The physics of light transmission through subwavelength apertures and aperture arrays

This article has been downloaded from IOPscience. Please scroll down to see the full text article.
(http://iopscience.iop.org/0034-4885/72/6/064401)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 89.133.195.186
The article was downloaded on 23/06/2012 at 09:48

Please note that terms and conditions apply.
The physics of light transmission through subwavelength apertures and aperture arrays

J Weiner

CePOF/IFSC/Universidade de São Paulo Av. Trabalhador São-carlense, 400 13566-590 São Carlos, SP, Brazil

E-mail: johweiner@gmail.com

Received 12 December 2008, in final form 10 March 2009
Published 27 May 2009
Online at stacks.iop.org/RoPP/72/064401

Abstract

The passage of light through apertures much smaller than the wavelength of the light has proved to be a surprisingly subtle phenomenon. This report describes how modern developments in nanofabrication, coherent light sources and numerical vector field simulations have led to the upending of early predictions from scalar diffraction theory and classical electrodynamics. Optical response of real materials to incident coherent radiation at petahertz frequencies leads to unexpected consequences for transmission and extinction of light through subwavelength aperture arrays. This paper is a report on progress in our understanding of this phenomenon over the past decade.

(Some figures in this article are in colour only in the electronic version)

Contents

1. Introduction 1
2. Early theories 2
   2.1. Kirchhoff scalar diffraction 2
   2.2. Bethe electromagnetic diffraction theory of small holes 3
   2.3. Kowarz scalar diffraction theory 3
3. Modern experiments and interpretations: claims and counter claims 4
4. Waves in metals 6
   4.1. Early history 6
   4.2. Electron gas model in 1D 6
4.3. Surface plasmon polaritons 7
4.4. Surface waves in subwavelength transmission 9
4.5. Résumé of surface waves and subwavelength transmission 14
4.6. Bloch mode analysis and FDTD simulations 15
4.7. Proposed physical explanation 16
5. Where are we and where are we going? 17
   5.1. What do we know? 17
   5.2. What do we not (yet) know? 17
   5.3. Prospects for the future 17

References 18

1. Introduction

Suppose we place an opaque, black, negligibly thick screen, aligned along an x–y plane, at the position \( z = 0 \) as indicated in figure 1(a). And further suppose that we send a plane wave incident on the screen, propagating along \( z \) from the left. What happens on either side of the screen? Because the screen is opaque and black there is no transmission and no reflection. Nothing happens. Now we open a small aperture in the screen, as shown in figure 1(b). Something happens; light transmits to the right side of the screen through the opening, and we ask the following question: What is the distribution of wave amplitude, direction, and phase on the right side of the screen? This seemingly simple and innocuous question has been the subject of over a hundred years of study, discussion and controversy!

This paper recounts efforts to find the answers first using simple scalar diffraction, then electromagnetic
vector fields and finally applying modern methods of numerical simulation to spatially map the vector field solutions of Maxwell’s equations. Along the way we will describe recent measurements of light transmission at optical wavelengths through apertures (slits and holes) of subwavelength dimension, fabricated with modern methods of nanotechnology. We will see that the ‘black, negligibly thick screen’ will have to be replaced by a real, dispersive material characterized by frequency-dependent permittivity and permeability, and that so doing opens up new answers to the original question of what the fields look like on the right side of the screen. These new answers include modes evanescent in z but propagating at the surface along x and y. They play a critical role in transmission and transmission \textit{extinction} through periodic aperture arrays milled into metal films. Finally we will find that despite huge advances in understanding and technology since the time of Gustav Kirchoff, there is still no universally agreed consensus on \textit{THE} answer to the original question. Since much of the current interest involves the interaction of light with subwavelength-nano-structures, an introduction to nano-optics in general \cite{1, 2}, and plasmonics in particular \cite{3}, can serve as general entry points to research in these areas.

2. Early theories

2.1. Kirchhoff scalar diffraction

Consider again figure 1(b). We seek a solution to the scalar wave equation in the hole at \( z = 0 \) and in the space on the right side of the black, opaque screen. We assume the solution will have the form \( \Psi(r, t) = \psi(r, t)e^{-i\omega t} \), where as usual \( r^2 = x^2 + y^2 + z^2 \), \( \omega \) is the wave frequency and \( t \) the time. The waves propagate in free space so that \( \omega = 2\pi v, \nu = c/\lambda_0 \) and \( k_0 = 2\pi/\lambda_0 \) with \( c \) the speed of light in free space, \( \lambda_0 \) the wavelength and \( k_0 \) the propagation parameter. The scalar Kirchhoff wave equation is

\[
\nabla^2 \psi + k_0^2 \psi = 0. \tag{1}
\]

The standard procedure to find suitable solutions is first to invoke Green’s theorem and write

\[
\psi(r) = \frac{1}{4\pi} \int_S \left[ \frac{\partial \psi}{\partial n} e^{ik_0 r} - \psi \frac{\partial}{\partial n} \left( \frac{e^{ik_0 r}}{r} \right) \right] ds, \tag{2}
\]

where \( S \) is the entire screen surface, imagined to extend to infinity along \( \pm x, \pm y \). Actually one must also imagine that the space on the right is closed by a surface joining the two ends of the screen as it extends to infinity. Then the sign of the normal \( n \) is positive for \( n \) pointing out of the enclosed volume at the surface. According to the theory of differential equations, if the values of \( \psi \) or if the values of \( \partial \psi / \partial n \) are known on \( S \), they are determined for all points interior to the surface (right side of screen). Because the screen is opaque we can set the value of \( \psi = 0 \) on its left side, but the small opening poses a problem. We do not know \textit{a priori} the value of \( \psi \) at the hole. The Kirchhoff approach to this problem is to assume an approximate solution with \( \psi = 0 \) and \( \partial \psi / \partial n = 0 \) on the right side of the screen, but at the hole \( \psi = \psi_0 \), the value of the unperturbed incident wave. Despite the fact that it is not entirely legitimate to invoke Green’s theorem for the solutions to equation (1) if discontinuities appear on the bounding surface (here we assume \( \psi = 0 \) on the screen and \( \psi = \psi_0 \) in the hole), this approach gives very good results as long as the characteristic dimension of the opening is large compared with \( \lambda_0 \). The reason is that most of the diffracted wave is distributed around the forward propagating direction along \( +z \) and the assumptions of \( \psi = 0 \) and \( \partial \psi / \partial n = 0 \) on the right side of the screen are largely respected. However, as the aperture becomes small compared with \( \lambda_0 \), the diffraction lobes bend over more and more toward the screen, and these assumptions lose their validity. Another more fundamental problem is that the wave solutions in which we are really interested must be solutions to Maxwell’s equations, a coupled set of vector field equations, not the scalar equation (1). This is not just a matter of applying the Kirchhoff equation to the E- and H-field components along each of the \( x, y, z \) coordinates. As was stated by Bouwkamp \cite{4}, ‘If the scalar Kirchhoff
2.2. Bethe electromagnetic diffraction theory of small holes

In an effort to address the shortcomings of Kirchhoff scalar diffraction, Bethe published in 1944 a new theory of electromagnetic wave diffraction by a small round hole in a perfectly conducting metal screen [5]. The paper was received at the Physical Review in January 1942 but not published until October 1944. A perfectly conducting, negligibly thick screen, while opaque, is different from our earlier screen. On the left side now appears incident and reflected light as well as transmitted waves through the hole. ‘Perfectly conducting’ means that the conductivity of the metal is infinite, that the incident \( E-M \) field does not penetrate the surface (no skin depth), and that the \( E \)-field component tangent to the screen and the \( H \)-field perpendicular to it vanish at the surface. Bethe’s strategy was first to find the \( E \)- and \( H \)-fields in the hole, using continuity and boundary conditions, and relying on the field amplitudes being essentially constant over the area of the hole. He found

\[
H_i(\text{hole}) = \frac{1}{2} H_0(x, y, z = 0...) \quad \text{and} \quad E_\perp(\text{hole}) = \frac{1}{2} E_0(x, y, z = 0...) \tag{3}
\]

where \( E_0, H_0 \) are the electric and magnetic field amplitude components of the incident wave impinging on the screen from the left, indicated by the \( z = 0... \) coordinate. With these in hand, the task was to find \( E(r), H(r) \) on the right side of the screen that satisfied Maxwell’s equations, was consistent with the fields in the hole and the ‘perfect conductor’ boundary conditions, \( E_\parallel(x, y, 0) = H_\perp(x, y, 0) = 0 \), on the right side of the screen. Since these boundary conditions must be strictly obeyed on the screen surface but not in the hole, we expect discontinuities in \( H_\perp \) and \( E_\parallel \) as we pass from the screen to the hole. The fields in the hole and the resulting field discontinuities at the hole boundary can be generated by fictitious magnetic charges and currents obeying Maxwell-like equations. The usual Maxwell–Faraday and Maxwell–Ampère laws (expressed in MKS units) transform such that

\[
\nabla \times E + \mu \frac{\partial H}{\partial t} = 0 \rightarrow \nabla \times E + \mu \frac{\partial H}{\partial t} = -J^*, \tag{4}
\]

\[
\nabla \cdot H = 0 \rightarrow \nabla \cdot H = \frac{1}{\mu_0} \rho^*, \tag{5}
\]

where \( J^*, \rho^* \) are the magnetic current and charge densities, respectively, and the magnetic permeability \( \mu \) is given by \( \mu = \mu_0 \mu_r \). In the last expression \( \mu_0 \) is the permeability of free space and \( \mu_r \) is the unitless relative permeability. As Bethe airily states [5], ‘It need hardly be pointed out that \( J^* \) and \( \rho^* \) have no physical meaning.’ The result of the calculation that is relevant here is the power through the hole integrated over the \( 2\pi \) solid angle into which the incident flux \( S_i \) can diffract.

\[
\frac{P_{\text{total}}}{\sigma_{\text{eff}}} = \frac{64}{27\pi} k^4 a^6 S_i \quad \text{with} \quad S_i = \frac{1}{2} c \varepsilon_0 E_i^2. \tag{6}
\]

The factor with the bracket underscore in equation (6) can be interpreted as an effective cross section \( \sigma_{\text{eff}} \) of the physical hole with area \( A = \pi a^2 \). We see that for holes much smaller than the wavelength of the light, \( ka \ll 1 \), the effective cross section for light transmission rolls off as \( \lambda^{-4} \). Kirchhoff scalar diffraction only falls off as \( \lambda^{-2} \); so for a given subwavelength hole Bethe’s result predicts much less light transmission. In a later review Bouwkamp [4] pointed out that, while the Bethe result is correct in the far field, the expressions for the \( E \)- and \( H \)-fields in the near field are in error, and he has proposed corrected near-field expressions. Since measurements of the transmitted intensity are almost always carried out in the far field we will not explore these near-field differences further. Note the extreme sensitivity of power transmission to the size of the hole in equation (6): it increases as the 6th power of the hole radius. We shall see later how this dependence figures importantly in later controversies over ‘extraordinary optical transmission’ (EOT).

Both the Kirchhoff and Bethe theories posit ideal material characteristics for the boundary-defining screen. Kirchhoff theory assumes a material opaque and perfectly black, while Bethe starts from an opaque, perfectly conducting boundary. In both cases only propagating waves carry diffracted power, and surface modes evanescent in \( z \) but propagating along the surface in \( x, y \) are excluded. In real metals we will see that surface waves play a very important role in light transmission, so it is worthwhile to consider propagating and evanescent modes within a relatively simple two-dimensional (2D) scalar diffraction theory. In 2D the scalar wave equations are equivalent to those derived from the electromagnetic theory, and so the results indicate what happens on real illuminated slits.

2.3. Kowarz scalar diffraction theory

The setup [6] of the problem is similar to figure 1 except that the round hole is replaced by a slit with long axis aligned along \( y \) and the waves emanating to the right are cylindrical rather than spherical. The screen is still an ideal opaque, perfectly black, negligibly thick material, and Kowarz invokes the Kirchhoff ‘approximate’ boundary conditions of zero field on the screen and incident field amplitude in the slit. Although the use of these boundary conditions has been termed a ‘pious hope’ by Bethe and ‘wishful thinking’ by Bouwkamp, the ‘pious hope’ in Kowarz’s study is that, even if the details of the near-field wave solutions are not accurate, the use of the Kirchhoff boundary conditions is good enough to determine the relative importance of evanescent and propagating modes as a function of the slit width aligned on \( x \) and located at \( z = 0 \). The wave equation is similar to equation (1) except \( \nabla^2 \) only spans \( x \) and \( z \).

\[(\nabla^2 + k^2) U(x, z) = 0. \tag{7}\]
The solutions $U(x, z)$ are resolved into homogeneous (propagating) or inhomogeneous (evanescent) waves. An ‘inhomogeneous’ wave in general is one whose surfaces of constant amplitude do not coincide with the surfaces of constant phase. In the particular case of a propagating medium with a real (loss free) index of refraction, an inhomogeneous wave propagates and attenuates in orthogonal directions, and in the setup of figure 1 a wave on the right side of the screen evanescent along $z$ but propagating on the surface $x–y$ is inhomogeneous.

$$U(x, z) = U_h(x, z) + U_i(x, z).$$

These solutions are expanded in a plane-wave basis using the ‘angular spectrum representation’ [7] and introduce $u_x$, a ‘direction cosine’ change of variable for $k_x$ and $k_z$.

$$U_h(x, z) = \int_{|u_x|\leq 1} \frac{1}{k} c(u_x) e^{i k u_x x} e^{i k_x (1-u_x^2)^{1/2} z} dk_x,$$

$$= \int_{|u_x|\leq 1} c(u_x) e^{i k u_x x} e^{i k_x (1-u_x^2)^{1/2} z} du_x,$$ (9)

and

$$U_i(x, z) = \int_{|u_x|> 1} c(u_x) e^{i k u_x x} e^{i k_x (1-u_x^2)^{1/2} z} du_x,$$ (10)

where

$$k^2 = k_x^2 + k_z^2,$$

$$k_x = ku_x,$$

$$k_z = ku_x = k[1 - u_x^2]^{1/2}$$ (11)

In equation (10) the integration runs over $k_x > k$ or $u_x > 1$ so that $k_z$ is pure imaginary and the waves are evanescent in $z$. The coefficients $c(u_x)$ can be determined and equations (9) and (10) rewritten

$$U_h(x, z) = \frac{K}{\pi} \int_{|u_x|\leq 1} \frac{\sin (k u_x a/2)}{u_x} e^{i k u_x x} e^{i k_x (1-u_x^2)^{1/2} z} du_x,$$ (12)

$$U_i(x, z) = \frac{K}{\pi} \int_{|u_x|> 1} \frac{\sin (k u_x a/2)}{u_x} e^{i k u_x x} e^{i k_x (1-u_x^2)^{1/2} z} du_x,$$ (13)

where $K$ is the amplitude of the incident plane wave. We are really interested in the solutions near the right surface of the screen ($z = 0$) and there they can be written

$$U_h(x, 0) = \frac{K}{\pi} \left\{ \text{Si} \left[ k \left(x + \frac{a}{2}\right)\right] - \text{Si} \left[ k \left(x - \frac{a}{2}\right)\right] \right\},$$ (14)

$$U_i(x, 0) = \frac{K}{\pi} \left\{ \text{Si} \left[ k \left(x + \frac{a}{2}\right)\right] - \text{Si} \left[ k \left(x - \frac{a}{2}\right)\right] \right\},$$ (15)

$$|x| > \frac{a}{2},$$

$$U_i(x, 0) = \frac{K}{\pi} \left\{ \pi - \text{Si} \left[ k \left(x + \frac{a}{2}\right)\right] + \text{Si} \left[ k \left(x - \frac{a}{2}\right)\right] \right\},$$ (16)

$$|x| \leq \frac{a}{2}.$$
gives rise to Bloch modes of the skin depth of the metal. Periodic structuring in the metal is clear in [10], but Treacy suspected that the success of ‘dynamic diffraction’ as another way of considering the problem of transmission through periodic holes or slit arrays. Treacy pointed out that the incident optical field oscillating at frequency $\omega$ will induce currents within the skin depth of the metal. Periodic structuring in the metal gives rise to Bloch modes of the $E-M$ field induced in the metal within the skin depth and consistent with this periodicity. Each of these Bloch modes will have an oscillating current associated with it. Treacy then invoked ‘interband scattering’ to distribute energy among these Bloch modes and from there to the propagating modes and surface waves at the aperture exit. Exactly how this redistribution happens is not made entirely clear in [10], but Treacy suspected that the success of ‘dynamic diffraction’ for interpreting x-ray scattering in crystals might be useful for understanding light transmission through these new fabricated periodic structures as well. Three years later Treacy published a much more complete presentation of dynamic diffraction [11] with an extensive introduction that delineated the relation between this approach and earlier interpretations which cast the transmission in terms of a ‘resonant’ excitation of surface plasmon waves. This trail-blazing paper pointed the way forward by emphasizing two critical factors in the proper analysis of the problem: (1) the $E-M$ field present on and just below the surface of a periodically structured metal resolves into Bloch modes that obey the periodic boundary conditions of the structure itself, independently of the wavelength of the incident light. In the simple case of a one dimensional (1D) periodic array (parallel slits or grooves) these modes run parallel to the surface in the metal and in the dielectric medium of the slit. Most of these Bloch modes are evanescently vanishing perpendicular to the boundary well within the metal skin depth. However propagating modes are also present; without them no light transmission would be observed. Linear combinations of all these Bloch modes, propagating and evanescent, define the transmitted $E-M$ field in the structure. (2) The relative mode populations are determined by matching the transmitted field, expanded in the Bloch-mode basis, to the incident and reflected fields at the surface. The matching conditions are the $H$- and $E$-field continuity conditions for field components parallel and perpendicular to the dielectric–metal interface. The complex coefficient of each member of the Bloch-mode expansion determines the relative importance of that mode, and the distribution of coefficient amplitudes is a sensitive function of the slit–slit distance around the minimum and maximum points of transmission. In the language of dynamic diffraction analysis surface plasmon modes are just one member of the Bloch expansion on an equal footing with all the other modes. Of course at array periodicities corresponding to transmission extrema they may emerge as

hole of radius $a$, normalized to the cross section of the hole $A = \pi a^2$, is

$$\frac{P}{A} = \frac{64}{\pi^2 27} k^4 a^4.$$  

(17)

With the experimental parameters of figure 3, the predicted efficiency of power transmission by the Bethe formula is 0.34 per cent, while Ebbesen et al [8] reported to have measured peak efficiencies of more than a factor of two. Therefore the enhancement of transmission over that expected from the conventional Bethe theory is about 600. In a follow-up paper [9] the same group reported peak transmission efficiencies ‘that are about 1000 times higher than that expected for subwavelength holes.’ A three-order-of-magnitude increase over the predictions of the Bethe formula would indeed be an extraordinary optical transmission. These early reports proposed that the transmission enhancements were due to a new phenomenon not taken into account in the scalar diffraction model of Kirchhoff or the electromagnetic ($E-M$) vector-field model of Bethe: the resonant excitation of surface plasmon (sp) waves supported by the periodic hole array structure in the metal film. We shall discuss the physics of the surface waves in general and surface plasmon waves in particular in the next section. About a year later Treacy suggested [10] ‘dynamic diffraction’ as another way of considering the problem of transmission through periodic hole or slit arrays. Treacy pointed out that the incident optical field oscillating at frequency $\omega$ will induce currents within the skin depth of the metal. Periodic structuring in the metal gives rise to Bloch modes of the $E-M$ field induced in the metal within the skin depth and consistent with this periodicity. Each of these Bloch modes will have an oscillating current associated with it. Treacy then invoked ‘interband scattering’ to distribute energy among these Bloch modes and from there to the propagating modes and surface waves at the aperture exit. Exactly how this redistribution happens is not made entirely clear in [10], but Treacy suspected that the success of ‘dynamic diffraction’ for interpreting x-ray scattering in crystals might be useful for understanding light transmission through these new fabricated periodic structures as well. Three years later Treacy published a much more complete presentation of dynamic diffraction [11] with an extensive introduction that delineated the relation between this approach and earlier interpretations which cast the transmission in terms of a ‘resonant’ excitation of surface plasmon waves. This trail-blazing paper pointed the way forward by emphasizing two critical factors in the proper analysis of the problem: (1) the $E-M$ field present on and just below the surface of a periodically structured metal resolves into Bloch modes that obey the periodic boundary conditions of the structure itself, independently of the wavelength of the incident light. In the simple case of a one dimensional (1D) periodic array (parallel slits or grooves) these modes run parallel to the surface in the metal and in the dielectric medium of the slit. Most of these Bloch modes are evanescently vanishing perpendicular to the boundary well within the metal skin depth. However propagating modes are also present; without them no light transmission would be observed. Linear combinations of all these Bloch modes, propagating and evanescent, define the transmitted $E-M$ field in the structure. (2) The relative mode populations are determined by matching the transmitted field, expanded in the Bloch-mode basis, to the incident and reflected fields at the surface. The matching conditions are the $H$- and $E$-field continuity conditions for field components parallel and perpendicular to the dielectric–metal interface. The complex coefficient of each member of the Bloch-mode expansion determines the relative importance of that mode, and the distribution of coefficient amplitudes is a sensitive function of the slit–slit distance around the minimum and maximum points of transmission. In the language of dynamic diffraction analysis surface plasmon modes are just one member of the Bloch expansion on an equal footing with all the other modes. Of course at array periodicities corresponding to transmission extrema they may emerge as
dominant players in achieving the matching conditions; but in general at any arbitrary slit separation many Bloch modes are needed to achieve the required match at the surface. Since much of the controversy surrounding the interpretation of the early experiments involves surface waves, it is worthwhile to digress to a brief description of their nature.

4. Waves in metals

4.1. Early history

Surface waves have engaged the attention of physicists and engineers since the beginning of the twentieth century, soon after the demonstration of wireless telegraphy by Nikola Tesla in 1893. It was obvious that long-distance signaling over the horizon would require some means for wave propagation to follow the curvature of the earth; and electromagnetic surface waves, at the boundary between earth and air (or sea and air) were leading candidates [12, 13]. Much later, in connection with the study of electron energy loss in solids, Pines proposed [14] that this energy loss be interpreted as wave excitation within an electron plasma. He coined the term ‘plasmon’ to indicate that the resulting plasma wave excitation exhibited finite quanta of energy. This plasmon model was then applied to thin metallic foils by Ritchie [15] to treat energy loss and electron momentum scattering, but it was Raether’s book [16] that brought into focus the study of plasmon surface waves at the boundary between a metal electron plasma and a dielectric medium.

It is important to distinguish two terms: ‘surface plasmon’ (sp) and ‘surface plasmon polariton’ (spp). A surface plasmon arises from the collective motion of the electron charge density at a frequency corresponding to the quantized harmonic oscillation of that charge density about the fixed positive charge centers in the metal crystal. This collective motion occurs in the volume as well as at the surface of a metal, but the resonant frequencies are not the same. The surface plasmon frequency \( \omega_{sp} \) and the bulk plasmon frequency \( \omega_p \) are related by

\[
\omega_p = \frac{1}{\sqrt{2}} \omega_{sp}
\]

since the restoring force applied to the electron density at the surface is only half the restoring force experienced by the electron density in the bulk. Excitation of bulk or surface plasmon waves is a collective, resonant process; and the group velocity of the wave approaches zero as the incident driving frequency approaches the resonant frequency. In contrast, a surface plasmon polariton (spp) is a mixed mode of an electromagnetic wave and a heavy-ion vibrational phonon wave. Normally these two modes of motion are only very weakly coupled; but when the phonon frequency approaches the electron oscillation frequency, the mixing of transverse phonon modes and \( E–M \) modes becomes very strong. An interface between the electron plasma gas and a dielectric medium supports these surface propagating waves with finite group velocity, and they can be described by a very simple model of harmonic motion in the electrons and the positively charged heavy ions.

4.2. Electron gas model in 1D

The motion of a scalar wave \( \psi \) with propagation parameter \( k = 2\pi/\lambda \) through any isotropic medium is described by equation (1)

\[
\nabla^2 \psi + k^2 \psi = 0.
\]

If the medium is ‘dispersive’ such that the dielectric constant \( \epsilon \) exhibits a frequency dependence, then \( k \) will also have a frequency dependence

\[
k(\omega) = \frac{\omega}{c} = \frac{\omega}{c} \sqrt{\epsilon(\omega)},
\]

where \( n \) is the index of refraction. The relation between \( k \) and \( \omega \) is termed the ‘dispersion relation’ and a plot of \( k \) versus \( \omega \), the ‘dispersion curve’. In order to obtain the dispersion relation for charge density waves in an electron plasma we begin by writing the dielectric constant in terms of the polarization field of the medium \( \mathbf{P} \) and the driving electric field \( \mathbf{E} \).

Using rationalized MKS units

\[
\epsilon(\omega) = 1 + \frac{\mathbf{P}}{\epsilon_0 \mathbf{E}},
\]

where \( \epsilon_0 \) is the permittivity of free space. The motion of a free electron \( e \) along a direction \( x \) subject to an electric field component \( E_x \) is governed by Newton’s second law,

\[
\frac{d^2 x}{dt^2} = -e E_x
\]

and if we assume that the field oscillates harmonically at frequency \( \omega \) we have

\[
x(t) = x e^{-i\omega t} \quad E_x(t) = E_x e^{-i\omega t}.
\]

Substituting into equation (20) we find

\[
x = \frac{e E_x}{m_e \omega^2} \quad \text{and the dipole } p_x = -e x = -\frac{e^2 E_x}{m_e \omega^2}.
\]

The polarization field \( P_x \) component along \( x \) is the density of dipoles

\[
P_x = -\frac{e^2 n_e}{m_e \omega^2} E_x,
\]

where \( n_e \) is the electron density. We can now write the dielectric constant (equation (19)) as

\[
\epsilon(\omega) = 1 + \frac{n_e e^2}{\epsilon_0 m_e \omega^2}
\]

and identify the ‘bulk plasmon frequency’ with

\[
\omega_{p}^2 = \frac{n_e e^2}{\epsilon_0 m_e}
\]

so that the dielectric constant can be written

\[
\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2}.
\]
The propagation region above the solid line) and the straight ‘light line’ $k = \omega/c$. In the shaded region below $\omega_p$, $k$ is imaginary and the wave penetrating the medium is evanescent. Light propagating from a dielectric and incident on the metal surface reflects at frequencies below $\omega_p$. As $\omega \gg \omega_p$, $k(\omega)$ approaches $k_0 = \omega/c$ asymptotically. The $E$-field oscillation frequency becomes so high that the electron motion can no longer follow it, and the electron plasma medium becomes essentially transparent. 

As $\omega > \omega_p$, $k(\omega)$ approaches zero, and the wavelength of propagation approaches infinity. The electrons throughout the medium oscillate collectively at the plasmon frequency. At frequencies $\omega < \omega_p$, $k(\omega)$ becomes pure imaginary and the wave $\psi = \psi_0 e^{i\omega t}$ becomes evanescent. The wave decays exponentially into the medium and penetrates only to the propagation depth $\delta = k^{-1}$. Figure 4 shows a plot of the bulk plasmon wave dispersion relation with the propagating region above $\omega_p$ and the ‘forbidden region’ below the bulk plasmon resonance where an incident wave only penetrates evanescently into the electron plasma gas.

4.3. Surface plasmon polaritons

In addition to plasmon waves propagating in the electron gas volume, they can also exist at the surface between the plasma and a dielectric. We can obtain the dispersion relations for these surface waves by writing the relevant Maxwell’s equations and the continuity conditions for $E-M$ field incident on the surface. We assume TM (transverse magnetic) polarization because surface waves with TE (transverse electric) cannot exist [17]. TM (TE) polarization means that the magnetic (electric) field component of the incident wave is aligned perpendicular to the plane of incidence. Starting with the two Maxwell curl equations and the constitutive relations appropriate for a dielectric material with dielectric constant $\epsilon$ and a nonmagnetic metal with permeability $\mu_0$, we write

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

where $D$, $E$, $B$, $H$ are the electric, magnetic, displacement and electric fields, respectively; and $H$, $B$ are the magnetic and magnetic induction fields, respectively. With the $E-M$ field incident on the surface as shown in figure 5, we write the field components as

$$H_y = H_0 e^{i(k_x x + k_z z - \omega t)}$$
$$E_x = E_0 e^{i(k_x x + k_z z - \omega t)}$$
$$E_z = E_0 e^{i(k_x x + k_z z - \omega t)}$$

and the continuity conditions at the boundary as

$$H_y^d = H_y^m$$
$$E_x^d = E_x^m$$
$$\epsilon_d E_z^d = \epsilon_m E_z^m$$

The superscripts $d$, $m$ refer to dielectric and metal. From equation (29) in the dielectric half-space we have

$$\frac{\partial H_y^d}{\partial z} = -\frac{\partial D_x^d}{\partial t} = -\epsilon_d \frac{\partial E_x^d}{\partial t}$$
$$\frac{\partial H_x^d}{\partial x} = \frac{\partial D_x^d}{\partial t} = \epsilon_d \frac{\partial E_x^d}{\partial t}$$

and substituting from equation (28)

$$\frac{k_x^d H_x^d}{\epsilon_d \omega} = E_x^d$$
$$\frac{k_y^d H_y^d}{\epsilon_d \omega} = -E_x^d$$

similarly in the metal

$$-\frac{k_x^m H_x^m}{\epsilon_m \omega} = E_x^m$$
$$\frac{k_y^m H_y^m}{\epsilon_m \omega} = -E_x^m$$
At the boundary between the dielectric and the metal the expressions on the left in equations (29) and (30) can be set equal; and we have, together with continuity of the $H$-field,

$$\frac{k^d}{\epsilon_d} H_y^d + \frac{k^m}{\epsilon_m} H_y^m = 0,$$

$$H_y^d - H_y^m = 0.$$

Equations (31) are a coupled set in $H_y^{d,m}$ with nontrivial solution only if the coefficients obey

$$\frac{k_m^m}{\epsilon_m} = -\frac{k_d^d}{\epsilon_d} \quad (32)$$

One each side of the boundary we also have expressions for the conservation of energy.

$$(k_x^d)^2 + (k_y^d)^2 = \left(\frac{\omega}{c}\right)^2 \epsilon_d,$$  

$$(k_x^m)^2 + (k_y^m)^2 = \left(\frac{\omega}{c}\right)^2 \epsilon_m.$$  

Using equation (32) to express the ratio of $k_x^d/k_x^m$ in terms of $\epsilon_d$ and $\epsilon_m$, and dividing equation (33) by equation (34) to eliminate $k_z$ results in the well-known expression for the propagation parameter of the complex propagation parameter $k_x$ for the surface plasmon polariton wave

$$k_x = \left(\frac{\omega}{c}\right) \sqrt{\frac{\epsilon_d \epsilon_m}{\epsilon_d + \epsilon_m}}.$$  

This result is valid assuming the dielectric constant of the metal $\epsilon_m$ is real. In fact all metals are somewhat lossy, and in the optical regime for high-conductivity metals such as gold and silver $\epsilon_m$ exhibits a relatively small imaginary term as well. The real $\epsilon_m$ must be replaced by the complex $\tilde{\epsilon}_m$, and the expression for the complex propagation parameter $k_x$ is then [16]

$$\tilde{k}_x = \left(\frac{\omega}{c}\right) \sqrt{\frac{\epsilon_d \text{Re}[\tilde{\epsilon}_m]}{\text{Re}[\tilde{\epsilon}_m] + \text{Im}[\tilde{\epsilon}_m]}} \left[1 + i \left(\frac{\omega}{c} \frac{\text{Re}[\tilde{\epsilon}_m]}{\epsilon_d + \epsilon_m}\right)^{3/2} \left(\frac{\text{Im}[\tilde{\epsilon}_m]}{2\text{Re}[\tilde{\epsilon}_m]}\right)^{1/2}\right].$$  

The scale length of interest in subwavelength holes and hole arrays is usually no more than $\sim 10 \mu m$, and over this range surface wave resistive losses in the visible and near-IR spectrum for gold and silver films are negligible. Therefore equation (35) is usually adequate.

In order to obtain the dispersion relation for the surface plasmon polaron, we use equation (24) for the frequency dependence of $\epsilon_m$ and substitute it into equation (35). The result is

$$k_x = \left(\frac{\omega}{c}\right) \sqrt{\frac{\omega^2 - \omega_p^2}{2\omega^2 - \omega_m^2}},$$  

and we see that as $\omega \to \omega_p/\sqrt{2}$, $k_{app} \to \infty$; and as $\omega \to 0$, $k_{app} \to k_0 = \omega/c$, the ‘light line.’ We can therefore add to figure 4 a sketch of the dispersion curve for the surface plasmon polaron as indicated in figure 6. So far we have only considered the electronic response of the metal to the incident $E$–$M$ field. But in addition to the conduction electrons, acting as a plasma-gas, the nuclei also experience wave motion in response to the driving field. These waves are called polarization waves, and they propagate either transversely or longitudinally to the direction of propagation [18]. The wave motion in the plasma, induced by an incident $E$–$M$ driving field, combines a conduction electron charge-density wave and the transverse polarization wave of the nuclei in the metal crystal. As $\omega \to \omega_p$, the period of the electron and nuclear wave motion runs over many crystal planes, and the propagation vector $k_{app} \to k_0$, the propagation vector of an $E$–$M$ wave in the dielectric. The wave propagates with linear dispersion very close to the dielectric ‘light line’. In between these two asymptotes the surface wave is mixed polarization and $E$–$M$ wave, and it is this hybrid character that has led to the label, ‘surface plasmon polaron.’ These ideas can be made a little more precise by modifying the dielectric constant of the plasma gas, equation (24), to include the motion of the nuclei of mass $M$.

$$\epsilon_m(\omega) = 1 - \frac{n_e e^2}{\epsilon_0 M \omega^2} - \frac{n_e e^2}{\epsilon_0 m_e \omega^2},$$  

where $\epsilon_m(\omega)$ is the metal-polariton dielectric constant, $n_e = n_e$ the density of nuclei and conduction electrons, $M$ the mass of the nuclei, and $m_e$ the electron mass. The extra term involving the nuclear mass in equation (38) arises by considering that the nuclear motion in the plasma is also harmonically driven by the external $E$–$M$ field and the total force is the sum of the restoring force on the nuclei and on the conduction electrons. By introducing the reduced mass, $1/\mu = 1/M + 1/m_e$, we see that this modified dielectric constant shifts the bulk plasmon resonance and surface plasmon resonance to slightly higher frequencies.
Setting $\varepsilon_{mp}(\omega)$ equal to zero to find the modified plasmon frequency $\omega_{mp}$ and modified surface plasmon frequency $\omega_{msp}$:

$$\omega_{mp}^2 = \frac{n e^2}{\mu} \quad \text{and} \quad \omega_{msp}^2 = \frac{1}{2} \omega_{mp}^2.$$

### 4.4. Surface waves in subwavelength transmission

#### 4.4.1. New theory results and interpretation

The initial suggestion by Ebbesen and coworkers that the unexpectedly high transmission was due to resonant excitation of surface plasmon modes was developed in an early theoretical study by Garcia-Vidal and co-workers [19]. They proposed that the transmission through a slit array in an optically thick metal film was due to two effects: (1) excitation of ‘coupled’ spp modes on both the entrance and exit surfaces of the film and (2) excitation of wave guide resonances within the slits themselves. They used a transfer matrix theory for numerical calculation [20] and an approximate quasianalytic model to extract a physical interpretation from the numerical results. This paper together with a number of subsequent theoretical studies [21–23] advanced the idea that a resonant excitation of Bloch-like spp normal modes on the surface of a periodic subwavelength structure was responsible for the peaks in transmission. This interpretation was called into question by Lezec and Thio [24] who presented new measurements that showed transmission spectra with much more modest enhancements (factors ‘consistently less than 7’) compared with the ‘extraordinary’ transmission enhancements (factors up to 1000) initially claimed by Ebbesen et al [8,9]. These new measurements showed that the transmission profile in subwavelength periodic hole arrays exhibited transmission suppression as well as enhancement, and Lezec and Thio proposed a new model based on Kowarz’s scalar diffraction study [6] that they called the ‘Composite Diffracted Evanescent Wave’ or the CDEW model. In this alternative CDEW model, waves appeared on the surface due to diffraction by subwavelength holes or slits, but a superposition of these surface modes resulted in a net wave amplitude that decreased as the inverse distance from the diffracting object with an effective range of no more than two or three optical periods. The spp modes, relatively long-range and stable against loss by radiation or phonon coupling, although included in the CDEW superposition, appeared to play no special role. The CDEW model accounted for features in the transmission profile of dielectric subwavelength arrays (where plasmons cannot exist) as well as metallic films, but the model was open to criticism because it was based on scalar diffraction rather than a true $E–M$ vector field diffraction, and the diffracting medium was the conventional opaque, nonreflecting screen of negligible thickness typical of early diffraction theories. The discrepancy in transmission enhancement can be explained by two factors: first, in the initial reports by Ebbesen et al the enhancement was calculated by normalizing to the flux transmitted according to the Bethe formula, equation (17) whereas Lezec and Thio normalized the transmission of an array of slits or holes to the measured transmission from a single array element. Second, in a later publication Thio [25] pointed out that in the initial reports the diameter of the holes was never verified and that in fact subsequent investigation (not detailed in [25]) revealed that the holes were significantly larger than their nominal size. Considering that, according to the Bethe calculation (equation (6)), the transmitted power varies as the sixth power of the hole radius, an underestimate of the true hole size could significantly affect the results. In addition to these new experimental results casting doubt on the central role of surface plasmon polaritons, new theoretical studies suggested a different interpretation for the role they do play. Cao and Lalanne [26], using an approach developed for the analysis of diffraction gratings, ‘rigorous coupled wave analysis,’ (RCWA) concluded that the transmission spectrum through subwavelength slit arrays should show a minimum at wavelengths equal to an integer multiple of the spp wavelength, and that therefore the spp mode played a ‘negative role’ in the transmission. This prediction of transmission minima was confirmed in a subsequent numerical simulation study on real metal (silver) films by Mansuripur and co-workers [27]. Using a finite-difference-time-domain (FDTD) technique this group solved Maxwell’s equations numerically in the vicinity of a slit array. While keeping the incident wavelength and slit width fixed, they systematically varied the distance between the slits as well as the metal film thickness. The simulations revealed wave-guide resonances within the slit channels as a function of metal film thickness and abrupt minima at an array pitch equal to an integral number of spp wavelengths. Furthermore, the transmission profile was found to vary quite rapidly in the vicinity of the minimum, with the overall form of the profile depending on the film thickness. Nevertheless, independent of film thickness, the transmission minimum was always found to be at the slit separation equal to an integer number of spp wavelengths. These results indicated that two kinds of interference effects may be operative: (1) interference between the incident wave and surface waves that strongly affect the spp amplitude and (2) Fabry–Perot-like resonances within the slits themselves. The overall transmission efficiency depends on both of these factors. Figure 7 summarizes the principal results of this study.

In a recent analysis of a groove-and-slit structure Ung and Sheng [28] have drawn attention to the presence of accumulated charge at the slit corners and interpreted the role of spp as agents promoting the accumulation of this oscillating dipolar charge that in turn radiates into the slit. They concluded that interference between the incident wave and the spp on the entrance surface, as a function of slit–groove separation, causes the enhancement (constructive interference) or extinction (destructive interference) of optical transmission through the slit elements. The question then is to understand how this destructive interference takes place just at the condition that the slits be separated by $n\lambda_{spp}$.

A Green’s tensor analysis of a single groove in a silver film has also underscored the importance of oscillating charge at the corners of discontinuities in the metal surface. Lévêque et al [29] have shown that the surface waves generated by a plane wave normally incident on a grooved silver film can be modeled by replacing the groove with a line dipole oriented in the direction of surface wave propagation. Figure 8 shows...
Figure 7. Numerical simulation results from [27]. Top panel shows transmission efficiency $\eta$ as a function of slit separation (abscissa) and silver film thickness (ordinate). For a given slit separation transmission efficiency shows maxima as film thickness is increased. At fixed thickness, transmission efficiency shows maxima and minima as arrays periodicity is increased. Note that the position of maxima varies with film thickness but position of the minima is always at an integer number of $\lambda_{spp}$. Bottom panel plots transmission minima and maxima over all thicknesses as a function of slit separation.

Figure 8. Figures adapted from [29], reprinted with permission from Lévéque G, Martin O J F and Weiner J 2007 Transient behavior of surface plasmon polaritons scattered at a subwavelength groove Phys. Rev. B 76 155418. Copyright (2007) by the American Physical Society. Left panel: Surface waves generated from groove source subject to normal incident plane wave illumination. Center panel: model schematic in which line dipole oriented as shown is substituted for the physical groove. Right panel: results of Green's tensor calculations using ‘exact’ numerical results (points) and dipole model (solid line). The numerical and model results superpose to within the width of the trace.

4.4.2. Recent results with single slits. In an attempt to settle the controversy over the role of surface plasmon waves, Gay et al [30] undertook measurements designed to isolate this
question from other effects due to the periodicity of array structures. To this end a series of single-slit–single-groove structures in a silver film were fabricated by focused ion beam (FIB) milling. The distance between the slit and the groove was systematically increased in 55 steps of 104 nm starting from a minimum slit–groove distance of about 400 nm. The basic idea was to test the behavior of the amplitude profile, and the phase of the surface wave generated at the groove by interfering it with the wave directly transmitted through the slit. An interference fringe pattern in the transmitted intensity contains amplitude, wavelength, and phase information of surface waves generated at the groove and propagating along the surface to the slit. Figure 9 shows a schematic of a typical structure and how the interference pattern appears in the far field. The results from these ‘front side’ measurements (with the groove directly illuminated by the incident plane wave) are shown in figure 10. The interference fringe clearly shows a rapid decrease in amplitude within 2–3 \( \mu \text{m} \) slit–groove distance followed by a leveling off to a persistent, apparently stable fringe out to the limit of the measurements, about 6 \( \mu \text{m} \). The initial rapid fall-off in amplitude together with high contrast in the fringe is consistent with a coherent diffracted surface wave composed of many \( k \)-modes. According to the CDEW model, these modes should dephase over 2–3 optical cycles of propagation along the surface, attenuating to a near-negligible amplitude over that interval. The measurements were clearly at variance with this prediction, but the wavelength of the observed persistent wave did not correspond to the spp wavelength predicted from equation (35) either. Subsequent measurements and simulations in silver [31] and gold films [32] confirmed that in fact, if the slit–groove distance had been extended out to about 10 \( \mu \text{m} \), the expected spp wavelength is indeed recovered. Although the CDEW model produces a satisfactory fit over the first few interference fringes, there is nothing in the model that predicts a persistent wave or the emergence of the spp as a surviving surface wave after the initial CDEW dephasing distance. Other theoretical studies applying a numerical implementation of rigorous coupled wave analysis (RCWA) diffraction theory [33], asymptotic analytic diffraction [34], finite difference time domain (FDTD) numerical solutions to Maxwell’s equations [31], or Green’s tensor analysis [29], produced calculations in reasonable accord with these experimental results. Of these various theoretical approaches, the Green’s tensor analysis provides a very clear, simple physical interpretation: a subwavelength groove or slit acts essentially as a line dipole driving the metal surface at the frequency of the incident plane wave. The response of the surface consists of a composite of evanescent and propagating modes all of which decay within 2–3 optical cycles, except for the spp mode which is inherently stable against radiative decay. The spp continues to propagate until resistive dissipation attenuates the amplitude. In the near IR spectral interval the characteristic decay length for the spp is \( \sim 100 \mu \text{m} \). Figure 11 compares the results of the Green’s tensor analysis to the experimentally measured points.
4.4.3. Recent results with slit arrays. In order to confront the question of how transmission minima and maxima depend on array periodicity, Pacifici et al [35] carried out a series of measurements on slit arrays in which the number of slits in the array varied from one to nine, and the array pitch was systematically varied in steps of 5 nm around the sensitive zone near $\lambda_{spp}$. The transmission efficiency $\eta$, defined as $|H_N|^2/|H_1|^2$, the ratio of the absolute value of the square of the $H$-field for transmission of an array of elements $N$ at pitch $p$ compared with the same measurement for a single slit, was determined as a function of $p$. The results are presented in the left panel of figure 12, and clearly show a minimum in transmission at slit separations equal to $\lambda_{spp}$ and $2\lambda_{spp}$. These measurements confirm conclusions from the RCWA calculation of the Lalanne group [26] and the FDTD simulations of the Mansuripur group [36]. The measurements also show that at intermediate slit separations the transmission is enhanced by up to a factor of $\sim 7$. The positions of the maxima are close to half-integer multiples of $\lambda_{spp}$, but there is clearly some shift in the position of the maximum and the overall shape of the transmission profile as a function of the number of elements in the slit array. For the case of the two-slit array the measurements were extended to include a third maximum and minimum in transmission. The right panel of figure 12 shows these results on a linear scale of transmission. In this simple two-slit case the transmission profile is readily recognized as that of a lossy Fabry–Perot resonator. The interpretation is that spps launched from each slit partially reflect at the opposite slit and set up the resonator interference. The finesse of this resonator is only about 3.4, and probably most of the spp surface wave energy is lost by scattering back into space at the slit edges. The solid lines traced through the measured points show the results of a simple model of surface wave reflection at the top surface, transmission through the slits and again surface wave reflection along the bottom surface and finally the detected transmitted radiation. In addition to the question of the position of the transmission maxima and minima, it is important to know to what extent each element of the array couples to all the others. Two limiting possibilities are: (1) surface waves couple collectively such that the incident light excites the periodic Bloch states covering the entire array, (2) array elements couple locally to nearest neighbors but are insensitive to the presence of array elements beyond. The first case predicts that increasing the number of array elements should increase the transmission at each array element. The second case predicts, for 1D slit arrays, that the per-slit transmission should not increase beyond a three-element array. Figure 13 plots the
per-slit transmission for slit arrays ranging from one to nine elements. The measurements were taken at slit separations corresponding to the maxima and minima of transmission. The results show that per-slit transmission increases up to about three or four array elements but subsequently exhibits ‘saturation’ with little augmentation of transmission intensity as more slits are added. The conclusion is that for 1D slit arrays, the elements are strongly interacting and only nearest neighbors on either side of any given slit contribute to the transmission. These conclusions are consistent with those from a finite element (FEM) numerical simulation study carried out on a series of slit arrays in which the transmission efficiency versus slit separation, and the per-slit transmission versus number of slits were investigated [37]. This behavior may not be universal and may depend on the structure of the array element. There is now evidence that other array forms (sinusoidal grating oscillation or positive ridges [38] rather than slits) may show interaction among array elements over a much greater range than nearest-neighbor. Finally it is worth noting that the intensity maxima, occurring approximately at half-integer multiples of $\lambda_{\text{spp}}$, are enhanced by about a factor of six or seven over the transmission through a single slit. While a transmission boost of this size is not insignificant, the original claims of ‘orders of magnitude’ appear in retrospect to have been somewhat exaggerated, partly because the measurements were normalized to the Bethe formula, equation (6), and partly because the holes were actually larger than the fabrication target.

**Figure 13.** Per-slit transmission efficiency as a function of the number of array elements. [35] reprinted figure with permission from Pacifici D, Lezec H J, Atwater H A and Weiner J 2008 Quantitative determination of optical transmission through subwavelength slit arrays in Ag films: role of surface wave interference and local coupling between adjacent slits Phys. Rev. B 77 115411. Copyright (2008) by the American Physical Society. The ‘max’ and ‘min’ numbers refer to the transmission at the peaks and valleys shown in the right panel of figure 12 for the two-slit case.

**Figure 14.** Adapted from [37]. Left panel: initial phase of the $H$-field component of the spp wave relative to the phase of the incident $H$-field at the slit center as a function of slit width at a fixed film thickness of 200 nm. The wavelength of the incident field is 800 nm. The three traces correspond to different structures. PEC (light gray trace) is for a perfect electrical conducting metal. FEM sym (narrow black trace) is for a gold film surrounded by air on each side. FEM asym (wide black trace) corresponds to a gold film with air on one side and a titanium glass structure on the other side. The FEM asym structure resembles the actual structure used in the Schouten et al. experiments [39]. The two vertical dashed lines indicate the range of slit widths which are subwavelength with respect to the incident wavelength of 800 nm. Right panel: the relative phase of the spp with respect to the incident $H$-field for the same three structures as a function of film thickness for a fixed slit width of 200 nm.

4.4.4. Phase of the spp. Although the preceding work clearly established the importance of the surface plasmon polaritons in light transmission through subwavelength arrays, the details of their excitation and interaction with other incident, reflected and transmitted optical modes were still far from clear. A possible clue toward a deeper understanding was to investigate the phase relationship between the incident wave and the spp. Stimulated by the two-slit experiment of Schouten et al [39], Janssen et al [37] carried out numerical simulations on single and multiple slit arrays. They determined the phase of the spp mode established on the incident air–metal surface at a distance along the surface sufficiently removed from the slit that transient surface waves were no longer present. This spp with its associated phase was then extrapolated back to the center of the slit and there the $H$-field component of the spp was compared with the phase of the $H$-field component of the incident wave. The results, shown in figure 14, demonstrate that for the structures involving a real metal (gold) the initial phase of the spp with respect to the incident $H$-field, at the position of the center of the slit, is close to $\pi$. The conclusion from this simulation study is consistent with an experiment carried out by the Toulouse group [40] in which they took
the single-slit–single-groove structure studied previously [30], and turned it around so that the groove was on the exit side, facing away from the incident light. They determined the ‘intrinsic phase’ between the wave directly transmitted through the slit and the surface wave by measuring the phase shift of the interference fringe as a function of the slit–groove distance. A schematic diagram of the setup is shown in figure 15. The phase shift between the directly transmitted reference wave and the secondary source at the groove consists of three terms: (1) the ‘intrinsic’ phase between launching propagating wave at the exit side of the slit and the launched surface wave at the same point, (2) the accumulated phase due to the optical path length between the slit and the groove traveled by surface wave and (3) the phase shift introduced by the groove depth prior to conversion of the surface wave back into a propagating mode. The groove acts essentially as a broadcast dipole antenna. When the oscillating electron density wave associated with the spp encounters the dielectric groove gap, an oscillating electric field forms across the gap; and the groove radiates as a line dipole antenna. The efficiency of radiation is related to the effective length of the antenna, i.e. the groove depth. When the groove depth is about equal to λ_{spp}/4 the phase shift introduced into the line dipole source is about 2π, and this ‘antenna length’ is therefore optimal for radiative efficiency. The 2π phase shift at the groove means that only the optical path length of the surface wave, a known quantity, and the intrinsic phase shift of the surface wave contribute to the total phase shift measured by the interference fringe. The right panel of figure 15 plots the intrinsic phase and the interference fringe contrast as a function of groove depth. A maximum in the contrast coincides with the peak radiation efficiency at the groove. From the plot we can see that when the contrast passes through a maximum the interference fringe phase passes through an inflexion point, the typical signature of a resonance condition, and we identify the inflexion point phase as the ‘intrinsic phase’. The measured phase is about 0.55π close to π/2. But how can this result be consistent with the simulation results of Janssen et al [37] who predicted a π shift, not π/2? The answer is that the experiment of Gay et al [40] is sensitive to the \(E\)-field component of the spp wave parallel to the surface, and this component is in quadrature with the \(H\)-field taken by Janssen et al as the reference. Therefore the results of the Toulouse group also exhibit a π phase shift with respect to the \(H\)-field component of the spp. Further evidence for the π phase shift comes from the Green’s tensor study of Lévêque et al [29] in which they determined a π phase shift between the vertical \(E\)-field component of the spp and the incident \(H\)-field component. It is easy to show that the vertical \(E\)-field of the spp is in quadrature with the \(E\)-field component parallel to the surface.

### 4.5. Résumé of surface waves and subwavelength transmission

So what have we learned so far about surface waves and their role in subwavelength transmission?

- Surface plasmon polariton modes play a critical role in transmission through subwavelength arrays because the transmission is extinguished when the array pitch equals an integer multiple of the spp wavelength.
- Transient surface modes are also present in the immediate vicinity of the launch sites, the material structures themselves. These transient modes damp away within a few optical cycles, but survive on the surface for distances comparable to or greater than the array pitch in structures of practical interest.
- Transmission enhancement through subwavelength slit arrays occurs when the array pitch is close to a half-integer multiple of the spp wavelength, but the position of the maximum is not as sharply defined as the minimum.
- The per-slit enhancement effect in the array is no more than about a factor of seven over transmission through a single slit.
Slit array elements interact strongly with their nearest neighbors, but weakly with array elements further removed.

Transmission enhancement is associated with the induction of strong oscillating charge dipoles at entrance slit corners and a slit geometry (width and depth) giving rise to a resonant Fabry–Perot cavity within the slit.

The magnetic field component spp mode is out of phase with respect to the magnetic field component of the TM polarized incident $E$–$M$ field.

The characteristics of transmission extinction—a sharp profile at a well-defined array pitch—together with the phase opposition of the incident and spp modes suggests field cancellation or destructive interference. But how does such an interference take place? What exactly interferes with what? The phase of the surface wave together with Bloch mode analysis, in the spirit of Treacy’s original dynamic diffraction theory [11], provide a physical picture.

### 4.6. Bloch mode analysis and FDTD simulations

In an important series of papers [17, 27, 36, 41–43] Mansuripur and his co-workers have developed a new modal analysis, complemented with numerical FDTD field simulations. This analysis leads to a physical interpretation of what the numerical field distributions really mean in terms of transmitted and reflected modes. This work culminated in a key paper [36] that analyzes the case of an infinite 1D slit array in a real metal (silver), infinitely deep. These two infinities permit two key simplifications: (1) a periodic array of infinite extent without the complication of end effects always present in real experiments with finite arrays, (2) slits of infinite depth avoid the Fabry–Perot effects that arise when modes propagating down the slit reflect back at the slit exit. The analysis consists first of finding the Bloch modes along the surface and within the metal that satisfy the periodic boundary conditions of the slit array and second of matching these modes to the incident $E$–$M$ at the air–metal boundary. The setup of the problem is shown in the left panel of figure 16. Waves below the plane of incidence propagating in the metal along $y$ must satisfy the boundary conditions of a periodic wave guide. They must therefore exhibit the translational symmetry of $y \rightarrow y + p$ and satisfy continuity conditions at the metal–air slit walls. In the $+z$ direction the waves in the metal must be evanescent but in the slit gaps there may be propagating and evanescent modes. The first few $H$-field components of these Bloch modes are shown in the right panel of figure 16. The first (top) mode shows amplitude concentrated within the slit and is the only propagating mode in the $+z$ direction. The second mode from the top is evanescent in $z$ and is identified with the surface plasmon polariton. The remaining modes are all evanescent and attenuate well within the skin depth of the metal. Once these transmitted modes have been determined they must be matched to the incident field and the reflected modes at the air–metal boundary. The matching conditions for the $H$- and $E$-fields (always in TM polarization) are written as

$$H_{zI} + \sum_{n=1}^{N} C_n^R H_{zR}^n = \sum_{n=1}^{N} C_n^T H_{zT}^n, \quad (39)$$

$$E_{yI} + \sum_{n=1}^{N} C_n^R E_{yR}^n = \sum_{n=1}^{N} C_n^T E_{yT}^n, \quad (40)$$

where $I, R, T$ denote incident, reflected, and transmitted fields, respectively. The index $n$ is the mode number and $N$ the total number of modes in the expansion. The matching is achieved by varying the $C$ coefficients until the matching error between the incident field and the reflected plus transmitted fields is minimized. In order to achieve a numerical precision such that the total reflected plus transmitted power is equal to the incident power to within 0.3% up to $N = 80$ modes are required. The results are reported in table 1 of [36] and reproduced in figure 17. They show that when the array pitch is not close to $\lambda_{\text{app}}$ many of the evanescent Bloch modes propagating within the metal skin depth are required to achieve a good match at the boundary. However when the slit separation is very close to $\lambda_{\text{app}}$ only two modes in

![Figure 16. Adapted from [36]. Left panel: setup of the 1D slit array problem. The slit width is fixed, $w = 100$ nm, but the pitch, $p$ is variable. The $E$–$M$ field is TM polarized with $\lambda_0 = 1000$ nm incident from $–z$ on the $x$–$y$ plane. The permittivity of silver is taken to be $\varepsilon = –48.8 + i2.99$. Right panel: plots of the $H$-component for the first few Bloch waves in the metal below the plane of incidence. The first mode at the top propagates in the $+z$ direction down the slits. All the other modes are standing wave propagating modes along $\pm y$ but evanescent along $+z$. The second mode from the top is identified with the surface plasmon polariton.](Image 107x633 to 293x765)
reflection and two modes in transmission are significantly populated. The two reflection modes consist of the normal, propagating reflection mode and an evanescent mode which is the exponentially decreasing 'wing' of the spp that propagates on the dielectric side of the boundary. The two transmission modes propagating within the metal skin depth consist of (1) the metal 'wing' of the spp mode (second from the top in the left panel of figure 16) and (2) the 'second harmonic' of this spp mode (third from the top in the same figure). Of these two evanescent modes the 'fundamental' spp (mode #2 in transmission) is dominant, the 'second harmonic' spp (mode #3 in transmission) contributing only about one tenth as much to the total transmitted field. As the slit separation approaches \( \lambda_{\text{spp}} \) we can see from the table, reproduced from [36] in figure 17, how mode population concentrates in just two modes in reflection and two modes in transmission. The coefficient \( C_{1T} \) of the propagating mode in transmission rapidly diminishes and reaches a minimum just as the pitch passes through \( \lambda_{\text{spp}} \). The modal analysis tells us that at the critical pitch \( p = \lambda_{\text{spp}} \), the only important surface mode is a standing wave spp and that the propagating mode through the slit essentially vanishes. However the modal analysis alone provides no physical explanation for why these modes behave as they do.

4.7. Proposed physical explanation

Remember that we argued in section 4.4.4 that the sharp minimum in transmission must be due to some kind of interference. It has recently been proposed [44] that the specific interference is between the magnetic field of the spp standing wave that sets up on the surface when \( p = \lambda_{\text{spp}} \) and the magnetic field due to the incident plane wave. This interference is due to the phase relation between the two harmonically time-varying fields and their relative amplitudes. Figure 18 shows the net standing wave \( E \)-field and \( H \)-field that arise from summing over all the populated modes determined by the modal analysis of [36]. These fields have a shape of essentially a half sine function with the maximum at the midpoint between the slits. According to the proposal of [44], when this \( H \)-field is superposed on the \( H \)-field standing wave arising from the incident plane wave at the \( x-y \) surface, the two time-harmonic \( H \)-fields oscillate \( \pi \) out of phase, and therefore with their directions along \( x \) opposed. At two points along \( y \), equidistant from the midpoint between the slits, the two \( H \)-fields will cancel, producing a null magnetic field at these two points. The disposition of the fields is diagrammed in figure 19. Null points of magnetic field imply null points of current and maximum points of charge accumulation. Thus when the array pitch just equals \( \lambda_{\text{spp}} \), the charge accumulation shifts from the slit edges to the null points approximately at the 1/4 and 3/4 wave points along \( y \) as indicated in figure 19. This charge shift minimizes the oscillating dipoles at the slit edges and leads to the minimum in transmission. As soon as the array pitch moves away from this critical value, many modes in transmission and reflection again come into play as indicated in figure 17. In principle the modal analysis provides phase information as well as the amplitude of each mode. Unfortunately [36] reported only the modal amplitudes, so the phase relations
5. Where are we and where are we going?

5.1. What do we know?

Most of what we know comes from the study of 1D structures such as slits, grooves and slit arrays. There are two reasons: the first is that the experiments are easier; there is no cutoff to a propagating mode through the structure, no matter how narrow the slit; and therefore for the same opening area, subwavelength slits transmit more light than round holes. The second reason is that the electromagnetic disturbance only propagates along one coordinate axis, perpendicular to the slit. For 2D structures such as hole arrays, waves are generated at the edges of the structure, and we know that these dipoles are generated by currents induced within the metal skin depth by the incident standing wave at the surface. We know that enhanced transmission in a slit array is no more than about a factor of 7 over transmission of an individual element of that array, and we know that slits in an array only couple with nearest neighbors. We know that for a slit structure the slit itself acts as a Fabry–Perot cavity and that this cavity is coupled to the surface waves on the incident and exit surfaces, but the exact nature of this coupling and how to design for optimum transmission is still not clear. Finally we know that the spp mode plays a key role in transmission extinction when the array pitch is just equal to surface plasmon wavelength, and we think we understand physically why this must be true. However even on this simple point the issue cannot be regarded as settled beyond a reasonable doubt.

5.2. What do we not (yet) know?

We do not understand the transmission profile as a function of array pitch. One of the principal reasons for the early confusion surrounding interpretation of the profile is that it can change very rapidly from a minimum to a maximum, assuming the shape of a ‘Fano profile’ [45]. The rapid rise to a maximum at an array pitch equal to λ_0, the free space wavelength (usually just slightly longer than λ_{spp}), has also been attributed [36] to the ‘Wood’s anomaly,’ [46] a grating pitch where evanescent modes on the surface, generated by normal incident light, convert to modes propagating at grazing incidence to the array surface. The Wood’s anomaly condition might plausibly be invoked to explain the destruction of the interference condition at λ_{spp} between the spp and incident plane standing waves, but it is difficult to understand how the Wood’s anomaly condition would lead to transmission enhancement through the slits. In any case it appears that the minimum is always at p = λ_{spp} (at least for slit arrays), but the amplitude and position of the maximum seems to depend on geometry, material, metal film thickness, and perhaps other variables poorly understood. Furthermore we do not know how to interpret physically what a Fano profile means for subwavelength transmission. The original Fano profile had its origins in quantum mechanical matter wave interference effects observed in the autoionization of double excited states of atoms [47], but it is not clear how to apply this interference picture to the transmission profile in subwavelength 1D or 2D arrays. The question that reappears is: What exactly is interfering with what? We do not know how to design an array for optimum transmission, especially in the case of 2D arrays. There is no doubt some interplay between the modes within the structures and modes on the surface, but a modal analysis analogous to the one carried out by the Mansuripur group in 1D is much harder to execute in 2D. Finally we do not yet understand the effect on the transmission profile of finite arrays. Theoretical treatments usually rely on periodic boundary conditions that assume essentially arrays of infinite extent. Increasing the number of array elements has been shown to affect the profile significantly [35] but we have no general understanding of how these changes come about.

5.3. Prospects for the future

We can say that we truly understand the physics of subwavelength transmission when we can a priori design structures to predetermined performance criteria and have them work as designed. Or in other words, when we have reduced the problem to engineering, the physics will have been...
mastered. In fact the analytical tools developed for circuit analysis and design—inductance, capacitance, admittance, impedance—may prove as useful in the optical regime as they have at radio and microwave frequencies. After all, a radio receiver is a subwavelength tuned circuit. A promising line of research [48, 49] is to learn how to apply circuit analysis to subwavelength inductive and capacitive elements and to analyze slit and hole arrays in terms of wave guides and transmission line theory. Another promising avenue is to fill the slit or hole gap with an active, nonlinear material. The presence of strong oscillating dipoles at the slit edges must mean very high electric fields across the dielectric gap. Materials that exhibit a nonlinear index of refraction could be used to implement transmission modulators at petahertz frequencies. Second-harmonic or frequency-sum generation are also obvious candidates for investigation. From an applications viewpoint, the ability to control optical transmission at petahertz frequencies with planar devices at the subwavelength scale opens the door to integrated, miniaturized, high-density, optical-bandwidth information processing with potentially significant advances for optical storage and telecommunications networks. While the physics of subwavelength transmission is a fascinating challenge to work out, the tool box required for the job, classical electrodynamics, has been available for over a century. For basic research in this area to have a lasting impact, we must be able to translate it into functional devices that really make a difference.

References

[34] Ung B and Sheng Y L 2008 Optical surface waves over metalo-dielectric nanostructures; Sommerfeld integrals revisited Opt. Express 16 9073–86
Opt. Express 14 11823–32


[48] Engheta N 2007 Circuits with light at nanoscales: optical nanocircuits inspired by metamaterials Science 317 1698–702