Physics 728

Radio Astronomy: Lecture #1

Prof. Dale E. Gary
NJIT

Introduction to Radio Astronomy

Overview of Radio Emission from Astronomical Objects

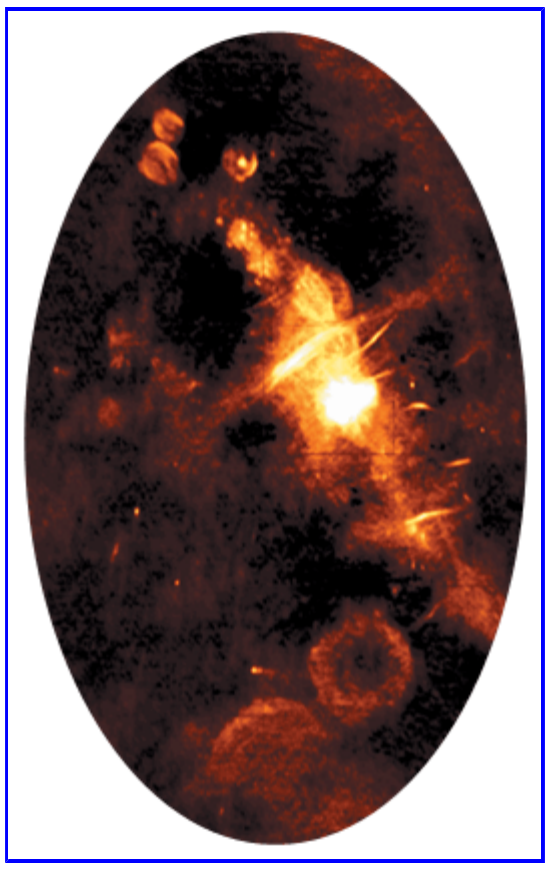
The Radio Sky

When we look at the sky at night with our unaided eyes, we see about 2000 stars of various levels of brightness, and if we are far from city lights we may see the faint band of the Milky Way, which is the light from billions of stars making up our galaxy. But if our eyes were able to see radiowaves, the sky might look like the image below.

(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation

It may appear similar to the starry sky, but in fact most of the point-like objects are not stars, but luminous radio galaxies billions of light years away. The larger sources are ionized clouds of hydrogen, or supernova remnants.

Looking toward the center of our galaxy, our radio eyes would see a large variety of strange features, most of which are not visible in other wavelengths.

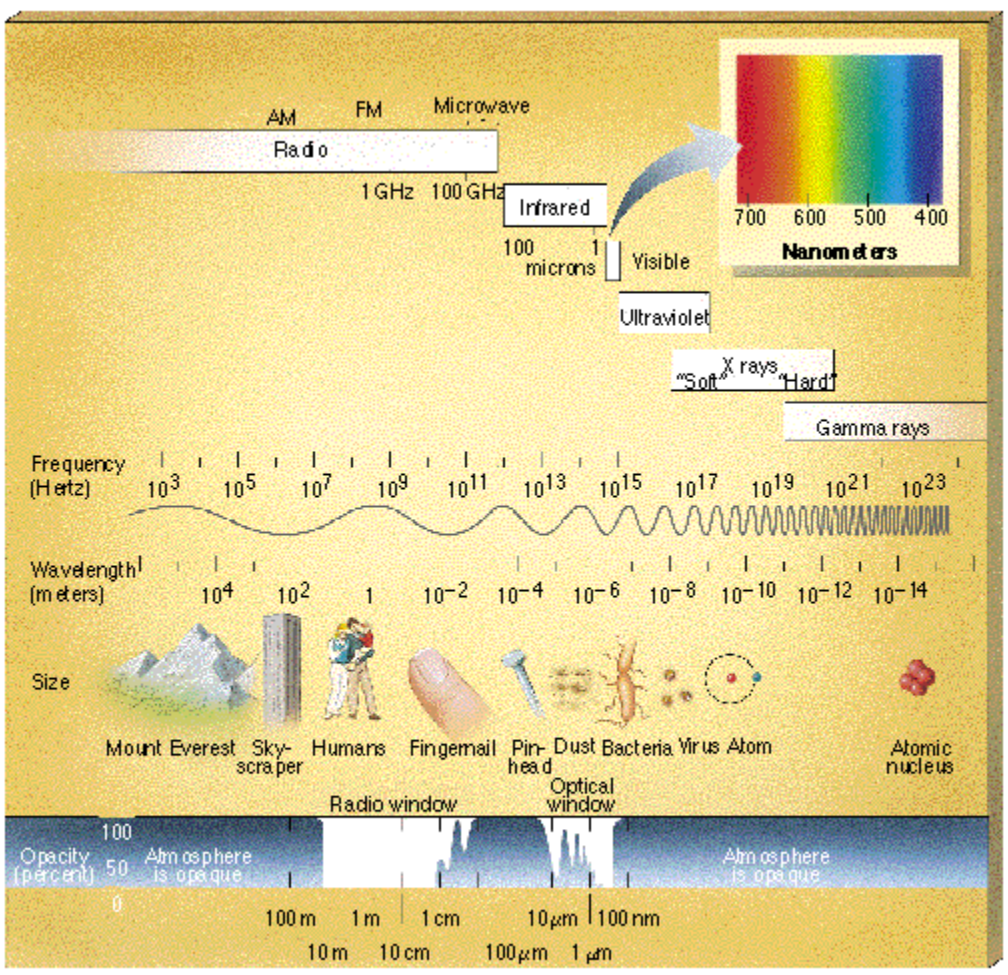


The Galactic Center - A Radio Mystery
Credit: N. E. Kassim, D. S. Briggs, T. J. W. Lazio, T. N. LaRosa, J. Imamura (NRL/RSD)

The Electromagnetic Spectrum

You should all be aware of the different types of radiation that make up the electromagnetic spectrum, but it is worthwhile to list the names of the different types of emission, in energy, or frequency) order:

- gamma rays (> ~1 MeV)
- hard X-rays (10-1000 keV)
- soft X-rays (1-10 A)
- EUV (~100 A)
- UV (~1000 A)
- visible (4000-7000 A -- 400-700 nm)
- near IR (~1 micron)
- IR (10 microns)
- THz (~100 microns--3000 GHz)
- submillimeter (300 GHz - 700 GHz)
- millimeter (30 GHz - 300 GHz)
- microwave (3 GHz - 30 GHz)
- decimeter (300 MHz - 3 GHz) ("cable" TV/UHF band)
- meterwave (30 MHz - 300 MHz) (TV/FM/HF band)
- dekameter (3 MHz - 30 MHz) (Shortwave
- AM band (0.5 MHz - 1.7 MHz) etc.



Note that the units change as we go from top to bottom--use energy units near the top, then switch to wavelength units, then switch to frequency units. This is purely a matter of convenience and convention. We could stick with energy, or wavelength, or frequency throughout, but the range of 6 or 7 decades makes it inconvenient to stick with one measure. The relationships among energy, frequency, and wavelength are, of course:

$$E = h\nu = hc/\lambda.$$

For the purposes of this course, we will be concentrating on techniques of interferometry and synthesis imaging that work for the range from submillimeter to dekameter, although there are practical difficulties at both extremes, and it is currently most common to use interferometry in the millimeter to meterwave range. There are on-going efforts to extend interferometry to both higher frequencies (submillimeter--ALMA, THz--NJIT project) and lower frequencies (space arrays).

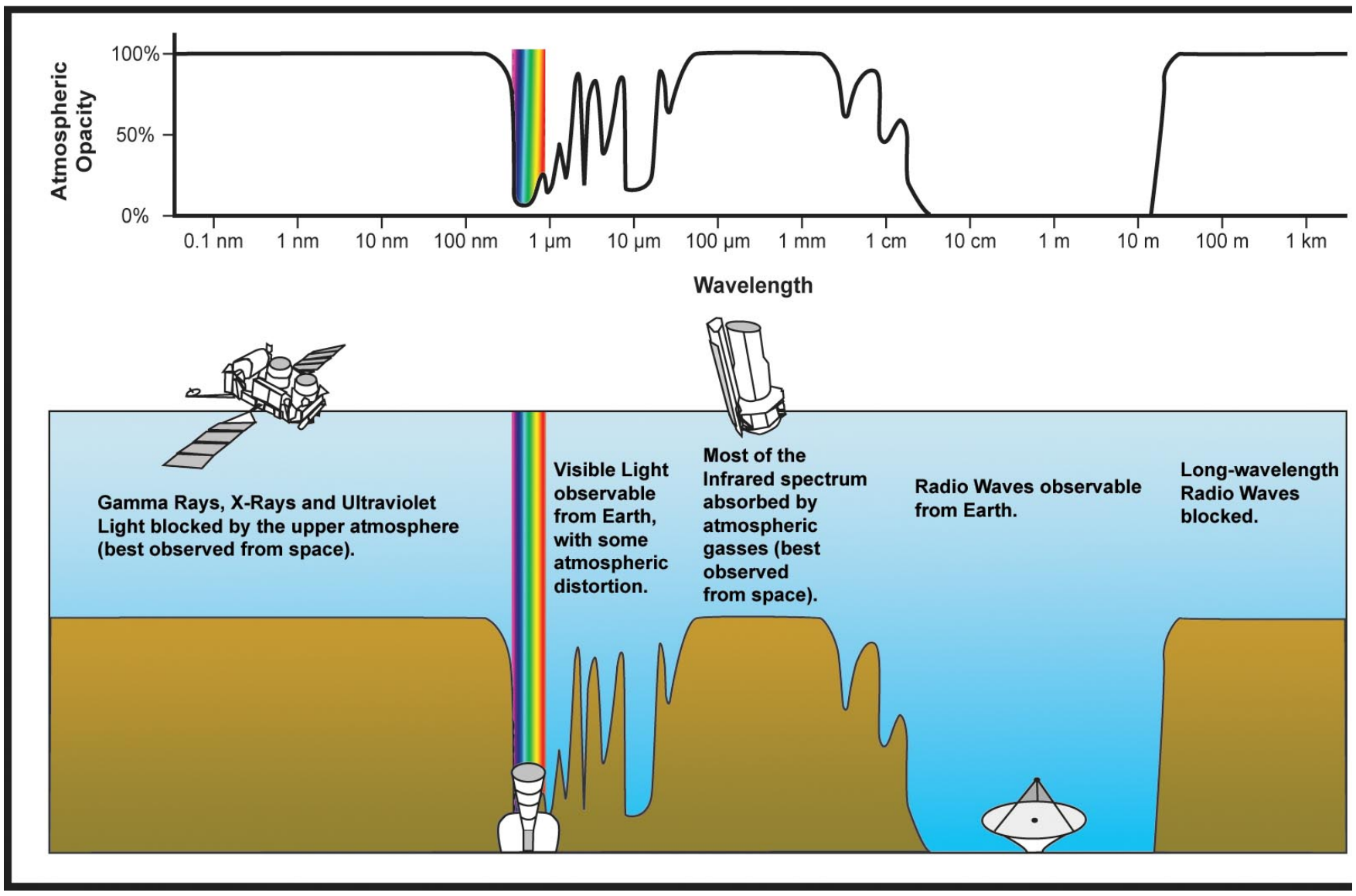
Why Observe At Radio Wavelengths?

There are many reasons why it is advantageous to observe at radio wavelengths.

Advantages of Radio

- Radio waves reach the ground
- Can observe objects or phenomena that are difficult or impossible to detect in other wavelength ranges
- Can use radio emission for quantatitive physical diagnostics of object parameters

The first reason is simply that it is possible to observe radio waves from the ground. As shown in the figure below, spacecraft are needed to observe astronomical objects in gamma rays, X-rays, UV, and IR, while ground observations are possible in the visible, some parts of the near IR, and the radio. NJIT has solar observatories exploiting all of these ground windows.



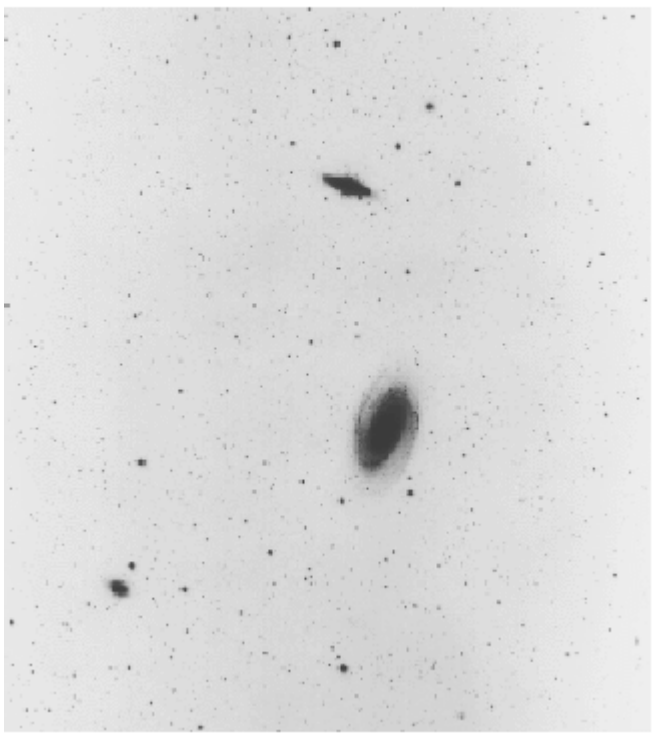
Credit: NASA/IPAC

Note that the window closes at the long-wavelength end of the spectrum--not because of the atmosphere, which remains transparent to long-wavelength radio waves--but rather due to the ionosphere, which reflects the radiation.

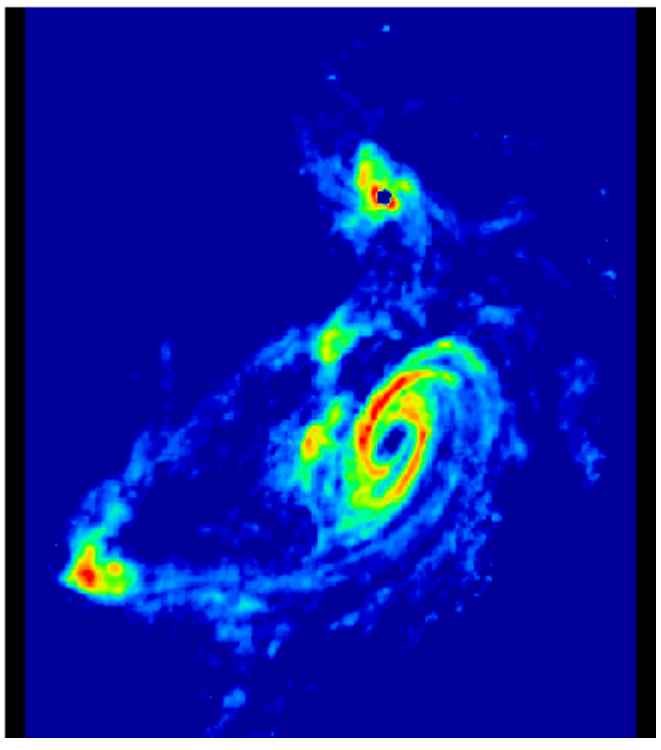
A second reason is that some objects and phenomena are invisible or hard to detect in other wavelengths, and can only be seen, or can be seen with greater sensitivity, in the radio. Here are a few of many many examples from which we could choose:

TIDAL INTERACTIONS IN M81 GROUP

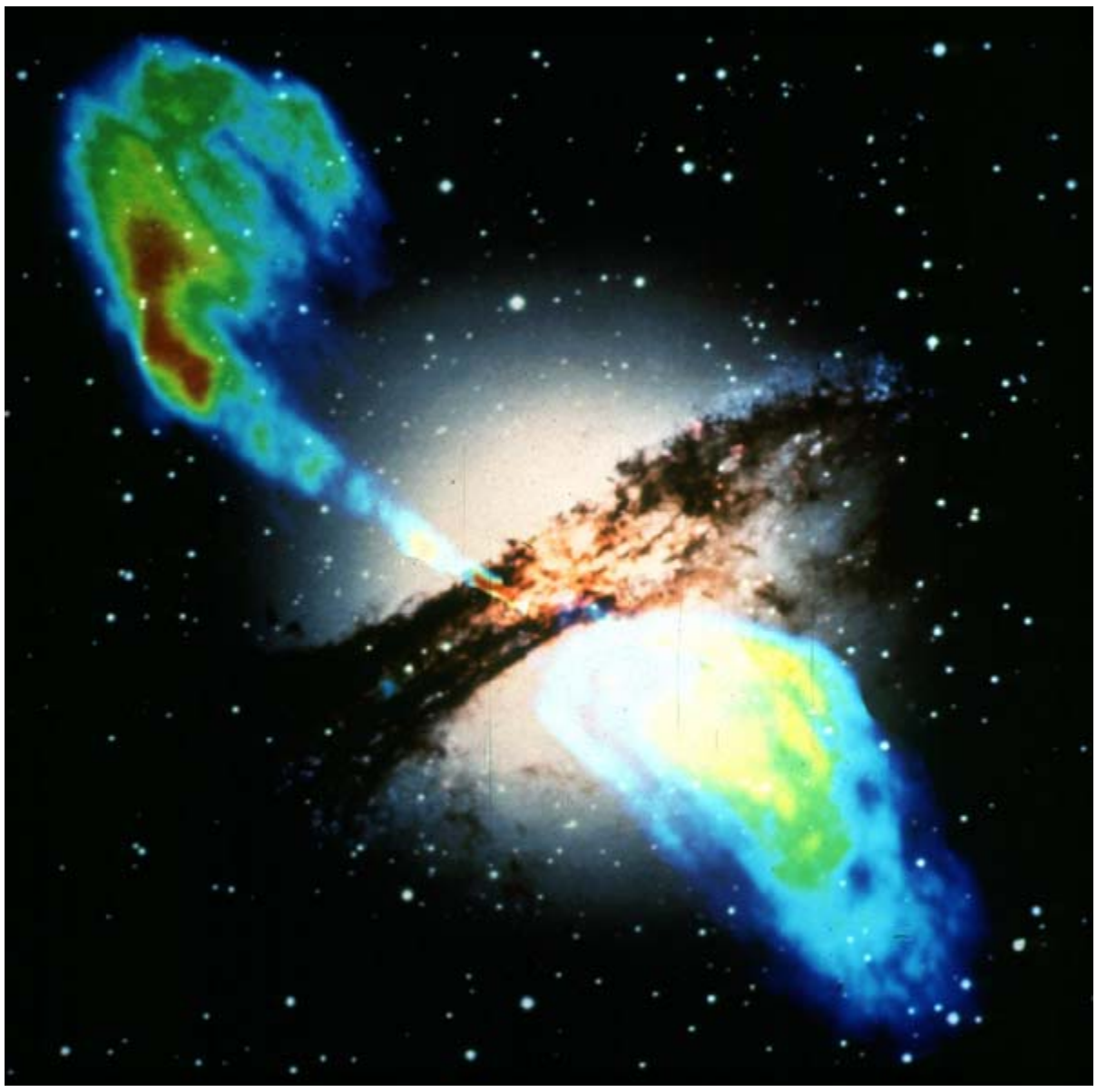
Stellar Light Distribution



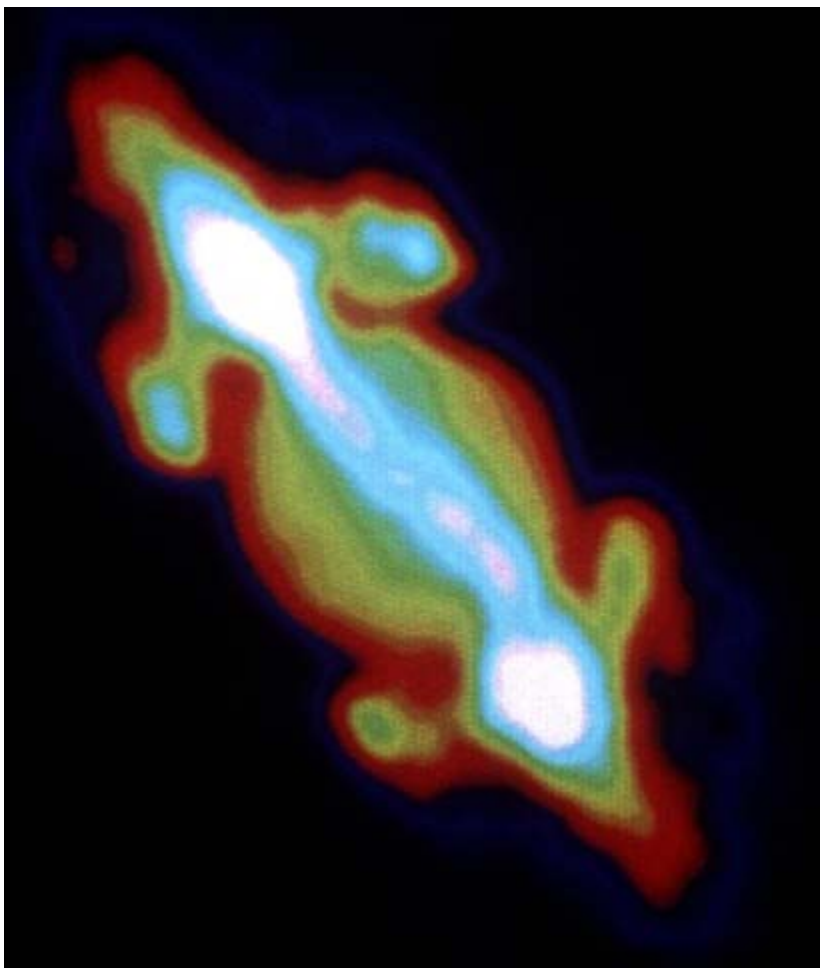
21cm HI Distribution

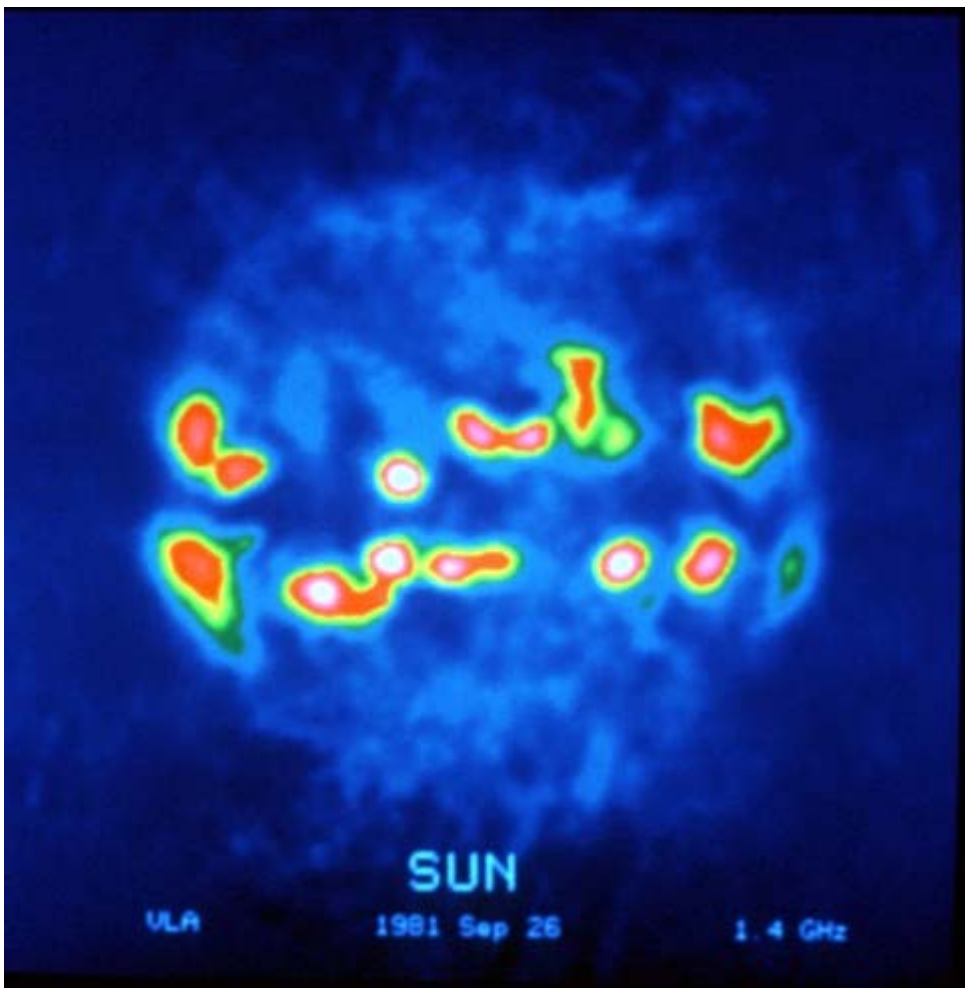


Neutral hydrogen traces interactions among galaxies in the M81 group.



Centaurus A -- peculiar galaxy with radio lobes. From HST web site.





The third important reason to explore astronomical objects in radio wavelengths is that the emission properties provide quantitative physical information about conditions in the source. We will see that radio emission is produced in a large number of ways. The low-energy radio photons are relatively easy to produce, which makes radio emission sensitive to a great many parameters. However, the number of mechanisms is itself a problem. Before one can use the emission to give information, one must first determine which radio emission mechanism is responsible for the emission. In practice, the most accurate way to determine the emission mechanism is to have ***spectral*** information, since different emission mechanisms have different characteristic spectral properties. In addition to helping to determine the emission mechanism, quantifying spectral properties such as peak brightness, peak frequency, spectral slopes, etc., also provides quantitative diagnostic parameters.

For all of these reasons and more, the radio range of wavelengths is as essential as gamma ray, X-ray, UV, optical, and IR for providing a complete picture of the physical nature of astronomical sources.

Overview of Radio Instrumentation

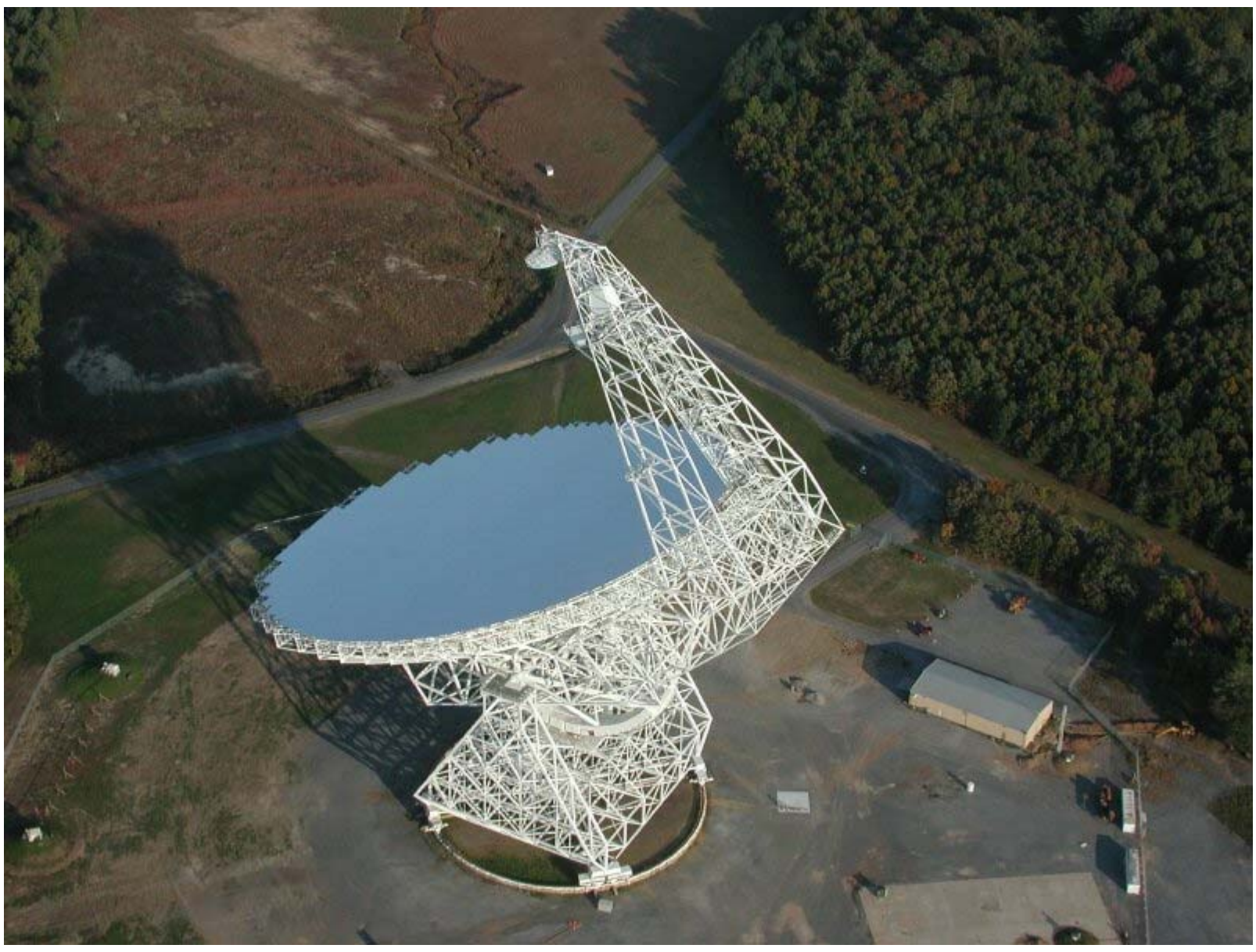
What Is Different About Radio Instrumentation?

Astronomical telescopes that work in the radio range look and operate very differently from the more familiar optical instrumentation. In fact, the "radio" range is so broad (6 or 7 orders of magnitude) that instruments at the low end of the frequency range look very different from those at the high end. We will take a brief look at the differences, using some existing or planned telescopes as examples.

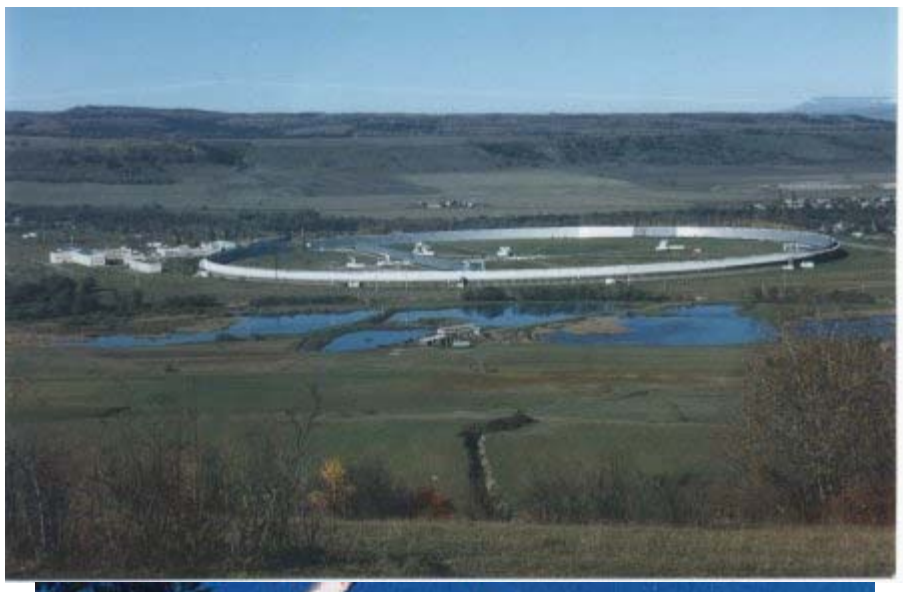
Single Element Instruments

The term "single element" means either single parabolic dishes, or in some cases single dipole elements. Here are a few pictures:

Arecibo: The largest single dish in the world, 306 m
(c) Cornell University / National Science Foundation

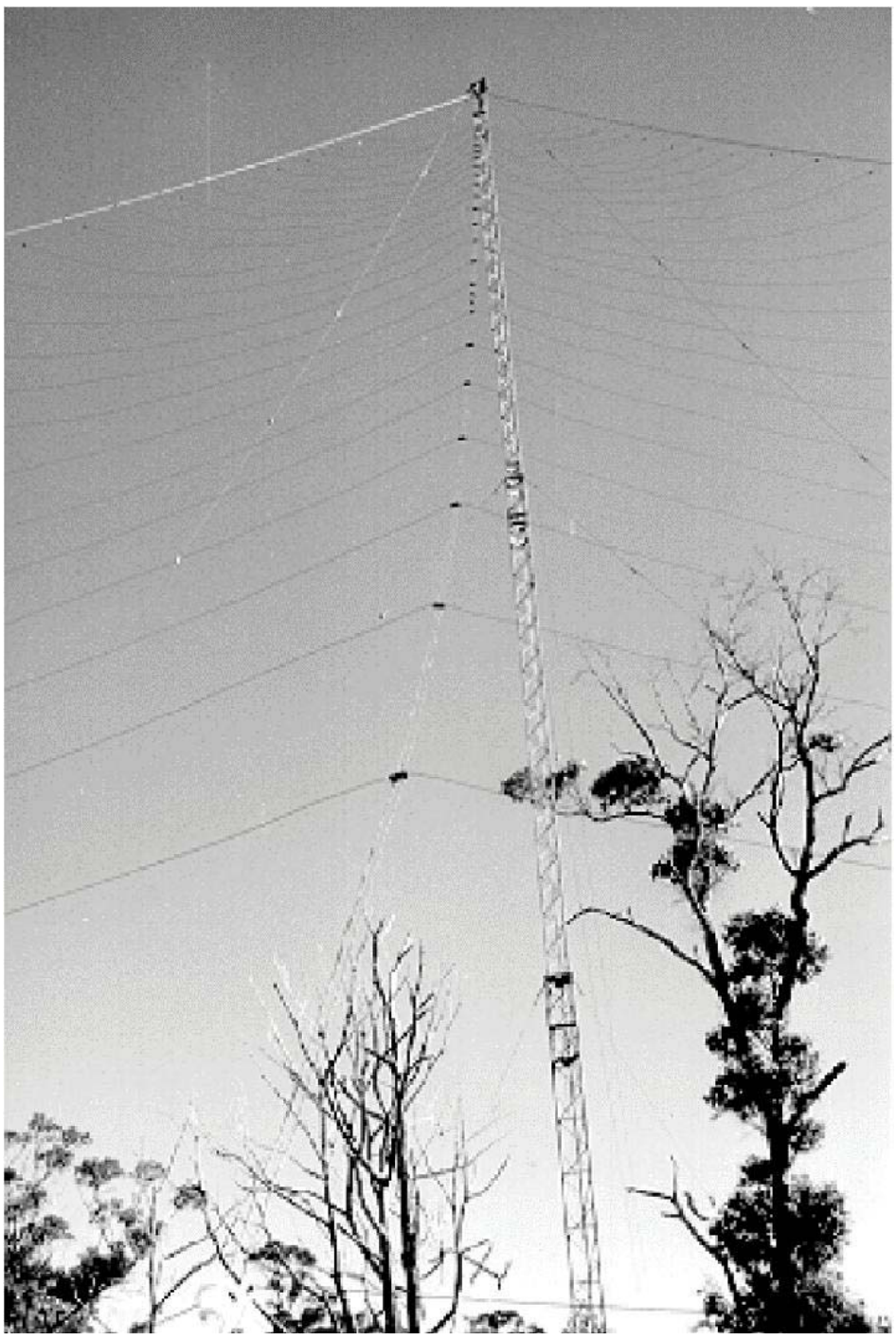


Green Bank Telescope (GBT): The largest fully steerable single dish in the world, 100 x 110 m
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation



RATAN 600: Diameter 600 m, part of a "dish" reflecting surface

Metsahovi: Large mm dish



Bruny Island Radio Spectrograph

Single radio elements have limited spatial resolution (the diffraction limit of the telescope). This diffraction limited resolution is proportional to wavelength (as it is also for optical telescopes, but it seems more extreme for radio telescopes due to the huge range of wavelengths over which they are typically used). The diffraction limit for a circular aperture of diameter D is $\theta \sim 1.22\ \lambda/D$, where θ is the angular diameter of the Airy Disk at the half-power point (the full-width-half-maximum, or FWHM) in radians. At a frequency of 5 GHz, even the Arecibo dish has an angular resolution of only about 50 arcseconds. The fully-steerable GBT has a resolution at this frequency of only 150 arcseconds.

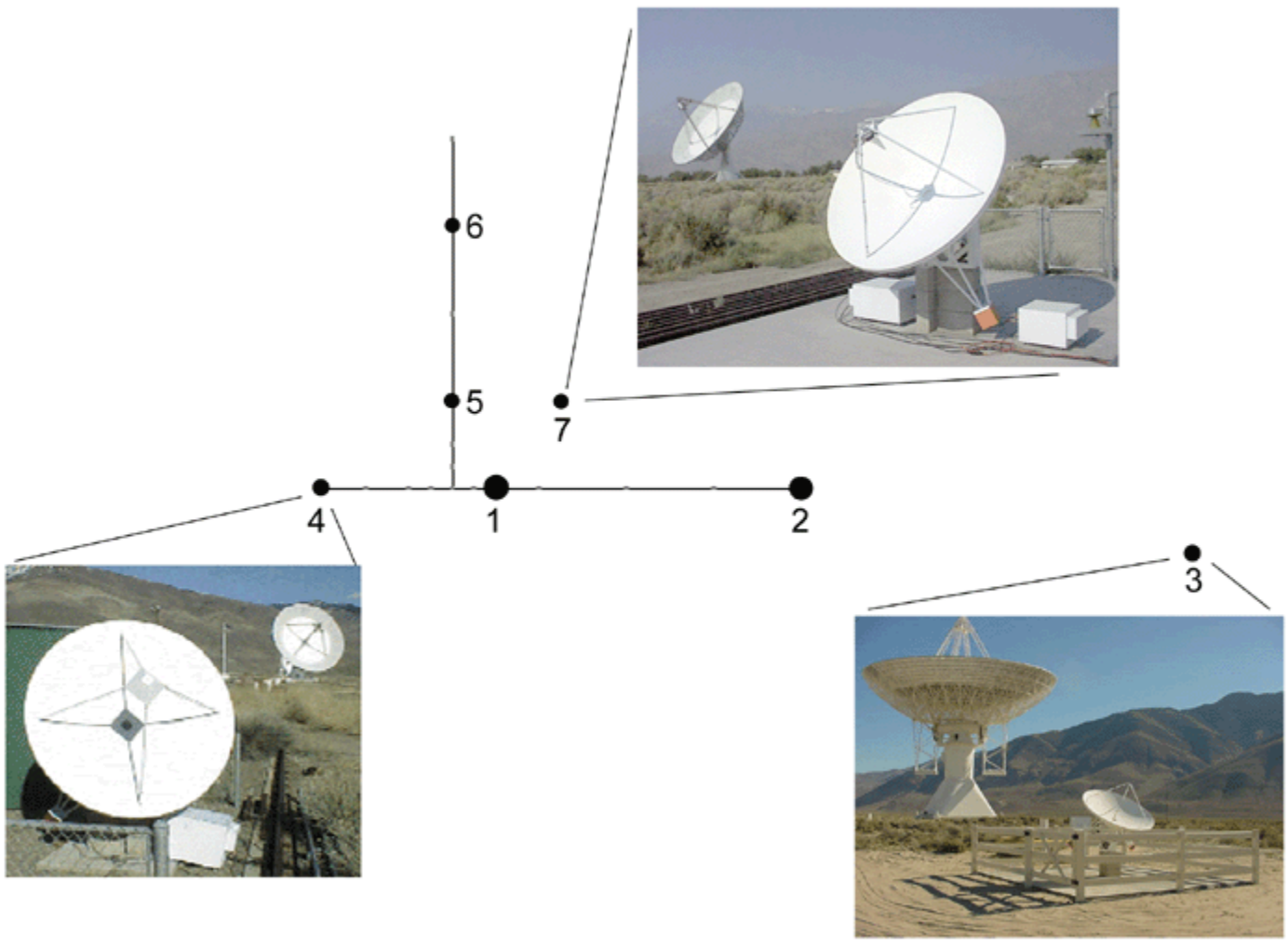
Because of the limited spatial resolution of single element telescopes, sophisticated techniques have been developed to combine single elements into multiple-element arrays, which work together to form a single telescope. In such arrays, the spatial resolution is determined not by the size of the individual elements, but rather by the maximum separation between elements, which is referred to as the ***baseline length***, B. With an interferometer, the diffraction limit is $\theta \sim \lambda/B$, where B can extend to many (even thousands of) km.

Interferometers

We now show some examples of interferometer arrays:

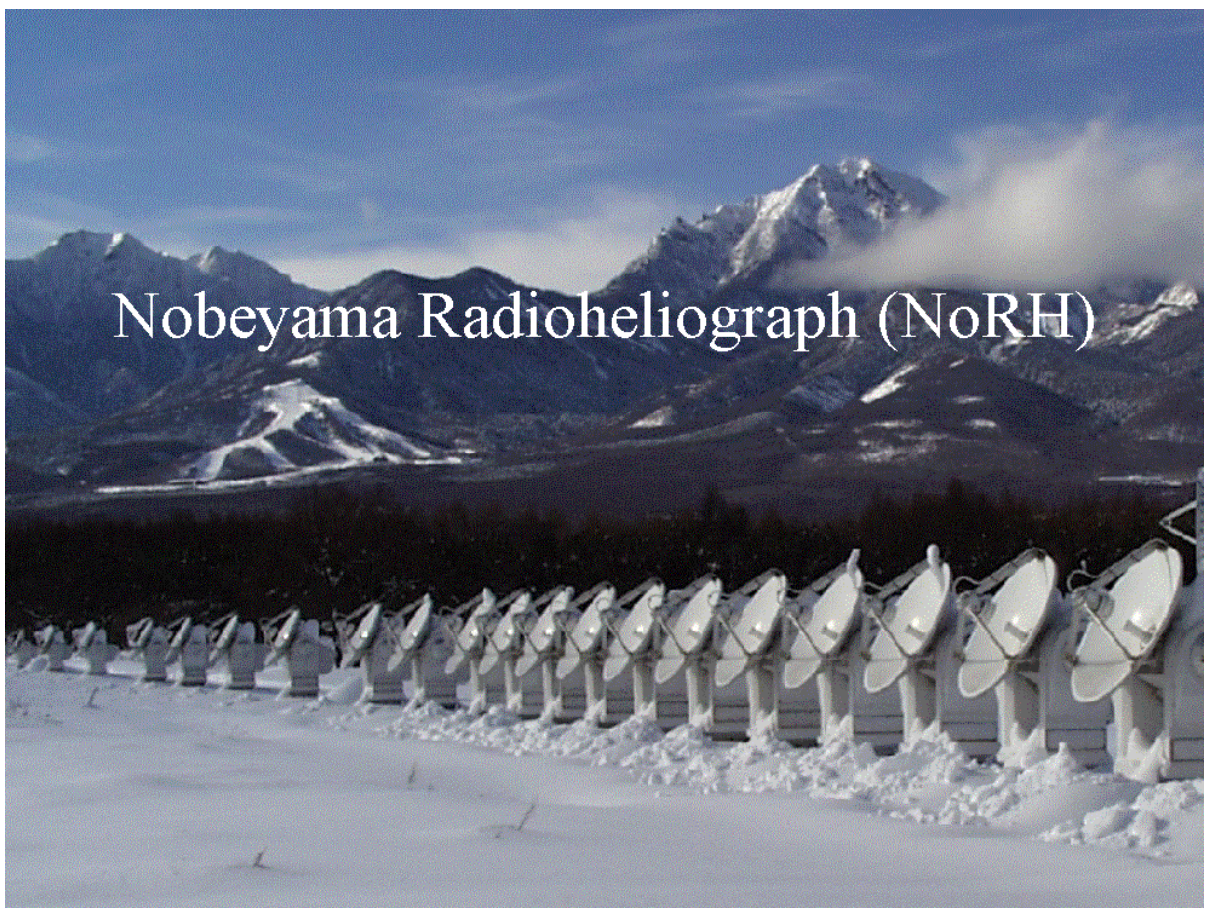
Close-up of VLA (Very Large Array)
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation

Aerial view of VLA in its most compact configuration.
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation



Owens Valley Solar Array (OVSA)

North arm of the Nancay Radioheliograph (Meudon, Observatoire de Paris)



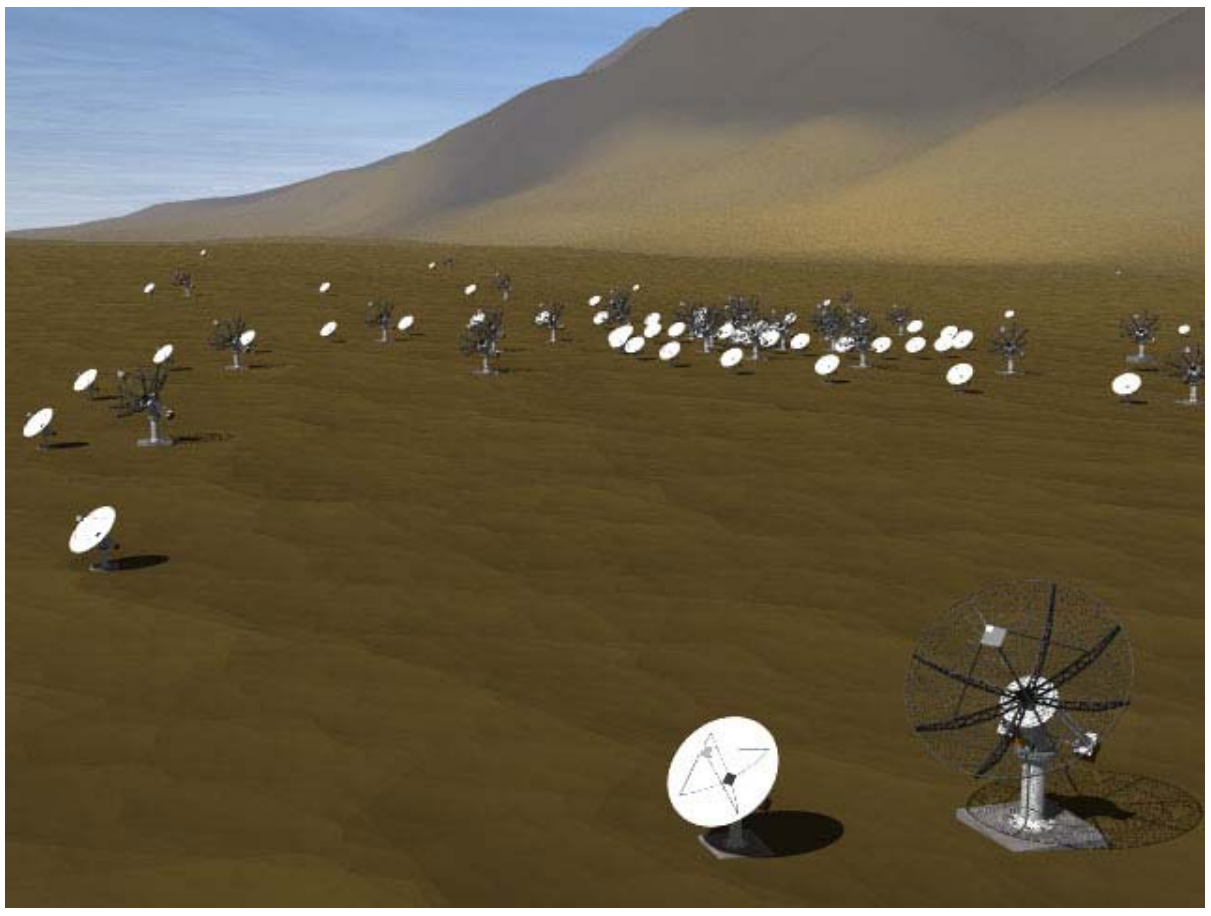
Nobeyama Radioheliograph (National Astronomical Observatory of Japan)

One of 10 antennas of the Very Long Baseline Array (VLBA) -- this one at OVRO.
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation

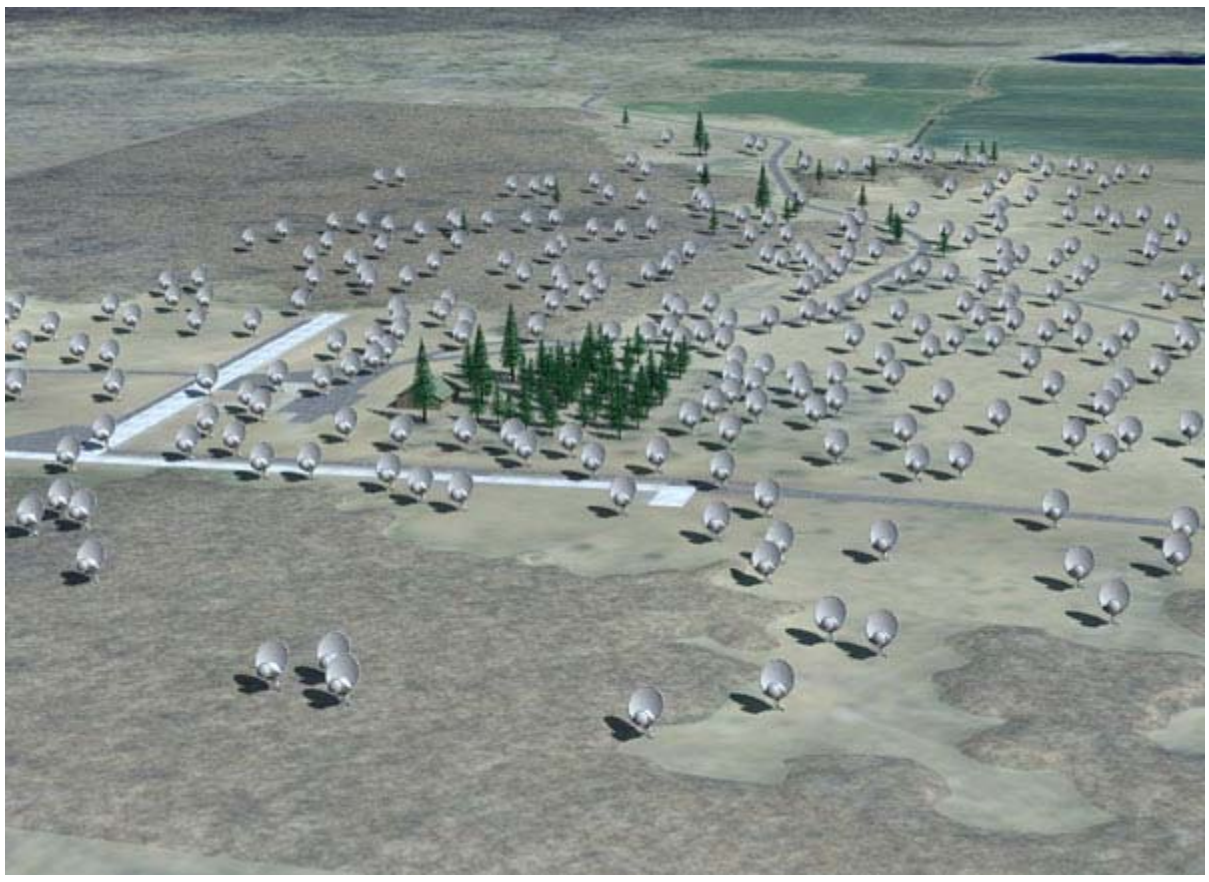
Site locations for the entire 10-station VLBA
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation

Future arrays, now under consideration/construction:

- Frequency Agile Solar Radiotelescope (FASR)



- Allen Telescope Array (ATA)



- Low Frequency Array (LOFAR)

- Atacama Large Millimeter Array (ALMA)

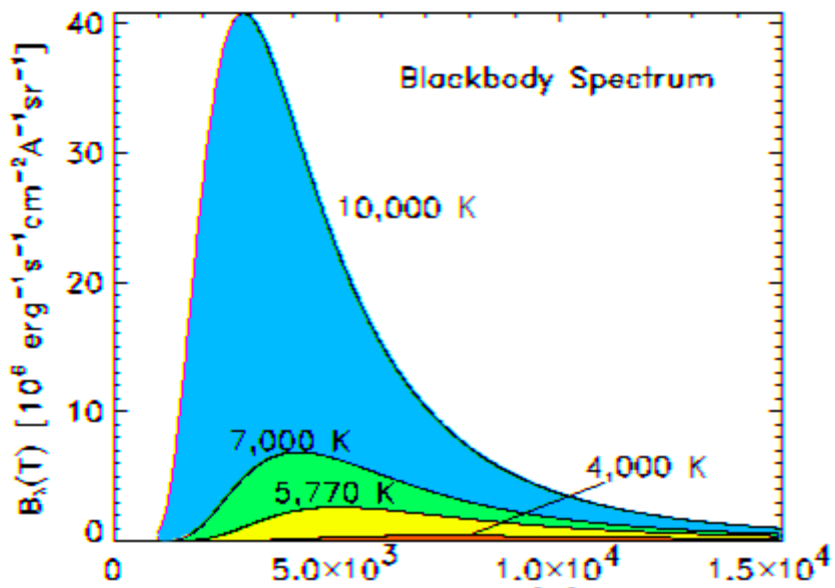
Brightness Temperature and Flux Density

Planck Function

You should all be familiar with the basics of black-body radiation. By simply observing how hot objects behave, 19th century scientists came up with the following empirical laws.

- the peak wavelength shifts proportional to temperature $\lambda_{max}T = 2.898\text{x}10^{-3}$ m K. (Wien Displacement Law)
- the intensity increases as the square of the frequency at low frequencies (Rayleigh-Jeans Law)
- the intensity decreases exponentially at high frequencies (Wien Law)
- the flux of radiation emitted by a blackbody increases as the fourth-power of the temperature (Stefan-Boltzmann Law)

A plot of this behavior for several temperatures as a function of wavelength is shown in the figure below.



The functional form that corresponds to these curves is called the ***Planck function***, which was derived by Max Planck by introducing the Planck constant $h = 6.63 \text{ x } 10^{-34}$ J s, and postulating that photon energies were quantized in units of $h\nu$. This was the beginning of the quantum theory, which was later extended to matter as well as radiation. This function has two forms, written below, which are related by $B_\lambda(T)\, d\lambda = B_\nu(T)\, d\nu$.

$$B_\lambda(T) = \frac{2hc^2/\lambda^5}{e^{hc/\lambda kT} - 1} \quad \text{(wavelength form)} \qquad (1)$$

$$B_\nu(T) = \frac{2h\nu^3/c^2}{e^{h\nu/kT} - 1} \quad \text{(frequency form)} \qquad (2)$$

- To derive the Wien displacement law, find the maximum of the function by setting $dB_\lambda(T) / d\lambda = 0$, to get $\lambda_{max}T = hc/5k = 2.898 \text{ x } 10^{-3}$ m
- To derive the Rayleigh-Jeans law, expand $e^{h\nu/kT}$ in $B_\nu(T)$ for $h\nu << kT$ to get $B_\nu(T) = 2kT\nu^2/c^2$
- To derive the Wien Law, expand $e^{h\nu/kT}$ in $B_\nu(T)$ for $h\nu >> kT$ to get $B_\nu(T) = (2h\nu^3/c^2)e^{-h\nu/kT}$
- To derive the Stefan-Boltzmann Law, integrate $B_\lambda(T)$ over all wavelengths--*hint: use the relation*

$$\int_0^{\inf} \frac{u^2 du}{e^u - 1} = \frac{\pi^4}{15}$$

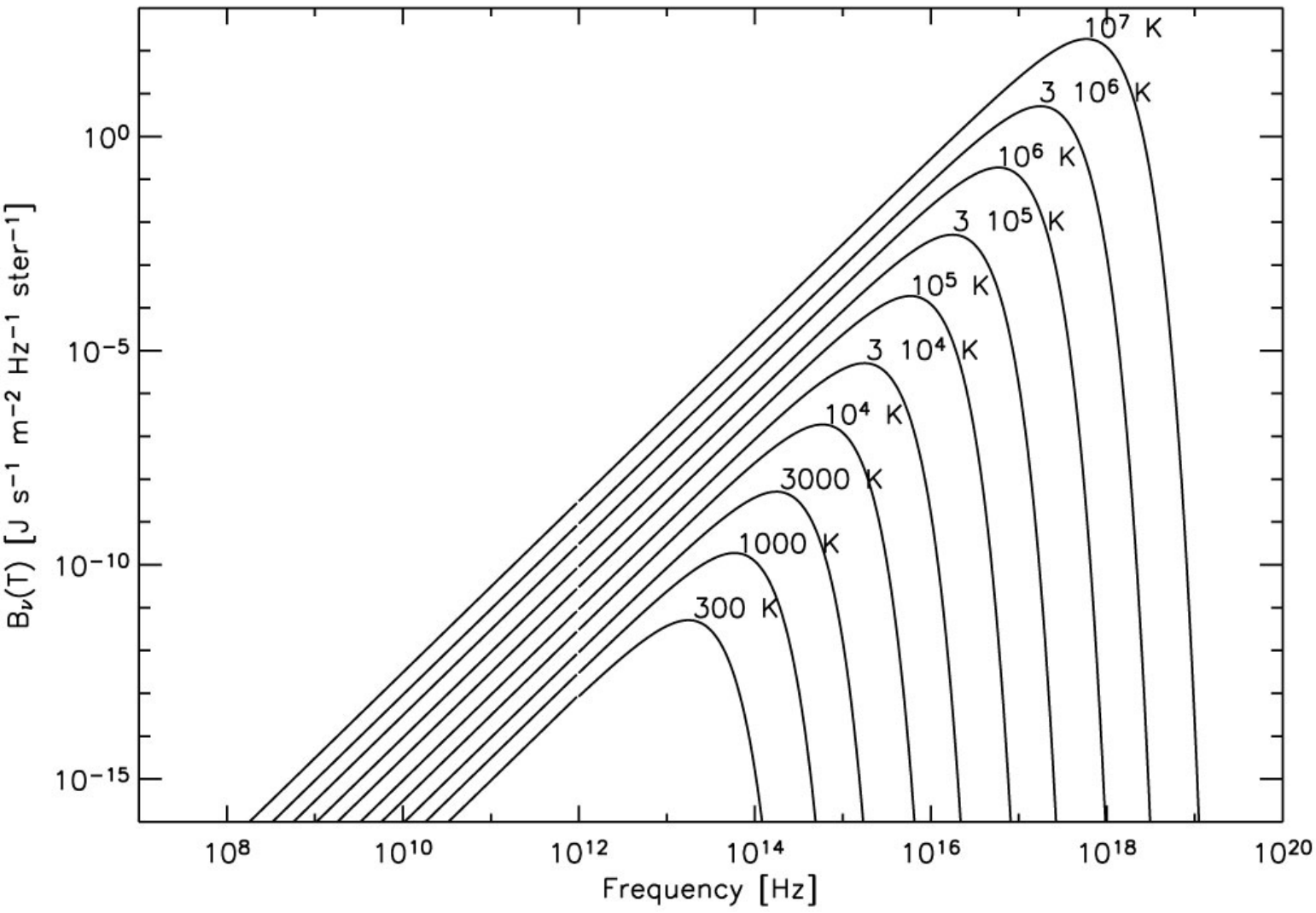
--to get the flux $F = \sigma T^4$, where $\sigma = 2\pi^5k^4/(15c^2h^3) = 5.669 \text{ x } 10^{-8}$ W/m^2/K^4 is the ***Stefan-Boltzmann constant***.

Rayleigh-Jeans Limit

The Rayleigh-Jeans limit is the Planck function in the limit of low-energy photons, $h\nu << kT$, which we argue is the relevant limit for any source that produces radio emission. To see this, take a relatively high radio frequency, say 100 GHz, and ask how cool the source must be in order that the above condition be violated (i.e. for $h\nu \sim kT$). We have

$$T = h\nu/k = (6.63 \text{ x } 10^{-34} \text{ J s})(1 \text{ x } 10^{11} \text{ s}^{-1})/1.38 \text{ x } 10^{-23} \text{ J/K} = 4.8 \text{ K !}$$

So even very cold sources at high frequencies still meet the Rayleigh-Jeans criterion. This turns out to be especially useful for radio astronomy, which we will discuss in a moment. But first, let's look at another plot of the Planck function, with axes suitable for a visual appreciation of the Rayleigh-Jeans limit.



By plotting $B_\nu(T)$ on a log-log plot, the part of the curve that obeys the Rayleigh-Jeans Law,

$$B_\nu(T) = 2kT\nu^2/c^2 \qquad (3),$$

is very obvious--it is the straight line portion with a slope of 2. Here you can see that changing the temperature over many orders of magnitude just increases the intensity linearly, and that it is valid over the entire range of radio frequencies all the way to THz (10^{12} Hz).

Surface Brightness (Intensity) and Flux Density

The ***monochromatic intensity*** $I(\nu)$ has units of J m^{-2} s^{-1} Hz^{-1} sr^{-1}, where sr = sterradians is the unit of ***solid angle***, $\Delta\Omega$. By comparing these units with the Planck function, we see that they are the same. The Planck function gives the monochromatic intensity of the blackbody that it represents. The ***intensity***, or ***surface brightness***, is then integrated over all frequencies:

$$I = \int I(\nu)d\nu \text{ (units: J m}^{-2}\text{ s}^{-1}\text{ sr}^{-1}\text{)}.$$

Integrate this again over angular area to get the ***flux*** F:

$$F = \int I\, d\Omega \quad \text{(units: J m}^{-2}\text{ s}^{-1}\text{, or W m}^{-2}\text{)}$$

which is just the power per unit area. In radio astronomy, we often discuss a related quantity called the ***flux density***, which is the monochromatic intensity (or the Planck function) integrated over solid angle:

$$S = \int I(\nu)\, d\Omega \quad \text{(units: W m}^{-2}\text{ Hz}^{-1}\text{)} \qquad (4)$$

In fact, the flux density is a fundamental quantity measured by radio telescopes, and is the basis for two different units:

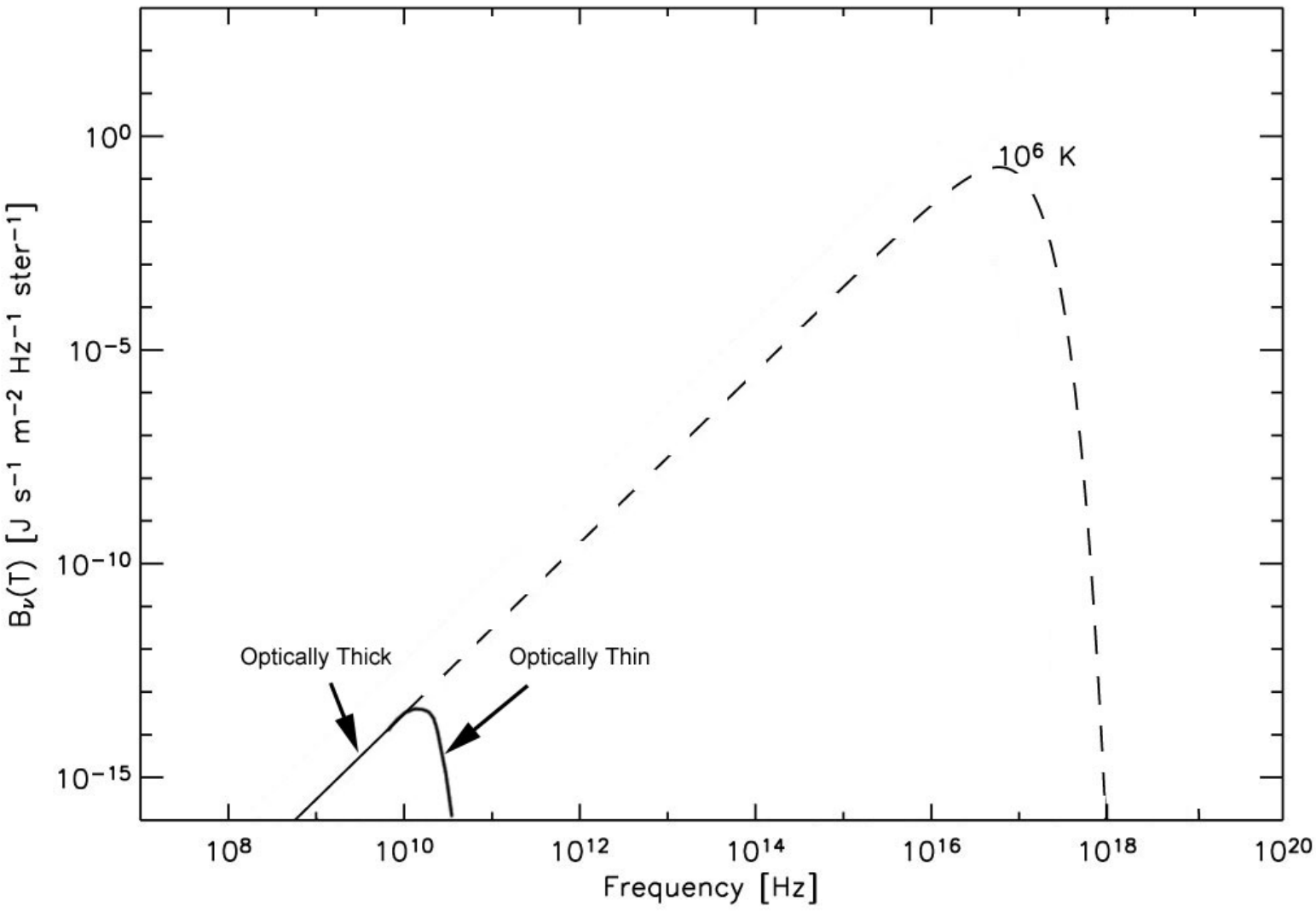
1 Jansky (Jy) = 10^{-26} W m^{-2} Hz^{-1}

1 Solar Flux Unit (sfu) = 10^{-22} W m^{-2} Hz^{-1} = 10000 Jy.

The quantity that a radio telescope measures is the flux density over some band $\Delta\nu$, so the strength of radio sources in the sky are often specified in Jy. Likewise, the strength of solar sources, especially solar radio bursts, are often specified in sfu.

Brightness Temperature

We are now ready to show a great conceptual simplification that the Rayleigh-Jeans limit gives to the discipline of radio astronomy. We have so far been talking about blackbodies, which are by definition optically thick and in thermal equilibrium. What if a source is not optically thick? In that case, its emission will appear weaker (lower intensity) than if it were optically thick. Whether or not a source is optically thick is a function of frequency. As it turns out, many radio-emitting plasmas are optically thick at low frequencies, but optically thin at high frequencies. In this case, the brightness follows the Planck function up to some frequency, then begins to fall away as it becomes more and more optically thin with frequency. Schematically, it looks something like this:



Radio spectrum for a 10^6 K plasma that is optically thick below about 10 GHz, and optically thin at higher frequencies. The brightness below 10 GHz corresponds to a million degree blackbody.

We will discuss optical depth in more detail in two weeks, when we discuss radiative transfer. For now, we just want to develop the idea of ***brightness temperature***.

In the Rayleigh-Jeans limit, a blackbody has a temperature given by the Rayleigh-Jeans Law, eq (3), i.e.

$$T = B_\nu(T)c^2/2k\nu^2$$

so as long as the plasma in the above figure is optically thick, we can use the brightness of the emission to determine the plasma temperature. But when it is optically thin, the brightness, or intensity, is less than the Planck function. Nevertheless, we can still talk about a brightness temperature, or the equivalent temperature that a blackbody would have in order to be as bright. The brightness temperature is the same as the true temperature only for an optically thick blackbody. We designate the brightness temperature as T_b. Using this notation, the flux density measured by a radio telescope becomes:

$$S = \int 2kT_b\nu^2/c^2 \, d\Omega = 2k\nu^2/c^2 \int T_b \, d\Omega \qquad (5)$$

where we have substituted B_ν for $I(\nu)$ in (4), and used (3). So the flux density measured by a radio telescope is just the brightness temperature integrated over the source, times some fundamental constants and frequency-squared.

So far, eq (5) pertains only to thermal emission, but we can extend it to all radio emission simply by considering non-thermal sources as having an ***effective temperature*** T_{eff}. For a single electron of energy E, its effective temperature is just its kinetic temperature $T_{eff} = E/k$.

To summarize, then, the brightness temperature is the equivalent temperature a black body would have in order to be as bright as the observed brightness. It is important to realize that this is a useful concept only for radiation that obeys the Rayleigh-Jeans Law.

One last point to make is the limit of the integral in eq (5). We earlier mentioned the resolution of a single dish antenna of diameter D, as $\theta \sim 1.22\ \lambda/D$. This is also the width of the field of view of the antenna--only a source in an area of the sky within this angular distance can be seen. The field of view

is also called the ***beam***. Let's look at some consequences of this.

- Unresolved (point) Source: If the antenna is pointing at a source with a very small size, the flux density that it measures is the brightness integrated over the entire source, i.e. the total flux density of the source. In this case, the integral (5) is to be taken over the source.
- Extended Source: If the antenna is pointing at an extended source, part of which falls outside the field of view, it will measure only the flux from the source within the beam. In this case, the integral (5) is to be taken over the beam size.

Physics 728 Radio Astronomy: Lecture #2 Prof. Dale E. Gary
NJIT

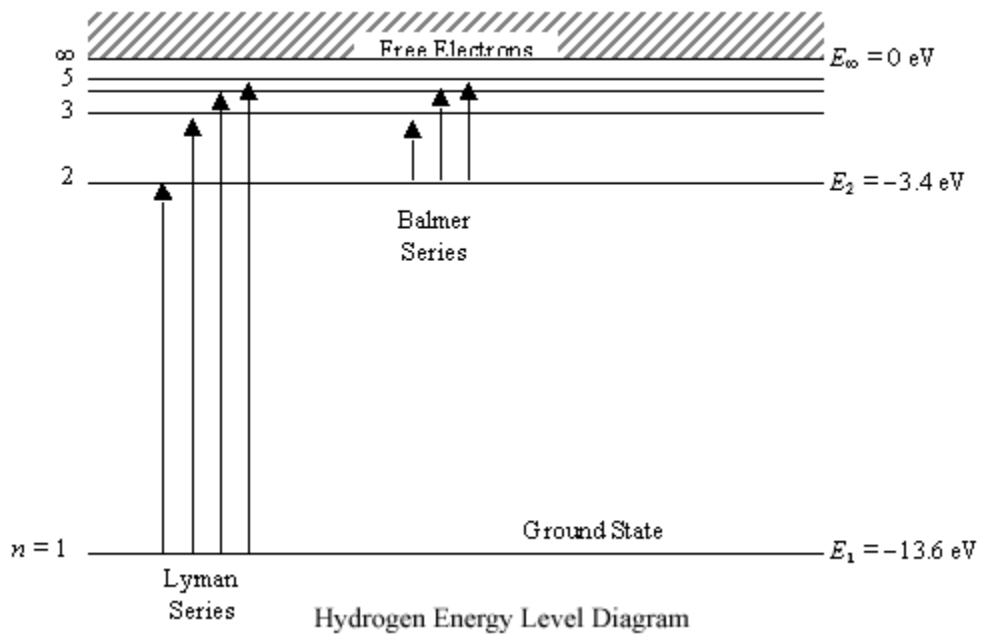
Radio Emission Mechanisms

Types of Radio Emission Mechanisms

- Atomic transitions
- Molecular transitions
- Free particle emission
  - Cerenkov emission
  - Bremsstrahlung (Coulomb collisions)
  - Magneto-bremsstrahlung (gyroemission)
  - Transition radiation
- Coherent free-particle emission (wave-particle interactions)
  - Plasma emission
  - Electron-cyclotron emission
- Wave emission (wave-wave interactions)

Atomic Transitions

You are familiar with atomic transitions and their role in optical and ultraviolet emission, but you may not be aware that they also are responsible for radio emission in an important regime where they are used to help us learn about interstellar space. We will not go into detail on this mechanism, which you should already know well, but recall the energy level diagram for hydrogen, shown below:



Hydrogen Energy Level Diagram

The difference between energy levels (energy states) in hydrogen corresponds to a wavelength

$$1/\lambda_{ab} = \nu_{ab}/c = (E_b - E_a)/ch = R_H[1/n_b^2 - 1/n_a^2]$$

where $R_H = (2\pi^2\mu e^4 k^2 Z^2)/ch^3 = 1.096776 \times 10^7\ \mathrm{m}^{-1}$ is the *Rydberg constant.* (More on the Rydberg Constant.) Here we use the reduced mass $\mu = (Am_p m_e)/(Am_p + m_e)$, and the value given is numerically correct only for hydrogen (atomic mass $A = 1$). We also use Z= atomic number = 1 for hydrogen, but I write it in terms of Z because it is valid for any atom with a single electron (e.g. He II). The transitions shown above correspond to absorption of energy by the atom, but reversing the arrows (transitions from higher to lower states) corresponds to emission. The energy differences corresponding to the transitions labeled Lyman Series correspond to UV emission. Those labeled Balmer Series are in the optical range (e.g. hydrogen alpha is the 3 to 2 transition). How would one get radio emission, say at 5 GHz?

Use n_a=110, n_a=109 to get

$$\nu_{ab} = R_H c[1/109^2 - 1/110^2] = 1.5 \times 10^{-6} R_H c = 5 \times 10^9\ \mathrm{Hz}.$$

This transition would be called the H109α transition, where α transitions are those for which $\Delta n = 1$. Transitions for $\Delta n = 2$ would be termed β transitions, and so on. So transitions in very high energy levels yield radio lines, called *Radio Recombination Lines*. Note that these transitions are so far from the atom that the same basic formula can work for higher atomic weight neutral atoms, e.g. He I, using the effective potential $eZ = 1$, so long as the appropriate reduced mass is used.

Radio recombination lines arise in *HII regions*, which are regions of the interstellar medium where the ordinarily neutral hydrogen gas is ionized by a nearby hot star, whose light is rich in UV photons. The UV photons ionize the atoms, and the electrons recombine and cascade downward toward the ground state. Some fraction (given by the Boltzmann and Saha equations) make the transition from 110 to 109, and emit a 5 GHz photon. Although the density of the interstellar medium is low (perhaps $10^4 \mathrm{cm}^{-3}$), the path length is long (~0.5 pc), so the emission can be observed as a weak radio line. Note that transitions around H40α are in the 100 GHz (millimeter) range, and around H600α are around 30 MHz range, so the entire radio

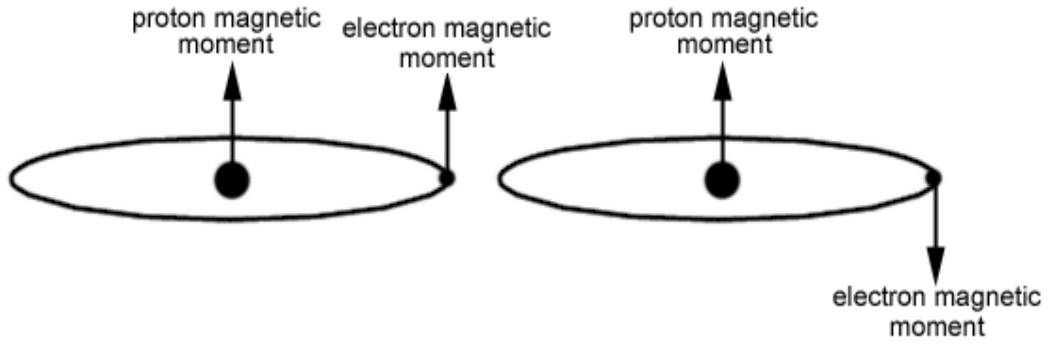
spectrum can provide recombination lines.

By studying these recombination lines, we learn the following about HII regions:

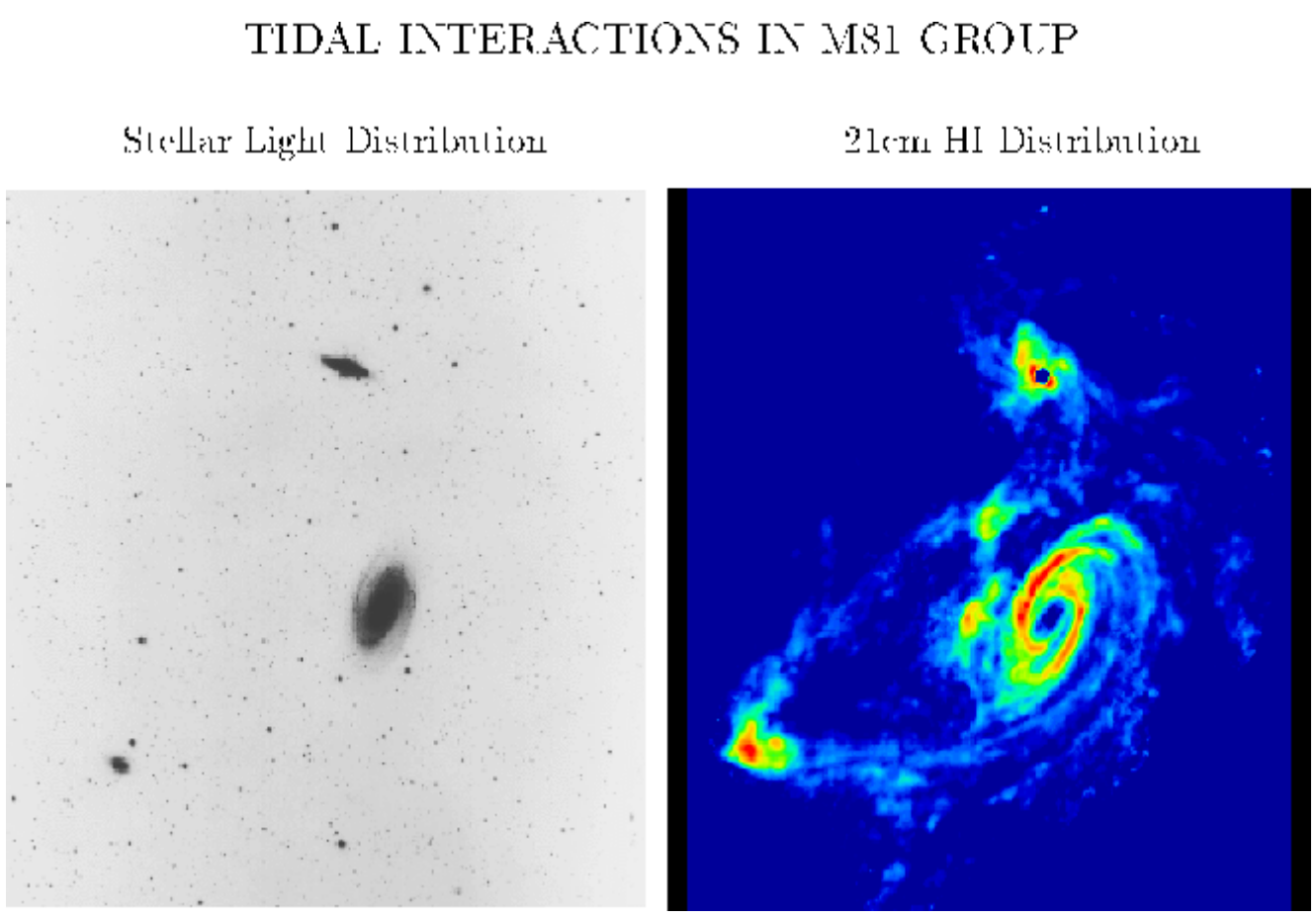
- temperature (from thermal width of lines and line ratios)
- density (from strength of lines and line ratios)
- composition (from He and C lines)
- velocities and motions (from doppler shifts)

Hyperfine Splitting and the 21 cm line

Another important transition is not a transition from one principle quantum state to another, but rather a spin flip of electrons in neutral hydrogen atoms. The ground state of hydrogen has slightly different energies for electrons with spin parallel and antiparallel to the proton, as shown below:



Which is the lowest energy state? Consider magnets, which want to be antiparallel, so the figure on the right represents the lower energy state. These magnetic moments are associated with the quantum property called spin, but the electron magnetic moment is opposite to the electron spin (due to the electron's negative charge). The lowest energy state is when the magnetic moments are anti-parallel, at which time the spins are parallel. Note that this spin flip is a magnetic dipole transition, and is highly forbidden by the quantum rules, but occurs spontaneously once every few million years for an isolated atom. Note that collisions in the interstellar medium occur far more often--about once every 400 years or so, and such collisions randomly cause spin flips. So there are atoms in about equal numbers in both energy states. Even though the spin flip is very rare for an individual atom, there is a truely staggering number of atoms along any line of sight, so the radio line produced by the spin flip is strong in all directions. The energy difference corresponds to a frequency of 1420 MHz, or a wavelength of 21 cm.



Neutral hydrogen traces interactions among galaxies in the M81 group.
(c) National Radio Astronomy Observatory / Associated Universities, Inc. / National Science Foundation

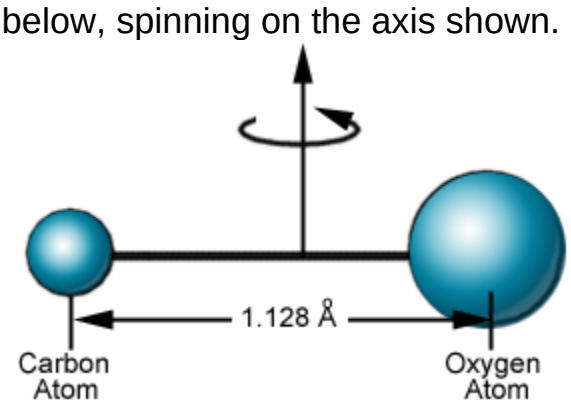
By studying the 21 cm line, we learn the following:

- velocity of neutral hydrogen clouds (from doppler shifts)
- rotation of our galaxy and other galaxies (from doppler shifts)
- distribution of neutral hydrogen in our galaxy and in other galaxies (from images)
- tidal interactions between galaxies (from images and doppler shifts)
- amount of gas in the interstellar medium (from strength of the lines)

Molecular Transitions

Consider a diatomic molecule such as CO, or O_2, or H_2. They can spin along different orthogonal axes, such as along the line joining the atoms, or perpendicular to this line. They can also vibrate. Under quantum mechanical rules, these rotational or vibrational modes are quantized, as you learned in

Statistical Mechanics. The general idea can be developed using classical physics, e.g. consider the molecule as a "bar-bell" as shown below, spinning on the axis shown.



The rotational kinetic energy is just $E = 1/2\ I\ \omega^2$, where $I = \mu r^2$ is the moment of inertia, and $\mu = mM / (m + M)$ is the reduced mass, as before. Thus, $E = 1/2\ \mu r^2 \omega^2$. But in quantum mechanics, the angular momentum $L = I\ \omega$ is quantized according to angular momentum quantum number J, $L = J\ h/2\pi$, so we find that

$$\omega^2 = (h/2\pi)^2\ \boldsymbol{J}.\boldsymbol{J} / \mu^2 r^4 = (h/2\pi)^2\ J\ (J + 1) / \mu^2 r^4$$

and finally

$$E = (h/2\pi)^2\ J\ (J + 1) / 2\mu r^2\ .$$

It is straightforward to show that the rotational transition J = 1 to 0 of the CO molecule results in a spectral line at frequency 115.2712 GHz, while the J = 2 to 1 transition gives a line at 230.5424 GHz, given that the intermolecular bond distance is 1.128 A. These are but two of many molecular lines that can be detected from clouds in interstellar space and in other galaxies to provide information on the distribution of cold, dense matter. Note that CO and other molecules are easily dissociated by radiation and collisions by hot gas atoms, so its presence implies very cold conditions. Typically molecular clouds have temperatures of order 100 K, and sometimes as cold as 10-30 K. They are associated with dust clouds, and it is the dust that protects the fragile molecules and helps to cool the cloud by radiation of IR.

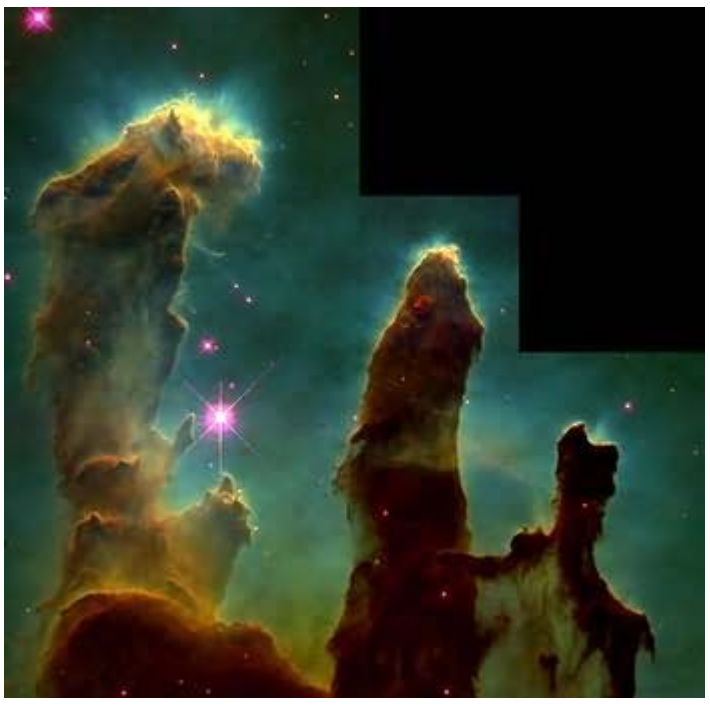


Image of M16 (Eagle Nebula) molecular cloud and star-forming region. From HST web site.

Free-Particle Emission

The emission types that we have discussed up to now are transitions between bound states, for which quantum mechanical rules are of the first importance. However, for astrophysical plasmas in a high-energy environment, such as in stars, or the tenuous atmosphere around stars, it is free particles that dominate the radio emission mechanisms. (Note that bound state transitions can still be important at UV and soft X-ray wavelengths in such high-energy plasmas.) This includes all radio emission from the Sun and in the interplanetary medium, including the Earth's magnetosphere and ionosphere.

Recall that a charge at rest or in uniform motion (in the non-relativistic limit) has an electrostatic electric field (cgs units):

$$E = q/r^2\ \boldsymbol{r},$$

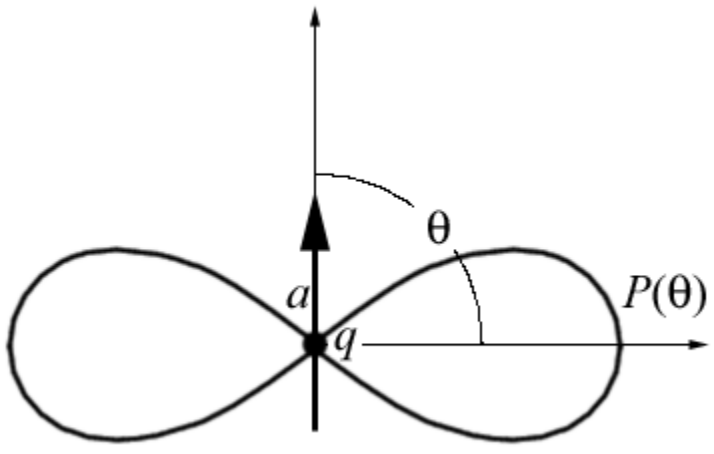
which is a radial field falling off as $1/r^2$. The power radiated is the square of the electric field, so the power from the radial part of an electrostatic field vanishes quickly, as $1/r^4$. However, when a transverse "kink" appears in the field, that transverse part falls off only as $1/r$, so the power falls off as $1/r^2$. So any phenomenon that produces a kink in the field will result in radiation. The typical way to produce a kink in the electrostatic field of a charge is to accelerate the charge. Accelerating charge aplet.

The Larmor Formula (see Jackson, 2nd edition, section 14.2) expresses the radiated power from an

accelerated charge per unit solid angle as:

$$dP/d\Omega = q^2a^2 \sin^2 \theta / 4\pi c^2 \qquad (1)$$

where q is the charge, a is the acceleration, and θ is the angle from the direction of the acceleration. Note that the "radiation pattern" is a dipole pattern, which looks like this:



and the pattern arises naturally from the geometry. A dipole antenna (where charges travel in simple harmonic motion up and down the antenna) has exactly this radiation pattern. In the case of relativistic motion of charged particles, the pattern is altered, as we will discuss shortly (see Jackson, 2nd edition, section14.3). To get the total power radiated, simply integrate (1) over all solid angle to get $P = 2/3\ q^2a^2 / c^2$.

What kind of charged particles are good at radiating? Note that the force required to produce an acceleration is $F = ma$, so the acceleration is $a = F/m$, and $dP/d\Omega \sim 1/m^2$. So electrons are better at radiating than protons by a factor $(m_p/m_e)^2 = 4 \times 10^6$! *For radio emission, we can assume that the only particles important for the emission are electrons*.

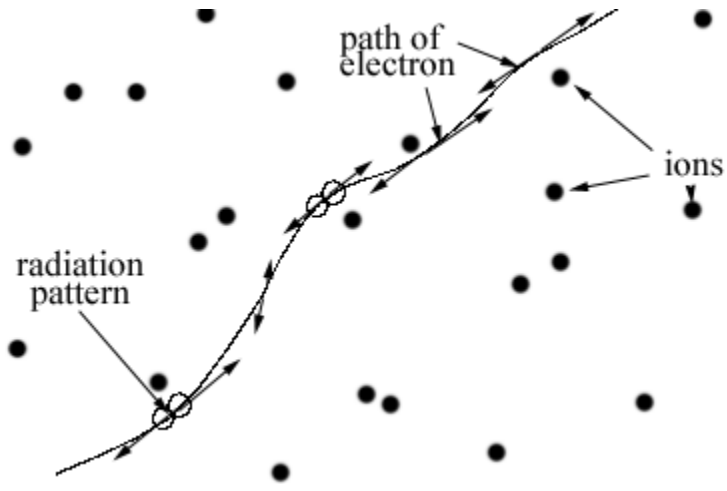
Cerenkov Emission

For completeness, we briefly mention that the velocity of a free charge moving in a medium can actually exceed the speed of light in the medium. This causes the light to radiate in a cone, exactly like sound waves from an aircraft that exceeds the speed of sound. This effect gives rise to Cerenkov emission, which is responsible for the blue glow given off by nuclear reactors that are surrounded by water, and for emission by cosmic rays in the Earth's atmosphere. This is one of the few mechanisms that can result in radiation without the need for the particle to accelerate. Cerenkov emission plays no role in radiation in plasmas, however, because the wave speed of light in an ionized medium is greater than c!

Bremsstrahlung

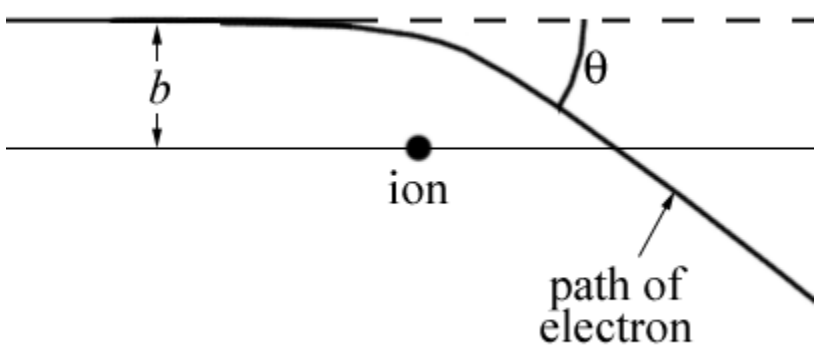
This is an important type of radiation in many astrophysical plasmas. The word is german for "braking radiation," as is due to accelerations caused by collisions between electrons and ions (Coulomb collisions). Recall that the Coulomb force is $\boldsymbol{F} = q\boldsymbol{E}$, so the force on an electron in the electrostatic field of a proton is an attractive force $\boldsymbol{F} = -e^2/r^2\boldsymbol{r}$. In head-on collisions, the electron can actually bounce off the proton, and lose most of its energy, or even reverse its direction. In this case, the energy of the photons emitted is typically in the X-ray range. The hard and soft X-rays from astrophysical plasmas, including the Sun, are a mixture of bound-bound atomic transitions (soft X-rays), nuclear line emission (hard X-rays/gamma rays), and electron bremsstrahlung. In the hard X-ray range (10-1000 MeV), the emission is purely electron bremsstrahlung.

Radio emission, in contrast, is due to relatively gentle, distant collisions of electrons with ions. Picture the plasma as a soup of electrons and ions, with the electrons zipping around at great speed through the relatively immobile ions. The path of an electron is continually wiggling due to slight deviations as it passes each ion, and each wiggle produces radiation in a broad spectrum (related to the energy lost or gained in each collision). Note that in the case of local thermodynamic equilibrium (LTE), the electron is also absorbing radiation in approximate equality with its energy loss. We will only discuss emission today, but will discuss absorption next time.



An electron moving through relatively immobile ions experiences many small accelerations, each one producing a radiation pattern along its direction of motion (perpendicular to its acceleration).

Let us look a bit more closely at a single collision. As shown in the figure below, the electron experiences a deviation of its straight line path by an angle θ, which depends on the speed of the electron and the distance of the encounter, which is called the impact parameter, b. Bremsstrahlung has the peculiarity that the amount of deviation (and hence the power radiated) is actually greater when the electron is moving more slowly, and smaller when v is high. We will see this when we get to the final form of the emissivity.



One can calculate the emission from this process, but it is rather involved (see Rybicki & Lightman, "Radiative Processes in Astrophysics," 1979; or the web discussion here.). The outline of the procedure is to use the small-angle-approximation, for which the electron deflection is negligible, and consider the motion along a straight line, where the separation between charges is $r = (b^2 + v^2t^2)^{1/2}$. One uses the dipole approximation to determine the net acceleration over the path, and thus the emission from a single electron for a single collision

$$dW(b)/d\omega = 8Z^2e^6/(3\pi c^3 m_e{}^2 v^2 b^2), \text{ for } b << v/\omega. \qquad (2)$$

Note that this depends on the impact parameter b, and we limit the solution to $b << v/\omega$ because collisions at a given b lead to emission only at $\omega < v/b$. To extend this to a volume flux of electrons, all with speed v, note that the rate of collisions on any one ion is just $n_e v$, and these electrons will fill the area with element or area $2\pi b\, db$. Then the total emission (energy) per unit time per unit volume per unit frequency range is

$$\frac{dW}{d\omega dV dt} = n_e n_i 2\pi v \int_{b_{min}}^{b_{max}} \frac{dW(b)}{d\omega} b\, db \qquad (3)$$

where the limits of the integral over impact parameter depend on the case to be evaluated. For any emission, there is a reasonable upper limit, where $b_{max} = v/\omega$. For radio emission (where the collision energy should be relatively low), we can take a lower limit as for a 90º deflection, which occurs for $b_{min} = 4Ze^2 / \pi m_e v^2$. For higher-energy emission (i.e. X-rays, with head-on collisions), a smaller impact parameter should be taken. When (2) is inserted into (3), we obtain

$$\frac{dW}{d\omega dV dt} = 16e^6 n_e n_i Z^2/(3c^3 m_e{}^2 v) \int_{b_{min}}^{b_{max}} \frac{db}{b}$$

$$= 16e^6 n_e n_i Z^2/(3c^3 m_e{}^2 v) \ln (b_{max}/b_{min}).$$

It is conventional to write this as

$$dW / d\omega dV dt = 16\pi e^6 n_e n_i Z^2/(3^{3/2} c^3 m_e{}^2 v)\, G_{ff}(T,\omega), \qquad (4)$$

where $G_{ff}(T,\omega)$ is the Gaunt factor:

$$G_{ff}(T,\omega) = (3^{1/2}/\pi) \ln [\pi (kT)^{3/2} / 2^{1/2} Z e^2 m_e{}^{1/2}\, \omega]. \text{ (c.f. Dulk 1985).}$$

where we have made the assumption that the electron velocities are the thermal velocities,

$$v = (2kT/m)^{1/2}.$$

Gaunt factors for many other cases can be calculated exactly (see Figure 5.2 of Rybicki and Lightman 1979).

Thermal Bremsstrahlung

The last step is to integrate the emissivity over a population of electrons, and in most cases for radio emission this is done explicitly for a thermal distribution (that is, the velocities are distributed according to a Maxwellian distribution). In other words, integrate (4) over all velocities. Let us use the symbol η_ν in place of the unwieldy $dW / d\nu dV dt$, which we call the emissivity. Then after integrating (and converting from ω to ν) we get:

$$\eta_\nu = (2^6\pi e^6/3m_e c^3)(2\pi/3m_e kT)^{1/2}\, n_e n_i Z^2\, G_{ff}(T,\nu).$$

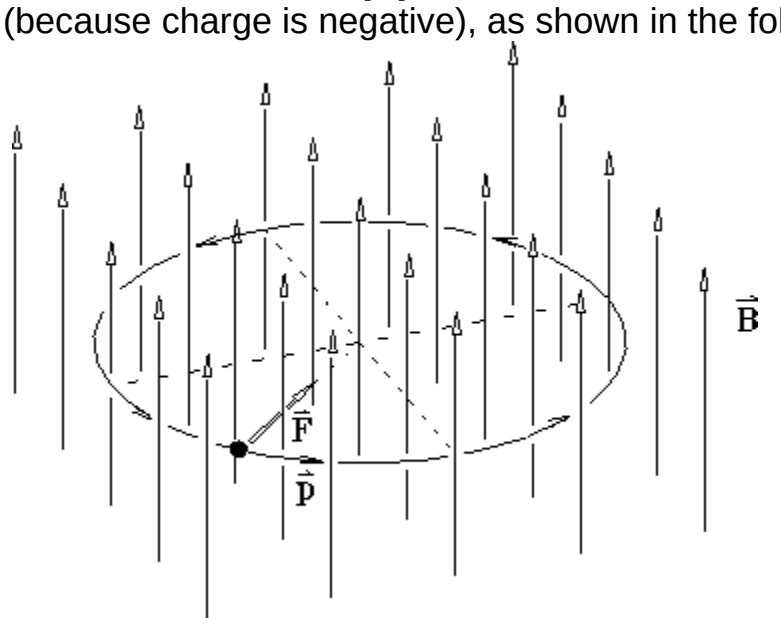
Note that for the solar corona, $n_e n_i Z^2 \sim n_e^2$, so the emissivity depends on the square of the electron density, and is inversely proportional to temperature (emissivity goes down for higher temperatures!). Also note that there is almost no frequency dependence--the only dependence is the weak (logarithmic) dependence in the Gaunt factor. We will see next time, when we introduce the absorption coefficient, that this is only true when the emission is optically thin.

Gyroemission

There is another way to accelerate free particles, by considering the effect of the magnetic field and looking at the magnetic part of the Lorentz force $\boldsymbol{F} = q\boldsymbol{E} + (q/c)\,\boldsymbol{v} \times \boldsymbol{B}$. For a plasma, there is usually no macroscopic electric field (except perhaps in current sheets), but often there is a non-negligible magnetic field (in which case the plasma is termed a ***magnetized plasma***). In this case, an electron of speed v will be accelerated perpendicular to both v and B, with magnitude

$$a = ev_{\rm perp}B/m_e c$$

in the *right-hand-rule sense* (because charge is negative), as shown in the following figure:



Put the thumb of your right hand in direction of B, and an electron (negatively charged) will gyrate in the direction of your fingers. A proton gyrates in the opposite direction.

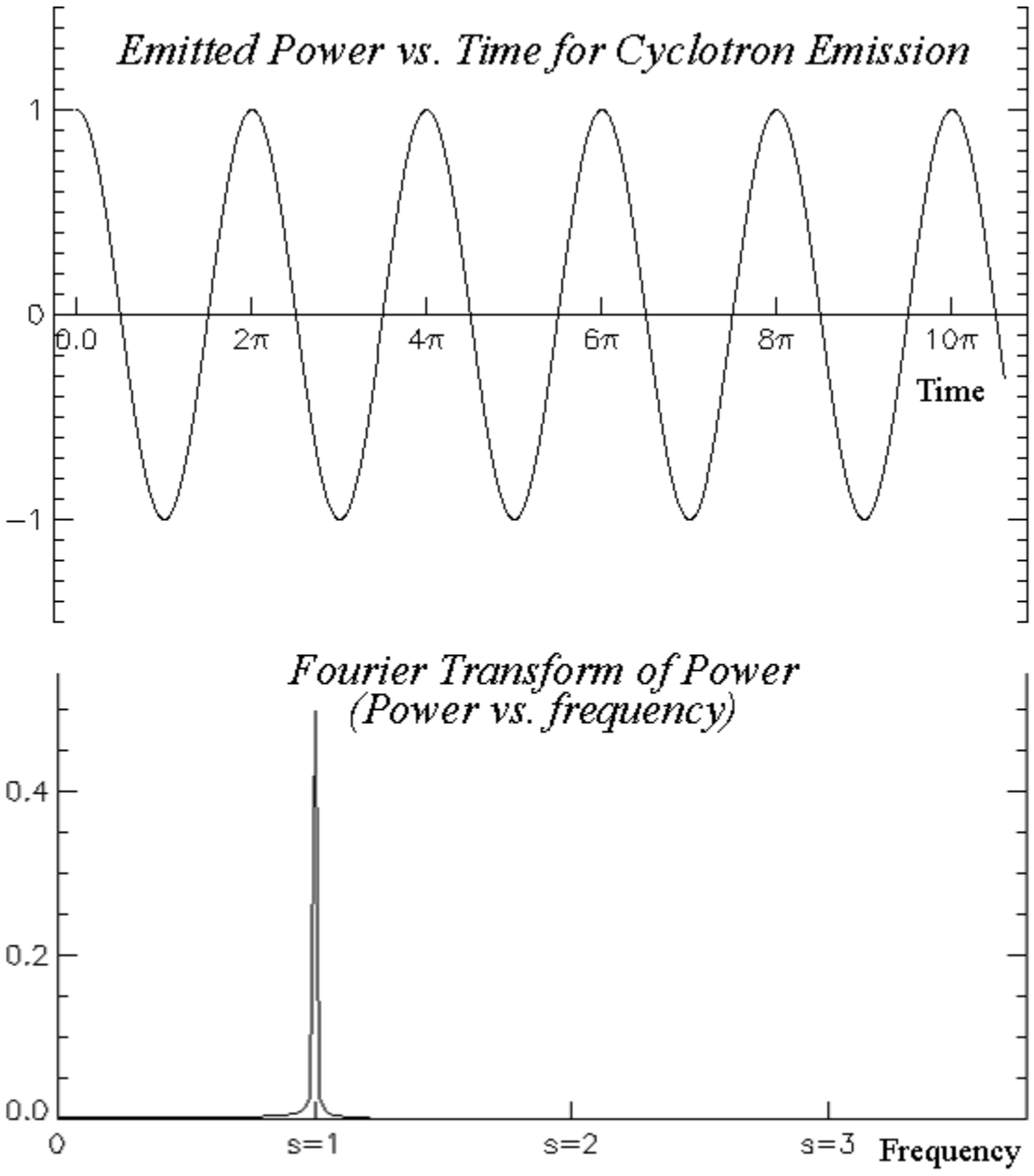
See this Java Applet for gyromotion.

Again, one can calculate the emission from this process, and again it is beyond the scope of this course. However, from the Larmor formula you can see that the power radiated will be

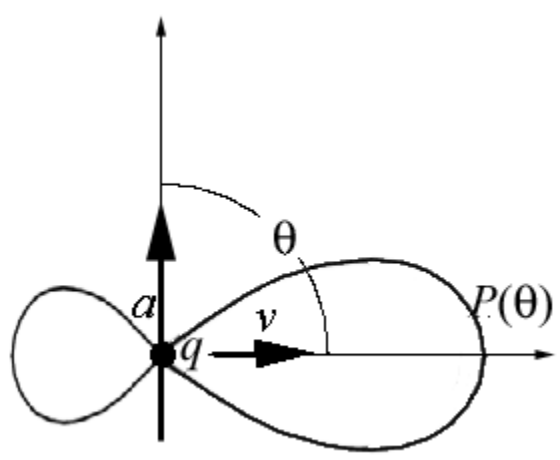
$$P = (2e^2/3c^3)(e^2B^2/m_e^2c^2)\, v^2_{\rm perp},$$

for the non-relativistic case. For the relativistic case, this formula should be multiplied by γ^2. Some of the properties of this emission, which will be very important for understanding solar emission from both active regions and flares, are:

- ***Cyclotron emission:*** In the non-relativistic case (low electron velocity, i.e. low temperature), the electrons gyrate at a fixed frequency, independent of their speed, called the gyrofrequency $\omega_B = eB/m_e c$, which depends *only* on magnetic field strength. The emission from a single electron, when viewed from afar, has a radiated power that varies sinusoidally, which gives rise to a cyclotron line at the gyrofrequency, $f_B = eB/2\pi m_e c = 2.8 \times 10^6\, B$ Hz (B in gauss).



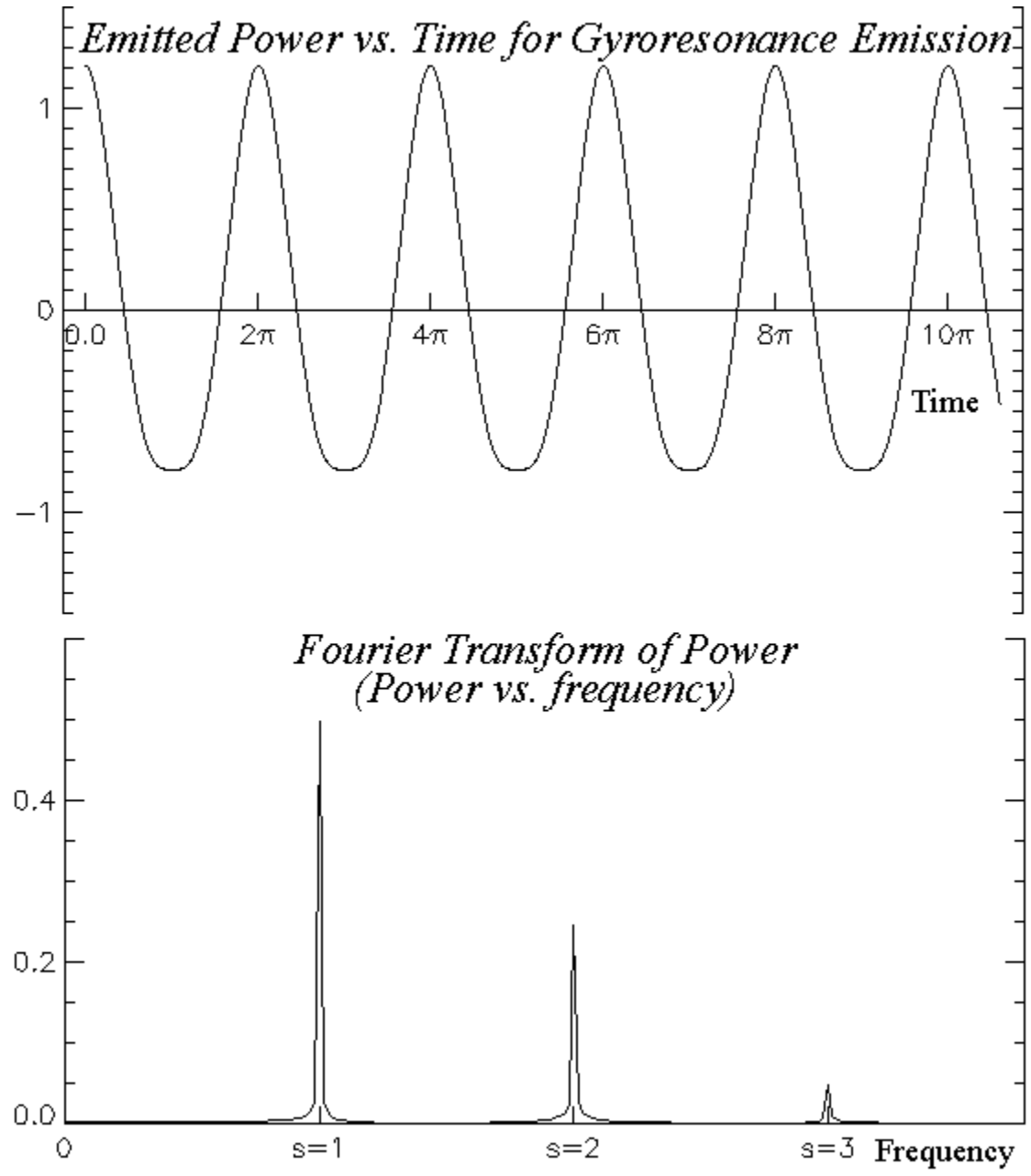
- ***Gyroresonance emission:*** At slightly higher electron velocity ($T \sim 10^5$-10^6 K), a relativistic effect comes in that changes the sinusoidal dipole pattern to a slightly asymmetric shape,



so that the radiated power peaks more strongly. This gives rise to harmonics of the cyclotron line (so-called gyroresonance lines),

$$f = sf_B = seB/2\pi m_e c = 2.8 \times 10^6 \, sB \text{ Hz} \quad (s = 1, 2, 3, ...).$$

This type of emission is responsible for bright coronal emission from solar active regions.



- ***Gyrosynchrotron emission:*** At mildly relativistic speeds (electron energies 100-300 keV), the effect becomes stronger, and the lines go up to harmonics 10-100. The lines also become broader (thermal broadening), so that they blend together into a continuum emission. This form of the emission is called gyrosynchrotron emission, and is the type of emission responsible for most radio emission from solar and stellar flares.
- ***Synchrotron emission:*** At highly relativistic speeds, the forward lobe becomes a narrow beam of width $1/\gamma^3$, and the emission comes in narrow pulses at the cyclotron frequency, beamed along the direction of motion. The pulses contain many many harmonics. This kind of emission is important in extreme energy environments such as black holes, neutron stars, and some extragalactic sources (generally associated with black holes).

These characteristics are for a single electron, or by extension a mono-energetic population of electrons. To determine the expected emissivity from a plasma, one must integrate the contribution of emission over a particular velocity distribution of electrons. There is a lot of effort in trying to determine an appropriate distribution to use for a given situation, and this is an area of active research. Typically there are two types of population that we consider: a thermal population (in which case the emissivity is expressed in terms of plasma parameters T, n_e, and magnetic field parameters B andθ), or a powerlaw distribution in energy (in which case the emissivity is expressed in terms of energy distribution parameters N, δ, and perhaps E_o, and magnetic field parameters B andθ).

In some cases, such as thermal gyroresonance emission, or synchrotron emission from an isotropic powerlaw distribution, the emissivity can be written down analytically. A good overview is given by Dulk (1985). Here is the emissivity in these two cases:

Gyroresonance emissivity for a thermal distribution of electrons:

$$\eta_\nu(s,\theta) = \pi^2 m_e/4c\ [\mu_\sigma d(\omega\mu_\sigma)/d\omega]^{-1}\beta^2\nu\nu_p{}^2(s^2/s!)\ [s^2\beta^2\sin^2\theta\ /2]^{s-1}[\beta\cos\theta]^{-1}$$
$$\text{x}\ \exp[-(1 - s\nu_B/\nu)^2/2\ \mu_\sigma{}^2\beta^2\cos^2\theta)](1 - \sigma|\cos\theta|)^2$$

Synchrotron emissivity for a powerlaw distribution of electrons:

$$\eta_\nu/BN = 1/2\ (\delta - 1)\ E_o{}^{\delta - 1}\ g(\delta)\ 3^{1/2}e^3/8\pi m_e c^2\ \sin\theta\ [(2m_e{}^2c^4/3\ \sin\theta)\ \nu/\nu_B]^{-(\delta - 1)/2}$$

Transition Radiation

To round-out our discussion of radio emission mechanisms from free particles, it is worthwhile to point out a relatively new mechanism proposed for astrophysical plasmas, although it is a well-known mechanism in devices. This mechanism shares one characteristic with Cerenkov emission, in the sense that no acceleration is required to produce the emission. Instead, the transverse component of

the electric field is produced by an inhomogeneity in the refractive index of the medium. To understand how the mechanism works, recall that the electrostatic field of a charged particle radiates outward with the group speed of light in the medium. Slight irregularities in these radiated field lines will occur when the group speed is not uniform. The mechanism was originally considered for the case of a density transition, hence the name *Transition Radiation*, but for astrophysical plasmas, where such density transitions would be smoothed out by motions of the hot particles, it is more appropriate to consider the case of density fluctuations.

In the case of such density fluctuations, perhaps due to wave turbulence in the plasma, it was first shown that the intensity of the radiation is very small and perhaps undetectable, for a reasonable level of turbulence. But recently Platinov and Fleishman have considered the case of emission near the plasma frequency. They found that there is a large enhancement of the emission due to the plasma frequency resonance, and so they gave emission in this regime the name *Resonant Transition Radiation*. We might expect an effect near the plasma frequency because the group speed of electromagnetic waves in a plasma approaches zero as the frequency of the waves approaches the plasma frequency.

Coherent Free-Particle Emission

The discussion so far as been about incoherent emission, in which electrons act separately to produce their emission. In this case, n particles produce n times the emission of one particle, and the photons emitted have no coherence--that is, no phase relationship. In this case the brightness temperature is related to the effective energy of the individual particles--the actual temperature in the case of a thermal distribution of particles, or the kinetic temperature for a non-thermal distribution.

It is also possible for electrons to act together to produce electromagnetic radiation, in which case the electrons all undergo acceleration in phase, acting together to produce photons that are in phase. In this case, the brightness temperature can far exceed the actual temperature or kinetic temperature of the electrons. In the case of the Sun, coherent bursts can reach brightness temperatures of 10^{15} K!

Formally, absorption of radiation is expressed as an imaginary part of the wave frequency, i.e. a purely oscillating wave has the form

$$E = E_o\, e^{i(kx-\omega t)}$$

while a damped wave has the form

$$E = E_o\, e^{i[kx-(\omega+i\gamma)t]} = E_o\, e^{i(kx-\omega t)}\, e^{-\gamma t}.$$

where γ is the damping rate. (Note that this can be converted to an equivalent absorption coefficient, κ, or emissivity loss per unit distance, by multiplying by the wave speed, $\gamma = \kappa\, \omega/k$.)

Under some circumstances, the damping rate can be negative, so that we have ***negative damping*** (or ***negative absorption***), i.e. wave growth. In this case, γ is called the ***growth rate***. Typically, different coherent mechanisms will produce this kind of linear growth over 3 or 4 e-folding times, and then the linear growth will saturate due to some non-linearity, such as running out of energy available for growth.

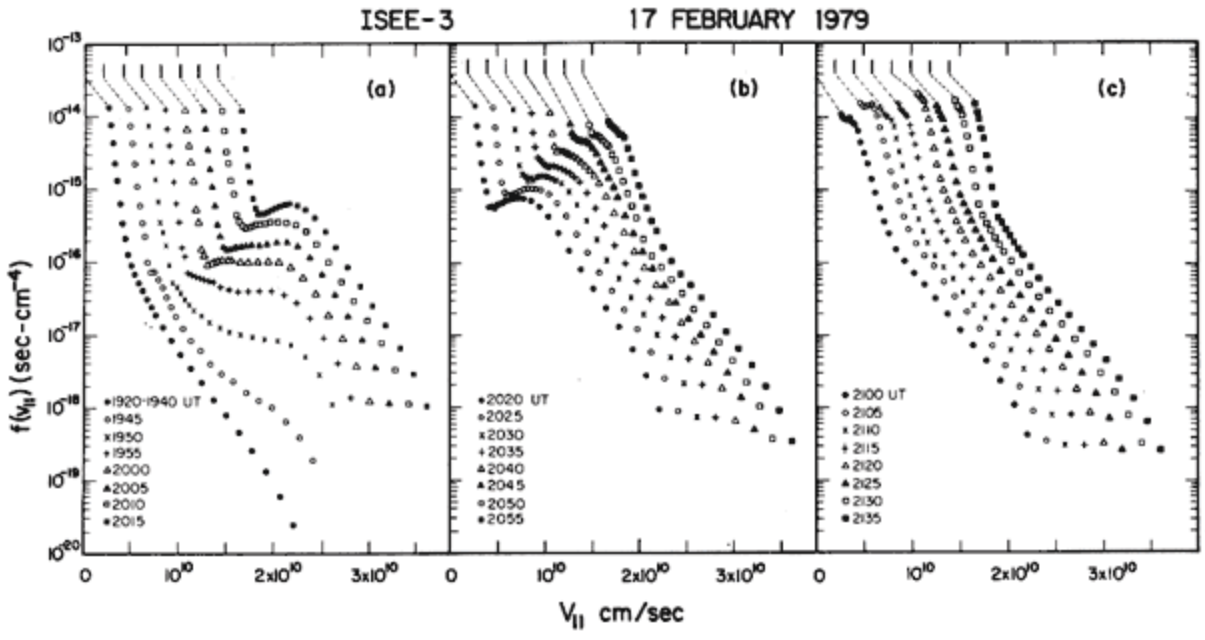
Plasma Emission

An ionized gas, i.e. a plasma, can support a wide range of wave modes. One mode, called a Langmuir wave, is so ubiquitous that it is called by the generic name plasma wave, or plasma oscillation, and the associated electromagnetic emission that it can produce is called plasma emission. The plasma frequency has a particularly simple form that depends only on the electron density:

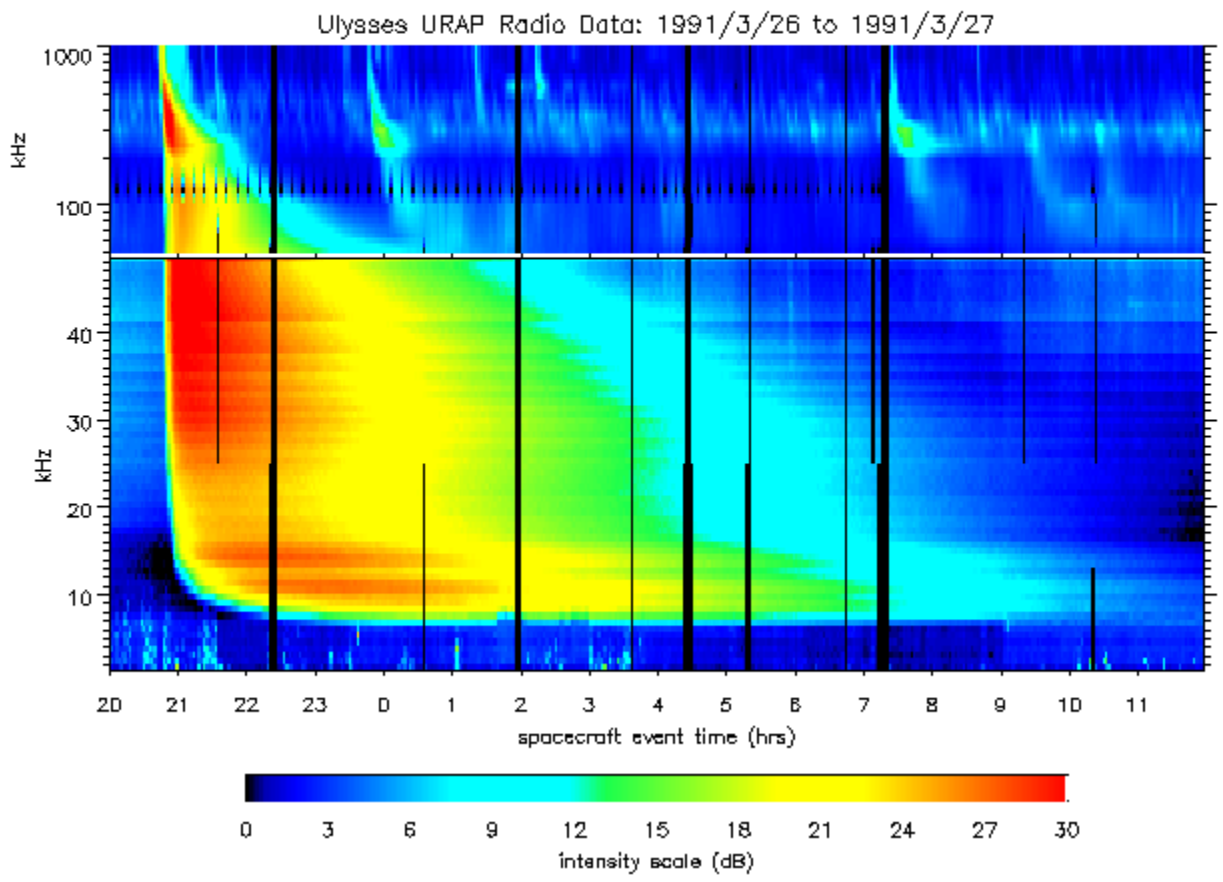
$$\nu_p = \omega_p/2\pi = (1/2\pi)[4\pi n_e e^2/m_e]^{1/2} \sim 9 \times 10^3\, n_e^{1/2},$$

where the units are cgs units, and so the electron density is expressed in units of cm^{-3}. This is the natural frequency of oscillation of a plasma, so any disturbance the provides energy to the plasma is likely to generate Langmuir waves at this frequency. Note that Langmuir waves themselves are longitudinal electrostatic waves, and are not observable from outside the plasma. However, under certain circumstances they can excite transverse (electromagnetic) waves at the same frequency that can escape.

As an important example, imagine a distribution of fast electrons in the so-called bump-on-tail distribution, as shown schematically below:



Here the electrons are said to form a beam, where the electrons stream along the magnetic field. The positive slope on the energy distribution is a source of free energy, and plasma waves can grow at the expense of the electron energies. These plasma waves can then scatter off ions (or couple to low-frequency ion-acoustic waves), to generate radio waves at the plasma frequency: $\omega_l + \omega_i = \omega_p$. In addition, two plasma waves going in nearly opposite directions can interact to produce radio waves at twice the plasma frequency: $\omega_l + \omega_l = 2\omega_p$. This is called "2nd-harmonic" emission, and is an example of the wave-wave interactions mentioned at the beginning of this lecture. Here is an example of radio emission produced by an electron beam traveling outward in the solar corona:



As the electrons stream outward, the electron density in the corona falls greatly, and the corresponding radio frequency of the emission goes to lower frequency. If we know the speed of the electron beam, we can deduce the density in the corona and interplanetary space. Conversely, if we know the radial density profile we can deduce the speed of the electrons. It is very important to understand that the intensity of the emission is only a function of the number of e-folding times over which growth occurs, and is not a diagnostic of the electron energies or other parameters of the source. This is a general property of coherent emission--intensities do not provide a useful diagnostic (at least not that we know how to interpret).

Electron-Cyclotron Maser (ECM) Emission

Like plasma emission, this mechanism uses a source of free-energy in the particle distribution to provide energy for wave growth, but this time the waves that develop may already be in the form of electromagnetic radiation. In this case, the source of free energy is not in the energy dependence of the particle distribution, but rather in the angular dependence, or pitch-angle distribution. The relevant geometry is downward-going electrons in a converging magnetic field. In this situation, electrons with a sufficiently large pitch angle will mirror, or reflect, but electrons with a smaller pitch angle can escape. The pitch angles smaller than the limit between mirrored and escaping particles are said to be in the loss-cone. The loss of particles in this cone set up a pitch-angle anisotropy, and under certain conditions the same electron-cyclotron (gyroresonance) waves that we met earlier, with frequency

$$\omega - kv_{\|} = s\Omega_B$$

see a distribution with a positive slope in their inertial frame. In the equation above, the gyrofrequency

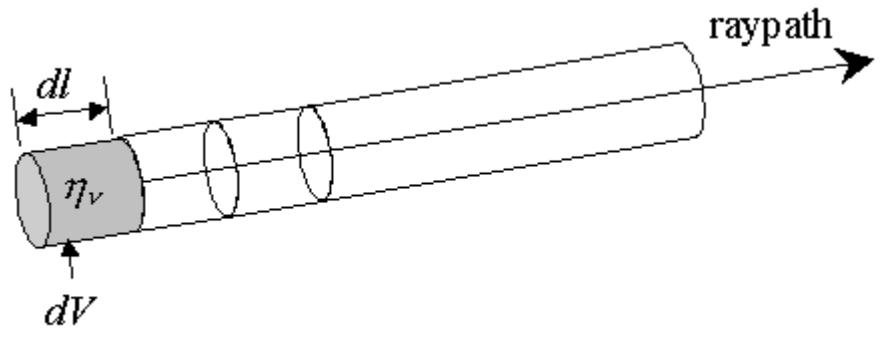
is $\Omega_B = eB/m_e c$, as before, but now we explicitly include a term $kv_{||}$, which represents the doppler shift of waves of wavenumber k seen by particles moving at speed $v_{||}$. Note that there is no positive slope in energy in the "rest" frame of the distribution, but there is for the wave frame. The result is that huge amounts of energy are available to go into production of radio waves near harmonics of the gyrofrequency. ECM emission is thought to be responsible for bursts seen in the outer atmosphere of Earth, called kilometric radiation, and in Jupiter's magnetosphere. It is also responsible for extremely bright, narrowband bursts from the Sun and stars, called spike bursts.

Physics 728 Radio Astronomy: Lecture #3 Prof. Dale E. Gary
NJIT

Radiative Transfer

Propagation of Radiation

We have discussed the generation of radiation by considering the ***volume emissivity***, η_ν, which is the energy per unit time per unit volume per unit frequency bandwidth per sterradian. Note that these units are the same as intensity, except it is per unit volume instead of per unit area. We will now discuss the total emission along a ***raypath*** through the plasma (loosely the same as the line of sight).



Along the raypath, more photons are added from each volume element, but also photons generated in one volume element can be absorbed in the same or another volume element. The increment in intensity, dI, in a volume element can be written as

$$dI = \underset{\text{source term}}{\eta_\nu dl} - \underset{\text{sink term}}{\kappa_\nu I dl}, \qquad (1)$$

where we introduce the ***absorption coefficient***, or ***opacity***, κ_ν, which is the fraction of intensity I absorbed per unit length.

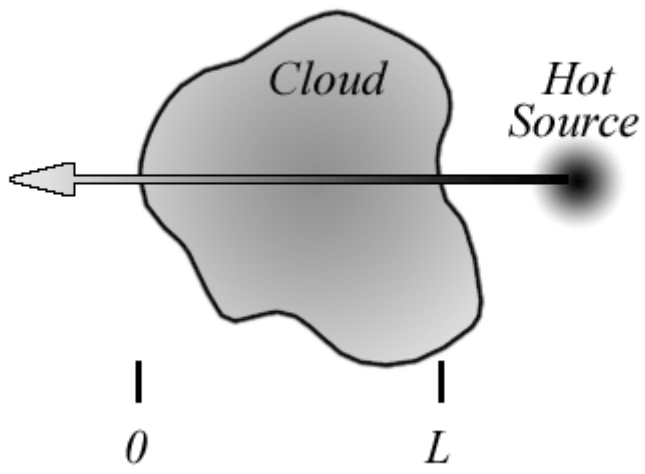
In thermodynamic equilibrium, Kirchoff's law states that emission and absorption occur at the same rate (the principle of detailed-balance). If this were not true--say absorption were greater than emission--then the gas would heat up and the gas would not be in thermodynamic equilibrium. If emission were greater than absorption, the gas would cool. In either case, ultimately the gas would come back into equilibrium then emission exactly balances absorption. In this special case, of course, $dI = 0$, and $I = B_\nu(T)$, the Planck function, so in the Rayleigh-Jeans limit equation (1) implies that

$$B_\nu(T) = \eta_\nu/\kappa_\nu = 2kT_e\nu^2/c^2. \qquad (2)$$

We will come back to this relationship between emissivity and absorption coefficient in a moment.

External Absorption

Consider a cool cloud with a hot source of radiation behind it. The radiation will enter the cloud with intensity I_o, but will be absorbed on passing through the cloud.



In this case, we are assuming that the cloud itself provides negligible emissivity, so equation (1) becomes

$$dI = -\kappa_\nu I dl,$$

where I is the intensity entering each volume element. This is a trivial differential equation whose solution is

$$L$$

$$I = I_o \exp\left(-\int_0 \kappa_\nu \, dl\right).$$

Note that the integration is taken along the line of sight *from the observer*. In the case where the absorption is constant, of course, κ_ν can be brought out of the integral. The integral quantity is a dimensionless quantity called the ***optical depth***, and designated τ_ν. Using this notation, the intensity exiting the cloud is just

$$I = I_o e^{-\tau}$$

The optical depth is a convenient way to refer to the "thickness" of a cloud. It basically measures how many e-foldings of intensity reduction the cloud's thickness represents. A cloud with $\tau_\nu >> 1$ is said to be optically thick, while a cloud with $\tau_\nu << 1$ is optically thin. Optical depth unity is thus an important dividing point between regimes.

Equation of Radiative Transfer

We can rearrange equation (1) to give a first-order ordinary differential equation (*the equation of radiative transfer*) for I, i.e.

$$dI/dl + \kappa_\nu I = \eta_\nu. \qquad (3)$$

This equation is solved by use of the integration factor $e^{-\int \kappa \, dl}$, to get

$$I = I_o e^{-\tau} + \eta_\nu/\kappa_\nu (1 - e^{-\tau}). \qquad (4)$$

Note that the first term is just the same solution we got in the previous section, and represents the contribution of an external source along the line of sight. The second term represents the contribution from the internal emission and absorption of the cloud. Now we come to the nice simplification that we can use in the Rayleigh-Jeans approximation:

$$T_b = T_o e^{-\tau} + T_e \, (1 - e^{-\tau}), \qquad (5 \text{ -- thermal source})$$

where T_b is the brightness temperature that we introduced in the first lecture, we have characterized the intensity of the external source in terms of an equivalent temperature T_o, and we have used equation (2) to relate the ratio of emissivity to absorption coefficient as the Planck function. We have essentially divided equation (4) by $2k\nu^2/c^2$. This last step, of course, assumes thermodynamic equilibrium, so it only holds for a thermal source. However, equation (2) is more general than the thermal case, and equation (5) still holds for nonthermal sources so long as we replace T_e with $T_{eff} = E/k$:

$$\eta_\nu/\kappa_\nu = 2kT_{eff}\nu^2/c^2. \qquad (2')$$

$$T_b = T_o e^{-\tau} + T_{eff} \, (1 - e^{-\tau}). \qquad (5' \text{ -- nonthermal source})$$

There is one additional approximation that is very useful. Recall the two regimes of optical depth, $\tau_\nu >> 1$ (optically thick), and $\tau_\nu << 1$ (optically thin). In these two regimes, equation (5') becomes:

$$T_b = T_{eff}, \qquad \tau_\nu >> 1 \text{ (optically thick regime)} \qquad (6a)$$

$$T_b = T_o(1 - \tau) + T_{eff}\tau. \qquad \tau_\nu << 1 \text{ (optically thin regime)} \qquad (6b)$$

In many cases, it is convenient to think about a given situation in terms of equations (6), although one should keep in mind the form of the more correct equations (5 and 5').

Opacity of Emission Mechanisms

In the previous lecture we wrote down some expressions for emissivity for certain emission mechanisms. However, in radio astronomy it is often simpler to use equations (5) or (6), which require the opacity of a given mechanism, instead of the emissivity. Note that equation (2') shows that these are quite simply related, and in fact I wrote down the emissivities in the previous lecture from expressions for opacity that I found in the literature, by converting from opacity to emissivity using equation (2').

We are now in a position to understand the radio spectrum due to various emission mechanisms by considering the expression for opacity for each mechanism, and the equations (6).

Thermal Bremsstrahlung

Let's start with thermal bremsstrahlung (free-free emission). The emissivity was given in Lecture 2 as:

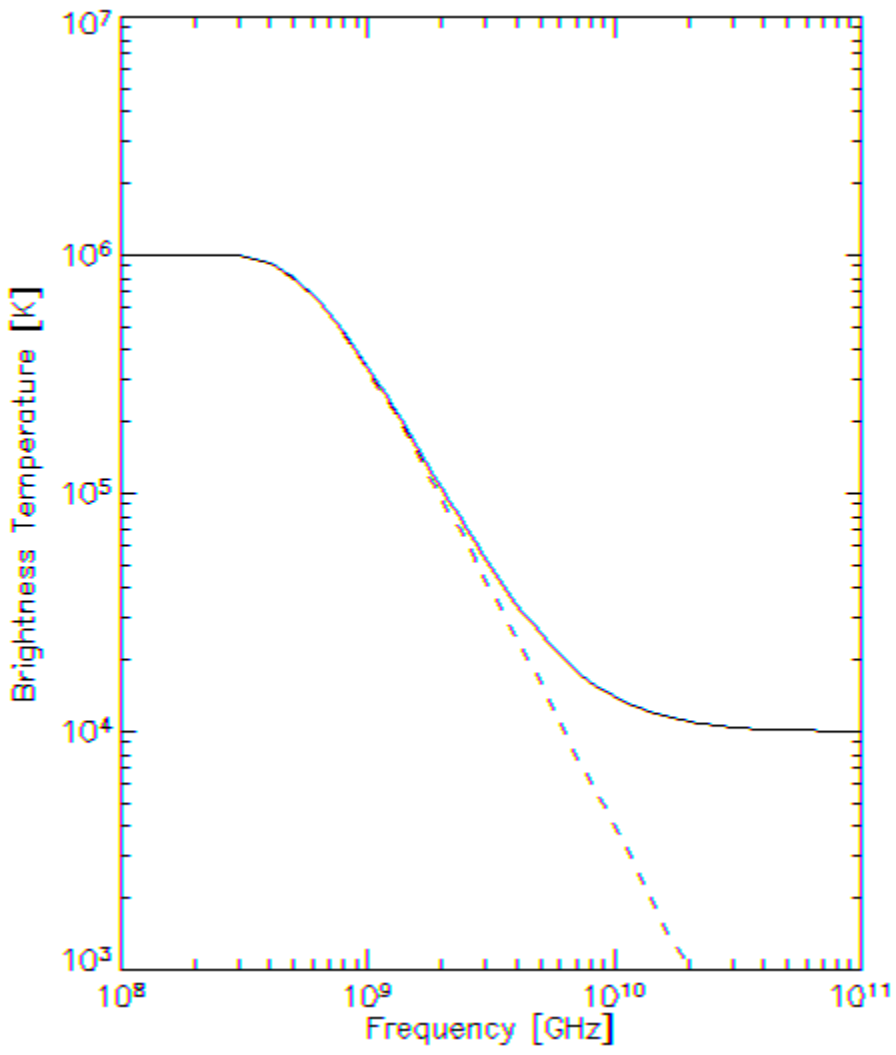
$$\eta_\nu = (2^6\pi e^6/3m_e c^3)(2\pi/3m_e kT)^{1/2}\, n_e n_i Z^2\, G_{ff}(T,\nu).$$

which we noted was independent of frequency except for a slight frequency dependence of the Gaunt factor $G_{ff}(T,\nu)$. Using (2), we have

$$\kappa_\nu = (1/3c)(2\pi/3)^{1/2}(\nu_p/\nu)^2[4\pi n_e \Sigma(n_i Z_i^2)e^4/m_e^{1/2}(kT)^{3/2}]G_{ff}(T,\nu)$$
$$= 9.78 \times 10^{-3} n_e \Sigma(n_i Z_i^2)/(\nu^2 T^{3/2}) \qquad (7)$$
$$\times \begin{cases} 18.2 + \ln T^{3/2} - \ln \nu & (T < 2 \times 10^5 \text{ K}) \\ 24.5 + \ln T - \ln \nu & (T > 2 \times 10^5 \text{ K}) \end{cases}$$

where two values for the Gaunt factor are given for conditions of the solar atmosphere. Note that the optical depth goes as ν^{-2}, so is greater at lower frequencies.

The complete spectrum, for conditions of the solar corona ($T = 10^6$ K, fully ionized hydrogen + 10% helium) will be found by inserting (7) into (5), assuming $\tau = \int \kappa\, dl = \kappa L$, where L is the relevant length of the line of sight through the corona, typically taken to be the density scale height $L = 0.1\, R_{sun}$. The figure below plots the brightness temperature spectrum, assuming that the corona overlies a 10,000 K, optically thick chromosphere.



Thermal Bremsstrahlung brightness temperature spectrum for an iso-thermal solar corona at 10^6 K, plus a 10,000 K chromosphere. The dashed line shows the result without a chromosphere.

This is the brightness temperature spectrum. The flux density spectrum is obtained from equation (5) of Lecture 1, where the brightness is integrated over the source size or the beam size, whichever is smaller.

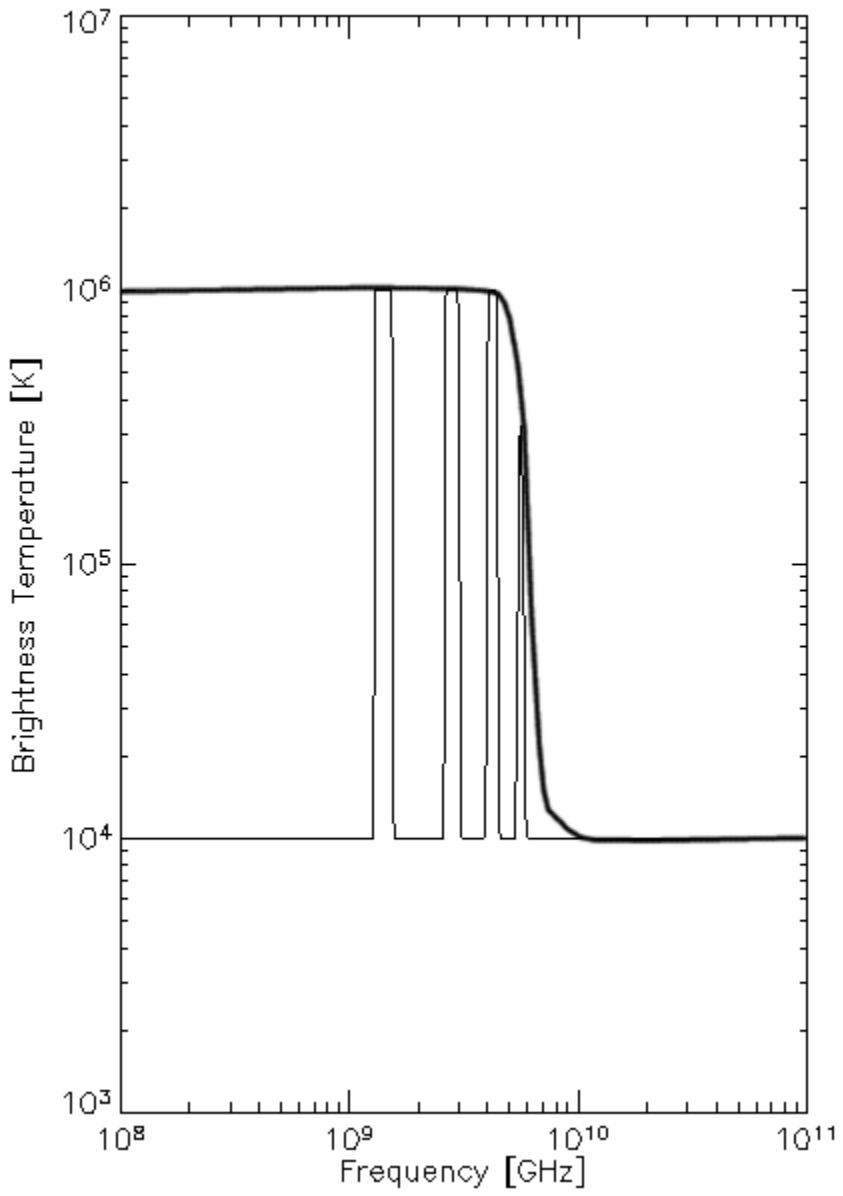
Gyroresonance Emission

In Lecture 2 we gave the emissivity for thermal gyroresonance emission. Here is the corresponding expression for opacity:

$$\kappa_\nu(s,\theta) = \pi^2/4c\, [\mu_\sigma d(\omega\mu_\sigma)/d\omega]^{-1}(\nu_p^2/\nu)(s^2/s!)\, [s^2\beta^2\sin^2\theta/2]^{s-1}[\beta\cos\theta]^{-1}$$
$$\times \exp[-(1 - s\nu_B/\nu)^2/(2\,\mu_\sigma^2\beta^2\cos^2\theta)](1 - \sigma|\cos\theta|)^2 \qquad (8)$$

where $\sigma = +1$ for o-mode (circular polarization in which the electric vector rotates in the direction opposite the direction of the spiraling electrons, and $\sigma = -1$ for x-mode (circular polarization in which the electric vector rotates in the same direction as the electrons), and μ_σ is the index of refraction for the mode. To use this with equation (5), one must multiply by an appropriate scale length. For the case of gyroresonance emission, the scale length is derived from the narrowness of the resonance around $s\nu_B$,

given by the exponential "line-shape" term in (8). An appropriate scale length is $L = 2L_B\beta \cos\theta$, where $L_B = B/\text{grad}(B)$ is the scale length of the magnetic field. Note that to determine the overall opacity, one calculates (8) separately for $s = 1, 2, 3...$, and then adds the contributions to get an overall opacity. A numerical calculation, given an atmospheric structure of n_e, T_e, B, and θ as a function of position along the line of sight is straightforward.

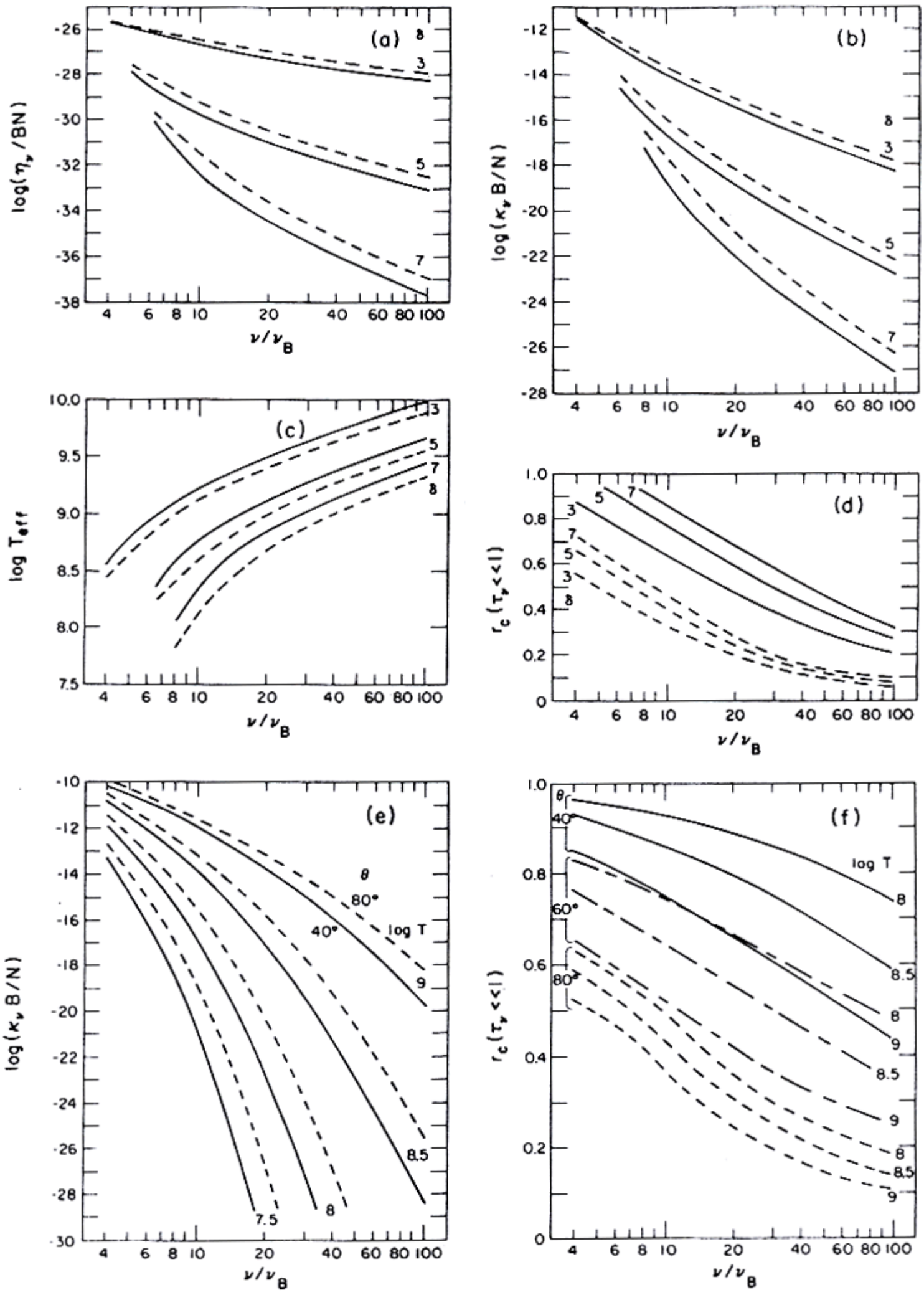


Thermal gyroresonance brightness temperature spectrum for an iso-thermal solar corona at 10^6 K, plus a 10,000 K chromosphere. The line spectrum is what we would see if the magnetic field were uniform at 500 G. The smooth spectrum shown by the heavy line is that expected in the more physical case of a smoothly varying magnetic field, with maximum field strength 500 G.

Gyrosynchrotron Emission

Recall that gyrosynchrotron emission is the same basic mechanism as gyroresonance, but for higher-energy electrons, and so the emission occurs at higher harmonics, typically in the range $s = 5$-100. These electrons may have a "super-hot" thermal distribution, or a power-law energy distribution, or any other physical distribution. Analytical expressions for the thermal case are possible but extremely complicated (see Dulk, 1985, pg 179). Analytical expressions for the power-law case are not possible, but numerical calculations show that the dependence of the emissivity and absorption coefficient are sufficiently simple over some parameter ranges that empirical expressions can be given.

Here are the plots of emissivity, opacity, effective temperature (ratio of emissivity to opacity), and degree of polarization (top four panels), for a power-law distribution:



From Dulk (1985), Ann. Rev. Astronomy & Astrophysics, **23**, 169. Solid lines are for q = 40o, while dashed lines are for q = 80o. The top four panels are for powerlaw electron distributions, while the bottom two panels are for a thermal distribution of electrons.

Note that these log-log plots show some curvature at low harmonics, but above about s=10 they are nearly straight lines (i.e. power-laws) in the various parameters. By fitting power-law functions to them, the following empirical equations were derived that give reasonably accurate results in the range s = 10-100:

$$\frac{\eta_\nu}{BN} \approx 3.3 \times 10^{-24} 10^{-0.52\delta} (\sin\theta)^{-0.43+0.65\delta} \left(\frac{\nu}{\nu_B}\right)^{1.22-0.90\delta},$$

$$\frac{\kappa_\nu B}{N} \approx 1.4 \times 10^{-9} 10^{-0.22\delta} (\sin\theta)^{-0.09+0.72\delta} \left(\frac{\nu}{\nu_B}\right)^{-1.30-0.98\delta},$$

$$T_{\rm eff} \approx 2.2 \times 10^{9} 10^{-0.31\delta} (\sin\theta)^{-0.36-0.06\delta} \left(\frac{\nu}{\nu_B}\right)^{0.50+0.085\delta},$$

$$r_c \approx 1.26\ 10^{0.035\delta} 10^{-0.071\cos\theta} \left(\frac{\nu}{\nu_B}\right)^{-0.782+0.545\cos\theta} \qquad (\tau_\nu \ll 1),$$

$$\nu_{\rm peak} \approx 2.72 \times 10^{3} 10^{0.27\delta} (\sin\theta)^{0.41+0.03\delta} (NL)^{0.32-0.03\delta} \times B^{0.68+0.03\delta}.$$

The plot of T_{eff} shows the optically thick spectral shape, while the plot of η_ν shows the optically thin spectral shape. I want to show how the spectral shape varies as the number of particles changes, which I will do interactively in Photoshop... Note that both η_ν and κ_ν are scaled by the number density, N, of radiating electrons, which makes sense. They are also shown scaled oppositely with respect to B, but I am not sure of the reason.

More on Propagation

At the beginning of this lecture we noted that we were considering propagation along a raypath. What do we mean by that? As the e-m waves propagate, they can refract due to the changing index of refraction in the plasma. We need to derive the general formula for the index of refraction. We have also mentioned several times that electromagnetic radiation can be in the x-mode or o-mode, which are two modes of opposite circular polarization. It is time to introduce something about that. A full treatment is well beyond the scope of this lecture, so I just want to introduce the subject and write down the full expressions for propagation of these two modes in a plasma.

Consider a geometry with propagation of an e-m wave (k direction) at some angle θ from the magnetic field, and define the plane containing k and B as the x-y plane. The part of B along k we will call B_L, and the part transverse to k we will call B_T. We can write down the equations of motion for electrons under the Lorentz force, and we can introduce the influence of collisions by considering a collision frequency ν_c, and write the average force (average rate of loss of momentum) as $-mv\nu_c$. Writing time derivatives with primes (e.g. $a_x = x''$ and $v_x = x'$), the three equations of motion for different components are:

$$mx'' = eE_x - ez'B_T/c - mx'\nu_c$$
$$my'' = eE_y - ez'B_L/c - my'\nu_c$$
$$mz'' = eE_z + ex'B_T/c - ey'B_L/c - mz'\nu_c$$

We then seek wave solutions to this set of coupled equations using the fourier method, and write the resulting equations in terms of the volume polarization, e.g. $P = n_e ex$. The solution is a bit complicated, and makes use of some relations that we would have to take time to develop, but the end result is the Appleton formula, which is the index of refraction in a plasma, written in terms of ratios of frequencies:

$$X = \omega_p^2/\omega^2;$$
$$Y = \Omega_B/\omega;$$
$$Z = \nu_c/\omega;$$

where ω_p^2 is the plasma frequency that we mentioned earlier, Ω_B is the gyrofrequency, and ν_c is the collision frequency. The index of refraction is:

$$n^2 = 1 - \frac{X}{1 - iZ - Y_T^2/2(1 - X - iZ) + \sigma[Y_T^4/4(1 - X - iZ)^2 + Y_L^2]^{1/2}} \qquad \text{(Appleton formula)}$$

where $\sigma = +1$ for o-mode and $\sigma = -1$ for x-mode (as before when we gave the expression for gyroresonance opacity). Here, Y_L and Y_T are the longitudinal and transverse parts Y, and involve the gyrofrequency for the corresponding components of B_L and B_T. When we can ignore collisions ($Z = 0$), the index of refraction becomes purely real and we have:

$$2X(1 - X)$$

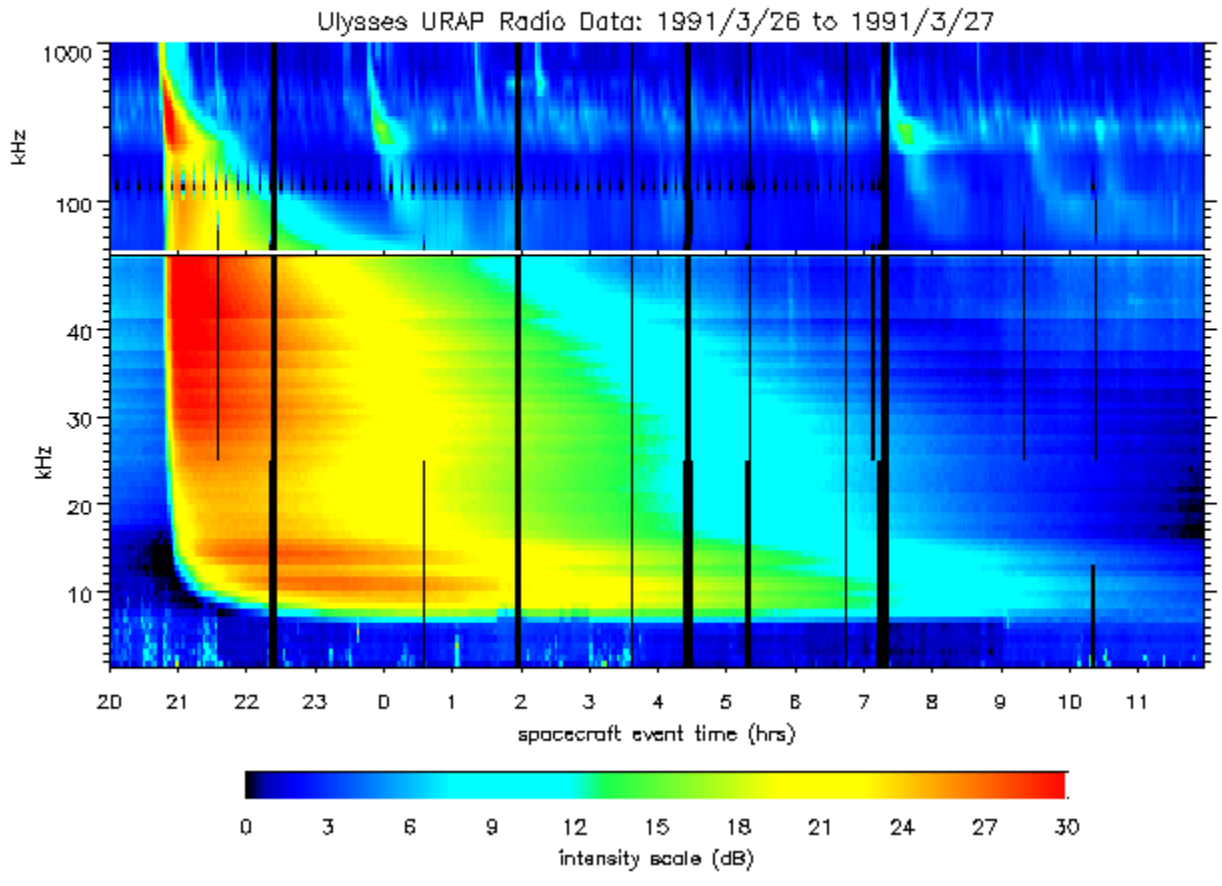
$$\mu^2 = 1 - \frac{}{2(1-X) - Y_T^2 + \sigma[Y_T^4 + 4(1-X)^2 Y_L^2]^{1/2}}$$ (Appleton-Hartree formula)

Let's look at some results of this formula:

No magnetic field and no collisions
In this case, the index of refraction becomes very simple:

$$\mu^2 = 1 - X = 1 - \omega_p^2/\omega^2$$

In most mediums you may be familiar with, you are used to the index of refraction being greater than one, but for a plasma (with no magnetic field) the index of refraction is always less than 1. Since the index of refraction is the ratio of speed of light in a vacuum to phase speed in the medium, this means that the phase speed in the medium exceeds the speed of light! But the group speed is less than the speed of light, so it does not violate special relativity. Note what happens when $\omega < \omega_p$. In this case, the index of refraction becomes imaginary ($\mu^2 < 0$), and the e-m wave is evanescent (cannot propagate). So there is a limit to how low a frequency the plasma can support, and waves at frequencies below the plasma frequency cannot propagate (see the figure below).



Notice that the emission from this type III burst, which is due to plasma emission at the local plasma frequency, stops abrubtly at about 7 kHz. That is because the Ulysses spacecraft is embedded in plasma whose plasma frequency is 7 kHz, so radio waves at frequencies less than this cannot reach the spacecraft. You can determine the local electron density from this fact.

As another example, if one sends a radio wave from the ground, vertically through the Earth's ionosphere, it will propagate upward initially with $\mu^2 = 1$, but as it enters the ionized medium (the plasma) μ will become smaller as the density increases (as ω_p increases). If the frequency of the wave is high enough (above about 8 MHz), μ will never reach zero, so although the group speed slows down (the signal is delayed), it still makes it through. However, if the frequency is below 8 MHz, the wave will propagate only until $\mu = 0$, at which point it will reflect and propagate downward. This is why shortwave radio signals skip off the ionosphere. If the propagation is not vertical, but is instead at some angle ϕ_o from the ground, then the reflection will occur earlier, when Snell's Law, $\mu \sin\phi = \mu_o \sin\phi_o$ is satisfied for $\phi = 90$. Thus, the reflection will occur at a value $\mu = \sin\phi_o$. So for radio waves propagating at some angle, the reflection will occur even at higher frequencies, up to 25 MHz or more. The ionosphere and the ground form a kind of waveguide for shortwave communication. It is worthwhile in this context to mention that when collisions are important, the radiation can be absorbed (due to imaginary part of μ^2). This occurs when solar X-rays ionize the lower atmosphere of Earth (the ionosphere, normally collisionless, extends to lower heights of higher density), and can cause loss of shortwave communication (so-called shortwave fadeout).

Magnetic field, no collisions
In this case, the plasma becomes birefringent, that is, there are two distinct modes in the plasma, again called the o-mode and the x-mode. Let's see where the index of refraction reaches zero (that is, where the waves become evanescent, or reflect):

$X = 1$ (o-mode)
$X = 1 - Y$ (x-mode, valid for $Y < 1$)
$X = 1 + Y$ (x-mode, valid for $Y > 1$)

The top two expressions are the reason that plasma emission tends to be o-mode polarized in the solar corona. Plasma emission, by its very nature, is generated near the plasma frequency, which is allowed for o-mode, since there $\mu = 0$, but is not imaginary. For x-mode, however, the second expression $X = 1 - Y$ says that frequencies $\omega < \omega_p + \Omega_B/2$ will be evanescent. Since the plasma emission is generated at $\omega = \omega_p$, this condition is fulfilled, and the x-mode is evanescent and thus not produced. There is a thermal width to the plasma frequency resonance, however, so in some cases some x-mode emission can escape, but only for small B.

In the example of reflection from the Earth's ionosphere, one must take into account the Earth's magnetic field. As a consequence, for normal propagation, o-mode polarized radiation will reflect at the layer where $X = 1$, (i.e. $\omega = \omega_p$, as before), but x-mode radiation will reflect at a lower layer, where $X = 1 - Y$.

Overview of the Emission and Propagation Process

To put all of this together, we have the following processes:

- emission from single, charged particles occurs typically due to accelerations by various forces, although any process that leads to transverse electric fields will lead to radiation.
  - electrons are far more efficient than ions for radiation
  - radiation process is defined by the cause of the accelerations
  - more than one radiation process can be active at once
- to determine the emissivity for a radiation process, one must integrate over a distribution of particles
  - thermal emission
  - nonthermal emission => powerlaw distribution or other, more complex distributions
- the radiation propagates through the plasma
  - may refract due to changing index of refraction
  - intensity increases along the raypath up to a point, but when the plasma becomes optically thick, the intensity is fixed at that due to the appropriate effective temperature.

Physics 728 Radio Astronomy: Lecture #4 Prof. Dale E. Gary NJIT

Primary Antenna Elements

Introduction

The antennas of an array have two main purposes:

- to intercept, or collect, the radiation and direct it to the receiver
- to restrict the field of view

The first function of the primary element can be expressed in terms of the antenna ***effective area***, $A(\nu, \theta, \phi)$, in units of m^2, where θ, ϕ are direction coordinates on the sky. If the brightness of a source is $I(\nu, \theta, \phi)$ in $\mathrm{W\ m^{-2}\ Hz^{-1}\ ster^{-1}}$, then the power collected, in W, is found by integrating over solid angle $\Delta\Omega$ and bandwidth $\Delta\nu$:

$$P = A(\nu, \theta, \phi)\ \underbrace{I(\nu, \theta, \phi)\ \Delta\Omega\ \Delta\nu}_{\text{flux density}\quad (\mathrm{W\ m^{-2}\ Hz^{-1}})}$$

flux $\quad (\mathrm{W\ m^{-2}})$

Associated with the collecting area (effective area) is the ***beam pattern***, also called the ***primary beam***, which is just the *Fourier Transform* of the aperture, as shown in the figure below.

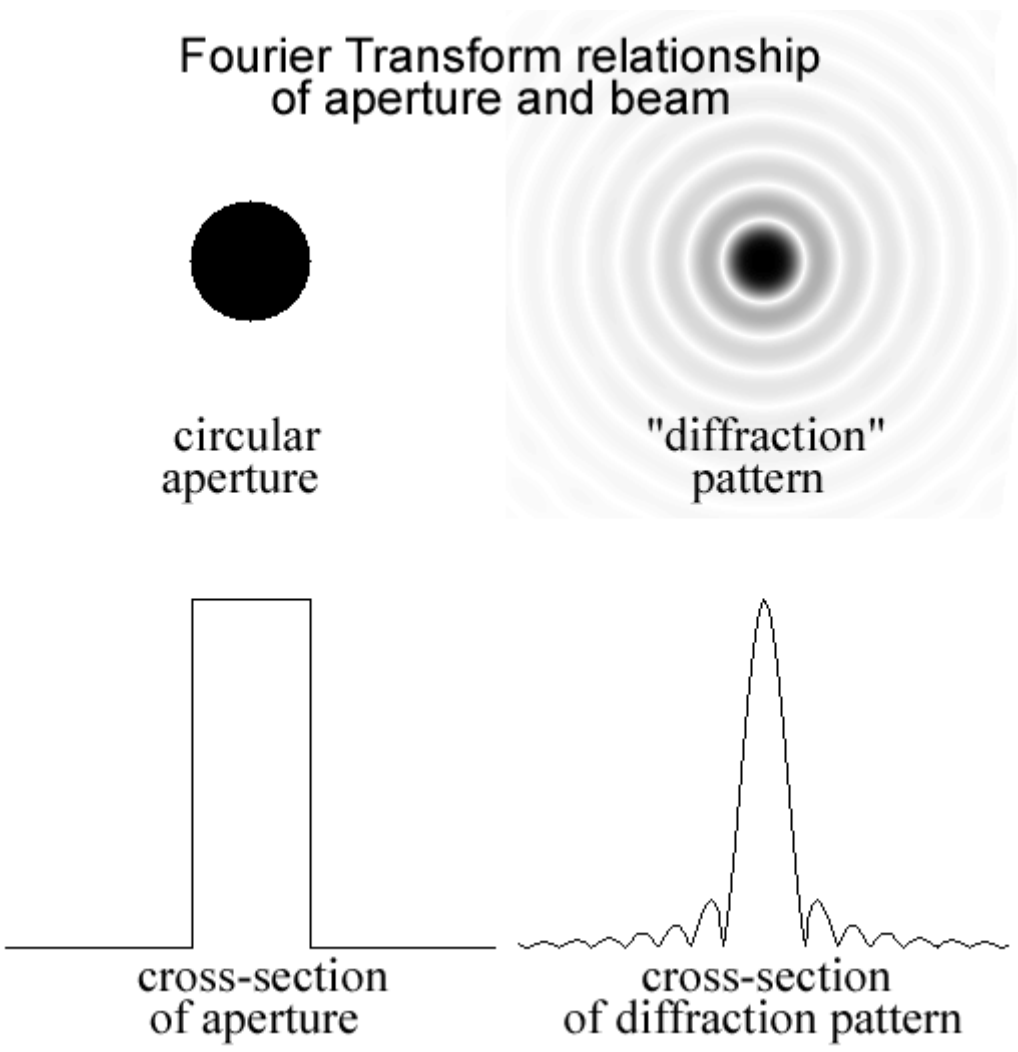


Figure 1.

We will come back to the primary elements, and the beam pattern, or primary beam, shortly. But first, let's take a closer look at Fourier Transforms.

Fourier Transform Relationship and Inverse

This is the first time we have explicitly met the Fourier Transform relationship, but since it occurs over and over in radio astronomy, it is worthwhile to look at it in some detail, in particular the Fast Fourier Transform. Those of you who have access to IDL (Interactive Data Language) may wish to experiment with the FFT function.

Given two spatial coordinates, x, y, in m, say, we consider the corresponding ***spatial frequencies***, u, v, in wavelengths, defined as $u = x/\lambda$, $v = y/\lambda$. Then $F(l,m)$ is the fourier transform of $f(u,v)$:

$$2\pi i(ul + vm)$$

$$F(l,m) = \int\int_{\text{aperture}} f(u,v)\, e \qquad du\, dv$$

where

- $f(u,v)$ = complex voltage distribution in the aperture (Fig. 1, upper left, assumes a flat plane wave)
- $F(l,m)$ = complex far-field voltage radiation pattern, with l,m = angular distances on the sky (pattern is shown in Fig. 1, upper right)

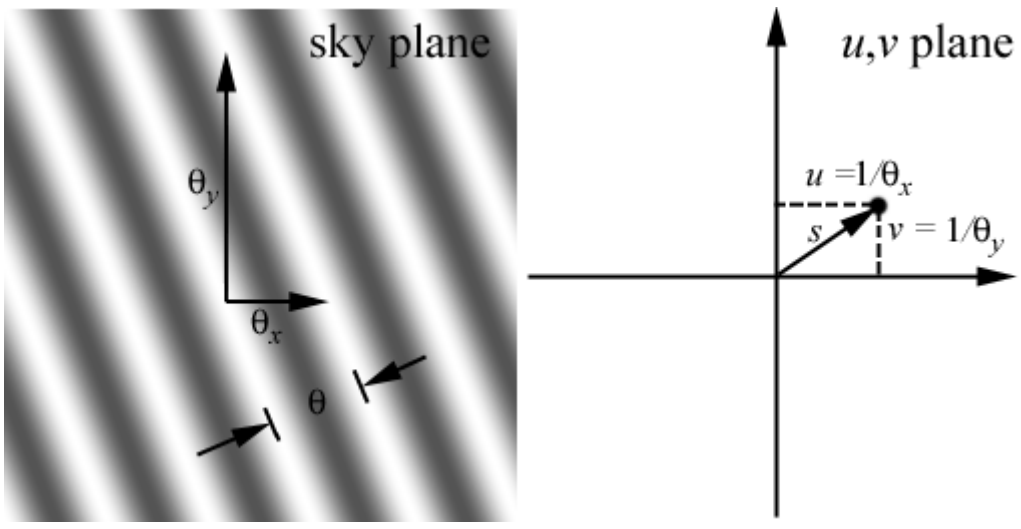
The inverse is written as

$$f(u,v) = \int\int_{\text{all sky}} F(l,m)\, e^{-2\pi i(ul + vm)} \qquad dl\, dm$$

and you would perform this inverse with the FFT by using fft(F,-1), where the -1 indicates the inverse Fourier Transform.

We are going to use Fourier Transforms further in discussing the response of pairs of antennas (interferometer baselines), which we will do in Lecture 6. Let's go ahead and introduce that now. A pair of antennas measures one point in the u,v plane, and the Fourier Transform of this point gives the familiar interference fringes on the sky.

These relationships are shown graphically in the figure below:



A point in the u,v plane a distance s from the origin has components u and v. In radio astronomy, this corresponds to a single baseline, or pair of antennas. The FT of this sampling corresponds to fringes in the sky plane, with angular separation θ = fringe spacing. The two corresponding angular coordinates are θ_l and θ_m, which are the fringe separations in the l and m angular directions.

The following terminology is used

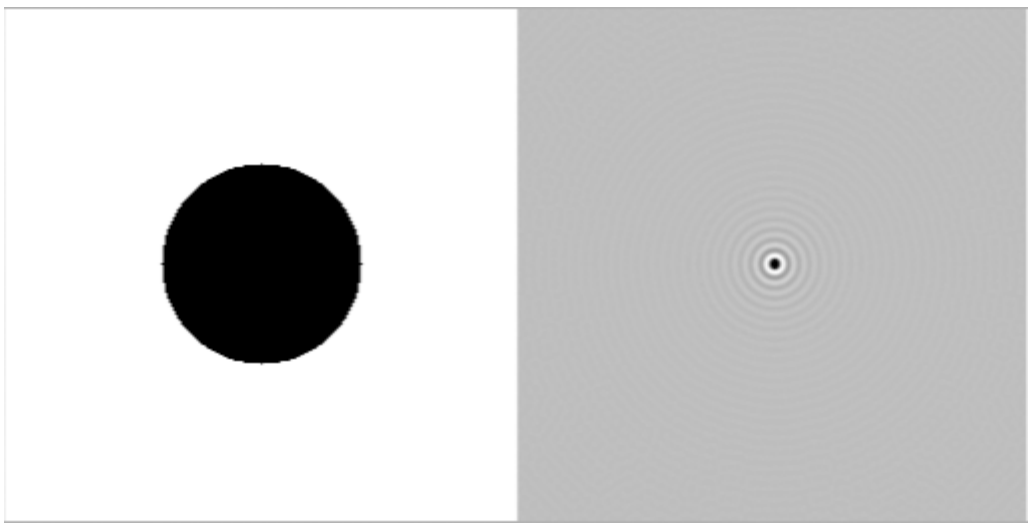
- $s = (u^2 + v^2)^{-1/2}$ = spatial frequency
- $s^{-1} = \theta$ = fringe spacing
- u = "x-component" of s
- v = "y-component" of s

When we use FFT's, of course, we must pixelize, or grid, the data into 2-d arrays that the computer can deal with. The first step, then, is to pick a pixel size for the u,v plane. Let's look at the antenna aperture problem, and consider an antenna of 6 m aperture. Note that we also have to know what frequency (wavelength) we want to consider. At 5 GHz, the wavelength is 6 cm, so 6 m is 100 wavelengths (we can say that the dish diameter is $D_\lambda = 100$). We might choose a pixel size corresponding 1 wavelength, so that our aperture will occupy a circle of radius 50 pixels (diameter 100 pixels). After we do the FFT, what is the corresponding pixel size in the sky plane (also called the map)?

If you step into the first pixel in the u,v plane, this corresponds to one cycle (one fringe) fitting across the entire map. The second pixel corresponds to two cycles fitting across the map, and so on. We said the fringe spacing was s^{-1}, so stepping one pixel ($s = \Delta s$) yields a fringe spacing of $\theta = N\Delta\theta = 1/\Delta s$ (radians), where N is the number of points in the array. Thus, for a u,v array of 256 points, with each pixel in the array representing 1 wavelength, we will have each pixel in the map corresponding to

$$\Delta\theta = (N\Delta s)^{-1} = 1/256 \text{ radian} = 13.4 \text{ arcmin}.$$

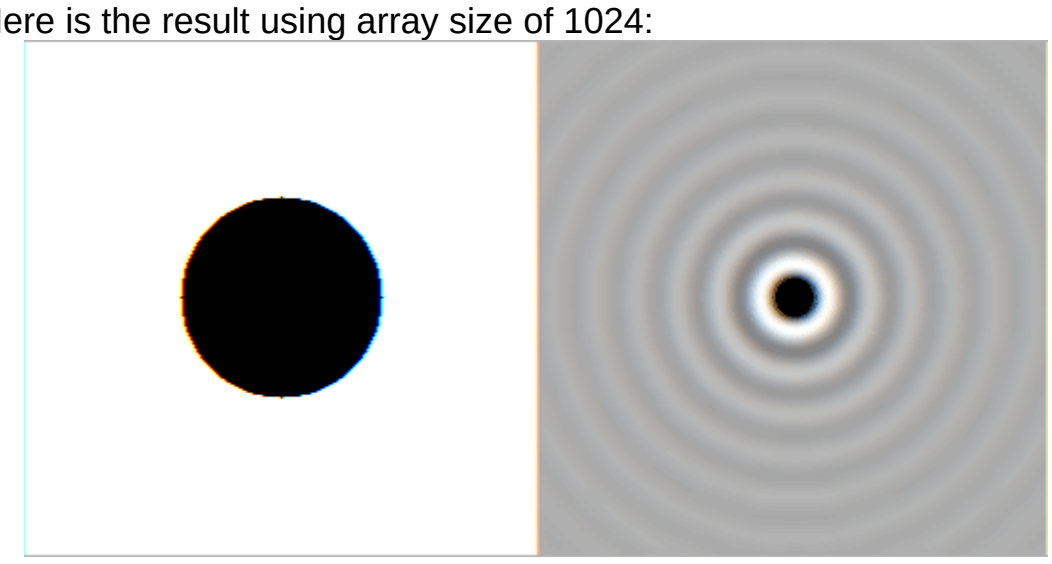
So how many pixels should the beam be? We said in the first lecture that the beam was $\theta \sim \lambda/D = 1/D_\lambda$, which is 0.01 radian, or 35 arcmin, so the beam will occupy less than three pixels! Here is the result:



ap = float(dist(256) le 50.) beam = fft(ap)

The IDL commands used to create the above two images, are shown just below each image.

So to see the beam shape in better resolution we have two choices--use a larger Δs, or use a larger array (larger N). Here is the result using array size of 1024:



ap = float(dist(1024) le 50.) beam = fft(ap)

The IDL commands used to create the above two images, are shown just below each image. This is just the inner 256x256 pixel part of each.

It may seem paradoxical that we can put the same size circular aperture into a larger array full of blank space, and get a higher resolution view of the beam shape, but that is the way it works. Note that we have the following useful relations:

$\Delta\theta = (N\Delta s)^{-1}$ (relation between pixel sizes in the two planes)
$N\Delta\theta = \Delta s^{-1}$ (relation between map sizes in the two planes)

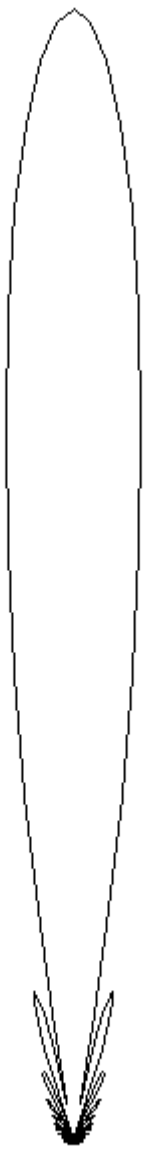
These are symmetrical:

$\Delta s = (N\Delta\theta)^{-1}$ (relation between pixel sizes in the two planes)
$N\Delta s = \Delta\theta^{-1}$ (relation between map sizes in the two planes)

Note also that the smallest fringe spacing possible is one in which every other pixel is a peak (i.e. $N/2$ fringes across the map). This is the Nyquist frequency, and corresponds to the largest useful distance in the u,v plane, e.g. $u = N/2$. If you try to go beyond $u = N/2$, the point "wraps around" and becomes $u = -N/2$.

Back to the Primary Beam

The "power pattern" of a single dish telescope is $|F(l,m)|^2$, i.e., the beam pattern is the square of the complex far-field voltage pattern, $|F(l,m)|^2 = F(l,m)F^*(l,m)$. Here is a plot of the angular pattern:



where you can see the main lobe and the side lobes. This pattern is, of course, azimuthally symmetric. The dish is maximally sensitive to radiation from the direction of the peak of the beam, but is also slightly sensitive to sources in the side lobes. If we write $A(\nu, \theta, \phi)$ for the collecting area, as before, then we can normalize to the peak value on-axis, $A(\nu, 0, 0) = A_o$ and consider the normalized beam pattern $A(\nu, \theta, \phi) = A(\nu, \theta, \phi) / A_o$. Then the beam solid angle is

$$\Omega_A = \int\int_{\text{all sky}} A(\nu, \theta, \phi)\, \Delta\Omega$$

Directivity

An isotropic antenna would have solid angle $\Omega_A = 4\pi$. The ***directivity*** (also called the ***gain*** of the antenna) is defined as the ratio of peak radiation intensity to average radiation intensity

$$D = 4\pi / \Omega_A$$

so a highly directive antenna has a small Ω_A.

Collecting Area vs. Physical Area

A_o is the collecting area when pointed directly at a source, yet this may not be the same as the physical area of the aperture. The ratio can be used to define an aperture efficiency $\eta < 1$

$$A_o = \eta A$$

which takes into account a variety of losses. We will come back to this shortly.

A fundamental relationship between A_o and Ω_A is

$$A_o \Omega_A = \lambda^2$$

which shows that as A_o gets bigger, Ω_A gets smaller (D = directivity increases).

Antenna Types

At wavelengths below ~ 1 m (300 MHz), simple wire antennas (e.g. dipoles, yagis, spirals, etc.) can be used. Note that we can *define* the collecting area of a dipole as

$$A_o = \lambda^2 / \Omega_A$$

and use the dipole radiation pattern, proportional to $\sin^2\theta$ to determine Ω_A. Other combinations, such as

a yagi antenna, have a different radiation pattern, but the relationship can still be used.

At short wavelengths, wires give too small a collecting area, so reflecting surfaces are used. We will concentrate on reflecting dishes from here on. A useful ***rule of thumb*** is that when the wavelength grows so large that a dish is less than 5 wavelengths in diameter, the dish is no longer competitive with the same size dipole. That means a 2 m dish, for example, is only effective to wavelengths shorter than 40 cm => frequencies higher than 750 MHz.

See Napier lecture for examples of types of antennas.

Antenna Performance

Aperture Efficiency

As we said a moment ago, the aperture efficiency $A_o = \eta A$ is made up of several factors. In general, we can write the efficiency as the product of several factors

$$\eta = \eta_{sf}\eta_{bl}\eta_s\eta_t\eta_{misc}$$

where

η_{sf} = reflector surface efficiency

η_{bl} = reflector blockage efficiency

η_s = feed spillover efficiency

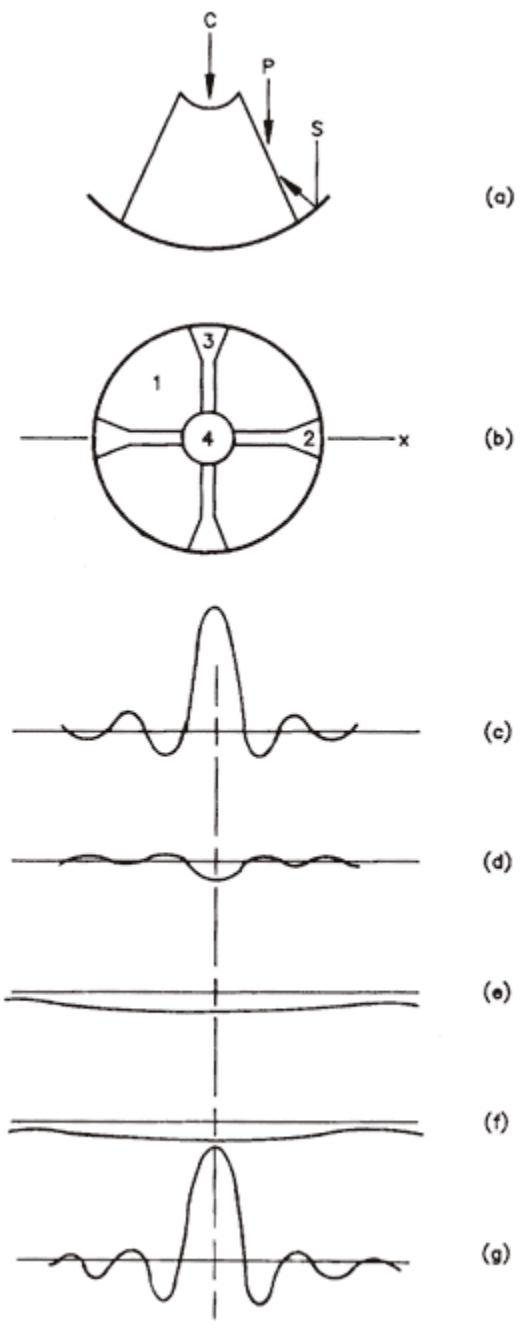
η_t = illumination efficiency

η_{misc} = losses due to reflector diffraction, feed position errors, feed mismatch, and others

See Napier lecture for discussion of reflector surface efficiency.

Aperture Blockage Efficiency

Many antenna designs (prime focus and cassegrain), require structures to hold the feed and/or subreflector in place, and these structures block part of the aperture. The situation is shown in the figure below, and can be analyzed by considering different parts of the structure separately, e.g. regions 1, 2, 3 and 4 in (b), and doing the Fourier Transform of each part. Thus, (c) is the FT of the unblocked aperture, (d) is the FT of part 2 (shown negative because it is a blockage), (e) is the FT of part 3, also negative, (f) is the FT of part 4, and finally (g) is the sum of the FT's. The reason one can separately treat each part is because ***the FT procedure is linear***. This is a very important property that we will use several times in the course.



The efficiency due to blockage is

η_{bl} = (1 - area blocked/total area)^2 ~ (1 – 2*area blocked/total area) for small blocked/total.

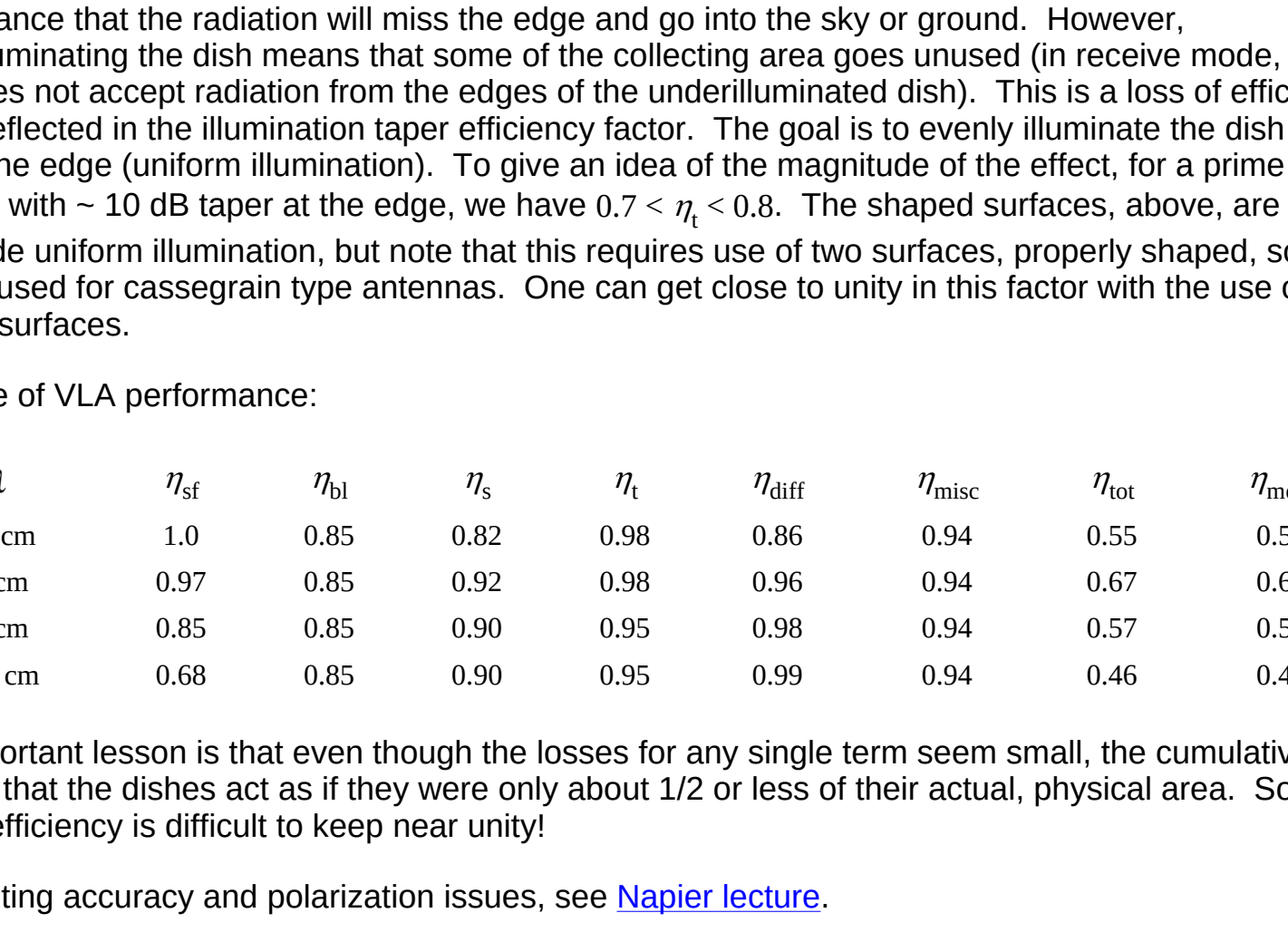
Note that the blockage affects the beam *shape* in addition to the aperture efficiency. Some designs are off-axis to reduce or eliminate blockage (but this may cause other unwanted effects).

Feed Spillover Efficiency

Consider the antenna as a transmitter. For a prime focus antenna, spillover efficiency is the fraction of radiated power intercepted by the reflector. Power not intercepted is lost. Telescopes for radioastronomy often have $F/D = 0.4$, so the half-angle that the dish presents as viewed from the feed is $\theta_R = \text{atan } 0.5/0.4 = 51^o$, so the total width is about 100^o. Thus, we need a broad feed pattern to illuminate the dish. Note that any spillover "sees" the ground in receiving mode, and that can increase the noise temperature, which we will discuss later.

The situation is different for a Cassegrain design, where the feed is at the reflector and the radiation has to hit the subreflector (secondary). In that case, θ_R is much smaller, so we need to use a more directive feed. In this case, spillover "sees" the sky.

Typically, $0.7 < \eta_s < 0.97$, with the higher values requiring "shaped" reflectors as in the figure below.



Illumination Taper Efficiency

We could certainly avoid spillover in the above consideration by under-illuminating the dish so that there is no chance that the radiation will miss the edge and go into the sky or ground. However, underilluminating the dish means that some of the collecting area goes unused (in receive mode, the feed does not accept radiation from the edges of the underilluminated dish). This is a loss of efficiency that is reflected in the illumination taper efficiency factor. The goal is to evenly illuminate the dish all the way to the edge (uniform illumination). To give an idea of the magnitude of the effect, for a prime focus antenna with ~ 10 dB taper at the edge, we have $0.7 < \eta_t < 0.8$. The shaped surfaces, above, are used to provide uniform illumination, but note that this requires use of two surfaces, properly shaped, so can only be used for cassegrain type antennas. One can get close to unity in this factor with the use of shaped surfaces.

Example of VLA performance:

λ	η_{sf}	η_{bl}	η_s	η_t	η_{diff}	η_{misc}	η_{tot}	η_{meas}
20 cm	1.0	0.85	0.82	0.98	0.86	0.94	0.55	0.51
6 cm	0.97	0.85	0.92	0.98	0.96	0.94	0.67	0.65
2 cm	0.85	0.85	0.90	0.95	0.98	0.94	0.57	0.52
1.3 cm	0.68	0.85	0.90	0.95	0.99	0.94	0.46	0.43

The important lesson is that even though the losses for any single term seem small, the cumulative effect is that the dishes act as if they were only about 1/2 or less of their actual, physical area. So the overall efficiency is difficult to keep near unity!

For pointing accuracy and polarization issues, see Napier lecture.

Physics 728 Radio Astronomy: Lecture #5 Prof. Dale E. Gary
NJIT

Front End Receiving System

Power vs. Temperature

The power level of the radiation (W) can be traced from its reception by the feed, through the receiving system. The "signal" is generally noise-like (white noise, containing all frequencies in the band). For convenience, which will become clear, we often consider the equivalent ***noise temperature*** corresponding to the power level

$$P = kT\,\Delta\nu$$

although we also refer to the power level in decibel milliwatts [dBm].

We can consider the power received by the antenna,

$$P_a = kT_a\Delta\nu, \qquad (1)$$

where T_a is the ***antenna temperature***, and the output power of the receiver as

$$P_{tot} = P_a + P_{sys} => T_{tot} = T_a + T_{sys},$$

where T_{sys} is the ***system temperature***, and represents the added noise of the system. It is a figure of merit, and should be kept as low as possible. We can break the system temperature into several contributions:

$$T_{sys} = T_{bg} + T_{sky} + T_{spill} + T_{loss} + T_{cal} + T_{rx},$$
$$\text{lump into } T_a \qquad \text{lump into } T_{rx}$$

where

T_{bg} = noise contribution from microwave and galactic backgrounds
T_{sky} = noise contribution from atmospheric emission
T_{spill} = noise contribution due to ground radiation (spillover and scattering)
T_{loss} = noise contribution due to losses in feed
T_{cal} = noise contribution due to injected noise
T_{rx} = receiver noise temperature

To give you an appreciation for the magnitude of these contributions, the following table gives the VLA performance as an example:

λ	T_{bg}	T_{sky}	T_{spill}	T_{loss}	T_{cal}	T_{rx}	T_{sys}
92 cm	25	3	15	7	5	70	125
20 cm	3	3	14	8	2	30	60
6 cm	3	3	7	5	2	30	50
3.6 cm	3	2	5	2	2	16	30
2 cm	3	8	6	13	6	80	116
1.3 cm	3	17	6	21	7	100	154

- Note: T_{bg}, T_{sky}, *and* T_{spill} vary with position on the sky.
- T_{bg} is the 3 K cosmic background (leftover emission from the Big Bang), except at the lowest frequency where radio emission from the galaxy becomes important.
- $T_{sky} = T_{sky}(1-e^{-\tau_o \csc E})$, where τ_o = zenith opacity, E = elevation angle. The value of T_{sky} is generally small except at high frequency where the atmosphere beocmes more opaque.

Noise From The Source (*i.e. the Signal*) T_a.

To see how important the unwanted T_{sys} is, let's compare it with a typical signal. Say we have a point source of flux density 1 Jy [= 10^{-26} W m^{-2} Hz^{-1}]. If observed with a radio telescope of 10 m diameter, what is T_a?

$$S = 2kT_a \nu^2/c^2\Delta\Omega\ . \qquad (2)$$

Note that $\Delta\Omega > \Delta\Omega_{source}$ so use $\Delta\Omega_{beam}$. From $\theta_{FWHM} \sim \lambda/D$

$$\Delta\Omega_{beam} = \pi/4\ (\theta_{FWHM})^2 = \pi/4\ (c^2/D^2\nu^2) \qquad (3)$$

Insert (2) into (1) and solve for T_a,

$$T_a = Sc^2/2k\nu^2\ 4/\pi\ (D^2\nu^2/c^2) = 4SD^2/2\pi k = 4(10^{-26})(10^2)/2\pi(1.38 \text{ x } 10^{-23}) = 0.04 \text{ K!}$$

So in order to measure a 1 Jy source, we have to measure 0.04 K against of order 100 K of system noise. If we make the antenna more directive (D larger) then the situation improves. The 100 m Bonn telescope, or the GBT antenna would see 4 K from a 1 Jy source. The 1000 ft Arecibo dish would see ~40 K. However, recall from last time that the effective area of an antenna is reduced by the efficiency, $A = \eta A_o$, with $\eta \sim 0.5$, so the problem is even worse. The appropriate antenna temperature expression for a source of flux density S is

$$T_a = S\eta A/2k. \qquad (4)$$

so Bonn has 1.5 K/Jy, Arecibo: 15 K/Jy. Clearly we are measuring very small signals.

Sensitivity of a Single Dish

We can take advantage of the fact that the signal will be correlated from one sample to the next, while the noise will not be, and "beat down the noise" by making making many sample measurements and adding them up. According to the Nyquist theorem, a time series of measurements of signal of bandwidth $\Delta\nu$, of duration τ, will contain $2\Delta\nu\tau$ independent samples, so the noise should go down by the square-root of the number of samples, or

$$\Delta T = T\ (2\Delta\nu\tau)^{-1/2}\ , \qquad (5)$$

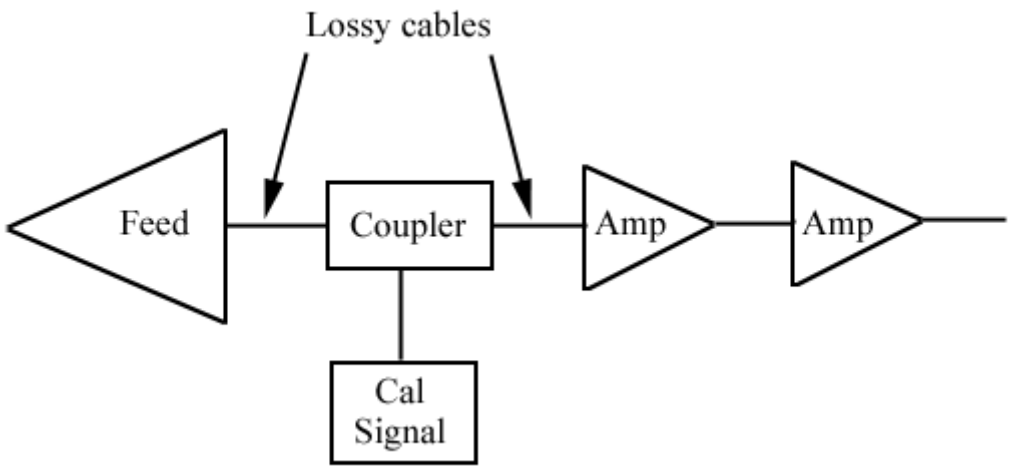
where T is the equivalent temperature of the signal plus noise. A more general expression, from Crane and Napier (1989) and Anantharamaiah (1989) [an earlier version of the NRAO Summer School book] is

$$\Delta T = \left[\frac{T_a^{\,2} + T_a T_{sys} + T_{sys}^{\,2}/2}{\Delta\nu\tau} \right]^{1/2} \qquad (6)$$

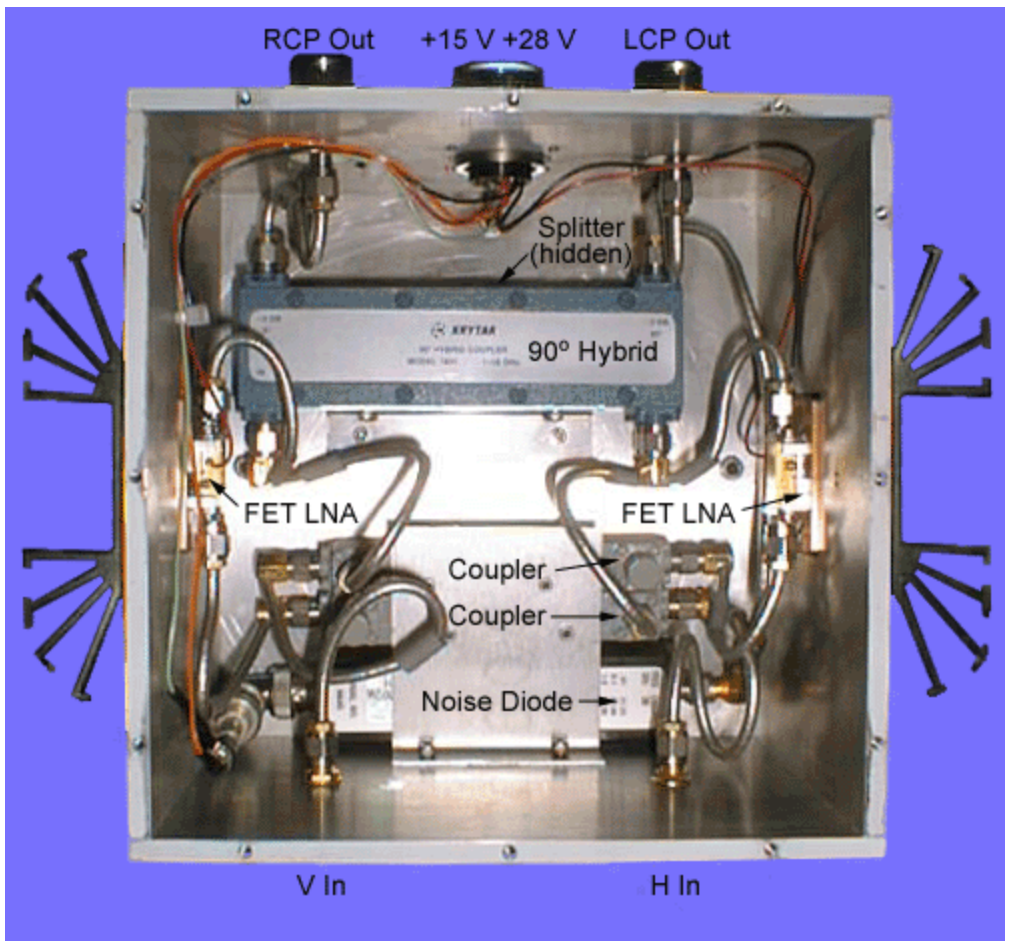
Normally when observing cosmic sources, we can ignore T_a relative to T_{sys}, which reduces to equation (5) with $T = T_{sys}$. However, what happens when we observe a strong source such as the Sun? The brightness temperature of the quiet Sun is about 10,000 K at, say, 10 GHz (recall homework problem set #1), so if we observe the Sun with a 25 m antenna with aperture efficiency ~ 0.5, the 10,000 K source will fill the beam and the antenna temperature will be 5,000 K. This is now much larger than the typical system temperature, so we say that the source dominates the noise, and the sensitivity of the antenna is now less than before. We will revisit this when we examine the sensitivity of an interferometer, and discuss the minimum noise level in radio images.

Sources of Noise in Receivers--the Front End

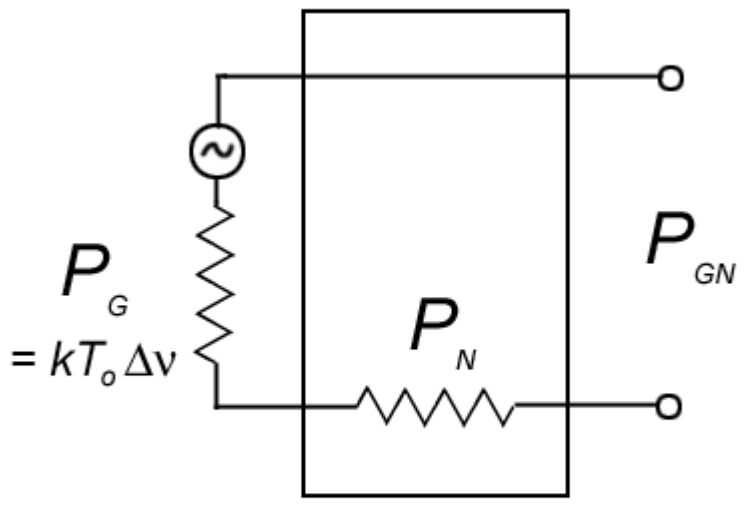
Let us now examine the signal path from the feed to the receiver. This chain is called the front end, and its characteristics dominate the noise of the system. A typical front end consists of the feed, some connecting cables, perhaps with filters or couplers, then a first stage low-noise amplifier (LNA), then typically a second stage LNA, followed by the receiver itself, as shown in the schematic below:



A front end that I recently built for Lucent Technologies (the Solar Radio SpectroPolarimeter) is shown in the figure below:



Before considering individual elements, we develop the concept of a general 2-port device, as shown by the rectangle in the schematic below:



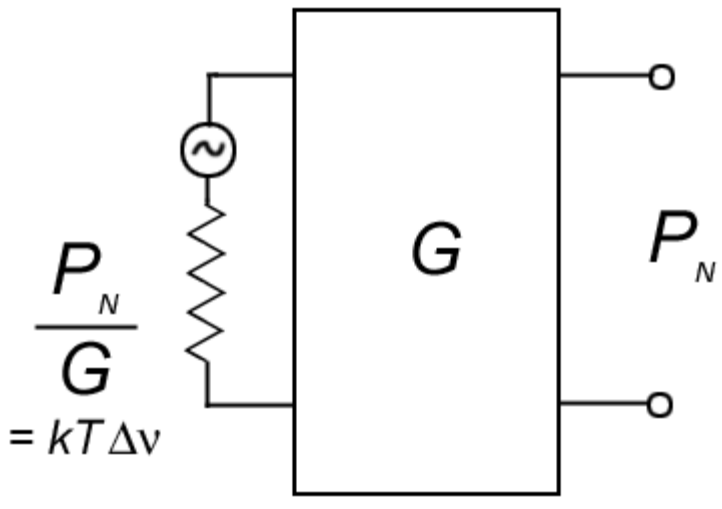
The input to this device is some hot resistor R, which generates noise $P_G = kT_o\Delta\nu$ by virtue of its temperature T_o = ambient (~290 K). The device has some internal resistance giving added noise P_N. We can describe the gain of the device as

$$G = P_{GN}/P_G,$$

so that the output power is

$$P_{GN} = GkT_o\Delta\nu.$$

Now replace the 2-port device with an ideal (lossless) device and adjust the input to give an output P_N (not P_{GN}). The temperature T of the external resistor required to attain output power P_N of the original device is called the ***noise temperature*** of the device. If the original device has no internal noise, then $P_N = 0$ and $T = 0$.



We define the ***noise figure***, NF, as

$$\mathrm{NF} = (P_{GN} + P_N) / P_{GN} \qquad (= 1 \text{ if no loss})$$

or

$$\mathrm{NF} = 1 + P_N / P_{GN} \;= 1 + GkT\Delta\nu / GkT_o\Delta\nu \;= 1 + T / T_o\,. \qquad (6)$$

So the noise figure is related to noise temperature by

$$T = (\mathrm{NF} - 1)\, T_o \qquad [T_o = \text{ambient temperature, usually considered to be 290 K}]$$

Often, NF is given in dB:

$$\mathrm{NF}_{\mathrm{dB}} = 10 \log \mathrm{NF}$$

Example: Miteq LNA noise figure.

Attenuators as 2-Port

One type of 2-port device is a "passive" attenuator. Transmission lines and cables between components have losses, and can be considered as an attenuator. The "gain" of an attenuator is $G = \varepsilon < 1$, and associated with this gain is the loss factor $L = 1/\varepsilon$.

As an example, a 3 dB attenuator has

$$L_{\mathrm{dB}} = 3 \text{ dB} \;\Rightarrow\; L = 2 \;\Rightarrow\; \varepsilon = 0.5.$$

As before, but replacing G with ε

$$P_{GN} = \varepsilon kT_o\Delta\nu = kT_o\Delta\nu / L.$$

The noise output of an attenuator is

$$P_N = (1 - \varepsilon)\, kT_{\mathrm{Phys}}\Delta\nu,$$

where T_{Phys} = physical temperature of cable or device.

Then from (6), the noise figure of an attenuator is

$$\mathrm{NF} = 1 + P_N / P_{GN} \;= 1 + (1 - \varepsilon)\, kT_{\mathrm{Phys}}\Delta\nu / \varepsilon kT_o\Delta\nu = 1 + (L - 1)T_{\mathrm{Phys}} / T_o\,.$$

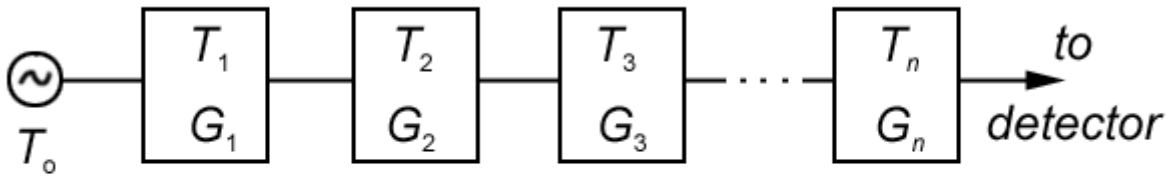
But recall that the noise temperature is

$$T = (\mathrm{NF} - 1)\, T_o = (L - 1)\, T_{\mathrm{Phys}}\,.$$

The meaning of this is that a 3 dB attenuator (loss factor $L = 2$) will contribute $T = T_{\mathrm{Phys}}$ to the noise temperature of the system, i.e. about 290 K, if it is put in the front end (before the first amplifier). Thus, the attenuator cuts the input signal by 3 dB (factor of 2), but *at the same time* it also introduces 290 K of noise into the system--a double whammy. But lossy components are sometimes used in low noise front ends, so how to we get away with it? To see that, we have to examine a series of 2-port devices.

Total System Temperature of a Series of 2-Port Devices

An actual system is just a linear chain of 2-port devices, as shown in the figure below:



T_i = noise temperatures

G_i = gains (some may be < 1)

where the input (i.e. from the feed) is shown as T_o, and each two-port device is labeled with its noise temperature and gain. Some of the gains may be less than one (i.e. a lossy cable or attenuator). The power output of the whole system will be

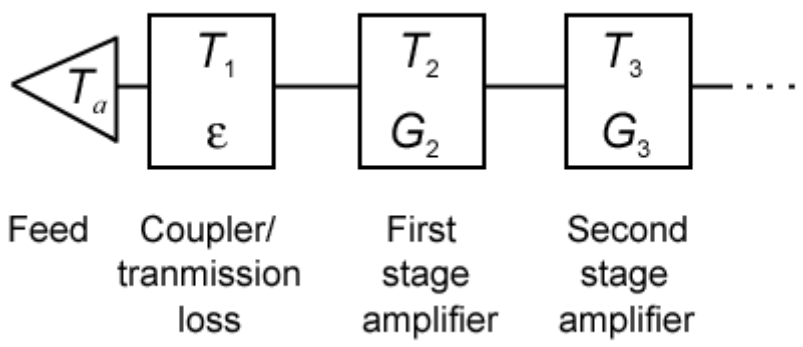
$$P = G_1G_2G_3...G_nkT_o\Delta\nu + G_1G_2G_3...G_nkT_1\Delta\nu + G_2G_3...G_nkT_2\Delta\nu + G_3...G_nkT_3\Delta\nu + ...$$

and the corresponding system temperature is

$$T = T_o + T_1 + T_2/G_1 + T_3/G_1G_2 + ... + T_n/G_1G_2...G_{n-1}\,.$$

You can see that the external temperature (the antenna temperature) just gets added to by all of the noise temperatures of the following devices, but each stage after the first stage gets divided by the total gains of the preceeding stages. This makes the first amplifier stage all-important.

Let's apply this to a real system, according to the block diagram below:



This is basically the block diagram for the system discussed at the beginning of this section. The total temperature of the system (from our introductory remarks above) is

$$T_{sys} = \underbrace{T_{bg} + T_{sky} + T_{spill} + T_{loss}}_{\text{lump into } T_a} + \underbrace{T_{cal} + T_{rx}}_{\text{lump into } T_{rx}},$$

where now we can identify the receiver temperature as

$$T_{rx} = (L-1)\,T_{\text{Phys}} + LT_2 + LT_3/G_2 + ...\ .$$

To be even more specific, say we have a system in which the cabling and coupler loss is 0.4 dB, => L = 1.10. Further, say the noise figure of the first amplifier is 2.5 dB =>T_2 = (NF – 1) T_{Phys} = (1.778 – 1) 290 K = 225 K, with gain G_2= 25 dB = 316. Finally, say the noise figure of the second amplifier is 8 dB => T_2 = (NF – 1) T_{Phys} = (6.31 – 1) 290 K = 1539 K. Plugging in these numbers (which are similar to the actual numbers for OVSA):

$$T_{rx} = (L-1)\,T_{\text{Phys}} + LT_2 + LT_3/G_2 = (1.10-1)\,290\text{ K} + (1.10)\,225\text{ K} + (1.10)1539\text{ K}/316$$
$$= 29\text{ K} + 247.5\text{ K} + 5.3\text{ K} = 282\text{ K}$$

Notice that the noise temperature of the first stage is all important, and as long as it has a lot of gain, following noise temperatures are not so important. Notice also that the loss term ahead of the first amplifier contributes to every following stage, not just the first, so if it is not small it will dominate. For example, there are two feeds used on the OVSA antennas, one of which has an intrinsic loss of 3 dB, while the other has very little loss. For homework, you will recalculate the receiver temperature using a loss of 3.4 dB.

Note that the noise figure, and noise temperature, are defined relative to the ambient temperature. What if we cool the front end (the lossy parts and the first stage amplifier) to, say, 30 K? Then the above receiver temperature would become:

$$T_{rx} = (1.10-1)\,30\text{ K} + (1.10)\,23\text{ K} + (1.10)1539\text{ K}/316$$
$$= 3\text{ K} + 25.3\text{ K} + 5.3\text{ K} = 33.6\text{ K}$$

so for the greatest sensitivity we will want to cool the receiver.

We repeat, however, that if we are observing a bright source like the Sun, the Sun itself sets the system temperature because the receiver temperature is small compared to the antenna temperature. So it does no good to cool the receivers for a solar radiotelescope, except in so far as it is needed for calibration on cosmic sources.

Saturation

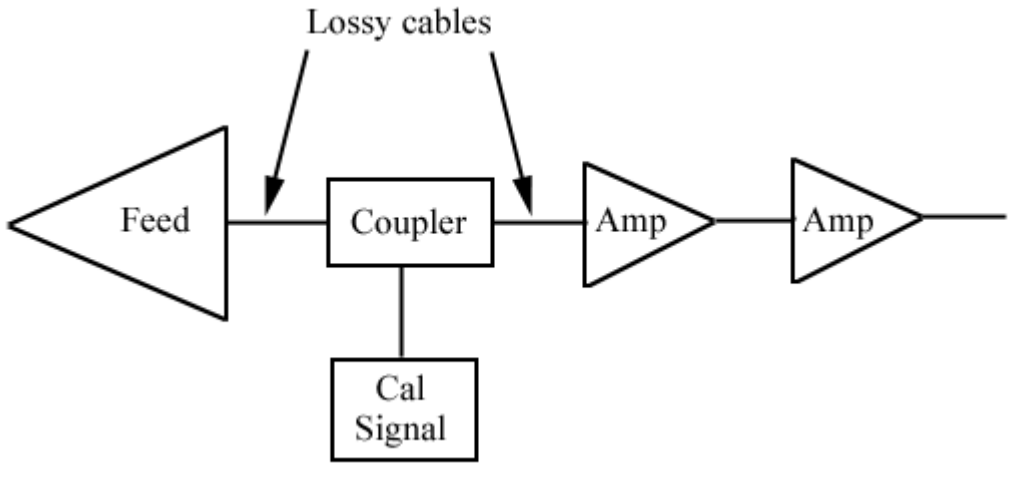
Another issue for solar work is the large and rapid changes in total flux density from the Sun. The largest flare might be of order 10^5 sfu, so let us see what this does to the signal out of the first amplifier of the OVSA system. The OVSA system is a broadband system, covering 1-18 GHz. First, what is the input power, P_a, due to a 10^5 sfu flare, observed with one of OVSA's 27 m antennas? We insert (4) into (1) to get

$$P_a = S\eta A\Delta\nu/2 ,$$

where $S = 10^5 10^{-22}$ W m^{-2} Hz^{-1} = 10^{-17} W m^{-2} Hz^{-1}, $\eta = 0.5$, $A = 572$ m^2, and $\Delta\nu = 1.7 \times 10^{10}$ Hz

$$P_a = (10^{-17})\ (0.5)\ (572)\ (1.7 \times 10^{10})/2 = 2.43 \times 10^{-5}\ \text{W} = 2.43 \times 10^{-2}\ \text{mW}.$$

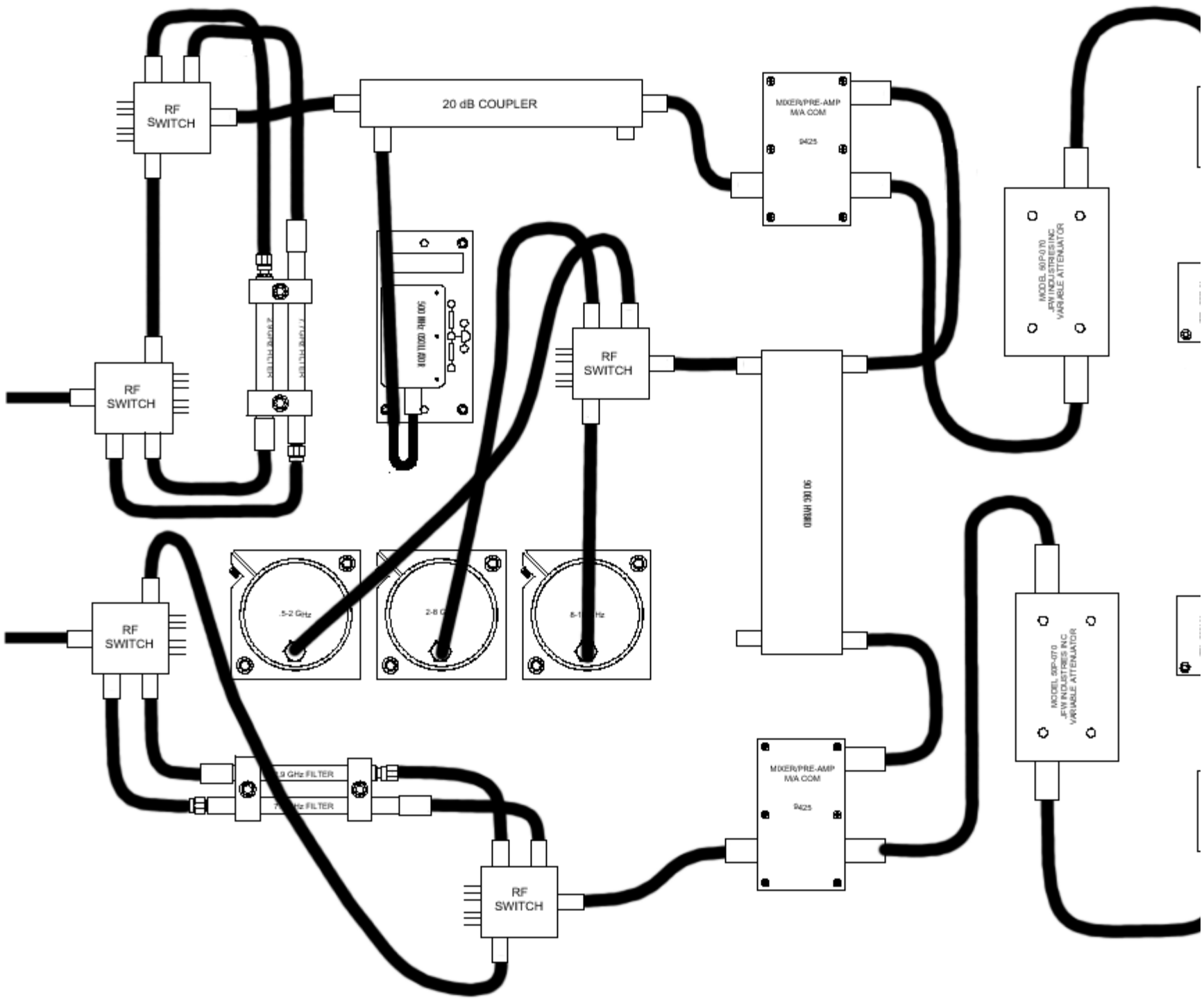
Let's convert this to dBm, and use dB everywhere to do the calculation. To get dBm, take the log of the power in mW and multiply by 10, so 2.43×10^{-2} mW becomes -16.1 dBm. The cable/coupler loss (for the linear feed) is -0.4 dB, the first stage amplifier gain is about 25 dB, so the output of the first stage amplifier is $-16.1 - 0.4 + 25$ dBm = 8.5 dBm. The first stage amplifier will have saturation problems at 10 dBm, so this is getting close to a problem with saturation. The second stage amplifier also has 25 dB of gain, but it is a "mixer/preamp" that combines the signal with a local oscillator signal and has a bandwidth of only 100 MHz. For this stage, the bandwidth is reduced by a factor of 17 GHz/100 MHz = 170, which is -22 dB, and it has a gain of 25 dB, for a net gain of 3 dB. Thus, the signal out of the mixer/preamp is 11.5 dBm, while saturation occurs at only 5 dBm. *We conclude that the second stage will saturate badly on large flares!* How do we solve this? We simply put a 10 dB attenuator between the first and second stage amplifiers. How much affect will this have on the receiver temperature?



−16.1 dBm	−0.4 dB	25 dB	3 dB	Net gain of each stage
−16.1 dBm	−16.5 dBm	8.5 dBm	11.5 dBm	Output power of each stage

Receiver

To end this lecture, let us look at what happens to the signal after it leaves the front end and enters the receiver. The figure below is the layout of the SRSP (Solar Radio Spectropolarimeter) receiver built by NJIT for Lucent Technologies.



This is a dual-channel receiver, with RCP and LCP signals being received separately. The signals enter on the left and exit on the right. The switches and filters on the left are to condition the signal to get around a problem with the mixer/preamps, which are just right of the center. The three circular devices are local oscillators, and three are required to cover the range 1-18 GHz. The frequency ranges are 1-2 GHz, 2-8 GHz, and 8-18 GHz. The operation will be explained in class. We will discuss more about receivers and following electronics when we get to correlators.

Physics 728 Radio Astronomy: Lecture #6 Prof. Dale E. Gary
NJIT

Fourier Synthesis Imaging

Interferometer Response to Plane Waves

The basic unit of an interferometer array is the single baseline between two antennas, as shown in Figure 1, below. A plane wave incident from directly overhead will arrive at the antennas in phase, and if the signal travels the same distance through cables to the correlator (shown with an X in the figure, because it is just a multiplication), the signal from the two antennas will give a maximum response.

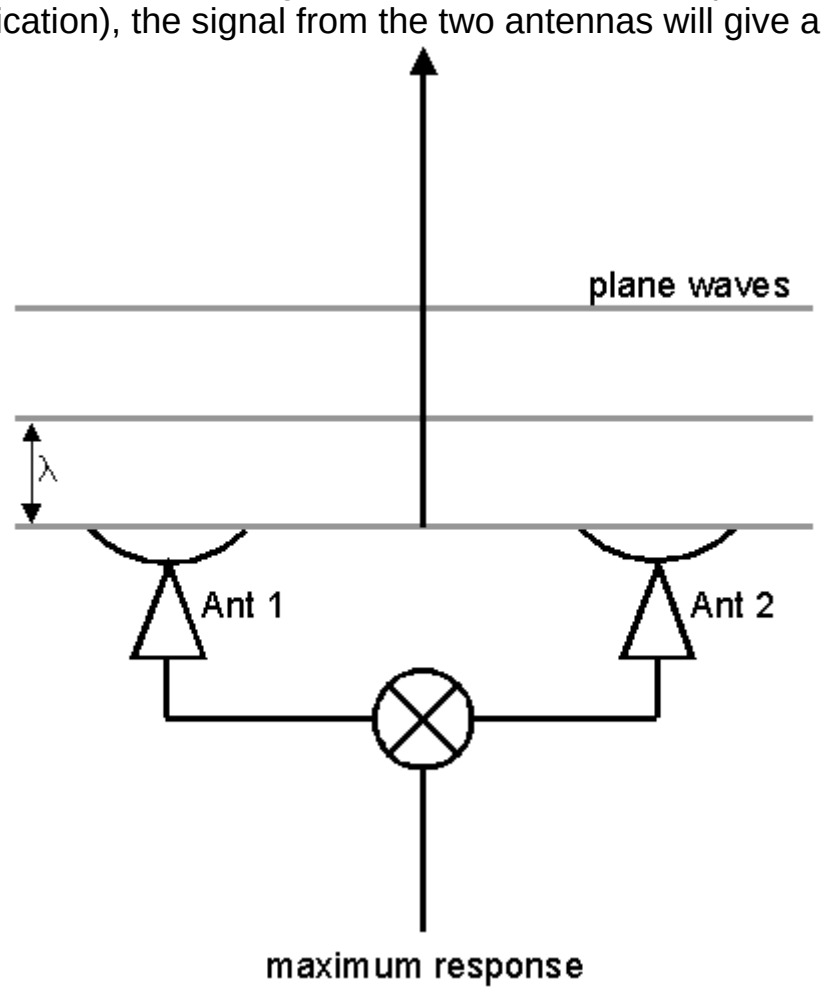


Figure 1: Geometry of an interferometer baseline, with plane waves incident vertically, of wavelength λ.

If the plane wave comes from a slightly different direction, it will arrive at the antennas slightly out of phase, and the response will be less, except that when the waves arise exactly 1 wavelength out of phase the response will again be a maximum, as shown in Figure 2.

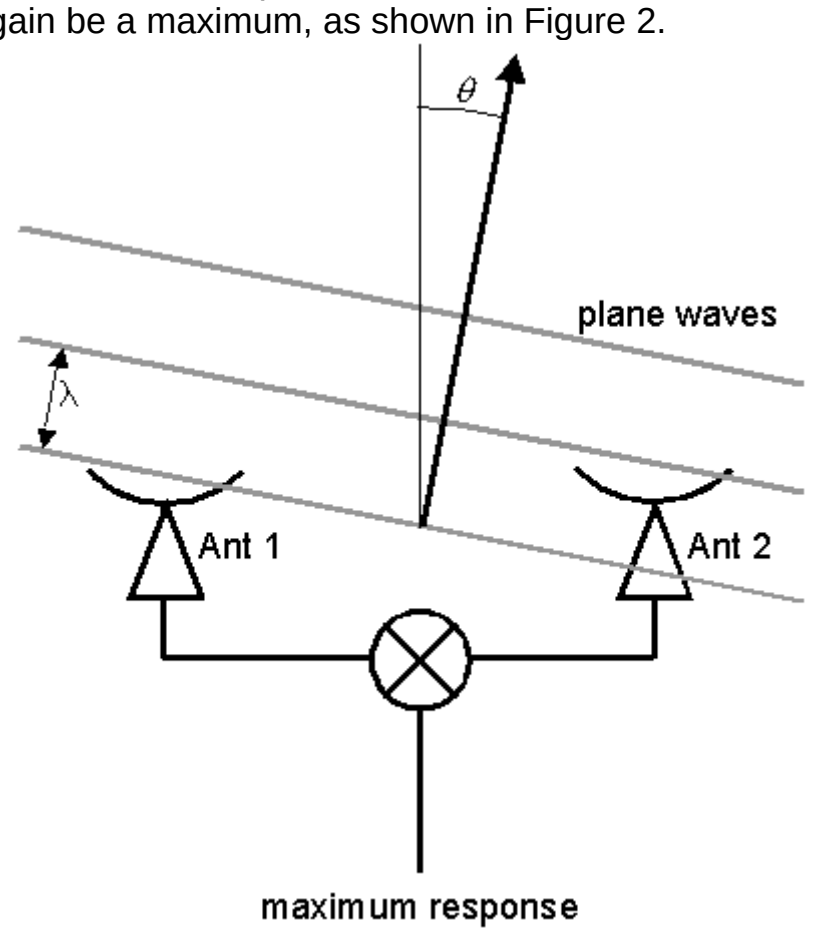
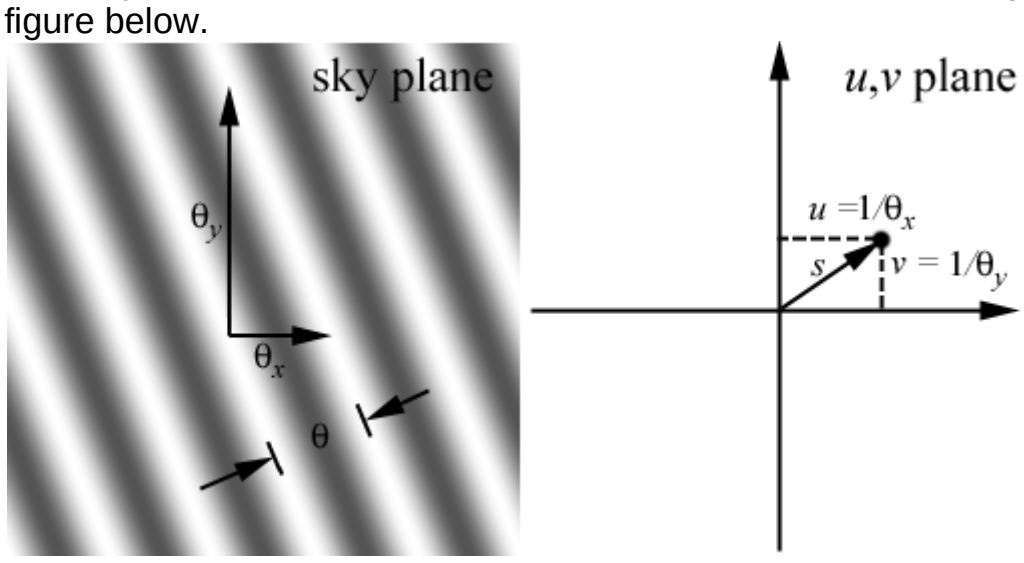


Figure 2: The same baseline as in FIgure 1, but for waves incident from an angle θ from the vertical. The waves arrive at the antennas again exactly in phase, because the angle is such that the difference in path length is λ.

The angle for which the secondary maximum occurs is

$$\theta = \sin^{-1}\lambda/B \qquad (1)$$

where B is the baseline length, or distance between antennas. The angle θ is also called the fringe spacing, as introduced in Lecture 4. Additional maxima occur at integer multiples of θ, to provide the fringe pattern across the sky. Recall from Lecture 4 that in two dimensions the fringe pattern looks like the left panel of the figure below.



If we have a truly monochromatic system, the fringes will extend all the way across the sky with the same amplitude. However, in practice our observation will be done over some bandwidth $\Delta\nu$. Since the spacing between fringes depends on wavelength, frequencies in one part of the band will have fringes with one spacing, and fringes in another part of the band will have another spacing, so that at large angles the frequencies start interfering with each other. At what value of θ will there be an extra pathlength equal to one wavelength (i.e. $n\lambda$ at one end of the band, and $(n+1)\lambda'$ at the other end)? For homework you will show that this occurs at

$$\theta = \sin^{-1}(c/B\Delta\nu)$$

From this result, for $\Delta\nu = 50$ MHz we have $\theta = 0.82^{o}$. So it is obvious that it is not possible to observe sources from anywhere except in a narrow range of angles directly overhead without addressing this problem of destructive interference across the band. The solution is to insert a delay, τ, in the signal path of one of the antennas to "steer" the ***phase center*** as shown in Figure 3. When this is done, the fringe spacing can be measured relative to this phase center direction rather than from directly overhead.

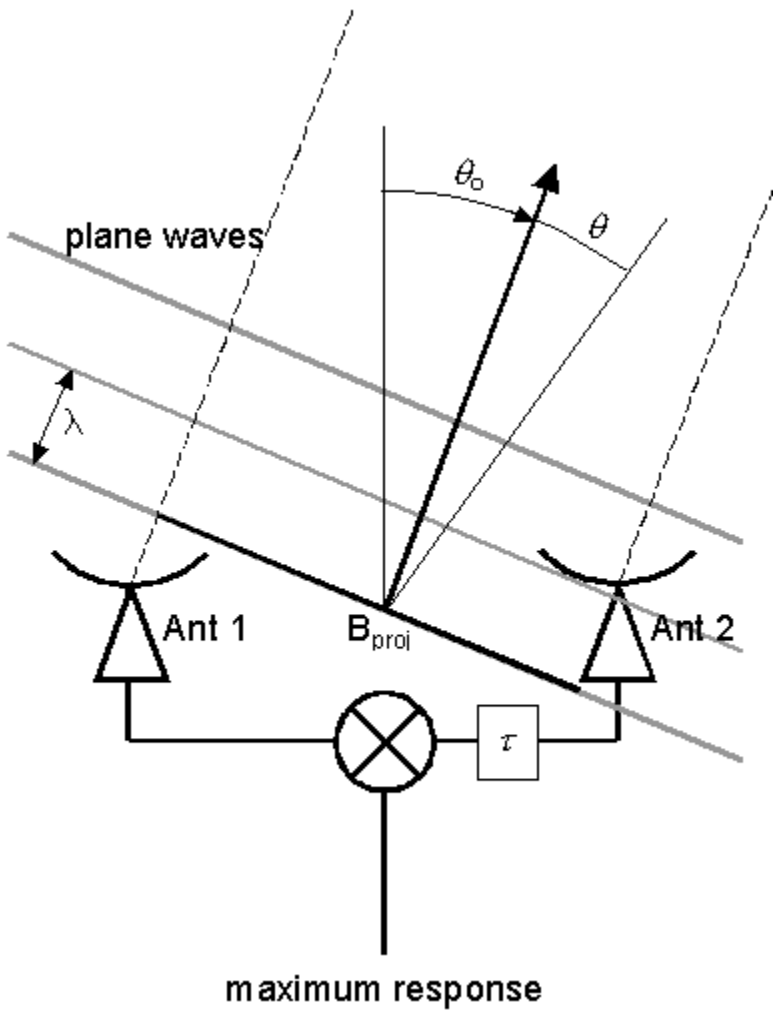


Figure 3: Geometry of an interferometer baseline where a delay τ is inserted in one antenna, in order to steer the phase center to a direction θ_o from the vertical λ.

However, our expression (1) requires modification, since now it is the projected distance between antennas that matters, not the distance along the ground. The fringe spacing is now given by

$$\theta = \sin^{-1}\lambda/B_{\text{proj}} \sim \lambda/B_{\text{proj}} \qquad (2)$$

where we make the approximation that θ is small and is expressed in radians. Note that θ is now

measured relative to the phase center direction, θ_o. The delay required to steer the phase center to the direction θ_o is called the ***tracking delay***, and must be continually updated to track the source across the sky.

The signal we have been discussing, with a maximum response at the phase center, is called the *cosine* component of the signal. It is also possible to apply a phase shift of p/2 to the signal from one of the antennas and measure this new signal, called the *sine* component. These two components are measured simultaneously. Using the notation

x = cosine component
y = sine component

we can determine the amplitude of the wavefront as

$$a = (x^2 + y^2)^{1/2}$$

and the phase of the wavefront as

$$\phi = \tan^{-1}(y/x) .$$

Interferometer Response in Terms of Sources

Point Source

Consider a point source in the sky, of amplitude a_1, located an angular distance $\theta_1 = 10$ arcsec from the phase center, as shown in Figure 4*a*. When measured with a baseline whose fringe spacing is $\theta = 30$ arcsec, the phase ϕ_1 of this source will be $\phi_1 = 2\pi\theta_1/\theta = 2\pi/3$, which is the angle shown in Figure 4*b*. The interferometer will measure a cosine component of $x = a_1 \cos(2\pi/3) = -a_1/2$, and a sine component of $y = a_1 \sin(2\pi/3) = 3^{1/2}a_1/2$. These are the components of the vector shown in the phasor representation of Figure 4*b*, and can be written more compactly as

$$a_1\, exp(i\phi_1) = a_1\, exp(i2\pi\theta_1/\theta) .$$

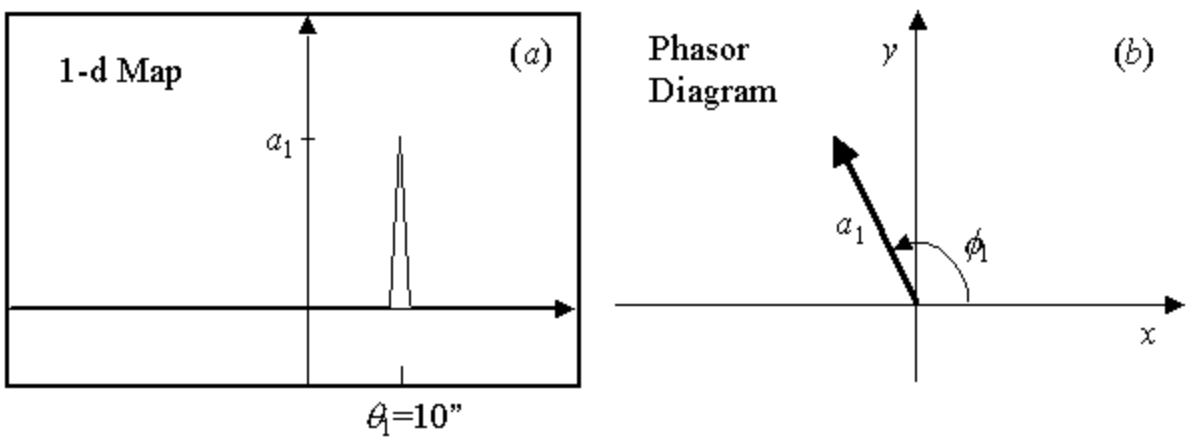


Figure 4: *a*) The spatial map of a point source of amplitude a_1 at spatial coordinate θ_1. *b*) The corresponding phasor diagram showing the interferometer response in terms of the amplitude a_1 and phase ϕ_1.

Two Point Sources

Now consider what happens when we add a second point source of amplitude a_2, located an angular distance $\theta_2 = -2$ arcsec from the phase center. This is equivalent to adding another vector to the phasor diagram, with phase $\phi_2 = 2\pi\theta_2/\theta = -2\pi/15$, as shown in Figure 5. Note that now the response of the interferometer is the *vector sum* of these two vectors, with resultant

$$a_r\, exp(i\phi_r) = a_1\, exp(i2\pi\theta_1/\theta) + a_2 exp(i2\pi\theta_2/\theta), \qquad (3)$$

so the cosine and sine components are $x = a_r \cos(\phi_r)$, and $y = a_r \sin(\phi_r)$, respectively.

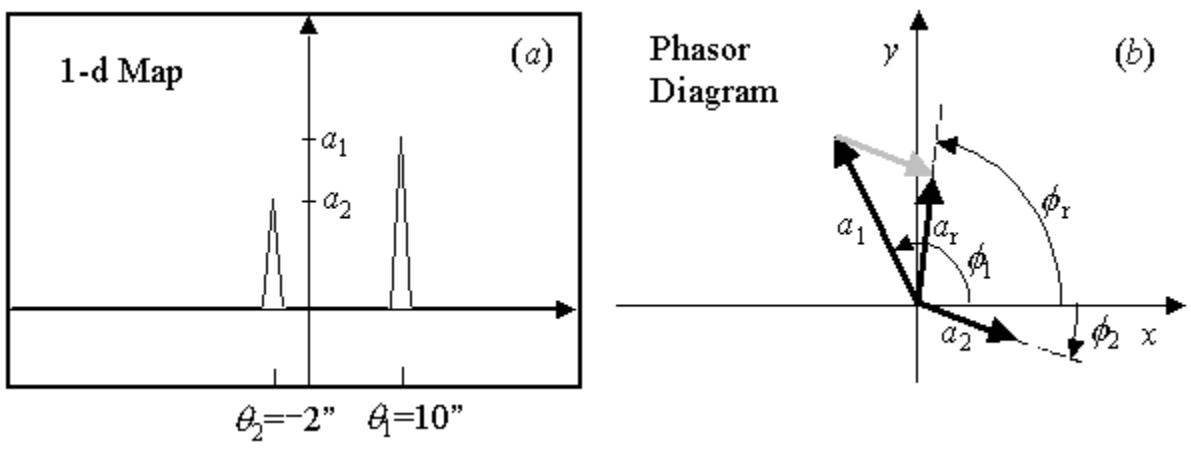


Figure 5: *a*) The spatial map of two point sources of amplitudes a_1 and a_2 at spatial coordinate θ_1=10" and θ_2=–2". *b*) The corresponding phasor diagram showing the interferometer response for the two sources, where the resultant of amplitude a_r and phase ϕ_r is the vector sum of the two individual responses.

A Continuous Distribution of Brightness

Any continuous distribution of brightness can be thought of as a collection of point sources. Consider a source of gaussian cross-section,

$$a(\theta_i) = A\, exp(-[(\theta_i - \theta_o)/\alpha]^2)\,, \qquad (4)$$

where the θ_i are the discrete spatial coordinates of size, say, $\Delta\theta = \theta_i - \theta_{i-1} = 1''$. Figure 6 shows the spatial distribution of equation (4) for a $1/e$ width $\alpha = 5''$, centered at spatial coordinate $\theta_o = 10''$.

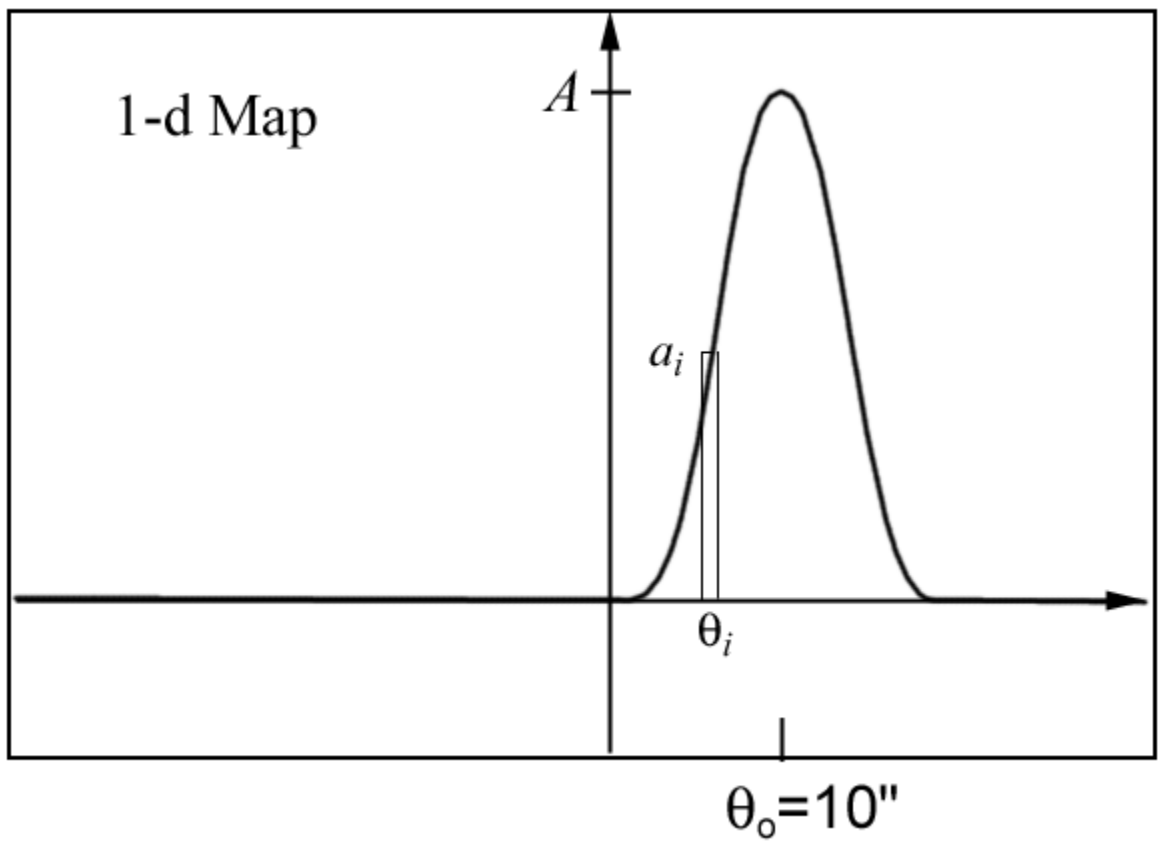


Figure 6: The spatial map of the gaussian function shown in eq. (4). The amplitude and position of one of the discrete 1" bins is shown.

By analogy with equation (3), it is obvious that the resultant signal measured by an interferometer baseline of fringe spacing θ, will be

$$a_r\, exp(i\phi_r) = \sum_i a_i exp(i2\pi\theta_i/\theta), \qquad (5)$$

where the summation is over the discrete 1" spatial bins. However, recall from Lecture 4 that the fringe spacing is the inverse of the spatial frequency, i.e. $\theta = 1/s$, and we identified coordinates u and v with the components of s, and the coordinates l and m with the sky components of angular position θ_i. Using these variables, equation (5) becomes

$$a_r\, exp(i\phi_r) = \sum_i a_i exp[i2\pi(ul + vm)]. \qquad (6)$$

The Complex Visibility and Fourier Transform Relation

Finally, we define the response of the interferometer as the ***complex visibility***, $V(u,v) = a_r exp(i\phi_r)$, define the sky brightness distribution as $I(l,m) = a_i$, and make the obvious generalization from a discrete summation to a continuous integral to obtain:

$$V(u,v) = \int I(l,m)\, exp[-i2\pi(ul + vm)]\, dl\ dm\ . \qquad (7)$$

Equation (7) demonstrates the Fourier Transform relationship between the sky brightness distribution $I(l,m)$ and the interferometer response (complex visibility) $V(u,v)$. In particular, if we measure the interferometer response $V(u,v)$, we can invert it (inverse Fourier Transform) to obtain the sky brightness distribution (the map):

$$I(l,m) = \int V(u,v)\, exp[i2\pi(ul + vm)]\, du\ dv\ . \qquad (8)$$

Obtaining u,v From An Antenna Array

A synthesis imaging radio instrument consists of a number of radio elements (radio dishes, dipoles, or other collectors of radio emission), which represent measurement points in u,v space. We now need to describe how to convert an array of dishes on the ground to a set of points in u,v space.

E, N, U coordinates to x, y, z

The first step is to determine a consistent coordinate system. Antennas are typically measured in units such as meters along the ground. We will use a right-handed coordinate system of East, North, and Up (E, N, U). These coordinates are relative to the local horizon, however, and will change depending on where we are on the spherical Earth. It is convenient in astronomy to use a coordinate system aligned with the Earth's rotational axis, for which we will use coordinates x, y, z as shown in Figure 7.

Conversion from (E, N, U) to (x, y, z) is done via a simple rotation matrix:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 & -\sin\lambda & \cos\lambda \\ 1 & 0 & 0 \\ 0 & \cos\lambda & \sin\lambda \end{bmatrix} \begin{bmatrix} E \\ N \\ U \end{bmatrix}$$

which yields the relations

$$x = -N \sin\lambda + U \cos\lambda$$
$$y = E$$
$$z = N \cos\lambda + U \sin\lambda$$

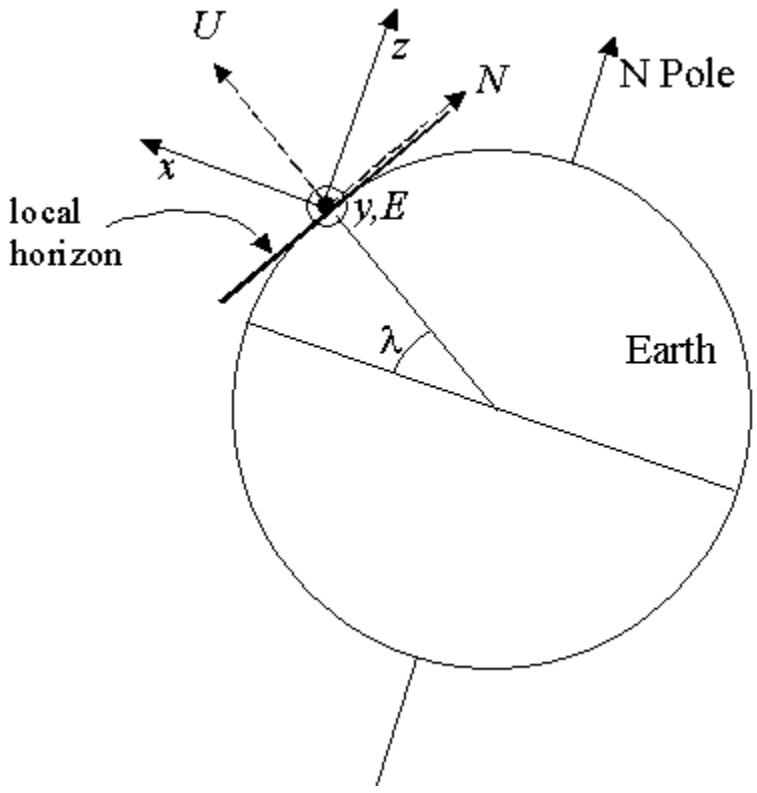


Figure 7: The relationship between E, N, U coordinates and x, y, z coordinates, for a latitude λ. The direction of z is parallel to the direction to the celestial pole. The directions y and E are the same direction.

Baselines and Spatial Frequencies

Note that the baselines are differences of coordinates, i.e. for the baseline between two antennas we have a vector

$$\boldsymbol{B} = (B_x, B_y, B_z) = (x_2 - x_1, y_2 - y_1, z_2 - z_1).$$

This vector difference in positions can point in any direction in space, but the part of the baseline that matters in calculating u,v is the component perpendicular to the direction θ_o (the phase center direction), which we called $B_{\rm proj}$ in Figure 3. Let us express the phase center direction as a unit vector $\boldsymbol{s}_o = (h_o, \delta_o)$, where h_o is the hour angle (relative to the local meridian) and δ_o is the declination (relative to the

celestial equator). Then $B_{\text{proj}} = \boldsymbol{B}.\boldsymbol{s}_o = B\cos\theta_o$.

Recall from Lecture 4 that the spatial frequencies u,v are just the distances expressed in wavelength units, so we can get the u,v coordinates from the baseline length expressed in wavelength units from the following coordinate transformation:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda}\begin{bmatrix} \sin h_o & \cos h_o & 0 \\ -\sin\delta_o\cos h_o & \sin\delta_o\sin h_o & \cos\delta_o \\ \cos\delta_o\cos h_o & -\cos\delta_o\sin h_o & \sin\delta_o \end{bmatrix}\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix}$$

Notice that we have introduced a spatial frequency w, which we must include to be accurate. However, if we limit our image to a small area of sky near the phase center (small angular coordinates l, m), then we can get away with considering only u,v coordinates. Ignoring causes distortion that is akin to projecting a section of the sky dome only a flat plane, as shown in Figure 8. The condition for this to be valid is $1/2\,(l^2 + m^2)w << 1$. For $w = 1000$ wavelengths for example, we could map out to about 1/30 radian or a little over 1 degree. As we use higher frequencies and/or longer baselines, the part of the sky we can map without distortion gets smaller. We will henceforth ignore the w coordinate and assume that we are measuring the sky in a small region near the phase center.

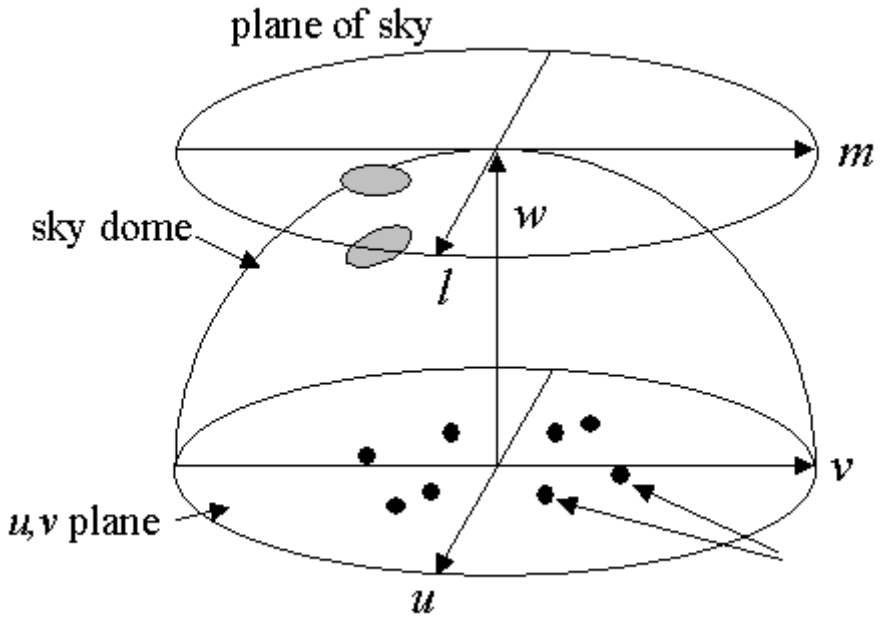


Figure 8: The sky dome and its approximation as a flat "plane of the sky" are shown. For baselines with a significant w component, the approximation of a flat sky causes distortions near the edges of the map. The gray oval on the plane of the sky is really on the curved sky dome.

Note that u,v depend on the hour angle, so as the Earth rotates and the source appears to move across the sky, the array samples different u,v at different times. Figure 9 shows an example with antenna locations (9a), corresponding u,v points at a single instant in time (9b), and the u,v points over many hours in time (9c). The u,v points trace out portions of ellipses, called u,v tracks, and sample more of the u,v plane. Making a map over a long period of time is called ***Earth Rotation Synthesis***.

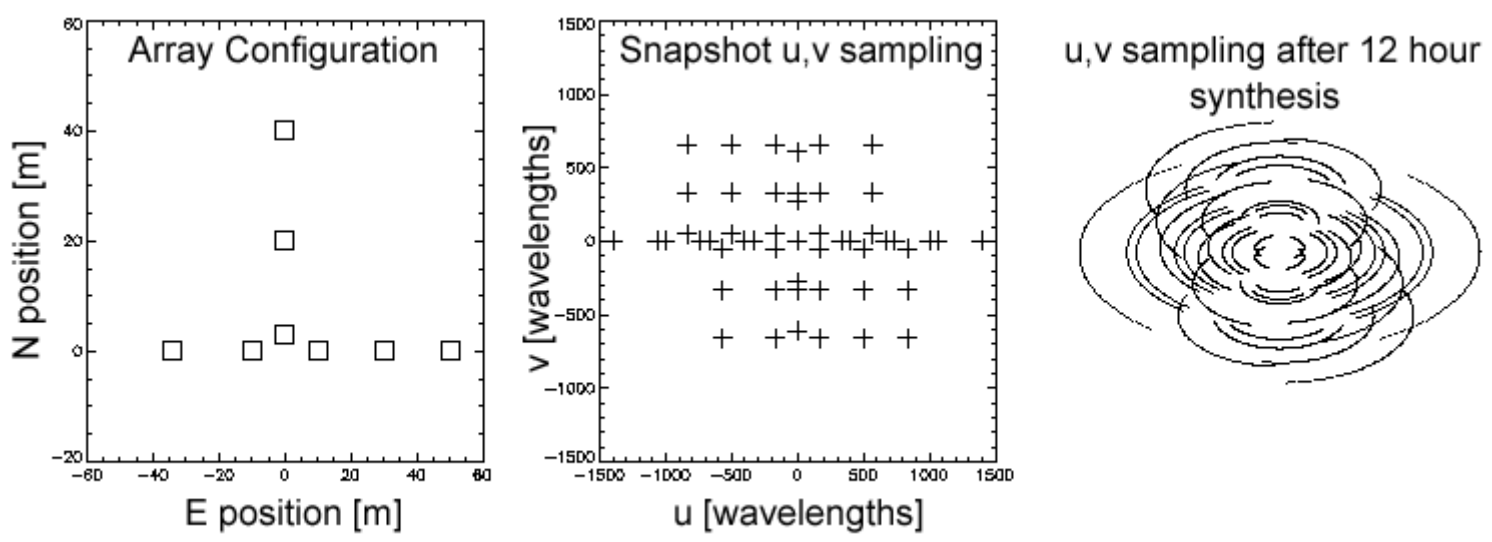


Figure 9: a) An example array configuration consisting of 8 antennas, with E and N antenna locations in meters. b) The u,v sampling that results from this array for a single time (snapshot), at zero hour angle. Each cross represents a baseline. There are N(N-1)/2 = 28 unique baselines, but there are twice this many points due to symmetry. c) The u,v sampling that results after a 12 hour integration. The rotating Earth causes the projected baselines to change with time, tracing out portions of elliptical paths.

Sampling the Visibilities

The way to think about the problem is to consider that for every sky brightness distribution $I(l,m)$ there

exists a visibility function $V(u,v)$ that is its Fourier Transform. This visibility function is everywhere a continuous complex function. An array of antennas with baselines $B_{\lambda,\mathrm{proj}} = (u_i,v_i)$ measures only certain of the set of continuous (u,v) in the visibility function, which we call the sampling function $S(u,v) = \sum_k \delta(u - u_k, v - v_k)$. Here $\delta(x,y)$ is the 2-d delta function, and the sum is taken over all baselines. The sampled visibility function $S(u,v)V(u,v)$ is the actual data provided by the array, and by performing the inverse Fourier Transform we recover the "dirty image" $I_D(l,m)$, given by

$$I_D(l,m) = \int S(u,v)V(u,v)\, exp[i2\pi(ul + vm)]\, du\ dv\ . \qquad (9)$$

Note that the dirty image uses only the sampled visibilities, and so does not contain full information about the sky brightness distribution. We must use some image reconstruction technique to recover the missing spacings. Note also that the actual visibility function measured by the instrument is $V'(u,v) = V(u,v) + N(u,v)$, where $N(u,v)$ is an unavoidable noise component.

We could use a "brute force" method and calculate the dirty image via the direct Fourier Transform, but this would require $2MN^2$ multiplications, where M is the number of samples (baselines*times, or u,v points) and N is the linear dimension of the image, in pixels. In practice, $M \sim N$ (to fill the u,v plane), so multiplications (and sin/cos evaluations) go as N^4.

Instead, we use the FFT, which requires of order $(N^2 \log_2 N)$ operations, but to use this requires ***gridding*** the u,v data, which we will discuss in more detail shortly.

Sampling Function

Recall the convolution theorem, which states that the Fourier Transform of a product of two functions is the convolution of the Fourier Transforms of the individual functions, and vice versa. If we denote the FT of a function F as $\underline{F}$, then

$$\underline{AB} = \underline{A} * \underline{B} \qquad \text{(convolution theorem)}$$
$$\underline{A * B} = \underline{A}\,\underline{B}\ .$$

where the * symbol denotes convolution. Application of the convolution theorem shows that

$$I_D = \underline{S(u,v)V(u,v)} = \underline{S(u,v)} * \underline{V(u,v)}$$

which can be interpreted as saying that the FT of the sampling function, $\underline{S(u,v)}$, is convolved with the brightness distribution $\underline{V(u,v)}$. We identify the FT of the sampling function, $\underline{S(u,v)}$, as the ***synthesized beam***, or point-spread-function, or diffraction pattern of the aperture:

$$\underline{S(u,v)} = B(l,m)$$

(not to be confused with the baselines, $B_{\lambda,\mathrm{proj}}$), which we obtain by filling in our sampled u,v points with $\mathrm{real\ part} = 1$ and $\mathrm{imaginary\ part} = 0$. This is exactly what we did with our exploration of the primary beam pattern for a single dish, if you will recall. Note that $B(l,m)$ is also the image we would obtain if we observed a unit point source at the phase center, since $\underline{\delta(l,m)} = V(u,v) = 1 + 0i$, and

$$I_D = B = \underline{S(u,v)V(u,v)} = \underline{S(u,v)} * \delta(l,m) = \underline{S(u,v)}$$

Note that this result is very general, and works just as well for optical systems, where the sampling function is called the ***point-spread-function***.

Making a Map (Inverting the Visibilities)

As we noted above, we invert the sampled visibilities to obtain the dirty image. We will do an example in class, to demonstrate the relationships shown in Figure 10. The upper row are the sky plane (l,m) representations of map, beam, and dirty map. The lower row are the corresponding u,v plane representations of visibility, sampling function, and sampled visibility. Panel (*f*) represents the actual measurements made by a radio array. Panel (*c*) represents the actual image from the array. There are standard image reconstruction techniques, which we will discuss next time, that essentially try to predict the missing visibilities in panel (*d*), to arrive at the true map (*a*).

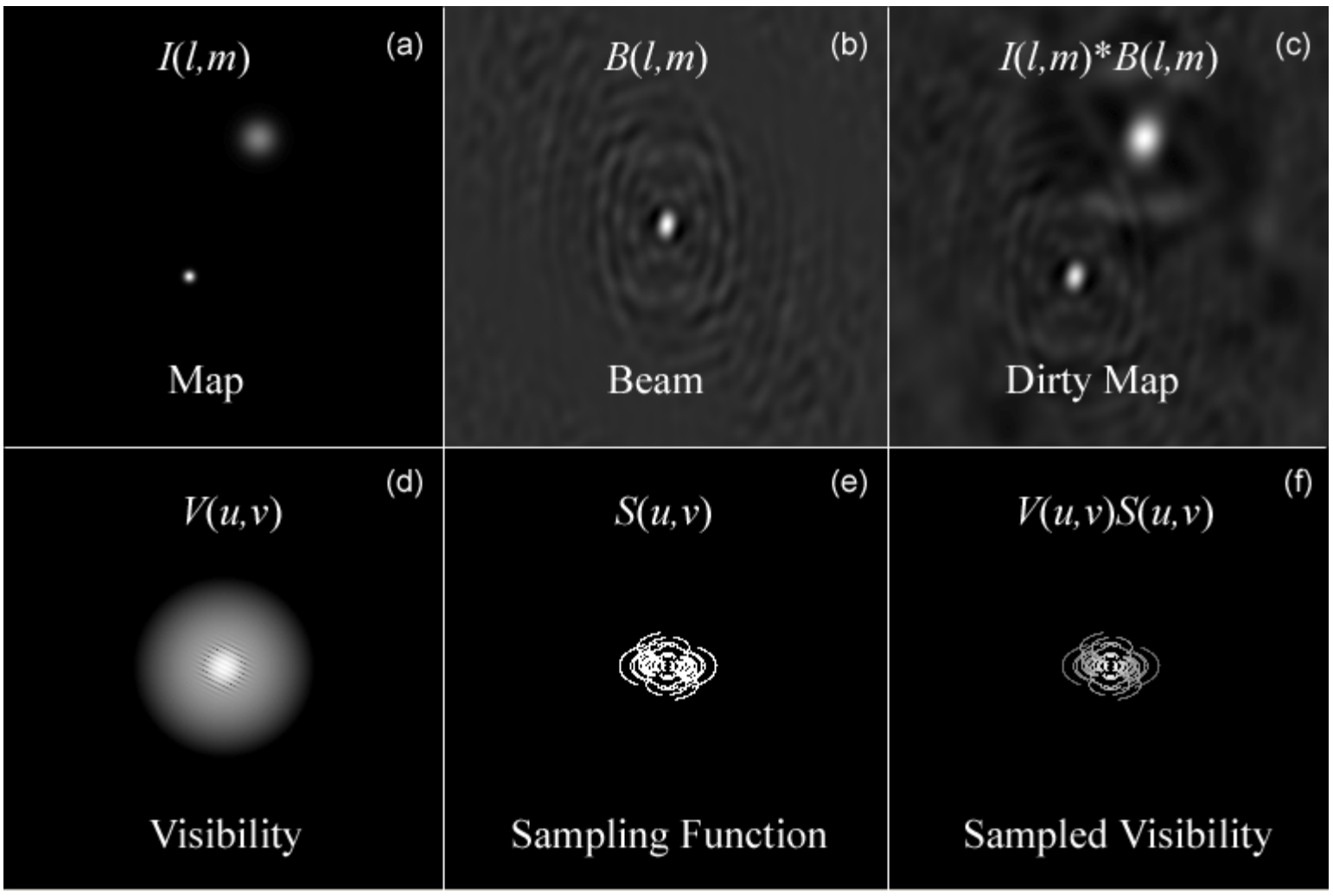


Figure 10: *a*) An example (model) sky map. *d*) The corresponding visibilities (Fourier Transform of the map). *c*) The synthesized beam, or point-spread-function, of a model antenna array. *e*) The sampling function of the array, whose Fourier Transform gives the beam in (*b*). *f*) The product of panels (*d*) and (*e*), representing the sampled visibilities. These are the actual measurements from the array. *c*) The dirty map that results from the Fourier Transform of the sampled visibilities. This is the same as the convolution of the map in (*a*) and the synthesized beam in (*b*).

Sampling Function Issues

Weighting Functions and Beam Shape

As with any image processing, it is possible and sometimes desirable to use weighting to control the point spread function. The synthesized beam may have large positive or negative sidelobes (the irregularities away from the center of Figure 10*b*) that we would like to minimize.

There are two important types of weighting commonly used in radio astronomy, called ***tapering*** and ***density weighting***. Tapering is used to suppress small-scale variations in the dirty map caused by incomplete sampling on the longest baselines (the largest u,v spacings). Often a gaussian taper

$$T(u,v) = \exp[-(u^2 + v^2)/\sigma^2]$$

is used, where σ is a measure of the width of the taper. Note that the shape of the beam is improved at the expense of spatial resolution. This is analogous to smoothing, and is also used routinely in optical image processing. Other forms of taper can be used--even "negative taper" where the outer u,v points are enhanced to improve spatial resolution.

The other important type of weighting is density weighting, in which the weighting factor is proportional to the number of u,v measurements in a given gridden cell. Two common choices are

$D_k = 1$ Natural weighting
$D_k = 1/N_S(k)$ Uniform weighting

where D_k is the weight to be applied to cell k, and $N_S(k)$ is the number of samples falling into cell k. It is typical for arrays to have a much larger number of short baselines that long ones. This is especially true of the Nobeyama array, which has a huge number of baselines of length d, and only a few larger ones:

```
. . . . . . . .   .   .       .       .       .
|--4d---|     |--4d---|       |--4d---|           Lots of 4d spacings
|---------------------23d----------------------|  Only one 23d spacing
```

This is made worse by the rate of angular change in the u,v plane when integrating over some time.

As a rule, uniform gridding gives the more pleasing result, with the full resolution of the array. However, when mapping an extended source a natural weighting gives smoother extended emission by minimizing the negative inner sidelobes. Other techniques can be used in later image reconstruction to influence the resolution and sensitivity to extended emission, which we will cover next time.

Gridding

In order to use the FFT algorithm, the measured u,v points must be gridded into the 2-d array to be inverted. As it turns out, the precise way in which this is done can greatly affect the macroscopic properties of the resulting image. The procedure can be thought of as a convolution, and we now look at some gridding convolution functions $C(u,v)$.

1. Pillbox
2. Truncated exponential
3. Truncated sinc function
4. Exponential times truncated sinc
5. Truncated spheroidal

These are truncated of necessity because they involve examining a range of FFT cells and seeing which u,v points lie within the search range--typically 6 or 8 cells. Each is truncated to a width of m cells so that

$C(u) = 0$ for $|u| > m\Delta u/2$. where Δu is the u,v plane cell spacing.

- Pillbox
  $C(u) = 1$ for $|u| < m\Delta u/2$
  0 otherwise

  This is equivalent to "nearest neighbor" if $m = 1$, and one simply sums the data in each cell. This is the fastest algorithm, but suffers the worst aliasing problems. Remember that $\underline{C}*\underline{V} = \underline{C}\,\underline{V}$, so the image $\underline{V}$ is multiplied by $\underline{C}$, which for a pillbox is just a sinc function.

- Truncated Exponential
  $C(u) = \exp[-(|u|/w\Delta u)^{\alpha}]$
  where typically $m = 6$, $w = 1$, and $\alpha = 2$ (gaussian). Then $\underline{C}$ is also a gaussian and falls off as a gaussian outside the map.

- Truncated Sinc
  $C(u) = \mathrm{sinc}(u/w\Delta u)$
  where typically $m = 6$, $w = 1$. If we use a width m = width of the map (expensive in compute time!), then $\underline{C}$ is a square wave and falls off sharply, which is the justification for this functional form.

I will not go through the others, but one can see that gridding is an important part of inverting the measurements to get the cleanest possible image. Again, gridding is forced upon us by use of the FFT. It completely disappears if we use the discrete Fourier Transform.

Complete Mathematical Representation

Define

$V(u,v)$ = actual visibility function
$V'(u,v)$ = measured (i.e. noisy) visibility function
$W(u,v) = T(u,v)\,D(u,v)$ = weighting function, including both taper and density weighting
$C(u,v)$ = gridding convolution function
$R(u,v)$ = Resampling function (bed of nails function, representing pixelization)

$$R(u,v) = \mathrm{III}(u/\Delta u, v/\Delta v) = \sum_j \sum_k \delta(j - u/\Delta u,\ k - v/\Delta v)$$

The final resampled visibility function is

$$V_R = R\,[C * (WV')]$$

and after FFT we obtain the "dirty image"

$$I_D = \underline{V}_R = \underline{R} * [\underline{C}\,(\underline{W} * \underline{V}')]$$

The convolution with $\underline{R}$ is the step that causes aliasing, so we choose a gridding convolution function $C(u,v)$ that minimizes aliasing.

Physics 728 Radio Astronomy: Lecture #7 Prof. Dale E. Gary
NJIT

Image Reconstruction

Image Restoration (Deconvolution)

Once we have the dirty image, I_D, we need to do more work to obtain a high quality image. There are two main approaches, and both are nonlinear, so are not easy to treat mathematically. The first is called ***CLEAN***, from the term used by its inventor Hogbom (1974, "Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines," *Astronomy & Astrophysics Supplement Series* **15**, 417.). The second is the Maximum Entropy Method (MEM).

Before detailing these methods, we should ask what range of spatial scales we can expect to recover in an image. The smallest source (hence related to pixel size) is given by $\Delta l = 1/2u_{max}$, $\Delta m = 1/2v_{max}$, while the largest source (hence related to the map size) is given by $N_l\Delta l = 1/u_{min}$, $N_m\Delta m = 1/v_{min}$. A rule of thumb for an appropriate map pixel size is $\Delta l = 1/3u_{max}$, $\Delta m = 1/3v_{max}$ in order to have three pixels per smallest fringe. There may be reasons to specify a larger map, for example to map out to the edge of the primary beam, and there may be other limitations to the field of view, such as due to bandwidth limitations (as in your homework problem) or violation of the flat plane of the sky assumption.

After gridding, our resampled visibility array has some cells populations and other cells empty (those that do not correspond to a u,v point measured with the array). The Fourier Transform inversion is not unique, because these unmeasured points could have *any* value without violating the data constraints. One particular solution, where all of the missing measurements are set to zero, is called the ***principle solution***, and of course this is the one that corresponds to the dirty image we have been discussing. However, this is not really a reasonable solution -- one would expect a more continuous distribution of visibilities. The goal of image restoration techniques is to find an algorithm that allows us to guess more reasonable values for the unmeasured points. *A priori* information is the key to choosing "reasonable" values -- for example, we may exclude negative flux values and attempt to maximize some measure of smoothness in the image.

CLEAN Algorithm

1. Find strength and location of brightest point in the image (or some previously defined portion of the image)
2. Subtract from the dirty image at this location the dirty beam B multiplied by peak strength and gain factor $\gamma < 1$. (Record the position and the subtracted flux.)
3. Go to (1) unless any remaining peak is below some user-specified level. What is left are the ***residuals***.
4. Convolve the accumulated point source model with an idealized CLEAN beam (i.e. add back components) usually an elliptical gaussian of the same size and shape as the inner part of the dirty beam.
5. Add the residuals from step (3) to the CLEAN image.

A number of similar algorithms have been developed, mostly to speed up the operation for certain types of imaging problems, but primarily they all do the same function.

CLEAN boxes
Step (1) can be restricted to smaller areas of the map if there is some reason to believe that the true emission is confined there. A practical example is shown with OVSA data.

Number of Iterations and Loop Gain
The algorithm subtracts some fraction of the source and sidelobes, and runs to some limit (or number of components N_{cl}). These are "knobs" to tweak that can affect the resulting map (which shows the nonuniqueness of hte solution). Using too high a gain tends to make extended, weak emission undetectable and noisy. Basically the brightness distribution is being decomposed into point sources, and the larger the gain the smaller the number of components. Using a high number of components means cleaning into the "noise" and wasting computer time. The values of gain and number of components must be chosen on a case by case basis, depending on the source and data quality.

The Special Problem of the Sun as a Source
When mapping the entire disk of the Sun, on which smaller size-scale sources are located, we have an extreme form of extended emission that would take a very long time to decompose into point sources. Thus, the usual CLEAN algorithm is not appropriate and antoher approach is needed. This is an example of a general approach in which models for the brightness distribution are used.

A good model for the solar disk is a uniform disk of appropriate size (the radius for example, is not the solar optical radius, but a slightly larger (and frequency dependent) radius of order 10" larger. One scales the disk to the appropriate flux level (perhaps as measured with another telescope), then does a

Fourier inversion of the model to obtain the expected u,v visibilities sampled with the array. One can then do a straight vector subtraction of the model visibilities, leaving residuals appropriate to a diskless Sun. One can then use normal cleaning techniques on this modified u,v database, and after making a clean map the disk model can be added back in. The best map will result when an accurate model is used.

Subtracting in the u,v Plane

As an example, I was once studying a rapidly varying flare star in the presence of a nearby radio-bright galaxy whose flux density was much stronger than the star I was trying to study. The flare star was so weak that it could not even be seen until the sidelobes of the galaxy were cleaned away.

In order to follow the rapid variations I needed to map once per minute of so, but over ~ 12 hourse this is a lot of maps to be CLEANed deeply. Worse, slight variations in cleaning could masquerade as stellar variations. What I did was to restrict the cleaning to the area around the galaxy, then used the clean components as a model of the galaxy. I then subtracted the model (the u,v visibilities of the inverse transformed model) from the u,v data, which created a u,v database containing only the star. I could then restrict my CLEAN to a small region of interest and quickly make maps of the variation. *This is all possible because of the linear nature of the Fourier Transform*. You can add and subtract in either the map plane or the u,v plane.

MEM Algorithm

The term Maximum Entropy comes from the similarity of the functional form of the "objective" function (the quantity to be maximized) to the statistical entropy in statistical mechanics. However, it is not really entropy and the analogy should not be taken too far. The objective function often used involves

$$H = -\sum_k I_k \ln I_k/M_k e \qquad I_k = \text{intensity in kth cell of image,}$$
$$M_k = \text{"default" image, or model}$$

which has two main attributes -- it produces a positive image everywhere, and the image is "smooth" in some sense. It is said that a solution that maximizes this objective function is the solution to missing u,v points that minimizes the amount of added information. The smoothness comes about by forcing a compression in pixel values, consistent with the data constraints. It is the logarithmic term that does this amplitude compression, and sometimes the objective function used involves a shortened form, $H = -\sum_k \ln I_k/M_k e$, which also works.

1. Start with a default map. This could be a flat map (a delta function in u,v space), or it could be the dirty map. This map becomes the model, M, by which the intensity in each pixel k is divided.
2. Calculate the entropy by summing the objective function above over the entire map. (In practice, algorithms minimize the "neg-entropy" $H' = -H = \sum_k I_k \ln I_k/M_k e$).
3. Calculate the χ^2 difference, over the entire map, between the model visibilities $\underline{V}(u_k,v_k)$, and the measured u,v data $V(u_k,v_k)$,

   $$\Delta = \chi^2 = \sum_k |V(u_k,v_k) - \underline{V}(u_k,v_k)|^2 \,/\, \sigma^2$$

   where σ^2 is a measure of the measurement errors.
4. Determine the objective function (the function to be maximized),

   $$r_i = \lambda H_i - \Delta_i$$

   where λ is a Langrange multiplier that affects the weighting of the entropy relative to the data, and i is the iteration number. Increasing λ makes the entropy more important and moves the solution away from the data more quickly, which may cause the algorithm to converge more quickly but can also lead to instability and divergence. Making λ smaller is safer, but can lead to a very large number of iterations before convergence.
5. To minimize the negative of the objective function, we determine the functional form for the derivative, which we want to bring closer to zero. This functional form provides an indication of whether we need to increase or decrease the model intensity I_k in each pixel in order to bring that pixel's contribution closer to the minimum.
6. We make those adjustments, which forms one iteration, and then repeat the entire process. We continue until the value of χ^2 reaches an acceptable limit.

The process then, is one of making slight changes to the brightness in each pixel of the model until the model both matches the data (the contribution of Δ in the objective function) and is sufficiently smooth (the contribution of the entropy term H). By changing the intensities in the map, we are putting non-zero values in the unmeasured u,v points in the u,v plane, but we may also be making changes to the measured points. If so, that will make χ^2 worse, but perhaps improve the entropy. The process will stop when we have filled in the unmeasured u,v in a manner that is consistent with the measured ones. The hope is that when we are done we will have a map that is fully consistent with the measured values, is everywhere as smooth as possible, and is everywhere positive.

The Receiving System for Interferometry

Introduction

We have seen what the receivers and front end (feed, amplifier, mixer) must do to sensitively detect weak radio emission. We have also looked at the way interferometry can be used to image the sky, and how pairs of antennas (baselines) measure Fourier components of the sky brightness distribution. We now will go back and look at what special aspects of the receiving system are needed to do interferometry.

Phaselock

The first part of the receiver is the heterodyning to bring the RF (radio frequency) down to a manageable intermediate frequency (IF) that we can work with it. This requires a mixer and local oscillator (frequency reference), as shown in Figure 1, below:

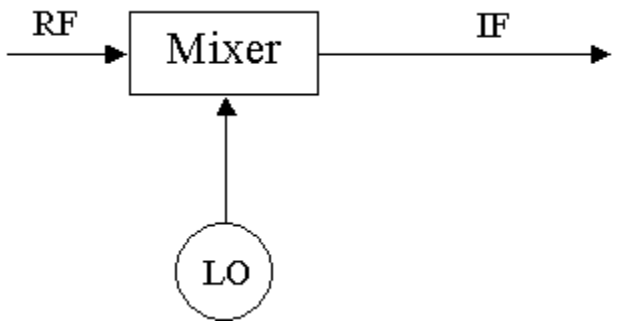


Figure 1: Heterodyne receiver, which uses a local oscillator (LO) operating at frequency wo, to tune to the desired radio frequency (RF) and mix with RF at a wide band of frequencies, and strip off a lower bandwidth section of intermediate frequency (IF) for further processing.

For interferometry, we must correlate the signals from two antennas, which requires a number of additional considerations. The main one is the ensure that the receivers of the two antennas are operating at exactly the same frequency. If one were on a frequency different by only 1 Hz, the resultant phase between the two would change by 360 degrees every second! To control the frequency of the two, we must use a phase lock system, whose block diagram might look like Figure 2.

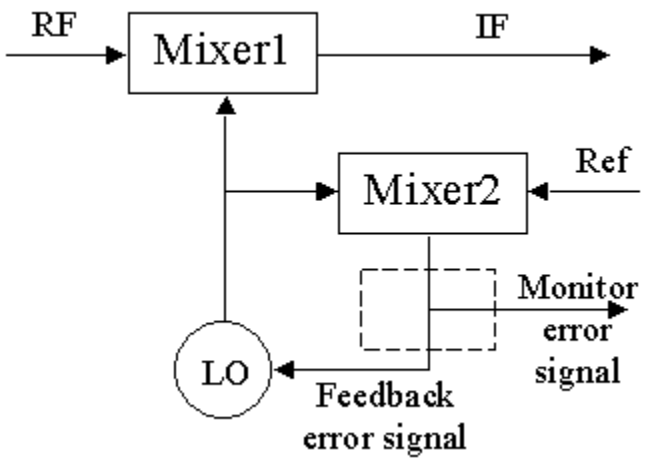


Figure 2: Adding a phase lock loop, which compares the LO output with an external reference frequency and sends an error signal back to the LO to keep it in perfect phase lock with the reference signal. Each antenna receives the Ref signal from the same source, so all receivers are locked to the same frequency.

Mixer 2 compares the LO to a frequency reference, which comes from the same frequency source for all antennas. Any error in the phase results in an error signal that is fed back to the oscillator to adjust its frequency to maintain exact frequency tuning.

Correlation

The IF signal from each receiver looks like a noise signal. Part of the waveform is really signal from the source, and part of it (perhaps the largest part) is noise. If they both look the same, how do we tell the difference? The source signal will be correlated between the two antennas, while the noise signal will not. This is illustrated with the simulated waveforms from two antennas, below:

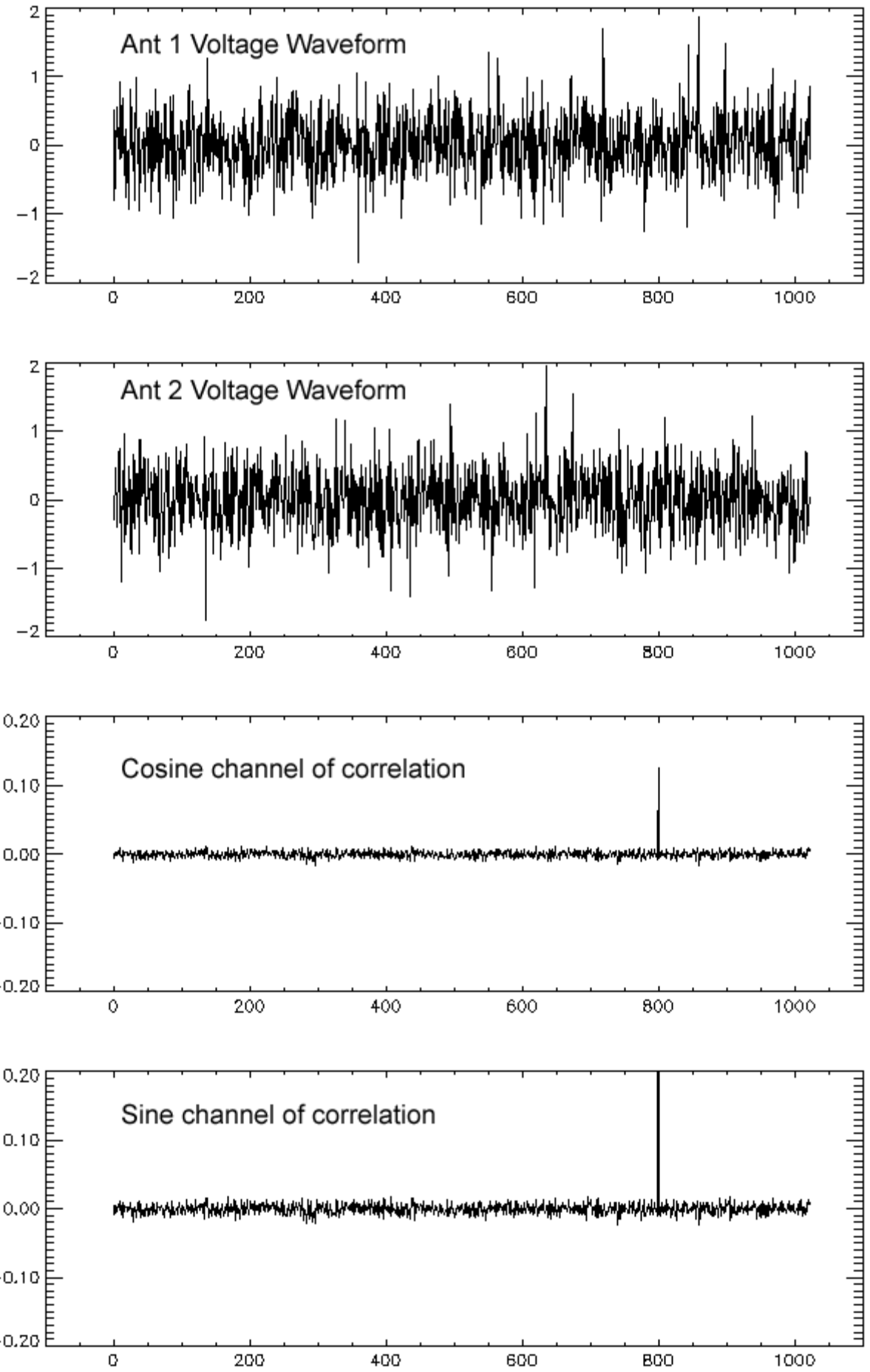


Figure 3: Two simulated voltage waveforms, with phase 30 degrees, with the waveform for antenna 1 shifted by 800 time samples. The noise level is 1/5 of the signal level in this example. The waveforms appear to have no relation to one another, but when correlated they give the plot in the third panel (cosine channel), which shows a good correlation (spike) at a time lag of 800 samples. Shifting the antenna 1 waveform by 90 degrees and performing the correlation again gives the result shown in the bottom panel (sine channel). The combination of the sine and cosine channels gives an amplitude of 0.268 and phase of 30.2 degrees. The correct values are 0.25 and 30 degrees.

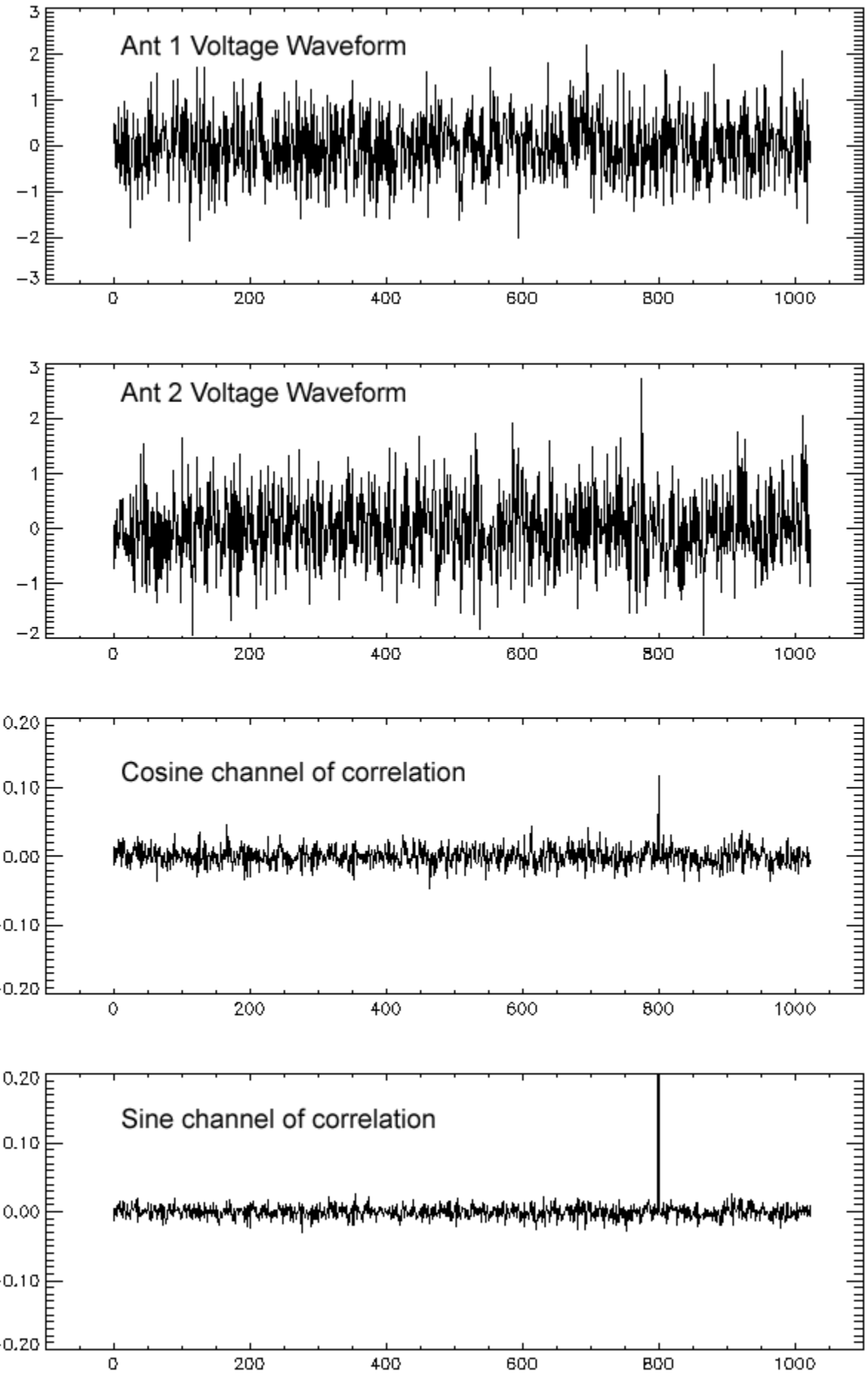


Figure 4: Two simulated voltage waveforms, with the same characteristics as for Figure 3, but now the noise level 5 times higher and is now equal to the signal level. Because the noise is uncorrelated, the correlated signal is hardly affected, and gives and amplitude of 0.245 and phase of 30.83 degrees, compared to the correct values of 0.25 and 30 degrees.

Given time varying voltages V_1 and V_2, the correlation is found by multiplying them, with one delayed by the geometrical delay $\tau_g = \boldsymbol{B} \cdot \boldsymbol{s}/c$, then averaging, i.e.

$$r = \langle V_1(t) V_2(t) \rangle$$

where $\langle\ \rangle$ denotes the expectation value, found by averaging over some integration time. Considering for the moment a monochromatic time-varying signal

$$V_1(t) = v_1 \cos[2\pi\nu(t-\tau_g)]$$
$$V_2(t) = v_2 \cos[2\pi\nu t]$$

we have

$$\begin{aligned} r &= \langle v_1 v_2 \cos[2\pi\nu(t-\tau_g)] \cos[2\pi\nu t] \rangle \\ &= v_1 v_2 \langle \cos^2(2\pi\nu t) \cos(2\pi\nu\tau_g) + \cos(2\pi\nu t) \sin(2\pi\nu t) \sin(2\pi\nu\tau_g) \rangle \\ &= v_1 v_2 \cos(2\pi\nu\tau_g) \end{aligned}$$

Since the geometrical delay τ_g changes due to the Earth's rotation, the relatively slowly varying cosine

term causes the oscillations that represent the motion of the source through the interferometer fringe pattern. In the old days of chart recorders, this fringe pattern was traced on paper and its amplitude and phase could be measured by hand. The rate of fringe oscillations is called the ***natural fringe rate***. The fringe frequency is

$$\nu_F = dw/dt = -\Omega_e u \cos\delta$$

where Ω_e is the Earth rotation rate, $w = (\boldsymbol{B}_\lambda \cdot \boldsymbol{s})$ is the spatial frequency corresponding to the projected baseline z-component given by the coordinate transformation from the previous lecture, u is the usual E, or y-component spatial frequency, and δ is the declination of the phase center. The geometry is as shown in Figure 5.

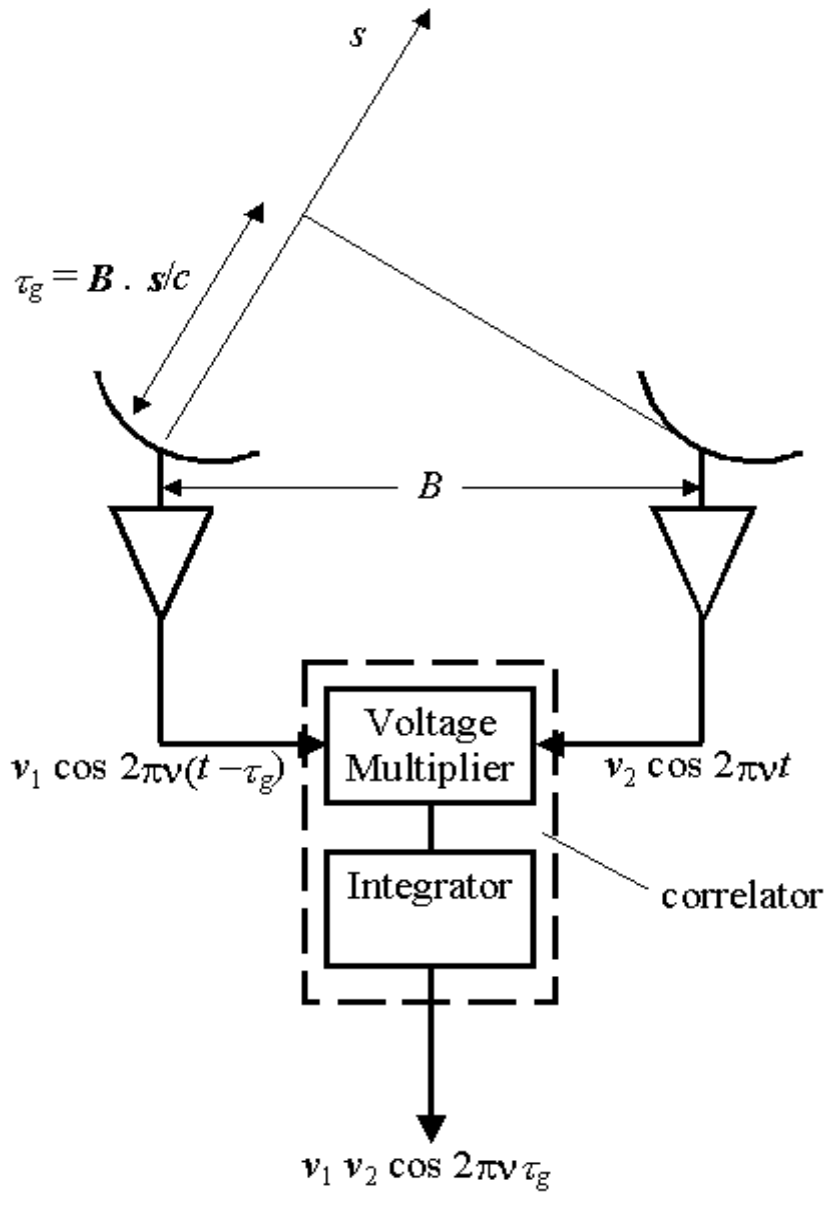


Figure 5: The geometry and block diagram leading to the measured cosine component of the correlation. Both the multiplier and the integrator are part of the device called the correlator. A refinement is shown in Figure 6.

A major refinement is to use a second correlator, and shift one of the signals by , so that both sine and cosine components are measured simultaneously, as shown below. The components in the dashed box in Figure 5 are indicated by each circled X in Figure 6.

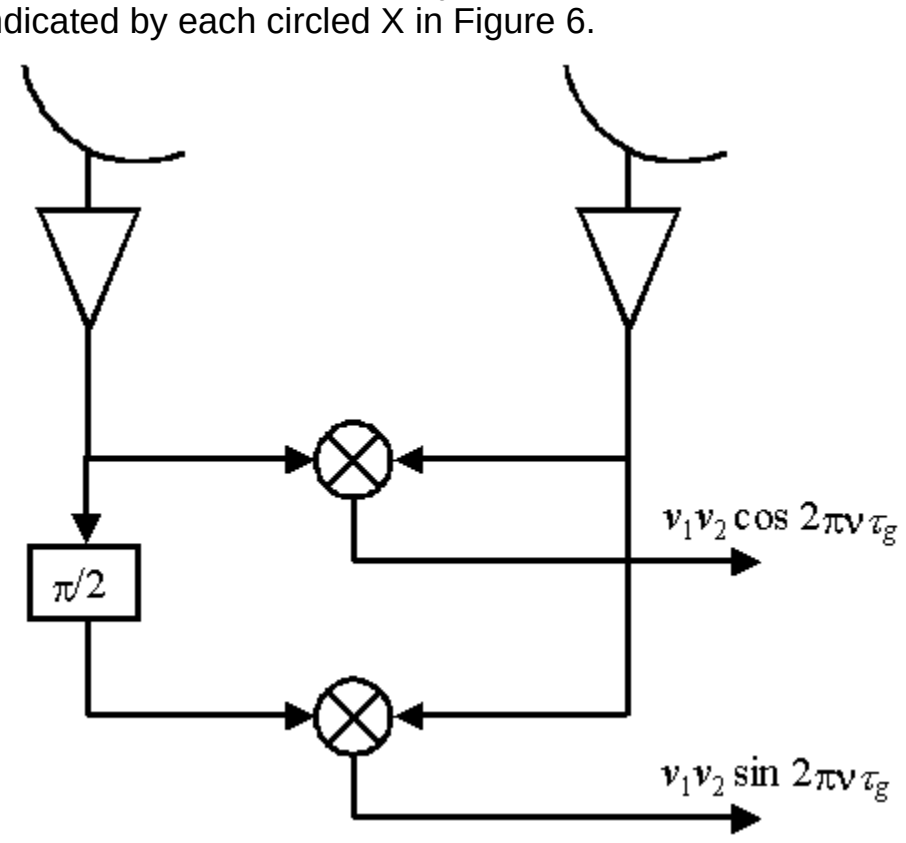


Figure 6: Inserting a phase shift of π/2 in one of the antennas and doing a second correlation allows both sine and cosine components to be measured simultaneously. These are recorded and become the complex visibility at spatial frequency u,v corresponding to the projected baseline between the

antennas.

The quantities measured out of the correlator are the real and imaginary parts of the complex visibility measured with the baseline, whose normalization is obtained by the calibration procedure, which we have not yet discussed. You may wonder how we accomplish the 90 degree phase shift over a finite IF bandwidth $\Delta\nu$. This is done in the OVSA receivers by changing the phase of the reference signal used to phase lock the local oscillator. Note that this is not equivalent to a time delay, which would shift the phase by different amounts for different frequencies, but rather it shifts the phase of each frequency separately. Some systems use the digital correlator to do the phase shifting. As it turns out, phase shifting, or synchronous detection, is also important for eliminating any DC offsets. If we multiply two waveforms with a DC offset, the offsets will give a non-zero signal even when there is no correlation in the signals. This is eliminated by periodically inverting the signal at the antenna, and then synchronously inverting the signal again at the correlator. In this way, the signals stay correlated while any unwanted DC offset gets inverted periodically and averages to zero.

To discuss correlators further, we will use the NRAO Summer School lecture on correlators.

Physics 728 Radio Astronomy: Lecture #8 Prof. Dale E. Gary
NJIT

More on Correlation

Digital Cross-Correlators

As we saw last time, the cross correlation of voltages from two antennas is

$$x_{ij} = \langle V_i V^*_j \rangle$$

where V_i is the complex voltage from antenna i, and V_j is the complex voltage from antenna j. This can be made a function of delay τ, by writing

$$x_{ij}(\tau) = \langle V_i(t) V^*_j(t+\tau) \rangle , \qquad (1)$$

that is, we delay $V_j(t)$ by a time τ and do the correlation (i.e. multiply and then integrate). Note that correlation is commonly used to find time delays in two waveforms, as seen in the figure below, from Hanaoka (1999):

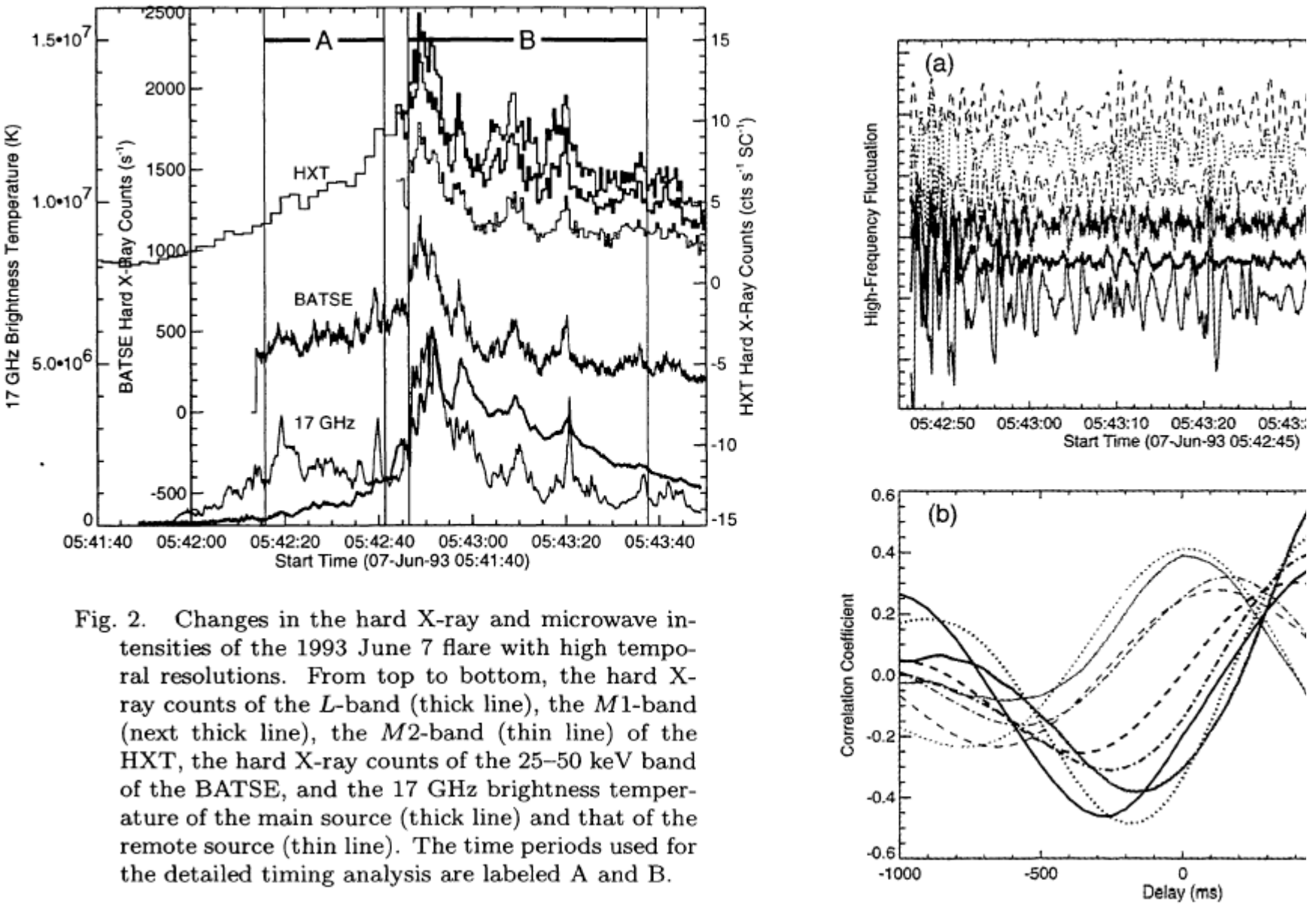


Fig. 2. Changes in the hard X-ray and microwave intensities of the 1993 June 7 flare with high temporal resolutions. From top to bottom, the hard X-ray counts of the L-band (thick line), the $M1$-band (next thick line), the $M2$-band (thin line) of the HXT, the hard X-ray counts of the 25–50 keV band of the BATSE, and the 17 GHz brightness temperature of the main source (thick line) and that of the remote source (thin line). The time periods used for the detailed timing analysis are labeled A and B.

Figure 1: Example of a cross-correlation to establish the time lag between two waveforms, in this case due to solar bursts measured in different wavelengths (hard X-rays and radio waves). The panel (b) shows the result of cross-correlation of waveforms in (a).

When $V_i(t)$ and $V_j(t)$ have a finite bandwidth, the ***delay pattern*** looks like this:

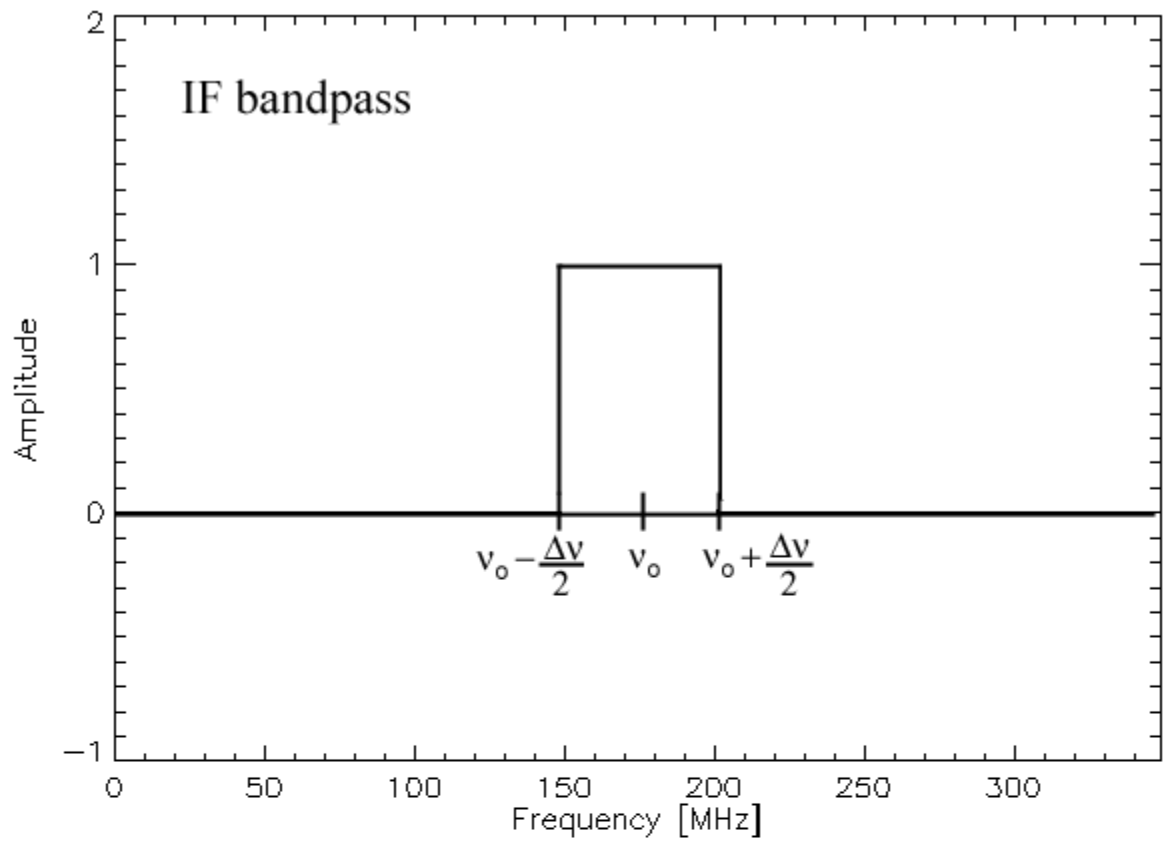


Figure 2: The relationship between the IF bandpass of each antenna and the delay pattern in the cross-correlation of signals from two antennas. The upper panel represents the spectrum of the signal $V_i(t)$ (equation 1) from one antenna, and the bottom panel represents the cross-correlation in eq. (1), $x_{ij}(\tau) = \langle V_i(t) V^*_j(t+\tau) \rangle$. The relationship between them is just the FT relation.

where the Fourier Transform of the IF bandpass gives the delay pattern. Note that the overall delay pattern is a cosine wave with period $1/\nu_o$, modulated by a sinc pattern of width $1/\Delta\nu$. This is a case where we could use what we learned in Lecture 6,

$$\underline{AB} = \underline{A} * \underline{B} \qquad \text{(convolution theorem)}$$
$$\underline{A * B} = \underline{A}\,\underline{B}\,.$$

Looking at the second equation, we can consider the square bandpass as a function A, which is a squarewave centered at $\nu = 0$, convolved with a function B, which is a delta function at $\nu = \nu_o$. The FT of this convolution is the product of the individual FTs, i.e. $\underline{A}$, which is a sinc function, times $\underline{B}$, which is a cosine function. So the convolution theorem really works!

For the Owens Valley Solar Array (OVSA), we need to determine the delay centers frequently, to ensure proper operation. This is done by observing a strong source and sweeping the delay through a range of values to find the peak of the delay pattern, which gives the optimum delay. (Will look at DLASCAN in I3542055.ARC) It is important that the delays in the system that correct for different cable lengths be known and compensated for, otherwise the amplitude of the correlation suffers considerable loss of efficiency. Once we know the delay centers (delays needed to correlate a source directly overhead), then it is easy to calculate the correct delays for any other location in the sky, i.e. the geometric delay $\tau_g = \boldsymbol{B} \cdot \boldsymbol{s}/c$. (Will look at Gelu's writeup on delay pattern, including phase.)

Last time we briefly mentioned digital correlator design, and discussed the two alternatives, FX and XF.

Let's repeat that in a little more detail. Recall that a single baseline gives a response as shown in Figure , below:

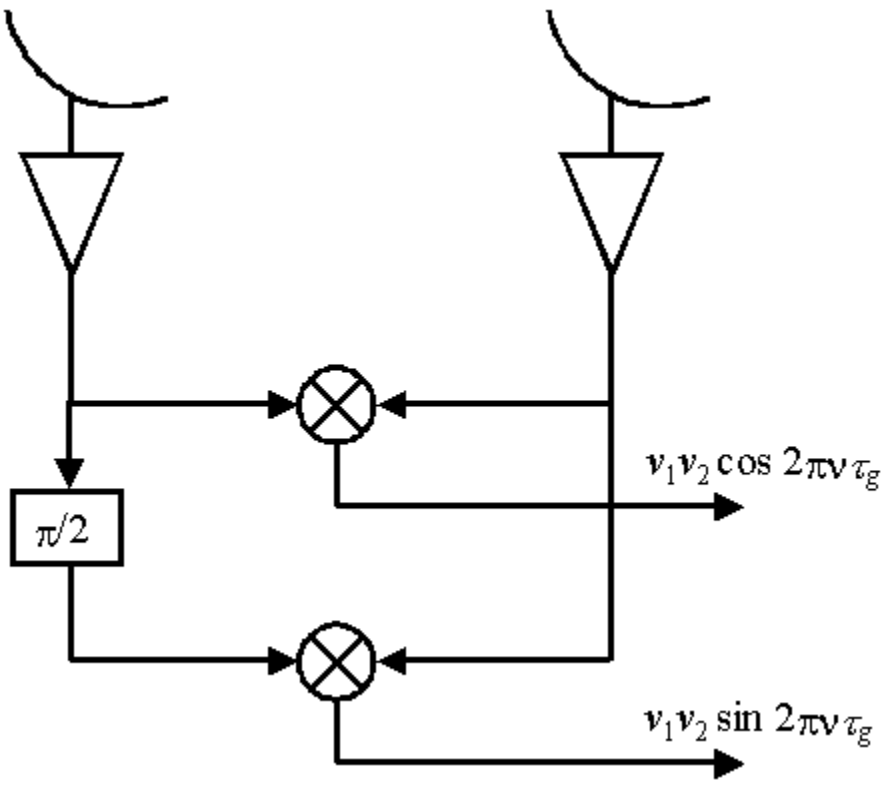


Figure 3: Inserting a phase shift of π/2 in one of the antennas and doing a second correlation allows both sine and cosine components to be measured simultaneously. These are recorded and become the complex visibility at spatial frequency u,v corresponding to the projected baseline between the antennas.

XF Correlator

We can combine the outputs into a single complex expression,

$$x_{ij}(\tau) = \langle V_i(t) V^*_j(t+\tau_{ij}) \rangle = \langle A_i(t) A^*_j(t+\tau_{ij}) \rangle \exp(i2\pi\nu_o\tau_{ij}) \qquad (2)$$

where we write the geometric delay for baseline i-j as $\tau_g = \tau_{ij}$, ν_o is the local oscillator frequency, and $A_i(t)$ is the slowly varying amplitude that, in figure 3, is written as v_i (integration represented by the <> brackets is implied in Figure 3). The exponential term in equation (2) represents a time variation that can be calculated from geometrical considerations and removed in software. Now the delay pattern in Figure 2 is obtained by slowly sweeping the geometric delay from below to above its optimum value, which takes of order 1/2 hour at OVSA. However, one can imagine delaying the signal in unit steps and doing many correlations simultaneously, to obtain the lower panel of Figure 2 instantaneously. Then one can do the inverse FT and obtain the ***cross-power spectrum***. This is the function of an XF, or lag correlator, which first does many correlations (the X part) and then does a FT (the F part). The goal is to get the visibilities for baseline i-j as a function of frequency. Note that OVSA does not do this, hence it obtains the visibility only at the optimum delay (zero lag), and is therefore called a continuum correlator. Figure 4 shows the X part of an XF correlator block diagram, which would then be followed by a hardware FFT:

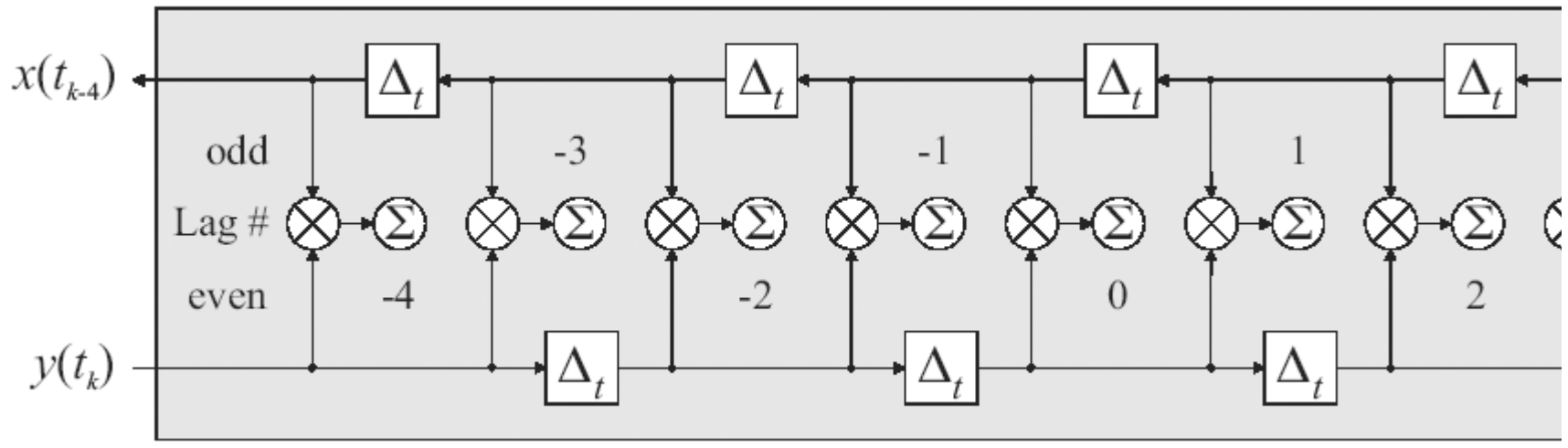


Figure 4: A lag correlator design. Note that the two antenna inputs come from opposite sides of the device, which allows for a symmetric design. The total lag between signals at each correlation is $n\Delta_t$, where the integer n is shown in each block. The S symbols denote integration. The output of each integrator is then $x_{ij}(\tau)$then, from equation (2). The FT of $x_{ij}(\tau)$ (not shown) gives the cross-correlation spectrum.

It is interesting to look at a little more detail of how such a correlator might be implemented. Figure 5, below, shows the block diagram of a larger part of the system, where the antenna signals x(t) and y(t) are digitized, then entered into "time demultiplexers," which are essentially shift registers. The data enters serially, so is at a high clock speed, but exits in parallel, so the correlator chips can operate at a lower clock speed. Each block labeled CC below, represents the entire lag correlator function shown in Figure 4, above, so they essentially subdivide the time between the correlator chip clock cycles. As an

example, say the digitizer is operating at 1000 MHz, and the time demultiplexers are 8-bit shift registers. The data come in at 1000 MHz, but go out at 1000/8 MHz = 125 MHz. Each line out of the time demultiplexer then represents a delay of 1/125 MHz = 8 nsec. The delays Δ_t in Figure 4 would then be 1 nsec each, and the correlator chips CC would give 8 lags. The total number of lags for each clock cycle of the larger block would be 64 lags.

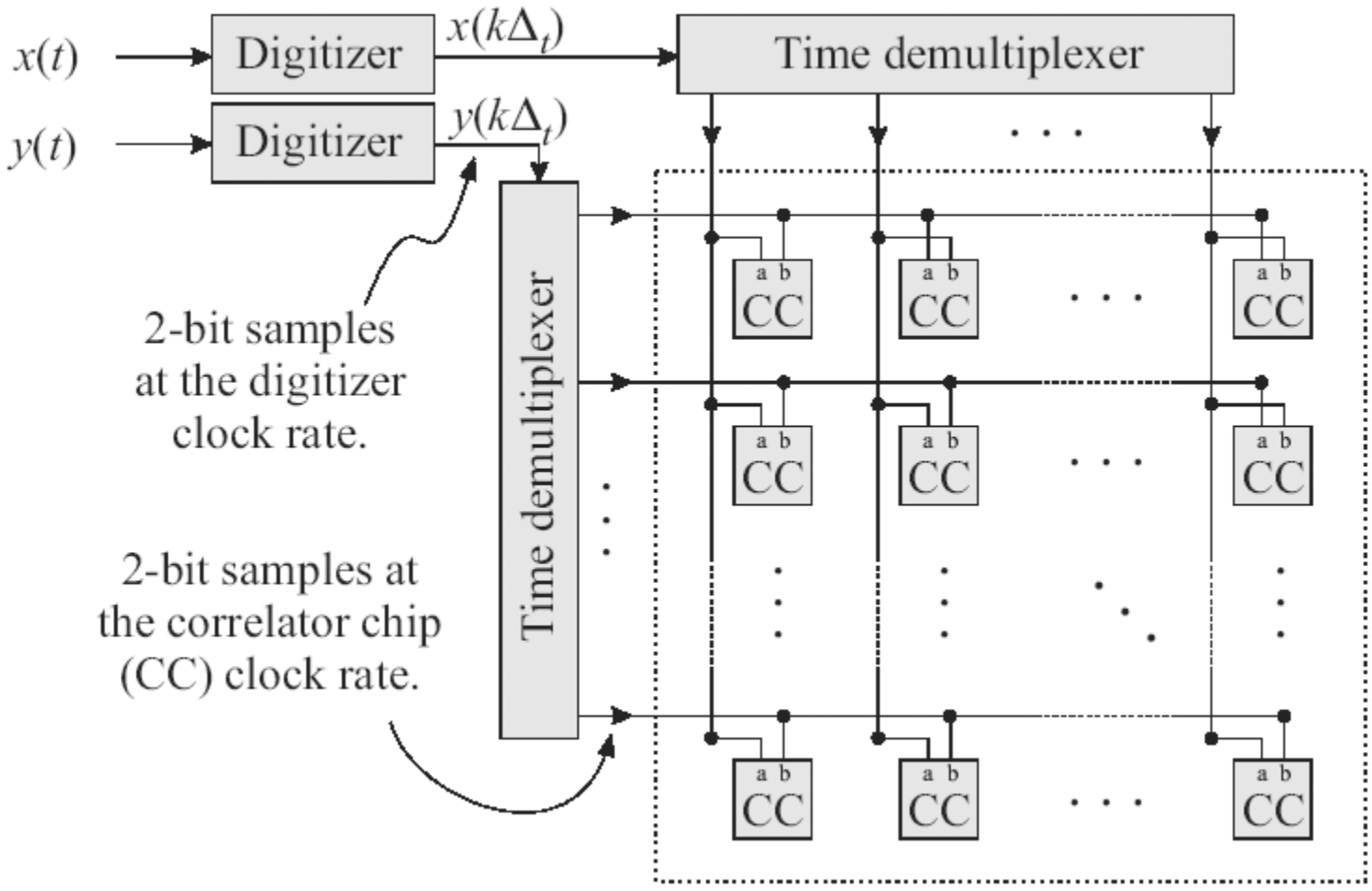


Figure 5: A larger part of the digital system, showing the digitizer step, and how the correlator chip, CC, shown in Figure 4 fits into the overall scheme. Historically, each CC block would be done on a single IC chip, designed for this special purpose and mass produced as an ASIC. The entire 8x8 area shown by the dotted line would be a printed circuit card. Note that this electronics performs the correlation for only a single baseline. For an array of N antennas, N*(N-1)/2 cards would be required. More recently, the entire block can be implemented on a single giant FPGA (Field Programmable Gate Array) chip, whose function can be programmed using the VHDL language, much like writing software.

We should also note that the lag correlator can be used to do autocorrelation, where a single from one antenna is correlated with itself (suitably lagged). After FT, this gives the auto-correlation (also called power) spectrum (and is a real quantity rather than a complex visibility).

Fractional Delay Errors

In any design, the geometrical delay must be compensated. For OVSA, this is done by switching in variable lengths of cable (called analog delay compensation), while in digital systems we can accomplish the same by using digital delays. In either case, the delays are digitized. For OVSA, the delays are digitized to 1 nsec (equivalent to about 1 foot cable sections). For the Caltech mm-array, operating at much higher frequency, the delays are digitized to 1/128 nsec! In all cases, however, the delays are digitized. Say we have a system with 1 nsec delay digitization, and that the system is set for perfect delay at the center of the sample. Then the delay will be off by a fractional time $-\varepsilon/2$ nsec at the start of the sample period and $+\varepsilon/2$ at the end of the sample period. For a wideband signal, this introduces a phase shift across the band that varies with frequency within the band, as can be seen from equation (2), i.e.

$$x_{ij}(\tau) = \langle V_i(t)V^*_j(t+\tau_{ij})\rangle = \langle A_i(t)A^*_j(t+(\varepsilon_i - \varepsilon_j))\rangle \exp(i2\pi\nu_o(\varepsilon_i - \varepsilon_j)) \qquad (2)$$

If uncompensated, the amplitude at the high frequency end of the band would be depressed. For this reason, correlators are often designed to incorporate phase rotation compensation, which is to say they shift the phase of the incoming signal both positively and negatively during the delay digitization period. This must also be digitized, and the accuracy depends on the number of bits of the correlation. For a three-level correlator, the phase rotator has three states, $+\phi$, $-\phi$, and 0.

FX Correlator

As we mentioned last time, an FX correlator does the same job as the XF correlator, but inverts the FT and correlation (X) functions, doing the frequency division first. It is perhaps simpler, conceptually. The idea is that the incoming signal is divided up into different frequencies using a filterbank, and each of these signals is then correlated with all of the signals at the same frequency among all of the other antennas. No lags are required, so the correlation job is simpler--each is essentially a continuum correlator. The phase correction due to delay errors is also a simpler job, because one simply has to shift the phase of each frequency channel in the appropriate way. The frequency division can be done with an FFT, but note that the frequency resolution of such an FFT is related to the number of samples being transformed. Fine frequency resolution requires a long time series. Even then, the Gibbs ringing phenomenon causes some problems, notably blending of signal into adjacent channels. There is a

relatively new development called polyphase filters, which has better characteristics in this regard, and can divide the incoming signal more cleanly into sub-bands.

The Frequency Agile Solar Radiotelescope may use an FX correlator design in order to solve the problem of interference. In an XF correlator, some broadband signal is cross-multiplied, and then Fourier transformed. In the presence of narrow-band interference (say some communication channel), the signal levels for the entire band may be quite high, saturating the digital electronics of the correlator. This destroys the measurement for all frequencies in the band. This can be overcome by using a many-bit correlator, but this means that every CC chip in Figure 5 has to operate on many bits, which is expensive. By using an FX correlator, one can digitize to many bits, and then channelize the input band into many sub-bands. Those channels with interference can then be removed and not correlated at all, while those without interference will have low amplitudes (relative to the interference channels), but can be sampled at much lower bit resolution (perhaps a 2-level, 1 bit comparator). The subsequence CC chips can all be 1 bit chips, making the implementation very inexpensive. The reason FASR faces this problem more than previous arrays is that it will operate over a very large bandwidth, and so will operate in parts of the radio spectrum where strong interference exists.

Spectral Line Imaging Issues

See the lecture by J. Hibbard at the NRAO Summer School. In order to do spectral line work, one must determine the individual bandpasses on each baseline (both amplitude and phase) and correct for the nonuniformities. One can then image at each point in a spectral line, and obtain a "data cube" containing two spatial and one spectral dimensions.

Physics 728

Radio Astronomy: Lecture #9

Prof. Dale E. Gary
NJIT

Calibration

Introduction

We have discussed how an interferometer array measures the complex visibilities of the source. This is a good opportunity for a review of that relationship. The basic need for calibration is to correct the measured visibilities $V'(u,v)$ to approximate as closely as possible the true visibilities $V(u,v)$.

The basic equation for a synthesis array is

$$V(u,v) = \int A(l,m)\, I(l,m)\, exp[-i2\pi(ul + vm)]\, dl\ dm\ . \qquad (1)$$

where (l,m) are the direction cosines with respect to the phase center,
(u,v) are the projected baseline coordinates in wavelengths, $u = B_{l,\lambda}$, $v = B_{m,\lambda}$,
$V(u,v)$ is the true visibility evaluated at u, v,
$A(l,m)$ is the normalized primary beam pattern (beam of a single antenna)
$I(l,m)$ is the brightness distribution of the source

Recall that the exponent comes from $\boldsymbol{B}_\lambda \cdot \boldsymbol{s} = ul + vm + wn = ul + vm + \tau_g \nu$, so when we say "with respect to the phase center," we mean that the group delay is compensated so that $\tau_g \nu = 0$.

We can rewrite this to emphasize the discrete sampling with antenna pair i,j at times t by

$$V_{ij}(t) = \int A(l,m)\, I(l,m)\, exp[-i2\pi(u_{ij}(t)l + v_{ij}(t)m)]\, dl\ dm\ . \qquad (2)$$

The term $ul + vm$ is the geometric phase difference $\Delta\phi$, produced by the geometric path length difference between antenna i and antenna j from the part of the source at location (l,m) relative to the phase center.

Recall that

$$\boldsymbol{B}_\lambda \cdot \boldsymbol{s} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} \sin h_o & \cos h_o & 0 \\ -\sin\delta_o \cos h_o & \sin\delta_o \sin h_o & \cos\delta_o \\ \cos\delta_o \cos h_o & -\cos\delta_o \sin h_o & \sin\delta_o \end{bmatrix} \begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix}$$

where the hour angle and declination (source coordinates) are h_o, δ_o, and the baseline lengths are B_x, B_y, B_z. Then the geometric phase difference at the phase center (the w term, at $l = m = 0$) is

$$\phi_g = 2\pi\, \tau_g \nu = (2\pi/\lambda)[B_x \cos\delta_o \cos h_o - B_y \cos\delta_o \sin h_o + B_z \sin\delta_o].$$

We can see what can affect the geometric phase by taking the differential of this expression:

$$\begin{aligned} d\phi_g = 2\pi\nu\, d\tau_g \ = (2\pi/\lambda)[&dB_x \cos\delta_o \cos h_o - dB_y \cos\delta_o \sin h_o + dB_z \sin\delta_o \\ &+ d\alpha_o \cos\delta_o (B_x \sin h_o + B_y \cos h_o) \\ &+ d\delta_o (-B_x \cos h_o \sin\delta_o + B_y \sin h_o \sin\delta_o + B_z \cos\delta_o)], \end{aligned} \qquad (3)$$

where we have used the relation between right ascension and hour angle: $h_o = \alpha_o - \mathrm{LST}$, so $dh_o = d\alpha_o$. Equation (2) shows how baseline errors (dB_x, dB_y, dB_z) and source position errors (α_o, δ_o) will affect the error in group delay $d\tau_g$ (or yield an error in phase $d\phi_g$). Note that a clock error is equivalent to a source position error $d\alpha_o$.

If we have a source whose position is known, we can use equation (3) to find the location of the antennas (this is called ***baseline determination***). The error in antenna position is largely independent of the baseline lengths. For example, say that we can measure $d\phi_g$ to within 1^o at 5 GHz ($\lambda = 6$ cm). Then we can measure dB_x, dB_y and dB_z to a precision of order

$$(1 / 360)\ 6\ \mathrm{cm} \sim 1 / 60\ \mathrm{cm}$$

even though $B = (B_x^2 + B_y^2 + B_z^2)^{1/2}$ = 5000 km or more (VLBI).

This has been used to verify and accurately measure motion of the Earth's crust due to plate tectonics.

Alternatively, once the baselines are accurately known, we can observe unknown sources and find their positions, (α_o, δ_o), to high accuracy. This does require long baselines, because to minimize ($d\alpha_o$, $d\delta_o$) we have to have large (B_x, B_y, B_z). So once again, VLBI is needed. Determining accurate source positions is called astrometry, and VLBI is the most accurate method for establishing a highly accurate coordinate system for the sky, of order 1 milliarcsecond in absolute position.

Complex Gains

A properly designed synthesis array will preserve a linear relationship between the measured visibilities $V'(u,v)$ and the actual visibilities $V(u,v)$, which in general can be written

$$V_{ij}(t) = G_{ij}(t)V_{ij}(t) + \varepsilon_{ij}(t) + \eta_{ij}(t) \qquad (4)$$

where we have broken the constant term into an offset $\varepsilon_{ij}(t)$ and a stochastic noise term $\eta_{ij}(t)$. The proportionality factor $G_{ij}(t)$ is called the ***complex gain*** for the ij-th baseline. All of the terms in this equation are "complex," but this is really just a convenient way of using one equation to represent the two outputs of each correlator -- the cosine and sine components (or real and imaginary).

The meaning of the complex gains can be seen by expressing them in terms of gains of individual antennas:

$$G_{ij}(t) = g_i(t)\, g_j^*(t) = a_i(t)\, a_j(t)\, \exp[i(\phi_i(t) - \phi_j(t))] \qquad (5)$$

where $a_i(t)$ is the multiplicative gain of the i-th antenna, and $\phi_i(t)$ is the phase of the i-th antenna. *When the array is optimized*, these numbers can be determined by measuring a point source at the phase center. Since the expected amplitude of such a source is constant, and the expected phase is zero, and differences in the a_i and any non-zero phase differences $\phi_i - \phi_j$ must be due to the antennas themselves. Note that for N antennas, we measure G_{ij} on $N\,(N-1)/2$ baselines, so for large N the problem is well over-determined. In this case we can use a least squares algorithm to solve for the antenna-based complex gains g_i, $i = 1, 2, \ldots N$.

However, first the array must be optimized, which requires some initial calibrations:

Antenna Pointing and Gain

Our equation for inverting the visibilities assumes that the normalized antenna primary beam distribution $A(l,m)$ is independent of time and identical for each antenna. The antenna tracking center must follow teh same intended position for all antennas (generally the same position as the phase tracking center, although this is not essential). Errors in tracking can cause reduced sensitivity and distortions in extended objects. For point sources, the tracking accuracy should be better than $0.1\theta_{\rm FWHM}$, where $\theta_{\rm FWHM}$ = angular size of FWHM of the primary beam. For extended objects (e.g. the Sun) that cover a large part of the primary beam, the pointing accuracy should be even better, say $\theta_{\rm FWHM}/20$. For OVSA at, say, 10 GHz, where the 27 m primary beam is 4.6 arcmin, this would require about 14" accuracy. For the 2 m antennas, the requirement is relaxed to about 3 arcmin.

The antenna pointing error is the difference between the actual pointing position and the desired one, and it has a complicated directional dependence due to

- misalignment of antenna rotation axes
- gravitational deformation (sagging)
- non-perpendicular axes
- atmospheric refraction
- differential heating of the structure
- wind loading

Often, rather than trying to describe axis and deformation errors based on engineering of the structure, an *ad-hoc* trigonometric dependence on the sky is determined using multiple measurements of sources around the sky. For the OVSA 27 m antennas, we point one antenna at the source and move the other in a prescribed pattern. (J069 arcfile)

For the OVSA 2 m antennas, which are not sensitive enough for this, we use the solar disk (which is nearly a point source for the 2 m primary beam). One problem is that this gives only a single declination track over a day, so it is not possible to optimize for pointing over the entire sky. (J067 arcfile)

Delay Calibration

We have already discussed delay calibration in some detail in Lecture 8, as well as the related ideas of bandpass calibration. Recall that an error in delay of only a few nanoseconds will cause the different parts of the observing band to partially interfere (different frequencies in the band arrive with different phase), leading to a drop in the amplitude. Obviously, we need to have the delays optimized.

Time and Place

The time of day and location of the antennas must be known to relatively high accuracy -- needed for determining the geometric delay. A clock error of 1 s, or a baseline error of a few cm, will cause a serious phase shift of the source over, say, 10 minutes. At OVRO, using a GPS clock and measuring baselines with cosmic source calibration (we will discuss this shortly), we get a time accuracy of << 1 ms, and baseline errors of about 3 mm. Therefore, these effects are not serious over a short time interval, but may still be problematic over 8 hours. This is one reason that we do phase calibration observations every ~ 2 hours.

Routine Corrections

Several additional corrections must be done, generally by the "on-line" computer in real time. These are:

1. System optimization
    1. level control, or gain control -- for VLA, this is done using an ALC (automatic leveling control) loop to keep the threshhold optimum on the digital correlator. At OVRO, descrete attenuation is switched in as necessary to avoid saturation (and non-linearity).
    2. Round trip phase measurement (to stabilize for phase changes due to temperature effects in the LO reference frequency).
    3. Pointing control (adjusting the antenna pointing according to the pointing parameters).
2. Timing corrections and calibrator fluxes
    1. The Earth does not rotate uniformly, due to wind, tide, and other global effects, precession and nutation. These parameters are available in terms of timing corrections from government sources.
    2. The fluxes of calibrators can change and must be monitored by specially designed radio telescopes for absolution flux calibration. We will see more about this shortly.
3. Path length changes due to the troposphere and ionosphere.
    1. The refractive index of the atmosphere differs from unity by about 1 part in 3000, so there is an additional delay in traversing the atmosphere of about 8 ns, or an "extra path length" of about 23 m at the zenith. This would not matter for a plane parallel atmosphere in which the antennas are alll at ground level, since only the relative phase matters. However, if we refer the phase to a calibrator at some other sky location (different zenith angle) we have to take account of the extra path length.

        $$\Delta L = (\Delta h + c\tau_g L \sec z / r_o) \sec z$$

        where z is the zenith angle, $L = 0.228\, P_{tot} + 6.3\, w$ is the zenith path length due to the atmosphere, P_{tot} is the total pressure at ground level, w is the vertical column water vapor content (cm), r_o is the Earth radius, and Δh is the height difference of the antennas. The second term is small, ~ 10^{-5} of baseline length, except for large z. The wet component is generally a small fraction of the dry component, but it is also variable over the sky, so $\Delta\phi$ is in fact dominated by the wet component.
    2. The ionosphere causes a frequency-dependent refraction that is proportional to λ^2, so is worse at low frequencies. The "extra" path length is

        $$L_i = -40\ \nu^{-2} N_e$$

        and is negative because the index of refraction $n < 1$ in an ionized medium. N_e is the electron column density in units of $10^{18}\mathrm{m}^{-2}$ (typically near unity for Earth's ionosphere). The differential path length is then

        $$\Delta L_i = c\tau_g L_i / (r_o \cos^2 z + 2H)$$

        where H is the height of the ionosphere (about 600 km).
4. Absorption by troposphere and ionosphere
    1. In addition to refraction, the atmosphere also absorbs radiation, which has two effects: reduces flux by $S = S_o\, exp(-\tau_a \sec z)$, increases noise (atmospheric emission) by $T_{atm}\,[1 - exp(-\tau_a \sec z)]$

        where $\tau_a = \alpha_0 + \alpha_1 P_v$ = atmospheric opacity and P_v = partial pressure of water vapor in millibars. τ_a can be measured using "tipping radiometers" which measure the sky brightness as a function of zenith angle. The water vapor line near 22 GHz is a favorite part of the spectrum to use for this purpose.

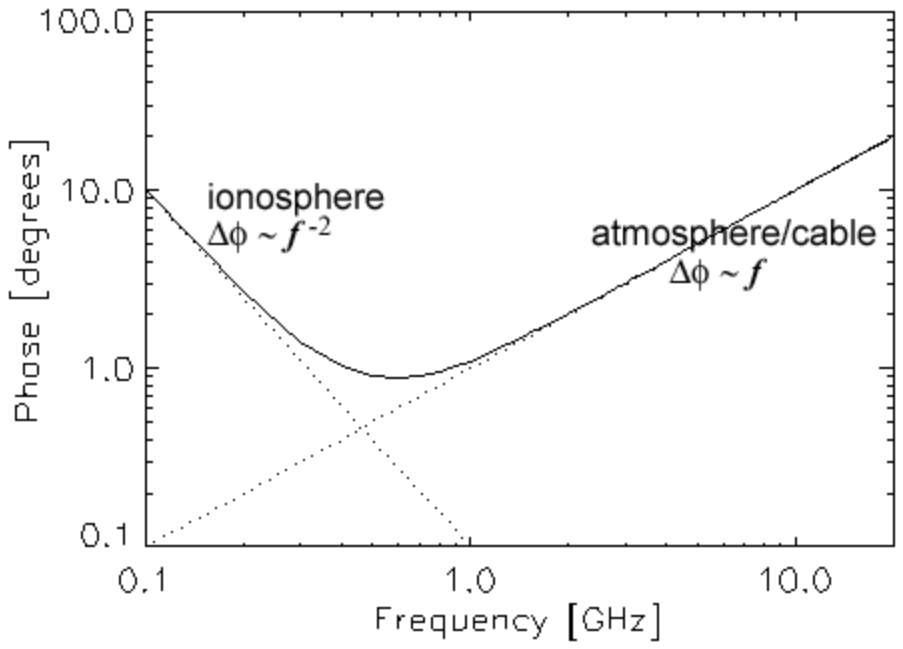


Figure 1: The dependence of phase errors due to the ionosphere and troposphere as a function of frequency. Phase errors as small as about 1 degree near 1 GHz can rise to 10-20 degrees at both high (20 GHz) and low (100 MHz) frequencies. The phase errors can be much greater than this under certain conditions.

The Offset and Noise Terms

Before getting to the final calibration, it is worthwhile to look at the unwanted offset and noise terms in equation (4). The offset term, $\varepsilon_{ij}(t)$, is generally kept small by design, and the amplitude part can be measured and subtracted by attenuating the signal. This is analogous to "dark current." The phase part is generally not a problem, since offsets on two antennas should not be correlated and it can be reduced to zero by using phase switching in the receiver.

The noise term, $\eta_{ij}(t)$, is part of the system and generally vastly ouweighs the signal, as we discussed in Lecture 5. But it can be reduced by averaging over some length of time. As we saw when discussing sensitivity, the noise is proportional to sqrt(t), where t is the integration time. Over this same period, the signal remains constant, so S/N increases.

An interesting question is, what kind of average should we perform? Often the data are represented as amplitude and phase, and one can either average amplitude and phase separately, or average as a vector average. A vector average may seem the best choice, but in many (most) cases the phase variations are uncorrelated with the amplitude variations (e.g. due to weather, temperature, or cable length changes). In this case, it is better to average amplitudes separately, and fit for any phase wind before determining phase).

Final Astrometric Calibration

We are fortunate that there are many point-like radio sources scattered around the sky and relatively well isolated from other confusing sources. When we measure such sources with an interferometer array, we can use the known visibilities (constant amplitude and zero phase) to calibrate the system. This is particularly useful because the signals traverse the entire system, including the ionosphere, troposphere, etc. It is not perfect, however, because generally the calibrator sources are in different sky locations than the source of interest -- unless the sources of interest is strong enough for ***self calibration***. We will discuss this in a moment.

As an example, let us take a look at the VLA calibrator manual. As you can see, there are lots of point sources, but they may not be point sources at all frequencies and size scales. Therefore, there are restrictions of baseline length (u,v spacing) and frequency. Note again that these sources can be variable. There is a nice database maintained by the VLA with a Java interface, to allow checking for variability.

Baseline-based Vs. Antenna-based Calibration

After suitable averaging, we are left with (for a point source)

$$G_{ij}(t) = V_{ij}(t)/S \qquad S = \text{flux of the point source}$$

where G_{ij} is the baseline-based gain (complex -- includes both amplitude and phase components). For a sufficient number of antennas, however, these are not all independent and we can do better. Let us write the baseline gain factor as

$$G_{ij}(t) = g_i(t)g_j^*(t)\, g_{ij}(t)$$

where the last term is a residual term very close to unity that basically keeps track of errors (usually < 1%). These complex numbers can be separated into amplitude and phase:

$$A_{ij}(t) = a_i(t)a_j(t)\, a_{ij}(t)$$
$$\Phi_{ij}(t) = \phi_i(t) - \phi_j(t) + \phi_{ij}(t).$$

Let us write $V_{ij}(t) = a_{ij}(t) \exp[i\phi_{ij}(t)]$; $V_{ij}(t) = A_{ij}(t) \exp[i\phi_{ij}(t)]$. Then for a point source the measured amplitudes and phases are

$$A_{ij}(t) = a_i(t)a_j(t)\, a_{ij}(t)\, S$$
$$\Phi_{ij}(t) = \phi_i(t) - \phi_j(t) + \phi_{ij}(t)$$

which can be solved for a_i and ϕ_i for all N antennas, provided g_{ij} is close to unity, using a least squares method (using logarithms for the amplitude terms). The error terms can then be checked to show if there are any problems. The errors are acceptable when

$$\phi_{ij}(t) = \Phi_{ij}(t) - \phi_i(t) - \phi_j(t) \; < 1^o$$
$$a_{ij}(t) = A_{ij}(t) / [a_i(t)a_j(t)\, S] \;\text{ within 1\% of 1.00.}$$

One advantage of antenna-based solutions is that they can be obtained even without all of the baselines. This is especially important for partially resolved or confused calibrators, as we saw with the restrictions in the VLA calibrator list. In this case one simply applies the solution on baselines that meet the criteria so that the source is a good point source. Note that this approach uses the fact that we have observed a point source calibrator. In fact, one can also do this on a more complicated source, as we will now describe in the context of ***Selfcal*** (self calibration).

Self Calibration

For our discussion of self calibration, we will use James Ulvestad's NRAO Summer School lecture notes.

Physics 728 Radio Astronomy: Lecture #10 Prof. Dale E. Gary
NJIT

Solar Radio Emission I

Introduction

We have basically covered the main issues one needs to know to understand how radio emission is produced, how it is measured with modern instrumentation, and how that instrumentation works. For the next three lectures we will look at some of the science that has been and can be done using radio astronomy techniques. We start with the Sun, since this object dominates the radio sky, and offers a laboratory to investigate the main types of radio phenomena at a high level of spatial and spectral detail.

Before launching into the radio science itself, it will be helpful to look briefly at the structure of the Sun. The Sun is a normal star of spectral type G2V, which means that it is burning hydrogen in its core, as it has been doing for the last 5 billion years, and as it will continue to do for about 5 billion years more.

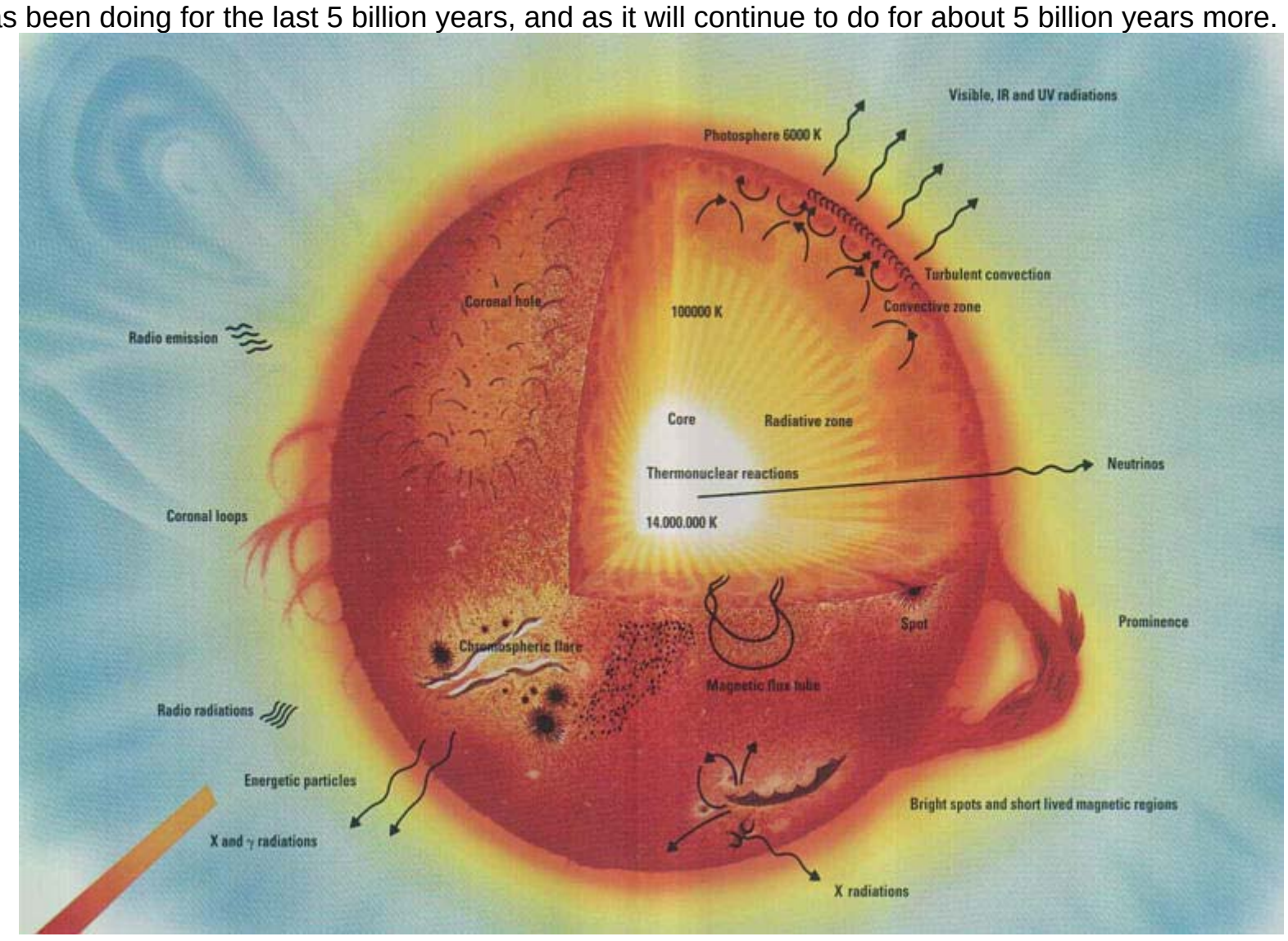


Figure 1: A schematic view of solar structure

The core temperature is about 14 million K, and the temperature falls off with distance from the core, eventually reaching the surface temperature of 5800 K. The photons generated in the core take about 1 million years to reach the surface, since they propagate outward in a random walk with a very short mean free path. Their point of last scattering is in the photosphere, after which they are finally free to stream out into space. Because they are in thermal equilibrium with the photosphere, they have a pure blackbody spectrum corresponding to the 5800 K temperature, but en route to space they encounter the more tenuous gas of the other layers of the solar atmosphere--the temperature minimum region, the chromosphere and the corona--so the solar spectrum shows many spectral lines. The fact that *these lines are mostly absorption lines* tells us that the temperature falls to yet lower values above the photosphere, to about 4500 K in the temperature minimum region. This is fully expected, but what is surprising is that above this height the temperature rises again, and in fact rises very steeply at about 2000 km above the photosphere to form a very hot (several million K), very tenuous plasma that we call the corona. The figure below shows the temperature profile starting just below the photosphere, through the temperature minimum region, the chromosphere, and the abrupt increase leading to the corona.

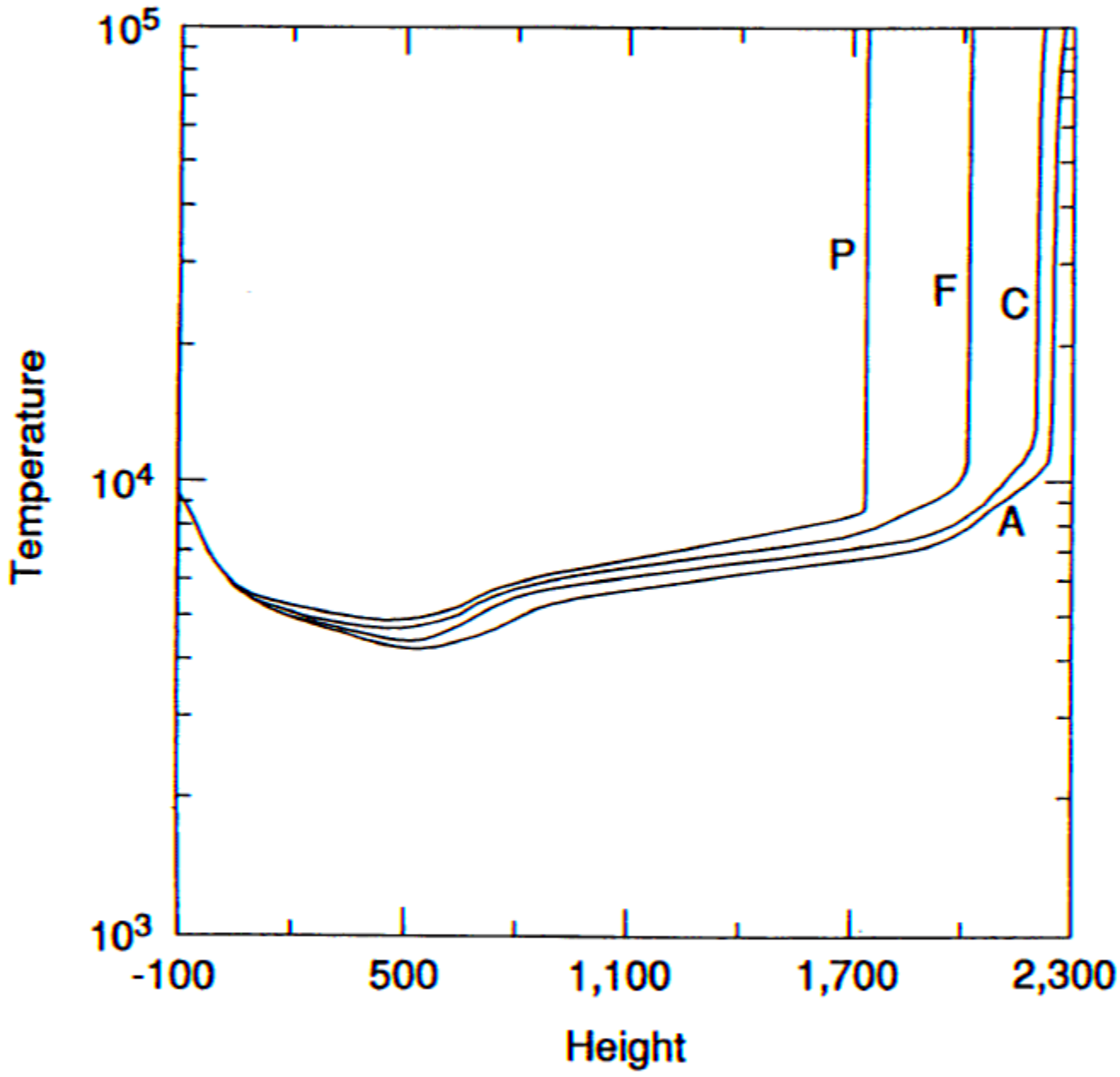


FIG. 3.—Temperature structure of our models A, C, F, and P. The height is measured in kilometers from the level; the temperature is in kelvins.

Figure 2: From Fontenla, Avrett & Loeser (1993)

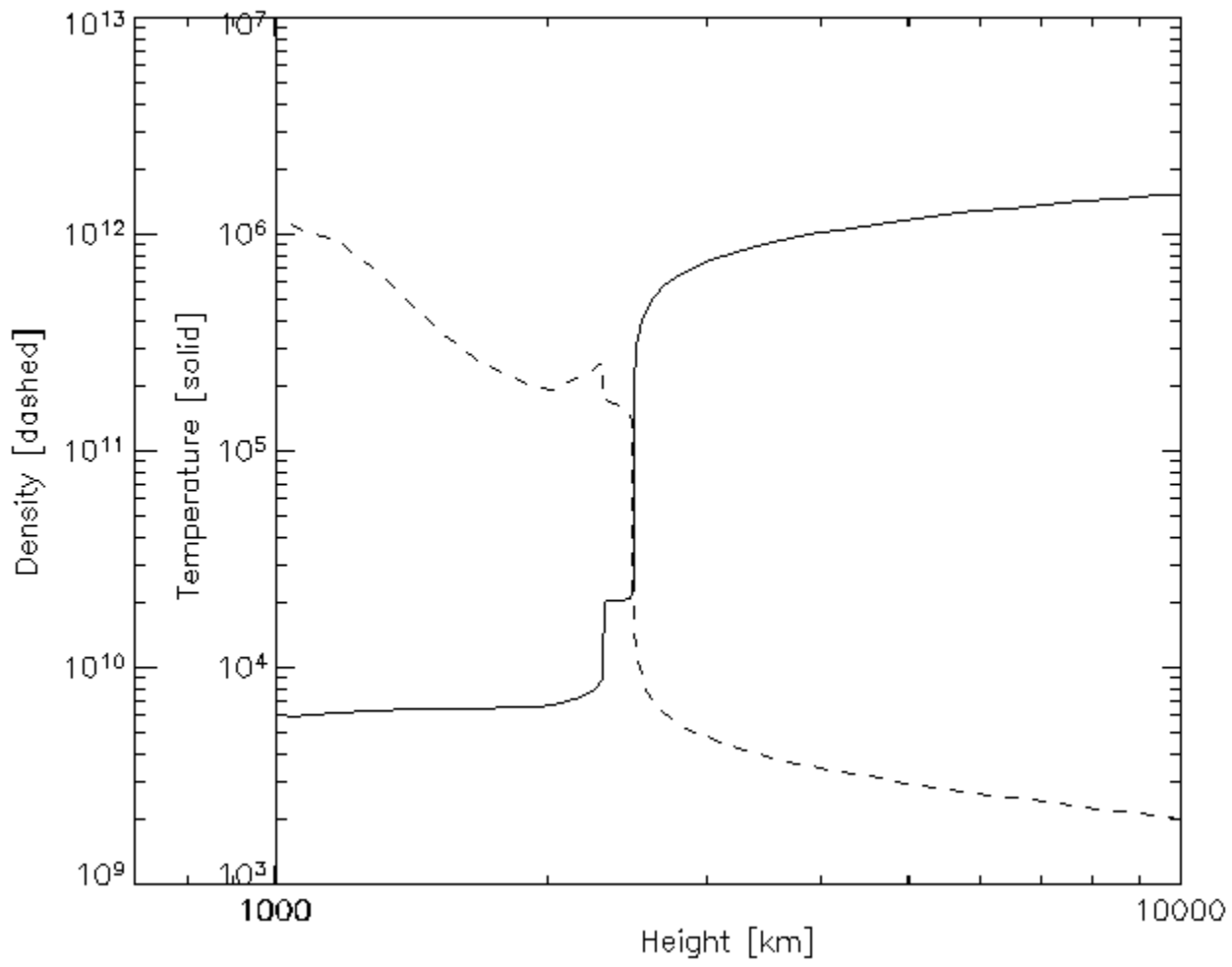


Figure 3: A similar plot to that of Figure 2, showing the temperature and density (for an older model showing the now discredited "Lyman α Plateau").

The appearance of the "quiet time" solar atmosphere at radio wavelengths is governed by the plasma parameters (temperature, density, and magnetic field strength) and the radiation mechanisms that generate the radio emission (free-free emission, gyroresonance emission, and plasma emission). The following figure shows the height versus frequency of three characteristic frequencies that we have already met--the plasma frequency, the gyrofrequency, and the frequency at which free-free emission reaches optical depth unity.

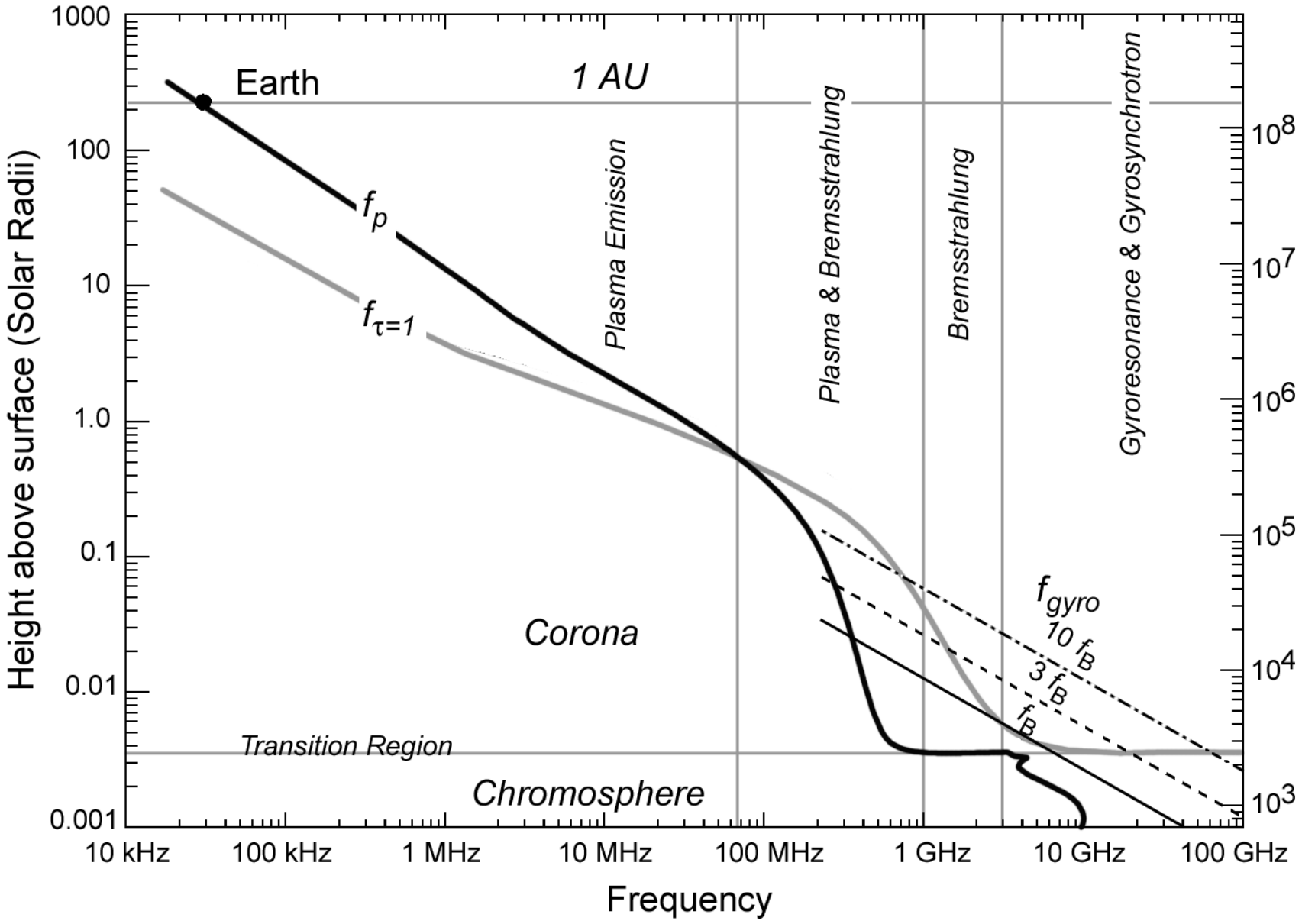


Figure 4: The highest curve in this plot specifies the type of emission mechanism that will dominate at different frequencies in the solar atmosphere. The curves are based on the dependence of different emission mechanisms on the plasma parameters of temperature, density, and magnetic field strength. The plot covers 7 orders of magnitude in frequency, and in height in the solar atmosphere.

Here is an animation showing the different appearance of the full Sun as a function of frequency. In reality it is only a set of four images, taken at the highest available resolution, with the images morphed from one image to another to give the impression of a large set of frequencies.

The Chromosphere

Because of the huge increase in electron density at the chromosphere, radio emission becomes optically thick due to free-free emission at heights higher than the solar minimum region, even at the highest frequencies. Therefore, radio observations pertain only to the upper chromosphere and higher. Let us start at the chromosphere and move outward in the solar atmosphere. As Figure 4 shows, this is also equivalent to starting at the highest frequencies and moving to lower frequencies. The figure below shows the Sun at very high (sub-mm) frequency.

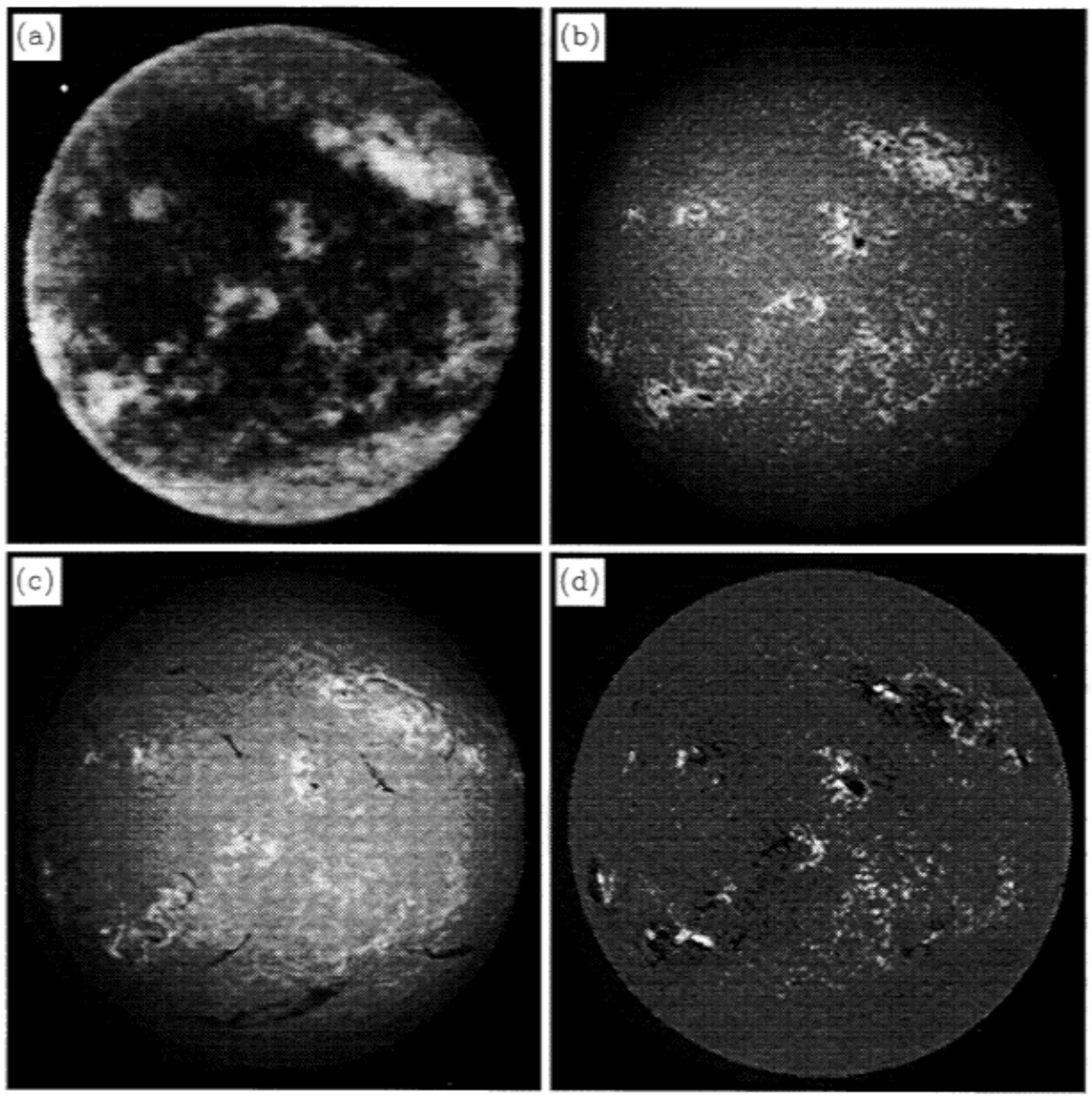


Figure 5: (*a*) The Sun at 850 microns wavelength (>300 GHz), showing the low contrast of features in the low chromosphere. The brightest features have a brightness temperature of about 6500 K, and the darkest features have a brightness temperature of about 5500 K. The other panels show (*b*) Ca II, K-line image, (*c*) H-alpha, and (*d*) a magnetogram, for comparison. From Bastian, Ewell, & Zirin (1993).

The figure below shows how the height of the radio Sun at 3 mm wavelength extends well beyond the visible edge (photosphere), and matches quite well with the tops of the *spicules*. This is from Belkora, Hurford, Gary and Woody (1992).

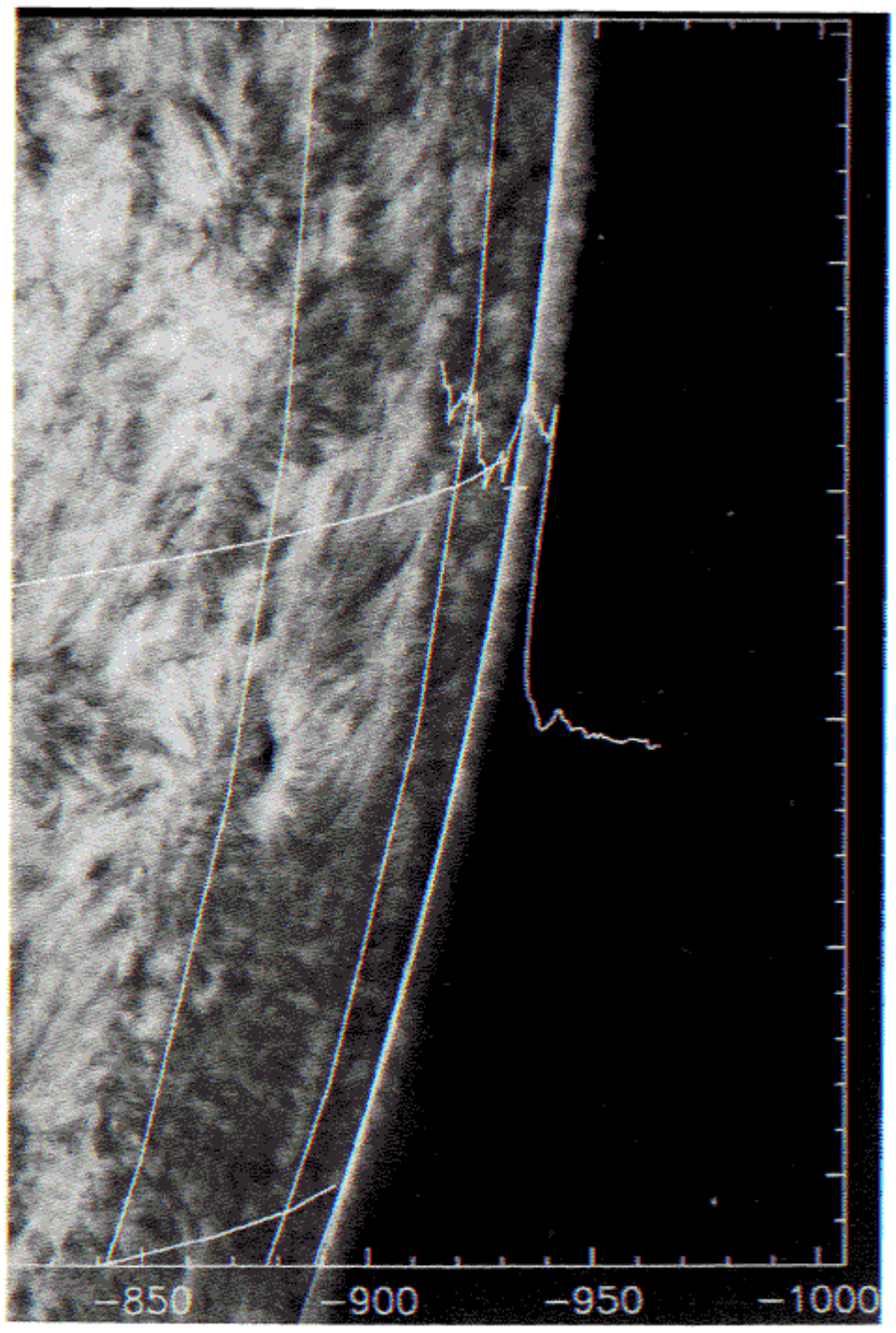


FIG. 6.—Off-band Hα photograph of the region near the limb, taken at BBSO on the day of the eclipse, overlaid with solar latitude and longitude lines at 10° intervals and with a plot of the 3 mm limb profile. *The vertical scale of the radio profile is arbitrary.* The plus sign shows the center of the beam. The spicules can be seen as a fuzzy edge above the optical limb, nearly at the same height as the sharp cutoff in the radio profile. The contrast of the Hα spicules was photographically enhanced for clarity.

Recent work has called into question the existence of a stationary temperature profile in the chromosphere, such as those indicated in Figures 2 and 3. This idea is spurred observationally by the fact that very cool regions of the atmosphere appear to exist in which CO (carbon monoxide) lines are seen. It is also suggested by time-dependent, dynamical models. Carlsson and Stein (1995) give the following plot showing a time-averaged "chromosphere" and the actual range of time-variable dynamical temperatures that went into the average.

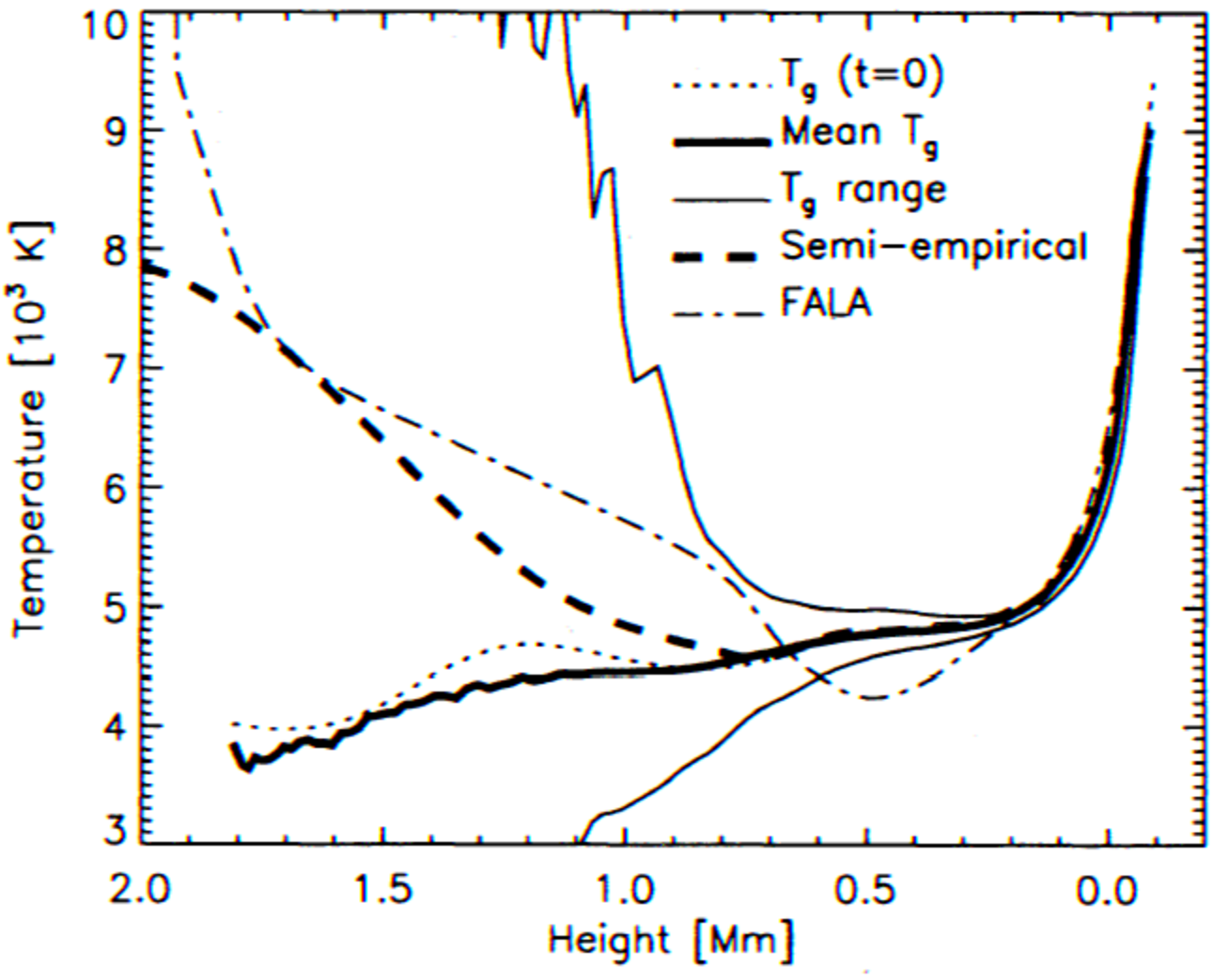


Figure 7: A comparison of the FAL model A with some temperature profiles of the lower solar atmosphere derived from time-dependent wave heating models of the chromosphere. The two thin solid curves show the extremes of temperature, while the thick line shows the time-averaged profile. The dot-dashed line is the FAL A model.

The chromosphere is dominated by the supergranular structure--large-scale convective cells of order 30" (200,000 km) in size. The gas rises in the center of the cells, moves to the edges, and descends. As a result, it tends to sweep relatively weak magnetic fields to the edges, where it collects to form the *chromospheric network structure*. Radio images show a good correspondence of the radio sources with the magnetic elements, as shown in Figure 8. Quantitative results are frustrated by the fact that such images from the VLA require an all-day (8-12 hour) synthesis, and so time-variability cannot be easily followed. There is some evidence that the individual magnetic elements "flicker" in brightness.

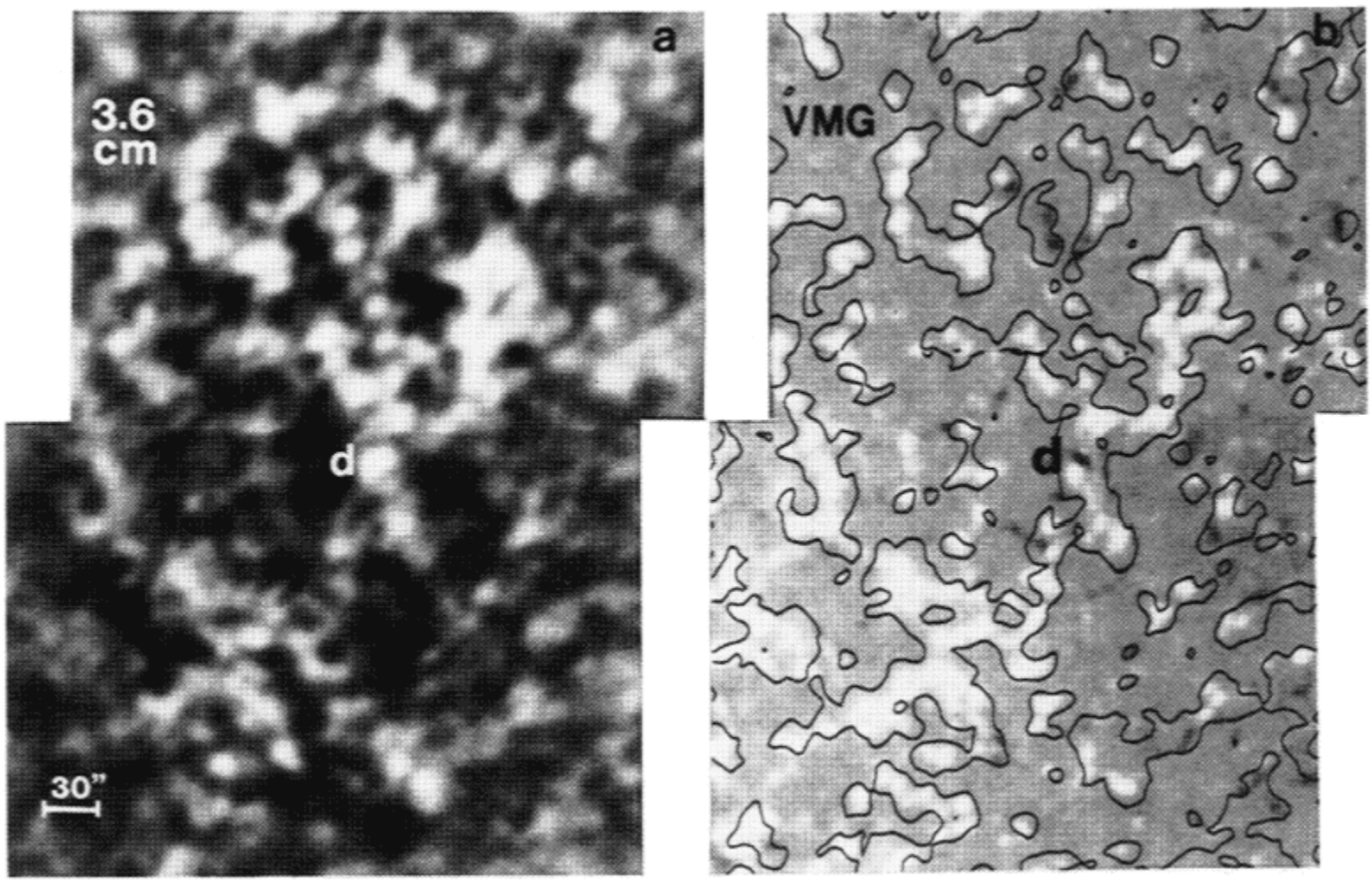
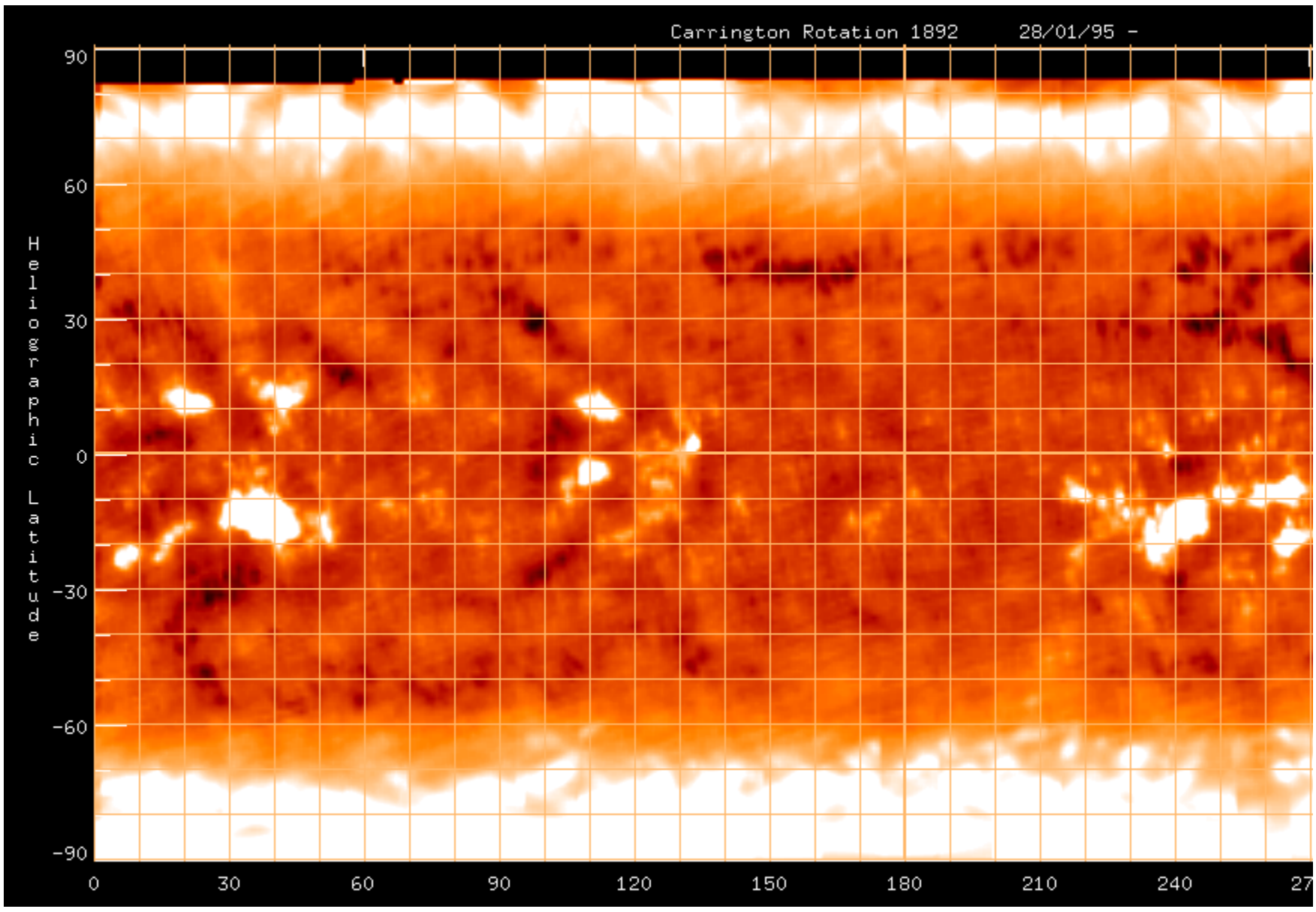


Figure 8: The panel on the left shows the radio image at 8.5 GHz (3.6 cm wavelength), showing the network struction. The corresponding magnetogram on the right shows a high degree of agreement, with each source overlaying a magnetic element of the chromospheric network. The feature marked with a "d" is the only one representing a magnetic dipole.

One still unresolved puzzle about the chromosphere is why at some frequencies (at least 10-100 GHz) the polar coronal holes appear brighter than the rest of the quiet Sun. There is some evidence that all coronal holes, even those not at the poles, are brighter. This means that the temperature of the lower chromosphere (that sampled by this range of frequencies) is greater at an equivalent optical depth. Meanwhile, coronal holes are *darker* than the rest of the quiet Sun at lower frequencies. Here is a

synoptic chart showing this effect at 17 GHz, from Nobeyama. The contrast in this image is enhanced, but the excess brightness is really only about 500-1000 K.



In addition to the ability of radio emission to give the temperature of the chromosphere, it can also tell us something about the magnetic field strength, derived from the magnetic field dependence of the polarization. Recall that free-free emission opacity was given in Lecture 3 as:

$$\kappa_\nu = (1/3c)(2\pi/3)^{1/2}(\nu_p/\nu)^2[4\pi n_e \Sigma(n_i Z_i^2)e^4/m_e^{1/2}(kT)^{3/2}]G_{ff}(T,\nu)$$
$$= 9.78 \times 10^{-3} n_e \Sigma(n_i Z_i^2)/(\nu^2 T^{3/2}) \qquad (1)$$
$$\times \begin{cases} 18.2 + \ln T^{3/2} - \ln \nu & (T < 2 \times 10^5 \text{ K}) \\ 24.5 + \ln T - \ln \nu & (T > 2 \times 10^5 \text{ K}) \end{cases}$$

However, this expression does not include the effect of the magnetic field. That can be included by changing the frequency from ν to $\nu + \sigma\nu_B |\cos\theta|$, where $\nu_B = 2.8 \times 10^6 B$ is the electron gyrofrequency, θ is the angle of B to the line of sight, and σ is +1 for o-mode, and -1 for x-mode. The expression (1) then becomes:

$$\kappa_\nu = 9.78 \times 10^{-3} n_e \Sigma(n_i Z_i^2)/[(\nu + \sigma\nu_B |\cos\theta|)^2 T^{3/2}] \qquad (2)$$
$$\times \begin{cases} 18.2 + \ln T^{3/2} - \ln \nu & (T < 2 \times 10^5 \text{ K}) \\ 24.5 + \ln T - \ln \nu & (T > 2 \times 10^5 \text{ K}) \end{cases}$$

where we ignore the slight dependence on ν_B in the logarithmic term. Recall that in order for free-free opacity to be important relative to gyroresonance opacity, we must have $\nu > s\nu_B$, where s is typically 3 or 4, otherwise gyroresonance opacity will dominate.

When we observe the chromosphere in the two senses of circular polarization (R and L), the expression is the same as (2), but without the absolute value on the $\cos\theta$ term. In other words, a given mode (i.e. x-mode), can correspond to either R or L polarization, depending on the sign of B.

Now, recall that for optically thin emission,

$$T_b = T_e \tau = T\kappa L.$$

We can express the degree of polarization as

$$P = (T_R - T_L)/(T_R + T_L) = (\kappa_R - \kappa_L)/(\kappa_R + \kappa_L) = (2\nu_B/\nu)\cos\theta. \qquad (3)$$

Thus, the degree of polarization is directly proportional to longitudinal (line-of-sight) component of

magnetic field strength, $B_l = B \cos\theta$. This gives a means to measure relatively weak magnetic fields in the chromosphere. Unfortunately, there is one more complication, because in fact the chromosphere is not optically thin. One might expect, then, that this would eliminate any net polarization, but note that the two modes become optically thick at different depths. The greater opacity in x-mode means it becomes optically thick higher in the chromosphere than the o-mode. Due to the temperature gradient in the chromosphere, the brightness temperature of the two modes is different (x-mode has a higher brightness temperature), so there is still a net polarization. Now, however, the polarization is due to the mode-dependence of the brightness temperature due to the temperature gradient, so we need a way to get the otherwise unknown temperature gradient. Luckily, if we have the brightness temperature measured at many frequencies one can use the slope of the radio spectrum, which itself is related to the temperature gradient. In fact, if the brightness temperature spectral index is written as n, then the expression (3) becomes

$$P = (n\nu_B/\nu)\cos\theta, \qquad (4)$$

which is a general expression since for an isothermal plasma the optically thin spectral index is $n = 2$, for which (4) reduces to (3). Can we really measure magnetic fields in this way? We do not really know at present, because we have not had an instrument capable of the required observations. We need excellent imaging to map the complicated structures of Figure 8, but we also need the images at many closely spaced frequencies to determine the spectral index. That is one motivation for building the Frequency Agile Solar Radiotelescope (FASR), which for the first time will have the required combination of imaging and spectral capability.

Active Region Corona

As we go higher in the solar atmosphere, the temperature rises steeply to millions of K, while the electron density falls greatly. This hot, tenuous plasma was first discovered through radio observations, was quite unexpected, and still remains unexplained. There have been many theories to try to explain it, such as wave energy coming from the surface and being deposited in the corona, but none of these theories seem to work. A favored mechanism now is dissipation of magnetic energy through many small flares (microflares), but current estimates show that there is not enough energy released in the visible events to account for the hot corona. There remains the possibility that even smaller events (nanoflares) might supply the needed energy, but so far they have not been shown to rise steeply enough in numbers to account for the needed energy.

The corona is everywhere hot, but certainly is hottest in active regions, which are regions associated with sunspots. Here we see that microflares and larger flaring events tend to concentrate, and it shows that flares are closely related to magnetic fields, that get stressed due to motions and new flux emergence to provide energy for the sudden releases that we call a flare. However, many aspects of this release of energy are still unknown. We will discuss flares and solar activity in the next lecture. For now, let's look at the general structure of active regions.

Physics 728 Radio Astronomy: Lecture #11 Prof. Dale E. Gary
NJIT

Solar Radio Emission II

Solar Activity

In addition to the structure of the non-flaring Sun, radio emission is exquisitely sensitive to flaring emission. Let us define such emission to be any brightening with time variability on scales shorter than a few hours. Of course, the Sun is continuously variable on all timescales, so this division is somewhat arbitrary, but it will suffice for our purposes. The brightening can take on a fantastic range of forms, from a slight increase due to heating (thermal emission) to 1000-fold increases in seconds or less. We have already discussed the radio emission mechanisms that are important -- free-free emission (bremsstrahlung), gyrosynchrotron emission, and plasma emission. Gyroresonance emission could also be produced due to heating, of course.

All of solar activity arises due to the solar magnetic field. There is a famous saying, attributed to various people, that *"If it did not have a magnetic field, the Sun would be as uninteresting an object as most astronomers believe it to be."* But fortunately for us, it does have a magnetic field, and so produces a wonderful variety of phenomena that are interesting to study -- and that are also important -- since solar activity affects us directly in many ways here on Earth.

The solar flare starts with a period of energy storage, called ***flare build-up***, that can occur over a period of days, but often results from the eruption of new magnetic flux from below the photosphere, which can take only hours. The stored energy takes the form of a non-potential magnetic field distribution. During this time, the changes take place in conditions of ***ideal MHD*** (ideal Magnetohydrodynamics), meaning that there is a balance of magnetic and gas pressures, and the field lines are "frozen in." Once conditions in the corona are right, the magnetic field can release its energy, sometimes in seconds, through a mysterious process called ***magnetic reconnection***. In this process, the field lines are cut (something normally impossible in ideal MHD) and reconnected to a lower-energy configuration that is closer to potential. The difference in energy between the original non-potential configuration and the resulting, more potential configuration is available for mass motions, acceleration of charged particles, and generation of waves.

What we call the flare can represent the immediate release of energy, the initial heating and acceleration of the charged particles, and all phenomena associated with the resulting mass motions and subsequent thermalization of the particles. One associated phenomenon is the Coronal Mass Ejection (CME), which is a magnetic bubble that becomes unstable and buoyant, leaving the Sun and propagating out into interplanetary space. There is a lot of argument about the relationships between flares and CMEs. The old assumption was that CMEs were the result of the flare, but timing studies are ambiguous and there is at least as much evidence that CMEs cause flares! In fact, both flares and CMEs can occur one without the other, so it is perhaps best to consider them related but separate phenomena.

Anatomy of a Flare

The figure below shows a schematic view of the relationships between the different components of a flare, somewhat late in its development. The point marked "acceleration site" is relatively high in the corona, and is the point where reconnection is taking place. Loops labeled MW and SXR are the locations of the microwave and soft X-ray sources, respectively. These are recently reconnected loops, with overlying loops being most recently formed. Here is a movie of 2D reconnection of the type illustrated. The part of the diagram at right shows a schematic of what actual radio observations show -- type III bursts going upward, RS (reverse slope) bursts going downward, and DCIM (decimetric pulsations) at higher frequencies. You can see that the region sampled by the decimetric range of frequencies (300 MHz to 3 GHz) is a very interesting one, since it covers the typical range of heights corresponding to the reconnection region. In fact, for historical reasons the part of the spectrum from about 400-1000 MHz has never had high-resolution imaging, so the spatial structure of sources in this band remains largely unknown. FASR will be the first to image this region in detail.

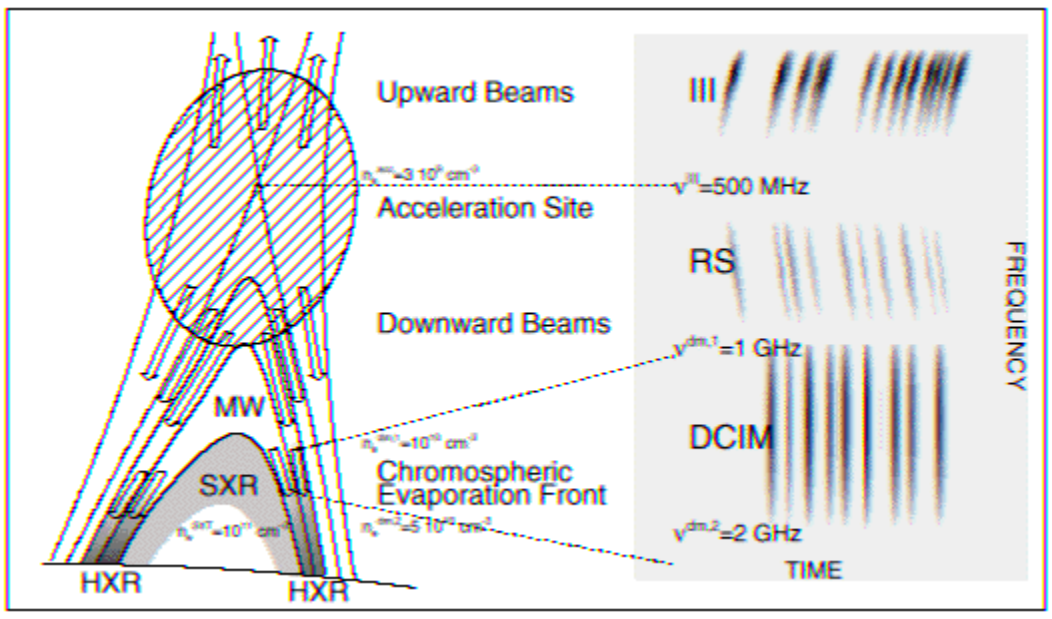
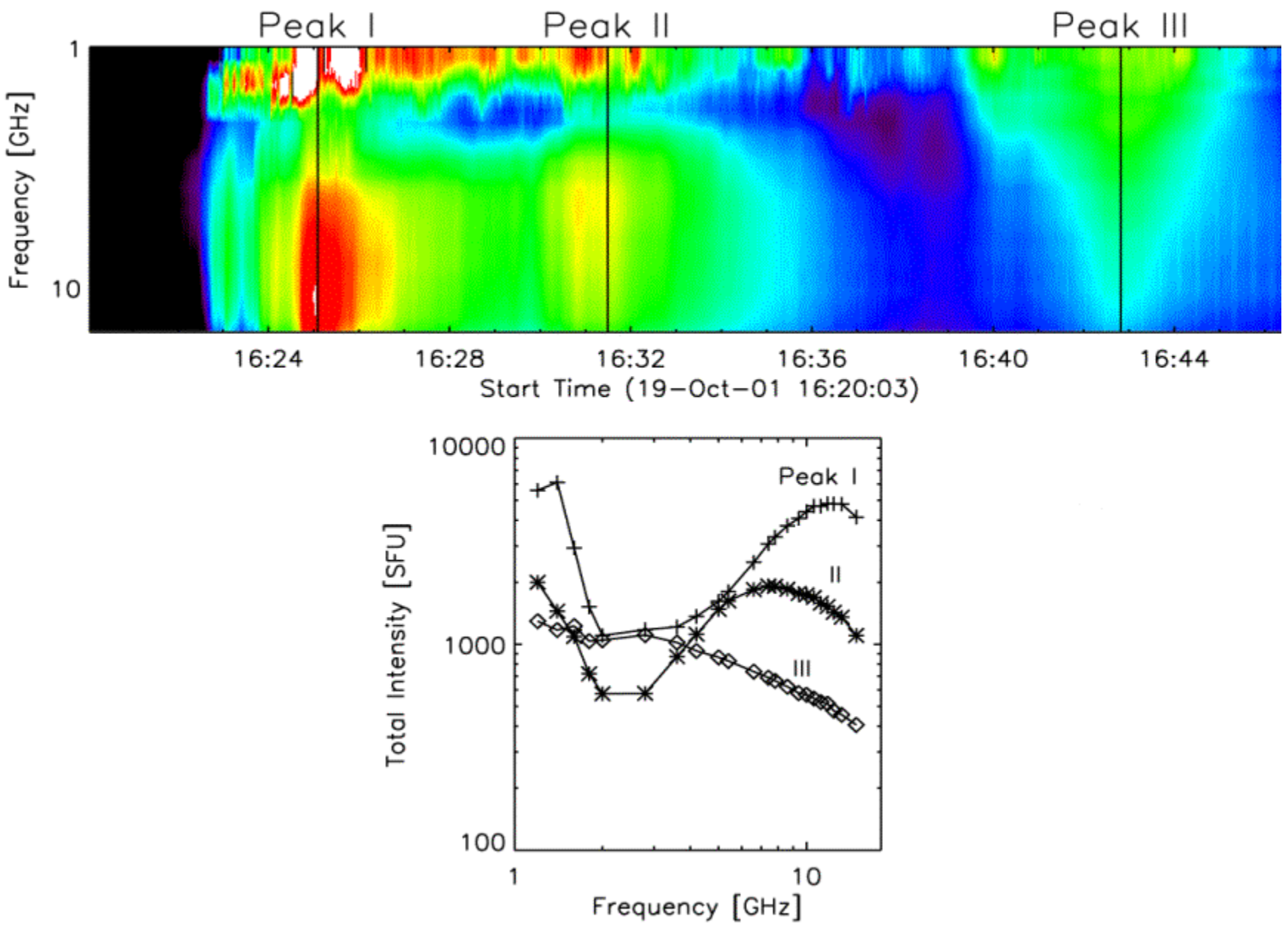
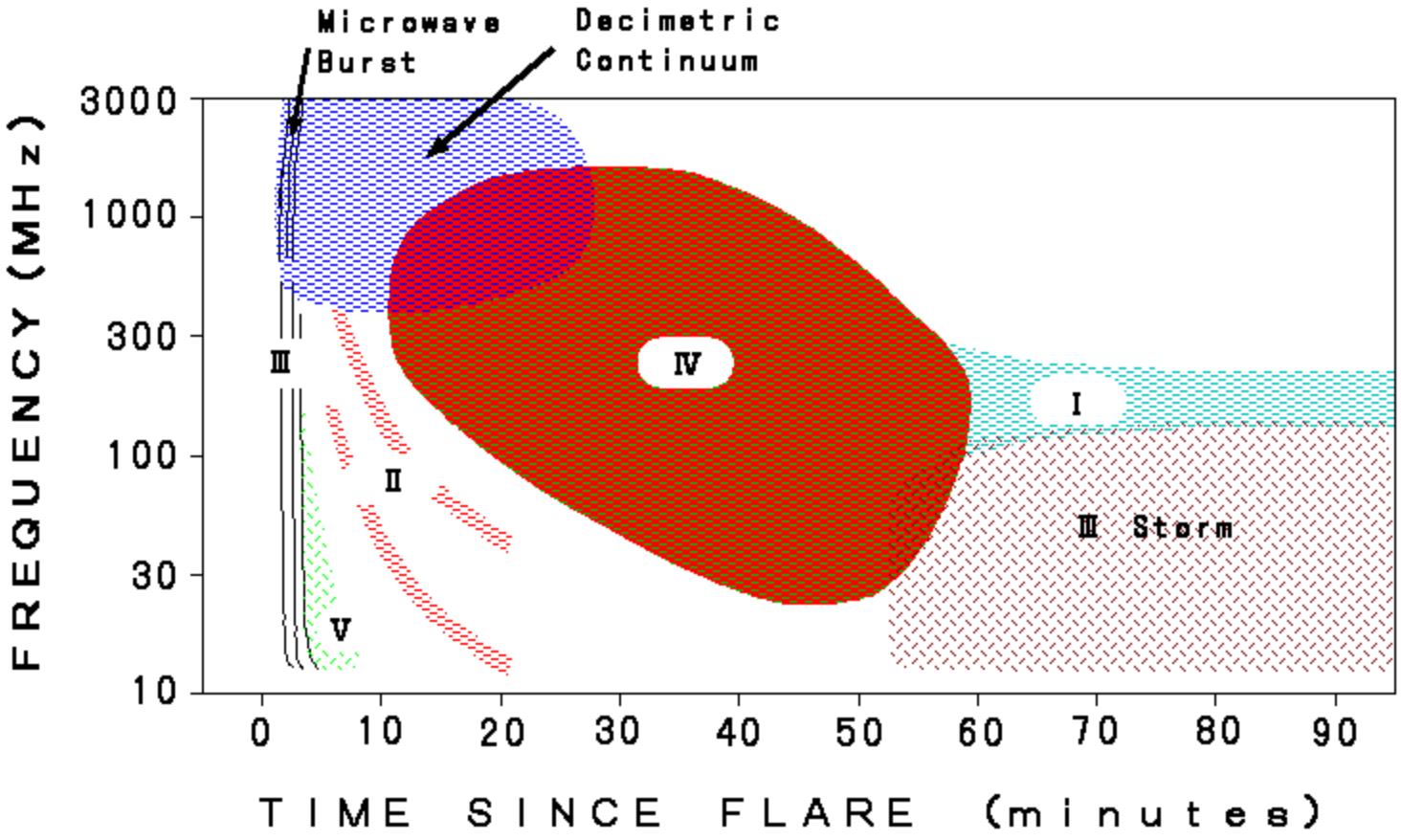


Figure 1: A schematic view of a solar flare

At still higher frequencies, electrons trapped on closed loops (the loops labeled MW) cause the gyrosynchrotron emission. Here is a radio dynamic spectrum from OVSA, showing the > 1 GHz part of this emission:



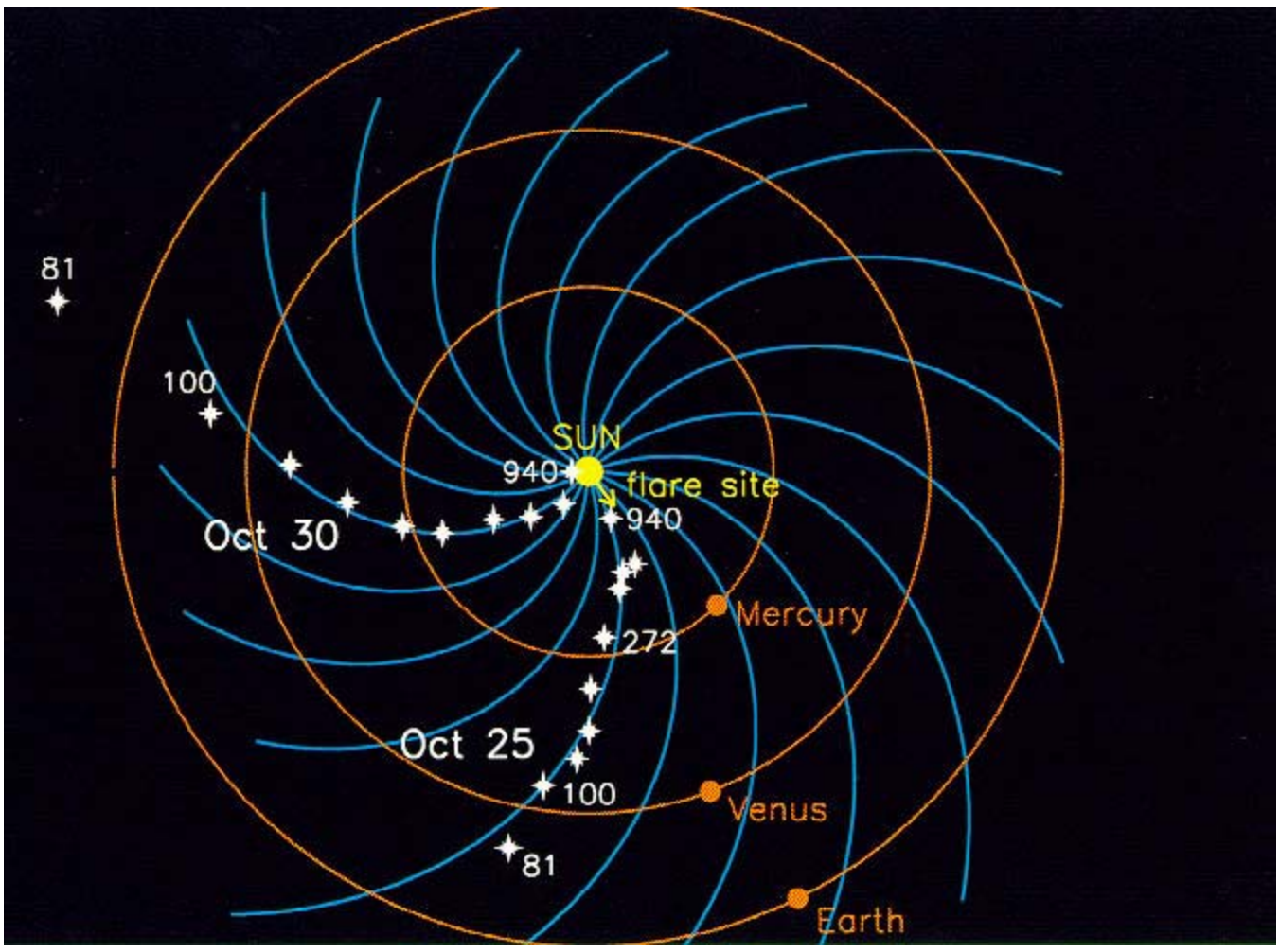
At still lower frequencies, the metric burst types are seen. These types were discovered in the 1950's, through the use of spectrographs, and each type has a unique signature in frequency and time. Here is a schematic diagram.



The types are:

- **Type I**
  - ***type I bursts*** or chains, narrowband, non-drifting bursts of short duration
    - cause is unknown, but emission appears to be plasma emission
  - ***type I storm*** -- a continuum emission with many type I bursts embedded, duration days
    - cause is unknown, but related to continuous reconnection above active region
- **Type II**
  - ***type II burst***, slowly drifting, often with fundamental/2nd harmonic structure, due to plasma emission
    - cause is a shock wave, propagating at 500-2000 km/s outward into the corona into interplanetary space *(also seen down to kilometric wavelengths)*.
- **Type III**
  - ***type III burst***, rapidly drifting, often with fundamental/2nd harmonic structure, due to plasma emission. The fundamental is highly o-mode polarized, and the 2nd harmonic is weakly (15%) x-mode polarized.
    - cause is a stream, or beam, of electrons moving at speed ~ $c/3$, propagating from low corona into interplanetary space *(also seen down to kilometric wavelengths)*.
  - ***type III storm*** -- a long lasting (up to a day or more) series of type III bursts, RS (reverse slope) bursts, reverse-drift pairs, and continuum.
- **Type IV**
  - ***stationary type IV*** -- broadband continuum emission, sometimes highly polarized, due to either plasma emission (o-mode polarized) or gyrosynchrotron emission (x-mode polarized).
    - cause is a plasmoid or high, filled loops of non-thermal particles
  - ***moving type IV*** -- a similar cause, but entrained in a CME or expanding arch.
- **Type V**
  - ***type V burst***, continuum emission following a type III burst, x-mode polarized (opposite sense to the associated type III)
    - cause is slower type III-like electrons in widely diverging magnetic fields, with both forward and counterstreaming langmuir waves, perhaps generated by previous passage of type III electrons.

One can observe type III bursts propagating all the way to Earth and beyond, as we saw in the earlier homework problem. By using a spacecraft equipped with a rotating dipole, the direction of the emission can be determined, and its frequency (proportional to square root of density) gives its distance from the Sun (using a density vs. distance model). Here is an example of tracing two type III bursts from the Sun out into the IP medium. You can see how the electrons trace the archimedian spiral (Parker spiral).

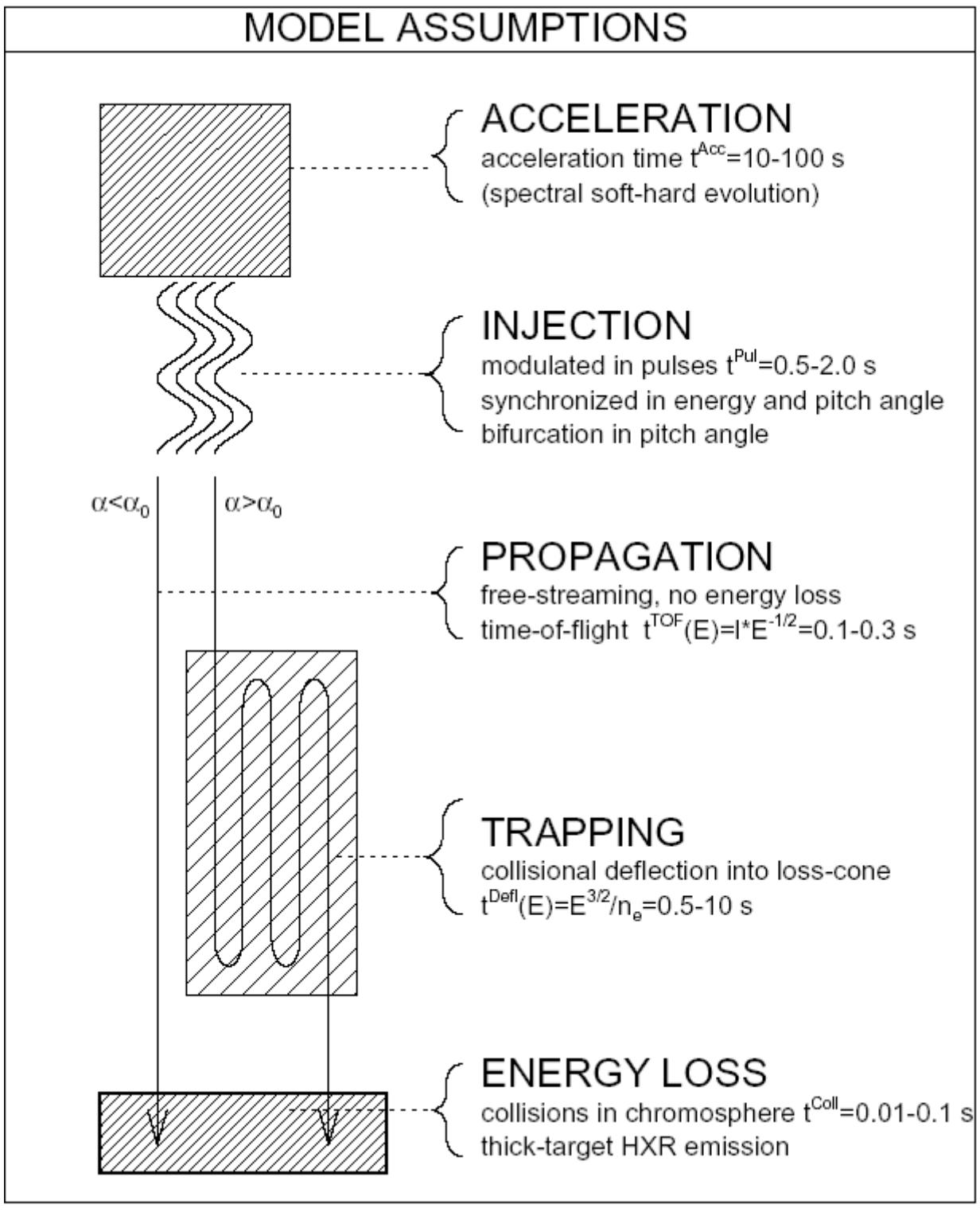


Gyrosynchrotron Emission

There are several useful diagnostics that can be deduced from gyrosynchrotron emission. As we discussed in Lecture 3, the emission is broadband emission, with optically thick emission at low frequencies (typically 2-5 GHz), optically thin emission at high frequencies (>10 GHz), and a peak around 5-10 GHz. The peak frequency is related to the magnetic field strength and number density, the optically thin spectral index is related to the electron powerlaw index, and the polarization tells us about the direction of the magnetic field. The sensitivity of microwave emission to these parameters makes it a useful complement to other measures of the accelerated electrons, i.e. hard X-rays.

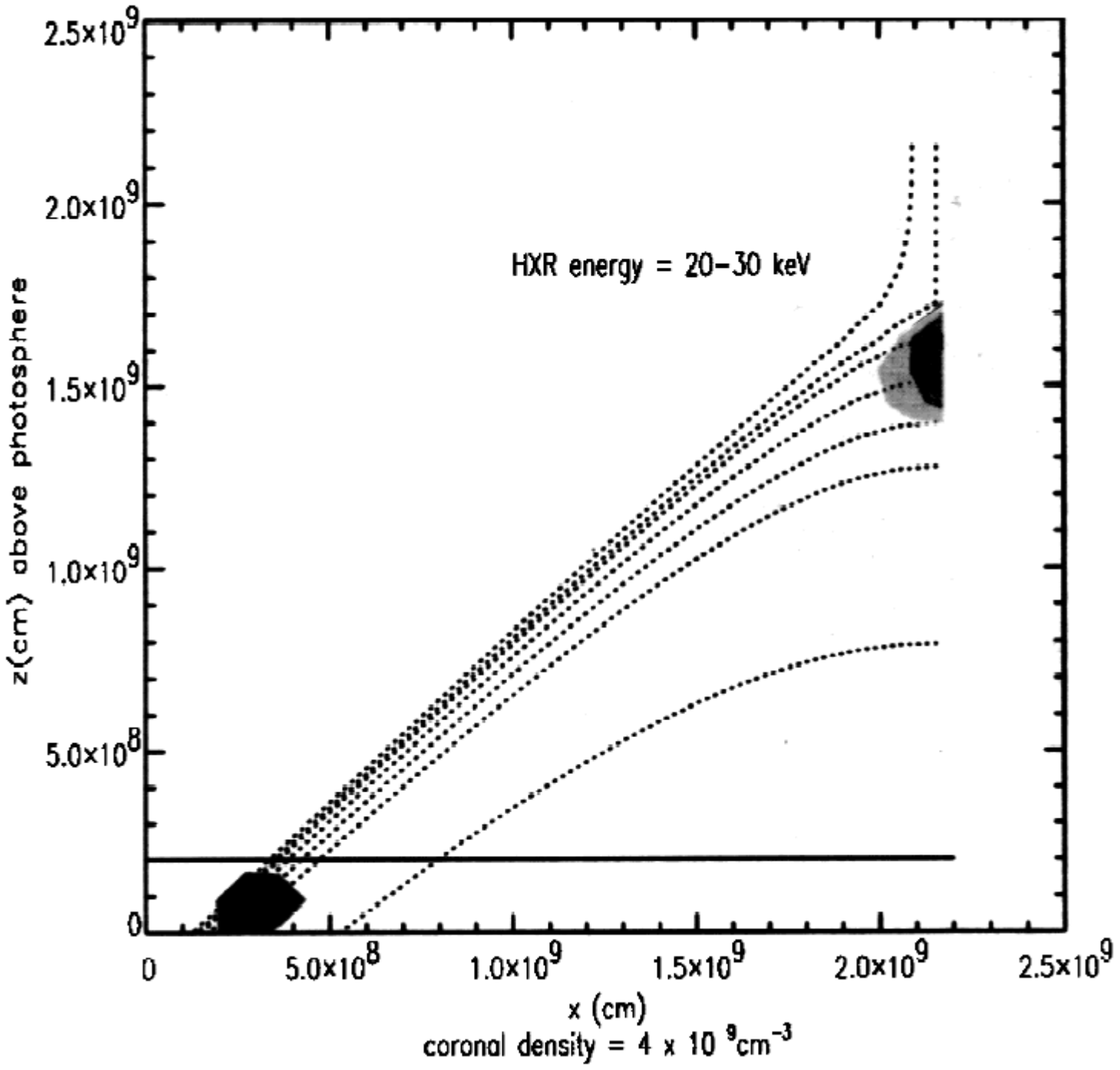
Acceleration, Trapping, and Precipitation

Before showing some examples, it is useful to examine what happens to charged particles after they have been accelerated. The goal is to understand the acceleration itself, which is fundamental to magnetic reconnection. At present it is largely a mystery how the reconnection process can accelerate so many particles in such a large volume in seconds. Let us leave the acceleration as a "black box," and see what can happen to the particles afterward.



From Aschwanden (1998)

The acceleration is connected to the emitting volume through a process called ***injection***. This allows the emitting volume to be separate from the acceleration volume. In addition, the injection into the volume can be modulated separately from the acceleration. After injection the electrons can go directly to the chromosphere (if they are in the loss-cone) and be lost, generating hard X-rays, or they can go into a magnetic trap for awhile, before ultimately escaping and suffering the same fate. The trapped electrons are the ones that produce the bulk of the microwave emission. A cartoon of the geometry is:



From Fletcher and Martens (1998)

Each of the above processes affects the time profile and energy distribution of the electrons, and those electrons ultimately produce the radio emission.

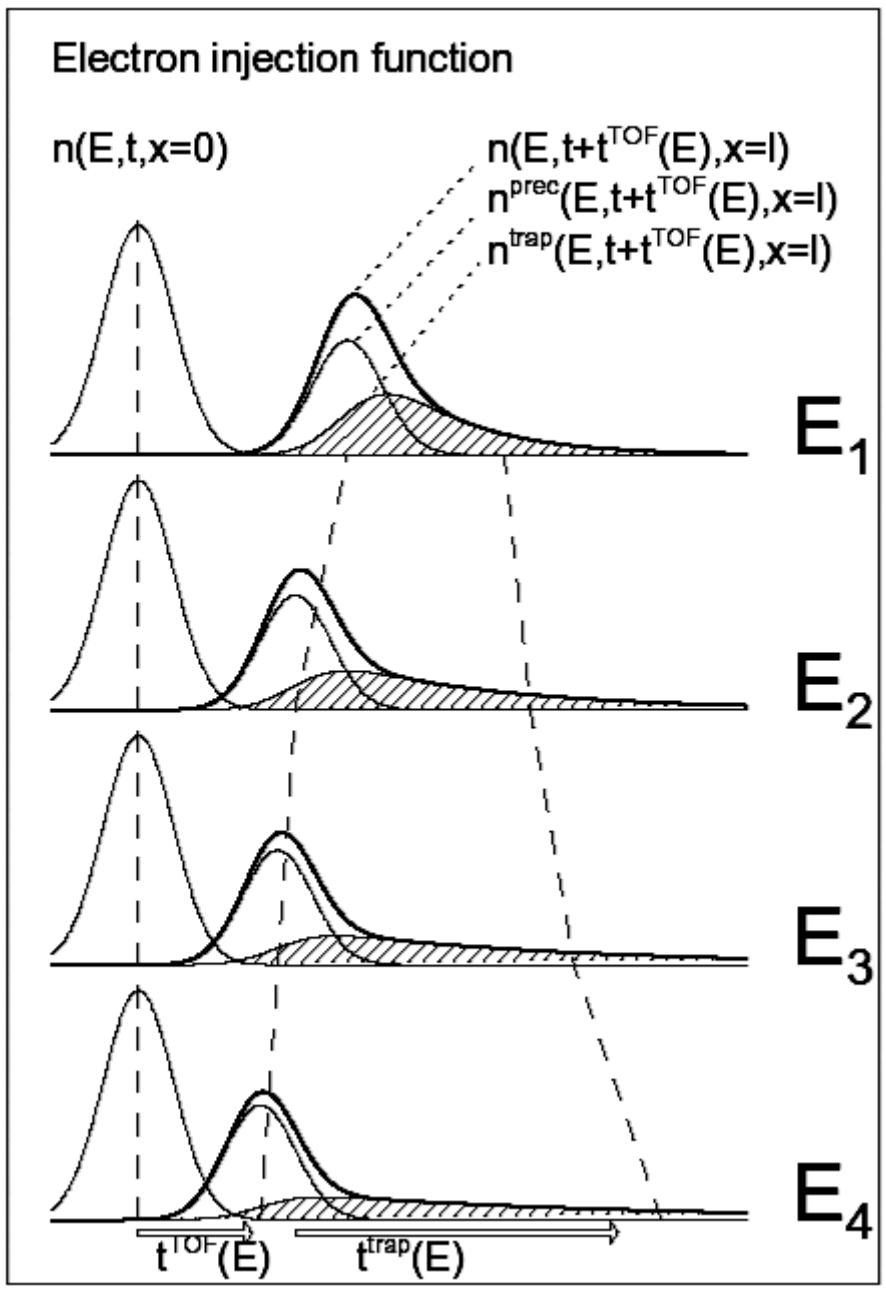
- The ***acceleration*** itself will have a time profile that can be arbitrary, somehow associated with the reconnection rate.
- The ***injection*** process
  - can be trivial, just a delta function that transmits the particles with the same time and energy characteristics as the acceleration,
  - or it can modulate the acceleration in both time and energy (including pitch angle).
- As shown on the left in the figure, the ***propagation*** can also be trivial, simply transmitting the electrons to the chromosphere with only a short delay. Typically some fraction of the electrons (those in the loss cone, with pitch angle $\alpha < \alpha_o$) will escape in this way. The loss cone is a function of the ratio of the magnetic field B_{inj} at the injection point (often considered the loop top) and the magnetic field at the B_{loss} loss point (the footpoint): $B_{loss} / B_{inj} = 1/\sin^2\alpha_o$.
- If both injection and propagation are trivial, then the hard X-rays produced in the chromosphere (***energy loss***) will mirror exactly the time and energy characteristics of the acceleration. This makes hard X-rays a particularly useful diagnostic of the acceleration process. This component of hard X-ray emission is called the ***direct precipitation*** component.

However, we have no reason to believe that the injection and propagation are trivial. We often see pulsations or modulations of the acceleration that can be attributed to the injection process. The propagation itself will be subject to other effects such as

- ***trapping***: Typically some fraction of the electrons (those not in the loss cone, with $\alpha > \alpha_o$) will be trapped for a time. Trapping results in an integration, where the product of acceleration and injection time profiles are convolved with an exponential decay time (the loss time).
- ***escape*** from the trap: This is due to particles being scattered into the loss cone. Such escaping particles will also produce hard X-ray emission when they strike the chromosphere, and this is called the ***secondary precipitation*** component. This can happen due to
  - Coulomb collisions (collisions between ions and electrons), in which case the collisions are energy dependent (larger effect on low-energy particles), which will flatten, or harden the electron energy spectrum over time. In this case if the original powerlaw spectral index is δ, the new spectral index after collisions will be $\delta - 3/2$.
  - Wave-particle interactions, in which case the energy dependence is more complicated,

depending on the wave modes involved, and the electron energy spectrum will in general not be a diagnostic of the accelerated spectrum (unless the wave mode interaction is understood in detail, theoretically).

The temporal relationships among injection, the direct precipitation, and the secondary precipition are shown in the following figure.



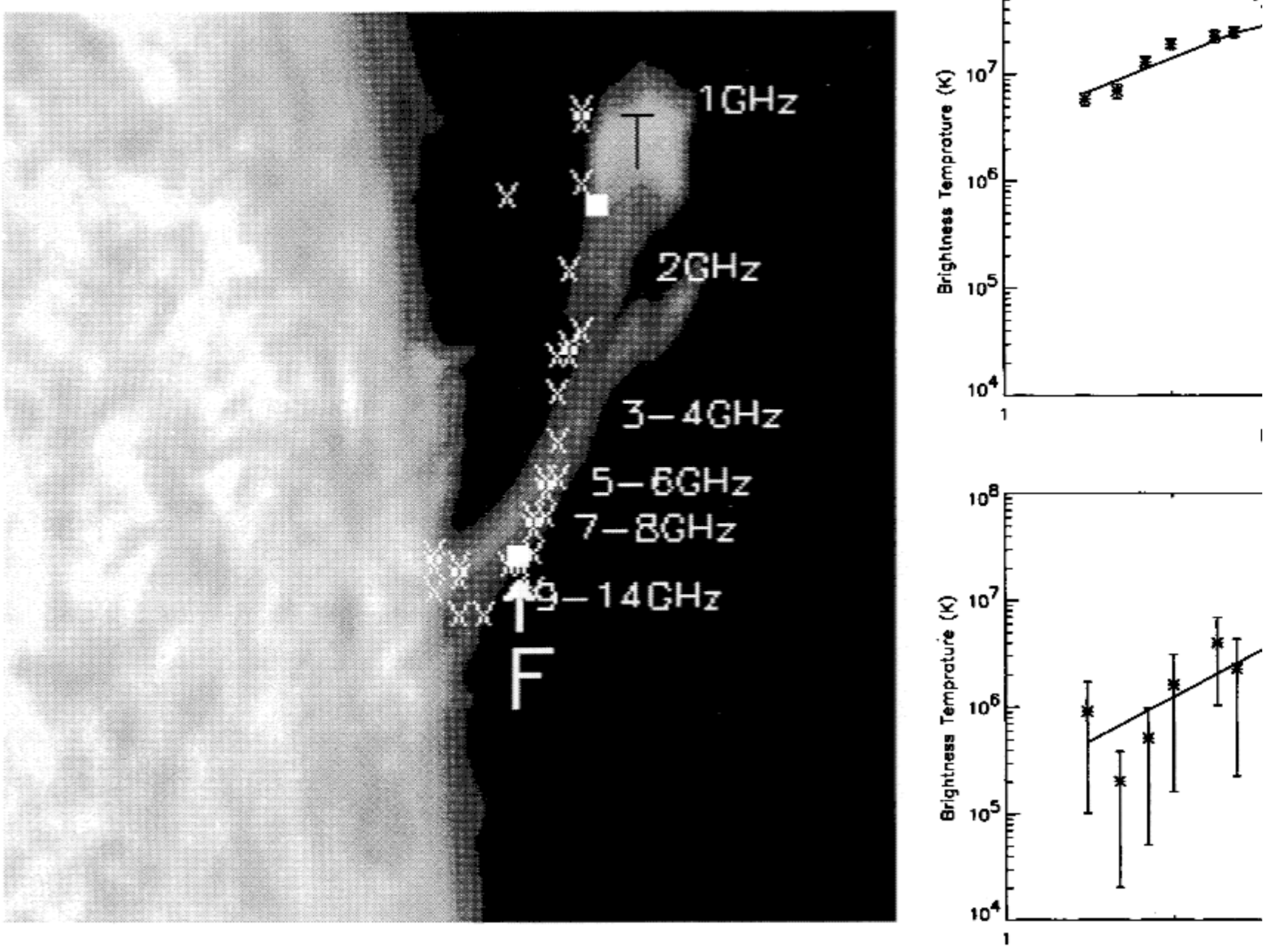
From Aschwanden (1998)

You can see that there are two opposite energy dependences. The TOF, or time-of-flight, delay decreases with increasing energy (they are simply moving faster) while the trapping time increases with increasing energy (because the higher energy particles suffer less from Coulomb collisions, and so do not scatter into the loss-cone so quickly.

Microwave Observations

It is within this framework that we now try to understand microwave emission from flares. The shaded component in the figure above is the hard X-ray signature from the slowly escaping electrons, but at the same time the trapped population (the ones that have not yet escaped) can be huge and it is that population that is producing the microwaves. Note that electrons much more readily produce microwave emission (which requires only small accelerations) than they do hard X-rays (which require large accelerations, essentially stopping the electrons). This means that microwaves are far more sensitive to high-energy electrons than are hard X-rays.

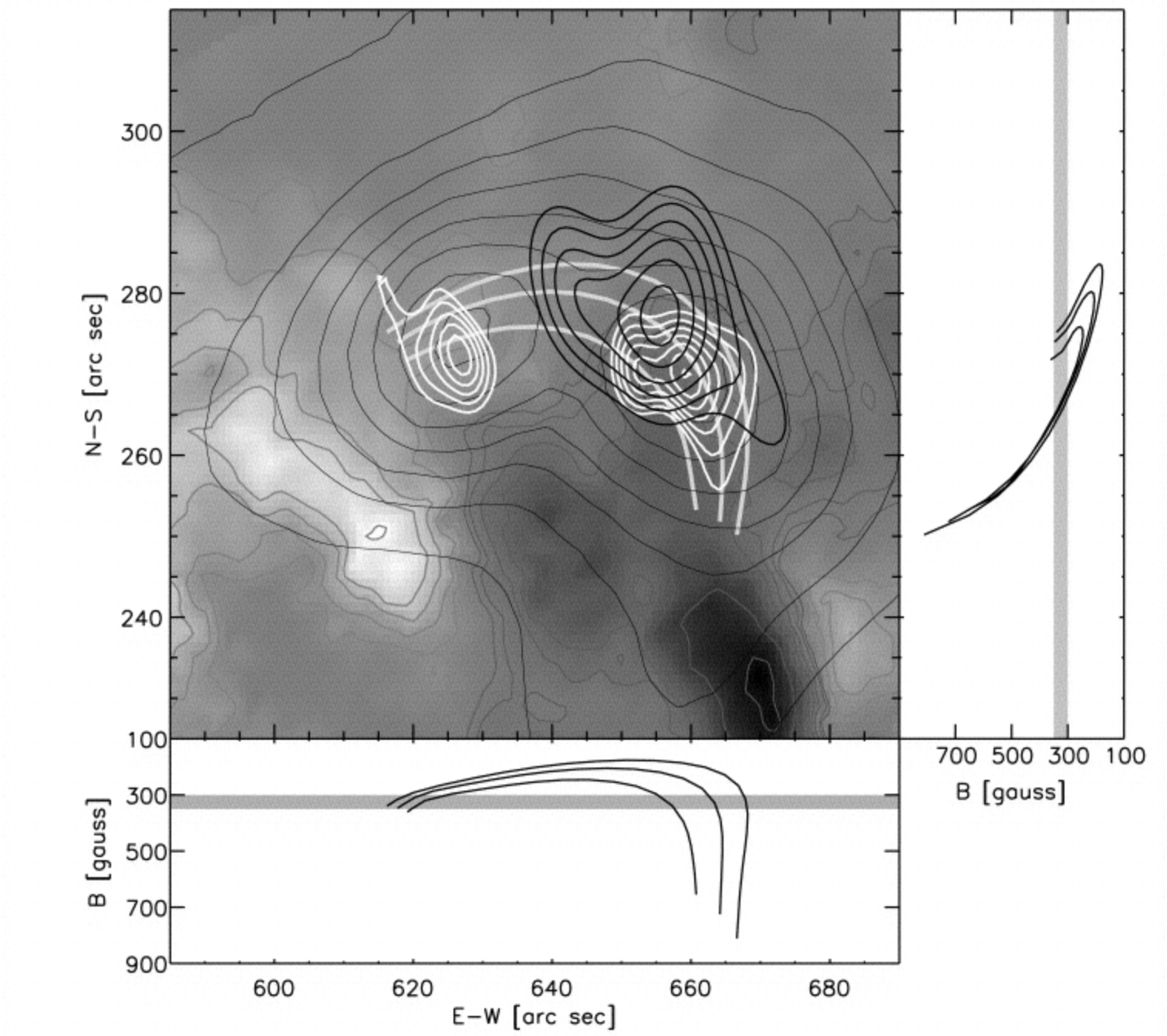
The frequency of the microwaves is proportional to the magnetic field, so when we make images at many frequencies we see a dispersion in position, essentially mapping out the magnetic field.



From Wang, Gary, Lim & Schwartz (1994)

When we make spatially resolved brightness temperature spectra at different points along the loop, we see that different parts of the loop look quite different. Again, the peak frequency is related to the magnetic field strength. The upper spectrum is taken at the location marked T, for "top", in the photo, while the lower spectrum is taken at the point marked F, for "footpoint". The H-alpha image shows a surge that occurred later than the flare, but we take the surge as marking the location of a large loop that was involved in the earlier flare. The footpoint may actually be the location of a separate, smaller flaring loop.

The figure below gives a good opportunity to see how trapping may be reflected in spatial observations.



From Lee, Gary & Shibasaki (2000)

Here we see three different frequencies, 17 GHz (thin contours, from Nobeyama), 10.6 GHz (white contours, from OVSA), and 5 GHz (thick black contours, from OVSA), along with expected field line locations from a photospheric field extrapolation. This is an asymmetric loop (the right side has much higher field strength than the left side), and we see that the electrons apparently mirror at some height above the footpoints on that side.

Some presentations from the FASR workshop, at http://www.ovsa.njit.edu/fasr/May_program.html :
Nobeyama (Shibasaki)
OVSA (Lee)
Time Scales (Aschwanden)

Physics 728 Radio Astronomy: Lecture #12 Prof. Dale E. Gary
NJIT

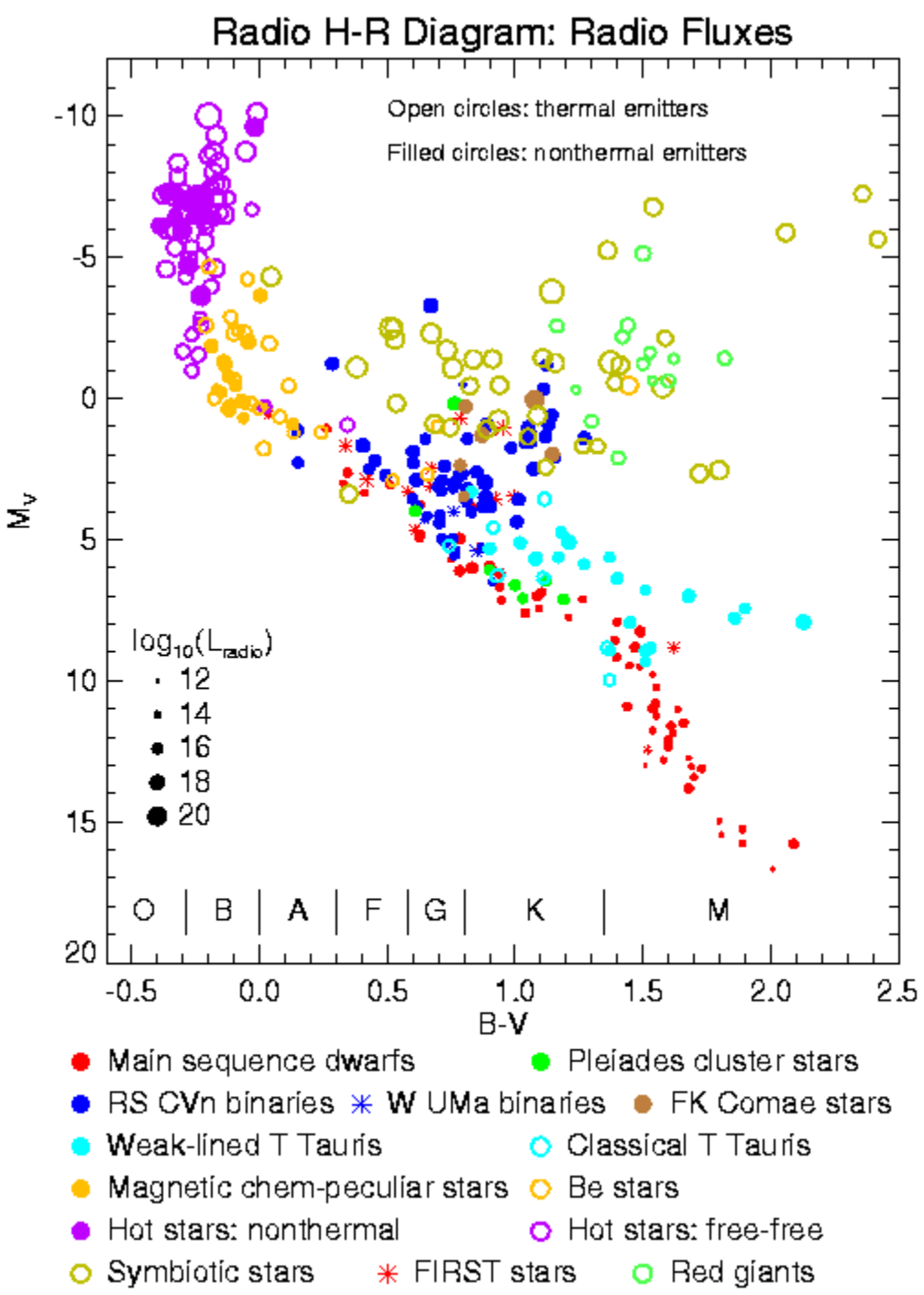
Astronomical Radio Emission

Solar System Objects

Planets and other solar system objects are general only thermal emitters (producing radio emission only due to black body radiation from their surfaces). Since they are generally cold (700 K for Mercury down to 30 K or so for Pluto), they are weak emitters. Here is Saturn's blackbody radio image. Jupiter is the main exception, since it has a very large magnetosphere (would be larger angular size than the Moon if we could see it in visible light), which traps high-energy electrons that then emit synchrotron radiation. We can also bounce radar signals off the nearby planets (Mercury and Mars), and image the echos.

Stars

A good place to start with stellar radio emission is to look at an H-R diagram.



This diagram is from Stephen White, at University of Maryland. The placement of the symbols is according to the star's classical visual magnitude vs. color (B-V), but the symbols themselves encode the information about the radio emission. Most of the main sequence and subgaint objects are nonthermal emitters (filled circles), while most of the gaints and many of the O-B stars are thermal emitters (simply because they are big). The blue circles near the center of the diagram are RS CVn binaries. These have a late-type subgiant "revved up" by tidal interactions with its close binary companion. The open circles just above the RS CVn ones are symbiotic stars, which again are binaries, but now with a compact companion (perhaps a black hole).

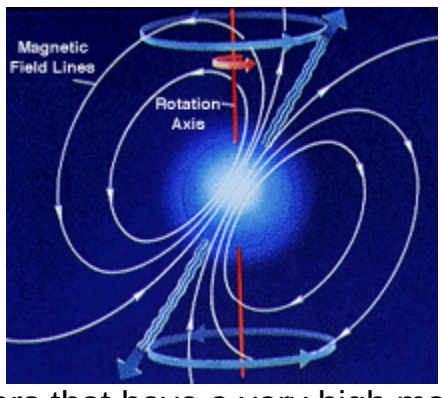
Note that the G, K and M dwarfs (red dots) are weak emitters. These objects are also flare stars, meaning that occasionally they have strong radio outbursts. Why should red dwarfs have large flares? You can calculate how strong the Sun's radio outbursts would appear if observed from the distance of the nearest stars, and you will find that they are barely detectable. Flare stars, on the other hand, have both optical and radio flares that are giant in comparison to solar flares. There is good evidence that such stars have a large fraction of their surfaces covered with "sunspots." This is probably due to their having fast rotation coupled with a fully convective interior, so that dynamo generation of magnetic fields is much larger than for the Sun. A few stars have so much activity that they can be said to have detectable "quiescent" radio emission all of the time (e.g. the star UV Cet). Here is a report of an intense flare on the flare star AD Leo, reaching a probably brightness temperature of 10^{13} K.

Pulsars

When a star's mass at the end of its life is $M > 1.4\ M_o$, electron degeneracy is no longer enough to keep gravity at bay, and matter is crushed to force inverse β decay

$p^+ + e^- \text{---> } n + \nu$. ($\nu$ is a neutrino)

so the protons and electrons are combined to form neutrons--a neutron star. The state of matter is a neutron degenerate gas. Degenerate objects have the peculiar property that with greater M they have smaller radii, up to ~ 3 M_o. Radii are typically 10-30 km.



Pulsars are rapidly rotating neutron stars that have a very high magnetic field (concentrated during the collapse of the core of a star into a neutron star) whose poles happens to be offset from the direction of the spin axis. If the spinning happens to bring the poles around to point at Earth, we will see these bright poles briefly as an intense radio emitting source. The emission is due to synchrotron emission of electrons in the high magnetic fields. The pulses, of course, are repeated on each spin. Pulsars spin at periods ranging from 4 s to 1.6 ms. Here is how they sound:

pulsar sounds

You can imagine the forces on the fastest pulsar. The surface speed is $v = 2\pi R / P$, where P is the pulsar period and R is the pulsar radius (about 10 km = 10^4 m):

$$v = 2\pi R / P = 2\pi(10^4 \text{ m}) / 0.0016 \text{ s} = 3.9 \text{ x } 10^7 \text{ m} > 0.1\ c.$$

Its centripetal acceleration is $a_c = v^2/R$, and the star will rip itself apart if this is greater than the gravitational acceleration holding the star together, $a_g = GM / R^2$, that is, if the surface velocity v is greater than $v = [GM/R]^{1/2}$. What is the smallest possible period P that a pulsar could have?

$$P = 2\pi R / v = 2\pi R / [GM/R]^{1/2} = 0.3 \text{ ms}$$

We can combine mass and radius and write this in terms of only one quantity, the density, as

$$P = 3.8 \text{ x } 10^5 \rho^{-1/2}.$$

With the fastest periods of about 2 ms, one can see that the density must be high.

Binary Pulsars and pulsar evolution

Pulsar Evolution

Galactic Objects

Cassiopeia A is a supernova remnant that has been studied with detailed maps over a couple of decades, now. One can actually see the expansion by watching the detailed images as a movie (unfortunately, the movies seem to be not working). What we are seeing is the outer envelope of an exploded star that is moving outward into the interstellar medium with high velocity, surrounded by a shock wave that is still heating material to emit X-rays.

SiO maser emission from stellar atmospheres, and water maser emission from H II and star formation regions, so the surprising variety of radio emission mechanisms. The SiO molecules surrounding some stars with extended, cool atmospheres is preferentially in the J = 1 spin state, and as radio emission at the right frequency (43 GHz) stimulates the transition from J =1 to 0, they emit another photon. This photon, along with the original one, proceed into the cloud of molecules and stimulates more transitions, giving rise to a very bright line emission in small regions. The direction of the magnetic field can be deduced from the direction of linear polarization of the emission. Likewise for the water (H_2O) maser, operating at 22 GHz.

Mapping H I In Galaxies

Atomic hydrogen (also called neutral hydrogen, and H I), shows the distribution of relatively cool gas in a galaxy. Comparing two different types of galaxy shows the extreme differences.

Nearby Spiral Galaxy

"Our 21 cm mosaic provides the most detailed view yet attained of neutral hydrogen in a spiral galaxy (other than the Milky Way). The observations are characterized by spatial resolution of 20 pc (5" at 840 kpc) and velocity sampling
of 1.3 km/s. For this reason, our database compares straightforwardly with the recent ATCA+Parkes surveys of the Large and Small Magellanic Clouds (Staveley-Smith et al. 1997, Stanimirovic et al. 1999, Kim et al. 1998). At the VLA, M33 was observed using six mosaic pointings in both the B (48 hr) and CS (6 hr) configurations. Our interferometric data has recently been complemented by ultra-sensitive total power observations obtained at WSRT, using the Dutch instrument in an auto-correlation mode whereby all 14 elements are employed as incoherent single-dishes.

"Figure 1 shows a color representation of our peak brightness temperature image, in which the hue has been assigned on the basis of velocity at peak ?B in each of the spectra. The pattern of galactic rotation dominates one's visual impression, but doesn't obscure significant localized motions, perhaps most apparent as abrupt color changes within the spiral arms. For this preliminary image, no masking of the cube has been applied. Instead, we preserved sensitivity by tapering to 40 pc resolution (10" FWHM). We are now developing methods to create a “multiresolution” version of this map, in which the beam size is position dependent and broadens to maintain signal-to-noise in faint regions such as the outer disk and interarm gaps."

Elliptical Galaxy

"VLA atomic hydrogen observations of the shell galaxy NGC 2865. The gas is shown as yellow contours on an optical image from the Digital Sky Survey. The main body of the NGC 2865 is typical of early type galaxies, but at fainter light levels the galaxy exhibits a peculiar morphology, with many shells, ripples and loops. The VLA spectral line observations shows gas within the main body of the elliptical, but also distributed in an extended ring around it."

Why do spirals have gas and dust in them, and ellipticals do not? The answer lies in our new understanding of how ellipticals form, through galaxy collisions and mergers. The stars in such a "collision" do not collide, but merely pass through each other. The gas and dust, however, does collide and ends up outside the galaxy.

Quasars and Lobe Radio Sources

Quasars are Active Galactic Nuclei (AGN), i.e. the centers of extremely active galaxies. For a time they were mysterious objects because they appear only like a faint star optically (the term quasar is short for quasi-stellar object), but recent observations show that they do have faint "nebulosity" around them, which is actually the light from the rest of the galaxy in which they are embedded. We now know that these are powered by supermassive black holes. They have extremely well collimated jets, seen both optically and in the radio (as below). The jets culminate in giant radio lobes (sometimes on only one side), which are many, many times the size of the parent galaxy.

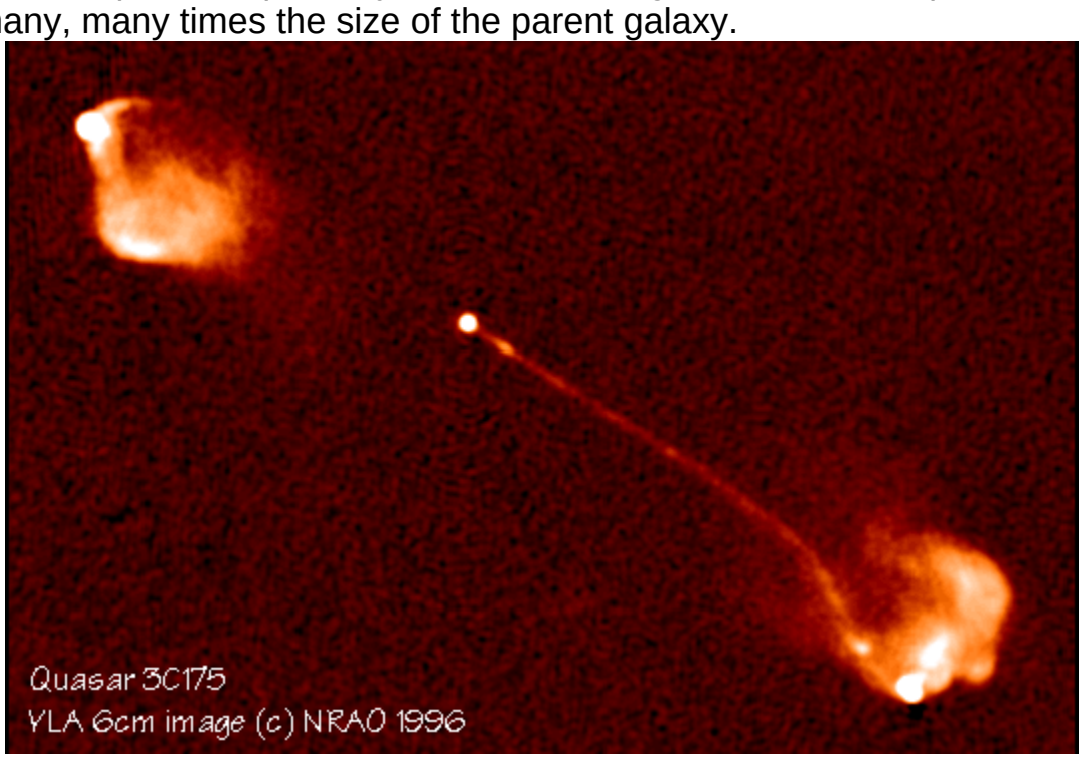


Image courtesy of NRAO/AUI

"This image shows the radio emission from relativistic streams of high energy particles generated by the quasar. This is a classic double-lobed radio source. Astronomers believe that the jets are fueled by material accreting onto a super-massive black hole at the center of the host galaxy (not shown in this image). The high energy particles are confined to remarkably well collimated jets, and are shot into extragalactic space at speeds approaching the speed of light, where they eventually balloon into massive radio lobes. The overall linear size is of the radio structure is 212 kpc (for a Hubble constant of 100 km/s/Mpc), which can be compared to a typical galaxy diameter of about 30 kpc. The quasar has double lobes with prominent hot spots, and has a narrow jet, but no counter-jet. Its possible that we only see the jet that is pointing toward us, which may be "Doppler boosted" in brightness when the particles emitting the radio radiation are moving toward us at close to the speed of light. The counter-jet would be moving away from us, and would thus not experience Doppler boosting. The jet brightens and bends as it enters its lobe."

One interesting phenomenon that one can observe in the jets is the presence of "superluminal"

sources. These are sources that appear to move at velocities as much as 45 times the speed of light! This is just an apparent speed, caused by the source moving very close to our line of sight at nearly the speed of light. In effect, we see time compressed, and so the source appears to be moving faster than c.