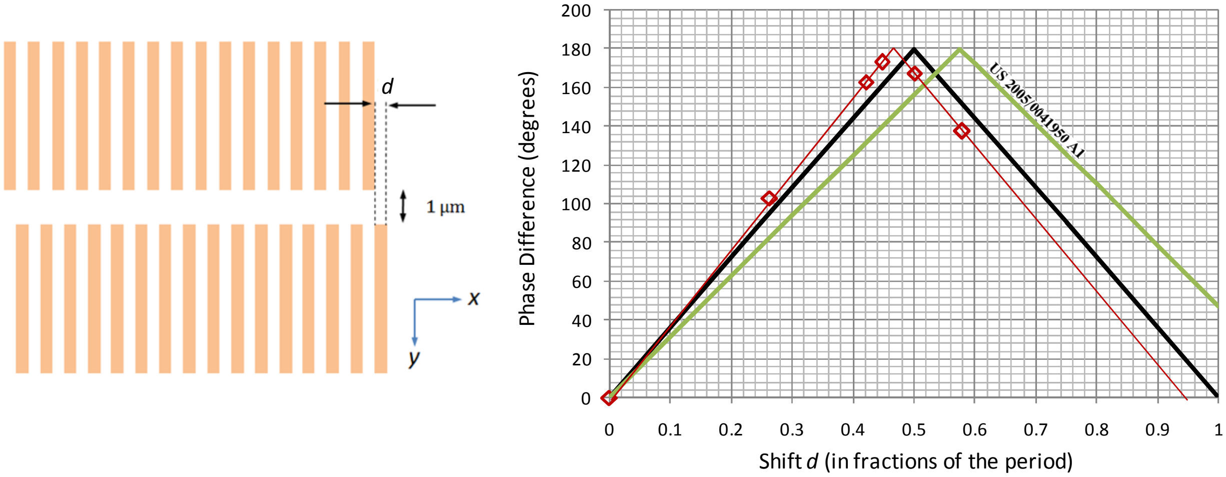
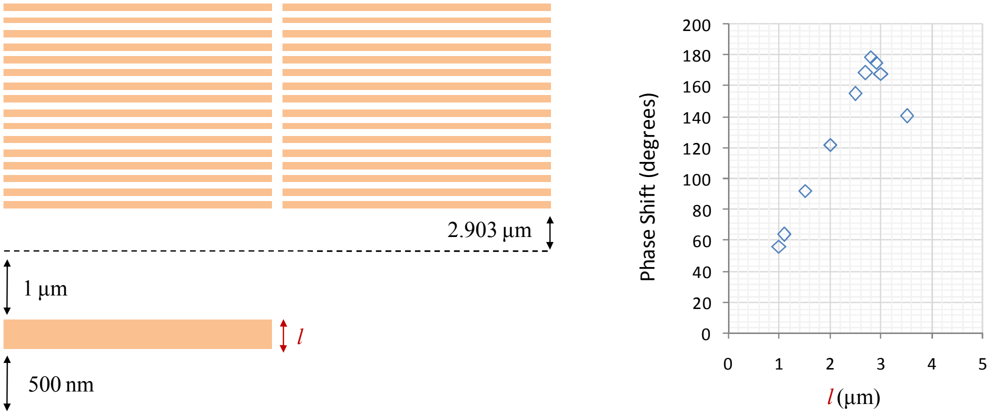


Figure 19: The effects of grating offset on phase shift according to intuition (black), back of the envelope calculations (green), and numerical modeling (red) which is the only accurate approach.

Modeling the propagation of light through the PSIM and its interaction with the metallic walls in FDTD is very time consuming because a very fine grid is required to guarantee minimal numerical dispersion over a distance of over 100 wavelengths, and one also has to resolve the skin depth of the fields in the metal (a few nanometers) over a distance of roughly 100 µm. Because these simulations are so demanding, some people in the field have sought other means of modeling the focusing action of PSIMs. Obviously beam propagation methods are not an option because the only provide intensity information, and we require vectorial information about the electric field at the focal plane to use as a source for the simulation of the NFT-media interaction. A popular approach has been to modify the method of stationary phase originally



l (µm)	Phase Shift
1.0	57°
1.5	92°
2.0	122°
2.5	155°
2.7	169°
2.8	179°
2.9	175°
3.0	168°
3.5	141°

Figure 20: A phase delay line in the form of a long grating tooth provides an alternative means of achieving the desired phase difference between the left and right arms of the grating.

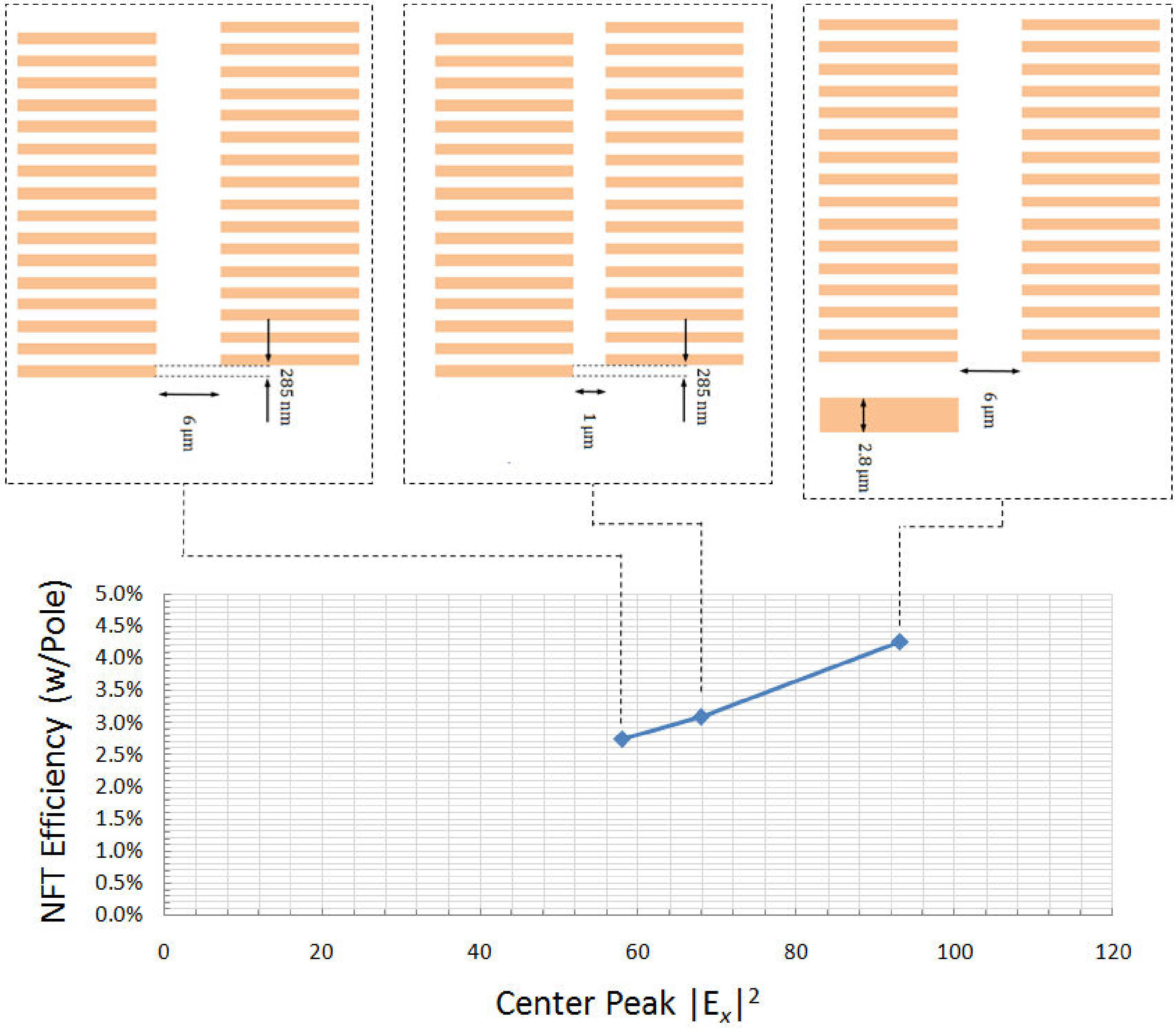


Figure 21: The impact of grating layout choices on the efficiency of the lollipop NFT.

developed by Wolf and Richards[61,62] for aplanatic systems, to obtain the electric field distribution at the focal plane of a PSIM, for an example see Fig.22. While this approach allows for quick calculation of the electric field pattern, it does not account for phase shifts at reflections with the mirror or scattering losses. Additional difficulties with diffraction of the light injected into the PSIM through the grating due to the split between the grating and the vertical offset, which are difficult to incorporate into the method of stationary phase ultimately render it useful only for rough qualitative estimates, such as the effect of incident beam spot-size on the grating, or split-grating spacing, but it cannot provide reliable enough field profiles for use in the NFT-media interaction simulation. We thus have to fall back on FDTD techniques when finalizing a PSIM design. Once the fields at the focal plane of the PSIM are known, the last segment of the device with the NFT and media stack can be modeled separately. The possibility of cascading fields from one simulation to the next allows one to quickly model different lollipop antenna variations

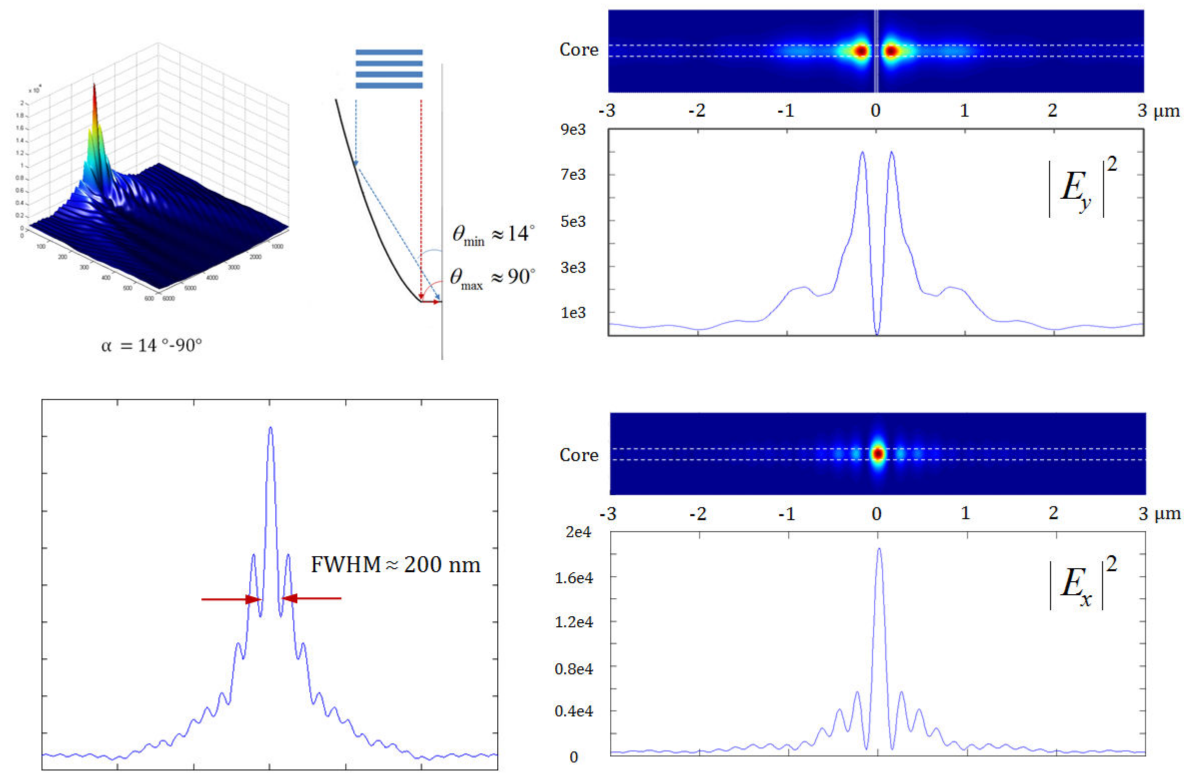


Figure 22: Example of electric field profile at the focal plane of a PASIM that can be obtained using a MATLAB implementation of the stationary phase method of Wolf and Richards.

and different media stack combination or write pole tip geometries rather quickly once the output fields from the PSIM are know. Before embarking on a parametric optimization of an NFT it is important to always check the validity of the modeling framework by first simulating structures with known behavior, such as those published in the literature, and making sure that the model recovers the experimental results (see Fig.24). For the lollipop NFT it is common to finely resolve the entire antenna and the adjacent media stack down to cells 1-2 nm on each side. The adjacent media should also be properly resolved (to say 5 nm cells) to capture the effects of side-lobes from stray light which could cause side-track erasure problems and thus must be minimized (see Fig.25). Typical shapes of the optical hot-spot in the recording media from the lollipop NFT are shown in Fig.12 for various lengths *l* of the peg. The MTO is fixed at 30 nm.

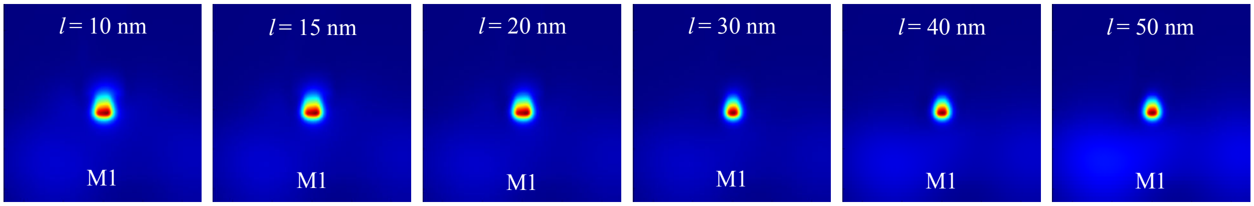


Figure 23: Typical shapes of the optical hot-spot in the recording media from the lollipop NFT for various lengths *l* of the peg. The MTO is fixed at 30 nm. The plots cover 500 nm × 500 nm.

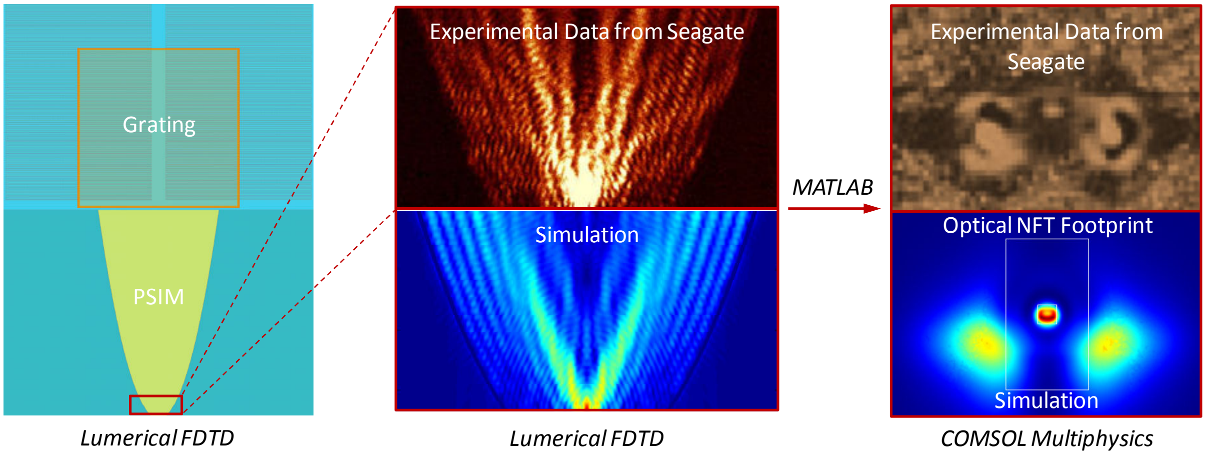


Figure 24: When working with a complicate modeling environment it is important to calibrate the model to known device structures experimental results before trying something new.

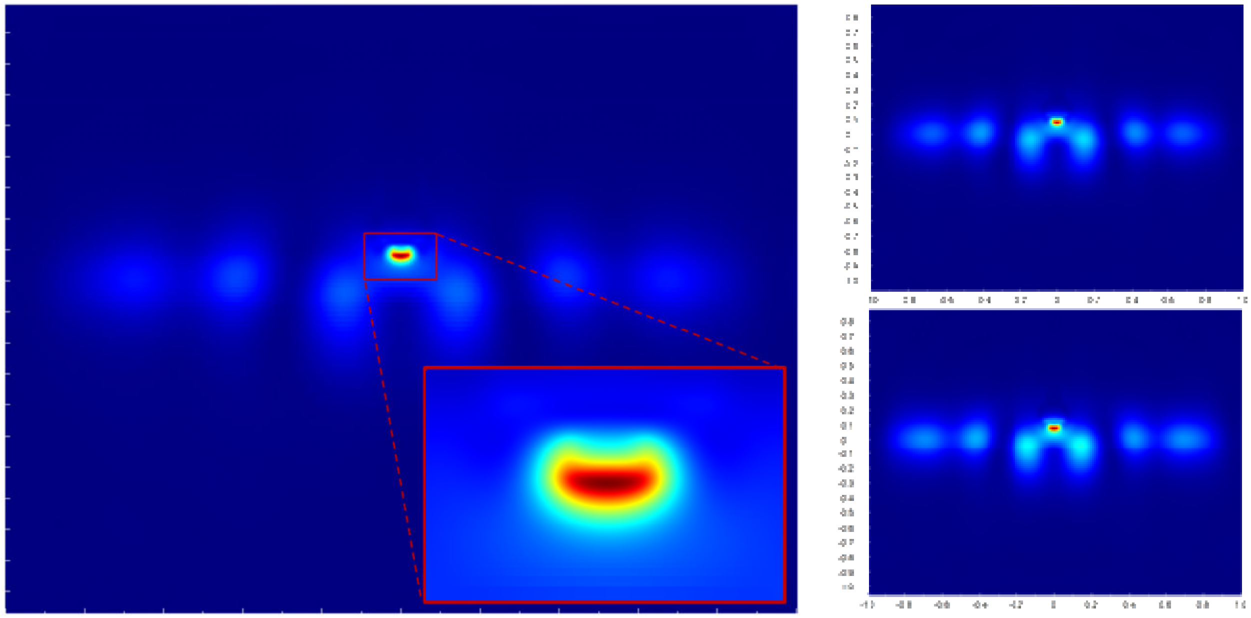


Figure 25: Main optical hot-spot in the recording media (enlarged inset to left image) and the accompanying side-lobes in the background. Poor lollipop antenna parameter choices can lead to undesirable strong side-lobes (right two images).

6.6 - Modeling the Hitachi Device

The Hitachi C-aperture is simpler to model than the Seagate lollipop antenna since the C-aperture interfaces with the first stage light condenser through a simple straight single mode dielectric waveguide. This allows us to omit the details of the first stage condenser (usually a tapered waveguide) from the model of the C-aperture design. Once a material system and waveguide cross-section dimensions are chosen, a numerical mode solver is used to extract the fields of the desired lowest order mode (in this case the TM_0-like mode shown in Fig.26) which is then used as a boundary condition in the simulation of the NFT-media interaction. The geometry of the C-aperture needs to be optimized for a given wavelength to guarantee good light throughput. The most important aspect is the thickness of the gold film in which the C-aperture is carved. Depending on the design details, the C-aperture itself may need to be back-filled with the dielectric used for the core of the feeding dielectric waveguide.

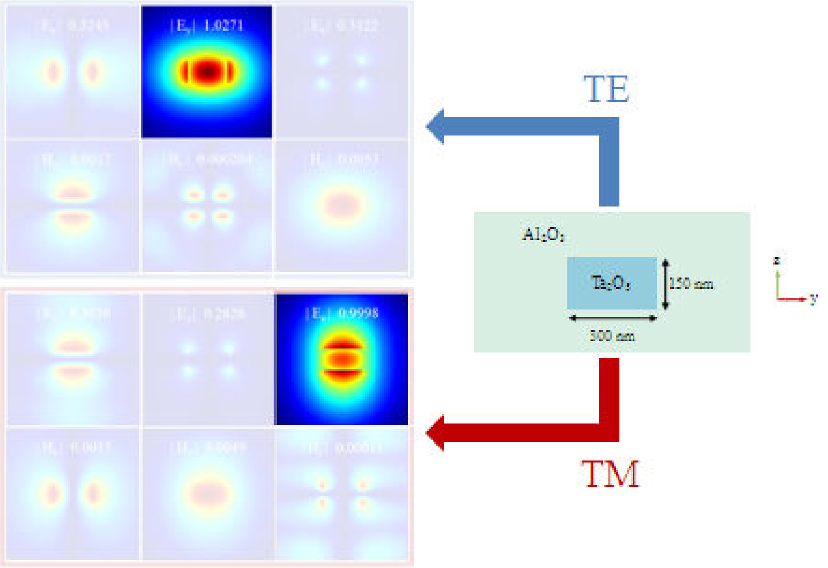


Figure 26: The waveguide that interface with the C-aperture supports both TE-like and TM-like modes. The C-aperture in a HAMR configuration requires TM illumination.

The detailed Hitachi design is illustrated in Fig.27. It features an abnormally large feeding waveguide and an unconventional gap between the feeding waveguide and the C-aperture for improved mode matching. The large waveguide was required by Hitachi in order to easily butt-couple light into it on a spin stand, but different illumination schemes can accommodate other waveguide designs, so large waveguide dimensions are not a strict requirement for the C-aperture NFT. Two problems with the C-aperture NFT that were not manifest in the Seagate design are that aperture designs are easy to fabricate using e-beam lithography and FIB milling, but are more difficult to mass produce using optical lithography. Also the feeding waveguide positioning in relation to the C-aperture requires a redesign[50] of the conventional magnetic write pole, and the availability of reliable high power (50-100 mW) TM polarized laser diodes, or some polarization rotating scheme implemented in passive integrated optics, neither of which are available today. Typical shapes of the optical hot-spot in the recording media from the C-aperture NFT are shown in Fig.28 for various thicknesses l of the gold film. The MTO is fixed at 30 nm.

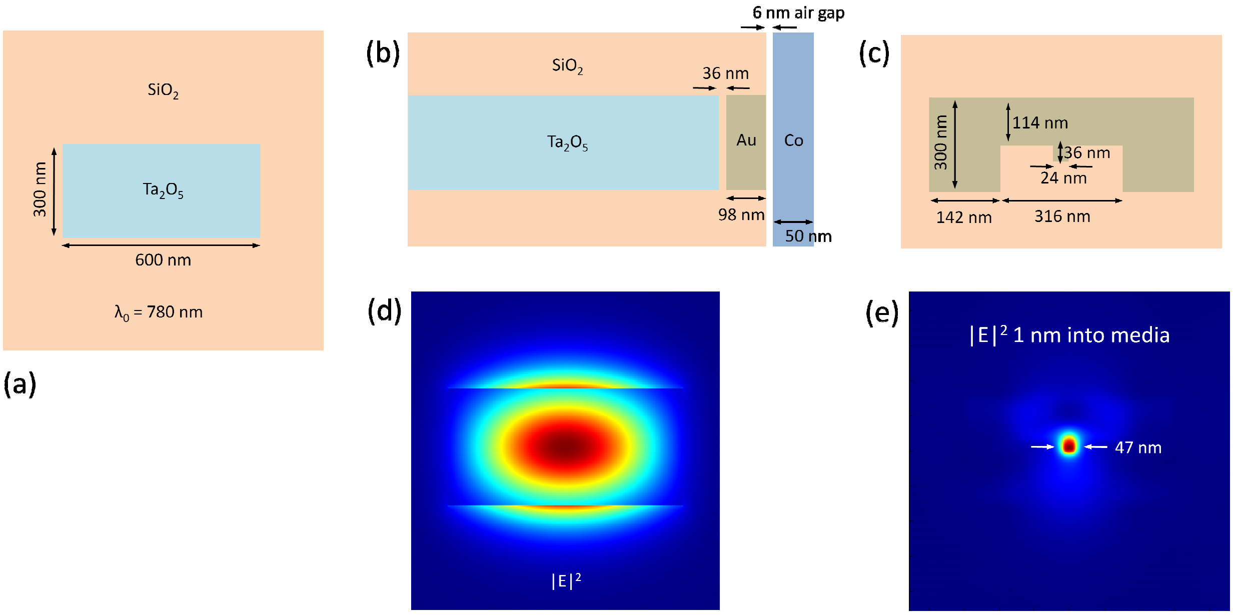


Figure 27: Schematic of the Hitachi C-aperture NFT design. (a) Transverse cross-section of the feeding dielectric waveguide and corresponding lowest mode TM-like excitation mode profile (d). (b) Longitudinal cross-section of waveguide/C-aperture interface. (c) Geometry of the C-aperture and corresponding resistive heating profile in the media (e).

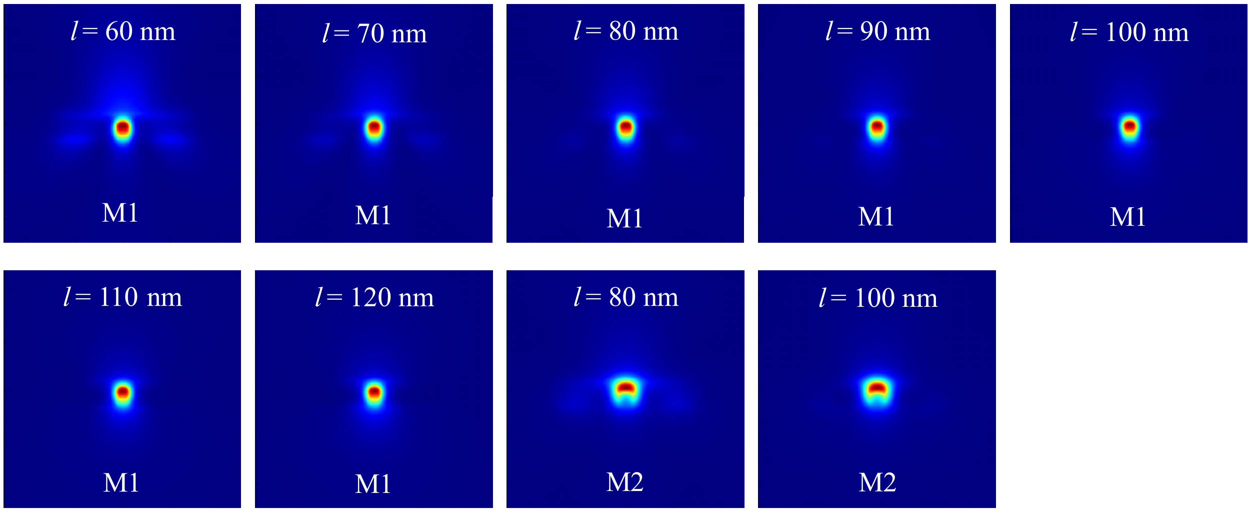


Figure 28: Typical shapes of the optical hot-spot in the recording media from the C-aperture NFT for various thicknesses l of the gold film. The MTO is fixed at 30 nm. Two types of media (denoted M1 and M2) are considered. The plots cover 500 nm 500 nm.

6.7 - Modeling the Impedance Matching Device

The impedance matching NFT follows naturally from the discussions in the first part of the dissertation and comprises a gold parallel plate waveguide of plate spacing 150 nm and width 300 nm which is tapered in both directions (~20° angle) to a smaller parallel plate waveguide with plate spacing 10 nm and plate width 20 nm. The latter parameter determines the cross-track FWHM of the optical spot in the media. Terminal waveguide dimensions can be made smaller so long as the width is always ≥ than the plate separation. The strong light confinement from the plasmonic parallel plate waveguide guarantees very good control over the optical spot size in the media. The purpose of the taper is to boost the wave impedance to try and match the impedance of the target magnetic grains in the media[55]. The impedance matching NFT is fed by a TM polarized single mode waveguide like the one used for the C-aperture. The core height and width need to be the same as the plate spacing and width of the plasmonic waveguide. The dielectric between the metallic plates in the plasmonic waveguide also has to be the same as that comprising the feeding waveguide core. The geometry of the impedance matching NFT is illustrated in Fig.23. The plasmonic mode at the narrow end of the NFT has a skin depth in the metal on the order of 10 nm for the configuration described here. This means that beyond 20 nm or so from the gold/dielectric interface, we can replace the gold with the magnetic write pole without affecting the propagation of the plasmonic mode and its light confinement, as illustrated in Fig.31. Typical shapes of the optical hot-spot in the recording media from the impedance matching NFT are shown in Fig.32 for various lengths *l* of the narrow plasmonic waveguide at the snout of the taper. Two types of media (M1 and M2) are considered. The MTO is 30 nm.

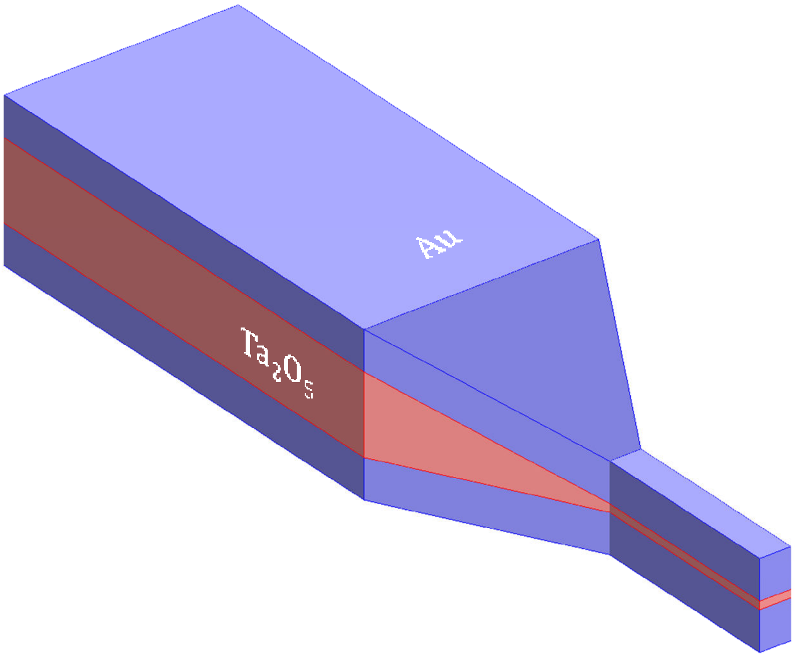


Figure 29: Geometry of the impedance matching NFT.

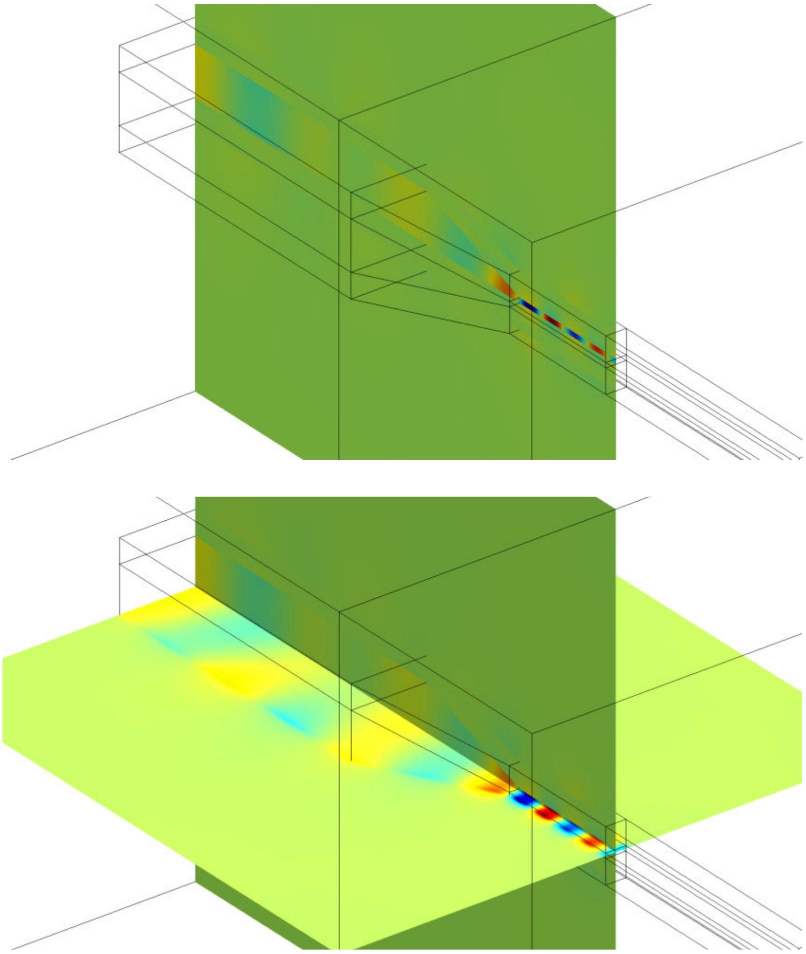


Figure 30: (top) Vertical slice of the electric field, and (bottom) vertical + horizontal slices of the electric field, illustrating the light confinement.

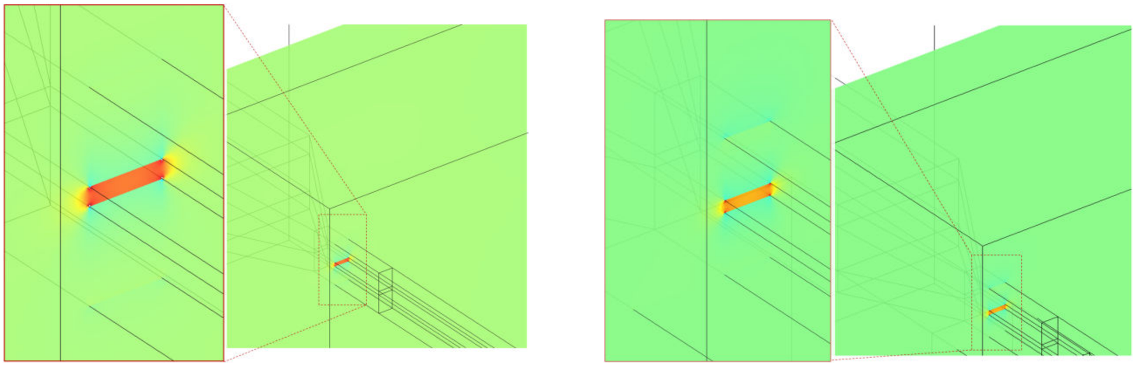


Figure 31: The confined field in the presence of a magnetic write pole with MTO = 30 nm (right) and the confined field in the absence of the write pole (left) are very similar since the device geometry is not sensitive to write pole proximity.

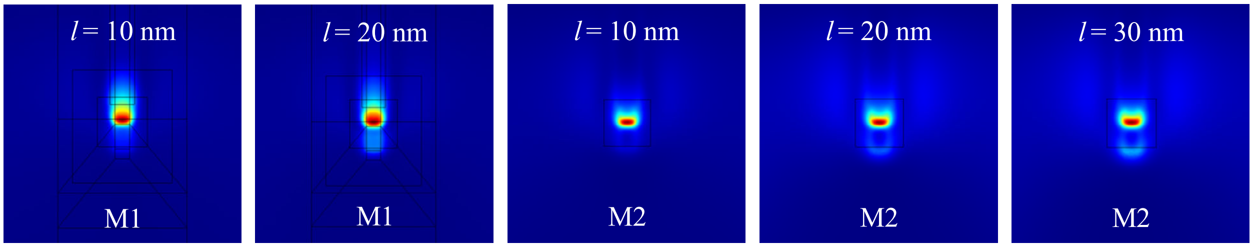


Figure 32: Typical shapes of the optical hot-spot in the recording media from the impedance matching NFT for various lengths l of the narrow plasmonic waveguide at the snout of the taper. Two types of media (denoted M1 and M2) are considered. The plots cover 500 nm × 500 nm.

6.8 - Modeling Results

We evaluate the performance of the various NFT designs by choosing a media stack (see Fig.33) and comparing *(i)* the cross-track FWHM of the optical hot-spot for a given minimum physical dimension, and *(ii)* the efficiency of power coupling from the NFT to the recording media. Results pertaining to the cross-track of the optical spot are presented by plotting the FWHM of the resistive heating profile versus the peg width (for the lollipop antenna), ridge width (for the C-aperture), and snout width (for the impedance matching NFT). Results pertaining to the NFT-media efficiency are presented by plotting the fraction of power reaching an NFT which is absorbed in a 100×100×8 nm^3 volume in the recording layer immediately adjacent the NFT, versus the length of the peg (for the lollipop antenna), thickness of the gold film (for the C-aperture), and length of the snout (for the impedance matching NFT). The reason for plotting efficiency versus peg/ridge/snout length is that the ABS plane is defined using mechanical polishing with an electronic lapping guide that has an under-lap over-lap error of about 10 nm, so that the sensitivity of the efficiency about the optimal peg/ridge/snout length is of interest and should be as small as possible to guarantee good device yield. Some results for the various NFTs are shown in Fig.28 where the MTO was set to 30 nm, the media stack had no diamond like carbon (DLC) layer, and the recording layer was taken as bulk FePt with optical properties $n = 3.3, k = 4.3$.

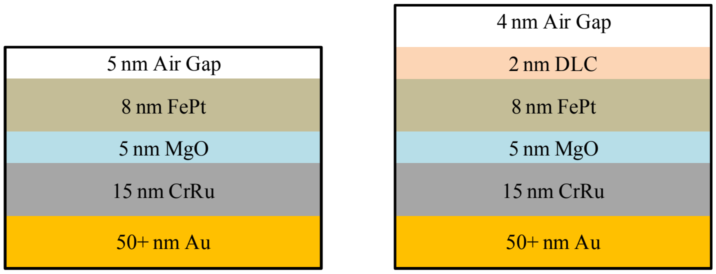


Figure 33: Typical media stacks used in modeling of the NFTs.

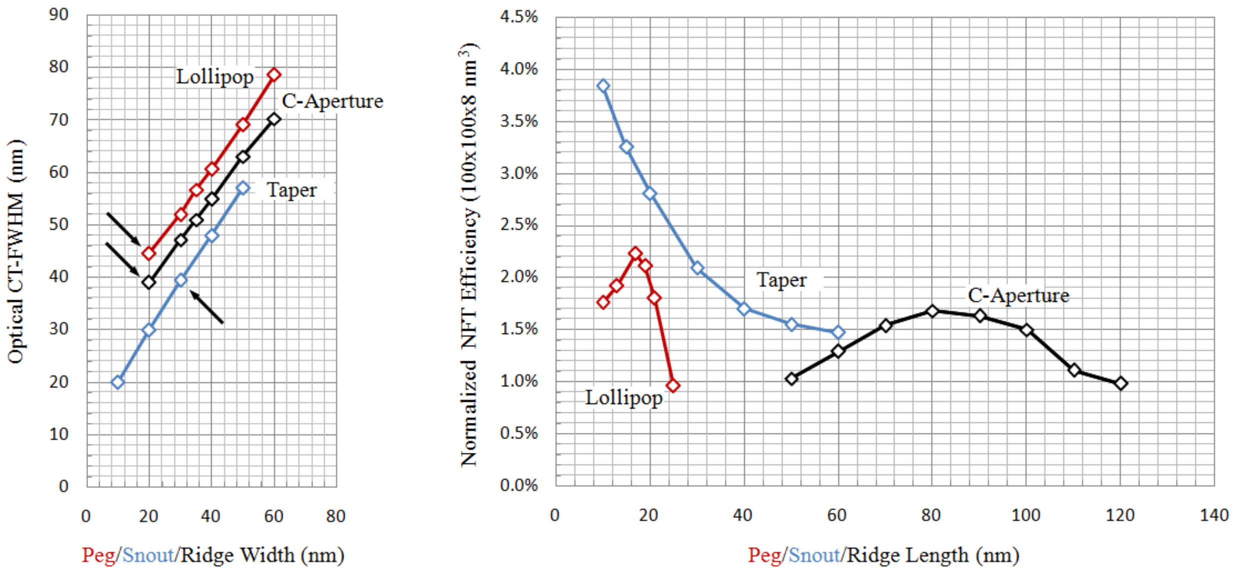


Figure 34: Some modeling results showing cross-track FWHM behavior (left) and NFT-media efficiency (right) for various NFTs. The impedance matching NFT is denoted as the taper NFT. Arrows point to the device widths used in the efficiency calculations.

Evidently the optical spot in the media is always wider than the minimum physical dimension of the NFT. The largest discrepancy is found for the lollipop antenna, which is not surprising since it provides the least confinement of all the NFTs considered here. The smallest discrepancy is found for the impedance matching NFT which naturally provides the most control over light confinement. The C-aperture optical spot cross-track is found in between those of the other two NFTs. In terms of invariance of efficiency with respect to polishing error the C-aperture performs the best, however its peak efficiency is the lowest out of the three NFTs. The impedance matching NFT has the second best tolerance to polishing error as well as the highest efficiency. The efficiency would be higher were it not for poor coupling between the feeding dielectric waveguide and the wide plasmonic waveguide, which is only roughly 20%. This could be improved by introducing an additional matching network as will be discussed later. The lollipop NFT is the most sensitive to polishing error and the lowest efficiency. The low efficiency is partly due to the fact that the base of the magnetic write pole is only 30 nm above the antenna, thus deteriorating its performance considerably. As the write pole is backed away the peak efficiency is found to rise to about 4%, but in practice the write pole will have to be present so this latter figure is of little significance. Particularly troubling for the lollipop antenna is that its peak efficiency corresponds to a peg length of ~20 nm, which means that if the sliders are over-lapped during fabrication, there is a good chance that the peg will be completely gone, and the NFT won't be at all functional. In the case of the other two types of NFT, polishing errors only result in diminished performance instead of dud bars. Despite its drawbacks, the lollipop antenna remains a fairly attractive NFT since it is the easiest to fabricate out of all the devices considered here.

6.9 - The Impact of Media Properties on Modeling Results

Consider the media stack shown in the left panel of Fig.33. It is conventional to assume the optical properties of bulk FePt ($n = 3.3, k = 4.3$) for the recording layer in the electromagnetic model. However, the recording layer (RL) actually consists of grains of FePt roughly 10 nm in diameter and 8 nm tall, embedded in a hard dielectric matrix of carbon or some other high-k dielectric. When Ellipsometry experiments are performed on the recording layer, it is found that for grains 10 nm in diameter, the effective optical properties are $n = 2.78, k = 1.76$. Compared to the values for bulk FePt which correspond to a metallic medium ($\varepsilon' = n^2 - k^2 < 0$), this media actually behaves macroscopically as a dielectric ($\varepsilon' = n^2 - k^2 > 0$). This means that using the Ellipsometry data in the model will remove any plasmonic interaction between the recording layer and the NFT which was present before. This result in reduced spreading of the optical fields in the recording layer, and consequently in a smaller cross-track of the optical spot size, which becomes comparable for all NFTs given they share the same minimum dimension, as illustrated in Fig.35.

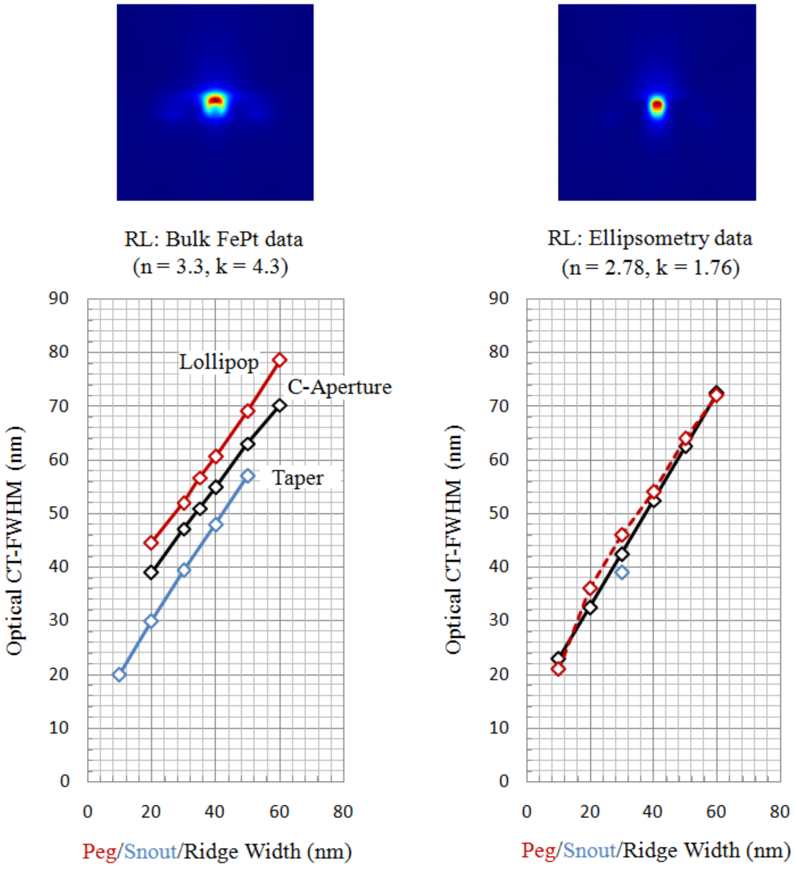


Figure 35: Cross-tracks of optical spot size in the media assuming bulk FePt (left) and Ellipsometry data (right) for the material layer in the model. Different choices of media models result in different optical profiles, as illustrated for a C-aperture in the top two panels.

The use of Ellipsometry data for the recording layer instead of values for bulk FePt also affects the NFT-media efficiency. Fig.36 show the efficiency trends for both media models as a function of the air gap between the NFT and the top of the media stack (assuming a 2 nm layer of DLC above the recording layer) for a lollipop antenna. The results obtained using Ellipsometry data are always better than those obtained assuming bulk values of FePt.

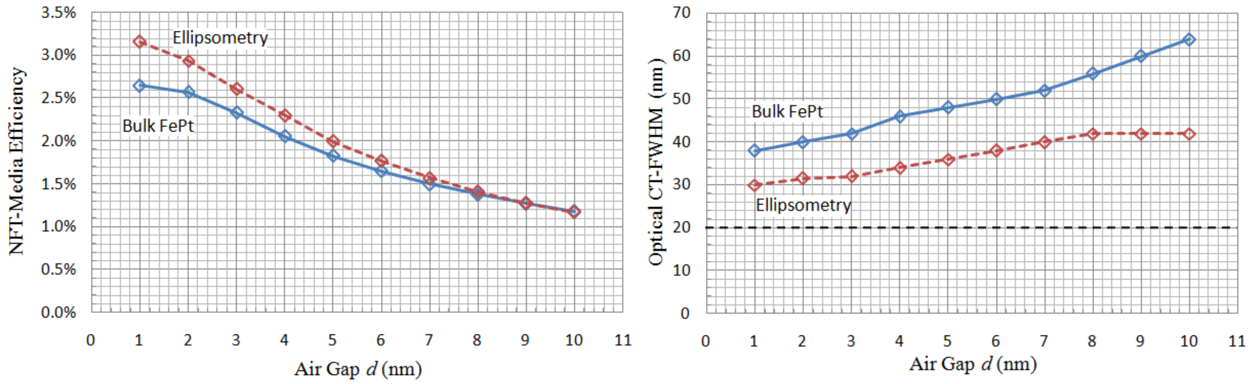


Figure 36: NFT-media efficiency (left) and cross-track of optical hot-spot in the media (right) for a lollipop antenna assuming Ellipsometry data and bulk FePt data for the recording layer. The black dashed line in the right figure represents the width of the lollipop peg.

Since the portion of the NFT that interfaces with the media is only on the order of 20x20 nm^2, it will interact with just a handful of grains, and will see the metallic nature of the recording layer. This suggests that using Ellipsometry data to capture the optical properties of the recording layer is not a good idea. To better capture the real situation in our simulations, we replace the region of the recording layer immediately adjacent to the NFT with a honeycomb of grains with a pitch of 4 nm and grain-to-grain spacing of 1 nm, corresponding to the media that will be required for 4 Tb/in^2 recording. This is illustrated in see Fig.37 for a C-aperture NFT.

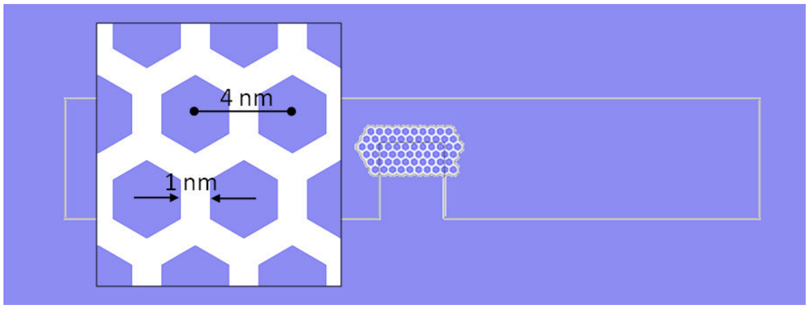


Figure 37: Discrete media model used in electromagnetic modeling of C-aperture. The grain structure of the recording layer is approximated by a honeycomb arrangement of FePt grains.

In the modeling we let the grains have the optical properties of bulk FePt since although small, the grain dimensions are quite large compared to the electron mean-free-path in an alloy, and thus surface effects can be neglected[63]. The results of the discrete media simulations are summarized in Fig.38 for a C-aperture NFT.

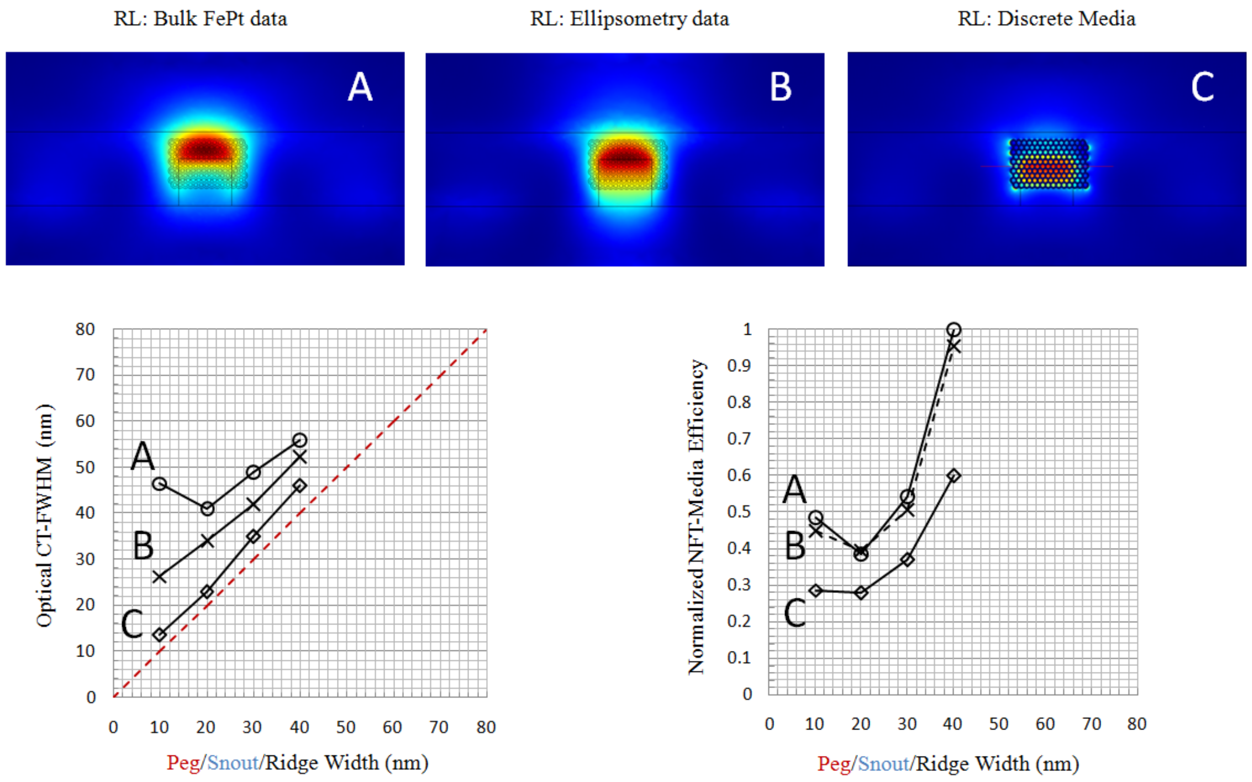


Figure 38: Comparison of optical hot-spot cross-track and NFT-media efficiency for recording layers comprising bulk FePt (A), media consistent with Ellipsometry data (B), and discrete FePt grains in a carbon matrix (C). The red dashed line shows the ideal cross-track FWHM. The NFT-media efficiency has been normalized to the largest value obtained for a ridge width of 40 nm.

We see that a recording layer of bulk FePt gives the broadest cross-track, and the highest efficiency. A recording layer consistent with Ellipsometry data (FePt grains of 10 nm diameter) gives a better cross-track and efficiency comparable to that of the bulk FePt case. The discrete media model (FePt grains of 4 nm diameter) gives the best cross-track, almost ideal, with only the grains immediately in front of the C-aperture ridge absorbing power from the NFT. Because the fill factor (FePt/area in media) is smaller than in the previous two cases, the efficiency is also smaller, at around 60% the value for the other two cases. This suggests that the cross-track optical blooming reported in the literature is an artifact of the simplistic media model which ignores the granular structure of the recording layer, and that the actual cross-track will be limited only by how closely the grains can approximate the profile of the minimum physical parameter of the NFT (i.e. the ridge with in the case of a C-aperture NFT). Smaller grains will thus always guarantee a smaller cross-track for a given NFT. Additionally we note that the broadening of the cross-track and corresponding boost in efficiency at low ridge widths observed in conventional recording layer models also vanish when using a discrete recording layer model, suggesting that smaller cross-tracks than previously anticipated may be possible. Similar results are obtained for the Seagate lollipop NFT as illustrated in Fig.39 below.

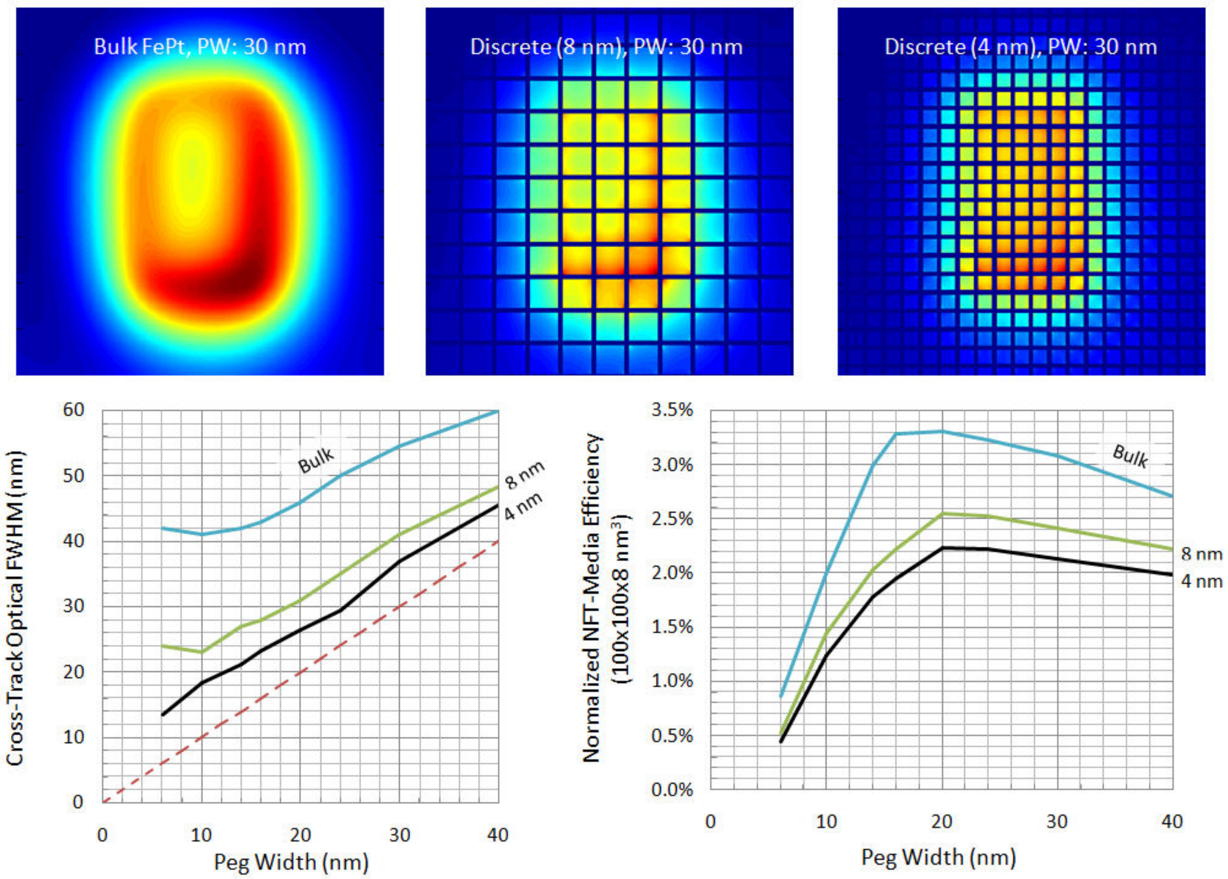
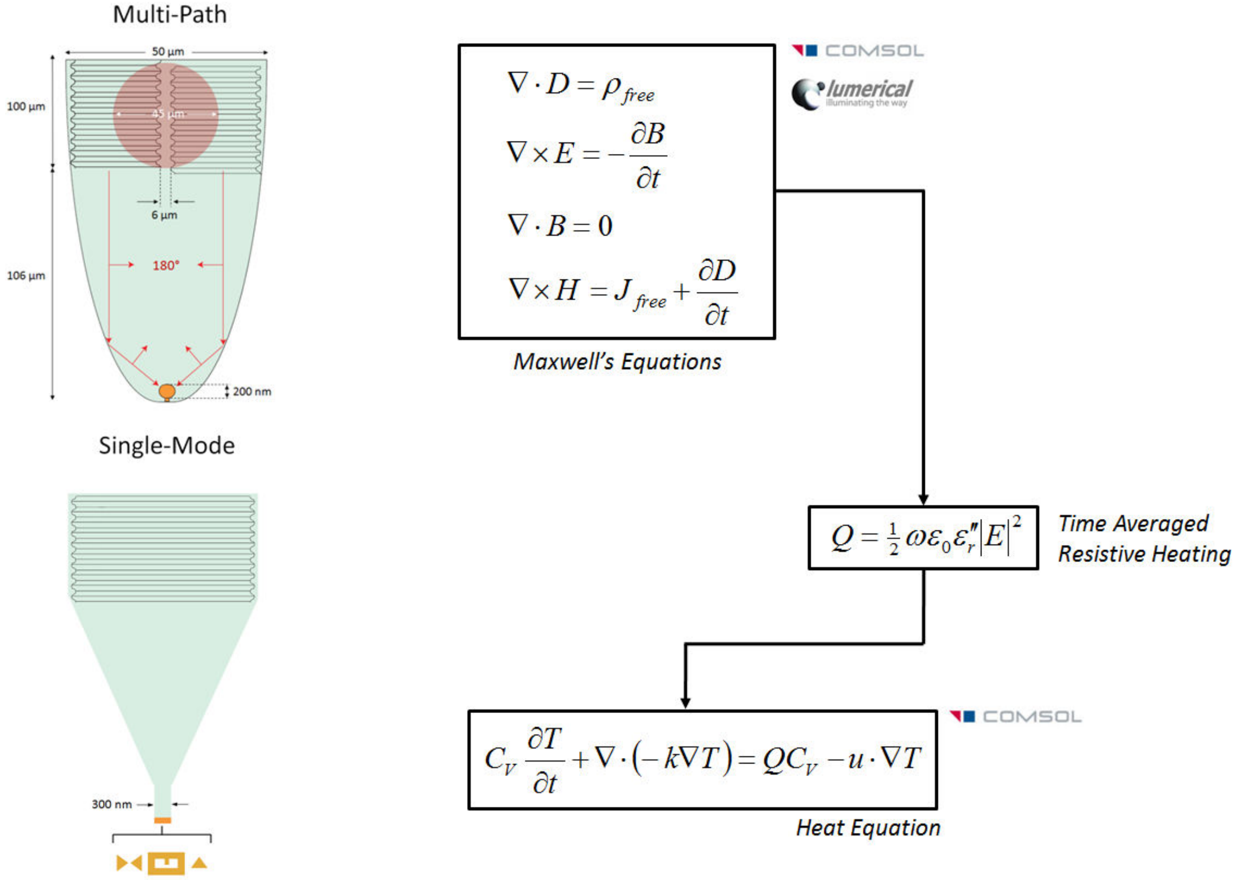


Figure 39: Comparison of optical hot-spot cross-track and NFT-media efficiency for recording layers comprising bulk FePt (left), and discrete FePt grains in a carbon matrix (center and right), for the lollipop NFT. The red dashed line in the left graph shows the ideal cross-track FWHM.

6.10 - Coupled Electromagnetic–Thermal Modeling

Ultimately the goal of HAMR modeling is to simulate heating the recording layer in the media stack. The electromagnetic simulations provide minimum optical cross-tracks and coupling efficiency, but ultimately the written cross-track will be broadened by thermal diffusion and the absolute power required for recording will be fixed by the thermal properties and interaction of the media stack and the NFT/slider system. To this end, we extract the resistive (Joule) heating in the media stack and NFT from the electromagnetic simulation and use it as the volumetric heat source in a thermal simulation to obtain the thermal profile in the recording media layer, and the peak temperatures in the NFT and recording layer for a given input laser power (details of the thermal problem are presented in Appendix K). This information is then handed over to Western Digital, Seagate, and Hitachi engineers who utilize it in their proprietary in-house modified Landau Lifshitz Gilber (LLG) micromagnetic code to simulate the writing process at the level of individual grains. Our framework for coupling the electromagnetic and thermal problems is summarized in Fig.40 below. NFTs that require single-mode illumination are the easiest to deal



Multi-Path
50 µm
100 µm
45 µm
6 µm
106 µm
180°
200 nm
Single-Mode
300 nm
COMSOL
lumerical
illuminating the way
$\nabla \cdot D = \rho_{free}$
$\nabla \times E = -\frac{\partial B}{\partial t}$
$\nabla \cdot B = 0$
$\nabla \times H = J_{free} + \frac{\partial D}{\partial t}$
Maxwell's Equations
$Q = \frac{1}{2}\omega\varepsilon_0\varepsilon_r''|E|^2$
Time Averaged Resistive Heating
$C_V \frac{\partial T}{\partial t} + \nabla \cdot (-k\nabla T) = QC_V - u \cdot \nabla T$
COMSOL
Heat Equation

Figure 40: Schematic summary of the electromagnetic-thermal modeling problem coupling.

with because as mentioned earlier, one only needs to model the NFT-media region, which can be separated from the first light collimation segment of the structure. The NFT-media region for single-mode illumination devices is usually sufficiently small that the problem can be solved with FEM in *COMSOL*, which seamlessly allows for coupling of the Maxwell equations and the heat equation as part of its proprietary *Multiphysics* modeling framework. One can thus very quickly solve both the electromagnetic problem and the thermal problem on the same grid, without incurring any additional numerical error that would arise from grid-to-grid mapping of the volumetric resistive heating profile. Relative motion between the NFT-slider assembly and the media stack can be accounted for through convective boundary conditions (see Appendix K) and one can solve for both steady-state temperature profiles in the media (conventionally used in LLG codes) and transient temperature profiles to model ramp up and ramp down at the beginning and end of recorded tracks. NFTs that require multipath illumination are more difficult to handle because the first stage of collimation cannot be isolated from the NFT-media interaction region. As mentioned earlier, these structures are too large to model with FEM and must be simulated using FDTD. One can however arrange to extract the fields in the vicinity of the NFT-media interaction region and map them to an equivalent problem in a FEM framework. Parameter tweaking and careful tuning of boundary conditions in the FEM code allows one to usually reproduce the field distribution from FDTD at the NFT and media in the FEM package.

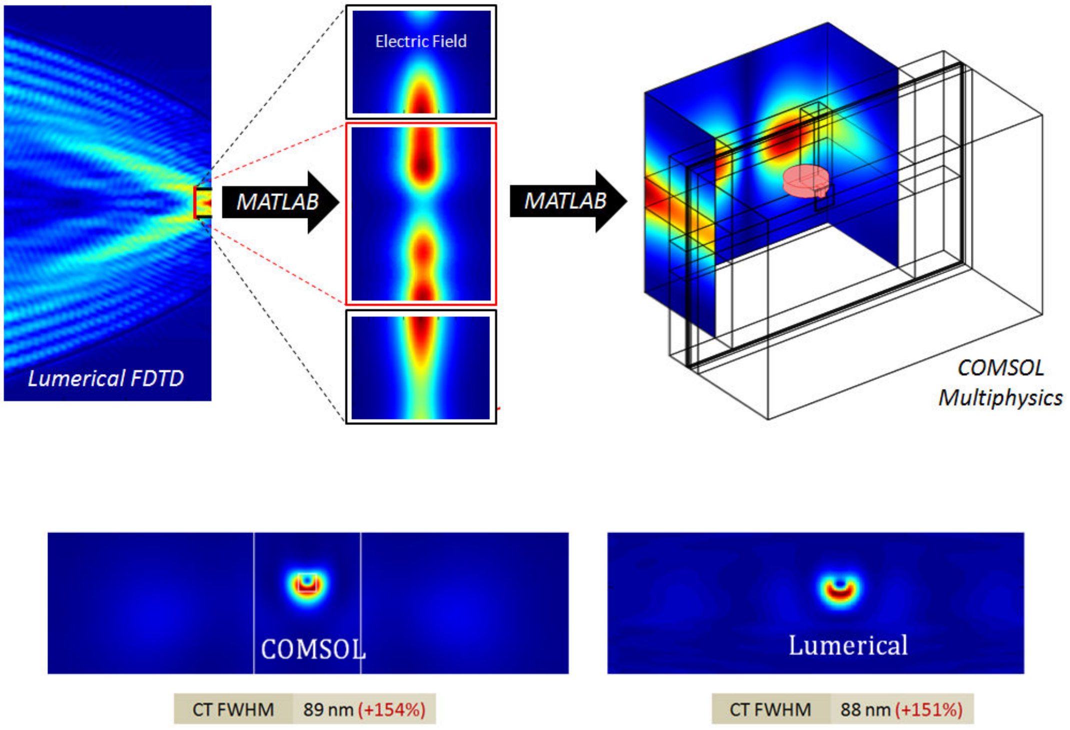


Figure 41: Mapping scheme to couple electromagnetic and thermal problems for multipath illumination NFTs like the lollipop antenna.

This allows one to model the electromagnetic problem in *Lumerical FDTD* and port it over to *COMSOL Multiphysics* for thermal modeling with FEM, as illustrated graphically in Fig.41.

6.11 - Conclusion

Over our two year collaboration with Western Digital Corporation and INSIC we have had the opportunity to leverage insights from the circuit analysis for metal-optics, presented in the first half of the dissertation, to impact design of optical antennas for use in heat assisted magnetic recording (HAMR). While the particular details of our collaboration were not discussed in this dissertation due to nondisclosure agreements, we have presented the modeling framework that we developed to accommodate the unconventional needs of the HAMR problem, and we have illustrated application of the required modeling principles and model flow for a handful of antenna designs that have appeared in the literature and that are being currently studied by other hard-disk drive (HDD) companies. We pointed out shortcomings of these designs, and proposed an alternate approach based on an optical voltage transformer which outperforms all existing optical antennas even before optimization of its first stage light collimation system. We concluded by mentioning some special considerations regarding the media stack models pertaining to the effect of media granularity on optical modeling results. Following our introduction of the granular media model at a pre-competitive INSIC review in 2010 it has been adopted by the top HDD companies as the standard for optical calculations, replacing the earlier and inadequate bulk recording medium layer standard that had been in place for over ten years prior.

Appendix A – Derivation of Surface Plasmon Dispersion Relation

Here we concern ourselves with surface plasmon modes at the interface between a semi-infinite metal and a semi-infinite dielectric, as illustrated in Fig.A1. It can be shown that surface plasmon modes are transverse magnetic in nature and arise at metal-dielectric interfaces where no independent charge or current densities exist. Taking this to be the case, and further assuming that all media are isotropic and display no intrinsic magnetic or electric polarization in the spectral region of interest, we may express the electric displacement, D, and the magnetic induction, B, simply as

$$D = \varepsilon E = \varepsilon_{i,m}\varepsilon_o E, \quad \text{and} \quad B = \mu H = \mu_o H,$$

where $\varepsilon = \varepsilon_{i,m}\varepsilon_0$ and $\mu = \mu_0$ are scalars, and the latter equality holds since we are working at optical frequencies. With the aforementioned assumptions and definitions, the general Maxwell equations in matter

$$\nabla \times E = -\,\partial B/\partial t$$

$$\nabla \times H = J + \partial D/\partial t$$

$$\nabla \cdot D = \rho \ , \nabla \cdot B = 0$$

simplify to

$$\nabla \times E = -\mu_o\, \partial H/\partial t$$

$$\nabla \times H = \varepsilon_{i,m}\varepsilon_o\, \partial E/\partial t$$

$$\nabla \cdot E = 0 \ , \ \nabla \cdot H = 0$$

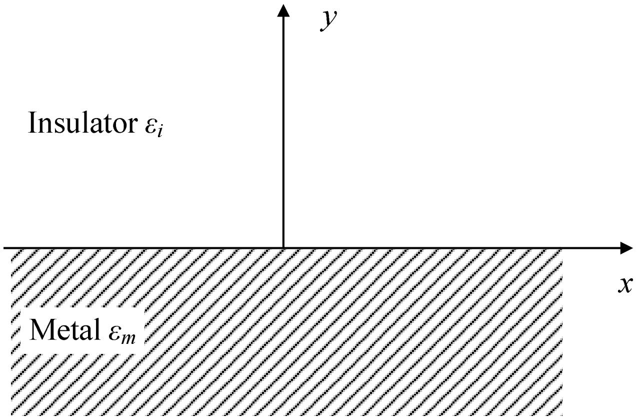


Figure A1: Reference coordinate system for discussions pertaining to surface plasmons at a single metal- insulator interface. Both metal and insulator are assumed to be semi-infinite in extent.

where, in Cartesian coordinates, the electric field vector, E, is defined as

$$E = E_x\hat{x} + E_y\hat{y} + E_z\hat{z}\ ,$$

and, in general, the field component amplitudes are themselves functions of space and time

$$E_i = E_i(x, y, z, t) \quad \text{for} \quad i = x, y, z.$$

In order to proceed further, we assume that, for a surface plasmon traveling in the positive x direction, referring to the coordinate system illustrated in Fig.1, the field component amplitudes take the form

$$E_i(x, y, z, t) = E_i(y)Re\{e^{j(kx-\omega t)}\}\ ,$$

where k is the propagation constant, $E_i(y)$ describes the spatial evolution of the ith's field component amplitude in the y direction, and no amplitude variation is considered in the z direction. The same applies for the rest of the vector quantities that appear in Maxwell's equation, polar and axial alike. Then, Faraday's Law, $\nabla \times E = -\mu_0\, \partial H/\partial t$, becomes

$$det\begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} = \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\hat{x} + \left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right)\hat{y} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\hat{z}$$

$$= j\omega\mu_o\left(H_x\hat{x} + H_y\hat{y} + H_z\hat{z}\right)$$

$$\rightarrow\ \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = j\omega\mu_o H_z$$

$$\rightarrow\ jkE_y - \frac{\partial E_x}{\partial y} = j\omega\mu_o H_z\ , \qquad (A.1)$$

where we used the fact that $\partial/\partial z \rightarrow 0$, and that for TM waves propagating along the x direction, the only nonzero field component amplitudes are $E_{x,y}$ and H_z. Similarly, Ampere's Law, $\nabla \times H = \varepsilon_{i,m}\varepsilon_o\, \partial E/\partial t$, becomes

$$det\begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ H_x & H_y & H_z \end{vmatrix} = \left(\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}\right)\hat{x} + \left(\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}\right)\hat{y} + \left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right)\hat{z}$$

$$= -j\omega\varepsilon_{i,m}\varepsilon_o\left(E_x\hat{x} + E_y\hat{y} + E_z\hat{z}\right)$$

$$\rightarrow \ \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = -j\omega\varepsilon_{i,m}\varepsilon_o E_x$$

$$\rightarrow \ \frac{\partial H_z}{\partial y} = -j\omega\varepsilon_{i,m}\varepsilon_o E_x \, , \qquad (A.2)$$

and

$$\rightarrow \ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = -j\omega\varepsilon_{i,m}\varepsilon_o E_y$$

$$\rightarrow \ jkH_z = j\omega\varepsilon_{i,m}\varepsilon_0 E_y \, . \qquad (A.3)$$

Next, we use relations $(A.2,3)$ to express H_z and E_x in terms of E_y:

from $(A.3)$ $\quad jkH_z = j\omega\varepsilon_{i,m}\varepsilon_o E_y \ \rightarrow \ H_z = \left(\omega\varepsilon_{i,m}\varepsilon_o/k\right)E_y \, , \qquad (A.4)$

from $(A.4)$ $\quad \partial H_z/\partial y = -j\omega\varepsilon_{i,m}\varepsilon_o E_x \ \rightarrow \ E_x = (j/k)\, \partial E_y/\partial y. \qquad (A.5)$

By substituting $(A.4)$ and $(A.5)$ into expression $(A.1)$ we obtain a second order linear partial differential equation in E_y:

$$jkE_y - \frac{\partial E_x}{\partial y} = j\omega\mu_o H_z \quad \rightarrow \quad jkE_y - \frac{\partial^2 E_y}{\partial y^2} = \frac{j\omega^2\varepsilon_{i,m}}{c^2 k} E_y$$

$$\rightarrow \ \left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_{i,m}\right) E_y = \frac{\partial^2 E_y}{\partial y^2} \, ,$$

where we used the relation $c = 1/\sqrt{\varepsilon_o\mu_o}$. To recapitulate, we have derived a partial differential equation that defines E_y implicitly, and we have equations that relate H_z and E_x to E_y. We thus only need to solve the partial differential equation for E_y in order to fully characterize the surface plasmon modes. To proceed, we define

$$K_{i,m} = \left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_{i,m}\right)^{1/2} , \qquad (A.6)$$

so that the partial differential equation takes the more tractable form

$$K_{i,m}^2 = \frac{\partial^2 E_y}{\partial y^2} \, .$$

It is easy then to see that the solution

$$E_y = Ae^{K_m y} \ \text{ for } \ y < 0 \, , \ E_y = Be^{-K_i y} \ \text{ for } \ y > 0$$

(where A and B are constants to be fixed by the boundary conditions) satisfies the differential equation and the physical requirement that all fields vanish at infinity. Next, we use the boundary conditions at the metal-dielectric interface to fix A and B. At the interface (which corresponds to the $y = 0$ plane), E_x, H_z and D_y have to be continuous. The continuity of D_y dictates that

$$\varepsilon_i B = \varepsilon_m A \qquad \rightarrow \qquad A = (\varepsilon_i/\varepsilon_m)B. \qquad (A.7)$$

The continuity of E_x dictates that

$$\frac{AK_m}{jk} = -\frac{BK_i}{jk} \qquad \rightarrow \qquad A = -B\frac{K_i}{K_m}. \qquad (A.8)$$

It is convenient then to define $B = 1$, $A = -K_i/K_m$. Then, by combining $(A.7{,}8)$, we find

$$-\frac{K_i}{K_m} = \frac{\varepsilon_i}{\varepsilon_m} \quad \rightarrow \quad \frac{K_i/\varepsilon_i}{K_m/\varepsilon_m} = -1$$

which, upon substitution of $(A.6)$ for $K_{i,m}$, yields the dispersion relation for surface plasmons at a single metal-dielectric interface:

$$\frac{K_i/\varepsilon_i}{K_m/\varepsilon_m} = -1 \qquad \rightarrow \qquad \frac{\left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_i\right)^{1/2} / \varepsilon_i}{\left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_m\right)^{1/2} / \varepsilon_m} = -1$$

$$\frac{k^2}{\varepsilon_m^2} - \frac{\omega^2}{c^2 \varepsilon_m} = \frac{k^2}{\varepsilon_i^2} - \frac{\omega^2}{c^2 \varepsilon_i}$$

$$k^2\left(\frac{1}{\varepsilon_m^2} - \frac{1}{\varepsilon_i^2}\right) = \frac{\omega^2}{c^2}\left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_i}\right)$$

$$k^2\left(\frac{\varepsilon_i^2 - \varepsilon_m^2}{\varepsilon_m^2 \varepsilon_i^2}\right) = \frac{\omega^2}{c^2}\left(\frac{\varepsilon_i - \varepsilon_m}{\varepsilon_m \varepsilon_i}\right)$$

$$k^2 = \frac{\omega^2}{c^2}(\varepsilon_i - \varepsilon_m)\left(\frac{\varepsilon_m \varepsilon_i}{\varepsilon_i^2 - \varepsilon_m^2}\right) = \frac{\omega^2}{c^2}(\varepsilon_i - \varepsilon_m)\frac{\varepsilon_m \varepsilon_i}{(\varepsilon_i - \varepsilon_m)(\varepsilon_i + \varepsilon_m)}$$

$$k^2 = \frac{\omega^2}{c^2}\frac{\varepsilon_m \varepsilon_i}{(\varepsilon_i + \varepsilon_m)}$$

$$k = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m \varepsilon_i}{\varepsilon_i + \varepsilon_m}} \qquad (A.9)$$

where, in general, ε_m and ε_i are both functions of frequency and take one complex values. For the case of free space as the dielectric, the dispersion relation reduces to

$$k = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m}{1 + \varepsilon_m}}.$$

Eqn. $(A.9)$ is simple analytical expression which gives the longitudinal wave-vector of a surface plasmon mode for a given set of material properties ε_m, ε_i at some working frequency ω. Since the dielectric constant of a metal is in general a complex quantity, so is the wave-vector k. In the region of interest for plasmonics, the dielectric constant ε_m of metal is primarily real and generally takes on negative values. On the other hand, the dielectric constant ε_i of dielectrics is always purely real and positive, so that if $Re[\varepsilon_m] \approx -\varepsilon_i$ then $\varepsilon_m + \varepsilon_i \to 0$, and the wave-vector k can become quite large.

Appendix B – Derivation of Parallel-Plate Surface Plasmon Dispersion Relation

Here we derive the dispersion relation for a parallel-plate waveguide surface plasmon mode (see geometry in Fig.B1). Since this is a natural extension of the single metal-insulator interface considered in Appendix A, the assumptions and initial derivation steps for the field equations here are the same as before, and will not be repeated.

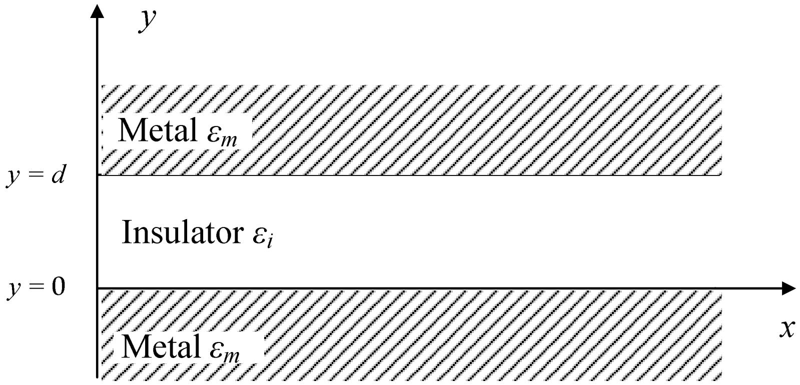


Figure B1: Reference coordinate system for discussions pertaining to surface plasmons in a parallel plate structure.

Following the procedure outlined in Appendix A, it is found that, in the case of the geometry at hand, the mathematical expression for E_y must again satisfy the partial differential equation

$$K_{i,m}^2 = \frac{\partial^2 E_y}{\partial y^2},$$

where, as before, $K_{i,m}$ is given by $(A.6)$. It is easy then to see that the solution

$$E_y = Ce^{-K_m y} \qquad \text{for } y > d$$

$$E_y = B_1 e^{-K_i y} + B_2 e^{K_i y} \quad \text{for } d > y > 0$$

$$E_y = Ae^{K_m y} \qquad \text{for } y < 0$$

(where A, B_1, B_2, and C are constants to be fixed by the boundary conditions) satisfies the differential equation and the physical requirement that all fields vanish at infinity. Next, we use the boundary conditions at the two metal-dielectric interfaces to fix A, B_1, B_2, and C. At both interfaces (which correspond to the $y = 0$ and the $y = d$ planes), E_x, H_z and D_y have to be continuous. To simplify the analysis, and without loss of generality, we immediately let $A = 1$.

Then, using the new solution along with relation $(A.5)$, we find the new form of E_x:

$$E_x = -(j/k)CK_m e^{-K_m y} \qquad \text{for } \; y > d$$

$$E_x = -(j/k)(B_1 K_i e^{-K_i y} - B_2 K_i e^{K_i y}) \qquad \text{for } \; d > y > 0$$

$$E_x = (j/k)K_m e^{K_m y} \qquad \text{for } \; y < 0$$

Clearly then, the continuity of E_x at the $y = 0$ interface requires

$$K_m = K_i(B_2 - B_1) \quad \rightarrow \quad B_2 = (K_m/K_i) + B_1 \, , \tag{B.1}$$

while the continuity of D_y at the $y = 0$ interface requires

$$\varepsilon_m = \varepsilon_i(B_1 + B_2) \, . \tag{B.2}$$

Combining $(B.1{,}2)$, we find

$$\varepsilon_m = \varepsilon_i(B_1 + B_2) = \varepsilon_i\left(B_1 + \left(\frac{K_m}{K_i} + B_1\right)\right)$$

$$= \varepsilon_i\left(2B_1 + \frac{K_m}{K_i}\right)$$

$$\rightarrow \quad B_1 = \frac{1}{2}\left(\frac{\varepsilon_m}{\varepsilon_i} - \frac{K_m}{K_i}\right)$$

$$\rightarrow \quad B_2 = \frac{1}{2}\left(\frac{\varepsilon_m}{\varepsilon_i} + \frac{K_m}{K_i}\right)$$

Furthermore, the continuity of E_x at the $y = d$ plane dictates that

$$K_i(B_1 e^{-K_i d} - B_2 e^{K_i d}) = CK_m e^{-K_m d},$$

which can be rearranged to give

$$C = \frac{K_i}{K_m}(B_1 e^{-K_i d} - B_2 e^{K_i d})e^{K_m d}$$

$$\rightarrow \quad C = \frac{K_i}{K_m}\left(\frac{1}{2}\left(\frac{\varepsilon_m}{\varepsilon_i} - \frac{K_m}{K_i}\right)e^{-K_i d} - \frac{1}{2}\left(\frac{\varepsilon_m}{\varepsilon_i} + \frac{K_m}{K_i}\right)e^{K_i d}\right)e^{K_m d}$$

We can now combine the expressions for B_1, B_2, and C to find the dispersion relation of surface plasmons in a parallel plate structure. This is done by considering the continuity of D_y at the $y = d$ plane:

$$\varepsilon_m C e^{-K_m d} = \varepsilon_i (B_1 e^{-K_i d} + B_2 e^{K_i d})$$

$$\frac{K_i}{K_m}(B_1 e^{-K_i d} - B_2 e^{K_i d}) = \frac{\varepsilon_i}{\varepsilon_m}(B_1 e^{-K_i d} + B_2 e^{K_i d})$$

$$\left(\frac{K_i}{K_m} - \frac{\varepsilon_i}{\varepsilon_m}\right) B_1 e^{-K_i d} = \left(\frac{K_i}{K_m} + \frac{\varepsilon_i}{\varepsilon_m}\right) B_2 e^{K_i d}$$

$$\left(\frac{K_i \varepsilon_m - K_m \varepsilon_i}{K_m \varepsilon_m}\right) B_1 e^{-K_i d} = \left(\frac{K_i \varepsilon_m + K_m \varepsilon_i}{K_m \varepsilon_m}\right) B_2 e^{K_i d}$$

$$\left(\frac{K_i \varepsilon_m - K_m \varepsilon_i}{K_m \varepsilon_m}\right)\left(\frac{\varepsilon_m}{\varepsilon_i} - \frac{K_m}{K_i}\right) = \left(\frac{K_i \varepsilon_m + K_m \varepsilon_i}{K_m \varepsilon_m}\right)\left(\frac{\varepsilon_m}{\varepsilon_i} + \frac{K_m}{K_i}\right) e^{2K_i d}$$

$$\frac{(K_i \varepsilon_m - K_m \varepsilon_i)^2}{(K_m \varepsilon_m)(K_i \varepsilon_i)} = \frac{(K_i \varepsilon_m + K_m \varepsilon_i)^2}{(K_m \varepsilon_m)(K_i \varepsilon_i)} e^{2K_i d}$$

$$e^{-2K_i d} = \left(\frac{K_i \varepsilon_m + K_m \varepsilon_i}{K_i \varepsilon_m - K_m \varepsilon_i}\right)^2$$

$$\rightarrow \quad e^{-K_i d} = \frac{K_i \varepsilon_m + K_m \varepsilon_i}{K_i \varepsilon_m - K_m \varepsilon_i}.$$

Upon substitution of $(A.6)$ for $K_{i,m}$ into the expression above, we obtain a transcendental equation in k that implicitly defines the desired dispersion relation for an arbitrary value of plate spacing d. This type of equation is not analytically tractable and requires numerical solution. In the expression above, we considered only the positive root because we are only concerned with even modes (*i.e.* field solutions with H_z symmetric with respect to the $y = d/2$ axis), seeing as these modes can be shown to provide for more optical confinement than the odd ones in a parallel plate geometry.

Appendix C – Equivalence of Surface Plasmon Dispersion

There is nothing so useless as doing efficiently that which should not be done at all.

–Peter F. Drucker

Here we show that the single plate surface plasmon dispersion relation derived in Appendix A follows from the parallel plate surface plasmon dispersion relation from Appendix B in the limit of large plate spacing $d \to \infty$. Recall from Appendix B the dispersion relation for the parallel plate case is

$$e^{-K_i d} = \left(\frac{K_i \varepsilon_m + K_m \varepsilon_i}{K_i \varepsilon_m - K_m \varepsilon_i}\right) \quad \text{with} \quad K_{i,m} = \left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_{i.m}\right)^{1/2},$$

where d is the plate spacing. As $d \to \infty$ this reduces to

$$K_i \varepsilon_m = -K_m \varepsilon_i$$

$$k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_i = \left(k^2 - \left(\frac{\omega}{c}\right)^2 \varepsilon_m\right) \frac{\varepsilon_i^2}{\varepsilon_m^2}$$

$$k^2 \left(1 - \frac{\varepsilon_i^2}{\varepsilon_m^2}\right) = \left(\frac{\omega}{c}\right)^2 \left(\varepsilon_i - \frac{\varepsilon_i^2}{\varepsilon_m}\right) = \left(\frac{\omega}{c}\right)^2 \left(\frac{\varepsilon_m \varepsilon_i - \varepsilon_i^2}{\varepsilon_m}\right)$$

$$k^2 \left(\frac{\varepsilon_m^2 - \varepsilon_i^2}{\varepsilon_m^2}\right) = \left(\frac{\omega}{c}\right)^2 \frac{\varepsilon_i (\varepsilon_m - \varepsilon_i)}{\varepsilon_m}$$

$$k^2 = \left(\frac{\omega}{c}\right)^2 \frac{\varepsilon_m \varepsilon_i (\varepsilon_m - \varepsilon_i)}{(\varepsilon_m^2 - \varepsilon_i^2)} = \left(\frac{\omega}{c}\right)^2 \frac{\varepsilon_m \varepsilon_i (\varepsilon_m - \varepsilon_i)}{(\varepsilon_m - \varepsilon_i)(\varepsilon_m + \varepsilon_i)}$$

$$\to \quad k = \frac{\omega}{c} \sqrt{\frac{\varepsilon_m \varepsilon_i}{\varepsilon_m + \varepsilon_i}}$$

which is the surface plasmon dispersion relation for a single plate from Appendix A.

Appendix D – Derivation of Capacitance per Unit Length for a Flat Metal Plate

Consider a 2D charge distribution σ_s in the plane, as shown in Fig.2(a):

$$\sigma_s = \sigma_o cos(kx), \quad \{x \in (-\infty, \infty), y \in (-W/2, W/2)\}$$

The potential V_o at the origin with respect to infinity is given by the usual integral over charge density:

$$V_o = \frac{1}{4\pi\varepsilon_o}\int_{-\infty}^{\infty} dx \int_{-W/2}^{W/2} dy \frac{\sigma_s}{\sqrt{x^2+y^2}} = \frac{1}{4\pi\varepsilon_o}\int_{-\infty}^{\infty} dx \int_{-W/2}^{W/2} dy \frac{\sigma_o cos(kx)}{\sqrt{x^2+y^2}}. \qquad (D.1)$$

The integration with respect to y may be evaluated using an indefinite integral from tables[12]:

$$\int_{-W/2}^{W/2} dy \frac{1}{\sqrt{x^2+y^2}} = 2\int_{0}^{W/2} dy \frac{1}{\sqrt{x^2+y^2}} = 2\left[ln\left|y+\sqrt{y^2+x^2}\right|\right]_{y=0}^{y=W/2}$$

$$= 2ln\left|\frac{W}{2}+\sqrt{\frac{W^2}{4}+x^2}\right| - 2ln|x|.$$

Inserting this integral into Eqn. $(D.1)$ results in a strongly oscillating term, multiplied by a logarithmic function of x:

$$V_o = \frac{1}{4\pi\varepsilon_o}\int_{-\infty}^{\infty} dx\, \sigma_o cos(kx)\cdot 2\left(ln\left|\frac{W}{2}+\sqrt{\frac{W^2}{4}+x^2}\right| - ln|x|\right). \qquad (D.2)$$

Changing variables to $kx = \theta$, this becomes

$$V_o = \frac{1}{4\pi\varepsilon_o k}\int_{-\infty}^{\infty} d\theta\, \sigma_o cos(\theta)\cdot 2\left(ln\left|\frac{kW}{2}+\sqrt{\frac{(kW)^2}{4}+\theta^2}\right| - ln|\theta|\right). \qquad (D.3)$$

The condition for treating Fig.3(a) as a one-dimensional wave on a plane is that the wavelength should be much shorter than the width W of the plane. Then $kW \gg 1$, and the first logarithm in the integrand becomes simply $ln|kW|$, a constant number which multiplies the oscillating cosine, and averages to zero. Then Eqn. $(D.3)$ simplifies to

$$V_o = -\frac{2\sigma_o}{4\pi\varepsilon_o k}\int_{-\infty}^{\infty} d\theta\, cos(\theta) ln|\theta|. \qquad (D.4)$$

Eqn. $(D.4)$ can be integrated by parts, with the integral converted to $\int_{-\infty}^{\infty}(sin\theta/\theta)\,d\theta$ which is equal[12] to $-\pi$. The peak voltage potential produced by all the surface charge is $V_o = \sigma_o/2\varepsilon_o k$. Since V_o represents the peak value of a cosine, $V(x) = (\sigma_o/2\varepsilon_o k)cos(kx)$. If we call C' the capacitance per unit length, then the surface charge per unit length becomes

$$C'V(x) = (C'\sigma_o/2\varepsilon_o k)cos(kx). \qquad (D.5)$$

We have an alternate expression for surface charge per unit length along the propagation direction, which is obtained by multiplying the surface charge density σ_s times width W:

$$\sigma_s W = \sigma_o W cos(kx). \qquad (D.6)$$

Requiring Eqns. $(D.5,6)$ to be equivalent, the capacitance per unit length along the propagation direction is: $C' = 2\varepsilon_o kW$.

Appendix E – Derivation of Inductance per Unit Length for a Flat Metal Plate

As the surface charge shown in Fig.2(a) oscillates in time, sinusoidal surface currents must flow in space. The surface currents produce a magnetic field $B(x,z)$ above the surface, spatially sinusoidal in the x-direction. The effective Faraday inductance L_F can be calculated[13] from $\int Bdxdz = L_F I$, where I is the surface current over the full metal width W, and the magnetic flux is obtained by integrating $\int Bdxdz$ above the metal surface in the $+z$-direction, and in the x-direction. Expressed as inductance/per unit length L' in the x-direction, the magnetic flux is $\int Bdxdz = \int L'_F I dx$. Equating the integrands yields $L'_F I = \int Bdz$. Thus a calculation of $\int Bdz/I$, allows us to determine L'_F, the Faraday inductance per unit length. we now show that $L'_F = \mu_o/2kW$.

In calculating $B(x,z)$, it is helpful to use the vector potential A, just as it was helpful to use the scalar potential V in calculating capacitance. Since all the currents flow in the x-direction, the only non-zero component is A_x:

$$A_x(z) = \frac{\mu_o}{4\pi}\int dx'dy\frac{J_s}{r} = \frac{\mu_o}{4\pi}\int_{-\infty}^{\infty} dx' \int_{-W/2}^{W/2} dy \frac{J_s}{\sqrt{x'^2 + y^2 + z^2}}$$

$$= \frac{\mu_o}{4\pi}\int_{-\infty}^{\infty} dx' \int_{-W/2}^{W/2} dy \frac{J_o cos(kx')}{\sqrt{x'^2 + y^2 + z^2}} . \qquad (E.1)$$

Which falls off as $1/r$ from the source of current, just as scalar potential falls off as $1/r$. The distinction between x and x', is that x' is the variable of integration of the current density, and x is the variable of integration of magnetic flux. The current density $J_s = J_o cos(kx')$ is expressed per unit area, and integrated per unit area. The magnetic field B can be derived from $B = \nabla \times A$, but the only non-zero component is $B_y = (\partial A_x/\partial z)$. Calculating the magnetic flux in the y-direction $\int(\partial A_x/\partial z)\, dzdx$ and integrating only in z, the magnetic flux simplifies to $dx[A_x(z)]_0^\infty$. An inspection of Eqn. $(E1)$ shows that $A_x(z=\infty) = 0$, but $A_x(z=0)$ is finite. This procedure is equivalent to using Stokes' Theorem for the magnetic flux:

$$\int B_y dxdz = \int (\nabla \times A_x)dxdz = \int A_x \cdot dl = A_x(z=0)dx\,,$$

where the only part of the contour integral that is non-zero is along the incremental path dx, at the metal surface $z = 0$. Therefore the magnetic flux is $A_x(z=0)dx$, which by using Eqn. $(E1)$ can be written:

$$dxA_x(0) = dx\frac{\mu_o}{4\pi}\int_{-\infty}^{\infty} dx' \int_{-W/2}^{W/2} dy \frac{J_o cos(kx')}{\sqrt{x'^2 + y^2}} . \qquad (E.2)$$

Eqn. $(E.2)$ is identical in structure to Eqn. $(D.1)$. The integration over y is an indefinite integral, and the integration over x' is a definite integral, as before. The integral reduces to

$$dxA_x(0) = dx\frac{(-2)\mu_o J_o}{4\pi k}\int_{-\infty}^{\infty} d\theta cos(\theta) ln|\theta| = dx\frac{\mu_o J_o}{2k}\,. \qquad (E.3)$$

Since Eqn. $(E.3)$ represents flux, then $dxA_x(0) = LI = L'dxI$, where I is the total current in the sheet $J_o W$. Then $dx(\mu_o J_o/2k) = L'dxJ_o W$. Cancelling equal terms on both sides of this equation, $L' = \mu_o/2kW$.

Appendix F – Derivation of NFT-Media Coupling Efficiency Integral

We present two separate derivations of the quantity that needs to be integrated in the recording medium to evaluate the local power dissipation. The first derivation is a conventional argument based on the Maxwell equations and Poynting's theorem. Subtracting the inner product of the magnetic field H with Faraday's Law from the inner product of the electric field E with Ampere's Law we obtain

$$E \cdot (\nabla \times H) - H \cdot (\nabla \times E) = J \cdot E + E \cdot \frac{\partial D}{\partial t} + H \cdot \frac{\partial B}{\partial t} \ .$$

It follows that

$$E \cdot J = -\nabla \cdot (E \times H) - E \cdot \frac{\partial D}{\partial t} - H \cdot \frac{\partial B}{\partial t}$$

$$E \cdot \sigma E = -\nabla \cdot (E \times H) - E \cdot \frac{\partial}{\partial t} \varepsilon_o \varepsilon_m E - H \cdot \frac{\partial B}{\partial t}$$

$$\sigma |E|^2 = -\nabla \cdot (E \times H) + j\omega \varepsilon_o \varepsilon_m |E|^2 - H \cdot \frac{\partial B}{\partial t}$$

where we took $J = \sigma E$ and assumed a harmonic time dependence $exp(-j\omega)$. Choosing to let $\sigma = 0$ and thus incorporating all Ohmic losses in the dielectric constant $\varepsilon_m = \varepsilon_m' + j\varepsilon_m''$, we have

$$0 = -\nabla \cdot (E \times H) + j\omega \varepsilon_o (\varepsilon_m' + j\varepsilon_m'') |E|^2 - H \cdot \frac{\partial B}{\partial t}$$

$$\omega \varepsilon_o \varepsilon_m'' |E|^2 = -\nabla \cdot (E \times H) + j\omega \varepsilon_o \varepsilon_m' |E|^2 - H \cdot \frac{\partial B}{\partial t} .$$

The term on the left-hand side represents Joule heating. The terms on the right-hand side represent (in the order they appear left to right) the power flow into the volume, the energy stored in the electric polarization and free-space electric field, and the energy stored in the free-space magnetic field. It follows that the time-averaged power dissipated in a volume V is simply

$$\bar{P} = \frac{1}{2} \int Re\left[\sigma_{optical}\right] |E|^2 dV = \frac{1}{2} \int \omega \varepsilon_o \varepsilon_m'' |E|^2 dV \ .$$

The alternative derivation is based on the circuit approach presented in the first part of the dissertation where we showed that the conductivity σ and dielectric constant $\varepsilon_m = \varepsilon_m' + j\varepsilon_m''$ of a material are related by

$$\sigma_{optical} = j\omega \varepsilon_o (1 - \varepsilon_m') + \omega \varepsilon_o \varepsilon_m'' \ ,$$

where the first term on the right-hand side is reactive and corresponds to kinetic inductance, and the second term is resistive and corresponds to Joule (Ohmic) heating. The time-averaged power dissipated in a volume V follows as

$$\bar{P} = \frac{1}{2}\int Re\left[\sigma_{optical}\right]|E|^2 dV = \frac{1}{2}\int \omega\varepsilon_o\varepsilon_m''|E|^2 dV\ .$$

Appendix G – Survey of Numerical Methods

The FDTD calculations were carried out in *Lumerical FDTD Solution* and the FEM calculations were carried out with *COMSOL Multiphysics, RF Module*. Both are commercially available software packages. While the FDTD method dicretizes the Maxwell equations, in FEM the equations are left intact but the solution is approximated through discretization. The difference in the two approaches manifests itself in FDTD relying on a mostly uniform grid, and FEM using a non-uniform mesh that is better suited for irregular geometries.

The FDTD method performs well when a uniform grid is required over a large volume such as when modeling wave propagation in the grating coupler or the PSIM. Because it relies on a time-stepping algorithm, the method is efficient in memory requirements and it is often the only option when dealing with very intricate media stacks. FDTD was used to model the lollipop design.

The FEM method arrives at a solution by inverting a large matrix representing a boundary value problem. The amount of physical random access memory (RAM) available limits the maximum matrix size and thus the maximum problem size. While FEM cannot handle some of the larger wave propagation problems solved using FDTD, when used on a problem that can be tackled by either method it is always faster, especially when the problem is well behaved and allows solution by iterative techniques. The C-aperture NFT can be modeled using either FDTD or FEM. Both approaches were used so as to calibrate the grid sizes to arrive at comparable results for a given problem type. The tapered metallic waveguide comprising the impedance matching NFT is best modeled using FEM since the slanted metal surfaces are approximated by staircasing in FDTD, partially spoiling the surface plasmon effects that rely on the clean interface between a metal and a dielectric. Below we present a table listing the computer resources on which the simulations were run.

MACHINE SPECS		MACHINE PERFORMANCE
 $21,000	(x2) Intel Xeon QC w5580 3.2 GHz 1333 MHz FSB (2 Processors, 8 cores, 16 threads) 144 GB DDR3 1066 MHz ECC (18 x 8 GB)	*COMSOL Multiphysics FEM* NFT + 4 Layer Media Stack ($8.5x10^6$ DOFs) Peak RAM Usage: 140 GB , Wall Time: 2 hrs *Lumerical FDTD* Grating Coupler + PSIM ($3.37x10^9$ Yee Cells) Peak RAM Usage: 60 GB , Wall Time: 6 days
 $27,579	(x8) AMD Opteron 8356 B3 2.3 GHz (8 Processors, 32 cores) 128 GB DDR2 667 MHz ECC (32 x 4 GB)	*COMSOL Multiphysics FEM* NFT + 4 Layer Media Stack ($8.5x10^6$ DOFs) Impossible, Not Enough RAM *Lumerical FDTD* Grating Coupler + PSIM ($3.37x10^9$ Yee Cells) Peak RAM Usage: 60 GB , Wall Time: 10 days
 $100,000	Sgi Scalable HPC Cluster (x20) Intel Xeon QC w5580 3.2 GHz 1333 MHz FSB (20 Processors, 80 cores, 160 threads) 80 GB DDR2 1333 MHz ECC (40 x 2 GB)	*COMSOL Multiphysics FEM* NFT + 4 Layer Media Stack ($8.5x10^6$ DOFs) Impossible, Not Enough RAM *Lumerical FDTD* Grating Coupler + PSIM ($3.37x10^9$ Yee Cells) Peak RAM Usage: 60 GB , Wall Time: 14 hrs

Appendix H – Review of the Calculus of Variations

The following is an adaptation of the material found in Chapter 9 of Boas.

Suppose that in solving a given physical problem you want the minimum values of a function $f(x)$. The equation $f^{'}(x) = 0$ is a necessary (but not a sufficient) condition for an interior minimum point. To find the desired minimum you would find all the values of x such that $f^{'}(x) = 0$, and then rely on the physics or on further mathematical tests to sort out the minimum points. We use the general term "stationary point" to mean simply that $f^{'}(x) = 0$ there; that is, stationary points include maximum points, minimum points, and points of inflection with horizontal tangent (saddle points). In the calculus of variations, we often state problems by saying that a certain quantity is to be minimized. However, what we actually always do is something similar to putting $f^{'}(x) = 0$; that is, we make the quantity stationary. The question of whether we have a maximum, minimum, or neither, is, in general, a difficult mathematical problem so we shall rely on the physics or geometry to answer it. Fortunately, in many applications, "stationary" is all that is required. Now what is the quantity that we want to make stationary? it is an integral

$$I = \int_{x_1}^{x_2} F(x, y, y')dx \quad \text{where } y' = \frac{dy}{dx} \tag{$H.1$}$$

and our problem is this: given the points (x_1, y_1) and (x_2, y_2) and the form of the function $F$of x, y, and y', find the curve $y = y(x)$ (passing through the given points) which makes the integral I have the smallest possible value (or stationary value). Before we try to do this, let us look at an example. suppose we want to find the equation of a curve $y = y(x)$ joining two points (x_1, y_1) and (x_2, y_2) in the plane so that the distance between the points measured along the curve (arc length) is a minimum. To find any arc length, we find $\int ds = \int \sqrt{dx^2 + dy^2}$ along the curve. But $\sqrt{dx^2 + dy^2} = \sqrt{1 + y'^2}$, so we want to minimize

$$I = \int_{x_1}^{x_2} \sqrt{1 + y'^2}dx \tag{$H.2$}$$

This is equation $(H.1)$ with $F(x, y, y') = \sqrt{1 + y'^2}$. Thus our problem is to find $y = y(x)$ which will make

$$I = \int_{x_1}^{x_2} \sqrt{1 + y'^2}dx$$

as small as possible. The $y(x)$ which does this is called an exrtemal. We want some way to represent algebraically all the curves passing through the given endpoints, but differing from the (as yet unknown) extremal by small amounts. We assume that all the curves have continuous second derivatives so that we can carry out the needed differentiations later. These curves are called varied curves; there are infinitely many of them as close as we like to the extremal. We construct a function representing these varied curves in the following way. Let $\eta(x)$ represent a function of x which is zero at x_1 and x_2 , and has a continuous second derivative in the interval x_1 to x_2 , but is otherwise completely arbitrary. We define the function $Y(x)$ by the equation

$$Y(x) = y(x) + \epsilon\eta(x) \qquad (H.3)$$

where $y(x)$ is the desired extremal and ϵ is a parameter. Because of the arbitrariness of $\eta(x)$, $Y(x)$ represents any (single-valued) curve (with continuous second derivative) you can draw through (x_1, y_1) and (x_2, y_2). Out of all these curves $Y(x)$ we want to pick the one curve that makes

$$I = \int_{x_1}^{x_2} \sqrt{1 + Y'^2}dx \qquad (H.4)$$

a minimum. Now I is a function of the parameter ϵ; when $\epsilon = 0, Y = y(x)$, the desired extremal. Our problem is then to make $I(\epsilon)$ take its minimum value when $\epsilon = 0$. In other words, we want

$$\left.\frac{\partial I}{\partial \epsilon}\right|_{\epsilon=0} = 0 \qquad (H.5)$$

Differentiating $(H.4)$ under the integral sign with respect to the parameter ϵ, we get

$$\frac{\partial I}{\partial \epsilon} = \int_{x_1}^{x_2} \frac{1}{2}\frac{1}{\sqrt{1 + Y'^2}}2Y'\frac{\partial Y'}{\partial \epsilon}dx \qquad (H.6)$$

Differentiating $(H.3)$ with respect to x, we get

$$Y'(x) = y'(x) + \epsilon\eta'(x) \qquad (H.7)$$

Then, from $(H.7)$ we have

$$\frac{\partial Y'(x)}{\partial \epsilon} = \eta'(x) \qquad (H.8)$$

We see from $(H.3)$ that putting $\epsilon = 0$ means putting $Y(x) = y(x)$. Then substituting $(H.8)$ into

$(H.6)$ and putting $\partial I/\partial \epsilon$ equal to zero when $\epsilon = 0$, we get

$$\left.\frac{\partial I}{\partial \epsilon}\right|_{\epsilon=0} = \int_{x_1}^{x_2} \frac{y'(x)\eta'(x)}{\sqrt{1+y'(x)^2}}dx \qquad (H.9)$$

we can integrate this by parts since we assumed that η and y have continuous second derivatives. Let

$$u = \frac{y'(x)}{\sqrt{1+y'(x)^2}} \qquad \therefore \partial u = \frac{d}{dx}\frac{y'(x)}{\sqrt{1+y'(x)^2}}dx$$

$$\partial v = \eta^{'}(x)dx \qquad \therefore v = \eta(x)$$

and

$$\left.\frac{\partial I}{\partial \epsilon}\right|_{\epsilon=0} = \left.\frac{y'(x)}{\sqrt{1+y'(x)^2}}\eta(x)\right|_{x_1}^{x_2} - \int_{x_1}^{x_2} \eta(x)\frac{d}{dx}\frac{y'(x)}{\sqrt{1+y'(x)^2}}dx = 0$$

The first term is zero because $\eta(x) = 0$ at the endpoints. In the second term recall that $\eta(x)$ is an arbitrary function. This means that

$$\frac{d}{dx}\frac{y'(x)}{\sqrt{1+y'(x)^2}} = 0 \qquad (H.10)$$

for otherwise we could select some function $\eta(x)$ so that the integrand would not be zero. Notice carefully here that we are not saying that when an integral is zero, the integrand is also zero; this is not true. What we are saying is that the only way $\int f(x)\eta(x)\partial x$ can always be zero for every $\eta(x)$ is for $f(x)$ to be zero. Integrating $(H10)$ with respect to x, we get

$$\frac{y'(x)}{\sqrt{1+y'(x)^2}} = constant$$

or $y^{'} = constant$. Thus the slope of $y(x)$ is constant, so $y(x)$ is a straight line as we expected. Now we could go through this process with every calculus of variations problem. It is much simpler to do the general problem once and for all and find a differential equation which we can use to solve later problems. The goal is to find the y which will make stationary the integral

$$I = \int_{x_1}^{x_2} F(x, y, y')dx \qquad (H.11)$$

where F is a given function. The $y(x)$ which makes I stationary is called an extremal whether I is a maximum or minimum or neither. The method is the one we have just used with the straight line. We consider a set of varied curves

$$Y(x) = y(x) + \epsilon\eta(x)$$

just as before. Then we have

$$I(\epsilon) = \int_{x_1}^{x_2} F(x, Y, Y')dx \qquad (H.12)$$

and we want $(\partial/\partial\epsilon)I(\epsilon) = 0$ when $\epsilon = 0$. Remembering that Y and Y' are functions of ϵ, and differentiating under the integral sign with respect to ϵ, we get

$$\frac{\partial I}{\partial \epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial Y}\frac{\partial Y}{\partial \epsilon} + \frac{\partial F}{\partial Y'}\frac{\partial Y'}{\partial \epsilon}\right) dx \qquad (H.13)$$

Substituting $(H.3)$ and $(H.7)$ into $(H.13)$, we have

$$\frac{\partial I}{\partial \epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial Y}\eta(x) + \frac{\partial F}{\partial Y'}\eta'(x)\right) dx \qquad (H.14)$$

we want $\partial I/\partial\epsilon = 0$ at $\epsilon = 0$; recall that $\epsilon = 0$ means $Y = y$. Then $(H.14)$ gives

$$\left.\frac{\partial I}{\partial \epsilon}\right|_{\epsilon=0} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y}\eta(x) + \frac{\partial F}{\partial y'}\eta'(x)\right) dx \qquad (H.15)$$

if y'' is continuous, we can integrate the second term by parts just as in the straight line problem:

$$u = \frac{\partial F}{\partial y'} \qquad \therefore \partial u = \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) dx$$

$$\partial v = \eta^{'}(x)dx \qquad \therefore v = \eta(x)$$

and

$$\int_{x_1}^{x_2} \frac{\partial F}{\partial y'}\eta'(x)dx = \left.\frac{\partial F}{\partial y'}\eta(x)\right|_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\eta(x)dx = 0 \qquad (H.16)$$

The integrated term is zero as before because $\eta(x)$ is zero at x_1 and x_2 .

Then we have

$$\left.\frac{\partial I}{\partial \epsilon}\right|_{\epsilon=0} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'}\right)\eta(x)dx = 0 \qquad (H.17)$$

as before, since $\eta(x)$ is arbitrary, we must have

$$\frac{d}{dx}\frac{\partial F}{\partial y'} - \frac{\partial F}{\partial y} = 0 \qquad (H.18)$$

This is the Euler (or Euler-Lagrange) equation. Any problem in the calculus of variations, then, is solved by setting up the integral which is to be stationary, writing what the function F is, substituting it into the Euler equation, and solving the resulting differential equation. As an example, let's find the geodesics in a plane gain, this time using the Euler equation. We are to minimize

$$\int_{x_1}^{x_2} \sqrt{1 + y'^2}dx$$

so we have $F = \sqrt{1 + y'^2}$. Then

$$\frac{\partial F}{\partial y'} = \frac{y'}{\sqrt{1 + y'^2}} \ , \qquad \frac{\partial F}{\partial y} = 0$$

and the Euler equation gives

$$\frac{d}{dx}\frac{y'}{\sqrt{1 + y'^2}} = 0$$

as we had before.

Lagrange's Equations

It is not necessary to restrict ourselves to problems with one dependent variable y. Recall that in ordinary calculus problems the necessary condition for a minimum point on $z = z(x)$ is $\partial z/\partial x = 0$; for a function of two variables $z = z(x, y)$, we have the two conditions $\partial z/\partial x = 0$ and $\partial z/\partial y = 0$. We have a somewhat analogous situation in the calculus of variations. Suppose that we are given an F which is a function of y, z, $\partial y/\partial x$, $\partial z/\partial x$, and x, and we want to find two curves $y = y(x)$ and $z = z(x)$ which make $I = \int F dx$ stationary. Then the value of the integral I depends on both $y(x)$ and $z(x)$ and you might very well guess that in this case we would have two Euler equations, one for y and one for z, namely

$$\frac{d}{dx}\frac{\partial F}{\partial y'} - \frac{\partial F}{\partial y} = 0$$

$$\frac{d}{dx}\frac{\partial F}{\partial z'} - \frac{\partial F}{\partial z} = 0 \qquad (H.19)$$

By carrying through calculations similar to those we used in deriving the single Euler equation for the one dependent variable case you can show that this guess is correct. If there are still more dependent variables (but one independent variable), then we write an Euler equation for each dependent variable. It is possible also to consider a problem with more than one independent variable, but the results are more complicated and less useful in application, so we shall not do it here.

There is a very important application of equations like $(H.19)$ to mechanics. In elementary physics, Newton's second law $\boldsymbol{f} = m\boldsymbol{a}$ is a fundamental equation. In more advanced mechanics, it is often useful to start from a different assumption (which can be proved equivalent to Newton's law). This assumption is called Hamilton's principle. It says that any particle or system of particles always moves in such a way that $I = \int_{t_1}^{t_2} L dt$ is stationary, where $L = T - V$ is called the Lagrangian; T is the kinetic energy, and V is the potential energy of the particle or system.

Appendix I – Review of Lagrange Multipliers

The following is an adaptation of the material found in Chapter 4.9 of Boas.

Consider a wire bent to fit the curve $y = 1 - x^2$. A string is stretched from the origin to a point (x, y) on the curve. Find (x, y) to minimize the length of the string. We want to minimize the distance $d = \sqrt{x^2 + y^2}$ from the origin to the point (x, y); this is equivalent to minimizing $f = d^2 = x^2 + y^2$. But x and y are not independent; they are related by the equation of the curve. This extra relation between the variables is what we refer to as a constraint. There are several ways to solve a problem like this. Some common methods are a) elimination, b) implicit differentiation, c) Lagrange multipliers. However, methods a) and b) can involve an enormous amount of algebra. We can shortcut this algebra by using the method of Lagrange multipliers or undetermined multipliers. In general, we want to find the maximum or minimum of a function $f(x, y)$, where x and y are related by an equation $\Phi(x, y) = constant$. Then f is really a function of one variable (say x). To find the maximum or minimum points of f, we set $\partial f/\partial x = 0$, or $\partial f = 0$. Since $\Phi = constant$, we get $\partial \Phi = 0$.

$$\partial f = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = 0$$

$$\partial \Phi = \frac{\partial \Phi}{\partial x} dx + \frac{\partial \Phi}{\partial y} dy = 0 \qquad (I.1)$$

Next, we multiply the $\partial \Phi$ equation by λ (this is the undetermined multiplier - we shall find its value later) and add it to the ∂f equation:

$$\left(\frac{\partial f}{\partial x} + \lambda \frac{\partial \Phi}{\partial x}\right) dx + \left(\frac{\partial f}{\partial y} + \lambda \frac{\partial \Phi}{\partial y}\right) dy = 0 \qquad (I.2)$$

We now pick λ so that

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial \Phi}{\partial y} = 0 \qquad (I.3)$$

That is, we pick $\lambda = -(\partial f/\partial y)/(\partial \Phi/\partial y)$, but it is not necessary to write it in this complicated form. In fact, this is exactly the point of the Lagrange multiplier λ ; by using the abbreviation λ for a complicated expression, we avoid some algebra. Then, from $(I.2)$ and $(I.3)$ we have

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial \Phi}{\partial x} = 0 \qquad (I.4)$$

Equations $(I.3), (I.4)$, and $\Phi = constant$ can now be solved for the three unknowns x, y, λ. We don't actually want the value of λ, but often the algebra is simpler if we do find it and use it in finding x and y which we do want. Note that equations $(I.3)$ and $(I.4)$ are exactly the equations we would write if we had a function

$$F(x,y) = f(x,y) + \lambda\Phi(x,y) \qquad (I.5)$$

of two independent variables x and y and we wanted to find its maximum and minimum values. Actually, of course, x and y are not independent; they are related by the Φ equation. However $(I.5)$ gives us a simple way of stating and remembering how to get equations $(I.3)$ and $(I.4)$. Thus we can state the method of Lagrange multipliers in the following way:

To find the maximum or minimum values of $f(x,y)$ when x and y are related by the equation $\Phi = constant$, form the function $F(x,y)$ as in (5) and set the two partial derivatives of Fequal to zero [equations (3) and (4)]. Then solve these two equations and equation $\Phi = constant$ for the three unknowns x, y, and λ.

As a simple illustration of the method we shall do our example problem by Lagrange multipliers. Here

$$f(x,y) = x^2 + y^2, \quad \Phi(x,y) = y + x^2 = 1$$

and we write the equations to minimize

$$F(x,y) = f + \lambda\Phi = x^2 + y^2 + \Phi(y + x^2)$$

namely,

$$\frac{\partial F}{\partial x} = 2x + \lambda 2x = 2x(1+\lambda) = 0$$

$$\frac{\partial F}{\partial y} = 2y + \lambda = 0$$

We solve these simultaneously with the equation $y + x^2 = 1$. From the first equation, either $x = 0$ or $\lambda = -1$. If $x = 0$, $y = 1$ from the Φ equation (and $\lambda = -2$). If $\lambda = -1$, then the second equation gives $y = 1/2$, and then the Φ equation gives $x^2 = 1/2$. Although it may not be apparent from this example, Lagrange multipliers simplify the work enormously in more complicated problems. Although we have only dealt with problems involving two variables (x and y), for a problem involving still more variables there are simply more equations, but no change in the method. Finally, to find the maximum or minimum of f subject to the conditions $\Phi_1 = constant$ and $\Phi_2 = constant$, we just define $F = f + \lambda_1\Phi_1 + \lambda_2\Phi_2$ and set each of the partial derivatives of F equal to zero. We then solve these equations and the Φ equations for the variables and the λ's.

Appendix J – Variational Methods for Symmetric Linear Operators

The following is an adaptation of the material found in Zienkiewicz & Morgan.

We showed that if we are presented with a variational principle in the form of a functional, then the corresponding Euler equation can always be determined. Normally, however, the behavior of a physical system is described in terms of a differential equation, and it is of interest to attempt to determine if a variational formulation of the problem is possible. We shall restrict our attention to the case of linear differential equations, as general rules for nonlinear equations are complicated.

A general linear differential equation may be written in the form

$$L[f] = s \quad \text{in } \Omega \tag{J.1}$$

where L is a linear operator and s is the source (a known function). The solution is required subject to the general boundary condition

$$\xi[f] + r = 0 \quad \text{on } \Gamma \tag{J.2}$$

where ξ is a linear operator and r a given function of position. Consider a set of functions θ which satisfy the homogeneous form of this boundary condition, that is,

$$\xi[\theta] = 0 \quad \text{on } \Gamma \tag{J.3}$$

The operator L is said to be symmetric (or self-adjoint) over the domain Ω with respect to this set of functions if, for any two members θ and ν of this set, we have that

$$\int \theta\, L[\nu] d\Omega = \int \nu\, L[\theta] d\Omega \tag{J.4}$$

additionally, a symmetric operator L is said to be positive definite over Ω with respect to this set of functions if, for any member θ of the set,

$$\int \theta\, L[\theta] d\Omega \geq 0 \tag{J.5}$$

with equality only if θ is identically zero in Ω. Following the definition of these properties of linear operators it is possible to produce the required variational principle. In general terms, the solution f of a self-adjoint linear differential equation $L[f] = s$ in a domain Ω corresponds to a stationary point for the functional

$$I[f] = \tfrac{1}{2}\langle f, L[f]\rangle - \langle f, s\rangle \tag{J.6}$$

where we denoted the scalar product $\int fg d\Omega = \langle f, g\rangle$. To prove this, let δf be a small variation of f. We will consider variations only up to a linear order in δf. We let δI denote the first-order variation of $I[f]$ when $f \to f + \delta f$ and say that $I[f]$ is stationary if $\delta I = 0, \forall \delta f$. Since f represents a minimum, the rate of change of I at f must be zero:

$$I[f + \delta f] = \tfrac{1}{2}\langle f + \delta f, L[f + \delta f]\rangle - \langle f + \delta f, s\rangle$$

$$= \tfrac{1}{2}\langle f, L[f]\rangle - \langle f, s\rangle + \tfrac{1}{2}\langle \delta f, L[f]\rangle + \tfrac{1}{2}\langle f, L[\delta f]\rangle - \langle \delta f, s\rangle + \tfrac{1}{2}\langle \delta f, L[\delta f]\rangle$$

$$= I[f] + \delta I + O((\delta f)^2). \qquad \text{(J.7)}$$

The first variation is the part that is linear in δf, that is,

$$\delta I = \tfrac{1}{2}\langle \delta f, L[f]\rangle + \tfrac{1}{2}\langle f, L[\delta f]\rangle - \langle \delta f, s\rangle \qquad \text{(J.8)}$$

In order for $I[f]$ to be stationary, the first variation must vanish. Now, L is self-adjoint, i.e., $\langle f, L[\delta f]\rangle = \langle \delta f, L[f]\rangle$ so the condition for I to be stationary becomes $\langle \delta f, L[f]\rangle - \langle \delta f, s\rangle = \langle \delta f, L[f] - s\rangle = 0$. Thus, for every admissible variation δf we have

$$\langle \delta f, L[f] - s\rangle = \int \delta f (L[f] - s) d\Omega = 0. \qquad \text{(J.9)}$$

Since δf is an arbitrary function, this requires that the residual $r = L[f] - s$ vanish everywhere in *Ω;* that is, that the differential equation $L[f] = s$ be satisfied. To summarize, the discussion above shows that we can solve the differential equation $L[f] = s$ by finding the function f that makes $I[f]$ stationary. The variational formulation gives a procedure, the Rayleigh-Ritz method, for finding approximate solutions of self-adjoint linear equations. It consists of the following steps:

- Approximate the solution f by an expansion in a finite set of basis (or trial) functions φ_i , $i = 1,2, \ldots, N$:

$$f(\boldsymbol{r}) = \sum_{i=1}^{N} f_i \varphi_i(\boldsymbol{r})$$

- Evaluate the quadratic variational form I as a function of the expansion coefficients

$$I(f_1, f_2, \ldots, f_N) = I[f] = \frac{1}{2}\langle f, L[f]\rangle - \langle f, s\rangle$$

$$= \frac{1}{2}\sum_i \sum_j f_i f_j \langle \varphi_i, L[\varphi_j]\rangle - \sum_i f_i \langle \varphi_i, s\rangle$$

$$= \frac{1}{2} \sum_i \sum_j L_{ij} f_i f_j - \sum_i s_i f_i$$

where $L_{ij} = \langle \varphi_i, L[\varphi_j] \rangle$ and $s_i = \langle \varphi_i, s \rangle$. Note that the matrix $\boldsymbol{L}$ is symmetric, $L_{i,j}$ =$L_{j,i}$, because the operator L is self-adjoint.

- Determine the expansion coefficients f_i by demanding that I be stationary with respect to all the coefficients:

$$0 = \frac{\partial I}{\partial f_k} = \frac{1}{2} \sum_j L_{kj} f_j + \frac{1}{2} \sum_i L_{ik} f_i - s_k = \sum_i L_{ki} f_i - s_k$$

This is a linear symmetric $N \times N$ system $\boldsymbol{Lf} = \boldsymbol{s}$ for the expansion coefficients. To summarize once more, we have shown that the solution of a linear differential equation $L[f] = s$ subject to the general boundary condition $\xi[f] + r = 0$ can be found by seeking the function f which makes a certain functional $I[f]$ stationary among the set of functions that satisfy the appropriate boundary conditions for the problem. If, however, we view the boundary condition as an additional constraint on the problem of making $I[f]$ stationary, then we can use an approach due to Lagrange in which the variation of a new functional $I_1[f, \lambda]$ is considered. The new functional is constructed as

$$I_1[f, \lambda] = I[f] + \int \lambda(\xi[f] + r) d\Gamma \tag{J.10}$$

where λ, known as a Lagrange multiplier, is a function of the space coordinates. The first variation in I_1 is then given by

$$\delta I_1[f, \lambda] = \delta I[f] + \int \delta\lambda(\xi[f] + r) d\Gamma + \int \lambda(\xi[\delta f]) d\Gamma. \tag{J.11}$$

When I_1 is stationary, that is, $\delta I_1 = 0$ for all variations $\delta f, \delta\lambda$ we have

$$\begin{aligned} \xi[f] + r = 0 &\quad \text{on } \Gamma \\ \xi[\delta f] = 0 &\quad \text{on } \Gamma \\ \delta I = 0 &\quad \text{in } \Omega \end{aligned} \tag{J.12}$$

where the last condition is equivalent to $L[f] - s = 0$, and so the function f which makes I_1 stationary is the solution of this equation which satisfies the boundary condition $\xi[f] + r = 0$ on Γ. If approximations for both f and λ are constructed in the usual manner as

$$f(\boldsymbol{r}) = \sum_{i=1}^{N} f_i \varphi_i(\boldsymbol{r})$$

$$\lambda(\boldsymbol{r}) = \sum_{i=1}^{N} \lambda_i \gamma_i(\boldsymbol{r}) \tag{J.13}$$

then the constants $\{f_i\,, \varphi_i\,:\; i = 1,2,\dots,N\}$ can be determined by the requirement that $I_1[f,\lambda]$ be stationary. Inserting the approximations into equation (J.10) gives

$$I_1[f,\lambda] \;= I[f] + \int \left[\sum_{i=1}^{N} \lambda_i\, \gamma_i\right]\left[\left(\sum_{i=1}^{N} f_i\, \xi[\varphi_i]\right) + r\right] d\Gamma \tag{J.14}$$

which is stationary with respect to variations in f_j, λ_j , provided that

$$\frac{\partial I_1}{\partial f_j} = \frac{\partial I}{\partial f_j} + \int \left[\sum_{i=1}^{N} \lambda_i\, \gamma_i\right] \xi[\varphi_j] d\Gamma = 0 \tag{J.15}$$

$$\frac{\partial I_1}{\partial \lambda_j} = \frac{\partial I}{\partial \lambda_j} + \int \gamma_i \left[\left(\sum_{i=1}^{N} f_i\, \xi[\varphi_i]\right) + r\right] d\Gamma = 0 \tag{J.16}$$

The first term on the right-hand side of equation (J.15) is given by the original variational principle, and we can write

$$0 = \frac{\partial I}{\partial f_j} = \sum_i L_{ji} f_i - s_j$$

The equation set (15)-(16) then becomes, in matrix form

$$\begin{bmatrix} \boldsymbol{L} & \boldsymbol{L}_1 \\ \boldsymbol{L}_1^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{s}_1 \end{bmatrix} \tag{J.17}$$

where

$$[\boldsymbol{L}_1]_{ij} = \int \gamma_j \xi[\varphi_i] d\Gamma \tag{J.18}$$

$$[\boldsymbol{s}_1]_i = -\int r\gamma_i d\Gamma \tag{J.19}$$

COMSOL Multiphysics does the same thing but uses the notation

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{N}_F \\ \boldsymbol{N}_F^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{\Lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \end{bmatrix}$$

with the nomenclature $\boldsymbol{L}$ → residual vector, $\boldsymbol{M}$ →constraint residual vector, $\boldsymbol{U}$ →solution vector, $\boldsymbol{K}$ →stiffness matrix, $\boldsymbol{N}_F$ →constraint Jacobian matrix, $\boldsymbol{\Lambda}$ →Lagrange multiplier vector. Note that this form of the problem only applies to linear self-adjoint operators in the event that all the constraints are ideal.

By interfacing *COMSOL Multiphysics* with MATLAB we can build a model using the COMSOL graphical user interface (GUI), solve it with an in-house algorithm, and then use the COMSOL GUI for post-processing. To do this, simply create a model within COMSOL, link to MATLAB, and then select *File→Export→FEM Structure as 'fem'*, and in the MATLAB command prompt type `fem1=csolver(fem)`, where `csolver` is defined as

```
function fem1=csolver(fem)

    fem1=fem;
    fem1.xmesh=meshextend(fem);

    [K,L,M,N]=assemble(fem);

    s=size(N);

    A=[K,N';N,sparse(s(1),s(1))];
    B=[L;M];

    X=A\B;

    U=X(1:s(2));

    fem1.sol=femsol(U);
```

The line '`X=A\B;`' can be replaced with a function call to an in-house solver of your choice. The solved model is then available for post-processing in COMSOL by selecting *File→Import→FEM Structure* and typing `fem1` at the prompt.

Appendix K – Modeling the Heat Equation

The goal of the HAMR thermal simulation is to find either the transient or the steady-state temperature profile in the recording media layer given that it is moving at some speed v relative to the NFT-slider assembly. Because the time scales of disk motion are glacially slow compared to the frequency of the light used to excite the NFT, one can solve the electromagnetic problem in isolation, obtain the volumetric Joule heating profile in the NFT and media stack, and then use it as a volumetric heat source term in the heat equation for the thermal problem. When solving the thermal problem, however, one has to account for the relative speed between the NFT and the media, as this will introduce finite conductive heat transfer and will generally affect the final thermal profile in the media compared to a static scenario. One way of tackling the problem of a moving load is to solve the time-dependent conductive heat equation over a very small time step (typically fractions of a nanosecond), then shift the solutions at the load by a distance vt where t is the duration of the step, and v is the relative linear velocity between the heating element and the load (40-80 m/s). The problem is then solved again and the process is repeated until the solution evolves to its steady state value, or reaches some other design point. This technique involves solving the conductive heat equation:

$$-\kappa\nabla^2 T + \rho C_p \frac{dT}{dt} = Q\,, \tag{K.1}$$

which is the result of two simple consideration:

1. The heat energy flux, $\mathbf{h}$, is always proportional to the temperature gradient, ∇T, and happens to be directly opposite to it (since heat flows from regions of high temperature to regions of low temperature), with the proportionality constant being the thermal conductivity κ. Mathematically this is expressed as

$$\mathbf{h} = -\kappa\nabla T\,. \tag{K.2}$$

2. The heat power introduced by a volumetric heat source, Q, in a volume of matter will either vacate the volume as heat flux through the surface enclosing it, or it will remain in the volume and result in a change in the temperature of the matter it encloses (this temperature change being dependent on the density, ρ, and volumetric heat capacity, C_p, of the matter at hand). Mathematically this is expressed as

$$\nabla\cdot\mathbf{h} + \rho C_p \frac{dT}{dt} = Q\,. \tag{K.3}$$

Substitution of equation (K.2) for $\mathbf{h}$ in equation (K.3) then yields the conductive heat equation (K.1). Although numerically and physically sound, this approach is quite time consuming if all one desires is the steady-state temperature profile at the load. In this case it is possible to alleviate the computational burden by treating the load as an incompressible fluid flowing at a constant velocity alongside the heating element. When cast this way, the problem requires that we introduce a convective term in the heat equation. The extra term arises once we replace dT/dt in $(\mathrm{K}.1)$ with the convective derivative

$$\frac{dT}{dt} = \frac{\partial T}{\partial t} + \frac{\partial x}{\partial t} \cdot \frac{\partial T}{\partial x} + \frac{\partial y}{\partial t} \cdot \frac{\partial T}{\partial y} + \frac{\partial z}{\partial t} \cdot \frac{\partial T}{\partial z} = \frac{\partial T}{\partial t} + (v \cdot \nabla)T\,, \tag{K.4}$$

to account for the relative motion in the frame of the heating element. The heat equation then becomes

$$-\kappa \nabla^2 T + \rho C_p \frac{\partial T}{\partial t} + \rho C_p (v \cdot \nabla)T = Q\,, \tag{K.5}$$

and in this form it can account for both conductive and convective heat flow. The *Convection and Conduction Thermal Module* in *COMSOL Multiphysics* is set up to solve this type of equation.

Bibliography

You can write the entire history of science in the last 50 years in terms of papers rejected by Science or Nature.

– Paul C. Lauterbur

[1] N. Engheta, A. Salandrino, A Alu, "Circuit elements at optical frequencies: nanoinductors, nanocapacitors, and nanoresistors", PRL 95, 095504 (2005).

[2] N. Engheta, "Circuits with light at nanoscales: optical nanocircuits inspired by metamaterials", Science 317, 1698 (2007).

[3] A. Ishikawa, T. Tanaka, S. Kawata, "Negative magnetic permeability in the visible light region", PRL 95, 237401 (2005).

[4] J. Zhou, T. Koschny, M. Kafesaki, E.N. Economou, J.B. Pendry, C.M. Soukoulis, "Saturation of the magnetic response of split-ring resonators at optical frequencies", PRL 95, 223902 (2005).

[5] C. Huang *et al*, "Study of plasmonic resonance in a gold nanorod with an LC circuit model," Optics Express, Vol. 17, No. 8, 6407, (2009).

[6] N.A. Krall, A.W. Trivelpiece, *Principles of Plasma Physics*, McGraw-Hill, p. 123 (1973).

[7] D.A. Cardwell, *Handbook of Super Conducting Materials*, CRC Press (2003).

[8] H.M. Barlow, A.L. Cullen, "Surface waves", Proc. IEE 100, 399 (1953).

[9] E.N. Economou, "Surface plasmons in thin films", Physical Review Vol. 182, No. 2, 539-554 (1969).

[10] H. Raether, *Surface plasmons on smooth and rough surfaces and gratings*, Springer Tracts in Modern Physics, Vol. 111 (1988).

[11] S.A. Maier, *Plasmonics: Fundamentals and Applications*, Springer (2007).

[12] H.B. Dwight, *Tables of Integrals and Other Mathematical Data*, MacMillan (1955).

[13] F.T. Ulaby, *Fundamentals of Applied electromagnetics*, Prentice Hall (2007).

[14] J.D. Jackson, *Classical Electrodynamics*, Wiley (1999).

[15] P. Ramo, J.R. Whinnery, Van Duzer, *Fields and Waves in Communication Electronics*, Wiley (1994).

[16] D.M. Pozar, *Microwave Engineering, 3rd Edition*, Wiley, (2005).

[17] N. Gershenfeld, *The Physics of Information Technology*, Cambridge Univ. Press (2000)

[18] P.B. Johnson, J.W. Christy, "Optical Constants of the Noble Metals", Phys. Rev. B, Vol. 6, No. 12, 4370 (1972).

[19] J. Conway, "Efficient Optical Coupling to the Nanoscale", Ph.D Dissertation (2006).

[20] T. Okamoto, "Near-field spectral analysis of metallic beads", from S. Kawata, Near-Field Optics and Surface Plasmon Polaritons, Springer Topics Appl. Phys. 81, 97-123 (2001).

[21] M.I. Stockman," Nanofocusing of Optical Energy in Tapered Plasmonic Waveguides", PRL 93, No. 13, 137404-1 (2004).

[22] L. Novotny & B. Hecht, *Principles of Nano-Optics*, Cambridge Press, (2006).

[23] D.F.P. Pile & D.K. Gramotnev, "Adiabatic and nonadiabatic nanofocusing of plasmons by tapered gap plasmon waveguides", APL 89, 041111, (2006).

[24] W.A. Challener, A.V. Itagi, *Near-Field Optics for Heat Assisted Magnetic Recording (Experiment, Theory, and Modeling)*, Springer (2009).

[25] P.J. Schuck et al, "Improving the mismatch between light and nanoscale objects with gold bowtie nanoantennas", PRL 94, 017402, (2005).

[26] A. Alu & N. Engheta, "Input impedance, nanocircuit loading, and radiation tuning of optical nanoantennas", PRL 101, 043901-1, (2008).

[27] A. Alu & N. Engheta, "Wireless at the Nanoscale: Optical Interconnects using Matched Nanoantennas", PRL 104, 213902, (2010).

[28] J.S. Huang *et al*, "Impedance matching and emission properties of nanoantennas in an optical nanocircuit", Nano Letters, Vol. 9, No. 5, 1897-1902, (2009).

[29] J. Wen, S. Romanov, and U. Peschel, "Excitation of plasmonic gap waveguides by nanoantennas", Optics Express, Vol. 17, 5925, (2009).

[30] P. Bharadwaj & L. Novotny, "Spectral dependence of single molecule fluorescence enhancement", Optics Express, Vol. 15, No. 21, 14266, (2007).

[31] S. Nie & S.R. Emory, "Probing single molecules and single nanoparticles by surface-enhanced Raman scattering", Science 275, 1102, (1997).

[32] K. Kneipp *et al*, "Single molecule detection using surface-enhanced Raman scattering (SERS)", PRL 78, 1667–1670, (1997).

[33] Y. Chu, M.G. Banaee & K.B. Crozier, "Double-Resonance Plasmon Substrates for Surface-Enhanced Raman Scattering with Enhancement at Excitation and Stokes Frequencies", ACS Nano 4, 2804–2810, (2010).

[34] P.L. Stiles, J.A. Dieringer, N.C. Shah & R.P. Van Duyne, "Surface-Enhanced Raman Spectroscopy", Annual Review of Analytical Chemistry (2008) 1, 601-626, (2008).

[35] C. Edward, Jordan, and K.G. Balmain, *Electromagnetic waves and radiating systems*, pp. 325-6 (1968).

[36] H.A. Wheeler, "Fundamental limitations of small antennas", Proc. of the I.R.E., pp. 1479-94 (1947).

[37] G. Dubost and J. Dupuy, "Effective area of an antenna", Electronic Letters, Vol. 12 (1976).

[38] R.P. Feynman, *Lectures on Physics*, Vol. I, Chapter 32 (1964).

[39] E.M. Purcell, Phys. Rev. 69, 681 (1946).

[40] C.A. Balanis, *Antenna Theory*, Wiley (2005).

[41] L. Novotny & B. Hecht, *Principles of Nano-Optics*, Cambridge Press, (2006).

[42] A. Kinkhabwala, *et. al.*, "Large single-molecule fluorescence enhancements produced by a bowtie nanoantenna", Nat. Phot., Vol. 3, (2009).

[43] E.K. Lau et al., “Enhanced modulation bandwidth of nanocavity light emitters,” Opt. Exp., Vol.17, No.10, (2009).

[44] T.W. McDaniel, "Ultimate limits to thermally assisted magnetic recording", J. Phys. Condens. Matter 17, R315-R332 (2005).

[45] W.A Challener *et al*, "Heat-assisted magnetic recording by a near-field transducer with efficient optical energy transfer", Nature Photonics 3, 220-224 (2009).

[46] C. Peng *et al*, "Surface-plasmon resonance of a planar lollipop near-field transducer", APL 94, 171106 (2009).

[47] C. Peng *et al*, "Focusing characteristics of a planar solid-immersion mirror", Applied Optics, Vol. 45, No. 8 (2006).

[48] W.A. Challener *et al*, "Miniature planar solid immersion mirror with focused spot less than a quarter wavelength", Optics Express, Vol. 13, No. 18 (2005).

[49] C. Peng *et al*, "Near-field optical recording using a planar solid immersion mirror", APL 87, 151105 (2005).

[50] B.C. Stipe *et al*, "Magnetic recording at 1.5 Pb m^{-2} using an integrated plasmonic antenna", Nature Photonics 4, 484-488 (2010).

[51] K. Sendur *et al*, "Ridge waveguide as a near field aperture for high density data storage", PRL 96, 2743 (2004).

[52] X. Shi *et al*, "A nano-aperture with 1000x power throughput enhancement for very small aperture laser system (VSAL)", Proc. SPIE, Vol. 4324, 320-27 (2002).

[53] W.A. Challener *et al*, "Light delivery for heat-assisted magnetic recording", Jpn. J. Appl. Phys., Vol. 42, 981-988 (2003).

[54] W.A. Challener *et al*, "Optical transducers for near field recording", Jpn. J. Appl. Phys., Vol. 45, no. 8B 6632-6642 (2006).

[55] M. Staffaroni, "A plasmonic transducer for near field recording", M.S. Thesis (2008).

[56] A. Bondeson, T. Rylander, P. Inlesrom, *Computational Electromagnetics*, Springer Texts in Applied Mathematics, Vol. 51 (2005).

[57] A. Taflove, S.C. Hagness, *Computational Electrodynamics, the Finite-Difference Time-Domain Method*, Artech House (2005).

[58] O.C. Zienkiewicz, K. Morgan, *Finite Elements & Approximation*, Dover (1983).

[59] C. Peng *et al*, "Input-grating couplers for narrow Gaussian beam: influence of groove depth", Optics Express, Vol. 12, No. 26, 6481 (2004).

[60] W.A. Challener *et al*, "Heat assisted magnetic recording with heat profile shaping", US 2005/0041950 A1 (2005).

[61] E. Wolf, "Electromagnetic diffraction in optical systems I. An integral representation of the image field", Proc. Royal Soc. London, Series A, Mathematical and Physical Sciences, 349 (1959).

[62] B. Richards, E. Wolf, "Electromagnetic diffraction in optical systems II. Structure of the image field in an aplanatic system", Proc. Royal Soc. London, Series A, Mathematical and Physical Sciences, 358 (1959).

[63] M.L. Boas, *Mathematical Methods in the Physical Sciences*, Wiley (2006).