Transmission of Light Through Small Elliptical Apertures (Part 1)

Masud Mansuripur, Armis R. Zakharian, and Jerome V. Moloney

The apertures of classical optics simply block those parts of an incident wavefront that fall outside the aperture, allowing everything else to go through intact. Moreover, multiple apertures act upon an incident beam independently of each other, polarization effects are usually negligible (i.e., scalar diffraction), and it is not necessary to keep track of both the electric- and the magnetic-field components of the beam.

All of the above assumptions break down when apertures shrink to dimensions comparable to or smaller than a wavelength. For example, transmission through two small adjacent apertures cannot be treated by assuming that only one aperture is open at a time, then adding the fields transmitted by the individual apertures. (This is because the electric charge and current distributions in the vicinity of one aperture are influenced by the radiation pattern of the other aperture.) Polarization effects are extremely important for small apertures as exemplified by the case of a normally incident beam, even when it goes through an elliptical aperture in a thin metal film: whereas in the case of polarization (i.e., E-field) parallel to the long axis of the ellipse there is negligible transmission, when the incident polarization is rotated 90° to point along the ellipse’s minor axis, the aperture transmits a substantial fraction of the incident light. Finally, to analyze the interaction of light with small apertures, it is generally necessary to keep track of both E and B components of the electromagnetic wave, as the modification of one of these fields produces non-trivial changes in the other field’s distribution.

This paper presents the results of computer simulations based on the Finite Difference Time Domain (FDTD) method for an elliptical aperture in a thin metal film illuminated by a normally incident, monochromatic plane wave. Both cases of incident polarization parallel and perpendicular to the long axis of the ellipse will be considered. We begin by developing an intuitive description of the behavior of the electromagnetic fields in each case, then present simulation results that exhibit patterns similar to those expected from this qualitative analysis. The simulations reveal, in quantitative detail, the amplitude and phase behavior of the E- and B-fields in and around the aperture.

Maxwell’s equations

In developing an intuitive understanding of the electromagnetic field distribution around an aperture, we rely heavily on the Maxwell’s divergence equations, $\nabla \cdot \mathbf{D} = \rho$ and $\nabla \times \mathbf{B} = 0$, where $\mathbf{D} = \varepsilon_0 \varepsilon \mathbf{E}$, $\mathbf{B} = \mu_0 \mathbf{H}$, and $\rho$ is the electric charge density. ($\varepsilon_0$ and $\mu_0$ are the permittivity and permeability of free-space, while $\varepsilon$ is the relative permittivity of the local environment.) The divergence-free nature of the magnetic field simply means that the B-field lines cannot be interrupted; they can go around in loops or they can form unbroken infinite lines, but they cannot originate, nor can they terminate, at specific points in space. A similar argument applies to the E-field lines, except in locations where electric charges exist. When charges are present, lines of D originate on positive charges and terminate on negative charges; everywhere else the D-lines can twist and turn in space, but they cannot start or stop.

The other two of Maxwell’s equations, $\nabla \times \mathbf{H} = \mathbf{J} + \partial \mathbf{D} / \partial t$ and $\nabla \cdot \mathbf{E} = \mathbf J - \partial \mathbf{B} / \partial t$, are necessary not only for generating the E and B fields from electrical currents (J is the local current density), but also to sustain these fields in source-free regions of space. When highly conducting media (e.g., metallic bodies) are present in a system, surface currents I, develop that support the magnetic field H immediately above the conducting surfaces. Aside from these electrical currents that act as sources of the H-field, time variations of the E-field are needed at each point of space to maintain the local B-field. In a similar vein, aside from electric charges that act as sources and sinks for the D-field, time variations of B are necessary to maintain the local E-field. The lines of the current density J remain divergence-free, except in those locations where they deposit electrical charges, that is, $\nabla \cdot \mathbf{J} = - \varepsilon_0 \partial \mathbf{E} / \partial t$.

Inside an electrical conductor $\mathbf{J} = \sigma \mathbf{E}$, where $\sigma$ is the conductivity of the material. Good conductors (e.g., metals) have large conductivities, which means that the E-field must all but vanish from the interior of such bodies. When the fields are oscillatory, any magnetic fields inside a good conductor will produce, by virtue of the Faraday law, $\nabla \times \mathbf{E} = - \partial \mathbf{B} / \partial t$, a local electric field. Since E-fields are not allowed inside a conductor, time-varying...
magnetic fields, being intimately associated with the electric fields, must also be absent. The interior of good conductors thus remains free of charges, currents, and time-varying electromagnetic fields. Charges and currents, however, can and do develop on the conductor’s surface, where they give rise to \( \mathbf{E} \) and \( \mathbf{B} \) fields in the vicinity of the surface outside the conductor.

The fifth equation of classical electrodynamics, the Lorentz law of force \( \mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \), expresses the force \( \mathbf{F} \) experienced by a particle of charge \( q \) and velocity \( \mathbf{v} \). This equation is occasionally useful in developing a qualitative picture of the current distribution in the vicinity of small apertures. For example, within the skin depth of a conductor, the directions of \( \mathbf{E} \) and \( \mathbf{B} \) would indicate the sense in which local surface currents are affected by the Lorentz force acting on the charge carriers. Typically, the \( \mathbf{E} \)-field is the dominant factor in this regard, as evidenced by the constitutive relation \( \mathbf{J} = \sigma \mathbf{E} \). Any transverse deflections of the current by the \( \mathbf{B} \)-field are generally neglected, unless the Hall conductivity of the medium is explicitly included in the constitutive relations.

**Radiation by an oscillating dipole**

With reference to Fig. 1(a), a static electric dipole \( \mathbf{p} \) creates, in its surrounding environment, electric-field lines that emerge from the positive pole and disappear into the negative pole. A periodically oscillating electric dipole emanates \( \mathbf{E} \)-field lines that reverse direction at half-period intervals. The constant speed of light in all directions in space then dictates that these \( \mathbf{E} \)-field reversals occur on spherical shells separated by a half-wavelength \( (\lambda/2) \) from their adjacent shells. The zero-divergence requirement imposed on the \( \mathbf{E} \)-lines by the first Maxwell equation thus requires the existence of the closed lines of field depicted in Fig. 1(b). The curl of the \( \mathbf{E} \)-field gives rise to \( \mathbf{B} \)-field lines that encircle the dipole in closed loops, sustaining the \( \mathbf{E} \)-field oscillations while simultaneously being generated by them. In the space between adjacent spherical shells separated by \( \lambda/2 \), the \( \mathbf{E} \)-lines are not parallel to these shell surfaces, but bend inward or outward as to maintain the divergence-free condition of the \( \mathbf{E} \)-field.

A static magnetic dipole \( \mathbf{m} \), shown in Fig. 1(c), is a closed loop of electrical current whose \( \mathbf{B} \)-field pattern is similar to the \( \mathbf{E} \)-field of an electric dipole. Figure 1(d) shows an oscillating magnetic dipole, which behaves in much the same way as an electric dipole does, albeit with a role reversal for \( \mathbf{E} \) and \( \mathbf{B} \). These examples indicate that by direct appeal to Maxwell’s equations, especially the divergence laws, it is possible to obtain an intuitive picture of the electromagnetic field distribution. In the discussions that follow, we will use the dipole radiation patterns sketched in Figs. 1(b) and 1(d) to elucidate the nature of transmission through subwavelength apertures in a thin metal film.

**Plane wave reflection from a (highly conducting) flat mirror**

Figure 2 shows the case of a normally incident plane wave on a perfect conductor (yellow slab at the bottom). The incident beam induces a surface current \( I_s \) in the conductor, which creates equal-amplitude plane waves propagating in the \( \pm Z \)-directions. In the half-space below the conductor, the induced and incident plane waves cancel each other out. In the half-space above the conductor, interference between the incident and reflected beams creates standing-wave fringes of the electric-field \( \mathbf{E} \) and the magnetic field \( \mathbf{B} \). The \( \mathbf{B} \)-field is strongest at the surface of the conductor, reversing sign at intervals of \( \Delta Z = \lambda/2 \), where its adjacent peaks are located. The peaks of the \( \mathbf{E} \)-field, also located at \( \lambda/2 \) intervals, are staggered relative to the \( \mathbf{B} \)-field peaks, thus coinciding with planes of vanishing magnetic field.

At the upper surface of the conductor, where the \( \mathbf{E} \)-field is zero, the \( \mathbf{B} \)-field is sustained by the surface current \( I_s \). (Although \( I_s \) is shown antiparallel to the standing-wave’s \( \mathbf{E} \)-field at \( \Delta Z = \lambda/4 \), in reality it is 90° behind this \( \mathbf{E} \)-field, reaching maximum when the \( \mathbf{E} \)-field directly above the surface is going through zero on its way to the peak.) In the half-space above the conductor, in the absence of any electrical charges and currents, the \( \mathbf{E} \)-field is sustained by the time-variations of the \( \mathbf{B} \)-field (\( \nabla \cdot \mathbf{E} = -\partial \mathbf{B}/\partial t \)), and vice versa (\( \nabla \cdot \mathbf{H} = \partial \mathbf{D}/\partial t \)).

In an imperfect conductor, where conductivity is large but finite, the \( \mathbf{E} \)- and \( \mathbf{B} \)-fields penetrate slightly beneath the surface, producing a Lorenz force on the moving charges that comprise the surface current. While the \( \mathbf{E} \)-field provides the current’s driving force, the magnetic component of the Lorenz force attempts to drive the surface current further down into the conductor (radiation pressure). In general, the surface current \( I_s \) need not be in-phase with the penetrating \( \mathbf{E} \)-field, since, at optical frequencies, the electrical conductivity \( \sigma \) is a complex number.

**Elliptical aperture illuminated with plane wave polarized along the long axis**

The presence of a small (subwavelength-sized) elliptical aperture in the system of Fig. 2 distorts the surface current \( I_s \) in the vicinity of the aperture by diverting the current’s path to avoid the hole, as shown in Fig. 3. The \( \mathbf{B} \)-lines within the fringe immediately above the mirror surface reorient in such a way as to remain perpendicular to the line of \( I_s \), thus bending toward the center of the aperture. The \( \mathbf{B} \)-lines directly above the aperture, lacking support from an underlying surface current, drop into the hole on the left side and re-emerge on the right side. (The \( \mathbf{B} \)-lines, of course, cannot break up because \( \nabla \cdot \mathbf{B} = 0 \) everywhere; they can only bend locally and change direction, but must remain continuous at all times.)
The lines of surface current $I_s$, that begin and end on the ellipse's sharp corners, deposit electric charges around these corners; these charges then act as sources and sinks for the $E$-lines in their neighborhood. Elsewhere, lack of any significant amount of charge means that the $E$-lines cannot break up, but rather they must twist and turn continuously as they adjust to the new environment created by the presence of the hole. The $E$-field in and around the aperture must be distributed in a way that would support the $B$-field (through the curl equations), but, because a parallel $E$-field cannot exist on conducting surfaces, it must also stay away from the interior walls of the hole. Figure 3 shows a possible way for the $E$-lines just above the aperture to dodge the side-walls and concentrate near the center, as they drop into the hole from above. The bundle of $E$-lines in the middle of the hole (parallel to the ellipse's long-axis) then acts as a source of circulating magnetic fields that wrap around the long axis ($\nabla \times \mathbf{H} = \partial \mathbf{D} / \partial t$), thus supporting the $B$-field above, below, and inside the aperture.

Figure 4(a) shows that, in the central XZ cross-section of the aperture, the $B$-lines above the aperture, without breaking up, thin down and sag toward and into the hole. Magnetic energy thus leaves the mid-section of the strong $B$-fringe above the hole and leaks into the hole and beyond. The behavior of the $E$-field in the central YZ-plane is depicted in Fig. 4(b). Here the strong fringe, which is not immediately above the aperture but a distance of $\Delta z = \lambda/4$ away, is squeezed laterally toward the hole's center, while, at the same time, leaking some of its energy into the aperture. Some of the $E$-lines originate or terminate on the charges deposited by the surface current $I_s$ on the sharp corners of the ellipse. (The dashed lines in Fig. 4(b) represent the bending of the $E$-field out of the YZ-plane toward charges that reside on the side-walls near these sharp corners.) Note that the charge polarity is such that the $E$-lines above have the same direction as those inside and below the aperture. It is important to recognize that the surface current $I_s$ lags $90^\circ$ behind the $E$-field of the first fringe. Thus, when the $E$-field directly above the aperture reaches its maximum along the negative Y-axis, $I_s$, which has been traveling in the positive Y-direction until that moment, has stopped and is beginning to reverse direction. This explains why the charges reach their maximum strength when the $E$-field immediately above the aperture is at a peak, and also clarifies the reasoning behind the polarity chosen for the charges in Fig. 4(b).

Aside from the incident beam, which is fixed at the outset, all other radiation in the system of Fig. 3 is generated by the surface currents $I_s$ (and the charges deposited by $I_s$ around the sharp corners of the aperture). The same is true of the system of Fig. 2, with its uniform current confined to the upper surface of the conductor. Any differences between the radiation fields in the systems of Fig. 2 and Fig. 3 must therefore arise from the difference between the two surface current distributions. Subtracting the (uniform) surface current of Fig. 2 from that of Fig. 3 yields the distribution sketched in Fig. 5(a). Far from the aperture, of course, the perturbation caused by the aperture is small and the two surface currents must cancel out. In the vicinity of the aperture we find two loops of current circulating in opposite directions, as well as positive and negative charges in those regions where the divergence of the local current is non-zero. As shown in Fig. 5(b), these circulating currents are equivalent to a pair of oppositely oriented magnetic dipoles $+\mathbf{m}$ and $-\mathbf{m}$ (i.e., a magnetic quadrupole, assuming their separation is much less than a wavelength); the charges localized on the aperture's sharp corners give rise to an oscillating electric dipole $\mathbf{p}$. Thus, adding the dipoles $\mathbf{p}$ and $\pm \mathbf{m}$ to the system of Fig. 2 should transform it over to the system of Fig. 3.

Figure 3. A small elliptical aperture in the system of Fig. 2, with its major axis parallel to the surface current $I_s$, distorts the current distribution by diverting its path to avoid the hole. The $B$-lines immediately above the surface bend toward and into the aperture, without breaking up. The $E$-field in and around the aperture gets redistributed in a way that supports the $B$-field while staying away from the long side-walls of the hole. The surface currents in the vicinity of the aperture deposit opposite charges around the sharp corners of the ellipse, causing the $E$-lines to break up at these corners.
Similarly, superposition of the nature of Fig. 3 seems to be the ± Fig. 5(b); the induced dipole port the interior side-walls of the aperture; the currents also help to sup-
dipoles is a consequence of the induced surface currents on the
their and the other to the left, so that everywhere inside the metal film 
the magnetic dipoles may have to tilt sideways, one to the right 
Fig. 5(b) must vanish within the body of the film. T o this end, 
immediately above the aperture.

Figure 5(c) shows that, in the vicinity of the aperture, the combined radiation pattern of the electric dipole and the magnetic quadrupole consists of a circulating B-field around the major axis of the ellipse, and an E-field pattern that tends to stay away from the long side-walls of the aperture.

In practice, the metallic film has a finite thickness, and the combined radiation by the dipole p and quadrupole ±m of Fig. 5(b) must vanish within the body of the film. To this end, the magnetic dipoles may have to tilt sideways, one to the right and the other to the left, so that everywhere inside the metal film their E- and B-fields will be cancelled by the corresponding fields of the electric dipole. Physically, the sideways tilt of the ±m dipoles is a consequence of the induced surface currents on the interior side-walls of the aperture; the currents also help to support the B-field adjacent to these side-walls; see Fig. 4(a).

All in all, the primary source of radiation through the aperture of Fig. 3 seems to be the ±m quadrupole depicted in Fig. 5(b); the induced dipole p in this system is relatively weak and plays a secondary role, namely, canceling the quadrupole's radiation inside the metal film. In general, quadrupolar sources are weak radiators, thus accounting for the weakness of transmission through an elliptical aperture illuminated by a plane wave whose polarization direction coincides with the major axis of the ellipse.

Figure 6 shows computed plots of Ex, Ey, Ez in an XY-plane located 20 nm above the surface of the conductor in the system of Fig. 3. The simulated conductor is a 124 nm-thick film of silver (n+iκ = 0.226 + i 6.99 at λ = 1.0 μm) having an 800 nm-long, 100 nm-wide elliptical aperture. The magnitude of each field component is plotted in the top row of Fig. 6, and the corresponding phase profile appears below it. For our purposes, the main utility of the phase distribution is to indicate the relative orientation of the various field components. For instance, if the phase of Ey at a given location happens to be φo, then if the phase of Ex at that location turns out to be equal (or nearly equal) to φo, we will know that Ex + Ey oscillates back and forth between the first and third quadrants of the XY-plane. However, if the phase of Ey hovers around φo ± 180°, then Ex + Ey oscillates between the second and fourth quadrants.

The E-field distribution of Fig. 6 is consistent with the qualitative behavior sketched in Figs. 3, 4(b), and 5(c). The Ez component bends the central field lines toward the middle of the aperture, and pushes the peripheral lines further away, thus ensuring that the long side-walls repel the parallel E-field. The Ez component is strengthened near the center of the aperture because the field lines are pushed upward and squeezed laterally toward the center. Finally, the Ez component confirms the presence of charges of opposite sign at and around the sharp corners of the aperture (V • D). These pictures are consistent with the presence of a weak electric dipole and a magnetic quadrupole in and around the elliptical aperture.
Computed amplitude and phase plots of $E_y$, $E_z$ in the central YZ-plane of the aperture are shown in Fig. 7. The bands of $E_y$ above the aperture are the standing-wave fringes created by the interference between the incident and reflected beams. The weak nature of transmission through the aperture is evident in the very small perturbation of the fringes, as they sag ever so slightly to fill the top of the aperture. The profile of $E_z$, once again, confirms the accumulation of electric charges around the sharp corners of the hole. Moreover, it shows that the charges on the top facet of the metal film, while much stronger than those on the bottom facet, have the same sign as the charges on the bottom; in other words, the top and bottom charges are both positive at one end of the ellipse, and both negative at the opposite end.

Figure 8 shows plots of $H_x$, $H_y$, $H_z$ in the XY-plane 20 nm above the surface of the conductor, while amplitude and phase plots of $H_x$ and $H_z$ in the central XZ-plane appear in Fig. 9.
As expected from the preceding discussion of Figs. 3 and 4, the magnetic fringe nearest the surface is seen to leak into the aperture by bending the **H**-lines near the corners of the ellipse toward the center and down into the hole.

Computed plots of **E**<sub>x</sub>, **E**<sub>y</sub>, **E**<sub>z</sub> in the XY-plane 20 nm below the conductor are shown in Fig. 10, and the corresponding **H**-field distributions appear in Fig. 11. While the profiles of these fields confirm the behavior expected from our earlier qualitative analysis, their small magnitudes testify to the weak nature of radiation by the ±m quadrupole (and the accompanying p dipole) induced by the incident beam in the vicinity of the aperture of Fig. 3.

Figure 12 shows distributions of the magnitude |**S**| of the Poynting vector in various cross-sections of the system of Fig. 3. The superimposed arrows on each plot show the projection of **S** in the corresponding plane. For instance, in the XZ cross-section depicted in Fig. 12(a), the arrows represent **S**<sub>y</sub>, **S**<sub>x</sub> + **S**<sub>z</sub>, whereas in the YZ cross-section of Fig. 12(b) the arrows correspond to the projection **S**<sub>y</sub>, **S**<sub>x</sub> + **S**<sub>z</sub> of the Poynting vector on the YZ-plane. The plots in Figs. 12(c) and (d) show the distributions of |**S**| in the XY-planes immediately above and below the aperture. In the absence of an aperture, **S** is essentially zero everywhere, as the reflected beam cancels out the incident beam’s energy flux. When the aperture is present, however, the fields are redistributed in such a way as to draw the incident optical energy toward the aperture. In the present case, the energy flow in from the periphery, fails to find a way through the aperture, bounces back and returns toward the source in the region directly above the aperture. In the process, several vortices are formed, where the incoming energy makes a sharp turnaround and heads back toward the source.

In Fig. 12(d) the Poynting vector **S** = |**S**|Re(**E**⋅**H**) at the bottom of the hole has a magnitude |**S**| ~ 2.5 × 10<sup>-6</sup> W/m<sup>2</sup>, consistent with the transmitted **E**-field of ~0.06 V/m and **H**-field of ~2.3 × 10<sup>-4</sup> A/m, considering the large phase difference of Δφ ~ 70° between the **E**- and **H**-fields near the bottom of the aperture; see Figs. 10 and 11. Since the incident plane-wave is assumed to have **E**<sub>0</sub> = 1.0 V/m, **H**<sub>0</sub> = |**E**<sub>0</sub>|/|**Z**<sub>o</sub>| = 2.65 × 10<sup>-3</sup> A/m (free-space impedance |**Z**<sub>o</sub>| ~ 377Ω), which correspond to an incident energy density ~1.32 × 10<sup>-3</sup> W/m<sup>2</sup>, the power transmission efficiency η at the center of the elliptical aperture of Fig. 3 is seen to be just under 0.2%. We will see next month, in part 2 of this column, that when the incident polarization is rotated 90° (to point along the minor axis of the ellipse), the transmission efficiency through the aperture increases to η ~ 93%, a nearly 500-fold improvement.

References:
7. The computer simulations reported in this article were performed by Sim3D, a product of MM Research, Inc., Tucson, Ariz.