from Exoplanets, S. Seager, ed. (Tucson: Univ. Arizona Press; 2010)

Direct Imaging of Exoplanets

Wesley A. Traub
Jet Propulsion Laboratory, California Institute of Technology

Ben R. Oppenheimer
American Museum of Natural History

A direct image of an exoplanet system is a snapshot of the planets and disk around a central star. We can estimate the orbit of a planet from a time series of images, and we can estimate the size, temperature, clouds, atmospheric gases, surface properties, rotation rate, and likelihood of life on a planet from its photometry, colors, and spectra in the visible and infrared. The exoplanets around stars in the solar neighborhood are expected to be bright enough for us to characterize them with direct imaging; however, they are much fainter than their parent star, and separated by very small angles, so conventional imaging techniques are totally inadequate, and new methods are needed. A direct-imaging instrument for exoplanets must (1) suppress the bright star's image and diffraction pattern, and (2) suppress the star's scattered light from imperfections in the telescope. This chapter shows how exoplanets can be imaged by controlling diffraction with a coronagraph or interferometer, and controlling scattered light with deformable mirrors.

1. INTRODUCTION

The first direct images of exoplanets were published in 2008, fully 12 years after exoplanets were discovered, and after more than 300 of them had been measured indirectly by radial velocity, transit, and microlensing techniques. This huge time lag occurred because direct imaging of exoplanets requires extraordinary efforts in order to overcome the barriers imposed by astrophysics (planet-star contrast), physics (diffraction), and engineering (scattering).

The structure of this chapter is as follows. Section 1 discusses the scientific purpose of direct imaging of exoplanets, and includes a glossary of terms. Section 2 discusses basic physical concepts, including brightness, contrast, wavefronts, diffraction, and photons. Section 3 discusses coronagraph and interferometer concepts. Section 4 addresses speckles and adaptive optics. Section 5 sketches recent results from exoplanet imaging and lists current projects. Section 6 outlines future prospects for exoplanet imaging on the ground and in space.

1.1. Exoplanet Images

We illustrate with three examples of direct imaging; as it happens, all three examples are of young, self-luminous objects. Figure 1 shows the dust ring and exoplanet Fomalhaut b by *Kalas et al.* (2008), in the visible. The central star was suppressed using a combination of methods described in this chapter: the rectangular-mask coronagraph (section 3.13) and angular differential imaging (section 4.12). Kalas et al. used the Hubble Space Telescope (HST) Advanced Camera for Surveys (ACS) in coronagraph mode. The planet is about 23 mag fainter than its star, and separated by 12.7 arcsec (98 AU at 7.7 pc distance). It was detected at two epochs, clearly showing common motion as well as orbital motion (see inset).

Figure 2 shows a near-infrared composite image of exoplanets HR 8799 b,c,d by *Marois et al.* (2008). The planets are at angular separations of 1.7, 1.0, and 0.6 arcsec from the star (68, 38, and 24 AU at 40 pc distance). The H-band planet-star contrasts (ratio of planet to star flux) are about 10^{-5}, i.e., roughly 12 mag fainter. The planets would be much

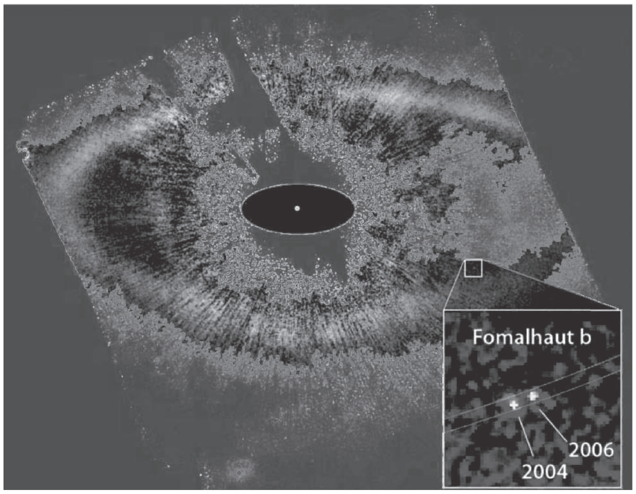


Fig. 1. Visible-wavelength image, from the Hubble Space Telescope, of the exoplanet Fomalhaut b. The planet is located just inside a large dust ring that surrounds the central star. Fomalhaut has been blocked and subtracted to the maximum degree possible.

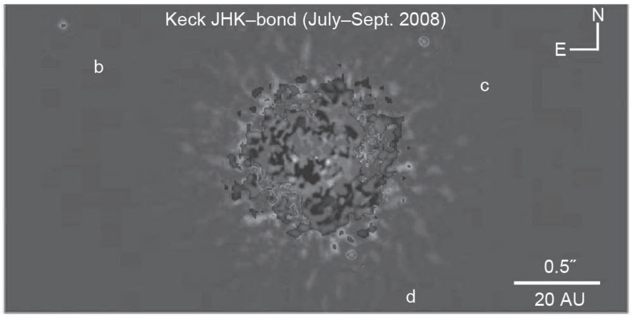


Fig. 2. Near-infrared image, from the Keck telescope, of the exoplanets HR 8799b, c, and d. The central star has been suppressed with angular differential imaging, coupled with adaptive optics. The splatter of dots in the center of this image is simply the small amount of leftover light from the central star that could not be subtracted by ADI, so it is an artifact.

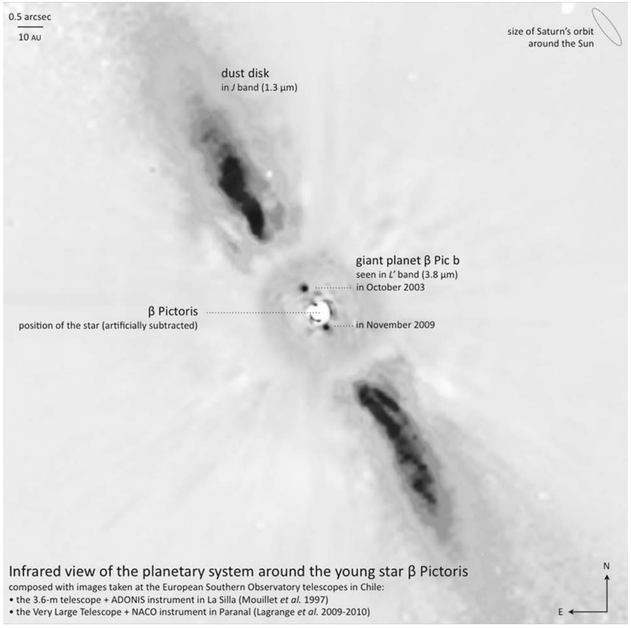
fainter were it not for their youth and consequent internal heat sources, putting their effective temperatures in the 1000 K regime (section 1.3). Marois et al. used the groundbased Gemini and Keck telescopes for these observations. Their techniques included minimizing diffraction using ADI, and minimizing atmospheric speckles using adaptive optics (section 4). If this system were instead the Sun and solar system, then Jupiter would be buried in the inner one-fourth of the speckle field, Earth would be in the inner one-twentieth radius, and both would be 4 to 5 orders of magnitude fainter than the speckles.

Figure 3 shows a near- and mid-infrared composite image of the β Pictoris system, from *Lagrange et al.* (2010). The star itself has been subtracted using a reference star image, and independently with ADI. The composite shows the edge-on dust disk plus the planet β Pic b at two epochs, 2003 (left and above) and 2009 (right and below). The proper motion of β Pic is north, not northwest, so the planet is co-moving, and not a background object. The planet age is ~12 m.y., much younger than the HR 8799 or Fomalhaut planets. The mass is ~9 M_{Jup}, and the semimajor axis in the 8–15 AU range, with a period as short as 17 years. The star is at 19 pc, and the star-planet separations shown are in the 0.3–0.5 arcsec range.

These images set the stage for our goal in this chapter, direct imaging of planets from Earth- to Jupiter-sized around nearby stars.

1.2. Exoplanet Spectra

The spectrum of an exoplanet tells us about its composition, clouds, thermal structure, and variability, as discussed in the chapter by Burrows and Orton for giant planets, and the chapter by Meadows and Seager for terrestrial planets. A direct image of an exoplanet permits us to obtain a spectrum, using a conventional spectrometer or an integral field spectrometer. The resolution will be low, because the planet is faint; however, since many molecular bands are intrinsically low-resolution features, we can still learn much about an atmosphere.



Infrared view of the planetary system around the young star β Pictoris
composed with images taken at the European Southern Observatory telescopes in Chile:
- the 3.6-m telescope + ADONIS instrument in La Silla (Mouillet *et al.* 1997)
- the Very Large Telescope + NACO instrument in Paranal (Lagrange *et al.* 2009-2010)

Fig. 3. Near- and mid-infrared composite image of β Pic b and the β Pic dust disk, from the ESO 3.6-m and Very Large Telescopes, with the star subtracted. The planet is shown at two epochs, 2003 and 2009, demonstrating co-moving position with the star as well as orbital motion.

1.3. Hot, Young, and Mature Exoplanets

Hot Jupiters have not yet been directly imaged, but their large thermal flux, 10^{-3} to 10^{-4} times the parent star, means that they will likely be imaged in the future. Their extreme closeness to the parent star requires extreme angular resolution, so the images will come from long-baseline interferometers, not from single-dish telescopes.

Young, self-luminous planets were the first to be directly imaged, because their high temperature and large size give them a strong, detectable flux, and their large distances from their parent stars makes them easier to see in the halo of atmospherically or instrumentally scattered star light. These young, self-luminous planets are likely to continue to be prime targets for detection in the near future, owing to this combination of favorable parameters. However, young planets cool off in a few tens of millions of years, so they will be found only around young stars, and not around nearby (older) stars.

A *mature exoplanet* may be defined here as one with an effective temperature that is roughly comparable to its star-planet equilibrium temperature (section 2.5). These planets, like those in the solar system or around mature nearby stars, will be fainter in the infrared than young, self-luminous planets, and therefore will require more sophisticated techniques to image them. In addition, most of them probably will be closer to their stars than the ones in HR 8799 and Fomalhaut, and will therefore potentially be detectable by single-dish telescopes, but will require the full power of the techniques in this chapter.

1.4. Radial Velocity, Transits, Lensing, and Astrometry

Currently, the techniques of radial velocity (RV), transits, and gravitational lensing are more productive than direct imaging. Remarkably, these techniques, including astrometry and direct imaging, perform nearly independent roles for exoplanet science, so each of them is valuable. Radial velocity has been very successful in measuring masses and periods of planets with masses greater than several Earths and in short-period orbits. Transits have been valuable in measuring the diameters and periods of giant planets, and in combined-light mode have measured temperature distributions, spectral features, and thermal inversions in two gas giant planets. Transits will also be valuable for determining mass and orbit statistics of distant planets, but its geometric bias precludes using it for the vast majority of nearby systems.

Ultimately, exoplanet science will require direct images and spectra of exoplanet systems. For this information, planets around nearby stars will be essential, because these systems will have larger apparent sizes and photon fluxes than more distant systems, and will therefore be relatively accessible to the techniques in this chapter. A combination of astrometry and imaging will provide the mass, period, orbit, and spectroscopic characterization for these planets, down to and including Earth-mass ones.

1.5. Solar System and Exoplanet Systems

There is a strong connection between solar system and exoplanet science. Until exoplanets were discovered in the early 1990s, it was widely thought that exoplanet systems would resemble our solar system. But with the discovery of hot Jupiters, it is now clear that our system is but one of many possible types. There are several points of comparison. Planetary migration and chaotic episodes are now thought to be common to all systems. Self-luminous planets are also common. Dust and debris structures are common as well. Our picture of the evolution of the solar system is strongly influenced by what we are learning about exoplanet systems. In particular, our picture of the evolution of Earth, and of life itself, may well depend on what we learn about habitable-zone terrestrial exoplanets. And for these planets especially, because of their faintness and small angular separations, the techniques in this chapter will be very important.

1.6. Glossary

Some of the terms used in this chapter are briefly defined here for reference:

Visible: wavelength range ~0.3–1.0 μm
Near-infrared: ~1.0–2.5 μm
Mid-infrared: ~2.5–10 μm
Far-infrared: ~10–200 μm
Photometry: broadband (~20%) flux measurement
Color: ratio of two broadband fluxes
Spectrum: narrowband (≲1%) flux measurement
Self-luminous planet: $T_{eff} \gg T_{equil}$
Mature planet: $T_{eff} \approx T_{equil}$
Terrestrial planet: $0.5\ M_\oplus \lesssim M_p \lesssim 10\ M_\oplus$
Gas giant planet: $10\ M_\oplus \lesssim M_p \lesssim 13\ M_{Jup}$
Habitable zone (HZ): liquid water possible on surface, $0.7\ AU \lesssim a/L_s^{1/2} \lesssim 1.5\ AU$
Wavefront: surface of constant phase of a photon
Ray: direction of propagation of photon, always perpendicular to wavefront
Diffraction: bending of wavefront around an obstacle
Scattering: diffraction from polishing or reflectivity errors, a source of speckles
Speckle: light pattern in image plane (coherent with star) from optical path differences in the beam
Coronagraph: telescope with internal amplitude and/or phase masks for imaging faint sources near a bright one
Occulter: coronagraph but with external mask.
Interferometer: two or more telescopes with coherently combined output
Nuller: coronagraph or interferometer using interference of wavefront to suppress a point source

2. FLUX AND PHOTON CONCEPTS

In this section we discuss the underlying equations and concepts needed to calculate flux and photon levels from exoplanets as well as nearby stars and zodi disks. We also discuss the semi-mysterious nature of photons, which are best thought of as waves in some contexts, but must be considered as particles in others; we try to remove the mystery.

2.1. Star Intensity

If we approximate a star as a blackbody of effective temperature T, then its *specific intensity* is the Planck function $B_\nu(T)$ where

$$B_\nu(T) = \frac{2h\nu^3}{c^2\left(e^{h\nu/kT} - 1\right)} \tag{1}$$

with units of erg/(s cm^2 Hz sr), and where, for a star or planet, the unit of area (cm^2) is in the plane of the sky, i.e., perpendicular to the line of sight, but not necessarily in the plane of the surface of the object.

It is sometimes convenient to use wavelength units instead of frequency units. From $B_\lambda d\lambda = B_\nu d\nu$ and $\lambda\nu = c$ we get

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5\left(e^{hc/\lambda kT} - 1\right)} \tag{2}$$

which has units of erg/(s cm^2 cm sr).

For the calculation of signal levels and signal to noise ratios we need to know the corresponding specific intensities in units of photons instead of ergs, i.e., $\dot{n}_\nu$ or $\dot{n}_\lambda$. The energy of a photon is hν, so we get $\dot{n}_\nu = B_\nu/h\nu$, or $\dot{n}_\lambda = B_\lambda\lambda/hc$, which leads to

$$\dot{n}_\nu = \frac{2\nu^2}{c^2\left(e^{h\nu/kT}-1\right)} \tag{3}$$

with units of photons/(s cm^2 Hz sr), and

$$\dot{n}_\lambda = \frac{2c}{\lambda^4\left(e^{hc/\lambda kT}-1\right)} \tag{4}$$

with units of photons/(s cm^2 cm sr).

For numerical calculations it is often convenient to insert numerical values of h, c, and k, and to express wavelengths in units of μm instead of cm, indicated by $\lambda_{\mu m}$, where

$$1 \text{ photon} = h\nu = \frac{1.986\times10^{-12}}{\lambda_{\mu m}}\text{erg} \tag{5}$$

Then the specific intensity $\dot{n}_\lambda(T_s)$ in photons is

$$\dot{n}_\lambda(T) = \frac{6\times10^{26}}{\lambda_{\mu m}^4\left(e^{14388/\lambda_{\mu m}T}-1\right)} \tag{6}$$

which has units of photons/(s cm^2 μm sr).

At a distance d from a star of radius r, such that the star appears to subtend a solid angle $\Omega = \pi(r/d)^2$ steradian (sr), the *photon flux* $\dot{N}_\lambda$ received is

$$\dot{N}_\lambda = \dot{n}_\lambda(T)\Omega \tag{7}$$

with units of photons/(s cm^2 μm), and likewise for $\dot{N}_\nu$.

Note that for light emitted by an object and subsequently collected by a telescope, or simply for light traversing an optical system, and in the absence of light loss by absorption or blockage, the *etendue,* i.e., the product of area and solid angle, is conserved. Thus the light emitted from the (sky plane) area (A_{star}) of a star, into the solid angle (Ω_{tel}) of a distant telescope, is related to the collecting area (A_{tel}) of the telescope and the solid angle (Ω_{star}) of the distant star

$$A_{star}\Omega_{tel} = A_{tel}\Omega_{star} \tag{8}$$

This explains the switching between the area and solid angle of star and telescope in the above equations.

Stellar flux is often expressed as a *radiant flux* $f_\lambda(m)$, which is a function of *apparent magnitude* m in a *standard spectral band*

$$f_\lambda(m) = 10^{a-0.4m} \tag{9}$$

with units of erg/(s cm^2 μm), outside Earth's atmosphere. For each standard spectral band there is an effective central wavelength λ_0, an effective bandwidth $\Delta\lambda$ (approximately the full width at half maximum, FWHM), and a corresponding value of a. The latter can differ by up to ±0.03, depending upon the calibration technique. A compilation of these parameters is given in Table 1.

Another common unit of flux density is the Jansky, where

$$\begin{aligned} 1\ \text{Jy} &= 10^{-26}\ \text{watt}/(\text{m}^2\text{Hz}) \\ &= \frac{3\times10^{-9}}{\lambda_{\mu m}{}^2}\text{erg}/(\text{s cm}^2\mu\text{m}) \\ &= \frac{1.51\times10^3}{\lambda_{\mu m}}\text{photon}/(\text{s cm}^2\mu\text{m}) \end{aligned} \tag{10}$$

Thus a zero-magnitude star has a flux density of about 3750 Jy at V, and 35 Jy at N. The photon densities are 1.03 × 10^7 at V, and 5.03 × 10^3 photons/(s cm^2 μm) at N.

2.2. Angular Separation

Kepler's third law says that a planet with semimajor axis a (AU) and eccentricity e has orbital period P (yr) where

$$P = a^{3/2}/M_s^{1/2} \tag{11}$$

Here M_s is in units of $M_\odot$ and $M_p \ll M_s$. If the distance from star to observer is d(pc), then the maximum angular separation between planet and star is

$$\theta = a(1+e)/d \tag{12}$$

with astronomical units: θ (arcsec), a (AU), and d (pc).

Some examples of angular separations (exoplanet – star) are given in Table 2, along with the required telescope diameters, occulter diameters, and interferometer baselines needed to suppress the star and directly image the exoplanet.

TABLE 1. Standard spectral bands.

Band	λ_0*	$\Delta\lambda$†	a‡
U	0.365	0.068	–4.38
B	0.44	0.098	–4.19
V	0.55	0.089	–4.43
R	0.70	0.22	–4.76
I	0.90	0.24	–5.08
J	1.22	0.26	–5.48
H	1.65	0.29	–5.94
K_s	2.16	0.32	–6.37
L	3.55	0.57	–7.18
M	4.77	0.45	–7.68
N	10.47	5.19	–9.02
Q	20.13	7.8	–10.14

U, B, V, R, and I data is from *Allen* (1991) and *Cox* (2000). J, H, K_s, L, M, N, and Q data is from *Cox* (2000).

*Effective wavelength in μm.

†Effective bandwidth (FWHM) in μm.

‡$\log_{10}$(f), where f has units of erg/(s cm^2 μm), at zero magnitude.

For the spectral types AFGKM, the approximate number of stars out to 10 and 30 pc is also noted.

2.3. Contrast of Planet

The spectrum of a planet is the sum of reflected starlight, thermal emission, and nonthermal features, as illustrated in Fig. 4 for the case of the Earth-Sun system as seen from a distance of 10 pc. The reflected and thermal continuum components are discussed in sections 2.4 and 2.5. Background light from zodiacal dust is discussed in section 2.6. A planet's color is discussed in section 2.7, and its absorption line spectrum in section 2.8. Nonthermal features (e.g., auroras) are expected to be faint, and are ignored here.

For direct imaging it is convenient to compare the brightness of a planet to its star, at any wavelength. The *contrast* C is defined to be the ratio of planet (p) to star (s) brightness, so we have

$$C = \frac{f_\lambda(p)}{f_\lambda(s)} = \frac{\dot{N}_\lambda(p)}{\dot{N}_\lambda(s)} \tag{13}$$

where C is a function of wavelength, the properties of the planet, and the apparent geometry of the planet-star system. Here f(p) is the sum of reflected and thermal fluxes.

The expected visible-wavelength contrast of typical Jupiter-like and Earth-like planets around nearby stars is shown in Fig. 5. We see that giant planets beyond the ice line will have typical contrasts on the order of 10^{-9} at visible wavelengths (see section 2.4), and separations of about 0.5 arcsec. Earth-like planets in the habitable zone will have contrasts of about 10^{-10} and separations of about 0.1 arcsec. As suggested by the limiting-case detection lines for several types of groundbased coronagraphs and the HST, these planets cannot be directly imaged by them. However, they could be imaged by a coronagraph in space, designed for this purpose, as discussed in this chapter.

TABLE 2. Angular separation examples.

Distance	10 pc	30 pc
Angular separation of planet at 1 AU (max)	100 mas	33 mas
Telescope diameter (min) at 0.5 μm	3.1 m	6.2 m
Occulter diameter at 0.5 μm	49 m	16 m
Interferometer baseline at 10 μm	21 m	62 m
Number of AFGKM stars	2, 11, 26, 42, 210	27 times greater

Angular separation θ is from equation (12), for the Earth-Sun system. Telescope diameter D is from $\theta = n\lambda/D$, where n = 3 is intermediate between the theoretical minimum for an internal coronagraph (n = 2) and an experimentally demonstrated value (n = 4). Occulter diameter is $D_O = 2\theta d_O$,where the distance between a telescope and its external occulter is d_O = 50,000 km. Interferometer baseline is $B = \lambda/\theta$. The number of stars is assumed to scale as d^3.

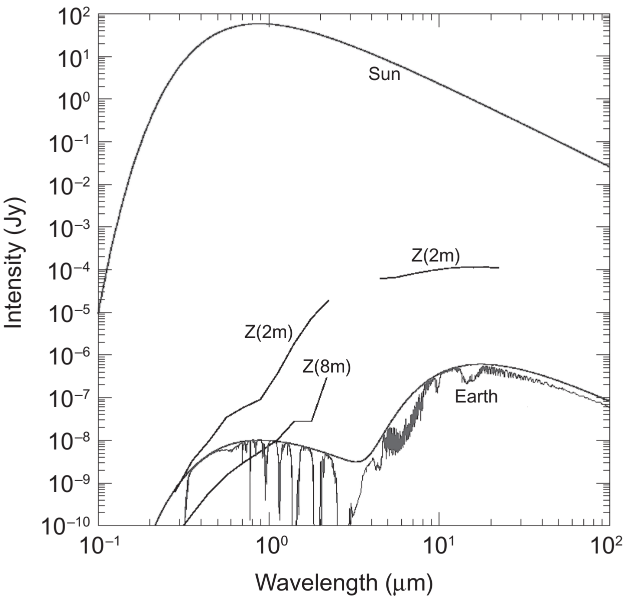


Fig. 4. Schematic spectrum of the Sun and Earth at 10 pc, in the visible and infrared (*Kasting et al.,* 2009). Here Earth at maximum elongation is at 0.1 arcsec (1 AU/10 pc) with a contrast of 10^{-10} in the visible and 10^{-7} in the mid-infrared (~10 μm). The exozodiacal light is sketched for small (2-m) and large (8-m) telescopes in the visible, and an interferometer with 2-m collectors in the infrared.

2.4. Visible Brightness of Planet

Reflected starlight from a planet is often assumed to follow a Lambert law, which states that the light that is incident on a surface, from any direction, is reflected uniformly in all directions, in the sense that the amount of light leaving an element of a surface is proportional to the projected area in the reflected direction. So to an observer, the apparent brightness of any given projected area of the illuminated surface of a planet is proportional to the amount of starlight hitting the surface within that apparent area.

The *phase angle* α of a planet is the planet-centered angle from star to observer. So α = 0 at superior conjunction with the planet behind the star, α = π/2 at quadrature (maximum elongation for a circular orbit), and α = π at inferior conjunction with the planet between the star and observer.

As an example, if the Moon were a Lambert reflector, then the full Moon (α = 0) would appear to be a uniformly bright object, with no limb darkening, but the quarter Moon (α = π/2) would appear to have a bright Sun-facing limb that tapers to zero intensity at the terminator, in proportion to the projected area toward the Sun (i.e., the cosine of the angle between the surface normal and the Sun). In practice, bare-rock bodies like Mars, Earth, and the Moon tend to be more uniformly bright than a Lambert surface, but cloudy planets like Venus and Jupiter tend to be closer to Lambertian.

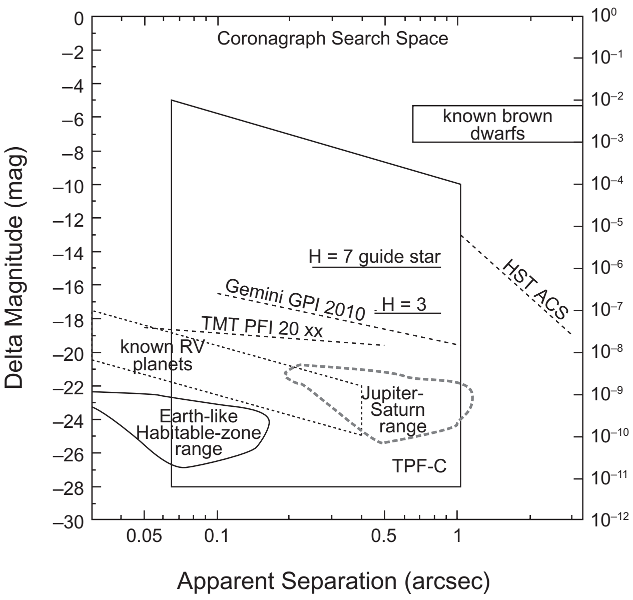


Fig. 5. Contrast vs. separation is shown for several types of companions, along with limits of the TPF-C coronagraph, as once planned. For example, the contrast of Earth and Jupiter twins, for many nearby stars, and for separations from about 20% to maximum elongation are shown.

The *geometric albedo* p of a planet is defined to be the ratio of planet brightness at $\alpha = 0$ to the brightness of a perfectly diffusing disk with the same position and apparent size as the planet. In other words, p is the ratio of the flux reflected toward an observer at zero phase angle to the flux from the star that is incident on the planet. The geometric albedo will in general be wavelength dependent. Numerical values of p_V, for the visible band, are listed in Table 3.

The reflected-light contrast of a planet can be written as

$$C_{vis} = p\phi(\alpha)\left(r_p/a\right)^2 \tag{14}$$

where $\phi(\alpha)$ is the phase law, sometimes called the integral phase function, at phase angle α, r_p is the planet radius, and a is the distance from planet to star, here simply written as the semimajor axis. For a Lambert sphere the phase law is

$$\phi(\alpha) = \left[\sin(\alpha) + (\pi - \alpha)\cos(\alpha)\right]/\pi \tag{15}$$

For example, in an edge-on system $\phi(0) = 1$ at superior conjunction, $\phi(\pi/2) = 1/\pi$ at maximum elongation, and $\phi(\pi) = 0$ at inferior conjunction.

This law is a good approximation for high-albedo planets such as Venus. There are no convenient expressions for other types of planet surfaces, such as rocky, low-albedo ones, but it is empirically observed that such objects tend to have relatively stronger reflection at zero phase angle, probably from a lack of shadowing on clumpy surfaces, compared to other angles. The net result is that the phase function tends to be smaller than the Lambert law, at angles away from zero. However, for lack of a better version, the Lambert law is often used for exoplanets.

For example, the visible contrasts of the Earth/Sun and Jupiter/Sun systems at maximum elongation, assuming that they reflect as Lambert spheres, are

$$C_{vis}(E) \simeq 2.1\times10^{-10} \tag{16}$$

$$C_{vis}(J) \simeq 1.4\times10^{-9} \tag{17}$$

These extreme contrasts, 10^{-10} or Δmag = 25 for Earth, and 10^{-9} or Δmag = 22.5 for Jupiter, are the driving forces behind nearly all the direct imaging discussion in this chapter. These huge brightness ranges, occurring in such close proximity on the sky, mean that scattered light in a telescope system, which is ignored in conventional astronomical imaging, now becomes an important experimental factor in isolating the light of a planet.

2.5. Infrared Brightness of Planet

The *Bond albedo* of a planet, A_{Bond}, is defined to be the ratio of total light reflected to total light incident, where "total" here means bolometric, i.e., integrated over all wavelengths, and the entire planet.

The emittance F of a black-body at *effective temperature* T is the total flow of radiation outward from a unit area of its surface, and is given by

$$F = \int B_\nu d\nu \cos(\vartheta) d\Omega = \int \pi B_\nu d\nu = \sigma T^4 \tag{18}$$

in units of erg/(s cm^2), where ϑ is the angle from the normal to the surface, $d\Omega = \sin(\vartheta)d\vartheta d\varphi$, φ is the azimuth around the

TABLE 3. Albedo and temperature.

Planet	a (AU)	p (visible geom. alb.)	A_{bond} (Bond alb.)	T_{equil}* (K)	T_{eff}† (K)
Mercury	0.387	0.138	0.119	433	433
Venus	0.723	0.84	0.75	231	231
Earth	1.000	0.367	0.306	254	254
Moon	1.000	0.113	0.123	269	269
Mars	1.524	0.15	0.25	210	210
Jupiter	5.203	0.52	0.343	110	124.4
Saturn	9.543	0.47	0.342	81	95.0
Uranus	19.19	0.51	0.290	58	59.1
Neptune	30.07	0.41	0.31	46	59.3

Data adapted from *de Pater and Lissauer* (2001, 2010).

*T_{equil} is calculated from A_{Bond}.
†T_{eff} is set equal to T_{equil} for terrestrial planets, but is measured for gas giants.

normal, and σ is the Stefan-Boltzman constant. For a star, this leads to the *luminosity* L, given by

$$L = 4\pi r_s^2 \sigma T^4 \tag{19}$$

where r_s is the radius of the star and T the effective temperature.

The flux from a star is diluted by a^{-2} by the time it reaches a planet at distance a. Of this, a fraction ($1-A_{Bond}$) is absorbed by the planet. The resulting *radiative equilibrium temperature* T_{equil} of a planet is determined by setting the incident flux equal to the radiated flux, assuming that the heat from the incident radiation is uniformly distributed over a fraction f of its total surface area, and that it radiates with an emissivity of unity. We find

$$T_{equil} = \left(\frac{1-A_{Bond}}{4f}\right)^{1/4} \left(\frac{r_s}{a}\right)^{1/2} T_s \tag{20}$$

Here f = 1 for a rapid rotator and f = 0.5 for a tidally locked or slowly rotating planet with no transfer of heat from the hot to cold side. The value of T_{equil} refers to the fraction f of area over which the heat is spread; this simple formulation assumes that none of the incident heat is distributed to the (1–f) of the remaining area, which would therefore be very cold.

The effective temperature T_{eff} of a planet is determined by fitting a blackbody curve to its experimentally measured infrared emission spectrum. If a planet has an internal heat source, then $T_{equil} < T_{eff}$; otherwise, these are equal.

Table 3 lists A_{Bond}, T_{equil}, and T_{eff} for solar system planets. The terrestrial planets have negligible internal heat sources so they have $T_{eff} = T_{equil}$, but the giant planets (save for Uranus) have significant internal heat, as measured in the thermal infrared. For example, Jupiter would have T_{equil} = 110 K from albedo alone, but internal heat pushes the observed effective temperature up to 124 K.

For the Earth we get T_{equil} = 254 K, which is representative of an effective radiating altitude (~40 km). The infrared optical thickness of the atmosphere below this level isolates the radiating level from the surface. The surface of Earth is roughly 288 K, i.e., 34 K warmer owing to the greenhouse effect of H_2O and CO_2, keeping it above the freezing point of water, on average.

The *habitable zone* (HZ) is defined as that range of distances from a star where liquid water can exist on the surface of a planet. For example, Earth has surface oceans and is therefore within its HZ. To extend this to terrestrial exoplanets requires knowing the factor ($1-A_{Bond}$)/f and the greenhouse effect for that planet. Absent this knowledge, we sometimes assume that a planet is like Earth in these respects, in which case the HZ for an Earth-like planet around another star will scale as $L^{1/2}$, giving

$$a(\text{HZ, E-like}) = (1\ \text{AU})(L/L_\odot)^{1/2} \tag{21}$$

The HZ in the solar system is approximately bounded by Venus and Mars. Early in the age of the solar system the luminosity of the Sun was 60% of its present value, so Venus may have been habitable then, before it experienced a runaway greenhouse effect that raised its surface temperature far above the liquid water range. Surprisingly, Mars may have been habitable at early times as well, if it had a sufficiently thick atmosphere with a strong greenhouse effect; however, it has since lost most of that atmosphere and is now well below the liquid water range. On this basis the HZ is empirically defined to be the range 0.7 AU to 1.5 AU, scaled by the square root of stellar luminosity.

The infrared contrast C_{IR} of a planet-star system is estimated by assuming that both bodies are uniformly luminous blackbodies. In this case the contrast depends only on the effective temperatures and radii, and not on the planet phase. We have

$$C_{IR}(\lambda) = \frac{B_\lambda(T_p) r_p^2}{B_\lambda(T_s) r_s^2} \tag{22}$$

and for a solar system twin we find these contrast values at a reference wavelength of λ = 10 μm

$$C_{IR}(E) \simeq 8.2 \times 10^{-8} \tag{23}$$

$$C_{IR}(J) \simeq 2.8 \times 10^{-8} \tag{24}$$

For example, if the Jupiter-Sun system were to be directly imaged from a distance of 10 pc, the intensities of these components would be about as shown in Fig. 6, which is similar to the case of the Earth-Sun system in Fig. 4, except for Jupiter being about an order of magnitude brighter in the visible and infrared. Note also the zodi brightness in both figures, as discussed in the next section.

2.6. Exozodi

The visible, reflected-light surface brightness of a zodiacal disk similar to the solar system disk was calculated by *Kuchner* (2004a) on the basis of visible and infrared observations of the local zodi, for the case of a Hong phase function, for a disk inclined at a median angle of 60°. The tabulated values are closely fit by a simple function given by

$$m_V = 22.1 + 5.6 \log(R_{AU}) \tag{25}$$

where m_V is the apparent brightness in units of V-band magnitudes, for a 1-arcsec^2 solid angle, and R_{AU} is the radius in the disk with units of AU, in the range R_{AU} = 0.1 to 4.5 AU. The V-band flux is then

$$F_{1-zodi} = 10^{-4.43-0.4m_V}\,\Omega_{as} \tag{26}$$

in units of erg = (s cm^2 μm). Here Ω_{as} is the solid angle of the telescope in square arcseconds, which from equation (50) is

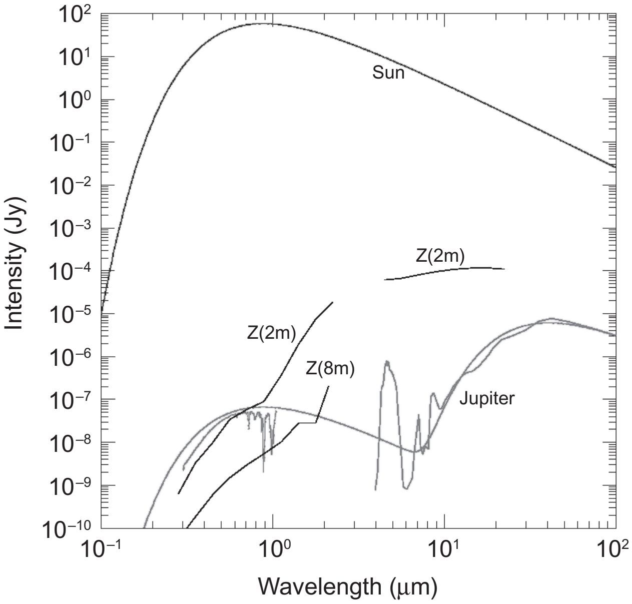


Fig. 6. Schematic spectrum of the Sun and Jupiter at 10 pc (*Kasting et al.,* 2009). Here Jupiter at maximum elongation is at 0.5 arcsec (5 AU/10 pc) with a contrast of 10^{-9} in the visible and 10^{-7} in the mid-infrared (~10 μm). The visible and infrared spectra are roughly approximated by blackbody spectra from reflected and emitted light; the prominent exception is the 4–5-μm peak, which corresponds to a spectral window on Jupiter, allowing us to see deeper and warmer levels compared to the cloud tops. The exozodi for 2-m and 8-m collectors is as in Fig. 4.

$$\Omega_{\mathrm{as}} \simeq \frac{\pi}{4}\left(\frac{\lambda}{D}206{,}000\right)^{2} \tag{27}$$

where the factor of 360 × 60 × 60/2π ≃ 206,000 converts radians to arcseconds.

As an example, equating the exozodi flux at 1 AU around a solar system twin at 10 pc, to the flux from an Earth at quadrature, and using the fact that the absolute magnitude of the Sun is $M_V = 4.82$, we find that the exozodi signal equals that of Earth for a telescope of diameter D = 2.4 m. Thus a telescope of this size or larger is needed to make Earth stand out from a solar-system-like zodi at 10 pc.

The thermal infrared intensity I can be modeled as a dilute blackbody $B_\lambda(T)$, with temperature T and optical depth τ specified empirically as a function of distance R_{AU} from the star

$$I(R_{AU}) = \tau(R_{AU})B_\lambda(T(R_{AU})) \tag{28}$$

where B is given by equation (2), for example. Here we assume the same disk as above, i.e., solar-system-zodi twin around Sun twin at 60° inclination. The optical depth is a weak function of radius, given by

$$\tau = 1.42 \times 10^{-7} R_{AU}^{-0.34} \tag{29}$$

The local temperature is a somewhat steeper function of radius, given by

$$T(R_{AU}) = 277 \times R_{AU}^{-0.467} \tag{30}$$

with units of Kelvins. For a nonsolar star, the temperature at 1 AU should be scaled as T_s, per equation (20).

If we ask the same question that we did for the optical range, namely what size telescope diameter D would give an exozodi signal equal to Earth at 10 pc (i.e., $\tau_E B_E \Omega_E = \tau_z B_z \Omega_z$, for $\Omega_z = (\pi/4)(\lambda/D)^2$), we find D = 105 m. This is larger than any future space telescope, and no ground telescope would be relevant because the warm background would be prohibitively large. So we see that the infrared zodi is going to be a bigger problem than the visible zodi. More realistically, we should use smaller (3-m) telescopes in an interferometer configuration (see section 3.18), but in this case the beam pattern of the interferometer will be a central nulling fringe projected onto the full zodi disk. Integrating over these distributions, the configuration of the Terrestrial Planet Finder Interferometer (TPF-I), for example, finds that the exozodi flux is about 100 times stronger than the Earth flux, again a large noise source, but one that might be workable.

References include *Kuchner* (2004a), for the Zodipic algorithm used in this section, and *Beichman et al.* (1999), for the TPF-I concept study.

2.7. Color

Exoplanets are faint, so the first direct images of them may be in broad photometric bands. The ratio of fluxes in two such bands, or equivalently the differences of magnitudes, give *color* information. A color-color diagram is shown in Fig. 7 for planets in the solar system. In the field of stellar astrophysics, color-color diagrams are a useful classification tool, and they could be for exoplanets as well, once we start obtaining direct images in photometric bands.

As an example, some sources of these colors are as follows. A rocky planet with little or no atmosphere tends to be relatively brighter in the red than in the blue, giving these surfaces a slightly red color, and explaining the clustering of points for Mercury, Moon, and Mars in the upper right of this diagram. A cloudy gas-giant planet with a substantial amount of gas-phase methane above its clouds, like Jupiter or Saturn, will be relatively faint in the red owing to the strong absorption bands of methane (see, for example, Table 4 for these band positions), and therefore its color will tend to look slightly blue, thus explaining the cluster of points in the blueward direction of this diagram. In ice giant planets, like Uranus and Neptune, the atmosphere is so cold that the clouds form at a relatively low level in the atmosphere, with a relatively large amount of methane above the clouds, producing almost total absorption of red light, and making the planet look significantly blue-green; this effect explains the extreme positions of these planets in this diagram. A fully cloud-covered terrestrial planet, like Venus, reflects with very little color compared to the Sun, but has a slight absorption at short visible wavelengths, possibly owing to a pigment in the sulfuric-acid cloud droplets. Earth is famously blue owing to strong Rayleigh scattering in its atmosphere (not ocean reflectivity), and therefore occupies a unique position off to the left in this diagram.

2.8. Spectroscopy

The immediate purpose of directly imaging an exoplanet is to measure its photon flux in broad and narrow wavelength bands. From these measurements we can characterize the planet in terms of mass, radius, effective temperature, age, temperature structure, molecular composition, clouds, rotation rate, and atmospheric dynamics. For Earth-like planets we can also search for habitability in terms of a surface temperature and pressure that permits liquid water, as well as signs of life, as evidenced by the presence of disequilibrium species such as coexisting oxygen (or ozone) and methane, and possibly the "red edge" reflective spectral signature from land plants. Small amounts of oxygen can be produced photochemically, and indeed we see oxygen on Mars, for example, but large amounts of oxygen cannot readily be produced (except perhaps in a runaway greenhouse situation where water is photodissociated and the hydrogen escapes, leaving oxygen), so a large amount of oxygen, such as on Earth, is a possible sign of life on a planet.

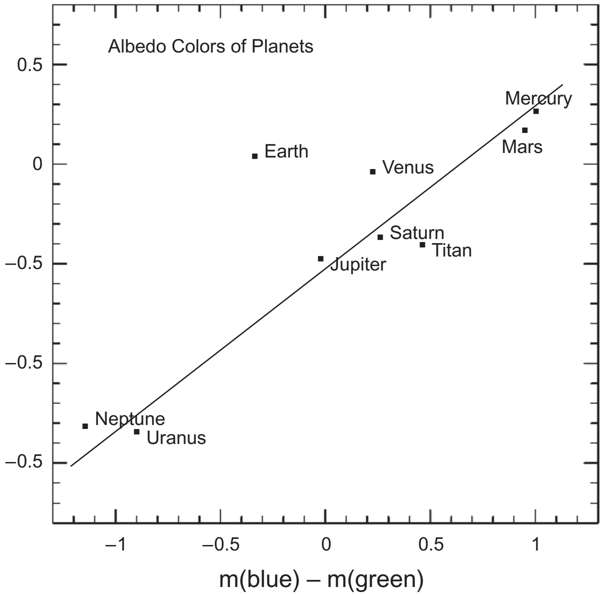


Fig. 7. A color-color diagram for planets in the solar system. The wavelength bands here are blue (0.4–0.6 μm), green (0.6–0.8 μm), and red (0.8–1.0 μm). This diagram shows that low-resolution color information can be valuable in classifying a planet (*Traub,* 2003).

As examples, Fig. 8 shows Earth's visible spectrum, as seen in Earthshine (light reflected from Earth to the dark side of the Moon, and back again to a groundbased telescope), where the Rayleigh scattering is strong, and bands of oxygen and water are prominent. Figure 9 shows the near-infrared Earthshine spectrum of Earth, in which water bands are very strong. Finally, Fig. 10 shows the thermal infrared spectrum of Earth. In all three of these cases, a simple model of Earth's atmosphere has been used to model the data, successfully reproducing the main features.

Exoplanets are faint, so we may expect that the time sequence of observations will be (1) detection in a convenient broad spectral band; (2) photometry in several broadbands, leading to a characterization by color; and (3) spectroscopy in narrow bands, leading to the identification of molecular bands and strong lines of atomic species. For Earth-like exoplanets,

TABLE 4. Spectral features of Earth.

Species	λ_0 (μm)*	$\Delta\lambda$ (μm)†	Depth‡
O_3	0.32	0.02	0.69
O_3	0.58	0.13	0.20
O_2	0.69	0.01	0.12
H_2O	0.72	0.02	0.37
CH_4	0.73	0.01	0.002
O_2	0.76	0.01	0.47
CH_4	0.79	0.03	0.001
H_2O	0.82	0.02	0.32
CH_4	0.89	0.03	0.002
H_2O	0.94	0.06	0.71
CH_4	1.00	0.05	0.011
CO_2	1.05	0.02	0.0006
H_2O	1.13	0.07	0.80
CO_2	1.21	0.03	0.01
O_2	1.27	0.02	0.15
H_2O	1.41	0.14	0.95
CO_2	1.59	0.14	0.03
CH_4	1.69	0.16	0.012
H_2O	1.88	0.18	0.97
CO_2	2.03	0.12	0.31
CH_4	2.32	0.29	0.009
H_2O	7.00	0.70	0.83
CH_4	7.65	0.59	0.09
N_2O	7.75	0.14	0.10
N_2O	8.52	0.37	0.02
CO_2	9.31	0.49	0.05
O_3	9.65	0.58	0.41
CO_2	10.42	0.65	0.04
CO_2	14.96	3.71	0.52
H_2O	20.49	7.64	0.21

Data adapted from *Des Marais et al.* (2002). Abundances are for present Earth.

[*] Central wavelength of feature.

[†] Approximate full-width at half-maximum.

[‡] Approximate depth of feature (e.g., 0.01 is a weak line, 0.95 strong) for Earth at quadrature, assuming a cloud-free atmosphere; if clouds are present, depths will be somewhat smaller.

Des Marais et al. (2002) listed all significant spectral features from about 0.3 to 100 μm wavelength, including the width and depth of each for a variety of abundances. They found that the highest resolution needed for an Earth twin is R = λ/Δλ = 70 for the O_2 band at 0.76 μm.

At visible and near-infrared wavelengths a coronagraph can utilize an *integral field spectrometer* (IFS) to provide simultaneous spectra of all pixels in the focal plane. At thermal-infrared wavelengths an interferometer has a single spatial pixel covering the entire star-planet system, so the flux in this pixel must be passed through a spectrometer before detection and image reconstruction.

2.9. Photons as Waves

A photon can be thought of as a particle when it is emitted from an atom on the surface of a star, and again when it is absorbed by a detector, but during its journey through space and through our optical instruments it is necessary to picture it as a wave. This empirical view is an expression of the famous wave-particle duality in quantum mechanics, as applied to photons.

To understand how a photon interacts with a telescope or interferometer, let us first consider how the photon gets from the star to our instruments. It is helpful to think of the light source, here a star, as a collection of many light-emitting atoms, but to visualize only one photon at a time as being emitted from that star, from a random location on the star (weighted by the surface intensity, of course). Each atom emits randomly in time and in phase with respect to each other atom; in other words, there is no coherence between one point and another on the surface of the star. Note that textbooks often speak about the partial coherence of light from a star, but this is an artifact of how we observe the star, not a property of the star itself.

This single photon propagates through space as a spherical expanding shell, with a thickness equal to the coherence length of the photon (speed of light times the lifetime of the emitting state of the atom), but not localized at any particular point or region on the sphere. We often assume that the wave is monochromatic, with a single wavelength and essentially infinite coherence length. The electric field of the photon is proportional to the real part of

$$e^{-i\left(\omega t-\vec{k}\cdot\vec{r}\right)} \tag{31}$$

times a constant and times r^{-2} where $\omega = 2\pi/f$, f is the time frequency of oscillation, t is time, $\vec{r}$ is distance from the emitting atom, $k = 2\pi/\lambda$, λ is the wavelength, and $\vec{k}/k$ is a direction vector from the atom to any point on the expanding sphere. We could use cos(X) instead of the real part of e^{iX} but the latter is more convenient for calculations. In thinking about diffraction we are entirely concerned with the interaction between the wavelength λ and the spatial dimensions of our apparatus, so we drop the time variation. Also, since the star is very distant compared to our instrument dimensions, we approximate the amplitude A of the electric field of the incident spherical wave as a plane wave

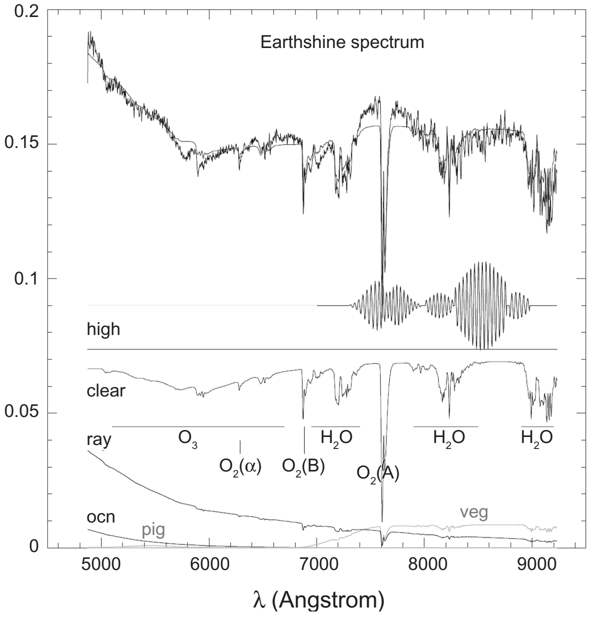


Fig. 8. The visible reflection spectrum of Earth, observed and modeled, along with the contributing spectral components from the clear atmosphere, clouds, Rayleigh scattering, and with weak contributions from the ocean as well as the red edge of land plants (*Woolf et al.,* 2006).

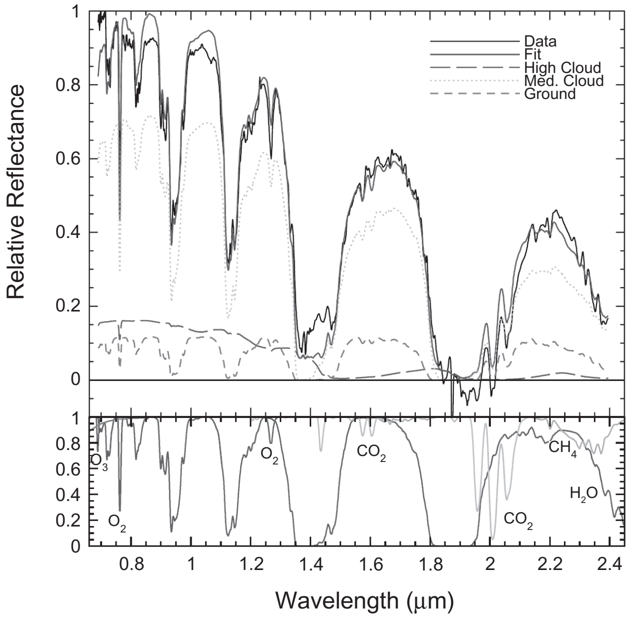


Fig. 9. The near-infrared reflection spectrum of Earth, observed and modeled, along with the contributing spectral components (*Turnbull et al.,* 2006). Gas-phase water is the dominant contributor.

$$A\left(\vec{k},\vec{x}\right)=A_0 e^{i\vec{k}\cdot\vec{x}} \tag{32}$$

where $\vec{x}$ is a local position vector centered in our instrument, and we usually set $A_0 = 1$.

The *wavefront* is defined as the surface containing all contiguous points in space at which the phase of this wave has the same value. Successive wavefronts are separated in space by λ, in phase by 2π, and in time by λ/c. If the medium is not a vacuum, but instead has an index of refraction n, then successive wavefronts are separated in space by λ/n, in phase by 2π, and in time by λ/nc, and the phase delay ϕ along a path of geometrical length z is $\phi = 2\pi nz/\lambda$.

In a medium of index of refraction n, a simple rule is to use a wavelength $\lambda = \lambda_0/n$ where λ_0 is the vacuum wavelength. For example, a wavefront that has passed through an ideal convex lens, thicker in the center than at the edges, will be delayed proportionately to the thickness of glass traversed, thus making it into a spherical converging wavefront.

The trick is now to think of this plane wave as falling on our instrument at all points equally, no matter how large an aperture, or how far apart one element of the aperture is from another. In other words, we can have one or many discrete entrance aperture elements, over as large an area as we wish, and the photon (wave) will somehow take notice of the arrangement and manage to "feel" the entire apparatus.

To visualize the interaction it is helpful to think of the Huygens wavelet picture, which is shown schematically in Fig. 11. Here, at every point on a wavefront, we imagine that little wavelets sprout and propagate outward, each on its own little spherical hemisphere. At a short time thereafter we add up all the electric fields from these propagating points. For a plane wave in free space, these wavelets will tend to cancel each other in all directions except for the single direction that is perpendicular to the wavefront. This direction defines the local *ray*, and is the basic element of geometrical optics.

For a plane wave passing through a finite-size aperture, the wavelets at the center of the aperture will tend to continue onward as before, but the wavelets near the edge will be able to propagate off to the side as well, because there are no canceling wavelets in the "shadow" of the aperture. This concept is the basis for the wave-optics picture that dominates in the following sections.

At any point inside the telescope or interferometer, if we wish to know the total electric field from the incident photon, we need only add up all the wavelets that could have reached that point, or in other words, calculate the sum of wavelet amplitudes *taking into account the phase of each wavelet* according to its distance of travel. We can write this summation as

$$A_{\text{out}}\left(\vec{x}\right)=\int M\left(\vec{x'}\right)A_{\text{in}}\left(\vec{x'}\right)e^{i\phi\left(\vec{x},\vec{x'}\right)}d\vec{x'} \tag{33}$$

where $A_{\text{out}}(\vec{x})$ is the amplitude of the total electric field at point $\vec{x}$ on the detector, $A_{\text{in}}(\vec{x'})$ is the incident electric field amplitude on the apparatus at point $\vec{x'}$ in the pupil, $M(\vec{x'})$ is a mask function that modifies the incident wavefront, and

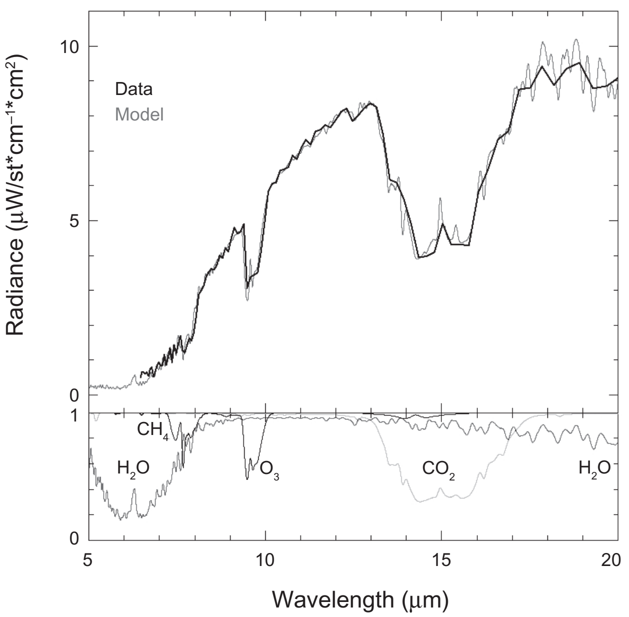


Fig. 10. The far-infrared thermal emission spectrum of Earth, observed and modeled, showing strong contributions from CO_2, O_3, and H_2O. Data (broken heavy line) is from the Thermal Emission Spectrometer, enroute to Mars, and fitted spectrum is from *Kaltenegger et al.* (2007).

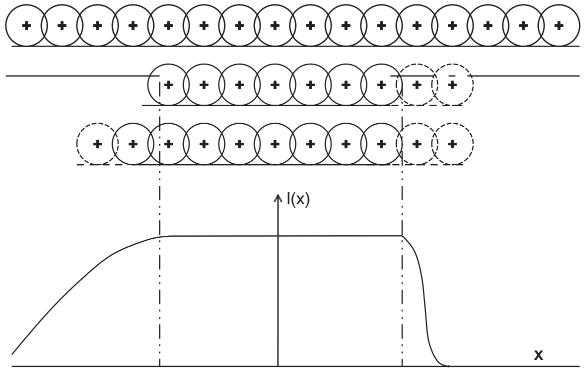


Fig. 11. Huygens' wavelets schematic. The incident wavefront approaches an opening, propagating from top to bottom. Before striking the opening the wave propagates by successive reformation of wavelets emerging from every point on the advancing wavefront, the sum of which, a short distance downstream, recreates a smooth, forward-moving plane wavefront. For illustration, this opening is drawn with a hard edge on the left but a soft semitransparent edge at the right; the geometrical shadow boundary is indicated by vertical dashed lines. At the hard edge, wavelets propagate beyond the geometrical shadow boundary. At the soft edge, wavelets are damped by the presence of adjacent weaker wavelets, with the net effect that the wavefront stays closer to the geometrical shadow boundary than in the hard-edge case.

$\phi(\vec{x}, \vec{x'})$ is the phase difference between the incident electric field at $\vec{x'}$ in the pupil and $\vec{x}$ in the detector, as measured along a minimum-time ray path. For example, a point-source star on the axis of our coordinate system will have $A_{in} = 1$ in the pupil plane, but if it is off-axis at angle $\theta \ll 1$ then $A_{in}(x') = e^{ik\theta x'}$. Also, we often have M = 1 in the pupil and M = 0 outside, although M can also be partially transparent, and can also have a phase delay of its own.

The intensity at point **x** on the detector is proportional to the magnitude-squared of the electric field

$$I(\vec{x}) = \left|A(\vec{x})\right|^2 \tag{34}$$

If the star has a finite angular size, then the intensity (not the electric field) from each point on its surface must be summed. If the detector pixel has a finite size, then the intensity over the area of the pixel must be summed as well. Finally, if the star has multiple wavelengths, then the intensity for each wavelength (weighted by the star spectrum and the transmission of the optics) must be summed.

2.10. Photons as Particles

The net intensity in each detector pixel is proportional to the *probability* that a photon will be detected in that pixel, for example, by the generation of a conduction electron in a CCD or CMOS detector. We can visualize the detection process by picturing all the wavefront segments of a photon collapsing to a single point in on the detector, no matter how large the aperture or how widespread the collection of subapertures. After many photons have passed through the apparatus, the measured intensity pattern will match the shape of the probability pattern.

The detected number of photons in a finite time interval will be given by the *Poisson process*

$$f(n, \bar{n}) = \bar{n}^n e^{-\bar{n}} / n! \tag{35}$$

where f is the probability that there will be exactly n electrons detected in a pixel, for the case where $\bar{n}$ is the expected average number of electrons. Here, of course, $\bar{n}$ is proportional to the calculated intensity distribution of diffracted light. The standard deviation of the number of detected events is $\sigma_n = \bar{n}^{1/2}$; this is called *photon noise* or *shot noise.* For large values of $\bar{n}$, say 10 or more, the probability distribution approaches a Gaussian or *normal process*

$$f(n, \bar{n}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(n-\bar{n})^2/2\sigma^2} \tag{36}$$

where again $\bar{n}$ is the mean and σ is the standard deviation, and in addition $\sigma = \bar{n}^{1/2}$.

These are important relations because directly imaged exoplanets are typically faint. For example, for an Earth twin at 10 pc we expect $\bar{n} \simeq 0.5$ photon/m^2 s^{-1} in a 10% bandwidth at visible wavelengths.

2.11. Photons in the Radio and Optical

Photons are photons, whether they have long (e.g., radio) or short (e.g., visible) wavelengths; however, there is a big difference in how they are detected. For example, we know that radio astronomers routinely use heterodyne detection, but optical astronomers do not. Why is this?

One part of the answer is Heisenberg's uncertainty principle, $\Delta E \Delta t \geq h/2\pi$. Suppose that we want to interfere an incident photon with one from a local oscillator. This will fundamentally amount to isolating it in time with an accuracy of a radian, or less, in phase, so $\Delta t \leq (\lambda/2\pi)/c = 1/(2\pi\nu)$. This gives us $\Delta E \geq h\nu$, which tells us that the uncertainty in the number of detected photons is greater than one, even though we have assumed that there is only one incoming photon in the first place. This quandry leads us to the second part of the answer, as follows.

We know from equations (46) and (49) that a wavefront incident on an opening of width D will be diffracted into an emerging beam of angular width approximately λ/D. So in two dimensions the product of area and solid angle is approximately

$$A\Omega = \lambda^2 \tag{37}$$

By time-reversal symmetry, this relationship applies to the emission process as well as the detection process. The conserved quantity, λ^2, defines a single electromagnetic mode. Applying this to the emission process, and using equation (7), we find that the photon rate $\dot{N}_\nu A$ from a blackbody, into solid angle Ω, and from area A, is

$$\begin{aligned} \dot{N}_\nu A &= \dot{n}_\nu A\Omega \\ &= \frac{2}{e^{h\nu/kT} - 1} \end{aligned} \tag{38}$$

photons per second per Hertz into AΩ. Now the uncertainty relation $\Delta p \Delta x \geq h$ for a photon, where its momentum is $p = h\nu/c$ and length is $x = ct$ gives

$$\Delta\nu\Delta t \geq 1 \text{ Hz sec} \tag{39}$$

for the product of a photon's frequency spread and total length in time. In addition, there are two polarization states possible. So using these minimum values we get the number of photons in the minimum area, minimum solid angle, minimum frequency bin, and minimum time interval, per polarization state, i.e., a single electromagnetic mode, as $n_{mode} = \dot{n}_\nu A\Omega\Delta\nu\Delta t/2$

$$n_{mode} = \frac{1}{e^{h\nu/kT} - 1} \tag{40}$$

As an example, suppose we are looking at a star or other object with a brightness temperature of T = 5000 K. Then in the visible, say $\lambda < 1$ μm, we get $n_{mode} < 0.1$ photon in a single electromagnetic mode, so we should only expect one photon at a time, on the average. However, at 10 μm we get three identical photons per mode, so heterodyne detection, with its added certainty of one photon, is just barely possible at this wavelength.

At longer wavelengths, say 1 cm, we get 3400 identical photons. This says that this photon is one of a group of 3400 others that are just like it, and are indistinguishable. Therefore we can have multiple radio antennas, each with its own receiver, completely independent of all the others, receiving some of these photons.

This explains why radio interferometers can be analyzed as if they were detecting classical waves, where each antenna can detect a small fraction of the classical wave. This is not how photons are detected, where once a given antenna detects a photon, the other antennas are automatically not allowed to detect that same photon. This is the fundamental reason why radio arrays work, because the array is showered with many identical photons. This is a consequence of stimulated emission in the blackbody source, whereby when one photon is emitted it stimulates many others to be emitted en route to leaving the source.

The Hanbury-Brown Twiss intensity interferometer is based on the above idea that there is a tendency for photons (as bosons) to arrive in pairs. This tendency is weak in the visible, but strong in the radio. The effect is strongest for a point source, as above, and it will be diluted for a source that appears to be resolved by the detecting apparatus. This is the basis for the demonstrated ability of the intensity interferometer to measure the angular diameter of stars.

3. CORONAGRAPH AND INTERFEROMETER CONCEPTS

In this section we discuss how to use coronagraphs and interferometers to observe exoplanets. There are many concepts for coronagraphs, most of which have been invented specifically for exoplanet observations. This is a very exciting field, with new ideas coming along at a fast pace, very little of which can be found in any optics textbook. This is all the more surprising, given that it was once thought that the only way to directly image an exoplanet was with an interferometer in space. Today we know that both approaches are viable, at least in an optical sense. And at the heart of the matter, interferometers and coronagraphs are essentially the same type of machine, balancing the amplitude and phase of one part of a wavefront against that of another part. This section should provide a good understanding of both kinds of telescopes.

3.1. Overview of Types

The types of coronagraphs and interferometers discussed in this section are listed in Table 5. This table is a short, representative list. A much more exhaustive list of types of coronagraphs and their theoretical properties is given in *Guyon et al.* (2006). Column 1 in Table 5 gives the name or names of a class of instrument. Column 2 labels each as primarily a coronagraph or interferometer, although this labeling is largely a matter of taste and history, rather than a fundamentally definable property; for example, in every coronagraph the rays or wavefront segments must be nearly perfectly phased so that they interfere with an extremely high degree of cancelation, so a coronagraph is in fact an interferometer with many contributing elements. If we understand interferometers, then by definition we should be able to understand coronagraphs, and vice versa. In column 3 we designate whether the instrument operates primarily in the pupil plane or the image plane, although here too the label is partially arbitrary, because in all cases some operation (i.e., an amplitude or phase adjustment, or both) is necessary in both planes. In an ideal theoretical view, there are almost no cases in which a deliberately manipulative operation takes place in a plane other than these; the exceptions are in real systems, where slight offsets of the plane of a mask or lens may occur for practical reasons, and in the case of the Talbot effect (section 4.11), where amplitude and phase can be mutually transformed.

3.2. How to Observe Exoplanets: Single Telescope

We now calculate the amplitude and intensity of an incident wavefront as observed by a simple telescope. The phase is calculated across a tilted surface in the pupil, oriented at an angle with respect to the incoming wavefront, at an angle projected onto the sky that corresponds to the point of interest in the focal plane of a perfect lens located in the plane of the pupil. For each such tilted surface there is a corresponding point in the focal plane, on a straight line from the equivalent point in the sky, through the center of the lens, to the focal plane.

The reason for this identification of a tilted surface with a point in the focal plane is that the ideal lens transfers its incident wavelet fronts along a tilted plane, independent of their individual phases, to a converging spherical wave, the convergent center of which is in the focal plane, off-axis by the tangent of the angle times the focal length.

The relative strength of an outgoing wave from one of these surfaces is determined by adding up all the wavelets on that surface. The phase at each point is $2\pi/\lambda$ times the distance between the input and output wavefronts.

3.3. Classical Single Pupil

Our telescope model is as follows. A plane wave from a point on a star is incident on a pupil plane that contains an

TABLE 5. Types of direct imaging instruments.

Name	Type	Main Plane
Pupil-masking	coron.	pupil
Pupil-mapping, PIAA	coron.	pupil
Lyot, Gaussian	coron.	image
Band-limited	coron.	image
Phase, vector vortex	coron.	image
Starshade	coron.	image
Keck Int., TPF-I, Darwin	int.	pupil
Visible nuller	coron/int.	pupil

imbedded ideal lens and is opaque elsewhere; this is shown schematically in Fig. 12. We call this plane number 1. The amplitude of the incident electric field is $A_1(x_1)$, where the subscript denotes this first plane, and the coordinates in this plane are (x_1,y_1). For simplicity of notation, we will work only in the x dimension whenever possible, and we will drop the subscripts wherever the meaning is otherwise clear.

The phase $\phi_1(x_1)$ of a wavelet in plane 1, as seen from a point at position θ_2 in plane 2, and therefore measured with respect to a reference plane tilted at angle θ_2 in the pupil, is

$$\phi_1\left(x_1\right) = 2\pi x_1 \sin\left(\theta_2\right)/\lambda \simeq 2\pi x_1 \theta_2/\lambda \qquad (41)$$

where $x_1 \sin(\theta_2)$ is the distance between the incoming wavefront from direction $\theta_0 = 0$ and the outgoing direction at angle θ_2, and we assume $\theta_2 \ll 1$.

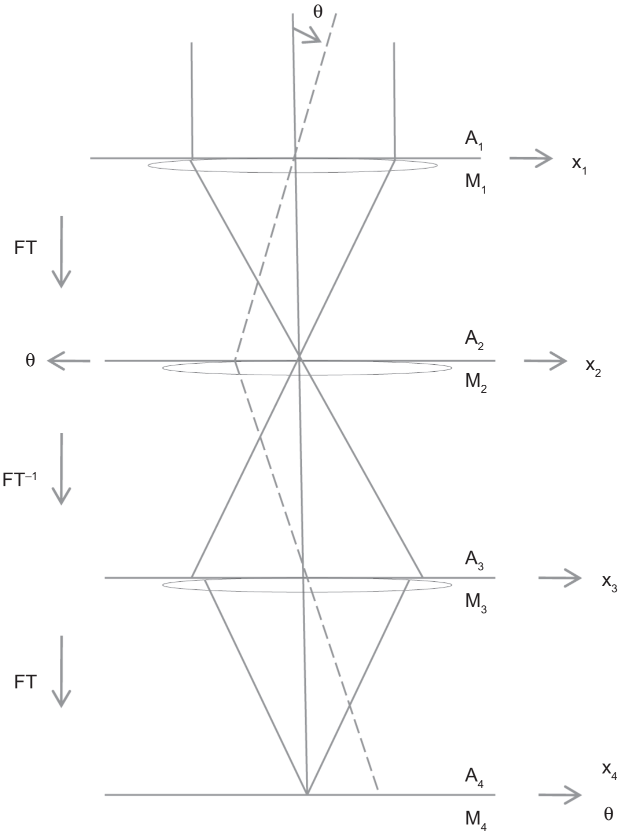


Fig. 12. Four coronagraph planes are shown: 1 is the input pupil; 2 is the image plane; 3 is a reimaged pupil; and 4 is a reimaged image plane. A simple telescope only uses planes 1 and 2, with the detector at plane 2. An internal coronagraph puts the detector at plane 4. The A_i are the incident and propagated electric field amplitudes falling on each plane. The M_i are masks in the planes. The transmitted amplitude just after each plane is M_iA_i. There is an ideal lens imbedded in each plane, with focal lengths such that the pupil in plane 1 is imaged on to plane 3, and the sky is imaged onto planes 2 and 4. The mask M_1 is the first pupil stop, and M_3 is often called the Lyot stop. In a conventional coronagraph mask M_2 could be a rectangular or Gaussian dark spot, blocking the central part of the star image. The mask M_4 can be thought of as the assembly of individual pixels in the detector. Linear positions in each plane are positive to the right, but by geometric optics the angle coordinate in plane 2 points left, while in plane 4 it points right. The star here is on-axis, at $\theta_0 = 0$.

The imbedded lens will focus the sum of wavelets that exit the pupil at angle θ_2 to a star image at a point in the image plane, i.e., plane number 2. The electric field in this plane is denoted $A_2(\theta_2)$.

The amplitude $A_2(\theta_2)$ in the image plane is the algebraic sum of all wavelets across the pupil

$$A_2\left(\theta_2\right) = \sum\left(\text{wavelets}\right) = \int_D A_1\left(x_1\right) e^{i\phi_1\left(x_1\right)} dx_1 \Big/ D \qquad (42)$$

where the integral is over the diameter D of the pupil, and $A_1(x_1)$ is the amplitude of the incoming wave in plane 1. Here the pupil is one-dimensional, but generally it is two-dimensional.

The divisor $\int_D dx_1 = D$ normalizes the righthand side by dividing out the area factor; for simplicity we will usually drop this normalization in most of the rest of this chapter, except where it improves the appearance of the result. Strictly speaking, the units of A_1 and A_2 should be the same, but in this chapter they are not, owing to the integral over the pupil; however, we shall retain this system in order to keep the notation simple. The correct factor in a result can often be calculated by applying conservation of energy between the input pupil and output image plane. Likewise, linear coordinates in each plane should be labeled x_1 for x, and x_2 for θ_2, but we prefer to let the physics dominate the math, and will often use x and θ where it is clear what is meant from the context. Regarding signs, we note that from geometric optics, a linear coordinate in the focal plane x_2 is related to the angle on the sky θ by $x_2 = -f \tan(\theta) \simeq -f\theta$, where f is the focal length of the telescope.

Inserting the approximate expression for $\phi_1(x_1)$ we get

$$A_2\left(\theta\right) = \int_D A_1\left(x_1\right) e^{i2\pi\theta x_1/\lambda} dx_1 \qquad (43)$$

This is the basic working equation for much of the imaging calculations that follow. Note that we use a "+" sign in the exponent, owing to our choice of coordinates; it is conventional to use a "–" sign, but the resulting intensities will be the same in either case.

In this case, where we retain only the linear approximation $\sin(\theta) \simeq \theta$ in the pupil plane, we speak of *Fraunhofer diffraction* (*Born and Wolf,* 1999, Chapter 8.3). The more exact, but more difficult to calculate, case is that of *Fresnel diffraction,* in which we retain higher-order terms. If the source and image are at infinity, as in Fig. 11 with the lenses as shown, these give identical results.

Note that the amplitude A(x) can be a real or complex number, where the real part is the magnitude of the electric field and the phase is the phase delay of the wavefront at that point. If the phase is a complex number, then the imaginary part is equivalent to a reduction of the amplitude, i.e., $e^{i(\phi_R(x)+i\phi_I(x))}$ is the same as $e^{-\phi_I(x)} \times e^{i\phi_R(x)}$, a reduced-amplitude wave.

For the case of a one-dimensional pupil, with an input amplitude $A_1(x_1) = 1$, we find

$$A_2(\theta) = \int_{-D/2}^{+D/2} e^{i2\pi x_1\theta/\lambda} dx_1 = \frac{\sin(\pi\theta D/\lambda)}{\pi\theta D/\lambda} D \tag{44}$$

The measured intensity is $I = |A|^2$, or

$$I_2(\theta) = \left[\frac{\sin(\pi\theta D/\lambda)}{\pi\theta D/\lambda}\right]^2 D^2 \tag{45}$$

The intensity pattern is thus the square of a $\text{sinc}(X) \equiv \sin(X)/X$ function, with a strong central peak at the point where the source star would have been imaged with geometrical optics (here $\theta_0 = 0$), and small secondary peaks.

The first zero is the solution of $I_2(\theta_{zero}) = 0$ and is given by

$$\theta_{zero} = \lambda/D \tag{46}$$

Here are two trivial examples. First, suppose we add a constant phase ϕ_0 across the aperture, for example, with a plane-parallel sheet of glass. The net amplitude in the focal plane is multiplied by $e^{i\phi_0}$ and the intensity is unchanged. Second, suppose the star is off-axis at angle θ_0, or that the telescope is mispointed by the same angle. Then the input wavefront is tilted by angle θ_0, and in the integral θ is replaced by $\theta-\theta_0$, and likewise in the expression for the intensity.

The corresponding results for a two-dimensional circular aperture are roughly similar, but to write the equations correctly we must pay attention to the scaling factor, which for most of this chapter we otherwise ignore. For a circular aperture of diameter D, the physical amplitude in the focal plane (*Born and Wolf,* 1999, Chapter 8.5.2) is

$$A_2(\theta) = \sqrt{I_0}\,\frac{2J_1(\pi D\theta/\lambda)}{\pi D\theta/\lambda} \tag{47}$$

and the corresponding intensity is $I_2 = |A_2|^2$, giving

$$I_2(\theta) = I_0\left[\frac{2J_1(\pi D\theta/\lambda)}{\pi D\theta/\lambda}\right]^2 \tag{48}$$

Here $J_1(X)$ is the Bessel function of first order, similar to a damped sine function. The first zero-intensity angle is

$$\theta_{zero} \simeq 1.22\lambda/D \tag{49}$$

which is the famous result for a clear circular aperture. The full-width at half-maximum (FWHM) of the intensity pattern is

$$\theta_{FWHM} \simeq 1.03\lambda/D \simeq \lambda/D \tag{50}$$

so this value is often referred to as the diameter of the diffraction-limited image of a point source. The relative intensities of the central and first four secondary maxima are 1.0, 0.017, 0.0042, 0.0016, and 0.00078 at respective values of 0.0, 1.6, 2.7, 3.7, 4.7 times λ/D.

The physical part of this result lies in the expression for I_0, which is

$$I_0 = \frac{EA_{pupil}}{\lambda^2 F^2} \tag{51}$$

where E is the rate of energy per unit area in the pupil plane, for example, $E = f_\lambda \Delta\lambda$ from equation (9), with $\Delta\lambda$ the wavelength range being observed. The term A_{pupil} is the area of the pupil, $\pi D^2/4$ for a circular pupil. The term F is the focal length of the system. Notice that I_0 has units of rate of energy per unit area in the focal plane, similar to E in the pupil plane. The I_0 factor in equation (51) can also be derived by applying conservation of energy in the pupil and image planes, given the shape of the diffraction pattern in the focal plane. As a note of caution, it is not always trivial to convert a one-dimensional diffraction result into a two-dimensional result, but the results in this paragraph show a method that should be useful in other contexts.

3.4. Fourier Optics Approximation

Suppose we write equation (43) above as

$$A_2(\theta) = \int_{-\infty}^{+\infty} M_1(x_1) A_1(x_1) e^{i2\pi\theta x_1/\lambda} dx_1 \tag{52}$$

where $M_1(x_1)$ is the pupil transmission function, here the top-hat or rectangular function, $M(x) = \text{rect}(x,D)$. The rectangular function is defined here as

$$\begin{aligned} \text{rect}(x,D) &= 1 \quad \text{if} -D/2 < x < +D/2 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{53}$$

Here again, $A_1(x_1)$ is the amplitude of the incident wave; however, now it can extend over all values of x_1, as befits an expanding spherical (here nearly flat) wavefront from an atom on a distant star. This gives us the following important result: The amplitude $A_2(\theta)$ of the electric field in the focal plane of a telescope is the Fourier transform of the function $M_1(x_1)A_1(x_1)$, the electric field transmitted by the pupil of the telescope.

Note: Given our physically inspired convention that θ is in the opposite direction of x_2, the above relation is a regular Fourier transform, not an inverse Fourier transform, as this relation is sometimes stated. The difference is not important, as long as it is consistent.

Suppose that the image plane (number 2) is transparent, and is immediately followed by a lens that has a focal length equal to half the distance from the image to pupil plane. From geometric optics we know that the input pupil plane (number 1) will be imaged, one to one, in a conjugate pupil plane (number 3) downstream. Note that in the Fraunhofer approximation, the Fourier-transform integral is a linear operator, and it is reversible, so that the light may be thought of as traveling in either direction; the amplitude in this second pupil plane will be the inverse Fourier transform of the amplitude in the image plane, to within a constant factor. Another way to see this is to notice that in the wavelet picture, the sum (or integral) that propagates from plane number 1 to plane number 2 should be exactly the same in propagating from plane number 2 to plane number 3; however, since $x_3 = -\theta f$, the coordinate in plane 2 is reversed in sign, so the wavelet phases reverse sign, and the Fraunhofer integral becomes an inverse Fourier transform. We denote the amplitude in the third plane by $A_3(x_3)$.

Thus we have the result

$$A_3\left(x_3\right) = \int_\theta A_2\left(\theta\right) e^{-i2\pi\theta x_3/\lambda} d\theta \tag{54}$$

Substituting and exchanging the order of integration we get

$$A_3\left(x_3\right) = \int_{x_1} M_1\left(x_1\right) A_1\left(x_1\right) \int_\theta e^{i2\pi\theta\left(x_1 - x_3\right)/\lambda} d\theta dx_1 \tag{55}$$

Now use the fact that

$$\int_{-\infty}^{+\infty} e^{i2\pi\left(x-x'\right)\theta/\lambda} d\theta = \delta\left(\left(x - x'\right)/\lambda\right) \tag{56}$$

where δ is the Dirac delta function, and that

$$\int_{-D/2}^{+D/2} \delta\left(\left(x - x'\right)/\lambda\right) dx = \lambda \ \mathrm{rect}\left(x', D\right) \tag{57}$$

We get

$$\begin{aligned} A_3\left(x_3\right) &= \lambda \int_{x_1} M_1\left(x_1\right) A_1\left(x_1\right) \delta\left(x_1 - x_3\right) dx_1 \\ &= \lambda M_1\left(x_3\right) A_1\left(x_3\right) \end{aligned} \tag{58}$$

which is an exact copy of the transmitted amplitude from the input pupil plane, to within a constant. This illustrates a general procedure whereby we can propagate light from one plane to another, using ideal lenses, infinite focal planes, and Fourier transforms.

We have used two simplifications in our picture of diffraction. The first simplification is that we have assumed that the pupil and image spaces are one-dimensional; expanding to the realistic case of two dimensions is in principle straightforward, but in practice often leads to more complex integrals; the net result is an increase in complexity with little gain in understanding. The second simplification is that we use the small angle approximation $\sin(x) \simeq x$, i.e., Fraunhofer diffraction; including the higher-order terms leads to Fresnel diffraction. Both topics are well covered in *Born and Wolf* (1999) and other standard texts.

3.5. Convolution Perspective on Imaging

A conceptually elegant way to view the operation of an imaging system, including a coronagraph, is to take advantage of the Fourier-transform relation between the planes in Fig. 12 and the fact that the Fourier transform of a product of functions is the convolution of the individual Fourier transforms, and also the Fourier transform of a convolution of two functions is the product of the individual Fourier transforms. In other words, $FT(f * g) = FT(f) \cdot FT(g)$, and also $FT(f \cdot g) = FT(f) * FT(g)$.

Referring to plane 1 in Fig. 12, we see that the input amplitude is $A_1(x_1)$, the mask is $M_1(x_1)$, and the output is M_1A_1.

At plane 2, the input amplitude is $FT[M_1A_1](x_2) = FT(M_1) * [FT(A_1)](x_2)$. The mask is $M_2(x_2)$. And the output is $M_2 \cdot [FT(M_1) * FT(A_1)](x_2)$.

At plane 3 the input is the FT^{-1} of the plane-2 output. We multiply the mask M_3 times that function, and apply the convolution rules again. This gives the output from plane 3 as $M_3 \cdot [FT^{-1}(M_2) * (M_1 \cdot A_1)]$.

At plane 4 the input is the FT of the plane-3 output. Substituting and simplifying we get the field at plane 4 to be $FT(M_3) * [M_2 \cdot FT(M_1A_1)]$.

The value of this picture will become clear when we look at individual coronagraph designs. The band-limited mask design will show clearly how this picture, and in particular the expression for the output of plane 3, can bypass difficult integrals to give a clear physical picture.

3.6. Imaging Recipes

We summarize the general case of propagation from plane 1 through plane 4 in Fig. 12 as a recipe for later reference, as follows.

The electric field incident on plane 1 is

$$A_1\left(x_1\right) = e^{i2\pi x_1\theta_0/\lambda} \tag{59}$$

for a point source located at angle θ_0.

A lens in plane 1 produces an electric field $A_2(\theta_2)$ incident on plane 2

$$A_2\left(\theta_2\right) = \int_{x_1} M_1\left(x_1\right) A_1\left(x_1\right) e^{i2\pi\theta_2 x_1/\lambda} dx_1 \tag{60}$$

where $M_1(x_1)$ is a mask on the output side of plane 1.

Likewise, a lens in plane 2 produces an electric field $A_3(x_3)$ incident on plane 3

$$A_3\left(x_3\right) = \int_{\theta_2} M_2\left(\theta_2\right) A_2\left(\theta_2\right) e^{-i2\pi x_3\theta_2/\lambda} d\theta_2 \tag{61}$$

where $M_2(\theta_2)$ is a mask on the output side of plane 2.

Finally, a lens in plane 3 produces an electric field $A_4(\theta_4)$ incident on plane 4

$$A_4(\theta_4) = \int_{x_3} M_3(x_3) A_3(x_3) e^{i2\pi\theta_4 x_3/\lambda} dx_3 \tag{62}$$

where $M_3(x_3)$ is a mask on the output side of plane 3.

3.7. Practical Considerations

Real optics can depart from ideal in several ways. One departure is that opaque baffles will diffract light slightly differently depending on whether the material of the stop is a metal or a dielectric. In practice this effect is mainly noticed in the immediate vicinity (a few wavelengths) of the stop. As an example, subwavelength diameter holes in a screen can have much greater transmission through a metal screen than the corresponding holes in a dielectric screen. A related example is that partially transmitting materials, as are used in some coronagraphs, will always have a phase shift associated with a given level of opacity, as determined by the Kramers-Kronig relation that connects the real (absorbing) and imaginary (phase shifting) parts of the index of refraction of the material.

More mundane considerations include the presence of scattering dust on optics, which can spoil a theoretically low contrast, and atmospheric phase fluctuations in laboratory experiments, which can amount to at least a wavelength or more of time-varying path, especially if there is a heating or cooling air flow nearby.

For coronagraphic telescopes that operate at extreme (planet-detecting) contrasts, the problem of *beam walk* becomes a factor. This arises if the telescope is body-pointed slightly away from a target star, and this error is compensated by the tip-tilt of a subsequent mirror, driven by a star-tracker (for example), thereby slightly shifting the beam transversely across the optics, and encountering a slightly different pattern of surface errors in those optical elements, thereby generating different speckle patterns (cf. section 4).

3.8. Off-Axis Performance

Stars have finite diameters, and telescopes have finite pointing errors. The performance of a coronagraph will be degraded by either effect, because the transmission of a coronagraph will generally increase in both cases, even if it is theoretically zero for an on-axis delta-function source. The degree to which even an ideal coronagraph will leak light is quantified in the concept of the *order of the null*.

The intensity leak is proportional to θ^n where θ is the off-axis angle and n is an integer power (*Kuchner,* 2004b, 2005). Since intensity is proportional to electric field squared, n is always even. Some examples are $n = 0$ (top-hat, disk phase knife); $n = 2$ (phase knife, four-quadrant phase mask); $n = 4$ (notch filter, band-limited mask, Gaussian, achromatic dual zone); $n = 8$ (band-limited, notch).

Given a functional form of the off-axis transmission of a mask, the effects of a finite-diameter star as well as a slightly mispointed telescope can be directly calculated by integrating the transmission function over the possibly offset disk of the star.

3.9. Effects of Central Obscuration, Spider, Segments

If a circular pupil of diameter D has a central obscuration of diameter d, then the summation of wavelets can be written as the sum over the larger pupil [amplitude $A_2(x_2,D)$] from equation (47) minus the sum over the smaller one [amplitude $A_2(x_2,d)$], and the intensity in plane 2 will be

$$I_2(\theta) = I_0\left[\frac{2J_1(\pi\theta D/\lambda)}{\pi\theta D/\lambda}A_D - \frac{2J_1(\pi\theta d/\lambda)}{\pi\theta d/\lambda}A_d\right]^2 \tag{63}$$

where A_D is the area of the larger pupil, and A_d is the area of the smaller, obscuring pupil. Notice that the weighting of the respective amplitudes is by area, not diameter. This case is an example of the need to get the correct energy-based coefficients of a diffraction pattern. Comparing this with the nonobscured case in equation (48), we see that the central core of the image is slightly sharper (i.e., slightly better angular resolution), but at the expense of significantly stronger diffraction rings around this core.

If a circular pupil of diameter D has a spider arm of width w placed across its center, or if the pupil is made up of a segmented mirror with a gap of width w, then the intensity pattern of a point source, in the focal plane 2 in a direction perpendicular to the spider or gap, will be

$$I_2(\theta) = I_0\left[\frac{2J_1(\pi\theta D/\lambda)}{\pi\theta D/\lambda}A_D - \frac{\sin(\pi\theta_w w/\lambda)}{\pi\theta_w w/\lambda}\frac{\sin(\pi\theta_D D/\lambda)}{\pi\theta_D D/\lambda}A_w\right]^2 \tag{64}$$

which is the square of the net amplitude of a circular clear pupil minus the amplitude of a rectangular blocked strip. Here the angular directions in the focal plane are θ_w in a direction parallel to the w dimension of the obscuration, θ_D in a direction parallel to the D dimension of the obscuration, and $\theta = \sqrt{\theta_w^2 + \theta_D^2}$, and $A_D = \pi D^2/4$ is the area of the full pupil, and $A_w = wD$ is the area of the obscuration. The strip obscuration or gap adds a surprisingly large diffracted intensity at large angles.

Equations (63) and (64) are examples of *Babinet's principle,* in which the amplitude resulting from an opaque part in a beam is represented as the negative of the amplitude from a transparent version of the opaque part.

As an example, suppose that we have a segmented primary mirror of total width D, made up from two adjacent segments, each of width D/2, and that there is a thin strip of width w overlying the joint between the segments. Examples are the adjacent segments of the Keck telescopes or the James Webb Space Telescope (JWST). Let us assume that there is a coronagraph that can suppress the central star and its diffraction pattern. Then the contrast of the diffracted spike compared to the (suppressed) central star intensity is $C(\text{spike}) \simeq (w/D)^2$. If we want this to be as faint as Earth in brightness, we need

$C = 10^{-10}$. If the segments are each D = 1 m wide, then we need $w \leq 10$ μm in width, about one-eighth the thickness of a human hair. This is much smaller than can be easily accomplished.

These examples show that diffracted light from an obscuration or gap in the pupil can generate a relatively large intensity at angles well away from the diffraction core of λ/D from a point-like star. Only a few types of coronagraphs are immune to these obscuring elements.

3.10. Pupil-Edge Apodization

One way to eliminate the diffraction side lobes of a pupil is to reduce the sharp discontinuity in the transmitted wavefront at the edge of the pupil. As we saw in the discussion of the single pupil with a sharp edge, the Huygens wavelets spread out dramatically at such an edge. An early suggestion was to taper the transmission of the pupil at the edges, to avoid a sharp change of transmission. For example, we could figuratively spray black paint on a telescope mirror so that the center was clear and the edge totally opaque. A more practical (and approximate) method is to surround the perimeter of a pupil with a lot of inward-pointing black triangles or similar pointed spikes, such that the azimuth average transmission drops smoothly from 1 at the center to 0 at the edge; this technique works surprisingly well, and can be implemented with ordinary tools.

Suppose we model this by a Gaussian intensity transmission function $e^{-(x/x_0)^2}$, which corresponds to a Gaussian amplitude function $e^{-(x/x_0)^2/2}$. We want this function to be small at the edge, so we assume that $x_0 < D/2$. The effective diameter D_{eff} is then roughly the FWHM of the intensity distribution, which is $D_{eff} \simeq x_0 2\ln(2)$. Inserting this amplitude into the one-dimensional equation for net amplitude, and making the approximation that $x_0 \ll D/2$, we find the normalized intensity pattern in the focal plane to be

$$I_2(\theta) = e^{-(2\pi\theta x_0/\lambda)^2} \tag{65}$$

This result shows that tapering the pupil, in the extreme case of strong tapering near the edges, can have a dramatic effect on the image of a point source, namely concentrating it in a tight image with no sidelobes. If we had integrated from 0 to D/2 instead of 0 to ∞ we would have obtained a similarly compact central peak, but with finite sidelobes. In the example shown, the intensity drops to 10^{-10} at an angular distance of about $\theta_{-10} \simeq 2\lambda/D_{eff}$, showing that in principle this is a powerful method of minimizing sidelobes. This technique is generally called *apodizing,* meaning to remove the feet. We could have used a cosine or other similar function, with roughly similar results. Obviously the technique can be extended to a more realistic circular aperture.

In practice, the Gaussian function, which tapers to zero on an infinite range, is replaced by a very similar-looking profile, a *prolate spheroid* function, defined on a finite range. The intensity pattern in the focal plane, with a prolate spheroid tapering of the pupil, can be made to drop to $\theta_{-10} \simeq 4\lambda/D$, which is still a dramatic feat.

3.11. Pupil-Masking Apodization

The concept of *pupil masking* is a practical version of the spray-paint apodization described above. In the pupil-masking apodization method, the pupil is covered by an opaque sheet that has tapered cutouts through which the wavefront can pass, as shown in Fig. 13. The cutouts are designed to transmit more light at the center of the pupil, and less at the (say) left and right edges. The corresponding projected left and right areas on the sky have faint diffracted light in the focal plane, so a planet could be detected in these areas. There are no sharp edges perpendicular to the left and right, so little diffraction. However in the orthogonal direction, say up and down, there are a lot of perpendicular edges, so a lot of light is diffracted in those directions. The search space on the sky is therefore limited to the projected areas with diffraction below a target threshhold, say 10^{-10}. The concept is so simple that it could be tested with paper and scissors at an amateur telescope. These pupil-masking types of stops have been tested in the laboratory, and have achieved dark zones as deep as about 10^{-7}, limited perhaps by minor imperfections in the mask edges. Also their transmission is relatively low, since much of the pupil is covered. Nevertheless these masks stand as a proof of principle that it is possible to beat the iron grip of diffraction, and they have inspired numerous other inventions.

References include *Kasdin et al.* (2003, 2005).

3.12. Pupil-Mapping Apodization

Another way to achieve a Gaussian-like amplitude distribution across a pupil is to rearrange the incoming rays, so to speak, so that they do not uniformly fill the pupil but rather crowd together near the center, and become sparse at the edges. This will make the amplitude of the electric field stronger at the center and weaker at the edges, but for visualization it is easiest to think of rays. Pupil-mapping is illustrated in Fig. 14.

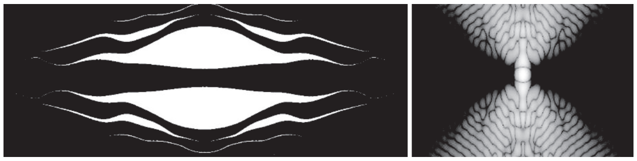


Fig. 13. **(a)** This optimized pupil mask has six openings within an elliptical envelope designed to match the pupil of TPF-C (section 6.9). **(b)** The corresponding image plane diffraction pattern, showing a strongly suppressed central star, and dark-hole areas (on the left and right) with residual intensities below a theoretical contrast of 10^{-10}. The IWA is $4\lambda/D$ and the throughput is 30%.

This effect can be created with two aspheric lenses (or mirrors). The first lens is shaped so that it converges the wavefront across most of the aperture but tapers to a flat piece of glass at the extreme edges so that those rays continue on parallel to their input direction. The second lens is placed well before the converging rays cross and shaped so as to make all the output rays parallel again, i.e., a diverging central part tapering to a flat piece of glass at the edges.

The resulting output beam will have the same diameter as the input beam but will be bright near the center and faint near the edge, in a Gaussian-like (prolate spheroid) amplitude distribution. Focusing this beam with a lens will generate a Gaussian-like bright star image with very faint ($\leq 10^{-10}$) sidelobes. The core image can theoretically be contained within a radius of about 2λ/D sky angle. The pair of lenses can be replaced by a pair of off-axis mirrors, and an off axis paraboloid can be incorporated into the second mirror so that a star image is formed without the need for an additional optic. These lenses or mirrors can also be manufactured at a small diameter and placed at the back end of a large off-axis telescope, at an image of the input pupil. Since this method uses all the input energy collected by the telescope, and forms a star image that is close to the theoretical limit of about 1λ/D in radius, it is theoretically an optimum type of coronagraph.

The pupil-mapping concept is also called *phase-induced amplitude apodization* (PIAA), a descriptive name reflecting the fact that the electric field amplitude is apodized near the edge of the pupil by manipulating the point-to-point phase of the incoming wavefront. In other words, slowing down the wavefront near the center of the beam causes it to converge locally, while the extreme edge of the beam is not slowed at all, and therefore continues on as if untouched.

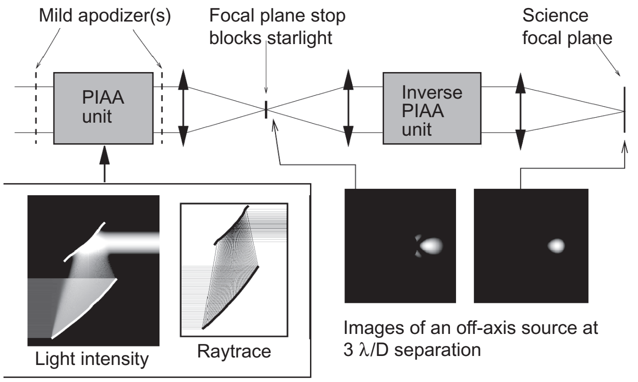


Fig. 14. Pupil-mapping (PIAA) schematic, showing in the lower left how a pair of specially shaped mirrors can reshape a uniformly bright input beam into a Gaussian-like beam (bright in the center, faint at the edges), which when focused by a lens produces a Gaussian-like star image (with extremely weak sidelobes), which in turn can be blocked by a focal-plane stop, so that any adjacent planet images are transmitted. The inverse PIAA unit is needed to reshape the strongly aberrated planet images, giving nearly normal planet images in the science focal plane.

In a practical system, the trick is to shape the outer 1% or so of the radius of the first mirror so as to spray all the light incident on this narrow annulus over a large part of the output pupil, forming the wing of the Gaussian-like distribution; this is difficult because it requires extremely accurate polishing of a relatively sharp radius of curvature into the edge of the optic.

The amplitude near the edge can also be controlled by inserting a thin opaque ring, or a fuzzy outer blocker, at a small loss of light. In practice this step appears to be necessary.

An opaque blocker should be placed at the image of the star, and the surrounding dark field, including any exoplanets, passed on to a detector. The image of an off-axis exoplanet at this point will be aberrated because the pupil mapping mirrors distort any off-axis point source, with the distortion getting worse with distance from the on-axis star; see section 3.23 for a general explanation of the this type of distortion. Fortunately, and surprisingly, this aberration can be eliminated by sending the image through a reversed set of lenses or mirrors.

References include the original paper (*Guyon,* 2003), a ray-trace theory of mirror shapes (*Traub and Vanderbei,* 2003), a proof that the inverse system will restore off-axis image shapes (*Vanderbei and Traub,* 2005), and recent laboratory results (*Guyon et al.,* 2009, 2010).

3.13. Lyot (Hard-Edge) Mask Coronagraph

Suppose that we build the simplest kind of coronagraph, a focusing lens of diameter D in the plane 1, a dark occulting mask of angular diameter θ_r in plane 2, followed by a lens that makes an image of the pupil in plane 3, a mask just after, and another lens that finally images the sky in plane 4. (Equivalently, the dark spot could be replaced by a transmitting hole in a mirror, which could be more convenient.)

The *hard-edge mask* $M_2(\theta_2)$ multiplies the amplitude according to

$$M_2(\theta_2) = 1 - \text{rect}(\theta_2, \theta_r) \qquad (66)$$

so that M is 0 (opaque) in the range ($-\theta_r/2$, $+\theta_r/2$) and 1 (transmitting) outside this range. This shape is also known as a *top-hat mask.*

This method was pioneered by B. Lyot for the purpose of imaging the faint ($<10^{-6}$) corona of the Sun. Today any such instrument for observing a faint source near a bright one is called a *coronagraph,* and the particular configuration that uses a hard-edge mask and stop (see below) is called a *Lyot coronagraph.*

We might expect that if the mask covers the central few Airy rings, we might block most of the light, or scatter it off to a large angle where it might be blocked in plane 3, and thus generate a greatly diminished star image (and diffraction pattern) in plane 4. Using equation (61), and an on-axis star, i.e., $A_1(x_1) = 1$, the amplitude $A_3(x_3)$ in plane 3 is

$$A_3(x_3) = \int_{\theta_2} M_2(\theta_2) A_2(\theta_2) e^{-i2\pi x_3 \theta_2/\lambda} d\theta_2 \qquad (67)$$

Then using equation (60) and taking the input pupil to be $M_1(x_1) = \text{rect}(x_1,D)$ we get

$$A_3\left(x_3\right) = \int_{\theta_2}\int_D e^{i2\pi\left(x_1-x_3\right)\theta_2/\lambda}\left[1-\text{rect}\left(\theta_2,\theta_r\right)\right]dx_1 d\theta_2 \quad (68)$$

where the integration range of θ_2 is ±∞ and the range of x_1 is ±D/2.

Using equation (56) for the delta function and the rectangular function from above, we find the amplitude in the second pupil plane to be

$$A_3\left(x_3\right) = \lambda\Bigg[\text{rect}\left(x_3,D\right) - \int_{-D/2}^{+D/2}\frac{\sin\left(\pi\left(x_1-x_3\right)\theta_r/\lambda\right)}{\pi\left(x_1-x_3\right)}dx_1\Bigg] \quad (69)$$

There is no simple expression for this amplitude, but by visualizing the terms we can see that the amplitude in plane 3 is a copy of that in plane 1 minus an oscillatory function of the position coordinate x_3. The amplitude is indeed diminished inside the range ±D/2, but there is now amplitude scattered outside this range.

We stop the light at the edge of the x_3 pupil from propagating further by inserting a *Lyot stop* here. This stop is an undersized image of the input pupil, but since the amplitude is small but finite at all radii, the diameter of the stop is a matter of judgement: A small diameter improves the rejection of the sidelobes in the x_4 image plane, but at the expense of overall throughput and angular resolution.

Thus the advantage of the Lyot coronagraph is its simplicity, but the disadvantage is that it will always have a finite leakage, so there is a limit to how small a contrast it can achieve.

Note that the amplitude will have a zero and a sharp discontinuity at x_3 = ±D/2, independent of wavelength, a characteristic of this (and the following) coronagraph.

References include *Lyot* (1933).

3.14. Gaussian Mask Coronagraph

A soft-edge mask to block the first star image should work better than a hard-edge one. An example is the Gaussian-shape amplitude mask

$$M_2\left(\theta\right) = 1 - e^{-\left(\theta/\theta_g\right)^2} \quad (70)$$

Following through as with the rectangular mask, we find the amplitude in plane 3 to be

$$A_3\left(x_3\right) = \lambda\left[\text{rect}\left(x_3,D\right) - \frac{1}{\sqrt{\pi}}\int_{z_-}^{z_+} e^{-z^2}dz\right] \quad (71)$$

where the integration limits are $z_\pm = \pi\theta_g(\pm D/2 - x_3)/\lambda$. This can be expressed in term of an error function, but a simple approximation, for illustration here, is to replace the integral by the magnitude of the integrand at the center of the range, multiplied by the width of the integration range (a crude box-car integration). This gives an approximate amplitude

$$A_3\left(x_3\right) \approx \lambda\left[\text{rect}\left(x_3,D\right) - \frac{\sqrt{\pi}\theta_g D}{\lambda}e^{-\left(\pi\theta_g x_3/\lambda\right)^2}\right] \quad (72)$$

Depending on the value of θ_g, which would nominally be in the neighborhood of a few times λ/D, this amplitude also has a small value inside the range ±D/2, a zero at the edge of that range, and more amplitude diffracted out beyond that range, which can then be removed with a hard-edge Lyot stop for M_3.

The diameter of this stop is a free parameter, the tradeoff being that a stop diameter less than D will reduce the background diffracted light in the second focal plane (good), but it will also reduce the light from the off-axis exoplanet (the light of which precisely fills the diameter D) and increase the diffracted diameter of its image (both bad). The reason for the latter is that this stop is now the effective diameter of the system for the exoplanet, and its diameter will determine the image size in plane 4, as can be verified by another Fourier transform.

3.15. Band-Limited Mask Coronagraph

The band-limited mask coronagraph is an evolutionary step beyond the rectangular and Gaussian-mask coronagraphs described above. The band-limited design is the answer to the question; can we find a masking pattern in plane 2 that minimizes the transmitted light from an on axis star, but at the same time allows an off-axis exoplanet image to pass? There are two extremes of answers to this question, absorbing masks (this section) and phase masks (section 3.16), and there are intermediate types that combine absorption and phase.

To illustrate the band-limited concept, we choose an amplitude mask with a periodic modulation

$$M_2\left(\theta\right) = c\left[1 - \cos\left(\theta/\theta_B\right)\right] \quad (73)$$

where c = 1/2 so that $0 \le M \le 1$, and θ_B is a scale factor that might chosen to be on the order of one to a few times λ/D so as to suppress the star's diffraction pattern out to the first one or several diffraction sidelobes. Note that this mask has no hard edges, so we might expect that it will not diffract light at large angles; in fact it is periodic, like a diffraction grating, so we might expect it to diffract light at a specific angle. Its spatial frequency range is limited, hence the name. The mask extends over many periods, in principle over $x_2 = \pm\infty$.

We calculate the amplitude of the electric field in the plane 3 using equation (61)

$$A_3\left(x_3\right) = \int_D\int_\theta M_2\left(\theta\right)e^{i2\pi\left(x_1-x_3\right)\theta/\lambda}d\theta dx_1 \quad (74)$$

Substituting, we get

$$A_3(x_3) = c\int_D\int_\theta e^{i2\pi(x_1-x_3)\theta/\lambda}\left[1-\cos(\theta/\theta_B)\right]d\theta dx_1 \quad (75)$$

Using cos(z) = (e^{iz} + e^{-iz})/2 and the delta function, and integrating over x_1 = (–D/2,+D/2), we find

$$\begin{aligned} A_3(x_3) = c\lambda\big[&\mathrm{rect}(x_3, D) - \\ &0.5\,\mathrm{rect}(x_3 - \lambda/(2\pi\theta_B), D) - \\ &0.5\,\mathrm{rect}(x_3 + \lambda/(2\pi\theta_B), D)\big] \end{aligned} \quad (76)$$

which is indeed zero over the central range of the pupil, with finite amplitude at the edges, and a zero at the exact edge. We can emphasize this by writing the amplitude as

$$\begin{aligned} A_3(x_3) = c\lambda\big[&0 \times \mathrm{rect}(x_3, D - \lambda/\pi\theta_B) + \\ &\mathrm{wiggle}(x_3 - D, \lambda/\pi\theta_B) + \\ &\mathrm{wiggle}(x_3 + D, \lambda/\pi\theta_B)\big] \end{aligned} \quad (77)$$

where the central part has zero amplitude, and the function called wiggle(x,Δx) is defined here by analogy with the rect(x,Δx) function.

This technique will work with any band-limited mask function, for example, $\sin^2$ (as above), $\sin^4$, 1–J_0, 1–sinc, 1–sinc^2, and (1–sinc^2)^2.

All these masks have greater average transparency away from the central dark region than the 1–cos mask. Here we work out the 1–sinc example

$$M_2(\theta) = c\left[1 - \frac{\sin(\theta/\theta_B)}{\theta/\theta_B}\right] \quad (78)$$

We need this relation first: $\int_0^\infty \sin(z)\cos(mz)/z dz$ is 0 if |m| > 1 but π/2 otherwise. The result is

$$\begin{aligned} A_3(x_3) = c\lambda\big[&0 \times \mathrm{rect}(x_3, D - \lambda/\pi\theta_B) + \\ &\mathrm{tilt}(x_3 - D, \lambda/\pi\theta_B) + \\ &\mathrm{tilt}(x_3 + D, \lambda/\pi\theta_B)\big] \end{aligned} \quad (79)$$

where the same central region has zero amplitude, and the ring of scattered light at the edges occupies the same width but has a tilted shape, tilt, similar to the wiggle function above. The amplitude is sketched in Fig. 15.

References include *Kuchner and Traub* (2002) and *Trauger and Traub* (2007).

3.16. Phase and Vector Vortex Mask Coronagraphs

Several types of coronagraphs depend on a pure phase manipulation of the photon in the focal plane.

A *phase mask* in the focal plane is four contiguous quadrants of transparent material onto which the star is focused at the symmetry point, with adjacent quadrants differing in optical thickness by a half-wavelength. The transmitted beam will be nulled out on axis, but an object imaged mainly in one of the quadrants will be transmitted.

A *scalar optical vortex* is a structural helical phase ramp, and generates a longitudinal phase delay by operating on both polarizations. The center of the optical vortex is a phase singularity in an optical field, which generates a point of zero intensity, resulting from a phase screw dislocation of the form $e^{il_p\psi}$, where l_p is called the topological charge, and ψ is the azimuthal coordinate.

A *vector optical vortex* is a space-varying birefringent mask that operates on the orthogonal components of polarization of the photon, such that a star focused at the center of the pattern will be nulled on axis. One version of this design is a set of engraved concentric annular groves in a glass plate, with the groove spacing smaller than a wavelength. The fineness of the grooves ensures that the diffracted light cannot go into side orders, but must continue to propagate in its original direction. However, the groove depth is designed such that the projected part of the photon's electric vector

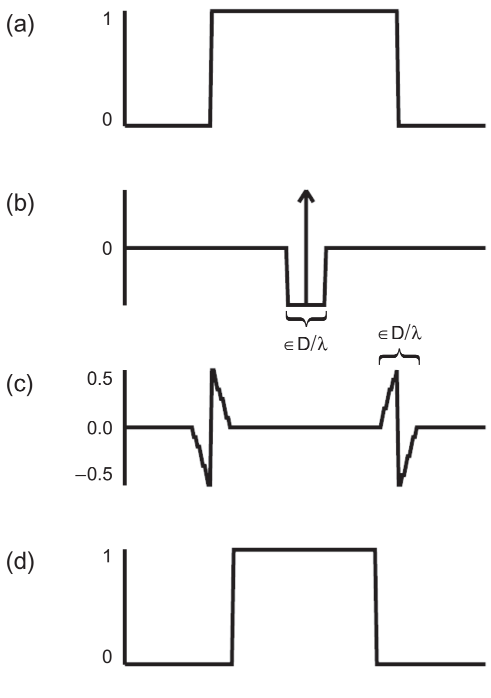


Fig. 15. Band-limited amplitudes for the 1–sinc mask. **(a)** Input pupil amplitude, at plane 1. **(b)** FT of mask, which, when convolved with the pupil, produces the amplitude distribution shown in **(c)**, equivalent to plane 3. **(d)** Transmission of Lyot stop in plane 3, transmitting the amplitude of the star in the central part of the plane (here zero owing to the action of the band-limited mask), and also transmitting the amplitude of any off-axis source, such as a planet. The diffracted light from the star is blocked by the slightly undersized Lyot pupil located just after plane 3.

that is parallel to the local groove direction sees a different optical path than the perpendicular vector component, and by the circularity of the design, a centered star image will effectively have half its amplitude delayed by a half wavelength with respect to the other half. The on-axis light is thus nulled, but of course energy is conserved so the light diffracts off at an angle to the incident beam. These grooves can be noncircularly symmetric, generating total delays of more than one full wavelength per rotation in azimuth. The number of half-wavelengths per turn is called the *topological charge* l_p of the design, with designs ranging from $l_p = 2$ (a minimum) up to $l_p = 6$ (limited somewhat by the complexity of the design). The difficulty of manufacturing subwavelength grooves or index of refraction spiral ramps has encouraged the search for another medium.

Liquid crystal polymers enable another type of vector optical vortex. Here the polymer molecules can be lined up so as to make a locally varying halfwave plate in which the optical axes rotate about the center of symmetry of the plate.

To show the action of a phase mask we illustrate with a one-dimensional example. Suppose that we put a phase mask in plane 2 with a phase discontinuity of π centered on the star image in the image plane. This mask can be written as

$$\begin{aligned} M_2(\theta_2) &= e^{+i\pi/2} \quad \text{if } \theta_2 > 0 \\ &= e^{-i\pi/2} \quad \text{if } \theta_2 < 0 \end{aligned} \tag{80}$$

So with a simple, on-axis star, and a pupil of diameter D in plane 1, equations (60) and (61) give an amplitude in plane 3 which is

$$A_3(x_3) = \frac{2\lambda}{\pi}\int_0^\infty \frac{\sin(z)}{z}\sin(mz)\,dz \tag{81}$$

where $m = 2x_3/D$, and which has the solution

$$A_3(x_3) = -\frac{\lambda}{2\pi}\ln\left(\frac{m-1}{m+1}\right)^2 \tag{82}$$

This amplitude has a sharp peak at the edge of the pupil ($m = 1$), but is not especially small inside the pupil, and therefore fails to be a good coronagraph. It is, however, an illustration of what a point source would do if it were centered on one of the arms of a four-quadrant phase mask, which is one of the reasons that this type is less ideal than the vector vortex type.

To move this calculation into a two-dimensional plane, let us replace x in any plane with (r, ψ), the radial and azimuthal coordinates, with appropriate subscripts for each plane. Following *Mawet* (2005, Appendix C), we find

$$A_3(r_3,\psi_3) = \frac{4e^{i2\psi_3}}{\lambda D}\int_0^\infty J_1(\pi D\theta/\lambda)\,J_2(2\pi r_3\theta/\lambda)\,d\theta \tag{83}$$

which illustrates the complexity of working in the full two-dimensional picture, but which fortuitously has a solution in Sonine's integral, giving

$$\begin{aligned} A_3(r_3,\psi_3) &= 0 \qquad \text{if } r_3 < D/2 \\ &= \frac{e^{i2\psi_3}}{\pi r_3^2} \quad \text{if } r_3 > D/2 \end{aligned} \tag{84}$$

where the dimensions are arbitrary.

This shows that a point-source star, centered on a vector vortex mask, will have zero amplitude transmitted into the second pupil, following the image plane, and that the star will be diffracted into a bright ring that peaks just outside the second pupil, and falls off quadratically with distance beyond that point. This is therefore an ideal coronagraph, making full use of the collecting pupil, and requiring a Lyot stop that exactly matches the input pupil.

References include *Mawet et al.* (2010).

3.17. External Occulters

An *external occulter* or *star shade* coronagraph is a concept in which a blocking mask is placed between the source and the telescope. The mask is made large enough to cover the star, but not so large as to obscure a nearby planet. In geometrical optics terms, the occulter must be larger than the telescope diameter, but have an angular radius smaller than the planet-star angular separation. For example, if $D_{tel} = 6$ m, and $\theta_{planet} = 0.1$ arcsec, then the occulter must be separated from the telescope by at least $z = D_{tel}/\theta_{planet} = 12{,}000$ km. Clearly, a telescope in low-Earth orbit will not work, but a drift-away or L2 orbit would be feasible, assuming that the positioning control can be accomplished.

The telescope-facing side of the occulter should look dark compared to the planet, so it must face away from the Sun. Thus we require that the angle between the occulter and the Sun be less than about 90°.

The occulter must move from target to target. For example, if there are about 200 potential targets in about 44,000 deg^2 of sky, then the average distance between targets is about 15°.

If the mask is circular, then wavelets from the edge will all have an equal optical path to points on the star-occulter axis, and there will be a bright central diffraction spike, the *Arago spot,* sometimes called the *Poisson spot.* To reduce the intensity of the Arago spot, the edge of the occulter must be softened, exactly as with a pupil or image mask for an internal coronagraph. Interestingly, this softening can be in the form of a moderate number of cut-outs around the edge of a circle, the structures between the cut-outs being called *petals,* as in the shape of a flower. Thus the occulter can be fabricated as a connected binary mask.

Suppose that we add a plane 0 in front of plane 1 in Fig. 12, with no imbedded lens. Then the originally proposed "hyper-Gaussian" mask is given by the continuous function

$$
\begin{aligned}
M_0\left(r_0\right) &= 0 && \text{if } r_0 < a \\
&= 1-\exp\left(-\left[\left(r_0-a\right)/b\right]^n\right) && \text{if } r_0 < r_0(\max) \\
&= 1 && \text{if } r_0 > r_0(\max)
\end{aligned}
\qquad (85)
$$

where M_0 is the rotationally averaged amplitude transmission. The value of r_0(max) is taken to be that for which the width of the petal is very small, e.g., ~1 mm.

Subsequently, significant improvements to the hyper-Gaussian shape have been made by optimizing the shape function, including accounting for a wide band of wavelengths, and a dark hole over the full diameter of a telescope, with margin for positioning. A typical functional shape for M_0 is a prolate spheroid. An example is shown in Fig. 16.

If a telescope of diameter D_{tel} is placed in the stellar shadow zone of an occulter of tip-to-tip diameter D_{occ}, and their separation is z, then the telescope will have a clear view of a planet that is separated from the star by an angle θ_{IWA} where

$$\theta_{IWA} = \left(D_{occ} + D_{tel}\right)/2z \qquad (86)$$

If diffraction is ignored, so that the shadow diameter is equal to the occulter diameter, and if we wish to have a tolerance margin of, say, ±1 m, for navigating the telescope into the shadow (so that $D_{occ} = D_{tel} + 2$ m here), and if we want an IWA of $\theta_{IWA} \simeq 0.1$ arcsec (to see an Earth around a Sun at 10 pc), and if we assume that the telescope is JWST (so $D_{tel} =$ 6.5 m), then we require $z > (D_{tel} + 1)/\theta_{IWA} \simeq 15{,}000$ km, and of course $D_{occ} = 8.5$ m.

In reality, diffraction into the shadow region forces these values to be much larger. Since the telescope is not in an image or pupil plane, Fraunhofer diffraction does not apply, and we must use the full power of Fresnel diffraction theory. In this framework, the Fresnel number N_F is conserved, where N_F is given by

$$N_F = D_{occ}^2/\lambda z \qquad (87)$$

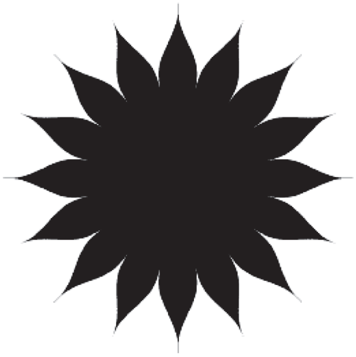


Fig. 16. An optimally-shaped star shade, from *Vanderbei et al.*, (2007), in which the ideal continuous apodization function is replaced by a 16-petal approximation with discrete (0 or 1) transmission.

The meaning of this is that the relative shape of the diffracted light in the shadow zone is the same if D_{occ} or λ or z are varied, while N_F is held constant.

As a specific example, we show in Table 6 the result of a calculation using an optimized shape for the occulter petals, for cases where the shadow intensity is 10^{-10} or less over the area of a circle of diameter D_{tel} + 2 m (to allow for a ±1 m navigation error). The Fresnel number for these examples is $N_F = 70$. We see that for JWST, for example, with $D_{tel} = 6.5$ m, we will need an occulter with tip-to-tip diameter $D_{occ} = 70$ m, at a distance of z = 140,000 km. For this case we can see a planet relatively close to its star, $\theta_{IWA} = 50$ mas, meaning that a search for Earth-like planets in the habitable zone of nearby stars could be possible. However, to do this, we will need a very large occulter at a very large distance, meaning that fuel for repositioning may become a limiting factor.

Another potential limiting factor is the accuracy requirement on the edge shape of the occulter. Errors in the shape will generate speckles in the focal plane of the telescope. A simulation has shown that edge errors on the order of 0.2 mm RMS can generate focal plane speckles at the 10^{-10} level, so this is approximately the tolerance of manufacturing and deployment of the petals.

References include *Cash* (2006), *Arenberg et al.* (2006), *Lyon et al.* (2007), *Vanderbei et al.* (2007), *Kasdin et al.* (2009), *Shaklan et al.* (2010), and *Glassman et al.* (2010).

3.18. How to Observe Exoplanets: Multiple Pupils

To obtain high angular resolution on a star-planet system we can use two or more separated telescopes instead of a single large one, increasing from a diameter D to a potentially much larger baseline B, and thereby reducing the angular resolution λ/B. A *nulling interferometer* is two or more telescopes arranged so as to collect segments of an incident wavefront and combine them with a half-wavelength path delay, so that the central star is largely canceled by balancing the electric fields and phases. Three examples that are especially relevant to direct imaging of exoplanets are the Keck Interferometer Nuller (KIN), the TPF-I, and the Large Binocular Telescope Interferometer (LBTI). Here we discuss the first two of these.

3.19. Beam Combination with an Interferometer

We distinguish here between nulling interferometers and imaging interferometers.

A nulling interferometer collects segments of a wavefront using several telescopes, sends these segments through delay

TABLE 6. Occulter examples.

D_{occ} (m)	z (km)	D_{tel} (m)	θ_{IWA}
70	140,000	6.5	50
50	72,000	4.0	72
37	39,000	2.4	98

lines to equalize their optical paths from the star, adds a half-wavelength extra delay to one or more of the paths, combines the beams directly on top of each other using semitransparent beam splitters, and focuses all the light onto a single detector. The net effect is that an on-axis star is canceled out by the half-wave delay, but that light from near the star is not canceled because its phase shift differs from π by an amount that increases with distance from the star, reaching 2π or effectively 0 wavelengths at an angle $\theta_0 = 0.5\lambda/B$. The situation can be pictured in terms of a fringe pattern projected on the sky, with the minimum-transmission point of the pattern centered on the star, preventing the star from being detected, while a disk around the star is allowed to be transmitted to the detector, multiplied by the fringe pattern.

An imaging interferometer is similar but adds a repeated and continuously changing delay such that the projected fringe pattern on the sky sweeps back and forth across the star and disk, say, and the transmitted light is recorded and later analyzed to extract the image by a deconvolution or data-fitting algorithm. We do not discuss this type any further in this chapter.

Returning to the nulling interferometer, we note that the examples discussed below are similar in many ways, yet different in that the KIN is designed to measure the zodiacal dust brightness, whereas TPF-I is designed to reject the zodi signal and search instead for point-source planets.

References include *Traub* (2000).

3.20. Interferometric Nulling

The thermal mid-infrared is an attractive spectral range for characterizing exoplanets, because it contains spectral features of H_2O, O_3, CO_2, and potentially CH_4 (see Table 4), and because the planet/star contrast is more favorable than in the visible (see Figs. 3 and 5).

However, for a given angular resolution, longer wavelengths require larger telescopes, or baselines, so for wavelengths 10–20 times greater than visible, conventional (~8 m) telescopes are not sufficient. Fortunately, long-baseline interferometers, with baselines on the order of B ~100 m, are able to do this.

The next two sections discuss an existing groundbased interferometer for exozodi observations, and a proposed space interferometer for exoplanet spectroscopy. The section that follows those is a note on the advantages and disadvantages of rearranging wavefront segments, a topic that is especially relevant to interferometers as well as to the pupil mapping coronagraph.

3.21. Nullers to Measure Zodiacal Light

The KIN was built to measure the zodiacal light in the 8–12 μm wavelength range around nearby stars, in preparation for the Terrestrial Planet Finder Coronagraph (TPF-C) and TPF-I. The KIN uses the two Keck 10-m telescopes, adjusted so that a target star is depressed by a factor of about 100 in intensity, allowing the surrounding zodi to be measured. This task was successfully completed. In operation, the KIN requires that the wavefront segments on the individual telescopes be flattened, using adaptive optics, and that the rapidly varying piston error of each segment be controlled to a fraction of a wavelength, using delay lines. Here we focus on the basic principle of nulling with the KIN.

The KIN has two major pupils (the two telescopes themselves), but four subpupils (the left and right halves of each telescope). The reason for splitting each telescope into two subpupils is that thermal emission from the sky above each telescope can be better suppressed. Figure 17 shows a schematic plan view of the pupils. Figure 18 shows sky coordinates of the target.

The KIN operates in the thermal infrared, so the telescope and sky emission are a huge, fluctuating background that needs to be removed. This is done by rapidly chopping between the three states shown in Table 7. In each state, the optical paths are adjusted to give the phases in this table. In particular, if we chop between states SZB and ZB, we will measure the star flux. Chopping between states ZB and B will give the zodi flux. The ratio of these results gives the contrast zodi/star. The output beams are sent through a prism so that the spectrum is split among 16 wavelength channels, and each is measured separately.

The combining phases here ignore the fact that an ideal beam splitter imparts a $\pi/2$ phase difference. We also treat the chopping as if it is the difference of two discrete states, whereas in fact the delay lines are scanned with a linear ramp that is slightly longer than the longest wavelength, and four measurements are made on the output intensity during this ramp, timed differently for each wavelength channel, from which the amplitude of the signal is extracted.

Numerically, a dust density corresponding to about 100 times the solar system level will produce a mid-infrared contrast at the KIN of about 10^{-2}, and in fact this is about the 1σ level of accuracy.

Each point in the sky, at radial angle θ from the optical axis, and at position angle α from the x axis, is a separate source, and is treated individually. Let us assume that the

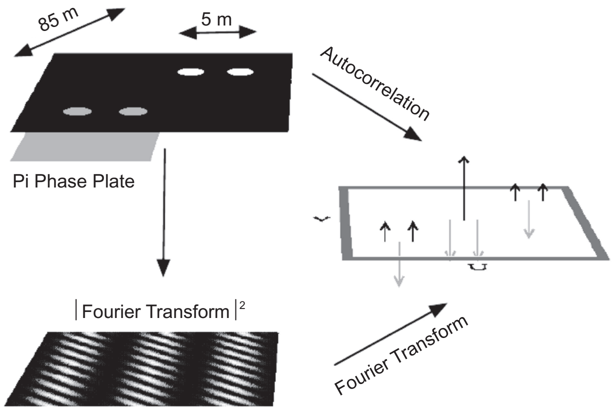


Fig. 17. Keck Nuller input is shown.

amplitude of the electric field from a photon from (θ, α) is $A_1(x_1) = e^{i\phi_1(x_1)}$, i.e., with magnitude unity and phase

$$\phi_1\left(x_1\right) = 2\pi\vec{\theta}\cdot\vec{r}/\lambda \qquad (88)$$

where $\vec{r}$ is a coordinate in the pupil, projected toward the star.

The four beams are combined with beamsplitters, effectively overlapping the wavefront segments directly on top of each other. The summed amplitudes give an output amplitude A_2 where

$$\begin{aligned}A_2 = \sum_{j=1}^{4} A_1\left(j,\vec{r}\right)M_1\left(j,\vec{r}\right) = \\ e^{i2\pi\left[\theta\cos(\alpha)(-B/2)+\theta\sin(\alpha)(+b/2)\right]/\lambda+\phi_A} + \\ e^{i2\pi\left[\theta\cos(\alpha)(-B/2)+\theta\sin(\alpha)(-b/2)\right]/\lambda+\phi_B} + \\ e^{i2\pi\left[\theta\cos(\alpha)(+B/2)+\theta\sin(\alpha)(+b/2)\right]/\lambda+\phi_C} + \\ e^{i2\pi\left[\theta\cos(\alpha)(+B/2)+\theta\sin(\alpha)(-b/2)\right]/\lambda+\phi_D}\end{aligned} \qquad (89)$$

Inserting the phases for each state, collecting terms, squaring to get intensities, and dropping a factor of 16, we find

$$I_2\left(SZB\right) = \cos^2\left(\pi\theta B\cos(\alpha)/\lambda\right)\cos^2\left(\pi\theta b\sin(\alpha)/\lambda\right) \qquad (90)$$

$$I_2\left(ZB\right) = \sin^2\left(\pi\theta B\cos(\alpha)/\lambda\right)\cos^2\left(\pi\theta b\sin(\alpha)/\lambda\right) \qquad (91)$$

$$I_2\left(B\right) = \cos^2\left(\pi\theta B\cos(\alpha)/\lambda\right)\sin^2\left(\pi\theta b\sin(\alpha)/\lambda\right) \qquad (92)$$

A sketch of the resulting fringes, i.e., the transmission pattern projected on the sky, is shown in Fig. 18. We see that the star (at θ = 0) is transmitted in state SZB, but nulled in states ZB and B, so chopping between these gives the star flux as

$$I_2\left(SZB\right) - I_2\left(ZB\right) = I\left(star\right) \qquad (93)$$

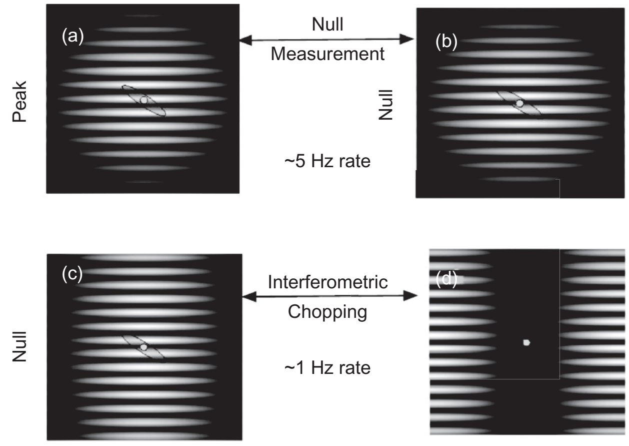


Fig. 18. Keck Nuller output is shown.

TABLE 7. KIN phase states.

State	ϕ_A	ϕ_B	ϕ_C	ϕ_D	Signal
SZB	0	0	0	0	star + zodi + background
ZB	0	0	π	π	zodi + background
B	0	π	π	2π	background

We also see that the zodi (at θ > 0) is transmitted in state ZB, but nulled in state B, so chopping between these gives the zodi, times the transmission pattern, as

$$\begin{aligned}I_2\left(ZB\right) - I_2\left(B\right) = \\ I\left(zodi\right)\sin^2\left(\frac{\pi\theta B\cos(\alpha)}{\lambda}\right)\cos\left(\frac{2\pi\theta b\sin(\alpha)}{\lambda}\right)\end{aligned} \qquad (94)$$

The beam pattern is known, and for a zodi disk that extends over several periods of the pattern the effective transmission is about a factor of 1/2. Therefore the contrast, zodi/star, is given by twice the ratio of equations (94) to (93).

References include *Colavita et al.* (2008) and *Barry et al.* (2008).

3.22. Nullers to Measure Exoplanets

The thermal infrared space missions TPF-I and Darwin, in the U.S. and Europe, are designed to detect and characterize exoplanets, down to and including Earth-like ones, around nearby stars. Here we describe the preferred method of operation that these projects have in common.

The TPF-I/Darwin schematic is identical to the KIN schematic. However, in TPF-I/Darwin the four collecting mirrors are mounted on four free-flying separated spacecraft, and the combining unit is another spacecraft at (x,y) = (0,0) but above the plane at $z \gg B$. The key fundamental difference is in the phases added to each beam, which are specified in Table 8.

Inserting these phases into equation (89) and working through as before, we find the following detected intensities, again dropping a factor of 16

$$I_2\left(left\right) = \cos^2\left(\frac{\pi\theta B\cos(\alpha)}{\lambda} + \frac{\pi}{4}\right)\sin^2\left(\frac{\pi\theta b\sin(\alpha)}{\lambda}\right) \qquad (95)$$

TABLE 8. TPF-I/Darwin phase states.

State	ϕ_A	ϕ_B	ϕ_C	ϕ_D	Signal
left	0	π	0 + π/2	π + π/2	null star + left zodi + left planet
right	0	π	0–π/2	π–π/2	null star + right zodi + right planet

$$I_2(\text{right}) = \sin^2\left(\frac{\pi\theta B\cos(\alpha)}{\lambda} + \frac{\pi}{4}\right)\sin^2\left(\frac{\pi\theta b\sin(\alpha)}{\lambda}\right) \quad (96)$$

In both states the star at $\theta = 0$ is nulled by the $\sin^2(0) = 0$ term. For a planet at any value of (θ, α) the signal is

$$I_2(\text{right}) - I_2(\text{left}) = I(\text{planet})\sin\left(\frac{2\pi\theta B\cos(\alpha)}{\lambda}\right)\sin^2\left(\frac{\pi\theta b\sin(\alpha)}{\lambda}\right) \quad (97)$$

So the planet will generate a signal given by its intensity times the beam pattern on the sky. As the array rotates about the line of sight, the pattern will sweep across the planet, producing a modulation pattern that is uniquely characteristic of the planet's brightness, and also its radial (θ) and azimuthal (α) position in the sky, so that a map of its position can be unambiguously reconstructed after one-half of a full rotation of the array.

Importantly, any symmetric brightness component will be removed by the chopping, so if the zodi is bright and symmetric, it will drop out of the signal stream. Specifically, if $I_{zodi}(\theta, \alpha) = I_{zodi}(\theta, \alpha + \pi)$, then the zodi signal is

$$I_2(\text{right}) - I_2(\text{left}) = 0 \quad (98)$$

This is helpful in detecting the planet, especially if the zodi is bright. Obviously asymmetries in the zodi will be detected, but since these might be generated by planets in the first place, this will be of value to measure.

Interestingly, by reprogramming the TPF-I chopping sequence we can easily measure either the symmetric part of the target signal or the asymmetric part. And in fact this could be done in a single chopping sequence if desired. Thus a full picture of the target can be built up. The fringe pattern scales with wavelength, so the output beam should be dispersed onto a detector array, the same as for KIN.

As an example, if the individual mirrors are 2 m in diameter, then at 10 μm wavelength the FWHM of each diffraction-limited beam pattern on the sky is $\lambda/D \simeq 1.0$ arcsec FWHM, which just barely will accommodate a Jupiter at 10 pc ($\theta \simeq 0.5$ arcsec radius). In each wavelength interval, all the light from the system falls on a single pixel. The modulation of that signal gives us a picture of the target at an angular resolution of $\lambda/2B \simeq 0.010$ arcsec, assuming a baseline of up to about $B = 100$ m, although in principle there is no limit to B.

References include *Beichman et al.* (1999) and *Cockell et al.* (2009).

3.23. Golden Rule

When thinking about an interferometer, and the image that could be formed in its focal plane, there is an important geometrical consideration that should be noticed: the *"golden rule"* of reimaging systems. This rule says that in order to have a wide field of view at a detector, the relative geometry of the input pupil must be preserved at the output pupil, to within a constant magnification factor. This rule was originally formulated for multi-telescope arrays such as the original Multi-Mirror Telescope (MMT) with its six primary mirrors, and it applies to later systems such as TPF-I.

There are three kinds of systems that do not obey this rule, and in each case the focal planes have extremely narrow fields of view. The first kind is the *pupil densification* system of a large array of telescopes in space, covering a baseline of several thousand kilometers, and phased up on a planet around a nearby star, for the purpose of making a true, spatially resolved image of that planet. In this concept the widely spaced collecting telescopes (the input pupil) are reimaged as a close-packed array (the output pupil), followed by an imaging lens. Because the two pupils are not related by a single scaling factor the output image has a very small field of view, which in this case is designed to be slightly larger than the diameter of the planet being imaged.

The second kind is the pupil-mapping coronagraph discussed in section 3.12, in which the output image of the first two mirrors has a field of view on the order of a few times λ/D, not sufficient to image a planetary system. Here the pupils are both circular, but the rays are rearranged, thereby essentially forcing a variable magnification factor between the two pupils, as a function of radial distance. However, in this case the image can be subsequently passed through a reversed set of optics, largely restoring the useful field of view.

The third kind is the family of interferometers discussed in sections 3.21 and 3.22 above. Here, by superposing the output pupils on top of each other, the exit pupil (a single opening) is clearly not a scale copy of the multiple openings in the entrance pupil pattern. This extreme case has a correspondingly tiny field of view, essentially just the diffraction beamwidth of the individual telesopes (λ/D).

References include *Traub* (1986), *Labeyrie* (1996), and *Pedretti et al.* (2000).

3.24. Visible Nuller

The visible nuller (Fig. 19) is a coronagraph-interferometer hybrid that can be used with segmented-mirror telescopes, such as the Thirty Meter Telescope (TMT), and is similar in plan view to the KIN and TPF-I, except that the baselines are more equal because they have to fit within a roughly circular primary mirror footprint. All the equations for KIN and TPF-I can be applied to the visible nuller, and in particular it can measure the symmetric as well as asymmetric parts of the target brightness distribution.

References include *Shao et al.* (2008).

4. SPECKLE CONCEPTS

All the discussions about methods in the preceding sections have dealt with idealized situations using perfect optics. In reality, nothing is perfect, and as work in exoplanet imaging has progressed, the effect of small variations in the photon propagation path lengths across the collected wavefronts has become a significant issue. In imaging, the effects of wavefront errors manifest themselves as *speckles.* Speckles appear as shown in Fig. 20. They are by far the dominant source of noise that must be controlled in order to image exoplanets.

4.1. Speckles from a Phase Step

The basic idea of speckles can be demonstrated with a simple example. Suppose that the wavefront incident on a telescope is advanced by a phase step ϕ/2 on one half of the pupil, and delayed by ϕ/2 on the other half. With reference to equation (60), the net amplitude in the focal plane becomes

$$A_2(\theta) = \int_0^{+D/2} e^{+i\phi/2} e^{i2\pi x_1\theta/\lambda} dx_1 + \int_{-D/2}^{0} e^{-i\phi/2} e^{i2\pi x_1\theta/\lambda} dx_1 \qquad (99)$$

The resulting amplitude in the focal plane is

$$A_2(\theta) = \frac{\sin(\pi D\theta/2\lambda)}{\pi D\theta/2\lambda} \cos(\pi D\theta/2\lambda + \phi/2) D \qquad (100)$$

If the wavefront has ϕ = 0, i.e., no phase jump, then we recover the standard diffraction result

$$A_2(\theta, \phi = 0) = \frac{\sin(\pi D\theta/\lambda)}{\pi D\theta/\lambda} D \qquad (101)$$

which is a single peak at the origin. However, if the total phase step is ϕ = π, i.e., a half-wavelength, then we get

$$A_2(\theta, \phi = \pi) = \frac{\sin^2(\pi D\theta/2\lambda)}{\pi D\theta/2\lambda} D \qquad (102)$$

which is a pair of peaks (speckles), each similar in width to the original single peak, and separated by about twice that width. In other words, where we once had a single image of the star, we now have two adjacent images. Smaller phase jumps will produce intermediate results, i.e., a pair of speckles but with unequal intensities and smaller separation. Clearly, we could continue to subdivide the pupil into smaller segments, producing about as many speckles as there are distinctive phase patches across the pupil.

4.2. Speckles from Phase and Amplitude Ripples

If we use an ideal telescope to image a star, and the wavefront from the star has been slightly distorted by possibly random phase and amplitude fluctuations, from an intervening atmosphere or from the telescope itself, then the natural result is a weakened star image surrounded by a halo of speckles. If the fluctuations are large, then the speckles will dominate and the star image will become just another speckle. If the telescope is a coronagraph, so the diffraction pattern is suppressed, then the speckles will certainly dominate.

Any continuous wavefront across the pupil can be represented by a sum of sine and cosine waves. For reference, we recall the standard result from Fourier analysis

$$A(x) = \sum_{n=0}^{\infty} \left(a_n \cos(2\pi nx/D) + b_n \sin(2\pi nx/D)\right) \qquad (103)$$

where A(x) is any real function on the interval x = (–D/2,+D/2). The cos( ) and sin( ) functions form an orthogonal basis set, and the coefficients are obtained by projecting A(x) onto this basis set and using the orthogonality. Multiplying both sides by cos(2πmx/D) or sin(–) and integrating we find

$$a_n = \frac{2}{D} \int_{-D/2}^{+D/2} A(x) \cos(2\pi nx/D) dx \qquad (104)$$

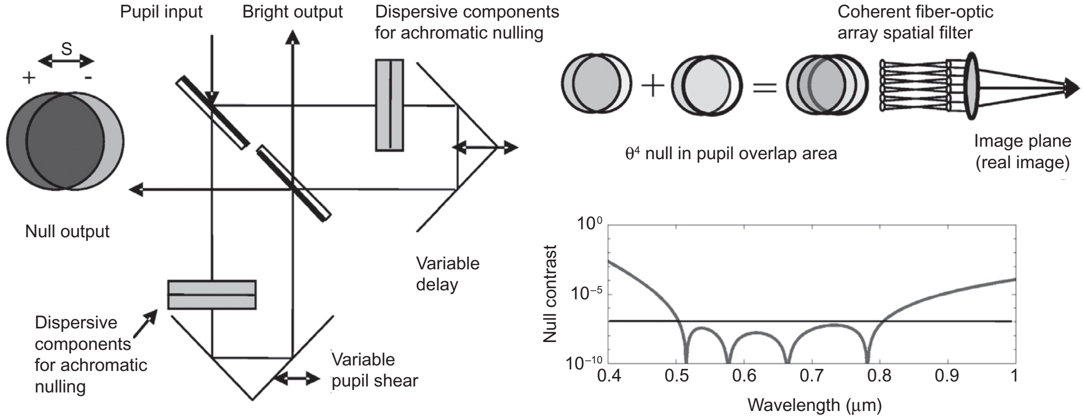

Pupil input
Bright output
Dispersive components for achromatic nulling
S
+
-
Null output
Variable delay
Dispersive components for achromatic nulling
Variable pupil shear
Coherent fiber-optic array spatial filter
+
=
θ⁴ null in pupil overlap area
Image plane (real image)
10⁰
10⁻⁵
10⁻¹⁰
Null contrast
0.4
0.5
0.6
0.7
0.8
0.9
1
Wavelength (µm)

Fig. 19. Visible Nuller schematic is shown. If the input pupil is a segmented mirror, then the shear s is set to a multiple of the segment spacing, in one pass through the interferometer, and set to a multiple along a different axis, in a second interferometer, and only the doubly overlapped segments are used in the fiber output stage.

$$b_n = \frac{2}{D}\int_{-D/2}^{+D/2} A(x)\sin(2\pi nx/D)dx \quad (105)$$

This method of representing functions in terms of a basis set can be extended to complex functions. Let A(x) be any complex function on the interval x = (–D/2,+D/2). Then A(x) can be expanded as

$$A(x) = \frac{1}{2}\sum_{n=-\infty}^{+\infty} c_n e^{i2\pi nx/D} \quad (106)$$

where the coefficients are

$$c_n = \frac{2}{D}\int_{-D/2}^{+D/2} A(x)e^{-i2\pi nx/D}dx \quad (107)$$

and n = 0, ±1, ±2, . . .

As an example, we could write $A(x) = A_0e^{i\phi(x)}$ where A_0 is a constant amplitude and $\phi(x)$ is a spatially-varying phase, e.g., linear for a tilted wavefront, etc. If ϕ is complex, then the imaginary part represents the spatial variation of amplitude, which could also be absorbed in the coefficient as $A(x) = A_0e^{-\mathrm{Im}(\phi)}$. If $\phi(x) \ll 1$ then the exponential can be expanded, giving $A(x) \simeq A_0 \times (1 + \phi(x))$, and in this case the phase $\phi(x)$ itself becomes the function that is expanded in terms of cos and sin basis functions. In this chapter we are dealing with wavefronts that are nearly perfect, e.g., $A(x) \simeq 1$, but have small departures from perfection, and these departures are the cause of speckles. It is for this reason that we view the cos and sin functions as frozen "ripples" on an otherwise flat "ocean" of amplitude.

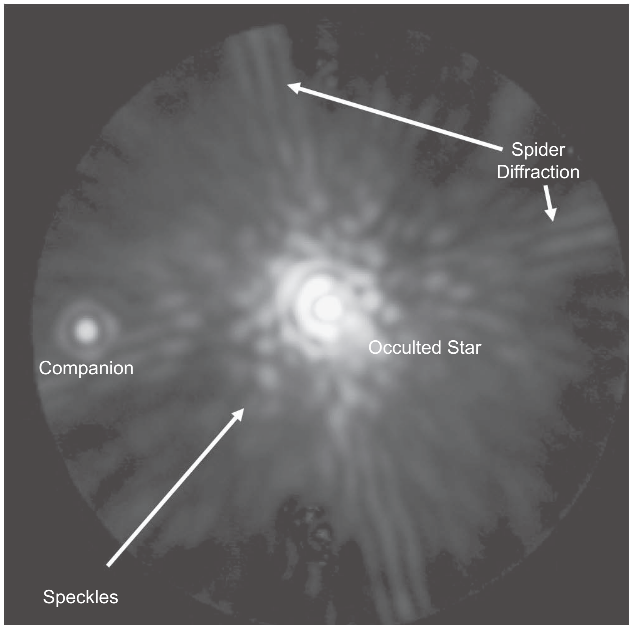


Fig. 20. An example of a classical Lyot-Coronagraph image of a nearby star in the near-IR. The star's PSF is created by occultation by the coronagraph's diffraction-limited rectangular focal plane mask, and a subsequent optimized pupil plane, or Lyot, stop. It shows the effect of spiders in the telescope and the resultant light propagated by a classical coronagraph. In addition, since this is a real observation, assisted by adaptive optics, the speckles have been frozen and are clearly visible. This star has a companion orbiting it, providing a PSF that is not occulted superimposed on the primary star's diffraction and speckle pattern.

Both methods can be easily extended to two-dimensional functions A(x, y), where, for example, A could be the wavefront across a two-dimensional pupil. In this case the ripples are two-dimensional, and can be visualized as a set of corrugated surfaces having spatial frequencies from one wave per diameter up to many waves per diameter, and with the corrugations arranged at all possible azimuthal orientations.

Since diffraction operates linearly on the electric field, it operates on each of these ripples independently, and we can sum the resulting amplitudes. We show in this section that the analysis of a single, generalized ripple across the wavefront provides deep insight into the origin of speckles, as well as clues as to how to reduce them in practice.

Recall that $A_1(x_1) = e^{-i\phi_1(x_1)}$ represents the amplitude of the electric field in our one-dimensional pupil, with range x_1 = (–D/2,+D/2). By Fourier analysis, the phase $\phi_1(x_1)$ can be written as the sum of potentially many sinusoidal ripples. Suppose a typical ripple has spatial period x_0, so that the phase of the wavefront can be written

$$\begin{aligned}\phi_1(x_1) = a\cos(2\pi x_1/x_0 + \alpha) + \\ ib\cos(2\pi x_1/x_0 + \beta)\end{aligned} \quad (108)$$

with units of radians. If the peaks and valleys of the ripple have values $\pm h_0$ (cm), then the corresponding amplitude of phase delay is

$$a = 2\pi h_0/\lambda \quad (109)$$

If the ripple also represents the patchy nature of scintillation or absorption across the wavefront, from a dark spot on the mirror, for example, then the transmitted field has an intensity ripple $e^{-2b\cos(2\pi x_1/x_0+\beta)}$; to see this, recall that a dark spot in a pupil can be represented by a real function, and is therefore expandable in a Fourier series, as we have shown above.

We find the amplitude in plane 2 to be

$$A_2(\theta) = \int_D e^{i\phi_1(x_1)}e^{i2\pi\theta x_1/\lambda}dx_1 \quad (110)$$

If we assume that the perturbation is small, $|\phi_1| \ll 1$, then we can expand the first exponential, giving

$$\begin{aligned}A_2(\theta) \simeq \int_D \Big[1 + i\Big(a\cos(2\pi x_1/x_0 + a) + \\ ib\cos(2\pi x_1/x_0 + \beta)\Big)\Big]e^{i2\pi\theta x_1/\lambda}\Big]dx_1\end{aligned} \quad (111)$$

Replacing cos(z) with $(e^{iz} + e^{-iz})/2$ allows us to integrate each term exactly. Then defining the well-known diffracted amplitude of a single star as

$$A_0(\theta) \equiv \frac{\sin(\pi\theta D/\lambda)}{\pi\theta D/\lambda} D \tag{112}$$

we find that the diffracted amplitude is the sum of a main peak, at the expected ($\theta_{main} = 0$) position of the star, plus two smaller peaks, one on each side, at $\theta_{speckle} = \pm\lambda/x_0$, where

$$\begin{aligned} A_2(\theta) = A_0(\theta) + \frac{1}{2}\left(iae^{i\alpha} - be^{i\beta}\right) A_0\left(\theta + \lambda/x_0\right) + \\ \frac{1}{2}\left(iae^{-i\alpha} - be^{-i\beta}\right) A_0\left(\theta - \lambda/x_0\right) \end{aligned} \tag{113}$$

The speckles are diffracted to either side, exactly as would be expected from the first orders of a diffraction grating with rulings spaced by the ripple period. So this is a very physically understandable picture. In a later section we will use this equation to show how a deformable mirror can null out one of these speckles, and indeed, a whole field of them.

The intensity pattern $I_2(\theta) = |A_2(\theta)|^2$ has six terms, and each is a function of θ

$$I_2 = I_0 + I_{(+1)} + I_{(-1)} + I_{(+2)} + I_{(-2)} + I_{(3)} \tag{114}$$

Here I_0 is the main peak, the central star image, defined above.

The next two terms are *speckles*

$$I_{(\pm1)} = \frac{1}{4}\left[a^2 + b^2 \pm 2ab\sin(\alpha - \beta)\right] I_0\left(\theta \pm \lambda/x_0\right) \tag{115}$$

These are the symmetrically placed speckles. Their intensities are equal if either phase errors or amplitude errors dominate, but if there is a mixture of these, the intensities can be unequal. Also note that these are exactly equal in shape to translated copies of the central peak, but scaled down.

The next two are *pinned speckles*

$$I_{(\pm2)} = \left[\mp a\sin(\alpha) - b\cos(\beta)\right]\sqrt{I_0(\theta)\, I_0\left(\theta \pm \lambda/x_0\right)} \tag{116}$$

These pinned speckles are located at the same place as the ordinary speckles, but they are scaled by the local intensity of the diffraction pattern from the main peak. In other words, they are effectively pinned to the preexisting diffraction rings, and they will show up as a locally enhanced or depressed intensity of the expected ring pattern, varying from point to point. As we will show below, they can be quite strong.

The last term, $I_{(3)}$, is negligible because is a crossproduct of the two speckles, so we drop it.

Here are some numerical examples. Suppose that a telescope in space has a small surface error, from perhaps a residual polishing tool imprint or a quilting from an eggcrate backing structure, so that the error is periodic. If the amplitude of this phase perturbations is $h_0 \simeq \lambda_0/100$, then the phase amplitude is $a = 2\pi/100 \simeq 0.06$ rad. The type 1 speckles will then have an intensity of about $I_{(\pm1)}/I_0 = a^2/4 \simeq 0.1\%$ times that of the star. The type 2 (pinned) speckle intensity will depend on the preexisting diffraction pattern; for example, if the Airy rings have an intensity that is 0.1% times the main star (around the fourth Airy ring), then the pinned speckles intensity will be about $I_{(\pm2)}/I_0 = 0.2\%$ times the star, i.e., brighter than the Airy ring itself. In addition, suppose that the reflectivity of the mirror is somewhat patchy, and can be represented by a reflectivity that varies from peak to average by about 1%, giving $b = -(1/2)\ln(0.99) \simeq 0.005$. Then the intensity of the type 1 speckles will be about $ab/2 \simeq 0.01\%$ of the main peak if both a phase and amplitude wave are present, or $b^2/4 \simeq 0.001\%$ if the amplitude ripple is present alone. If the spatial period of either of these ripples is $x_0 \simeq 100$ cm then the speckles will fall at an angle of about ±0.10 arcsec for visible wavelengths, roughly the location of Earth at 10 pc.

4.3. Speckles from Mirror Surfaces

Real mirrors have surface shape errors that get smaller in proportion to the size of region being considered. Thus the surface errors are not a white noise process, even though, for simplicity, we sometimes make that assumption, as we will do in deriving equations (120)–(125).

A common way to describe the shape error of a mirror surface, which leads directly to phase errors in the reflected wavefront, is through the *power spectral density* function PSD(k). Here k is the spatial frequency of a wave of wavelength Λ on the surface of the mirror. Recall from Fourier analysis that the sum of all such possible wavelengths can reproduce any arbitrary mirror shape error. We have that

$$k(\text{waves per cm}) = \frac{1}{\Lambda(\text{cm})} \tag{117}$$

Note that k is a two-dimensional quantity, e.g., $k = (k_x, k_y)$, which we will keep in mind as needed.

It is an empirical observation that large mirrors tend to have similar PSD functions, of the form

$$\text{PSD}(k) = \frac{A}{1 + (k/k_0)^n} \tag{118}$$

where $n \simeq 3.0 \pm 0.1$. This form applies to the residual shape of the mirror after low-order terms (typically focus, tip, tilt, coma, and astigmatism) have been subtracted, the reason being that these terms can be largely compensated by better alignment of a primary and secondary mirror system, and are somewhat independent of the intrinsic polished surface. Also, the low-order terms correspond to speckles within roughly 3λ/D, which could be blocked by a coronagraph, but the higher-order terms will contribute directly to speckles at larger angles, where we may expect to be imaging exoplanets.

The rms error of the surface is given by the area under the PSD curve, between specified spatial frequency limits, e.g., the same limits as given by the inner and outer range of a dark hole. We have

$$\begin{aligned} h^2_{rms} &= \iint PSD\left(k_x, k_y\right) dk_x \, dk_y \\ &= \iint PSD(k) 2\pi \, k \, dk \end{aligned} \tag{119}$$

The integral can be done analytically or numerically.

An analysis of the residual errors of the HST (2.4-m primary, plus secondary, combined), the Magellan (6.5 m), and a 1.5-m mirror shows that these all have roughly the same PSD shape and value, although it is also clear that more modern mirrors, polished with larger, shape-controlled lapping tools, will have smaller high-spatial-frequency errors.

Numerical values for HST are A = 6000 nm^2 cm^2, and k_0 = 0.040 cycles/cm. Integrating this over the range 0.02 to 0.28 cm^{-1}, corresponding to Λ = 3.6–50-cm wavelengths, we find h_{rms} = 7.5 nm. Thus the HST mirror surface has a $\lambda/84$ surface at the usual laser wavelength of λ = 633 nm.

References include *Hull et al.* (2003) and *Borde and Traub* (2006).

4.4. Speckles from the Atmosphere

For a groundbased telescope, observing a star through the atmosphere at visible wavelengths, the incident wavefront is typically distorted on all scales from kilometers to millimeters, with a PSD (see section 4.3) fall-off given by $n = 11/3 \simeq 3.7$, i.e., slightly faster than mirror-polishing errors. The bulk of the distortions take place on scales that lie between the *outer scale* L_0, typically 20 m, and the *inner scale* l_0, typically 0.4 cm.

The length r_0 is the diameter of a region over which the wavefront has an rms variation of about 1 rad (technically, 1.015 rad, by definition in the Kolmogorov model of turbulence). The value of r_0 is about 10 cm in the visible at a typical observatory, and scales as $\lambda^{6/5}$, so it is larger in the infrared. In the "frozen atmosphere" approximation, these patches of density fluctuations are carried along by the wind, so a typical timescale for wavefront change is $\tau_0 \simeq 0.31 \, r_0/V$ where V is a typical wind speed in the overlying atmosphere; if V = 10 m s^{-1} then τ_0 is about 3 ms. A groundbased image of a star therefore is made up of approximately $(D/r_0)^2$ speckles, churning on a timescale of τ_0, and spread over an angular diameter on the sky of about λ/r_0 or 1 arcsec in the visible, independent of telescope diameter.

The validity of the frozen flow hypothesis has been experimentally investigated by *Poyneer et al.* (2009). They find that speckles at groundbased telescopes can be modeled in terms of about one to three layers in the atmosphere, for over 70% of the time, with a median wind speed of about 10 m s^{-1}, but with a large range from 3 to 40 m s^{-1}. These observations also suggest that frozen flow accounts for about 30(±10)% of the total power that can be controlled by a deformable mirror (DM). These findings, together with the observation that the wind speeds vary by less than 0.5 m s^{-1} over 10-s intervals, suggests that about one-third of the speckle power could be reduced by an predictive algorithm.

References include *Poyneer et al.* (2009).

4.5. Speckle Suppression

Armed with a knowledge of how speckles are formed, we are now prepared to measure and suppress them. The following sections explain wavefront sensing, and how deformable mirrors are used to suppress speckles, singly as well as wholesale, arising from phase and amplitude variations across a pupil. The Talbot effect is examined to show how to suppress speckles from planes before or after the pupil plane. Some of the methods used today at groundbased and space telescopes are discussed, including ADI, SSDI, chromatic speckle suppression, and dual-mode polarimetric imaging. We conclude with a comparison of ground vs. space for direct imaging of exoplanets.

4.6. Wavefront Sensing and Control

At groundbased telescopes the atmosphere drives the incoming wavefront to an rms spatial variation of more than a wavelength, and in addition the wavefront varies rapidly in time, as discussed above. A *wavefront sensor* (WFS) is any device that allows us to measure the wavefront. The term *wavefront sensing and correction* (WFSC) applies when closed-loop corrections are applied.

The six main sources of uncertainty in a WFS are photon noise, chromaticity, aliasing, time delay, scintillation, and non-common-path errors.

Photon noise is obviously fundamental; it can be minimized by using a bright star and large values of r_0 and τ_0 on the ground, or their equivalent in space (e.g., from polishing errors and thermal drift).

Chromaticity arises from WFS systems in which the sensing wavelength is different from the science wavelength, and the wavefront errors are wavelength dependent; it can be minimized by using the same wavelength band for both purposes.

Aliasing arises when the WFS is sensitive to, but cannot distinguish between, a spatial frequency mode of the wavefront and an odd harmonic. For example, a pupil-based wavefront sensor with contiguous detecting elements, each of width w, is sensitive to a spatial wavelengths of size 2w as well as 2w/3.

Time delay occurs because a detector must integrate a signal for a finite length of time before it can be read out, and in addition the servo system has a finite bandwidth, which effectively adds more time delay to the correction signal. The choice of integration time depends on photon rate, the desired signal to noise ratio, and detector read noise.

Scintillation arises through the Talbot effect (section 4.11), in which wavefront phase perturbations generated in the upper atmosphere become intensity perturbations in the lower atmosphere. This causes *shadow bands* to flicker across a telescope pupil, which in turn will generate speckles in the image plane (see section 4.2).

Non-common-path errors arise when the WFS and the science focal plane are separated. For example, it is common for the WFS in a groundbased telescope to be fed by a *beamsplitter* that taps off part of the starlight and sends it to a sensor via an optical path that is different from the optical path to the focal plane. This method assumes that the WFS optics are perfect (or more generally, identical to the science optics), and do not add any wavefront ripples or speckles to the WFS system. Since we always use imperfect optics, this scheme is bound to fail at the level of the quality of the optics.

In principle, and in practice, it is much safer to detect speckles in the science focal plane, because this will avoid all six sources of uncertainty listed above. The decision as to how and where to put a WFS is determined by a combination of practical aspects of a given telescope, the desired level of wavefront correction, and the personal taste of the experimenter.

The Shack-Hartmann WFS is used at many telescopes. It is a simple system, easy to understand, but also probably the least effective. In this system a beamsplitter taps off part of the light, and a lens forms an image of the pupil. This pupil plane is filled with a large number of small lenses, sometimes formed by embossing a sheet of plastic. Each lenslet forms an image of the star onto a position-sensitive detector, e.g., multiple quad cells or a single CCD. A local tilt of the wavefront will produce a shift in the star image. Thus image position measurements can be converted to local wavefront slopes, and by patching together the slopes a full wavefront snapshot can be obtained. This information can then be used to drive a DM, and a closed-loop control established. The Shack-Hartmann WFS is sensitive to aliasing, because it uses a finite number of spatial sensing cells in the pupil plane, as discussed above.

A better WFS would be one that uses some of the bright starlight to interfere with the speckles. This is done in some systems by tapping off the bright star image, passing it through a spatial filter to make it nearly single-mode and therefore a smooth reference wavefront (using a small hole or ideally a single mode fiber), and interfering this with the speckle pattern; unfortunately, this kind of system is susceptible to non-common-path errors, rendering it less than ideal.

An even better WFS would use the star to interfere directly with the speckles, avoiding extra optical paths. This can be done by using a DM to diffract light out of the main star image and onto existing speckles, but adjusting the DM so that the phases cancel. This method is used in sections 4.8, 4.9, and 4.10 below.

References include *Guyon* (2005).

4.7. Contrast in a Dark Hole

The area in the focal plane over which speckles can be suppressed by a DM is called the *dark hole.* If the DM is placed at an image of the primary mirror, and the wavefront errors are entirely due to hills and valleys of that mirror, then it is clear that the DM simply needs to advance or retard the reflected wavefront in order to flatten it.

Suppose that a circular primary mirror, the pupil, is mapped onto a square DM, just filling it. The DM will be controlled to deform in such a way that a bumpy incident wavefront is reflected as a smooth wavefront, to the limit of control of the mirror. A typical DM has a thin glass facesheet, backed by an N × N square array of actuators that push or pull on the facesheet perpendicular to its surface.

In one dimension, it takes a string of four actuators (not two, as is often assumed) to approximate the shape of a single period of a sine or cosine wave, but if there are many such periods in a wave then we can get by with only about two actuators per period. This approximation breaks down in the case of a single period, because two actuators can approximate a cosine, but not a sine, wave, or the reverse, depending on where the actuators are located with respect to the wave.

Thus we can fit up to about N/2 periods of a wave with N actuators in one dimension. Also, there are N/2 whole waves, or modes, that can be approximated by a string of this length, i.e., 1 wave per diameter, 2 waves per diameter, . . . N/2 waves per diameter. By analogy, we assume that there are M modes in the full circular area of the pupil, where

$$M = \frac{\pi}{4}\left(\frac{N}{2}\right)^2 \tag{120}$$

Suppose that the average mode has an amplitude h_0, so that the speckle produced by this mode has relative intensity

$$I(\text{typ. speckle})/I_{\text{star}} \simeq a^2/4 = (\pi h_0/\lambda)^2 \tag{121}$$

where we used equations (109) and (115).

At each point in the pupil the net amplitude will be the sum of M complex vectors of average length h_0 but with random phases. This is exactly the random walk problem in two dimensions. Therefore the expected average amplitude will be $h_{rms} \simeq M^{1/2}h_0$, or

$$h_{rms} = \frac{\sqrt{\pi}Nh_0}{4} \tag{122}$$

So, from equation (121), writing the average contrast as

$$C \equiv \frac{I(\text{ave. speckle})}{I(\text{star})} = \left(\frac{\pi h_0}{\lambda}\right)^2 \tag{123}$$

we find

$$C = \pi\left(\frac{4h_{rms}}{N\lambda}\right)^2 \tag{124}$$

This says that to achieve a contrast C, with an otherwise perfect coronagraph, we need to control the N × N element DM with an accuracy such that the reflected wavefront has an RMS error of h_{rms} or better. The DM must be controlled to a surface error of $h_{rms}/2$, of course.

Inverting this relation, we can say that to achieve an average speckle contrast C we need to control the wavefront to a relative accuracy of

$$h_{\mathrm{rms}} = \frac{N\lambda\sqrt{C}}{4\sqrt{\pi}} \tag{125}$$

For example, if we desire C = 10^{-10}, and we have N = 64, then h_{rms} must be about λ/10,000. Thus the wavefront must be 100 times better than the typically "excellent" λ/100 wavefront.

In the visible, this means that the DM must control the reflected wavefront to an accuracy of $h_{\mathrm{rms}} \simeq 0.5$ Å. This may seem impossible, given that this is about half the radius of a Si or O atom; however, a typical (0.1 to 1.0 mm) DM element averages over many such atoms, and it is the average surface that counts here. In addition, we know from experiment (e.g., *Trauger and Traub,* 2007) that this is perfectly feasible.

It is useful to look at the DM as being a scattering grating device that can be commanded to generate a surface ripple that can diffract starlight to a specific target point in the focal plane. The phase of the controlled scatter can be adjusted by shifting the wave pattern on the DM from sine to cosine, for example. Thus the DM can be used to direct starlight to points in the focal plane, with the desired amplitude and phase so as to cancel starlight that arrived by other means. The DM is thus an extremely powerful device to have in the starlight beam, and it is required in all advanced adaptive optics systems. Note that this only works for light from the star itself; light from an exoplanet is not coherent with starlight, so the DM cannot use starlight to cancel an exoplanet, a fact we will use in section 4.10.

The angular radius of the dark hole is the maximum angle to which the high-frequency spatial period of the DM can scatter light. This angle is λ divided by two pistons of the DM, or D/(N/2), giving the angular radius as

$$\theta(\text{dark hole}) = \pm\frac{N\lambda}{2D} \tag{126}$$

Thus the maximum size of the dark hole is a square of angular size Nλ/D, which is a length of N resolution elements of the pupil. This square is centered on the star.

4.8. Single-Speckle Nulling

We show in this section that it is possible to make a speckle vanish by putting an appropriate pattern on an upstream DM. The method applies to a speckle that is caused by either a phase or amplitude perturbation. The method works for a speckle on one side of a star image, not both sides. To make many speckles vanish, see sections 4.9 and 4.10. To make speckles on both sides vanish, using two DMs, see section 4.11 on the Talbot effect.

Let us start with a simple case in which the one-dimensional pupil has a single ripple across it, formed by either a phase perturbation or an amplitude perturbation (or both). Recall that arbitrary shapes of such perturbations can be represented by a sum of cosine and sine functions, and that these can be visualized as the basis functions of phase ripples as well as absorption ripples. We assume that a coronagraph is present, so that it suppresses the central star and its diffraction pattern. Thus the telescope diffraction pattern is eliminated, and only scattered light from an imperfect wavefront remains. This is the case that was discussed in section 4.2. For that case we assumed that the pupil had a ripple given by

$$\begin{aligned}\phi_{\mathrm{pupil}}(x_1) = a\cos(2\pi x_1/x_0 + \alpha) + \\ ib\cos(2\pi x_1/x_0 + \beta)\end{aligned} \tag{127}$$

and we saw that the resulting amplitude in the image plane was

$$\begin{aligned}A(\theta) = A_0(\theta) + \frac{1}{2}\left(iae^{i\alpha} - be^{i\beta}\right)A_0(\theta + \lambda/x_0) + \\ \frac{1}{2}\left(iae^{-i\alpha} - be^{-i\beta}\right)A_0(\theta - \lambda/x_0)\end{aligned} \tag{128}$$

Suppose that we know the values of (a, α, b, β) (we will show how to estimate these later), and we wish to add a ripple to the DM to counteract the existing pupil ripple. If we add a ripple $-a\cos(2\pi x/x_0 + \alpha)$ we will clearly eliminate the a terms in the speckles, but the b terms will still remain; clearly, this is not sufficient.

However, suppose that we decide to eliminate (or null) the speckle at $\theta = \pm\lambda/x_0$, i.e., either the $A_0(\theta-\lambda/x_0)$ or the $A_0(\theta + \lambda/x_0)$ speckle. Let us add a ripple to the DM given by

$$\phi_{\mathrm{DM}}(x_1) = a'\cos(2\pi x_1/x_0 + \alpha') \tag{129}$$

Adding this to the existing ripple gives a net speckle amplitude that we set equal to zero

$$iae^{\pm i\alpha} + ia'e^{\pm i\alpha'} - be^{\pm i\beta} = 0 \tag{130}$$

We solve this by setting the real and imaginary parts each equal to zero. The solution for suppressing either of these speckles is

$$a' = -\sqrt{a^2 \pm 2ab\sin(\alpha - \beta) + b^2} \tag{131}$$

$$\tan(\alpha') = \frac{a\sin(\alpha) \pm b\cos(\beta)}{a\cos(\alpha) \mp b\sin(\beta)} \tag{132}$$

This shows that it is possible to suppress a speckle that arises from phase or absorption, or both together, using only the phase ripple of a DM, but only on one side of the star. Speckles originating from phase can be canceled on both sides of the star. Speckles originating from amplitude can be canceled on either one side of the star or the other side, but not both at the same time, since the speckle on the noncanceled side will get larger as the target speckle gets smaller.

For example, if there is no absorption, then b = 0, and we get $a' = -a$, and $\alpha' = \alpha$, which is physically logical. Likewise, if there is pure absorption, then a = 0, and we get $a' = -b$, and $\alpha' = \beta - \pi/2$.

4.9. Multispeckle Nulling

It is tedious to null speckles one by one. Not only are there a lot of speckles to null, but they are all coherent so that the intensity at a given point depends on that at neighboring points. In other words, there is coupling between the speckles, owing to their wings.

In this section we show, in principle, how to estimate the parameters of many speckles simultaneously (as was assumed for a single ripple in section 4.8). With this knowledge the DM can be set to null all speckles on one side of the star, using a method similar to that above. Here again, we assume a perfect coronagraph, but an imperfect residual wavefront.

Suppose that there are multiple speckles present in the pupil, perhaps from an imperfect primary mirror. A DM can null spatial wavelengths as short as two DM actuators in the pupil, so there are up to N/2 waves that need to be measured, in our usual one-dimensional case. (The two-dimensional case is similar.) Let us write the ripple in the pupil as

$$\phi_{\text{pupil}}(x_1) = \sum_{n=1}^{N/2} \left[a_n \cos(2\pi n x_1/D + \alpha_n) + i b_n \cos(2\pi n x_1/D + \beta_n)\right] \tag{133}$$

We measure the intensity of this (unknown) starting case at each of N points in the focal plane.

This method requires an extra pupil plane in addition to those shown in Fig. 12, so for this discussion let us assume that plane 1 is an image of the original telescope pupil.

If we now add a new, independent set of ripples (a_n', α_n') to the DM, the added phase will be

$$\phi_{\text{DM1}}(x_1) = \sum_{n=1}^{N/2} \left[a_n' \cos\left(2\pi n x_1 \big/ D + \alpha_n'\right)\right] \tag{134}$$

We measure the intensity in the focal plane at N points for this case as well.

We then add yet a different set of ripples

$$\phi_{\text{DM2}}(x_1) = \sum_{n=1}^{N/2} \left[a_n'' \cos\left(2\pi n x_1 \big/ D + \alpha_n''\right)\right] \tag{135}$$

and measure these intensities as well.

In both cases the parameters of the two added sets of ripples can be anything convenient, e.g., totally random ripples or a sharp delta function created by a single actuator. The main point is that the added ripples should be significantly different from the original set.

We now have three sets of intensities in the focal plane, on N pixels, therefore 3N data points. The number of unknowns (a_n, α_n, b_n, β_n) for N/2 waves is a total of 2N. Therefore we can use our 3N measurements to solve for 2N parameters.

The reason we need more measurements than parameters is because the intensity is the square of the amplitude, and therefore there are sign ambiguities that need to be resolved. If desired, yet another set of ripples and observations can be made, and a least-squares or singular-value decomposition solution found to the overdetermined set.

Once all the parameters of the original ripples have been measured, then the DM can be set to counteract them, on one side of the star, assuming that both phase and amplitude errors exist.

4.10. Speckle Energy Minimization

In practice, the multispeckle nulling method sketched above has several limitations: (1) intrinsic noise in the system, from photon noise as well as measurement noise, which limits our ability to perfectly measure the intensity pattern; (2) imperfect knowledge of the DM's response to an applied voltage; (3) higher-frequency ripples in the pupil, which can alias down into the dark hole; (4) the pupil phase drifting with time, during a several-minute measurement, from thermal expansion; (5) the need to observe a star over a finite range of wavelengths, e.g., a 10% or 20% band, but the solutions given above are only valid for a single wavelength; (6) the fact that both a sine and cosine ripple in the DM can only be generated out to an angle of half the radius of the dark hole, using the strict rule that four actuators are needed per wave.

Here is the basic idea of energy minimization. We go back to basics for a few steps, to clarify what we are doing. Suppose the amplitude at the input pupil is

$$A_1(x_1) = e^{i\phi_1(x_1)} \simeq 1 + i\phi_1(x_1) \tag{136}$$

where ϕ_1 can be complex. Expanding in a Fourier series we have a useful representation as

$$\phi_1(x_1) = \sum_{n=1}^{\infty} a_n \cos(2\pi n x_1/D + \alpha_n) + i\sum_{n=1}^{\infty} b_n \cos(2\pi n x_1/D + \beta_n) \tag{137}$$

Suppose there is a DM immediately after this pupil, with mask function

$$M_1(x_1) = e^{i\phi_{\text{DM}}(x_1)} \simeq 1 + i\phi_{\text{DM}}(x_1) \tag{138}$$

and where we expand in a finite Fourier series

$$\phi_{\text{DM}}(x_1) = \sum_{n=1}^{N/2} a'_n \cos(2\pi n x_1/D + \alpha'_n) \tag{139}$$

Then in plane 2, at the focus of the star, the amplitude $A_2(\theta)$ is

$$A_2(\theta) = \int_D M_1(x_1)\, A_1(x_1)\, e^{i2\pi x_1 \theta/\lambda} dx_1 \qquad (140)$$

Inserting the expressions for the phases, and keeping only terms to the first order, we get

$$A_2(\theta) = \int_D \left[1 + \phi_1(x_1) + \phi_{DM}(x_1)\right] e^{i2\pi x_1 \theta/\lambda} dx_1 \qquad (141)$$

Carrying this through will give $A_2(\theta)$ of a single star, as in equation (44), the diffraction pattern of the pupil (Airy rings), and the speckles from both the pupil and DM. Suppose that this is followed by a perfect coronagraph, which in essence allows us to delete the "1" from this expression, and to write the speckle amplitude as

$$\begin{aligned} A_{2,spec}(\theta) = \int_D \Bigg[& \sum_{n=1}^{\infty} a_n \cos(2\pi n x_1/D + \alpha_n) + \\ & i\sum_{n=1}^{\infty} b_n \cos(2\pi n x_1/D + \beta_n) + \\ & \sum_{n=1}^{n/2} a'_n \cos(2\pi n x_1/D + \alpha'_n) \Bigg] e^{i2\pi x_1 \theta/\lambda} dx_1 \end{aligned} \qquad (142)$$

Suppose that we wish to calculate the total speckle energy in the range from θ_{min} to θ_{max}, where these might be a target dark hole, i.e., a few times λ/D to $N/2$ times λ/D, for example. This total energy will be E_{spec} where

$$E_{spec} = \int_{\theta_{min}}^{\theta_{max}} \left|A_{2,spec}(\theta)\right|^2 d\theta \qquad (143)$$

As an extreme example, if we decide to calculate the total energy in the entire focal plane, i.e., θ from $-\infty$ to $+\infty$, after some work we find

$$\begin{aligned} E_{spec} = \Bigg[& \sum_{n=1}^{\infty} a_n^2 + 2\sum_{n=1}^{N/2} a_n a'_n \cos(\alpha_n - \alpha'_n) + \\ & \sum_{n=1}^{\infty} b_n^2 + \sum_{n=1}^{N/2} (a'_n)^2 \Bigg] D/2 \end{aligned} \qquad (144)$$

If we minimize this with respect to the parameters of the DM, we find that we need to set the DM as

$$\begin{aligned} a'_n &= -a_n \\ \alpha''_n &= \alpha_n \end{aligned} \qquad (145)$$

which cancels the wavefront distortion, up to the highest frequency allowed by the DM, but does not affect the absorption part of the wavefront. This is especially clear when we write the value of the minimum energy, which is then

$$E_{spec}(min) = \left[\sum_{n=N/2+1}^{\infty} a_n^2 + \sum_{n=1}^{\infty} b_n^2 \right] D/2 \qquad (146)$$

Here the first term is the sum of the power in all the high-frequency errors in the pupil, and the second term is all the absorption error terms, none of which are canceled.

A more useful exercise would be to calculate the total energy in the dark hole on one side of the star, as suggested above, or some other selected area. This would then use the DM to cancel both the delay and absorption terms in a finite window, as shown for a single speckle in equation (132). This is a work in progress.

References include *Borde and Traub* (2006) for the energy minimization method, and *Give'on et al.* (2006) for the electric field conjugation method.

4.11. Exploiting the Talbot Effect

The *Talbot effect* says that if a plane wave is incident on an infinitely wide periodic mask, then the transmitted wave will be periodically replicated downstream. At a distance midway between these replications, a pure phase disturbance will become a pure amplitude disturbance, and vice versa. This seemingly strange curiosity has important consequences in several areas.

Atmospheric scintillation is a familiar example. Suppose that an incident wavefront from a star passes through some turbulence near the boundary between the troposphere and the stratosphere. Breaking up the turbulence into Fourier components, we see that each component will impart a sinusoidal phase ripple onto the incident wave. At a distance Δz lower in the atmosphere (see equation (157)), the phase ripple will become an intensity ripple, or stellar scintillation.

The *Talbot carpet* is another example. Suppose that a plane wave is incident on a one-dimensional mask at $z = 0$ that has many small holes spaced by Λ in the x direction. Then, if we go to a downstream value of z, and ask what the condition is for wave transmitted by every n_Λ-th hole to have a path length that is the minimum distance plus n_λ wavelengths (in other words, an interference-created intensity maximum), from a right-triangle construction, we have the relation

$$z^2 + (n_\Lambda \Lambda)^2 = (z + n_\lambda \lambda)^2 \qquad (147)$$

Assuming that $\lambda \ll \Lambda$ we find

$$z = \frac{(n_\Lambda)^2}{n_\lambda} \frac{\Lambda^2}{2\lambda} \qquad (148)$$

where n_Λ and n_λ can be 1, 2, 3, . . . etc. There will be infinitely many such planes lying between $z = 0$ and $z = z_{TC}$, where

$$z_{TC} \equiv \frac{\Lambda^2}{2\lambda} \qquad (149)$$

This pattern of bright points is called the Talbot carpet.

We now ask, can we use this effect to compensate for intensity fluctuations in a pupil, by correcting the resulting downstream phase with a DM? The answer will be yes, but with a caveat. Let us start with an incident plane wave $A_1(x_1) = 1$ in the pupil, and a phase ripple imposed by a mask

$$M_1(x_1) = a\cos(2\pi x_1/\Lambda) \tag{150}$$

We now ask, what is the amplitude in a plane $A_2(x_2)$ downstream a distance z? Note there is no lens at plane 1, since we are allowing the wave to propagate freely in space, i.e., continuing on as a free wave, without a focusing lens. We sum up the wavelet contributions from plane 1, as we did for the Talbot carpet.

$$\begin{aligned} A_2(x_2,z) &= \sum(\text{wavelets}) \\ &= \int_{-\infty}^{+\infty} M_1(x_1)A_1(x_1)e^{i2\pi l/\lambda}\frac{1}{\sqrt{1}}dx_1 \end{aligned} \tag{151}$$

where l is the distance from x_1 to x_2

$$\begin{aligned} l &= \sqrt{z^2 + (x_1 - x_2)^2} \\ &\simeq z + \frac{(x_1 - x_2)^2}{2z} \end{aligned} \tag{152}$$

We use $1/\sqrt{l}$ in the integrand to account for the diminished wavelet amplitude with distance, to conserve energy in our two-dimensional space; since amplitude is less important than phase in this integral, we also use $l \simeq z$.

Now, if the phase ripple is weak, i.e., $a \ll 1$, we can expand $e^X \simeq 1 + X$, and use $e^{iX} = \cos X + i \sin X$, and proceed with the integration. We will need the Fresnel cosine integral, defined as

$$C(X) = \int_0^X \cos\left(\frac{\pi}{2}t^2\right)dt \tag{153}$$

where $C(\infty) = 1/2$.

After a bit of work, we find the amplitude in plane 2 to be

$$A_2(x_2,z) = e^{i2\pi z/\lambda}\sqrt{\frac{\lambda}{2}}\left[1 + ia\cos\left(\frac{2\pi x_2}{\Lambda}\right)e^{-i\pi\lambda z/\Lambda^2}\right] \tag{154}$$

The first factor is the familiar plane wave in the z direction. The second factor $(\sqrt{\lambda/2})$ is an artifact of our method of integration and can be ignored. The third term (in brackets) is the same as the input wave except for the additive periodic term in z. The corresponding intensity is

$$I_2(x_2,z) = \frac{\lambda}{2}\left[1 + 2a\cos\left(\frac{2\pi x_2}{\Lambda}\right)\sin\left(\frac{2\pi z}{z_T}\right)\right] \tag{155}$$

where the *Talbot distance* z_T is defined as

$$z_T = \frac{2\Lambda^2}{\lambda} \tag{156}$$

(Confusingly, z_{TC} differs from z_T by a factor of 4, a result of the fact that the former is generated by point sources, and the latter by a continuous wave source.)

We see that the wave that emerges from plane 1 reproduces itself at multiples of z_T, and that at the halfpoint between these reproductions an intensity pattern appears that also reproduces itself. We will have alternating planes of constant intensity (but varying phase), and varying intensity (but constant phase). The distance between adjacent planes of varying and constant intensity is Δz where

$$\Delta z = \frac{z_T}{4} = \frac{\Lambda^2}{2\lambda} \tag{157}$$

So if we go a distance Δz downstream from a pupil image, with no intervening optics, the intensity ripples will become phase ripples. Unfortunately, this distance is chromatic, i.e., it depends on wavelength. But at least we can see a method here to begin to reduce intensity ripples in the pupil.

Here is a numerical example. Suppose that the optics train of a telescope includes a reduced-diameter pupil plane (plane 3a, say), followed by a length Δz with no focusing optics, so that the beam propagates in a nominally parallel fashion to plane 3b. Suppose that there are amplitude variations in plane 3a that can be approximated by ripples of period length Λ. Let us place a DM in plane 3b, where the amplitude ripples will have turned into phase ripples, which we can cancel with the DM. Let us assume that the pupil diameter in plane 3a is such that the period of the disturbance is approximately equal to two actuators of the DM. If each actuator is 1 mm wide, then $\Lambda \approx 2$ mm. If the wavelength is $\lambda \simeq 0.5$ µm, then the separation between 3a and 3b will need to be about $\Delta z \simeq 4$ m, a large but not totally unrealistic length.

4.12. Angular Difference Imaging

If the wavefront sensing occurs in a plane that is different from the science focal plane, and if the speckles from the atmosphere and telescope pupil have been reduced in the wavefront sensing plane, there often remains a residual wavefront deformation in the science plane owing to a *non-common-path* problem. These speckles can be substantial, and since they arise from the telescope optics themselves, they can persist for a long time, typically many minutes or more. Unfortunately, a telescope speckle has the same appearance as an exoplanet, at a given wavelength, so strong, persistent speckles can easily overwhelm a faint exoplanet image.

The *angular differential imaging* (ADI) technique can overcome internal speckles from the telescope by simply rotating the telescope about the line of sight, or at an alt-az telescope by allowing the rotating Earth to rotate the apparent sky (except on the celestial equator). Since the detector remains fixed with respect to the telescope, the non-common-path speckles also remain fixed on the detector. Thus, subtracting a rotated image from a nonrotated one should eliminate the fixed-pattern speckles, allowing the exoplanet to be seen. Another name for this technique is *roll deconvolution,* a method that has had success on the HST.

References include *Marois et al.* (2006), *Hinkley et al.* (2007), and *Artigau et al.* (2008).

4.13. Simultaneous Spectral Differential Imaging

The technique of *simultaneous spectral differential imaging* (SSDI) is based on the fact that speckles are located at an angular distance from the star in proportion to their wavelength. So if images are taken at two or more wavelengths, and they are radially scaled to a common wavelength, then the difference of images should cause the fixed-pattern speckles to drop out. If an exoplanet is in the field it will show up as a radially shifting positive and negative feature.

An additional leverage factor arises if the exoplanet has a strong absorption feature in its spectrum, different from its star. For example, the methane band at 1.7 μm is very deep on some gas giants. The fact that the planet is relatively faint in this band gives it an extra handle for detection.

References include *Racine et al.* (1999), *Marois et al.* (2005), and *Biller et al.* (2006).

4.14. Chromatic Speckle Suppression

As an extension of the SSDI technique, one can use much higher spectral resolution to achieve superior speckle suppression. In this case, a coronagraph is outfitted with a hyperspectral imaging device, also sometimes referred to as an *integral field spectrograph.* Images are obtained at tens to hundreds of wavelengths simultaneously, usually over a single astronomical bandpass. The data forms a cube, with two spatial and one spectral axes. Speckles follow diagonally radiating paths through these data cubes, while real celestial objects will remain at the same spatial separation from the primary star. As such, in principle one can distinguish one wavelength's speckle pattern from another one and effectively remove the speckles without damaging the signal from a bona fide celestial object. The data processing methods are fairly complicated, even though the initial studies of this technique suggested relatively simple solutions for data processing.

References include *Sparks and Ford* (2002).

4.15. Dual-Mode Polarimetric Imaging

Perhaps the most successful of all speckle suppression techniques to date, *dual-mode polarimetric imaging* exploits the fact that in general starlight is very weakly polarized. If one is attempting to image an object or material around a star that exhibits large fractional polarization, the starlight and speckles can be removed with almost arbitrary precision. Images are formed using a Wollaston prism, which sends light with perpendicular polarization vectors in slightly different directions. Two images can formed and sensed simultaneously in this manner. When they are subtracted, only light that is actually polarized remains in the image. If the starlight is not polarized, it will be completely removed. Speckles are formed from unpolarized starlight. This technique has been used very successfully to image disks of dust that polarize light through the scattering process.

References include *Kuhn et al.* (2001), *Perrin et al.* (2004), and *Oppenheimer et al.* (2008).

4.16. Ground Versus Space Direct Imaging

Is it possible to directly image an Earth around a nearby star with a groundbased telescope? Or is it necessary to put an Earth-imaging telescope in space, above the atmosphere? In this section, we show an approach to answering this question.

For a large groundbased telescope, the overlying atmosphere will distort the incident wavefront by several wavelengths. We need to detect this distortion and remove it by reflecting the wavefront from a DM. The distortion may arise from several levels in the atmosphere, but for present purposes let us assume that it arises at a single level, such that we can image that layer on a DM, and remove the phase distortion without suffering any additional error, such as amplitude nonuniformity. Let us also assume that we can do this operation essentially instantaneously, without any time lag due to the measurement interval or the servo system. All these assumptions will be broken in real life, so the current calculation is optimistic in the sense that the real result will always be worse.

Let us start by assuming that we can detect the wavefront error with a Shack-Hartmann device. Suppose that about half the light in the pupil is split off and sent to an array of lenslets, each with diameter r_0. Assume that the local slope of the wavefront is α radians, approximately constant across the patch r_0. Each lenslet will focus the star in an image that has angular size λ/r_0. If there are n detected electrons in that image, we will be able to locate its centroid with an angular accuracy of about $\Delta\alpha = \lambda/(r_0\sqrt{n})$ radian. The uncertainty in the local measured slope of the wavefront will also be $\Delta\alpha$. The uncertainty h_{rms} in the delay of the wavefront over this patch is that error times the width of the patch, so

$$h_{rms} = \Delta\alpha \times r_0 = \frac{\lambda}{\sqrt{n}} \tag{158}$$

Let us assume that we can correct the wavefront to within this uncertainty.

From equation (125) we see that the resulting wavefront error will generate speckles whose intensity is a factor of C fainter than the star itself, according to

$$h_{rms} = \frac{N_{DM}\lambda\sqrt{C}}{4\sqrt{\pi}} \tag{159}$$

where N_{DM} is the number of DM elements per diameter D of the telescope. In our case we want to have at least one DM element per r_0 segment, so $N_{DM} = D/r_0$. Collecting terms we find the number of electrons needed is

$$n = \frac{16\pi r_0^2}{CD^2} \tag{160}$$

For a star of magnitude m, the number of electrons that we can collect is limited the number of photons in a volume determined by r_0 and τ_0, which gives

$$n = f_\lambda(m)\Delta\lambda A\tau_0 QE \tag{161}$$

where f is the flux density from equation (9), $\Delta\lambda$ is the bandwidth, A is the collecting area, τ_0 is the collecting time, and QE is the quantum efficiency.

Let us take $\Delta\lambda = 0.20\lambda$, $A = \pi r_0^2/4$, and $QE = 0.5$, representing the half of the incident light that is split off for the wavefront sensor, and assuming a perfect detector. The integration time is generally given by $\tau_0 = 0.31\ r_0/V$ where $V \simeq 500$ cm s^{-1} is the assumed wind speed, so τ_0 corresponds to the time that it takes the wind to move a patch of air of size r_0 about one-third of the way past a similarly sized collecting lenslet. Collecting terms we get

$$n = 10^{a-0.4m}(0.2\lambda)\left(\frac{\pi}{4}r_0^2\right)\left(0.31\frac{r_0}{V}\right)\left(\frac{QE\times\lambda}{2\times10^{-12}}\right) \tag{162}$$

where n has units of electrons, λ is in μm, r_0 is in cm, and V is in cm s^{-1}.

Equating the two expressions for n, and using the fact that r_0 scales with wavelength as

$$r_0(\lambda) = r_0(\lambda_V)(\lambda/\lambda_V)^{6/5} \tag{163}$$

and inserting numerical values, we find that the star magnitude m must be at least

$$m = 2.5\left[a + 10.69 + 3.2\log(\lambda_{\mu m}) + \log(CD_m{}^2)\right] \tag{164}$$

where D_m is the telescope diameter in meters.

Evaluating equation (164) for 30-m and 100-m telescopes, for a contrast $C = 10^{-10}$ appropriate for the Earth/Sun system, and for the BV RI J H K_s bands, we find that the result is nearly independent of band, giving

$$\begin{aligned} m(10^{-10}, 30\text{ m}) &\simeq -4.2 \pm 0.2 \\ m(10^{-10}, 100\text{ m}) &\simeq -1.4 \pm 0.2 \end{aligned} \tag{165}$$

This tells us that there are *no* stars bright enough to drive a servo system to achieve a speckle contrast as small as the ratio of Earth to Sun brightness, in any of these spectral bands, and even for an essentially perfect servo system. The detection might just barely be possible for a 100-m telescope, but even so there would be only a handful of near-infrared targets available. The bottom line is that to detect an Earth, and to characterize it, we will need to go to space, simply to eliminate unavoidably bright speckles from the turbulent atmosphere.

One might ask if a bright nearby guide star could be used for Earth detections instead of the target star. The answer is probably no, because atmospheric turbulence differs enough between stars that this level of compensation would be impossible.

One might also ask if laser guide stars could be used for Earth detections. Here again the answer is no, this time because the brightest laser guide star that has ever been used has an equivalent stellar magnitude of about +5, which, from equation (164), is very far from being bright enough.

If the target is self-luminous young Jupiters, with contrasts around $C \geq 10^{-8}$, then the limiting magnitudes in equation (165) become brighter by about 5 mag, so $m(10^{-8}, 30\text{ m}) \simeq 0.8$, for which a handful of near-infrared target stars might provide sufficiently large signals. Thus large ground-based telescopes should be able to detect self-luminous young Jupiters down to contrasts approaching 10^{-8}.

5. CURRENT PROJECTS

In the previous sections we described the myriad techniques needed to image exoplanets directly. These techniques are employed in many ways, often in combination, in current observational projects. In Table 9 we list many of the currently operating and proposed projects around the world that have exoplanet imaging and spectroscopy as a major part of their scientific justification. Here we briefly compare and contrast these many experiments. Please note that it is impossible in a paper of this nature to include every project in operation or proposed. Instead of being exhaustive and unabridged, our goal here is to provide examples of the applications of the techniques described in detail above. The reader can find additional information about these projects through their associated references.

As an indication of the current status of direct imaging of exoplanets, we note a few recent achievements here. In the laboratory, no one has yet published a coronagraph contrast close to, or better than, the value of 6×10^{-10} in monochromatic light, from *Trauger and Traub* (2007), which is rather surprising; this area clearly needs more work. On the bright side, we do have the wonderful images of HR 8799 b,c,d, β Pic b, and Fomalhaut b, as described in the introduction to this chapter. The former triad of young planets was recently directly imaged in the near-infrared with a vector vortex coronagraph using a relatively small (1.5 m) telescope pupil (*Serabyn et al.,* 2010), at K-band contrasts as low as 2×10^{-5} and as close as $2\lambda/D$; this is certainly an encouraging step forward.

This section and the next (section 6) are meant to provide a snapshot of the state-of-the-field at the time of publication. Of course the future projects likely will change in some respects. From a very general perspective the observational field of comparative exoplanetary science via direct detection is in a nascent stage. We are just now on the verge of routinely observing such objects with both photometry and spectroscopy. Much in this field will change as observations reveal the advantages or each technique described previously.

TABLE 9. Current, planned, and proposed projects for direct detection and study of exoplanets.

Project Name	First Light (year)	Telescope Diameter (m)	Optimal Wavelength (μm)	AO System Elements	Starlight Suppression Technique	Speckle Suppression Technique
Keck Imaging	2002	10.0 Keck-II	1.0–2.5	249	None	ADI
Gemini NIRI/ Altair Imaging	2004	7.98	1.0–2.5	177	None	ADI, SDI
Lyot Project	2004	3.63	1.2, 1.6	941	Lyot Coronagraph	Polarimetry
VLT/NACO	2005	8.0	4.0	177	Lyot Coronagraph	SDI
HiCIAO	2007	8.2 Subaru	1.2, 1.6, 2.2	188[*]	Lyot Coronagraph	None
MMT/AO	2008	6.5	5.0		Lyot Coronagraph	ADI
NICI	2008	7.987 Gemini-N	1.2, 1.6, 2.2, 3.8, 4.7	85[*]	Lyot Coronagraph	SDI or ADI
Project 1640	2008 2011	5.07 Palomar	1.640 1.640	249 3217	APL Coronagraph APL Coronagraph	Chromatic Chromatic Science-Arm
LBTI	2011	2 × 8.1 22.3 eff.	3.0–20.0	2 × 349	Interferometric Nulling	N/A
Gemini Planet Imager	2011	7.798 Gemini-S	0.95, 1.2, 1.6, 2.2	1579	APL Coronagraph	Chromatic-Science-Arm Polarimetry
SPHERE	2011	8.20 VLT	1.2, 1.6, 2.2 0.6, 0.8, 0.9	1312	APL or Phase-Mask Coronagraph	Chromatic Polarimetry
JWST	2015	6.5	5-27	108[†]	Lyot or Phase-Mask Coronagraph	PSF Subtraction and Chromatic
Planetscope	2015	1.00 Balloon	0.5–1.0	2304	Band-limited Coron.	Science-Arm
30–42 m Telescope	2018	30.0	1.0–2.5	≥3000	APL Coronagraph or Nulling Coronagraph	Science-Arm Chromatic
Probe Missions[§]	2019[‡]	1.0–2.0	0.5–1.0	2304	Band-limited, PIAA, Star Shade	Science-Arm
TPF-C	2024	3.5 × 8.0 elliptical	0.5–1.0	2304	PIAA, Band-Limited or Phase Coronagraph	Science-Arm Chromatic
TPF-O	2024	4	0.7–1.0	≈500	Occulter	
DaVINCI	2024[¶]	4 × 2.5 Single Spacecraft	1.0–13.0	~1000	Interferometric Nulling	N/A
TPF-I/ DARWIN	2028[¶]	4 × 4 Sep. Spacecraft	5.0–20.0	N/A	Interferometric Nulling	N/A
Large Space Telescope	2035[¶]	4.0 to 16.0	0.3–20.0	≈4000	Numerous	Unknown

[*] The HiCIAO and NICI instruments use a curvature-based bimorph mirror so actuator number is not directly comparable to those of the other systems.

[†] The MIRI instrument has active optics, not adaptive optics, for the JWST primary mirror with 18 segments with 6 actuators per segment.

[‡] The various proposed space projects have no certain launch dates nor is it known at the time this book was printed which, if any, of these missions would actually be constructed and flown into space.

[§] There are several different proposed NASA probe-class missions involving star shade occulters or internal coronagraphs and relatively small aperture telescopes with modest AO systems.

[¶] These projects are highly speculative and are not very well defined at this point.

A summary of the current dynamic range or contrast actually achieved by existing observations can be found in Fig. 21 . The projects that have achieved these results are described below.

5.1. HST Coronagraphs

HST has three coronagraphs, one each on these instruments: (1) Advanced Camera for Surveys, High Resolution Camera (ACS-HRC, from 2002 to 2007, now permanently lost); (2) Near Infrared Camera and Multi-Object Spectrometer (NICMOS, from 1997 to 1999, and 2002 to 2008, with a possible restart in the future); and (3) Space Telescope Imaging Spectrograph (STIS, from 1997 to 2004, and 2009 to present).

All three coronagraphs are of the simplest possible type, an opaque top-hat blocker in an image plane, followed by a Lyot mask in a pupil plane, and a detector in a final image plane.

ACS-HRC: The ACS coronagraph is in the uncorrected, aberrated beam. A glass sheet can be inserted near the focus (at the point of least confusion) with a choice between two simple top-hat metalized spots of diameter 1.8 and 3.0 arcsec. This point is upstream of the correcting optics, so the beam is aberrated. The spot sizes were chosen to reduce the diffracted light intensity to be less than the speckle intensity. A thin metal Lyot mask is inserted downstream, just in front of the aberration correction mirror, i.e., close to a pupil plane image. The Lyot mask blocks light from around the edges of the primary, secondary, and spider arms, reducing the throughput to about 48% of its value without the mask. Filters are available. The final image in the HRC is on a CCD detector.

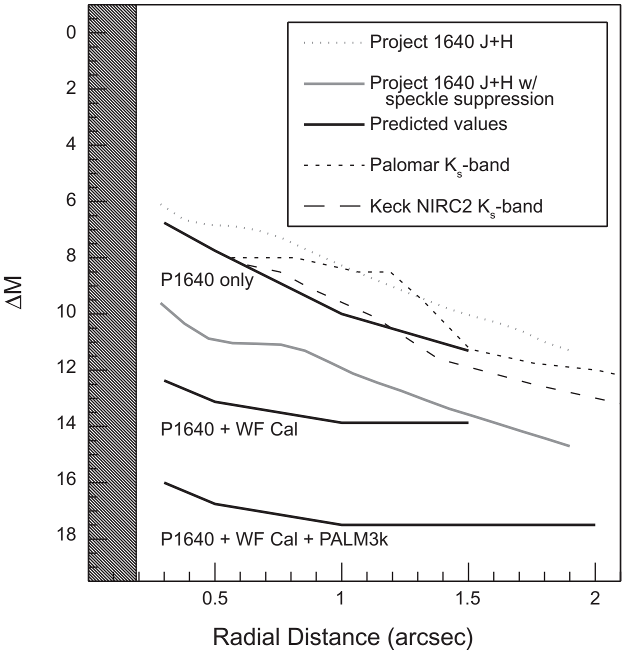


Fig. 21. Actual dynamic range or contrast achieved for existing systems. This is merely a selection of current achieved dynamic ranges, including Keck AO imaging, Palomar AO Lyot Coronagraph imaging, Project 1640 (with and without chromatic speckle suppression), Keck/Gemini ADI (e.g., *Marois et al.,* 2008), and expected performance of Project 1640 with the addition of a science-arm WFS and an extreme AO system (labeled WF Cal and PALM3k), and finally Lyot Project dual-imaging polarimetry. Dynamic range is expressed as the 5σ detection threshold for a point source as a function of radius and as a difference in magnitude with respect to the central star. Generally, ADI sees a gain of about 1 to 2 magnitudes over the speckle noise background, while dual-mode polarimetry can completely remove speckle noise at the level of 10^{-5} or better but only for detection of polarized objects. Image courtesy of S. Hinkley.

NICMOS: The NICMOS coronagraph is in the corrected beam. It comprises a 0.6-arcsec-diameter circular hole in a focal plane mirror, followed by a cryogenic Lyot mask of the primary, secondary, and spider diffraction in a pupil plane. The sky plane is reimaged on a near-infrared detector. The hole size was chosen to remove 93% of the encircled energy at H band, beyond which point the diffracted and scattered light profile flattens out.

STIS: The STIS coronagraph is in the corrected beam. The image-plane part comprises two orthogonal wedge-shaped blocking masks, where each wedge has a length of 50 arcsec and a width that ranges from 0.5 to 3.0 arcsec, and the wedges overlap at about their 1.25 arcsec width points. The pupil plane part has a Lyot mask of the outer edge of the primary mirror. The final image plane has a CCD detector. No filters are available, so the full 0.2 to 1.03 μm spectrum is imaged.

References include *Krist* (2004), and the HST Instrument Handbook descriptions of the ACS, NICMOS, and STIS coronagraphs, for which up-to-date versions are available on the web.

5.2. Keck and Gemini Imaging

A number of surveys have been conducted using adaptive optics and direct imaging without a coronagraph. In these cases either calibrator stars or an ADI or SDI technique is employed to reduce the starlight in the image. These techniques have been effective, particularly in the case of the star HR 8799 with its three companions that seem to be of planetary mass, one of which was also detected by HST/NICMOS. However, nondetections are in many respects equally important because they provide upper limits to the overall population of planets around nearby stars. For example, Gemini and its Altair AO system were used to conduct the "Gemini Deep Planet Survey" to search for planets at large separations (greater than ~1 arcsec) from their host stars. The survey used the ADI technique. No objects were found around 48 stars that were observed at least two times, and the conclusions are important: The fraction of stars with brown dwarf companions between 25 and 250 AU separations is between 2.2% and 0.4% at the 3σ confidence level. The upper limits on the fraction of stars with at least one planet of mass 0.5–13 M_{Jup} are 28% for the semimajor axis range of 10–25 AU, 13% for 25–50 AU, and 9.3% for 50–250 AU (also with a 3σ confidence).

References include *Marois et al.* (2008), *Janson et al.* (2010), *Herriot et al.* (2000), and *Lafreniere et al.* (2007, 2009).

5.3. Very Large Telescope NACO and Subaru HiCIAO

A number of surveys of similar nature were undertaken by other investigators using the VLT and Subaru telescopes. In these surveys a classical Lyot coronagraph is fitted behind an AO system. Typically these projects have much smaller fields of view than those described in the previous subsection, but they have more effective starlight suppression due to the presence of a coronagraph. The NACO instrument, which operates at 4 μm with a coronagraph and ADI starlight suppression, has been particularly effective in imaging objects that may be planets or brown dwarfs, not easily distinguishable when very young. The Subaru telescope is outfitted with a new AO coronagraphic imaging device called HiCIAO, which includes speckle suppression modes using dual-mode polarimetry and SDI. This is a very promising instrument currently in operation and producing images of objects in the brown dwarf mass regime (companions to GJ 758) but at temperatures that probably provide spectroscopic opportunities that will be very important for exoplanetary science. HiCIAO will likely find more objects similar to the companions of HR 8799.

References include *Lenzen et al.* (2003), *Rousset et al.* (2003), *Chauvin et al.* (2005), *Neuhauser et al.* (2005), *Tamura et al.* (2006), *Kasper et al.* (2009), *Lagrange et al.* (2009), and *Thalmann et al.* (2009).

5.4. Multiple Mirror Telescope

The MMT AO system operating at 5 μm has also been used to seek warm planets with direct imaging, no coronagraph, and speckle suppression using ADI and SDI. So far only stellar companions have been discovered, but this program is as competitive as the others mentioned so far, due to high-Strehl at the longer wavelengths. A number of experiments with advanced starlight suppression are also being conducted with this system, including phase apodization and nulling interferometry.

References include *Biller et al.* (2007), *Kenworthy et al.* (2007), *Liu et al.* (2007), *Kenworthy et al.* (2009), and *Mamajek et al.* (2010).

5.5. Lyot Project, Project 1640

This is a project that conducted a survey of over 100 nearby stars with the sensitivity to find brown dwarfs and warm, young gas giant planets. Using a Lyot coronagraph along with an extremely high-order AO system on the U.S. Air Force AEOS telescope, it was able to exploit polarimetric speckle suppression to achieve images with a contrast below 10^{-6}, showing a perturbed solar-system-scale disk around AB Aurigae, which may have a planet in formation at 100 AU separation. The project also set constraints on the brown dwarf population of companions through statistical analysis of nondetections. Project 1640, the successor to the Lyot Project at Palomar's 5-m Hale Telescope, combines AO correction with an apodized pupil and hard-edged mask coronagraph as well as an integral field hyperspectral imaging device, simultaneously obtaining coronagraphic images at 30 different wavelengths over the 1.0–1.8-μm range. This system has just begun surveying nearby stars and has found faint stellar companions already. It employs chromatic speckle suppression as well as ADI and SDI, being the first instrument to be able to attempt all three such speckle removal techniques on the same set of data.

References include *Hinkley et al.* (2008), *Oppenheimer et al.* (2008), *Wizinowich* (2008), *LeConte et al.* (2010), *Hinkley et al.* (2010), and *Zimmerman et al.* (2010).

6. FUTURE PROJECTS

Exoplanet imaging and spectroscopy is a growing field of research. Its promise has led to the proposal of many new types of advanced instruments and has been used to justify construction of new major facilities both on the ground and in space. Continuing the path through Table 9, we now describe some of the projects that have not yet begun routine observation, but are already in the process of being built. Finally, at the end of this section we describe a few of the more speculative projects on the decadal time horizon. Since those may substantially change, the discussion provided here is intentionally general. In order to provide a sense of scale in terms of cost, the implementation of all the projects listed in Table 9 is estimated by the authors to be over 100 billion USD. Clearly not all these projects will happen given the level of funding for this kind of research, but this number can be easily compared for scale to many other things that developed societies spend money on.

6.1. Large Binocular Telescope Interferometer

The Large Binocular Telescope Interferometer (LBTI) uses the nulling technique available to two-aperture systems. The stated goals for the project are to detect Jupiter-like planets around younger stars (<2 G.y.) in the solar neighborhood, and 3 M_{Jup} planets around solar-age stars. The system will have a very large field of view in comparison with most current and planned projects. It requires two separate adaptive optics systems as well as the nulling system. The expectation is that nulling at the level of about 10^{-7} will be achieved within a few diffraction elements of the central star. LBTI is expected to begin nulling operations in the fall of 2011.

References include *Hinz et al.* (2008) and *Hinz* (2009).

6.2. Project 1640 Phase II

Project 1640 is due to be upgraded to have a full 3217-actuator AO system for far superior wavefront control, and a second-stage wavefront sensor (called a "science-arm WFS") that obtains the wavefront distortion due to the optical system on a timescale of roughly 1 s. This science-arm WFS is designed to sense and control, through periodic feedback to the AO system, the long-lived speckles that are removed so efficiently by the polarimetric technique, allowing the chromatic speckle suppression to act on much fainter speckles at the 10^{-7} level. The system employs an apodized pupil Lyot coronagraph and the same hyperspectral imaging device as in the current

Phase I. This may allow objects as faint as 10^{-8} at 6λ/D to be detected and spectra to be extracted, although the field of view is only 4 arcsec wide. This begins to open the exoplanet characterization phase space to the study of many Jupiter-mass exoplanets. First light for this system is expected in 2011.

6.3. Gemini Planet Imager

The Gemini Planet Imager (GPI), with a goal of 10^{-8} contrast at 6λ/D in raw contrast, is a large instrument consisting of an AO system with 1500 actuators, a science-arm WFS, as well as a hyperspectral imager and a dual-mode polarimetry imager. This is perhaps the most ambitious of the systems currently being developed, employing, like Project 1640, an apodized pupil Lyot coronagraph for starlight suppression. The goals of the project are to provide images and spectra of roughly 100–200 young planets around nearby stars, with a similar field of view as Project 1640. First light is expected in 2011.

References include *Macintosh et al.* (2008).

6.4. Spectro-Polarimetric High-Contrast Exoplanet Research

The Spectro-Polarimetric High-Contrast Exoplanet Research (SPHERE) project at the VLT employs a relatively high-order 1312-actuator AO system with an apodized pupil Lyot coronagraph, and an integral field hyperspectral imager, as well as a separate dual-mode polarimetric imager. It will be possible to conduct both polarimetry in the shorter wavelengths and hyperspectral imaging at longer wavelengths simultaneously over a roughly 4-arcsec field of view. Due to see first light in 2011, this system is predicted to achieve a few times 10^{-7} contrast levels at 6λ/D, enough contrast to image and obtain spectra of several tens of warm, young planets with masses as small as Jupiter around the nearest 100–200 stars.

The four projects above are fully funded and very close to achieving first light around 2011. One can hope that by 2013 or so, low-resolution spectra and orbits of perhaps a hundred exoplanets will be obtained.

References include *Beuzit et al.* (2008).

6.5. James Webb Space Telescope

The JWST is another project that is currently funded and expected to see first light. There are two primary instruments being constructed for JWST that address exoplanet imaging directly. The NIRCAM instrument has several coronagraphic modes operating at 1–5 μm and should achieve contrasts of about 10^{-7} to within a few λ/D, which will allow imaging thermal emission from exoplanets. It employs apodization on the segments of the telescope to reduce speckles pinned to the diffraction pattern as well as several other options for starlight suppression. Grisms permit low-resolution spectroscopy. The TFI instrument has Fabry-Perot etalons operating from 1.5 to 2.4 μm and 3.1 to 5.0 μm, with several hard-edged circular occulting spots and a nonredundant masking mode. Although the inner working angles are larger for this instrument, it should provide performance similar to that of GPI and SPHERE. Another instrument of interest in direct imaging of exoplanets with JWST is the Mid-InfraRed Imager (MIRI), which has several coronagraphic options and images at wavelengths from 9 to 12 μm. This system is predicted to detect planets as cool as 300 K within an arcsecond of a star, using four-quadrant phase masks or a traditional Lyot-style coronagraph. The combination of all these instruments will provide a suite of measurements across a broad range of wavelengths.

References include *Boccaletti et al.* (2005), *Greene et al.* (2007), and *Krist et al.* (2009).

6.6. Planetscope

Planetscope is one of the few proposed balloon experiments that would directly image exoplanets. Using a small-aperture telescope and a coronagraph that includes a low-order AO system (primarily for fine guiding and optical defect correction), this type of project would benefit from getting above more than 99% of atmospheric turbulence. This, and other suborbital projects for exoplanets, hold the promise of delivering science results as well as demonstrating technology for future space missions.

References include *Traub et al.* (2008) and *Chen et al.* (2009).

6.7. Probe-Class Space Missions

At present there are a slew of proposed space missions for direct exoplanet imaging, each costing on the order of 600 to 1000 million USD, and termed *probe-class* (or *medium-class*) missions. These are generally missions that involve relatively small-aperture telescopes, or up to four smaller apertures (to permit a combination of nulling and coronagraphy). Some of these proposed systems use the PIAA technique or band-limited focal plane masks and hyperspectral imaging sensors. These projects typically have the goal of contrasts at the 10^{-9} level, sufficient to detect mature giant planets, but not Earths (unless there is one around a very nearby star). None of these missions will be launched for the next five years or so, but it is possible that one may enter development during that period.

These probe-class projects, in alphabetical order, and the respective lead authors, are ACCESS, a 1.5-m coronagraph (J. T. Trauger); DaVINCI, four 1.1-m visible nullers (M. Shao); EPIC, a 1.65-m visible nuller (M. Clampin); and PECO, a 1.4-m coronagraph (O. Guyon).

References include *Trauger et al.* (2008), *Shao et al.* (2008), *Lyon et al.* (2008), and *Guyon et al.* (2009).

6.8. Large Groundbased Observatories

Moving further into the future, the Thirty Meter Telescope (TMT) project, the Giant Magellan Telescope (GMT), and the European Extremely Large Telescope (E-ELT), all 30–42-m-diameter segmented aperture telescopes, have as core instruments planet detection and characterization projects. These systems face the difficulty of dealing with segmented apertures, which intrinsically diffract light (as do standard spider support structures in on-axis telescope designs). Most of the standard coronagraphic techniques are exceedingly inefficient on segmented mirror telescopes because the optical stops must mask off the segment edges. However, the visible nulling technique overcomes this issue by using a matrix of single-mode fibers mapped to the segments to clean up high-spatial-frequency wavefront errors, and a piston-only deformable mirror to map the segments into one coherent wavefront, as though it were a single aperture. In the process it interferometrically nulls the starlight, giving a peculiar spatially variable throughput. This requires many pointings to achieve sensitivity in the telescope's full field of view, but allows for use of the full aperture of the segmented mirror. Each of these projects expects first light in about 2018; planet detection instrumentation is not likely to be a first-light priority, but it may well be soon thereafter.

References include *Gilmozzi and Spyromilio* (2008), *Johns* (2008), *Nelson and Sanders* (2008), and *Shao et al.* (2008).

6.9. Large Space Missions

On the more distant horizon, several groups around the world have proposed UV/optical space telescopes on the scale of 4–16-m diameter. The most well-studied of these is TPF-C, with an 8-m × 3.5-m oval monolithic clear-aperture primary mirror and an internal coronagraph. TPF-C has had years of research behind its design and science program, and was deemed to be technically feasible. The prime goal of TPF-C is to carry out spectroscopic characterization of planets, at visible wavelengths, for planets down to and including Earth-twins.

Two other large telescopes have recently been suggested: the Advanced Technology Large Aperture Space Telescope (ATLAST), and the eXtrasolar Planet Characterizer (XPC or THEIA). ATLAST actually refers to a family of designs: an 8-m monolithic circular primary, a 9.2-m deployable segmented mirror, and a 16-m deployable segmented version. The 8-m clear-aperture version of ATLAST is similar to TPF-C, and could use its full pupil for an efficient coronagraph, but the segmented versions would require a visible nuller type of coronagraph. THEIA was originally studied as a possible way to combine a 4-m telescope with an internal and external coronagraph, simultaneously, but it was found to be not feasible and this aspect has been set aside in favor of the latter.

The external occulter concept (TPF-O) has been suggested as a way to utilize an existing telescope by adding a distant starshade, with the latter possibly being delivered to orbit (e.g., L2) independently of the former. Potential advantages include a lower cost for the sunshade itself (as compared to an internal coronagraph telescope, and assuming that the companion telescope is not included in the cost), and the ability to image planets closer to their parent star, compared to an internal coronagraph. Disadvantages include being limited to a narrow annulus on the sky, centered near 90° from the Sun, and a long slew time between target observations.

The original Earth-characterizing space instrument was the Terrestrial Planet Finder Interferometer (TPF-I), a thermal-infrared interferometer with formation-flying cryogenic collectors and combiners. A similar concept, Darwin, was developed in parallel, mainly in Europe. These projects have essentially merged in the sense that they are now technically essentially identical. Despite being first to develop, the TPF-I/Darwin concept is currently expected to be the last to be flown, after a TPF-C or similar type is flown, owing to the perceived complexity and cost of a system that requires five free-flying spacecraft, all at cryogenic temperatures.

These large-aperture telescopes would be designed to image planets as small as an Earth-twin, unless the majority of stars have zodiacal disks that are 10–100 times thicker than our own solar system's zodiacal dust. These devices, which in principle could be constructed now and launched within a few years, could begin to tackle the question of life in other planetary systems. The key to this is the detection of chemical abundances in a planet's atmosphere that are not consistent with thermochemical equilibrium and could only be generated by biological activity.

References include *Traub et al.* (2006), *Cash* (2006), *Lawson et al.* (2008), *Cady et al.* (2009), *Cockell et al.* (2009), *Kasting et al.* (2009), *Levine et al.* (2009), *Postman et al.* (2009), and *Soummer et al.* (2009).

6.10. Summary and Outlook

The effort to make direct images of exoplanets is a complicated and rather difficult endeavor. It involves clever optical techniques and exquisite control of the imaging system. Indeed, the systems about to begin collecting data in the next year or so are aiming for λ/1000 level wavefront control in devices that are dealing with light corrupted by the atmosphere as well as processed by 30–40 optical surfaces. These systems (GPI, SPHERE, P1640) will only be able to study the largest planets around relatively young stars. Yet, the future of this field is exciting. A handful of images have been obtained and spectra of those objects will be available soon.

Next-generation systems will likely provide upwards of 100 or 200 spectra of exoplanets within the following 4 to 5 years. Then the field of comparative planetary science will see a new vitality. For the first time in human history we will be able to compare what many different Jupiter-mass objects have in common, how they evolve through different ages, and whether they even do have anything in common.

In the more distant future, it seems likely that acquiring spectra, orbits, masses, and the other necessary characteristics of the much larger population of older planets in closer orbits to their stars, ones whose signature in our instruments is the result of reflected starlight, not internal radiation, will require spacebased experiments. We showed in section 4.16 that if we were to expand the sorts of groundbased instruments being built now to the regime where they would be sensitive to the majority of planets, they would require guide stars that are brighter than any that exist.

This implies, of course, that moving to space, where one can control wavefronts with far slower cadence, provides the obvious solution. Although we believe that this is true, one of the biggest mistakes a scientist can make is to assume that the primary mode of attacking a particular question is the only one. It is not impossible that the issue of extremely high-contrast imaging will be solved in an alternate way, without adaptive optics, perhaps, or with some other type of optical manipulation.

As such, one must remain optimistic, and in the end, the overwhelmingly compelling nature of the science of exoplanets, and how they directly relate to our own existence, means that the science will get done. Whatever mission, telescope, or technique is eventually used, perhaps even within the next 20 years, other planets similar to Earth with telltale signs of biological forcing of atmospheric chemistry will be discovered.

References include *Oppenheimer and Hinkley* (2009).

6.11. Epilogue: New Worlds, New Horizons

The Committee for a Decadal Survey issued their report "New Worlds, New Horizons in Astronomy and Astrophysics" in August 2010. This report, also known as Astro2010, specified priority science objectives for the decade 2012–2021 for all of ground- and spacebased astronomy and astrophysics in the U.S., but given the nature of our research, the implications are worldwide.

Astro2010 clearly gives exoplanets a high priority, as high as they can in the face of expected flat future budgets. Specifically, Astro2010 proposes a single flagship mission for the decade with dual science goals: dark energy and exoplanets. The exoplanet data will come from gravitational microlensing of stars toward the galactic bulge, giving us a census of planets with semimajor axes of roughly 1 AU and greater, to complement Kepler's census of planets at roughly 1 AU and smaller, for the purpose of getting the best possible estimate of the frequency of Earth-mass planets in habitable zones. Astro2010 also recommends developing exoplanet technology in the coming 5–10 years to lay the foundation for a future mission to study nearby Earth-like planets, with a possible new start for a flagship in the early 2020s.

Since direct imaging of nearby exoplanets is clearly a potential candidate for this mission, the present chapter is especially relevant. Our hope, as the authors, is that the methods we discuss here will inspire you, the reader, to even newer and better ways of directly imaging nearby exoplanets.

Acknowledgments. W.A.T. thanks NASA's Exoplanet Exploration Program as well as the Kavli Institute for Theoretical Physics (Exoplanet Program) for supporting and hosting part of this work. This research was supported in part by the National Science Foundation under Grant No. PHY05-51164. B.R.O. thanks the AMNH, FUTDI, and the Blavatnik Family Foundation for partial support. Part of the research in this chapter was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

Allen C. W. (1991) *Astrophysical Quantities.* Oxford Univ., Oxford.

Arenberg J. W., Lo A. S., Cash W., and Polidan R. S. (2006) New Worlds Observer: Using occulters to directly observe planets. In *Space Telescopes and Instrumentation I: Optical, Infrared, and Millimeter* (J. Mather et al., eds.), pp. 62651W. SPIE Conf. Series 6265, Bellingham, Washington.

Artigau E., Biller B. A., Wahhaj Z., Hartung M., Hayward T. L., et al. (2008) NICI: Combining coronagraphy, ADI, and SDI. In *Ground-based and Airborne Instrumentation for Astronomy II* (I. McLean and M. Casali, eds.), pp. 70141Z–70141Z–9. SPIE Conf. Series 7014, Bellingham, Washington.

Barry R. K., Danchi W. C., Traub W. A., Sokoloski J. L., Wisniewski J. P., et al. (2008) Milliarcsecond N-band observations of the nova RS Ophiuchi: First science with the Keck Interferometer Nuller. *Astrophys. J., 677,* 1253–1267.

Beichman C. A., Woolf N. J., and Lindensmith C. A. (1999) *The Terrestrial Planet Finder.* JPL Publication 99-3, Jet Propulsion Laboratory, Pasadena, California.

Beuzit J.-L., Feldt M., Dohlen K., Mouillet D., Puget P., et al. (2008) SPHERE: A planet finder instrument for the VLT. In *Ground-based and Airborne Instrumentation for Astronomy II* (I. McLean and M. Casali, eds.), pp. 701418–701418-12. SPIE Conf. Series 7014, Bellingham, Washington.

Biller B. A., Close L. M., Masciadri E., Lenzen R., Brandner W., et al. (2006) Contrast limits with the Simultaneous Differential Extrasolar Planet Imager (SDI) at the VLT and MMT. In *Advances in Adaptive Optics II* (B. Ellerbroek and D. Bonaccini, eds.), pp. 62722D. SPIE Conf. Series 6272, Bellingham, Washington.

Biller B. A., Close L. M., Masciadri E., Nielsen E., Lenzen R., et al. (2007) An imaging survey for extrasolar planets around 45 close, young stars with the Simultaneous Differential Imager at the Very Large Telescope and MMT. *Astrophys. J. Suppl., 173,* 143–165.

Boccaletti A., Baudoz P., Baudrand J., Reess J. M, and Rouan D. (2005) Imaging exoplanets with the coronagraph of JWSR/MIRI. *Adv. Space Res., 36,* 1099–1106.

Borde P. J. and Traub W. A. (2006) High-contrast imaging from space: Speckle nulling in a low-aberration regime. *Astrophys. J., 638,* 488–498.

Born M. and Wolf E. (1999) *Principles of Optics*. Cambridge Univ., Cambridge.

Cady E., Belikov R., Dumont P., Egerman R., Kasdin N. J., et al. (2009) Design of a telescope-occulter system for THEIA. *ArXiV preprints,* arXiv:0912.2938v1.

Cash W. (2006) Detection of Earth-like planets around nearby stars using a petal-shaped occulter. *Nature, 442,* 51–53.

Chauvin G., Lagrange A.-M., Dumas C., Zuckerman B., Mouillet D., et al. (2005) Giant planet companion to 2MASSW J1207334-393254. *Astron. Astrophys., 438,* L25–L28.

Chen P., Traub W. A., Kern B., and Matsuo T. (2009) Seeing in the stratosphere. *Bull. Am. Astron. Soc., 41,* 438.

Cockell C. S., Herbst T., Leger A., Absil O., Beichman C., et al. (2009) Darwin — An experimental astronomy mission to search for habitable planets. *Experimental Astron., 23,* 435–461.

Colavita M. M., Serabyn E., Booth A. J., Crawford S. L., Garcia-Gathright J. I., et al. (2008) Keck interferometer nuller upgrade. In *Optical and Infrared Interferometry* (M. Scholler et al., eds.), pp. 70130A–70130A-14.SPIE Conf. Series 7013, Bellingham, Washington.

Cox A. N., ed. (2000) *Allen's Astrophysical Quantities,* 4th edition. Springer, New York. 719 pp.

de Pater I. and Lissauer J. J. (2001) *Planetary Sciences*. Cambridge Univ., Cambridge.

de Pater I. and Lissauer J. J. (2010) *Planetary Sciences,* 2nd edition. Cambridge Univ., Cambridge.

Des Marais D. J., Harwit M. O., Jucks K. W., Kasting J. F., Lin D. N. C., et al. (2002) Remote sensing of planetary properties and biosignatures on extrasolar terrestrial planets. *Astrobiology, 2,* 153–181.

Gilmozzi R. and Spyromilio J. (2008) The 42-m European ELT: Status. In *Ground-based and Airborne Telescopes II* (L. M. Stepp and R. Gilmozzi, eds.), pp. 701219–701219-10. SPIE Conf. Series 7012, Bellingham, Washington.

Give'on A., Kasdin N. J., Vanderbei R. J., and Avitzour Y. (2006) On representing and correcting wavefront errors in high-contrast imaging systems. *J. Optical Soc. Am., A23,* 1063–1073.

Glassman T., Johnson A., Lo A., Dailey D., Shelton H., and Vogrin J. (2010) Error analysis on the NWO starshade. In *Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave* (J. M. Oschmann Jr. et al., eds.), 773150. SPIE Conf. Series 7731, Bellingham, Washington.

Greene T., Beichman C., Eisenstein D., Horner S., Kelly D., et al. (2007) Observing exoplanets with the JWST NIRCam grisms. In *Techniques and Instrumentation for Detection of Exoplanets III* (D. Coulter, ed.), pp. 66930G–66930G-10. SPIE Conf. Series 6693, Bellingham, Washington.

Guyon O. (2003) Phase-induced amplitude apodization of telescope pupils for extrasolar terrestrial planet imaging. *Astron. Astrophys., 404,* 379–387.

Guyon O. (2005) Limits of adaptive optics for high contrast imaging. *Astrophys. J., 629,* 592–614 (revised version available at *arXiv:astroph/0505086v2*).

Guyon O., Pluzhnik E. A., Kuchner M. J., Collins B., and Ridgway S. T. (2006) Theoretical limits on extrasolar terrestrial planet detection with coronagraphs. *Astrophys. J., 167,* 81–99.

Guyon O., Angel J. R. P., Belikov R., Egerman R., Gavel D., et al. (2009) Detecting and characterizing exoplanets with a 1.4-m space telescope: The Pupil mapping Exoplanet Coronagraphic Observer (PECO). In *Techniques and Instrumentation for Detection of Exoplanets IV* (S. Shaklan, ed.), pp. 74400F–74400F-10. SPIE Conf. Series 7440, Bellingham, Washington.

Guyon O., Pluzhnik E., Martinache F., Totems J., Tanaka S., et al. (2010) High-contrast imaging and wavefront control with a PIAA coronagraphic laboratory system validation. *Publ. Astron. Soc. Pac., 122,* 71–84.

Herriot G., Morris S., Anthony A., Derdall D., Duncan D., et al. (2000) Progress on Altair: The Gemini North adaptive optics system. In *Adaptive Optical Systems Technology* (P. Wizinowich, ed.), pp. 115–125. SPIE Conf. Series 4007, Bellingham, Washington.

Hinkley S., Oppenheimer B. R., Soummer R., Sivaramakrishnan A., Roberts L. C. Jr., et al. (2007) Temporal evolution of coronagraphic dynamic range and constraints on companions to Vega. *Astrophys. J., 654,* 633–640.

Hinkley S., Oppenheimer B. R., Brenner D., Parry I. R., Sivaramakrishnan A., et al. (2008) A new integral field spectrograph for exoplanetary science at Palomar. In *Adaptive Optics Systems* (N. Hubin, ed.), pp. 701519–701519-10. SPIE Conf. Series 7015, Bellingham, Washington.

Hinkley S., Oppenheimer B. R., Brenner D., Zimmerman N., Roberts L. C. Jr., et al. (2010) Discovery and characterization of a faint stellar companion to the A3V star zeta Virginis. *Astrophys. J., 712,* 421–428.

Hinz P. M. (2009) Detection of debris disks and wide orbit planets with the LBTI. In *Exoplanets and Disks: Their Formation and Diversity,* pp. 313–317. AIP Conf. Series 1158, American Institute of Physics, New York.

Hinz P. M., Solheid E., Durney O., and Hoffmann W. F. (2008) NIC: LBTI's nulling and imaging camera. In *Optical and Infrared Interferometry* (M. Schöller et al., eds.), pp. 701339–701339-12. SPIE Conf. Series 7013, Bellingham, Washington.

Hull T., Trauger J. T., Macenka S. A., Moody D., Olarte G., et al. (2003) Eclipse telescope design factors. In *High-Contrast Imaging for Exo-Planet Detection* (A. Schultz and R. Lyon, eds.), pp. 277–287. SPIE Conf. Series 4860, Bellingham, Washington.

Janson M., Bergfors C, Goto M., Brandner W., and Lafrenière D. (2010) Spatially resolved spectroscopy of the exoplanet HR 8799c. *Astrophys. J. Lett., 710,* L35–L38.

Johns M. (2008) Progress on the GMT. In *Ground-based and Airborne Telescopes II* (L. Stepp and R. Gilmozzi, eds.), pp. 70121B–70121B-15. SPIE Conf. Series 7012, Bellingham, Washington.

Kalas P., Graham J. R., Chiang E., Fitzgerald M. P., Clampin M., et al. (2008) Optical images of an exosolar planet 25 light-years from Earth. *Science, 322,* 1345–1348; erratum 19 Jan. 2009.

Kaltenegger L., Traub W. A., and Jucks K. W. (2007) Spectral evolution of an Earth-like planet. *Astrophys. J., 658,* 598–616.

Kasdin N. J., Vanderbei R. J., Spergel D. N., and Litman M. G. (2003) Extrasolar planet finding via optimal apodized-pupil and shaped-pupil coronagraphs. *Astrophys. J., 582,* 1147–1161.

Kasdin N. J., Belikov R., Beall J., Vanderbei R. J., Littman M. G., et al. (2005) Shaped pupil coronagraphs for planet finding: Optimization, manufacturing, and experimental results. In *Techniques and Instrumentation for Detection of Exoplanets II* (D. Coulter, ed.), pp. 128–136. SPIE Conf. Series 5905, Bellingham, Washington.

Kasdin N. J., Cady E. J., Dumont P. J., Lisman P. D., Shaklan S. B., et al. (2009) Occulter design for THEIA. *Techniques and Instrumentation for Detection of Exoplanets IV* (S. Shaklan, ed.), pp. 744005–744005-8. SPIE Conf. Series 7440, Bellingham, Washington.

Kasper M., Amico P., Pompei E., Ageorges N., Apai D., et al. (2009) Direct imaging of exoplanets and brown dwarfs with the VLT: NACO pupil-stabilized Lyot corongraphy at 4 µm. *Messenger, 137,* 8–13.

Kasting J., Traub W., Roberge A., Leger A., Schwartz A., et al. (2009) Exoplanet characterization and the search for life. *ArXiV preprints,* arXiv:0911.2936v1.

Kenworthy M. A., Codona J. L., Hinz P. M., Angel J. R. P., Heinze A., and Sivanandam S. (2007) First on-sky high-contrast imaging with an apodizing phase plate. *Astrophys. J., 660,* 762–769.

Kenworthy M. A., Mamjek E. E., Hinz P. M., Meyer M. R., Heinze A. N., et al. (2009) MMT/AO 5 µm imaging constraints on the existence of giant planets orbiting Fomalhaut. *Astrophys. J., 697,* 1928–1933.

Krist J. E. (2004) High contrast imaging with the Hubble Space Telescope: Performance and lessons learned. In *Optical, Infrared, and Millimeter Space Telescopes* (J. C. Mather, ed.), pp. 1284–1285. SPIE Conf. Series 5487, Bellingham, Washington.

Krist J. E., Balasubramanian K., Beichman C. A., Echternach P. M., Green J. J., et al. (2009) The JWST/NIRCam coronagraph mask design and fabrication. In *Techniques and Instrumentation for Detection of Exoplanets IV* (S. Shaklan, ed.), pp. 74400W–74400W-10. SPIE Conf. Series 7440, Bellingham, Washington.

Kuchner M. J. (2004a) A minimum-mass extrasolar nebula. *Astrophys. J., 612,* 1147–1151.

Kuchner M. J. (2004b) A unified view of coronagraph image masks. *ArXiV preprints,* arXiv:astro-ph/0401256v1.

Kuchner M. J. and Traub W. A. (2002) A coronagraph with a band-limited mask for finding terrestrial planets. *Astrophys. J., 570,* 900–908.

Kuchner M. J., Crepp J., and Ge J. (2005) Eighth-order image masks for terrestrial planet finding. *Astrophys. J., 628,* 466–473.

Kuhn J. R., Potter D., and Parise B. (2001) Imaging polarimetric observations of a new circumstellar disk system. *Astrophys. J. Lett., 553,* L189–L191.

Labeyrie A. (1996) Resolved imaging of extra-solar planets with future 10–100 km optical interferometric arrays. *Astron. Astrophys. Suppl. Ser., 118,* 517–524.

Lafreniere D., Doyon R., Marois C., Nadeau D., Oppenheimer, B. R., et al. (2007) The Gemini deep planet survey. *Astrophys. J., 670,* 1367–1390.

Lafreniere D., Marois C., Doyon R., and Barman T. (2009) HST/NICMOS detection of HR 8799b in 1998. *Astrophys. J. Lett., 694,* L148–L152.

Lagrange A.-M., Gratadour D., Chauvin G., Fusco T., Ehrenreich D., et al. (2009) A probable giant planet imaged in the beta Pictoris disk. *Astron. Astrophys., 493,* L21–L25.

Lagrange A.-M., Bonnefoy M., Chauvin G., Apai D., et al. (2010) A giant planet imaged in the disk of the young star beta Pictoris. *Science, 329,* 57.

Lawson P. R., Lay O. P., Martin S. R., Peters R. D., Gappinger R. O., et al. (2008) Terrestrial Planet Finder Interferometer: 2007–2008 progress and plans. In *Optical and Infrared Interferometry* (M. Schöller et al., eds.), pp. 70132N–70132N-15. SPIE Conf. Series 7013, Bellingham, Washington.

LeConte J., Soummer R., Hinkley S., Oppenheimer B. R., Sivaramakrishnan A., et al. (2010) The Lyot Project direct imaging survey of substellar companions: Statistical analysis and information from nondetections. *Astrophys. J., 716,* 1551–1565.

Lenzen R., Hartung M., Brandner W., Finger G., Hubin N. N., et al. (2003) NAOS-CONICA first on sky results in a variety of observing modes. In *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes* (M. Iye and A. F. M. Moorwood, eds.), pp. 944–952. SPIE Conf. Series 4841, Bellingham, Washington.

Levine M., Lisman D., Shaklan S., Kastin J., Traub W., et al. (2009) Terrestrial Planet Finder Coronagraph (TPF-C) flight baseline concept. *ArXiV preprints,* arXiv:0911.3200v1.

Liu W. M., Hinz P. M., Meyer M. R., Mamajek E. E., Hoffmann W. F., et al. (2007) Observations of Herbig Ae disks with nulling interferometry. *Astrophys. J., 658,* 1164–1172.

Lyon R. G., Clampin M., Melnick G., Tolls V., Woodruff R., and Vasudevan G. (2008) Extrasolar Planetary Imaging Coronagraph (EPIC): Visible nulling coronagraph testbed results. In *Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter* (J. Oschmann et al., eds.), pp. 701045–701045-7. SPIE Conf. Series 7010, Bellingham, Washington.

Lyot B. (1933) The study of the solar corona without an eclipse. *R. Astron. Soc. Canada, 27,* 225–234, 265–280.

Macintosh B. A., Graham J. R., Palmer D. W., Doyon R., Dunn J., et al. (2008) The Gemini Planet Imager from science to design to construction. In *Adaptive Optics Systems* (N. Hubin et al., eds.), pp. 701518–701518-13. SPIE Conf. Series 7015, Bellingham, Washington.

Mamajek E. E., Kenworthy M. A., Hinz P. M., and Meyer M. R. (2010) Discovery of a faint companion to Alcor using MMT/AO 5 μm imaging. *Astron. J., 139,* 919–925.

Marois C., Doyon R., Nadeau D., Racine R., Riopel M., et al. (2005) TRIDENT: An infrared differential imaging camera optimized for the detection of methanated substellar companions. *Publ. Astron. Soc. Pac., 117,* 745–756.

Marois C., Lafreniere D., Doyon R., Macintosh B., and Nadeau D. (2006) Angular differential imaging: A powerful high-contrast imaging technique. *Astrophys. J., 641,* 556–564.

Marois C., Macintosh B., Barman T., Zuckerman B., Song I., et al. (2008) Direct imaging of multiple planets orbiting the star HR 8799. *Science, 322,* 1348.

Mawet D., Riaud P., Absil O., and Surdej J. (2005) Annular groove phase mask coronagraph. *Astrophys. J., 633,* 1191–1200.

Mawet D., Serabyn E., Liewer K., Burruss R., Hickey J., and Shemo D. (2010) The vector vortex coronagraph: Laboratory results and first light at Palomar Observatory. *Astrophys. J., 709,* 53–57.

Nelson J. and Sanders G. H. (2008) The status of the Thirty Meter Telescope project. In *Ground-based and Airborne Telescopes II* (L. Stepp and R. Gilmozzi, eds.), pp. 70121A–70121A-18. SPIE Conf. Series 7012, Bellingham, Washington.

Neuhauser R., Guenther E. W., Wuchterl G., Mugrauer M., Bedalov A., and Hauschildt P. H. (2005) Evidence for a co-moving substellar companion of GQ Lup. *Astron. Astrophys., 435,* L13–L16.

Oppenheimer B. R. and Hinkley S. (2009) High-contrast imaging in optical and infrared astronomy. *Annu. Rev. Astron. Astrophys., 47,* 253–289.

Oppenheimer B. R., Brenner D., Hinkley S., Zimmerman N., Sivaramakrishnan A., et al. (2008) The solar-system-scale disk around AB Aurigae. *Astrophys. J., 679,* 1574–1581.

Pedretti E., Labeyrie A., Arnold L., Thureau N., Lardiere O., et al. (2000) First images on the sky from a hyper telescope. *Astron. Astrophys. Suppl. Ser., 147,* 285–290.

Perrin M. D., Graham J. R., Kalas P., Lloyd J. P., Max C. E., et al. (2004) Laser guide star adaptive optics imaging polarimitry of Herbig Ae/Be stars. *Science, 303,* 1345–1348.

Postman M., Traub W., Krist J., Stapelfeldt K., Brown R., et al. (2009) Advanced Technology Large-Aperture Space Telescope (ATLAST): Characterizing habitable worlds. *ArXiV preprints,* arXiv:0911.3841v1.

Poyneer L., van Dam M., and Veran J.-P. (2009) Experimental verification of the frozen flow atmospheric turbulence assumption with use of astronomical adaptive optics telemetry. *J. Optical Soc. Am., A26,* 833–846.

Racine R., Walker G. A. H., Nadeau D., Doyon R., and Marois C. (1999) Speckle noise and the detection of faint companions. *Publ. Astron. Soc. Pac., 111,* 587–594.

Rousset G., Lacombe F., Puget P., Hubin N. N., Gendron E., et al. (2003) NAOS, the first AO system of the VLT: On-sky performance. In *Adaptive Optical System Technologies II* (P. L. Wizinowich and D. Bonaccini, eds.), pp. 140–149. SPIE Conf. Series 4839, Bellingham, Washington.

Serabyn E., Mawet D., and Burruss R. (2010) An image of an exoplanet separated by two diffraction beamwidths from a star. *Nature, 464,* 1018–1020.

Shaklan S. B., Noecker M. C., Glassman T., Lo A. S., and Dumont P. J. (2010) Error budgeting and tolerancing of starshades for exoplanet detection. In *Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave* (J. M. Oschmann Jr. et al., eds.), 77312G. SPIE Conf. Series 7731, Bellingham, Washington.

Shao M., Bairstow S., Levine M., Vasisht G., Lane B. F., et al. (2008) DAVINCI, a dilute aperture visible nulling coronagraphic instrument. In *Optical and Infrared Interferometry* (M. Schöller et al., eds.), pp. 70132T–70132T-13. SPIE Conf. Series 7013, Bellingham, Washington.

Soummer R., Cash W., Brown R. A., Jordan I., Roberge A., et al. (2009) A starshade for JWST: Science goals and optimization. In *Techniques and Instrumentation for Detection of Exoplanets IV* (S. Shaklan, ed.), pp. 7440A–7440A-15. SPIE Conf. Series 7440, Bellingham, Washington.

Sparks W. B. and Ford H. C. (2002) Imaging spectroscopy for extrasolar planet detection. *Astrophys. J., 578,* 543–564.

Tamura M., Hodapp K., Takami H., Lyu A., Suto H., et al. (2006) Concept and science of HiCIAO: High contrast instrument for the Subaru next generation adaptive optics. In *Ground-based and Airborne Instrumentation for Astronomy* (I. McLean and M. Iye, eds.), p. 62690V. SPIE Conf. Series 6269, Bellingham, Washington.

Thalmann C., Carson J., Janson M., Goto M., McElwain M., et al. (2009) Discovery of the coldest imaged companion of a Sun-like star. *Astrophys. J. Lett., 707,* L123–L127.

Traub W. A. (1986) Combining beams from separated telescopes. *Appl. Optics, 25,* 528–532.

Traub W. A. (2000) Beam combination and fringe measurement. In *Principles of Long Baseline Interferometry* (P. R. Lawson, ed.), pp. 31–58. JPL Publ. 00-009, Jet Propulsion Laboratory, Pasadena, California. Available online at *http://olbin.jpl.nasa.gov/iss1999/coursenotes.html*.

Traub W. A. (2003) The colors of extrasolar planets. In *Scientific Frontiers in Research on Extrasolar Planets* (D. Deming and S. Seager, eds.), pp. 595–602. ASP Conf. Series 294, Astronomical Society of the Pacific, San Francisco.

Traub W. A. and Vanderbei R. J. (2003) Two-mirror apodization for high-contrast imaging. *Astrophys. J., 599,* 695–701.

Traub W. A., Levine M., Shaklan S., Kasting J., Angel J. R., et al. (2006) TPF-C: Status and recent progress. In *Advances in Stellar Interferometry* (J. Monnier et al., eds.), pp. 62680T. SPIE Conf. Series 6268, Bellingham, Washington.

Traub W., Chen P., Kern B., and Matsuo T. (2008) Planetscope: An exoplanet coronagraph on a balloon platform. In *Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter* (J. Oschmann et al., eds.), pp. 70103S–70103S-12. SPIE Conf. Series 7010, Bellingham, Washington.

Trauger J. T. and Traub W. A. (2007) A laboratory demonstration of the capability to image an Earth-like extrasolar planet. *Nature, 446,* 771–773.

Trauger J. T., Stapelfeldt K., Traub W., Henry C., Krist J., et al. (2008) ACCESS: A NASA mission concept study of an actively corrected coronagraph for exoplanet system studies. In *Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter* (J. Oschmann et al., eds.), pp. 701029–701029-11. SPIE Conf. Series 7010, Bellingham, Washington.

Turnbull M. C., Traub W. A., Jucks K. W., Woolf N. J., Meyer M. R., et al. (2006) Spectrum of a habitable world: Earthshine in the near-infrared. *Astrophys. J., 644,* 551–559.

Vanderbei R. J. and Traub W. A. (2005) Pupil mapping in two dimensions for high-contrast imaging. *Astrophys. J., 626,* 1079–1090.

Vanderbei R. J., Cady E., and Kasdin N. J. (2007) Optimal occulter design for finding extrasolar planets. *Astrophys. J., 665,* 794–798.

Wizinowich P., Dekany R., Gavel D., Max C., Adkins S., et al. (2008) W. M. Keck Observatory's next-generation adaptive optics facility. In *Adaptive Optics Systems* (N. Hubin et al., eds.), pp. 701511–701511-12. SPIE Conf. Series 7015, Bellingham, Washington.

Woolf N. J., Smith P. S., Traub W. A., and Jucks K. W. (2006) The spectrum of Earthshine: A pale blue dot observed from the ground. *Astrophys. J., 574,* 430–433.

Zimmerman N., Oppenheimer B. R., Hinkley S., Brenner D., Parry I. R., et al. (2010) Parallactic motion for companion discovery: An M-dwarf orbiting Alcor. *Astrophys. J., 709,* 733–740.