Atmospheric Optics

Craig F. Bohren
Pennsylvania State University, Department of Meteorology, University Park, Pennsylvania, USA
Phone: (814) 466-6264; Fax: (814) 865-3663; e-mail: bohren@ems.psu.edu

Abstract
Colors of the sky and colored displays in the sky are mostly a consequence of selective scattering by molecules or particles, absorption usually being irrelevant. Molecular scattering selective by wavelength – incident sunlight of some wavelengths being scattered more than others – but the same in any direction at all wavelengths gives rise to the blue of the sky and the red of sunsets and sunrises. Scattering by particles selective by direction – different in different directions at a given wavelength – gives rise to rainbows, coronas, iridescent clouds, the glory, sun dogs, halos, and other ice-crystal displays. The size distribution of these particles and their shapes determine what is observed, water droplets and ice crystals, for example, resulting in distinct displays.

To understand the variation and color and brightness of the sky as well as the brightness of clouds requires coming to grips with multiple scattering: scatterers in an ensemble are illuminated by incident sunlight and by the scattered light from each other. The optical properties of an ensemble are not necessarily those of its individual members.

Mirages are a consequence of the spatial variation of coherent scattering (refraction) by air molecules, whereas the green flash owes its existence to both coherent scattering by molecules and incoherent scattering by molecules and particles.

Keywords
sky colors; mirages; green flash; coronas; rainbows; the glory; sun dogs; halos; visibility.

1 Introduction

Atmospheric optics is nearly synonymous with light scattering, the only restrictions being that the scatterers inhabit the atmosphere and the primary source of their illumination is the sun. Essentially all light we see is scattered light, even that directly from the sun. When we say that such light is unscattered we really mean that it is scattered in the forward direction;

hence it is *as if* it were unscattered. Scattered light is radiation from matter excited by an external source. When the source vanishes, so does the scattered light, as distinguished from light emitted by matter, which persists in the absence of external sources.

Atmospheric scatterers are either molecules or particles. A particle is an aggregation of sufficiently many molecules that it can be ascribed macroscopic properties such as temperature and refractive index. There is no canonical number of molecules that must unite to form a *bona fide* particle. Two molecules clearly do not a quorum make, but what about 10, 100, 1000? The particle size corresponding to the largest of these numbers is about 10^{-3} μm. Particles this small of water substance would evaporate so rapidly that they could not exist long under conditions normally found in the atmosphere. As a practical matter, therefore, we need not worry unduly about scatterers in the shadow region between molecule and particle.

A property of great relevance to scattering problems is *coherence*, both of the array of scatterers and of the incident light. At visible wavelengths, air is an array of incoherent scatterers: the radiant power scattered by N molecules is N times that scattered by one (except in the forward direction). But when water vapor in air condenses, an incoherent array is transformed into a coherent array: uncorrelated water molecules become part of a single entity. Although a single droplet is a coherent array, a cloud of droplets taken together is incoherent.

Sunlight is incoherent but not in an absolute sense. Its lateral coherence length is tens of micrometers, which is why we can observe what are essentially interference patterns (e.g., coronas and glories) resulting from illumination of cloud droplets by sunlight.

This article begins with the color and brightness of a purely molecular atmosphere, including their variation across the vault of the sky. This naturally leads to the state of polarization of skylight. Because the atmosphere is rarely, if ever, entirely free of particles, the general characteristics of scattering by particles follow, setting the stage for a discussion of atmospheric visibility.

Atmospheric refraction usually sits by itself, unjustly isolated from all those atmospheric phenomena embraced by the term scattering. Yet refraction is another manifestation of scattering, coherent scattering in the sense that phase differences cannot be ignored.

Scattering by single water droplets and ice crystals, each discussed in turn, yields feasts for the eye as well as the mind. The curtain closes on the optical properties of clouds.

2
Color and Brightness of Molecular Atmosphere

2.1
A Brief History

Edward Nichols began his 1908 presidential address to the New York meeting of the American Physical Society as follows: ''In asking your attention to-day, even briefly, to the consideration of the present state of our knowledge concerning the color of the sky it may be truly said that I am inviting you to leave the thronged thoroughfares of our science for some quiet side street where little is going on and you may even suspect that I am coaxing you into some

blind alley, the inhabitants of which belong to the dead past.''

Despite this depreciatory statement, hoary with age, correct and complete explanations of the color of the sky still are hard to find. Indeed, all the faulty explanations lead active lives: the blue sky is the reflection of the blue sea; it is caused by water, either vapor or droplets or both; it is caused by dust. The true cause of the blue sky is not difficult to understand, requiring only a bit of critical thought stimulated by belief in the inherent fascination of all natural phenomena, even those made familiar by everyday occurrence.

Our contemplative prehistoric ancestors no doubt speculated on the origin of the blue sky, their musings having vanished into it. Yet it is curious that Aristotle, the most prolific speculator of early recorded history, makes no mention of it in his *Meteorologica* even though he delivered pronouncements on rainbows, halos, and mock suns and realized that ''the sun looks red when seen through mist or smoke.'' Historical discussions of the blue sky sometimes cite Leonardo as the first to comment intelligently on the blue of the sky, although this reflects a European bias. If history were to be written by a supremely disinterested observer, Arab philosophers would likely be given more credit for having had profound insights into the workings of nature many centuries before their European counterparts descended from the trees. Indeed, Möller [1] begins his brief history of the blue sky with Jakub Ibn Ishak Al Kindi (800–870), who explained it as ''a mixture of the darkness of the night with the light of the dust and haze particles in the air illuminated by the sun.''

Leonardo was a keen observer of light in nature even if his explanations sometimes fell short of the mark. Yet his hypothesis that ''the blueness we see in the atmosphere is not intrinsic color, but is caused by warm vapor evaporated in minute and insensible atoms on which the solar rays fall, rendering them luminous against the infinite darkness of the fiery sphere which lies beyond and includes it'' would, with minor changes, stand critical scrutiny today. If we set aside Leonardo as *sui generis*, scientific attempts to unravel the origins of the blue sky may be said to have begun with Newton, that towering pioneer of optics, who, in time-honored fashion, reduced it to what he already had considered: interference colors in thin films. Almost two centuries elapsed before more pieces in the puzzle were contributed by the experimental investigations of von Brücke and Tyndall on light scattering by suspensions of particles. Around the same time Clausius added his bit in the form of a theory that scattering by minute bubbles causes the blueness of the sky. A better theory was not long in coming. It is associated with a man known to the world as Lord Rayleigh even though he was born John William Strutt.

Rayleigh's paper of 1871 marks the beginning of a satisfactory explanation of the blue sky. His scattering law, the key to the blue sky, is perhaps the most famous result ever obtained by dimensional analysis. Rayleigh argued that the field E_s scattered by a particle small compared with the light illuminating it is proportional to its volume V and to the incident field E_i. Radiant energy conservation requires that the scattered field diminish inversely as the distance r from the particle so that the scattered power diminishes as the square of r. To make this proportionality dimensionally homogeneous requires the inverse square of a quantity with the dimensions of

length. The only plausible physical variable at hand is the wavelength of the incident light, which leads to

$$E_s \propto \frac{E_i V}{r\lambda^2}. \quad (1)$$

When the field is squared to obtain the scattered power, the result is Rayleigh's inverse fourth-power law. This law is really only an often – but not always – very good approximation. Missing from it are dimensionless properties of the particle such as its refractive index, which itself depends on wavelength. Because of this *dispersion*, therefore, nothing scatters exactly as the inverse fourth power.

Rayleigh's 1871 paper did not give the complete explanation of the color and polarization of skylight. What he did that was not done by his predecessors was to give a law of scattering, which could be used to test quantitatively the hypothesis that selective scattering by atmospheric particles could transform white sunlight into blue skylight. But as far as giving the agent responsible for the blue sky is concerned, Rayleigh did not go essentially beyond Newton and Tyndall, who invoked particles. Rayleigh was circumspect about the nature of these particles, settling on salt as the most likely candidate. It was not until 1899 that he published the capstone to his work on skylight, arguing that air molecules themselves were the source of the blue sky. Tyndall cannot be given the credit for this because he considered air to be *optically empty*: when purged of all particles it scatters no light. This erroneous conclusion was a result of the small scale of his laboratory experiments. On the scale of the atmosphere, sufficient light is scattered by air molecules to be readily observable.

2.2 Molecular Scattering and the Blue of the Sky

Our illustrious predecessors all gave explanations of the blue sky requiring the presence of water in the atmosphere: Leonardo's "evaporated warm vapor," Newton's "Globules of water," Clausius's bubbles. Small wonder, then, that water still is invoked as the cause of the blue sky. Yet a cause of something is that without which it would not occur, and the sky would be no less blue if the atmosphere were free of water.

A possible physical reason for attributing the blue sky to water vapor is that, because of selective *absorption*, liquid water (and ice) is blue upon transmission of white light over distances of order meters. Yet if all the water in the atmosphere at any instant were to be compressed into a liquid, the result would be a layer about 1 cm thick, which is not sufficient to transform white light into blue by selective absorption.

Water vapor does not compensate for its hundredfold lower abundance than nitrogen and oxygen by greater scattering per molecule. Indeed, scattering of visible light by a water molecule is slightly *less* than that by either nitrogen or oxygen.

Scattering by atmospheric molecules does not obey Rayleigh's inverse fourth-power law exactly. A least-squares fit over the visible spectrum from 400 to 700 nm of the *molecular scattering coefficient* of sea-level air tabulated by Penndorf [2] yields an inverse 4.089th-power scattering law.

The molecular scattering coefficient β, which plays important roles in following sections, may be written

$$\beta = N\sigma_s, \quad (2)$$

where N is the number of molecules per unit volume and σ_s, the scattering cross section (an average because air is

a mixture) per molecule, approximately obeys Rayleigh's law. The form of this expression betrays the incoherence of scattering by atmospheric molecules. The inverse of β is interpreted as the scattering *mean free path*, the average distance a photon must travel before being scattered.

To say that the sky is blue because of Rayleigh scattering, as is sometimes done, is to confuse an agent with a law. Moreover, as Young [3] pointed out, the term Rayleigh scattering has many meanings. Particles small compared with the wavelength scatter according to the same law as do molecules. Both can be said to be Rayleigh scatterers, but only molecules are necessary for the blue sky. Particles, even small ones, generally diminish the vividness of the blue sky.

Fluctuations are sometimes trumpeted as the "real" cause of the blue sky. Presumably, this stems from the fluctuation theory of light scattering by media in which the scatterers are separated by distances small compared with the wavelength. In this theory, which is associated with Einstein and Smoluchowski, matter is taken to be continuous but characterized by a refractive index that is a random function of position. Einstein [4] stated that "it is remarkable that our theory does not make *direct* use of the assumption of a discrete distribution of matter." That is, he circumvented a difficulty but realized it could have been met head on, as Zimm [5] did years later.

The blue sky is really caused by scattering by molecules – to be more precise, scattering by bound electrons: free electrons do not scatter selectively. Because air molecules are separated by distances small compared with the wavelengths of visible light, it is not obvious that the power scattered by such molecules can be added. Yet if they are completely uncorrelated, as in an ideal gas (to good approximation the atmosphere is an ideal gas), scattering by N molecules is N times scattering by one. This is the only sense in which the blue sky can be attributed to scattering by fluctuations. Perfectly homogeneous matter does not exist. As stated pithily by Planck, "a chemically pure substance may be spoken of as a vacuum made turbid by the presence of molecules."

2.3 Spectrum and Color of Skylight

What is the spectrum of skylight? What is its color? These are two different questions. Answering the first answers the second but not the reverse. Knowing the color of skylight we cannot uniquely determine its spectrum because of *metamerism*: A given perceived color can in general be obtained in an indefinite number of ways.

Skylight is not blue (itself an imprecise term) in an absolute sense. When the visible spectrum of sunlight outside the earth's atmosphere is modulated by Rayleigh's scattering law, the result is a spectrum of scattered light that is neither solely blue nor even peaked in the blue (Fig. 1). Although blue does not predominate spectrally, it does predominate perceptually. We perceive the sky to be blue even though skylight contains light of all wavelengths.

Any source of light may be looked upon as a mixture of white light and light of a single wavelength called the *dominant wavelength*. The *purity* of the source is the relative amount of the monochromatic component in the mixture. The dominant wavelength of sunlight scattered according to Rayleigh's law is about 475 nm, which lies solidly in the blue if we take this to mean light with wavelengths between 450 and 490 nm. The purity of this

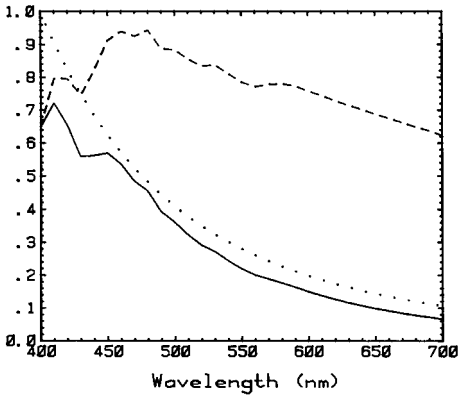

Fig. 1 Rayleigh's scattering law (dots), the spectrum of sunlight outside the Earth's atmosphere (dashes), and the product of the two (solid curve). The solar spectrum is taken from Thekaekara, M. P., Drummond, A. J. (**1971**), *Nat. Phys. Sci.* **229**, 6–9 [6]

scattered light, about 42%, is the upper limit for skylight. Blues of real skies are less pure.

Another way of conveying the color of a source of light is by its *color temperature*, the temperature of a blackbody having the same perceived color as the source. Since blackbodies do not span the entire gamut of colors, not all sources of light can be assigned color temperatures. But many natural sources of light can. The color temperature of light scattered according to Rayleigh's law is infinite. This follows from Planck's spectral emission function $e_{b\lambda}$ in the limit of high temperature,

$$e_{b\lambda} \approx \frac{2\pi ckT}{\lambda^4}, \quad \frac{hc}{\lambda} \ll kT, \qquad (3)$$

where h is Planck's constant, k is Boltzmann's constant, c is the speed of light *in vacuo*, and T is absolute temperature. Thus, the emission spectrum of a blackbody with an infinite temperature has the same functional form as Rayleigh's scattering law.

2.4 Variation of Sky Color and Brightness

Not only is skylight not pure blue, but its color and brightness vary across the vault of the sky, with the best blues at zenith. Near the astronomical horizon the sky is brighter than overhead but of considerably lower purity. That this variation can be observed from an airplane flying at 10 km, well above most particles, suggests that the sky is inherently nonuniform in color and brightness (Fig. 2). To understand why requires invoking multiple scattering.

Multiple scattering gives rise to observable phenomena that cannot be explained solely by single-scattering arguments. This is easily demonstrated. Fill a blackened pan

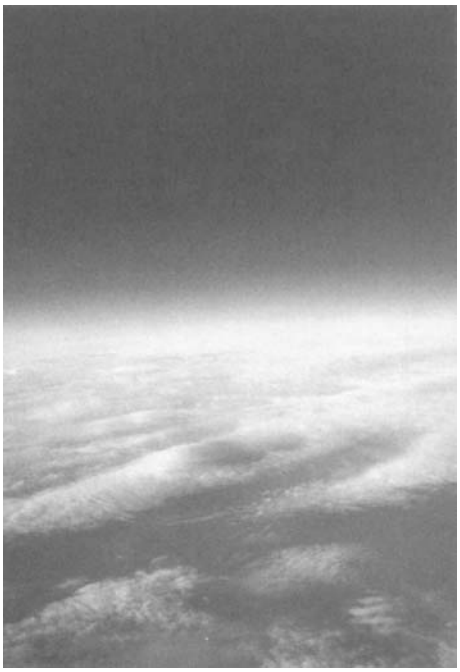

Fig. 2 Even at an altitude of 10 km, well above most particles, the sky brightness increases markedly from the zenith to the astronomical horizon

with clean water, then add a few drops of milk. The resulting dilute suspension illuminated by sunlight has a bluish cast. But when more milk is added, the suspension turns white. Yet the properties of the scatterers (fat globules) have not changed, only their *optical thickness*: the blue suspension being optically thin, the white being optically thick.

Optical thickness is physical thickness in units of scattering mean free path, and hence is dimensionless. The optical thickness τ between any two points connected by an arbitrary path in a medium populated by (incoherent) scatterers is an integral over the path:

$$\tau = \int_1^2 \beta \, ds. \qquad (4)$$

The *normal optical thickness* τ_n of the atmosphere is that along a radial path extending from the surface of the Earth to infinity. Figure 3 shows τ_n over the visible spectrum for a purely molecular atmosphere. Because τ_n is generally small compared with unity, a photon from the sun traversing a radial path in the atmosphere is unlikely to be scattered more than once. But along a tangential path, the optical thickness is about 35 times greater (Fig. 4), which leads to several observable phenomena.

Even an intrinsically black object is luminous to an observer because of *airlight*, light scattered by all the molecules and particles along the line of sight from observer to object. Provided that this is uniformly illuminated by sunlight and that ground reflection is negligible, the airlight radiance L is approximately

$$L = GL_0(1 - e^{-\tau}), \qquad (5)$$

where L_0 is the radiance of incident sunlight along the line of sight with optical thickness τ. The term G accounts for geometric reduction of radiance because of scattering of nearly monodirectional sunlight in all directions. If the line of sight is uniform in composition, $\tau = \beta d$, where β is the scattering coefficient and d is the physical distance to the black object.

If τ is small ($\ll 1$), $L \approx GL_0\tau$. In a purely molecular atmosphere, τ varies with wavelength according to Rayleigh's law; hence the distant black object in such an atmosphere is perceived to be

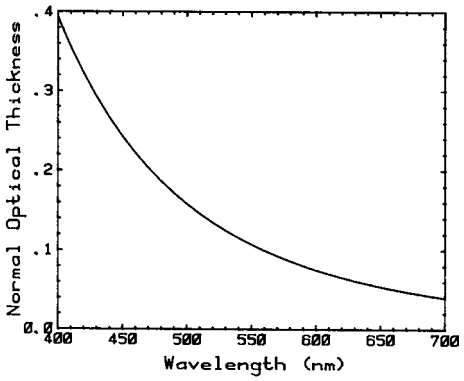

Fig. 3 Normal optical thickness of a pure molecular atmosphere

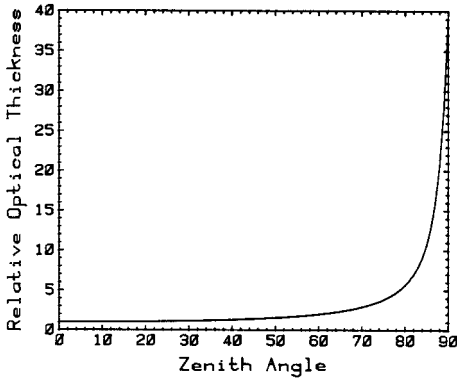

Fig. 4 Optical thickness (relative to the normal optical thickness) of a molecular atmosphere along various paths with zenith angles between 0° (normal) and 90° (tangential)

bluish. As τ increases so does L but not proportionally. Its limit is GL_0: The airlight radiance spectrum is that of the source of illumination. Only in the limit $d = 0$ is $L = 0$ and the black object truly black.

Variation of the brightness and color of dark objects with distance was called *aerial perspective* by Leonardo. By means of it we estimate distances to objects of unknown size such as mountains.

Aerial perspective belongs to the same family as the variation of color and brightness of the sky with zenith angle. Although the optical thickness along a path tangent to the Earth is not infinite, it is sufficiently large (Figs. 3 and 4) that GL_0 is a good approximation for the radiance of the horizon sky. For isotropic scattering (a condition almost satisfied by molecules), G is around 10^{-5}, the ratio of the solid angle subtended by the sun to the solid angle of all directions (4π). Thus, the horizon sky is not nearly so bright as direct sunlight.

Unlike in the milk experiment, what is observed when looking at the horizon sky is not multiply scattered light. Both have their origins in multiple scattering but manifested in different ways. Milk is white because it is weakly absorbing and optically thick, and hence all components of incident white light are multiply scattered to the observer even though the blue component traverses a shorter average path in the suspension than the red component. White horizon light has escaped being multiply scattered, although multiple scattering is why this light is white (strictly, has the spectrum of the source). More light at the short-wavelength end of the spectrum is scattered *toward* the observer than at the long-wavelength end. But long-wavelength light has the greater likelihood of being transmitted to the observer without being scattered *out of* the line of sight. For a long optical path, these two processes compensate, resulting in a horizon radiance spectrum which is that of the source.

Selective scattering by molecules is not sufficient for a blue sky. The atmosphere also must be optically thin, at least for most zenith angles (Fig. 4) (the blackness of space as a backdrop is taken for granted but also is necessary, as Leonardo recognized). A corollary of this is that the blue sky is not inevitable: an atmosphere composed entirely of nonabsorbing, selectively scattering molecules overlying a nonselectively reflecting earth need not be blue. Figure 5 shows calculated spectra of the zenith sky over black ground for a molecular atmosphere with the present normal optical thickness as well as for hypothetical atmospheres 10 and 40 times thicker. What we take to be inevitable is accidental: If our atmosphere were much thicker, but identical in composition, the color of the sky would be quite different from what it is now.

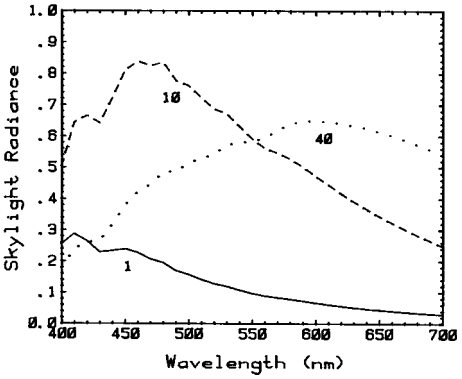

Fig. 5 Spectrum of overhead skylight for the present molecular atmosphere (solid curve), as well as for hypothetical atmospheres 10 (dashes) and 40 (dots) times thicker

2.5 Sunrise and Sunset

If short-wavelength light is preferentially scattered out of direct sunlight, long-wavelength light is preferentially transmitted in the direction of sunlight. Transmission is described by an exponential law (if light multiply scattered back into the direction of the sunlight is negligible):

$$L = L_0 e^{-\tau}, \tag{6}$$

where L is the radiance at the observer in the direction of the sun, L_0 is the radiance of sunlight outside the atmosphere, and τ is the optical thickness along this path.

If the wavelength dependence of τ is given by Rayleigh's law, sunlight is *reddened* upon transmission: The spectrum of the transmitted light is comparatively richer than the incident spectrum in light at the long-wavelength end of the visible spectrum. But to say that transmitted sunlight is reddened is not the same as saying it is red. The perceived color can be yellow, orange, or red, depending on the magnitude of the optical thickness. In a molecular atmosphere, the optical thickness along a path from the sun, even on or below the horizon, is not sufficient to give red light upon transmission. Although selective scattering by molecules yields a blue sky, reds are not possible in a molecular atmosphere, only yellows and oranges. This can be observed on clear days, when the horizon sky at sunset becomes successively tinged with yellow, then orange, but not red.

Equation (6) applies to the radiance only in the direction of the sun. Oranges and reds can be seen in other directions because reddened sunlight illuminates scatterers not lying along the line of sight to the sun. A striking example of this is a horizon sky tinged with oranges and pinks in the direction *opposite* the sun.

The color and brightness of the sun changes as it arcs across the sky because the optical thickness along the line of sight changes with solar zenith angle Θ. If the Earth were flat (as some still aver), the transmitted solar radiance would be

$$L = L_0 e^{\tau_n / \cos\Theta}. \tag{7}$$

This equation is a good approximation except near the horizon. On a flat earth, the optical thickness is infinite for horizon paths. On a spherical earth, optical thicknesses are finite although much larger for horizon than for radial paths.

The normal optical thickness of an atmosphere in which the number density of scatterers decreases exponentially with height z above the surface, $\exp(-z/H)$, is the same as that for a uniform atmosphere of finite thickness:

$$\tau_n = \int_0^\infty \beta \, dz = \beta_0 H, \tag{8}$$

where H is the *scale height* and β_0 is the scattering coefficient at sea level. This equivalence yields a good approximation even for the tangential optical thickness. For any zenith angle, the optical thickness is given approximately by

$$\frac{\tau}{\tau_n} = \sqrt{\frac{R_e^2}{H^2}\cos^2\Theta + \frac{2R_e}{H} + 1} - \frac{R_e}{H}\cos\Theta, \tag{9}$$

where R_e is the radius of the Earth. A flat earth is one for which R_e is infinite, in which instance Eq. (9) yields

the expected relation

$$\lim_{R_e \to \infty} \frac{\tau}{\tau_n} = \frac{1}{\cos \Theta}. \qquad (10)$$

For Earth's atmosphere, the molecular scale height is about 8 km. According to the approximate relation Eq. (9), therefore, the horizon optical thickness is about 39 times greater than the normal optical thickness. Taking the exponential decrease of molecular number density into account yields a value about 10% lower.

Variations on the theme of reds and oranges at sunrise and sunset can be seen even when the sun is overhead. The radiance at an observer an optical distance τ from a (horizon) cloud is the sum of cloudlight transmitted to the observer and airlight:

$$L = L_0 G(1 - e^{-\tau}) + L_0 G_c e^{-\tau}, \qquad (11)$$

where G_c is a geometrical factor that accounts for scattering of nearly monodirectional sunlight into a hemisphere of directions by the cloud. If the cloud is approximated as an isotropic reflector with reflectance R and illuminated at an angle Φ, the geometrical factor G_c is $\Omega_s R \cos\Phi/\pi$, where Ω_s is the solid angle subtended by the sun at the Earth. If $G_c > G$, the observed radiance is redder (i.e., enriched in light of longer wavelengths) than the incident radiance. If $G_c < G$, the observed radiance is bluer than the incident radiance. Thus, distant horizon clouds can be reddish if they are bright or bluish if they are dark.

Underlying Eq. (11) is the implicit assumption that the line of sight is uniformly illuminated by sunlight. The first term in this equation is airlight; the second is transmitted cloudlight. Suppose, however, that the line of sight is shadowed from direct sunlight by clouds (that do not, of course, occlude the distant cloud of interest). This may reduce the first term in Eq. (11) so that the second term dominates. Thus, under a partly overcast sky, distant horizon clouds may be reddish even when the sun is high in the sky.

The zenith sky at sunset and twilight is the exception to the general rule that molecular scattering is sufficient to account for the color of the sky. In the absence of molecular absorption, the spectrum of the zenith sky would be essentially that of the zenith sun (although greatly reduced in radiance), hence would not be the blue that is observed. This was pointed out by Hulburt [7], who showed that absorption by ozone profoundly affects the color of the zenith sky when the sun is near the horizon. The Chappuis band of ozone extends from about 450 to 700 nm and peaks at around 600 nm. Preferential absorption of sunlight by ozone over long horizon paths gives the zenith sky its blueness when the sun is near the horizon. With the sun more than about 10° above the horizon, however, ozone has little effect on the color of the sky.

3
Polarization of Light in a Molecular Atmosphere

3.1
The Nature of Polarized Light

Unlike sound, light is a vector wave, an electromagnetic field lying in a plane normal to the propagation direction. The polarization state of such a wave is determined by the degree of correlation of any

two orthogonal components into which its electric (or magnetic) field is resolved. Completely polarized light corresponds to complete correlation; completely unpolarized light corresponds to no correlation; partially polarized light corresponds to partial correlation.

If an electromagnetic wave is completely polarized, the tip of its oscillating electric field traces out a definite elliptical curve, the *vibration ellipse*. Lines and circles are special ellipses, the light being said to be linearly or circularly polarized, respectively. The general state of polarization is elliptical.

Any beam of light can be considered an incoherent superposition of two collinear beams, one unpolarized, the other completely polarized. The radiance of the polarized component relative to the total is defined as the *degree of polarization* (often multiplied by 100 and expressed as a percentage). This can be measured for a source of light (e.g., light from different sky directions) by rotating a (linear) polarizing filter and noting the minimum and maximum radiances transmitted by it. The degree of (linear) polarization is defined as the difference between these two radiances divided by their sum.

3.2 Polarization by Molecular Scattering

Unpolarized light can be transformed into partially polarized light upon interaction with matter because of different changes in amplitude of the two orthogonal field components. An example of this is the partial polarization of sunlight upon scattering by atmospheric molecules, which can be detected by looking at the sky through a polarizing filter (e.g., polarizing sunglasses) while rotating it. Waxing and waning of the observed brightness indicates some degree of partial polarization.

In the analysis of any scattering problem, a plane of reference is required. This is usually the *scattering plane*, determined by the directions of the incident and scattered waves, the angle between them being the *scattering angle*. Light polarized perpendicular (parallel) to the scattering plane is sometimes said to be vertically (horizontally) polarized. Vertical and horizontal in this context, however, are arbitrary terms indicating orthogonality and bear no relation, except by accident, to the direction of gravity.

The degree of polarization P of light scattered by a tiny sphere illuminated by unpolarized light is (Fig. 6)

$$P = \frac{1 - \cos^2 \theta}{1 + \cos^2 \theta}, \tag{12}$$

where the scattering angle θ ranges from 0° (forward direction) to 180° (backward direction); the scattered light is partially linearly polarized perpendicular to the scattering plane. Although this equation

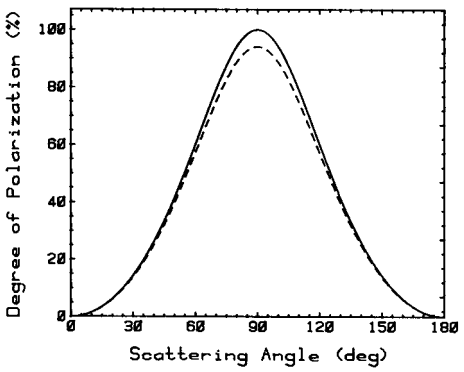

Fig. 6 Degree of polarization of the light scattered by a small (compared with the wavelength) sphere for incident unpolarized light (solid curve). The dashed curve is for a small spheroid chosen such that the degree of polarization at 90° is that for air

is a first step toward understanding polarization of skylight, more often than not it also has been a false step, having led countless authors to assert that skylight is completely polarized at 90° from the sun. Although $P = 1$ at $\theta = 90°$ according to Eq. (12), skylight is never 100% polarized at this or any other angle, and for several reasons.

Although air molecules are very small compared with the wavelengths of visible light, a requirement underlying Eq. (12), the dominant constituents of air are not spherically symmetric.

The simplest model of an asymmetric molecule is a small spheroid. Although it is indeed possible to find a direction in which the light scattered by such a spheroid is 100% polarized, this direction depends on the spheroid's orientation. In an ensemble of randomly oriented spheroids, each contributes its mite to the total radiance in a given direction, but each contribution is partially polarized to varying degrees between 0 and 100%. It is impossible for beams of light to be incoherently superposed in such a way that the degree of polarization of the resultant is greater than the degree of polarization of the most highly polarized beam.

Because air is an ensemble of randomly oriented asymmetric molecules, sunlight scattered by air never is 100% polarized. The intrinsic departure from perfection is about 6%. Figure 6 also includes a curve for light scattered by randomly oriented spheroids chosen to yield 94% polarization at 90°. This angle is so often singled out that it may deflect attention from nearby scattering angles. Yet, the degree of polarization is greater than 50% for a range of scattering angles 70° wide centered about 90°.

Equation (12) applies to air, not to the atmosphere, the distinction being that in the atmosphere, as opposed to the laboratory, multiple scattering is not negligible. Also, atmospheric air is almost never free of particles and is illuminated by light reflected by the ground. We must take the atmosphere as it is, whereas in the laboratory we often can eliminate everything we consider extraneous.

Because of both multiple scattering and ground reflection, light from any direction in the sky is not, in general, made up solely of light scattered in a single direction relative to the incident sunlight but is a superposition of beams with different scattering histories, hence different degrees of polarization. As a consequence, even if air molecules were perfect spheres and the atmosphere were completely free of particles, skylight would not be 100% polarized at 90° to the sun or at any other angle.

Reduction of the maximum degree of polarization is not the only consequence of multiple scattering. According to Fig. 6, there should be two *neutral points* in the sky, directions in which skylight is unpolarized: directly toward and away from the sun. Because of multiple scattering, however, there are three such points. When the sun is higher than about 20° above the horizon there are neutral points within 20° of the sun, the *Babinet point* above it, the *Brewster point* below. They coincide when the sun is directly overhead and move apart as the sun descends. When the sun is lower than 20°, the *Arago point* is about 20° above the antisolar point, the direction opposite the sun.

One consequence of the partial polarization of skylight is that the colors of distant objects may change when viewed through a rotated polarizing filter. If the sun is high in the sky, horizontal airlight will have a fairly high degree of polarization. According to the previous section,

airlight is bluish. But if it also is partially polarized, its radiance can be diminished with a polarizing filter. Transmitted cloudlight, however, is unpolarized. Because the radiance of airlight can be reduced more than that of cloudlight, distant clouds may change from white to yellow to orange when viewed through a rotated polarizing filter.

4
Scattering by Particles

Up to this point we have considered only an atmosphere free of particles, an idealized state rarely achieved in nature. Particles still would inhabit the atmosphere even if the human race were to vanish from the Earth. They are not simply by-products of the ''dark satanic mills'' of civilization.

All molecules of the same substance are essentially identical. This is not true of particles: They vary in shape and size, and may be composed of one or more homogeneous regions.

4.1
The Salient Differences between Particles and Molecules: Magnitude of Scattering

The distinction between scattering by molecules when widely separated and when packed together into a droplet is that between scattering by incoherent and coherent arrays. Isolated molecules are excited primarily by incident (external) light, whereas the same molecules forming a droplet are excited by incident light and by each other's scattered fields. The total power scattered by an incoherent array of molecules is the sum of their scattered powers. The total power scattered by a coherent array is the square of the total scattered field, which in turn is the sum of all the fields scattered by the individual molecules. For an incoherent array we *may* ignore the wave nature of light, whereas for a coherent array we *must* take it into account.

Water vapor is a good example to ponder because it is a constituent of air and can condense to form cloud droplets. The difference between a sky containing water vapor and the same sky with the same amount of water but in the form of a cloud of droplets is dramatic.

According to Rayleigh's law, scattering by a particle small compared with the wavelength increases as the sixth power of its size (volume squared). A droplet of diameter 0.03 μm, for example, scatters about 10^{12} times more light than does one of its constituent molecules. Such a droplet contains about 10^7 molecules. Thus, scattering per molecule as a consequence of condensation of water vapor into a coherent water droplet increases by about 10^5.

Cloud droplets are much larger than 0.03 μm, a typical diameter being about 10 μm. Scattering per molecule in such a droplet is much greater than scattering by an isolated molecule, but not to the extent given by Rayleigh's law. Scattering increases as the sixth power of droplet diameter only when the molecules scatter coherently in phase. If a droplet is sufficiently small compared with the wavelength, each of its molecules is excited by essentially the same field and all the waves scattered by them interfere constructively. But when a droplet is comparable to or larger than the wavelength, interference can be constructive, destructive, and everything in between, and hence scattering does not increase as rapidly with droplet size as predicted by Rayleigh's law.

The figure of merit for comparing scatterers of different size is their scattering cross section per unit volume, which, except for a multiplicative factor, is the scattering cross section per molecule. A scattering cross section may be looked upon as an effective area for removing radiant energy from a beam: the scattering cross section times the beam irradiance is the radiant power scattered in all directions.

The scattering cross section per unit volume for water droplets illuminated by visible light and varying in size from molecules (10^{-4} μm) to raindrops (10^3 μm) is shown in Fig. 7. Scattering by a molecule that belongs to a cloud droplet is about 10^9 times greater than scattering by an isolated molecule, a striking example of the virtue of cooperation. Yet in molecular as in human societies there are limits beyond which cooperation becomes dysfunctional: Scattering by a molecule that belongs to a raindrop is about 100 times less than scattering by a molecule that belongs to a cloud droplet. This tremendous variation of scattering by water molecules depending on their state of aggregation has profound observational consequences. A cloud is optically so much different from the water vapor out of which it was born that the offspring bears no resemblance to its parents. We can see through tens of kilometers of air laden with water vapor, whereas a cloud a few tens of meters thick is enough to occult the sun. Yet a rainshaft born out of a cloud is considerably more translucent than its parent.

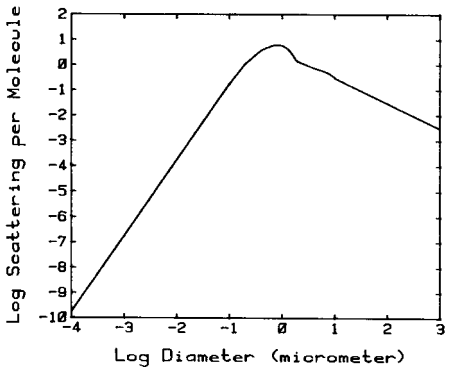

Fig. 7 Scattering (per molecule) of visible light (arbitrary units) by water droplets varying in size from a single molecule to a raindrop

4.2 The Salient Differences between Particles and Molecules: Wavelength Dependence of Scattering

Regardless of their size and composition, particles scatter approximately as the inverse fourth power of wavelength if they are small compared with the wavelength and absorption is negligible, two important caveats. Failure to recognize them has led to errors, such as that yellow light penetrates fog better because it is not scattered as much as light of shorter wavelengths. Although there may be perfectly sound reasons for choosing yellow instead of blue or green as the color of fog lights, greater transmission through fog is not one of them: Scattering by fog droplets is essentially independent of wavelength over the visible spectrum.

Small particles are selective scatterers; large particles are not. Particles neither small nor large give the reverse of what we have come to expect as normal. Figure 8 shows scattering of visible light by oil droplets with diameters 0.1, 0.8, and 10 μm. The smaller droplets scatter according to Rayleigh's law; the larger droplets (typical cloud droplet size) are nonselective. Between these two extremes are droplets (0.8 μm) that scatter long-wavelength light more than short-wavelength. Sunlight or moonlight seen

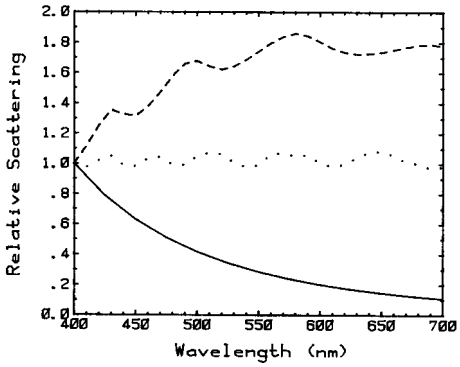

Fig. 8 Scattering of visible light by oil droplets of diameter 0.1 μm (solid curve), 0.8 μm (dashes), and 10 μm (dots)

through a thin cloud of these intermediate droplets would be bluish or greenish. This requires droplets of just the right size, and hence it is a rare event, so rare that it occurs once in a blue moon. Astronomers, for unfathomable reasons, refer to the second full moon in a month as a blue moon, but if such a moon were blue it would be only by coincidence. The last reliably reported outbreak of blue and green suns and moons occurred in 1950 and was attributed to an oily smoke produced in Canadian forest fires.

4.3 The Salient Differences between Particles and Molecules: Angular Dependence of Scattering

The angular distribution of scattered light changes dramatically with the size of the scatterer. Molecules and particles that are small compared with the wavelength are nearly isotropic scatterers of unpolarized light, the ratio of maximum (at 0° and 180°) to minimum (at 90°) scattered radiance being only 2 for spheres, and slightly less for other spheroids. Although small particles scatter the same in the forward and backward hemispheres, scattering becomes markedly asymmetric for particles comparable to or larger than the wavelength. For example, forward scattering by a water droplet as small as 0.5 μm is about 100 times greater than backward scattering, and the ratio of forward to backward scattering increases more or less monotonically with size (Fig. 9).

The reason for this asymmetry is found in the singularity of the forward direction. In this direction, waves scattered by two or more scatterers excited solely by incident light (ignoring mutual excitation) are always in phase regardless of the wavelength and the separation of the scatterers. If we imagine a particle to be made up of N small subunits, scattering in the forward direction increases as N^2, the only direction for which this is always true. For other directions, the wavelets scattered by the subunits will not necessarily all be in phase. As a consequence, scattering in the forward direction increases with size (i.e., N) more rapidly than in any other direction.

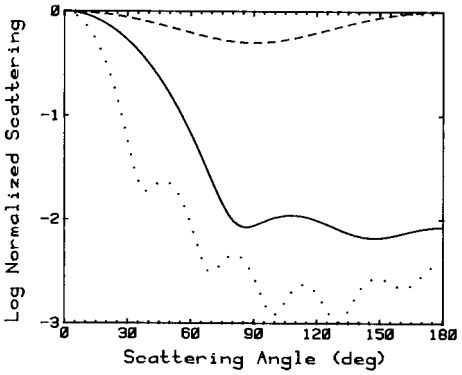

Fig. 9 Angular dependence of scattering of visible light (0.55 μm) by water droplets small compared with the wavelength (dashes), diameter 0.5 μm (solid curve), and diameter 10 μm (dots)

Many common observable phenomena depend on this forward-backward asymmetry. Viewed toward the illuminating sun, glistening fog droplets on a spider's web warn us of its presence. But when we view the web with our backs to the sun, the web mysteriously disappears. A pattern of dew illuminated by the rising sun on a cold morning seems etched on a windowpane. But if we go outside to look at the window, the pattern vanishes. Thin clouds sometimes hover over warm, moist heaps of dung, but may go unnoticed unless they lie between us and the source of illumination. These are but a few examples of the consequences of strongly asymmetric scattering by single particles comparable to or larger than the wavelength.

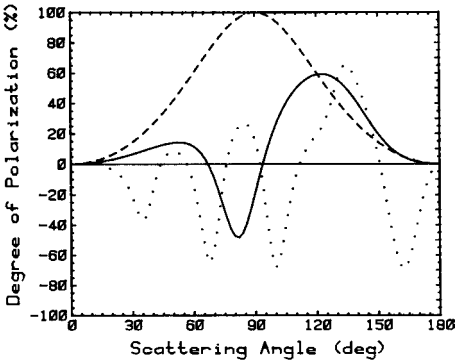

Fig. 10 Degree of polarization of light scattered by water droplets illuminated by unpolarized visible light (0.55 μm). The dashed curve is for a droplet small compared with the wavelength; the solid curve is for a droplet of diameter 0.5 μm; the dotted curve is for a droplet of diameter 1.0 μm. Negative degrees of polarization indicate that the scattered light is partially polarized parallel to the scattering plane

4.4

The Salient Differences between Particles and Molecules: Degree of Polarization of Scattered Light

All the simple rules about polarization upon scattering are broken when we turn from molecules and small particles to particles comparable to the wavelength. For example, the degree of polarization of light scattered by small particles is a simple function of scattering angle. But simplicity gives way to complexity as particles grow (Fig. 10), the scattered light being partially polarized parallel to the scattering plane for some scattering angles, perpendicular for others.

The degree of polarization of light scattered by molecules or by small particles is essentially independent of wavelength. But this is not true for particles comparable to or larger than the wavelength. Scattering by such particles exhibits *dispersion of polarization*: The degree of polarization at, say, 90° may vary considerably over the visible spectrum (Fig. 11).

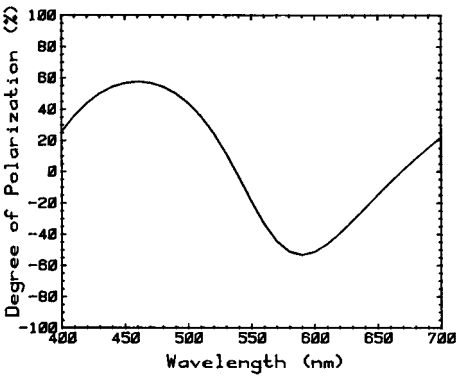

Fig. 11 Degree of polarization at a scattering angle of 90° of light scattered by a water droplet of diameter 0.5 μm illuminated by unpolarized light

In general, particles can act as polarizers or retarders or both. A polarizer transforms unpolarized light into partially polarized light. A retarder transforms polarized light of one form into that of another (e.g.,

linear into elliptical). Molecules and small particles, however, are restricted to roles as polarizers. If the atmosphere were inhabited solely by such scatterers, skylight could never be other than partially linearly polarized. Yet particles comparable to or larger than the wavelength often are present; hence skylight can acquire a degree of ellipticity upon multiple scattering: Incident unpolarized light is partially linearly polarized in the first scattering event, then transformed into partially elliptically polarized light in subsequent events.

Bees can navigate by polarized skylight. This statement, intended to evoke great awe for the photopolimetric powers of bees, is rarely accompanied by an important caveat: The sky must be clear. Figures 10 and 11 show two reasons – there are others – why bees, remarkable though they may be, cannot do the impossible. The simple wavelength-independent relation between the position of the sun and the direction in which skylight is most highly polarized, an underlying necessity for navigating by means of polarized skylight, is obliterated when clouds cover the sky. This was recognized by the decoder of bee dances himself von Frisch, [8]: ''Sometimes a cloud would pass across the area of sky visible through the tube; when this happened the dances became disoriented, and the bees were unable to indicate the direction to the feeding place. Whatever phenomenon in the blue sky served to orient the dances, this experiment showed that it was seriously disturbed if the blue sky was covered by a cloud.'' But von Frisch's words often have been forgotten by disciples eager to spread the story about bee magic to those just as eager to believe what is charming even though untrue.

4.5 The Salient Differences between Particles and Molecules: Vertical Distributions

Not only are the scattering properties of particles quite different, in general, from those of molecules; the different vertical distributions of particles and molecules by themselves affect what is observed. The number density of molecules decreases more or less exponentially with height z above the surface: $\exp(-z/H_m)$, where the molecular scale height H_m is around 8 km. Although the decrease in number density of particles with height is also approximately exponential, the scale height for particles H_p is about 1–2 km. As a consequence, particles contribute disproportionately to optical thicknesses along near-horizon paths. Subject to the approximations underlying Eq. (9), the ratio of the tangential (horizon) optical thickness for particles τ_{tp} to that for molecules τ_{tm} is

$$\frac{\tau_{tp}}{\tau_{tm}} = \frac{\tau_{np}}{\tau_{nm}}\sqrt{\frac{H_m}{H_p}}, \qquad (13)$$

where the subscript t indicates a tangential path and n indicates a normal (radial) path. Because of the incoherence of scattering by atmospheric molecules and particles, scattering coefficients are additive, and hence so are optical thicknesses. For equal normal optical thicknesses, the tangential optical thickness for particles is at least twice that for molecules. Molecules by themselves cannot give red sunrises and sunsets; molecules need the help of particles. For a fixed τ_{np}, the tangential optical thickness for particles is greater the more they are concentrated near the ground.

At the horizon the relative rate of change of transmission T of sunlight with zenith

angle is

$$\frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}\Theta} = \tau_n \frac{R_e}{H}, \qquad (14)$$

where the scale height and normal optical thickness may be those for molecules or particles. Not only do particles, being more concentrated near the surface, give disproportionate attenuation of sunlight on the horizon, but they magnify the angular gradient of attenuation there. A perceptible change in color across the sun's disk (which subtends about 0.5°) on the horizon also requires the help of particles.

Fig. 12 Because of scattering by molecules and particles along the line of sight, each successive ridge is brighter than the ones in front of it even though all of them are covered with the same dark vegetation

5 Atmospheric Visibility

On a clear day can we really see forever? If not, how far can we see? To answer this question requires qualifying it by restricting viewing to more or less horizontal paths during daylight. Stars at staggering distances can be seen at night, partly because there is no skylight to reduce contrast, partly because stars overhead are seen in directions for which attenuation by the atmosphere is least.

The radiance in the direction of a black object is not zero, because of light scattered along the line of sight (see Sec. 2.4). At sufficiently large distances, this airlight is indistinguishable from the horizon sky. An example is a phalanx of parallel dark ridges, each ridge less distinct than those in front of it (Fig. 12). The farthest ridges blend into the horizon sky. Beyond some distance we cannot see ridges because of insufficient contrast.

Equation (5) gives the airlight radiance, a radiometric quantity that describes radiant power without taking into account the portion of it that stimulates the human eye or by what relative amount it does so at each wavelength. Luminance (also sometimes called *brightness*) is the corresponding photometric quantity. Luminance and radiance are related by an integral over the visible spectrum:

$$B = \int K(\lambda) L(\lambda)\, \mathrm{d}\lambda, \qquad (15)$$

where the luminous efficiency of the human eye K peaks at about 550 nm and vanishes outside the range 385–760 nm.

The *contrast* C between any object and the horizon sky is

$$C = \frac{B - B_\infty}{B_\infty}, \qquad (16)$$

where B_∞ is the luminance for an infinite horizon optical thickness. For a uniformly illuminated line of sight of length d, uniform in its scattering properties, and

with a black backdrop, the contrast is

$$C = -\frac{\int KGL_0 \exp(-\beta d)\, d\lambda}{\int KGL_0\, d\lambda}. \qquad (17)$$

The ratio of integrals in this equation defines an average optical thickness:

$$C = -\exp(-\langle\tau\rangle). \qquad (18)$$

This expression for contrast reduction with (optical) distance is mathematically, but not physically, identical to Eq. (6), which perhaps has engendered the misconception that atmospheric visibility is reduced because of attenuation. Yet as there is no light from a black object to be attenuated, its finite visual range cannot be a consequence of attenuation.

The distance beyond which a dark object cannot be distinguished from the horizon sky is determined by the *contrast threshold*: the smallest contrast detectable by the human observer. Although this depends on the particular observer, the angular size of the object observed, the presence of nearby objects, and the absolute luminance, a contrast threshold of 0.02 is often taken as an average. This value in Eq. (18) gives

$$-\ln|C| = 3.9 = \langle\tau\rangle = \langle\beta d\rangle. \qquad (19)$$

To convert an optical distance into a physical distance requires the scattering coefficient. Because K is peaked at around 550 nm, we can obtain an approximate value of d from the scattering coefficient at this wavelength in Eq. (19). At sea level, the molecular scattering coefficient in the middle of the visible spectrum corresponds to about 330 km for "forever": the greatest distance at which a black object can be seen against the horizon sky assuming a contrast threshold of 0.02 and ignoring the curvature of the earth.

We also observe contrast between elements of the same scene, a hillside mottled with stands of trees and forest clearings, for example. The extent to which we can resolve details in such a scene depends on sun angle as well as distance.

The airlight radiance for a nonreflecting object is Eq. (5) with $G = p(\Theta)\Omega_s$, where $p(\Theta)$ is the probability (per unit solid angle) that light is scattered in a direction making an angle Θ with the incident sunlight and Ω_s is the solid angle subtended by the sun. When the sun is overhead, $\Theta = 90^\circ$; with the sun at the observer's back, $\Theta = 180^\circ$; for an observer looking directly into the sun, $\Theta = 0^\circ$.

The radiance of an object with a finite reflectance R and illuminated at an angle Φ is given by Eq. (11). Equations (5) and (11) can be combined to obtain the contrast between reflecting and nonreflecting objects:

$$C = \frac{F e^{-\tau}}{1 + (F-1)e^{-\tau}},$$
$$F = \frac{R\cos\Phi}{\pi\, p(\Theta)}. \qquad (20)$$

All else being equal, therefore, contrast decreases as $p(\Theta)$ increases. As shown in Fig. 9, $p(\Theta)$ is more sharply peaked in the forward direction the larger the scatterer. Thus, we expect the details of a distant scene to be less distinct when looking toward the sun than away from it if the optical thickness of the line of sight has an appreciable component contributed by particles comparable to or larger than the wavelength.

On humid, hazy days, visibility is often depressingly poor. Haze, however, is not water vapor but rather water that has ceased to be vapor. At high relative humidities, but still well below

100%, small soluble particles in the atmosphere accrete liquid water to become solution droplets (haze). Although these droplets are much smaller than cloud droplets, they markedly diminish visual range because of the sharp increase in scattering with particle size (Fig. 7). The same number of water molecules when aggregated in haze scatter vastly more than when apart.

6
Atmospheric Refraction

6.1
Physical Origins of Refraction

Atmospheric refraction is a consequence of molecular scattering, which is rarely stated given the historical accident that before light and matter were well understood refraction and scattering were locked in separate compartments and subsequently have been sequestered more rigidly than monks and nuns in neighboring cloisters.

Consider a beam of light propagating in an optically homogeneous medium. Light is scattered (weakly but observably) laterally to this beam as well as in the direction of the beam (the forward direction). The observed beam is a coherent superposition of incident light and forward-scattered light, which was excited by the incident light. Although refractive indices are often defined by ratios of phase velocities, we may also look upon a refractive index as a parameter that specifies the phase shift between an incident beam and the forward-scattered beam that the incident beam excites. The connection between (incoherent) scattering and refraction (coherent scattering) can be divined from the expressions for the refractive index n of a gas and the scattering cross section σ_s of a gas molecule:

$$n = 1 + \tfrac{1}{2}\alpha N, \tag{21}$$

$$\sigma_s = \frac{k^4}{6\pi}|\alpha|^2, \tag{22}$$

where N is the number density (not mass density) of gas molecules, $k = 2\pi/\lambda$ is the wave number of the incident light, and α is the polarizability of a molecule (induced dipole moment per unit inducing electric field). The appearance of the polarizability in Eq. (21) but its square in Eq. (22) is the clue that refraction is associated with electric fields whereas lateral scattering is associated with electric fields squared (powers). Scattering, without qualification, often means incoherent scattering in all directions. Refraction, in a nutshell, is coherent scattering in a particular direction.

Readers whose appetites have been whetted by the preceding brief discussion of the physical origins of refraction are directed to a beautiful paper by Doyle [9] in which he shows how the Fresnel equations can be dissected to reveal the scattering origins of (specular) reflection and refraction.

6.2
Terrestrial Mirages

Mirages are not illusions, any more so than are reflections in a pond. Reflections of plants growing at its edge are not interpreted as plants growing into the water. If the water is ruffled by wind, the reflected images may be so distorted that they are no longer recognizable as those of plants. Yet we still would not call such distorted images illusions. And so is it with mirages. They are images noticeably different from what they would be in the absence of atmospheric

refraction, creations of the atmosphere, not of the mind.

Mirages are vastly more common than is realized. Look and you shall see them. Contrary to popular opinion, they are not unique to deserts. Mirages can be seen frequently even over ice-covered landscapes and highways flanked by deep snowbanks. Temperature *per se* is not what gives mirages but rather temperature gradients.

Because air is a mixture of gases, the polarizability for air in Eq. (21) is an average over all its molecular constituents, although their individual polarizabilities are about the same (at visible wavelengths). The vertical refractive index gradient can be written so as to show its dependence on pressure p and (absolute) temperature T:

$$\frac{d}{dz}\ln(n-1) = \frac{1}{p}\frac{dp}{dz} - \frac{1}{T}\frac{dT}{dz}. \qquad (23)$$

Pressure decreases approximately exponentially with height, where the scale height is around 8 km. Thus, the first term on the right-hand side of Eq. (23) is around 0.1 km^{-1}. Temperature usually decreases with height in the atmosphere. An average lapse rate of temperature (i.e., its decrease with height) is around 6 °C/km. The average temperature in the troposphere (within about 15 km of the surface) is around 280 K. Thus, the magnitude of the second term in Eq. (23) is around 0.02 km^{-1}. On average, therefore, the refractive-index gradient is dominated by the vertical pressure gradient. But within a few meters of the surface, conditions are far from average. On a sun-baked highway your feet may be touching asphalt at 50 °C while your nose is breathing air at 35 °C, which corresponds to a lapse rate a thousand times the average. Moreover, near the surface, temperature can increase with height. In shallow surface layers, in which the pressure is nearly constant, the temperature gradient determines the refractive index gradient. It is in such shallow layers that mirages, which are caused by refractive-index gradients, are seen.

Cartoonists by their fertile imaginations unfettered by science, and textbook writers by their carelessness, have engendered the notion that atmospheric refraction can work wonders, lifting images of ships, for example, from the sea high into the sky. A back-of-the-envelope calculation dispels such notions. The refractive index of air at sea level is about 1.0003 (Fig. 13). Light from empty space incident at glancing incidence onto a uniform slab with this refractive index is displaced in angular position from where it would have been in the absence of refraction by

$$\delta = \sqrt{2(n-1)}. \qquad (24)$$

This yields an angular displacement of about 1.4°, which as we shall see is a rough upper limit.

Trajectories of light rays in nonuniform media can be expressed in different ways. According to Fermat's principle of least

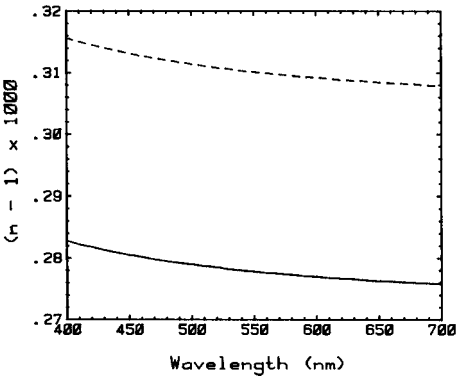

Fig. 13 Sea-level refractive index versus wavelength at −15 °C (dashes) and 15 °C (solid curve). Data from Penndorf, R. (**1957**), *J. Opt. Soc. Am.* **47**, 176–182 [2]

time (which ought to be extreme time), the actual path taken by a ray between two points is such that the path integral

$$\int_1^2 n\,\mathrm{d}s \tag{25}$$

is an extremum over all possible paths. This principle has inspired piffle about the alleged efficiency of nature, which directs light over routes that minimize travel time, presumably freeing it to tend to important business at its destination.

The scale of mirages is such that in analyzing them we may pretend that the Earth is flat. On such an earth, with an atmosphere in which the refractive index varies only in the vertical, Fermat's principle yields a generalization

$$n\sin\theta = \text{constant} \tag{26}$$

of Snel's law, where θ is the angle between the ray and the vertical direction. We could, of course, have bypassed Fermat's principle to obtain this result.

If we restrict ourselves to nearly horizontal rays, Eq. (26) yields the following differential equation satisfied by a ray:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} = \frac{\mathrm{d}n}{\mathrm{d}z}, \tag{27}$$

where y and z are its horizontal and vertical coordinates, respectively. For a constant refractive-index gradient, which to good approximation occurs for a constant temperature gradient, Eq. (27) yields parabolas for ray trajectories. One such parabola for a constant temperature gradient about 100 times the average is shown in Fig. 14. Note the vastly different horizontal and vertical scales. The image is displaced downward from what it would be in the absence of atmospheric refraction; hence the designation *inferior* mirage. This is the familiar highway mirage, seen over highways warmer than the air above them. The downward angular displacement is

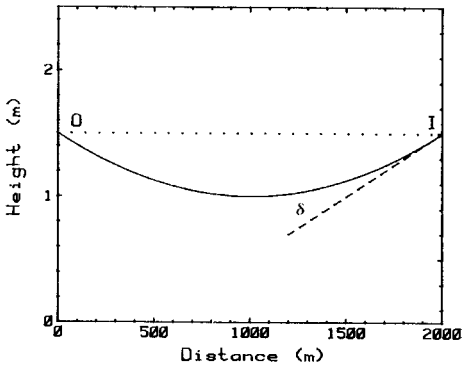

Fig. 14 Parabolic ray paths in an atmosphere with a constant refractive-index gradient (inferior mirage). Note the vastly different horizontal and vertical scales

$$\delta = \frac{1}{2} s \frac{\mathrm{d}n}{\mathrm{d}z}, \tag{28}$$

where s is the horizontal distance between object and observer (image). Even for a temperature gradient 1000 times the tropospheric average, displacements of mirages are less than a degree at distances of a few kilometers.

If temperature increases with height, as it does, for example, in air over a cold sea, the resulting mirage is called a *superior mirage*. Inferior and superior are not designations of lower and higher caste but rather of displacements downward and upward.

For a constant temperature gradient, one and only one parabolic ray trajectory connects an object point to an image point. Multiple images therefore are not possible. But temperature gradients close to the ground are rarely linear. The upward transport of energy from a hot surface occurs by molecular conduction through a stagnant boundary

layer of air. Somewhat above the surface, however, energy is transported by air in motion. As a consequence, the temperature gradient steepens toward the ground if the energy flux is constant. This variable gradient can lead to two observable consequences: magnification and multiple images.

According to Eq. (28), all image points at a given horizontal distance are displaced downward by an amount proportional to the (constant) refractive index gradient. A corollary is that the closer an object point is to a surface, where the temperature gradient is greatest, the greater the downward displacement of the corresponding image point. Thus, nonlinear vertical temperature profiles may magnify images.

Multiple images are seen frequently on highways. What often appears to be water on the highway ahead but evaporates before it is reached is the inverted secondary image of either the horizon sky or horizon objects lighter than dark asphalt.

6.3 Extraterrestrial Mirages

When we turn from mirages of terrestrial objects to those of extraterrestrial bodies, most notably the sun and moon, we can no longer pretend that the Earth is flat. But we can pretend that the atmosphere is uniform and bounded. The total phase shift of a vertical ray from the surface to infinity is the same in an atmosphere with an exponentially decreasing molecular number density as in a hypothetical atmosphere with a uniform number density equal to the surface value up to height H.

A ray refracted along a horizon path by this hypothetical atmosphere and originating from outside it had to have been incident on it from an angle δ below the horizon:

$$\delta = \sqrt{\frac{2H}{R}} - \sqrt{\frac{2H}{R} - 2(n-1)}, \qquad (29)$$

where R is the radius of the Earth. Thus, when the sun (or moon) is seen to be on the horizon it is actually more than halfway below it, δ being about 0.36°, whereas the angular width of the sun (or moon) is about 0.5°.

Extraterrestrial bodies seen near the horizon also are vertically compressed. The simplest way to estimate the amount of compression is from the rate of change of angle of refraction θ_r with angle of incidence θ_i for a uniform slab

$$\frac{d\theta_r}{d\theta_i} = \frac{\cos\theta_i}{\sqrt{n^2 - \sin^2\theta_i}}, \qquad (30)$$

where the angle of incidence is that for a curved but uniform atmosphere such that the refracted ray is horizontal. The result is

$$\frac{d\theta_r}{d\theta_i} = \sqrt{1 - \frac{R}{H}(n-1)}, \qquad (31)$$

according to which the sun near the horizon is distorted into an ellipse with aspect ratio about 0.87. We are unlikely to notice this distortion, however, because we expect the sun and moon to be circular, and hence we see them that way.

The previous conclusions about the downward displacement and distortion of the sun were based on a refractive-index profile determined mostly by the vertical pressure gradient. Near the ground, however, the temperature gradient is the prime determinant of the refractive-index gradient, as a consequence of which the sun on the horizon can take on shapes

Fig. 15 A nearly triangular sun on the horizon. The serrations are a consequence of horizontal variations in refractive index

more striking than a mere ellipse. For example, Fig. 15 shows a nearly triangular sun with serrated edges. Assigning a cause to these serrations provides a lesson in the perils of jumping to conclusions. Obviously, the serrations are the result of sharp changes in the temperature gradient – or so one might think. Setting aside how such changes could be produced and maintained in a real atmosphere, a theorem of Fraser [10] gives pause for thought. According to this theorem, ''In a horizontally (spherically) homogeneous atmosphere it is impossible for more than one image of an extraterrestrial object (sun) to be seen above the astronomical horizon.'' The serrations on the sun in Fig. 15 are multiple images. But if the refractive index varies only vertically (i.e., along a radius), no matter how sharply, multiple images are not possible. Thus, the serrations must owe their existence to horizontal variations of the refractive index, a consequence of gravity waves propagating along a temperature inversion.

6.4 The Green Flash

Compared to the rainbow, the green flash is not a rare phenomenon. Before you dismiss this assertion as the ravings of a lunatic, consider that rainbows require raindrops as well as sunlight to illuminate them, whereas rainclouds often completely obscure the sun. Moreover, the sun must be below about 42°. As a consequence of these conditions, rainbows are not seen often, but often enough that they are taken as the paragon of color variation. Yet tinges of green on the upper rim of the sun can be seen every day at sunrise and sunset given a sufficiently low horizon and a cloudless sky. Thus, the conditions for seeing a green flash are more easily met than those for seeing a rainbow. Why then is the green flash considered to be so rare? The distinction here is that between a rarely observed phenomenon (the green flash) and a rarely observable one (the rainbow).

The sun may be considered to be a collection of disks, one for each visible wavelength. When the sun is overhead, all the disks coincide and we see the sun as white. But as it descends in the sky, atmospheric refraction displaces the disks by slightly different amounts, the red less than the violet (see Fig. 13). Most of each disk overlaps all the others except for the disks at the extremes of the visible spectrum. As a consequence, the upper rim of the sun is violet or blue, its lower rim red, whereas its interior, the region in which all disks overlap, is still white.

This is what would happen in the absence of lateral scattering of sunlight. But refraction and lateral scattering go hand in hand, even in an atmosphere free of particles. Selective scattering by atmospheric molecules and particles causes the color

of the sun to change. In particular, the violet-bluish upper rim of the low sun can be transformed to green.

According to Eq. (29) and Fig. 13, the angular width of the green upper rim of the low sun is about 0.01°, too narrow to be resolved with the naked eye or even to be seen against its bright backdrop. But depending on the temperature profile, the atmosphere itself can magnify the upper rim and yield a second image of it, thereby enabling it to be seen without the aid of a telescope or binoculars. Green rims, which require artificial magnification, can be seen more frequently than green flashes, which require natural magnification. Yet both can be seen often by those who know what to look for and are willing to look.

7
Scattering by Single Water Droplets

All the colored atmospheric displays that result when water droplets (or ice crystals) are illuminated by sunlight have the same underlying cause: light is scattered in different amounts in different directions by particles larger than the wavelength, and the directions in which scattering is greatest depends on wavelength. Thus, when particles are illuminated by white light, the result can be angular separation of colors even if scattering integrated over all directions is independent of wavelength (as it essentially is for cloud droplets and ice crystals). This description, although correct, is too general to be completely satisfying. We need something more specific, more quantitative, which requires theories of scattering.

Because superficially different theories have been used to describe different optical phenomena, the notion has become widespread that they are caused by these theories. For example, coronas are said to be caused by diffraction and rainbows by refraction. Yet both the corona and the rainbow can be described quantitatively to high accuracy with a theory (the Mie theory for scattering by a sphere) in which diffraction and refraction do not explicitly appear. No fundamentally impenetrable barrier separates scattering from (specular) reflection, refraction, and diffraction. Because these terms came into general use and were entombed in textbooks before the nature of light and matter was well understood, we are stuck with them. But if we insist that diffraction, for example, is somehow different from scattering, we do so at the expense of shattering the unity of the seemingly disparate observable phenomena that result when light interacts with matter. What is observed depends on the composition and disposition of the matter, not on which approximate theory in a hierarchy is used for quantitative description.

Atmospheric optical phenomena are best classified by the direction in which they are seen and by the agents responsible for them. Accordingly, the following sections are arranged in order of scattering direction, from forward to backward.

When a single water droplet is illuminated by white light and the scattered light projected onto a screen, the result is a set of colored rings. But in the atmosphere we see a mosaic to which individual droplets contribute. The scattering pattern of a single droplet is the same as the mosaic provided that multiple scattering is negligible.

7.1
Coronas and Iridescent Clouds

A cloud of droplets narrowly distributed in size and thinly veiling the sun (or moon)

can yield a spectacular series of colored concentric rings around it. This corona is most easily described quantitatively by the Fraunhofer diffraction theory, a simple approximation valid for particles large compared with the wavelength and for scattering angles near the forward direction. According to this approximation, the differential scattering cross section (cross section for scattering into a unit solid angle) of a spherical droplet of radius a illuminated by light of wave number k is

$$\frac{|S|^2}{k^2}, \tag{32}$$

where the *scattering amplitude* is

$$S = x^2 \frac{1+\cos\theta}{2} \frac{J_1(x\sin\theta)}{x\sin\theta}. \tag{33}$$

The term J_1 is the Bessel function of first order and the size parameter $x = ka$. The quantity $(1+\cos\theta)/2$ is usually approximated by 1 since only near-forward scattering angles θ are of interest.

The differential scattering cross section, which determines the angular distribution of the scattered light, has maxima for $x\sin\theta = 5.137,\ 8.417,\ 11.62,\ \ldots$ Thus, the dispersion in the position of the first maximum is

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} \approx \frac{0.817}{a} \tag{34}$$

and is greater for higher-order maxima. This dispersion determines the upper limit on drop size such that a corona can be observed. For the total angular dispersion over the visible spectrum to be greater than the angular width of the sun (0.5°), the droplets cannot be larger than about 60 μm in diameter. Drops in rain, even in drizzle, are appreciably larger than this, which is why coronas are not seen through rainshafts. Scattering by a droplet of diameter 10 μm (Fig. 16), a typical cloud droplet size, gives sufficient dispersion to yield colored coronas.

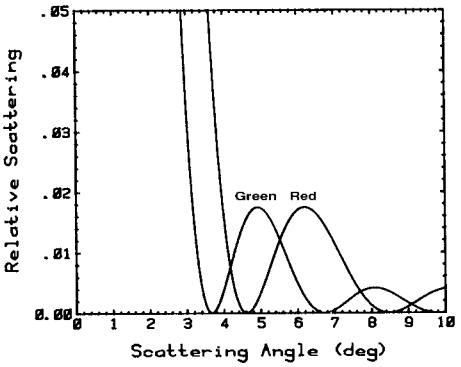

Fig. 16 Scattering of light near the forward direction (according to Fraunhofer theory) by a sphere of diameter 10 μm illuminated by red and green light

Suppose that the first angular maximum for blue light (0.47 μm) occurs for a droplet of radius a. For red light (0.66 μm) a maximum is obtained at the same angle for a droplet of radius $a + \Delta a$. That is, the two maxima, one for each wavelength, coincide. From this we conclude that coronas require narrow size distributions: if cloud droplets are distributed in radius with a relative variance $\Delta a/a$ greater than about 0.4, color separation is not possible.

Because of the stringent requirements for the occurrence of coronas, they are not observed often. Of greater occurrence are the corona's cousins, iridescent clouds, which display colors but usually not arranged in any obviously regular geometrical pattern. Iridescent patches in clouds can be seen even at the edges of thick clouds that occult the sun.

Coronas are not the unique signatures of spherical scatterers. Randomly oriented ice columns and plates give similar patterns according to Fraunhofer theory [11]. As a practical matter, however, most coronas

probably are caused by droplets. Many clouds at temperatures well below freezing contain subcooled water droplets. Only if a corona were seen in a cloud at a temperature lower than $-40\,^{\circ}\mathrm{C}$ could one assert with confidence that it must be an ice-crystal corona.

7.2 Rainbows

In contrast with coronas, which are seen looking toward the sun, rainbows are seen looking away from it, and are caused by water drops much larger than those that give coronas. To treat the rainbow quantitatively we may pretend that light incident on a transparent sphere is composed of individual rays, each of which suffers a different fate determined only by the laws of specular reflection and refraction. Theoretical justification for this is provided by van de Hulst's ([12], p. 208) *localization principle*, according to which terms in the exact solution for scattering by a transparent sphere correspond to more or less localized rays.

Each incident ray splinters into an infinite number of scattered rays: externally reflected, transmitted without internal reflection, transmitted after one, two, and so on internal reflections. At any scattering angle θ, each splinter contributes to the scattered light. Accordingly, the differential scattering cross section is an infinite series with terms of the form

$$\frac{b(\theta)}{\sin\theta}\frac{\mathrm{d}b}{\mathrm{d}\theta}. \tag{35}$$

The *impact parameter* b is $a\sin\Theta_i$, where Θ_i is the angle between an incident ray and the normal to the sphere. Each term in the series corresponds to one of the splinters of an incident ray. A *rainbow angle* is a singularity (or *caustic*) of the differential scattering cross section at which the conditions

$$\frac{\mathrm{d}\theta}{\mathrm{d}b} = 0, \quad \frac{b}{\sin\theta} \neq 0 \tag{36}$$

are satisfied. Missing from Eq. (35) are various reflection and transmission coefficients (Fresnel coefficients), which display no singularities and hence do not determine rainbow angles.

A rainbow is not associated with rays externally reflected or transmitted without internal reflection. The succession of rainbow angles associated with one, two, three ... internal reflections are called *primary, secondary, tertiary* ... rainbows. Aristotle recognized that "Three or more rainbows are never seen, because even the second is dimmer than the first, and so the third reflection is altogether too feeble to reach the sun (Aristotle's view was that light streams outward from the eye)". Although he intuitively grasped that each successive ray is associated with ever-diminishing energy, his statement about the nonexistence of tertiary rainbows in nature is not quite true. Although reliable reports of such rainbows are rare (unreliable reports are as common as dirt), at least one observer who can be believed has seen one [13].

An incident ray undergoes a total angular deviation as a consequence of transmission into the drop, one or more internal reflections, and transmission out of the drop. Rainbow angles are angles of minimum deviation.

For a rainbow of any order to exist,

$$\cos\Theta_i = \sqrt{\frac{n^2-1}{p(p+1)}} \tag{37}$$

must lie between 0 and 1, where Θ_i is the angle of incidence of a ray that gives a rainbow after p internal reflections and

n is the refractive index of the drop. A primary bow therefore requires drops with refractive index less than 2; a secondary bow requires drops with refractive index less than 3. If raindrops were composed of titanium dioxide ($n \approx 3$), a commonly used opacifier for paints, primary rainbows would be absent from the sky and we would have to be content with only secondary bows.

If we take the refractive index of water to be 1.33, the scattering angle for the primary rainbow is about 138°. This is measured from the forward direction (solar point). Measured from the antisolar point (the direction toward which one must look in order to see rainbows in nature), this scattering angle corresponds to 42°, the basis for a previous assertion that rainbows (strictly, primary rainbows) cannot be seen when the sun is above 42°. The secondary rainbow is seen at about 51° from the antisolar point. Between these two rainbows is *Alexander's dark band*, a region into which no light is scattered according to geometrical optics.

The colors of rainbows are a consequence of sufficient dispersion of the refractive index over the visible spectrum to give a spread of rainbow angles that appreciably exceeds the width of the sun. The width of the primary bow from violet to red is about 1.7°; that of the secondary bow is about 3.1°.

Because of its band of colors arcing across the sky, the rainbow has become the paragon of color, the standard against which all other colors are compared. Lee and Fraser [14, 15], however, challenged this status of the rainbow, pointing out that even the most vivid rainbows are colorimetrically far from pure.

Rainbows are almost invariably discussed as if they occurred literally in a vacuum. But real rainbows, as opposed to the pencil-and-paper variety, are necessarily observed in an atmosphere, the molecules and particles of which scatter sunlight that adds to the light from the rainbow but subtracts from its purity of color.

Although geometrical optics yields the positions, widths, and color separation of rainbows, it yields little else. For example, geometrical optics is blind to *supernumerary bows*, a series of narrow bands sometimes seen below the primary bow. These bows are a consequence of interference, and hence fall outside the province of geometrical optics. Since supernumerary bows are an interference phenomenon, they, unlike primary and secondary bows (according to geometrical optics), depend on drop size. This poses the question of how supernumerary bows can be seen in rain showers, the drops in which are widely distributed in size. In a nice piece of detective work, Fraser [16] answered this question.

Raindrops falling in a vacuum are spherical. Those falling in air are distorted by aerodynamic forces, not, despite the depictions of countless artists, into teardrops but rather into nearly oblate spheroids with their axes more or less vertical. Fraser argued that supernumerary bows are caused by drops with a diameter of about 0.5 mm, at which diameter the angular position of the first (and second) supernumerary bow has a minimum: interference causes the position of the supernumerary bow to increase with decreasing size whereas drop distortion causes it to increase with increasing size. Supernumerary patterns contributed by drops on either side of the minimum cancel, leaving only the contribution from drops at the minimum. This cancellation occurs only near the tops of rainbows, where supernumerary bows are seen. In the vertical parts of a rainbow, a

horizontal slice through a distorted drop is more or less circular, and hence these drops do not exhibit a minimum supernumerary angle.

According to geometrical optics, all spherical drops, regardless of size, yield the same rainbow. But it is not necessary for a drop to be spherical for it to yield rainbows independent of its size. This merely requires that the plane defined by the incident and scattered rays intersect the drop in a circle. Even distorted drops satisfy this condition in the vertical part of a bow. As a consequence, the absence of supernumerary bows there is compensated for by more vivid colors of the primary and secondary bows [17]. Smaller drops are more likely to be spherical, but the smaller a drop, the less light it scatters. Thus, the dominant contribution to the luminance of rainbows is from the larger drops. At the top of a bow, the plane defined by the incident and scattered rays intersects the large, distorted drops in an ellipse, yielding a range of rainbow angles varying with the amount of distortion, and hence a pastel rainbow. To the knowledgeable observer, rainbows are no more uniform in color and brightness than is the sky.

Although geometrical optics predicts that all rainbows are equal (neglecting background light), real rainbows do not slavishly follow the dictates of this approximate theory. Rainbows in nature range from nearly colorless fog bows (or cloud bows) to the vividly colorful vertical portions of rainbows likely to have inspired myths about pots of gold.

7.3 The Glory

Continuing our sweep of scattering directions, from forward to backward, we arrive at the end of our journey: *the glory*. Because it is most easily seen from airplanes it sometimes is called the *pilot's bow*. Another name is *anticorona*, which signals that it is a corona around the antisolar point. Although glories and coronas share some common characteristics, there are differences between them other than direction of observation. Unlike coronas, which may be caused by nonspherical ice crystals, glories require spherical cloud droplets. And a greater number of colored rings may be seen in glories than in coronas because the decrease in luminance away from the backward direction is not as steep as that away from the forward direction. To see a glory from an airplane, look for colored rings around its shadow cast on clouds below. This shadow is not an essential part of the glory, it merely directs you to the antisolar point.

Like the rainbow, the glory may be looked upon as a singularity in the differential scattering cross section Eq. (35). Equation (36) gives one set of conditions for a singularity; the second set is

$$\sin\theta = 0, \quad b(\theta) \neq 0. \tag{38}$$

That is, the differential scattering cross section is infinite for nonzero impact parameters (corresponding to incident rays that do not intersect the center of the sphere) that give forward (0°) or backward (180°) scattering. The forward direction is excluded because this is the direction of intense scattering accounted for by the Fraunhofer theory.

For one internal reflection, Eq. (38) leads to the condition

$$\sin\Theta_i = \frac{n}{2}\sqrt{4 - n^2}, \tag{39}$$

which is satisfied only for refractive indices between 1.414 and 2, the lower refractive index corresponding to a grazing-incidence

ray. The refractive index of water lies outside this range. Although a condition similar to Eq. (39) is satisfied for rays undergoing four or more internal reflections, insufficient energy is associated with such rays. Thus, it seems that we have reached an impasse: the theoretical condition for a glory cannot be met by water droplets. Not so, says van de Hulst [18] in a seminal paper. He argues that 1.414 is close enough to 1.33 given that geometrical optics is, after all, an approximation. Cloud droplets are large compared with the wavelength, but not so large that geometrical optics is an infallible guide to their optical behavior. Support for the van de Hulstian interpretation of glories was provided by Bryant and Cox [19], who showed that the dominant contribution to the glory is from the last terms in the exact series for scattering by a sphere. Each successive term in this series is associated with ever larger impact parameters. Thus, the terms that give the glory are indeed those corresponding to grazing rays. Further unraveling of the glory and vindication of van de Hulst's conjectures about the glory were provided by Nussenzveig [20].

It sometimes is asserted that geometrical optics is incapable of treating the glory. Yet the same can be said for the rainbow. Geometrical optics explains rainbows only in the sense that it predicts singularities for scattering in certain directions (rainbow angles). But it can predict only the angles of intense scattering, not the amount. Indeed, the error is infinite. Geometrical optics also predicts a singularity in the backward direction. Again, this simple theory is powerless to predict more. Results from geometrical optics for both rainbows and glories are not the end but rather the beginning, an invitation to take a closer look with more powerful magnifying glasses.

8
Scattering by Single Ice Crystals

Scattering by spherical water drops in the atmosphere gives rise to three distinct displays in the sky: coronas, rainbows, and glories. Ice particles (crystals) also can inhabit the atmosphere, and they introduce two new variables in addition to size: shape and orientation, the second a consequence of the first. Given this increase in the number of degrees of freedom, it is hardly cause for wonder that ice crystals are the source of a greater variety of displays than are water drops. As with rainbows, the gross features of ice-crystal phenomena can be described simply with geometrical optics, various phenomena arising from the various fates of rays incident on crystals. Colorless displays (e.g., sun pillars) are generally associated with reflected rays, colored displays (e.g., sun dogs and halos) with refracted rays. Because of the wealth of ice-crystal displays, it is not possible to treat all of them here, but one example should point the way toward understanding many of them.

8.1
Sun Dogs and Halos

Because of the hexagonal crystalline structure of ice it can form as hexagonal plates in the atmosphere. The stable position of a plate falling in air is with the normal to its face more or less vertical, which is easy to demonstrate with an ordinary business card. When the card is dropped with its edge facing downward (the supposedly aerodynamic position that many people instinctively choose), the card somersaults in a helter-skelter path to the ground. But when the card is dropped with its face parallel to the ground, it rocks back and forth gently in descent.

A hexagonal ice plate falling through air and illuminated by a low sun is like a 60° prism illuminated normally to its sides (Fig. 17). Because there is no mechanism for orienting a plate within the horizontal plane, all plate orientations in this plane are equally probable. Stated another way, all angles of incidence for a fixed plate are equally probable. Yet all scattering angles (deviation angles) of rays refracted into and out of the plate are not equally probable.

Figure 18 shows the range of scattering angles corresponding to a range of rays incident on a 60° ice prism that is part of a hexagonal plate. For angles of incidence less than about 13°, the transmitted ray is totally internally reflected in the prism. For angles of incidence greater than about 70°, the transmittance plunges. Thus, the only rays of consequence are those incident between about 13° and 70°.

All scattering angles are not equally probable. The (uniform) probability distribution $p(\theta_i)$ of incidence angles θ_i is related to the probability distribution

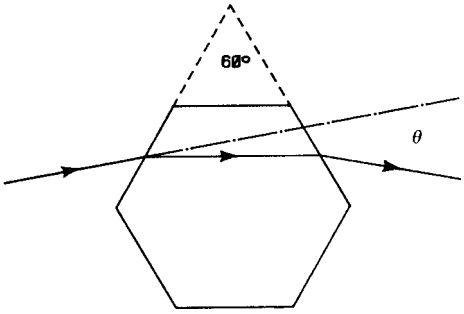

Fig. 17 Scattering by a hexagonal ice plate illuminated by light parallel to its basal plane. The particular scattering angle θ shown is an angle of minimum deviation. The scattered light is that associated with two refractions by the plate

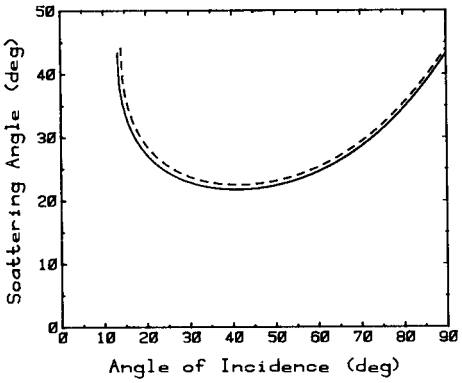

Fig. 18 Scattering by a hexagonal ice plate (see Fig. 17) in various orientations (angles of incidence). The solid curve is for red light, the dashed for blue light

$P(\theta)$ of scattering angles θ by

$$P(\theta) = \frac{p(\theta_i)}{\mathrm{d}\theta / \mathrm{d}\theta_i}. \qquad (40)$$

At the incidence angle for which $\mathrm{d}\theta/\mathrm{d}\theta_i = 0$, $P(\theta)$ is infinite and scattered rays are intensely concentrated near the corresponding angle of minimum deviation.

The physical manifestation of this singularity (or caustic) at the angle of minimum deviation for a 60° hexagonal ice plate is a bright spot about 22° from either or both sides of a sun low in the sky. These bright spots are called *sun dogs* (because they accompany the sun) or *parhelia* or *mock suns*.

The angle of minimum deviation θ_m, hence the angular position of sun dogs, depends on the prism angle Δ (60° for the plates considered) and refractive index:

$$\theta_m = 2\sin^{-1}\left(n\sin\frac{\Delta}{2}\right) - \Delta. \qquad (41)$$

Because ice is dispersive, the separation between the angles of minimum deviation for red and blue light is about 0.7° (Fig. 18),

somewhat greater than the angular width of the sun. As a consequence, sun dogs may be tinged with color, most noticeably toward the sun. Because the refractive index of ice is least at the red end of the spectrum, the red component of a sun dog is closest to the sun. Moreover, light of any two wavelengths has the same scattering angle for different angles of incidence if one of the wavelengths does not correspond to red. Thus, red is the purest color seen in a sun dog. Away from its red inner edge a sun dog fades into whiteness.

With increasing solar elevation, sun dogs move away from the sun. A falling ice plate is roughly equivalent to a prism, the prism angle of which increases with solar elevation. From Eq. (41) it follows that the angle of minimum deviation, hence the sun dog position, also increases.

At this point you may be wondering why only the 60° prism portion of a hexagonal plate was singled out for attention. As evident from Fig. 17, a hexagonal plate could be considered to be made up of 120° prisms. For a ray to be refracted twice, its angle of incidence at the second interface must be less than the critical angle. This imposes limitations on the prism angle. For a refractive index 1.31, all incident rays are totally internally reflected by prisms with angles greater than about 99.5°.

A close relative of the sun dog is the 22° halo, a ring of light approximately 22° from the sun (Fig. 19). Lunar halos are also possible and are observed frequently (although less frequently than solar halos); even moon dogs are possible. Until Fraser [21] analyzed halos in detail, the conventional wisdom had been that they obviously were the result of randomly oriented crystals, yet another example of jumping to conclusions. By combining optics and aerodynamics, Fraser showed that if ice crystals are small enough to be randomly oriented by Brownian motion, they are too small to yield sharp scattering patterns.

But completely randomly oriented plates are not necessary to give halos, especially ones of nonuniform brightness. Each part of a halo is contributed to by plates with a different tip angle (angle between the normal to the plate and the vertical). The transition from oriented plates (zero tip angle) to randomly oriented plates occurs over a narrow range of sizes. In the transition region, plates can be small enough to be partially oriented yet large enough to give a distinct contribution to the halo. Moreover, the mapping between tip angles and azimuthal angles on the halo depends on solar elevation. When the sun is near the horizon, plates can give a distinct halo over much of its azimuth.

Fig. 19 A 22° solar halo. The hand is not for artistic effect but rather to occlude the bright sun

When the sun is high in the sky, hexagonal plates cannot give a sharp halo but hexagonal columns – another possible form of atmospheric ice particles – can. The stable position of a falling column is with its long axis horizontal. When the sun is directly overhead, such columns can give a uniform halo even if they all lie in the horizontal plane. When the sun is not overhead but well above the horizon, columns also can give halos.

A corollary of Fraser's analysis is that halos are caused by crystals with a range of sizes between about 12 and 40 μm. Larger crystals are oriented; smaller particles are too small to yield distinct scattering patterns.

More or less uniformly bright halos with the sun neither high nor low in the sky could be caused by mixtures of hexagonal plates and columns or by clusters of bullets (rosettes). Fraser opines that the latter is more likely.

One of the by-products of his analysis is an understanding of the relative rarity of the 46° halo. As we have seen, the angle of minimum deviation depends on the prism angle. Light can be incident on a hexagonal column such that the prism angle is 60° for rays incident on its side or 90° for rays incident on its end. For $n = 1.31$, Eq. (41) yields a minimum deviation angle of about 46° for $\Delta = 90°$. Yet, although 46° halos are possible, they are seen much less frequently than 22° halos. Plates cannot give distinct 46° halos although columns can. Yet they must be solid and most columns have hollow ends. Moreover, the range of sun elevations is restricted.

Like the green flash, ice-crystal phenomena are not intrinsically rare. Halos and sun dogs can be seen frequently – once you know what to look for. Neuberger [22] reports that halos were observed in State College, Pennsylvania, an average of 74 days a year over a 16-year period, with extremes of 29 and 152 halos a year. Although the 22° halo was by far the most frequently seen display, ice-crystal displays of all kinds were seen, on average, more often than once every four days at a location not especially blessed with clear skies. Although thin clouds are necessary for ice-crystal displays, clouds thick enough to obscure the sun are their bane.

9
Clouds

Although scattering by isolated particles can be studied in the laboratory, particles in the atmosphere occur in crowds (sometimes called *clouds*). Implicit in the previous two sections is the assumption that each particle is illuminated solely by incident sunlight; the particles do not illuminate each other to an appreciable degree. That is, clouds of water droplets or ice grains were assumed to be optically thin, and hence multiple scattering was negligible. Yet the term cloud evokes fluffy white objects in the sky, or perhaps an overcast sky on a gloomy day. For such clouds, multiple scattering is not negligible, it is the major determinant of their appearance. And the quantity that determines the degree of multiple scattering is optical thickness (see Sec. 2.4).

9.1
Cloud Optical Thickness

Despite their sometimes solid appearance, clouds are so flimsy as to be almost nonexistent – except optically. The fraction of the total cloud volume occupied by water substance (liquid or solid) is about 10^{-6} or less. Yet although the mass density of clouds is that of air to within

a small fraction of a percent, their optical thickness (per unit physical thickness) is much greater. The number density of air molecules is vastly greater than that of water droplets in clouds, but scattering per molecule of a cloud droplet is also much greater than scattering per air molecule (see Fig. 7).

Because a typical cloud droplet is much larger than the wavelengths of visible light, its scattering cross section is to good approximation proportional to the square of its diameter. As a consequence, the scattering coefficient [see Eq. (2)] of a cloud having a volume fraction f of droplets is approximately

$$\beta = 3f\frac{\langle d^2\rangle}{\langle d^3\rangle}, \tag{42}$$

where the brackets indicate an average over the distribution of droplet diameters d. Unlike molecules, cloud droplets are distributed in size. Although cloud particles can be ice particles as well as water droplets, none of the results in this and the following section hinge on the assumption of spherical particles.

The optical thickness along a cloud path of physical thickness h is βh for a cloud with uniform properties. The ratio $\langle d^3\rangle/\langle d^2\rangle$ defines a mean droplet diameter, a typical value for which is 10 μm. For this diameter and $f = 10^{-6}$, the optical thickness per unit meter of physical thickness is about the same as the normal optical thickness of the atmosphere in the middle of the visible spectrum (see Fig. 3). Thus, a cloud only 1 m thick is equivalent optically to the entire gaseous atmosphere.

A cloud with (normal) optical thickness about 10 (i.e., a physical thickness of about 100 m) is sufficient to obscure the disk of the sun. But even the thickest cloud does not transform day into night. Clouds are usually translucent, not transparent, yet not completely opaque.

The scattering coefficient of cloud droplets, in contrast with that of air molecules, is more or less independent of wavelength. This is often invoked as the cause of the colorlessness of clouds. Yet wavelength independence of scattering by a single particle is only sufficient, not necessary, for wavelength independence of scattering by a cloud of particles (see Sec. 2.4). Any cloud that is optically thick and composed of particles for which absorption is negligible is white upon illumination by white light. Although absorption by water (liquid and solid) is not identically zero at visible wavelengths, and selective absorption by water can lead to observable consequences (e.g., colors of the sea and glaciers), the appearance of all but the thickest clouds is not determined by this selective absorption.

Equation (42) is the key to the vastly different optical characteristics of clouds and of the rain for which they are the progenitors. For a fixed amount of water (as specified by the quantity fh), optical thickness is inversely proportional to mean diameter. Rain drops are about 100 times larger on average than cloud droplets, and hence optical thicknesses of rain shafts are correspondingly smaller. We often can see through many kilometers of intense rain whereas a small patch of fog on a well-traveled highway can result in carnage.

9.2 Givers and Takers of Light

Scattering of visible light by a single water droplet is vastly greater in the forward ($\theta < 90^\circ$) hemisphere than in the backward ($\theta > 90^\circ$) hemisphere (Fig. 9). But water droplets in a thick cloud illuminated by sunlight collectively scatter

much more in the backward hemisphere (reflected light) than in the forward hemisphere (transmitted light). In each scattering event, incident photons are deviated, on average, only slightly, but in many scattering events most photons are deviated enough to escape from the upper boundary of the cloud. Here is an example in which the properties of an ensemble are different from those of its individual members.

Clouds seen by passengers in an airplane can be dazzling, but if the airplane were to descend through the cloud these same passengers might describe the cloudy sky overhead as gloomy. Clouds are both givers and takers of light. This dual role is exemplified in Fig. 20, which shows the calculated diffuse downward irradiance below clouds of varying optical thickness. On an airless planet the sky would be black in all directions (except directly toward the sun). But if the sky were to be filled from horizon to horizon with a thin cloud, the brightness overhead would markedly increase. This can be observed in a partly overcast sky, where gaps between clouds (blue sky) often are noticeably darker than

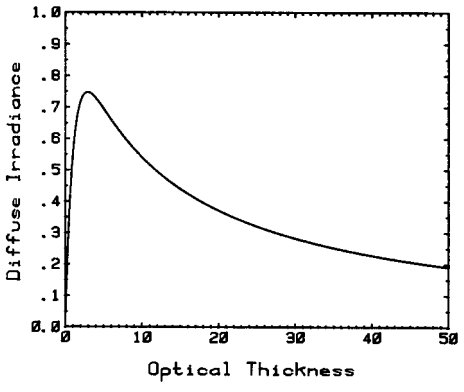

Fig. 20 Computed diffuse downward irradiance below a cloud relative to the incident solar irradiance as a function of cloud optical thickness

their surroundings. As so often happens, more is not always better. Beyond a certain cloud optical thickness, the diffuse irradiance decreases. For a sufficiently thick cloud, the sky overhead can be darker than the clear sky.

Why are clouds bright? Why are they dark? No inclusive one-line answers can be given to these questions. Better to ask, Why is that particular cloud bright? Why is that particular cloud dark? Each observation must be treated individually; generalizations are risky. Moreover, we must keep in mind the difference between brightness and radiance when addressing the queries of human observers. Brightness is a sensation that is a property not only of the object observed but of its surroundings as well. If the luminance of an object is appreciably greater than that of its surroundings, we call the object bright. If the luminance is appreciably less, we call the object dark. But these are relative rather than absolute terms.

Two clouds, identical in all respects, including illumination, may still appear different because they are seen against different backgrounds, a cloud against the horizon sky appearing darker than when seen against the zenith sky.

Of two clouds under identical illumination, the smaller (optically) will be less bright. If an even larger cloud were to appear, the cloud that formerly had been described as white might be demoted to gray.

With the sun below the horizon, two identical clouds at markedly different elevations might appear quite different in brightness, the lower cloud being shadowed from direct illumination by sunlight.

A striking example of dark clouds can sometimes be seen well after the sun has set. Low-lying clouds that are not illuminated by direct sunlight but are

seen against the faint twilight sky may be relatively so dark as to seem like ink blotches.

Because dark objects of our everyday lives usually owe their darkness to absorption, nonsense about dark clouds is rife: they are caused by pollution or soot. Yet of all the reasons that clouds are sometimes seen to be dark or even black, absorption is not among them.

Glossary

Airlight: Light resulting from scattering by all atmospheric molecules and particles along a line of sight.

Antisolar Point: Direction opposite the sun.

Astronomical Horizon: Horizontal direction determined by a bubble level.

Brightness: The attribute of sensation by which an observer is aware of differences of luminance (definition recommended by the 1922 Optical Society of America Committee on Colorimetry).

Contrast Threshold: The minimum relative luminance difference that can be perceived by the human observer.

Inferior Mirage: A mirage in which images are displaced downward.

Irradiance: Radiant power crossing unit area in a hemisphere of directions.

Lapse Rate: The rate at which a physical property of the atmosphere (usually temperature) decreases with height.

Luminance: Radiance integrated over the visible spectrum and weighted by the spectral response of the human observer. Also sometimes called *photometric brightness*.
.5.5

Mirage: An image appreciably different from what it would be in the absence of atmospheric refraction.

Neutral Point: A direction in the sky for which the light is unpolarized.

Normal Optical Thickness: Optical thickness along a radial path from the surface of the earth to infinity.

Optical Thickness: The thickness of a scattering medium measured in units of photon mean free paths. Optical thicknesses are dimensionless.

Radiance: Radiant power crossing a unit area and confined to a unit solid angle about a particular direction.

Scale Height: The vertical distance over which a physical property of the atmosphere is reduced to $1/e$ of its value.

Scattering Angle: Angle between incident and scattered waves.

Scattering Coefficient: The product of scattering cross section and number density of scatterers.

Scattering Cross Section: Effective area of a scatterer for removal of light from a beam by scattering.

Scattering Plane: Plane determined by incident and scattered waves.

Solar Point: The direction toward the sun.

Superior Mirage: A mirage in which images are displaced upward.

Tangential Optical Thickness: Optical thickness through the atmosphere along a horizon path.

References

Many of the seminal papers in atmospheric optics, including those by Lord Rayleigh, are bound together in Bohren, C. F. (Ed.) (1989), *Selected Papers on Scattering in the Atmosphere*, Bellingham, WA: SPIE Optical Engineering Press. Papers marked with an asterisk are in this collection.

[1] Möller, F. (1972), Radiation in the atmosphere, in D. P. McIntyre (Ed.), *Meteorological Challenges: A History*. Ottawa: Information Canada, pp. 43–71.
[2]* Penndorf, R. (**1957**), *J. Opt. Soc. Am.* **47**, 176–182.
[3]* Young, A. T. (**1982**), *Phys. Today* **35**(1), 2–8.
[4] Einstein, A. (**1910**), *Ann. Phys. (Leipzig)* **33**, 175; English translation in Alexander, J. (Ed.) (1926), *Colloid Chemistry*, Vol. I. New York: The Chemical Catalog Company, pp. 323–339.
[5] Zimm, B. H. (**1945**), *J. Chem. Phys.* **13**, 141–145.
[6] Thekaekara, M. P., Drummond, A. J. (**1971**), *Nat. Phys. Sci.* **229**, 6–9.
[7]* Hulburt, E. O. (**1953**), *J. Opt. Soc. Am.* **43**, 113–118.
[8] von Frisch, K. (1971), *Bees: Their Vision, Chemical Senses, and Language*, revised edition, Ithaca, NY: Cornell University Press, p. 116.
[9] Doyle, W. T. (**1985**), *Am. J. Phys.* **53**, 463–468.
[10]* Fraser, A. B. (**1975**), *Atmosphere* **13**, 1–10.
[11] Takano, Y., Asano, S. (**1983**), *J. Meteor. Soc. Jpn.* **61**, 289–300.
[12] van de Hulst, H. C. (1957), *Light Scattering by Small Particles*. New York: Wiley-Interscience.
[13] Pledgley, E. (**1986**), *Weather* **41**, 401.
[14] Lee, R., Fraser, A. (**1990**), *New Scientist* **127**(September), 40–42.
[15] Lee, R. (**1991**), *Appl. Opt.* **30**, 3401–3407.
[16]* Fraser, A. B. (**1983**), *J. Opt. Soc. Am.* **73**, 1626–1628.
[17] Fraser, A. B. (**1972**), *J. Atmos. Sci.* **29**, 211, 212.
[18]* van de Hulst, H. C. (**1947**), *J. Opt. Soc. Am.* **37**, 16–22.
[19]* Bryant, H. C., Cox, A. J. (**1966**), *J. Opt. Soc. Am.* **56**, 1529–1532.
[20]* Nussenzveig, H. M. (**1979**), *J. Opt. Soc. Am.* **69**, 1068–1079.
[21]* Fraser, A. B. (**1979**), *J. Opt. Soc. Am.* **69**, 1112–1118.
[22] Neuberger, H. (1951), *Introduction to Physical Meteorology*. University Park, PA: College of Mineral Industries, Pennsylvania State University.

Further Reading

Minnaert, M. (1954), *The Nature of Light and Colour in the Open Air*. New York: Dover Publications, is the bible for those interested in atmospheric optics. Like accounts of natural phenomena in the *Bible*, those in Minnaert's book are not always correct, despite which, again like the *Bible*, it has been and will continue to be a source of inspiration.

A book in the spirit of Minnaert's but with a wealth of color plates is by Lynch, D. K., Livingston, W. (1995), *Color and Light in Nature*. Cambridge, UK: Cambridge University Press.

A history of light scattering, From Leonardo to the Graser: Light Scattering in Historical Perspective, was published serially by Hey, J. D. (**1983**), *S. Afr. J. Sci.* **79**, 11–27, 310–324; Hey, J. D. (**1985**), *S. Afr. J. Sci.* **81**, 77–91, 601–613; Hey, J. D., (**1986**), *S. Afr. J. Sci.* **82**, 356–360. The history of the rainbow is recounted by Boyer, C. B. (1987), *The Rainbow*. Princeton, NJ: Princeton University Press.

A unique, beautifully written and illustrated treatise on rainbows in science and art, both sacred and profane, is by Lee, R. L., Fraser, A. B. (2001), *The Rainbow Bridge*, University Park, PA: Penn State University Press.

Special issues of *Journal of the Optical Society of America* (August 1979 and December 1983) and *Applied Optics* (20 August 1991 and 20 July 1994) are devoted to atmospheric optics.

Several monographs on light scattering by particles are relevant to and contain examples drawn from atmospheric optics: van de Hulst, H. C. (1957), *Light Scattering by Small Particles*. New York: Wiley-Interscience; reprint (1981), New York: Dover Publications; Deirmendjian, D. (1969), *Electromagnetic Scattering on Polydispersions*. New York: Elsevier; Kerker, M. (1969), *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press; Bohren, C. F., Huffman, D. R. (1983), *Light Scattering by Small Particles*. New York: Wiley-Interscience; Nussenzveig, H. M. (1992), *Diffraction Effects in Semiclassical Scattering*. Cambridge, UK: Cambridge University Press.

The following books are devoted to a wide range of topics in atmospheric optics: Tricker, R. A. R. (1970), *Introduction to Meteorological Optics*. New York: Elsevier; McCartney, E. J. (1976), *Optics of the Atmosphere*. New York: Wiley; Greenler, R. (1980), *Rainbows, Halos, and Glories*. Cambridge, UK: Cambridge University Press. Monographs of more limited scope are those by Middleton, W. E. K. (1952), *Vision Through the Atmosphere*. Toronto: University of Toronto Press; O'Connell, D. J. K. (1958), *The Green Flash and Other Low Sun Phenomena*. Amsterdam: North Holland; Rozenberg, G. V. (1966), *Twilight: A Study in Atmospheric Optics*. New York: Plenum; Henderson, S. T. (1977), *Daylight and its Spectrum*, (2nd ed.), New York: Wiley; Tricker, R. A. R. (1979), *Ice Crystal Haloes*. Washington, DC: Optical Society of America; Können, G. P. (1985), *Polarized Light in Nature*. Cambridge, UK: Cambridge University Press; Tape, W. (1994), *Atmosphere Halos*. Washington, DC: American Geophysical Union.

Although not devoted exclusively to atmospheric optics, Humphreys, W. J. (1964), *Physics of the Air*. New York: Dover Publications, contains a few relevant chapters. Two popular science books on simple experiments in atmospheric physics are heavily weighted toward atmospheric optics: Bohren, C. F. (1987), *Clouds in a Glass of Beer*. New York: Wiley; Bohren, C. F. (1991), *What Light Through Yonder Window Breaks?* New York: Wiley.

For an expository article on colors of the sky see Bohren, C. F., Fraser, A. B. (**1985**), *Phys. Teacher* **23**, 267–272.

An elementary treatment of the coherence properties of light waves was given by Forrester, A. T. (**1956**), *Am. J. Phys.* **24**, 192–196. This journal also published an expository article on the observable consequences of multiple scattering of light: Bohren, C. F. (**1987**), *Am. J. Phys.* **55**, 524–533.

Although a book devoted exclusively to atmospheric refraction has yet to be published, an elementary yet thorough treatment of mirages was given by Fraser, A. B., Mach, W. H. (**1976**), *Sci. Am.* **234**(1), 102–111.

Colorimetry, the often (and unjustly) neglected component of atmospheric optics, is treated in, for example, Optical Society of America Committee on Colorimetry (1963), *The Science of Color*. Washington, DC: Optical Society of America. Billmeyer, F. W., Saltzman, M. (1981), *Principles of Color Technology*, (2nd ed.), New York: Wiley-Interscience. MacAdam, D. L. (1985), *Color Measurement*, (2nd ed.), Berlin: Springer.

Understanding atmospheric optical phenomena is not possible without acquiring at least some knowledge of the properties of the particles responsible for them. To this end, the following are recommended: Pruppacher, H. R., Klett, J. D. (1980), *Microphysics of Clouds and Precipitation*. Dordrecht, Holland: D. Reidel. Twomey, S. A. (1977), *Atmospheric Aerosols*. New York: Elsevier.

Interferometry

Parameswaran Hariharan
School of Physics, University of Sydney, Australia
Phone: (612) 9413 7159; Fax: (612) 9413 7200; e-mail: hariharan_optics@hotmail.com

Katherine Creath
Optineering, Tucson, Arizona, USA
Phone: (520) 882-2950; Fax: (520) 882-6976; e-mail: kcreath@ieee.org

Abstract
This article reviews the field of interferometry. It begins by outlining the fundamentals of two-beam and multiple-beam interference. The rest of the article discusses the applications of interferometry for measurement of length, optical testing, fringe analysis, interference microscopy, interferometric sensors, interference spectroscopy, nonlinear interferometers, interferometric imaging, space-time and gravitation, holographic interferometry, Moiré techniques, and speckle interferometry. Each section provides a referenced (and cross-referenced) overview of the application area.

Keywords
interferometry; interference; interferometers; optical testing; optical metrology; nondestructive testing.

1 Introduction

Optical interferometry uses the phenomenon of interference between light waves to make extremely accurate measurements. The interference pattern contains, in addition to information on the optical paths traversed by the waves, information on the spectral content of the light and its spatial distribution over the source.

Young was the first to state the principle of interference and demonstrate that the summation of two rays of light could give rise to darkness, but the father of optical interferometry was undoubtedly Michelson. Michelson's contributions to interferometry, from 1880 to 1930, dominated the field to such an extent that optical interferometry was regarded for many years as a closed chapter. However, the last four decades have seen an explosive growth of interest in interferometry due to several new developments.

The most important of these was the development of the laser, which made available, for the first time, an intense source of light with a remarkably high degree of spatial and temporal coherence. Lasers have removed most of the limitations imposed by conventional light sources and have made possible many new techniques, including nonlinear interferometry.

Another development that has revolutionized interferometry has been the application of electronic techniques. The use of photoelectric detector arrays and digital computers has made possible direct measurements of the optical path difference at an array of points covering an interference pattern, with very high accuracy, in a very short time.

Light scattered from a moving particle has its frequency shifted by an amount proportional to the component of its velocity in a direction determined by the directions of illumination and viewing.

Lasers have made it possible to measure this frequency shift and, hence, the velocity of the particles, by detecting the beats produced by mixing the scattered light and the original laser beam.

Another major advance has been the use of single-mode optical fibers to build analogs of conventional two-beam interferometers. Since very long optical paths can be accommodated in a small space, fiber interferometers are now used widely as rotation sensors. In addition, since the length of the optical path in such a fiber changes with pressure or temperature, fiber interferometers have found many applications as sensors for a number of physical quantities.

In the field of stellar interferometry, it is now possible to combine images from widely spaced arrays of large telescopes to obtain extremely high resolution. Interferometry is also being applied to the detection of gravitational waves from black holes and supernovae.

Holography (see HOLOGRAPHY) is a completely new method of imaging based on optical interference. Holographic interferometry has made it possible to map the displacements of a rough surface with an accuracy of a few nanometers, and even to make interferometric comparisons of two stored wavefronts that existed at different times. Holographic interferometry and a related technique, speckle interferometry (see SPECKLE AND SPECKLE METROLOGY), are now used widely in industry for nondestructive testing and structural analysis.

The applications outlined above provide a glimpse of the many areas of optics that use interferometry. This article is meant to provide an overview. More detail is available in the cross-referenced articles as well as in the lists of works cited and further reading.

2 Interference and Coherence

When two light waves are superimposed, the resultant intensity depends on whether they reinforce or cancel each other. This is the well-known phenomenon of interference (see WAVE OPTICS).

If, at any point, the complex amplitudes of two light waves, derived from the same monochromatic point source and polarized in the same plane, are $A_1 = a_1 \exp(-i\phi_1)$ and $A_2 = a_2 \exp(-i\phi_2)$, the intensity (or the irradiance, units W m^{-2}) at this point is

$$\begin{aligned} I &= |A_1 + A_2|^2 \\ &= I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos(\phi_1 - \phi_2), \end{aligned} \tag{1}$$

where I_1 and I_2 are the intensities due to the two waves acting separately and $\phi_1 - \phi_2$ is the difference in their phases.

The visibility V of the interference fringes is defined by the relation

$$\begin{aligned} \mathrm{V} &= \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \\ &= \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2}. \end{aligned} \tag{2}$$

Interference effects can be observed quite easily by viewing a transparent plate illuminated by a point source of monochromatic light. In this case, interference takes place between the waves reflected from the front and back surfaces of the plate.

For a ray incident at an angle θ_1 on a plane-parallel plate (thickness d, refractive index n) and refracted within the plate at an angle θ_2, the optical path difference between the two reflected rays is

$$\Delta p = 2nd \cos\theta_2 + \frac{\lambda}{2}, \tag{3}$$

since an additional phase shift of π is introduced by reflection at one of the surfaces. The interference fringes are circles centered on the normal to the plate (fringes of equal inclination, or Haidinger fringes).

With a collimated beam, the interference fringes are contours of equal optical thickness (Fizeau fringes). The variations in the phase difference observed can represent variations in the thickness or the refractive index of the plate. A polished flat surface can be compared with a reference flat surface, by placing them in contact and observing the fringes of equal thickness formed in the air film between them. Introduction of a small tilt between the test and reference surfaces produces a set of almost straight and parallel fringes. Any deviations of the test surface from a plane are seen as a departure of the fringes from straight lines. The errors of the test surface can then be evaluated, as shown in Fig. 1, by measuring the maximum deviation (Δx) of a fringe from a straight line as well as the spacing between successive fringes (x). Each fringe corresponds to a change in the optical path difference of half a wavelength.

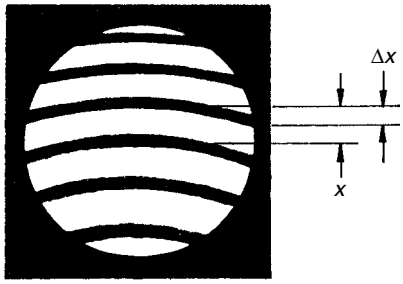

Fig. 1 Evaluation of the errors of a polished flat test surface by interference

2.1 Localization of Fringes

An extended monochromatic source can be considered as an array of independent point sources. Since the light waves from these sources take different paths to the point where interference is observed, the elementary interference patterns produced by any two of them will not, in general, coincide. Interference fringes are then observed with maximum contrast only in a particular region (the region of localization).

With a plane-parallel plate, the interference fringes are localized at infinity. With a wedged thin film, and near-normal incidence, the interference fringes are localized in the wedge.

2.2 Coherence

A more detailed analysis [1] shows that the interference effects observed depend on the degree of correlation between the wave fields at the point of observation. The intensity in the interference pattern is given by the relation

$$I = I_1 + I_2 + 2(I_1 I_2)^{1/2}|\gamma| \times \cos(\arg \gamma + 2\pi\nu\tau), \qquad (4)$$

where I_1 and I_2 are the intensities of the two beams, ν is the frequency of the radiation, τ is the mean time delay between the arrival of the two beams and γ is the (*complex*) degree of coherence between the wave fields.

With two beams of equal intensity, the visibility of the interference fringes is equal to $|\gamma|$, with a maximum value of 1 when the correlation between the wave fields is complete.

The correlation between the fields at any two points, when the difference between the optical paths to the source is small

enough for effects due to the spectral bandwidth of the light to be neglected, is a measure of the spatial coherence of the light. If the size of the source and the separation of the two points are very small compared to their distance from the source, it can be shown that the complex degree of coherence is given by the normalized two-dimensional Fourier transform of the intensity distribution over the source (see FOURIER AND OTHER TRANSFORM METHODS).

Similarly, the correlation between the fields at the same point at different times is a measure of the temporal coherence of the light and is related to its spectral bandwidth. With a point source (or when interference takes place between corresponding elements of the original wavefront), the visibility of the fringes as a function of the delay is the Fourier transform of the source spectrum.

To make this analysis complete, we must also take into account the polarization effects. In general, for maximum visibility, the beams must start in the same state of polarization (see POLARIZED LIGHT, BASIC CONCEPTS OF) and interfere in the same state of polarization. For natural (unpolarized) light, the optical path difference must be the same for all polarizations. The effects of deviations from these conditions, which can be quite complex, have been discussed by [2].

3 Two-beam Interferometers

Two methods are used to obtain two beams from a common source.

In wavefront division, two beams are isolated from separate areas of the primary wavefront. This technique was used in Young's experiment and in the Rayleigh interferometer.

More commonly, two beams are derived from the same portion of the primary wavefront (amplitude division) using a beam splitter (a transparent plate coated with a partially reflecting film), a diffraction grating or a polarizing prism.

3.1 The Michelson Interferometer

In the Michelson interferometer, a single beam splitter is used, as shown in Fig. 2, to divide and recombine the beams. However, to obtain interference fringes with white light, the two optical paths must contain the same thickness of glass. Accordingly, a compensating plate (of the same thickness and the same material as the beam splitter) is introduced in one beam.

The interference pattern observed is similar to that produced in a plate ($n = 1$) bounded by one mirror and the image of the other mirror produced by reflection from the beam splitter. With an extended source, the interference fringes are circles localized at infinity (fringes of equal

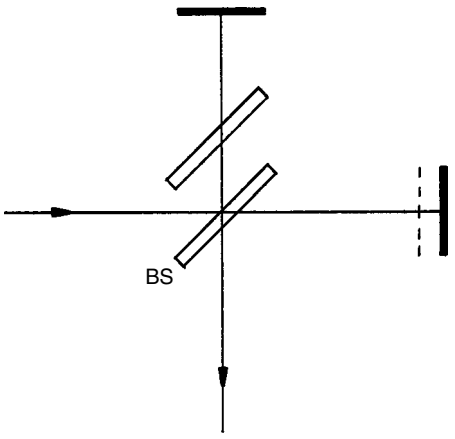

Fig. 2 The Michelson interferometer

inclination). With collimated light (the Twyman–Green interferometer), straight, parallel fringes of equal thickness (Fizeau fringes) are obtained.

3.2 The Mach–Zehnder Interferometer

As shown in Fig. 3, the Mach–Zehnder interferometer (MZI) uses two beam splitters to divide and recombine the beams.

The MZI has the advantage that each optical path is traversed only once. In addition, with an extended source, the region of localization of the fringes can be made to coincide with the test section. The MZI has been used widely to map local variations of the refractive index in wind tunnels, flames, and plasmas.

A variant, the Jamin interferometer, along with the Rayleigh interferometer, is commonly used to measure the refractive index of gases and mixtures of gases. Accurate measurements of the refractive index of air are a prerequisite for interferometric measurements of length (see Sect. 5).

3.3 The Sagnac Interferometer

In one form of the Sagnac interferometer, as shown in Fig. 4, the two beams traverse

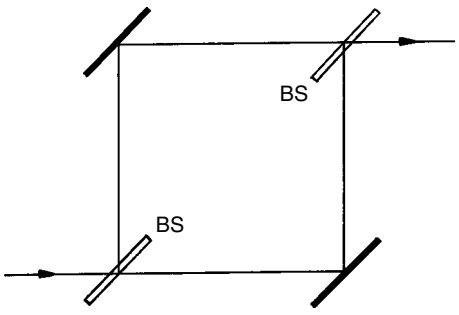

Fig. 3 The Mach–Zehnder interferometer

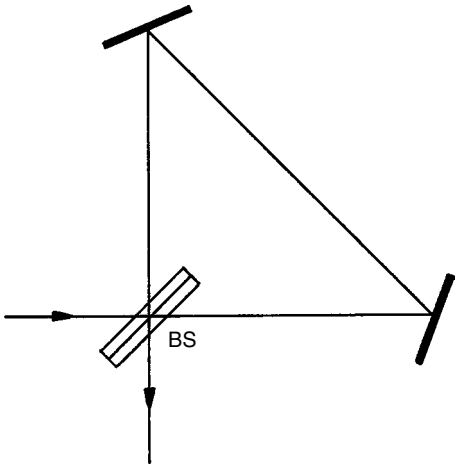

Fig. 4 The Sagnac interferometer

exactly the same path in opposite directions. However, with an odd number of reflections in the path, the wavefronts are laterally inverted with respect to each other.

Since the optical paths traversed by the two beams are always very nearly equal, fringes can be obtained easily with an extended, white-light source. Modified forms of the Sagnac interferometer are used for rotation sensing, since the rotation of the interferometer with an angular velocity Ω about an axis making an angle θ with the normal to the plane of the beams introduces an optical path difference

$$\Delta p = \left(\frac{4\Omega A}{c}\right)\cos\theta, \qquad (5)$$

between the two beams, where A is the area enclosed by the beams and c is the speed of light.

4 Multiple-beam Interference

With two highly reflecting surfaces, we have to take into account the effects

of multiply reflected beams (see WAVE OPTICS). The intensity in the interference pattern formed by the transmission is

$$I_T(\phi) = \frac{T^2}{1 + R^2 - 2R\cos\phi}, \qquad (6)$$

where R and T are, respectively, the reflectance and transmittance of the surfaces and $\phi = (4\pi/\lambda)nd\cos\theta_2$. As the reflectance R increases, the intensity at the minima decreases, and the bright fringes become sharper.

The finesse, defined as the ratio of the separation of adjacent fringes to the full width at half maximum (FWHM) of the fringes (the separation of points at which the intensity is equal to half its maximum value), is

$$F = \frac{\pi R^{1/2}}{1 - R}. \qquad (7)$$

The interference fringes formed by the reflected beams are complementary to those obtained by transmission.

4.1 Fringes of Equal Chromatic Order

With a white-light source, interference fringes cannot be seen for optical path differences greater than a few micrometers. However, if the reflected light is examined with a spectroscope, the spectrum will be crossed by dark bands corresponding to interference minima. With a thin film (thickness d, refractive index n), if λ_1 and λ_2 are the wavelengths corresponding to adjacent dark bands, we have

$$d = \frac{\lambda_1\lambda_2}{2n|\lambda_2 - \lambda_1|}. \qquad (8)$$

With two highly reflecting surfaces enclosing a thin film, very sharp fringes of equal chromatic order (FECO fringes) can be obtained, using a white-light source and a spectrograph. FECO fringes permit measurements with a precision of $\lambda/500$.

A major application of FECO fringes has been to study the microstructure of surfaces [3, 4]. However, the test surface must be coated with a highly reflective coating.

5 Measurement of Length

One of the earliest applications of interferometry was in measurements of lengths. Because of the limited distance over which interference fringes could be observed with conventional light sources, Michelson had to perform a laborious series of comparisons to measure the number of wavelengths of a spectral line in the standard meter. The extremely narrow spectral bandwidth of light from a laser has led to the development of a number of interferometric techniques for direct measurements of large distances. The values obtained for the optical path length are divided by the value of the refractive index of air, under the conditions of measurement, to obtain the true length.

5.1 Electronic Fringe Counting

If an additional phase difference of $\pi/2$ is introduced between the beams in one half of the field, two detectors can provide signals in quadrature to drive a bidirectional counter. These signals can also be processed to obtain an estimate of the fractional interference order [5].

5.2 Heterodyne Interferometry

In the Hewlett–Packard interferometer [6], a He-Ne laser is forced to oscillate

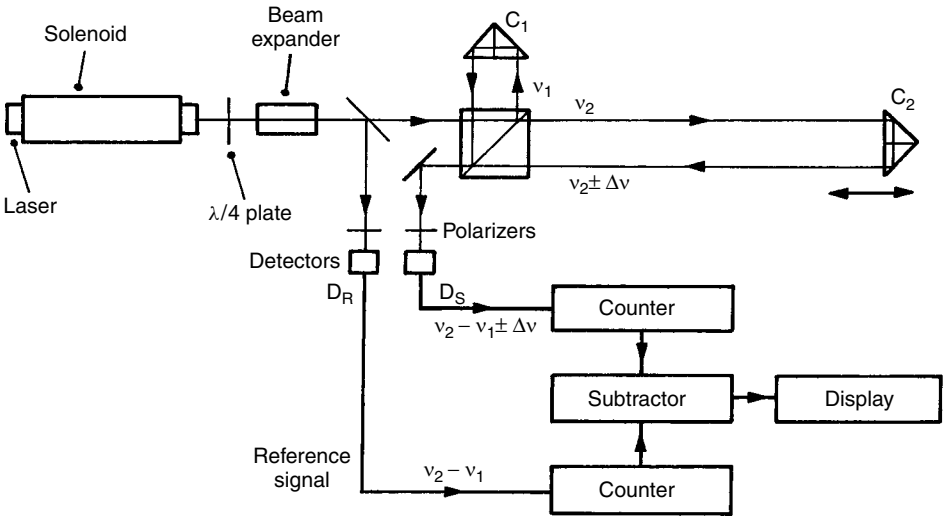

Fig. 5 Fringe-counting interferometer using a two-frequency laser (after Dukes, J. N., Gordon, G. B. (**1970**), *Hewlett-Packard J.* **21**, 2–8 [6]. © Hewlett–Packard Company. Reproduced with permission)

simultaneously at two frequencies separated by about 2 MHz by applying a longitudinal magnetic field. As shown in Fig. 5, these two frequencies that have opposite circular polarizations pass through a $\lambda/4$ plate that converts them to orthogonal linear polarizations.

A polarizing beam splitter reflects one frequency to a fixed cube-corner, while the other is transmitted to a movable cube-corner. Both frequencies return along a common axis and, after passing through a polarizer set at 45°, are incident on a photodetector. The beat frequencies from this detector and a reference detector go to a differential counter. If one of the cube-corners is moved, the net count gives the change in the optical path in wavelengths.

Very small changes in length can be measured by heterodyne interferometry. In one technique [7], a small frequency shift is introduced between the two beams, typically by means of a pair of acousto-optic modulators operated at slightly different frequencies. The output from a detector viewing the interference pattern contains a component at the difference frequency, and the phase of this heterodyne signal corresponds to the phase difference between the interfering beams.

In another technique, the two mirrors of a Fabry–Perot interferometer are attached to the two ends of the sample, and the wavelength of a laser is locked to a transmission peak [8]. A change in the separation of the mirrors results in a change in the wavelength of the laser and, hence, in its frequency. These changes can be measured with high precision by mixing the beam from the laser with the beam from a reference laser, and measuring the beat frequency.

5.3 Two-wavelength Interferometry

If an interferometer is illuminated simultaneously with two wavelengths λ_1 and λ_2, the envelope of the fringes yields the interference pattern that can be obtained with

a synthetic wavelength

$$\lambda_s = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|}. \qquad (9)$$

One way to implement this technique is with a CO_2 laser, which is switched rapidly between two wavelengths, as one of the mirrors of an interferometer is moved over the distance to be measured. The output signal from a photodetector is then processed to obtain the phase difference at any point [9].

5.4 Frequency-modulation Interferometry

Absolute measurements of distance can be made with a semiconductor laser by sweeping its frequency linearly with time [10]. If the optical path difference between the two beams in the interferometer is L, one beam reaches the detector with a time delay L/c, and they interfere to yield a beat signal with a frequency

$$f = \left(\frac{L}{c}\right)\left(\frac{df}{dt}\right), \qquad (10)$$

where df/dt is the rate at which the laser frequency varies with time.

5.5 Laser-feedback Interferometry

If, as shown in Fig. 6, a small fraction of the output of a laser is fed back to it by an external mirror, the output of the laser varies cyclically with the position of the mirror [11]. A displacement of the mirror by half a wavelength corresponds to one cycle of modulation.

A very simple laser-feedback interferometer can be set up with a single-mode semiconductor laser. An increased measurement range and higher accuracy can be obtained by mounting the mirror on a piezoelectric translator, and using an active feedback loop to hold the optical path constant [12].

6 Optical Testing

A major application of interferometry is in testing optical components and optical systems (see OPTICAL METROLOGY).

6.1 Flat Surfaces

The Fizeau interferometer (see Fig. 7) is used widely to compare a polished flat surface with a standard flat surface without placing them in contact and risking damage to the surfaces. Measurements are made on the fringes of equal thickness formed with collimated light in the air space separating the two surfaces.

To determine absolute flatness, it is possible to use a liquid surface as a reference [13]; however, a more often-used method is to test a set of three nominally flat surfaces in pairs. Errors of each of the three surfaces can be evaluated using this technique without the need for a known standard flat surface [14–16].

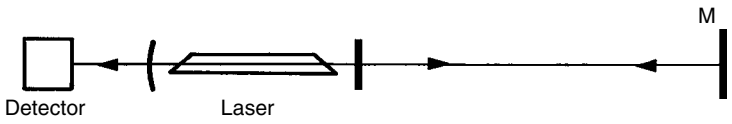

Fig. 6 Laser-feedback interferometer

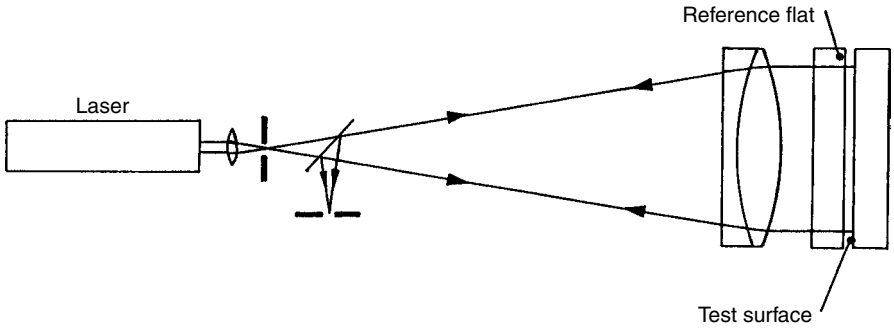

Fig. 7 Fizeau interferometer used to test flat surfaces

6.2 Homogeneity

The homogeneity of a material can be checked by preparing a plane-parallel sample and placing it in the test path of the interferometer. The effects of surface imperfections and systematic errors can be minimized by submerging the sample in a refractive-index matching oil and making measurements with and without the sample in the test path [17].

6.3 Concave and Convex Surfaces

The Fizeau and Twyman–Green interferometers can be used to test concave and convex surfaces [18, 19]. Typical test configurations for curved optical surfaces are shown in Fig. 8.

6.4 Prisms

Figure 9 shows a test configuration for a 60° prism. With a prism having a roof

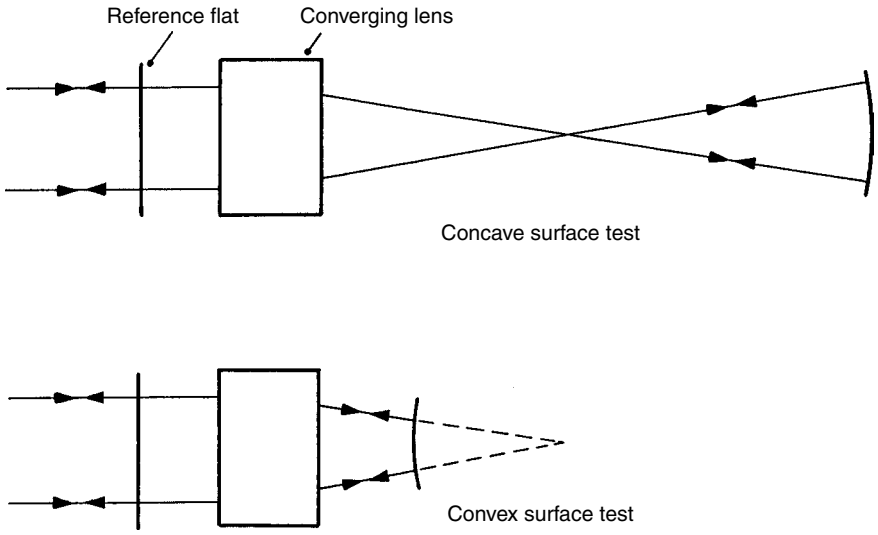

Fig. 8 Fizeau interferometer used to test concave and convex surfaces

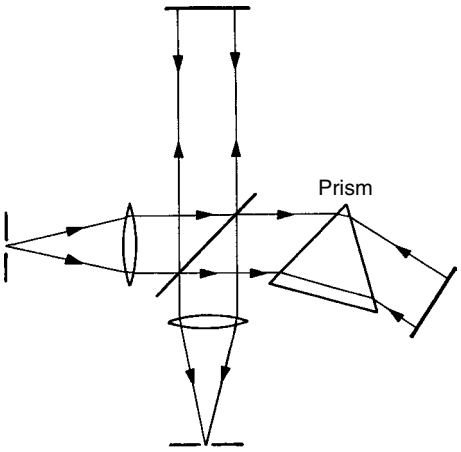

Fig. 9 Twyman–Green interferometer used to test a prism

angle of 90°, the beam is retro-reflected back through the system.

6.5 Aspheric Surfaces

Problems can arise in testing an aspheric surface against a spherical reference wavefront because the fringes in some parts of the resulting interferogram may be too closely spaced to be resolved. One way to solve this problem is to use a compensating null-lens [20]; another is to use a computer-generated hologram to produce a reference wavefront matching the desired aspheric wavefront [21, 22]. Shearing interferometry is yet another way to reduce the number of fringes in the interferogram [23, 24].

6.6 Optically Rough Surfaces

One way to test fine ground surfaces, before they are polished, is by infrared interferometry with a CO_2 laser at a wavelength of 10.6 μm [25]. A simpler alternative, with nominally flat surfaces, is to use oblique incidence [26]. Another means of measuring these surfaces uses scanning white-light techniques similar to those described in Sect. 8.3.

6.7 Shearing Interferometers

Shearing interferometers, in which interference takes place between two images of the test wavefront, have the advantage that they eliminate the need for a reference surface of the same dimensions as the test surface [23, 24]. With a lateral shear, as shown in Fig. 10(a), the two images undergo a mutual lateral displacement. If the shear is small, the wavefront aberrations can be obtained by integrating the phase data from two interferograms with orthogonal directions of shear. With a radial shear, as shown in Fig. 10(b), one of the images is contracted or expanded with respect to the

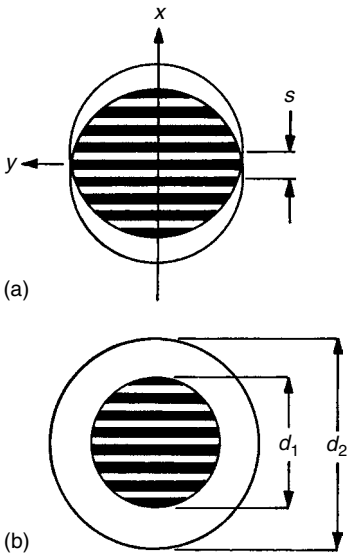

Fig. 10 Images of the test wavefront in (a) lateral and (b) radial shearing interferometers

other. If the diameter of one image is less than (say) 0.3 of the other, the interferogram obtained is very similar to that obtained with a Fizeau or Twyman–Green interferometer.

Other forms of shear, such as rotational, inverting or folding shears, can also be used for specific applications.

6.8 The Point-diffraction Interferometer

As shown in Fig. 11, the point-diffraction interferometer [27] consists of a pinhole in a partially transmitting film placed at the focus of the test wavefront.

The interference pattern formed by the test wavefront, which is transmitted by the film, and a spherical reference wave produced by diffraction at the pinhole corresponds to a contour map of the wavefront aberrations. Both wavefronts traverse the same path, making this compact interferometer insensitive to vibrations.

7 Fringe Analysis

High-accuracy information can be extracted from the fringe pattern, including the calculation of aberration coefficients (see OPTICAL ABERRATIONS; [28]), by using an electronic camera interfaced with a computer to measure and process the intensity distribution in the interference pattern. Several methods are available for this purpose (see [29]).

7.1 Fringe Tracking and Fourier Analysis

Early approaches to fringe analysis were based on fringe tracking [30]. In order to analyze a single fringe pattern, it is desirable to introduce a tilt between the interfering wavefronts so that a large number of nominally straight fringes are obtained. The shape of the fringe will be modified by the errors of the test wavefront. Fourier analysis of the fringes

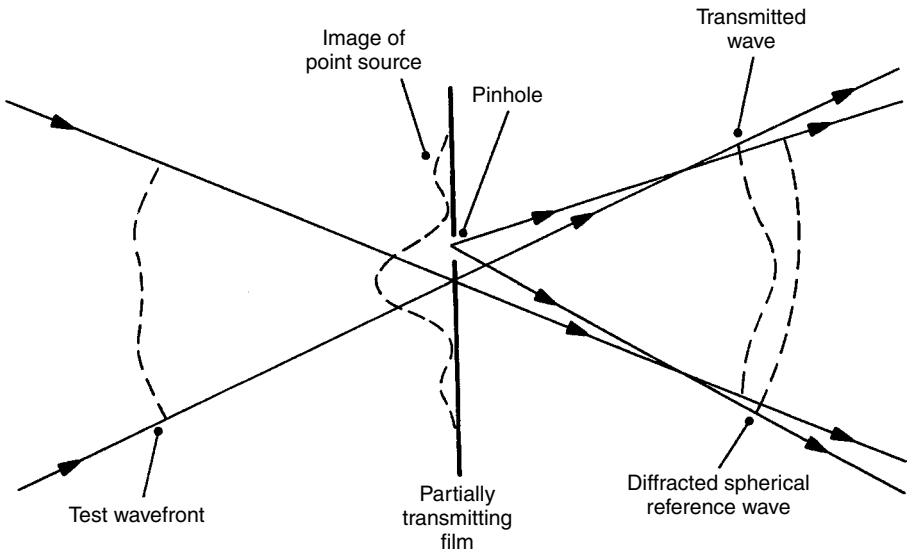

Fig. 11 The point-diffraction interferometer [27]

can then determine the test wavefront deviation [31–33].

7.2 Phase-shifting Interferometry

Direct measurement of the phase difference between the beams at a uniformly spaced array of points offers many advantages. In order to determine the phase of the wavefront at each data point, at least three interferograms are required. The phase difference between the interfering beams is usually varied linearly with time, and the intensity signal is integrated at each point over a number of equal phase segments covering one period of the sinusoidal output signal. This technique is often simplified by adjusting the phase difference in equal steps.

The most common way to accomplish the phase shift between the object and reference beams is by changing the optical path difference between the beams through a shift of the reference mirror along the optical axis. Other ways include tilting a glass plate, moving a grating, frequency-shifting, or rotating a half-wave plate or analyzer. Typically, intensity information from four or five interferograms are used to calculate the original phase difference between the wavefronts on a point-by-point basis [34, 35]. The repeatability of measurements is around $\lambda/1000$.

Because the phase-calculation algorithm utilizes an arctangent function, which does not yield any information on the integral interference order, it is necessary to use a phase unwrapping procedure to detect changes in the integral interference order and remove discontinuities in the retrieved phase [29, 36].

Figure 12 shows a three-dimensional plot of the errors of a flat surface produced by an interferometer using a digital phase-measurement system.

Normally, to implement such a phase unwrapping procedure, it is necessary to have at least two measurements per fringe spacing; this constraint, limits the phase gradients that can be measured. However, techniques are available that can be used in special situations, with some *a priori*

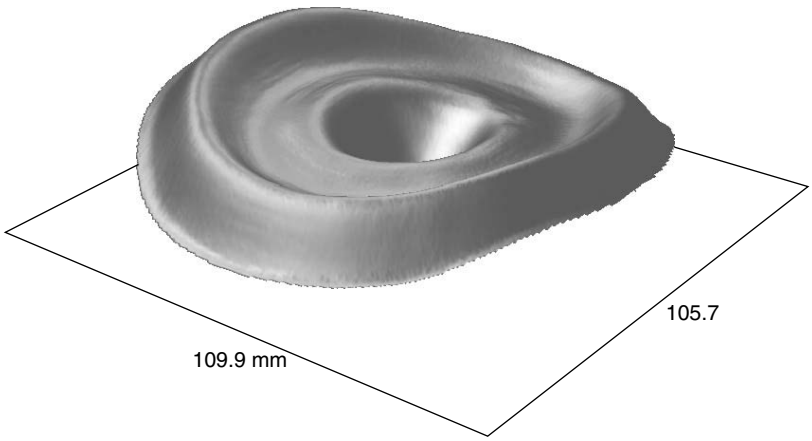

Fig. 12 Three-dimensional plot of the errors of a flat surface (326 nm Peak-to-Valley) obtained with a phase-measurement interferometer (Courtesy of Veeco Instruments Inc.)

knowledge about the test surface, to work around this limitation [37].

7.3 Determining Aberrations

With most optical systems, it is then convenient to express the deviations of the test wavefront as a linear combination of Zernike circular polynomials in the form

$$W(\rho,\theta) = \sum_{k=0}^{n}\sum_{l=0}^{k} \rho^k \times (A_{kl}\cos l\theta + B_{kl}\sin l\theta), \qquad (11)$$

where ρ and θ are polar coordinates over the pupil, and $(k-l)$ is an even number. If the optical path differences at a suitably chosen array of points are known, the coefficients A_{kl} and B_{kl} can be calculated from a set of linear equations [38, 39].

8 Interference Microscopy

Interference microscopy provides a noncontact method for studies of surfaces as well as a method for studying living cells without the need to stain them.

Two-beam interference microscopes have been described using optical systems similar to the Fizeau and Michelson interferometers. For high magnifications, a suitable configuration is that described by Linnik [40] in which a beam splitter directs the light onto two identical objectives; one beam is incident on the test surface, while the other is directed to the reference mirror.

8.1 The Mirau Interferometer

The Mirau interferometer permits a very compact optical arrangement. As shown schematically in Fig. 13, light from an illuminator is incident through the microscope objective on a beam splitter. The transmitted beam falls on the test surface, while the reflected beam falls on an aluminized spot on a reference surface. The two reflected beams are recombined at the same beam splitter and return through the objective.

As shown in Fig. 14, very accurate measurements of surface profiles can be made using phase shifting. With a rough surface, the data can be processed to obtain the rms surface roughness and the autocovariance function of the surface [41].

8.2 The Nomarski Interferometer

Common-path interference microscopes use polarizing elements to split and recombine the beams [42]. In the Nomarski interferometer (see Fig. 15), two polarizing

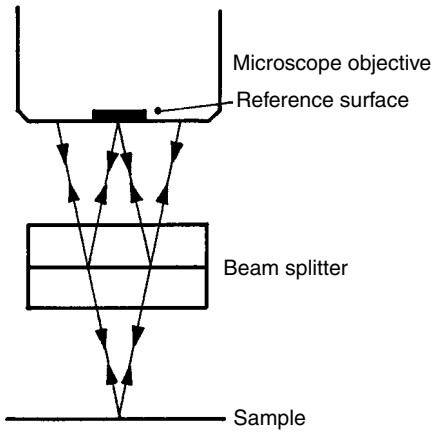

Fig. 13 The Mirau interferometer

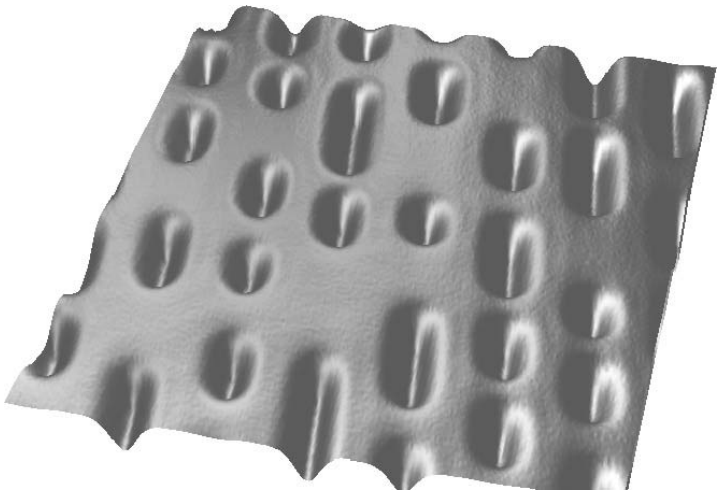

Fig. 14 Pits (approximately 90 nm deep) on the surface of a mass-replicated CD-ROM. 11 μm × 13 μm field of view (Courtesy of Veeco Instruments Inc.)

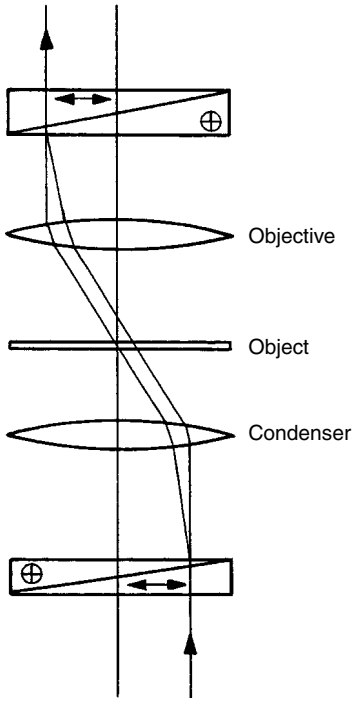

Fig. 15 The Nomarski interferometer

prisms (see MICROSCOPY) introduce a lateral shear between the two beams.

With small isolated objects, two images are seen covered with fringes that map the phase changes introduced by the object. With larger objects, the interference pattern is a measure of the phase gradients, revealing edges and local defects.

The use of phase-shifting techniques to extract quantitative information from the interference pattern has been described by [43].

8.3 White-light Interferometry

With monochromatic light, ambiguities can arise at discontinuities and steps producing a change in the optical path difference greater than a wavelength. This problem can be overcome by using a broadband (white-light) source. When the surface is scanned in height and the corresponding variations in intensity at each point are recorded, the height position corresponding to equal optical paths at which the visibility of the fringes

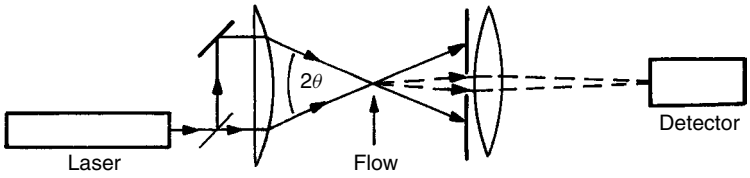

Fig. 16 Laser–Doppler interferometer for measurements of flow velocities

is a maximum yields the height of the surface at that point.

The method most commonly used to recover the fringe visibility function from the fringe intensity is by digital filtering [44]. More recent techniques combine phase-shifting techniques with signal demodulation [45, 46]. More detail on these techniques is available in OPTICAL METROLOGY.

Another technique that can be used with white light is spectrally resolved interferometry [47, 48]. A spectroscope is used to analyze the light from each point on the interferogram. The optical path difference between the beams at this point can then be obtained from the intensity distribution in the resulting channeled spectrum (see Sect. 4.1).

White-light interferometry techniques used for biomedical applications are generally referred to as optical coherence tomography or coherence radar (BIOMEDICAL IMAGING TECHNIQUES; [49, 50]).

9 Interferometric Sensors

Interferometers can be used as sensors for several physical quantities.

9.1 Laser–Doppler Interferometry

Laser–Doppler interferometry [51] makes use of the fact that light scattered by a moving particle has its frequency shifted.

In the arrangement shown in Fig. 16, two intersecting laser beams making angles $\pm\theta$ with the direction of observation are used to illuminate the test field.

The frequency of the beat signal observed is

$$\Delta\nu = \frac{2\nu \sin\theta}{\lambda}, \qquad (12)$$

where ν is the component of the velocity of the particle in the plane of the beams at right angles to the direction of observation.

Simultaneous measurements of the velocity components along orthogonal directions can be made by using two pairs of illuminating beams (with different wavelengths) in orthogonal planes.

It is also possible to use a self-mixing configuration for velocimetry, in which the light reflected from the moving object is mixed with the light in the laser cavity. A very compact system has been described by [52] using a laser diode operated near threshold with an external cavity to ensure single-frequency operation.

Very small vibration amplitudes can be measured by attaching one of the mirrors in an interferometer to the vibrating object. If the reflected beam is made to interfere with a reference beam with a fixed-frequency offset, the time-varying output contains, in addition to a component at the offset frequency (the carrier), two sidebands [53]. The vibration amplitude is

given by the relation

$$a = \left(\frac{\lambda}{2\pi}\right)\left(\frac{I_s}{I_c}\right), \qquad (13)$$

where I_c and I_s are, respectively, the power in the carrier and the sidebands.

9.2 Fiber Interferometers

Since the length of the optical path in an optical fiber changes when it is stretched, or when its temperature changes, interferometers in which the beams propagate in single-mode optical fibers (see FIBER OPTICS and SENSORS, OPTICAL) can be used as sensors for a number of physical quantities [54]. High sensitivity can be obtained, because it is possible to have very long, noise-free paths in a very small space.

In the interferometer shown in Fig. 17, light from a laser diode is focused on the end of a single-mode fiber and optical fiber couplers are used to divide and recombine the beams. Fiber stretchers are used to shift and modulate the phase of the reference beam. The output goes to a pair of photodetectors, and measurements are made with a heterodyne system or a phase-tracking system [55].

It is also possible, as shown in Fig. 18, to use a length of a birefringent single-mode fiber, in a configuration similar to a Fizeau interferometer, as a temperature-sensing element [56].

The outputs from the two detectors are processed to give the phase retardation between the waves reflected from the front and rear ends of the fiber. Changes in temperature of 0.0005 °C can be detected with a 1-cm-long sensing element.

Measurements of electric and magnetic fields can also be made with fiber interferometers by bonding the fiber sensor to a piezoelectric or magnetostrictive element. In addition, it is possible to multiplex several optical fiber sensors in a single system to make measurements of various quantities at a single location or even at different locations (see [57, 58]).

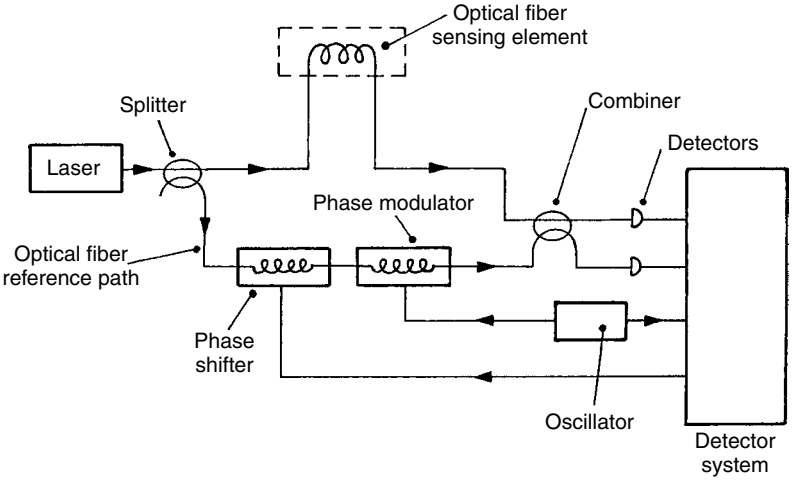

Fig. 17 Interferometer using a single-mode fiber as a sensing element. (Giallorenzi, T. G., Bucaro, J. A., Dandridge, A., Sigel, Jr G. H., Cole, J. H., Rashleigh, S. C., Priest, R. G. (**1982**), *IEEE J. Quantum Electron.* **QE-18**, 626–665 [55] © IEEE, 1982. Reproduced with permission.)

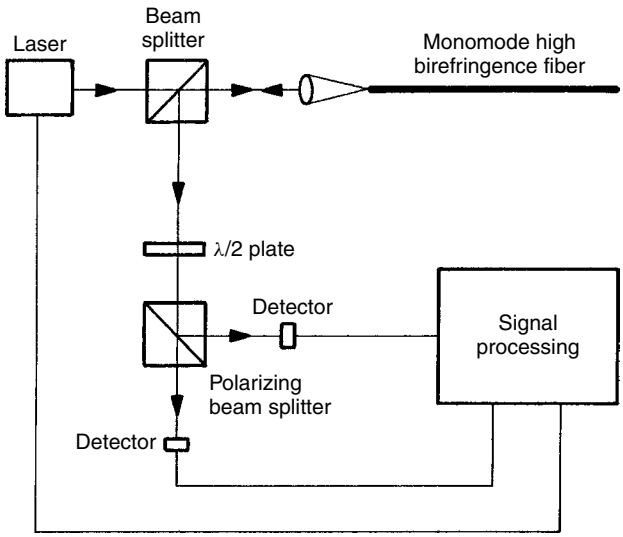

Fig. 18 Interferometer using a birefringent single-mode fiber as a sensing element [56]

9.3 Rotation Sensing

Another application of fiber interferometers is in rotation sensing [59, 60]. The configuration used, in which the two waves traverse a closed multiturn loop in opposite directions, is the equivalent of a Sagnac interferometer (see Sect. 3.3). A typical system is shown in Fig. 19 [61].

10 Interference Spectroscopy

Interferometric techniques are now used widely in high-resolution spectroscopy (see SPECTROSCOPY, LASER) because they offer, in addition to higher resolution, a higher throughput.

10.1 Etendue of an Interferometer

The throughput of an optical system is proportional to a quantity known as its etendue (see OPTICAL RADIATION SOURCES AND STANDARDS).

In the optical system shown in Fig. 20, the effective areas A_S and A_D of the source and detector are images of each other.

The etendue of the system is

$$E = A_S \Omega_S = A_D \Omega_D, \qquad (14)$$

where Ω_S is the solid angle subtended by the lens L_S at the source and Ω_D is the solid angle subtended by the lens L_D at the detector.

Since the etendue of a conventional spectroscope is limited by the entrance slit, a much higher etendue can be obtained with an interferometer.

10.2 The Fabry–Perot Interferometer

The Fabry–Perot interferometer (FPI) [62] uses multiple-beam interference between two flat, parallel surfaces coated with

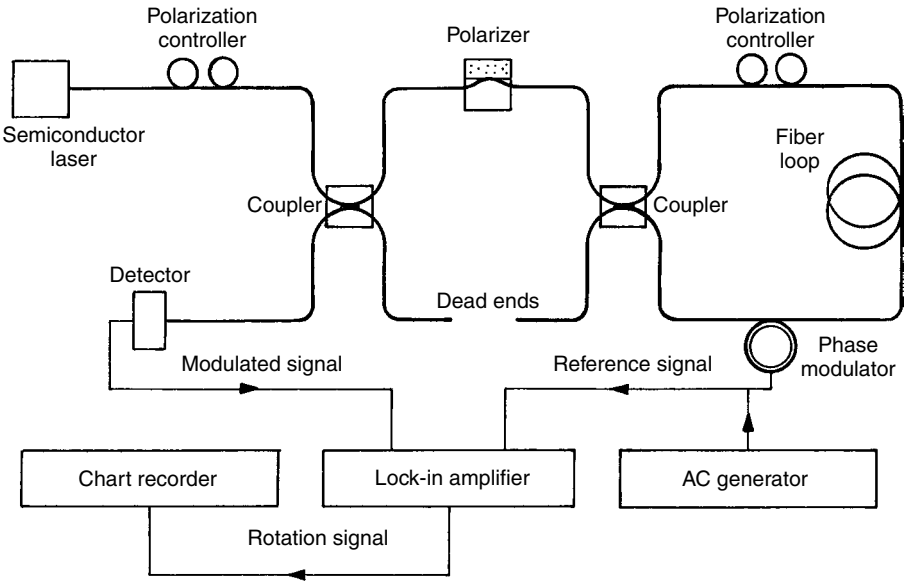

Fig. 19 Fiber interferometer for rotation sensing [61]

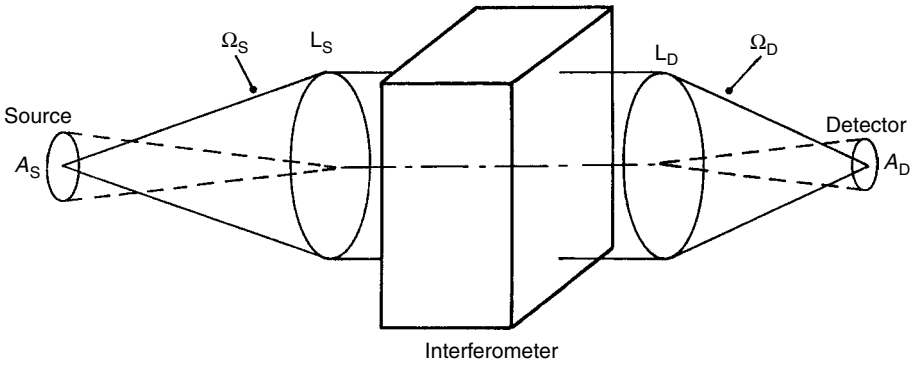

Fig. 20 Etendue of an interferometer

semitransparent, highly reflecting coatings. With a fixed spacing d, any single wavelength produces a system of sharp, bright rings (fringes of equal inclination), defined by Eq. (6), centered on the normal to the surfaces.

For any angle of incidence, with a broadband source, it also follows from Eq. (6) that the separation of successive intensity maxima corresponds to a wavelength difference $\lambda^2/2nd$. This wavelength difference, known as the free spectral range (FSR), is the range of wavelengths that can be handled by the FPI without successive interference orders overlapping.

The resolving power of the FPI is obtained by dividing the FSR by the finesse (see Eq. 7) and is given by the relation

$$R = \frac{\lambda}{\Delta\lambda} = \left(\frac{2nd}{\lambda}\right)\left[\frac{\pi R^{1/2}}{1-R}\right], \qquad (15)$$

where $\Delta\lambda$ is the half-width of the peaks, and R is the reflectance of the surfaces.

One way to overcome the limited FSR of the FPI is by imaging the fringes onto the slit of a spectrometer, but this procedure limits the etendue of the system. A better way is to use two or more FPIs, with different values of d, in series.

Another important characteristic of an FPI is the contrast factor, defined by the ratio of the intensities of the maxima and minima, which is

$$C = \left[\frac{1+R}{1-R}\right]^2. \qquad (16)$$

For applications such as Brillouin spectroscopy, in which a weak spectrum line may be masked by the background due to a neighboring strong spectral line, a much higher contrast factor can be obtained by passing the light several times through the same FPI [63].

A much higher throughput can be obtained with the confocal Fabry–Perot interferometer shown in Fig. 21 that uses two spherical mirrors separated by a distance equal to their radius of curvature, so that their foci coincide. Since the optical path difference is independent of the angle of incidence, a uniform field is obtained. An extended source can be used, and the transmitted intensity is recorded as the separation of the plates is varied [64].

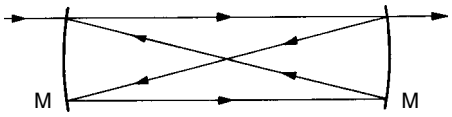

Fig. 21 Confocal Fabry–Perot interferometer

10.3 Wavelength Measurements

Accurate measurements of the wavelength of the output from a tunable laser, such as a dye laser, can be made with an interferometric wavelength meter.

In the dynamic wavelength meter shown in Fig. 22, a beam from the dye laser as well as a beam from a reference laser, whose wavelength is known, traverse the same two paths. The wavelength of the dye laser is determined by counting fringes simultaneously at both wavelengths, as the end reflector is moved [65].

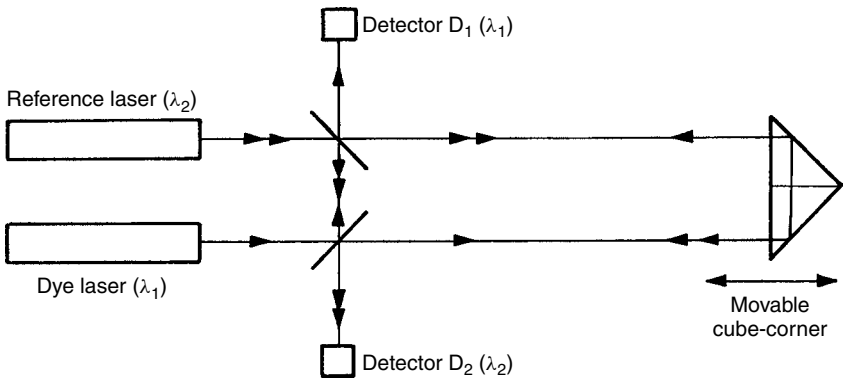

Fig. 22 Dynamic wavelength meter [65]

In a simpler arrangement [66], the fringes of equal thickness formed in a wedged air film are imaged on a linear detector array. The spacing of the fringes is used to evaluate the integral interference order, and their position to determine the fractional interference order.

10.4 Laser Linewidth

The extremely small spectral bandwidth of the output from a laser can be measured by mixing, as shown in Fig. 23, light from the laser with a reference beam from the same laser that has undergone a frequency shift and a delay [67].

10.5 Fourier-transform Spectroscopy

Major applications of Fourier-transform spectroscopy include measurements of infrared absorption spectra as well as emission spectra from faint astronomical objects (see SPECTROMETERS, ULTRAVIOLET AND VISIBLE LIGHT and SPECTROMETERS, INFRARED).

With a scanning spectrometer, the total time of observation T is divided between, say, m elements of the spectrum. Since in the infrared, the main source of noise is the detector, the signal-to-noise (S/N) ratio is reduced by a factor $m^{1/2}$. This reduction in the S/N ratio can be avoided by varying the optical path difference in an interferometer linearly with time, in which case each element of the spectrum generates an output modulated at a frequency that is inversely proportional to its wavelength. It is then possible to record all these signals simultaneously (or, in other words, to multiplex them) and then, by taking the Fourier transform of the recording (see FOURIER AND OTHER TRANSFORM METHODS), to recover the spectrum [68–70].

As shown in Fig. 24, a Fourier-transform spectrometer is basically a Michelson interferometer illuminated with an approximately collimated beam from the source whose spectrum is to be

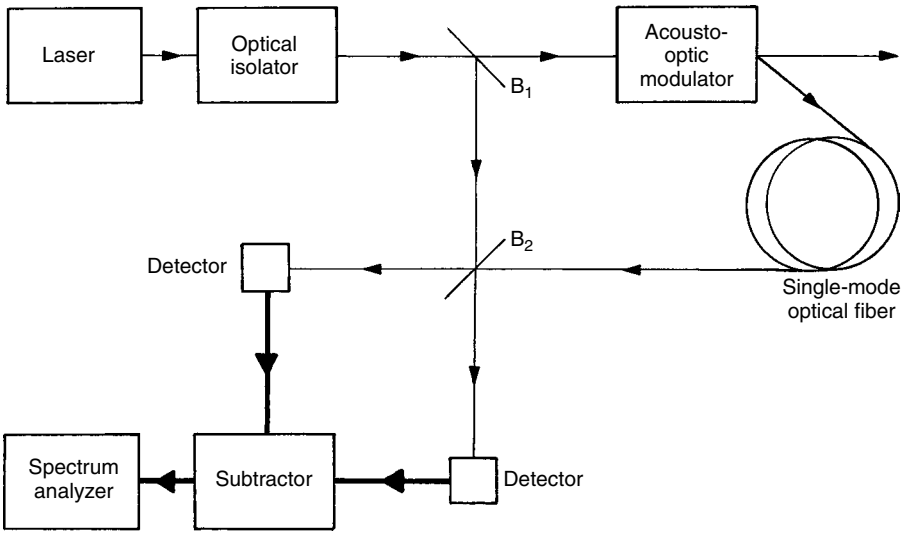

Fig. 23 Measurement of laser linewidth by heterodyne interferometry [67]

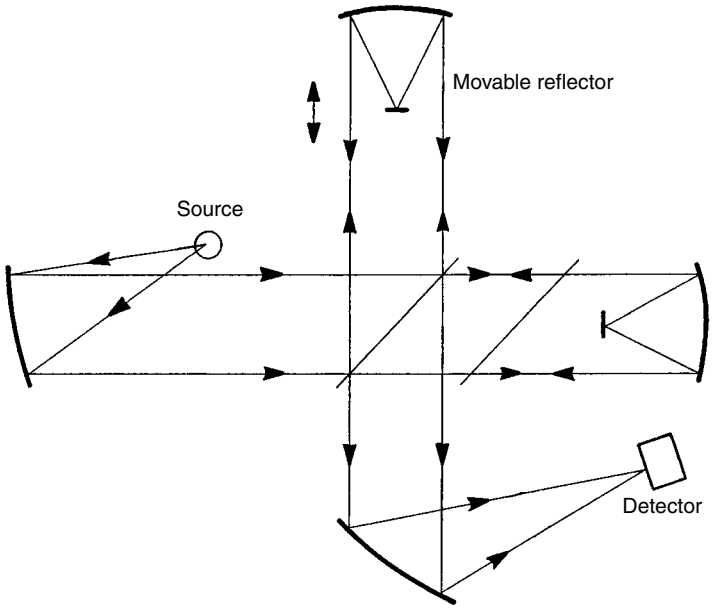

Fig. 24 Fourier-transform spectrometer

recorded. The mirrors are often replaced by "cat's eye" reflectors to minimize problems due to tilting as they are moved.

The interferogram is then sampled at a number of equally spaced points. To avoid ambiguities (aliasing) the change in the optical path difference between samples must be less than half the shortest wavelength in the spectrum. Finally, the spectrum is computed using the fast Fourier-transform algorithm [71]. Errors in the computed spectrum can be reduced by a process called apodization, where the interference signal is multiplied by a symmetrical weighting function whose value decreases gradually with the optical path difference.

11
Nonlinear Interferometers

The high light intensity available with pulsed lasers has opened up completely new areas of interferometry based on the use of nonlinear optical materials (see NONLINEAR OPTICS).

11.1
Second-harmonic Interferometers

Second-harmonic interferometers produce a fringe pattern corresponding to the phase difference between two second-harmonic waves generated at different points in the optical path from the original wave at the fundamental frequency.

Figure 25 is a schematic of an interferometer using two frequency-doubling crystals that can be considered as an analog of the Mach–Zehnder interferometer [72].

In this interferometer, the infrared beam from a Q-switched Nd:YAG laser ($\lambda_1 = 1.06\ \mu m$) is incident on a frequency-doubling crystal, and the green ($\lambda_2 =$

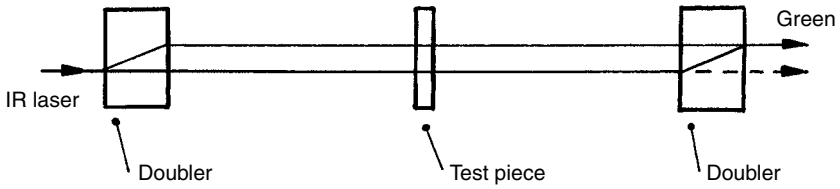

Fig. 25 Second-harmonic interferometer using two frequency-doubling crystals

0.53 μm) and infrared beams emerging from this crystal pass through the test piece. At the second crystal, the infrared beam undergoes frequency doubling to produce a second green beam that interferes with the one produced at the first crystal. The interference order at any point is

$$N(x, y) = \frac{(n_2 - n_1)\, \mathrm{d}(x, y)}{\lambda_2}, \qquad (17)$$

where $\mathrm{d}(x, y)$ is the thickness of the test specimen at any point (x, y) and n_1 and n_2 are its refractive indices for infrared and green light, respectively.

Other types of interferometers that are analogs of the Fizeau, Twyman, and point-diffraction interferometers have also been described [73, 74].

11.2 Phase-conjugate Interferometers

In a phase-conjugate interferometer, the test wavefront is made to interfere with its conjugate, eliminating the need for a reference wave and doubling the sensitivity.

In the phase-conjugate interferometer shown in Fig. 26, which can be regarded as an analog of the Fizeau interferometer, a partially reflecting mirror is placed in front of a single crystal of barium titanate, which functions as a self-pumped phase-conjugate mirror [75, 76].

An interferometer in which both mirrors have been replaced by a single-phase conjugator has the unique property that the field of view is normally completely dark and is unaffected by misalignment or by air currents. However, because of the delay in the response of the phase

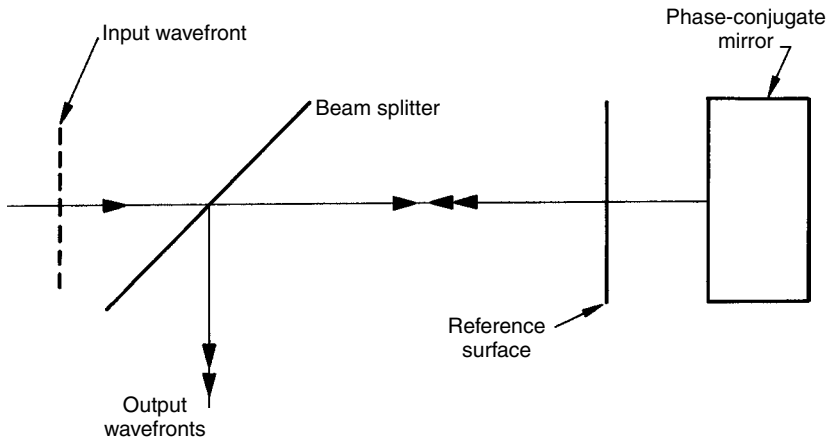

Fig. 26 Phase-conjugate interferometer [76]

conjugator, any sudden local change in the optical path produces a bright spot that slowly fades away [77].

11.3
Measurement of Nonlinear Susceptibilities

A modified Twyman–Green interferometer can be used for measurements of nonlinear susceptibilities. A system that can be used to measure the relative phase shift between two-phase conjugators, as well as the ratio of their susceptibilities, and yields high sensitivity, even with weak signals, has been described by [78].

12
Interferometric Imaging

Interferometric imaging started with the development of techniques to measure the diameters of stars that could not be resolved with conventional telescopes (see ASTRONOMICAL TELESCOPES AND INSTRUMENTATION).

Michelson's stellar interferometer [79] used the fact that the angular diameter of a star can be calculated from observations of the visibility of the fringes in an interferometer using light from the star reaching the surface of the earth at two points separated by a known distance.

If we assume the star to be a uniform circular source with an angular diameter 2α, and D is the separation of two mirrors receiving the light from the star and feeding it to a telescope, as shown in Fig. 27, the visibility of the fringes is

$$V = \frac{2J_1(u)}{u}, \tag{18}$$

where $u = 2\pi\alpha D/\lambda$ and J_1 is a first-order Bessel function. The fringe visibility drops

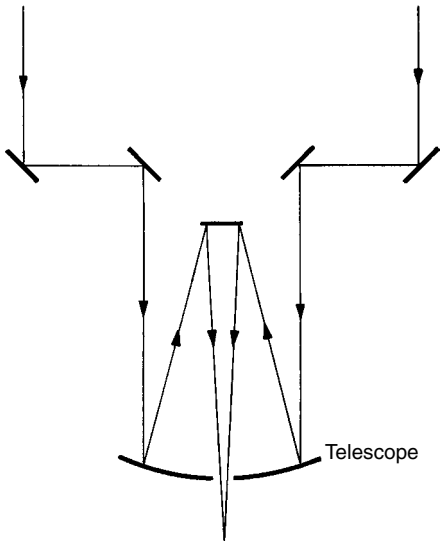

Fig. 27 Michelson's stellar interferometer

to zero when

$$D = \frac{1.22\lambda}{2\alpha}. \tag{19}$$

Measurements with Michelson's stellar interferometer over baselines longer than 6 m presented serious difficulties because of the difficulty of maintaining the optical path difference between the beams stable and small enough not to affect the visibility of the fringes. However, modern detection, control, and data handling techniques have made possible a new version of Michelson's stellar interferometer [80] designed to make measurements over baselines up to 640 m.

12.1
The Intensity Interferometer

The problem of maintaining the equality of the two paths was minimized in the intensity interferometer, which used measurements of the degree of correlation between the fluctuations in the outputs of two photodetectors at the foci of two

large light collectors separated by a variable distance [81, 82].

The actual instrument used light collectors operated with separations up to 188 m. With a bandwidth of 100 MHz it was only necessary to equalize the two optical paths to within 30 cm, but, because of the narrow bandwidth, measurements could only be made on 32 of the brightest stars.

12.2 Heterodyne Stellar Interferometers

In these instruments, as shown in Fig. 28, light from the star is received by two telescopes and mixed with light from a laser at two photodiodes. The resulting heterodyne signals are multiplied in a correlator. The output signal is a measure

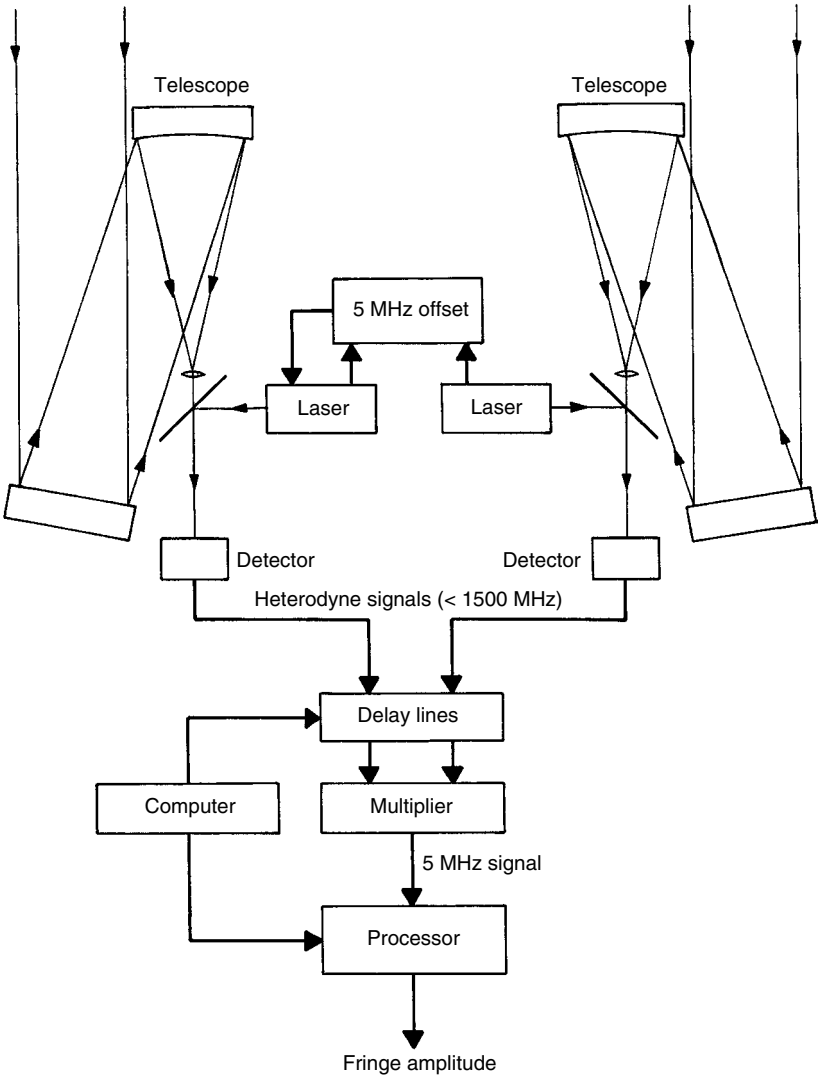

Fig. 28 Schematic of a heterodyne stellar interferometer [83]

of the degree of coherence of the wave fields at the two photo detectors [83].

As with the intensity interferometer, this technique only requires the two optical paths to be equalized to within a few centimeters; however, the sensitivity is higher, since it is proportional to the product of the intensities of the laser and the star. An infrared interferometer comprising two telescopes with an aperture of 1.65 m has been constructed, capable of yielding an angular resolution of 0.001 second of arc [84]. Larger telescopes are planned (see ASTRONOMICAL TELESCOPES AND INSTRUMENTATION).

12.3 Stellar Speckle Interferometry

Stellar speckle interferometry makes use of the fact that, due to local inhomogeneities in the earth's atmosphere, the image of a star produced by a large telescope, when observed under high magnification, has a speckle structure [85] (see also ASTRONOMICAL TELESCOPES AND INSTRUMENTATION). However, individual speckles have dimensions corresponding to a diffraction-limited image of the star. Reference [86] showed that a high-resolution image could be extracted from a number of such speckled images recorded with sufficiently short exposures to freeze the speckles.

While the angular resolution that can be obtained by speckle interferometry is limited by the aperture of the telescope, it has been applied successfully to a number of problems, including the study of close double stars.

12.4 Telescope Arrays

The ultimate objective would be the ability to produce high-resolution images of stars. Unfortunately, with a two-element interferometer, it is only possible to obtain information on the fringe amplitude because the value of the phase is affected by instrumental and atmospheric effects. However, with a triangular array, the closure phase is determined only by the coherence function. As the number of elements increases, the image becomes better constrained [87].

Some images have already been obtained from a large, multielement interferometer (the Cambridge Optical Aperture Synthesis Telescope) [88] and several other telescope arrays are nearing completion (see ASTRONOMICAL TELESCOPES AND INSTRUMENTATION).

13 Space-time and Gravitation

Michelson's classical experiment to test the hypothesis of a stationary ether showed an effect that was less than one-tenth of that expected. This experiment was repeated by [89], with a much higher degree of accuracy, by locking the frequency of a He-Ne laser to a resonance of a thermally isolated Fabry–Perot interferometer mounted along with it on a rotating horizontal granite slab. When the frequency of this laser was compared to that of a stationary, frequency-stabilized laser, the frequency shifts were found to be less than 1 part in 10^6 of those expected with a stationary ether.

13.1 Gravitational Waves

It follows from the general theory of relativity that binary systems of neutron stars, collapsing supernovae, and black holes should be the sources of gravitational waves.

Because of the transverse quadrupolar nature of a gravitational wave, the local distortion of space-time due to it stretches space in a direction normal to the direction of propagation of the wave, and shrinks it along the orthogonal direction. This local strain could, therefore, be measured by a Michelson interferometer in which the beam splitter and the end mirrors are attached to separate, freely suspended masses [90, 91].

Theoretical estimates of the intensity of gravitational radiation due to various possible events, suggest that a sensitivity to strain of the order of 10^{-21} over a bandwidth of a kilohertz would be needed. This would require an interferometer with unrealistically long arms.

One way to obtain a substantial increase in sensitivity is, as shown in Fig. 29, by using two identical Fabry–Perot cavities, with their mirrors mounted on freely suspended test masses, as the arms of the interferometer. The frequency of the laser is locked to a transmission peak of one interferometer, while the optical path length in the other is continually adjusted so that its peak transmittance is also at this frequency [92].

A further increase in sensitivity can be obtained by recycling the available light. Since the interferometer is normally adjusted so that observations are made on a dark fringe, to avoid overloading the detector, most of the light is returned toward the source and is lost. If this light is reflected back into the interferometer in the right phase, by an extra mirror placed in the input beam, the amount of light traversing the arms of the interferometer can be increased substantially [93].

Two other techniques that can be combined with these techniques for obtaining even higher sensitivity are signal recycling [94] and resonant side-band extraction [95].

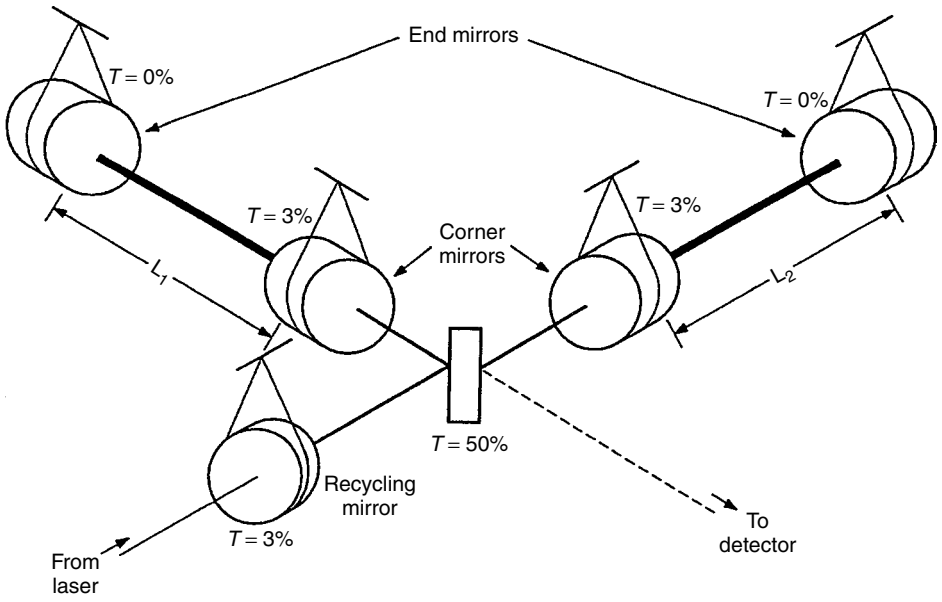

Fig. 29 Schematic of an interferometric gravitational wave detector [91]

13.2
LIGO

The laser interferometer gravitational observatory (LIGO) project [96, 97] involves the construction of three laser interferometers with arms up to 4-km long, two at one site and the third at another site separated from the first by almost 3000 km. The test masses and the optical paths in these interferometers are housed in a vacuum. Correlating the outputs of the three interferometers should make it possible to distinguish the signals due to gravitational waves from the bursts of instrumental and environmental noise.

13.3
Limits to Measurement

The limit to measurements of such small displacements is ultimately related to the number of photons n that pass through the interferometer in the measurement time. The resulting uncertainty in measurements of the phase difference between the beams has been shown to be [98]

$$\Delta\phi \geq \frac{1}{2\sqrt{n}}, \tag{20}$$

and is known as the standard quantum limit (SQL).

14
Holographic Interferometry

Holography (see HOLOGRAPHY and OPTICAL TECHNIQUES FOR MECHANICAL MEASUREMENT) makes it possible to store and reconstruct a perfect three-dimensional image of an object. The reconstructed wave can then be made to interfere with the wave generated by the object to produce fringes that contour, in real time, any changes in the shape of the object. Alternatively, two holograms can be recorded with the object in two different states and the wavefronts reconstructed by these two holograms can be made to interfere.

Since holographic interferometry makes it possible to measure, with very high precision, changes in the shape of objects with rough surfaces, it has found many applications including nondestructive testing and vibration analysis [99–101].

14.1
Strain Analysis

The phase difference at any point (x, y) in the interferogram is given by the relation (see Fig. 30)

$$\Delta\phi = \mathbf{L}(x, y) \cdot (\mathbf{k}_1 - \mathbf{k}_2) = \mathbf{L}(x, y) \cdot \mathbf{K}, \tag{21}$$

where $\mathbf{L}(x, y)$ is the vector displacement of the corresponding point on the surface of the object, $\mathbf{k}_1$ and $\mathbf{k}_2$ are the propagation vectors of the incident and scattered light, and $\mathbf{K} = \mathbf{k}_1 - \mathbf{k}_2$ is known as the sensitivity vector [102].

To evaluate the vector displacements (out-of-plane and in-plane), it is convenient to use a single direction of observation and

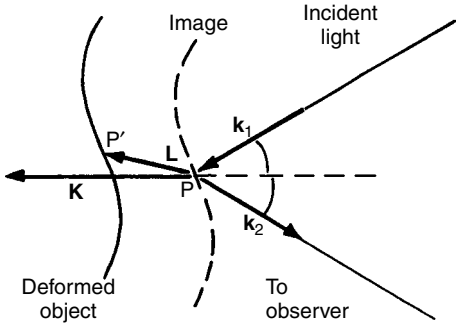

Fig. 30 Phase difference produced by a displacement of the object

record four holograms with the object illuminated from two different angles in the vertical plane and two different angles in the horizontal plane. Phase shifting is used to obtain the phase differences at a network of points [103]. These values can then be used, along with information on the shape of the object, to obtain the strains.

14.2 Vibration Analysis

One way to study the vibrating objects is to record a hologram of the vibrating object with an exposure time that is much longer than the period of the vibration [104]. The intensity at any point (x, y) in the image is then given by the relation

$$I(x, y) = I_0(x, y) J_0[\mathbf{K} \cdot \mathbf{L}(x, y)], \qquad (22)$$

where $I_0(x, y)$ is the intensity with the stationary object, J_0 is a zero-order Bessel function and $\mathbf{L}(x, y)$ is the amplitude of vibration of the object. The fringes observed (time-average fringes) are contours of equal vibration amplitude, with the dark fringes corresponding to the zeros of the Bessel function.

Alternatively, a hologram can be recorded of the stationary object, and the real-time interference pattern obtained with the vibrating object can be viewed using stroboscopic illumination. A brighter image can be obtained by recording the hologram with stroboscopic illumination, synchronized with the vibration cycle, and viewing the interference fringes formed with the stationary object, using continuous illumination. Phase-shifting techniques can then be used to map the instantaneous displacement of the vibrating object (see Fig. 31) [105].

14.3 Contouring

The simplest method of contouring an object is by recording two holograms with the object illuminated from slightly different angles. More commonly used techniques are two-wavelength contouring and two-refractive-index contouring.

In two-wavelength contouring [106], a telecentric lens system is used to image

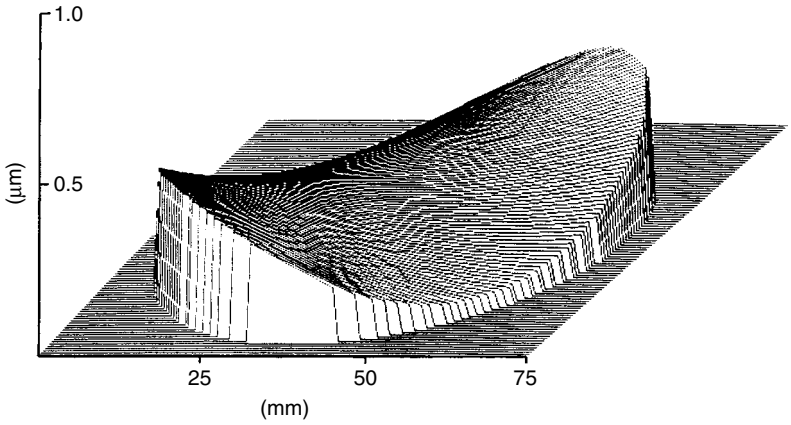

Fig. 31 Three-dimensional plot of the instantaneous displacement of a metal plate vibrating at 231 Hz obtained by stroboscopic holographic interferometry using phase shifting[105]

the object on the hologram plane and exposures are made with two different wavelengths, λ_1 and λ_2. When the hologram is illuminated with one of the wavelengths (say λ_2), fringes are seen contouring the reconstructed image, separated by an increment of height

$$|\delta z| = \frac{\lambda_1 \lambda_2}{2(\lambda_1 - \lambda_2)} \qquad (23)$$

In two-refractive-index contouring [107], the object is placed in a cell with a glass window and imaged by a telecentric system.

Two holograms are recorded on a plate placed near the stop of the telecentric system with the cell filled with liquids having refractive indices n_1 and n_2, respectively. Contours are obtained with a spacing

$$|\delta z| = \frac{\lambda}{2(n_1 - n_2)}. \qquad (24)$$

Digital phase-shifting techniques can be used with both these methods of holographic contouring. Figure 32 shows a three-dimensional plot of a wear mark on a flat surface obtained by phase shifting, using the two-refractive-index technique [108].

15 Moiré Techniques

Moiré techniques complement holographic interferometry and can be used where a contour interval greater than 10 μm is required [109] (see also OPTICAL TECHNIQUES FOR MECHANICAL MEASUREMENT).

A simple way to obtain Moiré fringes is to project interference fringes (or a grating) onto the object and view it through a grating of approximately the same spacing. The contour interval is determined by the fringe (grating) spacing and the angle between the illumination and viewing directions. Phase shifting is possible by shifting one grating or the projected fringes.

15.1 Grating Interferometry

The in-plane displacements of nearly flat objects can be measured with

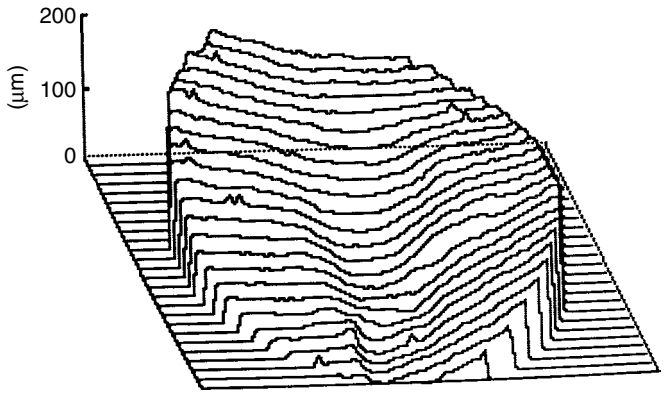

Fig. 32 Three-dimensional plot of a wear mark on a flat surface obtained by phase shifting, using the two-refractive-index technique [108]

submicrometer sensitivity by grating interferometry [109, 110].

A reflection grating is attached to the object under test with a suitable adhesive and illuminated by two coherent beams symmetrical to the grating normal. The interference pattern produced by the two diffracted beams reflected from the grating yields a map of the in-plane displacements. Polarization techniques can be used for phase shifting.

16
Speckle Interferometry

The image of an object illuminated by a laser is covered with a stationary granular pattern known as a speckle pattern (see SPECKLE AND SPECKLE METROLOGY). Speckle interferometry [111] utilizes interference between the speckled image of an object illuminated by a laser and a reference beam derived from the same laser. Any change in the shape of the object results in local changes in the intensity distribution in the speckle pattern. If two photographs of the speckled image are superimposed, fringes are obtained, corresponding to the degree of correlation of the two speckle patterns that contour the changes in shape of the surface [112].

Speckle interferometry can be a very simple way of measuring the in-plane displacements using an optical system in which the surface is illuminated by two beams making equal but opposite angles to the normal.

16.1
Electronic Speckle Pattern Interferometry (ESPI)

Measurements can be made at video rates using a TV camera interfaced to a computer [111, 113]. As shown in Fig. 33, the object is imaged on a CCD array along with a coaxial reference beam. This technique was originally known as electronic speckle pattern interferometry (ESPI).

If an image of the object in its original state is subtracted from an image of the object at a later stage, regions in which the speckle pattern has not changed, corresponding to the condition

$$\mathbf{K} \cdot \mathbf{L}(x, y) = 2m\pi, \qquad (25)$$

where m is an integer, appear dark, while regions where the pattern has changed are covered with bright speckles [114, 115].

This technique is also known as TV holography.

16.2
Phase-shifting Speckle Interferometry

Each speckle can be regarded as an individual interference pattern and the

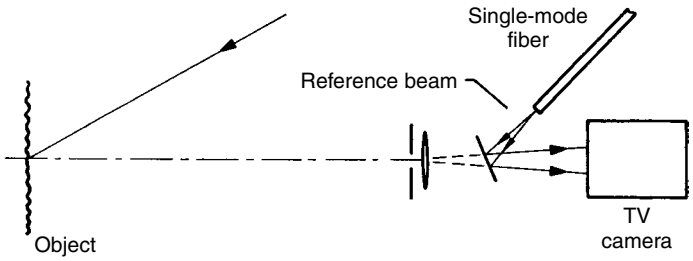

Fig. 33 System for ESPI

phase difference between the two beams at this point can be measured by phase shifting, with the object before and after an applied stress. The result of subtracting the second set of values from the first is then a contour map of the deformation of the object [116, 117].

Further developments in speckle interferometry with phase-measurement techniques, high-resolution CCD arrays, and real-time processing have created many techniques encompassing what is now most commonly known as digital holography (see OPTICAL TECHNIQUES FOR MECHANICAL MEASUREMENT and HOLOGRAPHY).

16.3 Vibrating Objects

If the period of the vibration is small compared to the exposure time or the scan time of the camera, the contrast of the speckles at any point is given by the expression

$$C = \frac{\{1 + 2\alpha J_0^2[\mathbf{K} \cdot \mathbf{L}(x, y)]\}^{1/2}}{1 + \alpha}, \qquad (26)$$

where α is the ratio of the intensities of the reference beam and the object beam, $\mathbf{K}$ is the sensitivity vector and $\mathbf{L}(x, y)$ is the vibration amplitude at that point. Regions corresponding to the zeros of the J_0 Bessel function appear as dark fringes.

Phase-shifting techniques can also be applied to the analysis of vibrations [118].

Glossary

Beam splitter: An optical element that divides a single beam of light into two beams of the same wave form.

Coherence: A complex quantity whose magnitude denotes the correlation between two wave fields; its phase denotes the effective phase difference between them.

Degree of Coherence: The value of the coherence expressed as a fraction of that for complete correlation between the wave fields.

Fringes of Equal Inclination: Interference fringes created from two collinear interfering beams having wavefronts with different radii of curvature.

Fringes of Equal Thickness: Interference fringes created with collimated beams when the optical path difference depends only on the thickness and refractive index.

Interference Order: The number of wavelengths in the optical path difference between two interfering beams.

Interferogram: The varying part of an interference pattern, after subtracting any uniform background.

Moiré Fringes: Relatively coarse fringes produced by the superposition of two fine fringe patterns with slightly different spacings or orientations.

Optical Path Difference: The difference in the optical path length between two interfering beams.

Optical Path Length: The product of the refractive index of the medium traversed by a beam and the length of the path in the medium.

Optical Phase: The resultant phase of a light beam after allowing for changes due to the optical path traversed and reflection at any surfaces.

Phase Shifting: A technique that shifts the phase of one interfering beam relative to the other in order to determine optical path difference from the intensity in an interference fringe pattern.

Region of Localization: The region in which interference fringes are observed with maximum contrast.

References

[1] Born, M., Wolf, E. (1999), *Principles of Optics*. Cambridge: Cambridge University Press.

[2] Leonhardt, K. (1981), *Optische Interferenzen*. Stuttgart: Wissenschaftliche Vergesselschaft.

[3] Tolansky, S. (1970), *Multiple Beam Interferometry of Surfaces and Thin Films*. Oxford: Oxford University Press.

[4] Bennett, J. M., Mattson, L. (1989), *Introduction to Surface Roughness and Scattering*. Washington, DC: Optical Society of America.

[5] Smythe, R., Moore, R. (**1984**), *Opt. Eng.* **23**, 361–364.

[6] Dukes, J. N., Gordon, G. B. (**1970**), *Hewlett-Packard J.* **21**, 2–8.

[7] Crane, R. (**1969**), *Appl. Opt.* **8**, 538–542.

[8] Jacobs, S. F., Shough, D. (**1981**), *Appl. Opt.* **20**, 3461–3463

[9] Matsumoto, H. (**1986**), *Appl. Opt.* **25**, 493–498.

[10] Kubota, T., Nara, M., Yoshino, T. (**1987**), *Opt. Lett.* **12**, 310–312.

[11] Ashby, D. E. T. F., Jephcott, D. F. (**1963**), *Appl. Phys. Lett.* **3**, 13–16.

[12] Yoshino, T., Nara, M., Mnatzakanian, S., Lee, B. S., Strand, T. C. (**1987**), *Appl. Opt.* **26**, 892–897.

[13] Bünnagel, R. (**1956**), *Z. Angew. Phys.* **8**, 342–350.

[14] Schultz, G., Schwider, J. (1976), in E. Wolf (Ed.), *Progress in Optics*, Vol. 13. Amsterdam: North Holland, pp. 93–167.

[15] Fritz, B. S. (**1984**), *Opt. Eng.* **23**, 379–383.

[16] Hariharan, P. (**1998**), *Proc. SPIE* **3479**, 2–13.

[17] Schwider, J. (1990), in E. Wolf (Ed.), *Progress in Optics*, Vol. 29. Amsterdam: North Holland, pp. 271–359.

[18] Mantravadi, M. V. (1992a), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 1–49.

[19] Malacara, D. (1992a), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 51–94.

[20] Offner, A., Malacara, D. (1992), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 427–454.

[21] Loomis, J. S. (**1980**), *Opt. Eng.* **19**(5), 679–685.

[22] Creath, K., Wyant, J. C. (1992), in D. Malacara (Ed.), *Optical Shop Testing*, (2nd ed.), New York: John Wiley & Sons, pp. 599–652.

[23] Mantravadi, M. V. (1992b), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 123–172.

[24] Malacara, D. (1992b), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 173–206.

[25] Kwon, O., Wyant, J. C., Hayslett, C. R. (**1980**), *Appl. Opt.* **19**, 1862–1869.

[26] Birch, K. G. (**1973**), *J. Phys. E: Sci. Instrum.* **6**, 1045–1048.

[27] Smartt, R. N., Steel, W. H. (**1975**), *Jpn. J. Appl. Phys.* **14**(Suppl. 14-1), 351–356.

[28] Wyant, J. C., Creath, K. (1992), in R. R. Shannon, J. C. Wyant (Eds.), *Applied Optics and Optical Engineering*, Vol. XI. San Diego: Academic Press, pp. 1–54.

[29] Robinson, D. W., Reid, G. T. (Eds.) (1993), *Interferogram Analysis: Digital Processing Techniques for Fringe Pattern Measurement*. London: IOP Publishing.

[30] Yatagai, T. (1993), in D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Techniques*. Bristol: IOP Publishing, pp. 72–93.

[31] Takeda, M., Ina, H., Kobayashi, S. (**1982**), *J. Opt. Soc. Am.* **72**, 156–160.

[32] Takeda, M. (**1990**), *Ind. Metrol.* **1**, 79–99.

[33] Kujawinska, M. (1993), in D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Techniques*. Bristol: IOP Publishing, pp. 141–193.

[34] Creath, K. (1988), in E. Wolf (Ed.), *Progress in Optics*, Vol. XXVI. Amsterdam: Elsevier, pp. 349–393.

[35] Creath, K. (1993), D. W. Robinson, G. T. Reid (Eds.), *Interferogram Analysis: Digital Fringe Pattern Measurement Techniques*. Bristol: IOP Publishing, pp. 94–140.
[36] Ghiglia, D. C., Pritt, M. D. (1998), *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software*. New York: John Wiley.
[37] Greivenkamp, J. E., Bruning J. H. (1992), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 501–598.
[38] Malacara, D., DeVore, S. L. (1992), in D. Malacara (Ed.), *Optical Shop Testing*. New York: John Wiley, pp. 455–499.
[39] Malacara, D., Servin, M., Malacara, Z. (1998), *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.
[40] Linnik, V. P. (**1933**), *C. R. Acad. Sci. URSS* **1**, 18.
[41] Wyant, J. C., Creath, K. (**Nov 1985**), *Laser Focus/Electro Opt*. 118–132.
[42] Françon, M., Mallick, S. (1971), *Polarization Interferometers: Applications in Microscopy and Macroscopy*. London: Wiley-Interscience.
[43] Cogswell, C. J., Smith, N. I., Larkin, K. G., Hariharan, P. (**1997**), *Proc. SPIE* **2984**, 72–81.
[44] Lee, B. S., Strand, T. C. (**1990**), *Appl. Opt.* **29**, 3784–3788.
[45] Hariharan, P., Roy, M. (**1994**), *J. Mod. Opt.* **41**, 2197–2201.
[46] Larkin, K. G. (**1996**), *J. Opt. Soc. Am. A.* **17**, 832–843.
[47] Calatroni, J., Guerrero, A. L., Sainz, C., Escalona, R. (**1996**), *Opt. Laser Technol.* **28**, 485–489.
[48] Sandoz, P., Tribillon, G., Perrin, H. (**1996**), *J. Mod. Opt.* **43**, 701–708.
[49] Schmit, J. M. (**1999**), *IEEE J. Selected Top. Quantum Electron.* **5**(2), 1205–1215.
[50] Häusler, G., Lindner, M. W. (**1998**), *J. Biomed. Opt.* **3**(1), 21–31.
[51] Durst, F., Melling, A., Whitelaw, J. H. (1976), *Principles and Practice of Laser-Doppler Anemometry*. London: Academic Press.
[52] de Groot, P. J., Gallatin, G. M. (**1989**), *Opt. Lett.* **14**, 165–167.
[53] Puschert, W. (**1974**), *Opt. Commun.* **10**, 357–361.
[54] Culshaw, B. (1984), *Optical Fibre Sensing and Signal Processing*. London: Peregrinus.
[55] Giallorenzi, T. G., Bucaro, J. A., Dandridge, A., Sigel, Jr G. H., Cole, J. H., Rashleigh, S. C., Priest, R. G. (**1982**), *IEEE J. Quantum Electron.* **QE-18**, 626–665
[56] Leilabady, P. A., Jones, J. D. C., Corke, M., Jackson, D. A. (**1986**), *J. Phys. E: Sci. Instrum.* **19**, 143–146.
[57] Farahi, F., Gerges, A. S., Jones, J. D. C., Jackson, D. A. (**1988a**), *Electron. Lett.* **24**, 54–55.
[58] Farahi, F., Jones, J. D. C., Jackson, D. A. (**1988b**), *Electron. Lett.* **24**, 409–410.
[59] Vali, V., Shorthill, R. W. (**1976**), *Appl. Opt.* **15**, 1099–1100.
[60] Ezekiel, S., Arditty, H. J. (Eds.) (1982), *Fiber-Optic Rotation Sensors and Related Technologies*. New York: Springer-Verlag.
[61] Bergh, R. A., Lefevre, H. C., Shaw, H. J. (**1981**), *Opt. Lett.* **6**, 502–504.
[62] Vaughan, J. M. (1989), *The Fabry-Perot Interferometer*. Bristol: Adam Hilger.
[63] Sandercock, J. R.(**1970**), *Opt. Commun.* **2**, 73–76.
[64] Fork, R. L., Herriott, D. R., Kogelnik, H. (**1964**), *Appl. Opt.* **3**, 1471–1484.
[65] Kowalski, F. V., Hawkins, R. T., Schawlow, A. L. (**1976**), *J. Opt. Soc. Am.* **66**, 965–966.
[66] Snyder, J. J. (1977), in J. L. Hall, J. L. Carsten (Eds.), *Laser Spectroscopy III*. Berlin: Springer-Verlag, pp. 419–420.
[67] Abitbol, C., Gallion, P., Nakajima, H., Chabran, C. (**1984**), *J. Opt. (Paris)* **15**, 411–418.
[68] Vanasse, G. A., Sakai, H. (1967), in E. Wolf (Ed.), *Progress in Optics*, Vol. 6. Amsterdam: North Holland, pp. 261–330.
[69] Bell, R. J. (1972), *Introductory Fourier Transform Spectroscopy*. New York: Academic Press.
[70] Chamberlain, J. (1979), *The Principles of Interferometric Spectroscopy*. Chichester: Wiley.
[71] Kauppinen, J., Partanen, J. (2001), *Fourier Transforms in Spectroscopy*,. New York: John Wiley.
[72] Hopf, F. A., Tomita, A., Al-Jumaily, G. (**1980**), *Opt. Lett.* **5**, 386–388.
[73] Hopf, F. A., Cervantes, M. (**1982**), *Appl. Opt.* **21**, 668–677.
[74] Liepmann, T. W., Hopf, F. A. (**1985**), *Appl. Opt.* **24**, 1485–1488.
[75] Feinberg, J. Hellwarth, R. W. (**1981**), *J. Opt. Soc. Am.* **71**(12), 1563.

[76] Gauthier, D. J., Boyd, R. W., Jungquist, R. K., Lisson, J. B., Voci, L. L. (**1989**), *Opt. Lett.* **14**, 323–325.
[77] Anderson, D. Z., Lininger, D. M., Feinberg, J. (**1987**), *Opt. Lett.* **12**, 123–125.
[78] Saltiel, S. M., Van Wonterghem, B., Rentzepis, P. M. (**1989**), *Opt. Lett.* **14**, 183–185.
[79] Michelson, A. A., Pease, F. G. (**1921**), *Astrophys. J.* **52**, 249–259.
[80] Davis, J., Tango, W. J., Booth, A. J., Minard, R. A., Owens, S. M., Shobbrook, R. R. (**1994**), *Proc. SPIE* **2200**, 231–241.
[81] Hanbury Brown, R., Twiss, R. Q. (**1954**), *Philos. Mag.* **19**, 10–11.
[82] Hanbury Brown, R. (1974), *The Intensity Interferometer*. Cambridge: Cambridge University Press.
[83] Johnson, M. A., Betz, A. L., Townes, C. H. (**1974**), *Phys. Rev. Lett.* **33**, 1617–1620.
[84] Bester, M., Danchi, W. C., Townes, C. H. (**1990**), *Proc. SPIE* **1237**, 40–48.
[85] Dainty, J. C. (Ed.) (1984), *Laser Speckle and Related Phenomena*, Topics in Applied Physics, Vol. 9, (2nd ed.), Berlin: Springer.
[86] Labeyrie, A. (1976), in E. Wolf (Ed.), *Progress in Optics*, Vol. 14, Amsterdam: North Holland, pp. 49–87.
[87] Baldwin, J. E., Haniff, C. A., Mackay, C. D., Warner, P. J. (**1986**), *Nature* **320**, 595–597.
[88] Baldwin, J. E., Beckett, M. G., Boysen, R. C., Burns, D., Buscher, D. F., Cox, G. C., Haniff, C. A., Mackay, C. D., Nightingale, N. S., Rogers, J., Scheuer, P. A. G., Scott, T. R., Tuthill, P. G., Warner, P. J., Wilson, D. M. A., Wilson, R. W. (**1996**), *Astron. Astrophys.* **306**, L13–L16.
[89] Brillet, A., Hall, J. L. (**1979**), *Phys. Rev. Lett.* **42**, 549–552.
[90] Saulson, P. R. (1994), *Fundamentals of Interferometric Gravitational Wave Detectors*. Singapore: World Scientific.
[91] Weiss, R. (**1999**), *Rev. Mod. Phys.* **71**, S187–S196.
[92] Drever, R. W. P. (1983), in N. Deruelle, T. Piran (Eds.), *Gravitational Radiation*. Amsterdam: North Holland, pp. 321–338.
[93] Strain, K. A., Meers, B. J. (**1991**), *Phys. Rev. Lett.* **66**, 1391–1394.
[94] Meers, B. J. (**1988**), *Phys. Rev. D* **38**, 2317–2326.
[95] Mizuno, J., Strain, K. A., Nelson, P. G., Chen, J. M., Schilling, R., Rüdiger, A., Winkler, W., Danzmann, K. (**1993**), *Phys. Lett. A* **175**, 273–276.
[96] Abramovici, A., Althouse, W. E., Drever, R. W. P., Gürsel, Y., Kawamura, S., Raab, F. J., Shoemaker, D., Sievers, L., Spero, R. E., Thorne, K. S., Vogt, R. E., Weiss, R., Whitcomb, S. E., Zucker, M. E. (**1992**), *Science* **256**, 325–333.
[97] Barish, B. C., Weiss, R. (**Oct 1999**), *Phys. Today* 44–50.
[98] Caves, C. M. (**1980**), *Phys. Rev. Lett.* **45**, 75–79.
[99] Vest, C. M. (1979), *Holographic Interferometry*. New York: John Wiley.
[100] Rastogi, P. K. (Ed.) (1994). *Holographic Interferometry*. Berlin: Springer-Verlag.
[101] Kreis, T. (1996), *Holographic Interferometry: Principles and Methods*. Berlin: Akademie-Verlag.
[102] Ennos, A. E. (**1968**), *J. Phys. E: Sci. Instrum.* **1**, 731–734.
[103] Hariharan, P., Oreb, B. F., Brown, N. (**1983**), *Appl. Opt.* **22**, 876–880.
[104] Powell, R. L., Stetson, K. A. (**1965**), *J. Opt. Soc. Am.* **55**, 1593–1598.
[105] Hariharan, P., Oreb, B. F. (**1986**), *Opt. Commun.* **59**, 83–86.
[106] Haines, K. A., Hildebrand, B. P. (**1965**), *Phys. Lett.* **19**, 10–11.
[107] Tsuruta, T., Shiotake, N., Tsujiuchi, J., Matsuda, K. (**1967**), *Jpn. J. Appl. Phys.* **6**, 661–662.
[108] Hariharan, P., Oreb, B. F. (**1984**), *Opt. Commun.* **51**, 142–144.
[109] Patorski, K. (1993), *Handbook of the Moiré Fringe Technique*. Amsterdam: Elsevier.
[110] Post, D., Han, D., Ifju, P. (1994), *High Sensitivity Moiré*. New York: Springer.
[111] Jones, R., Wykes, C. (1989), *Holographic and Speckle Interferometry*. Cambridge: Cambridge University Press.
[112] Leendertz, J. A. (**1970**), *J. Phys. E: Sci. Instrum.* **3**, 214–218.
[113] Lokberg, O. J., Slettemoen, G. A. (1987), in R. R. Shannon, J. C. Wyant (Eds.), *Applied Optics and Optical Engineering*, Vol. X. San Diego: Academic Press, pp. 455–504.
[114] Butters, J. N., Leendertz, J. A. (**1971**), *Opt. Laser Technol.* **3**, 26–30.
[115] Macovski, A., Ramsey, S. D., Schaefer, I. F. (**1971**), *Appl. Opt.* **10**, 2722–2727.
[116] Creath, K. (**1985**), *Appl. Opt.* **24**, 3053–3058.

[117] Robinson, D. W., Williams, D. C. (**1986**), *Opt. Commun.* **57**, 26–30.
[118] Stetson, K. A., Brohinsky, W. R. (**1988**), *J. Opt. Soc. Am.* **5**, 1472–1476.

Further Reading

Adrian, R. J. (Ed.) (1993), *Laser Doppler Velocimetry*, Vol. MS78. Bellingham: SPIE.
Brown, G. M. (Ed.) (2000), *Modern Interferometry*, Selected SPIE Papers on CD-ROM, Vol. 15. Bellingham: SPIE.
Hariharan, P. (1985), *Optical Interferometry*. Sydney: Academic Press.
Hariharan, P. (Ed.) (1991), *Interferometry*, Vol. MS28. Bellingham: SPIE.
Hariharan, P. (1992), *Basics of Interferometry*. San Diego: Academic Press.
Hariharan, P. (1995), *Optical Holography*. Cambridge: Cambridge University Press.
Hariharan, P., Malacara-Hernandez, D. (Eds.) (1995), *Interference, Interferometry and Interferometric Metrology*, Vol. MS110. Bellingham: SPIE.
Lawson, P. R. (Ed.) (1997), *Long Baseline Stellar Interferometry*, Vol. MS139. Bellingham: SPIE.
Meinischmidt, P., Hinsch, K. D., von Ossietzky, C., Sirohi, R. S. (Eds.) (1996), *Electronic Speckle Pattern Interferometry: Principles and Practice*, Vol. MS132. Bellingham: SPIE.
Sirohi, R. S., Hinsch, K. D., von Ossietzky, C. (Eds.) (1998), *Holographic Interferometry: Principles and Techniques*, Vol. MS144. Bellingham: SPIE.
Steel, W. H. (1983), *Interferometry*. Cambridge: Cambridge University Press.
Udd, E. (Ed.) (1991), *Fiber Optic Sensors: An Introduction for Engineers and Scientists*. New York: John Wiley & Sons.
Udd, E., Tatam, R. P. (Eds.) (1994), *Interferometry '94: Interferometric Fiber Sensing: 16–20 May 1994, Warsaw, Poland*, SPIE Proceedings No. 2341, Bellingham: SPIE.

Laser Cooling and Trapping of Neutral Atoms

Harold J. Metcalf
Department of Physics, State University of New York, Stony Brook, N.Y. 11794-3800, USA
Phone: (631) 632-8185; Fax: (631) 632-8176; e-mail: hmetcalf@notes.cc.sunysb.edu

Peter van der Straten
Debye Institute, Department of Atomic and Interface Physics, Utrecht University, 3508 TA Utrecht, The Netherlands
Phone: 31-30-2532846; Fax: 31-30-2537468; e-mail: P.vanderStraten@phys.uu.nl

Abstract
This article presents a review of some of the principal techniques of laser cooling and trapping that have been developed during the past 20 years. Its approach is primarily experimental, but its quantitative descriptions are consistent in notation with most of the theoretical literature.

Keywords
laser cooling; atom trapping; optical lattice; Bose–Einstein condensation.

1 Introduction

The combination of laser cooling and atom trapping has produced astounding new tools for atomic physicists [1]. These experiments require the exchange of momentum between atoms and an optical field, usually at a nearly resonant frequency. The energy of light $\hbar\omega$ changes the internal energy of the atom, and the angular momentum $\hbar$ changes the orbital angular momentum ℓ of the atom, as described by the well-known selection rule $\Delta\ell = \pm 1$. By contrast, the linear momentum of light $p = \hbar\omega/c = \hbar k$ cannot change the internal atomic degrees of freedom, and therefore must change the momentum of the atoms in the laboratory frame. The force resulting from this momentum exchange between the light field and the atoms can be used in many ways to control atomic motion, and is the subject of this article.

1.1 Temperature and Entropy

The idea of ''temperature'' in laser cooling requires some careful discussion and disclaimers. In thermodynamics, temperature is carefully defined as a parameter of the state of a closed system in thermal equilibrium with its surroundings. This, of course, requires that there be thermal contact, that is, heat exchange, with the environment. In laser cooling, this is clearly not the case because a sample of atoms is always absorbing and scattering light. Furthermore, there is essentially no heat exchange (the light cannot be considered as heat even though it is indeed a form of energy). Thus, the system may very well be in a steady state situation, but certainly not in thermal equilibrium, so the assignment of a thermodynamic ''temperature'' is completely inappropriate.

Nevertheless, it is convenient to use the label of temperature to describe an atomic sample whose average kinetic energy $\langle E_k \rangle$ has been reduced by the laser light, and this is written simply as $k_B T/2 = \langle E_k \rangle$, where k_B is the Boltzmann's constant (for the case of one dimension, 1-D). It must be remembered that this temperature assignment is absolutely inadequate for atomic samples that do not have a Maxwell–Boltzmann velocity distribution, whether or not they are in thermal contact with the environment; there are infinitely many velocity distributions that have the same value of $\langle E_k \rangle$ but are so different from one another that characterizing them by the same ''temperature'' is a severe error. (In the special case where there is a true damping force, $F \propto -v$, and where the diffusion in momentum space is a constant independent of momentum, solutions of the Fokker–Planck equation can be found analytically and can lead to a Maxwell–Boltzmann distribution that does indeed have a temperature.)

Since laser cooling decreases the temperature of a sample of atoms, there is less disorder and therefore less entropy. This seems to conflict with the second law of thermodynamics, which requires the entropy of a closed system to always increase with time. The explanation lies in the consideration of the fact that in laser cooling, the atoms do not form a closed system. Instead, there is always a flow of laser light with low entropy into the system and fluorescence with high entropy out of it. The decrease of entropy of the atoms is accompanied by a much larger increase in entropy of the light field. Entropy considerations for a laser beam are far from trivial, but recently it has been shown that the entropy lost by the atoms

is many orders of magnitude smaller than the entropy gained by the light field.

1.2 Phase Space Density

The phase space density $\rho(\vec{r}, \vec{p}, t)$ is defined as the probability that a single particle is in a region $d\vec{r}$ around $\vec{r}$ and has momentum $d\vec{p}$ around $\vec{p}$ at time t. In classical mechanics, $\rho(\vec{r}, \vec{p}, t)$ is just the sum of the $\rho(\vec{r}, \vec{p}, t)$ values of each of the N particles in the system divided by N. Since the phase space density is a probability, it is always positive and can be normalized over the six-dimensional volume spanned by position $\vec{r}$ and momentum $\vec{p}$. For a gas of cold atoms, it is convenient to choose the elementary volume for $\rho(\vec{r}, \vec{p}, t)$ to be $\hbar^3$, so it becomes the dimensionless quantity

$$\rho_\phi = n\lambda_{\mathrm{deB}}^3, \tag{1}$$

where λ_{deB} is the deBroglie wavelength of the atoms in the sample and n is their spatial density.

The Liouville theorem requires that ρ_ϕ cannot be increased by using conservative forces. For instance, in light optics one can focus a parallel beam of light with a lens to a small spot. However, that simply produces a high density of light rays in the focus in exchange for the momentum part of ρ_ϕ because the beam entering the lens is parallel but the light rays are divergent at the focus.

For classical particles, the same principle applies. By increasing the strength of the trapping potential of particles in a trap, one can increase the density of the atoms in the trap but at the same time, the compression of the sample results in a temperature increase, leaving the phase space density unchanged.

In order to increase the phase space density of an atomic sample, it is necessary to use a force that is not conservative, such as a velocity-dependent force. In laser cooling, the force on the atoms can be a damping force, that is, always directed opposite to the atomic velocity, so that the momentum part of ρ_ϕ increases. This process arises from the irreversible nature of spontaneous emission.

2 Optical Forces on Neutral Atoms

The usual form of electromagnetic forces is given by $\vec{F} = q(\vec{\mathcal{E}} + \vec{v} \times \vec{B})$, but for neutral atoms, $q = 0$. The next order of force is the dipole term, but this also vanishes because neutral atoms have no inherent dipole moment. However, a dipole moment can be induced by a field, and this is most efficient if the field is alternating near the atomic resonance frequency. Since these frequencies are typically in the optical range, dipole moments are efficiently induced by shining nearly resonant light on the atoms.

If the light is absorbed, the atom makes a transition to the excited state, and the subsequent return to the ground state can be either by spontaneous or by stimulated emission. The nature of the optical force that arises from these two different processes is quite different and will be described separately.

The spontaneous emission case is different from the familiar quantum-mechanical calculations using state vectors to describe the system. Spontaneous emission causes the state of the system to evolve from a pure state into a mixed state and so the density matrix is needed to describe it. Spontaneous emission is an essential ingredient for the dissipative nature of the optical forces.

2.1 Radiative Optical Forces

In the simplest case–the absorption of well-directed light from a laser beam–the momentum exchange between the light field and the atoms results in a force

$$\vec{F} = \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \hbar\vec{k}\gamma_p, \qquad (2)$$

where γ_p is the excitation rate of the atoms. The absorption leaves the atoms in their excited state, and if the light intensity is low enough that they are much more likely to return to the ground state by spontaneous emission than by stimulated emission, the resulting fluorescent light carries off momentum $\hbar k$ in a random direction. The momentum exchange from the fluorescence averages zero, so the net total force is given by Eq. (2).

The excitation rate γ_p depends on the laser detuning from atomic resonance $\delta \equiv \omega_\mathrm{l} - \omega_\mathrm{a}$, where ω_l is the laser frequency and ω_a is the atomic resonance frequency. This detuning is measured in the atomic rest frame, and it is necessary that the Doppler-shifted laser frequency in the rest frame of the moving atoms be used to calculate γ_p. In Sect. 2.3.3, we find that γ_p for a two-level atom is given by the Lorentzian

$$\gamma_p = \frac{s_0\gamma/2}{1 + s_0 + [2(\delta + \omega_\mathrm{D})/\gamma]^2}, \qquad (3)$$

where $\gamma \equiv 1/\tau$ is an angular frequency corresponding to the natural decay rate of the excited state. Here, $s_0 = I/I_\mathrm{s}$ is the ratio of the light intensity I to the saturation intensity $I_\mathrm{s} \equiv \pi hc/3\lambda^3\tau$, which is a few mW cm^{-2} for typical atomic transitions. The Doppler shift seen by the moving atoms is $\omega_\mathrm{D} = -\vec{k}\cdot\vec{v}$ (note that $\vec{k}$ opposite to $\vec{v}$ produces a positive Doppler shift for the atoms). The force is thus velocity-dependent and the experimenters' task is to exploit this dependence to the desired goal, for example, optical friction for laser cooling.

The maximum attainable deceleration is obtained for high intensities of light. High-intensity light can produce faster absorption, but it also causes equally fast stimulated emission; the combination produces neither deceleration nor cooling. The momentum transfer to the atoms by stimulated emission is in the opposite direction to what it was in absorption, resulting in a net transfer of zero momentum. At high intensity, Eq. (3) shows saturation of γ_p at $\gamma/2$, and since the force is given by Eq. (2), the deceleration saturates at a value $\vec{a}_\mathrm{max} = \hbar\vec{k}\gamma/2M$.

2.2 Dipole Optical Forces

While detuning $|\delta| \gg \gamma$, spontaneous emission may be much less frequent than stimulated emission, unlike the case of the dissipative radiative force that is necessary for laser cooling, given by Eqs. (2) and (3). In this case, absorption is most often followed by stimulated emission, and seems to produce zero momentum transfer because the stimulated light has the same momentum as the exciting light. However, if the optical field has beams with at least two different $\vec{k}$-vectors present, such as in counterpropagating beams, absorption from one beam followed by stimulated emission into the other indeed produces a nonzero momentum exchange. The result is called the dipole force, and is reversible and hence conservative, so it cannot be used for laser cooling.

The dipole force is more easily calculated from an energy picture than from a momentum picture. The force then derives

from the gradient of the potential of an atom in an inhomogeneous light field, which is appropriate because the force is conservative. The potential arises from the shift of the atomic energy levels in the light field, appropriately called the "light shift", and is found by direct solution of the Schrödinger equation for a two-level atom in a monochromatic plane wave. After making both the dipole and rotating wave approximations, the Hamiltonian can be written as

$$\mathcal{H} = \frac{\hbar}{2}\begin{bmatrix} -2\delta & \Omega \\ \Omega^* & 0 \end{bmatrix} \tag{4}$$

where the Rabi frequency is $|\Omega| = \gamma\sqrt{s_0/2}$ for a single traveling laser beam. Solution of Eq. (4) for its eigenvalues provides the dressed state energies that are light-shifted by

$$\omega_{\mathrm{ls}} = \frac{\left[\sqrt{|\Omega|^2 + \delta^2} - \delta\right]}{2}. \tag{5}$$

For sufficiently large detuning $|\delta| \gg |\Omega|$, approximation of Eq. (5) leads to $\omega_{\mathrm{ls}} \approx |\Omega|^2/4\delta = \gamma^2 s_0/8\delta$.

In a standing wave in 1-D with $|\delta| \gg |\Omega|$, the light shift ω_{ls} varies sinusoidally from node to antinode. When δ is sufficiently large, the spontaneous emission rate may be negligible compared with that of stimulated emission, so that $\hbar\omega_{\mathrm{ls}}$ may be treated as a potential U. The resulting dipole force is

$$\vec{F} = -\vec{\nabla}U = -\frac{\hbar\gamma^2}{8\delta I_{\mathrm{s}}}\vec{\nabla}I, \tag{6}$$

where I is the total intensity distribution of the standing-wave light field of period $\lambda/2$. For such a standing wave, the optical electric field (and the Rabi frequency) at the antinodes is double that of each traveling wave that composes it, and so the total intensity $I_{\max}$ at the antinodes is four times that of the single traveling wave.

2.3 Density Matrix Description of Optical Forces

2.3.1 Introduction

Use of the density matrix ρ for pure states provides an alternative description to the more familiar one that uses wave functions and operators but adds nothing new. Its equation of motion is $i\hbar(\mathrm{d}\rho/\mathrm{d}t) = [\mathcal{H}, \rho]$, and can be derived directly from the Schrödinger equation. Moreover, it is a straightforward exercise to show that the expectation value of any operator $\mathcal{A}$ that represents an observable is $\langle\mathcal{A}\rangle = \mathrm{tr}\,(\rho\mathcal{A})$.

Application of the Ehrenfest theorem gives the expectation value of the force as $\langle F\rangle = -\mathrm{tr}\,(\rho\nabla\mathcal{H})$. Beginning with the two-level atom Hamiltonian of Eq. (4), we find the force in 1-D to be

$$\langle F\rangle = \hbar\left(\frac{\partial\Omega}{\partial z}\rho_{eg}^* + \frac{\partial\Omega^*}{\partial z}\rho_{eg}\right). \tag{7}$$

Thus, $\langle F\rangle$ depends only on the off-diagonal elements $\rho_{eg} = \rho_{ge}^*$, terms that are called *the optical coherences*.

2.3.2 Open Systems and the Dissipative Force

The real value of the density matrix formalism for atom-light interactions is its ability to deal with open systems. By not including the fluorescent light that is lost from an atom-laser system undergoing cooling, a serious omission is being made in the discussion above. That is, the closed system of atom plus laser light that can be described by Schrödinger wave functions and is thus in a pure state, undergoes evolution to a "mixed" state by virtue of the spontaneous emission. This omission can be rectified by simple ad hoc additions

to the equation of motion, and the result is called the optical Bloch equations (OBE). These are written explicitly as

$$\begin{aligned}
\frac{d\rho_{gg}}{dt} &= +\gamma\rho_{ee} + \frac{i}{2}(\Omega^*\tilde{\rho}_{eg} - \Omega\tilde{\rho}_{ge}) \\
\frac{d\rho_{ee}}{dt} &= -\gamma\rho_{ee} + \frac{i}{2}(\Omega\tilde{\rho}_{ge} - \Omega^*\tilde{\rho}_{eg}) \\
\frac{d\tilde{\rho}_{ge}}{dt} &= -\left(\frac{\gamma}{2} + i\delta\right)\tilde{\rho}_{ge} + \frac{i}{2}\Omega^*(\rho_{ee} - \rho_{gg}) \\
\frac{d\tilde{\rho}_{eg}}{dt} &= -\left(\frac{\gamma}{2} - i\delta\right)\tilde{\rho}_{eg} + \frac{i}{2}\Omega(\rho_{gg} - \rho_{ee}),
\end{aligned} \tag{8}$$

where $\tilde{\rho}_{eg} \equiv \rho_{eg}e^{-i\delta t}$ for the coherences.

In these equations, the terms proportional to the spontaneous decay rate γ have been put in "by hand", that is, they have been introduced into the OBEs to account for the effects of spontaneous emission. The spontaneous emission is irreversible and accounts for the dissipation of the cooling process. For the ground state, the decay of the excited state leads to an increase of its population ρ_{gg} proportional to $\gamma\rho_{ee}$, whereas for the excited state, it leads to a decrease of ρ_{ee}, also proportional to $\gamma\rho_{ee}$. These equations have to be solved in order to evaluate the optical force on the atoms.

2.3.3 Solution of the OBEs in Steady State

In most cases, the laser light is applied for a period long compared to the typical evolution times of atom-light interaction, that is, the lifetime of the excited state $\tau = 1/\gamma$. Thus, only the steady state solution of the OBEs have to be considered, and these are found by setting the time derivatives in Eq. (8) to zero. Then the probability ρ_{ee} to be in the excited state is found to be

$$\rho_{ee} = \frac{\gamma_p}{\gamma} = \frac{s_0/2}{1 + s_0 + (2\delta/\gamma)^2} = \frac{s/2}{1+s}, \tag{9}$$

where $s \equiv s_0/[1 + (2\delta/\gamma)^2]$ is the off-resonance saturation parameter. The excited-state population ρ_{ee} increases linearly with the saturation parameter s for small values of s, but for s of the order of unity, the probability starts to saturate to a value of 1/2. The detuning dependence of γ_p (see Eq. 3) showing this saturation for various values of s_0 is depicted in Fig. 1.

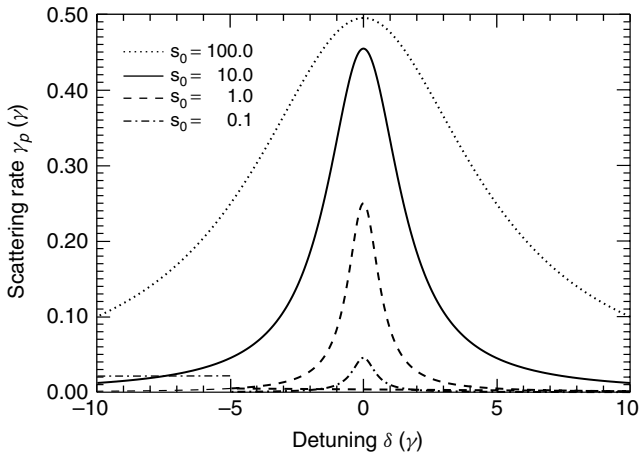

Fig. 1 Excitation rate γ_p as a function of the detuning δ for several values of the saturation parameter s_0. Note that for $s_0 > 1$, the line profiles start to broaden substantially from power broadening

2.3.4 Radiative and Dipole Forces

Some insight into these forces emerges by expressing the gradient of the Rabi frequency of Eq. (7) in terms of a real and imaginary part so that $(\partial\Omega/\partial z) = (q_r + iq_i)\Omega$. Then Eq. (7) becomes

$$F = \hbar q_r(\Omega\rho_{eg}^* + \Omega^*\rho_{eg}) + i\hbar q_i(\Omega\rho_{eg}^* - \Omega^*\rho_{eg}) \quad (10)$$

Thus, the first term of the force is related to the dispersive part of the atom-light interaction, whereas the second term is related to the absorptive part of the atom-light interaction.

To appreciate the utility of the separation of $\nabla\Omega$ into real and imaginary parts, consider the interaction of atoms with a traveling plane wave $\mathcal{E}(z) = \mathcal{E}_0(e^{i(kz-\omega t)} + \text{c.c.})/2$. In this case, $q_r = 0$ and $q_i = k$, and so the force is caused only by absorption. The force is given by $F_{sp} = \hbar k\gamma\rho_{ee}$ and is the radiative force of Eqs. (2) and (3).

For the case of counterpropagating plane waves, there is a standing wave whose electric field is $\mathcal{E}(z) = \mathcal{E}_0\cos(kz)(e^{-i\omega t} + \text{c.c.})$. Thus, $q_r = -k\tan(kz)$ and $q_i = 0$, so there is only the dispersive part of the force, given by

$$F_{dip} = \frac{2\hbar k\delta s_0 \sin 2kz}{1 + 4s_0\cos^2 kz + (2\delta/\gamma)^2}. \quad (11)$$

This replaces Eq. (6) for the dipole force and removes the restriction $|\delta| \gg |\Omega|$, thereby including saturation effects. Even though the average of this force over a wavelength vanishes, it can be used to trap atoms in a region smaller than the wavelength of the light.

2.3.5 Force on Moving Atoms

In order to show how these forces can be used to cool atoms, one has to consider the force on moving atoms. For the case of atomic velocities that are small compared with γ/k, the motion can be treated as a small perturbation in the atomic evolution that occurs on the time scale $1/\gamma$. Then the first-order result is given by

$$\frac{d\Omega}{dt} = \frac{\partial\Omega}{\partial t} + v\frac{\partial\Omega}{\partial z} = \frac{\partial\Omega}{\partial t} + v(q_r + iq_i)\Omega. \quad (12)$$

For the case of atoms moving in a standing wave, this results in the same damping force as Eq. (13) below.

3 Laser Cooling

3.1 Slowing Atomic Beams

Among the earliest laser cooling experiments was the deceleration of atoms in a beam [2]. The authors exploited the Doppler shift to make the momentum exchange (hence the force) velocity dependent. It worked by directing a laser beam opposite an atomic beam so that the atoms could absorb light, and hence momentum $\hbar k$, very many times along their paths through the apparatus as shown in Fig. 2 [2, 3]. Of course, excited-state atoms cannot absorb light efficiently from the laser that excited them, so between absorptions they must return to the ground state by spontaneous decay, accompanied by the emission of fluorescent light. The spatial symmetry of the emitted fluorescence results in an average of zero net momentum transfer from many such fluorescence events. Thus, the net force on the atoms is in the direction of the laser beam, and the maximum deceleration is limited by the spontaneous emission rate γ.

Since the maximum deceleration $\vec{a}_{max} = \hbar\vec{k}\gamma/2M$ is fixed by atomic parameters, it is straightforward to calculate the minimum

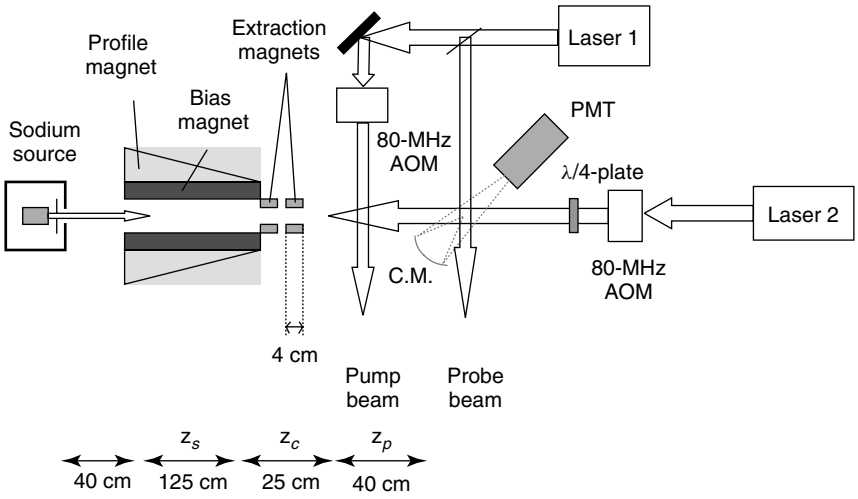

Fig. 2 Schematic diagram of apparatus for beam slowing. The tapered magnetic field is produced by layers of varying length on the solenoid

stopping length $L_{\min}$ and time $t_{\min}$ for the rms velocity of atoms $\bar{v} = 2\sqrt{k_B T/M}$ at the source temperature. The result is $L_{\min} = \bar{v}^2/2a_{\max}$ and $t_{\min} = \bar{v}/a_{\max}$. In Table 1 are some of the parameters for slowing a few atomic species of interest from the peak of the thermal velocity distribution.

Maximizing the scattering rate γ_p requires $\delta = -\omega_D$ in Eq. (3). If δ is chosen for a particular atomic velocity in the beam, then as the atoms slow down, their changing Doppler shift will take them out of resonance. They will eventually cease deceleration after their Doppler shift has been decreased by a few times the power-broadened width $\gamma' = \gamma\sqrt{1+s_0}$, corresponding to Δv of a few times $v_c = \gamma'/k$. Although this Δv of a few m s^{-1} is considerably larger than the typical atomic recoil velocity $v_r = \hbar k/M$ of a few cm s^{-1}, it is still only a small fraction of the average thermal velocity $\bar{v}$ of the atoms, such that significant further cooling or deceleration cannot be accomplished.

Tab. 1 Parameters of interest for slowing various atoms. The stopping length L_{min} and time t_{min} are minimum values. The oven temperature T_{oven} that determines the peak velocity is chosen to give a vapor pressure of 1 Torr. Special cases are H at 1000 K for dissociation of H_2 into atoms, and He in the metastable triplet state, for which two rows are shown: one for a 4-K source and another for the typical discharge temperature

Atom	T_{oven} *[K]*	$\bar{v}$ *[m s^{-1}]*	L_{min} *[m]*	t_{min} *[ms]*
H	1000	5000	0.012	0.005
He*	4	158	0.03	0.34
He*	650	2013	4.4	4.4
Li	1017	2051	1.15	1.12
Na	712	876	0.42	0.96
K	617	626	0.77	2.45
Rb	568	402	0.75	3.72
Cs	544	319	0.93	5.82

In order to achieve deceleration that changes the atomic speeds by hundreds of m s^{-1}, it is necessary to maintain $(\delta + \omega_D) \ll \gamma$ by compensating such large changes of the Doppler shift. This can be

done by changing ω_D through the angular dependence of $\vec{k} \cdot \vec{v}$, or changing δ either via ω_l or ω_a. The two most common methods for maintaining this resonance are sweeping the laser frequency ω_l along with the changing ω_D of the decelerating atoms [4–6], or by spatially varying the atomic resonance frequency with an inhomogeneous d.c magnetic field to keep the decelerating atoms in resonance with the fixed frequency laser [2, 3, 7].

The use of a spatially varying magnetic field to tune the atomic levels along the beam path was the first method to succeed in slowing atoms [2, 3]. It works as long as the Zeeman shifts of the ground and excited states are different so that the resonant frequency is shifted. The field can be tailored to provide the appropriate Doppler shift along the moving atom's path. A solenoid that can produce such a spatially varying field has layers of decreasing lengths. The technical problem of extracting the beam of slow atoms from the end of the solenoid can be simplified by reversing the field gradient and choosing a transition whose frequency decreases with increasing field [9].

For alkali atoms such as sodium, a time-of-flight (TOF) method can be used to measure the velocity distribution of atoms in the beam [8]. It employs two additional beams labeled pump and probe from Laser 1 as shown in Fig. 2. Because these beams cross the atomic beam at 90°, $\omega_D = -\vec{k} \cdot \vec{v} = 0$, and they excite atoms at all velocities. The pump beam is tuned to excite and empty a selected ground hyperfine state (hfs), and it transfers more than 98% of the population as the atoms pass through its 0.5 mm width. To measure the velocity distribution of atoms in the selected hfs, this pump laser beam is interrupted for a period of $\Delta t = 10 - 50$ μs with an acoustic optical modulator (AOM). A pulse of atoms in the selected hfs passes the pump region and travels to the probe beam. The time dependence of the fluorescence induced by the probe laser, tuned to excite the selected hfs, gives the time of arrival, and this signal is readily converted to a velocity distribution. Figure 3 shows the measured velocity distribution of the atoms slowed by Laser 2.

3.2 Optical Molasses

3.2.1 Doppler Cooling

A different kind of radiative force arises in low intensity, counterpropagating light beams that form a weak standing wave. It is straightforward to calculate the radiative force on atoms moving in such a standing wave using Eq. (3). In the low intensity case where stimulated emission is not important, the forces from the two light beams are simply added to give $\vec{F}_{OM} =$

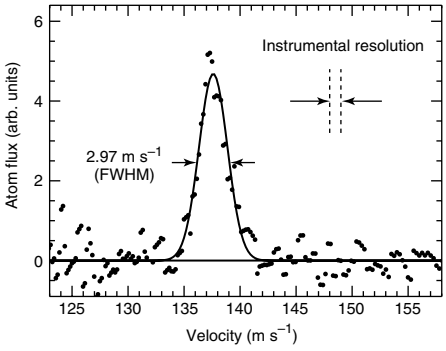

Fig. 3 The velocity distribution measured with the TOF method. The experimental width of approximately $\frac{1}{6}(\gamma/k)$ is shown by the dashed vertical lines between the arrows. The Gaussian fit through the data yields an FWHM (full width at half maximum) of 2.97 m s^{-1} (figure taken from Molenaar, P. A., vander Straten, P., Heideman, H. G. M., Metcalf, H. (**1997**), *Phys. Rev. A* **55**, 605–614)

$\vec{F}_+ + \vec{F}_-$, where $\vec{F}_\pm$ are found from Eqs. (2) and (3). Then the sum of the two forces is

$$\vec{F}_{\mathrm{OM}} \cong \frac{8\hbar k^2 \delta s_0 \vec{v}}{\gamma(1 + s_0 + (2\delta/\gamma)^2)^2} \equiv -\beta\vec{v}, \tag{13}$$

where terms of order $(kv/\gamma)^4$ and higher have been neglected. The force is proportional to velocity for small enough velocities, resulting in viscous damping for $\delta < 0$ [10, 11] that gives this technique the name "optical molasses" (OM).

These forces are plotted in Fig. 4. For $\delta < 0$, this force opposes the velocity and therefore viscously damps the atomic motion. The force $\vec{F}_{\mathrm{OM}}$ has maxima near $v \approx \pm\gamma\sqrt{s_0 + 1}/2k$ and decreases rapidly for larger velocities.

3.2.2 Doppler Cooling Limit

If there were no other influence on the atomic motion, all atoms would quickly decelerate to $v = 0$ and the sample would reach $T = 0$, a clearly unphysical result. In laser cooling and related aspects of optical control of atomic motion, the forces arise because of the exchange of momentum between the atoms and the laser field. These necessarily discrete steps of size $\hbar k$ constitute a heating mechanism that must be considered.

Since the atomic momentum changes by $\hbar k$, their kinetic energy changes on an average by at least the recoil energy $E_{\mathrm{r}} = \hbar^2 k^2/2M = \hbar\omega_{\mathrm{r}}$. This means that the average frequency of each absorption is at least $\omega_{\mathrm{abs}} = \omega_{\mathrm{a}} + \omega_{\mathrm{r}}$. Similarly, the energy $\hbar\omega_{\mathrm{a}}$ available from each spontaneous decay must be shared between the outgoing light and the kinetic energy of the atom recoiling with momentum $\hbar k$. Thus, the average frequency of each emission is $\omega_{\mathrm{emit}} = \omega_{\mathrm{a}} - \omega_{\mathrm{r}}$. Therefore, the light field loses an average energy of $\hbar(\omega_{\mathrm{abs}} - \omega_{\mathrm{emit}}) = 2\hbar\omega_{\mathrm{r}}$ for each scattering event. This loss occurs at a rate of $2\gamma_p$ (two beams), and the energy is converted to atomic kinetic energy because the atoms

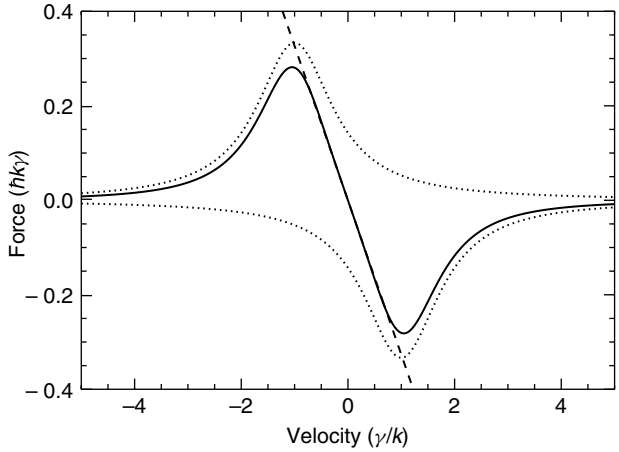

Fig. 4 Velocity dependence of the optical damping forces for 1-D optical molasses. The two dotted traces show the force from each beam, and the solid curve is their sum. The straight line shows how this force mimics a pure damping force over a restricted velocity range. These are calculated for $s_0 = 2$ and $\delta = -\gamma$, so there is some power broadening evident

recoil from each event. The atomic sample is thereby heated because these recoils are in random directions.

The competition between this heating and the damping force of Eq. (13) results in a nonzero kinetic energy in steady state, where the rates of heating and cooling are equal. Equating the cooling rate, $\vec{F}_{\mathrm{OM}} \cdot \vec{v}$, to the heating rate, $4\hbar\omega_{\mathrm{r}}\gamma_p$, we find the steady state kinetic energy to be $(\hbar\gamma/8)(2|\delta|/\gamma + \gamma/2|\delta|)$. This result is dependent on $|\delta|$, and has a minimum at $2|\delta|/\gamma = 1$, whence $\delta = -\gamma/2$. The temperature found from the kinetic energy is then $T_{\mathrm{D}} = \hbar\gamma/2k_{\mathrm{B}}$, where T_{D} is called the Doppler temperature or the Doppler cooling limit. For ordinary atomic transitions, T_{D} is typically below 1 mK.

Another instructive way to determine T_{D} is to note that the average momentum transfer of many spontaneous emissions is zero, but the rms scatter of these about zero is finite. One can imagine these decays as causing a random walk in momentum space, similar to Brownian motion in real space, with step size $\hbar k$ and step frequency $2\gamma_p$, where the factor of 2 arises because of the two beams. The random walk results in an evolution of the momentum distribution as described by the Fokker–Planck equation, and can be used for a more formal treatment of laser cooling. It results in diffusion in momentum space with diffusion coefficient $D_0 \equiv 2(\Delta p)^2/\Delta t = 4\gamma_p(\hbar k)^2$. Then the steady state temperature is given by $k_{\mathrm{B}}T = D_0/\beta$. This turns out to be $\hbar\gamma/2$ as above for the case $s_0 \ll 1$ when $\delta = -\gamma/2$. This remarkable result predicts that the final temperature of atoms in OM is independent of the optical wavelength, atomic mass, and laser intensity (as long as it is not too large).

3.2.3 Atomic Beam Collimation – One-dimensional Optical Molasses – Beam Brightening

When an atomic beam crosses a 1-D OM as shown in Fig. 5, the transverse motion of the atoms is quickly damped while the longitudinal component is essentially unchanged. This transverse cooling of

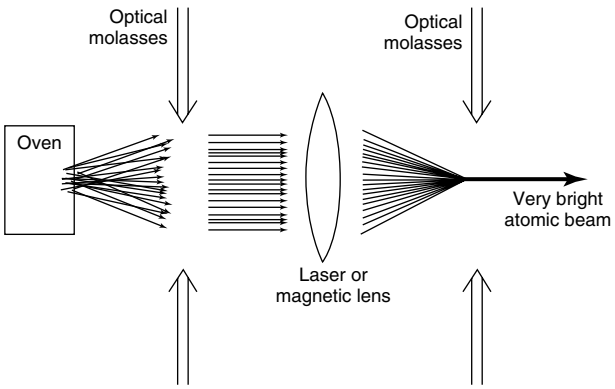

Fig. 5 Scheme for optical brightening of an atomic beam. First, the transverse velocity components of the atoms are damped out by an optical molasses, then the atoms are focused to a spot, and finally the atoms are recollimated in a second optical molasses (figure taken from Sheehy, B., Shang, S. Q., van der Straten, P., Metcalf, H. (**1990**), *Chem. Phys.* **145**, 317–325)

an atomic beam is an example of a method that can actually increase its brightness (atoms/s-sr-cm^2) because such active collimation uses dissipative forces to compress the phase space volume occupied by the atoms. By contrast, the usual focusing or collimation techniques for light beams and most particle beams is restricted to selection by apertures or conservative forces that preserve the phase space density of atoms in the beam.

This velocity compression at low intensity in one dimension can be easily estimated for two-level atoms to be about $v_c/v_D = \sqrt{\gamma/\omega_r} \equiv \sqrt{1/\varepsilon}$. Here v_D is the velocity associated with the Doppler limit for laser cooling discussed above: $v_D = \sqrt{\hbar\gamma/2M}$. For Rb, v_D = 12 cm s^{-1}, $v_c = \gamma/k \simeq 4.6$ m s^{-1}, $\omega_r \simeq 2\pi \times 3.8$ kHz, and $1/\varepsilon \simeq 1600$. (The parameter ε characterizes optical forces on atoms.) Including two transverse directions along with the longitudinal slowing and cooling discussed above, the decrease in three-dimensional 3-D phase space volume for laser cooling of an Rb atomic beam from the momentum contribution alone can exceed 10^6. Clearly optical techniques can create atomic beams enormously more times intense than ordinary thermal beams and also many orders of magnitude brighter.

3.2.4 Experiments in Three-dimensional Optical Molasses

By using three intersecting orthogonal pairs of oppositely directed beams, the movement of atoms in the intersection region can be severely restricted in all 3-D, and many atoms can thereby be collected and cooled in a small volume.

Even though atoms can be collected and cooled in the intersection region, it is important to stress that this is *not* a trap (see Sect. 4 below), that is, atoms that wander away from the center experience no force directing them back. They are allowed to diffuse freely and even escape, as long as there is enough time for their very slow diffusive movement to allow them to reach the edge of the region of intersection of the laser beams. Since the atomic velocities are randomized during the damping time $M/\beta = 2/\omega_r$, atoms execute a random walk in position space with a step size of $2v_D/\omega_r = \lambda/(\pi\sqrt{2\varepsilon}) \cong$ few μm. To diffuse a distance of 1 cm requires about 10^7 steps or about 30 s [13, 14].

In 1985, the group at Bell Labs was the first to observe 3-D OM [11]. Preliminary measurements of the average kinetic energy of the atoms were done by blinking off the laser beams for a fixed interval. Comparison of the brightness of the fluorescence before and after the turnoff was used to calculate the fraction of atoms that left the region while it was in the dark. The dependence of this fraction on the duration of the dark interval was used to estimate the velocity distribution and hence the temperature. This method, which is usually referred to as *release and recapture*, is specifically designed to measure the temperature of the atoms, since the usual way of measuring temperatures cannot be applied to an atomic cloud of a few million atoms. The result was consistent with T_D as calculated from the Doppler theory, as described in Sect. 3.2.2.

Later a more sensitive ballistic technique was devised at NIST that showed the astounding result that the temperature of the atoms in OM was very much lower than T_D [15]. These experiments also found that OM was less sensitive to perturbations and more tolerant of alignment errors than was predicted by Doppler theory. For example, if the intensities of the two counterpropagating laser beams forming

an OM were unequal, then the force on the atoms at rest would not vanish, but the force on the atoms with some nonzero drift velocity *would* vanish. This drift velocity can be easily calculated by using unequal intensities s_{0+} and s_{0-}, to derive an analog of Eq. (13). Thus, atoms would drift out of an OM, and the calculated rate would be much faster than observed by deliberately unbalancing the beams in the experiments [16].

3.3 Cooling Below the Doppler Limit

3.3.1 Introduction

It was an enormous surprise to observe that the ballistically measured temperature of the Na atoms was as much as 10 times *lower* than $T_D = 240\ \mu K$ [15], the temperature minimum calculated from theory. This breaching of the Doppler limit forced the development of an entirely new picture of OM that accounts for the fact that in 3-D, a two-level picture of atomic structure is inadequate. The multilevel structure of atomic states, and optical pumping among these sublevels, must be considered in the description of 3-D OM.

In response to these surprising measurements of temperatures below T_D, two groups developed a model of laser cooling that could explain the lower temperatures [17, 18]. The key feature of this model that distinguishes it from the earlier picture is the inclusion of the multiplicity of sublevels that make up an atomic state (e.g., Zeeman and hfs). The dynamics of optically pumping the moving atoms among these sublevels provides the new mechanism for producing ultralow temperatures [19].

The dominant feature of these models is the nonadiabatic response of moving atoms to the light field. Atoms at rest in a steady state have ground-state orientations caused by optical pumping processes that distribute the populations over the different ground-state sublevels. In the presence of polarization gradients, these orientations reflect the local light field. In the low-light-intensity regime, the orientation of stationary atoms is completely determined by the ground-state distribution; the optical coherences and the excited-state population follow the ground-state distribution adiabatically.

For atoms moving in a light field that varies in space, optical pumping acts to adjust the atomic orientation to the changing conditions of the light field. In a weak pumping process, the orientation of moving atoms always lags behind the orientation that would exist for stationary atoms. It is this phenomenon of nonadiabatic following that is the essential feature of the new cooling process.

Production of spatially dependent optical pumping processes can be achieved in several different ways. As an example, consider two counterpropagating laser beams that have orthogonal polarizations, as discussed below. The superposition of the two beams results in a light field having a polarization that varies on the wavelength scale along the direction of the laser beams. Laser cooling by such a light field is called polarization gradient cooling. In a 3-D OM, the transverse wave character of light requires that the light field always has polarization gradients.

3.3.2 Linear ⊥ Linear Polarization Gradient Cooling

One of the most instructive models for discussion of sub-Doppler laser cooling was introduced in Ref. [17] and very well described in Ref. [19]. If the polarizations of two counterpropagating laser beams are identical, the two beams interfere

and produce a standing wave. When the two beams have orthogonal linear polarizations (same frequency ω_l) with their $\hat{\mathcal{E}}$ vectors perpendicular (e.g., $\hat{x}$ and $\hat{y}$), the configuration is called lin $\perp$ lin or lin-perp-lin. Then the total field is the sum of the two counterpropagating beams given by

$$\begin{aligned}\vec{\mathcal{E}} &= \mathcal{E}_0\,\hat{x}\cos(\omega_l t - kz) + \mathcal{E}_0\hat{y}\cos(\omega_l t + kz)\\ &= \mathcal{E}_0[(\hat{x}+\hat{y})\cos\omega_l t\cos kz\\ &\quad + (\hat{x}-\hat{y})\sin\omega_l t\sin kz]. \end{aligned} \tag{14}$$

At the origin, where $z = 0$, this becomes

$$\vec{\mathcal{E}} = \mathcal{E}_0(\hat{x}+\hat{y})\cos\omega_l t, \tag{15}$$

which corresponds to linearly polarized light at an angle $+\pi/4$ to the x-axis. The amplitude of this field is $\sqrt{2}\mathcal{E}_0$. Similarly, for $z = \lambda/4$, where $kz = \pi/2$, the field is also linearly polarized but at an angle $-\pi/4$ to the x-axis.

Between these two points, at $z = \lambda/8$, where $kz = \pi/4$, the total field is

$$\begin{aligned}\vec{\mathcal{E}} = \mathcal{E}_0\Big[&\hat{x}\sin\Big(\omega_l t + \frac{\pi}{4}\Big)\\ &- \hat{y}\cos\Big(\omega_l t + \frac{\pi}{4}\Big)\Big]. \end{aligned} \tag{16}$$

Since the $\hat{x}$ and $\hat{y}$ components have sine and cosine temporal dependence, they are $\pi/2$ out of phase, and so Eq. (16) represents circularly polarized light rotating about the z-axis in the negative sense. Similarly, at $z = 3\lambda/8$ where $kz = 3\pi/4$, the polarization is circular but in the positive sense. Thus, in this lin $\perp$ lin scheme the polarization cycles from linear to circular to orthogonal linear to opposite circular in the space of only half a wavelength of light, as shown in Fig. 6. It truly has a very strong polarization gradient.

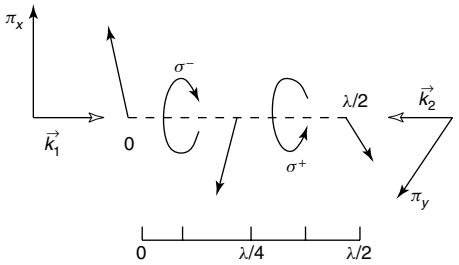

Fig. 6 Polarization gradient field for the lin $\perp$ lin configuration

Since the coupling of the different states of multilevel atoms to the light field depends on its polarization, atoms moving in a polarization gradient will be coupled differently at different positions, and this will have important consequences for laser cooling. For the $J_g = 1/2 \rightarrow J_e = 3/2$ transition (one of the simplest transitions that show sub-Doppler cooling [20]), the optical pumping process in purely σ^+ light drives the ground-state population to the $M_g = +1/2$ sublevel. This optical pumping occurs because absorption always produces $\Delta M = +1$ transitions, whereas the subsequent spontaneous emission produces $\Delta M = \pm 1,\ 0$. Thus, the average $\Delta M \geq 0$ for each scattering event. For σ^--light, the population is pumped toward the $M_g = -1/2$ sublevel. Thus, atoms traveling through only a half wavelength in the light field, need to readjust their population completely from $M_g = +1/2$ to $M_g = -1/2$ and back again.

The interaction between nearly resonant light and atoms not only drives transitions between atomic energy levels but also shifts their energies. This light shift of the atomic energy levels, discussed in Sect. 2.2, plays a crucial role in this scheme of sub-Doppler cooling, and the changing polarization has a strong influence on the light shifts. In the low-intensity limit of two laser beams, each of intensity $s_0 I_s$, the

light shifts ΔE_g of the ground magnetic substates are given by a slight variation of the approximation to Eq. (5) that accounts for the multilevel structure of the atoms. We write

$$\Delta E_g = \frac{\hbar s_0 C_{ge}^2 \gamma^2}{8\delta}, \qquad (17)$$

where C_{ge} is the Clebsch–Gordan coefficient that describes the coupling between the particular levels of the atom and the light field.

In the present case of orthogonal linear polarizations and $J = 1/2 \rightarrow 3/2$, the light shift for the magnetic substate $M_g = 1/2$ is three times larger than that of the $M_g = -1/2$ substate when the light field is completely σ^+. On the other hand, when an atom moves to a place where the light field is σ^-, the shift of $M_g = -1/2$ is three times larger. So, in this case, the optical pumping discussed above causes a larger population to be there in the state with the larger light shift. This is generally true for any transition $J_g \rightarrow J_e = J_g + 1$. A schematic diagram showing the populations and light shifts for this particular case of negative detuning is illustrated in Fig. 7.

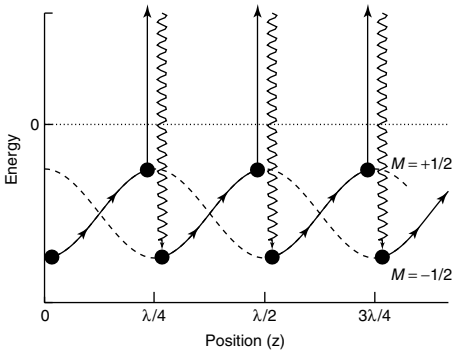

Fig. 7 The spatial dependence of the light shifts of the ground-state sublevels of the $J = 1/2 \Leftrightarrow 3/2$ transition for the case of lin $\perp$ lin polarization configuration. The arrows show the path followed by atoms being cooled in this arrangement. Atoms starting at $z = 0$ in the $M_g = +1/2$ sublevel must climb the potential hill as they approach the $z = \lambda/4$ point where the light becomes σ^- polarized, and they are optically pumped to the $M_g = -1/2$ sublevel. Then they must begin climbing another hill toward the $z = \lambda/2$ point where the light is σ^+ polarized and they are optically pumped back to the $M_g = +1/2$ sublevel. The process repeats until the atomic kinetic energy is too small to climb the next hill. Each optical pumping event results in the absorption of light at a frequency lower than emission, thus dissipating energy to the radiation field

3.3.3 Origin of the Damping Force

To discuss the origin of the cooling process in this polarization gradient scheme, consider atoms with a velocity v at a position where the light is σ^+-polarized, as shown at the lower left of Fig. 7. The light optically pumps such atoms to the strongly negative light-shifted $M_g = +1/2$ state. In moving through the light field, atoms must increase their potential energy (climb a hill) because the polarization of the light is changing and the state $M_g = 1/2$ becomes less strongly coupled to the light field. After traveling a distance $\lambda/4$, atoms arrive at a position where the light field is σ^--polarized, and are optically pumped to $M_g = -1/2$, which is now lower than the $M_g = 1/2$ state. Again, the moving atoms are at the bottom of a hill and start to climb. In climbing the hills, the kinetic energy is converted to potential energy, and in the optical pumping process, the potential energy is radiated away because the spontaneous emission is at a higher frequency than the absorption (see Fig. 7). Thus, atoms seem to be always climbing hills and losing energy in the process. This process brings to mind a Greek myth, and is thus called ''Sisyphus laser cooling''.

The cooling process described above is effective over a limited range of atomic velocities. The force is maximum for atoms that undergo one optical pumping process while traveling over a distance $\lambda/4$. Slower atoms will not reach the hilltop before the pumping process occurs and faster atoms will travel a longer distance before being pumped toward the other sublevel, so $\Delta E/\Delta z$ is smaller. In both cases, the energy loss is smaller and therefore the cooling process less efficient. Nevertheless, the damping constant β for this process is much larger than for Doppler cooling, and therefore the final steady state temperature is lower [17, 19].

In the experiments of Ref. [21], the temperature was measured in a 3-D molasses under various configurations of the polarization. Temperatures were measured by a ballistic technique, in which the flight time of the released atoms was measured as they fell through a probe a few centimeters below the molasses region. The lowest temperature obtained was 3 μK, which is a factor 40 below the Doppler temperature and a factor 15 above the recoil temperature of Cs.

3.3.4 The Limits of Sisyphus Laser Cooling

The extension of the kind of thinking about cooling limits in the case of Doppler cooling to the case of the sub-Doppler processes must be done with some care, because a naive application of similar ideas would lead to an arbitrarily low final temperature. In the derivation in Sect. 3.2.2, it is explicitly assumed that each scattering event changes the atomic momentum p by an amount that is a small fraction of p and this clearly fails when the velocity is reduced to the region of $v_r \equiv \hbar k/M$.

This limitation of the minimum steady state value of the average kinetic energy to a few times $E_r \equiv k_B T_r/2 = M v_r^2/2$ is intuitively comforting for two reasons. First, one might expect that the last spontaneous emission in a cooling process would leave atoms with a residual momentum of the order of $\hbar k$, since there is no control over its direction. Thus, the randomness associated with this would put a lower limit of $v_{min} \sim v_r$ on such cooling. Second, the polarization gradient cooling mechanism described above requires that atoms be localizable within the scale of $\sim \lambda/2\pi$ in order to be subject to only a single polarization in the spatially inhomogeneous light field. The uncertainty principle then requires that these atoms have a momentum spread of at least $\hbar k$.

The recoil limit discussed here has been surpassed by evaporative cooling of trapped atoms [22] and two different optical cooling methods, neither of which can be described in simple notions. One of these uses optical pumping into a velocity-selective dark state [23]. The other one uses carefully chosen, counterpropagating, laser pulses to induce velocity-selective Raman transitions, and is called *Raman cooling* [24].

4 Traps for Neutral Atoms

In order to confine any object, it is necessary to exchange kinetic for potential energy in the trapping field, and in neutral atom traps, the potential energy must be stored as internal atomic energy. Thus, practical traps for ground-state neutral atoms are necessarily very shallow compared with thermal energy because the energy-level shifts that result from convenient size fields are typically considerably smaller than $k_B T$ for $T = 1$ K. Neutral atom trapping therefore depends

on substantial cooling of a thermal atomic sample, and is often connected with the cooling process. In most practical cases, atoms are loaded from magneto-optical traps (MOTs) in which they have been efficiently accumulated and cooled to mK temperatures (see Sect. 4.3), or from optical molasses, in which they have been optically cooled to μK temperatures (see Sect. 3.2).

The small depth of typical neutral atom traps dictates stringent vacuum requirements because an atom cannot remain trapped after a collision with a thermal energy background gas molecule. Since these atoms are vulnerable targets for thermal energy background gas, the mean free time between collisions *must* exceed the desired trapping time. The cross section for destructive collisions is quite large because even a gentle collision (i.e., large impact parameter) can impart enough energy to eject an atom from a trap. At pressure P sufficiently low to be of practical interest, the trapping time is $\sim (10^{-8}/P)$ s, where P is in Torr.

4.1 Dipole Force Optical Traps

4.1.1 Single-beam Optical Traps for Two-level Atoms

The simplest imaginable optical trap consists of a single, strongly focused Gaussian laser beam (see Fig. 8) [25, 26] whose intensity at the focus varies transversely with r as

$$I(r) = I_0 e^{-2r^2/w_0^2}, \qquad (18)$$

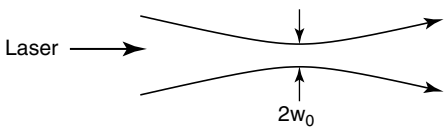

Fig. 8 A single focused laser beam produces the simplest type of optical trap

where w_0 is the beam waist size. Such a trap has a well-studied and important macroscopic classical analog in a phenomenon called *optical tweezers* [27–29].

With the laser light tuned below resonance ($\delta < 0$), the ground-state light shift is negative everywhere, but largest at the center of the Gaussian beam waist. Ground-state atoms, therefore, experience a force attracting them toward this center, given by the gradient of the light shift, which is found from Eq. (5), and for $\delta/\gamma \gg s_0$ is given by Eq. (6). For the Gaussian beam, this transverse force at the waist is harmonic for small r and is given by

$$F \simeq \frac{\hbar\gamma^2}{4\delta}\frac{I_0}{I_s}\frac{r}{w_0^2}e^{-2r^2/w_0^2}. \qquad (19)$$

In the longitudinal direction, there is also an attractive force but it is more complicated and depends on the details of the focusing. Thus, this trap produces an attractive force on the atoms in three dimensions.

Although it may appear that the trap does not confine atoms longitudinally because of the radiation pressure along the laser beam direction, careful choice of the laser parameters can indeed produce trapping in 3-D. This can be accomplished because the radiation pressure force decreases as $1/\delta^2$ (see Eqs. 2 and 3), but by contrast, the light shift and hence the dipole force decreases only as $1/\delta$ for $\delta \gg \Omega$ (see Eq. 5). If $|\delta|$ is chosen to be sufficiently large, atoms spend very little time in the untrapped (actually repelled), excited state because its population is proportional to $1/\delta^2$. Thus, a sufficiently large value of $|\delta|$ produces longitudinal confinement and also maintains the atomic population primarily in the trapped ground state.

The first optical trap was demonstrated in Na with light detuned below the D-lines [26]. With 220 mW of dye laser light

tuned about 650 GHz below the atomic transition and focused to an $\sim$10 μm waist, the trap depth was about $15\hbar\gamma$ corresponding to 7 mK.

Single-beam dipole force traps can be made with the light detuned by a significant fraction of its frequency from the atomic transition. Such a far-off-resonance trap (FORT) has been developed for Rb atoms using light detuned by nearly 10% to the red of the D_1 transition at $\lambda = 795$ nm [30]. Between 0.5 and 1 W of power was focused to a spot about 10 μm in size, resulting in a trap 6 mK deep where the light-scattering rate was only a few hundreds per second. The trap lifetime was more than half a second.

There is a qualitative difference when the trapping light is detuned by a large fraction of the optical frequency. In one such case, Nd : YAG light at $\lambda = 1064$ nm was used to trap Na whose nearest transition is at $\lambda = 596$ nm [31]. In a more extreme case, a trap using $\lambda = 10.6$ μm light from a CO_2 laser has been used to trap Cs whose optical transition is at a frequency $\sim$12 times higher ($\lambda = 852$ nm) [32]. For such large values of $|\delta|$, calculations of the trapping force cannot exploit the rotating wave approximations as was done for Eqs. (4) and (5), and the atomic behavior is similar to that in a DC field. It is important to remember that for an electrostatic trap Earnshaw's theorem precludes a field maximum, but that in this case there is indeed a local 3-D intensity maximum of the focused light because it is not a static field.

4.1.2 **Blue-detuned Optical Traps**

One of the principal disadvantages of the optical traps discussed above is that the negative detuning attracts atoms to the region of highest light intensity. This may result in significant spontaneous emission unless the detuning is a large fraction of the optical frequency such as the Nd : YAG laser trap [31] or the CO_2 laser trap [32]. More important in some cases is that the trap relies on Stark shifting of the atomic energy levels by an amount equal to the trap depth, and this severely compromises the capabilities for precision spectroscopy in a trap [33].

Attracting atoms to the region of *lowest* intensity would ameliorate both these concerns, but such a trap requires positive detuning (blue), and an optical configuration having a dark central region. One of the first experimental efforts at a blue detuned trap used the repulsive dipole force to support Na atoms that were otherwise confined by gravity in an optical "cup" [34]. Two rather flat, parallel beams detuned by 25% of the atomic resonance frequency were directed horizontally and oriented to form a V-shaped trough. Their Gaussian beam waists formed a region $\simeq$1 mm long where the potential was deepest, and hence provided confinement along their propagation direction as shown in Fig. 9. The beams were the $\lambda = 514$ nm and $\lambda = 488$ nm from an argon laser, and

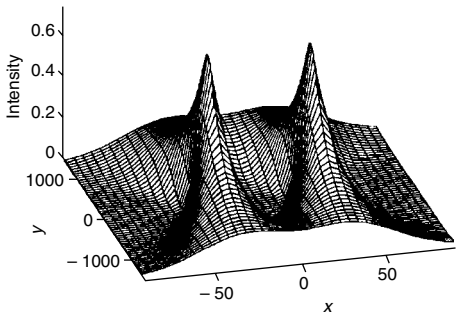

Fig. 9 The light intensity experienced by an atom located in a plane 30 μm above the beam waists of two quasi-focused sheets of light traveling parallel and arranged to form a V-shaped trough. The *x* and *y* dimensions are in μm (figure taken from Davidson, N., Lee, H. J., Adams, C. S., Kasevich, M., Chu, S. (**1995**), *Phys. Rev. Lett.* **74**, 1311–1314)

the choice of the two frequencies was not simply to exploit the full power of the multiline Ar laser, but also to avoid the spatial interference that would result from the use of a single frequency.

Obviously, a hollow laser beam would also satisfy the requirement for a blue-detuned trap, but conventional textbook wisdom shows that such a beam is not an eigenmode of a laser resonator [35]. Some lasers can make hollow beams, but these are illusions because they consist of rapid oscillations between the TEM_{01} and TEM_{10} modes of the cavity. Nevertheless, Maxwell's equations permit the propagation of such beams, and in the recent past there have been studies of the LaGuerre–Gaussian modes that constitute them [36–38]. The several ways of generating such hollow beams have been tried by many experimental groups and include phase and amplitude holograms, hollow waveguides, axicons or related cylindrical prisms, stressing fibers, and simply mixing the TEM_{01} and TEM_{10} modes with appropriate cylindrical lenses.

An interesting experiment has been performed using the ideas of Sisyphus cooling (see Sect. 3.3) with evanescent waves combined with a hollow beam formed with an axicon [39]. In the previously reported experiments with atoms bouncing under gravity from an evanescent wave field [40, 41], they were usually lost to horizontal motion for several reasons, including slight tilting of the surface, surface roughness, horizontal motion associated with their residual motion, and horizontal ejection by the Gaussian profile of the evanescent wave laser beam. The authors of Ref. [39] simply confined their atoms in the horizontal direction by surrounding them with a wall of blue-detuned light in the form of a vertical hollow beam. Their gravito-optical surface trap cooled Cs atoms to $\simeq 3\ \mu K$ at a density of $\simeq 3 \times 10^{10}/cm^3$ in a sample whose $1/e$ height in the gravitational field was only 19 μm. Simple ballistics gives a frequency of 450 bounces per second, and the ≃6-s lifetime (limited only by background gas collisions) corresponds to several thousand bounces. However, at such low energies, the deBroglie wavelength of the atoms is $\simeq 1/4\ \mu m$, and the atomic motion is no longer accurately described classically, but requires deBroglie wave methods.

4.2 Magnetic Traps

4.2.1 Introduction

Magnetic trapping of neutral atoms is well suited for use in very many areas, including high-resolution spectroscopy, collision studies, Bose–Einstein condensation (BEC), and atom optics. Although ion trapping, laser cooling of trapped ions, and trapped ion spectroscopy were known for many years [42], it was only in 1985 that neutral atoms were first trapped [43]. Such experiments offer the capability of the spectroscopic ideal of an isolated atom at rest, in the dark, available for interaction with electromagnetic field probes.

Because trapping requires the exchange of kinetic energy for potential energy, the atomic energy levels will necessarily shift as the atoms move in the trap. These shifts can severely affect the precision of spectroscopic measurements. Since one of the potential applications of trapped atoms is in high-resolution spectroscopy, such inevitable shifts must be carefully considered.

4.2.2 Magnetic Confinement

The Stern–Gerlach experiment in 1924 first demonstrated the mechanical action of inhomogeneous magnetic fields on

neutral atoms having magnetic moments, and the basic phenomenon was subsequently developed and refined. An atom with a magnetic moment $\vec{\mu}$ can be confined by an inhomogeneous magnetic field because of an interaction between the moment and the field. This produces a force given by

$$\vec{F} = \vec{\nabla}(\vec{\mu} \cdot \vec{B}) \tag{20}$$

since $E = -\vec{\mu} \cdot \vec{B}$. Several different magnetic traps with varying geometries that exploit the force of Eq. (20) have been studied in some detail in the literature. The general features of the magnetic fields of a large class of possible traps has been presented [44].

W. Paul originally suggested a quadrupole trap composed of two identical coils carrying opposite currents (see Fig. 10). This trap clearly has a single center in which the field is zero, and is the simplest of all possible magnetic traps. When the coils are separated by 1.25 times their radius, such a trap has equal depth in the radial (x-y plane) and longitudinal (z-axis) directions [44]. Its experimental simplicity makes it most attractive, both because of ease of construction and of optical access to the interior. Such a trap was used in the first neutral atom trapping experiments at NIST on laser-cooled Na atoms for times exceeding 1 s, and that time was limited only by background gas pressure [43].

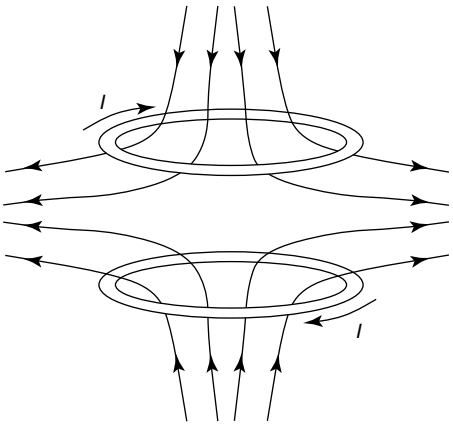

Fig. 10 Schematic diagram of the coil configuration used in the quadrupole trap and the resultant magnetic field lines. Because the currents in the two coils are in opposite directions, there is a $|\vec{B}| = 0$ point at the center

The magnitude of the field is zero at the center of this trap, and increases in all directions as

$$B = \nabla B \sqrt{\rho^2 + 4z^2}, \tag{21}$$

where $\rho^2 \equiv x^2 + y^2$ and the field gradient is constant (see Ref. [44]). The field gradient is constant along any line through the origin, but has different values in different polar directions because of the '4' in Eq. (21). Therefore, the force of Eq. (20) that confines the atoms in the trap is neither harmonic nor central, and orbital angular momentum is not conserved.

The requisite field for the quadrupole trap can also be provided in two dimensions by four straight currents as indicated in Fig. 11. The field is translationally invariant along the direction parallel to the currents, so a trap cannot be made this way without additional fields. These are provided by end coils that close the trap, as shown.

Although there are very many different kinds of magnetic traps for neutral particles, this particular one has played a special

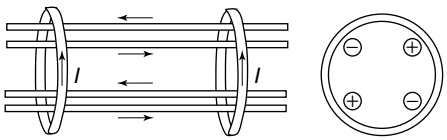

Fig. 11 The Ioffe trap has four straight current elements that form a linear quadrupole field. The axial confinement is accomplished with end coils as shown. These fields can be achieved with many different current configurations as long as the geometry is preserved

role. There are certain conditions required for trapped atoms not to be ejected in a region of zero field such as occurs at the center of a quadrupole trap (see Sects. 4.2.3 and 4.2.4). This problem is not easily cured; the Ioffe trap has been used in many of the BEC experiments because it has $|\vec{B}| \neq 0$ everywhere.

4.2.3 Classical Motion of Atoms in a Quadrupole Trap

Because of the dependence of the trapping force on the angle between the field and the atomic moment (see Eq. 20), the orientation of the magnetic moment with respect to the field must be preserved as the atoms move about in the trap. Otherwise, the atoms may be ejected instead of being confined by the fields of the trap. This requires velocities low enough to ensure that the interaction between the atomic moment $\vec{\mu}$ and the field $\vec{B}$ is adiabatic especially when the atom's path passes through a region where the field magnitude is small and therefore the energy separation between the trapping and nontrapping states is small. This is especially critical at the low temperatures of the BEC experiments. Therefore energy considerations that focus only on the trap depth are not sufficient to determine the stability of a neutral atom trap; orbit and/or quantum state calculations and their consequences must also be considered.

For the two-coil quadrupole magnetic trap of Fig. 10, stable circular orbits of radius ρ_1 in the $z = 0$ plane can be found classically by setting $\mu \nabla B = Mv^2/\rho_1$, so that $v = \sqrt{\rho_1 a}$, where $a \equiv \mu \nabla B/M$ is the centripetal acceleration supplied by the field gradient (cylindrical coordinates are appropriate). Such orbits have an angular frequency of $\omega_T = \sqrt{a/\rho_1}$. For traps of a few centimeter size and a few hundred Gauss depth, $a \sim 250\ \mathrm{m\ s^{-2}}$, and the fastest trappable atoms in circular orbits have $v_{\max} \sim 1\ \mathrm{m\ s^{-1}}$ so $\omega_T/2\pi \sim 20$ Hz. Because of the anharmonicity of the potential, the orbital frequencies depend on the orbit size, but in general, atoms in lower-energy orbits have higher frequencies.

For the quadrupole trap to work, the atomic magnetic moments must be oriented with $\vec{\mu} \cdot \vec{B} < 0$ so that they are repelled from regions of strong fields. This orientation must be preserved while the atoms move around in the trap even though the trap fields change directions in a very complicated way. The condition for adiabatic motion can be written as $\omega_Z \gg |\mathrm{d}B/\mathrm{d}t|/B$, where $\omega_Z = \mu B/\hbar$ is the Larmor precession rate in the field.

Since $v/\rho_1 = v\nabla B/B = |\mathrm{d}B/\mathrm{d}t|/B$ for a uniform field gradient, the adiabaticity condition is

$$\omega_Z \gg \omega_T. \tag{22}$$

More general orbits must satisfy a similar condition. For the two-coil quadrupole trap, the adiabaticity condition can be easily calculated. Using $v = \sqrt{\rho_1 a}$ for circular orbits in the $z = 0$ plane, the adiabatic condition for a practical trap ($\nabla B \sim 1$ T/m) requires $\rho_1 \gg (\hbar^2/M^2 a)^{1/3} \sim 1\ \mu$m as well as $v \gg (\hbar a/M)^{1/3} \sim 1\ \mathrm{cm\ s^{-1}}$. Note that violation of these conditions (i.e., $v \sim 1\ \mathrm{cm\ s^{-1}}$ in a trap with $\nabla B \sim 1$ T/m) results in the onset of quantum dynamics for the motion (deBroglie wavelength $\approx$ orbit size).

Since the nonadiabatic region of the trap is so small (less than $10^{-18}\ \mathrm{m^3}$ compared with typical sizes of $\sim$2 cm, corresponding to $10^{-5}\ \mathrm{m^3}$), nearly all the orbits of most atoms are restricted to regions where they are adiabatic. Therefore, most of such laser-cooled atoms stay trapped for

many thousands of orbits corresponding to several minutes. At laboratory vacuum chamber pressures of typically 10^{-10} torr, the mean free time between collisions that can eject trapped atoms is ~2 min, so the transitions caused by nonadiabatic motion are not likely to be observable in atoms that are optically cooled.

4.2.4 Quantum Motion in a Trap

Since laser and evaporative cooling have the capability to cool atoms to energies where their deBroglie wavelengths are on the scale of the orbit size, the motional dynamics must be described in terms of quantum mechanical variables and suitable wave functions. Quantization of the motion leads to discrete bound states within the trap having $\vec{\mu} \cdot \vec{B} < 0$, and also a continuum of unbound states having $\vec{\mu} \cdot \vec{B}$ with opposite sign.

Studying the behavior of extremely slow (cold) atoms in the two-coil quadrupole trap begins with a heuristic quantization of the orbital angular momentum using $Mr^2\omega_T = n\hbar$ for circular orbits. The energy levels are then given by

$$E_n = \frac{3}{2} E_1 n^{2/3}, \quad \text{where}$$
$$E_1 = (Ma^2\hbar^2)^{1/3} \sim h \times 5\ \text{kHz}, \quad (23)$$

For velocities of optically cooled atoms of a few cm s^{-1}, $n \sim 10 - 100$. It is readily found that $\omega_Z = n\omega_T$, so that the adiabatic condition of Eq. (22) is satisfied only for $n \gg 1$.

These large-n bound states have small matrix elements coupling them to the unbound continuum states [45]. This can be understood classically since they spend most of their time in a stronger field, and thus satisfy the condition of adiabaticity of the orbital motion relative to the Larmor precession. In this case, the separation of the rapid precession from the slower orbital motion is reminiscent of the Born–Oppenheimer approximation for molecules.

On the other hand, the small-n states, whose orbits are confined to a region near the origin where the field is small, have much larger coupling to the continuum states. Thus, they are rapidly ejected from the trap. The transitions to unbound states resulting from the coupling of the motion with the trapping fields are called *Majorana spin flips*, and effectively constitute a ''hole'' at the bottom of the trap. The evaporative cooling process used to produce very cold, dense samples reduces the average total energy of the trapped atoms sufficiently that the orbits are confined to regions near the origin and so, such losses dominate [44, 45].

There have been different solutions to this problem of Majorana losses for confinement of ultracold atoms for the BEC experiments. In the JILA-experiment, the hole was rotated by rotating the magnetic field, and thus, the atoms do not spend sufficient time in the hole to make a spin flip. In the MIT experiment, the hole was plugged by using a focused laser beam tuned to the blue side of atomic resonance, which expelled the atoms from the center of the magnetic trap. In the Rice experiment, the atoms were trapped in an Ioffe trap, which has a nonzero field minimum. Most BEC experiments are now using the Ioffe trap solution.

4.3 Magneto-optical Traps

4.3.1 Introduction

The most widely used trap for neutral atoms is a hybrid employing both optical and magnetic fields to make a magneto-optical trap (MOT), first demonstrated in

1987 [46]. The operation of an MOT depends on both inhomogeneous magnetic fields and radiative selection rules to exploit both optical pumping and the strong radiative force [46, 47]. The radiative interaction provides cooling that helps in loading the trap and enables very easy operation. MOT is a very robust trap that does not depend on precise balancing of the counterpropagating laser beams or on a very high degree of polarization.

The magnetic field gradients are modest and have the convenient feature that the configuration is the same as the quadrupole magnetic traps discussed in Sect. 4.2.2. Appropriate fields can readily be achieved with simple, air-cooled coils. The trap is easy to construct because it can be operated with a room-temperature cell in which alkali atoms are captured from the vapor. Furthermore, low-cost diode lasers can be used to produce the light appropriate for many atoms, so the MOT has become one of the least expensive ways to make atomic samples with temperatures below 1 mK.

Trapping in an MOT works by optical pumping of slowly moving atoms in a linearly inhomogeneous magnetic field $B = B(z)$ (see Eq. 21), such as that formed by a magnetic quadrupole field. Atomic transitions with the simple scheme of $J_g = 0 \rightarrow J_e = 1$ have three Zeeman components in a magnetic field, excited by each of three polarizations, whose frequencies tune with the field (and therefore with position) as shown in Fig. 12 for 1-D. Two counterpropagating laser beams of opposite circular polarization, each detuned below the zero-field atomic resonance by δ, are incident as shown.

Because of the Zeeman shift, the excited state $M_e = +1$ is shifted up for $B > 0$, whereas the state with $M_e = -1$ is shifted down. At position z' in Fig. 12,

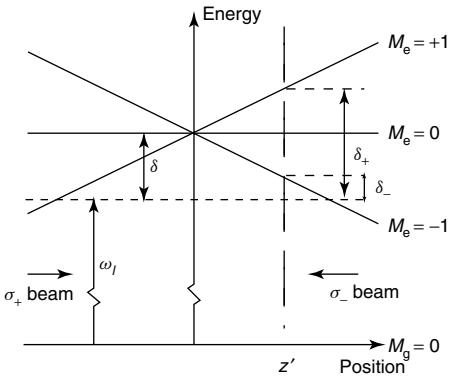

Fig. 12 Arrangement for an MOT in 1-D. The horizontal dashed line represents the laser frequency seen by an atom at rest in the center of the trap. Because of the Zeeman shifts of the atomic transition frequencies in the inhomogeneous magnetic field, atoms at $z = z'$ are closer to resonance with the σ^- laser beam than with the σ^+ beam, and are therefore driven toward the center of the trap

the magnetic field, therefore, tunes the $\Delta M = -1$ transition closer to resonance and the $\Delta M = +1$ transition further out of resonance. If the polarization of the laser beam incident from the right is chosen to be σ^- and correspondingly σ^+ for the other beam, then more light is scattered from the σ^- beam than from the σ^+ beam. Thus, the atoms are driven toward the center of the trap where the magnetic field is zero. On the other side of the center of the trap, the roles of the $M_e = \pm 1$ states are reversed and now more light is scattered from the σ^+ beam, again driving the atoms toward the center.

So far, the discussion has been limited to the motion of atoms in 1-D. However, the MOT scheme can easily be extended to 3-D by using six instead of two laser beams. Furthermore, even though very few atomic species have transitions as simple as $J_g = 0 \rightarrow J_e = 1$, the scheme works for any $J_g \rightarrow J_e = J_g + 1$ transition. Atoms that scatter mainly from the σ^+

laser beam will be optically pumped toward the $M_g = +J_g$ substate, which forms a closed system with the $M_e = +J_e$ substate.

4.3.2 Cooling and Compressing Atoms in an MOT

For a description of the motion of atoms in an MOT, consider the radiative force in the low intensity limit (see Eqs. 2 and 3). The total force on the atoms is given by $\vec{F} = \vec{F}_+ + \vec{F}_-$, where $\vec{F}_\pm$ can be found from Eqs. (2) and (3), and the detuning $\delta_\pm$ for each laser beam is given by $\delta_\pm = \delta \mp \vec{k} \cdot \vec{v} \pm \mu' B/\hbar$. Here, $\mu' \equiv (g_e M_e - g_g M_g)\mu_B$ is the effective magnetic moment for the transition used. Note that the Doppler shift $\omega_D \equiv -\vec{k} \cdot \vec{v}$ and the Zeeman shift $\omega_Z = \mu' B/\hbar$ both have opposite signs for opposite beams.

The situation is analogous to the velocity damping in an OM from the Doppler effect as discussed in Sec. 3.2, but here the effect also operates in position space, whereas for molasses it operates only in velocity space. Since the laser light is detuned below the atomic resonance in both cases, compression and cooling of the atoms is obtained simultaneously in an MOT.

When both the Doppler and Zeeman shifts are small compared to the detuning δ, the denominator of the force can be expanded as for Eq. (13) and the result becomes

$$\vec{F} = -\beta\vec{v} - \kappa\vec{r}, \qquad (24)$$

where the damping coefficient β is defined in Eq. (13). The spring constant κ arises from the similar dependence of $\vec{F}$ on the Doppler and Zeeman shifts, and is given by $\kappa = \mu' \beta \nabla B/\hbar k$

The force of Eq. (24) leads to damped harmonic motion of the atoms, where the damping rate is given by $\Gamma_{MOT} = \beta/M$ and the oscillation frequency $\omega_{MOT} = \sqrt{\kappa/M}$. For magnetic field gradients $\nabla B \approx 0.1\ \mathrm{T\ m^{-1}}$, the oscillation frequency is typically a few kHz, and this is much smaller than the damping rate that is typically a few hundred kHz. Thus, the motion is overdamped, with a characteristic restoring time to the center of the trap of $2\Gamma_{MOT}/\omega_{MOT}^2 \approx$ several milliseconds for typical values of detuning and intensity of the lasers.

4.3.3 Capturing Atoms in an MOT

Although the approximations that lead to Eq. (24) for force hold for slow atoms near the origin, they do not apply for the capture of fast atoms far from the origin. In the capture process, the Doppler and Zeeman shifts are no longer small compared to the detuning, so the effects of the position and velocity can no longer be disentangled. However, the full expression for the force still applies and the trajectories of the atoms can be calculated by numerical integration of the equation of motion [48].

The capture velocity of an MOT is serendipitously enhanced because atoms traveling across it experience a decreasing magnetic field just as in beam deceleration described in Sect. 3.1. This enables resonance over an extended distance and velocity range because the changing Doppler shift of decelerating atoms can be compensated by the changing Zeeman shift as atoms move in the inhomogeneous magnetic field. Of course, it will work this way only if the field gradient ∇B does not demand an acceleration larger than the maximum acceleration a_{max}. Thus, atoms are subject to the optical force over a distance that can be as long as the trap size, and can therefore be slowed considerably.

The very large velocity capture range v_{cap} of an MOT can be estimated by using $F_{max} = \hbar k \gamma/2$ and choosing a maximum size of a few centimeters for the beam

diameters. Thus, the energy change can be as large as a few K, corresponding to $v_{cap} \sim 100\ m\ s^{-1}$ [47]. The number of atoms in a vapor with velocities below v_{cap} in the Boltzmann distribution scales as v_{cap}^4, and there are enough slow atoms to fall within the large MOT capture range even at room temperature, because a few K includes 10^{-4} of the atoms.

4.3.4 Variations on the MOT Technique

Because of the wide range of applications of this most versatile kind of atom trap, a number of careful studies of its properties have been made [47, 49–56], and several variations have been developed. One of these is designed to overcome the density limits achievable in an MOT. In the simplest picture, loading additional atoms into an MOT produces a higher atomic density because the size of the trapped sample is fixed. However, the density cannot increase without limit as more atoms are added. The atomic density is limited to $\sim 10^{11}\ cm^{-3}$ because the fluorescent light emitted by some trapped atoms is absorbed by others.

One way to overcome this limit is to have much less light in the center of the MOT than at the sides. Simply lowering the laser power is not effective in reducing the fluorescence because it will also reduce the capture rate and trap depth. But those advantageous properties can be preserved while reducing fluorescence from atoms at the center if the light intensity is low only in the center.

The repumping process for the alkali atoms provides an ideal way of implementing this idea [57]. If the repumping light is tailored to have zero intensity at the center, then atoms trapped near the center of the MOT are optically pumped into the "wrong" hfs state and stop fluorescing. They drift freely in the "dark" at low speed through the center of the MOT until they emerge on the other side, into the region where light of both frequencies is present and begin absorbing again. Then they feel the trapping force and are driven back into the "dark" center of the trap. Such an MOT has been operated at MIT [57] with densities close to $10^{12}/cm^3$, and the limitations are now from collisions in the ground state rather than from multiple light scattering and excited-state collisions.

5 Optical Lattices

5.1 Quantum States of Motion

As the techniques of laser cooling advanced from a laboratory curiosity to a tool for new problems, the emphasis shifted from attaining the lowest possible steady state temperatures to the study of elementary processes, especially the quantum mechanical description of the atomic motion. In the completely classical description of laser cooling, atoms were assumed to have a well-defined position and momentum that could be known simultaneously with arbitrary precision. However, when atoms are moving sufficiently slowly that their deBroglie wavelength precludes their localization to less than $\lambda/2\pi$, these descriptions fail and a quantum mechanical description is required. Such exotic behavior for the motion of whole atoms, as opposed to electrons in the atoms, had not been considered before the advent of laser cooling simply because it was too far out of the range of ordinary experiments. A series of experiments in the early 1990s provided dramatic evidence for these new quantum states of motion of neutral atoms, and

led to the debut of deBroglie wave atom optics.

The quantum description of atomic motion requires that the energy of such motion be included in the Hamiltonian. The total Hamiltonian for atoms moving in a light field would then be given by

$$\mathcal{H} = \mathcal{H}_{\text{atom}} + \mathcal{H}_{\text{rad}} + \mathcal{H}_{\text{int}} + \mathcal{H}_{\text{kin}}, \quad (25)$$

where $\mathcal{H}_{\text{atom}}$ describes the motion of the atomic electrons and gives the internal atomic energy levels, $\mathcal{H}_{\text{rad}}$ is the energy of the radiation field and is of no concern here because the field is not quantized, $\mathcal{H}_{\text{int}}$ describes the excitation of atoms by the light field and the concomitant light shifts, and $\mathcal{H}_{\text{kin}}$ is the kinetic energy operator of the motion of the center of mass of the atoms. This Hamiltonian has eigenstates of not only the internal energy levels and the atom-laser interaction that connects them, but also that of the kinetic energy operator $\mathcal{H}_{\text{kin}} \equiv P^2/2M$. These eigenstates will therefore be labeled by quantum numbers of the atomic states as well as the center of mass momentum p. An atom in the ground state, $|g; p\rangle$, has an energy $E_g + p^2/2M$, that can take on a range of values.

In 1968, V.S. Letokhov [58] suggested that it is possible to confine atoms in the wavelength-size regions of a standing wave by means of the dipole force that arises from the light shift. This was first accomplished in 1987 in 1-D with an atomic beam traversing an intense standing wave [59]. Since then, the study of atoms confined to wavelength-size potential wells has become an important topic in optical control of atomic motion because it opens up configurations previously accessible only in condensed matter physics using crystals.

The limits of laser cooling discussed in Sect. 3.3.4 suggest that atomic momenta can be reduced to a "few" times $\hbar k$. This means that their deBroglie wavelengths are equal to the optical wavelengths divided by a "few". If the depth of the optical potential wells is high enough to contain such very slow atoms, then their motion in potential wells of size $\lambda/2$ must be described quantum mechanically, since they are confined to a space of size comparable to their deBroglie wavelengths. Thus, they do not oscillate in the sinusoidal wells as classical localizable particles, but instead occupy discrete, quantum mechanical bound states [60], as shown in the lower part of Fig. 13.

The basic ideas of the quantum mechanical motion of particles in a periodic potential were laid out in the 1930s with the Kronig–Penney model and Bloch's theorem, and optical lattices offer important opportunities for their study. For example,

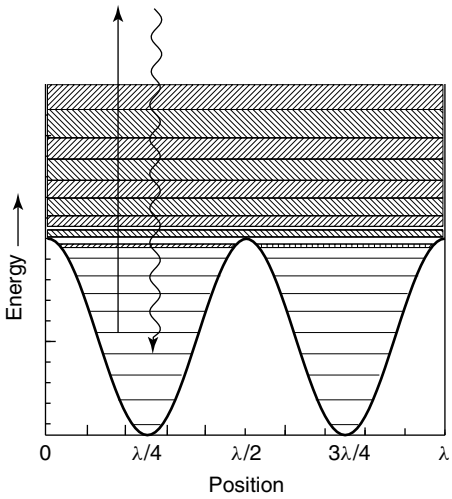

Fig. 13 Energy levels of atoms moving in the periodic potential of the light shift in a standing wave. There are discrete bound states deep in the wells that broaden at higher energy, and become bands separated by forbidden energies above the tops of the wells. Under conditions appropriate to laser cooling, optical pumping among these states favors populating the lowest ones as indicated schematically by the arrows

these lattices can be made essentially free of defects with only moderate care in spatially filtering the laser beams to assure a single transverse mode structure. Furthermore, the shape of the potential is exactly known and does not depend on the effect of the crystal field or the ionic energy level scheme. Finally, the laser parameters can be varied to modify the depth of the potential wells without changing the lattice vectors, and the lattice vectors can be changed independently by redirecting the laser beams. The simplest optical lattice to consider is a 1-D pair of counterpropagating beams of the same polarization, as was used in the first experiment [59].

Of course, such tiny traps are usually very shallow, so loading them requires cooling to the μK regime. Even atoms whose energy exceeds the trap depth must be described as quantum mechanical particles moving in a periodic potential that display energy band structure [60]. Such effects have been observed in very careful experiments.

Because of the transverse nature of light, any mixture of beams with different $\vec{k}$-vectors necessarily produces a spatially periodic, inhomogeneous light field. The importance of the ''egg-crate'' array of potential wells arises because the associated atomic light shifts can easily be comparable to the very low average atomic kinetic energy of laser-cooled atoms. A typical example projected against two dimensions is shown in Fig. 14.

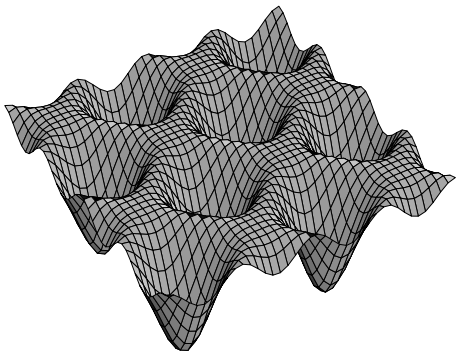

Fig. 14 The ''egg-crate'' potential of an optical lattice shown in two dimensions. The potential wells are separated by $\lambda/2$

Atoms trapped in wavelength-sized spaces occupy vibrational levels similar to those of molecules. The optical spectrum can show Raman-like sidebands that result from transitions among the quantized vibrational levels [61, 62] as shown in Fig. 15. These quantum states of atomic motion can also be observed by stimulated emission [62, 63] and by direct RF spectroscopy [64, 65].

5.2 Properties of 3-D Lattices

The name ''optical lattice'' is used rather than optical crystal because the filling fraction of the lattice sites is typically only a few percent (as of 1999). The limit arises because the loading of atoms into the lattice is typically done from a sample of trapped and cooled atoms, such as an MOT for atom collection, followed by an OM for laser cooling. The atomic density in such experiments is limited by collisions and multiple light scattering to a few times 10^{11} cm^{-3}. Since the density of lattice sites of size $\lambda/2$ is a few times 10^{13} cm^{-3}, the filling fraction is necessarily small. With the advent of experiments that load atoms directly into a lattice from a BEC, the filling factor can be increased to 100%, and in some cases it may be possible to load more than one atom per lattice site [66, 67].

In 1993 a very clever scheme was described [68]. It was realized that an n-dimensional lattice could be created by only $n+1$ traveling waves rather than

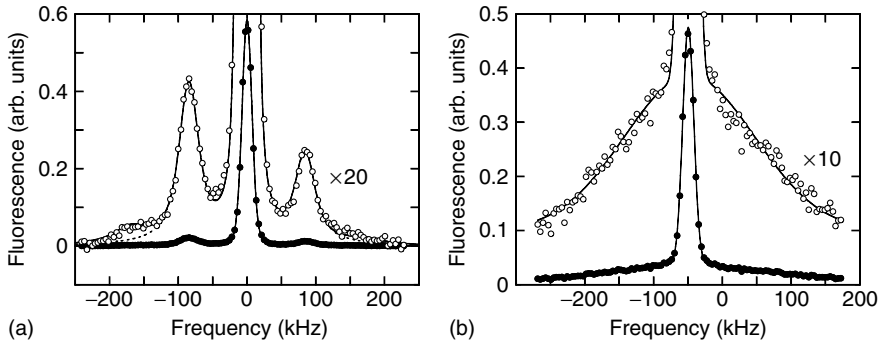

Fig. 15 (a) Fluorescence spectrum in a 1-D lin ⊥ lin optical molasses. Atoms are first captured and cooled in an MOT, and then the MOT light beams are switched off leaving a pair of lin ⊥ lin because of spontaneous emission of the atoms to the same vibrational state from which they are excited, whereas the sideband on the left (right) is due to spontaneous emission to a vibrational state with one vibrational quantum number lower (higher) (see Fig. 13). The presence of these sidebands is a direct proof of the existence of the band structure. (b) Same as (a) except that the 1-D molasses is $\sigma^+ - \sigma^-$, which has no spatially dependent light shift and hence no vibrational states (figure taken from Jessen, P. S., Gerz, C., Lett, P. D., Phillips, W. D., Rolston, S. L., Spreeuw, R. J. C., Westbrook, C. I. (**1992**), *Phys. Rev. Lett.* **69**, 49–52)

$2n$. The real benefit of this scheme is that in case of phase instabilities in the laser beams, the interference pattern is only shifted in space, but the interference pattern itself is not changed. Instead of producing optical wells in 2-D with four beams (two standing waves), these authors used only three. The $\vec{k}$-vectors of the coplanar beams were separated by $2\pi/3$, and they were all linearly polarized in their common plane (not parallel to one another). The same immunity to vibrations was established for a 3-D optical lattice by using only four beams arranged in a quasi-tetrahedral configuration. The three linearly polarized beams of the 2-D arrangement described above were directed out of the plane toward a common vertex, and a fourth circularly polarized beam was added. All four beams were polarized in the same plane [68]. The authors showed that this configuration produced the desired potential wells in 3-D.

5.3 Spectroscopy in 3-D Lattices

The group at NIST developed a new method that superposed a weak probe beam of light directly from the laser upon some of the fluorescent light from the atoms in a 3-D OM, and directed the light from these combined sources onto on a fast photodetector [70]. The resulting beat signal carried information about the Doppler shifts of the atoms in the optical lattices [61]. These Doppler shifts were expected to be in the sub-MHz range for atoms with the previously measured 50-μK temperatures. The observed features confirmed the quantum nature of the motion of atoms in the wavelength-size potential wells (see Fig. 15) [15].

The NIST group also studied atoms loaded into an optical lattice using of laser light from the spatially ordered array [71]. They cut off the laser beams that formed the lattice, and before the

atoms had time to move away from their positions, they pulsed on a probe laser beam at the Bragg angle appropriate for one of the sets of lattice planes. The Bragg diffraction not only enhanced the reflection of the probe beam by a factor of 10^5, but by varying the time between the shut-off of the lattice and turn-on of the probe, they could also measure the ''temperature'' of the atoms in the lattice. The reduction of the amplitude of the Bragg-scattered beam with time provided some measure of the diffusion of the atoms away from the lattice sites, much like the Debye–Waller factorDebye–Waller factor in X-ray diffraction.

5.4 Quantum Transport in Optical Lattices

In the 1930s, Bloch realized that applying a uniform force to a particle in a periodic potential would not accelerate it beyond a certain speed, but instead would result in Bragg reflection when its deBroglie wavelength became equal to the lattice period. Thus, an electric field applied to a conductor could not accelerate electrons to a speed faster than that corresponding to the edge of a Brillouin zone, and that at longer times the particles would execute oscillatory motion. Ever since then, experimentalists have tried to observe these Bloch oscillations in increasingly pure and/or defect-free crystals.

Atoms moving in optical lattices are ideally suited for such an experiment, as was beautifully demonstrated in 1996 [69]. The authors loaded a 1-D lattice with atoms from a 3-D molasses, further narrowed the velocity distribution, and then instead of applying a constant force, simply changed the frequency of one of the beams of the 1-D lattice with respect to the other in a controlled way, thereby creating an accelerating lattice. Seen from the atomic reference frame, this was the equivalent of a constant force trying to accelerate them. After a variable time t_a, the 1-D lattice beams were shut off and the measured atomic velocity distribution showed beautiful Bloch oscillations as a function of t_a. The centroid of the very narrow velocity distribution was seen to shift in velocity space at a constant rate until it reached

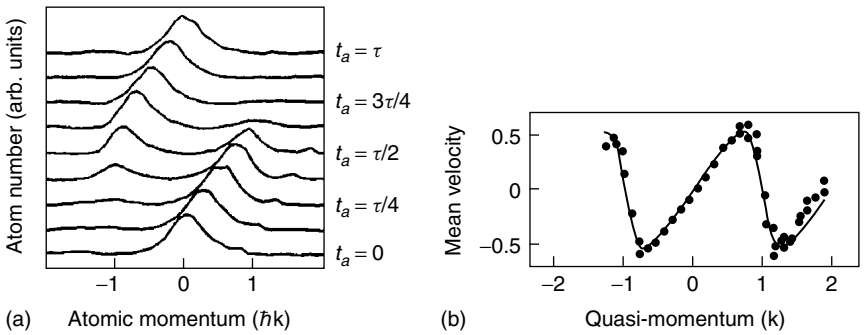

Fig. 16 Plot of the measured velocity distribution *verses*. time in the accelerated 1-D lattice. (a) Atoms in a 1-D lattice are accelerated for a fixed potential depth for a certain time t_a and the momentum of the atoms after the acceleration is measured. The atoms accelerate only to the edge of the Brillouin zone where the velocity is $+\nu_r$, and then the velocity distribution appears at $-\nu_r$. (b) Mean velocity of the atoms as a function of the quasi-momentum, that is, the force times the acceleration time (figure taken from Ben Dahan, M., Peik, E., Reichel, J., Castin, Y., Salomon, C. (**1996**), *Phys. Rev. Lett.* **76**, 4508–4511)

$v_r = \hbar k/M$, and then it vanished and reappeared at $-v_r$ as shown in Fig. 16. The shape of the ''dispersion curve'' allowed measurement of the ''effective mass'' of the atoms bound in the lattice.

6 Bose–Einstein Condensation

6.1 Introduction

In the 1920s, Bose and Einstein predicted that for sufficiently high phase space density, $\rho_\phi \sim 1$ (see Sect. 1.2), a gas of atoms undergoes a phase transition that is now called Bose–Einstein condensation. It took 70 years before BEC could be unambiguously observed in a dilute gas. From the advent of laser cooling and trapping, it became clear that this method could be instrumental in achieving BEC.

BEC is another manifestation of quantization of atomic motion. It occurs in the absence of resonant light, and its onset is characterized by cooling to the point where the atomic deBroglie wavelengths are comparable to the interatomic spacing. This is in contrast with the topics discussed in Sect. 5 where the atoms were in an optical field and their deBroglie wavelengths were comparable to the optical wavelength λ.

Laser cooling alone is inherently incapable of achieving $\rho_\phi \sim 1$. This is easily seen from the recoil limit of Sect. 3.3.4 that limits λ_{deB} to λ/''few''. Since the cross section for optical absorption near resonance is $\sigma \sim \lambda^2$ near $\rho_\phi \sim 1$, this limit of $\lambda_{deB} \sim \lambda$/''few'' results in the penetration depth of the cooling light into the sample being smaller than λ. Thus, the sample would have to be extremely small and contain only a few atoms, hardly a system suitable for investigation.

Temperatures lower than the recoil limit are readily achieved by evaporative cooling, and so all BEC experiments employ it in their final phase. Evaporative cooling is inherently different from the other cooling processes discussed in Sect. 3, and hence discussed here separately.

Since the first observations in 1995, BEC has been the subject of intense investigation, both theoretical and experimental. No attempt is made in this article to even address, much less cover, the very rich range of physical phenomena that have been unveiled by these studies. Instead, we focus on the methods to achieve $\rho_\phi \sim 1$ and BEC.

6.2 Evaporative Cooling

Evaporative cooling is based on the preferential removal of those atoms from a confined sample with an energy higher than the average energy followed by a rethermalization of the remaining gas by elastic collisions. Although evaporation is a process that occurs in nature, it was applied to atom cooling for the first time in 1988 [72].

One way to think about evaporative cooling is to consider cooling of a container of hot liquid. Since the most energetic molecules evaporate from the liquid and leave the container, the remaining molecules obtain a lower temperature and are cooled. Furthermore, it requires the evaporation of only a small fraction of the liquid to cool it by a considerable amount.

Evaporative cooling works by removing the higher-energy atoms as suggested schematically in Fig. 17. Those that remain have much lower average energy (temperature) and so they occupy a smaller volume near the bottom of the

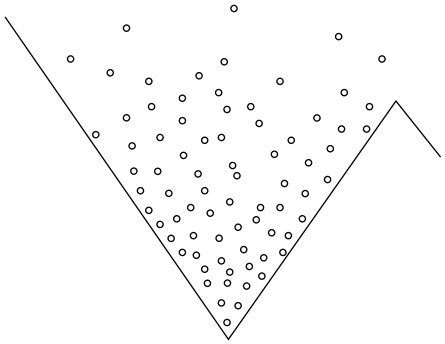

Fig. 17 Principle of the evaporation technique. Once the trap depth is lowered, atoms with energy above the trap depth can escape and the remaining atoms reach a lower temperature

trap, thereby increasing their density. For trapped atoms, it can be achieved by lowering the depth of the trap, thereby allowing the atoms with energies higher than the trap depth to escape, as discussed first by Hess [73]. Elastic collisions in the trap then lead to a rethermalization of the gas. This technique was first employed for evaporative cooling of hydrogen [72, 74–76]. Since both the temperature and the volume decrease, ρ_ϕ increases.

Recently, more refined techniques have been developed. For example, to sustain the cooling process the trap depth can be lowered continuously, achieving a continuous decrease in temperature. Such a process is called *forced evaporation* and is discussed in Sect. 6.3 below.

6.2.1 Simple Model

This section describes a simple model of evaporative cooling. Since such cooling is not achieved for single atoms but for the whole ensemble, an atomic description of the cooling process must be replaced by thermodynamic methods. These methods are completely different from the rest of the material in this article, and will therefore remain rather elementary.

Several models have been developed to describe this process, but we present here the simplest one [77] because of its pedagogical value [22]. In this model, the trap depth is lowered in one single step and the effect on the thermodynamic quantities, such as temperature, density, and volume, is calculated. The process can be repeated and the effects of multiple steps added up cumulatively.

In such models of evaporative cooling, the following assumptions are made:

1. The gas behaves sufficiently ergodically, that is, the distribution of atoms in phase space (both position and momentum) depends only on the energy of the atoms and the nature of the trap.
2. The gas is assumed to begin the process with $\rho_\phi \ll 1$ (far from the BEC transition point), and so it is described by classical statistics.
3. Even though $\rho_\phi \ll 1$, the gas is cold enough that the atomic scattering is pure (s-wave) quantum mechanical, that is, the temperature is sufficiently low that all higher partial waves do not contribute to the cross section. Furthermore, the cross section for elastic scattering is energy-independent and is given by $\sigma = 8\pi a^2$, where a is the scattering length. It is also assumed that the ratio of elastic to inelastic collision rates is sufficiently large that the elastic collisions dominate.
4. Evaporation preserves the thermal nature of the distribution, that is, the thermalization is much faster than the rate of cooling.
5. Atoms that escape from the trap neither collide with the remaining atoms nor exchange energy with them. This is called *full evaporation*.

6.2.2 Application of the Simple Model

The first step in applying this simple model is to characterize the trap by calculating how the volume of a trapped sample of atoms changes with temperature T. Consider a trapping potential that can be expressed as a power law given by

$$U(x,y,z) = \varepsilon_1 \left|\frac{x}{a_1}\right|^{s_1} + \varepsilon_2 \left|\frac{y}{a_2}\right|^{s_2} + \varepsilon_3 \left|\frac{z}{a_3}\right|^{s_3}, \quad (26)$$

where a_j is a characteristic length and s_j the power, for a certain direction j. Then the volume occupied by trapped atoms scales as $V \propto T^{\xi}$ [78], where

$$\xi \equiv \frac{1}{s_1} + \frac{1}{s_2} + \frac{1}{s_3}. \quad (27)$$

Thus, the effect of the potential on the volume of the trapped sample for a given temperature can be reduced to a single parameter ξ. This parameter is independent of how the occupied volume is defined, since many different definitions lead to the same scaling. When a gas is held in a 3-D box with infinitely high walls, then $s_1 = s_2 = s_3 = \infty$ and $\xi = 0$, which means that V is independent of T, as expected. For a harmonic potential in 3-D, $\xi = 3/2$; for a linear potential in 2-D, $\xi = 2$; and for a linear potential in 3-D, $\xi = 3$.

The evaporative cooling model itself [77] starts with a sample of N atoms in volume V having a temperature T held in an infinitely deep trap. The strategy for using the model is to choose a finite quantity η, and then (1) lower the trap depth to a value $\eta k_B T$, (2) allow for a thermalization of the sample by collisions, and (3) determine the change in ρ_ϕ.

Only two parameters are needed to completely determine all the thermodynamic quantities for this process (the values after the process are denoted by a prime). One of these is $\nu \equiv N'/N$, the fraction of atoms remaining in the trap after the cooling. The other is γ (This γ is not to be confused with the natural width of the excited state.), a measure of the decrease in temperature caused by the release of hot atoms and subsequent cooling, modified by ν, and defined as

$$\gamma \equiv \frac{\log(T'/T)}{\log(N'/N)} = \frac{\log(T'/T)}{\log \nu}. \quad (28)$$

This yields a power-law dependence for the decrease in temperature caused by the loss of the evaporated particles, that is, $T' = T\nu^{\gamma}$. The dependence of the other thermodynamic quantities on the parameters ν and γ can then be calculated.

The scaling of $N' = N\nu$, $T' = T\nu^{\gamma}$, and $V' = V\nu^{\gamma\xi}$ can provide the scaling of all the other thermodynamic quantities of interest by using the definitions for the density $n = N/V$, the phase space density $\rho_\phi = n\lambda_{\mathrm{deB}}^3 \propto nT^{-3/2}$, and the elastic collision rate $k_{\mathrm{el}} \equiv n\sigma v \propto nT^{1/2}$. The results are given in Table 2. For a given value of η, the scaling of all quantities depends only on γ. Note that for successive steps j, ν has to be replaced with ν^j.

Tab. 2 Exponent q for the scaling of the thermodynamic quantities $X' = X\nu^q$ with the reduction ν of the number of atoms in the trap

Thermodynamic variable	*Symbol*	*Exponent q*
Number of atoms	N	1
Temperature	T	γ
Volume	V	$\gamma\xi$
Density	n	$1-\gamma\xi$
Phase space density	ρ	$1-\gamma(\xi+3/2)$
Collision rate	k	$1-\gamma(\xi-1/2)$

The fraction of atoms remaining is fully determined by the final trap depth η for a given potential characterized by the trap parameter ξ. In order to determine the change of the temperature in the cooling process, it is necessary to consider in detail the distribution of the atoms in the trap, and this is discussed more fully in Refs. [1, 22, 77].

6.2.3 Speed of Evaporation

So far, the speed of the evaporative cooling process has not been considered. If the trap depth is ramped down too quickly, the thermalization process does not have time to run its course and the process becomes less efficient. On the other hand, if the trap depth is ramped down too slowly, the loss of particles by inelastic collisions becomes important, thereby making the evaporation inefficient.

The speed of evaporation can be found from the principle of detailed balance [22]. Its application shows that the ratio of the evaporation time and the elastic collision time is

$$\frac{\tau_{\mathrm{ev}}}{\tau_{\mathrm{el}}} = \frac{\sqrt{2}e^{\eta}}{\eta}. \tag{29}$$

Note that this ratio increases exponentially with η.

Experimental results show that $\sim$2.7 elastic collisions are necessary to rethermalize the gas [79]. In order to model the rethermalization process, Luiten et al. [80] have discussed a model based on the Boltzmann equation where the evolution of the phase space density $\rho(\vec{r}, \vec{p}, t)$ is calculated. This evolution is not only caused by the trapping potential, but also by collisions between the particles. Only elastic collisions, whose cross section is given by $\sigma = 8\pi a^2$ with a as the scattering length, are considered. This leads to the Boltzmann equation [81].

6.2.4 Limiting Temperature

In the models discussed so far, only elastic collisions have been considered, that is, collisions where kinetic energy is redistributed between the partners. However, if part of the internal energy of the colliding partners is exchanged with their kinetic energy in the collision, then it is inelastic. Inelastic collisions can cause problems for two reasons: (1) the internal energy released can cause the atoms to heat up and (2) the atoms can change their internal states, and the new states may no longer be trapped. In each case, such collisions can lead to trap loss and are therefore not desirable.

Apart from collisions with the background gas and three-body recombination, there are two inelastic processes that are important for evaporative cooling of alkali atoms: dipolar relaxation and spin relaxation. The collision rate nk_{dip} for them at low energies is independent of velocity [82]. Since the elastic collision rate is given by $k_{\mathrm{el}} = n\sigma v_{\mathrm{rel}}$, the ratio of good (= elastic) to bad (= relaxation) collisions goes down when the temperature does. This limits the temperature to a value T_e near which the ratio between good and bad collisions becomes unity, and T_e is given by

$$k_{\mathrm{B}} T_e = \frac{\pi M k_{\mathrm{dip}}^2}{16\sigma^2}. \tag{30}$$

The limiting temperature for the alkalis is of the order of 1 nK, depending on the values of σ and k_{dip}.

In practice, however, this ratio has to be considerably larger than unity, and so the practical limit for evaporative cooling occurs when the ratio is $\sim 10^3$ [22]. In the model of Ref. [80], the authors discuss different strategies for evaporative cooling. Even for the strategy of the

lowest temperature, the final temperature is higher than T_e.

The collision rate between atoms with one in the excited state (S + P collisions) is also much larger at low temperatures than the rate for such collisions with both atoms in the ground state (S + S collisions). Since S + P collisions are generally inelastic and the inelastic energy exchange generally leads to a heating of the atoms, increasing the density increases the loss of cold atoms. To achieve BEC, resonant light should therefore be avoided, and thus laser cooling is not suitable for achieving BEC.

6.3 Forced Evaporative Cooling

In all the earliest experiments that achieved BEC, the evaporative cooling was "forced" by inducing rf transitions to magnetic sublevels that are not bound in the magnetic trap. Atoms with the highest energies can access regions of the trap where the magnetic field is stronger, and thus their Zeeman shifts would be larger. A correspondingly high-frequency rf field would cause only these most energetic atoms to undergo transitions to states that are not trapped, and in doing so, the departing atoms carry away more than the average energy. Thus, a slow sweep of the rf frequency from high to low would continuously shave off the high-energy tail of the energy distribution, and thereby continuously drive the temperature lower and the phase space density higher. The results of evaporative cooling from the first three groups that have obtained BEC have shown that using this rf shaving technique, it is much easier to select high-energy atoms and waste them than it is to cool them.

For the evaporation of the atoms, it is important that atoms with an energy above the cutoff are expelled from the trap. By using RF-evaporation, one can expel the atoms in all three dimensions equally and thus obtain a true 3-D evaporation. In the case of the time orbiting potential (TOP) trap, the atoms are evaporated along the outer side of the cloud that is exposed to the highest magnetic field on the average. This is a cylinder along the direction of rotation axis of the magnetic field and thus the evaporation takes place in 2-D.

Once the energy of the atoms becomes very small, the atoms sag because of gravity and the outer shell of the cloud is no longer at a constant magnetic field. Atoms at the bottom of the trap have the highest energy and thus the evaporation becomes 1-D. In case of harmonic confinement, $U_{\text{trap}} = U''z^2/2$, the equipotential surface is at $z \approx \sqrt{2\eta k_B T/U''}$. Now, the gravitational energy is given by $U_{grav} = mgz$ and thus the limiting temperature for 1-D evaporation to take place is given by [22]

$$k_B T < \frac{2\eta(mg)^2}{\mu B''} \tag{31}$$

For a curvature of $B'' = 500\ \text{T m}^{-2}$, the limiting temperature becomes 1 μK for ^{7}Li, 10 μK for ^{23}Na, and 150 μK for ^{87}Rb. Below this temperature, evaporation becomes less efficient.

In the three experiments that obtained BEC for the first time in 1995, the problem of this "gravitational sag" was not known, but it did not prevent the experimentalist from observing BEC. The solution used in those experiments was because of the light mass (^{7}Li), tight confinement (^{23}Na), and TOP trap (^{87}Rb). In the last case, the axis of rotation is in the z-direction and thus the evaporation always remains 2-D. Table 3 shows typical values of ρ_ϕ for various situations.

Tab. 3 Typical numbers for the phase space density as obtained in the experiments aimed at achieving BEC. The different stages of cooling and trapping the atoms are discussed in the text

Stages	*T*	λ_{deB}	*n*	ρ_ϕ
Oven	300 °C	0.02 nm	10^{10} cm^{-3}	10^{-16}
Slowing	30 mK	2 nm	10^{8} cm^{-3}	10^{-12}
Pre-cooling	1 mK	10 nm	10^{9} cm^{-3}	10^{-9}
Trapping	1 mK	10 nm	10^{12} cm^{-3}	10^{-6}
Cooling	1 μK	0.3 μm	10^{11} cm^{-3}	3×10^{-3}
Evaporation	70 nK	1 μm	10^{12} cm^{-3}	2.612

7 Conclusion

In this article, we have reviewed some of the fundamentals of optical control of atomic motion. The reader is cautioned that this is by no means an exhaustive review of the field, and that many important and current topics have been omitted. Much of the material here was taken from our recent textbook [1], and the reader is encouraged to consult that source for the origin of many of the formulas presented in the present text, as well as for further reading and more detailed references to the literature.

Glossary

Atomic Beam Slowing: Using laser light to slow down the velocities of atoms in a beam.

Bad Collisions: Inelastic scattering of atoms leading to loss of atoms from the trap during evaporation.

Beam Brightening: Increasing the brightness of an atomic beam by using laser cooling.

Bloch Oscillations: The oscillatory motion of particles moving through a periodic potential, predicted in the 1930s by Bloch for electrons in solid state.

Bose–Einstein Condensation (BEC): New state of matter theoretically described in the 1920s by Bose and Einstein and experimentally first observed in dilute gases in 1996, where the interatomic spacing is smaller than the deBroglie wavelength of the atoms.

Brightness: The number of atoms emitted from a source or in a beam per unit of time, per unit of solid angle, and per unit of area of the source.

Capture Range: The range in velocity, where the optical forces are effective.

Damping Force: The optical force on atoms that leads to damping of their velocity.

Dark-spot Magneto-optical Trap: Magneto-optical trap, where in the center of the trap the atoms are not kept in the cycling transition, which reduces the loss of atoms due to inelastic collisions.

Dipole Force Trap: Trap using the dipole optical force to trap atoms.

Dipole Forces: The optical force on the atom caused by the gradient of the light shift of the atom.

Doppler Cooling Limit: Limit of cooling atoms given by the Doppler theory.

Doppler Cooling Techniques: Using the Doppler shift and the detuning of the light from atomic resonance in order to cool atoms.

Doppler Theory: Theory describing the cooling of two-level atoms by exploiting the Doppler shift of the atoms in combination with the detuning of the laser light from resonance.

Ehrenfest Theorem: Theorem by Ehrenfest making the correspondence between classical relations, like $F = -\mathrm{grad}\, V$, and their quantum-mechanical counterpart, in this case $\langle F \rangle = -\mathrm{grad}\langle H \rangle$.

Entropy: Thermodynamic measure of the disorder in the system.

Evaporative Cooling: The preferential removal of high-energy atoms from a gas sample, thereby reducing the temperature of the remaining atoms.

Far-off-resonance Traps: Dipole force trap, where the light is far detuned from resonance, which minimizes the scattering of light due to spontaneous emission.

Fokker-Planck Equation: Equation describing the evolution of the atomic momentum distribution under the combined influence of force and diffusion.

Good Collisions: Elastic scattering of atoms leading to the thermalization of the atoms during evaluation.

Laser Cooling: Using the interaction between laser light and atoms to reduce the average kinetic energy of the atoms.

lin ⊥ lin Cooling: Polarization gradient cooling, where the two beams have perpendicular, linear polarization.

Magnetic Traps: Traps where atoms are trapped due to their magnetic moment by an inhomogeneous magnetic field.

Magneto-optical Trap (MOT): Combination of light forces and inhomogeneous magnetic field, leading to cooling and trapping of atoms.

Majorana Losses: Loss of atoms from magnetic traps, where the atoms in the center of the trap undergo a motion-induced transition to an untrapped state.

Multilevel Atoms: Atoms that have degenerate magnetic substates in ground and excited state, which are coupled by laser light.

Optical Bloch Equations (OBE): Relations describing the evolution of the state vector of the internal state of the atom due to the interaction with laser light.

Optical Forces: The force induced on the atom by laser light.

Optical Lattice: The periodic trapping potential for atoms created by the interference of laser beams.

Optical Molasses: Viscous damping of the atomic velocities by the use of counterpropagating laser beams.

Phase Space Density: Probability of finding a particle in a certain region of space, with a certain range of velocities.

Polarization Gradient Cooling: The use of polarization gradients to cool atoms below the Doppler limit.

Polarization Gradient: The interference pattern that occurs when two counterpropagating laser beams have different polarization. The resulting polarization of the light field is not constant in space, that is, shows a gradient.

Quadrupole Trap: The simplest magnetic trap, where the magnetic field is generated by two separated identical coils carrying opposite currents.

Radiative Forces: The optical force on the atom caused by the scattering of light.

Recoil Limit: Limit in laser cooling determined by the recoil by one photon on the atom.

Saturation Intensity: Intensity of the light where the rate for spontaneous emission and the rate for stimulated emission are equal.

Saturation Parameter: Ratio between the intensity of the laser light and the saturation intensity.

σ^+-σ^- Cooling: Polarization gradient cooling, where the two beams have opposite, circular polarization.

Sisyphus Cooling: The continuous transfer of kinetic energy to potential energy in laser cooling, where the potential energy is radiated away by spontaneously emitted photons. A special case of Sisyphus cooling is polarization gradient cooling to sub-Doppler temperature.

Spontaneous Emission: The return of the atom from the excited state to the ground state by the emission of one photon in a random direction.

Stimulated Emission: The return of the atom from the excited state to the ground state by the emission of one photon in the direction of the laser light.

Sub-Doppler Cooling: Using laser cooling techniques beyond the Doppler techniques to cool atoms below the Doppler limit.

s-wave Scattering: Scattering of atoms at low temperatures, where only the lowest-order partial wave is effective.

Temperature: A measure of the average kinetic energy of the atoms, that is, $k_B T/2 = \langle E_k \rangle$.

Time-of-flight (TOF) Method: Technique to determine the flight time of atoms over a well-defined flight path in order to detect the atomic velocity.

Time-orbiting Potential (TOP) trap: Magnetic quadrupole trap, where the zero of the magnetic field oscillates in space in order to reduce trap loss due to Majorana flips.

Two-level Atoms: Hypothetical atoms that only have one nondegenerate ground state and one nondegenerate excited state, which are resonant with the laser light.

References

[1] Metcalf, H. J., vander Straten, P. (1999), *Laser Cooling and Trapping*. New York: Springer-Verlag.

[2] Phillips, W., Metcalf, H. (**1982**), *Phys. Rev. Lett.* **48**, 596–599.

[3] Prodan, J., Phillips, W., Metcalf, H. (**1982**), *Phys. Rev. Lett.* **49**, 1149–1153.

[4] Prodan, J., Phillips, W. (**1984**), *Prog. Quant. Electron.* **8**, 231–235.

[5] Ertmer, W., Blatt, R., Hall, J. L., Zhu, M. (**1985**), *Phys. Rev. Lett.* **54**, 996–999.

[6] Watts, R., Wieman, C. (**1986**), *Opt. Lett.* **11**, 291–293.

[7] Bagnato, V., Lafyatis, G., Martin, A., Raab, E., Ahmad-Bitar, R., Pritchard, D. (**1987**), *Phys. Rev. Lett.* **58**, 2194–2197.

[8] Molenaar, P. A., vander Straten, P., Heideman, H. G. M., Metcalf, H. (**1997**), *Phys. Rev. A* **55**, 605–614.

[9] Barrett, T. E., Dapore-Schwartz, S. W., Ray, M. D., Lafyatis, G. P. (**1991**), *Phys. Rev. Lett.* **67**, 3483–3487.

[10] Dalibard, J., Phillips, W. (**1985**), *Bull. Am. Phys. Soc.* **30**, 748.

[11] Chu, S., Hollberg, L., Bjorkholm, J., Cable, A., Ashkin, A. (**1985**), *Phys. Rev. Lett.* **55**, 48–51.

[12] Sheehy, B., Shang, S. Q., vander Straten, P. Metcalf, H. (**1990**), *Chem. Phys.* **145**, 317–325.

[13] Gould, P., Lett, P., Phillips, W. D. (1987), New measurement with optical molasses, in W. Persson, S. Svanberg, (Eds.), *Laser Spectroscopy VIII*, Berlin: Springer, pp. 64–67.

[14] Hodapp, T., Gerz, C., Westbrook, C., Futrlehner, C., Phillips, W. (**1992**), *Bull. Am. Phys. Soc.* **37**, 1139.

[15] Lett, P., Watts, R., Westbrook, C., Phillips, W., Gould, P., Metcalf, H. (**1988**), *Phys. Rev. Lett.* **61**, 169–172.

[16] Lett, P. D., Watts, R. N., Tanner, C. E., Rolston, S. L., Phillips, W. D., Westbrook, C. I. (**1989**), *J. Opt. Soc. Am. B* **6**, 2084–2107.

[17] Dalibard, J., Cohen-Tannoudji, C. (**1989**), *J. Opt. Soc. Am. B* **6**, 2023–2045.

[18] Ungar, P. J., Weiss, D. S., Chu, S., Riis, E. (**1989**), *J. Opt. Soc. Am. B* **6**, 2058–2071.

[19] Cohen-Tannoudji, C., Phillips, W. D. (**1990**), *Phys. Today* **43**, 33–40.

[20] Gupta, R., Padua, S., Xie, C., Batelaan, H., Metcalf, H. (**1994**), *J. Opt. Soc. Am. B* **11**, 537–541.

[21] Salomon, C., Dalibard, J., Phillips, W. D., Clairon, A., Guellati, S. (**1990**), *Europhys. Lett.* **12**, 683–688.

[22] Ketterle, W., Vandruten, N. J. (**1996**), *Adv. At. Mol. Opt. Phys.* **37**, 181–236.

[23] Aspect, A., Arimondo, E., Kaiser, R., Vansteenkiste, N., Cohen-Tannoudji, C. (**1988**), *Phys. Rev. Lett.* **61**, 826–829.

[24] Kasevich, M., Chu, S. (**1992**), *Phys. Rev. Lett.* **69**, 1741–1744.

[25] Ashkin, A. (**1970**), *Phys. Rev. Lett.* **24**, 156–159.

[26] Chu, S., Bjorkholm, J., Ashkin, A., Cable, A. (**1986**), *Phys. Rev. Lett.* **57**, 314–317.

[27] Ashkin, A. (**1980**), *Science* **210**, 1081–1088.

[28] Ashkin, A., Dziedzic, J. M. (**1985**), *Phys. Rev. Lett.* **54**, 1245–1248.

[29] Ashkin, A., Dziedzic, J. M. (**1987**), *Science* **235**, 1517–1520.

[30] Miller, J. D., Cline, R. A., Heinzen, D. J. (**1993**), *Phys. Rev. A* **47**, R4567–R4570.

[31] Adams, C. S., Lee, H. J., Davidson, N., Kasevich, M., Chu, S. (**1995**), *Phys. Rev. Lett.* **74**, 3577–3580.

[32] Takekoshi, T., Knize, R. J. (**1996**), *Opt. Lett.* **21**, 77–79.

[33] Metcalf, H., Phillips, W. (**1986**), *Metrologia* **22**, 271–278.

[34] Davidson, N., Lee, H. J., Adams, C. S., Kasevich, M., Chu, S. (**1995**), *Phys. Rev. Lett.* **74**, 1311–1314.

[35] Siegman, A. (1986), *Lasers*. Mill Valley: University Sciences.

[36] Simpson, N., Dholakia, K., Allen, L., Padgett, M. (**1997**), *Opt. Lett.* **22**, 52–54.

[37] McGloin, D., Simpson, N., Padgett, M. (**1998**), *App. Opt.* **37**, 469–472.

[38] Beijersbergen, M. (1996), Phase Singularities in Optical Beams. Ph.D. Thesis, University Leiden, Leiden, The Netherlands.

[39] Ovchinnikov, Yu. B., Manek, I. Grimm, R. (**1997**), *Phys. Rev. Lett.* **79**, 2225–2228.

[40] Aminoff, C. G., Steane, A. M., Bouyer, P., Desbiolles, P., Dalibard, J., Cohen-Tannoudji, C. (**1993**), *Phys. Rev. Lett.* **71**, 3083–3086.

[41] Kasevich, M. A., Weiss, D. S., Chu, S. (**1990**), *Opt. Lett.* **15**, 607–609.

[42] Wineland, D., Itano, W., Bergquist, J., Bollinger, J. (1985), Trapped Ions and Laser Cooling. NBS Technical Note 1086. Washington, DC: US Govt. Printing Office.

[43] Migdall, A., Prodan, J., Phillips, W., Bergeman, T., Metcalf, H. (**1985**), *Phys. Rev. Lett.* **54**, 2596.

[44] Bergeman, T., Erez, G., Metcalf, H. (**1987**), *Phys. Rev. A* **35**, 1535.

[45] Bergeman, T. H., Balazs, N. L., Metcalf, H., Mcnicholl, P., Kycia, J. (**1989**), *J. Opt. Soc. Am. B* **6**, 2249–2256.

[46] Raab, E., Prentiss, M., Cable, A., Chu, S., Pritchard, D. (**1987**), *Phys. Rev. Lett.* **59**, 2631–2634.

[47] Metcalf, H. (**1989**), *J. Opt. Soc. Am. B* **6**, 2206–2210.

[48] Molenaar, P. (1995), Photoassociative Reactions of Laser-Cooled Sodium. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands.

[49] Cornell, E. A., Monroe, C., Wieman, C. E. (**1991**), *Phys. Rev. Lett.* **67**, 2439–2442.

[50] Steane, A. M., Foot, C. J. (**1991**), *Europhys. Lett.* **14**, 231–236.

[51] Steane, A. M., Chowdhury, M., Foot, C. J. (**1992**), *J. Opt. Soc. Am. B* **9**, 2142–2158.

[52] Wallace, C. D., Dinneen, T. P., Tan, K. Y. N., Kumarakrishnan, A., Gould, P. L.,

Javanainen, J. (**1994**), *J. Opt. Soc. Am. B* **5**, 703–711.
[53] Walker, T., Sesko, D., Wieman, C. (**1990**), *Phys. Rev. Lett.* **64**, 408–411.
[54] Sesko, D. W., Walker, T. G., Wieman, C. E. (**1991**), *J. Opt. Soc. Am. B* **8**, 946–958.
[55] Gibble, K. E., Kasapi, S., Chu, S. (**1992**), *Opt. Lett.* **17**, 526–528.
[56] Lindquist, K., Stephens, M., Wieman, C. (**1992**), *Phys. Rev. A* **46**, 4082–4090.
[57] Ketterle, W., Davis, K. B., Joffe, M. A., Martin, A., Pritchard, D. E. (**1993**), *Phys. Rev. Lett.* **70**, 2253–2256.
[58] Letokhov, V. S. (**1968**), *JETP Lett.* **7**, 272.
[59] Salomon, C., Dalibard, J., Aspect, A., Metcalf, H., Cohen-Tannoudji, C. (**1987**), *Phys. Rev. Lett.* **59**, 1659–1662.
[60] Castin, Y., Dalibard, J. (**1991**), *Europhys. Lett.* **14**, 761–766.
[61] Jessen, P. S., Gerz, C., Lett, P. D., Phillips, W. D., Rolston, S. L., Spreeuw, R. J. C., Westbrook, C. I. (**1992**), *Phys. Rev. Lett.* **69**, 49–52.
[62] Verkerk, P., Lounis, B., Salomon, C., Cohen-Tannoudji, C., Courtois, J. Y., Grynberg, G. (**1992**), *Phys. Rev. Lett.* **68**, 3861–3864.
[63] Lounis, B., Verkerk, P., Courtois, J. Y., Salomon, C., Grynberg, G. (**1993**), 1-D *Europhys. Lett.* **21**, 13–17.
[64] Gupta, R., Padua, S., Xie, C., Batelaan, H., Bergeman, T., Metcalf, H. (**1992**), *Bull. Am. Phys. Soc.* **37**, 1139.
[65] Gupta, R., Padua, S., Bergeman, T., Metcalf, H. (1993), Search for motional quantization of laser-cooled atoms, in E. Arimondo, W. Phillips, F. Strumia (Eds.), *Laser Manipulation of Atoms and Ions, Proceedings of Fermi School CXVIII, Varenna*, Amsterdam: North Holland.
[66] Anderson, B. P., Kasevich, M. A. (**1998**), *Nature* **282**, 1686–1689.
[67] Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W., Bloch, I. (**2002**), *Nature* **415**, 39–44.
[68] Grynberg, G., Lounis, B., Verkerk, P., Courtois, J. Y., Salomon, C. (**1993**), *Phys. Rev. Lett.* **70**, 2249–2252.
[69] Ben Dahan, M., Peik, E., Reichel, J., Castin, Y., Salomon C. (**1996**), *Phys. Rev. Lett.* **76**, 4508–4511.
[70] Westbrook, C. I., Watts, R. N., Tanner, C. E., Rolston, S. L., Phillips, W. D., Lett, P. D., Gould, P. L. (**1990**), *Phys. Rev. Lett.* **65**, 33–36.
[71] Birkl, G., Gatzke, M., Deutsch, I. H., Rolston, S. L., Phillips, W. D. (**1995**), *Phys. Rev. Lett.* **75**, 2823–2826.
[72] Masuhara, N., Doyle, J. M., Sandberg, J. C., Kleppner, D., Greytak, T. J., Hess, H. F., Kochanski, G. P. (**1988**), *Phys. Rev. Lett.* **61**, 935.
[73] Hess, H. F. (**1986**), *Phys. Rev. B* **34**, 3476.
[74] Doyle, J. M., Sandberg, J. C., Yu, I. A., Cesar, C. L., Kleppner, D., Greytak, T. J. (**1991**), *Phys. Rev. Lett.* **67**, 603.
[75] Luiten, O. J., Werij, H. G. C., Setija, I. D., Reynolds, M. W., Hijmans, T. W., Walraven, J. T. M. (**1993**), *Phys. Rev. Lett.* **70**, 544–547.
[76] Setija, I. D., Werij, H. G. C., Luiten, O. J., Reynolds, M. W., Hijmans, T. W., Walraven, J. T. M. (**1993**), *Phys. Rev. Lett.* **70**, 2257–2260.
[77] Davis, K. B., Mewes, M. O., Ketterle, W. (**1995**), *App. Phys. B* **60**, 155–159.
[78] Bagnato, V., Pritchard, D. E., Kleppner, D. (**1987**), *Phys. Rev. A* **35**, 4354.
[79] Monroe, C., Cornell, E., Sackett, C., Myatt, C., Wieman, C. (**1993**), *Phys. Rev. Lett.* **70**, 414.
[80] Luiten, O. J., Reynolds, M. W., Walraven, J. T. M. (**1996**), *Phys. Rev. A* **53**, 381.
[81] Huang, K. (1963), *Statistical Mechanics*. New York: Wiley.
[82] Landau, L. D., Lifshitz, E. M. (1958), *Quantum Mechanics (Non-Relativistic Theory)*. Oxford: Pergamon Press.

Lasers, Semiconductor

Nagaatsu Ogasawara
University of Electro-Communications, Department of Electronics Engineering, Tokyo, Japan

Ryoichi Ito
Meiji University, Department of Physics, Kawasaki, Japan
Phone/Fax: +81-44-934-7425; e-mail: ito-ryoichi@sam.hi-ho.ne.jp

Abstract
Following a brief history of semiconductor lasers, this article describes their operation principles in relation to their device structures and materials. Their fundamental operation characteristics, including the static, spectral, and dynamic aspects, are then presented. Finally, the current state of the art in semiconductor laser technology is described by reviewing up-to-date devices incorporating new structures and new functions.

Keywords
semiconductor laser; laser diode; photonics; compact light source; heterostructure; fiber-optic telecommunications; optical disk.

1 Introduction

Semiconductor lasers have now grown to be key components in modern photonics technology, most notably as light sources in fiber-optic telecommunications systems and optical disk systems. They have also found an increasing number of applications ranging from instruments such as bar code scanners and computer printers to solid-state laser pumps and measuring and sensoring equipment for engineering use.

Semiconductor lasers have the following features that distinguish them from other lasers.

1. *Compactness*: The typical size of a laser chip is $300 \times 200 \times 100\,\mu m^3$. The small chip contains all the ingredients of a laser structure: a resonator, a waveguide, an active medium, and a *p-n* junction to pump the active medium.

2. *High efficiency*: Semiconductor lasers can be driven by low electrical power [(several tens of mA) × (1–2 V)]. The efficiency of converting electrical power into optical power is several tens of percent, which should be compared with that of 0.1% or less for gas and other lasers.

3. *Capability for high-speed direct modulation*: Light output can be modulated at frequencies of 10 GHz or more simply by modulating the pumping current.

4. *Wide emission spectrum*: The emission wavelength is determined by the band-gap energy of the active-layer material. The blue-near-infrared region is covered by III–V compounds. The infrared region (3–30 μm) is covered by IV–VI compounds.

5. *High reliability*: Semiconductor lasers have become highly reliable devices because of the remarkable improvement in crystal quality, although rapid degradation was a great problem at the early stage of their development.

6. *Sensitivity to temperature*: The threshold current and the lasing wavelength strongly depend on temperature. This is a disadvantage for most applications. On the other

hand, wavelength tunability is useful in spectroscopic research.

This article is intended to serve as a useful reference for those who wish to grasp the basic concepts of semiconductor lasers and utilize them in engineering and research activities. Following this introduction is a brief history of semiconductor lasers. Next, the operation principles of semiconductor lasers are described in relation to their structures and materials. The fourth section presents a general view of fundamental operation characteristics covering the static, spectral, and dynamic aspects. The fifth section gives the current state of the art in semiconductor laser technology by reviewing up-to-date devices incorporating new structures and new functions. Because of the limited space, citation of published papers has been kept to a minimum. Several books listed in the (Further Reading) section can be consulted both for further study and for access to the relevant published papers.

2 History

The advent of semiconductor injection lasers dates back to 1962, only two years later than the first achievement of laser action in ruby, when stimulated emission from forward-biased GaAs diodes was demonstrated [1–3]. The stimulated emission resulted from the radiative recombination of electrons and holes in the direct-gap semiconductor GaAs. The electrons and holes were injected by forward bias into the depletion region in the vicinity of the *p-n* junction. Optical feedback was provided by polished facets perpendicular to the junction plane. The demonstration of GaAs lasers prompted the exploration of many other III–V and IV–VI compound semi-conductors. Unfortunately, these early homojunction devices, consisting of a single semiconductor, were not considered for serious applications since continuous operation at room temperature was not feasible because of their high threshold current densities for lasing (several tens of kA cm^{-2}).

A breakthrough was brought about in 1970 by employing a double heterostructure grown by liquid-phase epitaxy [4, 5]. In the double heterostructure, stimulated emission occurred only within a thin active layer of GaAs sandwiched between *p*- and *n*-doped AlGaAs layers that have a wider band gap. The threshold current density was dramatically reduced to several kA cm^{-2} or less, and continuous operation at room temperature was eventually accomplished. The invention of the double heterostructure was obviously the most important step in the history of the development of semiconductor lasers toward practical utility. For the invention of heterostructure, Alferov and Kroemer were honored with a Nobel Prize in physics in 2000, while Alferov, Hayashi and Panish shared the Kyoto Prize in 2001.

There still remained, however, a troublesome problem of degradation; the reliability of the early double-heterostructure devices was very poor and many of them stopped their continuous operation within a few or several tens of minutes. Fortunately, intensive investigations of the degradation mechanism led to the identification of the causes of major failure: the recombination-enhanced growth of dislocations, and facet erosion [6 Part B, Chapter 8]. These studies have revealed that a long operating life is obtained by eliminating dislocation generation during material growth and device processing and

by coating the mirror facets with dielectric thin films.

Another task undertaken in parallel with the improvement of reliability was the stabilization of the optical mode in the lateral transverse direction (parallel to the junction plane). It had been recognized by the late 1970s that the lateral-mode instability often observed at higher output powers was detrimental to most applications and that such an instability could be effectively suppressed by incorporating a built-in refractive-index profile in the lateral direction [7]. A large number of laser structures have been proposed and demonstrated to introduce lateral variations of the refractive index.

As a result of these research and development efforts during the first two decades, the maturity of AlGaAs lasers reached the stage of practical use in printers and disk players early in the 1980s. Meanwhile, the rapid reduction of transmission loss in optical fibers achieved during the 1970s stimulated intensive development efforts on GaInPAs lasers, which have emission wavelengths (1.3–1.55 μm) in the low-loss and dispersion-free region of silica fibers. During the first half of the 1980s, GaInPAs lasers emerged as indispensable light sources in telecommunications systems using silica fibers (see OPTICAL COMMUNICATIONS).

Progress has continued to date in both device structures and materials. In the following, let us look at some major topics.

The idea of utilizing periodic gratings incorporated in a laser medium as a means of optical feedback was proposed by Kogelnik and Shank in [8]. Distributed-feedback and distributed-Bragg-reflector semiconductor lasers utilizing this principle have since become the major device structures used in long-haul, high bit-rate telecommunications systems because of the stable single–longitudinal-mode oscillation they exhibit even under high-speed modulation.

Recent advances in growth technology of semiconductor ultrathin layers by molecular-beam epitaxy and metal-organic vapor-phase epitaxy have also created a new class of laser structures. Quantum-well lasers incorporating such ultrathin active layers (~20 nm or less) have been shown to display a variety of superior laser characteristics, such as low-threshold currents and wide modulation bandwidths, which are attributed to the size quantization of electrons in the ultrathin active layers [9]. Quantum-well structures as well as multiple quantum-well structures are now commonly employed in most semiconductor lasers.

In the visible spectrum, GaInP lasers were first commercialized as red-emitting lasers (0.63–0.69 μm) in 1988 by Sony, Toshiba, and NEC. They are now extensively used in DVD systems. Blue–green laser emission was achieved first in ZnCdSe lasers [10] and then in InGaN lasers [11, 12]. InGaN blue lasers are being employed in second-generation DVD players (see DATA STORAGE, OPTICAL).

3 Structures, Materials, and Operation Principles

3.1 Double Heterostructure

A semiconductor laser is a diode, as shown schematically in Fig. 1. When a forward current is passed through the *p*-*n* junction, electrons and holes injected into the active region from the *n* and *p* regions respectively, recombine to emit radiation. The wavelength λ of the radiation is

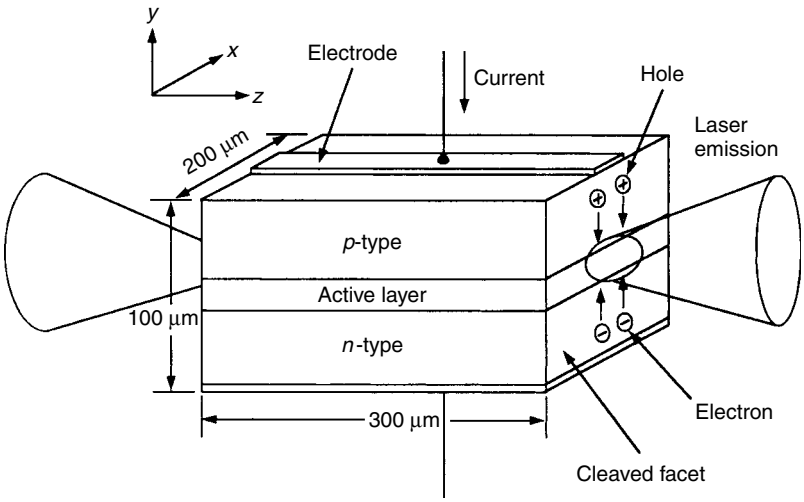

Fig. 1 Schematic illustration of a double-heterostructure semiconductor laser

basically determined by the relation $\lambda \sim hc/E_g$, where h is the Planck constant, c is the light velocity, and E_g is the band-gap energy of the active-region material.

For low injection currents, light is emitted through spontaneous emission. In order for lasing to take place, a sufficiently high concentration of carriers must be accumulated within the active region to induce population inversion. This is effectively accomplished by adopting a double-heterojunction (DH) structure, where a thin active layer, typically ~0.1 μm thick, but as thin as 10 nm in quantum-well lasers, is sandwiched between *n*- and *p*-type cladding layers, which have wider band gaps than the active layer. Electrons and holes injected into the active layer through the heterojunctions are confined within the thin active layer by the potential barriers at the heteroboundaries, as illustrated in Fig. 2(a). The DH structure forms an efficient optical waveguide as well, because of the refractive-index difference between the active and cladding layers shown in Figs. 2(b) and 2(c). Thus, the DH structure facilitates the interaction needed for laser action between the optical field and the injected carriers.

Let us use a GaAs(active)/$Al_xGa_{1-x}As$ (cladding) DH structure as an example. The dependence of the band-gap energy of $Al_xGa_{1-x}As$ on the AlAs mole fraction x can be approximated for $x < 0.42$ (direct-gap region) by [6 Part A, p. 193].

$$E_g = 1.424 + 1.247x \text{ eV}. \tag{1}$$

The energy-gap difference between the active and cladding layers ΔE_g is divided into the band discontinuities ΔE_c in the conduction band and ΔE_v in the valence band according to [13].

$$\Delta E_c \sim 0.62 \Delta E_g \tag{2}$$

and

$$\Delta E_v \sim 0.38 \Delta E_g. \tag{3}$$

The efficiency of the carrier confinement depends strongly on the magnitude of the band discontinuity. The AlAs mole fraction x of the cladding layers is usually chosen to be greater than that of the active

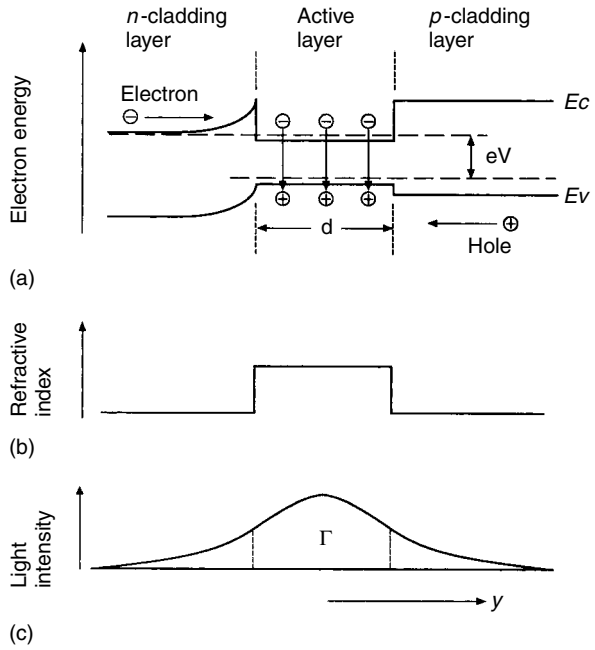

Fig. 2 Diagram illustrating carrier confinement and waveguiding in a double heterostructure (a) Energy-band diagram at high forward bias; (b) refractive-index distribution; (c) light intensity distribution. E_c and E_v: The edges of the conduction and valence bands. V: Applied voltage. d: Active-layer thickness (~0.1 μm). Γ: Confinement factor

layer by ~0.3. This gives rise to a ΔE_c approximately 10 times as large as the room-temperature thermal energy, which is sufficient to suppress electron diffusion over the heterobarriers. The hole leakage current is less important because of the smaller diffusion constant for holes.

The characteristics of a three-layer slab waveguide are conveniently described in terms of the normalized waveguide thickness D, defined as

$$D = \left(\frac{2\pi}{\lambda}\right) d\sqrt{\eta_a^2 - \eta_c^2}, \qquad (4)$$

where η_a and η_c are the refractive indices of the active and cladding layers respectively and d is the active-layer thickness. For example, the condition that a waveguide supports only the lowest-order fundamental mode is expressed as

$$D < \pi. \qquad (5)$$

Meanwhile, the refractive index of $Al_xGa_{1-x}As$, $\eta(x)$, can be approximated for a light wavelength of ~0.9 μm by [6, Part A, p. 45]

$$\eta(x) = 3.590 - 0.710x + 0.091x^2. \qquad (6)$$

Thus, for a waveguide with a GaAs active layer surrounded by $Al_{0.3}Ga_{0.7}As$ layers and a wavelength λ of 0.87 μm, the condition of Eq. (5) simply becomes

$$d < 0.36\,\mu m. \qquad (7)$$

Since the thickness d is typically 0.1 to 0.2 μm, the single, fundamental-mode condition is almost always satisfied in practical devices.

The confinement factor Γ, defined as the fraction of the electromagnetic energy of the guided mode that exists within the active layer, is an important parameter representing the extent to which the waveguide mode is confined to the active layer. Γ for a fundamental mode is approximately given by [14]

$$\Gamma \sim \frac{D^2}{2 + D^2}. \tag{8}$$

For a $GaAs/Al_{0.3}Ga_{0.7}As$ waveguide with $d = 0.1$ μm, $\Gamma \sim 0.27$.

3.2 Fabry–Pérot Cavity

In addition to the optical gain, optical feedback is the other ingredient for laser oscillation. This is provided by a pair of mirror facets at both ends of the devices. These facets are normally formed simply by cleaving the crystal. Since the refractive index of major semiconductor laser materials is ∼3.6, the reflectivity of the mirror is ∼0.3. This is very low compared to other types of lasers but is still sufficient to provide optical feedback in semiconductor lasers.

3.3 Lasing Threshold Condition

When a sufficient number of electrons and holes is accumulated to form an inverted population, the active region exhibits optical gain and can amplify light passing through it since stimulated emission overcomes interband absorption. The condition for self-sustained laser oscillation to occur is that the light makes a full round trip in the cavity without attenuation; that is, the optical gain should equal the losses both inside the cavity and through the partially reflecting end facets. Thus, the gain coefficient at threshold g_{th} is given by the relation

$$\Gamma g_{th} = \underbrace{\Gamma\alpha_a + (1-\Gamma)\alpha_c + \alpha_s}_{\alpha_i} + \frac{1}{L}\ln\frac{1}{R}. \tag{9}$$

Here, α_a and α_c denote the losses in the active and cladding layers respectively, due to free-carrier absorption. α_s accounts for scattering loss due to heterointerfacial imperfections. The first three loss terms on the right-hand side combined are termed *internal loss* α_i and add up to 10 to 20 cm^{-1} [6, Part A, p. 174–176]. The reflection loss $L^{-1}\ln R^{-1}$ due to output coupling (∼40 cm^{-1} for $L \sim 300$ μm, $R \sim 0.3$) is normally the largest among the loss terms. Despite the fact that both the internal and the reflection losses are exceptionally large as compared to other lasers, lasing is brought about by virtue of the high optical gain in semiconductors.

It is not an easy task to analyze precisely the optical gain spectrum in semiconductor lasers. Fortunately, however, we can utilize a phenomenological linear relationship between the maximum gain g (the peak value of the gain spectrum for a given carrier density) and the injected carrier density n,

$$g(n) = \frac{\partial g}{\partial n}(n - n_t), \tag{10}$$

to a good approximation [15–17]. Here, $\partial g/\partial n$ is termed *differential gain*, and n_t denotes the carrier density required to achieve transparency where stimulated emission balances against interband absorption corresponding to the onset of

population inversion. For GaAs lasers,

$$\frac{\partial g}{\partial n} \sim 3.5 \times 10^{-16}\ \mathrm{cm}^2 \qquad (11)$$

and

$$n_t \sim 1.5 \times 10^{18}\ \mathrm{cm}^{-3}. \qquad (12)$$

Substituting Eqs. (10) to (12) into Eq. (9) and assuming that $\Gamma = 0.27$, $\alpha_i = 10\ \mathrm{cm}^{-1}$, and $L^{-1} \ln R^{-1} = 40\ \mathrm{cm}^{-1}$, we get a threshold carrier density n_{th} of $\sim 2 \times 10^{18}\ \mathrm{cm}^{-3}$.

The threshold current density J_{th} is expressed as

$$J_{th} = \frac{\mathrm{edn}_{th}}{\tau_s}, \qquad (13)$$

where τ_s is the carrier lifetime due to spontaneous emission. Assuming that $\tau_s = 2$–4 ns and $d = 0.1\ \mu\mathrm{m}$, we obtain a threshold current density J_{th} or $\sim 1\ \mathrm{kA\ cm}^{-2}$.

3.4 Stripe Geometry

Most modern semiconductor lasers adopt a stripe geometry, where current is injected only within a narrow region beneath a stripe contact several μm wide, in order to keep the threshold current low and to control the optical field distribution in the lateral direction. As compared with broad-area lasers, where the entire laser chip is excited, the threshold current of the stripe-geometry lasers is reduced roughly proportional to the area of the contact.

Figure 3(a) shows the simplest form of the stripe geometry [18]. Lasing occurs in a limited region of the active layer beneath the stripe contact where a high density of current flows. Such lasers are termed *gain-guided lasers* because the optical intensity distribution in the lateral direction (x direction) is determined by the gain profile produced by carrier density distribution.

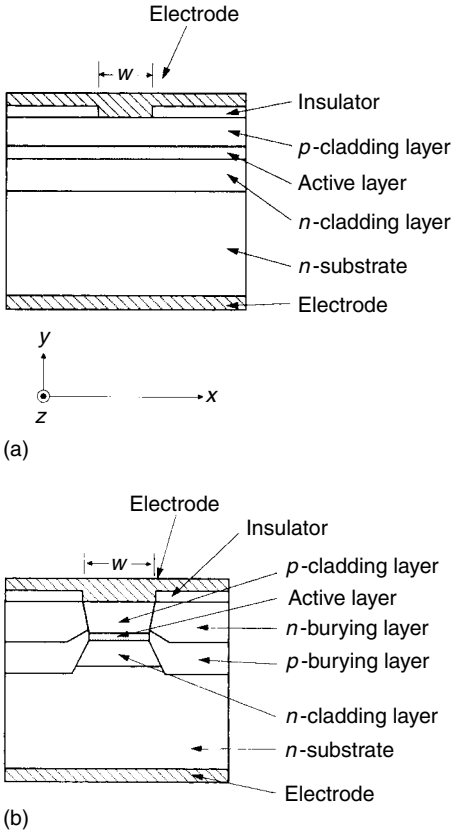

Fig. 3 Two representative examples of stripe-geometry lasers (front view of the end facet) (a) Stripe-contact laser; (b) Buried heterostructure (BH) laser. *W*: Stripe width (2–10 μm)

Devices incorporating a built-in refractive-index variation in the lateral direction are termed *index-guided lasers*. Figure 3(b) shows the structure of a buried heterostructure (BH) laser [19] as a representative example of index-guided lasers. The active region is surrounded by materials with lower refractive indices in both the vertical (y) and lateral (x) transverse directions, thus forming a waveguide structure in both directions. In the BH laser, the

width of current flow is delineated by the *p-n* junction formed by the burying layers, which is reverse biased when the active region is forward biased.

3.5 Materials and Emission Wavelengths

Semiconductor-laser materials that have been studied range over III–V, IV–VI, and II–VI compounds as listed in Table 1.

3.5.1 III–V Compounds

III–V compounds are the most important and popular laser materials [6, Part B, Chapter 5]. In particular, AlGaAs (active layer)/AlGaAs (cladding layer)/GaAs (substrate) lasers emitting at 0.7 to 0.9 μm and GaInP/AlGaInP/GaAs lasers emitting at 0.63 to 0.69 μm are extensively used in optical disk systems and laser printers. GaInPAs/InP/InP lasers emitting at 1.2 to 1.6 μm are used in fiber-optic communications systems. GaInN/AlGaN/sapphire lasers emitting at 0.38 to 0.45 μm are starting to be used in second-generation DVD players. All of them exhibit low-threshold, room-temperature continuous wave (cw) operation with high reliability.

The common features among them are as follows:

1. The active layer consists of direct-gap materials.
2. Binary compounds (GaAs, InP) are used as the substrate except for GaInN/AlGaN/sapphire lasers, for which GaN substrate will be used in the future.
3. The active and cladding layers have nearly the same lattice constant as the substrate.

Tab. 1 Major semiconductor-laser materials and their emission wavelengths

Materials (active/cladding/substrate)	*Emission wavelengths [μm]*
III–V compounds	
AlGaAs/AlGaAs/GaAs	0.7–0.9
GaInPAs/InP/InP	1.2–1.6
GaInP/AlGaInP/GaAs	0.66–0.69
GaInPAs/GaInP/GaAs or GaPAs	0.65–0.9
GaInPAs/AlGaAs/GaAs	0.62–0.9
AlGaAsSb/AlGaAsSb/GaSb	1.1–1.7
GaInAsSb/AlGaAsSb/GaSb or InAs	2–4
InPAsSb/InPAsSb/GaSb or InAs	2–4
GaInAs(strained)/AlGaAs/GaAs	0.9–1.1
GaInPAs(strained)/GaInPAs/InP	~1.55
GaInN/AlGaN/sapphire	0.38–0.45
IV–VI compounds	
PbSnTe/PbSnSeTe/PbTe	6–30
PbSSe/PbS/PbS	4–7
PbEuTe/PbEuTe/PbTe	3–6
II–VI compounds	
ZnCdSe/ZnSSe/GaAs	~0.5

The use of direct-gap semiconductors, featuring efficient radiative recombination, as the active-region material (**1**) is essential in achieving laser action. Features **2** and **3** are related to the crystal quality of the hetero-interfaces. To minimize the generation of lattice defects that may impair the device reliability, the lattice parameter of the active and cladding layers should be matched to that of the high-quality binary substrate (the lattice-matching condition). The lattice parameter of $Al_xGa_{1-x}As$ is essentially independent of alloy concentration x. This is the only exception in ternary alloys, arising from the fortuity that GaAs and AlAs have virtually the same lattice parameter, the difference being as small as ~0.1%. In quaternary alloys like $Ga_xIn_{1-x}P_yAs_{1-y}$ and $Al_xGa_yIn_{1-x-y}P$, the band-gap energy can be tuned in a certain range for a fixed lattice parameter by choosing appropriate pairs of x and y values. The design of lattice-matched DH structures is facilitated by taking advantage of this degree of freedom in quaternary alloys. In some quantum-well structures, a certain degree of strain in the active material caused by inevitably or deliberately introduced slight lattice mismatching is used to provide better laser performance.

3.5.2 IV–VI Compounds

PbSnTe and PbSSe lasers emitting in the infrared region (3–30 μm) [20] are used in high-resolution gas spectroscopy and in air pollution monitoring. These lasers operate only at cryogenic temperatures. However, an appreciable spectral tuning is feasible by simply changing temperature and injection current, which is extremely desirable for spectroscopy.

3.5.3 II–VI Compounds

ZnCdSe/ZnSSe/ZnMgSSe/GaAs quantum-well lasers emitting in the blue–green spectral region emerged around 1991 as a result of success in overcoming the difficulty of *p*-type doping into these materials. [10, 21] Despite extensive efforts, however, these lasers have never attained device life times exceeding several hundred hours and hence have not been put into practical use.

4 Fundamental Characteristics

In this section, several aspects of fundamental characteristics are described. Numerical examples are given based on AlGaAs ($\lambda \sim 0.8$ μm) and GaInPAs ($\lambda \sim 1.3$ μm) lasers.

4.1 Light–Current Characteristics

Figure 4 shows an example of output power (P) versus DC injection current (I) characteristics. The ordinate is the light power emitted from one end facet, and essentially the same power is emitted from the other end facet.

The threshold current at room temperature is typically several tens of mA. The dependence of the threshold current I_{th} on device temperature T is phenomenologically expressed as

$$I_{th} \propto \exp\left(\frac{T}{T_0}\right), \qquad (14)$$

where T_0 is a constant referred to as the characteristic temperature and takes a value of 100 to 150 and 50 to 80 K for AlGaAs and GaInPAs lasers, respectively.

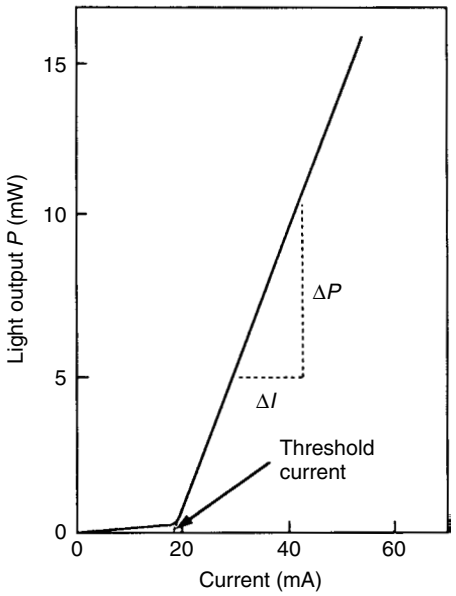

Fig. 4 Light output versus injection current curve in an AlGaAs laser

The light power emitted by injection above threshold is expressed as

$$P = \hbar\omega_l \frac{I - I_{\text{th}}}{e} \eta_{\text{i}} \frac{\frac{1}{2}L^{-1}\ln(1/R)}{\alpha_{\text{i}} + L^{-1}\ln(1/R)}. \quad (15)$$

Here, $(I - I_{\text{th}})/e$ is the rate of excess carrier injection beyond the threshold. η_{i} is the internal quantum efficiency representing the fraction of injected carriers that recombine radiatively and generate photons of energy $\hbar\omega_l$. The factor $1/2L^{-1}\ln R^{-1}/(\alpha_{\text{i}} + L^{-1}\ln R^{-1})$ represents the fraction of the generated photons that are coupled out of the cavity through one of the end facets. From Eq. (15), the differential efficiency η_D, defined as the slope of output power versus current curve above threshold, is written as

$$\eta_D = \frac{\Delta P}{\Delta I} = \frac{\hbar\omega_l}{e} \eta_{\text{i}} \frac{\frac{1}{2}L^{-1}\ln(1/R)}{\alpha_{\text{i}} + L^{-1}\ln(1/R)}. \quad (16)$$

The differential external quantum efficiency η_{ext}, defined as the ratio of the number of photons coupled out to the number of carriers injected, is given by

$$\eta_{\text{ext}} = \frac{\Delta P/\hbar\omega_l}{\Delta I/e} = \eta_{\text{i}} \frac{\frac{1}{2}L^{-1}\ln(1/R)}{\alpha_{\text{i}} + L^{-1}\ln(1/R)}. \quad (17)$$

Above threshold, the stimulated emission predominates over the nonradiative processes and $\eta_{\text{i}} \sim 1$. Typical values of $L = 300\,\mu\text{m}$, $R = 0.3$, and $\alpha_{\text{i}} = 20\,\text{cm}^{-1}$ give $\eta_{\text{ext}} \sim 0.3$. Since $\hbar\omega_l \sim E_{\text{g}}$, $\eta_D \sim (E_{\text{g}}/e)\eta_{\text{ext}}$. For GaAs lasers ($E_{\text{g}} = 1.424$ eV), we get $\eta_D \sim 0.4$ W/A.

The maximum rating for output power is typically 5 to 30 mW. Figure 5 shows schematically the three phenomena that determine the maximum power rating.

1. *Kink*: The nonlinearity in the curve of light versus current caused by lateral-mode instability is referred to as a kink. Since a kink is often associated with a beam-profile shift and the generation of intensity noise, lasers cannot be used

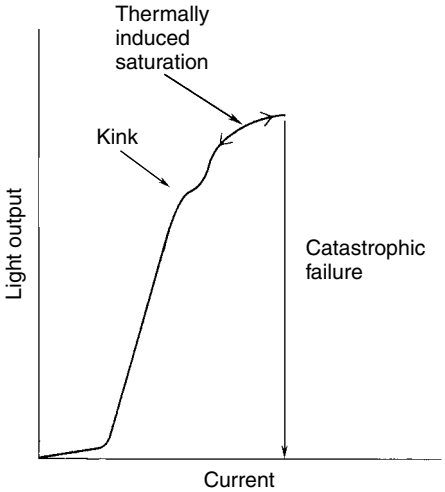

Fig. 5 Schematic illustration of the three phenomena occurring at high injection levels

in the kink region for most applications. The lateral-mode instability arises from the optical power dependence of the refractive index. By incorporating a refractive-index profile in the lateral direction to form a waveguide, the power level for kink generation can be appreciably increased.

2. *Catastrophic optical damage*: For optical power densities higher than several MW cm^{-2}, the end facets of semiconductor lasers may be melted. This results in a sudden decrease of output power and failure of the device. The catastrophic damage is considered to be brought about by the enhanced light absorption associated with surface states at the end facets. The importance of this process seems to depend on materials and photon energies. In GaInPAs lasers ($\lambda \sim 1.3$ μm), the catastrophic optical damage practically does not occur.

3. *Thermally induced saturation*: Heating of the junction under high current injections raises the threshold current. This may lead to saturation of output power with an increase in injection current, especially in lasers with lower T_0 values.

4.2 Beam Profile and Polarization

Typical radiation intensity patterns of laser diodes are shown in Figs. 6 and 7. Shown in Fig. 6 are the near-field patterns, that is, the spatial distributions of optical intensity on the end facet in the directions (a) perpendicular and (b) parallel to the junction plane. The far-field pattern (Fig. 7) is the angular distribution of radiant intensity measured at distances several mm or more away from the facet, which is mathematically a Fourier transform of the near-field pattern. The single-lobed radiation patterns show that the transverse mode of the device is single and fundamental. Transverse-mode-stabilized lasers incorporating appropriate waveguide structures exhibit such beam profiles stably up to reasonable output power levels.

Normally, the angular width or beam divergence (the full width at half maximum) perpendicular to the junction, $\theta_\perp$ (20°–60°), is larger than that parallel to the junction $\theta_\perp$ (10°–30°) since the near-field spot is an ellipsoid because of the large difference between the stripe width

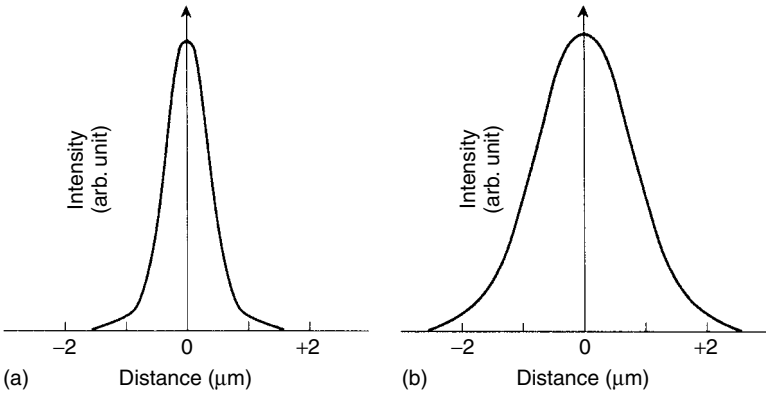

Fig. 6 Near-field optical intensity patterns (a) perpendicular; (b) parallel to the junction plane

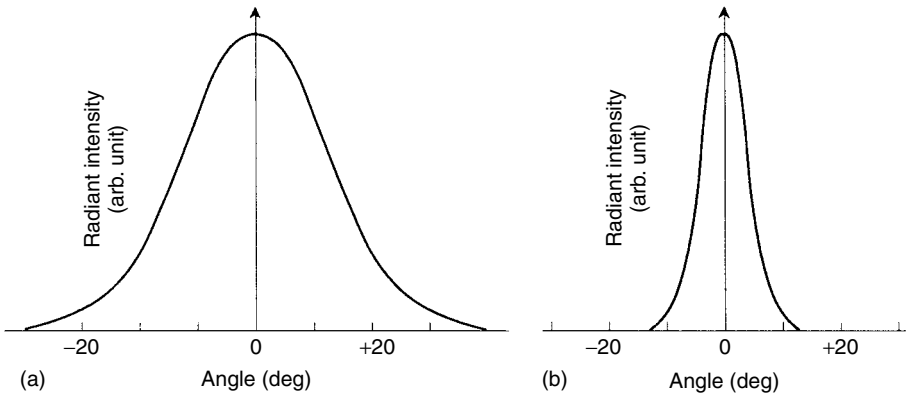

Fig. 7 Far-field radiant intensity patterns (a) perpendicular; (b) parallel to the junction plane

(several μm) and the active-layer thickness (~0.1 μm). $\theta_\perp$ can be reduced by adopting thinner active layers; then, most of the optical energy penetrates into the cladding layers, increasing the spot size perpendicular to the junction.

The output beam from gain-guided lasers can be considerably astigmatic; the beam waist perpendicular to the junction plane is located essentially on the end facet, while the virtual beam waist along the junction plane is formed in the cavity at a position several tens of μm behind the end facet. This is brought about by the difference of the waveguiding mechanism between the two directions. In contrast, in index-guided lasers, the beam waist, both parallel and perpendicular to the junction plane, is located within several μm of the facet. The astigmatism should be taken into consideration when the beam is coupled into lenses.

The laser beam is linearly polarized along the junction plane. This is because, in a slab waveguide, the facet reflectivity for the TE mode is higher than that for the TM mode [22]. Since the gain needed to reach threshold depends on the facet reflectivity (see Eq. (9)), the TE mode is selected for oscillation in DH lasers because of the higher reflectivity.

4.3 Spectral Characteristics

Figure 8 shows the variations of lasing spectrum with output power in an index-guided AlGaAs laser. At lower powers, the laser oscillates in several longitudinal modes. Here, the longitudinal-mode wavelength λ_N is determined by the relation

$$\left(\frac{1}{2}\frac{\lambda_N}{\eta}\right) N = \mathrm{L} \quad (N : \text{positive integers}), \tag{18}$$

where η is the effective refractive index of the waveguide and L is the cavity length. The mode spacing $\Delta\lambda$ is given by

$$\Delta\lambda = \frac{\lambda^2}{2\eta_g L}, \tag{19}$$

where $\eta_g = \eta - (\partial\eta/\partial\lambda)\lambda$ is the group index. $\Delta\lambda$ is typically 3 Å ($\lambda = 0.8$ μm)–8 Å ($\lambda = 1.3$ μm) for $L = 300$ μm.

As injection is increased, the laser power tends to concentrate in a single-longitudinal mode, while the power in the

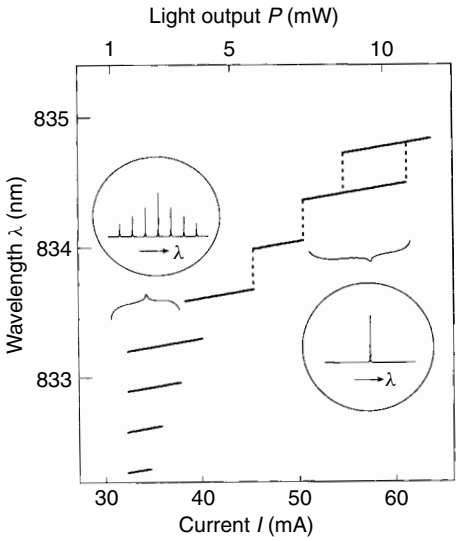

Fig. 8 Lasing wavelength versus injection current characteristics in an AlGaAs laser

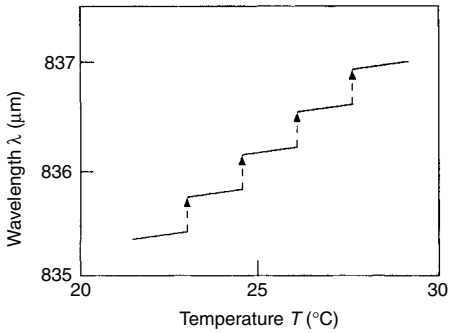

Fig. 9 Temperature dependence of the lasing wavelength

remaining modes saturates. This is common among index-guided lasers. Gain-guided lasers, in contrast, tend to exhibit multiple–longitudinal-mode spectra even at higher powers.

It can be seen in Fig. 8 that the lasing wavelength shifts toward longer wavelengths as the output power is increased. This is induced by the temperature rise in the active region with the increase in injection current. Shown in Fig. 9 is an example of the wavelength shift due to a heat sink temperature change at a fixed injection current. Each longitudinal-mode shifts at a rate of 0.5 Å K^{-1} ($\lambda \sim$ 0.8 μm)–0.8 Å K^{-1} ($\lambda \sim 1.3$ μm) because of the temperature dependence of the refractive index. In addition, the lasing wavelength jumps toward longer wavelengths, as temperature is raised, at a rate of 2 Å K^{-1} ($\lambda \sim 0.8$ μm)–5 Å K^{-1} ($\lambda \sim 1.3$ μm). This is caused by the temperature dependence of the band-gap energy.

At ranges of temperature and bias current where the oscillation mode jumps to a neighboring mode, it is often observed that the lasing is randomly switched between the two longitudinal modes. The random switching is associated with a large-intensity noise since the output power is different between modes by (0.1–1%). Some devices exhibit hysteresis in the lasing wavelength versus temperature and/or injection-current characteristics. These phenomena have been interpreted in terms of nonlinear mode coupling among longitudinal modes [17].

The spectral linewidth of a single-longitudinal mode is inversely proportional to the output power to a good approximation. The product of the linewidth Δf and the output power P is typically in the range of 1 to 100 MHz mW. This value is 10 to 50 times larger than Δf_{ST} obtained from the well-known Schawlow–Townes formula. An analysis shows that the enhancement is expressed as [23]

$$\Delta f = \Delta f_{ST}(1 + \alpha^2), \qquad (20)$$

where α is the linewidth enhancement factor defined as

$$\alpha = -\frac{4\pi}{\lambda}\frac{(\partial\eta/dn)}{(\partial g/dn)} \qquad (21)$$

by the derivatives of the refractive index η and the gain g with respect to the carrier density n. α represents the magnitude of the amplitude-phase coupling, inherent in semiconductor lasers, originating from the strong dependence of the refractive index on the carrier density. The linewidth is determined not only by the direct phase fluctuation caused by spontaneous emission but there also exists an additional contribution from the amplitude fluctuation since it is coupled into the phase fluctuation through the carrier density fluctuation.

α takes a value of 2 to 6 depending on the active-region material, the injected carrier density, and the lasing photon energy [24], while in most gas and solid-state lasers, α can be virtually taken to be zero. This is because the gain spectrum in semiconductor lasers based on a band-to-band transition is asymmetric with respect to the gain-peak frequency (corresponding to the laser frequency) and, as a consequence, the associated refractive index in the vicinity of the lasing frequency varies appreciably with the injected carrier density. In contrast, in ordinary lasers where the lasing transition takes place between two discrete levels, the gain spectrum is symmetric and the associated refractive-index dispersion crosses zero at the lasing frequency. The nonzero value of this parameter affects a number of semiconductor laser characteristics including linewidth broadening, lateral-mode instability, and frequency chirping.

4.4 Dynamic Characteristics

The capability of direct-current modulation is one of the important advantages of semiconductor lasers. The laser response to a stepwise current pulse is schematically shown in Fig. 10. There is a delay time τ_d for the turn-on of lasing because a finite time is required before the injection carriers are accumulated to form population inversion. τ_d is approximated by [25]

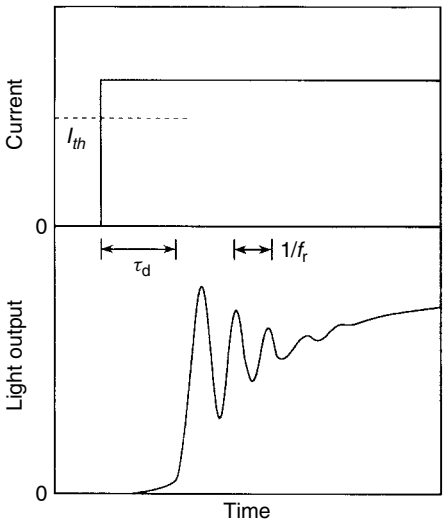

Fig. 10 Laser response to a staircase injection current. τ_d: Turn-on delay time. f_r: Relaxation oscillation frequency

$$\tau_d = \tau_s \sqrt{\frac{I_{th}}{I}} \tanh^{-1} \sqrt{\frac{I_{th}}{I}}, \qquad (22)$$

where τ_s is the carrier lifetime at threshold ($\sim$2 ns), I_{th} is the threshold current, and I is the current pulse height. τ_d can be reduced by prebiasing the laser. Then τ_d is expressed by the DC bias current I_0 ($< I_{th}$) as

$$\tau_d = \tau_s \sqrt{\frac{I_{th}}{I}} \left(\tanh^{-1} \sqrt{\frac{I_{th}}{I}} - \tanh^{-1} \sqrt{\frac{I_0}{I}} \right). \qquad (23)$$

After the onset of laser oscillation, relaxation oscillations are generated for a

time interval of several nanoseconds before the steady state is attained. The relaxation oscillation occurs because of the resonant interaction between photons and carriers; the energy stored in the cavity is transferred back and forth between the photon and carrier subsystems. Therefore, the periodic oscillation of the laser output is accompanied by a modulation of carrier density. The temporal variation of gain spectrum caused by the carrier density modulation leads to a multiple–longitudinal-mode oscillation. Furthermore, the frequency of each individual mode is modulated because of the carrier-density–dependent refractive index, leading to a broadening of the time-averaged spectral linewidth. This phenomenon is referred to as *frequency chirping*. The magnitude of the chirping is governed by the linewidth enhancement factor α.

When a laser is biased above threshold, the bandwidth of the dynamic response of the laser is essentially determined by the relaxation oscillation frequency f_r. Let $\Delta I(\omega)e^{i\omega t}$ be a small-amplitude sinusoidal current superimposed on a DC bias I_0 ($>I_{th}$). Then the photon density inside the active region is made up of a DC component S_0 and an AC component $\Delta S(\psi)e^{i\omega t}$. The magnitude of the modulated component $\Delta S(\omega)$ is analyzed to be [26]

$$\Delta S(\omega) = \frac{\Gamma \tau_p \omega_r^2}{-\omega^2 + i\omega\Omega + \omega_r^2} \frac{\Delta I(\omega)}{eV_a}. \quad (24)$$

Here, the angular relaxation oscillation frequency ω_r is given by

$$\omega_r = 2\pi f_r = \sqrt{\frac{S_0}{\tau_p}\frac{c}{\eta_g}\frac{\partial g}{\partial n}}. \quad (25)$$

Ω represents the damping of the relaxation oscillation and is given by

$$\Omega = \frac{1}{\tau_s} + S_0 \frac{c}{n_g}\frac{\partial g}{\partial n}. \quad (26)$$

τ_p is the photon lifetime of the cavity ($\sim$2 ps) expressed as

$$\tau_p^{-1} = \frac{c}{\eta_g}\left[\alpha_i + \frac{1}{L}\ln\left(\frac{1}{R}\right)\right]. \quad (27)$$

V_a denotes the active-region volume. The magnitude of the modulated component of output power $\Delta P(\omega)$ is proportional to $\Delta S(\omega)$; that is,

$$\Delta P(\omega) = \frac{1}{2}\frac{c}{\eta_g}\hbar\omega_l \frac{1}{L}\ln\left(\frac{1}{R}\right)\Delta S(\omega)\frac{V_a}{\Gamma}. \quad (28)$$

Figure 11 shows an example of relative modulation response $|\Delta P(\omega)/\Delta P(0)| = \omega_r^2/[(\omega^2 - \omega_r^2)^2 - \omega^2\Omega^2]^{1/2}$. A flat response at modulation frequencies less than f_r is followed by a peak at f_r and a sharp drop at modulation frequencies exceeding f_r. Therefore f_r is the uppermost useful modulation frequency. f_r is proportional to the square root of output power

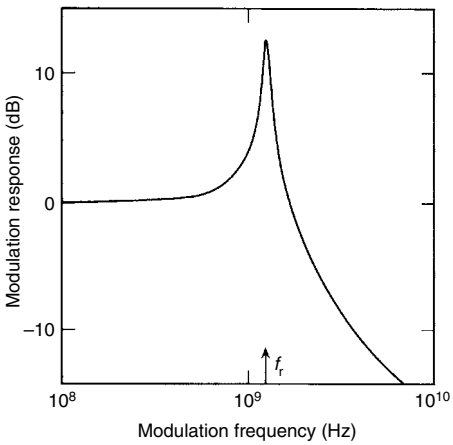

Fig. 11 Calculated small-signal modulation response. f_r is chosen to be 1.3 GHz

P and typically takes a value of 3 to 5 GHz at $P = 10$ mW. In practice, however, the modulation bandwidth may be limited by the electrical parasitics associated with a specific device structure that leads to a rolloff in the modulation response. The cutoff frequency due to the parasitic rolloff varies 3 to 30 GHz depending on the device structure.

5
New Structures and Functions

5.1
Quantum-well Lasers

Quantum-well (QW) lasers have an active region composed of ultrathin layers (~20 nm thick or less) forming narrow potential wells for injected carriers. Illustrated in Fig. 12 are the two representative examples of the active-region structures in AlGaAs QW lasers. Lasers comprising multiple wells (Fig. 12a) are termed *multiple-quantum-well* (MQW) lasers [27, 28]. Shown in Fig. 12(b) is the active region of a GRIN (graded index)–SCH (separate confinement heterostructure)–SQW (single-quantum-well) laser [27, 28]. In the latter structure, the injected carriers are confined in the SQW, while the laser light is guided by the GRIN waveguide, thus ensuring an effective interaction between carriers and light without resort to multiple-well structure.

The quantization of electronic states in the well gives rise to a variety of superior laser characteristics over the conventional lasers with bulk active region. A major benefit is the reduction of the threshold current. Figure 13 helps us understand the low-threshold characteristics of QW lasers, where the energy distributions of the density of states and the injected carriers are compared between a bulk material and a QW structure. As the parabolic density of states in the bulk material changes into the staircase density of states in the QW structure, the energy distribution of the injected carriers narrows. The narrower energy distribution of the injected carriers leads to a narrower gain spectrum, with a higher peak gain value for a given carrier

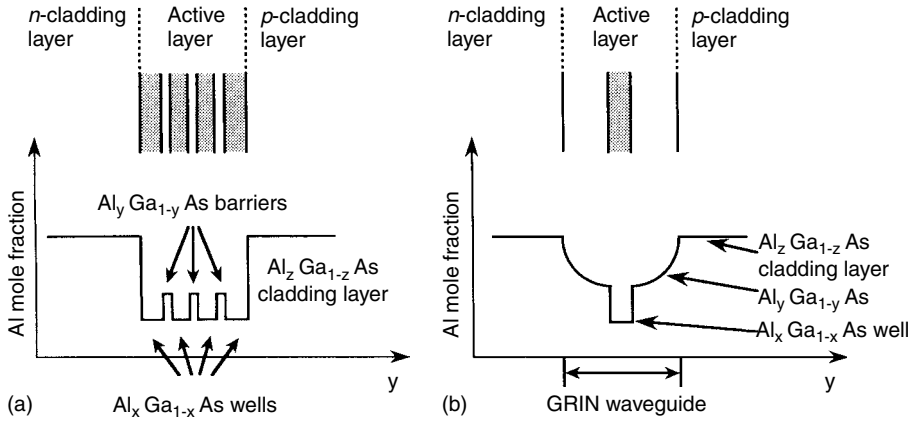

Fig. 12 Two representative examples of active-region structures in AlGaAs quantum-well (QW) lasers (a) Multiple-quantum well (MQW) laser; (b) Graded index (GRIN)–separate confinement heterostructure (SCH)–single-quantum-well (SQW) laser

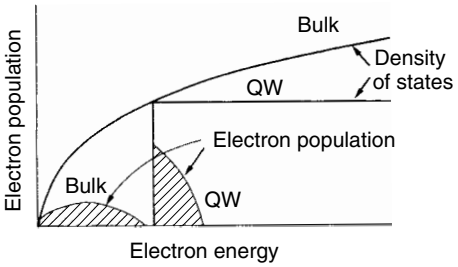

Fig. 13 Comparison of energy distribution of injected carriers between bulk and quantum-well active regions

density. Because of the staircase density of states, however, optical gain in QW lasers saturates for higher injection levels. Since the saturated gain value depends on the number of QW layers, SQW is preferable for high-Q cavities while MQW structure should be employed for lossy cavities. A threshold current as low as 0.55 mA has been demonstrated in a high-reflectivity coated ($R \sim 0.8$) GRIN-SCH-SQW laser [29]. The narrowing of the gain spectrum in QW lasers as compared to bulk lasers is accompanied by an enhancement of $\partial g/\partial n$ (nearly doubled) and a reduction of α (see Eq. (21)). This leads to an enhancement of modulation bandwidths (see Eq. (25)) and a reduction of linewidths (see Eqs. (20) and (21)) under both DC and pulsed excitation [30].

The DH structure is generally formed by successive epitaxial growth of materials lattice-matched to the substrate. The lattice-matching condition imposes severe limitations on the combination of DH structure materials. However, the limitation is alleviated to a certain extent if the grown layer is thin enough to be elastically strained without any generation of dislocations. Applying this principle to the active region, we can obtain lasers whose emission wavelengths cannot be realized by lattice-matched systems. In QW lasers comprising a strained $Ga_xIn_{1-x}As$ well and AlGaAs cladding layers grown on a GaAs substrate, laser emission at 0.9 to 1.1μm can be realized by varying the composition x and the well width. This wavelength region corresponds to the gap between AlGaAs and GaInPAs laser wavelengths and is fit for such applications as the excitation of Er-doped fibers and blue–green light generation by the second-harmonic generation technique.

Furthermore, it is predicted theoretically that the strain is beneficial in obtaining such good laser properties as low-threshold currents, high-relaxation oscillation frequencies, and low linewidth-enhancement factors [31, 32]. This is related to the strain-induced alleviation of the asymmetry in the density of states between the conduction and valence bands. The strain effects on high-speed modulation characteristics in GaInPAs lasers are now being studied experimentally.

5.2 Distributed-feedback and Distributed Bragg-reflector Lasers

In place of mirror facets in Fabry–Pérot cavities, periodic gratings incorporated within laser waveguides can be utilized as a means of optical feedback. Integrated optical feedback from the periodic grating provides strong wavelength selectivity. Devices incorporating the grating in the pumped region are termed *distributed-feedback* (DFB) lasers (Fig. 14a), while those incorporating the grating in the passive region are termed *distributed Bragg-reflector* (DBR) lasers (Fig. 14b) [33]. By virtue of the strong wavelength selectivity, DFB and DBR lasers oscillate in a single-longitudinal mode even under high-speed modulation, in contrast to Fabry–Pérot lasers, which

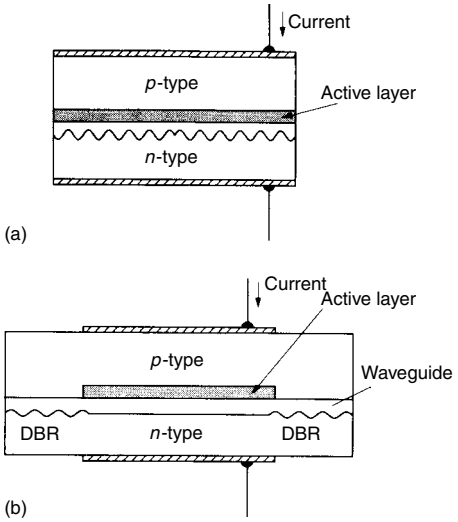

Fig. 14 (a) Distributed-feedback (DFB); (b) distributed Bragg-reflector (DBR) lasers

exhibit multiple–longitudinal-mode oscillation when pulsed rapidly. This is an advantageous feature for optical data transmission using fibers because the spectral width determines the maximum bit rate transmitted in the presence of fiber-chromatic dispersion.

The carrier-density dependence of the refractive index can be exploited to provide wavelength tunability to the periodic gratings [34]. Illustrated in Fig. 15 is a tunable laser utilizing this principle, which can be tuned continuously over a wavelength range of several nanometers. Here, the Bragg wavelength of the DBR structure, at which the reflection loss is minimized, is tuned by the injection current I_3. The current I_2 is adjusted so that one of the resonant wavelengths of the cavity is brought into coincidence to the Bragg wavelength. Thus, by choosing appropriate pairs of I_2 and I_3, the oscillation wavelength can be varied continuously without mode jumping. Tunable lasers are expected to be used in future wavelength-division multiplexing communications systems.

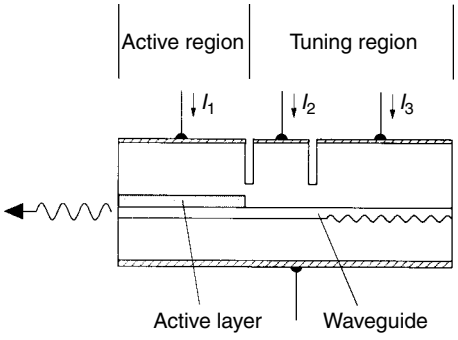

Fig. 15 Wavelength-tunable distributed Bragg-reflector (DBR) laser

5.3
Semiconductor-laser Arrays

The major drawbacks of semiconductor lasers, the relatively low output powers, and the conspicuous beam divergence can be alleviated by integrating multiple laser stripes into an array as shown in Fig. 16. The stripes are closely spaced so that the radiation from neighboring stripes is coupled to form coherent modes of the entire array. The array modes, often referred to as supermodes, are phase-locked combinations of the individual stripe modes and are characterized by the phase relationship between the optical fields supported by adjacent stripes. If an array is properly designed so that

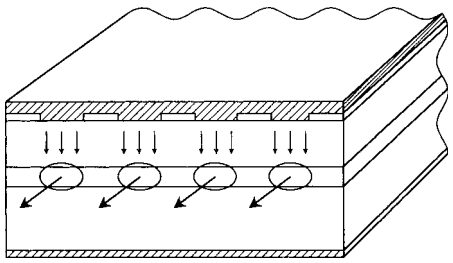

Fig. 16 Semiconductor-laser array

one particular supermode, which is a uniphase superposition of the individual stripe modes (so called 0° shift mode), is excited, a single-lobed, low-divergence beam is attained; the divergence angle $\theta_{\parallel}$ is approximated by the relation $\theta_{\parallel} \sim \lambda/Ns$, where s is the center-to-center separation of adjacent stripes and N is the number of stripes.

A serious problem in the performance of laser arrays is their liability to multisupermode oscillation; the difference among the modal gains of the individual supermodes is so small that the lateral-mode behavior of an array is susceptible to spatial hole burning and temperature distribution in the lateral direction induced at high-excitation levels. The multisupermode oscillation is generally accompanied by a multilobed output beam, which greatly reduces the utility of arrays for most potential applications. The maximum output power for the single-lobed beam operation is typically 300 to 500 mW.

5.4 Surface-emitting Lasers

Conventional semiconductor lasers utilizing cleaved end facets as a means of optical feedback are not fit for monolithic integration. For monolithic integration, it is desirable to have laser output normal to the wafer surface. Illustrated in Fig. 17 is the basic configuration of a surface-emitting laser incorporating two mirrors parallel to the surface to form a vertical cavity typically 5 to 10 μm long [35]. It is crucially important to increase the reflectivity of mirrors, since the reflection loss otherwise becomes very high because of the very short gain region. Bragg reflectors composed of multilayered semiconductors or dielectrics can be exploited to provide reflectivities exceeding 95%. Carrier injection is performed through a pair of electrodes in the upper and lower surfaces to excite a cylindrical active region typically of 10-μm diameter, 1 to 3 μm thick.

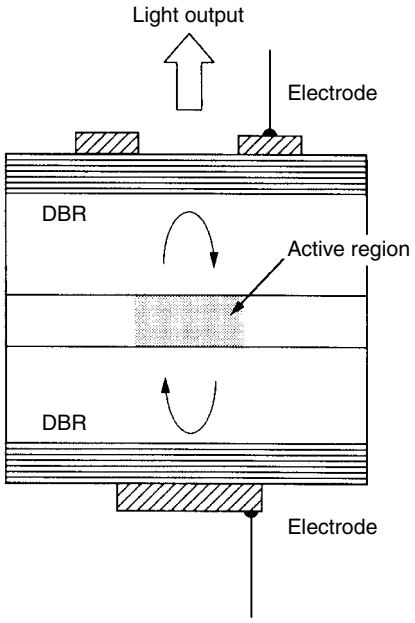

Fig. 17 Vertical-cavity Surface-emitting Laser

The surface-emitting laser has several advantages over the conventional edge-emitting laser, such as the capability of being integrated on a wafer to form a two-dimensional densely packed laser array, the feasibility of stable single–longitudinal-mode operation because of the large mode spacing due to the short cavity length, and the attainability of a narrow circular output beam.

6 Summary

We have reviewed the history and the state of the art of semiconductor lasers. We note

here that their successful development is the result of contributions of scientists and engineers from a number of fields, including semiconductor physics, quantum electronics, crystal growth, processing technologies, and system applications, as is most notably represented by the application to the fiber-optic communications systems.

Ongoing research and development activities indicate that the same will be true in the future. There is growing interest in improving laser characteristics by miniaturizing device structures. The natural extension of the QW lasers will be quantum wire and box lasers, where we can take advantage of the two- and three-dimensional quantization of electronic systems. Researchers in quantum optics are interested in the physics of microcavity semiconductor lasers, which have cavity lengths comparable to light wavelength. Such a microcavity will drastically alter the nature of spontaneous emission and can lead to ultralow-threshold ($\sim$1 μA or less), low-noise, and high-speed lasers. In order to convert the microcavity lasers as well as the quantum wire and box lasers from a laboratory curiosity to a practical light source, further progress in growth and processing technology of semiconductor microstructures is indispensable. Integration of these superior devices will find new application fields such as optical interconnection in large-scale-integration circuits and two-dimensional information processing.

Glossary

Active Layer: A layer, typically $\sim$0.1 μm thick in which stimulated emission occurs.

Cladding Layers: Layers that sandwich an active layer made of materials that have wider band gaps than the active layer, p and n doped to facilitate carrier injection into the active layer.

Distributed Bragg Reflector Laser: A laser incorporating a periodic grating at both ends of the cavity as a means of optical feedback.

Distributed-feedback Laser: A laser incorporating a periodic grating in the vicinity of the active layer as a means of optical feedback.

Double Heterojunction Structure: An active-region structure in which an active layer of one material is sandwiched between two cladding layers of another material.

Gain-guided Laser: A laser in which the optical field distribution in the transverse direction is determined by the gain profile produced by carrier density distribution.

Index-guided Laser: A laser in which the optical field distribution in the transverse direction is determined by a built-in refractive-index profile.

Linewidth-enhancement Factor: The ratio between the carrier-induced variations of the real and imaginary parts of susceptibility $\chi(n)$, i.e.,

$$\alpha = \frac{\partial[\mathrm{Re}\chi(n)]/\partial n}{\partial[\mathrm{Im}\chi(n)]/\partial n}, \qquad (29)$$

where n is the carrier density. An equivalent expression is given in Eq. (21).

Phase-locked Laser Array: An array of stripe lasers spaced closely so that a

phase-locked oscillation among the laser stripes is attained.

Quantum-well Laser: A laser that has ultrathin active layers forming narrow potential wells for injected carriers and giving rise to a size quantization of the electronic states.

Stripe-geometry Laser: A laser in which the current is injected only within a narrow region beneath a stripe contact several μm wide.

Surface-emitting Laser: A laser that emits light normal to the wafer surface.

References

[1] Hall, R. N., Fenner, G. E., Kingsley, J. D., Soltys, T. J., Carlson, R. O. (**1962**), *Phys. Rev. Lett.* **9**, 366–368.

[2] Nathan, M. I., Dumke, W. P., Burns, G., Dill Jr, F. H., Lasher, G. (**1962**), *Appl. Phys. Lett.* **1**, 62–64.

[3] Quist, T. M., Rediker, R. H., Keyes, R. J., Krag, W. E., Lax, B., McWhorter, A. L., Zeigler, H. J. (**1962**), *Appl. Phys. Lett.* **1**, 91–92.

[4] Alferov, Zh. I., Andreev, V. M., Garbuzov, D. Z., Zhilyaev, Y. V., Morozov, E. P., Portnoi, E. L., Trofim, V. G. (**1970**), *Fiz. Tekh. Poluprovodn.* **4**, 1826–1829. [English Translation: (**1971**), *Sov. Phys. Semicond.* **4**, 1573–1575.]

[5] Hayashi, I., Panish, M. B., Foy, P. W., Sumski, S. (**1970**), *Appl. Phys. Lett.* **17**, 109–111.

[6] Casey Jr, H. C., Panish, M. B. (1978), *Heterostructure Lasers*. Parts A and B, New York: Academic Press.

[7] Lang, R. (**1979**), *IEEE J. Quantum Electron.* **QE-15**, 718–726.

[8] Kogelnik, H., Shank, C. V. (**1971**), *Appl. Phys. Lett.* **18**, 152–154.

[9] Arakawa, Y., Yariv, A. (**1986**), *IEEE J. Quantum Electron* **QE-22**, 1887–1899.

[10] Haase, M. A., Qiu, J., Depuydt, J. M., Cheng, H. (**1991**), *Appl. Phys. Lett.* **59**, 1272–1274.

[11] Nakamura, S., Senoh, M., Nagahama, S., Iwasa, N., Yamada, T., Matsusita, T., Kiyoku, H., Sugimoto, Y. (**1996**), *Jpn. J. Appl. Phys.* **35**, L74–L76.

[12] Akasaki, I., Sota, S., Sakai, H., Tanaka, T., Koike, M., Amano, H. (**1996**), *Electron. Lett.* **32**, 1105–1106.

[13] Kroemer, H. (**1986**), *Surf. Sci.* **174**, 299–306.

[14] Botez, D. (**1978**), *IEEE J. Quantum Electron.* **QE-14**, 230–232.

[15] Yamamoto, Y., Saita, S., Mukai, T. (**1983**), *IEEE J. Quantum Electron.* **QE-19**, 47–58.

[16] Yamada, M. (**1983**), *IEEE J. Quantum Electron.* **QE-19**, 1365–1380.

[17] Ogasawara, N., Ito, R. (**1988**), *Jpn. J. Appl. Phys.* **27**, 607–626.

[18] Dyment, J. C. (**1967**), *Appl. Phys. Lett.* **10**, 84–86.

[19] Tsukada, T. (**1974**), *J. Appl. Phys.* **45**, 4899–4906.

[20] Horikoshi, Y. (1985), in W. T. Tsang (Ed.), *Semiconductors and Semimetals*, Vol. 22. Part C, Orlando, FL: Academic Press, Chap. 3.

[21] Ohkawa, K., Karasawa, T., Mitsuya, T. (**1991**), *Jpn. J. Appl. Phys.* **30**, L152–L155.

[22] Ikegami, T. (**1972**), *IEEE J. Quantum Electron.* **QE-8**, 470–476.

[23] Henry, C. H. (**1982**), *IEEE J. Quantum Electron.* **QE-18**, 259–264.

[24] Osinski, M., Buus, J. (**1987**), *IEEE J. Quantum Electron.* **QE-23**, 9–29.

[25] Chinone, N., Ito, R., Nakada, O. (**1974**), *IEEE J. Quantum Electron.* **QE-10**, 81–84.

[26] Yariv, A. (1989), *Quantum Electronics*. New York: Wiley, Chap. 11.

[27] Tsang, W. T. (**1981a**), *Appl. Phys. Lett.* **39**, 786–788.

[28] Tsang, W. T. (**1981b**), *Appl. Phys. Lett.* **39**, 134–137.

[29] Lau, K. Y., Derry, P. L., Yariv, A. (**1988**), *Appl. Phys. Lett.* **52**, 88–90.

[30] Uomi, K., Sasaki, S., Tsuchiya, T., Okai, M., Aoki, M., Chinone, N. (**1990**), *Electron. Lett.* **26**, 52–53.

[31] Suemune, I., Coldren, L. A., Yamanishi, M., Kan, Y. (**1988**), *Appl. Phys. Lett.* **53**, 1378–1380.

[32] Ohtoshi, T., Chinone, N. (**1989**), *IEEE Photon. Technol. Lett.* **1**, 117–119.

[33] Mroziewicz, B., Bugajski, M., Nakwaski, W. (1991), *Physics of Semiconductor Lasers*. Amsterdam: North Holland, Chap. 6.

[34] Yoshikuni, Y. (1991), in Y. Yamamoto (Ed.), *Coherence, Amplification, and Quantum Effects in Semiconductor Lasers*. New York: Wiley, Chap. 4.

[35] Iga, K., Koyama, F., Kinoshita, S. (**1988**), *IEEE J. Quantum Electron.* **QE-24**, 1845–1855.

Further Reading

Agrawal, G. P., Dutta, N. K. (1986), *Long-Wavelength Semiconductor Lasers*. New York: Van Nostrand Reinhold.

Casey Jr, H. C., Panish, M. B. (1978), *Heterostructure Lasers*, Parts A and B, New York: Academic Press.

Kressel, H., Butler, J. K. (1977), *Semiconductor Lasers and Heterojunction LEDs*. New York: Academic Press.

Mroziewicz, B., Bugajski, M., Nakwaski, W. (1991), *Physics of Semiconductor Lasers*. Amsterdam: North Holland.

Nakamura, S., Chichibu, S. F. (2000), *Nitride Semiconductor Blue Lasers and Light Emitting Diodes*. London: Taylor & Francis.

Nakamura, S., Fasol, G., Pearton, S. J. (2000), *The Blue Laser Diode: The Complete Story*. New York: Springer.

Peterman, K. (1988), *Laser Diode Modulation and Noise*. Dordrecht: Kluwer Academic Publishers.

Thompson, G. H. B. (1980), *Physics of Semiconductor Laser Devices*. Chichester: Wiley.

Yamamoto, Y. (Ed.) (1991), *Coherence, Amplification, and Quantum Effects in Semiconductor Lasers*. New York: Wiley.