A Search for Interstellar Communications at Optical Wavelengths

A thesis presented by

Charles Michael Coldwell

to

Department of Physics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the subject of

Physics

Harvard University

Cambridge, Massachusetts

June 2002

A Search for Interstellar Communications at Optical Wavelengths

by

Charles Michael Coldwell

Abstract

The Harvard-Smithsonian optical SETI project is a search for intentional transmissions from intelligent extraterrestrial civilizations. A plausible scenario for these transmissions is developed which concludes that they could consist of very short pulses of visible light. The search targets stars similar to our own Sun (F, G and K type dwarves) nearby in our galaxy. A group at Princeton with whom we are collaborating has started simultaneously observing the same stars; this has reduced the number of background events to zero. 8,206 stars have been observed at Harvard and 1,088 at Princeton; so far no events been seen which present the assumed characteristics of an extraterrestrial intelligent origin.

Thesis Supervisor: Paul Horowitz
Title: Professor of Physics

The author can be contacted at `coldwell@frank.harvard.edu`

For Cindy

Acknowledgments

First and foremost, I would like to thank my advisor, Paul Horowitz, who has been so much more than a mentor over the years. His generosity knows no limits, and his fascination with physics and technology is both genuine and infectious. His laboratory is so much more than a facility for doing research in physics; it is more like a playground for people who love to figure out how things work. It would be hard to imagine a better advisor.

Costas Papaliolios has been a constant presence and a tremendous help to me during my career in graduate school. He has always had the uncanny ability to "cut the Gordian knot": to find the simple solutions to thorny problems that stymie his peers. It is fashionable to call this "thinking outside of the box" these days; personally I think he's just a really smart guy. But Cos is much more than a powerful intellect, he's also a kind, generous, sympathetic and funny man, not to mention a damn good shot at pool.

This experiment couldn't have happened without the phenomenal contributions of Jonathan Wolff, who conjured up the entire apparatus in one summer almost faster than the design was forming. Jonathan is the embodiment of technical brilliance; a quick study who knows no fear when it comes to new challenges.

Our collaborators at the Smithsonian Astrophysical Observatory, Robert Stefanik, Joseph Zajac, Joseph Caruso, David Latham and Guillermo Torres put in countless sleepless nights at the telescope making the observations that form our corpus of data, not to mention daily performing small miracles to keep our observatory going on a shoestring budget. The importance of their participation in making this program possible cannot be overstated. Particular kudos go to Joe Zajac for letting me hack apart his code to support simultaneous observations at Princeton.

Our collaborators in Princeton, David Wilkinson, Norman Jarosik, Ed Groth, Natalie Deffenbaugh, Alexander Willman and many more deserve special thanks for the incredible job they did restoring their telescope to working condition and providing us with the ultimate solution to our systematics.

Anne Sung endured the tedious learning curve to master the CAD tools necessary to rebuild the instrument electronics as a printed circuit board so that we could provide a copy for our collaborators in Princeton. She did a beautiful job; if I ever see her signature on another silkscreen, I'll know it's a good board.

Andrew Howard came into our group with great enthusiasm for the optical SETI program and will carry it forward from here. It could not be in better hands. He has been particularly helpful to me in making corrections and suggestions for this thesis, and completely indispensible to the Harvard OSETI program.

Doug Mink at the Harvard-Smithsonian Center for Astrophysics provided most of the astrometric data used in this thesis. He has written a suite of very powerful computer programs for extracting interesting statistics about star positions, proper motions and parallaxes and was very generous with his time in digging out the particular values that I was interested in.

Cindy Hancox has been a sympathetic ear and pillar of support for me; this thesis is dedicated to her.

For moral and technical support thanks to Jonathan Weintroub. Thanks also to Darren Leigh for encouragement (or was that harassment?), the Harvard thesis LaTeX document class and for blazing the trail with the first Ph.D. in SETI.

Although none of them were personally involved in this experiment, I would also like to thank the hundreds (thousands?) of computer programmers who contributed to the excellent free software tools which made this experiment possible, especially those responsible for the GNU/Linux operating system (Richard Stallman, Linus Torvalds, Alan Cox and many more), the PostgreSQL database engine, the TeX and LaTeX document formatting systems (Donald Knuth and Leslie Lamport) and the SM plotting package (Robert Lupton and Patricia Monger).

Chapter 1

Hypothesis

1.1 Background

Historically, SETI experiments have focussed on the microwave part of the electromagnetic spectrum. This is no doubt due in large part to the influence of the original paper by Cocconi and Morrison [5] proposing a search for interstellar communications in the microwave region of the electromagnetic spectrum. Their suggestion for the optimum channel was based largely on considerations of signal attenuation in planetary atmospheres, and on technological limitations of the time (1959). The optical and near-infrared regimes are dismissed without much discussion: "The bandwidths which seem physically possible in the near-visible or gamma-ray domains demand either very great power at the source or very complicated techniques. The wide radio-band from, say, 1 Mc. to 10^4 Mc./s, remains as the rational choice."

Little did the authors know that the "complicated techniques" required for communication at optical and near-infrared wavelengths would be perfected here on Earth as little as two years after the publication of their paper, when in 1960 Arthur L. Schawlow and Charles H. Townes were awarded a patent for the invention of the laser. A working ruby laser operating at 690 nm was demonstrated in the same year. A year later in 1961, Townes would co-author a paper [18] proposing that exactly that which Cocconi and Morrison had dismissed so lightly was actually an imminently practical means of interstellar communication. In the years that followed, the

revolution in terrestrial communications that ensued as a result of the invention of the laser added weight to their words: visible wavelengths were discovered to be an excellent medium for transmitting information.

1.2 Optical communication

Communication techniques at visible wavelengths differ qualitatively from those used at microwave wavelengths. One of the most obvious differences is that there is very little transmission of signals at visible wavelengths "on the air"; rather, the signals are piped around in optical fibers. There are a number of reasons for this. First, signals at visible wavelengths do not have the penetration properties of radio signals (e.g. they don't go through walls). Furthermore, signals at visible wavelengths can be significantly degraded by the earth's atmosphere, in particular when passing through turbulent regions or weather. Finally, there is a great deal of background light at visible wavelengths which can interfere with the signal you are trying to transmit.

Naturally, the first two of these concerns do not affect signals traversing interstellar space, where there are no walls and no atmosphere. However, background light, in particular that which emanates from the star that the transmitting civilization is orbiting, will remain a concern for interstellar communication. Ignoring absorption lines, the Sun's spectrum can be approximated by a thermal blackbody at 5800°K, which reaches its maximum at visible wavelengths (one could argue that this is why evolution made these wavelengths visible). In order to communicate at wavelengths in the visible, the transmitting civilization must find a way to outshine their star, and the receiving civilization must find a way to distinguish starlight from signal.

The total power output from the Sun is about 4×10^{26} Watts, isotropicallly distributed, with most of this output in the visible part of the spectrum. The entire present power consumption of the human race is only about 10^{13} Watts [12]. If the transmitting civilization has access to similar resources then clearly they will not be able to outshine their star continuously in all directions. But one fact is immediately obvious: that for a fixed total energy the signal can be made arbitrarily brighter by

Name	Nova	NIF	Mercury	Helios	Nike	Sombrero
Date	1997	(2004)	1999	(2015)	1997	(2015)
Type	SSL	SSL	SSL	SSL	Gas	Gas
Gain medium	Nd:glass	Nd:glass	Yb:S-FAP	Yb:S-FAP	KrF	KrF
Pump	Lamp	Lamp	Diode	Diode	E-beam	E-beam
Pulse energy	~ 0.1 MJ	~ 2 MJ	0.1 kJ	~ 2 MJ	~ 2 kJ	~ 2 MJ
Pulse duration	$\sim$ nsec	$\sim$ nsec	$\sim$ nsec	$\sim$ nsec	$\sim$ nsec	$\sim$ nsec
Pulse rep rate	0.001 Hz	0.001 Hz	10 Hz	10 Hz	0.01 Hz	10 Hz
Wavelength	353 nm	353 nm	1047 nm	347 nm	248 nm	248 nm
Efficiency	0.1%	0.5%	10%	10%	$\sim$1.5%	$\sim$7%

Table 1.1: Properties of state-of-the-art high power lasers currently or soon to be available on Earth. Taken from [14]

making the duration shorter.

Just how short and bright could the pulse be? As is commonplace in SETI research, we look for examples on Earth at the risk of anthropomorphizing the transmitting civilization. Research into laser fusion has led to the development of incredibly powerful carbon dioxide lasers such as Helios [4] capable of delivering a 10 kJ, 0.5 ns pulse at 10.6 μm, or 2×10^{13} watts. The "National Ignition Facility" at Lawrence Livermore National Laboratory was set up by the Department of Energy with the goal of producing intense pressures and temperatures for simulating the conditions of a thermonuclear explosion [17]. The goal of the NIF is to produce a neodymium-glass laser (called "Nova") capable of producing 3–4 MJ in a 10 ns pulse, or about 3.5×10^{14} watts. These lasers and some others are listed in Table 1.2 which was taken from [14], and which shows the current state of the art in terrestrial lasers, as well as some projections into the near future.

It seems reasonable from the values in Table 1.2 to adopt as our model transmitter a laser capable of producing a 1 MJ pulse of 1 ns duration at 1 μm wavelength and of repeating this performance at 10 Hz. This implies a peak power output of 10^{15} watts, which is still eleven to twelve orders of magnitude short of outshining the sun. The rest will have to come from reciprocal bandwidth and directivity.

1.3 Directivity

Directivity is achieved by using a telescope. The illumination pattern for a circular aperture (such as a parabolic reflector uniformly illuminated by a plane wave) is the familiar Fraunhofer diffraction pattern

$$I(\theta) = 4 \left[\frac{J_1\left[(\pi D/\lambda)\sin\theta\right]}{(\pi D/\lambda)\sin\theta} \right]^2 \tag{1.1}$$

where D is the mirror diameter, λ is the wavelength, θ is the angle to the axis of the parabola, and $I(\theta)$ is normalized so that $I(0) = 1$. The half-power beam width can be found by solving

$$4 \left[\frac{J_1(x)}{x} \right]^2 = \frac{1}{2} \tag{1.2}$$

which has solution $x \approx 1.6134$. The full width at half maximum is just $2 \times x$

$$\frac{\pi D}{\lambda}\sin\theta = 3.2268 \rightarrow \sin\theta = 3.2268 \times \frac{\lambda}{\pi D} \rightarrow \theta \approx \frac{\lambda}{D} \tag{1.3}$$

giving a beam solid angle of

$$\Omega = \pi \left(\frac{\theta}{2} \right)^2 = \frac{\pi}{4} \frac{\lambda^2}{D^2} \tag{1.4}$$

The gain of a 100% efficient antenna is the ratio of the solid angle of an isotropic radiator to the solid angle of the antenna beam[13]

$$G = \frac{4\pi}{\Omega} = 16\frac{D^2}{\lambda^2} \tag{1.5}$$

For a telescope such as the Keck, with a diameter of 10 meters working at a wavelength of 1μm, $\theta \approx 10^{-7}$, which gives a beam solid angle of $\Omega = \frac{\pi}{4} \times 10^{-14}$ steradians or about 10^{-2} square arcseconds and a gain of

$$G = 16 \times 10^{14} = 152\,\text{dB} \tag{1.6}$$

4

Positions	0.77/0.64 mas (RA/dec)
Proper motions	0.88/0.74 mas/yr (RA/dec)
Parallaxes	0.97 mas

Table 1.2: Hipparcos median precisions

Now we can multiply the power output of our laser (10^{15} Watts) by the gain (16×10^{14}) and compare the result (16×10^{29} Watts EIRP[1]) to the output of the Sun, (4×10^{26} Watts) and discover that we have outshone our star by a factor of 4000 (36 dB).

Note that, in this analysis, we have not made any assumptions about the bandwidth of the receiver. In other words, the hypothetical transmitting apparatus outshines the star *even if the receiver admits flux at all wavelengths*. At visible wavelengths, the directivity that is achievable with large telescopes is so great that it is not necessary for the receiver to be wavelength specific.

1.4   Astrometry and Pointing

The use of a high-gain antenna puts a substantial burden on the transmitting civilization. Using a beam with an angular width of 10^{-7} radians means it is essential to know exactly where your target is in order to hit it.

Once again, we must use the best practice available here on Earth to estimate how well the other party could measure the positions of other stars such as our own. The best astrometric catalog produced so far on Earth is the Hipparcos Catalog [2], and the relevant precisions are shown in Table 1.2. Our requirement of a pointing precision of 10^{-7} radians (the width of the beam) translates to 20 mas (milli-arcseconds), and as can be seen from the table, the Hipparcos catalog gives positions that are much better than this.

However, the requirement is more demanding than just knowing the apparent positions of stars to high precision since the the stars in our galaxy are in motion with respect to each other. The apparent position of a star 100 light-years away

---

[1] "Equivalent Istropic Radiated Power"

is where it was 100 years ago, not where it will be 100 years hence when a signal arrives. The secular change in the angular position of the star on the celestial sphere is called the "proper motion" and is tabulated in star catalogs. Proper motions are difficult to measure to high accuracy since they can only be derived after many years of high precision observations of the positions of stars. As can be seen from Table 1.2, the ~ 0.75 mas error in proper motion accumulates to an error of the same size of the beam (20 mas) after about 25 years, so targets farther than 25 light-years away cannot be hit with the required accuracy.

Furthermore, knowing the positions and proper motions to high accuracy is still not enough. At a distance of 100 light-years (10^{18} meters), a beam with an angular width of 10^{-7} radians has only spread to cover an area with a diameter of 10^{11} meters, or about 1 AU (the distance from the Earth to the Sun). A typical value for the proper motion of a star at a distance of 100 ly ($= 31$ pc) is .25 as/yr, which corresponds to a velocity of about 7.75 AU/yr ($= 31$ pc $\times$ .25 as/yr) perpendicular to the line of sight. Therefore, at a distance of 100 light-years, a target star can be expected to spend no more than $1/7.75 = 0.13$ yr or about a month and a half in a beam that is pointed in a fixed direction with respect to the transmitting civilization's star. Since the starlight took 100 years to get from target to transmitter and the signal will take 100 years to get from transmitter to target, this means that the distance to the star must be known to a precision of about 0.065%.

Once again, Table 1.2 shows that on Earth the parallaxes of stars are only known to about 1 mas. Distances are derived from parallaxes as follows

$$R\,(\text{pc}) = \frac{1\,(\text{AU})}{p\,(\text{as})} \tag{1.7}$$

where R is the distance to the star in parsecs and p is the measured parallax in arcseconds. Therefore the error in the distance measurement is related to the error in the parallax measurement by

$$\Delta R = \frac{1}{p^2}\Delta p \rightarrow \frac{\Delta R}{R} = \frac{\Delta p}{p} \leq 0.00065 \tag{1.8}$$

Referring again to Table 1.2, we see that $\Delta p \approx 10^{-3}$ means that $p \approx 1.55$ as or $R \approx 0.65$ pc, or 2.1 light-years.

The uncertainties in position, proper motion and parallax can be combined as follows. Suppose ϕ is the apparent position of the target star now, μ is its proper motion (in arcseconds/year) and p is its parallax (in arcseconds). Then the distance to the star is $1/p$ (in parsecs) and the time the star light took to travel from the star to the observer is $1/pc$ (in years if c is measured in parsec/yr, i.e. $c = 1/3.26$). Therefore μ/pc is how far the star moved while its light was in transit so μ/pc must be added to the apparent position of the star to calculate where it is now and another μ/pc to calculate where it will be when the signal arrives. Therefore the transmitting telescope must be pointed at the position

$$\phi + \frac{2\mu}{cp} \tag{1.9}$$

for the signal to reach the star, where μ/cp has units of arcseconds. Standard error propagation methods give the pointing uncertainty σ in terms of the errors in ϕ, μ and p as

$$\sigma^2 = (\Delta\phi)^2 + \left(\frac{2\mu}{cp}\right)^2 \left[\left(\frac{\Delta\mu}{\mu}\right)^2 + \left(\frac{\Delta p}{p}\right)^2\right] \tag{1.10}$$

In equation 1.10, there are three terms which sum in quadrature due to uncertainties in position, proper motion and parallax respectively. We can calculate the relative importance of each of these terms as follows. Since the proper motion is a change in angular position, $\mu \propto 1/R$, where R is the distance to the star. For a star at a distance of 100 light years ($R = 31$ pc) the average proper motion $\mu = .250$ arcseconds per year, so in general $\mu = 7.75/R$. Using $p = 1/R$ this gives

$$\frac{2\mu}{cp} \approx 50 \text{ as} \tag{1.11}$$

Substituting this and $\Delta\phi = \Delta\mu = \Delta p \approx 1$ mas (from Table 1.2) into equation 1.10 gives

$$\sigma^2 = (10^{-3})^2 \left[1 + (50)^2 \left[\left(\frac{R}{7.75}\right)^2 + \left(\frac{R}{1}\right)^2\right]\right] \tag{1.12}$$

7

The expression in square brackets above contains the three terms due to the uncertainties in position, proper motion and parallax, respectively. As expected, the term due to the uncertainty in parallax is the largest in the error. Furthermore, the terms due to uncertainties in proper motion and parallax grow proportional to the distance between transmitter and receiver; this effectively sets the transmitter's range. Taking into account all of the error terms and using the best astrometry currently available on Earth gives a transmitter range of ~ 0.35 parsecs or 1.15 light years. The closest star to our Sun, Alpha Centauri, is at a distance of 4.3 light years.

However, astrometry is not a static field here on Earth, and advances made during the next ten to twenty years could significantly increase our transmitting range. For example, the astrometric sub-array at the Navy Prototype Optical Interferometer in Flagstaff, Arizona is designed to produce star positions accurate to a few milliarcseconds with a limiting magnitude of 10 [11]. This does not improve significantly on the Hipparcos accuracy, but since the NPOI is ground-based it can make observations over a much longer period of time than the Hipparcos satellite could, which significantly improves measurements of the proper motions of stars. A quantum leap in astrometric precision is expected after the launch of NASA's Space Interferometry Mission (SIM) which has a design goal of making astrometric measurements with a precision of 4 microarcseconds [16]. This has the potential to improve the proper motion and parallax measurements, and therefore the transmission range, proportionally, i.e. by about three orders of magnitude.

Knowing where to point is not enough, however; you must also be able to build a telescope with the required mechanical pointing precision. The typical pointing quality of a modern major astronomical telescope (such as the MMT on Mount Hopkins in Arizona) is 1.5 arcseconds, or about 7.5×10^{-6} radians. The pointing requirement of 10^{-7} radians is therefore about 75 times better than typical terrestrial practice. Nonetheless, it is reasonable to believe that an improvement in pointing accuracy of two orders of magnitude is within the reach of a determined engineer. Improvements in pointing beyond the "seeing disk" of about 0.25 arcseconds (at very good astronomical sites) are expensive and difficult to justify if the objective is to observe

but not to transmit. Therefore, there is reason to believe that modern astronomical telescopes do not point as well as they could, but only as well as they have to.

Interstellar communication at optical wavelengths is not something that human civilization is capable of at this point in our history. Interestingly, the limitation is not the power of the lasers or the diameters of the telescopes we can construct, but rather our limited knowledge of positions, distances and proper motions of the stars in our neighborhood. Improvements in the measurements of these quantities over the next ten to twenty years are expected to lift this limitation.

1.5   Adaptive Aperture

So far the working model for the transmitter's strategy has assumed that the objective is to outshine the star by a fixed ratio. This is not necessarily the optimal transmission strategy. The hypothetical laser capable of delivering 1 MJ in 1 ns through diffraction limited beam from a 10 meter aperture results in a signal-to-noise ratio of 36 dB that is far higher than is typical in communications.

Suppose instead that the pulse energy and the beam size on target are held fixed. The angular size of the beam is $\theta = \lambda/D$ where D is the diameter of the transmitter's aperture which will be adjusted for a constant beam size on the target. The beam size on target is simply

$$k = R\frac{\lambda}{D} \tag{1.13}$$

where R is the range to the target. To hold k fixed the transmitter simply adjusts D so that $D = \lambda R/k$.

In this case, instead of having a constant signal-to-noise ratio at every target, the photon fluence is fixed instead. The total number of photons in a 1 μm, 1 MJ, 1 ns pulse is

$$N_p = \frac{E\lambda}{hc} = \frac{1}{hc} = 5 \times 10^{24} \tag{1.14}$$

If the size of the beam at the target is held fixed at 10 AU (a radius much larger than

the habitable zone in our solar system) then the photon fluence is given by

$$\frac{5 \times 10^{24}}{1.7 \times 10^{24}} \approx 3 \, \mathrm{m}^{-2} \tag{1.15}$$

within 1 ns. For a signal to noise ratio of at least one, the number of laser photons must be equal to or greater than the number of stellar photons. This sets a minimum distance between transmitter and receiver. The receiver must be far enough away from the transmitter's star so that the stellar photon flux has fallen below $3 \times 10^9 \, \mathrm{s}^{-1} \mathrm{m}^{-2}$. The photon flux from the Sun on the Earth is about $10^{20} \, \mathrm{s}^{-1} \mathrm{m}^{-2}$ at a distance of 1 AU, and it falls off proportional to $1/R^2$ with increasing distance. The flux from the Sun is down to the required level at a distance of about $3 \times 10^5 \, \mathrm{AU} \approx 1 \, \mathrm{pc}$, which is less than the distance to the closest star from the Sun (Alpha Centauri at 1.3 pc). Therefore this minimum distance would not be a limitation in practice.

Using this technique, the angular size of the beam is not fixed, but varies with the range to the target as

$$\theta = \frac{\lambda}{D} = \frac{k}{R}. \tag{1.16}$$

It is still a requirement that the angular size of the beam, θ, must be larger than the typical astrometric error σ in order for the transmitter to hit the target. However, with an adaptive aperture both the beam size and the astrometric error vary with range R in such a way that the range is greatly extended.

To see how this works, recall that the pointing requirement is that the angular size of the beam must be greater than the astrometric error

$$\theta = \frac{k}{R} \geq \sigma \tag{1.17}$$

If k is measured in AU and R in parsecs then σ has units of arcseconds. At large distances, the uncertainties in proper motion and parallax dominate σ and it grows $\propto R$ (see Equation 1.10). Let ϵ be the "fundamental angular measurement error", for example 10^{-3} arcseconds for the Hipparcos data (see Table 1.2). Then at large R

Equation 1.10 can be approximated by

$$\sigma \approx (56R)\epsilon \tag{1.18}$$

and we have

$$\frac{k}{R} \geq \sigma \Rightarrow \frac{k}{56R^2} \geq \epsilon \Rightarrow R \leq \sqrt{\frac{k}{56\epsilon}} \tag{1.19}$$

To get some idea of the scale of R, we can plug in $k = 10\,\mathrm{AU}$ and $\epsilon = 1\,\mathrm{mas}$ and find

$$R \leq 13.4\,\mathrm{pc} = 44\,\mathrm{ly} \tag{1.20}$$

This is a substantial improvement in range over the values derived in the previous section (0.35 parsec = 1.15 light years) assuming a fixed angular beam size. Clearly, the adaptive aperture is favored if range is the figure of merit for a transmission strategy. Since the number of potential communicators grows as the range cubed, there is every reason to believe that this is the case.

The range of either strategy can be increased to an arbitrarily large amount if the transmitter chooses to sweep the beacon over the range of angles bounded by the pointing uncertainty. The penalty for using this approach is an increase in the amount of time required to reach all targets within a certain distance. The number of potential targets at a given distance R is proportional to R^2 and the pointing uncertainty is proportional to R. The area that the telescope must cover is proportional to the pointing uncertainty squared. Therefore, the time required for the transmitter to illuminate all the targets at a distance R is proportional to R^4, and the time required to illuminate all targets within a distance R grows as R^5. Therefore, this technique does not seem like a practical method coping with large astrometric error if the transmitter's goal is to reach a large number of stars, and it is seen that astrometric errors still set the outer limit on communication range.

Table 1.3: Extinction in the Johnson-Cousins Photometric Bands

Band	λ	$A(\lambda)$	$T(\lambda)$
U	0.365	1.560	0.238
B	0.440	1.310	0.299
V	0.550	1.000	0.398
R	0.700	0.749	0.502
I	0.900	0.479	0.643
J	1.250	0.282	0.771
H	1.650	0.176	0.850
K	2.200	0.108	0.905

1.6 Extinction and scattering: Interstellar Dust

Studies of pulsars have shown that radio pulses will be broadened as they traverse interstellar space due to fluctuations in the plasma electron density in the interstellar medium [7]. Visible and infrared radiation are essentially immune to these effects; however, shorter wavelengths will be affected by scattering and absorption by interstellar grains. These grains are mostly carbon and silicate dust particles a fraction of a micron in size which will scatter and absorb light of commensurate wavelengths. The density of dust in the galaxy is not uniform, and the resulting scattering and absorption depends a great deal on the galactic latitude and longitude of the source[1]; in what follows, typical values are used.

Considering first just the absorption, the fraction of radiated power remaining after one kiloparsec can be expressed as

$$T(\lambda) = 10^{-\frac{2}{5} \times A(\lambda)} \tag{1.21}$$

where $A(\lambda)$ is the extinction in magnitudes at wavelength λ (in μm) and $A(0.550) = 1.0$.[1] [15] [6]. Table 1.3 shows data taken from [3]; these data are plotted in Figure 1.1. As can be seen from both table 1.3 and figure 1.1, the transparency of the interstellar medium increases with increasing wavelength. This trend continues until the galaxy becomes essentially transparent at 10 μm.

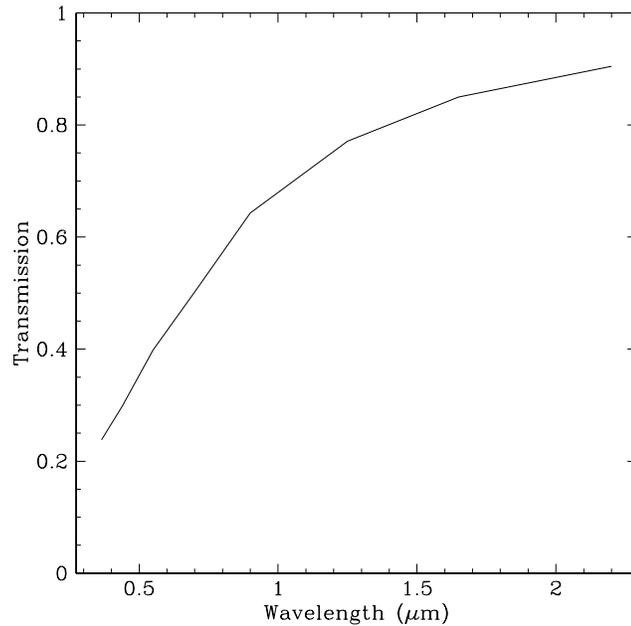Visible light is also scattered by interstellar grains. For wavelengths between

Figure 1.1: Transmission vs wavelength: showing the fraction of light remaining after traversing one kiloparsec of interstellar space as a function of wavelength.

0.03 and 1 μm, the effects of scattering on short pulses can be divided into three components [6]:

1. an unscattered component

2. a forward scattered component due to large grains

3. a diffuse component due to wide-angle scattering

An instrument sensitive to narrow pulses will only detect the unscattered component. An optical pulse is significantly distorted by *any* scattering; a scattering diameter of one arcsecond for a source at ten kiloparsecs would broaden a pulse to over two seconds [6]. Studies of scattered starlight (either the diffuse galactic light scattered in the galactic disk or reflection nebulae) have determined that the albedo (ratio of scattering to absorption cross sections) of the dust is about ~ 0.5.[1], so scattering is not a negligible effect compared to absorption.

The conclusion is that optical pulses may only be detectable from sources relatively nearby, say within a kiloparsec. This is probably less restrictive than the astrometric requirement. Here on Earth, the quality of our astrometry would limit us to transmitting within a distance of about 13.4 parsecs (see Sections 1.4 and 1.5). Nonetheless, it would seem more worthwhile to focus search efforts on stars within a kiloparsec, and this is what we have done.

1.7 Infrared

There are other reasons besides the transparency of the interstellar medium to prefer the infrared for interstellar communications. For a fixed laser energy E the number of photons produced goes as

$$N_{\text{laser}} = \frac{E}{h\upsilon} = \frac{E\lambda}{hc} \propto \lambda \tag{1.22}$$

whereas for a Sun-like star (temperature $\sim 10^3 - 10^4\,^\circ\text{K}$), the infrared wavelengths are well into the Rayleigh-Jeans part of the blackbody spectrum where the radiated power within a given bandwidth goes as

$$P \propto \frac{1}{\lambda^2} \tag{1.23}$$

and therefore the number of photons from the star goes as

$$N_{\text{star}} \propto \frac{1}{\lambda} \tag{1.24}$$

and the signal to noise ratio for a fixed laser energy as

$$\frac{N_{\text{laser}}}{N_{\text{star}}} \propto \lambda^2, \tag{1.25}$$

pushing the advantage back toward longer wavelengths.

These advantages have to be traded off against the advantages of shorter wave-

lengths, of course. As has already be shown, the gain of the transmitting antenna goes $\propto 1/\lambda^2$, an effect which pushes the advantage back toward shorter wavelengths. At much longer wavelengths ($\nu \sim 1\,\text{GHz}$), the galactic synchrotron radiation produces considerable background noise, and the dilute plasma of free electrons in the galaxy creates a dispersive medium which broadens pulses. Therefore the range of wavelengths is contrained on the long end as well. Townes has made an argument based on noise lower bounds enforced by quantum mechanics that the optimal wavelength depends on the receiver construction, with heterodyne receivers favoring wavelengths around $3\,\text{m}$, and photon counting detectors favoring wavelengths around $1\,\mu\text{m}$[19]. Since no frequency band can be uniquely and unambigously identified as optimal for interstellar communication, it seems that the best strategy is to search all of them as the technology to do so becomes available.

1.8   Backgrounds

Even if a signal successfully traverses the hazards of interstellar space, there remains the possibility that it would not be recognized as intentional because of some sort of natural astrophysical or atmospheric phenomenon. The issue of astrophysics on very short timescales has been investigated by Dravins [8]; the conclusion is that there are no known astrophysical phenomena which could cause nanosecond pulses of visibile radiation of the required intensity. The requirements for a natural astrophysical system to generate a nanosecond pulse of the required intensity are quite demanding. It must either be coherent, or at most centimeters in size and emit more than a solar luminosity in equivalent isotropic radiated power in visible radiation. Although coherent astrophysical sources of the required luminosity do exist (such as the water maser in the galaxy NGC4258), none of these emit short pulses in the visible spectrum.

Another possible background are cosmic rays and the detritus that remains after they collide with the upper atmosphere. Approximately 75% of the particles that survive to reach sea level are muons [9], with a mean energy of 2 GeV and a differntial energy spectrum proportional to E^{-2} up to about 1 TeV where it steepens to

$E^{-3.6}$. The total flux is about $10^2 \, \text{m}^{-2} \text{s}^{-1} \text{sterad}^{-1}$ The muons have an zenith angle distribution proportional to $\cos^2 \theta$ which transitions to $\sec \theta$ at 1 TeV.

A charged particle radiates if its velocity is greater than the phase velocity of light in the medium through which it propagates. This Cherenkov radiation is beamed down a narrow cone with an opening angle $\theta_C = \arccos\left(\frac{1}{\beta n}\right)$ where β is the speed of the particle relative to the speed of light in vacuum. The muon mass is about 106 MeV and the index of refraction in air at one atmosphere is 1.000293, which means muons with energies greater than about 4.4 GeV ($\beta = 0.999707$) will emit Cherenkov radiation as they pass through the atmosphere.

Fortunately, the induced Cherenkov pulse is too diffuse to be collected by a telescope with a narrow angle of admittance. Although a typical TeV muon does produce a short (5 ns) optical pulse with about 30 photons/m^2 falling on the base of the narrow light cone (~ 150m on the ground), the source appears diffuse with a FWHM of about 2°, much larger than the field of view of the telescope which would only see $\sim 2 \times 10^{-4}$ photons per flash [10].

1.9   Pileup

Another possible background are statistical fluctuations in the flux of stellar photons which could accidentally "pile up" during a short enough period of time to be confused with an intentional pulse. We can compute exactly how often to expect accidental pileups as a function of the photon rate.

Let $u = rt$ be the average number of photons expected during a time t if the overall average rate is r photons per second. The Poisson distribution gives the probability of exactly n photons

$$P_n(u) = \frac{u^n e^{-u}}{n!} \tag{1.26}$$

and therefore the probability of m or more photons is given by

$$Q_m(u) = \sum_{n=m}^{\infty} P_n(u) \tag{1.27}$$

Since $P_n(u)$ is a properly normalized probability we have

$$1 = \sum_{n=0}^{\infty} P_n(u) = \sum_{n=0}^{m-1} P_n(u) + \sum_{n=m}^{\infty} P_n(u) \qquad (1.28)$$

and therefore

$$Q_m(u) = 1 - \sum_{n=0}^{m-1} P_n(u) = 1 - e^{-u} \sum_{n=0}^{m-1} \frac{u^n}{n!} \qquad (1.29)$$

An approximate value for Q_m can be estimated by noticing that the second term on the right hand side of equation (1.29) is the product of two factors, e^{-u} and

$$\sum_{n=0}^{m-1} \frac{u^n}{n!} \qquad (1.30)$$

which is an $m - 1$ order approximation to e^u. In particular,

$$\sum_{n=0}^{m-1} \frac{u^n}{n!} = e^u - \frac{u^m}{m!} + O(u^{m+1}) \qquad (1.31)$$

and therefore

$$Q_m(u) = 1 - e^{-u} \sum_{n=0}^{m-1} \frac{u^n}{n!} \approx 1 - e^{-u} \left(e^u - \frac{u^m}{m!} \right) = \frac{u^m e^{-u}}{m!} = P_m(u) \approx \frac{u^m}{m!} \qquad (1.32)$$

which shows that, to leading order in u, the probability of m or more photons is equal to the probability of *exactly* m photons.

In our experiment (described in detail in the next chapter), the light gathered by the telescope is divided in half by a beamsplitter, and each half is used to illuminate a detector. For our purposes, the question of pileup can be phrased precisely: if the photon arrivals from the star at our telescope are goverened by the Poisson distribution for some rate r, what is the rate of events in which we get two or more photons on each detector simultaneously?

Assuming a 50-50 beamsplitter, the requirement is for at least four photons (two for each of two detectors) to come into the telescope during a detector resolving time t. If the single photon rate is r, each single photon event represents an opportunity for three-or-more photons to arrive during the interval t immediately after. Therefore

the rate of four-or-more photon events must be

$$r \times Q_3(rt) \tag{1.33}$$

where Q_3 is given, both exactly and approximately, by equation (1.32). This result is the rate of four-or-more photon events in the telescope; it must be multiplied by an overall factor representing the effect of the beamsplitter to get the rate of two-or-more photon events in both of two detectors.

As was shown in equation (1.32), the probability of m-or-more photon events, Q_m can be approximated to leading order in rt by the probability of exactly-three-photon events, P_m. In this case, the rate of four-or-more photon events can be approximated to leading order in rt by the rate of exactly-four-photon events. Each of these four photons has a choice of two paths to take at the beamsplitter, and for a 50-50 beamsplitter both choices are equally likely. The beamsplitter factor then is the same as the probability of getting exactly two heads in four tosses of a fair coin: the probability of a given configuration of two heads and two tails, $(1/2)^4$, multiplied by the number of such configurations that are possible:

$$\begin{pmatrix} 4 \\ 2 \end{pmatrix} \left(\frac{1}{2}\right)^4 = \frac{3}{8} \tag{1.34}$$

The final result for the rate of two photon pileups in both detectors is therefore, to leading order in rt (using equation (1.32))

$$\frac{3}{8} r \frac{(rt)^3}{3!} = \frac{r^4 t^3}{16} \tag{1.35}$$

For comparison, we can calculate the rate of simultaneous two-photon events in two beams of intensity $r/2$. The rate of two-photon events in one beam of intensity $r/2$ is found by modifying equation (1.33), $r/2 \times Q_1(rt/2)$. Each two photon event in one beam represents an opportunity for two or more photons to arrive in the other beam within a time t, which has a probability of $Q_2(rt/2)$. Since the beams are

18

interchangeable, there is an additional overall factor of two:

$$2 \times r/2 \times Q_1(rt/2) \times Q_2(rt/2) \approx 2 \times \frac{r^4 t^3}{2^4 2!} = \frac{r^4 t^3}{16} \tag{1.36}$$

Which is the same result derived in equation (1.35).

It is interesting to see if this result can be generalized to the case of simultaneous m-photon pileups in both of two detectors. In other words, for an arbitrarily large photon pileup, is it the same to consider a beam of intensity r divided by a 50-50 beamsplitter onto two detectors as to consider two beams of intensity $r/2$ each with its own detector?

Considering first the case of $2m$ photons in the telescope being divided up in a 50-50 beamsplitter, the rate of $2m$ pileups in the telescope is

$$r \times Q_{2m-1}(rt) \approx \frac{r^{2m} t^{2m-1}}{(2m-1)!} \tag{1.37}$$

and the beamsplitter factor is given by

$$\binom{2m}{m} \left(\frac{1}{2}\right)^{2m} = \frac{(2m)!}{(m!)^2 2^{2m}} \tag{1.38}$$

giving an overall rate of

$$\frac{2m}{(m!)^2} \times \frac{r^{2m} t^{2m-1}}{2^{2m}} \tag{1.39}$$

Now consider two simultaneous m-photon pileups in two beams of intensity $r/2$. The rate of m-photon pileups in one of the beams is

$$r/2 \times Q_{m-1}(rt/2) \approx \frac{r^m t^{m-1}}{2^m (m-1)!} \tag{1.40}$$

and the probability of m or more photons in the other beam is

$$Q_m(rt/2) \approx \frac{r^m t^m}{2^m m!} \tag{1.41}$$

19

There is an overall factor of 2 since the beams are interchangeable, giving a rate of

$$\frac{2}{m!(m-1)!} \times \frac{r^{2m}t^{2m-1}}{2^{2m}} = \frac{2m}{(m!)^2} \times \frac{r^{2m}t^{2m-1}}{2^{2m}} \tag{1.42}$$

which is exactly the same result as in equation (1.39) for the case of single beam divided by a beamsplitter.

Going back to the four photon case, for a typical photo-electron rate of 10^4 per second and a detector resolving time of twenty nanoseconds, we get a pileup rate of $(10^4)^4(20 \times 10^{-9})^3/16 = 5 \times 10^{-9}$ per second, so on average such a pileup event will happen every 2×10^8 seconds or about every six and a third years of integration time. It can therefore be concluded with great confidence that accidental pileup of two photons in both detectors essentially never happens.

1.10 Transmission Strategy

The hypothesis we are testing is that an ETI might choose to initiate an interstellar communication channel with its neighbors by illuminating them with a powerful, narrowly focussed laser. The advantages compared to radio wavelengths for the transmitter are higher bandwidth, higher gain antennas and lower dispersion in the interstellar medium. The advantages to the receiver are the ease with which a receiving apparatus can be built, and the fact that it can be a broadband receiver so that a carrier frequency does not have to be guessed.

Suppose then, that the ETI illuminates all of the F, G and K -type dwarf stars within 100 light-years of its own. In the neighborhood of our Sun, this means about 1000 stars. The most advantageous strategy for the transmitter would be to use an adaptive aperture to keep the beam size fixed at the target. In doing so, the transmitter can be sure to outshine its own star by a ratio of from 1 to 35 dB. If the transmitting apparatus has a 10 Hz repetition rate, then all of the 1000 nearby stars can be cycled through in 100 seconds, producing a pattern of bright flashes that would be recognizeable to an observer of the transmitter's star.

The receiver's problem, then, is to construct an apparatus that can distinguish between these bright flashes and the background light of the transmitter's star. That is the objective of our instrument.

Chapter 2

Experiment

In the previous chapter a scenario for interstellar optical communication was described in which the transmitting civilization is obliged to create lasers of enormous power and point them with pinpoint precision. By doing so, they make the receiving civilization's job much easier: it must only be able to distinguish a short pulse of laser light from the background produced by the transmitter's star. It is this objective that has guided the design of our experiment.

2.1 Hybrid Avalanche Photodiodes

The property of an ultrashort laser pulse that distinguishes it from the background light from the star is that the former delivers a relatively large number of photons in a very short period of time whereas the latter almost never does (the possibility of stellar photon pileup is discussed at length in the previous chapter). Therefore, to begin with, we need a detector with a short resolving time that is capable of cleanly distinguishing between multiphoton and single photon events.

The hybrid avalanche photodiode is just such a detector. An HAPD consists of an avalanche photodiode behind a photocathode (as seen by incoming photons) with a vacuum between them and a high voltage (-7500 V) across them. Photoelectrons emitted by the photocathode are accelerated by this field and injected into the avalanche photodiode, where they deposit their energy in the depletion layer creating
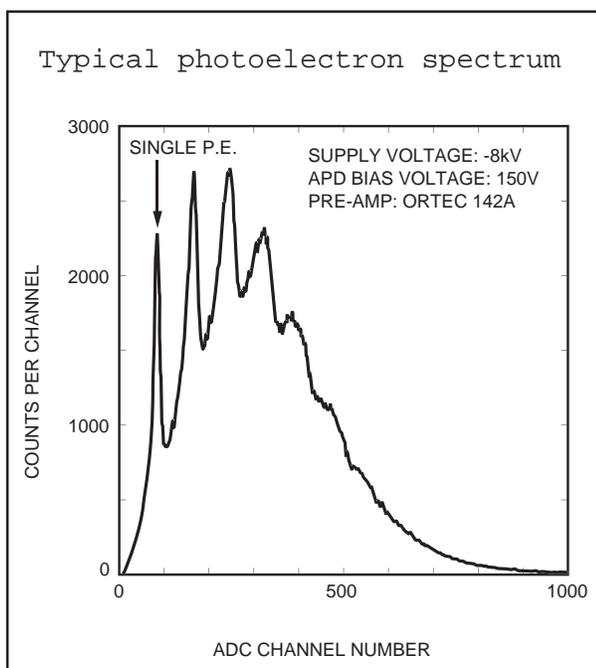
Figure 2.1: R7110U hybrid APD pulse height distribution

electron-hole pairs. Since the excitation energy of one electron-hole pair is 3.6 eV, a gain of about 2100 is expected. In addition to this, there is an avalanche multiplication in the diode which gives another factor of 20 gain, for about 46 dB total gain in the device. The important property of the HAPD, and what distinguishes it from a conventional photomultiplier tube, is that the actual gain is does not vary much from the expected gain.

The HAPD chosen for this experiment is the Hamamatsu R7110U. The R7110U is a compact cylinder about two centimeters in diameter and two centimeters high. These devices had only recently become available on the market at the time the experiment was being developed, and to some extent the experiment served as a beta test for the devices. Figure 2.1, taken from the manufacturer's catalog, shows a pulse height distribution for the R7110U phototube. It is clear from this Figure that a discriminator set at the same voltage as ADC channel 125 would cleanly distinguish between single photon events and multiphoton events.

Like all photodetectors, the HAPDs have a background "dark count" (signal in
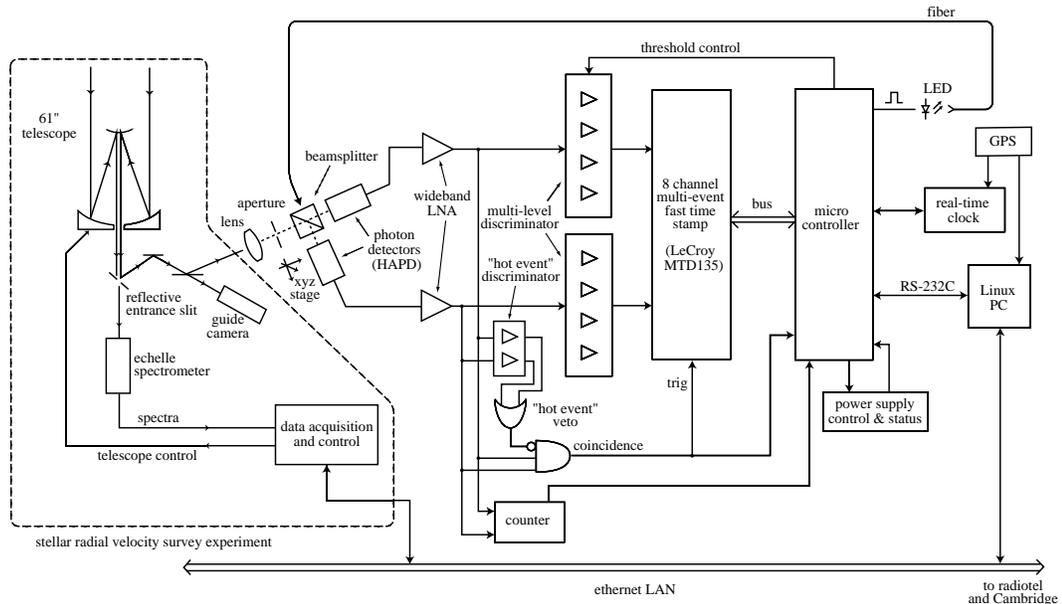
Figure 2.2: Block Diagram of the Experiment

the absence of any light). In addition, probably due to the high voltage bias applied on such a small package detector, there are frequent (5–10 Hz) "hot events": huge bipolarity outputs with a ringing aftermath extending up to several hundred nanoseconds. Because of this hot event background, it would be very difficult to notice an occasional transient signal by looking at the output of a single detector. Instead, we divide the light from our telescope with a beamsplitter, illuminate two identical detectors, and require a simultaneous output from both detectors for an event to be considered real. We have also implemented a hot event veto (described in detail later) that takes advantage of the fact that the hot events are bipolar but the photo-electron events are not.

2.2 Implementation

The block diagram in Figure 2.2 shows the experiment. The 61-inch Wyeth Telescope at Harvard University's Agassiz Station Observatory in the town of Harvard,

24

Massachusetts has been used for a stellar radial velocity survey for the past twenty years. The radial velocity survey uses an echelle spectrograph to measure the spectra of stars, and due to the narrow slit on the entrance to this instrument, a large fraction of the light collected by the telescope for the radial velocity survey was (and still is) thrown away. Our experiment piggy-backs on the radial velocity survey by using most of this previously wasted light. The arrangement has proven to be very advantageous because of the operational support by the people who run the radial velocity survey (see the Acknowledgements section).

As shown in the block diagram, roughly a quarter of all of the light collected by the 61-inch telescope ends up in our instrument (the "OSETI box"). The light in the OSETI box falls on a beamsplitter (after passing through a focusing lens and a focal-plane aperture that defines the sensitive area and reduces the amount of sky light that falls on the detectors) whose outputs fall on two HAPDs. This is the heart of the experiment, and all of the design decisions follow from it: light from the telescope illuminates two detectors that are capable of cleanly distinguishing a multiphoton event from a single photon event. If both detectors simultaneously register a multiphoton event, the circumstances of this event are stored for later analysis. Everything else in the experiment is subsidary to, and in support of, this function.

The implementation of the coincidence detection is as follows (and is shown in the block diagram, Figure 2.3). The outputs of the HAPDs are passed through two gain stages (one inverting and one non-inverting) and then each is input into an identical array of four comparators. The comparator threshholds within an array are set at increasing voltages which correspond to two, four, eight and sixteen photo-electron events at the detectors. The outputs of the lowest comparators are AND-ed together to generate the coincidence trigger.

As has already been discussed, the HAPDs produce a fairly high number of "hot events," which have the property that the output spikes and then rings through a few oscillations before returning to normal. These oscillations can be used to implement a hot event veto: if either detector output goes negative after a positive pulse, then
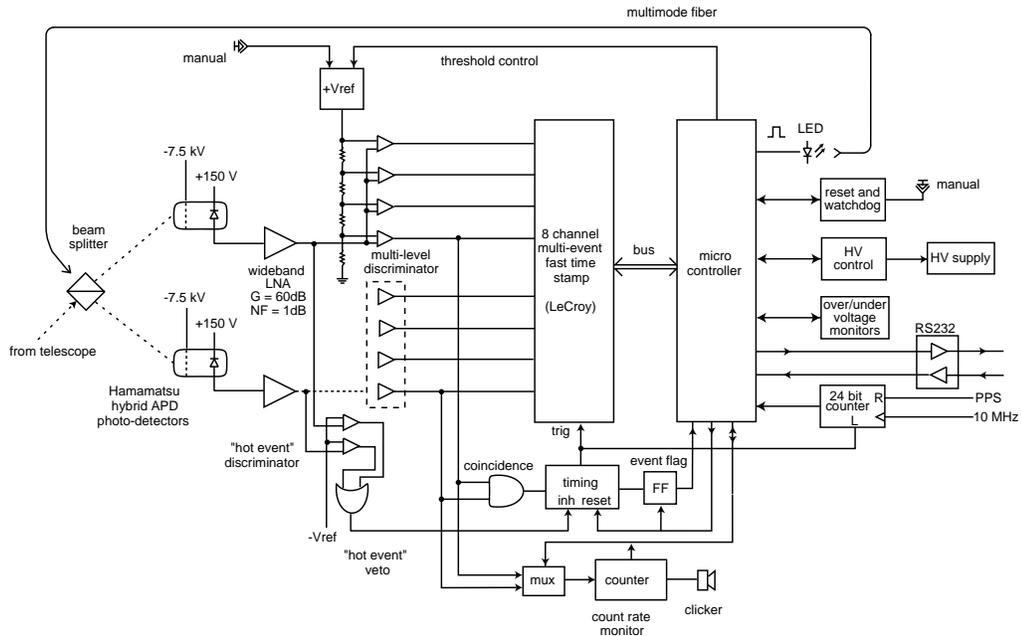
25

Figure 2.3: Electrical Block Diagram

that event is vetoed. The implementation of the hot event veto is simply another pair of comparators (one for each detector) which compare the detector outputs to a negative voltage reference. These comparator outputs generate the hot event veto.

In general, we would like to have more information about these events than simply that they happened and they weren't vetoed. This is the reason for implementing an array of four comparators per detector instead of just using a single pair set for the lowest threshold. We want to know how high the output voltages went and when they went there. This is the function of the LeCroy MTD135 eight channel fast timestamp ASIC. The MTD135 is an extremely fast (capable of a time resolution of 0.5 nanoseconds, although we use 0.625 nanoseconds) chip that generates timestamps every time one of its inputs makes a high-to-low or low-to-high transistion. The MTD135 has storage capacity to hold the last sixteen timestamps on each channel. The coincidence trigger, which is the AND-ed output of the lowest comparator from each comparator array delayed by about 300 nanoseconds, stops the MTD135, freezing the last few timestamps in memory to be read out by the microcontroller. The
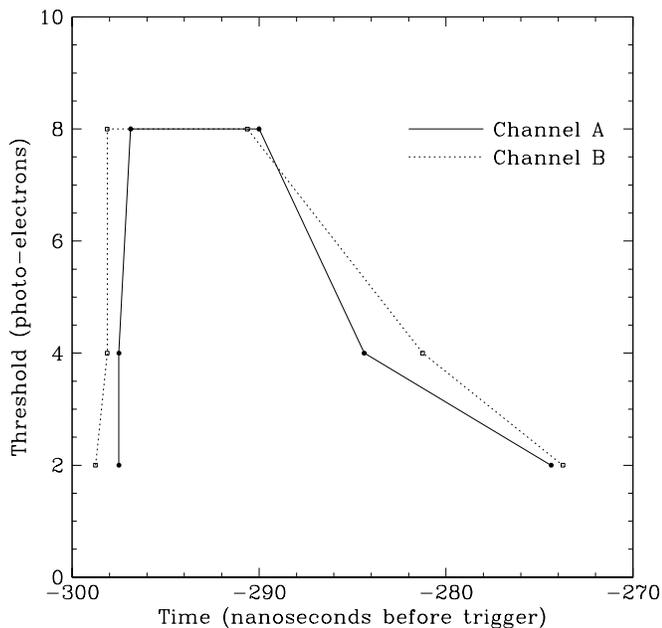
Figure 2.4: Coincidence reconstructed from MTD135 timestamps

microcontroller passes these timestamps on to the computer where the event is logged.

Figure 2.4 shows a typical event, recorded during an observation of the star HD285830 on November 1, 1998 (night of October 31 – November 1) at 08:45:54 UTC, reconstructed from the MTD135 timestamps. In this case, the output of both detectors crossed the comparator threshholds set for two, four and eight photoelectrons about 298 nanoseconds before the trigger event which stopped the MTD135. The output fall time for the detector is seen to be somewhat slower than the rise time, as expected.

2.3   Diagnostics

In addition to the basic functionality of being able to detect and record coincident events from both detectors, the experimental apparatus includes circuitry for self-diagnosis and monitoring. In fact, each observation begins with a series of diagnostic tests to ensure that the instrument is functioning normally. If any of these tests fail,

then the observation does not proceed.

It is worth digressing on the subject of the design philosophy that guided the firmware (microcontroller program) and its relationship to the software that runs on the computer and controls the experiment (the so-called "OSETI daemon"). The firmware design is "transactional", by which it is meant that commands are received from the computer (via a standard RS-232 serial port), executed immediately, and their results returned to the computer (via the same RS-232 serial port). The microcontroller does nothing autonomously except to report coincidences when they happen. So, for example, if one of the power supply voltage monitors indicates that the supply has gone out of its nominal range of acceptable values, the microcontroller will nonetheless start an observation *if the computer instructs it to do so*. Furthermore, all of the commands and responses are ASCII encoded, so that it is possible to perform diagnostics on the apparatus by connecting a dumb terminal to it and simulating by hand the commands that the OSETI daemon would have generated, and then reading the microcontroller's response to these commands.

The diagnostics phase of the observation consists of a set of transactions between the OSETI daemon program on the computer and the microcontroller in the apparatus. To begin with, the computer must be certain that there is a microcontroller listening on the serial port to deal with the commands it will send; to do this it just sends an empty line (a single ASCII carriage return). The microcontroller should respond with a string containing the name and version number of the apparatus, as of this writing "`OSETIbox 2.0`."

The apparatus contains six power supplies providing ±5 volts, +24 volts, +15 volts, -7500 volts (the photocathode supply) and +150 volts (the avalanche photodiode supply). Five of these supplies are monitored by MC34161 universal voltage monitors. These ICs contain "window comparators," which are really two comparators with open collector outputs and with thresholds above and below the nominal supply voltage defining an acceptable range for the supply, which indicate when the supply is within this range. The outputs of the monitors on the +24, +15 and -5 volt supplies are wired together to produce a single output which is available to the

microcontroller. The other two monitors provide independent information on the status of the +5 and -7500 volt power supplies; the +150 volt supply is not monitored. Therefore, in total there are three bits returned by the microcontroller in response to a query for voltage monitor status: the first gives the status of the +5 volt supply, the second gives the combined status of the +24, +15 and -5 volt supplies and the third gives the status of the -7500 volt supply.

If the microcontroller is responding and the power supplies are all within their nominal range, the next (perhaps most important) test is to check to see if the photodetectors can see light. As can be seen in the electrical block diagram (Figure 2.3), the microcontroller can flash an LED whose output is carried by an optical fiber to the otherwise unused fourth port of the beamsplitter. This generates a flash in both of the two detectors, which should be reported as a coincidence by the microcontroller.

If all the tests up to this point are passed successfully, then we can be confident that all of the essential components of the system are functioning properly. The next two diagnostics provide information about the instrument performance on the particular target on the particular night that the observation takes place. As can be seen in the block diagram, Figure 2.3, the microcontroller can control the comparator thresholds, in particular, it can lower them so that the lowest comparator threshold is at the level expected for a single photoelectron. The outputs of the two lowest comparators are steered through a multiplexer (also under the control of the microcontroller) into a sixteen bit counter, which can be read out by the microcontroller.

The procedure is to lower the thresholds, wait for 0.3 seconds and then read out the counters. This "low threshold count rate" gives some information about how bright the star is. Next, the thresholds are returned to their normal, higher levels (where the lowest comparator threshhold corresponds to two photoelectrons) and once again the count rate is measured by allowing the counters to accumulate for 0.3 seconds and then reading them out. This is the "high threshold count rate". In all of the development that follows in this thesis, "count rates" will always refer to the measurements made during this phase of the diagnostics, before the integration on the star begins.

2.4   Computers

The communication that takes place between the microcontroller and the Linux PC, which runs the OSETI experiment, is recorded in an ASCII log file stored on the Linux PC. This computer also talks to the Sun workstation that runs the radial velocity survey so that it knows when the telescope is on target, and what the target is. That information is transmitted to from the radial velocity workstation to the Linux PC over the observatory's local area network at the beginning of the observation, and is, in fact, the event that triggers the diagnostics phase and the integration that follows.

The software that runs on the Linux PC (the "OSETI daemon") has been kept very simple. The minimum requirements are that it must be able to talk to the instrument through a serial port, to the main observatory computer over the network, and it must record everything it does in a file for later analysis. Everything has been designed for maximum transparency: the serial protocol, network protocol and log file format are all in human-readable ASCII, a design choice which makes it easy to do diagnostics on any one of these subsystems independently by simulating the others. For example, to test the OSETI daemon, you do not need to involve the radial velocity survey's Sun workstation even though this is normally how the daemon is activated. Instead, you can simply telnet to the TCP port that the daemon is listening on from any computer (including the OSETI Linux PC itself) and type the magic activation word "start". Furthermore, to test the OSETI instrument itself, the involvement of the OSETI Linux PC is not required; you can simply connect a dumb terminal (or another computer running a terminal emulator) and type the commands that the OSETI daemon would have sent. And finally, no special software is required to read and understand the English text in the log file.

Figure 2.5 is a flow chart showing how the OSETI daemon works. Actions taken by the daemon itself are in boxes and events outside of the daemon that can cause it to respond label the arrows. Briefly, the way the system works is that the OSETI daemon listens for incoming connections from the network on TCP port 8001 (the box labelled "listening"). After accepting the connection, if the daemon receives the
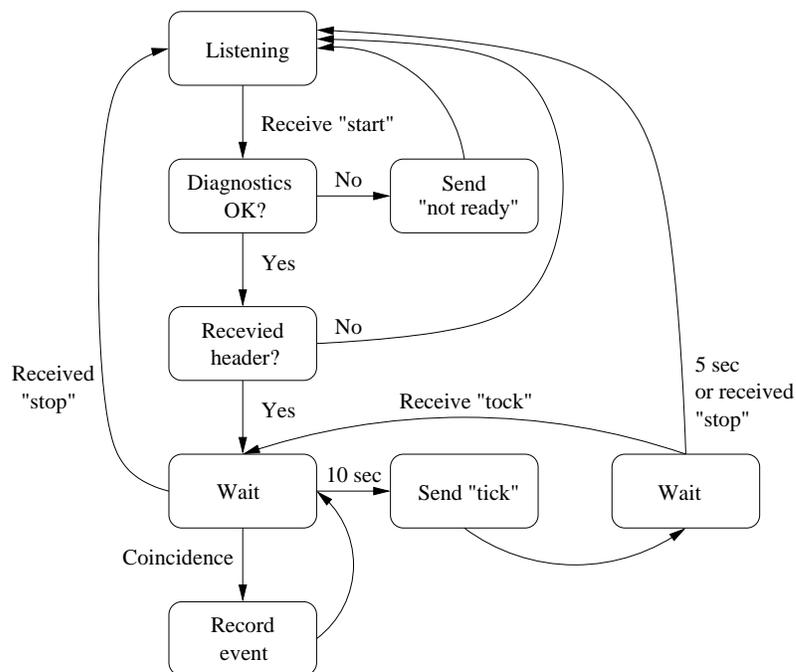
Figure 2.5: OSETI daemon flow chart

ASCII string "start" from the network, it puts the instrument through the series of diagnostics described in the previous section. All actions are recorded into a log file (normally stored in `/var/log/osetilog.YYYYMMDD`, where `YYYYMMDD` is replaced with the year, month and day of the previous local noon). If any of these diagnostic tests fail, then the daemon will send "not ready" back through the network connection. Otherwise, it will send the word "ready". Then the daemon expects to receive the header information (containing the name and magnitude of the target star, observer's name and weather conditions, etc.). The daemon does not interpret this header at all, but just copies it verbatim into the log file. After receiving and recording the header, the daemon activates timestamping by the MTD135 and waits. If the microcontroller reports a coincidence, the circumstances are recorded into the log file. Otherwise, the daemon continues to wait until it receives the word "stop" from its network connection, indicating that the integration is finished. A very simple "keepalive" protocol runs during the integration to make sure that the computer that started the integration has not failed during the integration. Every ten seconds, the OSETI

31

daemon sends the word "tick" out on its network connection. If it does not receive the response "tock" within five seconds, then the observation is terminated and the daemon goes back to waiting for an incoming network connection.

The simple protocol shown in Figure 2.5 was very easy to hook into the existing radial velocity survey software. The software that runs the radial velocity survey instrument was modified slightly so that it would generate the network events necessary to start and stop the OSETI daemon and to participate in the "keepalive" protocol. This design meant the supporting the OSETI experiment did not require the radial velocity survey observers to make any changes to their normal operating procedures.

2.5 Database

Reading the English text in the log file is an excellent way of diagnosing problems with the system; however, it is not a practical way of handling the massive amounts of data accumulated over three years of observations. For that, what is really needed is a database. Furthermore, the process of retrieving the data from the observatory and storing it in the database should be as automatic as possible. Finally, it should be possible for an investigator to answer simple questions about the data without needing to know implementation details about the database. All of these desiderata were achieved relatively easily by taking advantage of some standard Unix utilities and the PostgreSQL database backend (a free software project).

First, the log file that is stored on the OSETI computer at the observatory must be transferred to the server in Cambridge. This is done by a Unix shell script run automatically once at day at 12:05 PM. The same shell script then executes a relatively complicated Perl script that parses the log file into the database and then emails a summary to the project investigators. As header and file formats have changed over the years it has become necessary to make this Perl script extremely adaptable both to parse the new formats and to maintain backward compatibility with the old formats. Doing this in an elegant and maintainable fashion was perhaps the most difficult programming challenge in this experiment.

Once the data are in the database, it becomes possible to manipulate and analyze them in almost any way imaginable. However, doing so requires the investigator to have a good command of the database language SQL ("Structured Query Language") as well as another programming language that it will be embedded into (such as Perl or C). Since this is a fairly substantial burden it forms a barrier between the investigators and the data. Therefore, additional Perl scripts were written to enable access to the database via the web. Although this method is much less flexible than an interactive command line session with the database, it is much more widely accessible and can easily provide the most common database queries.

2.6   Modifications for second observatory

As will be described in detail later, during normal operations of our experiment we get between zero and a few (typically a handful) coincidences per night of observing. An opportunity for implementing an excellent means of coping with these background events presented itself when a group of investigators in the physics department at Princeton University (D. Wilkinson, E. Groth, N. Jarosik, *et al*) expressed an interest in participating in our OSETI program.

The idea is to have both observatories pointed at the same target at the same time. Then the pair of observatories works analogously to the pair of detectors in the instrument: any event seen by both observatories simultaneously is much more likely to be real than an event only seen by one or the other. To get some idea of the noise rejection improvement, suppose that the event rate at one observatory is about 5 per hour = 1.4×10^{-3} per second (the average value for the Harvard observatory at this writing is 0.4 per hour) and that the timing uncertainty is 1 microsecond. Then the rate of accidental simultaneous events at both observatories is just $(1.4 \times 10^{-3})^2(1 \times 10^{-6}) = 2 \times 10^{-12}$ per second, or about once every sixteen thousand years.

Furthermore, if it is possible to time the events at each observatory with sufficient precision, the distance between the observatories can be used as a timing baseline.
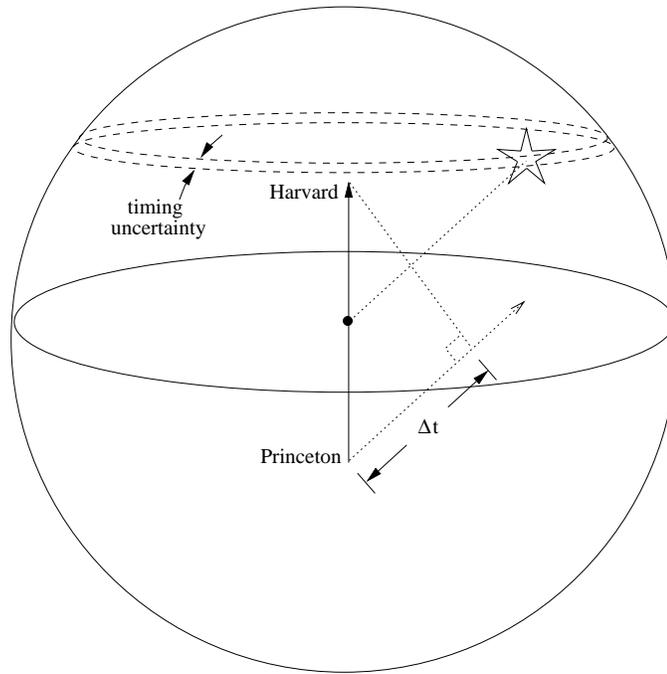
Figure 2.6: Two observatory timing geometry

Figure 2.6 shows the geometry. In the figure, the vector from Princeton to Harvard has been chosen as the axis of the celestial sphere. Light from stars above the plane perpendicular to this axis that goes through the center of the celestial sphere (the "equator") will arrive at Harvard before Princeton and light from stars below this plane will arrive at Princeton before Harvard. The time difference between a signal arriving at Harvard and the same signal arriving at Princeton, Δt, is just the inner product of the vector from Princeton to Harvard (with distance measured in light travel time) with the unit vector from the center of the celestial sphere to the position of the star. This difference is a maximum for stars that are on the "poles" of the celestial sphere shown in Figure 2.6, and its value at maximum is just equal to the light travel time from Harvard to Princeton, or about 1.6 milliseconds. Furthermore, for any given time difference less than this maximum, there is a circle on the celestial sphere perpendicular to and centered on the Princeton-Harvard vector and every point on this circle satisfies the condition that the inner product of its unit position vector with the Princeton-Harvard vector is equal to the given time difference. This circle is

34

broadened into an annulus by the timing uncertainty in the experiment. Therefore, all points within this annulus can be considered to "satisfy" a given Δt, but *no other points on the celestial sphere do.*

Therefore, if the two observatories see nearly simultaneous events, we can be more confident that the source of the events was extra-terrestrial if the annulus on the celestial sphere corresponding to the measured time difference between the two observatories contains the star that was being observed. The better the timing accuracy at the two observatories, the narrower the annulus becomes and the more confident we can be that a signal satisfying all these conditions is really from the star we were observing. If D is the light travel time from Princeton to Harvard and θ is the angle between the Princeton-Harvard vector and the star's position vector, then

$$\Delta t = D \cos \theta \tag{2.1}$$

so that

$$\frac{d(\Delta t)}{d\theta} = -D \sin \theta \tag{2.2}$$

which gives the relationship between the angular uncertainty (the width of the annulus) and the timing uncertainty

$$d\theta = \frac{-d(\Delta t)}{D \sin \theta} \tag{2.3}$$

The singularity at $\theta = 0$ has no serious practical implications since the vector between Princeton and Harvard goes below the horizon.

From the discussion above, it is clearly desirable for each observatory in a two-observatory configuration to have a high quality timestamp referred to universal time for every coincidence event. These universal timestamps are not to be confused with the timestamps generated by the MTD135, which are referred to the trigger events but not to any globally significant time standard. Therefore, additional circuitry was required to provide time stamps referred to an external, universal time standard.

The availability of inexpensive GPS reference clocks and frequency standards

makes it easy to implement a timestamp of the kind required. Off-the-shelf models can produce a very stable 10 MHz and pulse-per-second signal as well as a serial ASCII data stream for synchronizing a computer clock. Simply put (refer again to Figure 2.3), the modifications made to the OSETI instrument were to add a 24-bit counter that is clocked by the 10 MHz from the GPS receiver, reset by the pulse-per-second, and latched by the coincidence trigger. This provides us with a UTC timestamp with a 100 nanosecond granularity modulo one second which the microcontroller in the instrument can read out at its leisure. To determine which second, we use the serial data stream from the GPS receiver to synchronize the system clock on the computer running the OSETI daemon to UTC to within few milliseconds. When the microcontroller reports a coincidence to the computer, it records the system time with a precision of one millisecond. Since we know it should take the microcontroller no more than some tens of milliseconds to transmit the coincidence data to the computer, we can effectively generate a full timestamp referred to UTC for each coincidence.

Since the OSETI instruments used at both observatories are nearly identical, we re-implemented the circuit in Figure 2.3 as a printed circuit board (it had originally been executed in wire-wrap by Jonathan Wolff; Anne Sung laid out the printed circuit), and then modified it to support the precise timing requirement by adding a daughterboard that holds the counter and line receivers for the 10 MHz and pulse-per-second signals from the GPS receiver. One copy of the populated printed circuit was delivered to the group at Princeton and installed in their instrument.

In addition to the modifications to the electronics already described, observing simultaneously at two observatories required some modifications to the software infrastructure at both ends. Both observatories have computer controlled telescopes, although the design of the software at each observatory is radically different. However, since the OSETI instruments were functionally identical, it was desirable to use the same OSETI daemon software at both locations.

The modification to the Princeton telescope software basically amounted to adding enough network functionality for it to act as a "server" for the "client" at Harvard.

Once again, the server listens on a TCP socket for an incoming connection on port 8002, when it receives a connection request it is accepted and data are read from the resulting connection socket until an end-of-file marker is reached. These data contain the name and coordinates of the star to which the Harvard observatory is moving, formatted in human-readable ASCII. The Princeton computer parses the coordinates and then presents the telescope operator with the option to acquire the new star or ignore the request entirely.

If the telescope operator chooses to acquire the new target, then the telescope is automatically slewed to the new position and the telescope software opens a new network connection (this time as the client) to a copy of the OSETI daemon which happens to be running on the same computer. The Princeton telescope software uses the same OSETI daemon protocol (shown in Figure 2.5) that is used by the Harvard telescope software, so no modifications of the OSETI daemon were necessary.

The modification to the Harvard software consisted of adding the network client functionality to send target coordinates to Princeton when the Harvard telescope operator decides to move to a new target. Because the Harvard telescope could spend as little as two minutes integrating on bright stars, it was necessary to send the coordinates at the earliest possible opportunity, even before the integration begins. In practice, it has turned out that the Princeton telescope is usually on target *before* the Harvard telescope because their system is more automatic.

Chapter 3

Analysis

3.1 Introduction and nomenclature

At the time of this writing, the targeted optical SETI program has been running for three and a half years (since October of 1998) at Harvard, and seven months (since October of 2001) at Princeton. During that time, the two observatories have accumulated a total of over 150 days of integration time on the sky, making 28,000 observations of 8,000 different stars. All of the data acquired during the entire program are stored in a single database occupying 130 megabytes.

As has already been described, the data collected by this experiment are the circumstances of coincident outputs of the two phototubes and the results of several diagnostic tests run at the beginning of every observation. These data are kept in the database with information about the integration time, the name of the target and some information about the weather at the observatory during the observation. In the analysis that follows, the following nomenclature will be used (Figure 3.1 is a diagram of the relationships among the terms listed below).

1. An **observation** consists of two phases, a **diagnostic phase** during which a set of measurements are made to ensure the instrument is functioning properly followed by an **integration** lasting from two to thirty minutes during which triggers (see below) are recorded.
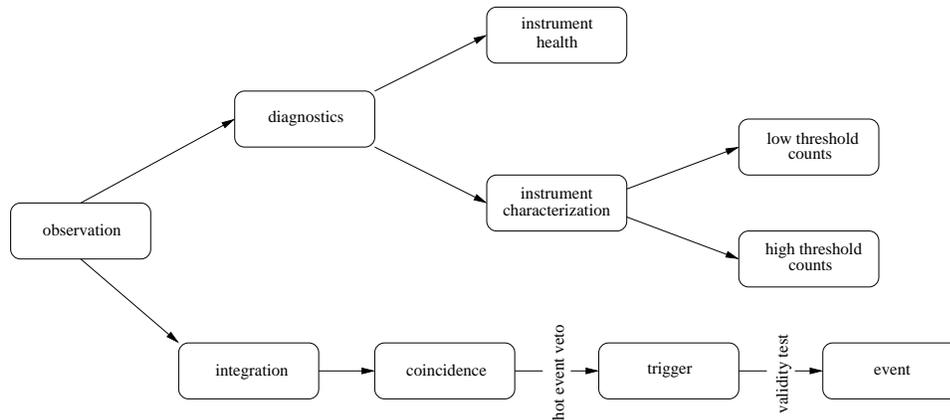
Figure 3.1: Diagram showing the relationship between the terms used in analyzing the OSETI data.

2. The diagnostic phase consists of two parts. The first is a set of tests to verify that the instrument is functioning properly. This includes making sure that power supplies are within their nominal voltage ranges and that the detectors can see a short flash of light generated by an LED. The second part is a set of measurements used to characterize the instrument performance at the time of the observation.

3. **High** and **low thresholds**: Each phototube output is amplified and then input into an array of four comparators (see Figures 2.2 and 2.3). Within each array, the comparator thresholds are exponentially increasing: if the lowest comparator is set for a voltage V_0, then the remaining three will be at $2V_0$, $4V_0$ and $8V_0$. The overall scale of these thresholds, V_0, can be switched between two settings called the high and low threshold settings. The high threshold setting is the one used during integrations. It is set such that V_0 is greater than the amplified single-photo-electron output of the detector, and therefore none of the comparators should produce any output for single-photo-electron events at this setting. The low threshold setting is only used during the diagnostic phase of an observation. It is set such that V_0 is less than the amplified single-photo-electron output of the detector, and therefore the lowest comparator output

39

should toggle for every single-photo-electron output at this setting.

4. **Low** and **high threshold count rates**: The output from the lowest comparator in each comparator array is connected to a counter. During the diagnostic phase of an observation, we measure the number of counts accumulated during two intervals lasting 0.1 and 0.2 seconds for each of the two detectors and for each of the two threshold settings for a total of eight numbers. These numbers are combined as follows.

   The low threshold setting counts the number of outputs exceeding a single-photo-electron threshold from each of the detectors during each of the two time intervals. Therefore, a measure of how much light is getting into our apparatus is the sum of the number of counts for the two intervals and for the two phototubes divided by 0.3 seconds; this is called the low threshold count rate.[1]

   At the high threshold setting, the lowest comparator output should not be exceeded by single-photo-electron outputs from the detectors. Since this is the setting used during integration, if both detectors simultaneously exceed the lowest threshold of their respective comparator arrays at this setting it will cause a trigger.[2] A measure of how often we should expect accidental triggers is the sum of the number of counts for both of the time intervals for each detector separately. This is the high threshold count rate for each detector. The geometric mean of the two high threshold count rates is used if a single high threshold count rate for an observation is needed.

5. **Hot events** are abnormal outputs from the detectors characterized by a very large output (well in excess of the high threshold comparator setting) followed by a few oscillations of ringing. A veto circuit should prevent simultaneous

---

[1]The largest legitimate value that could be obtained by summing the values of a 16-bit counter during 0.1 and 0.2 second intervals is $(32767 + 65535)/0.3 \approx 3.3 \times 10^5$. The counter behaves oddly when it rolls over; the manifestation of this is that one of the two intervals shows a count near 65535 when the actual number is much smaller. This shows up in the scatter plots that follow as a cluster of points near $2.18 \times 10^5 = 65535/0.3$.

[2]Assuming that the event passes the "hot event" veto.

hot events from becoming a trigger; however, they will be included in the high threshold count rate because the counter is not affected by the veto.

6. **Coincidences, triggers and events**: When both detector outputs simultaneously exceed the lowest comparator threshold this is called a coincidence. If the coincidence is not vetoed, the MTD135 is triggered, and the time stamps of the comparator crossings will be recorded and stored into the database. This is called a trigger. Because there is a 300 nanosecond delay between the initial threshold crossing and the trigger, it is possible for threshold crossings during this interval to overwrite the LIFO memory in the MTD135. This generally only happens when there is a high rate of hot events. The result is an amalgam of several unrelated events and no single waveform can be reconstructed from the time stamps. Otherwise, if the time stamps can be assembled into waveforms for each detector and these waveforms overlap and peak within one threshold of each other, then the trigger is considered an event. All events are triggers, but only those triggers satisfying our validity tests are events.

We also define a trigger rate to be the number of triggers recorded during an observation divided by the integration time. The event rate is defined analogously.

7. **Visual magnitude**: The visual magnitude of a star is a logarithmic measure of its brightness that is tabulated in standard astronomical catalogs. It is defined to be proportional to the logarithm of the photon flux F (photons per second per square meter) through a Johnson-Cousins photometric V filter (centered at 550 nanometers with a width of about 180 nanometers) as

$$\log_{10} F = -\frac{2}{5}m + \text{const.} \tag{3.1}$$

The constant can be determined as follows. Standard astronomical references [3] give the flux for a magnitude zero type A0V star as 3.75×10^{-9} erg cm^{-2} s^{-1} Å^{-1}. The width of a Johnson-Cousins V filter is 1800Å, so we expect about (3.75 ×

$10^{-9}) \times 1800 = 6.75 \times 10^{-6}$ erg cm^{-2} integrated over the entire V band. A 550 nm photon has an energy of about 3.62×10^{-12} ergs (about 2 eV), giving 1.87×10^{10} photons per second per square meter for a magnitude zero star. Therefore, the relationship between photon flux F in the V band (photons per square meter per second) and visual magnitude m is

$$\log_{10} F = -\frac{2}{5}m + 10.3 \qquad (3.2)$$

As will be explored in detail below, at Harvard we get about 1.5 triggers and 0.4 events per hour of observation, and at Princeton we get about 11 triggers and 0.4 events per hour of observation. We have good reason to believe that all of these events are an instrumental background. Unfortunately, there is no way to distinguish an individual background event from a real event based on the data from the instrument alone. This is an important factor in determining the sensitivity of the experiment to intermittent transmissions.

The major objective of this analysis is to try to discover if any relationships exist between the behavior of the instrument as measured during the diagnostic phase of the observation and the rate of triggers and events during the integration. If such relationships exist, they could shed light on the physical origin of these background events and provide us with an indication of how to eliminate them.

3.2   Harvard data: counts versus magnitude

At the beginning of every observation, the low threshold count rate is measured. Since this value is measured with the comparator thresholds lowered to single-photo-electron levels, it is expected to reflect the brightness of the star being observed. In particular, it should fall exponentially as the visual magnitude of the stars observed increases (see Equation 3.2). Some deviation from this ideal curve is expected because the detector sensitivity is peaked at about 450 nanometers, and has a width of 250 nanometers, so it is both bluer and wider than the photometric V filter that is used to define visual
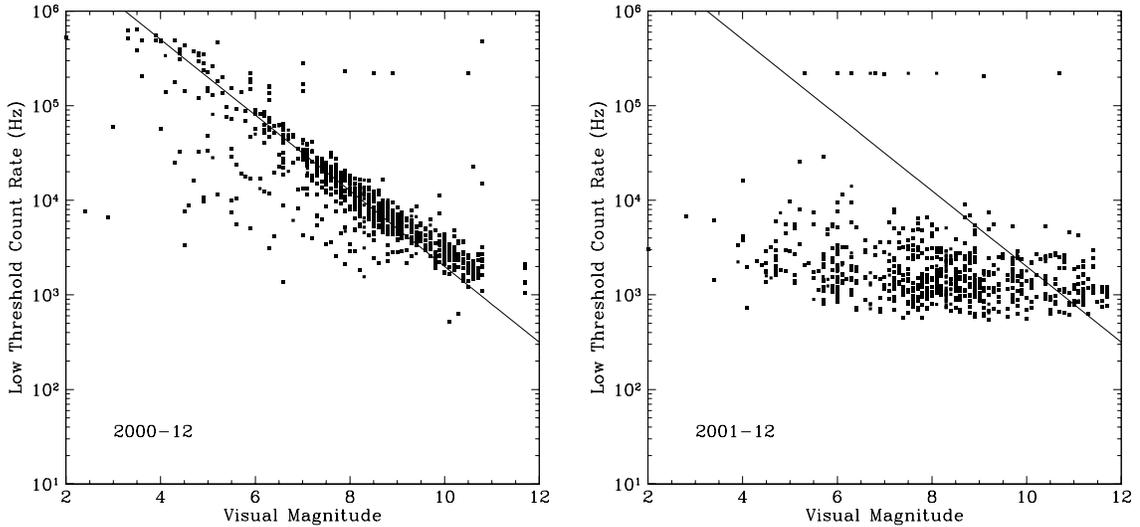
Figure 3.2: Scatter plots of low threshold count rate versus visual magnitude from observations made at the Harvard observatory during December 2000 (left) and December 2001 (right). The line corresponds to $\log_{10} F = -\frac{2}{5}m + 7.3$.

magnitudes. Therefore, the count rate from stars of equal visual magnitude will depend on their spectral type. Furthermore, the detectors always produce a certain rate of dark counts even when they are not illuminated. For very faint targets, this dark count rate can dominate the photo-electron counts from starlight. Since this dark count rate depends on external factors (notably temperature), in introduces additional variability in the low threshold count rate for fainter stars.

Figure 3.2 shows two scatter plots of the low threshold count rate versus visual magnitude from observations made at the Harvard observatory during the months of December 2000 and December 2001 (note the logarithmic y axes). Each point in these plots represents the data from a single observation. It is clear from these plots that the system was working as expected in December 2000, but not at all as expected a year later. As can be seen from Appendix A, which contains plots such as the ones in Figure 3.2 for every month of the OSETI program so far, this transition from data that show a correlation between count rate and visual magnitude to data that do not has happened at least twice during our program.

An investigation into this problem revealed that our instrument had gone out of alignment with the slit spectrograph used by the radial velocity survey. Since the telescope operators always guide the telescope so as to illuminate their slit, this meant that our instrument was not being illuminated. This misalignment persisted from June of 1999 through November of 2000 and from June of 2001 through May of 2002, during which time observations accumulating 2,160 hours of integration took place. This leaves 1,212 hours of integration from observations during the periods of good alignment.

For purposes of finding extraterrestrial signals, it is of course essential to have the instrument aligned with the telescope. However, any background present in the absence of starlight will remain when the instrument is aligned properly. Therefore, for purposes of studying the background it is reasonable to include the data from the misaligned periods, but for purposes of setting a limit on the prevalence of civilizations transmitting signals according to our scenario, only the data from the well aligned periods can be used.

The diagonal line in the plots in Figure 3.2 corresponds to the equation

$$\log_{10} F = -\frac{2}{5}m + 7.3 \qquad (3.3)$$

We can use this formula to make a rough estimate of the ratio of the number of photons detected to the number of photons entering the aperture of the telescope. Our instrument should get about one quarter of the light from the 61 inch diameter circular aperture (the rest goes to the radial velocity survey instrument and guide camera), corresponding to an effective collecting area of 0.47 square meters. Equation 3.2, rewritten as

$$F = 10^{10.3 - 2m/5} \qquad (3.4)$$

gives the expected flux in photons per square meter per secondas a function of magnitude, so the expected flux entering the aperture of our telescope is

$$F = 0.47 \times 10^{10.3 - 2m/5} \approx 10^{10 - 2m/5} \Rightarrow \log_{10} F = -\frac{2}{5}m + 10 \qquad (3.5)$$

so the constant term in Equation 3.3 corresponds to an overall efficiency of $10^{-2.7} = 0.2\%$

As can be seen in the block diagram, Figure 2.2, starlight is reflected by four mirrors, divided by two beamsplitters and focussed by one lens before reaching our detectors. A reasonable value for the efficiencies is about 85% for the mirrors and about 92% for the beamsplitters and the lens. The quantum efficiency of the detectors is about 10% averaged over their bands, so the expected overall efficiency is roughly

$$(0.85)^4(0.92)^3(0.10) \approx 4\% \tag{3.6}$$

This leaves a discrepancy of a factor of 16 between the expected efficiency and the measured efficiency. Such a low efficiency is a serious problem since it implies a proportional increase in the required transmitter power for a detection. However, it is not clear why the efficiency is so low. The telescope primary mirror was recoated in October of 2000, not long before the data in the left panel of Figure 3.2 were taken, and so it seems unlikely that mirror inefficiencies are to blame. This discrepancy will be investigated further.

3.3 Harvard data: seasonal variations

Figure 3.3 shows a plot of average trigger rate versus time for the three and a half years of the targeted optical SETI program at Harvard. The average values plotted in the figure were calculated by summing the number of triggers and independently summing the integration times of observations until the integration time exceeded thirty hours. The average trigger rate was then calculated as the ratio of number of triggers divided by integration time and the process started over again. As can be seen from the figure, the trigger rate goes up by a factor of 30–40 during the summer months, and then back down again in winter.

One possible explanation that was investigated was that the APD gain was going up in summer because of warmer temperatures. This temperature coefficient is a well
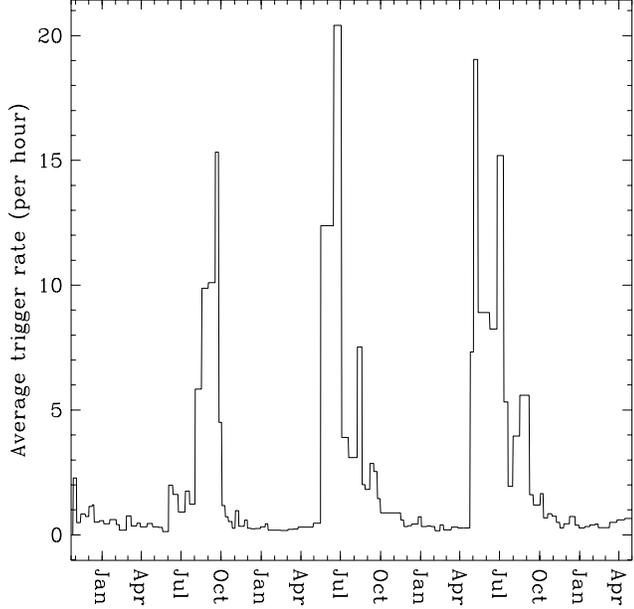
Figure 3.3: Seasonal variation in the trigger rate from October 1998 until May 2002

known phenomenon of APDs, and can be cancelled out by temperature compensating the APD bias voltage. High voltage supplies designed for use with APDs typically have this temperature compensation built in, and ours is no exception. In order to verify that this temperature compensation was working as intended, we examined the HAPD outputs on a hot day in summer to make sure that the waveforms (and particularly their peak values) had not changed since the previous winter; they had not. Prior to the collaboration with Princeton, we had not found a good way of dealing with the data from the summers apart from just throwing it away.

The more likely explanation is that this seasonal variation is a manifestation of the well-known sensitivity of phototubes to humidity. The high voltage bias on the photocathode, -7,500 volts, can lead to sparking and corona discharge during periods of high humidity. Since this seasonal variation was discovered, some considerable effort has been expended in trying to keep the instrument sealed from infiltration by humidity, and to pump dry nitrogen through it. These efforts do not seem to have had a significant impact on the summertime trigger rate, unfortunately.

46

3.4 Harvard data: triggers and events

Recall that the MTD135 LIFO memory that holds the timestamps until they are read out by the microcontroller can be overwritten during the 300 nanosecond window of opportunity between the coincidence and the trigger (see Figures 2.3 and 3.1). When this happens, the timestamps as read out by the microcontroller usually make no sense at all since they are really an amalgam of the several events that happened during the 300 nanosecond window. There are a number of validity tests that can be imposed on the timestamps to exclude most of these triggers from consideration. These can be summarized as three requirements: that it must be possible to reconstruct *some* waveform like the ones shown in figure 2.4 for each of the two detector outputs; that the waveforms must overlap in time; and that the highest thresholds exceeded by each must be within one threshold level of the other.

If, as seems reasonable, we assume that all of the triggers recorded so far are false alarms, then a measure of how effective the validity tests are is the ratio of the number of events to the total number of triggers. As of this writing, the total number of triggers in the Harvard data (including data from the misaligned periods) is 7536 and the number considered to be events is 2312, or about 31%. An interesting question in view of the seasonal variation in the trigger rate presented in the previous section is: is it the case that a fixed percentage of triggers pass all of our validity tests, or does it vary seasonally or perhaps from one target to another? One might hope that as the trigger rate goes up in the summer the ratio of events to triggers would go down to compensate, but this does not appear to be the case.

Figure 3.4 shows the average ratio of events to triggers at various points in time during the Harvard OSETI program from October 1998 until May 2002 (including data from the misaligned periods). The figure was generated by adding up the number of events and independently the number of triggers from observations in chronological order until the total integration time exceeded thirty hours, then plotting the ratio and starting again. The ratio stays within one and a half standard deviations of the global average value of 0.31 with the exception of two periods around September of 1999
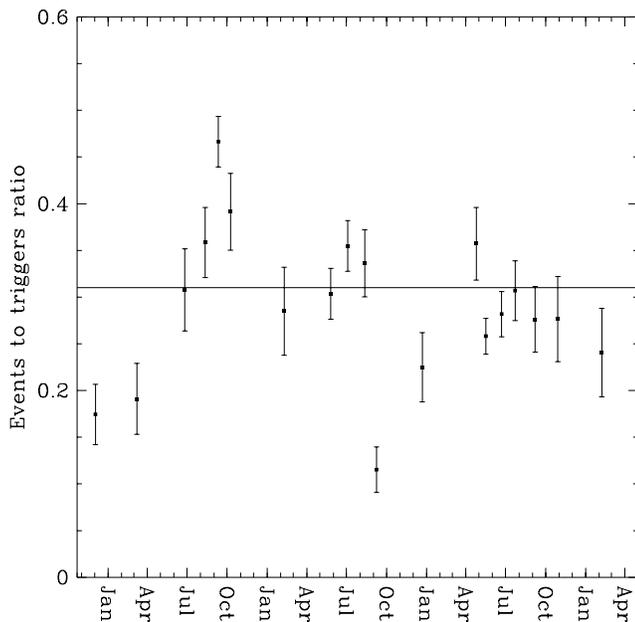
Figure 3.4: Averaged events/triggers ratio versus time during the OSETI program. Each point represents an average over observations adding up to thirty hours of total integration time. The horizontal line marks the global average value of this ratio, 0.31.

and September of 2000 when is it significantly higher and then lower, respectively. Since these are opposite behaviors at the same time of year, the conclusion must be that there is no seasonal dependence.

The question of seasonal dependence can also be addressed indirectly. Suppose that we select only observations where there was exactly one trigger. If our hypothesis that the validity tests always mark a fixed percentage of the triggers as events is right, then we would expect that 31% of these observations would record the trigger as an event and the remaining 69% would record it as just a trigger. Carrying on, if we select only observations where there were exactly 2 triggers, then we would expect the weighted average would be 0.31, that is

$$\left\langle \frac{\text{events}}{\text{triggers}} \right\rangle = \frac{0 \times N_0 + 1 \times N_1 + 2 \times N_2}{2 \times (N_0 + N_1 + N_2)} = 0.31 \tag{3.7}$$

48

where N_i is the number of these observations recording i events. In general, for the set of observations with n triggers, we expect that

$$\frac{\sum_{i=0}^{n} i N_i}{n \sum_{i=0}^{n} N_i} = 0.31 \qquad (3.8)$$

This is just the ratio of the number of events to the number of triggers where attention has been restricted to those observations returning a total of n triggers. Since observations made during the summer get a much larger number of triggers, a seasonal dependence in the ratio of events to triggers should manifest itself as a departure from the expected value as n gets larger.

Figure 3.5 is a plot of Equation 3.8 for all of the Harvard data. The plot shows that the ratio of events to triggers does *not* depend on the number of triggers per observation, whereas figure 3.3 shows quite clearly that the number of triggers per observation *does* depend on the season (there are more triggers in the summer even though there are fewer observations due to the shorter nights). Once again, the conclusion is that the ratio of events to triggers also does not vary seasonally the way the overall number of triggers does.

To paraphrase, although any real signal would have to pass our validity tests, it seems that quite a few that aren't real also pass. If our interest is in finding real signals, then the validity tests will eliminate more than two-thirds of the background. On the other hand, if our interest is in studying the background, then we might as well choose to include the triggers that don't pass the validity tests, since excluding them only reduces the number of triggers available for analysis without altering their underlying statistics.

3.5 Harvard data: triggers versus counts

Figure 3.6 shows two scatter plots of the Harvard trigger rate versus the low and high threshold count rates. There is no obvious correlation between the two quantities shown in either plot. This result was expected, since the low threshold count rate
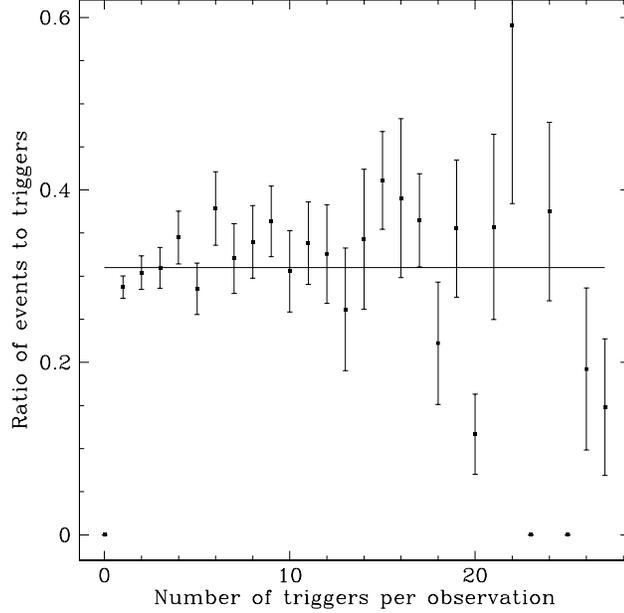
Figure 3.5: Ratio of events to triggers versus number of triggers per observation. The horizontal line marks the global average ratio, 0.31

should reflect the magnitude of the star, and the high threshold count rate should reflect the rate of hot events. Neither of these should cause triggers because stellar photon pileup is very rare (even very bright stars should rarely produce accidental triggers), and hot events are vetoed before they can become triggers.

However, this result forces us to come up with a different explanation for the seasonal variation in the trigger rate at Harvard than that it is caused by an increase in the number of hot events during the summers. If this were true then there *should* be a correlation between the high threshold count rate and the trigger rate.

The question of whether there exists a seasonal variation in the hot event rate can be addressed directly by looking for a corresponding variation in the high threshold count rate. Figure 3.7 shows a plot of high threshold count rate averaged over ten observations versus time. There is some seasonal variation, it is not nearly as strong as the variation in trigger rate shown in figure 3.3.

The conclusion is that, at least for the Harvard data, whatever is causing the
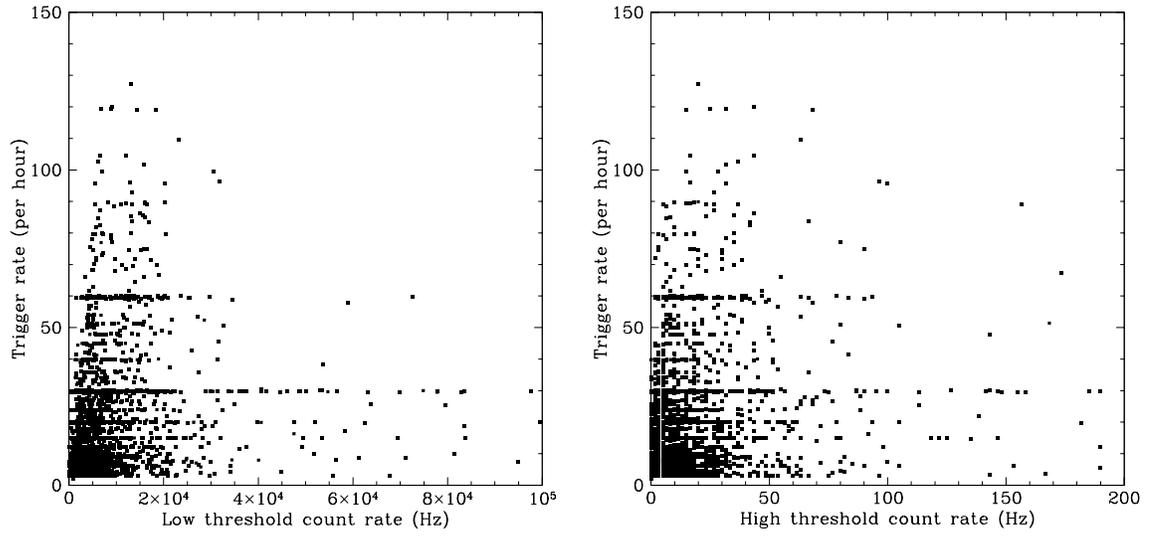
Figure 3.6: Scatter plots of trigger rate versus low (left) and high (right) threshold count rates for the Harvard data
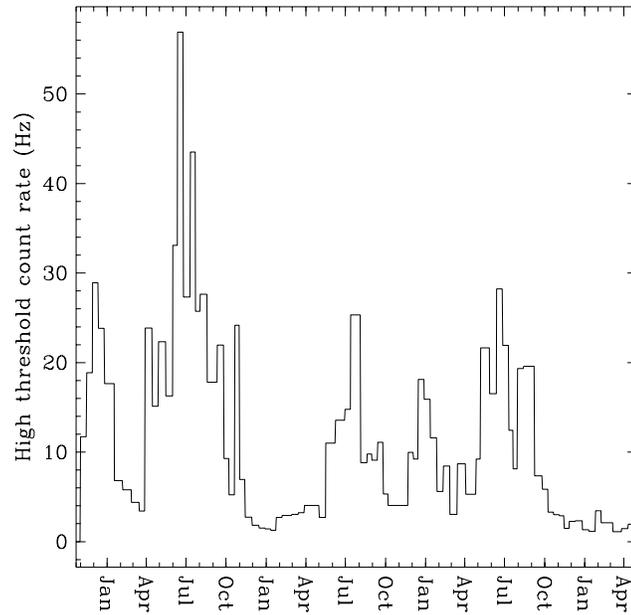

Figure 3.7: High threshold count rate, averaged over ten observations, versus time (October 1998 to May 2002). Compare this to the seasonal variation in trigger rate in figure 3.3

triggers is not showing up in either the high or low threshold count rates. The most likely explanation is that the count rates, and especially the high threshold count rate, are measured during intervals that are too short (0.3 seconds) to catch a statistically significant number of the events that show up during the much longer integration periods (2–30 minutes). Since the hot event rate is usually about 5–10 Hz, you would expect to count between zero and three such events during a 0.3 second time period. So the high threshold count rate as it is currently measured, is *always* limited by small-number statistics.

3.6   Princeton data: counts versus magnitude

Figure 3.8, shows a scatter plot of low threshold counts versus visual magnitude for the Princeton data. The horizontal line of points near 2.2×10^5 Hz is the same counter rollover artifact already discussed. Other than that, the points cluster relatively close to the line

$$\log_{10} F = -\frac{2}{5}m + 6.5 \tag{3.9}$$

shown in the figure. The collecting area of a 36-inch (diameter) aperture is about 0.66 square meters, so from equation 3.2 we would expect the constant term in equation 3.9 to be about 10.1, which indicates a system efficiency of about 0.02%. The expected efficiency is about 75 times greater than this. An adequate explanation for this discrepancy has not yet been found. There are taps on the detector outputs available on the front panel of the instrument, and when a laboratory counter was connected to these the measured efficiency was about 1.5%, so a likely explanation is that the low comparator thresholds are too high.

3.7   High versus low threshold count rates

Figure 3.9 shows scatter plots of high versus low threshold count rates from both observatories side by side. A significant difference between the two observatories is that the data from Princeton show a clearer correlation between the measured low
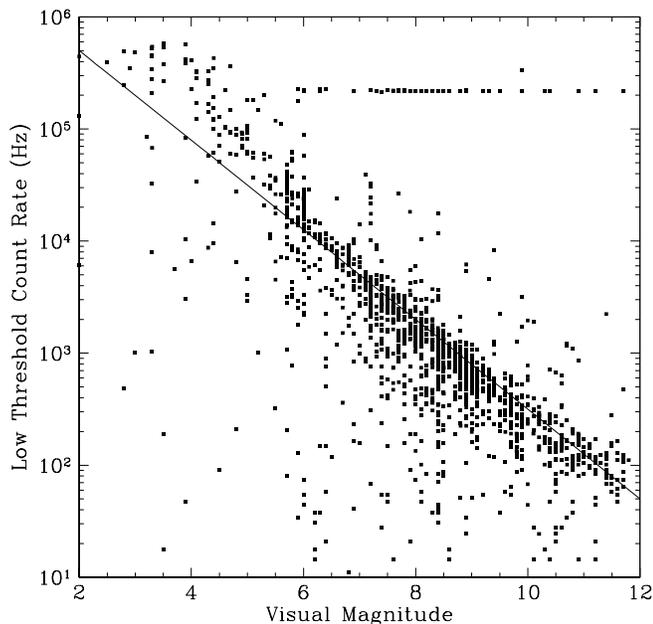
Figure 3.8: Scatter plot of low threshold count rate versus visual magnitude for the Princeton data. The horizontal line is a counter artifact.

and high threshold count rates (N.B., the Harvard plot does not include data from the misaligned periods). Once again, the cluster of points near the x axis at 2.2×10^5 Hz is an artifact of the rollover of the 16-bit counter.

If our belief that the high threshold count rates reflect the rate of "hot events" is correct, then there should be no correlation with the low threshold count rate. Of course, any detector output which exceeds the high threshold setting will also exceed the low one and therefore all high threshold counts are also low threshold counts. However, there are typically orders of magnitude more low threshold counts, i.e. the low threshold count rate should be dominated by single-photo-electron events whereas the high threshold count rate should be dominated by hot events, and we do not believe the two types of events are correlated.

Clearly, although this model might apply to the Harvard data on the left in Figure 3.9, something else is going on in the Princeton data on the right. The most likely explanation is that some single-photo-electron events at Princeton are exceed-
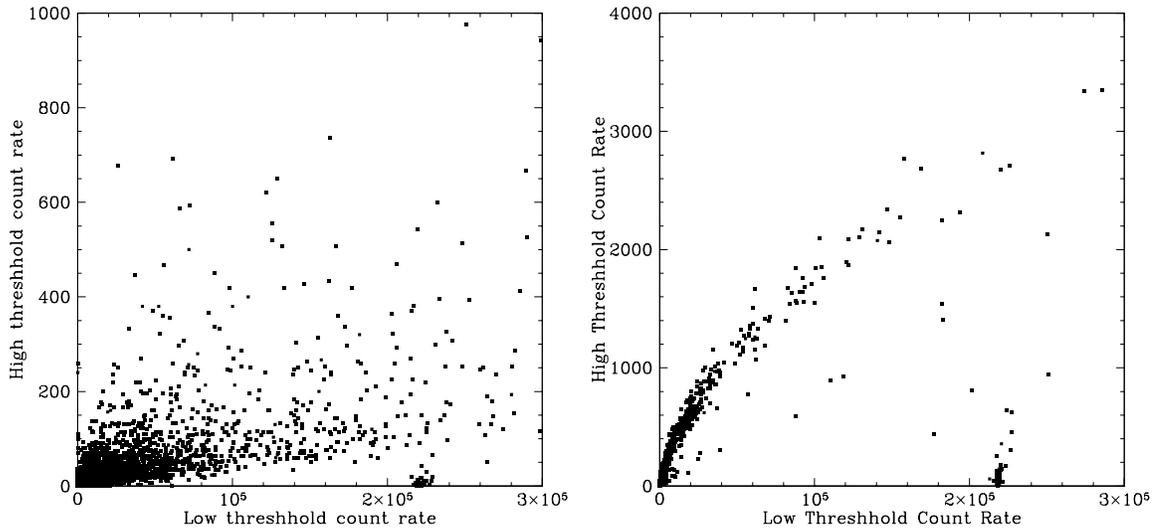
53

Figure 3.9: Scatter plots of Harvard (left) and Princeton (right) high versus low threshold count rates. The y-axis scale has been expanded on the Harvard plot.

ing the high threshold setting, i.e., this threshold has been set too low. This theory is reinforced by the fact that the Princeton high threshold count rates are typically four times higher than those from Harvard, whereas the low threshold count rates are about the same. (Note that, at Harvard we get 1/4 of the light from a 61-inch aperture, or an effective 30-inch aperture, and Princeton has all of the light from a 36-inch aperture, so one would expect the low threshold rates to be within 50% of each other based on aperture size alone.)

3.8 Princeton data: triggers versus counts

Furthermore, if the high threshold counts are caused by hot events, then they should *not* correlate with the triggers if the hot event veto is working. On the other hand, if the high threshold count rates are dominated by events which are not vetoed, then a correlation would be expected.

Suppose that the high threshold count rate is c and is the same for both detectors. Each count in one detector provides an opportunity for a simultaneous count in
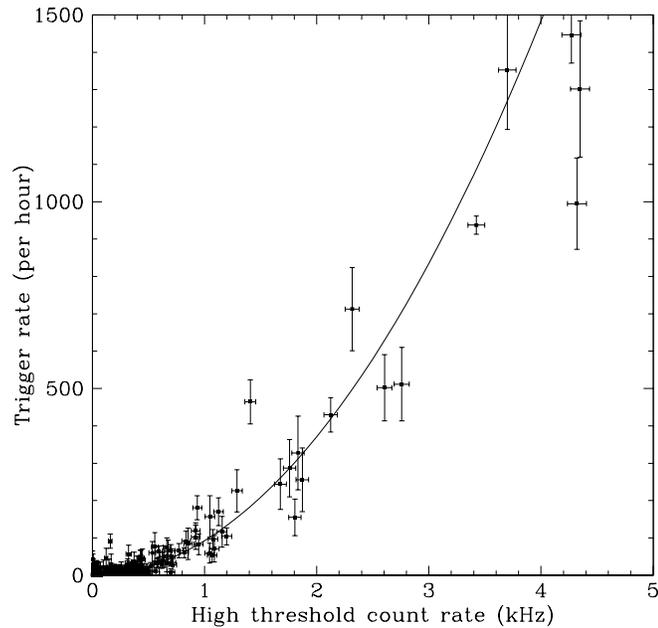
Figure 3.10: Scatter plot showing trigger rate versus high threshold count rate for the Princeton data. Each point represents a single observation. The parabola is a fit to the data, explained below.

the other detector. This opportunity lasts lasts a time τ, the longest time between two events that the apparatus considers to be "simultaneous" (typically of order 25 nanoseconds). Therefore, there is an "opportunity time" of $c\tau$ seconds per second. The probability of having a count in the other detector during any time interval is just c times the duration of the interval, and therefore the rate of simultaneous counts (i.e. triggers) in both detectors is $c^2\tau$ per second.

The important point to draw from this analysis is that we expect the simultaneous event rate to be proportional to the individual event rate squared, where the constant of proportionality is the aperture window. Figure 3.10 shows a scatter plot of trigger rate versus high threshhold count rate for all of the Princeton data so far, with a parabola drawn for reference. It certainly does not strain credulity much to claim that there is such a quadratic dependence of trigger rate on count rate from looking at this plot. However, in order to turn this observation into something quantitatively useful, we need to be able to fit a parabola to the data we have and use the result to

predict how many triggers we expect from new observations. The uncertainty in the fit for the parabola has a direct bearing on the uncertainty (width) of the probability distribution for triggers.

The equation of the relevant parabola is

$$h = kc^2 \tag{3.10}$$

where c is the count rate, h is the trigger rate and k is a constant parameter that we want to fit. It is important to emphasize that both h and c are rates that we cannot measure directly. What we measure is the number of events (triggers and counts) that happened during a given time interval. This implies that there are uncertainties in both axes in figure 3.10, a fact that will manifest itself in the form of a convolution integral below.

From a Bayesian perspective, fitting the parameter k is equivalent to finding the posterior distribution

$$p(k|n_1, n_2, n_3, \ldots) \tag{3.11}$$

where n_i is the number of triggers recorded during observation i. To get started, focus on $p(k|n)$ for a single observation; including the data from all observations is straightforward from there.

As usual, the starting point is Bayes' theorem

$$p(k|n) = \frac{P(n|k)p(k)}{P(n)} \tag{3.12}$$

The choice of prior distribution $p(k)$ is almost irrelevant because of the large number of observations involved. If we wanted to be pedantic, the correct choice would be the maximum entropy prior that satisfies the constraint that $0 \leq k < \infty$, namely

$$p(k)dk = \frac{dk}{k}. \tag{3.13}$$

However, it will be easier to use a uniform prior on k since the calculations that follow

will have to be done numerically.

The probability distribution for n can be treated as a normalization constant since this is exactly what we get by marginalizing out k in the numerator

$$P(n) = \int_0^\infty P(n|k)p(k)dk \tag{3.14}$$

Therefore, we can ignore $P(n)$ as long as we are careful to keep all of our posterior distributions normalized. The problem therefore reduces to finding $P(n|k)$.

We can assume that both types of events (triggers or counts) are described by Poisson statistics. Suppose we know h, then as usual the probability of finding n triggers during a time T is

$$P(n|h) = \frac{(hT)^n e^{-hT}}{n!} \tag{3.15}$$

Equation 3.10 shows how to put this expression in terms of the count rate

$$P(n|k, c) = \frac{(kc^2T)^n e^{-kc^2T}}{n!} \tag{3.16}$$

The problem is that we are *not* given c, but only the fact that we had m counts during the time interval t (this is the source of the uncertainty in the abscissa of figure 3.10). However, we can eliminate the nuisance parameter c by using marginalization if we have a probability density for the rate c given m.

Therefore, we need an expression for the posterior probability density of a Poisson rate c given that m events happened during a time t. The starting point is Bayes theorem

$$p(c|m) = \frac{P(m|c)p(c)}{P(m)} \tag{3.17}$$

In this case, the denominator can be considered to be a normalization constant found by integrating the numerator over all possible values for c

$$P(m) = \int_0^\infty P(m|c)p(c)dc. \tag{3.18}$$

57

$P(m|c)$ is given by the Poisson formula

$$P(m|c) = \frac{c^m e^{-c}}{m!} \qquad (3.19)$$

and $p(c)$ is given by the maximum entropy prior satisfying the constraint that $0 \leq c < \infty$

$$p(c) = \frac{dc}{c} \qquad (3.20)$$

The numerator in Bayes' theorem is therefore

$$P(m|c)p(c) = \frac{c^{m-1} e^{-c}}{m!} \qquad (3.21)$$

and we can find the denominator by integrating it from 0 to ∞

$$P(m) = \frac{1}{m!} \int_0^\infty c^{n-1} e^{-c} dc = \frac{1}{m!} \Gamma(m) = \frac{1}{m} \qquad (3.22)$$

so the exact posterior probability distribution for c given m is

$$p(c|m) = \frac{c^{m-1} e^{-c}}{(m-1)!}. \qquad (3.23)$$

Since we are given m (this is precisely what it measured by the high threshhold count rate diagnostic), we can use this to write

$$P(n|k, m) = \int_0^\infty P(n|k, c)p(c|m)dc \qquad (3.24)$$

which becomes

$$P(n|k, m) = \frac{(kT)^n t^m}{n!(m-1)!} \int_0^\infty e^{-(tc + kTc^2)} c^{2n+m-1} dc \qquad (3.25)$$

An expression for the integral in terms of "parabolic cylinder functions" can be found in Gradshteyn and Ryzhik, 3.462 number 1. However, since the parabolic cylinder functions are themselves defined in terms of integrals which do not have known an-
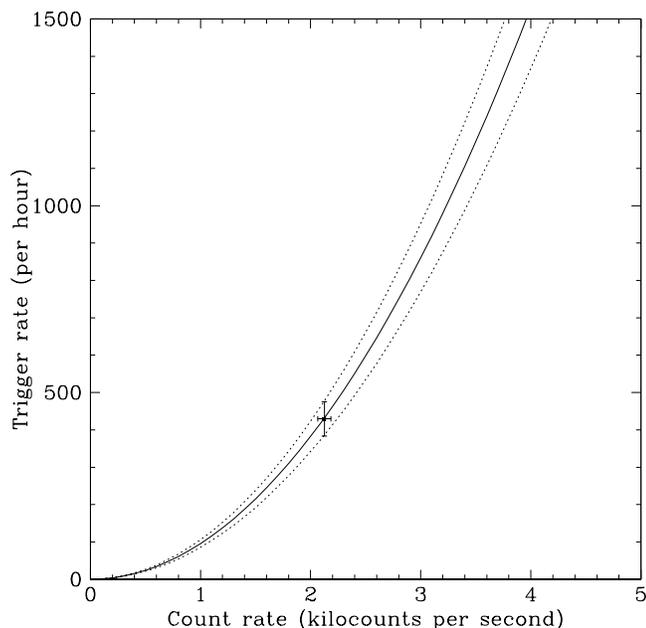
Figure 3.11: Fit to a single observation (Princeton number 1325)

alytic solutions, this amounts to nothing more than expressing an integral that I cannot evaluate in terms of another integral that I cannot evaluate. At such times, it is best to bring numerical methods to bear.

Figure 3.11 shows the result of applying the numerical fit to a single observation (Princeton observation number 1325, of HD098230, a 4.87 magnitude F8.5V type star), with the solid line representing the most probable value of k, and the dashed lines representing $k \pm 1\sigma$. The error bars shown are the 1σ errors in the Poisson rate estimator derived in a previous section, namely $\sigma = \sqrt{m}/t$ in the x direction and $\sigma = \sqrt{n}/T$ in the y direction. Figure 3.12 shows the values of the probability density calculated at discrete values of k by numerical methods (points) overlaid by best-fit gaussian distribution (solid line). The numerical values detemined by the fit were $k = 95.645 \pm 10.146$. Notice that the numerical probability density for k does differ slightly from the gaussian fit; it is not exactly centered about its mean. However, the difference is small.

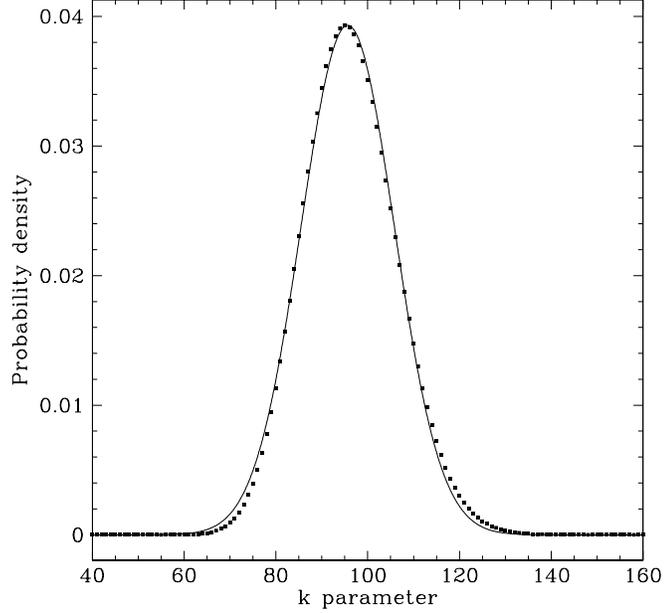To carry on with the analysis, we need a method for combining the data from all

Figure 3.12: Fit to a single observation (Princeton number 1325)

the observations. If we use a uniform prior on k, then Bayes theorem can be written

$$p(k|n_1, n_2, n_3, \ldots) \propto P(n_1, n_2, n_3, \ldots |k) \qquad (3.26)$$

where the constant of proportionality is found by requiring that the posterior distribution for k should be normalized. Since the observations are independent of one another,

$$P(n_1, n_2, n_3, \ldots |k) \propto P(n_1|k)P(n_2|k)P(n_3|k)\ldots \qquad (3.27)$$

In order to avoid numerical difficulties, it is easier to work with the logarithm of the right hand side (the so-called "log-likelihood")

$$\log P(n_1, n_2, n_3, \ldots |k) = \text{const.} + \log P(n_1|k) + \log P(n_2|k) + \log P(n_3|k)\ldots \quad (3.28)$$

The numerical implementation is now clear. The computer program keeps a one dimensional array of double-precision floating-point values, indexed by the parameter

60

k, where each array element holds the log-likelihood for the corresponding value of the parameter. For each observation, we have recorded the number of high threshhold counts m, the number of triggers n and the integration time T. The high threshhold counts are always measured during a time period, t, lasting 600 milliseconds. To incorporate the data from a given observation we compute

$$n \log(kT) + m \log(t) - \log(n!) - \log((m-1)!) + \log \int_0^\infty e^{-(tc+kTc^2)} c^{2n+m-1} dc \quad (3.29)$$

for every value of k and add it into the corresponding log-likelihood array element.

A couple of observations can save a lot of numerical difficulties. First, a very good approximation for the logarithm of a factorial is calculated using the Stirling approximation

$$\log(n!) = \frac{1}{2} \log(2\pi) + \left(n + \frac{1}{2}\right) \log n - n + \frac{1}{12n} \quad (3.30)$$

which has an error less than $1/(360n^3)$ and relieves us of the need to calculate the factorials of large integers.

Secondly, the majority of the contribution to the integral in 3.29 comes from the vicinity of the maximum of the probability density for c, Equation 3.23. A function of the form $e^\mu \mu^m$ reaches its maximum at $\mu = m$, and will fall to a value a e^{-10} times smaller when $\mu = nm$ where

$$n - \log n = 1 + \frac{10}{m}. \quad (3.31)$$

Since m is the number of counts, we will always have $m \geq 1$ in the data we are analyzing, so the left side of the equation above can be conservatively approximated by 11, giving $n \approx 14$. Therefore the upper bound in the integral in 3.29 can be approximated by $c = 14m/t$. A similar argument shows that the lower bound can be set to a value m/n where

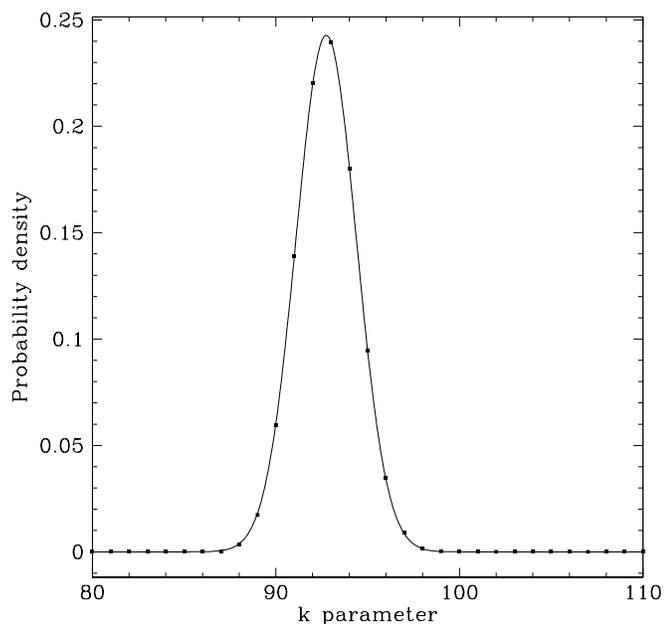$$\frac{1}{n} + \log n = 1 + \frac{10}{m} \leq 11 \quad (3.32)$$

Figure 3.13: Gaussian distribution fit to numerical posterior for k parameter

giving $n \approx 60,000$. In the Princeton data, $m \leq 4,000$, so $m/60,000$ is not larger than zero by a large enough amount to save any computation time. Therefore we leave the lower bound at zero.

Finally, $p(k|n)$ is computed by exponentiating and then normalizing the log-likelihood function, we can add any constant we want to the log-likelihood function. In particular, we can add a constant that will bring all of the exponents into the range of values that can be expressed by double-precision floating-point numbers (roughly 2.3×10^{-308} to 1.7×10^{308} for positive numbers).

The net result of all these numerical exercises is the parabola shown already in figure 3.10, and the probability density for the k parameter shown in 3.13. The numerical value for k is 92.7 ± 1.6. A gaussian fit is shown overlaid on the points that were computed numerically; this time the agreement is very good.

Finally, we can examine what implications this analysis has for the instrument's aperture time. It was already shown that if the singles rate is c, the expected simultaneous rate is $h = c^2\tau$. Therefore the parameter k is the instrument aperture time

τ, where c is measured in kilocounts per second and h is measured in triggers per hour

$$92.740 = 3.6 \times 10^9 \tau \to \tau \approx 26\text{ns} \qquad (3.33)$$

To confirm this value, we set up a benchtop experiment using one of the spare printed circuit boards that implement the block diagram in figure 2.3 (excluding the HAPDs and low noise amplifiers). This circuit board was stimulated with a fast pulse generator, and the experiment confirmed that our coincidence aperture is about 25 nanoseconds.

There is another, more important conclusion that can be drawn from the quadratic dependence of the Princeton trigger rate on the high threshold count rate, and that is that the high threshold count rate is *not* dominated by hot events, as we would have expected. Hot events should be vetoed, so they are not expected to contribute much to the trigger rate, but they should dominate the high threshold count. Once again, the interpretation points to comparator thresholds set too low on the Princeton instrument.

3.9 Princeton data: triggers and events

Whereas the Harvard observatory has recorded 7536 triggers during a total of 141 days of integration time, the Princeton observatory has recorded 10,682 in only 16 days of integration time. Of these, only 313, or 3%, have been marked as events. Some of the Princeton observations show remarkably large numbers of triggers, such as observation number 937, a 1.5 hour integration that recorded 3271 triggers, of which only 77 passed the validity tests.

Because the program has only been running for seven months, it is impossible to look for a seasonal variation in the ratio of events to triggers in the Princeton data. However, we can still check the consistency of the 3% ratio of events to triggers as a function of the number of triggers recorded during an observation. This is plotted in Figure 3.14, which shows that the Princeton ratio is also fairly consistent (compare Figure 3.5).
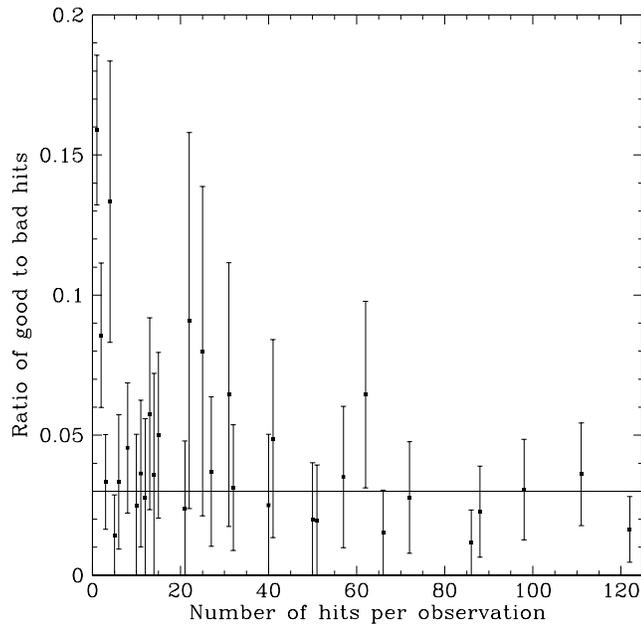
Figure 3.14: Ratio of events to triggers versus number of triggers per observation. The horizontal line marks the global average ratio, 0.31

3.10   Results of the analysis

Uncovering the misalignment of the Harvard instrument was by far the most important result of the analysis. However, a number of other important conclusions can be drawn as well. First, although the two observatories are meant to be functionally equivalent, there are a number of important differences between them:

1. The Harvard instrument gets 1/4 of the light from a 61-inch aperture telescope, which is an effective 30.5-inch aperture, whereas the Princeton instrument gets all of the light from a 36-inch aperture. This means that the Princeton instrument should see about 40% more light, but in practice it seems to get about six times less (see Figures 3.2, 3.8 and Equations 3.3, 3.9). Even so, the Harvard system efficiency of 0.4% seems about five times lower than is usual for this sort of astronomical facility.

2. The Harvard data shows no correlation between high and low threshold count

rates nor between the trigger rate and high or low threshold count rates, whereas the Princeton data shows a strong correlation between high and low threshold count rates, and a weaker but still credible correlation between the trigger rate and high threshold count rate (see Figures 3.6, 3.9 and 3.10). This strongly implies that the ultimate origins of the triggers seen at the two different observatories are different.

3. Both observatories seem to mark a fixed percentage of the triggers as events (i.e., they satisfy all validity tests), but the percentages differ by an order of magnitude with Harvard accepting 30% but Princeton only 3%. Once again, this seems to point to different origins for the triggers at each observatory.

Although an adequate explanation for the triggers we do see was not developed, some doubt was cast on some of our formerly favorite theories — that they were caused by hot events passing through the veto (due to the lack of correlation with high threshold count rates), or that they might be the result of single-photo-electron events exploring the long tail of the pulse height distribution (due to the lack of correlation with low threshold count rates). The trigger rate at the Harvard observatory is not strongly correlated with the count rates measured by the diagnostics, and the Harvard instrument continued to record triggers even during the periods of misalignment. Conversely, the Princeton triggers do seem to go up significantly for brighter stars, probably indicating that the comparator high threshold setting is too low. Furthermore the Princeton high and low threshold count rates are strongly correlated and the low threshold count rate does not seem to register all the outputs from the detectors, indicating the the low threshold setting is too high.

Chapter 4

Conclusions

In an experiment such as this where the detection of a signal is expected to be a very rare event, it is critical to continuously monitor a set of diagnostics which ensure that the apparatus is functioning correctly since the presence of the sought-after signal cannot be depended on as an indication of a working system. To a large extent this principle has guided the design of the instrument, and a considerable part of the electronics that support the hybrid APDs have no other function than to make sure that the detectors are working as intented. Nonetheless we were caught unawares by a failure mode that we hadn't anticipated.

The targeted optical SETI experiment is based on the reasonable assumption that signals from extraterrestrial civilizations will come from the vicinity of stars, in particular stars not too dissimilar from our own. The background starlight is more than just a potential source of accidental pileups, it is also the one unmistakable marker that the instrument is on target.

4.1 Photometry

Standard astronomical catalogs provide a tremendous amount of data about the stars in our survey; however, only the data included in the radial velocity survey header (the coordinates and visual magnitudes) is available to our instrument in real time and included in the database for post-analysis. If, in addition, the spectral classifications

of the stars were readily available, it would be possible to calculate the expected photon flux to good precision and check it against the measured photon flux during the diagnostic phase of an observation. This type of diagnostic would have caught the misalignment of the Harvard instrument early on. Even now, it is still difficult to tell if a given observation considered in isolation is producing the expected flux because we do not have a means of measuring the dark count rate in our detectors during the diagnostic phase of an observation. The dark count rate can be as much as the flux from a magnitude eight star.

An additional difficulty is that at present, single photon counts are only accumulated for 0.3 seconds during the diagnostic phase which is not very long if the objective is good photometry. Although this interval could be extended, a much better approach would be to use the entire integration period for photometry, since this gives the highest quality photometry possible subject to the constraint that the integration time is chosen by the radial velocity survey observers. This would require a significant modification of the existing circuit, since an additional independent set of comparators permanently set with single photo-electron thresholds would be needed. It is a proposal worth considering, however.

High quality photometry would require more than just a longer exposure to the starlight, however. In addition to measuring the count rate on target, we need to measure the count rate off target. The off target count rate has three components: dark counts, hot events and sky brightness. Sky brightness can be distinguished from dark counts and hot events by pointing the telescope at blank sky and measuring the count rate with the instrument shuttered and then uncovered. Dark counts can be distinguished from hot events by measuring the number of counts with the comparator thresholds set for single photo-electrons (dark counts) and then multiple photo-electrons (hot events) with the instrument shuttered.

Figure 4.1 shows a scatter plot of the measured low threshold count rate versus visual magnitude for the month of May, 2001. The notable feature of this plot is the count rate remains high, around 10^4–10^5 Hz, even for stars with increasing magnitudes above about magnitude seven. Clearly, these counts must represent dark counts and
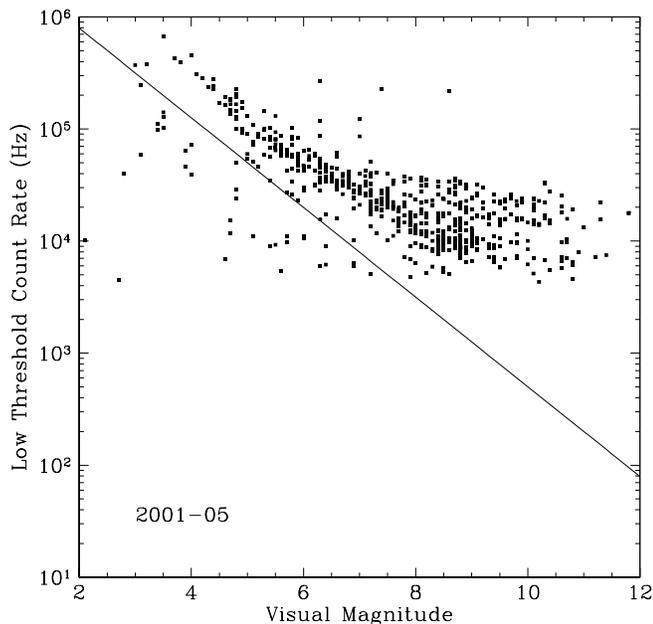
Figure 4.1: Scatter plot of low threshold count rate verses visual magnitude for observations from the Harvard observatory during May, 2001.

hot events, not stellar photons, whether or not the instrument was misaligned at the time the observations were made. However, the detectors are not saturated by these count rates, as the much higher rates for brighter stars reveal. What is clear, however, is that it would be impossible to determine if the detector is illuminated or not from only the low threshold count rate since the dark counts and hot events produce a low threshold count rate at least as high as the expected photon rate. For stars around magnitude eight, the difference between a shuttered and illuminated detector should be about a factor of two in the count rate, but without a shutter this difference cannot be measured. Furthermore, without knowledge of the spectral classification of the target we cannot even compute the *expected* count rate to within the required factor of two.

All of these measurements would require significant modifications of the instrument (for example, there is no shutter in the current configuration) and the observing routine, and high quality photometry was never our objective. However, high-quality

photometry considered as a side-effect of a complete and thorough detector characterization before each observation would be a big help for understanding our background. A number of very important questions have been left unanswered by the analysis, not least of which is what is the ultimate source of the triggers we see? It certainly seems that there are different mechanisms at work at the different observatories, and that at the Harvard observatory this mechanism shows a regular seasonal variation. Unfortunately, without more complete information about the condition of the detectors during these observations, it is very difficult to find a satisfactory answer. For example, the high threshold count rate measured during the dianostic phase of each observation is supposed to provide information about the hot event rate. However, since this rate is measured only during 0.3 seconds, it usually counts from zero to at most a few hot events. This means the underlying rate is measured very poorly, and something as essential as determining if this rate shows a seasonal variation corresponding to the trigger rate variation is very hard to do because of the small-number statistics involved.

4.2 Triggers and events

As was shown, the validity tests that we enforce on the triggers seem to pass a fairly large number of false positives. In the case of the Harvard observatory, almost a third of the triggers which we believe to be background are marked as events; at Princeton it is a much smaller percentage of a much larger number of triggers.

A significant limitation on our ability to improve these validity tests is the limited voltage resolution of the comparator array (which is used as a nonlinear two bit analog–to–digital converter) and the shallow LIFO depth of the MTD135 (maximum depth is sixteen events). This means that our ability to digitize and store the output waveforms from the detectors is fairly crude (see Figure 2.4 for an example). As was discussed, the hot events in the detectors have characteristic waveforms that are very different from the normal output when seen on an oscilloscope, and this was the basis of the hot event veto circuitry built into the instrument. If the summertime triggers

also had a characteristic waveform, and these waveforms could be stored in more detail, it would be much easier to distinguish the real photons from the background.

There are essentially four theories we are entertaining to try to explain the summertime increase in the trigger rate.

1. The number of hot events is going up during the summer. This is difficult to detemine from the data available because the high threshold count rate, which should be a reflection of the hot event rate, is measured for too short a time interval.

2. The gain of the APD is going up during the summer. There is a well known temperature dependence in the gain of an APD, but our power supply should be compensating for this. We have looked for a summertime gain increase by checking the detector outputs on an oscilloscope, and none was seen.

3. There is an increase in the amount of corona discharge and sparking in the detectors (which have a -7500 volt bias applied to them) during the summer because of increased humidity. This is the most likely explanation, although it is troubling that our efforts to keep the instrument dry seem to have had little effect on the trigger rate during summer.

4. The number of dark counts is going up during the summers. It is well known that the dark count rate in phototubes is temperature dependent; this is why they are often cooled in astronomical instruments. Futhermore, it can be seen from the plots in Appendix A that during the misaligned period from June 2001 until February of 2002 the number of low threshold counts went down by more than an order of magnitude as the temperatures cooled from summer to winter. However, it is not at all clear why dark counts, which normally do not exceed the high threshold setting, would cause triggers.

None of these theories is completely satisfactory with the information we have on hand at the moment. A deeper investigation of the problem seems called for,

perhaps involving cooling the detectors as well as controlling the humidity of their environment.

4.3   Two observatory operations

Although having a thorough understanding of the instrumental backgrounds and controlling them to the largest degree possible is the best experimental practice, two observatory operation makes it so easy to cleanly distinguish between the real triggers and the background seen at each observatory individually that controlling it becomes less important. It is hard to imagine any improvement that could be made at either observatory in isolation which would lead to as much of a reduction in the noise as using both in tandem does.

A number of new operational issues come up with a second observatory. One of the main goals of the design of the original, single observatory system was to minimize the burden on the radial velocity survey observers. However, coordinating observations at two geographically separated locations necessarily requires tighter integration of operations and some interaction between the observers at the two instruments. It is no longer pratical to try to maintain the illusion that the OSETI instrument "runs in the background" while the radial velocity survey runs the same way it always has.

Since the Princeton observer only receives notification when the Harvard observer moves to a new target, he has no way to tell when, for example, weather conditions at Harvard have deteriorated to the point so that observations have shut down for the night. To some extent, these sorts of problems can be solved with a telephone; however, it is a nuisance for the Harvard observer to be burdened with the obligation to call Princeton before every fifteen minute break.

A number of possible solutions have been discussed. One particularly attractive one would be to use voice-over-IP to implement a simple two-way intercom between the two observatories which could be activated simply by keying a microphone.
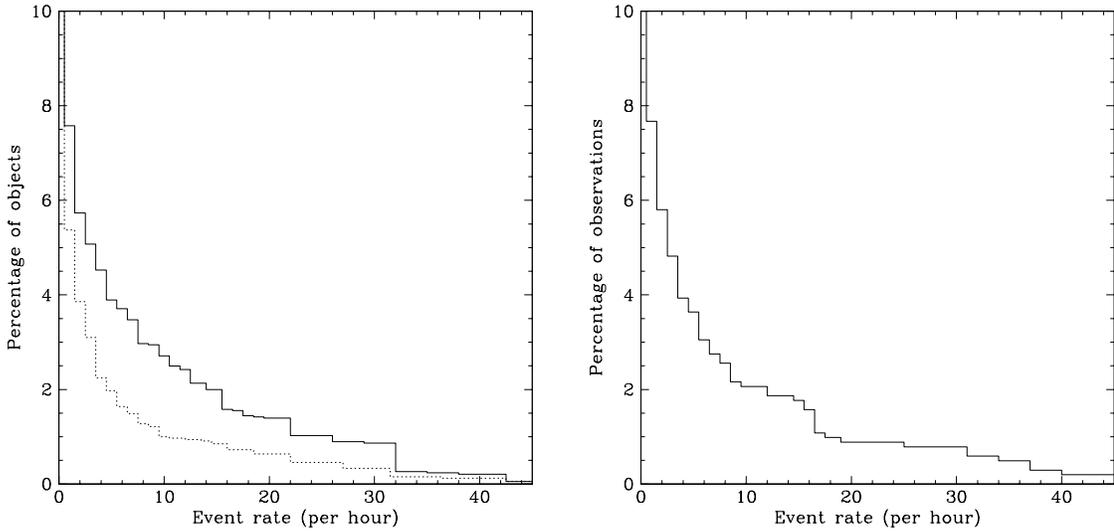
Figure 4.2: Histogram of the percentage of targets which recorded a given event rate or higher as a function of event rate for Harvard data (left) and Princeton data (right). In the left panel, the solid line shows data from the aligned period, the dashed line from the misaligned period.

4.4 Prevalence of transmitting civilizations

The Harvard data excluding the misaligned periods contains 8,555 observations of 3,802 objects with the average integration lasting 506 seconds. During these observations, 1,755 triggers and 467 events were recorded. We can use this information to set an approximate limit on the number of civilizations transmitting according to our model.

Suppose that the transmitter is illuminating the N stars nearest to his own, and that his transmitter can transmit F times per second. Then the time that each receiver has to wait between successive transmissions is N/F. N grows proportionally to the range cubed, and for a range R of 100 light years $N \approx 1000$, so $N = 10^{-3}R^3$. Define

$$\eta = 10^3 \frac{F}{R^3};$$

(4.1)

η is the event rate on Earth due to a signal from this transmitter.

The Figure 4.4 shows histograms of the percentage of targets which recorded a given event rate or higher as a function of the event rate. In the left panel, the solid line represents data from the Harvard observatory during periods of good alignment and the dashed line represents data from the misaligned period; the right panel represents all of the Princeton data. All Harvard summertime data was discarded because of the seasonal increase in the trigger rate; there is no summertime data from the Princeton observatory yet. The two histograms in the left panel include roughly the same number of individual objects: 3,802 from the "good" Harvard data and 3,293 from the "bad". The Princeton data includes a total of 1,017 distinct objects.

At both observatories, about 8% of the objects, recorded one event per hour or more. If we make the assumption that all of the events recorded so far are false alarms, then Figure 4.4 sets the limit on the sensitivity of the instrument in terms of η. In particular, we must have $\eta \geq 40$ events per hour in order to be 99% confident of a detection since less than 1% of our targets recorded as many triggers per hour accidentally. On the other hand, the fact that we did not see any event rates above 45 per hour allows us to set a limit on η for the stars within our target list. In other words, if there is an ETI transmitting to Earth from one of the stars in our target list, then $\eta < 45$ per hour for this transmitter. Since our target list is fairly complete out to 100 light years, we can say furthermore that $F < 12.5$Hz for any ETI within this distance.

This limit on F can be used to set an upper bound on the power used for transmission by ETIs within 100 light years of Earth. At the Harvard observatory, the overall system efficiency of 0.2% means that in order for us to receive four photons (the minimum requirement to exceed the comparator thresholds), 2,000 must be delivered to our effective aperture. We get 1/4 of the light from a 61-inch telescope, so our effective aperture diameter is 30.5-inches, or 0.47 square meters. This means that the incoming fluence of photons would have to be about 4,255 per square meter. As has already been shown in Section 1.5, the transmission strategy which is least demanding on the transmitter's astrometry is to illuminate a constant area *at the target*, adjusting the beam size for range as necessary. If we assume that the illuminated area

is circular with a diameter of 10 AU (an area of 1.75×10^{24} square meters), then the required output from the transmitter is 7.5×10^{27} photons. At a wavelength of 550 nanometers (2.25 eV), this corresponds to an energy of 2.7 gigajoules per nanosecond pulse. The average power required to transmit at 12.5 Hz is therefore 34 gigawatts, which is the electrical output of about 34 large nuclear reactors. This upper bound must be qualified with the observation that if a civilization with these resources at their disposal chose to expend them with lower energy pulses at a higher repetition rate then their pulses might not contain enough photons for us to detect them.

4.5 Final thoughts

In many respects our optical SETI program has been a pioneer. We were the first targeted search for pulses transmissions in optical wavelengths, and the first to use two observatories to veto false positives. As the Harvard optical SETI program moves forward, lessons learned from this experiment will be incorporated into the design of future experiments, notably the all sky survey instrument which is currently in an advanced stage of development. This instrument will implement a number of the recommendations made here, including a more detailed digitization of the detector waveforms and better photometry by monitoring low and high threshold count rates continuously during integration. The detector system being developed for the all sky survey is conceptually just a multi-pixel version of the one used in this experiment, but this seemingly simple additional requirement engenders a vastly more sophisticated system. The only reasonable way to implement such a system is to build a custom integrated circuit, which meant that we were free to design it with exactly the functionality that is needed.

The biggest advantage the all sky survey will have is the increase in sky coverage. If, as Freeman Dyson suggests, interstellar communications will not necessarily come from the vicinity of stars, then any complete SETI program will have to fill in the spaces between. Our targeted program which observed 8,000 stars with a field of view of one arcminute, covered less than 0.005% of the sky; the rest will have to come later.

Appendix A

Misalignment

This appendix contains plots of low threshold count rate versus visual magnitude for all of the observations made during the first three and a half years of the Harvard optical SETI program, grouped chronologically by month. The instrument has been removed from and then replaced on the telescope a number of times during this period to make the telescope available for other instruments and for maintenance on the telescope. Table A displays the information we have on these events from the director's log files (thanks to Robert Stefanik for providing this information). "Astronomy 191" is an undergraduate course at Harvard which uses a CCD detector on the 61-inch telescope, requiring removal of the echelle spectrograph and with it our optical SETI instrument. The effect of the mirror recoating in October 2000 is especially noticeable in the plots that follow; the signal strength improved by almost an order of magnitude over previous well-aligned periods.

Table A.1: Removal and replacement of the optical SETI instrument

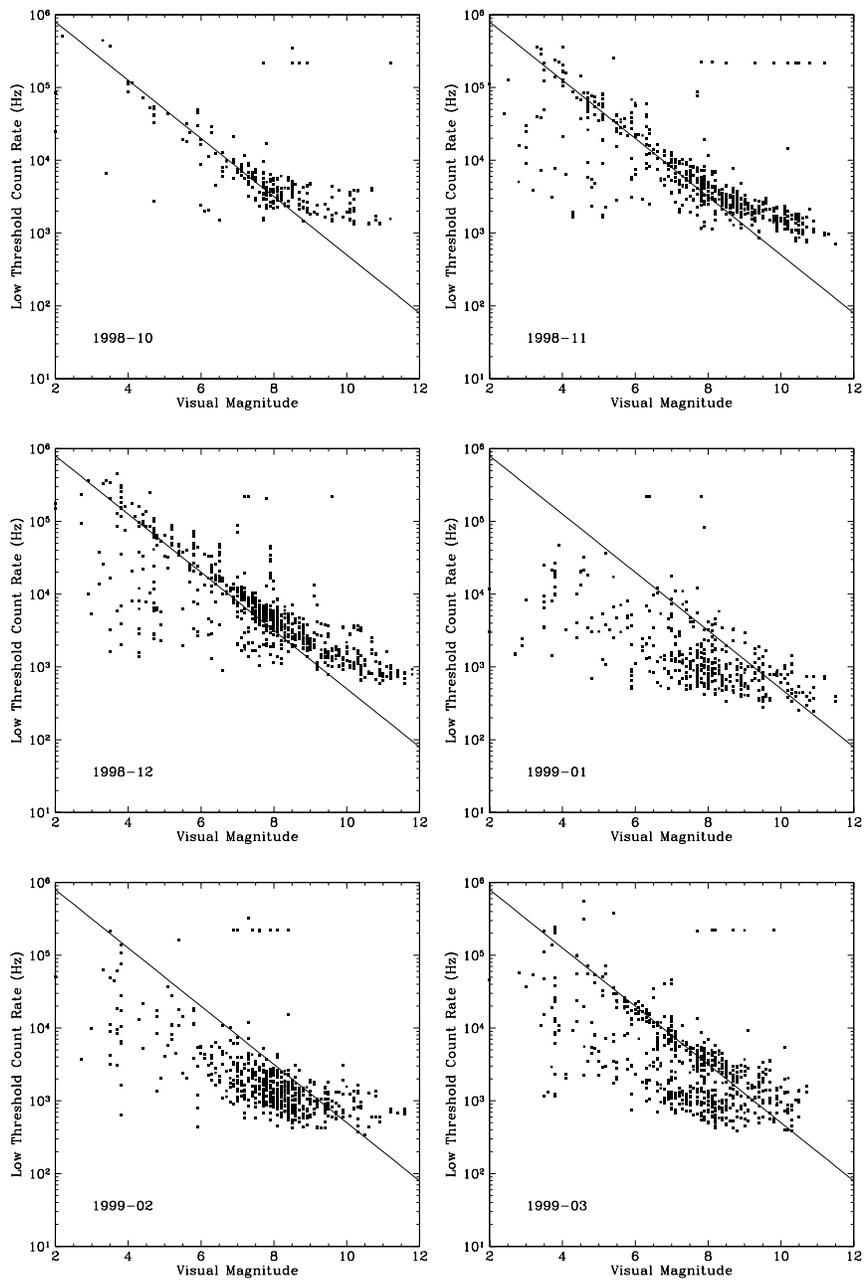| Removed	Replaced	Comments
17 February 1999	23 February 1999	Astronomy 191
10 April 2000	15–23 April 2000	Astronomy 191
9 October 2000	24 October 2000	Mirror recoating
13 April 2001	20 April 2001	Astronomy 191
16 February 2002	2–3 March 2002	Astronomy 191

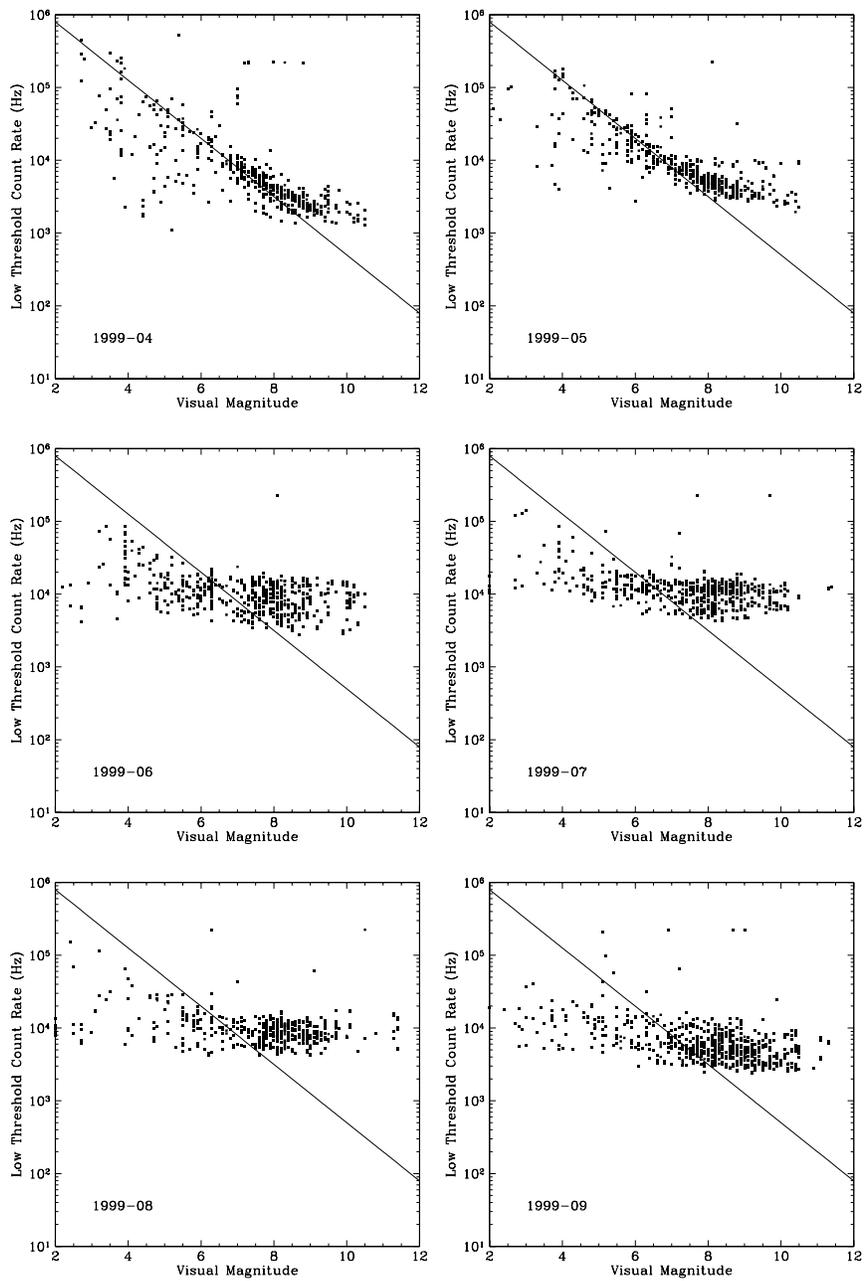Figure A.1: October 1998 through March 1999

76

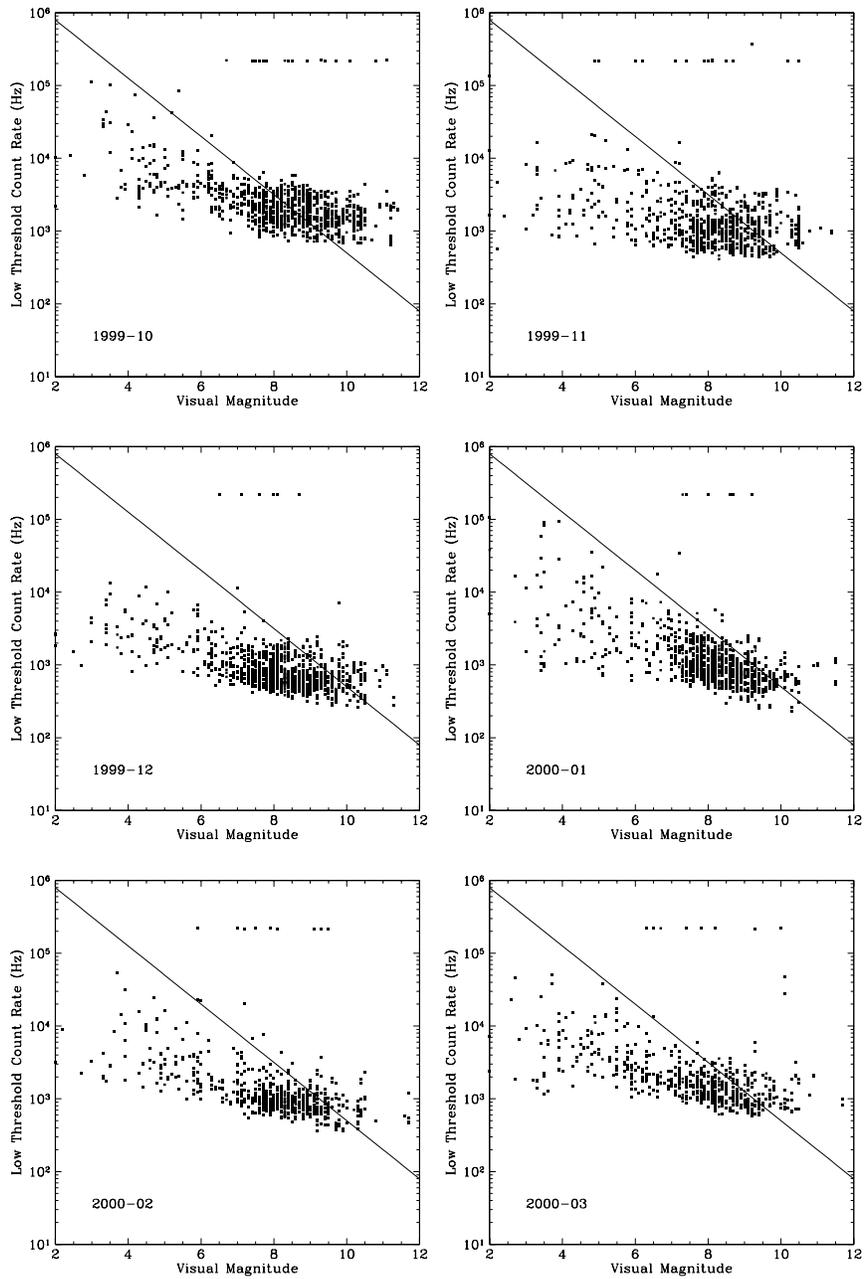Figure A.2: April 1999 through September 1999

77

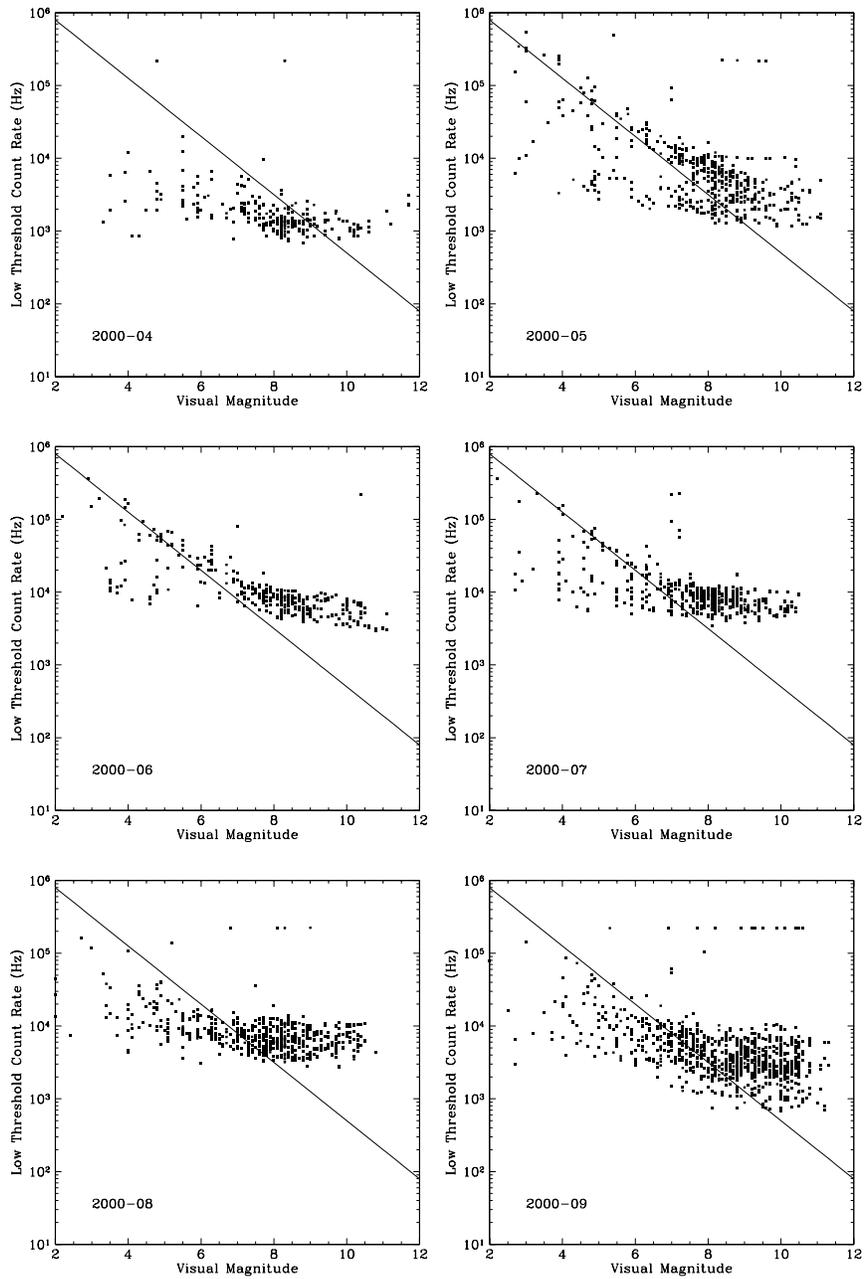Figure A.3: October 1999 through March 2000

78

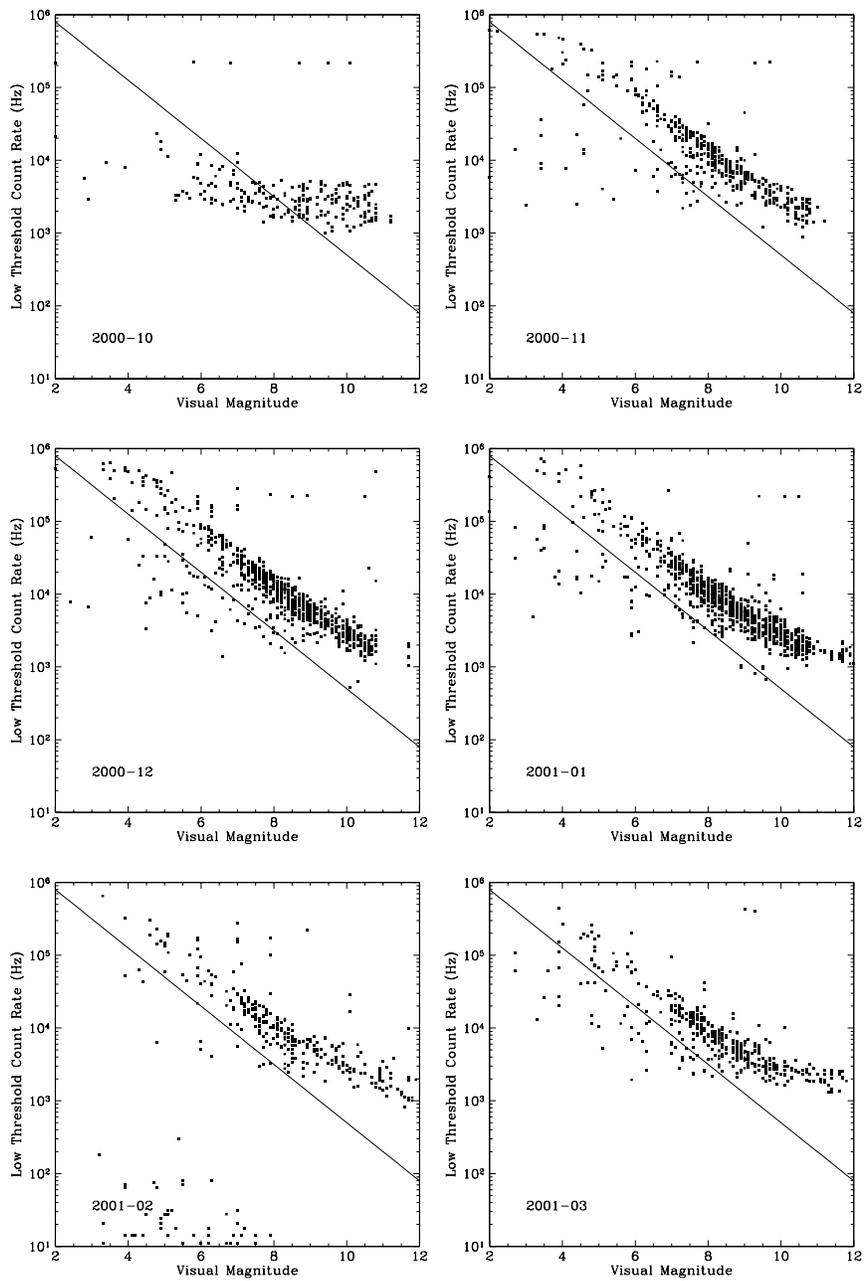Figure A.4: April 2000 through September 2000

79

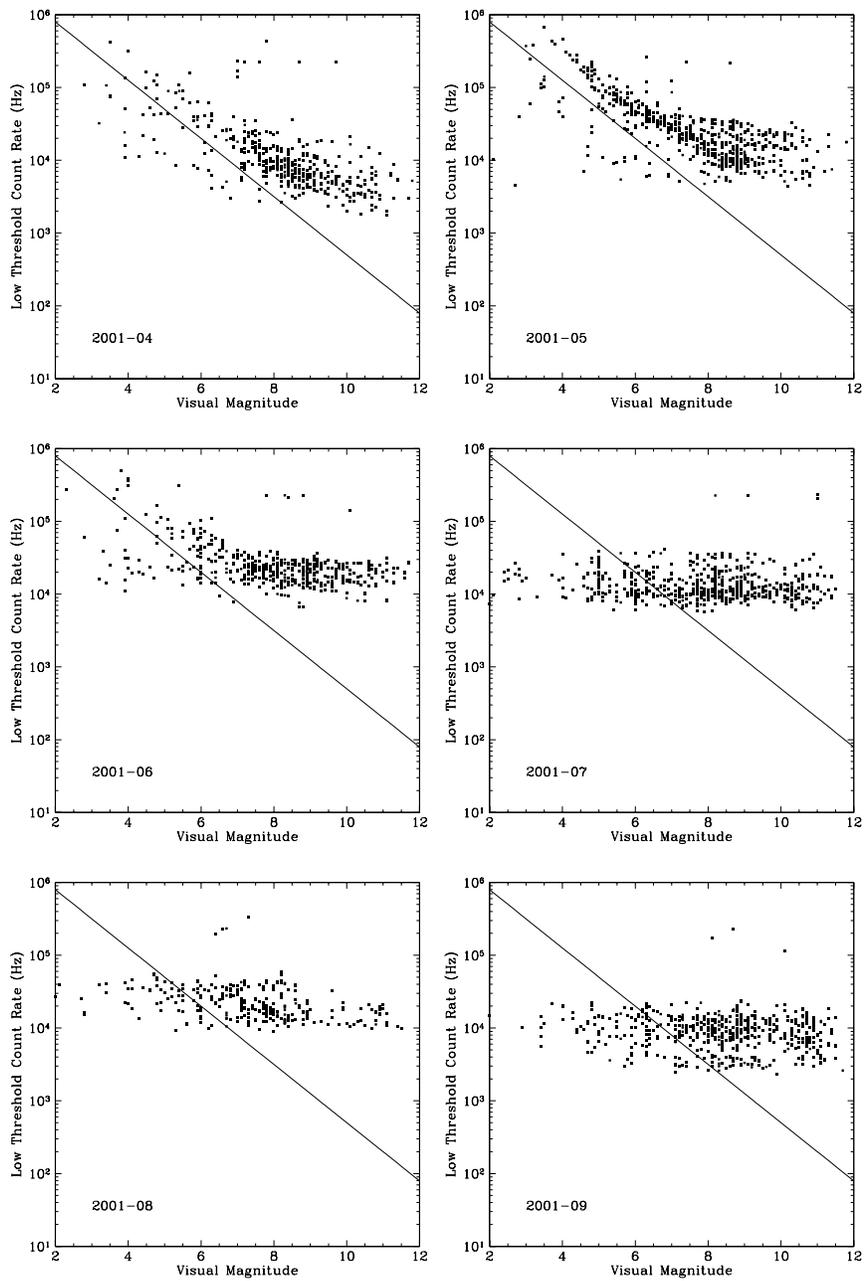Figure A.5: October 2000 through March 2001
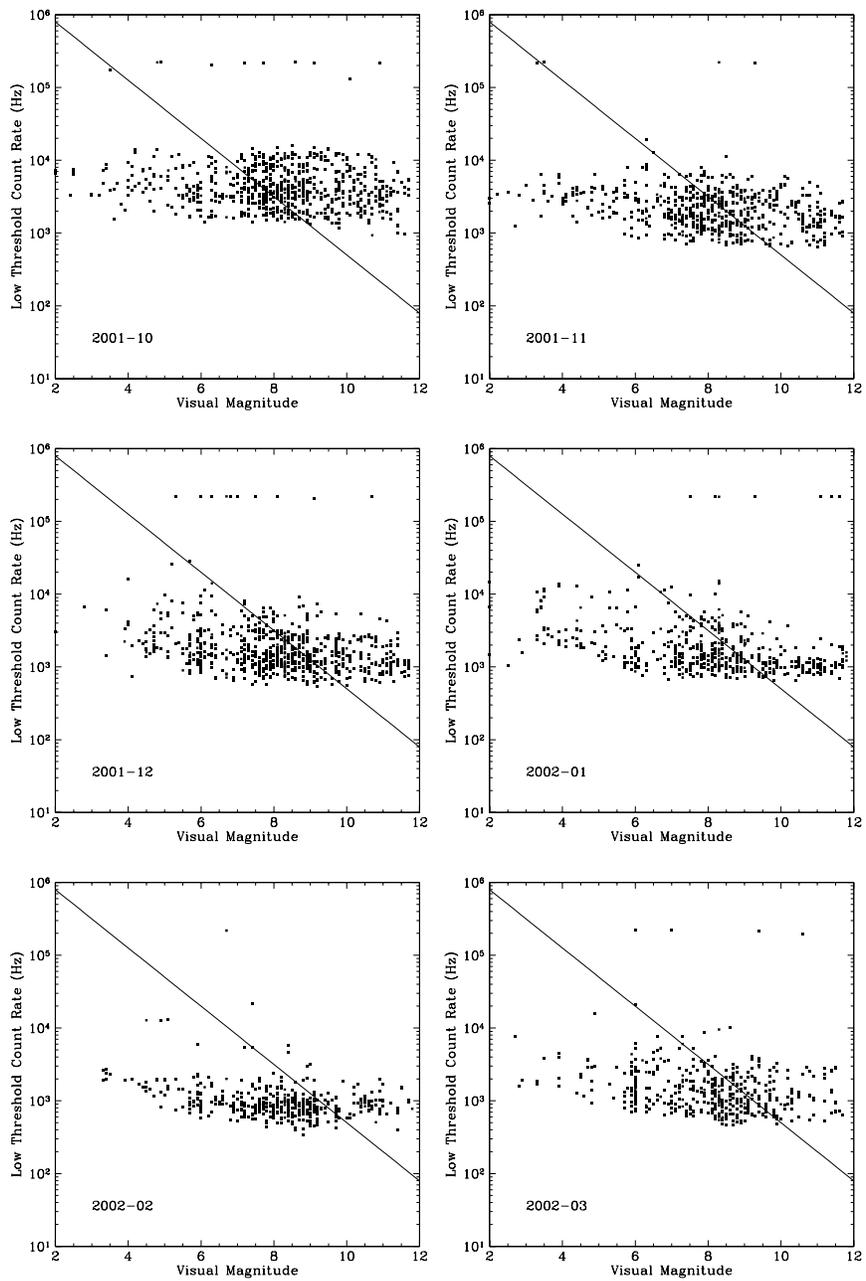
Figure A.6: April 2001 through September 2001

81

Figure A.7: October 2001 through March 2002

Bibliography

[1] Per A. Aannestad and Edward M. Purcell. Interstellar grains. *Annual Review of Astronomy and Astrophysics*, 11:309–362, 1973.

[2] European Space Agency. *The Hipparcos and Tycho Catalogs, Volume 1: Introduction and Guide to the Data*. ESA Publications Division, Noordwijk, The Netherlands, 1997.

[3] C. W. Allen. *Astrophysical Quantities, 4th Edition*. Springer, Berlin, 2000.

[4] R. L. Carlson, J. P. Carpenter, D. E. Casperson, R. B. Gibson, R. P. Godwin, R. F. Haglund, J. A. Hanlon, E. L. Jolly, and T. F. Stratton. Helios: A 15 tw carbon dioxide laser-fusion facility. *IEEE Journal of Quantum Electronics*, 17(9):1662–1678, 1981.

[5] Giuseppe Cocconi and Philip Morrison. Searching for interstellar communications. *Nature*, 184(4690):844–846, 1959.

[6] J. M. Cordes. Pulse broadening in the optical and ir from interstellar grains. *SETI 2020: The Report of the SETI Science and Technology Working Group*, 1999.

[7] J. M. Cordes and T. J. Lazio. Interstellar scattering effects on the detection of narrow-band signals. *Astrophysical Journal*, 376:123, 1991.

[8] D. Dravins. Astrophysics on the shortest timescales. *The Messenger*, 78:9–19, 1994.

[9] Particle Data Group. Review of particle properties. *Physical Review D*, 50:1173–1200, 1994.

[10] Andrew Howard and Paul Horowitz. Is there rfi in pulsed optical seti. 2001.

[11] Donald J. Hutter. The usno astrometric interferometer. *Proceedings of the SPIE*, 2200:81–88, 1994.

[12] Nicolai Kardashev. Soviet astronomy. *Astronomical Journal*, 8, 1964.

[13] John D. Kraus. *Antennas*. McGraw-Hill, New York, second edition, 1988.

[14] William F. Krupke. An update on high power lasers. *SETI Science and Technology Working Group Meeting Number 2*, 1998.

[15] Kenneth R. Lang. *Astrophysical Formulae, vol. 1, 3rd Edition*. Springer, Berlin, 1999.

[16] Robert A. Laskin and Jeffrey W. Yu. Overview and status of the space interferometry mission (sim) technology development. *Proceedings of the SPIE*, 4006:1014–1028, 2000.

[17] General Accounting Office. National ignition facility: Management and oversight failures caused major cost overruns and schedule delays. *GAO Report*, GAO(RCED-00-271).

[18] R. N. Schwartz and C. H. Townes. Interstellar and interplanetary communication by optical masers. *Nature*, 190(4772):205–208, 1961.

[19] C. H. Townes. At what wavelengths should we search for signals from extraterrestrial intelligence? *Proceedings of the National Academy of Science*, 80:1147–1151, 1983.