Tunneling time, the Hartman effect, and superluminality: A proposed resolution of an old paradox

Herbert G. Winful

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

Accepted 1 September 2006

editor: G.I. Stegeman

Abstract

The issue of tunneling time is replete with controversy and paradoxes. The controversy stems from the fact that many tunneling time definitions seem to predict superluminal tunneling velocities. One prediction, termed the Hartman effect, states that the tunneling time becomes independent of barrier length for thick enough barriers, ultimately resulting in unbounded tunneling velocities. Experiments done with “single photons”, classical light waves, and microwaves all show this apparent superluminality. The origin of these paradoxical effects has been a mystery for decades. In this article, we review the history of tunneling times starting with the early work of MacColl, Hartman, and Wigner. We discuss some of the tunneling time definitions, with particular emphasis on the phase time (also known as the group delay or Wigner time) and the dwell time. The key experiments are reviewed. We then discuss our recent work, which suggests that the group delay in tunneling is not a transit time as has been assumed for decades. It is, in reality, a lifetime and hence should not be used to assign a speed of barrier traversal. We show how this new understanding along with the concept of energy storage and release resolves all the outstanding tunneling time paradoxes.

© 2006 Elsevier B.V. All rights reserved.

PACS: 03.65.Xp; 03.75.Lm; 42.70.Qs; 42.50.Xa

Contents

1. Introduction ......................................................................................................... 2

2. Tunneling time definitions ............................................................................................. 3
   2.1. A brief history of tunneling time ................................................................................... 3
   2.2. Stationary state tunneling ......................................................................................... 3
   2.3. Group delay or phase time ........................................................................................ 5
   2.4. Dwell time ..................................................................................................... 7
   2.5. Relation between group delay and dwell time ................................................................. 7
   2.6. Flux delays .................................................................................................... 9

3. The Hartman effect and superluminality ................................................................................. 11

4. Electromagnetic analogs ............................................................................................... 12
   4.1. Electromagnetic dwell time and group delay ................................................................. 13
   4.2. Waveguide below cutoff .......................................................................................... 14
   4.3. Photonic band gap structure ....................................................................................... 17
   4.4. Frustrated total internal reflection ............................................................................. 19

E-mail address: arrays@eecs.umich.ed.
1. Introduction

There have been numerous published reviews in recent years on the subject of tunneling time [1–10]. Indeed, this journal published one such review less than 2 years ago [10]. Given the profusion and currency of reviews on the subject, one might question the need for yet another one so soon. The answer is quite simple. Within the last 4 years, a radically different point of view of tunneling times has emerged, one that is capable of resolving all the thorny paradoxes that have dogged the subject for decades, and one that explains all the experimental results. It is our hope that by presenting this new understanding in this forum, the nature of the discussion of tunneling time will be fundamentally changed.

The question of how long it takes a particle to tunnel through a potential barrier is one that has occupied physicists since the early days of quantum mechanics [11,12]. The search for a general answer to that question has turned up a large number of tunneling time definitions, some of which suggest that the tunneling process is superluminal or faster than the speed of light [4]. The fact that many experimenters have reported measuring superluminal tunneling velocities has only served to add fuel to the controversies surrounding tunneling time [13–20]. For a thorough discussion of these tunneling times the classic 1989 review by Hauge and Stövneng [1] cannot be bettered. That review ends with this statement, which we have taken as our motivation: “At this stage one could choose to continue the search for a general answer to the question posed . . . Alternatively, one could turn to tunneling experiments now in progress with the aim of thoroughly understanding the temporal aspects of the individual experiments. At the present time, the latter strategy seems to us the more promising one.” In this review we have chosen to follow this latter strategy. We conduct a critical analysis of the key experiments that report superluminal tunneling velocities in order to determine what it is that they actually measure. What we find is surprising. Contrary to the widely accepted belief, our conclusion is that no one has ever measured a superluminal group velocity in barrier tunneling. Furthermore, we find that even the theoretical predictions of superluminal tunneling group velocity are based on an untested and unspoken assumption: that the group delay in barrier tunneling is a traversal time. We will show that this assumption is false and that the theoretical and measured tunneling times are lifetimes as opposed to transit times. This new interpretation of group delay makes it possible to resolve such hitherto intractable mysteries as the Hartman effect, the lack of dependence of group delay on barrier length for thick barriers [21–23].

After a brief history of the subject we turn to the definition of the group delay and the dwell time, the two tunneling times that are considered relatively well established [1]. We mention some of the other proposed tunneling times and then show how the group delay is related to the dwell time. The Hartman effect is introduced as a general feature of these
tunneling time definitions and one that has been commonly taken to imply superluminal tunneling velocities [2,4–10]. The tunneling of electromagnetic wave packets is then shown to be analogous to quantum mechanical tunneling. Several key tunneling experiments with electromagnetic waves and acoustic waves are described and it is pointed out that all these measurements are in the regime of quasi-statics: pulse length much greater than the barrier length, a necessary condition for distortionless tunneling. We then critically examine the notion of superluminal tunneling velocity and the commonly accepted “reshaping” argument for apparent superluminality. We conclude this review with our new interpretation of the group delay and demonstrate how it explains the Hartman effect and the anomalously short delay times seen in both single- and multiple-barrier tunneling.

2. Tunneling time definitions

2.1. A brief history of tunneling time

Shortly after the discovery of quantum mechanical tunneling, Condon in 1930 posed the question of the speed of the tunneling process [11]. MacColl [12] carried out an approximate wave-packet analysis of the time-dependent Schrödinger equation that suggested that tunneling takes no appreciable time. The problem lay dormant for 30 years until technological advances made possible thin film devices based on quantum tunneling. Hartman [21] revisited the question in an effort to understand the frequency limitations for tunneling devices involving metal–insulator–metal thin film sandwiches. Using the method of stationary phase previously applied to scattering problems by Eisenbud [24], Bohm [25], and Wigner [26], he obtained an analytical expression for the time delay (the group delay or phase time) in barrier tunneling that suggested a finite but short traversal time that saturates with distance. Also in the 1960s Baz’ proposed timing potential scattering events by use of a quantum-mechanical clock based on Larmor precession of a spin in the presence of a magnetic field [27]. Rybachenko [28] applied this idea to the specific case of tunneling through a barrier and obtained the so-called Larmor times \( \tau_{LT} \) and \( \tau_{LR} \) for transmitted and reflected particles, respectively. The more recent surge of activity in tunneling time physics probably dates back to the 1982 proposal by Büttiker and Landauer [29] to measure tunneling time by modulating the height of a barrier. They inferred a “traversal time” by noting the critical frequency at which the tunneling particle could no longer adiabatically follow the oscillating potential. For thick barriers this Büttiker–Landauer time coincides with a semiclassical result obtained by dividing the barrier length by the magnitude of the imaginary velocity under the barrier. At about the same time Büttiker applied Smith’s definition [30] of a scattering lifetime or dwell time to the tunneling problem and also refined Rybachenko’s Larmor clock idea [31]. Other subsequent approaches to the tunneling time problem have led to complex times [32,33] whose real and imaginary parts reduce under certain circumstances to some of the previously mentioned definitions.

In the early nineties, experiments based on electromagnetic analogs of quantum tunneling by Enders and Nimtz [13], Steinberg et al. [14], and Spielmann et al. [15] were providing support for the group delay or phase time even when they seemed to imply superluminal group velocities. In this review, we take the point of view that the group delay is a well-established quantity since the measured delay times agree with the theoretical predictions provided by the method of stationary phase. Of course the experiments are designed to measure exactly this quantity (since they follow the peak of a wave packet) and thus it is not quite fair to claim that the other tunneling times have not been proven. They have simply not been tested. No one has done an experiment with a modulated barrier or measured the Larmor times associated with a tunneling particle in the presence of a magnetic field. What is available is a wealth of data on group delays. What is needed now is a thorough understanding of the meaning of those measured delays and how they shed light on the dynamics of quantum tunneling. Because the dwell time will be shown to be identical to the group delay for electromagnetic tunneling through photonic band gaps, it is on those two tunneling time definitions that we will focus in this review.

2.2. Stationary state tunneling

We begin by collecting some relevant results from the stationary state description of quantum tunneling. The one-dimensional (1D) tunneling configuration considered here is shown in Fig. 1. In the stationary state description, a particle of definite energy \( E \) and momentum \( \hbar k \) is incident from the left upon a real potential barrier \( V(x) \) that occupies the region \( 0 < x < L \). In the regions \( x < 0 \) and \( x > L \) where \( V(x) = 0 \), the energy and momentum are related through \( E = \hbar^2 k^2 / 2m \), where \( m \) is the mass of the particle. Whereas a classical particle is totally reflected by this potential
barrier when $E < V$, quantum mechanically there is a finite, albeit small, probability that the particle will tunnel through the barrier and end up in the region $x > L$. The probability of this event is measured by the magnitude squared of the barrier transmission coefficient $T = |T| e^{i\phi}$. The particle is much more likely to be reflected, with a probability given by the magnitude squared of the reflection coefficient $R = |R| e^{i\phi}$. The particle is described by the wave function $\psi(x, t)$ which satisfies the time-dependent Schrödinger equation

$$\left[ \frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} - V(x) \right] \psi(x, t) = -i\hbar \frac{\partial}{\partial t} \psi(x, t). \quad (1)$$

For stationary states the wave function separates into a time-independent part and a complex exponential time factor:

$$\psi(x, t) = \psi_E(x) \exp(-iEt/\hbar)$$

with $\psi_E(x)$ a solution of the time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m} \frac{d^2\psi_E}{dx^2} + (V - E)\psi_E = 0. \quad (2)$$

Henceforth, we will drop the subscript $E$ that signifies stationary energy eigen-states and distinguish them from time-dependent wave functions by including the variable $t$ in the latter. To the left of the barrier the wave function consists of incident and reflected plane waves:

$$\psi_I = e^{ikx} + Re^{-ikx}, \quad (3a)$$

while to the right we have a pure traveling wave

$$\psi_{III} = Te^{ikx}. \quad (3b)$$

Inside the barrier the wave function is $\psi_{II}(x; k)$, a function that includes forward (to be transmitted) and backward (to be reflected) components. The transmission and reflection coefficients are found by requiring that the wave function and its derivative with respect to $x$ be continuous at 0 and $L$.

For a rectangular barrier, $V(x) = V_0$, a constant in the region between 0 and $L$. In this case the wave function in the barrier region is

$$\psi_{II} = Ce^{-\kappa x} + De^{\kappa x}, \quad (4)$$

where $\kappa = \sqrt{2m(V_0 - E)/\hbar}$, $C = (1 - ik/\kappa)e^{\kappa L}/2g$, $D = (1 + ik/\kappa)e^{-\kappa L}/2g$, $A = (\kappa/k - k/\kappa)/2$, and $g = \cosh \kappa L + iA \sinh \kappa L$. The transmission coefficient is

$$T = \frac{e^{-iL}}{\cosh \kappa L + iA \sinh \kappa L}. \quad (5a)$$
Fig. 2. Transmission of a rectangular potential barrier of height $V_0$ versus energy $E/V_0$. Here $\sqrt{2mV_0L/\hbar} \equiv \gamma L = 3\pi$.

Fig. 3. Probability density as a function of position for a stationary state.

and the reflection coefficient is

$$ R = \frac{-i(\kappa/k + k/\kappa) \sinh \kappa L}{2(\cosh \kappa L + i\lambda \sinh \kappa L)}. \quad (5b) $$

Fig. 2 shows the transmission probability $|T|^2$ versus energy for a moderately “opaque” barrier whose strength is measured by $\sqrt{2mV_0L/\hbar} \equiv \gamma L = 3\pi$. The transmission probability is about $6.5 \times 10^{-6}$ for $E = V_0/2$ and jumps to unity around $E = V_0$. Fig. 3 shows the probability density $P(x) = |\psi(x)|^2$, which corresponds to a partial standing wave in front of the barrier, a near-exponential decay inside the barrier, and a traveling wave at the exit.

2.3. Group delay or phase time

The stationary state tunneling solution exists everywhere for all time and does not directly reveal information about tunneling dynamics. However, one can construct a spatially localized wave packet by summing over a band of stationary states with different energies:

$$ \psi(x, t) = \int_E f(E - E_0)\psi_E(x)e^{-iE_t/\hbar} \, dE. \quad (6) $$

Here $f(E - E_0)$ is a sharply peaked energy distribution such as a Gaussian centered at the mean energy $E_0$. Such a wave packet, prepared somewhere to the left of the barrier, propagates to the right with a group velocity $v = \hbar^{-1}\partial E/\partial k = \hbar k/m$. Upon colliding with the barrier the incident wave packet disappears and is replaced by a reflected wave packet and
transmitted packet as shown in Fig. 4. The transmitted wave packet is described by

$$\psi(x, t) = \int_E f(E - E_0) |T(E)| \exp[i\phi_t(E) + ikx - iEt/\hbar] \, dE. \quad (7)$$

If the width of the energy distribution is sufficiently narrow, the magnitude of the transmission coefficient is approximately constant over the range of the integral and hence the wave packet does not suffer any distortion or reshaping. According to the method of stationary phase, the peak of the transmitted wave-packet is located where the phase of the transmitted wave packet is stationary. Upon setting the energy derivative of the phase equal to zero, we find that a wave-packet peak appears at position $L$ with a delay given by

$$\tau_{gt} = \hbar \frac{\partial}{\partial E} (\phi_t + kL), \quad (8)$$

if the peak of the incident wave packet is at $x = 0$ at $t = 0$. Similarly, the group delay for the reflected wave packet is

$$\tau_{gr} = \hbar \frac{\partial}{\partial E} \phi_r. \quad (9)$$

It can be shown that for a symmetric barrier $\tau_{gt} = \tau_{gr} \equiv \tau_g$ [34]. A symmetric barrier is thus characterized by a single group delay that describes the occurrence of the peaks of both transmitted and reflected wave packets. For an asymmetric barrier we can define a bidirectional group delay [23] as the weighted average of transmission and reflection group delays, the weights being the probabilities of transmission and reflection:

$$\tilde{\tau}_g = |T|^2 \tau_{gt} + |R|^2 \tau_{gr}. \quad (10)$$

The group delays as defined are actually asymptotic quantities that apply to completed tunneling events with distinct wave packets measured far from the barrier as illustrated in Fig. 4c. During its approach to the barrier, the incoming wave packet interferes with the part of itself that has already been reflected. The resulting distortion makes it difficult to locate an incident “peak”. However, one can extrapolate the time that a freely propagating incident wave packet traveling with velocity $v = \hbar k/m$ would have arrived at $x = 0$ in the absence of reflections. Similarly, the transmitted wave packet at some distance from the barrier can be extrapolated backwards to the exit $x = L$. The group delays thus defined in Eqs. (8) and (9) should therefore be understood as asymptotic, extrapolated quantities.

It should be noted that the group delay does not claim to be a traversal time. This is because the incident and transmitted wave packets are different entities. All we can say is that at $t = 0$, the extrapolated peak of the incident wave packet is at $x = 0$. At $t = \tau_{gt}$ the extrapolated peak of the transmitted wave packet is at $x = L$. There is no implication
that the incident peak has propagated to the exit. We cannot say where the transmitted wave packet is at \( t = 0 \) and hence cannot say that the group delay measures the time it takes a wave packet to travel from input to output.

Büttiker and Landauer [29] have argued strenuously that the group delay is not a physically meaningful quantity with which to characterize tunneling dynamics. This is because of possible wave-packet distortion that may occur during tunneling and also because of the fact that transmitted and incident packets are not related by a simple causal translation. A counter argument against these statements is that numerical solutions of the time-dependent Schrödinger equation [35,36] as well as experiments to be discussed in Section 6 confirm the status of the group delay as marking the occurrence of the peak of an undistorted transmitted wave packet. Thus, the group delay has a clear physical significance. Its interpretation as a traversal time, however, is not justified since the object that arrives at the exit is not the same object that enters at the input.

2.4. Dwell time

The dwell time was first introduced by Smith in a potential scattering context [30]. He defined it as the difference between the time spent by a particle in the region of the scattering potential and the time spent in the same region in the absence of the scattering potential. Operationally it is determined by dividing the excess number of particles in the scattering region \( \langle I \rangle \) by the incident particle flux \( F \):

\[
Q = \frac{\langle I \rangle}{F} = \lim_{R \to \infty} \left[ \int_{0}^{R} (|\psi(x)|^2 - 2/v) \, dx \right]_{av},
\]

where some averaging device is used to eliminate an oscillatory term that is important at low energy.

In the context of tunneling, the dwell time as we now know it was written down by Büttiker as [31]

\[
\tau_d = \frac{\int_{0}^{L} |\psi(x)|^2}{j_{in}},
\]

where \( j_{in} = \hbar k/m \) is the incident flux and \( \psi(x) \) is the stationary state wave function. For an incident free particle described by a unit amplitude plane wave \( \exp(ikx) \), the flux is also equal to the group velocity \( v = \hbar k/m \). The dwell time does not distinguish transmitted particles from reflected particles since it is a property of an entire wave function with forward and backward components [1]. It is the time spent in the barrier region, averaged over all incoming particles, regardless of whether a particle is ultimately transmitted or reflected [37]. To quote Buttiker and Landauer [29], “This time is the average dwell time of a particle in the barrier, and is not the traversal time, if most of the particles are reflected.” As such, it cannot be used to define a traversal velocity for either transmitted or reflected channels alone.

We show later that the group delay in tunneling has exactly the same status as the dwell time: a characteristic of an entire wave function with transmitted and reflected components.

One might question the meaning of a dwell time defined using stationary state wave functions since stationary states last for an infinite amount of time. However, it has been shown by Leavens and Aers [37] that the above expression for the dwell time is equivalent to an integral over the norm of a time-dependent wave packet (normalized to unity) over the barrier for all time:

\[
\tau_d = \int_{0}^{\infty} \int_{0}^{L} dt \, dx |\psi(x, t)|^2.
\]

This is an integral over all time of the probability of finding the particle inside the barrier region \([0, L]\) and hence tells us the dwell or sojourn time in the barrier regardless of whether the particle is transmitted or reflected at the end of its stay.

2.5. Relation between group delay and dwell time

Even though the group delay is defined in terms of the energy derivative of a phase shift and the dwell time by an integral over a probability density, the two quantities are not unrelated. In fact, they are practically equal for energies above the barrier height and differ by a self-interference delay for below-barrier energies. The relation between \( \tau_g \) and \( \tau_d \) can be obtained either through a lengthy but explicitly time-dependent wave-packet analysis as done by Hauge et al.
or, more simply, by using general properties of stationary state scattering solutions of the Schrödinger equation, as done by others [23,39]. The result is [23]

\[ |T|^2 \tau_{gt} + |R|^2 \tau_{gr} = \int_{0}^{L} \frac{|\psi(x)|^2 \, dx}{j_{in}} - \frac{\text{Im}(R)}{k} \hbar \frac{\partial k}{\partial E}. \]  

Eq. (13) is a simple and general result that unifies two of the major tunneling times. The quantity on the left-hand side is \( \tilde{\tau}_g \), the bidirectional group delay of Eq. (10). The first term on the right-hand side is the dwell time. The second term is a self-interference delay

\[ \tau_i = -\hbar \text{Im}(R) \, d \ln k/dE = -\text{Im}(R)/kv \]  

that comes from the overlap of incident and reflected waves in front of the barrier. As the wave-packet tunnels through the barrier, part of the incident packet interferes with a portion that has already been reflected [1,23,38]. This term is of great importance at low energy \((E \rightarrow 0)\) when the particle spends most of its time dwelling in front of the barrier, interfering with itself, held up in a standing wave. We can add this dwell time in front of the barrier (which can be positive or negative) to the strictly positive dwell time inside the barrier to obtain a generalized dwell time \( \tilde{\tau}_d \). Eq. (13) thus states that the bidirectional group delay is equal to the generalized dwell time:

\[ \tilde{\tau}_g = \tau_d + \tau_i = \tilde{\tau}_d. \]  

If the barrier is symmetric, then the bidirectional group delay is the same as the group delay in reflection or in transmission: \( \tilde{\tau}_g = \tau_{gt} = \tau_{gr} \equiv \tau_g \) and Eq. (15) becomes

\[ \tau_g = \tau_d + \tau_i. \]  

It is clear from Eq. (14) that the self-interference delay depends on the dispersion in front of the barrier. A similar term also shows up in the tunneling of electromagnetic waves, which we take up in Section 4. That term, however, disappears in the photonic band gap structure where the approach to the barrier is characterized by non-dispersive plane wave propagation. In that case the dwell time is exactly equal to the group delay. The self-interference delay also vanishes at barrier resonances where the reflection coefficient goes to zero. We remark that the general relation between group delay and dwell time has also been proven for relativistic particles described by the Dirac equation [40].

It is often stated without proof [1] that a necessary (but not sufficient) requirement for meaningful tunneling traversal times in transmission \((\tau_T)\) and in reflection \((\tau_R)\) is that they satisfy

\[ \tau_d = |T|^2 \tau_T + |R|^2 \tau_R. \]  

The justification for this criterion is largely based on classical thinking regarding mutually exclusive events. In quantum mechanics, a wave packet can be both transmitted and reflected. As pointed out by Landauer and Martin [3], quantum mechanically one does not sum probabilities; it is the complex amplitudes that are summed. Eq. (16a) neglects the possibility of interference between the amplitudes for reflection and transmission. For group delays, the exact quantum mechanical relation to the dwell time is that given below

\[ \tau_d = |T|^2 \tau_{gt} + |R|^2 \tau_{gr} - \tau_i. \]  

Since this follows directly from the Schrödinger equation without approximation, it is a more fundamental relation than Eq. (16a). It is of course always possible to define tunneling times that satisfy Eq. (16a). For example, since \(|R|^2 + |T|^2 = 1\) for a lossless barrier, we can multiply \( \tau_i \) in (16b) by \(|R|^2 + |T|^2\) and obtain

\[ \tau_d = |T|^2 (\tau_{gt} - \tau_i) + |R|^2 (\tau_{gr} - \tau_i). \]  

Eq. (16a) naturally follows if we define \( \tau_T = \tau_{gt} - \tau_i \) and \( \tau_R = \tau_{gr} - \tau_i \). These times happen to coincide with the Larmor times \( \tau_T^L \) and \( \tau_R^L \) which are just the group delays minus self-interference delays [1]. For a symmetric barrier these times equal the dwell time so that Eq. (17) conveys no information.

The relation between group delay, dwell time, and self-interference delay is easily checked for a rectangular barrier. From Eq. (5a) the total phase of the transmitted wave is found to be

\[ \phi_0 = \phi_i + kL = -\tan^{-1}(A \tanh \kappa L). \]
The energy derivative of $\phi_0$ yields the group delay

$$\tau_g = \frac{mL \cos^2 \phi_0}{\hbar k} \left[ \left( \frac{k}{\kappa} + \frac{\kappa}{k} \right)^2 \tanh \frac{\kappa L}{k} - \left( \frac{k^2}{\kappa^2} - 1 \right) \text{sech}^2 \kappa L \right],$$

where $\cos^2 \phi_0 = 1/(1 + A^2 \tanh^2 \kappa L)$. The dwell time is found by integrating the probability density and is given by

$$\tau_d = \frac{mL \cos^2 \phi_0}{\hbar k} \left[ \left( \frac{1 + \frac{k^2}{\kappa^2}}{\kappa^2} \right) \tanh \frac{\kappa L}{k} - \left( \frac{k^2}{\kappa^2} - 1 \right) \text{sech}^2 \kappa L \right].$$

The self-interference delay as calculated from Eq. (14) is

$$\tau_i = \frac{mL \cos^2 \phi_0}{\hbar k} \left[ \left( 1 + \frac{k^2}{\kappa^2} \right) \tanh \frac{\kappa L}{k} \right].$$

Clearly, the sum of Eqs. (20) and (21) yields the group delay of Eq. (19). Fig. 5 shows the three times plotted as a function of the normalized energy. The group delay and dwell time are obviously identical in the “classical” region $E > V_0$.

Numerical solutions of the time-dependent Schrödinger equation by Collins et al. [35] have shown that for wave packets that are narrow in momentum space it is the group delay that predicts the occurrence of a peak in the probability density at the exit of the barrier. Fig. 6 shows the delay obtained from the numerical solutions compared with the group delay and dwell time calculated from the stationary state wave functions.

2.6. Flux delays

The dwell time was defined as the stored probability (number of particles) within the barrier divided by the incident flux of particles. It measures the time it takes to empty the barrier of the accumulated stored particles, or equivalently, the time the incident flux has to be turned on and act in order to produce the expected accumulation of stored particles. It describes an escape process through both the reflection and transmission channels. It is also possible to define certain flux delays that measure the time taken to empty the barrier through either channel alone. The net particle flux
Fig. 6. Numerical results for the arrival time of the peak of a Gaussian wave packet (*) compared to analytical group delay (solid curve) and dwell time (dot-dashed curve). (From Ref. [35])

(incident minus reflected) is given by

\[ j_1 = \text{Re}[\psi^*(\hbar/im)\partial\psi/\partial x], \] (22)

where \( \psi \) is the stationary state solution in any of the three regions. This yields a constant (independent of position) flux, which is the net transmitted flux. Various authors have calculated a tunneling “transit” time by dividing the stored probability by the transmitted flux, [41–43]:

\[ \tau_t = \frac{\int_0^L |\psi(x)|^2}{j_1}. \] (23)

We prefer to call it the transmitted flux delay since it is not necessarily a transit time. It follows from a fluid-mechanics inspired definition of a local velocity field \( v(x) \) which is related to the local particle density \( \rho(x) = \psi(x)^*\psi(x) \) through \( j_1 = \rho v \). A delay time is then given by \( \int dx/v \) which is the flux delay of Eq. (23). This delay time however, is a property of an entire wave function made up of forward going and backward going components [44,45] and hence cannot be considered a traversal time for transmitted particles only. Similarly, one can define a reflected flux delay by dividing the stored probability by the reflected flux [46]:

\[ \tau_r = \frac{\int_0^L |\psi(x)|^2}{|j_r|}, \] (24)

where \( j_r = -|R|^2\hbar k/m \). This delay time is likewise not a transit time. Since

\[ j_{in} = j_1 + |j_r|, \] (25)

a division by the probability \( \int_0^L |\psi(x)|^2 \, dx \) yields the relation [46]

\[ \frac{1}{\tau_d} = \frac{1}{\tau_t} + \frac{1}{\tau_r}. \] (26)

Since the dwell time is a sojourn time in the barrier for both reflected and transmitted particles, its inverse has the clear meaning of an escape rate through both channels. Eq. (26) suggests one way (certainly not the only one) in which this escape rate may be distributed between the two channels: \( 1/\tau_t \) being the escape rate through the transmission channel and \( 1/\tau_r \) the escape rate through the reflection channel.
A relation between these flux delays and the group delays can be obtained by dividing Eq. (17) by the dwell time $\tau_d$. It reads

$$\frac{1}{\tau_t} = \frac{(\tau_{gt} - \tau_i)}{\tau_t} + \frac{(\tau_{gr} - \tau_i)}{\tau_r},$$

(27)

since $\tau_d/|T|^2 = \tau_t$ and $\tau_d/|R|^2 = \tau_r$. For a symmetric barrier $\tau_g - \tau_i = \tau_d$ and hence Eq. (27) reduces to Eq. (26).

3. The Hartman effect and superluminality

Hartman analyzed the temporal aspects of tunneling by writing down solutions of the time-dependent Schrödinger equation in terms of a superposition integral over stationary state solutions weighted by a Gaussian momentum distribution function as in Eq. (6). The integral over energy extended from zero to infinity and hence had contributions from non-tunneling components with $E > V_0$ as well as the tunneling components with $E < V_0$. The center of the narrow energy distribution of the incident wave packet is at $E_0 < V_0$. Without explicitly evaluating the integrals he could infer certain properties of the transmitted wave packet by examining the magnitude and phase of the integrand. What he found was that for thin barriers the transmitted packet has essentially the same form as the incident packet and that its delay (obtained by the method of stationary phase) is greater than the “equal time”, the time to traverse a distance of free space equal to the barrier thickness. For thicker barriers the peak of the transmitted packet is shifted to somewhat higher energies as a result of the filtering action of the barrier. Most importantly, in what has become known as the Hartman effect, he found that the delay time becomes independent of barrier thickness and is shorter than the equal time. For very thick barriers the below-barrier components are suppressed so that the propagating above-barrier components begin to dominate, at which point the delay time begins to increase again (“classical region”). The transition between the plateau region and the classical region has been characterized quantitatively by Brouard et al.\[47,48\].

For a sufficiently narrow wave packet, if $E_0$ is not too close to $V_0$, the integral over energies can be truncated at the barrier height. There will be no above-barrier contributions and the group delay saturates even as the barrier length goes to infinity. Furthermore, for sufficiently narrow energy distributions, the filtering effect is negligible and the mean energy of the transmitted and incident wave packets is the same. Fig. 7 shows the saturation behavior of the group delay as predicted by Eq. (19). Here we have taken $E = V_0/2$ so that

$$\tau_g = \frac{2 \tanh \kappa L}{\kappa v}.$$ 

(28)

In the limit $\kappa L \to \infty$, we obtain the limiting group delay (for all $E$)

$$\tau_{g\infty} = \frac{2}{\kappa v},$$

(29)

independent of length, which is the hallmark of the Hartman effect. Note that for a symmetric barrier the reflection group delay also saturates to the same value.

The name “Hartman effect” appears to have been first used in the review by Olkhovsky et al.\[2\]. If the group delay is taken as the transit time across the distance $L$, the implication is that the particle travels with a group velocity

$$v_g = \frac{L}{\tau_g}.$$ 

If $\tau_g$ becomes a constant while the barrier length increases, this “group velocity” then increases with length. Since there is no restriction on barrier length, ultimately the group velocity will exceed the speed of light in vacuum. Of course the transmission of the barrier is also decreasing exponentially with length, which means that ultimately as the transmission approaches zero, the “group velocity” approaches infinity! It is the presence of these rather large velocities implied by the Hartman effect that has led to the ongoing controversies regarding superluminality and causality in barrier tunneling. Because of this apparent superluminality of the group delay, there are some who dismiss it as a relevant time scale for the tunneling process. This is part of the motivation for the ongoing search for other tunneling times.
The dwell time also exhibits the Hartman effect. From Eq. (20), we see that in the limit $\kappa L \to \infty$,

$$\tau_{\infty} = \frac{2}{\kappa v} \left( \frac{E}{V_0} \right).$$ (30)

It is also present in the relativistic expressions for group delay and dwell time as calculated for the Dirac equation [40] and hence it is not an artifact arising from the use of the non-relativistic Schrödinger equation. It appears in electromagnetic and acoustic analogs as well. A generalization to multiple barriers [10] (“generalized Hartman effect”), where the tunneling time is believed to be independent of barrier separation will be discussed in Section 12.2.

The effect of dissipation on the Hartman effect has been considered theoretically [49] and experimentally [50]. It is found that for increasing dissipation the Hartman effect disappears and the group delay increases with barrier length as would be expected for classical propagation. Since tunneling is a quantum interference effect, anything such as dissipation that decreases the possibility of interference will destroy it. Physically, the presence of dissipation introduces an imaginary component to the purely real evanescent decay constant $\kappa$ so that the field within the barrier will have the form $\exp(-\kappa z + i\gamma z)$. This results in a $z$-dependent phase shift and hence a length-dependent delay.

The Hartman effect is at the heart of the tunneling time conundrum. Its origin has been a mystery for decades [51–53]. Its resolution would be of fundamental importance as it would lead to conclusive answers regarding superluminality and the nature of barrier tunneling. In Section 11, we show how a re-interpretation of the group delay in tunneling makes it possible to resolve this mystery.

4. Electromagnetic analogs

Tunneling is a wave property that manifests itself for all kinds of waves, be they matter waves, electromagnetic waves, or sound waves. In fact, the time-independent Schrödinger equation for quantum particles and the Helmholtz equation for electromagnetic waves are identical in form. The Helmholtz equation for wave propagation in a bulk inhomogeneous medium is

$$\nabla^2 \tilde{E} + [n(x, y, z)\omega/c]^2 \tilde{E} = 0,$$ (31)
where $\tilde{E}$ is a scalar component of the electric field, $n(x, y, z)$ is the refractive index at the angular frequency $\omega$, and $c$ is the speed of light in vacuum. A formal analogy can be made with the time-independent Schrödinger equation

$$\nabla^2 \psi + \left[ \frac{2m}{\hbar^2}(E - V(x, y, z)) \right] \psi = 0$$

if we set $n(x, y, z)\omega/c = [2m(E - V(x, y, z))/\hbar^2]^{1/2}$. Depending on the material or structure that supports the waves it is possible to have evanescent waves, which decay very much like the quantum mechanical wave functions in forbidden regions. These occur in regions where the effective refractive index is imaginary and correspond to the absence of real wave propagation beyond some cutoff temporal or spatial frequency. The commonly used electromagnetic barriers include waveguides with a narrowed section, dielectric-filled waveguides with an air gap, periodic dielectric structures (also known as photonic band gap structures), and closely spaced dielectric prisms coupled through frustrated total internal reflection (FTIR). Before discussing these electromagnetic barriers in some detail, we first consider some general results regarding electromagnetic tunneling times.

### 4.1. Electromagnetic dwell time and group delay

With a few minor modifications, the general results for tunneling time obtained for quantum particles also apply to the tunneling of electromagnetic waves. For electromagnetic structures the dwell time is defined as the time average stored energy within the barrier divided by the input power:

$$\tau_d = \frac{\langle U \rangle}{P_{in}}.$$  \hspace{1cm} (32)

Here, both the stored energy and input power are averaged over a cycle of the high-frequency electromagnetic wave. The stored energy includes both electric and magnetic field contributions: $\langle U \rangle = \langle U_e \rangle + \langle U_m \rangle$, which quantities are given by

$$\langle U_e \rangle = \frac{1}{4} \int_V E \cdot E^* \frac{\partial V}{\partial \omega} \, dv, \quad \langle U_m \rangle = \frac{1}{4} \int_V H \cdot H^* \frac{\partial \mu}{\partial \omega} \, dv.$$  \hspace{1cm} (33)

These relations hold for general dispersive media where both the dielectric permittivity $\varepsilon(\omega)$ and magnetic permeability $\mu(\omega)$ are functions of the angular frequency $\omega$. In most cases of interest, the tunneling pulses are sufficiently narrowband that material dispersion can be neglected compared to the structural dispersion introduced by, for example, the waveguide or photonic band gap structure. The stored energies thus reduce to the commonly used forms

$$\langle U_e \rangle = \frac{1}{4} \int_V \varepsilon E \cdot E^* \, dv, \quad \langle U_m \rangle = \frac{1}{4} \int_V \mu H \cdot H^* \, dv.$$  \hspace{1cm} (33)

It should be noted that the electric and magnetic stored energies are not necessarily equal for an arbitrary electromagnetic structure. It should also be realized that the dwell time in electromagnetic tunneling, like the quantum version, does not differentiate between transmission and reflection channels. The stored energies in the dwell time expression include both forward and backward contributions integrated over the barrier region. The dwell time is thus a property of an entire wave function with transmitted and reflected components and hence cannot be used to assign a traversal time for forward transit.

The group delay is defined as the derivative of the phase of the transmitted wave with respect to angular frequency:

$$\tau_g = \frac{d\phi_0}{d\omega},$$  \hspace{1cm} (34)

where $\phi_0 = \beta L + \phi_t$ is the total phase of the transmitted wave, $\beta$ is the propagation constant, and $\phi_t$ is the phase of the transmission function. When dealing with the slowly varying envelope of a high-frequency electromagnetic wave, the relevant quantity in the group delay is just the envelope phase $\phi_0 = \phi_t$. It will be seen that within the slowly varying envelope approximation the group delay and dwell time are identical. In what follows, we shall see that the electromagnetic group delay and dwell time also exhibit the Hartman effect in that they saturate with increasing barrier length.
4.2. Waveguide below cutoff

The waveguide below cutoff was first proposed by Hupert as an electromagnetic model for quantum mechanical tunneling [54,55]. It is the closest physical analog of quantum tunneling since it shares the same dispersion relation. In the geometry of Fig. 8a, a central waveguide filled with air or some dielectric with refractive index \( n_2 \) is connected to two other uniform waveguides filled with a different dielectric material of refractive index \( n_1 \). With \( n_2 < n_1 \), the operating frequency is chosen such that the central waveguide operates below cutoff while the other two guides are above cutoff. The central waveguide thus acts as a potential barrier in the manner shown in Fig. 1. Huppert actually carried out cw experiments on such a structure and mapped out the field distribution within the guide with the use of a moving probe [54]. The result is similar to the sketch shown in Fig. 3: a standing wave in guide I, an evanescent wave in guide II, and a traveling wave in guide III. This geometry permits analytical solutions for the field distributions as well as for the group delay and dwell time [56,57].

An important result obtained by Winful is that the group delay for this electromagnetic barrier is equal to the dwell time plus a self-interference delay, in exact analogy to the quantum mechanical tunneling barrier [57]. For simplicity we assume the waveguides support only the dominant TE\(_{10}\) mode. The electric and magnetic fields for this mode can be written

\[
\begin{align*}
E(x, z, \omega) &= \hat{y} \sin(\eta x) \psi(z), \\
H_t(x, z, \omega) &= \hat{x} \sin(\eta x) \frac{i}{\omega \mu_0} \frac{d\psi}{dz}, \\
H_z(x, z, \omega) &= -\hat{z} \cos(\eta x) \frac{i\eta}{\omega \mu_0} \frac{d\psi}{dz},
\end{align*}
\]

where \( \eta = \pi/a \) is the eigenvalue of the transverse mode, \( a \) is the width of the guide, \( z \) is the propagation direction, and a harmonic time dependence of \( \exp(-i\omega t) \) has been assumed. The wave function \( \psi \) satisfies the Helmholtz equation

\[
\frac{d^2 \psi}{dz^2} + \beta^2 \psi = 0,
\]

where the propagation constant \( \beta \) is given by

\[
\beta^2 = n_1^2 k_0^2 - \eta^2
\]
with \( k_0 = \omega/c \). There is a cutoff angular frequency \( \omega_{c1} = \eta c/n_1 \) below which \( \beta \) is imaginary and the waveguide does not support propagating modes. We will assume operation above this cutoff frequency for the waveguides I and III connected to the junction II which can be an arbitrary lossless scatterer with a reflection coefficient \( R = |R| e^{i\phi} \) and transmission coefficient \( T = |T| e^{i\delta} \). The relation between group delay and dwell time for this electromagnetic scattering problem can be derived in the same manner as done for the quantum problem, except that one starts with the variational theorem [58]

\[
\oint_S \left[ \frac{\partial E}{\partial \omega} \times H^* + E^* \times \frac{\partial H}{\partial \omega} \right] ds = 4i\langle U \rangle,
\]

(37b)

where \( \langle U \rangle \) is the total time average stored energy in region II and the surface integral is carried out over the metal walls of the waveguide and the planes \( z = 0 \) and \( L \) that bound that region. The result of the calculation is [57]

\[
\tau_g \equiv \frac{d\phi_0}{d\omega} = \frac{\langle U \rangle}{P_{in}} + \frac{\text{Im}(R)}{\beta} \left( \frac{\beta}{\omega} - \frac{d\beta}{d\omega} \right),
\]

(38)

where \( P_{in} = \epsilon_0 |E_0|^2 A c^2 \beta/4\omega \) is the time-averaged incident power, \( E_0 \) is the amplitude of the incident mode in region I, \( A \) is the waveguide cross-sectional area, and \( \text{Im}(R) \) is the imaginary part of the reflection coefficient. Because of the symmetry of the barrier, the reflection group delay \( \tau_r = d\phi_0/d\omega \) is equal to the transmission group delay \( \tau_g \).

The first term on the right-hand side in Eq. (38) is the dwell time \( \tau_d \). The second term is a self-interference term arising from the overlap between incident and reflected waves in the region before the obstacle:

\[
\tau_i = \frac{\text{Im}(R)}{\beta} \left( \frac{\beta}{\omega} - \frac{d\beta}{d\omega} \right).
\]

(39)

It depends not only on the reflectivity of the obstacle but also on the dispersion in the connecting waveguides. It vanishes if the waveguides are dispersionless \( (d\beta/d\omega = \beta/\omega) \) since in that case the interference pattern (envelope) travels with the same velocity as the phase fronts and there is no extra delay. This is the case for the symmetric photonic barrier immersed in air, where we find in the next section that \( \tau_i = 0 \). It also vanishes when the reflection coefficient is zero (at transmission resonances) or is purely real. With use of the complex Poynting theorem for lossless media, the self-interference delay can also be written [57]

\[
\tau_i = \frac{\langle U_m \rangle - \langle U_e \rangle}{P_{in}} \left( \frac{v_{g1}^0}{v_{p1}^0} - 1 \right).
\]

(40)

Here \( v_{p1}^0 = \omega/\beta \) is the phase velocity in the region outside the junction while \( v_{g1}^0 = d\omega/d\beta \) is likewise the group velocity outside the junction, assuming an infinite waveguide. The self-interference delay is thus seen to be proportional to the reactive stored energy, the difference between magnetic and electric stored energies. In this form one recognizes an intimate connection between group delay and stored energy:

\[
\tau_g = \frac{\langle U \rangle}{P_{in}} + \frac{\langle U_m \rangle - \langle U_e \rangle}{P_{in}} \left( \frac{v_{p1}^0}{v_{g1}^0} - 1 \right).
\]

(41)

In their analysis of finite photonic band gap structures, D’Aguanno et al. also find a similar term proportional to the difference between stored magnetic and electric energies [59,60].

For completeness we collect here the detailed expressions for the stored energies and delay times for the case where the central waveguide is below cutoff so that the propagation constant in that region becomes an attenuation constant \( \kappa \) given by

\[
\kappa^2 = \eta^2 - n_2^2 k_0^2.
\]

(42)

The cutoff angular frequency of the barrier is \( \omega_{c2} = \eta c/n_2 \). The stop band is the frequency region \( \omega_{c1} < \omega < \omega_{c2} \) where \( \omega_{c2} = (n_1/n_2)\omega_{c1} \). The field inside the barrier is a solution of the Helmholtz equation (36) with \( \beta^2 \) replaced by \( -\kappa^2 \).
\[ \psi_{II} = Ce^{-\kappa z} + De^{\kappa z}, \]

where \( C = (1 - i\beta/\kappa)e^{\kappa L}/2g \), \( D = (1 + i\beta/\kappa)e^{-\kappa L}/2g \), \( g = \cosh \kappa L + i\Delta \sinh \kappa L \), and \( \Delta = (\kappa/\beta - \beta/\kappa)/2 \). The transmission and reflection coefficients are

\[ T = e^{-i\beta L}/g, \]
\[ R = -i((\kappa/\beta + \beta/\kappa) \sinh \kappa L)/2g. \]

The phase of the transmitted wave is

\[ \phi_0 = \arg(T) + \beta L = -\tan^{-1}(\Delta \tanh \kappa L). \]

From this we find the group delay

\[ \tau_g = \frac{d\phi_0}{d\omega} = \frac{L}{v_g^0} \left\{ \frac{\omega_1^2}{\omega^2} \left( \frac{\beta}{\kappa} + \frac{\kappa}{\beta} \right)^2 \tanh \kappa L - \frac{n_2^2}{n_1^2} \left( \frac{\beta^2}{\kappa^2} - 1 \right) \text{sech}^2 \kappa L \right\}, \]

where

\[ \cos^2 \phi_0 = \frac{1}{1 + \Delta^2 \tanh^2 \kappa L} \]

and

\[ v_g^0 = (c/n_1)\sqrt{1 - (\omega_1/\omega)^2}. \]

The time average stored energy in the barrier is

\[ \langle U \rangle = \left( \frac{\varepsilon_0 |E_0|^2 A}{4} \right) \frac{\cos^2 \phi_0}{2} \left\{ \frac{n_1^2}{k_0^2} \left( \frac{\kappa^2 + \beta^2}{\kappa^2} \right) \tanh \kappa L - \frac{\beta^2}{\kappa^2} \left( \frac{\beta^2}{\kappa^2} - 1 \right) \text{sech}^2 \kappa L \right\}, \]

while the net reactive stored energy is

\[ \langle U_m \rangle - \langle U_\omega \rangle = \left( \frac{\varepsilon_0 |E_0|^2 A}{4} \right) \frac{\cos^2 \phi_0}{2} \left( \frac{\kappa^2 + \beta^2}{k_0^2} \right) \tanh \kappa L. \]

When divided by the input power, these stored energies yield the dwell time (Eq. (32))

\[ \tau_d = \frac{L}{v_g^0} \left\{ \frac{\cos^2 \phi_0}{2} \left( \frac{\omega_1^2}{\omega^2} \left( 1 + \frac{\beta^2}{\kappa^2} \right) \frac{\tanh \kappa L}{\kappa L} - \frac{n_2^2}{n_1^2} \left( \frac{\beta^2}{\kappa^2} - 1 \right) \text{sech}^2 \kappa L \right\} \]

and the self-interference time (Eq. (40))

\[ \tau_i = \frac{L}{v_g^0} \left\{ \frac{\cos^2 \phi_0}{2} \left( \frac{\omega_1^2}{\omega^2} \left( 1 + \frac{\kappa^2}{\beta^2} \right) \frac{\tanh \kappa L}{\kappa L} \right) \right\}. \]

Their sum yields the group delay, which is seen to be identical to Eq. (47).

Fig. 9 shows the group delay (solid line), the dwell time, and the interference delay. It is seen that outside the stop band the dwell time is identical to the group delay. The two differ in the stop band where the reflectivity is high. The interference delay diverges as the incident wave approaches cutoff whereas the dwell time goes to zero. This is because the incident wave spends all of its time being reflected by the barrier and nothing penetrates. The times are normalized by \( \tau_0 = L/c \), the transit time of a light front across a distance \( L \) in vacuum. Normalized delays less than \( L/c \) have been called superluminal. However, these are not propagation delays and should not be associated with velocities. Fig. 9
Fig. 9. Group delay (solid line), dwell time (dashed line) and self-interference delay (dotted line) for an electromagnetic waveguide below cutoff. (From Ref. [57])

Fig. 10. Photonic band gap structure (PBG) consisting of a dielectric with a periodic spatial modulation of the refractive index.

shows that it is possible to operate in the tunneling regime for a range of frequencies such that the group delay equals the dwell time. This regime is where the slowly varying envelope approximation holds.

From Eq. (47) it is evident that the group delay saturates with increasing barrier length. This is the electromagnetic analog of the quantum Hartman effect [21,2] which has been taken to imply superluminal, and indeed, infinite velocities of propagation for tunneling wavepackets [2–10]. From Eqs. (52) and (53), we obtain in the limit \( L \to \infty \),

\[
\tau_d = \frac{2}{\kappa v_0 g_1} \left( \frac{\omega c_1}{\omega} \right)^2 \frac{k^2}{k^2 + \beta^2}, \quad \tau_i = \frac{2}{\kappa v_0 g_1} \left( \frac{\omega c_1}{\omega} \right)^2 \frac{\beta^2}{k^2 + \beta^2}, \quad \tau_g = \tau_d + \tau_i = \frac{2}{\kappa v_0 g_1} \left( \frac{\omega c_1}{\omega} \right)^2,
\]

(54)

all of which are independent of length. Thus all three times, group delay, dwell time, and self-interference delay exhibit the Hartman effect.

4.3. Photonic band gap structure

The photonic band gap structure is the electromagnetic analog of the Kronig–Penney model in quantum mechanics. It is a dielectric structure whose refractive index varies periodically with distance along the propagation direction with a period comparable to the light wavelength (Fig. 10). Examples are the multilayer dielectric mirrors [14,15] and fiber Bragg gratings [61] used in some important experimental tests of tunneling time theories. These structures act as photonic barriers for a narrow band of frequencies that approximately satisfy the Bragg condition. Consider, for simplicity, a refractive index variation of the form

\[
n(z) = n_0 + n_1 \cos(2\beta_0 z),
\]

(55)
The complex electric and magnetic fields within the structure are taken as a sum of forward and backward waves

\[
\frac{d^2 \tilde{E}}{dz^2} + \omega^2/c^2 (n_0^2 + 2n_0 n_1 \cos 2\beta_0 z) \tilde{E} = 0. \tag{56}
\]

The complex electric and magnetic fields within the structure are taken as a sum of forward and backward waves

\[
E(z, t) = E_F(z, t)e^{i(\beta_0 z - \omega_0 t)} + E_B(z, t)e^{-i(\beta_0 z + \omega_0 t)}, \tag{57}
\]

\[
H(z, t) = (1/\eta)[E_F(z, t)e^{i(\beta_0 z - \omega_0 t)} - E_B(z, t)e^{-i(\beta_0 z + \omega_0 t)}], \tag{58}
\]

where \(E_F\) and \(E_B\) are the forward and backward components of the field envelopes, and \(\eta = \sqrt{\mu/\varepsilon}\) is the intrinsic impedance of the unperturbed medium. For photonic band gap structures, the carrier wave is always a propagating wave. Propagation through a photonic band gap thus differs in this fundamental way from tunneling through a quantum barrier. It is only the modulation or envelope that is evanescent within the stop band. Within the slowly varying envelope approximation, use of Eqs. (57) in (56) leads to the following coupled-mode equations for the forward and backward fields [22]:

\[
\frac{dE_F}{dz} - i\frac{\Omega}{v} E_F = i\kappa E_B, \tag{59a}
\]

\[
\frac{dE_B}{dz} - i\frac{\Omega}{v} E_B = -i\kappa E_F. \tag{59b}
\]

Here \(\kappa = n_1 n_0 \omega B/2c\) is a coupling constant related to the strength of the refractive index perturbations, \(\Omega = \omega - \omega_B\), and \(v = c/n_0\). With the boundary condition \(E_F(0) = E_0\) and \(E_B(L) = 0\) the field solutions are

\[
E_F(z) = E_0[\gamma \cosh \gamma(z - L) + i(\Omega/v) \sinh \gamma(z - L)]/g, \tag{60a}
\]

\[
E_B(z) = -i[E_0 \kappa \sinh \gamma(z - L)]/g, \tag{60b}
\]

where \(\gamma = \sqrt{\kappa^2 - (\Omega/v)^2}\) and \(g = \gamma \cosh \gamma L - i(\Omega/v) \sinh \gamma L\). The barrier amplitude transmission coefficient is

\[
T = E_F(L)/E_0 = (\gamma/|g|)e^{i\phi_t}, \tag{61}
\]

the phase of which is given by

\[
\phi_t = \tan^{-1}[(\Omega/v) \tanh \gamma L].
\]

The stop band is the frequency region \(|\Omega| < \kappa v\) within which the envelope fields are hyperbolic functions. We thus define a cutoff frequency \(\Omega_c = \kappa v\), which characterizes the width of the stop band. From the phase of the transmission coefficient we obtain the group delay

\[
\tau_g = \frac{d\phi_t}{d\Omega} = \tau_0 \frac{(\kappa^2/\gamma^2)(\tanh \gamma L)/\gamma L - (\Omega/\gamma v)^2 \sech^2 \gamma L}{1 + (\Omega/\gamma v)^2 \tanh^2 \gamma L}, \tag{62}
\]

where \(\tau_0 = L/v\).

It has been shown that the group delay for the photonic band gap structure is identical to the dwell time within the slowly varying envelope approximation [22]. From the field solutions the time average stored energy in the structure is found to be

\[
U = U_0 \frac{(\kappa^2/\gamma^2)(\tanh \gamma L)/\gamma L - (\Omega/\gamma v)^2 \sech^2 \gamma L}{1 + (\Omega/\gamma v)^2 \tanh^2 \gamma L}, \tag{63}
\]
where \( U_0 = \frac{1}{2} \varepsilon_0 \varepsilon_0^2 E_0^2 A L \) is the energy stored in a barrier-free region of the same length. Dividing by the incident power \( P_i = (1/2) \varepsilon_0 \varepsilon_0 \varepsilon_0^2 |E_0|^2 A \), we obtain the dwell time

\[
\tau_d = \frac{U}{P_i} = \tau_0 \left[ \frac{(\kappa^2/\gamma^2)(\tanh \gamma L)/(\gamma L) - (\Omega/\gamma v)^2 \text{sech}^2 \gamma L}{1 + (\Omega/\gamma v)^2 \tanh^2 \gamma L} \right],
\]

which is identical to the group delay. (For very short structures a more exact treatment reveals a negligible self-interference correction which vanishes as \( 1/\omega_c \), where \( \omega_c \) is the carrier frequency \([60]\).) The identity between the group delay (phase time) and the dwell time for a photonic band gap structure is a key result that makes it possible to resolve the apparent superluminality seen in tunneling time experiments. As with all other barriers the group delay saturates as \( L \to \infty \), here reaching the limit

\[
\tau_{g\infty} = \tau_{d\infty} = 1/\gamma v.
\]

At the Bragg frequency \( \Omega = 0 \) and hence the limiting group delay becomes

\[
\tau_{g\infty} = 1/\kappa v = 1/\Omega_c.
\]

This is a general relation that says that the limiting group delay for any filter is the inverse of the cutoff angular frequency. Equivalently, the delay-bandwidth product is given by

\[
\Omega_c \tau_g = 1.
\]

### 4.4. Frustrated total internal reflection

Fig. 11 illustrates a tunneling configuration referred to as frustrated total internal reflection (FTIR). When a beam passing through a glass prism is incident at the glass–air interface beyond the critical angle \( \theta_c \) it undergoes total internal reflection. The field in the air region decays exponentially with distance along the normal \( x \) and propagates along \( y \). If a second prism is brought into close proximity (within a wavelength) to the first one, energy will tunnel across the gap and a propagating field will be excited in the second prism. For a TE- or s-polarized wave, the field in all three regions can be written as

\[
\tilde{E}(x, y, t) = \psi(x) e^{i(k y \sin \theta_c)},
\]
where $k = n\omega/c$ is the wavenumber in glass. Use of this field in the Helmholtz equation leads to

\[
\frac{d^2\psi}{dx^2} + (\omega/c)^2[n^2 \sin^2 \theta] \psi = 0 \quad \text{in the glass regions},
\]

\[
\frac{d^2\psi}{dx^2} + (\omega/c)^2[1 - n^2 \sin^2 \theta] \psi = 0 \quad \text{in the air gap}.
\]

Here, the coefficients of $\psi$ are the squares of the $x$-components of the wave vectors in the different regions. This is a 2D tunneling problem which can be mapped onto the 1D quantum tunneling problem under certain conditions [62]. An important role is played by the Goos–Haenchen shift, the lateral shift of a finite beam upon total internal reflection. The group delay in FTIR has been calculated by several authors for both TE and TM polarizations [62–64]. Here too the group delay and dwell time saturate with the separation between the prisms, a manifestation of the Hartman effect.

5. Quasi-static dynamics of tunneling

In a sense, true tunneling is a fairly slow process requiring wave packets whose temporal envelopes evolve on a time scale longer than the transit time of a light front (traveling at $c$) across the barrier. In other words, the spatial extent $\delta x$ of the wave packet must exceed the barrier length $L$ as depicted in Fig. 12. The interaction is therefore a quasi-static one [18,65] in which the system is always close to steady state, except for an initial turn-on transient that occurs long before the bulk of the wave packet arrives. [See Section 14.2 for a discussion of this transient.] The quasi-static nature of the tunneling process arises from the requirement that the spectrum of the incident wave packet be narrow compared to the height of the barrier or the width of the stop band, as shown in Fig. 13a. The magnitude of the transmission coefficient is thus approximately constant over the entire spectrum of the wave packet and the phase is a linear function of frequency. Under these conditions there is no distortion and the shape of the transmitted wave packet faithfully follows that of the incident. On the other hand, if the spectrum of the wave packet is broad, as shown in Fig. 13b, the transmitted pulse will be significantly distorted since it has substantial spectral components outside the stopband, in regions where the transmission amplitude is not uniform and the phase is nonlinear. Those spectral components outside the stop band do not tunnel but instead “fly” over the barrier. We will see in Section 6 that every tunneling experiment or simulation that has reported superluminal propagation has been done under quasi-static conditions.

The quasi-static requirement for pulses longer than the barrier can be understood by examining a photonic band gap structure with a stop band of width $\Omega_c/2\pi$ (Hz), where $\Omega_c = \kappa v$. Tunneling without distortion requires that the pulse width $\tau_p$ satisfy the condition

\[
1/\tau_p \ll \kappa v,
\]
Fig. 13. (a) Spectrum of a narrow band pulse and transmission function of a typical barrier. For true, distortionless tunneling, pulse spectrum must be much narrower than stopband and (b) spectrum of a broadband pulse compared to transmission function of a barrier. The spectral components outside the stopband do not tunnel. Such a pulse will be distorted.

i.e. the pulse spectrum must fit inside the stop band. Thus, the spatial extent of the pulse $v\tau_p$ must obey the inequality

$$v\tau_p \gg \frac{2\pi}{\kappa}.$$ 

In other words the pulse spatial length should be at least 6 decay lengths. A true barrier (in the opaque limit of interest) should be at least a couple of decay lengths, i.e. $\kappa L \geq 2$. On the other hand, for measurable transmission, the barrier should not exceed four or five decay lengths since $T \sim \exp(-2\kappa L)$. Indeed, for $\kappa L = 5$, $T \sim 0.004\%$ which is negligible. Thus a true barrier with finite transmission will have a length in the range

$$\frac{2}{\kappa} < L < \frac{5}{\kappa}.$$
Fig. 14. (a) Incident (dotted), transmitted (solid) and reference (dot-dashed) pulses for a barrier of strength $\kappa L = 5$ and (b) snapshots of the intra-barrier energy density distribution for a narrowband tunneling pulse. The peak of the incident pulse occurs at $t = 15$. In (a) the tunneled pulse has been normalized by its peak value of 0.00018. (From Ref. [65])

Hence, the pulse length $2\pi/\kappa$ exceeds the barrier length for distortionless tunneling in barriers with finite transmission. A pure evanescent wave is necessarily a quasi-static excitation for which the entire structure responds as a unit, with every spatial point in phase.

For a photonic band gap structure the time-dependent coupled-mode equations for the forward and backward waves in the structure are [18]:

\[
\begin{align*}
\frac{\partial E_F}{\partial z} + \frac{1}{v} \frac{\partial E_F}{\partial t} &= i\kappa E_B e^{-i2\beta z}, \\
\frac{\partial E_B}{\partial z} - \frac{1}{v} \frac{\partial E_B}{\partial t} &= -i\kappa E_F e^{i2\beta z},
\end{align*}
\]

(66a)

(66b)

where $\Delta \beta = n_0(\omega - \omega_B)/c$. Numerical solutions of these propagation equations have been used to explore the dynamics of pulse tunneling within the photonic band gap [18,65]. Fig. 14a shows an incident narrowband Gaussian pulse
multiplies the first derivative of the field envelope is the group delay. For the forward field this group delay actually
to time. These quasi-static solutions are in excellent agreement with the exact numerical solutions. The quantity that
the above barrier components. In this case one can actually follow a peak through the barrier (Fig. 15b).
shorter incident pulses, significant distortion occurs (as shown in Fig. 15a) because of the non-uniform transmission
that the incident peak has traveled to the exit since over 99.999% of the incident energy is in the reflected pulse. For
never appears in the barrier. The transmitted pulse is a much diminished replica of the incident pulse but one cannot say
incident field and deviates only slightly from exponential behavior at the barrier exit. It is important to note that a peak
decay behavior predicted from the steady-state analysis. This steady-state profile moves up and down slowly with the
intra-barrier dynamics reveals that the incident peak does not actually propagate through the barrier and hence input and output peaks are not related by a simple causal translation. Fig. 14b shows snapshots of the spatial distribution of intra-barrier energy density taken at successive instants of 0.5 time units near the peak of the incident pulse. These snapshots show the exponential decay behavior predicted from the steady-state analysis. This steady-state profile moves up and down slowly with the incident field and deviates only slightly from exponential behavior at the barrier exit. It is important to note that a peak never appears in the barrier. The transmitted pulse is a much diminished replica of the incident pulse but one cannot say that the incident pulse has traveled to the exit since over 99.999% of the incident energy is in the reflected pulse. For shorter incident pulses, significant distortion occurs (as shown in Fig. 15a) because of the non-uniform transmission and the above barrier components. In this case one can actually follow a peak through the barrier (Fig. 15b).

For narrowband pulses the tunneling process is a quasi-steady-state phenomenon in which the field envelope throughout the barrier can follow the slow variations of the input envelope with little phase lag. In this quasi-static limit we can obtain approximate solutions to the coupled-mode equations for arbitrary input pulse profiles by expanding the complex amplitudes of the sinusoidal solutions to first order in the frequency parameter $\frac{\Omega}{\kappa v}$ and performing an inverse Fourier transform, whereupon $i\Omega \rightarrow -\partial / \partial t$. The resulting solutions are

\begin{align}
E_F(z, t) &= \frac{\cosh \kappa(z - L)}{\cosh \kappa L} \left\{ A(t) - \frac{1}{\kappa v} \left[ \tanh \kappa L + \tanh \kappa(z - L) \right] A'(t) \right\}, \tag{67a} \\
E_B(z, t) &= -i \frac{\sinh \kappa(z - L)}{\cosh \kappa L} \left\{ A(t) - \frac{\tanh \kappa L}{\kappa v} A'(t) \right\}. \tag{67b}
\end{align}

Here $A(t)$ is the envelope of the incident pulse as measured at $z = 0$ and the primes denote derivatives with respect to time. These quasi-static solutions are in excellent agreement with the exact numerical solutions. The quantity that multiplies the first derivative of the field envelope is the group delay. For the forward field this group delay actually depends on position $z$ within the barrier. However, the only physically accessible field is the transmitted field at the exit

\[ E_F(L, t) = \frac{1}{\cosh \kappa L} \left\{ A(t) - \frac{\tanh \kappa L}{\kappa v} A'(t) \right\}. \]

Since this is just $E_F(L, t) \approx T_0 A(t - \tau_g)$, in the quasi-static approximation every part of the delayed incident pulse experiences the same steady-state transmission $T_0$.

The quasi-static nature of tunneling described here for electromagnetic waves also holds for quantum tunneling. Simulations of the tunneling of both non-relativistic and relativistic particles exhibit the same features seen above. Krekora et al. [66] have shown through a series of snapshots of the wave function solutions of the Dirac equation how a relativistic particle tunnels. Fig. 16a shows incident, reflected and transmitted wave packets. Here, the spatial extent of the wave packet is an order of magnitude greater than the barrier width. It is clear that the duration of the tunneling event is determined purely by the length of the wave packet. The transmitted and reflected wave packets are replicas of the incident wave packet (one greatly diminished) and no distortion or reshaping is seen. The snapshots of the probability density distribution reveal the static distribution, which simply moves up and down, following the incident wave packet with only a small delay.

6. Tunneling time experiments

There have been very few experiments designed to measure the tunneling time of quantum particles. Because it is not possible to monitor a single tunneling electron without altering its state, the measurements have been indirect and
the results are difficult to interpret [67]. The tunneling of classical wave packets, however, can be monitored directly in a non-invasive manner. As a result, since the early 1990s, there have been many electromagnetic, optical, and acoustic experiments to determine tunneling time. We should note at the outset that despite the many reports of measured superluminal group velocities, tunneling time experiments actually measure delays or distances. They never measure velocity. Velocities are always inferred. Anomalously short delay times have been measured. The question then is this: what are these times that are measured? Here we review some experiments in the optical, microwave, and acoustic domains. We focus on those experiments that yield quantitative results, demonstrate key predictions of tunneling theory, or are of historical interest. These experiments do show that the group delay properly describes the occurrence of a diminished transmitted pulse peak and that there is little or no distortion or reshaping of a narrowband incident pulse. Several of the experiments also confirm the Hartman effect, the saturation of group delay with distance. It is the interpretation of these experiments that has been controversial. In this section we will review these experiments. In subsequent sections, we critique the usual interpretation of these delays as a manifestation of superluminal velocity and show how a new alternative interpretation explains all the experimental results without any implications of superluminal barrier traversal.
A common feature of all the time domain measurements is that they monitor the arrival time of a peak of a pulse, or in the case of the single-photon experiments, the simultaneous detection of a tunneling and reference photon. The assumption is then made that the observed delay is a transit time. This classically motivated transit time cannot possibly apply since the transmitted pulse is not the same entity as the incident pulse. Since greater than 99% of the incident wave-packet energy is reflected, there is no sense in which one can say that the incident pulse has traveled to the exit. One can always measure the arrival time of something. However, before this arrival time is related to a “transit time” one must know the departure time of the thing that arrived. The measured group delay is therefore not a transit time. Operationally, the group delay measures the time at which the intensity at the exit reaches a peak, relative to the time at which the intensity at the input would have reached a peak in the absence of a barrier. Those temporal events are not necessarily related by propagation.

6.1. Optical experiments

(i) Steinberg, Kwiat, and Chiao (1993): The first tunneling time experiment at optical frequencies is the now classic work of Steinberg et al. [14]. In a paper that stimulated much of the current interest in superluminal propagation, they reported measuring a group velocity for single-photon tunneling that was as large as $1.7c$. The evidence for such a group velocity is indirect since the experiment is an interferometric one that measures distances and the number of photon counts in coincidence at a pair of detectors.

The experimental setup is shown in Fig. 17. A continuous wave (cw) ultraviolet (UV) laser produces pairs of correlated nearly identical photons through the process of parametric down conversion in the nonlinear crystal KDP.
The two cw beams of photons are sent along two different paths to a 50:50 beam splitter in an arrangement known as a Hong–Ou–Mandel interferometer [69]. In the absence of the multilayer coating that acts as a photonic barrier, the length difference between the two paths can be adjusted by translating the “trombone prism” in one arm of the interferometer. If two photons arrive simultaneously (within their coherence time of 20 fs) at the beam splitter, a quantum interference effect of the two-photon probability amplitudes causes them to pair up and head off in one direction or the other. Only one of the detectors A or B will register a click, meaning that the coincidence count will be zero. The number of coincidence counts as a function of the displacement of the prism is shown as the upper trace in Fig. 18 [14,68]. It exhibits a minimum when the path length difference is zero. The width of the dip shown is related to the coherence time of the photon packets and hence this experiment measures the duration of the wave packets. The photonic barrier used in this experiment was an 11-layer dielectric mirror with quarter-wave layers of alternating high \( (n = 2.22) \) and low \( (n = 1.41) \) refractive index. When this photonic barrier is inserted in one leg, it changes the length difference between the two paths. Steinberg et al. found that they have to lengthen the path containing the photonic barrier in order to equalize the group delays in the two different arms. The lower curve in Fig. 18 shows that the minimum of the coincidence count rate has shifted to the left by a distance \( c\delta \tau \) which yields a relative delay between the two photons.
of $\delta \tau = -1.47$ fs. This correlation measurement thus determines the time shift in the occurrence of a peak of a photon flux that has passed through air and that of a photon flux that has passed through the photonic barrier. It should be noted that the transmission of the photonic barrier is only 1% so that the photon flux that reaches the beam splitter with the barrier in place is much weaker than what would have been measured in its absence. The photonic barrier is 1.1 $\mu$m thick, which means that the group delay for a reference light beam traveling the equivalent distance in free space is $\tau_0 = d/c = 3.6$ fs. The measured group delay for the transmitted photons in the presence of the barrier is thus $\tau_g = 2.13$ fs. The group delay predicted by the method of stationary phase is 1.7 fs and hence the agreement between measured and calculated group delays is fair. The authors then assume that the measured group delay is a transit time in the classical sense and that the tunneling photons follow a path of length $d$. They calculate a classically motivated group velocity of $v_g = d/\tau_g = 1.7c$ and this result is taken as evidence that tunneling single photons travel with superluminal group velocity. The superluminality is attributed to a pulse-reshaping phenomenon although, as emphasized by the authors, the transmission probability of the barrier is a very flat function of frequency near midgap so that there is essentially no distortion of the wave-packet (even though the experiment does not really determine wave-packet shape).

It should also be noted that the wave-packet width of 20 fs (the correlation time) greatly exceeds the 3.6-fs transit time of a light front traversing the same distance in the absence of the barrier. This means that quasi-static conditions hold. This is the case in all situations of distortionless tunneling. It should also be stressed that the experiment actually uses cw beams. The “wave packets” are simply photon fluctuations whose duration is the coherence time. Although this experiment is often cited as proof that single photons tunnel with superluminal group velocity, we will show in Section 11 that the results can be explained without invoking faster-than-light propagation. Also, the “single-photon” aspect of this experiment only plays a role in the detection process. As pointed out by the authors, the tunneling part of the experiment is completely describable by the classical Maxwell equations since it is a linear process [4]: “Propagation effects are then governed by the classical wave equations, and quantization merely affects detection statistics and higher-order effects.” The measured delay is in reasonable agreement with the prediction of the group delay as calculated from the frequency derivative of the transmission phase shift, a classical calculation based on Maxwell’s equations. The tunneling part of the experiment is thus governed by the same physics that applies to all the other photonic barrier experiments described below.

(ii) Spielmann, Szöpcs, Stingl, and Krausz (1994): Spielmann et al. [15] performed the first pulsed tunneling time measurements at optical frequencies using photonic band gap materials. The pulses used had a width (FWHM) of only 12 fs and hence nonlinear autocorrelation techniques were needed to detect them (Fig. 19). The photonic barriers consisted of alternating quarter wave layers ($\lambda/4 \approx 0.2$ $\mu$m) of low index ($L$, fused silica) and high index ($H$, titanium dioxide) materials arranged in the following manner: (substrate) -- ($HL)^n$ -- (air), with $n = 3, 5, 7, 9, 11$. The transmittivities in the middle of the stop band ranged from 0.3 for the thinnest sample to $2 \times 10^{-4}$ for the thickest sample. They measured the difference in group delays for a pulse weakly transmitted through a barrier and a reference pulse traversing the same distance in the absence of a barrier. The results are shown in Fig. 20. The measured delay time differences become negative for barriers with more than 5 pairs of quarter-wave layers. The authors infer from this that the group velocity exceeds $c$ for those cases. The most striking result is that the group delay, obtained by adding $n \times 83$ fs to the delay difference, becomes independent of the number of elements, thus confirming the Hartman effect. For the thickest barriers, some narrowing of the transmitted pulses was observed, which may be explainable by the effects of non-uniform spectral transmission. In fact the bandwidth of the pulses was sufficiently wide that some spectral components were transmitted outside the gap region leading to the rise of a pedestal on the pulse. The pulse width of 12 fs is equivalent to a free-space wave-packet length of 3.6 $\mu$m. This is longer than the barrier lengths of 1.2, 2.0, and 2.8 $\mu$m, for the 3, 5, and 7-period structures thus satisfying the quasi-static tunneling requirements. For the 9- and 11-period structures, the pulse length is equal to or shorter than the barrier length and hence some distortion is to be expected as demonstrated in simulations of the time-dependent coupled-mode equations [65].

As in the Steinberg et al. experiment, the measured delay time difference exceeds that predicted by the stationary phase calculation. Here the discrepancy is about 1.5 fs. This could be due to the insufficiently narrow bandwidth of the pulses used. Calculations aimed at resolving some of these discrepancies have been presented by Laude and Tournois [70] and Pereyra [71].

Except for the thinnest samples, for which some pulse narrowing was observed, this experiment also operated in the quasi-static regime, meaning that the pulse length exceeded the sample thickness. These authors also interpret the measured group delays as transit times: “The measured transit time is found to be paradoxically short (implying superluminal tunneling) and independent of the barrier thickness for opaque barriers…” However, it should be noted
that as much as 99.9999% of the incident pulse is actually reflected and hence it is incorrect to say we are observing the transit time of an incident pulse from entrance to exit. It is certainly not the transit time of the incident pulse, nor is it the transit time of an input “peak” since the peak does not travel through the barrier as discussed in Section 5.
All that can be said with certainty is that in the presence of the barrier, the very weak measured intensity at the exit reaches a maximum sooner than the very large measured intensity does in the absence of the barrier. There is no a priori reason to relate this time-to-maximum to a transit time.

(iii) Longhi, Marano, Laporta, and Belmonte (2001): By far the cleanest and most quantitative pulsed tunneling time experiments done in the optical domain are those of Longhi et al. [61]. Because the pulses used were relatively long (380 ps) they could be measured directly using a detection system with an overall temporal resolution of 15 ps. Thus, it was confirmed directly that there was no distortion or “reshaping” of the transmitted pulse. The barrier consisted of a 2-cm-long silica fiber Bragg grating. The strength of the periodic refractive index variation was $\kappa L = 140 m^{-1}$, resulting in a coupling strength of $\gamma L = 2.8$. The background refractive index is $n_0 = 1.452$ so that the group velocity in the unperturbed medium is $v = 2.065 \times 10^8 m/s$. The transit time in the unperturbed medium is thus 97 ps while the free space transit time over that distance is 67 ps. The pulse length (380 ps) is thus much greater than the free space transit time (67 ps) thus satisfying the quasistatic conditions. The group delay at the Bragg frequency is calculated from the expression $\tau_g = \frac{\tanh(\kappa L)}{\kappa L}$ to be 34 ps. The difference between the reference group delay (97 ps) and the tunneling group delay is thus predicted to be 63 ps. This is exactly what is observed in the experiment. The center wavelength of the pulse was 1.5 $\mu m$ and could be tuned by a few nanometers. It should be noted that the transmission at that wavelength is 1.5%. Use of an erbium-doped fiber amplifier EDFA at the exit amplified the transmitted pulse to a comparable amplitude with the reference pulse. In essence the amplifier acts as an automatic gain control, which equalizes all the signals. Although operating in saturation mode, the distortion introduced by the amplifier was negligible. Fig. 21 shows the calculated and measured transmission function as well as the calculated group delay and the group delay measured with a cw input and a modulation transfer technique. Fig. 22 shows the reference pulse and the amplified tunneled pulse.
The experiment shows impressive quantitative agreement between measured group delays and analytical calculations of the group delay based on the method of stationary phase. This group delay is also identical to the dwell time.

Again it should be noted that the experiment described above measures group delays. Following common practice, Longhi et al. then infer group velocities by assuming that the group delay is a transit time and then dividing the barrier length by the measured delay. The result of that exercise is a “group velocity” greater than 1.97 times the vacuum speed of light. Longhi et al. also plotted the group velocity versus barrier reflectivity (Fig. 23). The figure implies that the group velocity approaches infinity as the reflectivity (transmittivity) of the barrier approaches unity (zero). That of course immediately raises the question: why should a pulse travel faster and faster just because a barrier has been made more repulsive? Why should a pulse travel with infinite velocity when the transmission probability approaches zero?

(iv) Balcou and Dutriaux (1997): In 1968, Agudin[72] proposed a steady-state technique to measure the tunneling group delay in frustrated total internal reflection (FTIR). The scheme is based on the fact that a beam suffers a lateral displacement along the interface as it tunnels across a gap. Fig. 11 is a schematic showing a narrow beam undergoing total internal reflection at an interface between a dense dielectric (say, glass) and a less dense dielectric (say, air). If another dielectric medium is brought into close proximity some distance $d$ from the first interface, some energy from the incident beam can tunnel across the air gap and end up in the second dielectric. The reflected and transmitted beams are displaced some distance $D$ along the interface. Inside the air gap, the electromagnetic fields decay exponentially with distance along the $z$ direction (normal to the interface) while they propagate with an $\exp(ik_y y)$ phase factor along the $y$ direction. Conservation of momentum (phase matching) requires that the $y$-component $k_y$ of the wave vector be the same across the interfaces. Assuming a propagation velocity $v_y$ along the interface, the temporal delay between the incidence of the beam at $A$ and the emergence of reflected and transmitted beams at a distance $D$ along the $y$ direction is given by $\tau_g = D/v_y$, where $v_y = \partial \omega / \partial k_y$. For an incident angle of $\theta_i$, $k_y = (n \omega/c) \sin \theta_i$ hence

$$\tau_g = D \left( \frac{n \sin \theta_i}{c} \right).$$

Since all the quantities in parenthesis are known, a measurement of the displacement $D$ then yields the group delay. In practice, it is easier to measure a related quantity $D_\perp$ which is the shift between the incident beam axis and the transmitted beam axis in a direction normal to the two beams.

Balcou and Dutriaux [73] carried out such an experiment using two right-angled prisms of fused silica ($n = 1.409$) and a cw Gaussian beam of wavelength 3.39 $\mu$m. Because any finite beam has a spread of wave vectors, it is necessary to take into account the differential transmission for different incident angles in the beam. When this is properly accounted for, it turns out that there is a shift in the mean wave vector of the transmitted beam, which corresponds to a slight change in the angle of the transmitted beam from the value $\theta_i$ predicted for perfectly collimated beams. Balcou and Dutriaux showed that this angular deviation $\delta \theta_i$ could be related to the imaginary part of a complex time proposed by
Pollak and Miller. Referring to this time as the “loss time” $\tau_L$, they show that

$$\tau_L = \left( \frac{2z_R}{nc\cos(2\theta_i)} \right) \delta \theta,$$

where $z_R$ is the Rayleigh length of the incident beam. Balcou and Dutriaux measured the beam shift and angular displacement as a function of the width of the air gap for both TE and TM polarized light. From these measurements they inferred the group delays in reflection and transmission for both polarizations as well as the loss time in transmission. The results are shown in Fig. 24a–f. It is seen that the group delays for both polarizations saturate with barrier width, thus confirming the Hartman effect for frustrated total internal reflection. It is also evident that the group delays in reflection are equal to the group delays in transmission as predicted for a symmetric barrier. Furthermore, the error in the reflection group delay is much smaller than the error in the transmission delay owing to the fact that the reflected beam intensity greatly exceeds the transmitted beam intensity, allowing for a much better signal-to-noise ratio (SNR). Indeed for any symmetric barrier one is better off measuring reflection group delay than transmission group delay since it yields the same result with better SNR. The group delays for the TM polarized beam are also larger than the delays for the TE polarized beam. For thick enough barriers (in the saturation regime) the transmission delay is much shorter than the delay that would have been suffered by a beam traversing the same distance in the absence of a barrier. The authors infer from this that the tunneling beam traverses the barrier with superluminal group velocity. In fact,
for a width of 20 $\mu$m the authors would infer a group velocity of $5 \times 10^8$ m/s for a measured delay of 40 fs. It should be kept in mind however, that this is a cw measurement and that nothing is being transported at that speed.

The “loss times” are relatively insensitive to the polarization state of the incident beam. For thin barriers the loss times are relatively flat and then increase linearly with barrier thickness without saturating (Fig. 24c,f). The authors further infer that the loss times imply subluminal effective traversal velocities. They compare the loss times with the predictions of the semiclassical time. They find that the data points are parallel to the semiclassical times. However, there is a rather large discrepancy of about 200 fs between the two times. Notwithstanding this obvious disagreement with the semiclassical time, the authors reach the surprising conclusion that “our results substantiate the semiclassical time as the most relevant to describe the physics of tunneling”. They reach this conclusion on the basis of an argument that a tunneling time should not depend on boundary conditions, “which are not really part of tunneling”. This argument, unfortunately is not correct. As Chiao and Steinberg [4] point out, tunneling indeed depends on boundary conditions. In fact, for something like the multilayer periodic structure, tunneling comes about through the many dielectric reflections at the interfaces of the multilayers. These reflections are certainly boundary effects. Balcou and Dutriaux also make the point that the most important question is how long the tunneling wave couples to other degrees of freedom inside the barrier. However, that coupling should depend on how much energy makes it into the barrier, a quantity that certainly depends on boundary conditions, i.e. whether a beam is TE or TM polarized. We conclude therefore that their conclusion regarding the relevance of the semiclassical time are incorrect. Their experiment should really be seen as a beautiful confirmation of the Hartman effect and of the equality of transmission and reflection group delays for symmetric barriers. It should be noted that Carniglia and Mandel [74] in an earlier experiment had noted the saturation of the transmission phase shift. This result can also be seen as indirect confirmation of the Hartman effect.

6.2. Microwave, radio-frequency, and terahertz experiments

(i) Ranfagni, Mugnai, Fabeni, and Pazzi (1991): The first electromagnetic test of the various tunneling time definitions was by Ranfagni et al. [75]. Their barrier consisted of a rectangular metallic waveguide with a constricted segment, which supported evanescent waves. The input pulses were approximately step functions. Because of their sharp rise, the bandwidth of these pulses was quite large. Furthermore, most of the experimental results were for operation above cutoff. Very few results were obtained in the interesting tunneling region below cutoff. The few measurements in this region appeared to agree with the Buttiker–Landauer time, though any agreement should be considered coincidental since the experiments did not involve the modulation of a barrier but the tracking of a pulse feature. Overall, these experiments did not have the accuracy of the later ones and the authors could not conclusively differentiate between competing tunneling time definitions. A refinement to their setup permitted measurements farther into the region below cutoff [76]. They measured both the delay of a step input and the delay of the envelope of a beat signal between two closely spaced frequencies. These measurements of delay did show agreement with the group delay.

(ii) Enders and Nimtz (1992): Enders and Nimtz [13,77,78] carried out a frequency-domain (steady state) experiment to test the predictions of superluminal barrier traversal. They used rectangular metal waveguides with a narrowed section such that waves are evanescent in that region. By using a network analyzer they measured the magnitude and phase of the transmission coefficient as a function of frequency for barrier lengths of $L = 40, 60, 80,$ and 100 mm. They found that the transmission phase shift was practically independent of the length of the cutoff section in agreement with the Hartman prediction for the analogous quantum mechanical barrier. The magnitude of the transmission coefficient was as low as 0.00064 for the longest barrier. The frequency-domain transmission data was then transformed to the time domain by means of discrete Fourier transforms to create a hypothetical tunneling pulse. For the longest barrier the group delay was 130 ps as compared to a 333-ps delay for a reference pulse traveling the same distance in vacuum. The group delays for all four barrier lengths were identical, again providing experimental confirmation of the Hartman effect. Enders and Nimtz concluded that the barrier traversal velocity in the narrowed waveguide is superluminal. It should be noted however that this was a completely cw experiment, that nothing was actually transported at greater than the speed of light, and that what was measured was a frequency-dependent phase shift. The actual data upon which they base all their conclusions regarding superluminal transport is shown in Fig. 25. The “pulse” was synthesized mathematically through a Fourier transform with a Kaiser–Bessel weighting function and had a FWHM of 3 ns. This corresponds to a spatial extent of 90 cm, much greater than the 10-cm length of the thickest barrier and in line with the quasi-static requirement for distortionless tunneling. Enders and Nimtz [78] also tried direct time-domain pulse
tunneling experiments using step input pulses with very short rise times. The results however were not definitive and suffer from interpretational problems because of the presence of significant frequency components above cutoff. It must be noted then, that the only really clear group delay measurements carried out by this group involved cw excitations. All that can be said is that a frequency-dependent phase shift of cw electromagnetic waves has been measured. That frequency-dependent phase shift is independent of the length of the evanescent region.

In their paper entitled “Zero-time tunneling of evanescent mode packets” Enders and Nimtz [78] claimed that for opaque barriers the traversal time was zero: “an instantaneous traversal of electromagnetic packets through space seems possible”. They based this claim on the fact that adding further length segments did not increase the delay and thus the pulse must have covered the extra distance in no time. We show later that this independence on length is what one would expect for a lifetime and not a transit time.

(iii) Mojahedi, Schamiloglu, Hegeler, and Malloy (2000): Mojahedi and coworkers have also carried out microwave tunneling experiments using 1D photonic band gap structures made of stacks of dielectric slabs [79–81]. Both frequency-domain and time-domain measurements were reported. In the frequency domain they measured the magnitude and phase of the transmission function versus frequency. Group delay was calculated by taking the frequency derivative of the transmission phase shift. From these delay calculations, a group velocity was inferred by dividing the barrier length by the group delay. Within the stop band they inferred group velocities as large as 2.1c. In the time-domain experiments, they used microwave pulses of width 9.1 ns (FWHM). For a photonic barrier of length 22.75 cm they inferred a tunneling group velocity of 2.3c from group delay measurements. It should be noted that this experiment is also in the quasi-static regime since the spatial length of the pulse is 2.73 m, more than an order of magnitude greater than the barrier length. A notable feature of the pulsed experiment is that there was very little distortion of the tunnelled pulse. A very slight broadening (about 2%) was observed.

(iv) Hache and Poirier (2002): Hache and coworkers [82] have carried out delay measurements at radio frequencies on an electromagnetic periodic structure made of coaxial cables with alternating characteristic impedances. The structures act as filters with a stop band centered around 10 MHz. They used Gaussian pulses of width 4 µs to measure group delays of order 150 ns within the stop band of a 120-m-long structure. From such delay measurements they
inferred group velocities of around $3c$. Here too it should be noted that the spatial extent of the pulse is 1200 m, an order of magnitude greater than the barrier length. This experiment is also therefore a quasi-stationary one. The title “Long-range superluminal pulse propagation” is somewhat misleading since the “propagation distance” is only one-tenth the pulse length! Locating the peak of such a long pulse in a region of high attenuation is a process that is subject to large errors as can be seen from the data in Fig. 26. The authors explain the superluminal group velocities with the usual reshaping argument that the front of the pulse is attenuated less than the rear. However, the transmitted pulses do not show any evidence of such preferential transmission. In fact, if anything, the front of the pulse appears suppressed relative to the rear (Fig. 27).

(v) J. J. Carey, Zawadzka, Jaroszynski, and Wynne (2000): In a paper entitled “Non-causal time response in frustrated total internal reflection?” Carey et al. [83] reported measuring tunneling times of single-cycle terahertz pulses (pulse width 0.8 ps) through an air gap between two Teflon prisms. They claimed that “both the phase and energy of the pulse travel faster than the speed of light in vacuum.” We should note that the claim of energy velocity being greater than the speed of light in vacuum could not possibly be true since most of the incident energy is in fact reflected. The energy that gets through the barrier is always less than what would have been in a freely propagating pulse at the barrier exit. The authors also made a number of statements that suggested that causality was violated in this
experiment. We quote some of these statements: “Theoretical analysis of the experiments shows that the time-response function for electromagnetic waves propagating in the air gap is non-causal.” “Theoretical analysis shows that in FTIR part of the incoming pulse travels backwards in time. This apparently violates the principle of causality in a way that is different from the shadow effect.” “The theory used to describe these data appears to show that the principle of causality does not apply in FTIR.” In a Comment on this paper, Mochan and Brudny [84] pointed out that the apparent violation of causality stemmed from an improper specification of a path for the tunneling wave as well as the neglect of the transverse spatial extent of the field. Carey et al. had assumed that an evanescent wave follows a path normal to the prism face. However, under total internal reflection, one cannot assign a path to the tunneling wave since the “angle” of transmission is an imaginary quantity. In their response [85] the authors appear to soften their earlier claims saying “we never argued that causality is violated although it may appear that some features of the pulse cross the gap instantaneously or indeed backwards in time.” In another experiment which we discuss next, the group of Grischkowsky [86] demonstrated conclusively that there is nothing acausal in the FTIR tunneling of terahertz pulses and that Carey et al. erred in their analysis and interpretation.

(vi) M.T. Reiten, Grischkowsky, Cheville (2001): Reiten et al. [86] carried out a similar terahertz experiment and obtained results that directly contradict the earlier claims of non-causal propagation. The experimental setup is shown in Fig. 28. The pulses used are typical broadband terahertz pulses with spectral content from near-DC to 3 THz and a nominal pulsewidth of order 1 ps. The transmitted pulses were measured for several different prism separations (Fig. 29) and compared with a completely causal theory. The results show that the transmitted field peaks earlier. However, at the widest separation, the transmitted energy is approximately 22,000 times smaller than the energy in a reference pulse that does not tunnel. The spectrum of the transmitted field is shifted to lower frequencies as the higher frequencies are rejected. This leads to significant broadening of the transmitted pulse. The pulse length as measured at the exit exceeds the width of the air gap. In fact at the widest gap of 1000 μm the 0.2 THz peak of the transmitted spectrum corresponds to a wavelength of 1.5 mm, which is longer than the barrier. The barrier acts essentially as a lumped element. We are again in the quasistatic regime. Reiten et al. prove that their complex transfer function is causal by showing that its real and imaginary parts satisfy the Kramers–Kronig relations. They point out that since one is unable to measure or define a path for the tunneling wave through the system, a group velocity cannot be assigned. We will show that the time delay seen is simply a measure of the lifetime of stored energy in the Fabry–Perot resonator formed by the gap.

6.3. Acoustic experiments

Tunneling is inherently a wave phenomenon and hence, like diffraction and interference, can occur with any kind of wave, from quantum mechanical matter waves to electromagnetic and acoustic waves. Guided by this reasoning, a number of workers have investigated the issue of tunneling time of acoustic waves and phonons from both theoretical and experimental perspectives [87–91]. These studies have revealed phenomena analogous to the Hartman effect (independence of tunneling time on barrier length) and the superluminality seen with electromagnetic waves. Researchers have reported experimental results on tunneling acoustic waves with headlines such as “Breaking the sound barrier” [91], implying that tunneling sound waves travel with supersonic velocities. Here, we will review some
Fig. 29. (a) Transmitted terahertz pulses for different gap lengths and (b) calculated and measured spectra for transmitted terahertz pulses in FTIR experiment. (From Ref. [86])
of the key experiments on acoustic wave tunneling. As in other tunneling experiments, it will be noted that the pulses used are much longer than the length of the acoustic barrier so that the interaction occurs under quasistatic conditions. Secondly, all these experiments measure a delay time. A group velocity is then inferred by dividing the barrier length by the measured delay.

(i) Yang, Page, Liu, Cowan, Chan, Sheng (2002): Yang et al. [90] studied the tunneling of ultrasound pulses through 3D phononic crystals. Their goal was to determine whether the tunneling time of ultrasound waves is independent of the length of the phononic crystal and whether the tunneling time is less than the transit time in a barrier-free region. Their phononic crystal consists of a close-packed face-centered-cubic (fcc) array of 0.8-mm-diameter tungsten carbide beads in water. Along the [1 1 1] direction, the structure possesses a phononic band gap between 0.8 and 1.2 MHz. The input acoustic pulses had a bandwidth that exceeded the width of the stopband and hence considerable distortion of the transmitted pulses was seen. Group delay data was obtained by digitally filtering the output pulse using a narrow Gaussian filter of bandwidth 0.01 MHz. This filtered output was compared with a similarly filtered signal transmitted through the substrate without the phononic crystal sample. The delay of the peak of the tunneling pulse is shown in Fig. 30 as a function of sample thickness. The experimental delay times saturate as the sample thickness is increased, in agreement with the theoretical predictions. This can be seen as a manifestation of the Hartman effect for sound waves. There is a discrepancy between the actual experimental values and the theoretical predictions, which has been attributed to the presence of absorption. The delay was also found to be inversely proportional to the width of the stopband so that

\[ \tau_g \Omega_c \sim 1. \]

This agrees with the result obtained in Section 4 for the limiting group delay of a photonic band gap structure: 

\[ \tau_g = 1/\Omega_c. \]

The authors then convert the group delays to group velocities by dividing sample length by the measured delays. The result of this procedure is shown in Fig. 31. The group velocity thus determined is seen to increase with sample thickness and eventually exceeds the maximum sound velocity in any of the bulk materials. Again the discrepancy between theory and experiment is ascribed to absorption. The conclusion of Yang et al. is that the transmission of pulses through the phononic band gap involves tunneling because the group velocity increases with barrier thickness. In other words, they assert, as done by many others, that a signature of tunneling is a group velocity that increases with barrier thickness.

(ii) Robertson, Ash, and McGaugh (2002): Another tunneling experiment with acoustic waves has been carried out by Robertson and coworkers [91]. A distinguishing feature of this experiment compared to the one by Yang et al. is that the bandwidth of the acoustic pulses used was sufficiently narrow that their spectrum was entirely contained within the stop band of the acoustic band gap structure. As a result the transmitted pulses were undistorted. The pulse length of 20 ms is much greater than the transit time of 1.19 ms through the unperturbed waveguide hence the interaction is a quasi-static one. The acoustic band gap structure consists of a waveguide of standard 3/4-in PVC plumbing pipe to which is attached a sidebranch array of shorter waveguides spaced 20.5 cm apart (Fig. 32). This periodic structure serves as a filter with several stop bands. The stopband used in these experiments extends from 650 to 1100 Hz.
A Gaussian modulated pulse at carrier frequency 850 Hz was transmitted through a straight guide without sidebranches to serve as a reference. The pulse was then sent through filters with 3, 4, and 5 sidebranches and the delay determined from the arrival time of the peak of the pulse or of the centroid. The measured arrival times were 0.85, 0.85, and 0.89 ms for the three, four, and five sidebranch filters, respectively. It is seen that these delay times are relatively insensitive to the length of the barrier, a manifestation of the Hartman effect. **Fig. 33** shows the direct and tunneled pulses recorded for the four-sidebranch filter. The peak of the tunneled pulse (scaled by a factor of 32) precedes the direct pulse by
a slight but perceptible amount (0.94 ms). This shift is a very small fraction of the pulse width. More importantly, the tunneled pulse does not appear to be distorted or reshaped.

Based on the measured delays and the filter lengths $L$, the authors infer group velocities of 480*, 720, and 920 m/s for the three ($L = 0.41$ m), four ($L = 0.65$ m), and five ($L = 0.82$ m) element arrays, respectively, by dividing the lengths by the delays. [* Note: the group velocity for the 0.41-m filter is misprinted as 695 m/s in the original article.] Given that the speed of sound in the unperturbed waveguide is 344 m/s, the authors suggest that they have broken the sound barrier by sending acoustic pulses through an acoustic band gap. These “supersonic” velocities are ascribed to a pulse-reshaping phenomenon even though the tunneled pulse replicates the shape of the incident pulse.

6.4. Summary of experimental findings

The optical, microwave, and acoustic tunneling time experiments have confirmed the following facts:

- The group delay (phase time) indeed describes the time at which the transmitted pulse peaks at the exit.
- The group delay for the (much attenuated) pulse transmitted by the barrier is shorter than that of a pulse traversing the same length of free space.
- The group delay in reflection equals the group delay in transmission for a symmetric barrier.
- The group delay saturates with barrier length (Hartman effect).
- There is no reshaping or shortening of the transmitted pulse.
- Distortionless tunneling is a quasi-static phenomenon requiring pulses longer than the barrier.
- Both narrowband pulsed and cw measurements yield the same value for the group delay.

Any viable theory of tunneling time must offer a consistent explanation of all of these experimental findings. We will now address the reshaping argument that has become the standard explanation of all these “superluminal group velocities” and show why it does not explain the experimental results.
7. The “reshaping” argument

The existence of “superluminal” group velocities in tunneling has been attributed to a reshaping phenomenon in which the barrier transmits the early parts of the incident pulse and rejects the later parts, acting in essence as a time-dependent shutter [4,14,16]. This argument was first put forth by Chiao et al. [16] in analogy to similar arguments used by Garrett and McCumber [92] to explain superluminality in allowed propagation through an absorbing medium. It is useful to quote a recent semi-popular article that describes the current understanding: “In all cases, the pulse that emerges from the tunneling process is greatly attenuated, and ‘front-loaded’—only the leading edge of the incident pulse survives the tunneling event without being severely attenuated to the point that it cannot be detected. If we measure the speed by the peak of the pulse, it looks faster than the incident pulse [19].” Fig. 34 taken from that article shows a schematic of this reshaping description. A consequence of the reshaping argument is that the transmitted pulse is narrowed [93,94] and that its peak is made up only of the leading parts of the incident pulse. [See Section 14 for a debate on this issue.]

The reshaping argument has been accepted as the explanation for apparent superluminality in spite of the fact that there is no experimental evidence in its favor. A barrier that only transmits the early parts of a pulse and rejects the latter parts would necessarily distort an incident pulse. However, the exact numerical solutions of the coupled mode equations for a photonic barrier show that the transmission is the same for all parts of the delayed input pulse, at least over the detectable bulk of the pulse [18,65]. This constancy of the transmission is also in agreement with experimental observations for pulses whose bandwidth is narrow compared to the stopband of the barrier. Indeed, it is curious that most experimenters point to the absence of distortion or reshaping in their tunneled pulses at the same time as they offer up a reshaping argument as the explanation for the observed short delay [14,82]. It should be borne in mind that tunneling is a quasi-static process requiring pulses whose duration greatly exceeds the transit time of a light front (propagating at c) across the barrier. Under these conditions, the transmission of the barrier is in steady state and every part of the delayed pulse experiences the same transmission after an initial brief transient that lasts about one transit time. This transient occurs in the far wings of the incident pulse and has nothing to do with the actual tunneling process. Fig. 35a shows the incident and transmitted pulses for a long pulse tunneling through a barrier of strength $kL = 4$. In Fig. 35b, the two normalized pulses are overlaid so that their shapes can be compared. On this scale their shapes are identical. Notwithstanding claims to the contrary, there is no reshaping seen in theory or experiment. For a slowly varying input pulse, every portion of the main part of the transmitted pulse (after an initial transient) is delayed by the
Fig. 35. (a) Incident, reflected, and reference pulses for tunneling through a photonic band gap and (b) the normalized tunneled pulse overlaid with the incident pulse. Note the absence of “reshaping” or distortion on this scale.

same amount from the incident pulse. Any distortion or reshaping is due to the higher-order terms in the expansion of the transmission phase. Those distortion terms go to zero in the limit that the pulse length approaches infinity and yet the delay stays finite. It is not a mechanism for the prompt appearance of the transmitted peak. A pulse that is sufficiently long will not experience any reshaping. Thus “reshaping” cannot be seen as an essential part of tunneling dynamics. It is rather a sign of an approximation gone wrong. “Reshaping” cannot explain why the group delay measured for narrowband pulses is the same as that calculated from the transmission phase shift of cw waves. For a cw input there is no “early part” to be transmitted and a “later part” to be rejected. Furthermore, no one has been able to explain the Hartman effect in any quantitative way on the basis of reshaping arguments. It does not explain why the group delay in reflection equals the group delay in transmission for a symmetric barrier and why reflected and transmitted pulses have the same shape as the incident pulse. If the early part of a pulse is transmitted and the later part rejected, then as we make the pulse longer and longer we would have a short transmitted pulse (the early part) and a very long reflected pulse (all the rest of the long incident pulse). This is not what is seen in numerical simulations, which show that reflected and transmitted pulses have identical shapes.

Another flaw in the reshaping argument can be stated thus: if the barrier transmits only the early parts and rejects the latter parts of a pulse, then as we make a pulse longer and longer, the transmitted pulse should stay the same since
only the front part contributes to it. In fact, as also pointed out by Landauer [17], if the transmitted pulse only comes from the leading edge of the incident pulse, then if you have a pulse that is one day long (cw input) nothing should get through after the first picosecond or so. Yet we know that there is a cw transmission which is comparable to the transmission observed with the typical pulses used in experiments, pulses that are quasi-cw. Fig. 36 illustrates this flaw in the reshaping argument. The dotted curve is the hypothetical transmitted pulse made up of the leading parts of the incident Gaussian pulse shown. Also shown is a cw input that turns on in the same manner as the Gaussian pulse. According to the reshaping argument, since only the leading edge gets through the barrier, the transmitted pulse should look the same as before, with all the subsequent cw parts rejected. This is difficult to reconcile with the fact that the barrier indeed has a cw transmission so that the output in steady state would equal the peak value of the transmitted pulse shown.

It does take time for the barrier reflectivity to build up to its steady-state value. That build up process, however, occurs in the far wings of the input pulse, long before the main part of the pulse arrives. That part of the pulse, the front or “turn on” part contains the high-frequency components, which do not tunnel because they lie outside the stop band. That portion is necessarily “reshaped”. However, that is not what is normally meant by the reshaping of a pulse as it does not lead to any forward shift of the pulse peak. Furthermore, that portion has nothing to do with the tunneling process. The front propagates at $c$ and the reflectivity builds up in a couple of transit times. [See discussion on transients in Section 14.]

Our conclusion from an examination of all the experimental evidence is that there is no reshaping of the transmitted pulse and hence the observed short group delay is not due to “pulse reshaping.” This conclusion is supported by numerical simulations of the propagation equations.

8. The problem with superluminal tunneling velocities

As we have seen, theory predicts and experiment confirms that the group delay in tunneling can be paradoxically short. Based on these calculated or measured group delays, group velocities have been inferred by dividing the width of the barrier $L$ by the delay $\tau_g$: $v_g = L/\tau_g$. This is a practice that is almost universally followed in this field and one that, on the face of it, is perfectly innocuous. This procedure results in superluminal and indeed unbounded group velocities as the delay saturates at a constant value while the barrier length increases. However, there is a major assumption hidden in this simple procedure. The assumption is that the group delay is a transit time or propagation delay, the time it takes a well-defined object to travel from point $A$ to point $B$, passing through every point in between. Classical notions of velocity assume a well-defined path or trajectory linking points $A$ and $B$. The problem is that quantum mechanics does not permit us to specify one in classically forbidden regions. Similarly, classical evanescent waves in forbidden regions
A

\[z = 0\]

(t = 0)

(a)

A

B

C

\[z = L\]

(t = \tau)

(b)

Fig. 37. (a) The classically motivated notion of transit time requires that the particle “A” at \(z = L\) be the same as the particle “A” at \(z = 0\) and (b) in tunneling there are three different objects involved: incident wave packet (pachyderm A), transmitted wave packet (B), and reflected wave packet (pachyderm C). The figure illustrates the relative scale of the three objects.

cannot be assigned one definite path. The particle or wave packet follows all causally accessible paths, paths which interfere with each other. Indeed it is possible to derive the group delay by summing over all these paths [93–95]. Hence, taking the group delay in tunneling as a transit time from A to B along the path of length \(L\) is a flawed assumption. We must stress the point that no one has ever proven that the group delay in barrier tunneling is a transit time: it is only an assumption.

There are other reasons to question this assumption. Buttiker and Landauer [29,96] have pointed out that there is no obvious causative relation between an input pulse peak and a transmitted pulse peak. In fact, even before this work, Hartman had noted that in the case of quantum tunneling, dispersion in the propagation up to the barrier causes the higher momentum components of the wave packet to reach the barrier ahead of the peak of the incident wave packet. The transmitted packet can therefore attain a peak even before the incident peak arrives at the barrier since its peak is now formed largely by those higher momentum components. Clearly then, in the case of dispersive propagation leading up to the barrier, an incident peak does not necessarily turn into a transmitted peak. This observation should have settled the issue once and for all. However, in the experiments with photonic band gap structures, the propagation of the optical pulses is non-dispersive in the approach to the barrier. The pulses are not chirped and hence every part of the pulse has the same frequency content. The argument of Hartman that the faster frequency components reach the barrier first does not apply. Here there are other reasons to question the notion that the input peak propagates through the barrier and becomes the output peak. Numerical simulations have shown definitively that the pulse peak does not even enter the barrier so that input and output peaks are not related by a simple causal translation [18,65].

Classical notions of transit time require that the incident particle and transmitted particle be the same entity. A particle “A” enters a region (for our purposes here assumed 1D) at \(z = 0\) at time \(t = 0\). At time \(t = \tau\) that same particle “A” leaves the region at \(z = L\). We are then justified in calling \(\tau\) the transit time of the particle “A” from \(z = 0\) to \(L\). When applied to wave packets the concept of transit time requires that the wave packet at \(z = L\) be substantially the same as the one at \(z = 0\). In tunneling there are three different objects involved: an incident pulse A, a transmitted pulse B, and a reflected pulse C. The scale of these objects is shown in Fig. 37. When one speaks of a “transit time” what is it that has made the transit? It is certainly not the incident pulse A. In tunneling, the transmitted pulse is not the incident pulse that has propagated to the exit. In fact it is the reflected pulse that is substantially the same as the incident pulse, containing more than 99.9% of the incident pulse energy. The transmitted energy of 0.1% of the incident energy cannot be used to mark the “transit time” of the incident pulse. The “transit time” is also not the transit time of the transmitted pulse B since B does not exist at \(z = 0\). The classically motivated concept of transit time simply has no meaning in a tunneling context when most of the incident pulse is reflected. To loosely speak of a “transit time” in the context of tunneling is
inappropriate without specifying exactly what it is that has made the transit. Some speak of the transit time of the peak of the wave packet, and yet, as seen in simulations, the peak of the pulse does not travel through the barrier.

The list of arguments against the group delay being a transit time is a long one. It should be obvious that since the group delay saturates with barrier length, it cannot be a transit time unless we assume that the wave packet is smart enough to adjust its velocity by just the right amount to cover the increased distance in the same amount of time. If we interpret the group delay as a transit time, we are faced with a group velocity that increases with barrier length. But how does the wave packet know that the barrier length has been increased and so it should speed up? Another reason to question the identification of group delay with transit time is the fact that for symmetric barriers the transmission and reflection group delays are identical. There is no obvious reason why that should be the case since transmitted and reflected wave packets would generally travel different distances if the group delay is a transit time.

Finally, for the electromagnetic structures of interest the group delay is identical to the dwell time, a time that is certainly not a traversal time [1,29,31,37]. As discussed in Section 2, the dwell time is a property of an entire wave function and does not distinguish between transmitted and reflected components. For particles, it is the time spent in the barrier region averaged over all incoming particles regardless of whether they are transmitted or reflected. It is clearly not a transit time when most of the incident particles are reflected and hence no one would propose the dwell time as a candidate for traversal time. It is interesting that Chiao and Steinberg [4] state in their review article: “The dwell time may appear unsatisfactory as a candidate for several reasons. Foremost, it is a characteristic of an entire wave function, comprising both transmitted and reflected portions”. What has hitherto not been realized is that the group delay in tunneling has exactly the same status since the transmission coefficient upon which it is based results from the interference between forward and backward waves within the barrier and is a characteristic of an entire wave function. Indeed, without the backward (reflected) component there would be no net transmitted flux. The group delay and dwell time have therefore the same status and are, in fact, identical for a photonic band gap structure. Neither of them is a transit time and neither should be related to a “group velocity”.

The association of a tunneling group delay with a group velocity leads to a major logical conundrum. Group velocity in quantum mechanics is understood to be the velocity with which a massive particle travels [25]. As such, we cannot allow superluminal or infinite values since relativity forbids such particles to travel faster than the speed of light. Explanation of the superluminal group velocity as a “reshaping” phenomenon also beg the question: can particles travel faster than light? In their interpretation of the single-particle tunneling experiment, Steinberg, et al appear to answer that question in the affirmative, at least in the case when those particles are photons [16]: “A photon tunneling through the barrier is therefore most likely to arrive before a photon traveling unimpeded at the speed of light. Our experiment confirmed this prediction.” Granted that photons are massless particles and need not obey any speed limit, yet the experiments are designed to test predictions of superluminal propagation for massive particles described by the Schrödinger equation. The inference then is that such particles also travel faster than light. But detectors detect energy and what gets transmitted through the barrier is always less than what would have been there at the same instant in the absence of the barrier.

The difficulties brought about by this interpretation are endless. For example, a purely evanescent wave has the following form in an infinite medium:

$$e^{-kz \cos \omega t}.$$  

For such a wave the disturbance at every spatial point moves up and down in phase: it merely stands and waves. There is no phase shift. Because there is no phase shift, over any distance there is no delay. The group delay is zero. This zero group delay has been taken to mean that evanescent waves travel with infinite velocity [97]. That, by itself, is an oxymoron: an evanescent wave, by definition, is a non-propagating entity. If it is traveling it cannot be evanescent.

It is this identification of group delay with a transit time and a forward group velocity that ignites the whole debate about causality and the possibility of superluminal information transfer. If this delay is not a transit time then the issue of causality is a moot one. On the subject of superluminal communication, what is not mentioned in the debates is that the rate of information transfer is determined by system bandwidth and not by group velocity. True tunneling is an extremely narrowband process, one that does not permit any rapid modulation. Any attempt to increase the rate of information transfer through the rapid modulation of a tunneling carrier wave is doomed to failure, as the modulation sidebands will certainly extend beyond the stopband of the system.

It has also been argued that superluminal photons cannot be used to transmit information because the photons are emitted at random times. But the very foundation of information theory is the notion of randomness, of uncertainty,
of probability. If we were absolutely certain of what was going to be transmitted we would not gain any information upon its receipt. Thus, we must conclude that if photons indeed can propagate with superluminal velocity, they must be able to convey information, notwithstanding arguments to the contrary. We cannot have it both ways: the photons (as particles) travel superluminally or they do not. If the photons (particles) are traveling with superluminal velocity, they should be able to transmit information superluminally and the statistical nature of their emission is part of the uncertainty that is inherent to any communication system.

Group velocity is meaningless in a situation where a wave packet is much longer than the propagation region whereas group delay remains a meaningful, well-defined concept.

9. The meaning of group delay in barrier tunneling

If the tunneling group delay is not a transit time, then what is it? What is this delay time that is measured in tunneling time experiments? Our answer to this question is this: the group delay in tunneling is a lifetime and not a transit time [22,57,98]. For electromagnetic pulses it is a lifetime of stored energy leaking out of both ends of the barrier. For quantum wave packets it is the lifetime of a transient scattering state, or equivalently the storage time of probability density within the barrier plus any time spent “dwelling” in front of the barrier.

There have been earlier attempts to attach a meaning other than transit time to the group delay in tunneling. Campi and Cohen [99] suggest that “what is calculated is not the transit time of a particle traversing a given distance, but is in fact the phase delays encountered by the packet at the barrier surfaces.” While this is technically true, it does not offer any deeper insight into the mechanism responsible for the delay. That work appears to have been totally ignored in the ensuing four decades. Landauer and others have argued that phase is accumulated as the wave functions adjust in relative size at the boundary. Yet the group delay is still interpreted by them as a transit time [3]. Nimtz takes the view that the group delay is purely due to phase accumulated at the barrier entrance and that the wave packet spends zero time inside the barrier [7,78]. This cannot be taken too seriously since both quantum and electromagnetic calculations do show a field or probability amplitude within the barrier.

The first suggestion that the group delay in tunneling is a lifetime and not a transit time was contained in Winful’s Optics Express paper of 2002 [22]. That paper argued that the group delay is identical to the dwell time and hence is related to the Q or quality factor of an electromagnetic cavity. It should be noted that Smith in his original definition of the dwell time called it Q. “It will be remarked that the definition of the lifetime in terms of the ratio of particles trapped to flux in or out is reminiscent of the definition of the Q of an oscillating system in electromagnetic theory” [30].

In the presence of reflections the group delay and the dwell time both relate to the simultaneous escape of energy through both ends of the barrier. Neither of these times can be assigned to just the transmitted pulse or just the reflected pulse, in the sense of the time it takes a well-defined pulse to travel from A to B. During tunneling the pulse or wave packet is interacting simultaneously with both boundaries of the barrier (Fig. 38). Incident, reflected, and transmitted fields are all connected by the field stored briefly within the barrier. Indeed, the group delay is just the lifetime of stored energy escaping through both ends of the barrier. It is a cavity lifetime.

To see the connection to cavity lifetime, first recall the standard definition of the Q of a cavity [58]:

\[ Q = \frac{\omega U}{P_d}, \tag{68} \]

which is the time average stored energy divided by the power dissipated per cycle \( P_d \). As a result of dissipation, the cavity mode has a finite lifetime, a \( 1/e \) lifetime of stored energy which is given by [58]

\[ \tau_c \equiv \frac{Q}{\omega} = \frac{U}{P_d}. \tag{69} \]

For a cavity with no internal losses the power dissipated is the power that escapes through the ends of the cavity. At steady state, this power lost equals the incident power, or \( P_d = P_i \). Thus the cavity lifetime can be written

\[ \tau_c = \frac{U}{P_i} = \tau_d, \]
which shows that the cavity lifetime and the dwell time are identical. Furthermore, for a symmetric barrier all three quantities, the group delay, the dwell time, and the cavity lifetime are one and the same object. The dwell time is an ensemble average and not the time spent by any single particle. The same is true of the group delay. As Collins et al. [35] have noted, a semiclassical interpretation requires that the group delay be seen as an ensemble average, not as the delay of a single particle. The self-interference delay, being a dwell time, is also an ensemble average.

For photonic band gap structures the group delay and dwell time are one and the same entity [22,65] and neither of them should be considered a candidate for a transit time in tunneling. As shown in Section 4, the group delay and dwell time for a photonic band gap structure obey the equality

\[ \tau_g = \frac{\phi_t}{\phi_t} = \frac{U}{P_{\text{in}}} = \tau_d. \]

We cannot overstress the importance of this relation. The group delay is thus seen as the length of time the incident photon flux \( P_{\text{in}} \) has to act in order to build up the accumulated photon density or stored energy \( U \) within the barrier. Under quasi-static conditions this time is also the cavity lifetime, the lifetime of stored energy escaping through both ends of the barrier. The rate of escape of this stored energy is just

\[ \frac{1}{\tau_g} = \frac{P_{\text{in}}}{U}. \]

Since \( P_{\text{in}} = P_t + P_r \), this total rate can be written

\[ \frac{1}{\tau_g} = \frac{P_t}{U} + \frac{P_r}{U} = \frac{1}{\tau_t} + \frac{1}{\tau_r}, \]

which is seen as the sum of the rate of escape through the transmission channel alone and the rate through the reflection channel alone. Numerical simulations have shown that the group delay is indeed the \( 1/e \) lifetime of stored energy leaking out of both ends [98].

It is easy to see why group delay in tunneling is not a traversal time for forward transit through a barrier. For a symmetric barrier the inverse of the group delay is equal to the sum of the inverse of the transmission flux delay and
the inverse of the reflection flux delay. The group delay is thus associated with fluxes in both forward and backward directions. In the absence of self-interference delay the group delay is identical to the dwell time, which does not distinguish between reflected and transmitted particles. The limiting group delay is just the inverse of the width of the stop band. This is what characterizes the lifetime or response time of the system.

10. Origin of the Hartman effect

The origin of the Hartman effect, the saturation of group delay with barrier length, has been a mystery for decades. The phenomenon exists in the tunneling of all kinds of waves, be they matter waves, electromagnetic waves, or sound waves. Experimentally it has been observed using electromagnetic waves and sound waves and there is no question as to its existence. If one interprets the group delay in tunneling as a transit time then the Hartman effect naturally leads to superluminal and unbounded group velocities. On the other hand, if the group delay in tunneling is not a transit time but a lifetime then the Hartman effect is very easy to explain. Such an explanation was first put forth by Winful in a 2002 paper that specifically addressed the tunneling of electromagnetic waves [22].

Simply put, the origin of the Hartman effect is the saturation of stored energy with barrier length. Since the group delay is proportional to stored energy, it saturates as the stored energy saturates. As discussed earlier, tunneling is a quasi-static phenomenon requiring wave packets that are much longer than the barrier. As a result, at any instant in time the field distribution in the barrier is approximately the steady-state distribution. For the photonic band gap structure the energy density for light tuned to the middle of the stop band is

$$\langle u \rangle = \frac{1}{2} \varepsilon |E_0|^2 \frac{cosh 2 \kappa (z - L)}{cosh^2 \kappa L}.$$  

(70)

This distribution is essentially an exponential decay with distance. In the absence of self-interference effects the group delay is just the time-averaged stored energy divided by the input power. This delay is the time it takes for the integrated stored energy to respond to a small change in input power, or equivalently, the time the incident flux has to act to produce the accumulated stored energy. Under quasi-static conditions it is the lifetime of stored energy leaking out at both ends [98]. Because of the exponential decay of the energy density with distance, beyond a 1/e distance it does not matter how long the barrier is: practically all the energy is stored within that 1/e distance. Since the barrier is much shorter than the pulse, it effectively acts as a lumped element with respect to the pulse, much like a capacitor. Fig. 39 shows the normalized stored energy and the normalized group delay versus barrier length for a photonic band gap structure. The Hartman effect, the saturation of group delay (or cavity lifetime) with barrier length is thus completely explained by the saturation of stored energy.

Fig. 39. Origin of the Hartman effect: normalized group delay and normalized stored energy showing saturation with barrier length.
For quantum particles the Hartman effect arises from the saturation of the integrated probability density or particle number under the barrier [23]. There are actually two terms in the group delay. One is the dwell time, the time it takes to empty the barrier of the stored density. The other is the self-interference delay, which is a dwell time in the region in front of the barrier. This self-interference delay is proportional to the imaginary part of the reflection coefficient, which can be written [23]

\[
\frac{\hbar k}{m} \text{Im}(R) = -\frac{1}{\hbar} \int_0^L \left[ \frac{\hbar^2}{2m} \left| \frac{d\psi}{dx} \right|^2 + (V - E) |\psi|^2 \right] dx.
\]

In the limit \( L \to \infty \), the probability density inside the barrier is simply the decaying exponential \( |\psi|^2 \sim \exp(-2\kappa x) \). It can be seen that the dwell time, the reflection coefficient, and the self-interference delay are proportional to the integrated probability density in this exponential limit. This integrated probability density saturates with barrier length and hence the reflection coefficient, dwell time, self-interference delay, and phase time all saturate. As \( L \to \infty \), we find

\[
\tau_d = \frac{2}{\kappa v} \left( \frac{E}{V_0} \right), \quad \tau_i = \frac{2}{\kappa v} \left( \frac{V_0 - E}{V_0} \right), \quad \tau_g = \tau_d + \tau_i = \frac{2}{\kappa v}.
\]

Chiao and Steinberg associate this time with the time taken to reach two \( 1/e \) penetration depths for a particle traveling with velocity \( v \). However, it is really the sum of a dwell time within the barrier plus a dwell time due to self-interference in front of the barrier. For the case of electromagnetic tunneling through a waveguide below cutoff, one encounters a similar self-interference term, which is proportional to the difference between stored electric and magnetic energies. Both of these energies saturate with barrier length and hence the Hartman effect in that case is also explainable on the basis of stored electric and magnetic energy [57].

The notion of stored energy along with the realization that the group delay is not a transit time thus makes it possible to resolve the paradox of the Hartman effect. This explanation of the Hartman effect has now been applied to the saturation of group delay for long-wavelength phonons in semiconductor heterostructures. Villegas et al. find that the lack of dependence of the tunneling time on system size can be explained by the saturation of the stored vibrational energy within the heterostructure [88,89]. These ideas have also been used to resolve another conundrum, the “generalized Hartman effect” which had been described for the case of multiple-barrier tunneling [100–102].

11. Reinterpretation of tunneling time experiments

With this interpretation of the group delay as a cavity lifetime, it is now possible to explain every aspect of tunneling experiments without appealing to superluminal group velocities. In the typical pulsed tunneling time experiment, a long pulse of electromagnetic energy is sent through a barrier-free region of length \( L \). The arrival time of the peak of this pulse at a detector is used as a reference time. Let the group delay in traversing the barrier-free region be \( \tau_0 \). This delay is an actual transit time since the transmitted pulse is the same entity as the incident pulse. A barrier of length \( L \) is next inserted in the path of the pulse. The arrival time of the peak of a transmitted pulse is then compared with the reference time. Let the group delay in the presence of the barrier be \( \tau_1 \). As we have explained before, the group delay measured here is not the transit time of the incident pulse since the pulse is actually reflected and only a tiny fraction of the incident energy makes it through the barrier. What is detected is the leakage of stored energy from the barrier. The various experiments done at optical frequencies only differ in the methods used to detect this leaked energy. They all involve comparison of the tunneled pulse with the reference pulse either through direct detection or through some correlation technique.

Wave propagation in any medium (including vacuum) proceeds through the storage and release of energy. Consider first the reference pulse. Fig. 40a shows a snap shot of the energy density in the pulse as a function of position at the instant when the peak of the pulse arrives at the input plane. At that instant the flux of energy crossing that plane is the incident power \( P_i \). The shaded area shows the stored energy in the region \( 0 < z < L \) for that value of incident flux. Some time later, that flux of energy leaves the region. For the reference pulse, all the entering energy leaves the region at the exit plane \( z = L \) since there is no reflection or absorption. The time it takes for all the energy to leave is given by the stored energy divided by the rate at which energy enters: \( \tau_0 = U_0/P_i \). The time average stored energy in the transparent region of volume \( V = LA \) is just \( U_0 = (1/2)\varepsilon_0 n_0^2 E_0^2 AL \). The net energy flux transmitted in the forward
Fig. 40. (a) Snapshot of a pulse traversing a region of free space. The shaded area shows the stored energy corresponding to the peak incident power and (b) snapshot of a pulse interacting with a barrier. The shaded area shows the stored energy corresponding to the peak incident power. In front of the barrier is shown only the incident power but not the reflected power. (From Ref. [98])

direction through this lossless, reflectionless region is equal to the input power \( P_i = (1/2) \varepsilon_0 n_0 c |E_0|^2 A \). Upon dividing \( U_0 \) by \( P_i \) we obtain \( n_0 L / c \equiv \tau_0 \), the time it takes for all the energy stored in the region of length \( L \) to leave that region in the direction of the net flux and with velocity \( v = c / n_0 \). Here, because all the energy that enters leaves later in the forward direction, one can infer a sensible velocity \( v = L / \tau_0 \).

Now consider a pulse incident on a barrier. It should be noted right away that unlike the case of the reference pulse the incident and transmitted pulses are not the same entity and so the classically motivated “transit time” does not apply here. The incident pulse creates a cavity field, which is made up of a sum of forward and backward propagating
components that have undergone various amounts of multiple scattering within the barrier. This cavity field then gives rise to a transmitted field and a reflected field. The transmitted pulse is not the delayed incident pulse. It is the released barrier field. One cannot associate a given temporal point within the transmitted or reflected pulse with a given temporal point in the incident pulse. Again we calculate the stored energy in the barrier at the moment the pulse peak arrives at the input. From Fig. 40b, it is clear that this stored energy represented by the shaded area is much smaller than in the case of free space propagation. The energy density decays almost exponentially with distance. It consists of a forward and backward component whose sum, just inside the barrier exceeds the incident energy density. Most of the stored energy leaves the barrier in the backward direction and a small amount is transmitted in the forward direction. The group delay, the time it takes for this stored energy to escape through both forward and backward channels is simultaneously 
\[ \tau_1 = \frac{U}{P_i} \].
Because the stored energy is much smaller than in the free space case, for the same input power the delay time for this stored energy is much less than in the free space case. Note that this delay is not the time it takes for the input peak to propagate to the exit since the pulse does not really propagate through the barrier. The incident pulse modulates the stored energy in the barrier. The stored energy responds to this modulation with a finite delay. What is really measured is the lifetime of stored energy escaping through both ends: the cavity lifetime. This explains why the group delay is shorter for stronger barriers (stronger barriers store less energy), why it saturates with barrier length (stored energy saturates with barrier length), and why it is less than the free space delay (barrier stores less energy than free space for the same input power).

For the photonic band gap structure, the group delay and dwell time are one and the same quantity [22]. At the center of the stop band this delay is given by
\[ \tau_g = \tau_d = \tau_0 \frac{\tanh \frac{\kappa L}{\kappa L}}{L} \],
where \( \tau_0 = \frac{L}{\nu} \) is the barrier-free delay. Since the delay is given by the ratio of stored energy to input power, the ratio of the barrier delay to free propagation delay is just the ratio of stored energies:
\[ \frac{\tau_g}{\tau_0} = \frac{U}{U_0} = \frac{\tanh \frac{\kappa L}{\kappa L}}{\kappa L} \].

The parameters of the Longhi et al. [61] experiment yield \( \kappa L = 2.8 \) from which we find that the barrier only stores 35% of the energy that would be stored in an equivalent volume without a barrier. This is all the experiment is saying. It does not imply that anything is being transported at faster than the speed of light. It takes 34 ps to empty the energy stored in the barrier through both ends (with most of it leaving in the backward direction) compared to the 97 ps it takes to empty the energy stored in the equivalent region of free space, with all of it exiting in the forward direction. This calculation yields a delay time difference of 63 ps, which is exactly what is observed in the experiment.

This explanation applies to all the reported superluminal tunneling experiments including the “single-photon” measurements of Steinberg et al. [14]. As noted by the authors themselves [4], for purely linear phenomena such as tunneling, single photons exhibit the same behavior as classical pulses. “Propagation effects are then governed by the classical wave equations, and quantization merely affects detection statistics and higher order effects.” Steinberg et al. describe these photons as 20-fs wave packets, a duration that corresponds to their correlation time. They also go on to note that the existence of “wave packets” for photons is controversial since a position operator does not exist for photons. It would be desirable to have an explanation for the SKC results that does not appeal to photon “wave packets”.

To explain the SKC results without invoking wave packets, we consider the signal and idler photons as simply modes of the electromagnetic field that are generally independent and hence do not exhibit second-order interference. They can nevertheless exhibit fourth-order interference, which is monitored through the Hong–Ou–Mandel interferometer. The detection probabilities depend on the relative phase between signal and idler photons. However, the phase accumulated by a mode in any region of space is proportional to the energy stored in that region. This is easily seen from the relation between group delay and stored energy
\[ \frac{\partial \phi_0}{\partial \Omega} = \frac{U}{P_m} \].
The phase accumulated is thus given by

\[ \phi_0 = \frac{U}{P_{in}} \Omega, \]

where we have set an arbitrary constant phase to zero. The phase is linear in frequency, which means that in the time domain there is a pure delay without distortion. This is why both cw and pulsed measurements yield the same result for the group delay. More importantly, for a given input power the phase is proportional to the stored energy. In the stop band of a photonic band gap structure the stored energy is reduced below its free space value (as a result of destructive interference) and this leads to a reduction in accumulated phase below what would have been gained through free propagation. In order to make up for the loss of phase, the path length external to the barrier has to be increased, thereby adding some propagation phase. The phase lag in the barrier is due to the finite response time of the structure acting as a lumped element. The phase lag in free space results from the need to transport stored energy out of the region before fresh energy can enter. It should be noted that the experiment simply measured the shift in a mirror position and not a velocity. This shift corresponds to the difference in stored energies and does not imply that anything was transported with superluminal velocity.

In sum, tunneling experiments measure the lifetime of stored energy escaping simultaneously through both ends of a barrier. For the same incident power, the less energy stored, the shorter the group delay or lifetime. If the incident energy is mostly reflected, the stored energy will be substantially less than what would have been stored in the equivalent length of free space. As a result, the group delay in the presence of the barrier will be less than the delay in its absence. This is why the transmitted pulse peaks sooner. The tunneling experiments should thus be seen as an elegant way to measure the non-resonant cavity lifetime. Since the peak does not propagate through the barrier and since the transmitted pulse is not the same entity as the incident pulse, it does not make sense to say that the incident pulse propagated with superluminal group velocity through the barrier. A decent barrier transmits much less than 1% of the incident energy. On that scale it is as if we sent in an elephant and out came an ant. We would not say the elephant has traveled to the exit.

12. Double-barrier tunneling

The phenomenon of tunneling through two or more potential barriers (Fig. 41) is the basis for the operation of important technological devices such as the resonant tunneling diode [101,102]. In a manner analogous to the transmission of light through a Fabry–Perot resonator [103], the transmission of electrons through a sequence of barriers exhibits maxima when the spacing between the barriers is roughly an integer multiple of one-half the deBroglie wavelength of the electron. For electron energies close to these resonant values, electronic tunneling devices display negative conductivity and high-frequency oscillations [101]. Because the ultimate response time of such devices may be limited by the tunneling time, there have been many numerical and analytical studies of the dynamics of resonant tunneling [104–112]. At resonance the tunneling time has afforded few surprises. However, for off-resonance tunneling, an interesting conundrum has recently emerged which has been named the “generalized Hartman effect” [113,114].
This relates to an apparent lack of dependence of the group delay on the separation between barriers. Here, we show that the interpretation of group delay in tunneling as a lifetime immediately resolves all paradoxes in non-resonant tunneling and is consistent with what is known for resonant tunneling.

12.1. Resonant tunneling

As discussed in Section 2, the bi-directional group delay in scattering by a localized, lossless potential is given by

\[ \tilde{\tau}_g = \tau_d + \tau_i, \]

where \( \tau_d \) is the dwell time and \( \tau_i = -\text{Im}(R)/k \) is the self-interference delay. This result holds for any number of barriers. At transmission resonances the reflection coefficient is zero and hence the self-interference term disappears. In that case the group delay and dwell time are identical and either quantity can be used to give a unique value for the delay time in tunneling. Liu [106] and Collins et al. [105] calculated the group delay in resonant tunneling by evaluating the energy derivative of the transmission phase shift. An exact calculation of the transmission of a symmetrical double-barrier structure yields [106]

\[
T^{-1} = \frac{V^2}{4E(V - E)} (\sinh \kappa L_0)^2 e^{i k L} + \left( \cosh \kappa L_0 + \frac{i(V - 2E)}{2\sqrt{E(V - E)}} \sinh \kappa L_0 \right)^2 e^{-i k L},
\]  

(71)

where \( \kappa = \sqrt{2m(V - E)/\hbar^2} \) and \( k = \sqrt{2mE/\hbar^2} \). The transmission function and the energy derivative of its phase are shown in Fig. 42. Around the resonance the transmission function can be approximated as a Lorentzian (also referred to as a Breit–Wigner form)

\[
T = \frac{\Gamma}{(E - E_0) + i\Gamma},
\]

(72)
where $\Gamma$ is the half-width at half-maximum of the resonance line, $E$ is the energy of the incident electron and $E_0$ is the energy of the first metastable resonant state. The energy derivative of the transmission phase shift yields

$$\tau_g = \frac{\hbar \Gamma}{(E - E_0)^2 + \Gamma^2}. \quad (73)$$

This approximate result agrees very well with the exact result and also with numerical simulations. At exact resonance the delay time is

$$\tau_g = \frac{\hbar}{\Gamma}, \quad (74)$$

which is just the lifetime of the quasi-bound state. Thus, it is seen that at resonance the group delay is interpreted as a lifetime of a metastable state. From the approximate results it is seen that the lifetime increases linearly with barrier separation and exponentially with barrier thickness. The excitation of the resonance requires wave packets that are narrow in $k$-space and hence very broad in spatial extent. Note that the delay off resonance is a couple of orders of magnitude shorter than the on-resonance group delay.

12.2. Off-resonant tunneling: “superluminality” and the “generalized Hartman effect

Olkhovsky et al. [113] first suggested that in non-resonant tunneling through two successive barriers separated by an intermediate free region $R$, “the total traversal time does not depend not only on the barrier widths (the so-called “Hartman effect”), but also on the $R$ width: so that the effective velocity in the region $R$, between the two barriers, can be regarded as practically infinite.” This puzzling phenomenon has been termed “the generalized Hartman effect.” The structure has been referred to as a “destroyer of space” since the particle appears to make the voyage from input to output as if the intervening space did not exist. This effect has been extended to multiple barriers [115] and arrays of delta function potentials [116,117] and there are also experimental results that apparently confirm its existence [118]. Numerical solutions of Maxwell’s equations [97] apparently suggest instantaneous traversal of the inter-barrier region in agreement with early electromagnetic experiments of Nimtz [119]. Related work on semiconductor superlattices also predicts infinite traversal velocities at band edges [120].

While the Hartman effect for a single barrier is readily understood through the saturation of stored energy, the “generalized Hartman effect” is more troublesome since the waves in the interbarrier region are oscillatory and do not decay. The stored energy between the barriers should not saturate with barrier separation. It turns out, however, that this latter effect, the lack of dependence of group delay on separation, is an artifact. The group delay does increase linearly with barrier separation.

The double-barrier structure is analogous to a Fabry–Perot resonator with mirrors of reflectivity $R$ separated by a distance $L$. For an incident field of amplitude $E_0$, frequency $\omega$, and wavenumber $k = n\omega/c$ the total phase of the transmitted wave is [100]

$$\phi(k) = kL + \tan^{-1}[R \sin 2kL/(1 - R \cos 2kL)],$$

which yields the group delay [100,121]

$$\tau_g = \frac{d\phi}{d\omega} = \frac{1 - R^2}{1 + R^2 - 2R \cos(2kL)} \frac{L}{v}, \quad (75)$$

with $v = n/c$. On the other hand, the time average stored energy in the cavity is [100]

$$\langle U \rangle = \frac{\varepsilon_0 ALT (1 + R)|E_0|^2}{2(1 + R^2 - 2R \cos 2kL)}. \quad (76)$$

Upon dividing by the incident power we obtain the dwell time

$$\tau_d = \frac{1 - R^2}{1 + R^2 - 2R \cos(2kL)} \frac{L}{v}. \quad (77)$$
Thus the group delay and dwell time are identical for a Fabry–Perot resonator. It is interesting to note that Yu et al. [122] previously obtained a similar identity by using a statistical argument for the dwell time, which averages the time spent by all the transmitted photons. Since the dwell time is just the cavity lifetime we see that group delay, dwell time, and cavity lifetime are one and the same object for the Fabry–Perot cavity.

The group delay is proportional to the time average stored energy. The stored energy depends on the round trip phase shift $\theta = 2kL$ seen by the wave as it bounces back and forth within the cavity. Under resonant conditions, $\theta = 2m\pi \ (m = 1, 2, \ldots)$, the recirculating fields add up in phase thereby enhancing the stored energy and increasing the storage time. The group delay or cavity lifetime at resonance is

$$\tau_{on} = \left( \frac{1 + R}{T} \right) \frac{L}{v}, \quad (78)$$

which can be made arbitrarily large compared to the transit time $L/v$ as the mirror transmission $T \to 0$. On the other hand, when $\theta = (2m + 1)\pi$, anti-resonant conditions obtain. Because of destructive interference between the recirculating phasors the stored energy is reduced below the value it would have had in the absence of the mirrors. Under these conditions the group delay is

$$\tau_{off} = \left( \frac{T}{1 + R} \right) \frac{L}{v}. \quad (79)$$

This delay is always shorter than the cavity transit time and can be made arbitrarily small as $T \to 0$. Indeed, when $T = 0$ the group delay is zero. This of course does not mean that a pulse was transmitted through the Fabry–Perot in zero time. It simply means that no energy was stored between the mirrors, all of it was reflected and no power was transmitted. We also see that the group delay is not independent of barrier separation but increases at a linear rate with that separation, in proportion to the stored energy. The rate of increase is proportional to the transmittivity $T$ of the first mirror. It is definitely not a transit time but the cavity lifetime under non-resonant conditions.

- The lack of dependence of group delay on barrier separation claimed by Olkhovsky et al. is actually not true for any finite level of transmission. It is only when the transmission is exactly zero that the group delay becomes independent of separation. This of course is not of any interest since the transmission is zero and one could not measure any transmitted light with which to assign a group delay. At anti-resonance, for any finite level of transmission, the group delay will increase linearly with barrier separation, in proportion to the stored energy.
We should note that the group delay for the double barrier does not display any limiting behavior with respect to barrier separation. The problem with the original calculation stems from the procedure of taking the limit of infinite barrier length (and hence zero transmission) before exploring the dependence on separation.

A more serious question that must be answered is this: why is the predicted group delay so short? The interpretation of the group delay as a lifetime makes it possible to show that there is no superluminal propagation involved in double-barrier tunneling. The calculated delay of the output peak is just the lifetime of energy stored in the interbarrier region. That delay is proportional to the stored energy, which is also proportional to the cavity length. For the same input power, the less energy stored, the less time it takes to release, with most of the energy released in the backward (reflection) direction.

12.3. A re-interpretation of “superluminal” double-barrier tunneling experiments

There is an optical experiment on double-barrier tunneling that has been taken as confirmation of the generalized Hartman effect and as a demonstration of group velocities as large as 5c [118]. We disagree with that interpretation and see it rather as a beautiful demonstration of the dependence of cavity lifetime on frequency and on cavity length. The experiment simulates quantum-mechanical double-barrier tunneling by using two fiber-Bragg gratings of length $L_0$ separated by a uniform fiber region of length $L$. The transmission of such a structure exhibits Fabry–Perot resonances within a broad stop band (Fig. 44). The group delay also exhibits peaks and valleys that closely follow the Fabry–Perot transmission maxima and minima. The experimental transmission and group delay curves were obtained in the frequency domain using cw lasers and a phase-shift technique. For time-domain measurements of delay, pulses of 1.3 ns duration and carrier frequency tuned to an anti-resonance were sent to double-barrier FBG’s with different separations. The spatial extent of the pulse (39 cm) is much greater than the barrier separations (ranging from 18 to 47 mm) and is also greater than the length of the entire double-barrier structure. The pulsed experiments are thus in the quasi-stationary regime described earlier.

Fig. 45a shows the amplified intensity of the transmitted part ($\sim 0.8\%$) of a 1.3 ns pulse that encounters a double barrier compared to the same pulse transmitted through a barrier-free-region of the same length. The barrier reflects 99.2% of the incident energy. The transmitted portion reaches a maximum 248 ps before the reference pulse attains a maximum at the exit. Fig. 45b shows the measured group delay versus barrier separation for five different double barriers. The dashed line is the delay that would be measured for a wave packet traversing a barrier-free region of the same total length:

$$\tau_1 = \frac{L}{v} + \frac{2L_0}{v}.$$ (80)
The solid line is a fit to the double-barrier group delay expression given by

$$
\tau_g = \frac{\sqrt{S L}}{v} + \frac{2L_0 \tanh 2\kappa L_0}{v} \frac{2\kappa L_0}{L_0^2},
$$

where $\kappa$ is the grating coupling constant and

$$S = \frac{1}{\cosh^2(2\kappa L_0)} = \left[ \frac{T}{1+R} \right]^2
$$

is the overall power transmission coefficient of the two cascaded gratings under anti-resonant conditions and at the Bragg frequency. In the last equality, $T = 1/\cosh^2 \kappa L_0$ is the transmissivity and $R = \tanh^2 \kappa L_0$ is the reflectivity of a single grating of length $L_0$.

Note that, contrary to what is predicted as the “generalized Hartman effect” the delay is not independent of barrier separation. It increases at a slow rate with separation. At any separation the group delay for the tunneling pulse is much shorter than that of the free space pulse. However, this does not mean that the tunneling pulse travels faster. At any instant in time, the transmitted power is due to the barrier releasing energy stored from a time one group delay time before. A short delay simply means that there was very little energy stored from the very long pulse. The ratio of the stored energy in the region between the barriers to that in the same region without barriers should yield the ratio of group delays. From Eq. (75) it is seen that the ratio of the stored energies at anti-resonance is given by $T/(1+R) = \sqrt{S}$. Thus, a measurement of the ratio of the slopes of the delay versus $L$ plots in Fig. 45b yields the ratio of stored energies as well as the overall transmissivity. The inter-barrier delay increases from a value of about $32 \text{ ps}$ to $50 \text{ ps}$ as the barrier separation is increased from $10$ to $50 \text{ mm}$. This yields a slope of $4.5 \times 10^{-10} \text{ s/m}$. The slope of the dashed line is about $4.75 \times 10^{-9} \text{ s/m}$. From their ratio we find a power transmission of about $S = 0.9\%$. This is not far from the value of $0.8\%$ given in Ref. [118] as the minimum transmission of the cascaded gratings. Since more than $99\%$ of the incident energy is reflected, the measured delay cannot be associated with a forward traversal time.

**Resonant and non-resonant tunneling with acoustic waves:** van der Biest et al. [123] have studied resonant and non-resonant tunneling of acoustic waves through a double-barrier made of *phononic* crystals, periodic structures that
forbid the propagation of acoustic waves within a certain band of frequencies. Three-dimensional resonant tunneling structures were created by inserting a 7.05-mm-thick aluminum slab between two identical phononic crystals composed of an fcc arrangement of tungsten carbide beads in water. Two-dimensional structures, made of an array of steel rods in water, were also investigated. These structures were excited with acoustic pulses with center frequencies at 500 kHz and at 1 MHz and the transmitted signals monitored as a function of frequency. The transmitted amplitude displays the typical resonant behavior as shown in figure. Their measurements of group delay (Fig. 46) show very short group delays off resonance and very long delays on resonance. They interpret the short delays as due to an ultrafast or “supersonic” transit velocity as large as $9v_{\text{water}}$, where $v_{\text{water}}$ is the velocity of sound in the water matrix within which the phononic crystals are embedded. Again it should be noted that these experiments measured delays and not velocities. The correct interpretation of these delays is the lifetime of stored energy leaking out of both ends. Indeed, the authors recognize the delay on resonance as an indication that “a long lifetime can be associated with each resonance”. The off-resonance delay is also a lifetime, only shortened by destructive interference. The measured delays are proportional to the transmission, which in turn is proportional to the stored energy.

To summarize, the group delay in double-barrier tunneling is a linear function of barrier separation. This means that the “generalized Hartman effect” does not exist. The reduced group delay in off-resonant transmission is not due to a superluminal group velocity but is a result of a reduced cavity lifetime when operating under non-resonant conditions. Thus, the interpretation of group delay as a lifetime makes it possible to explain the anomalously short times seen in double-barrier tunneling.

13. Conclusions

We have examined the meaning of the group delay in barrier tunneling and reviewed the optical, electromagnetic, and acoustic experiments that measure this important quantity. Our conclusion is that the group delay (phase time) in tunneling is a lifetime and not a transit time. Group velocity is not a meaningful concept in this context and we find no evidence for superluminality in barrier tunneling. The interpretation of the group delay in tunneling as a lifetime rather than a transit time makes it possible to resolve all the known paradoxes in the physics of tunneling time. In particular it allows for an explanation of the Hartman effect and the rather short delays seen in single and double-barrier tunneling. Since the group delay is only a small fraction of the wave-packet length, the duration of the tunneling process has to be the length of the wave packet. The conclusions presented here go against the grain of current thinking and may either reignite the tunneling time controversy or help settle the issue once and for all.
14. Discussion

Because the subject of this review is an old and contentious one, it is expected that there will be many opposing viewpoints. Part of my motivation in writing the review is to reignite debate on the subject. An anonymous referee who reviewed this article raised questions which I felt were best addressed outside the main body of the text. Since those questions are likely to be among those raised by other physicists, I have collected those questions and my responses in this Discussion section. The referee’s comments are in italics and labeled AR (for Anonymous Referee). My responses are labeled HW.

14.1. Treatment of the reshaping argument

AR: On page 40, “A consequence of the reshaping argument is that the transmitted pulse is narrowed.” I don’t believe this to be a necessary consequence of the reshaping interpretation. It is a consequence of certain dispersion relations, and is true or false irrespective of whether one explains the phenomenon as reshaping. Near a transmission minimum, the bandwidth of a pulse is obviously increased by filtering. If the dispersion is sufficiently flat, this will narrow the pulse width. This is true whether or not one uses the reshaping argument, and fails if the dispersion is not flat enough or if the transmission probability does not have a simple (e.g., quadratic) form. The author in general seems to depict the reshaping argument as a straw man. The word “reshaping” is indeed unfortunate, as the authors who support it seem to say simultaneously that the packet is “reshaped” and that its “shape does not change.” Poor choice of words indeed. Yet one understands what they mean. The “peak is made up only of the leading parts of the incident pulse,” as Winful writes, but molded into the shape of the original peak of the incident pulse. It does not “necessarily distort an incident pulse.” As discussed by Japha & Kurizki and by Chiao & Steinberg, one form of reshaping can be through a Taylor expansion. A transmitted field with a memory effect (which is of course what a dispersive barrier leads to) may be proportional not to the incident field but rather to the incident field plus some constant times its time derivative, or in general a term with more such expressions. Since the time derivative generates translations in time, each point on this curve will be “reshaped” and yet the resulting output pulse will be nothing more or less than a “translated” copy of the original curve. I think that it is perfectly useful for a new paper to argue against perceived wisdom on the subject, and to present a new approach, but I think this paper treats the “reshaping” argument less than fairly, and would be stronger if this were rectified.

HW: The referee disputes my statement “A consequence of the reshaping argument is that the transmitted pulse is narrowed.” The narrowing of a pulse as a consequence of the “reshaping” mechanism is not something I invented. The paper by Japha and Kurizki which he and I cite as one important source of the reshaping argument makes several connections between reshaping and pulse narrowing [94]:

- “We identify the universal mechanism that is responsible for superluminal (faster-than-light) traversal times as well as the narrowing of wave packets transmitted through various non-dissipative media.”
- “The latter experiment has also revealed a remarkable feature, namely, that the temporal width of the transmitted wave packet is strongly narrowed down.”
- “Indeed, such reshaping explains pulse narrowing and superluminal pulse traversal in absorbing [10] (or amplifying [11]) media…”
- “Is there a common mechanism for superluminal time delays and wave-packet narrowing, which applies to both EM pulses in dielectric structures and relativistic massive particles in potential barriers [4]?”
- “If ω lies in a dip of the spectral transmission curve I_{tr}(ω), then \( \partial^2[\ln I_{tr}] / \partial \omega^2 > 0 \) and the pulse will be narrowed.”
- “The temporal narrowing effect will be most salient when \( \Delta t \sim (\partial^2 / \partial \omega^2)[\ln(I_{tr}(ω))] \), provided \( \Delta t \) is large enough to allow overlap of successive wave packets.”
- “This effect is seen to be sensitive to coherence: the phase incoherence… exponentially diminishing the narrowing in (8).”
- “Destruction of the back half of \( ψ_{in} \) by interference also makes the transmitted pulse narrower, because it consists mostly of the forward tail of \( ψ_{in} \).”
- “Reduced intensity, superluminal time delay, and temporal narrowing of pulses transmitted through such media follow, as in layered structures from destructive interference…”
• “Our theory has demonstrated that the universal mechanism of predominantly destructive interference between accessible causal paths \[12\] is responsible for transmission attenuation, superluminal delay times, and wave-packet narrowing.”

I suggest that the reader will be forgiven for seeing a connection between the reshaping argument and wave-packet narrowing.

**AR:** Near a transmission minimum, the bandwidth of a pulse is obviously increased by filtering. If the dispersion is sufficiently flat, this will narrow the pulse width.

**HW:** It is not at all obvious to me that the bandwidth (frequency spread) of a pulse can be increased by filtering, a process that removes frequencies. Tunneling occurs in the stop band where the transmission is flat. A sufficiently narrowband pulse will have a spectral width much narrower than the stop band. No filtering of the pulse will occur and the bandwidth will not increase. Indeed, in the limit that the pulse spectral width goes to zero, the group delay remains finite and there is obviously no filtering or reshaping, yet the possibility of anomalously short (“superluminal”) delays remains.

**AR:** The author in general seems to depict the reshaping argument as a straw man.

**HW:** I am not sure why the referee thinks I depict the reshaping argument as a straw man. When I looked up the definition of straw man I found “An argument or opponent set up so as to be easily refuted or defeated.” In the review I merely report what the reshaping argument says, as extracted from published texts. It is hardly my fault that the argument as published is easily refuted or defeated. It also does not help that there is no experimental evidence in its favor. It is my responsibility to lay this all out so that the reader can make up his own mind.

**AR:** It does not necessarily distort an incident pulse.”

**HW:** If a barrier or shutter transmits only the early parts of a pulse and rejects the later parts (as stated in many published descriptions of the reshaping argument) how can it not distort the pulse?

**AR:** As discussed by Japha & Kurizki and by Chiao & Steinberg, one form of reshaping can be through a Taylor expansion. A transmitted field with a memory effect (which is of course what a dispersive barrier leads to) may be proportional not to the incident field but rather to the incident field plus some constant times its time derivative, or in general a term with more such expressions. Since the time derivative generates translations in time, each point on this curve will be “reshaped” and yet the resulting output pulse will be nothing more or less than a “translated” copy of the original curve.

**HW:** The use of a Taylor series expansion to justify the reshaping argument explanation of superluminality is not all convincing. One could also very well take a pulse propagating a distance \( L \) in free space with input amplitude \( A(0, t) \) at \( z = 0 \). At \( z = L \), the amplitude is a delayed version of the input and can be written as \( A(L, t) = A(0, t - \tau) \). This delayed pulse is identical in shape to the input pulse. We would certainly not ascribe free space delay to a “reshaping phenomenon”. Yet we can certainly expand the delayed pulse in a Taylor series:

\[
A(0, t - \tau) = A(0, t) - \tau A'(0, t) + 1/2\tau^2 A''(0, t) + \cdots \tag{D1}
\]

Sure, the first derivative term induces a time translation. If we neglect the higher-order terms, of course the pulse will be reshaped somewhat, the amount of reshaping dependent on the size of the neglected terms. If we keep all the terms in the expansion the delayed pulse is identical to the input pulse. Thus reshaping cannot be seen as a mechanism for pure delay. The Taylor series “explanation” of the reshaping argument for superluminality could just as well be applied to free space propagation for which we find no superluminality. It just doesn’t hold water. I can always walk into a classroom and state that “a pulse propagates through free space by performing a Taylor expansion.” While mathematically true, I doubt that my students would just simply buy that as the physical mechanism for pulse propagation.

In fact, an examination of the Taylor series expansion of the transmitted pulse shows that the reshaping argument which says “the barrier transmits the early parts of the pulse and rejects the later parts” is quite wrong. By that argument the barrier transmission is higher for the rising part of the incident pulse than for the trailing part. But in reality, what happens is exactly the opposite. We can all agree, as the referee states, that “A transmitted field with a memory effect (which is of course what a dispersive barrier leads to) may be proportional not to the incident field but rather to the incident field plus some constant times its time derivative . . . “. This is what the first two terms of the Taylor expansion give us

\[
A(L, t) = T_0 A(0, t - \tau) \approx T_0 [A(0, t) - \tau A'(0, t)], \tag{D2}
\]
where $T_0$ is the steady state (constant, time-independent) transmission function. Now suppose $A(0,t)$ is a smooth bell-shaped pulse like a Gaussian. On the rising edge of the pulse $A'(0,t) > 0$. Since $\tau$ is a positive number, we see that on the rising edge, a positive quantity is subtracted from the incident field to yield the transmitted field. Therefore, on the rising edge, the transmission is actually less than the steady-state transmission and not more, as the reshaping argument would have us believe. On the other hand, for the trailing edge of the pulse $A'(0,t) < 0$. A positive quantity is added to the incident field. Thus, on the trailing edge the transmission is higher than the steady-state value. This is the manifestation of any memory effect, a time lag. The output follows the input with a time lag. While the input is rising, the output is still holding on to the earlier higher values dictated by the earlier input. If the delay is small and there is little distortion the transmitted pulse can be found simply by multiplying the delayed input pulse by the steady-state transmission function. In short, the Taylor series justification of the reshaping argument predicts the wrong behavior for the transmitted pulse.

14.2. Transient response in tunneling

**AR:** On page 40, I am confused by this discussion of the short transient, especially given the experimental results which observe only a single undistorted peak, as Winful stresses. As for transmission being in steady state, I believe it is again either Japha & Kurizki or Steinberg & Chiao who made the point about Fabry Perot or dielectric stack style interference that each part of the pulse interferes with reflected portions of earlier parts. During the rising edge of a pulse, this interference is less effective. The question of whether or not one is in steady state depends on the timescale over which the input field is growing, relative to the “storage time” in the structure. Therefore, for a Gaussian pulse, there will be some particular time at which one crosses over into steady state, and this will depend on the pulse width. I believe this is also the problem with the strange argument on page 42 that even with a daylong pulse, only the first picosecond should be transmitted. The arguments presented elsewhere have been more rigorous than that, and one should look at the local slope of the incident wave.

**Response**

**HW:** Regarding the short transient, if we assume that there is an instant in time $t_0$ at which the pulse is “turned on”, then for $t < t_0$, $E = 0$ and for $t \geq t_0$, $E > 0$. Let us call the step in the field at that instant the “front”. Note that the “front” is essentially the first measurable field that emanates from the source. It is in the distant wings of any smoothly varying pulse like a Gaussian. In practice our ability to detect a “front” depends on the sensitivity of our measurement apparatus. In a digital computer simulation of pulse tunneling the pulse is truncated at a point that is a few pulse widths away from the pulse peak. For example, in the simulations on page 22 the $1/e^2$ intensity half-width of the incident Gaussian pulse is 4 units and the pulse is truncated at 15 units from the peak where the intensity (relative to the peak) is $7.81 \times 10^{-7}$. While negligible for all practical purposes, it still provides a marker with which to check such things as propagation speed and causality. No matter how small, this first non-zero signal is a step in the field amplitude. Let us therefore follow a unit step function as it propagates through a barrier. Fig. D1 shows the evolution of a step function through the barrier. The front travels at $c$ and does not feel the presence of the barrier. Behind this front, the small dielectric reflections begin to build up. By the time the front reaches the exit, an exponentially decaying quasi-standing wave has built up behind the front. If we examine the transmitted field we see that the first measurable field appears at $t = 1$ (in units of the transit time) as expected for a causal system. This is the front, which has a transmission of 100%. Within a time of about 0.2 units after the arrival of the front, the transmission first drops to zero since the reflectivity has built up by now and most of the incoming energy has been dumped out of the structure. This is the time scale that characterizes the group delay or storage time. Because of the sharp turn on there is some overshoot and ringing, but by $t = 2$, the transmission has settled down to near its steady-state value of $1.3 \times 10^{-3}$. This is the time scale that we refer to as the “transient”. It is no more than a few (3–4) transit times, even for the sharpest possible turn on.

For a pulse such as the Gaussian in Fig. D2 with $\tau_p = 6$, the peak of the pulse is at $t = 24$ while the front is at $t = 0$ where its value is exp($-16$) = $1.125 \times 10^{-7}$, imperceptible on the scale of the figure. The transient behavior described for the step function above occurs in the time frame $0 < t \ll 10$, long before the bulk of the pulse begins to arrive (roughly around $t = 10$). Fig. D2 shows the incident pulse (dotted), the tunneled pulse (solid) normalized by its peak value of $1.3 \times 10^{-3}$, and a reference pulse (dashed). The delay of the reference pulse peak is 1 unit while that of the tunneled pulse is 0.25. In order to observe the transient behavior it is necessary to magnify the scale of the figure by about a thousand fold. Fig. D3 shows the region $0 < t < 5$ of Fig. D2 magnified by $10^4$. On this scale we see the
arrangement of the transmitted front at $t = 1$ with an intensity of about $10^{-7}$, same as in the incident front. Following this is the sharp drop and the transient oscillations seen with the step input. All this transient behavior is occurring in the distant, barely measurable wings of the pulse. By the time the main rising part of the pulse arrives at the input the system is in quasi-steady state and can be characterized by its steady-state transmission of $1.3 \times 10^{-3}$ not the initial transient transmission of 100%. It is true that the barrier transmits the “front” of the pulse and rejects (mostly) the rest, but the front has absolutely nothing to do with tunneling. The front does not tunnel. It propagates through essentially
free space. Tunneling occurs way behind the front, after the steady-state reflectivity has built up. Indeed tunneling is what goes on after about \( t = 3 \). The short transient is barely measurable which is why experiments observe just a single peak.

By the same token, it makes no sense to say that the peak of the transmitted pulse is made up of the leading parts of the pulse. The leading part (the front) has left the barrier long before the main part of the pulse has even arrived at the input. Furthermore, there is no way, even in principle, to associate a part of the transmitted pulse with any particular part of the incident pulse. A one-to-one correspondence can only be made between the transmitted front and the incident front.

Indeed the storage time (which I have shown is nothing but the group delay) is much shorter than the pulse width in the quasistatic limit (which is the only limit for which it makes sense to talk about true tunneling). Recall that the limiting group delay is \( \tau_g = 1/\Omega_c \). On the other hand, the true tunneling without distortion requires that the pulse width satisfy \( \tau_p \gg 1/\Omega_c \). Clearly, the incident pulse is much longer than the storage time so the system is essentially in steady state.

The “strange” argument about a daylong pulse is actually due to Rolf Landauer. I agree with him. The arguments presented about reshaping have by-and-large not been as nuanced as the referee would suggest. The general picture that is bandied about in published articles is that the barrier transmits the early parts of the pulse and rejects the later parts. If you apply that to the step function shown (our day-long pulse), then indeed only the leading edge is transmitted, a transmission which, as argued above has nothing to do with tunneling. No one has mentioned a need to examine the local slope of the pulse. Even if we examine the local slope, the reshaping argument is still wrong since it predicts the exact opposite of what really happens in tunneling (see prior discussion on the Taylor series and the slope of the incident pulse.)

14.3. Tunneling of quantum particles

AR: On page 42 and beyond, the author tries to argue that nothing is traveling. In some sense, this is similar to the point many authors in the field try to make in order to rescue causality. But he subtly moves from quantum mechanics to optics at this point. One agrees that if something has made the transit, “it is certainly not the incident pulse A.” But in quantum mechanics, a single particle will be detected in one place or another. On some occasions, is it not the incident electron E or the incident photon P which was previously on one side and later on the other side? Certainly, speaking only of waves one can avoid this issue, but that is perhaps why it is in the context of quantum mechanics rather than any other wave theory that the issue has been so controversial.
Why the group delay is not a transit time

I make the point that the transmitted pulse is not the same entity as the incident pulse. The referee agrees with me on this point, at least as far as electromagnetic wave packets are concerned. However, when it comes to quantum particles, he appears to suggest that one can associate a transmitted particle with an incident particle: “But in quantum mechanics, a single particle will be detected in one place or another. On some occasions, is it not the incident electron E or the incident photon P which was previously on one side and later on the other side?”

I totally disagree with the referee on this one. Quantum mechanics has nothing to say about the history of a single, identifiable particle. In other words, there is no way, even in principle to mark an electron on the input side of the barrier at \(t = 0\) and then detect that same electron arriving at the exit of the barrier at some later time. The referee’s idea of labeling an electron E or photon P reminds me of this passage from Peter Pesic’s very nice book Seeing Double [124]: “As Schrödinger emphasized, not only are all electrons exactly alike in their observable characteristics, there is no way one electron could be marked (by being painted red, say) so that it could be distinguished from the others. One must let go of the preconception that one could pick out a certain electron (say, Ben) either by noticing some characteristic feature or by marking it, so that one could follow its subsequent career.” Later on he shows how this “identicality” of electrons makes the assignment of a unique trajectory impossible: “If that is so, then an electron does not have a trajectory or history in the normal sense, for that also requires being able to label each one and follow it through time.”

We must conclude that in barrier tunneling, the transmitted particle is not necessarily the same as the incident particle. For that reason, the classically motivated notion of transit time fails completely. The above argument also means that idea of the “single-photon tunneling time” is flawed. There is no way to measure the “transit time” of a single tunneling photon.

Let me try to make this even more clear. Group delay, the quantity at issue here, has no meaning for a single particle. To see this, consider a single particle represented by the wave packet A to the left of the barrier in Fig. 4. This wave packet simply tells us the probability of finding the particle (say, Ben) at some point to the left of the barrier. Suppose we make a position measurement on the particle at \(t = 0\). That measurement collapses the wave function so that the particle is now localized at some point \(x_0\). After interacting with the barrier, the particle is now represented by two wave packets B and C. Because the initial measurement collapsed the wave function, the two wave packets B and C are completely decoupled from A. A post-tunneling measurement by a detector at the exit may collapse these wave packets to some position \(x_1\) to the right of the barrier. The particle’s position \(x_1\) has no connection to the particle’s position \(x_0\). Only the ensemble average location of the wave packets B and C is related to that of wave packet A. Thus the group delay is not the delay experienced by any particular particle. It is certainly not the transit time of Ben from \(x_0\) to \(x_1\).

14.4. Why the group delay is not a transit time

AR: On page 43, the author argues that the group delay is identical to the dwell time, and seems to suggest that since the latter is “certainly not a traversal time,” neither is the former? In the case of a particle in the classical limit traversing a barrier, both the group delay and the dwell time tend to \(L/v = mL/\hbar k\). Would one not interpret this as the transit time? I don’t think the fact that the (integral) definition of dwell time leads to the same mathematical result is any argument against interpreting it as a transit time.

Response

HW: I am not quite sure I understand what the referee means by classical limit of a particle traversing a barrier since quantum tunneling is a purely quantum mechanical effect. If the referee really means the motion of particles of mass \(m\) traversing distance \(L\) in the absence of a barrier then we have classically allowed motion for which a classical correspondence can be made. In that case both the group delay and dwell time for the wave packet indeed yield \(\tau = mL/\hbar k\) which can be identified with a classical traversal time \(L/v\). Note two very important points here: (1) The motion is classically allowed, and (2) There are no reflections, meaning that there is only one exit channel and a unidirectional flow of particles. This is the only condition under which the dwell time and group delay can be associated with a transit time from A to B.

AR: I don’t think the fact that the (integral) definition of dwell time leads to the same mathematical result is any argument against interpreting it as a transit time.
HW: That is not the only reason for rejecting the transit time interpretation of group delay. As I point out, there is no proof that the group delay in tunneling is a transit time. It is only an assumption. It is understood that the dwell time is not a transit time since it does not differentiate between transmission and reflection channels. What has hitherto not been realized is that the group delay in tunneling has exactly the same status in that it does not differentiate between reflection and transmission channels. You may protest that we calculate the transmission group delay from the transmission phase shift. But the transmission phase shift comes from the solution of the steady-state Helmholtz or Schrödinger equation with the appropriate boundary conditions. This steady-state solution contains both forward (to be transmitted) and backward (to be reflected) components. Indeed, without the backward component there would be no net transmitted flux: tunneling requires the presence of both forward and backward components simultaneously. Thus the transmitted wave function contains within its amplitude and phase the effects of the backward going field as well. The transmission group delay does not describe unidirectional flow. Neither does the reflection group delay. And in fact both delays are equal (for a symmetric barrier). The relation between group delay and dwell time is not an accident: they are both properties of an entire field with transmitted and reflected components that cannot be disentangled.

14.5. Detection of tunneling particles

AR: At the bottom of the same page 44, “detectors detect energy and what gets transmitted through the barrier is always less...” Indeed, on average less is transmitted. But a quantum particle is in the end detected or not. A photon detector captures $\hbar \omega$ of energy even if the transmission is 0.1%; it never absorbs a quantum of $\hbar \omega/1000$.

HW: The referee argues that a quantum particle is either detected or not. This is presumably in connection with the notion that because the wave packet transmitted through the barrier peaks sooner than the reference wave packet, it must mean that the tunneling particle arrives sooner. Quantum mechanics deals with probabilities and the probability of detecting a particle at any given moment at a certain point beyond a barrier is always orders of magnitude less in the presence of the barrier than in its absence. The detector monitoring the non-tunneling particles will always click first, compared to the one monitoring the tunneling particles. When a particle is eventually detected beyond the barrier, there is no way to compare its “transit time” with that of a particle that traveled in a barrier free region, given that we do not know when the particle entered the barrier. It is true that detectors capture whole quanta but the fact is that we can only ask the question “what is the rate at which quanta are arriving at my detector?” The answer would be that the arrival rate with the barrier in place is orders of magnitude less than in its absence. We cannot ask how long a particular quantum took to traverse the barrier (see point made in 14.3).

14.6. How the lifetime argument explains apparent superluminality

AR: Starting on page 44, I would appreciate a more complete description of how the lifetime explanation leads to an output pulse that (a) has the same rise & fall time, i.e., is symmetric; (b) has the same shape as the incident pulse; and (c) doesn’t depend on the thickness of the barrier. These are all among the author’s central points, and in the ensuing pages, he tried to make them clear, and by the end, I think I began to glimpse his argument, but it could be more cogently presented. I suppose that since the lifetime is shorter than the pulse length, the output shape is dominated by the latter? I suppose this is also why one doesn’t see a rise time dominated by the pulse shape and a fall time dominated by the lifetime? Is there another (non-quasistatic) regime where one would see this? The hardest argument to buy is how the apparent superluminal transmission can be explained by calling it a lifetime. If the barrier were pointlike, then certainly a lifetime explanation would be a nice explanation of a delay. But from “the light arrived at $t = 0$, lived for 1 s, and left at $t = 1$ s” to “the light arrived at $x = 0$, lived for 1 s, and left at $t = 1$ s from $x = 1000 \text{ km}$” there are still 1000 km which have been swept under the rug! Somehow a particle arrives on one side but leaves from the other, and Winful wants to ignore the issue of how it gets from one side to the other, talking instead of how long it lives in the intermediate regime. If the bus driver refuses to let me stay on his bus for more than 5 min, I don’t think this means I’ll get to Tipperary any faster... Seriously, later on the author makes some arguments about how the entire field under the barrier grows or shrinks as a whole, and I believe this is the fundamental point. As there are no propagating waves under the barrier, this charging and discharging is spatially “instantaneous.” But I don’t think that point is made early enough, and a reader would probably find the point easier to follow if the issue of “what happened to the transit?” were dealt with earlier in this section.
Response

HW: Imagine a very long cavity resonator, say a Fabry–Perot. This resonator supports modes which are standing waves. In a standing wave every spatial point moves up and down in phase. When the resonator is operating in a single mode, the entire system pulsates and throbs like one big oscillator. Here is a distributed system acting like a lumped element. The standing wave field is

\[ E(z, t) = A \cos \omega t \cos 2kz, \]

with a time averaged intensity

\[ I = I_0 \cos 2kz. \]

Suppose we now modulate this standing wave so that

\[ I(t) = I_0(t) \cos 2kz. \]

Every spatial part of this standing wave will follow the modulation in phase, with no delay if the modulation is slow compared to the round trip time in the cavity. In fact, an observer at the exit \( L \) will see the same modulation as at the input. For an infinitely long barrier the standing wave mode is an evanescent field with an exponential distribution. This entire spatial distribution oscillates up and down with no delay in response to a slow modulation. Thus we are used to spatial distributions that respond “instantaneously”, with no delay. In fact, what should be surprising is that there is a delay in a finite-length tunneling barrier. The delay is a result of the fact that in a finite structure some energy can escape through the boundaries and has to be replenished.

To answer the questions raised in (a), (b), and (c):

(a) The output pulse is symmetric (has same rise and fall time) because the lifetime is much shorter than the pulse length. Pulse asymmetry can be seen when the lifetime is comparable to the pulse width, as at transmission resonances.

(b) The output pulse has the same shape as the input pulse because the input simply modulates the stored energy, which is proportional to the instantaneous value of the slowly varying input power. Every portion of the transmitted pulse experiences the same delay \( \tau_g = U(t)/P_{in}(t) \).

(c) For thick enough barriers the group delay is independent of barrier thickness (assuming pulse length exceeds barrier length) because the stored energy saturates with barrier length. The group delay is just the lifetime of this stored energy, most of it exiting the barrier at the input end (where most of the energy is stored).

Even though the barrier is not point like (lumped element) it can be characterized by a lifetime since it supports standing waves. Resonant cavities have lifetimes even though they are not point like objects. Once a mode has been established in the cavity, the entire cavity is characterized by the oscillations of that mode.

AR: Somehow a particle arrives on one side but leaves from the other, and Winful wants to ignore the issue of how it gets from one side to the other, talking instead of how long it lives in the intermediate regime. If the bus driver refuses to let me stay on his bus for more than 5 min, I don’t think this means I’ll get to Tipperary any faster...

HW: I will pursue your bus analogy further in order to clarify this lifetime issue. The group delay or lifetime is not the time any one passenger spends on the bus. It is the time it takes to empty the bus. So, suppose two buses leave from Cork and head for Tipperary. You get on bus A while your twin brother gets on bus B. On your bus A there are 100 passengers while on bus B there are only 10 passengers. Both buses arrive at the Tipperary bus stop at about the same time and the first passenger from each bus steps off. A dinner of corned beef and cabbage awaits the passengers of each bus when all the passengers off that bus have disembarked. Who is likely to be eating first: you or your twin brother? Clearly, your twin brother, on the bus with only 10 passengers will most certainly be dining sooner than you.

This also explains why the group delay is the same whether you go from Cork to Tipperary or from Cork to Galway. It only depends on the number of passengers on the bus and not the distance traveled. It is the time it takes to empty the bus. The more passengers aboard, the longer it takes.

At this point I should also reiterate the point I made earlier that quantum mechanics has nothing to say about the motion of a single identifiable particle. To quote Liboff, in *Introductory Quantum Mechanics* [125] “In general, we may note the fundamental rule that quantum mechanics does not delineate the trajectory of a single particle.” We cannot mark a particular electron, say, Ben, at the barrier entrance, detect Ben again some time later at the exit and say
“Aha we have measured Ben’s group delay.” It is best to think in terms of wave packets. After all, tunneling is a wave phenomenon. As a wave packet approaches the barrier, the probability density for finding a particle inside the barrier slowly rises. At some instant the wave function attains a maximum at the input. . .or, at least it would have attained a maximum at the input in the absence of reflections from the barrier. If you think in terms of an ensemble of particles, by the time any particle arrives at the input there is already a distribution of particles inside the barrier. . . an exponential distribution with most of the particles bunched within a 1/e decay distance from the input. There is only a tiny trickle of particles near the exit. Neglecting any delays due to self-interference on the approach to the barrier, we can ask, at what time does a maximum in the probability density occur at the exit? It is the time it takes for the output flux to rise to the steady-state value dictated by the new input conditions. It is not the time taken by any one particle to go from input to output.

Lifetime is a property of the whole structure. Thus you cannot say “the light arrived at \( x = 0 \), lived for 1 s, and left at \( t = 1 \text{s from } x = 1000 \text{ km} \)” First of all the “light” that leaves from \( x = 1000 \text{ km} \) at \( t = 1 \text{s} \) is not the same light that entered at \( x = 0 \) at \( t = 0 \). If you recall the discussion regarding fronts and the short transient period, by the time the bulk of the pulse arrives at the input, the barrier is already filled with light arranged in an exponentially decreasing energy density along the barrier. When this new light enters, the total stored energy adjusts its value to be consistent with the new input conditions. This adjustment time is the 1/e lifetime of stored energy escaping through both ends of the barrier (with most of it escaping in the backward direction). Again, once the exponential standing wave (evanescent mode) has been established within the barrier, the newly arriving light modulates this stored energy and thus the amount of flux that escapes through the boundaries. Other than the front, it is not possible to track a particular point within an incoming pulse as it tunnels through the barrier. Because of the multiple reflections, once any light enters, it gets all mixed up, scrambled, so that we cannot look at the transmitted pulse and say, aha this portion of the transmitted pulse entered the barrier at such and such a time.

**AR:** Somehow a particle arrives on one side but leaves from the other, and Winful wants to ignore the issue of how it gets from one side to the other, talking instead of how long it lives in the intermediate regime.

**HW:** I hope by now it is clear that we cannot ask how any single particle gets from one side of the barrier to the other. The dwell time and group delay are not transit times and cannot be used for this purpose. I suggest that one meaningful quantity to characterize “transit time” in tunneling is the length of the wave packet. This is because the wave packet is much longer than the barrier. The time scale of the interaction (and transit!), on an ensemble average basis, is therefore set by the wave-packet length and not by the barrier length. Basically, when the uncertainty in particle position greatly exceeds the barrier length (which is the case for true tunneling) the notion of the transit time of any single particle through the barrier is meaningless.

References