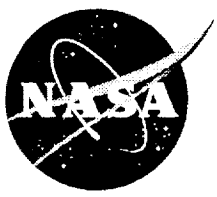
NASA Reference Publication 1391

December 1996

Current Perspectives in High Energy Astrophysics

Jonathan F. Ormes, Editor



NASA Reference Publication 1391

1996

Current Perspectives in High Energy Astrophysics

Jonathan F. Ormes, Editor
Goddard Space Flight Center
Greenbelt, Maryland

National Aeronautics and Space Administration

Goddard Space Flight Center
Greenbelt, Maryland

This publication is available from the NASA Center for AeroSpace Information,
800 Elkridge Landing Road, Linthicum Heights, MD 21090-2934, (301) 621-0390.

Preface

This book is based on a set of lectures given by members of the staff of the Laboratory for High Energy Astrophysics at NASA's Goddard Space Flight Center. The lectures were given at the University of Maryland as part of an astronomy course in High Energy Astrophysics given first during the fall of 1994 and in revised form during the winter of 1996. The course was designed for beginning graduate students in astronomy or physics.

The topics in the book are organized by the distance to the astronomical object or phenomenon being discussed; it starts nearby and works its way to the distant universe. Following a general introductory chapter, it begins with the nearest star, the sun, and works its way out through the Milky Way galaxy to phenomena found in extragalactic space. The history of energetic particles in the galaxy can be found: stars collapse and explode as supernova making remnants seen in X-rays and producing shock waves that accelerate cosmic rays. Cosmic rays interact with interstellar gas and dust to produce gamma-rays. Radioactive elements produced by stellar synthesis are distributed throughout the interstellar medium and can be observed via their radioactive decay gamma-ray emission lines. High-energy radiation from extragalactic space is shown to be dominated by active galactic nuclei and diffuse emissions therefrom. X-rays allow us to trace the distribution and history of matter, both visible and invisible, in the universe at large. The final chapter brings us full circle: gamma-ray bursts may come from the distant universe, from a halo around our galaxy, or, perhaps, from the distant regions of our own solar system.

This is an appropriate time for such a course because much new has been learned recently from missions such as the Compton Gamma-Ray Observatory (CGRO) and other 1990s missions such as the joint German/British/American Roentgen Satellite (ROSAT) and the Japanese/American Advanced Satellite for Cosmology and Astrophysics (ASCA). These new missions have made discoveries and advanced understanding of the X-ray and gamma-ray radiation from a variety of astronomical objects: active galactic nuclei including blazars and Seyfert galaxies, pulsars, the interstellar medium and the sun. They have solved some mysteries from the 1970s such as revealing the identity of the gamma-ray source known as Geminga and have opened new mysteries such as whether or not gamma-ray bursts are coming from the galaxy or from the distant reaches of the universe. It is hoped that the book can be used as a textbook or for background reading in astronomy or astrophysics courses. This first edition will be published in limited quantities and distributed for review and comment by our colleagues and their students. Is the material clear and easily understood? Are the Chapters coherent individually, and do they fit appropriately into the whole? What is too repetitive? What is left out? What about the length? Your comments are eagerly solicited. If the feedback is positive, and students find the book useful, it will be revised and a second more widely distributed edition will be published.

The editor wishes to thank the Chapter authors first and foremost for their excellent contributions, but also for their patience with the many step process of correcting the manuscript and making the style uniform. The editor was assisted early on in this process by Ms. Cheryl Madison and for the past six months by Mr. George Hilton. Without his dedicated efforts and patience with the authors, and theirs with him and me, this book would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

HIGH ENERGY ASTROPHYSICS IN THE 1990s

JONATHAN F. ORMES
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

High energy astrophysics is a space-age discipline that has taken a quantum leap forward in the 1990s. The observables are photons and particles that are unable to penetrate the atmosphere and can only be observed from space or very high altitude balloons. This chapter presents an overview and summary of the results that will be presented in more detail in the remaining chapters. The lectures presented as the chapters of this book are based on the results from the Compton Gamma-Ray Observatory (CGRO) and Advanced Satellite for Cosmology and Astrophysics (ASCA), missions to which the Laboratory for High Energy Astrophysics at NASA's Goddard Space Flight Center made significant hardware contributions. These missions study emissions from very hot plasmas, nuclear processes and high energy particle interactions in space. Results to be discussed include gamma-ray beaming from active galactic nuclei (AGN), gamma-ray emission from pulsars, radioactive elements in the interstellar medium, X-ray emission from clusters of galaxies, and the progress being made to unravel the gamma-ray burst mystery. The recently launched X-ray Timing Explorer (XTE) and prospects for upcoming Astro-E and Advanced X-ray Astronomy Satellite (AXAF) missions are also discussed.

1.1 INTRODUCTION

Imagine yourselves as super-men and women with X-ray and gamma-ray vision looking at the sky from above the atmosphere. You would see a very different looking sky than is seen on a dark night from Earth. You would see beams flashing by and sources blinking on and off on all time scales in a wide variety of colors. "Stars" would appear and disappear, sometime reappearing again, often not. It would be a "laser show" worthy of the best that Disneyland can produce.

Why would physicists be interested in the field of study known as High Energy Astrophysics? The radiations that make up the field are photons in the X-ray and gamma-ray bands of the electromagnetic spectrum, as well as the energetic particles that produce this radiation. X-rays are formed by thermal bremstrahlung in million degree plasmas and by some of the nonthermal processes that characterize the gamma-ray band: high energy interactions between particles and electromagnetic fields. These interactions include synchrotron radiation by electrons in magnetic fields, Compton interactions between energetic electrons and photons, and other collisional processes that result in radiation from interaction decay products such as $\pi^o s$. The physics involves beams of relativistic particles and blobs of plasma accelerated to nearly the speed of light. It involves matter in highly condensed states and in magnetic fields from μ-*gauss* to 10^{13} *gauss*.

X-rays and gamma-rays are absorbed in the Earth's upper atmosphere and, hence, must be observed from platforms in space. These are "photon limited" observations in which arrival direction and energy are recorded for each individual quantum. The differential spectra of photons and energetic particles, dN/dE, generally fall rapidly with increasing energy. Often, however, the energy emitted, $E \times dN/dE$, can peak in the X-ray or gamma-ray band indicating that for certain types of objects, high energy processes may dominate the energetics and, hence, the evolution of an astronomical object.

X-rays are generally meant to include photons with energy of 0.1 to 30 kilovolts (keV), a range in which k- and l-shell atomic transitions produce line emission indicative of the presence of heavy nuclei. The lines are generally present on top of a thermal continuum from astrophysical plasmas with typical temperatures in the range of 10^5 to 10^7 K. These photons can be detected individually by measuring the energy they deposit in a solid or gas where they transfer their energy to atomic electrons via Compton collisions and subsequently produce free electrons or ion-electron pairs. The electric signals can be amplified and manipulated to measure the amplitude of the energy deposit. Any photon from a nuclear transition or interaction is called a gamma-ray. Thicker targets of solid or heavy liquids are required to efficiently absorb low energy gamma-rays; above 30 megavolt (MeV) the photons are detected via pair production. Astrophysical sources in the 10 keV to 30 MeV band are the most poorly explored because of technical problems detecting the photons and the high backgrounds from natural radioactivity.

This introductory chapter is organized by starting with the highest energies and working downward in energy and simultaneously moving towards wave bands that emit ever-increasing flux. It is mission oriented in the sense that the newest results have derived from the missions described. The discussion is meant to whet the appetite for subsequent chapters that describe the physics and astrophysics in more detail. In it we review some of the more interesting recent results of missions making observations in these bands and describe some of the future prospects in this field.

The book is organized by astronomical object or phenomenon, and begins with the nearest star, our own Sun, and works its way through the Galaxy to extragalactic objects ever more distant from the solar system. It ends with the phenomena of gamma-ray bursts; we do not know where they come from. The subsequent chapters are all based on lectures given by the authors as part of a course on high energy astrophysics given at the University of Maryland during fall 1993 and winter 1996 terms.

Chapter 2 describes the Sun as a laboratory for the study of nuclear gamma-ray line from the collisions of energetic particles accelerated in the solar magnetic fields. Chapter 3 is the story of galactic cosmic rays and the view of them throughout the galaxy taken in gamma-rays. Chapter 4 describes the supernova remnants that fill the interstellar medium and may provide the environment for the acceleration of the cosmic rays as described in Chapter 5. Chapter 6 covers the discovery of 511 keV annihilation radiation from the galactic center and our current understanding of this phenomenon. Chapters 7 and 8 deal with pulsars and compact binary systems, respectively, drawing on results from very recent X-ray and gamma-ray observations. Chapter 9 expands on the exciting and powerful new field of X-ray spectroscopy only briefly alluded to in this chapter. Chapter 10 explores more fully the new understanding of the gamma-ray sky. Chapters 11 and 12 deal with active galactic nuclei. Chapters 13, 14, and 15 cover various aspects of the X-ray background tracing the distribution of the emitting plasmas with emphasis in Chapter 15 on the emissions from clusters of galaxies. The last chapter describes the mystery of the phenomena of gamma-ray bursts.

Table 1.1: CGRO Instrument Parameters

Instrument	Energy Range (MeV)	Field of View Location	Source
BATSE	0.03 - 1.2	$4\pi sr$	$\sim 2°$
OSSE	0.06 - 10	$4° \times 11°$	——
COMPTEL	1 - 30	1 sr	~ 10 *arcmin*
EGRET	20 - 30,000	0.6 sr	~ 5 *arcmin*

1.2 GAMMA-RAY BURSTS

1.2.1 Compton Gamma-Ray Observatory

The Compton Gamma-Ray Observatory was launched in April 1991. The instruments and their primary characteristics are given in Table 1.1 (Gehrels, Chipman, and Kniffen, 1993). This mission is the first extensive examination of gamma-rays from space over the energy range from tens of kilovolts to tens of gigavolts, six decades in energy. All the instruments onboard CGRO are in excellent health and still taking data more than 5 years after its launch. Energetic Gamma-Ray Experiment Telescope (EGRET) the only experiment with a lifetime-limiting consumable, has enough spark chamber gas to operate on an intermittent basis for another year or two. The satellite should be capable of lasting well past the turn of the millennium.

First we discuss the results from EGRET, the highest energy experiment. The most dramatic and surprising discovery made in "100 MeV photon astronomy" has been the observation of gamma-rays coming from a class of active galactic nuclei (AGN) known as Blazars and their cousins, the BL Laecerta objects. The luminosity (energy output) of these objects in the gamma-ray band is impressive; at 4 billion light years distance, 3C279 was emitting 10^48 erg/s, assuming isotropic emission. The radiation is characterized by time variability on scales as short as a day. Blazars emit blobs of plasma with apparent motions of speeds greater than the speed of light. The most dramatic time variability is shown in Figure 1.1 from an observation of 3C279 in June 1991 shortly after the launch of the satellite (Hartman et al., 1992, Kniffen et al., 1993). The bulk of the energy from these objects is being emitted as gamma-ray photons. Figure 1.2 shows the emitted energy per unit energy; the gamma-ray emission clearly dominates, at least for those times when there is a gamma-ray outburst. The size of the emitting object is limited by the light crossing time. For an object at a redshift distance, Z, the size of the object, R, is given by

$$R \leq \frac{c\Delta T}{(1+Z)} \tag{1.1}$$

The luminosity, L_γ, of the object and its size along with

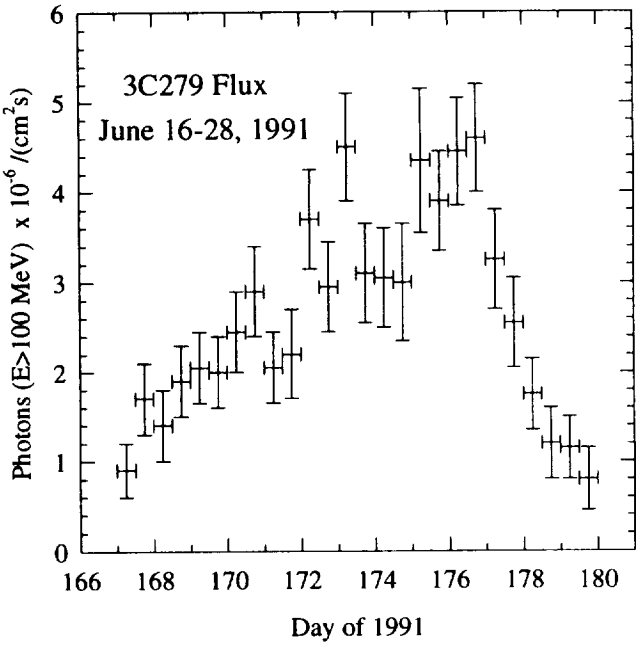


Figure 1.1: Gamma-ray time variability of quasar 3C279 in June 16-28, 1991 (Kniffen et al.1993).

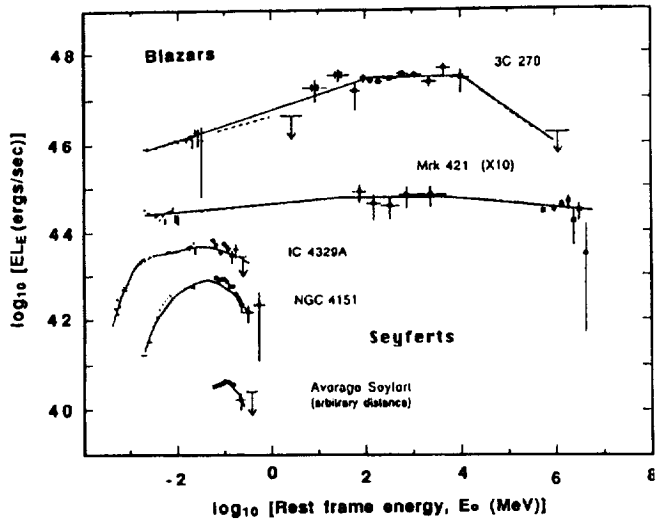


Figure 1.2: The energy emitted per decade by active galactic nuclei identified as gamma-ray emitters. Notice the very different spectra for the Blazars and the Seyfert galaxies (reproduced with permission, Dermer and Gehrels, 1995).

the cross section of the photons can be used to determine the optical depth, τ, of the object

$$\tau = \left[\frac{L\gamma}{R}\right]\left[\frac{\sigma_T}{m_e c^3}\right] \tag{1.2}$$

The optical depth turns out to be about 10 for 3C279

with ΔT of a day if the radiation is assumed to be isotropic, a number that would drive the energy requirements even higher. The way out is to assume the gamma-radiation is also beamed. That hypothesis has proven to be consistent with further observations, as the number of detected AGN has confirmed the association of gamma-rays with those objects whose radio beams are preferentially pointed towards the Earth. The total number of high-energy gamma-ray sources now identified with specific AGN is about 50 (Fichtel et al., 1994, von Montigny et al., 1995). Many other EGRET sources may also be AGN. The presence of beamed photons implies the organized acceleration of particles. These objects are the most powerful particle accelerators known. The energy to drive them comes ultimately from the infall of material onto a massive ($10^8 M_\odot$) compact object, probably a massive black hole.

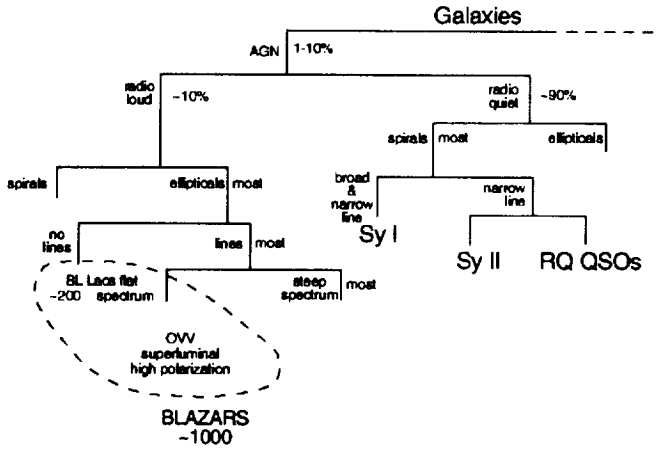


Figure 1.3: A categorization of active galactic nuclei according to type which, indicates the relative fractions of the various kinds (reproduced with permission, Gehrels and Dermer, 1994).

The OSSE experiment discovered another class of AGN, namely Seyfert galaxies that emit 100 keV photons, as shown in Figure 1.2. The spectra fall sharply at higher energies. These AGN are not powerful radio sources. The gamma-rays could be made by Compton scattering of optical and ultraviolet photons from relativistic electrons in their accretion disks (Dermer and Gehrels, 1995). Many people are attempting to integrate the understanding of gamma-ray photons from AGN into the unified picture of AGN that has emerged from earlier X-ray studies of these objects (Mushotzky, Done, and Pounds, 1993). Gehrels and Dermer characterize the various types of AGN as shown in Figure 1.3. In this picture, there are the of order of 1000 identified Blazars and BL Lacs that probably emit 100 MeV gamma-rays. Many more should be observable in hard X-rays. As far as AGN go, as a whole, the conclusion from CGRO might be "What you see depends on where you see it from!".

1.2.2 Pulsars

Another class of astrophysical object in which the gamma-radiation is important is a pulsar. Pulsars are believed to be highly magnetized, very rapidly rotating neutron stars with magnetic axis aligned at an angle to the spin axis as shown in Figure 1.4. They look very different

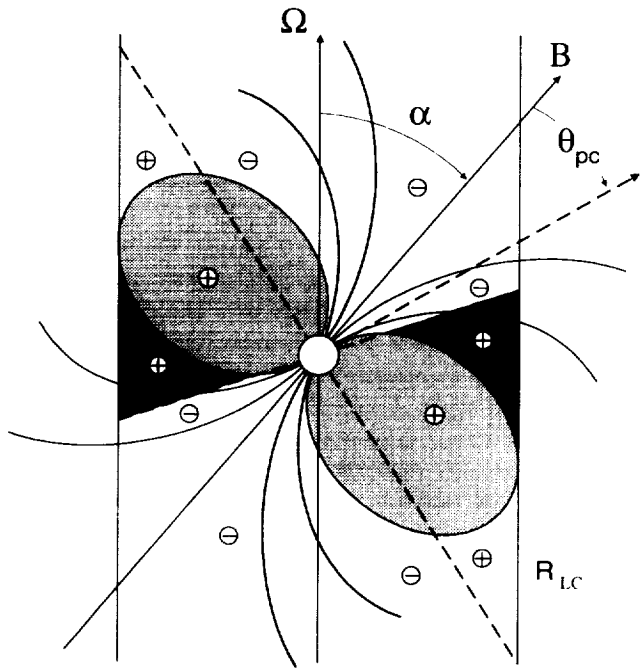


Figure 1.4: Pulsar schematic. In general, the angle between the rotation axis and the magnetic dipole axis of the neutron star is not known.

in detail when observed in different energy bands, as can be seen in Figure 1.5. Recent work (e.g., Romani and Yadigaroglu, 1994) attempts to explain those different light curves as manifestations of viewing the pulsars from different aspects. They base their explanation on the so-called "outer gap" model (Cheng, Ho, and Ruderman, 1986) in which the acceleration of particles, presumably electrons, takes place near the speed of light cylinder where the magnetic field corotating with the neutron star is moving with nearly the speed-of-light. On the other hand, others (Daugherty and Harding, 1982, 1994 and Sturner and Dermer, 1994) place the acceleration of particles near the polar caps. These fascinating objects are a complete physics laboratory by themselves. There is a relativistically rotating plasma being dragged around by a "spinning top" made of condensed neutrons. The object accelerates particles to relativistic energies and emits a beam of radiation with emission throughout the electromagnetic spectrum. They are the only objects

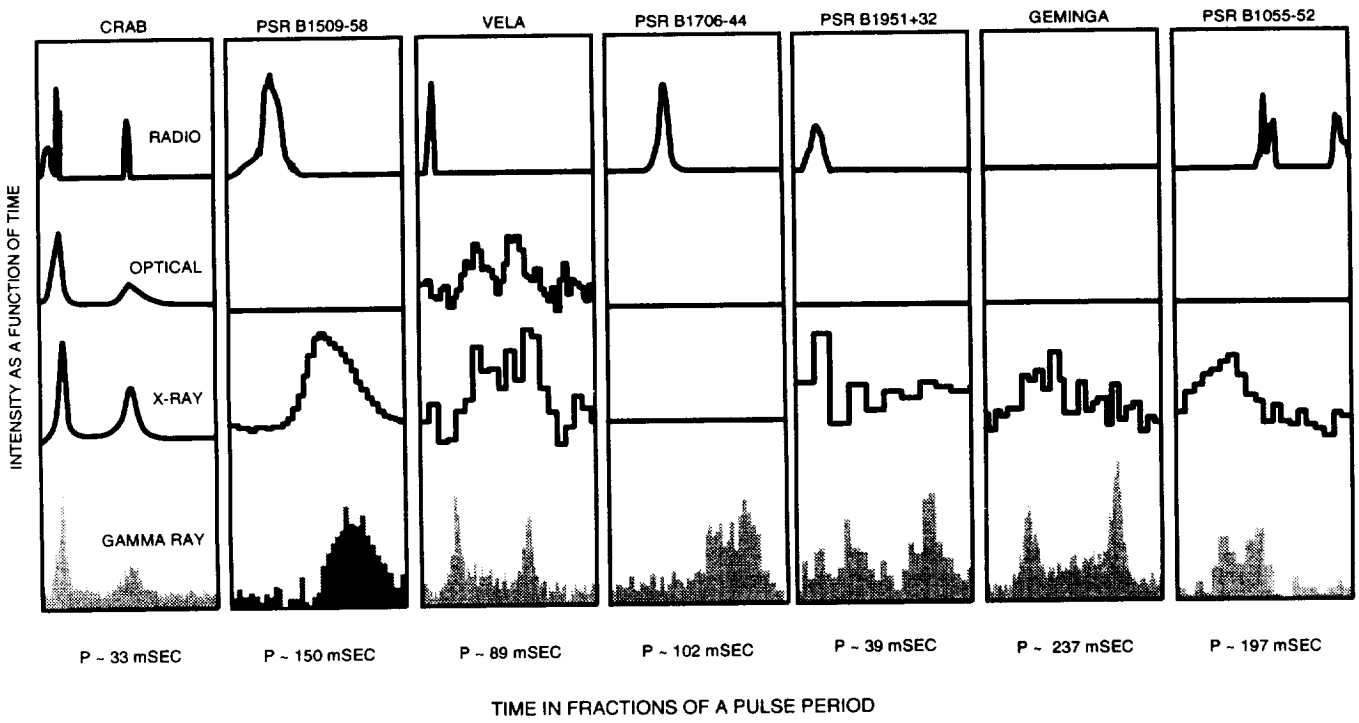


Figure 1.5: Light curves of the six known gamma-ray pulsars. Crab: Manchester, 1971, (radio); Groth, 1975, (optical); Rappaport et al., 1971, (X-ray); Nolan et al., 1993. B1509-58: Ulmer et al., 1993, (radio and gamma-ray); Kawai et al., 1991, (X-ray). Vela: Kanbach et al., 1994, (radio and gamma-ray); Wallace et al., 1977, (optical); Ogelman, 1994, (X-ray). PSR B1706-44: Thompson et al., 1992. Geminga: Bertsch et al., 1992, (gamma-ray), Halpern and Ruderman, 1993, (X-ray). PSR B1055-52: Fierro et al., 1993, (radio and gamma-ray); Ogelman and Finley, 1993, (X-ray).

Table 1.2: Gamma-Ray Pulsar Characteristics

Pulsar	Period(s)	Age(yr)	Luminosity ($erg\ s^{-1}$)	Efficiency
Crab	0.033	1300	4.5×10^{35}	0.001
B1509-58	0.150	1550	1.9×10^{35}	0.010
Vela	0.089	11,000	2.3×10^{34}	0.003
B1706-44	0.102	17,000	3.7×10^{35}	0.011
B951+32	0.040	110,000	1.4×10^{34}	0.004
Geminga	0.237	340,000	2.6×10^{33}	0.080
B1055-52	0.197	530,000	5.0×10^{33}	0.150

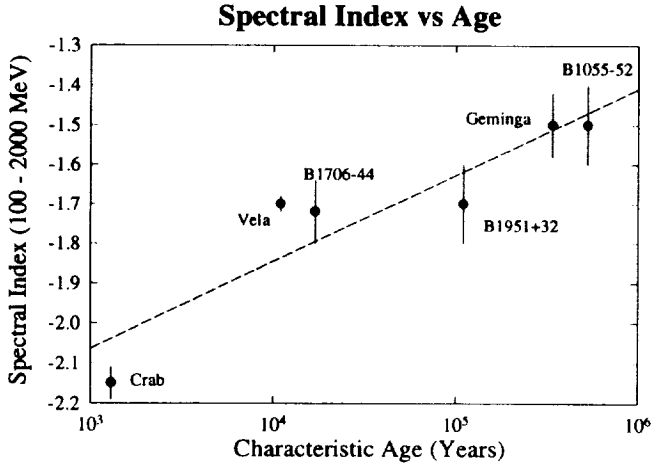


Figure 1.6: The spectral index of the gamma-ray emission from pulsars as a function of their apparent age.

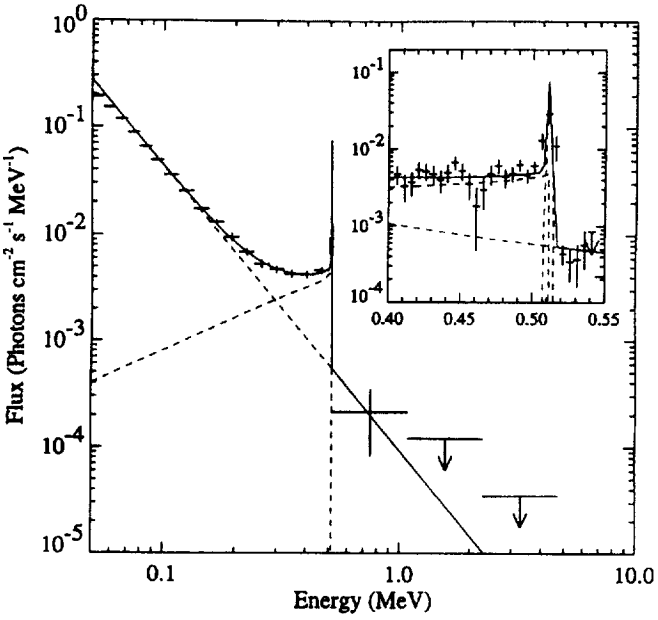


Figure 1.7: Spectrum of the galactic center region in 0.511 MeV annihilation radiation from OSSE.

that we can prove lose energy to gravitational radiation when they are paired in binary systems (Hulse and Taylor, 1975).

Those pulsars from which gamma-rays have been detected are listed in Table 1.2. They are ordered according to apparent age derived from the ratio of their periods to their rate of changes in period ($\frac{p}{dp/dt}$). The latter is the spin-down rate caused by the loss of angular momentum. One can define a gamma-ray efficiency as the ratio of it's luminosity to the energy available from it's spin down. As can be seen in the last column of the table, this efficiency is very strongly correlated with the age. The spectrum of photons from pulsars is typically a power law spectrum, $dN_\gamma/dE = kE^{-\gamma}$. From Figure 1.6, one sees that the spectral index, γ, decreases with increasing age of the pulsar. This implies that a larger fraction of its decreasing spin-down energy is going into higher energy photons as pulsars age. One of the important stages in the amazing life of a pulsar is its emergence as a bright source flashing its beam of gamma-rays at the Earth 5 to 30 times per second.

1.2.3 Gamma-Ray Spectroscopy

The first direct proof that stars build elements and spread them into the interstellar medium comes from the HEAO-3 observation of 1.809 MeV gamma-rays from the decay of ^{26}Al into an excited state of ^{26}Mg (Mahoney et al., 1984). The COMPTEL experiment on CGRO has mapped this radiation in the galactic plane. There is considerable structure, and the overall distribution is somewhat puzzling. It is "clumpy" and does not match well the distribution of the supernova remnants, novae, Wolf-Rayet stars, or red giants that might be expected to be the source of the recently synthesized ^{26}Al (Diehl et al.1994).

CGRO's OSSE experiment has also mapped the distribution of 0.511 keV positron annihilation radiation from the galactic center first discovered by a balloon-borne experiment (see Chapter 6). The OSSE team has observed this source extensively. The spectrum in Figure 1.7 shows the narrow positron emission line and the positronium continuum towards lower energies. OSSE scanned over the galactic center with its 4 degree by 0.5 degree slit and showed that the distribution of the radiation is inconsistent with the distribution of any of the postulated sources of the radiation such as novae. The best fit is a point source at the galactic center combined with sources distributed along the galactic plane.

The COMPTEL instrument on CGRO has discovered line emission at 4.4 and 6.1 MeV coming from the Orion complex. This radiation was predicted in the 70s to be the result of interactions of energetic carbon and oxygen nuclei (Meneguzzi and Reeves, 1975; Ramaty, Kozlovsky, and Lingenfelter, 1979). It is presumably coming from ions accelerated at the shocks formed by the interaction of the winds from the many Wolf-Rayet stars in the complex (e.g., Nath and Biermann, 1994).

CGRO will continue to observe novae, supernovae remnants, AGN, etc. for the next several years. In the latter part of the decade, ESA's International Gamma-Ray Astrophysics Laboratory, Integral, will be launched. It will provide imaging and a much more sensitive spectroscopy instrument. With Integral's launch, gamma-ray spectroscopy can be expected to become a richer and more complex discipline that will explore nuclear emissions throughout the galaxy.

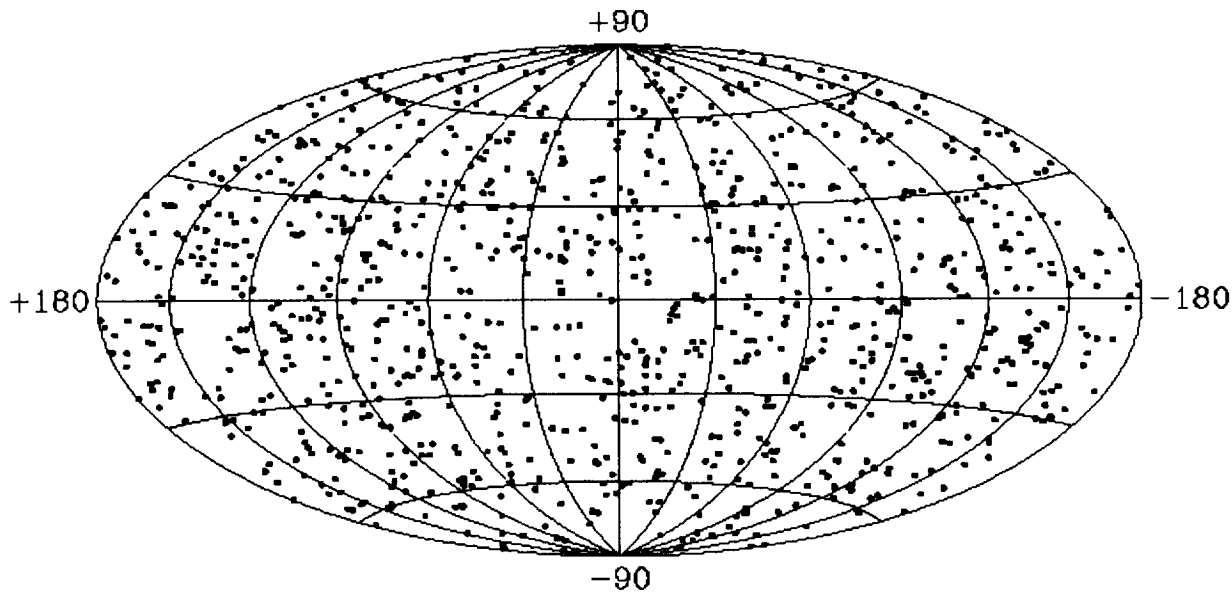


Figure 1.8: All sky map of the first 1000 gamma-ray bursts observed by BATSE on the Compton Gamma-Ray Observatory from April 21, 1991, to May 27, 1994.

1.3 Gamma-Ray Bursts

The most unexpected results from the CGRO have come from its smallest instrument, BATSE. With a set of 8 detectors with 2000 cm^2 sensitive area located on each of the corners of CGRO, BATSE can observe all the sky not occulted by the Earth and, by comparing the response of the 8 detectors can, in real time, determine the position of a gamma-ray burst on the sky with a precision of a few [1] degrees. Between the launch on April 21, 1991, and May 27, 1994, 1000 bursts were observed. The distribution on the sky is shown in Figure 1.8 (Meegan et al., 1992, Fishman, 1994). There is no statistical deviation from isotropy. The bursts range in duration from *ms* to several seconds. Some "light curves" are rich with structure; others are quite simple. If the burst sources were isotropically distributed in space, their intensity vs. size distribution (log N vs. log S) should follow a -3/2 power law. This holds for the largest bursts, but not the smallest, as shown in Figure 1.9. This was known before the launch of CGRO and, along with arguments about energetics, the lack of identifiable bright optical counterparts, and the apparent detection of cyclotron lines in the spectra, led to the tentative association of burst sources with neutron stars. The relative paucity of small bursts was cited as evidence for an edge of the distribution. It was thought that BATSE would confirm the deviation from -3/2 power law in the log(N) vs. log(S) distribution, which it did, and that it would find evidence for the shape of the neutron star distribution in the galaxy, which it did not.

So where are we in understanding gamma-ray bursts? There is clear evidence that the distribution has an "edge", but it's isotropy says we must be at the center of the distribution. This means that the burst sources are either very local or cosmological. The sensitivity of BATSE is such that if the sources are galactic, the scale size is sufficiently large that the solar system is near the center of the distribution; i.e., the sources are distributed

[1] "We would also like to point out that the first observational evidence for an association of gamma-ray bursts with high velocity neutron stars was the coincidence of the March 5th, 1979 burst error box with the edge and not the center of the supernova remnant N49 in the Large Magellanic Cloud, as reported by Tom Cline and his cohorts fifteen years ago. This identification caused a large controversy, since it implied that burst sources could be outside the galaxy and, therefore, more powerful than first assumed. There could be no doubt about Tom's belief in this identification when he attached the personalized license plate "N49" to his car (or mobile stereo system, as he called it) for all to see. In an ironic twist on the excitement created by this first counterpart object in an error box, some today would consider the Large Magellanic Cloud much too close to be the source of gamma-ray bursts."- From the Preface to "High Velocity Neutron Stars and Gamma-Ray Bursts" AIP Conf. Proc. 366, eds, R. Rothschild and R.Lingenfelter, 1995.

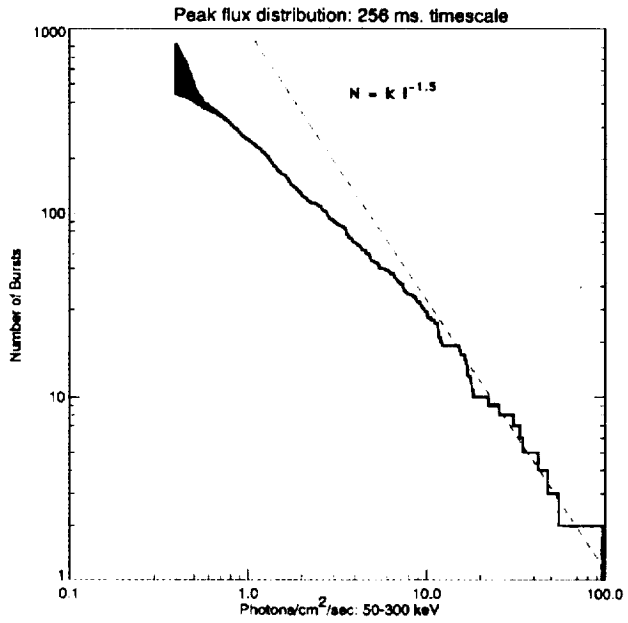


Figure 1.9: The number of bursts observed by BATSE is shown as a function of their intensity. Note the absence of small bursts. The solid band for the very lowest intensity bursts is a consequence of uncertainty in the noise near threshold. Intensity is integrated throughout the range 50-399 keV.

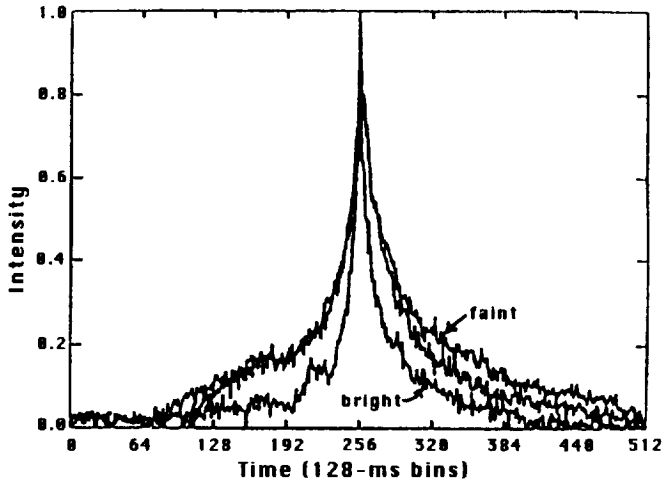


Figure 1.10: Average time profiles for gamma-ray bursts.

through a large region compared to the 8 kpc distance from the Sun to the center of the galaxy. On this size scale, one might see effects of the merging of the Milky Way sources with the distribution around Andromeda, our nearest large neighbor. An Andromeda bulge in the distribution is not observed, but whether such a bulge would be observable has been called into question (e.g., see Duncan, Li, and Thompson, 1993). The distribution could be very local, something associated with our solar system, but there is no known source candidate on this size scale.

Perhaps the burst sources are located at "cosmological" distances (Paczynski, 1986). This is suggested by the logN-log S distribution and the isotropy. However, the absence of any correlation with clusters of galaxies like the Virgo cluster means putting the burst sources at redshift distances, Z, of 1 or more. If this is the case, the energetics require that the sources are very dramatic events such as the coalescence of two neutron stars (Narayan, Paczynski, and Piram, 1992).

Norris and collaborators (Norris et al.1994; Nemiroff et al., 1994) have shown that the burst "light curves" (e.g., durations, time between peaks, corresponding spectral effects) show a correlation with size that is consistent with the redshift effects from placing the sources at redshifts of $Z > 1$ (around $Z = 2$). The results obtained by aligning the burst peaks are shown in Figure 1.10. Great care is taken to make the signal-to-noise ratio the same for the bright and dim bursts in this comparison. Other analyses have shown that the spectral hardness is correlated with intensity, and it evolves throughout each burst and even throughout the substructure of the bursts. The statistical analyses have been repeated and confirmed by others using different statistics, so there is general agreement that the time dilation-like and spectral evolution effects are probably real. The debate rages as to whether the effects are cosmological or something intrinsic about the sources themselves.

But there are problems placing these bursts at cosmological distances. The energy required of a source is of the order of 10^{51} erg. The spectra from coalescing neutron stars are predicted to be "black bodies," but the observed spectra are not. They extend well above 1 MeV in the cases for which observations are available.

The "water has been further muddied" by the observation by EGRET of delayed, very high energy photons from a burst that occurred on Feb. 17, 1994 (Hurley et al., 1994). It was a large burst and was within the EGRET field of view. As shown in Figure 1.11, the burst was observed in time coincidence with a burst recorded on the Ulysses spacecraft. The remarkable thing was the observation by EGRET of continuing high energy emission an hour later when the satellite could again see the source as it reemerged from behind the Earth. Among the observed photons was one with the highest energy recorded from any gamma-ray burst, 25 GeV. Such photons are rare. The probability of observing a photon with > 20 GeV within 1 square degree in position and within 1 hour in time is of the order of 10^{-5}! Explaining these delayed high energy photons is not easy (Meszaros, and Rees, 1994).

Perhaps the origin of gamma-ray bursts can ulti-

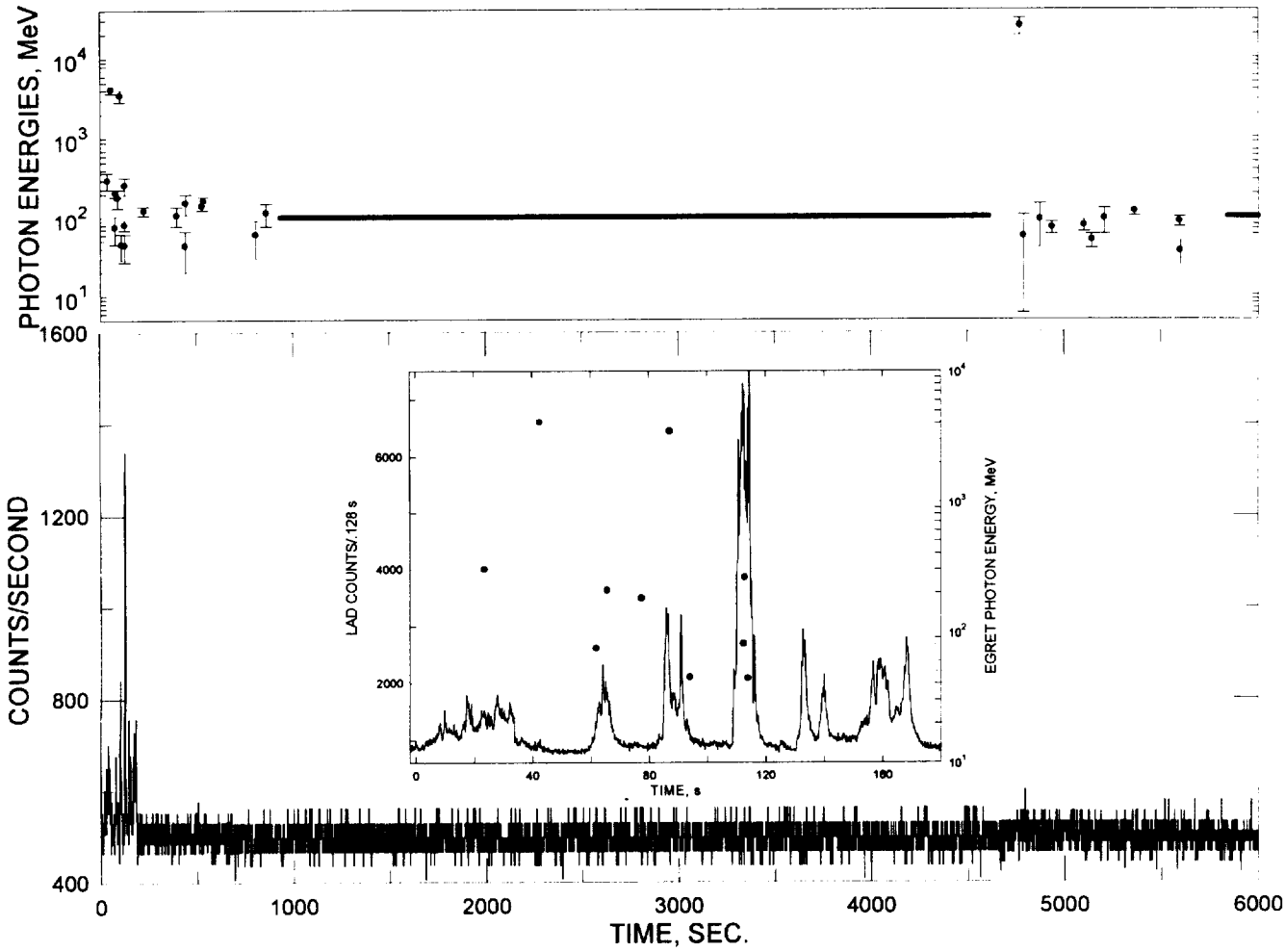


Figure 1.11: Upper panel shows the energy of each individual photon observed by EGRET from the Feb. 17, 1994, gamma-ray burst. The bottom panel shows the count rate from the detector on Ulysses at the same time. Note the EGRET photons after the eclipse.

mately be explained by the more conventional neutron star model. Within the past 2 years, it has been proved that a subset of gamma-ray bursts, the so-called soft gamma repeaters, are associated with neutron stars and supernova remnants (Kulkarni and Frail, 1993). There are three identified repeaters. Two are associated with plerion (filled with radio emission) supernova remnants. The third has a pulsar that is apparently ejected from a nearby neutron star. The implied velocity of the neutron star is of the order of 1000 $km\ s^{-1}$. There is growing realization that high velocity neutron stars may be common and have formed a 100 kpc halo around the galaxy. In this case, our galaxy might perturb the halo around Andromeda and vice-versa, perhaps isotropizing the distribution. These neutron stars are postulated to be driven to large velocity by slight asymmetries in the explosion of the star as it became a supernova. If this proves to be correct, one might say that the galaxy is surrounded by a halo of neutrino-accelerated neutron stars or "rocket stars."

Prior to the launch of CGRO, neutron stars were the generally accepted sources of gamma-ray burst (GRB). One piece of evidence was the presence of spectral features that could be associated theoretically with cyclotron lines from very high magnetic fields. Especially important were features seen at 20 and 40 keV by the Japanese Ginga satellite. Ginga saw a 20 keV feature in one of the ten bursts it observed. Of these, five bursts were strong enough for Ginga to see a 40 keV feature, and such a feature was observed once. BATSE has seen no line features. The strong enough bursts from BATSE number 10 for the 20 keV and 24 for the 40 keV feature. There is a 10% chance that the Ginga and BATSE results are consistent, so more observations are clearly

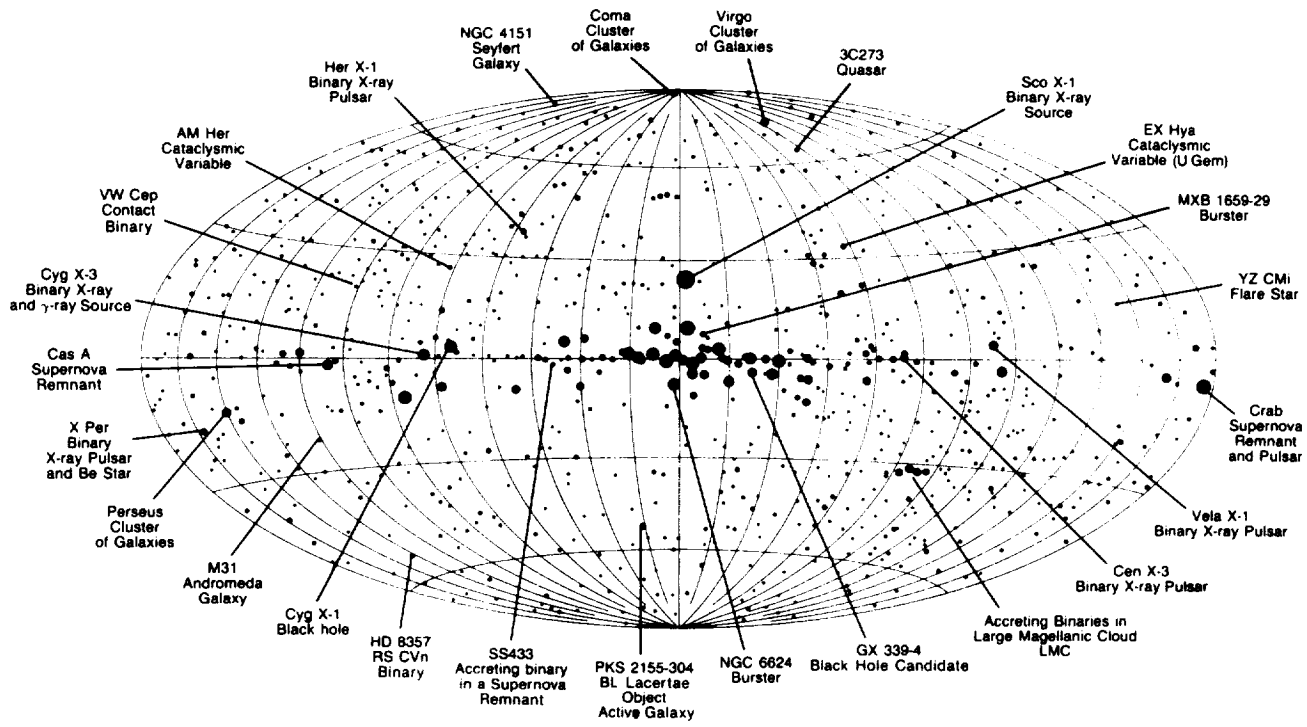


Figure 1.12: The X-ray sky. Map of the X-ray sky known in 1987 plotted in galactic coordinates. Note the concentration toward the galactic center and galactic plane. Bright sources are shown as larger circles. A number of individual sources are identified, both by their respective catalogue designations and by source type. (With the permission of Kent S. Wood and the Naval Research Laboratory, as well as of Jay M. Pasachoff and W. B. Saunders, Co.)

called for in the 10 - 100 keV part of the spectrum. This discrepancy may prove to be real, calling into question the interpretation that some bursts show evidence for cyclotron lines.

So the origin of gamma-ray bursts remains one of the most fascinating puzzles in high energy astrophysics yet to be solved. Whether we are seeing something from the more distant Universe or the taillight beams from neutron star rockets flying around the galaxy, these are interesting objects deserving of future intense study.

1.4 X-RAY ASTRONOMY

The 1990s will be an extremely important decade of progress in high energy astrophysics. The early part will be famous for the discoveries of the CGRO, and the latter part, for progress in X-ray astronomy, radiation from 0.1 to 30 keV.

The X-ray sky, circa 1987, is shown in Figure 1.12 (Wood, 1987). Most of the objects known in astronomy are represented here. Galactic objects include pulsars, supernova remnants, cataclysmic variables, flare stars, and the strongest sources, binary stellar systems. Thermal Bremstrahlung X-rays are a natural emission of plasmas at temperatures between 10^6 and 10^8 Kelvins and such plasmas are ubiquitous. Extragalactic objects are well represented by AGN including quasars and Seyfert galaxies and, especially important cosmologically, the clusters of galaxies. In many classes of objects, e.g., active galaxies, clusters of galaxies, accreting neutron stars, and supernova remnants, a large fraction of the energy is radiated in the X-ray band.

The Einstein X-ray observatory which flew in the late 1970s' discovered diffuse X-ray emission from 80 million degree gas filling the intergalactic space in clusters of galaxies. This decade began with the Granat mission (Russia) and the Broad Band X-Ray Telescope (BBXRT) with its revolutionary thin foil X-ray optics. These optics were further utilized by the Japanese on their ASCA mission launched in 1993 (Tanaka, Inoue, and Holt, 1994). Still to come in this decade are the ESA SAX mission (Italy and The Netherlands) and Spectrum-Roentgen-Gamma (Russia, U.K., Denmark, Germany, US, and Italy), followed by the AXAF and Astro-E (Japan and US).

1.4.1 X-Ray Timing Studies

The most recently launched high energy astrophysics mission, launched in 1995, is the XTE(Bradt et al., 1993). This mission is intended to explore astrophysical clocks in the Universe. The objects being studied and some of the physical processes are indicated in Figure 1.13. A schematic of the spacecraft is shown in Figure 1.14. Its detectors have a field of view of 1 degree and are sensitive to hard X-rays from 2 keV to 200 keV. Studies will range from temporal structure on scales of 1 ms deep in the potential wells of white dwarfs, neutron stars, and black holes to systems with periods of many months, even years. XTE will study chaotic phenomena like accretion and periodic systems like binaries and pulsars. It also has an all sky monitor that can respond to transient events and alert ground controllers to repoint the satellite to observe them in detail as soon as possible. The satellite can be pointed at a typical transient event within a few hours.

GRO J1744-28, better known as the "bursting pulsar," is probably the one you are thinking of. Almost immediately following its launch, XTE was pointed to look at a new, previously unknown, X-ray source that had been discovered by BATSE. The source is GRO J1744-28, better known as the "bursting pulsar. BATSE is able to use the shadow of the Earth to tell where a source is, and XTE was pointed in that direction and found the precise position of the source. XTE was able to study the pattern of time variability in the source and see the signature of matter moving around a neutron star at the inner edge of its surrounding accretion disk. This is the closest view yet obtained at any wavelength of a region of space where the effects of gravity are so profound.

1.4.2 X-ray spectroscopy

X-ray spectroscopy is about to become "of age." In the years since the Einstein Observatory was launched, the technology has developed so that nondispersive techniques obtain resolution competitive with dispersive techniques. In a photon-limited astronomy this makes observations possible in a much shorter time; the energy of every photon can be precisely determined.

The large-area thin-foil mirrors employed first on BBXRT and currently on ASCA help to make possible "faster cameras," and spectroscopy can now be done on a large number of X-ray sources (Serlemitsos, 1988). Figure 1.15 shows a spectrum of Tycho supernova remnant taken with BBXRT (Petre et al., 1993). It is clear that the spectrum contains emission lines from heavy elements in the hot plasma. The abundances can be compared to solar system abundances, and in the case of Tycho, the abundances for the heavy elements are approximately 0.5 solar.

The results from ASCA are even more impressive.

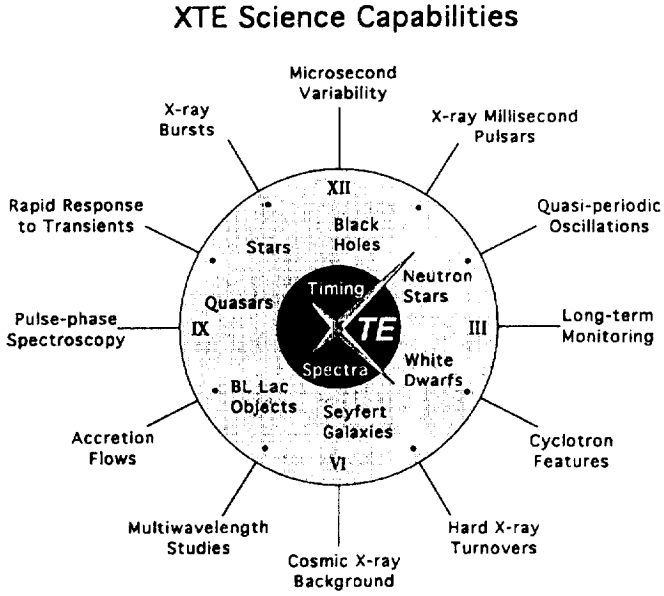


Figure 1.13: The objects to be studied and the scientific issues to be addressed by the X-ray Timing Explorer (XTE).

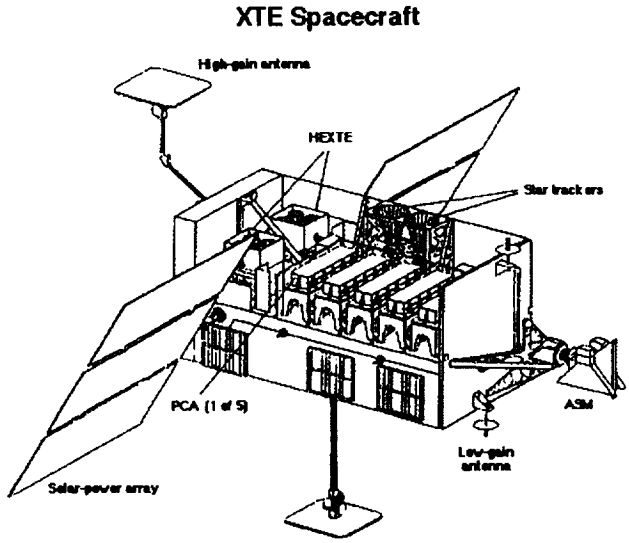


Figure 1.14: The X-ray Timing Explorer spacecraft

An example is shown in Figure 1.16. Results from observations of the cataclysmic variable system, a binary consisting of a white dwarf and a normal star, indicate that the different lines cannot be fit with a single temperature (Ishida et al., 1994). This indicates that the plasma in not in thermal equilibrium and that there is element fractionation in the plasma.

It is now possible to measure Doppler velocities using X-ray lines. Red- and blue-shifted iron lines measured by ASCA from the two jets in SS433 are shown in Figure

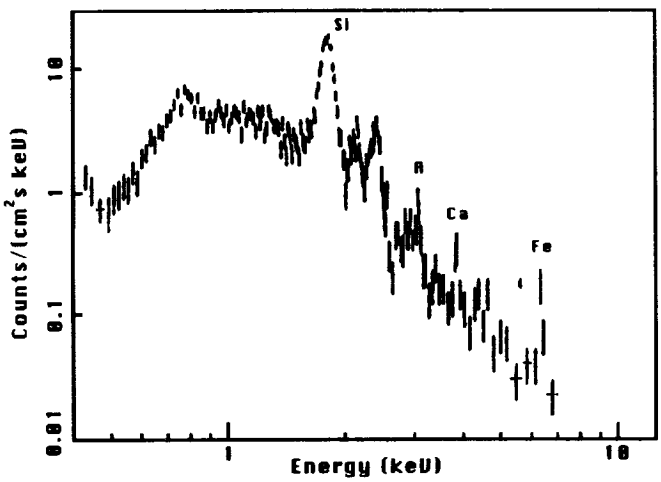


Figure 1.15: X-ray spectrum of the Tycho supernova remnant taken by the BBXRT telescope on the Shuttle's flight of Astro-1.

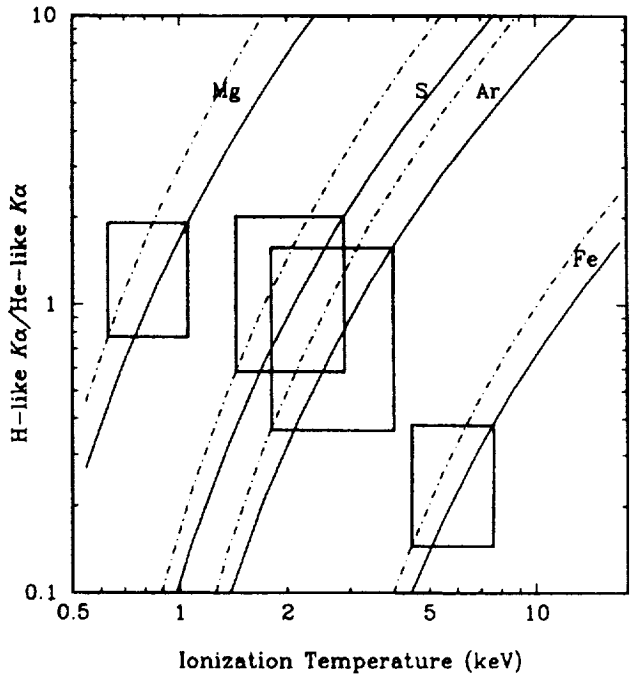


Figure 1.16: Ratio of H-like to He-like Ka emission lines from the cataclysmic variable binary stellar system EX Hydra as a function of ionization temperature.

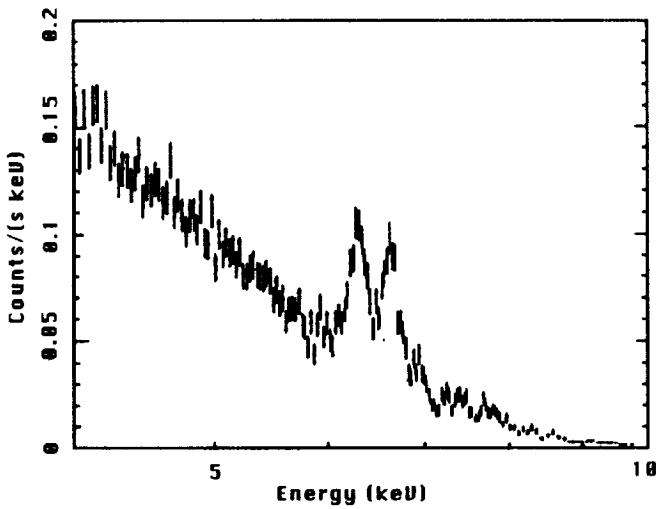


Figure 1.17: Red- and blue-shifted iron lines from SS433 beams as observed by ASCA.

1.17 (Mirabel and Rodriguez, 1994; Kotani et al.1994). X-ray spectroscopy is now becoming as powerful a tool as optical spectroscopy. (You will recall that optical spectroscopy enabled the discovery of the second most abundant element in the Universe by observing spectral lines from the Sun.) The comparison of X-ray and optical redshifts for distant clusters shown in Figure 1.18 turned up an error in the optical redshift for the cluster in Ophiuchus (Mushotzky, 1994).

An improved version of the ASCA optics will be used for the next Japanese-American X-ray astronomy mission now under development, Astro-E. This mission will also carry a new high resolution spectroscopy device known as a quantum calorimeter (Kelley et al., 1993). It is so named because it determines the energy of each incident X-ray photon by measuring the temperature rise in an X-ray absorbing heavy metal. When temperatures of milli-Kelvins are maintained, the resistance of the metal varies as a strong function of its temperature. The resistance can be measured electronically as the device heats up and then cools off. These devices will have resolution, Δ E = 10 eV, a factor of 3 to 10 better than is currently being achieved on ASCA. Plasmas between 0.3 and 10 keV should be rich with line emission reflective of the ionization states, the elemental abundances in the plasmas, and the motions of the plasma deep within the potential wells of binary stellar, neutron star, and black hole systems.

The ASCA has proven the power of spatially resolved spectroscopy, a powerful new tool for studying hot plasmas in astrophysical settings. For example, the ability to resolve the nebula of the supernova remnant SN1006 into a thermal spectrum from the remnant's shock-heated interior and a presumably synchrotron spectrum in a surrounding shell has provided the first evidence for the acceleration of particles to energies greater than 1 TeV (10^12 eV) in an astrophysical environment (Koyama et al., 1995). ASCAs other results are too many and detailed to describe further here. The reader is referred to Chapter 4 for results on super-

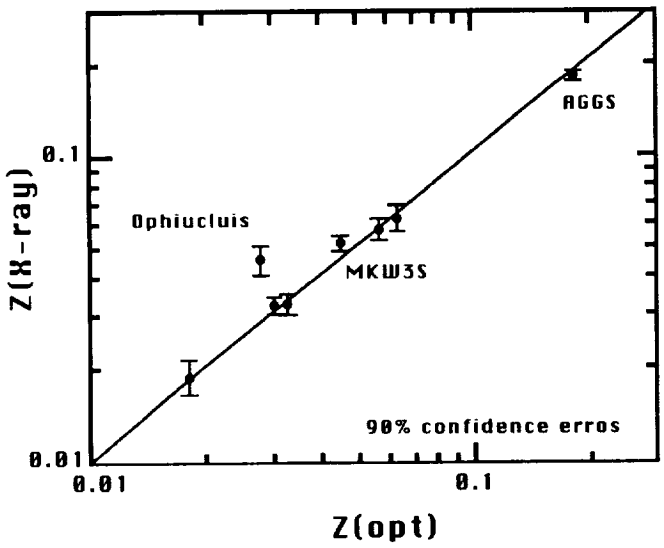


Figure 1.18: Comparison of ASCA X-ray redshifts with optical redshifts.

nova remnants, to Chapter 9 for spectroscopy of discrete sources, and to Chapter 13 for a discussion of the measurements of the hot gas surrounding clusters of galaxies and their implications for the distribution of both dark and baryonic matter and the elemental abundances of the latter.

1.4.3 X-ray Imaging

No discussion of the prospects for X-ray astronomy would be complete without mention of the Advanced X-ray Astronomy Facility, an imaging X-ray telescope that will do for X-ray astronomy what the Hubble Space Telescope is now doing for optical astronomy. Its mirrors are grazing incidence reflectors. They are nested sets of parabolas and hyperbolas made with the smoothest surfaces ever polished.

The capabilities and scientific prospects of AXAF are legion, and a new version of a reference book such as this that is written a decade from now will be full of its exciting discoveries. The reader is referred to the review of AXAF (Zombeck, 1983) for more information.

1.5 CONCLUSION

The beginning of this decade has been characterized by important new discoveries using observations at gamma-ray wavelengths. Some of them have been described briefly in this paper. The remainder of the decade will be characterized by a renaissance in X-ray astronomy, the results of which cannot be predicted. But the added understanding of astrophysical objects that comes from observations at all wavelengths has been proven by the results from CGRO. While X-ray and gamma-ray astrophysics are characterized by low fluxes of individual quanta, each quantum carries a significant energy. The death throes of the objects may be found in their high energy emissions.

1.6 REFERENCES

Bertsch, D.L., Dame, T.M., Fichtel, C.E., Hunter, S.D., Sreekumar, P., Stacy, J. G., and Thaddeus, P., 1993, ApJ. 405, L21.

Bertsch, D.L., et al., 1992, Nature 357, 306.

Bradt, H.V., Rothschild, R.E., and Swank, J.H., 1993, Astron. and Astrophysics: Suppl. 97, 355.

Cheng, K.S., Ho, and Ruderman, M.A., 1986, ApJ, 300, 500.

Dermer, C.D., and Gehrels, N., 1995, ApJ, 447, 103.

Daugherty, J.K., and Harding, A.K., 1982, ApJ, 252, 337.

Daugherty, J.K., and Harding, A.K., 1994, ApJ, 429, 325.

Diehl, et al., 1994, Astron. and Astrophy. ApJ, Suppl. 92, 429.

Duncan, Li, and Thompson, D.J., 1993, Compton Gamma-Ray Observatory, AIP Conference Proceedings 280, eds: M. Friedlander, N. Gehrels, and D.J. Macomb, (NY: AIP), 1074.

Fichtel, C.E., et al., 1994, ApJ, Suppl. 94, 551.

Fierro, J.M., et al., 1993, ApJ, 413, L27.

Fishman, G., 1994, NASA/MSFC release.

Gehrels, N., Chipman, and Kniffen, D.A., 1993, ApJ, Suppl. 92, 351.

Gehrels, N. and Dermer, C.D., 1994, private communication.

Groth, E.J., 1975, ApJ, Suppl. 293, 431.

Halpern, J.P., and Ruderman, M.A., 1993, ApJ, 415, 286.

Hartman, R.C., et al., 1992, ApJ, 385, L1.

Hulse, R.A. and Taylor, J.H., 1975, ApJ, 195, L51.

Hurley, K., et al., 1994, Nature, 372, 652.

Ishida, et al., 1994, Publ. Astron. Soc. Japan, 46, L81.

Kanbach, G., et al., 1994, Astron. and Astrophy. 289, 855.

Kawai, N., et al., 1991, ApJ, 383, L65.

Kelley, et al., 1993, Journal of Low Temperature Physics, 93, 225.

Kniffen, D.A., et al., 1993, ApJ, 411, 133.

Kotani, et al., 1994, Publ. Astron. Soc. Japan, 46, L147.

Koyama, K., Petre, R., Gotthelf, E.V., Hwang, U., Matsuura, M., Ozaki, M., and Holt, S.S., 1995, Nature, 378, 255.

Kulkarni, S.R., and Frail, 1993, Nature, 365, 33.

Leventhal, M., MacCullum, and Stang, 1978, ApJ, 225, L11.

Mahoney, W.A., et al., 1984, ApJ, 286, L578.

Manchester, R.N. 1971, ApJ, 163, L61.

Meegan et al., 1992, Nature, 355, 143.

Meneguzzi and Reeves, 1975, Astron. and Astrophy. 40, 91.
Meszaros, P., and Rees, 1994, MNRAS, 269, L41.
Mirabel, I.F., and Rodriguez, L.F., 1994, Nature, 371, 46.
Mushotzky, R.F., 1994, private communication.
Mushotzky, R.F., Done, and Pounds, 1993, Annual Rev. of Astronomy and Astrophysics, 31, 717.
Narayan, Paczynski, and Piram, 1992, ApJ, 395, L83.
Nath and Biermann, 1994, MNRAS, in press.
Nemiroff, R.J., et al., 1994, ApJ, 435, L133.
Nolan, P.L., et al., 1993, ApJ, 409, 697.
Norris, J.P., et al., 1994, ApJ, 424, 540.
Ögelman, H.B., and Finlay, 1993, ApJ, 413, L31.
Ögelman, H.B., 1994, in "Lives of Neutron Stars" NATO ASI, Kemer, Turkey, in press.
Paczynski, B., 1986, ApJ, 308, L43.
Petre, R., et al., 1993, Proc. 10th Int'nl Colloquium UV and X-ray Spectroscopy of Laboratory and Astrophysical Plasmas, eds. Silver and Kahn, 424.
Ramaty, R., Kozlovski, and Lingenfelter, 1979, ApJ, Supp., 40, 487.
Rappaport, S., Bradt, H., and Mayer, W., 1971, Nature, 229, 40.
Romani, R.W., and Yadigaroglu, I.-A., 1994, ApJ, 438, 314.
Serlemitsos, P.J., et al., 1988, Appl. Opt. 27, No. 8, 1'447
Sturner, S.J. and Dermer, C.D., 1994, ApJ, 420, L79.
Tanaka, Y., Inoue, and Holt, 1994, Publ. Astron. Soc. Japan, 46, L37.
Thompson, D.J., et al., 1992, Nature, 359, 615.
Ulmer, M.P., et al., 1993, ApJ, 413, 738.
von Montigny, et al., 1995, ApJ, in press.
Wallace P.T., et al., 1977, Nature, 266, 692.
Wood, K.S., 1987, in "Contemporary Astronomy" ed. J.M. Pasachoff, Saunders College Publishing, The Dryden Press, 426.
Zombeck, 1983, Adv. in Space Research, 2, 259, see also special issue, 1987, Astro. Lett. and Communications, 26, 1 and 1994 CfA preprint #4003, International School of Space Science Course on X-ray Astronomy, L'Aquila, Italy, Aug. 29-Sept. 10, 1994.

Chapter 2

GAMMA-RAY CONTINUUM AND MILLIMETER WAVE EMISSIONS FROM SOLAR FLARES

REUVEN RAMATY
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

We present a theoretical discussion on the methods for correlating gamma-ray continuum and millimeter wave emissions from solar flares. Both of these emissions are produced by electrons of similar energies, typically around 1 MeV, interacting in the solar atmosphere. The gamma-rays are bremsstrahlung and the millimeter waves are gyrosynchrotron radiation. Thin and thick target interaction and radiation models for the accelerated particles are presented. The comparison of the models with observations of gamma-rays and 80 GHz radio emission from the 4 June 1991 flare provides information on a variety of parameters, including the magnetic field in the corona and the variation of the accelerated electron spectrum during the course of the flare.

2.1 INTRODUCTION

The interactions of accelerated particles with matter and magnetic fields in the solar atmosphere produce a variety of electromagnetic emissions generally referred to as nonthermal radiations to distinguish them from radiations produced by particles in thermal or quasithermal equilibrium. Although acceleration mechanisms were not discussed in this lecture, we mention here that the leading processes in flares are thought to be acceleration by shock waves (Ellison and Ramaty, 1985), acceleration due to gyroresonant interactions with plasma turbulence (Miller and Steinacker, 1992) and acceleration by large scale electric fields (Benka and Holman, 1994). The reader should also consult the earlier review on the problem by Forman, Ramaty, and Zweibel (1986).

Nonthermal flare radiations are observed both at high energies, as hard X-rays and gamma-rays, and at very low energies, as radio waves extending in frequency up to the millimeter region. The bulk of the hard X-rays are bremsstrahlung (braking radiation) produced by accelerated electrons interacting with both neutral and ionized gas in the solar atmosphere (e.g., Dennis, 1985). This bremsstrahlung extends from about 20 keV, where it merges with the thermal bremsstrahlung produced by hot flare plasma, well into the gamma-ray region (Rieger, 1989). In the energy region from about 0.4 to 8 MeV there are nuclear lines superimposed on the bremsstrahlung continuum. This nuclear line emission results from the interactions of accelerated protons, α particles and heavier nuclei with the ambient solar atmosphere. As nuclear line emission was not discussed in this lecture, the reader is referred to the reviews by Ramaty and Murphy (1987), Mandzhavidze and Ramaty (1993) and Ramaty and Mandzhavidze (1994).

A large variety of nonthermal radio emission are observed from solar flares (e.g., Dulk, 1985). At the highest frequencies, i.e., in the microwave and millimeter regions, and in particular for relatively large flares, the observed radio emission is gyrosynchrotron radiation produced by electrons in the energy range from several hundreds of keV to several MeV. Since the same electrons also produce bremsstrahlung gamma-ray continuum, there is a close connection between these two types of nonthermal solar flare emissions. This was the topic of the present lecture.

A series of 6 very large (X-class) solar flares occurred in June 1991 and gamma-ray emission has been observed from all of them. These very large flares also produced the most intense millimeter emission ever observed from the Sun. The details of these observations can be found in Ramaty et al., 1994. In the present paper we explore the relationship between the gamma-ray continuum and the millimeter wave emission observed around the peak of the impulsive phase of the 4 June flare. We are interested in the highest frequency radio emission because the associated gamma-ray emission is necessarily available only for large flares and for these the radio emission will be optically thin only at the highest frequencies.

We discuss accelerated particle interaction models in Sec.2.2, bremsstrahlung in Sec.2.3, gyrosynchrotron radiation in Sec.2.4, and we compare the calculation with the data in Sec.2.5.

2.2 ACCELERATED PARTICLE INTERACTION MODELS

The transport and interactions of accelerated particles at the Sun are governed by the complex magnetic structures that are known to exist in the solar atmosphere. Images of the solar corona, obtained by observing the Sun in soft and hard X-rays (e.g., by instruments on the YOHKOH satellite; e.g., Enome and Hirayama, 1994), reveal loop like structures. The magnetic field lines that define these loop are anchored in the photosphere and extend into the corona. These loops are thought to play a dominant role in the transport and interactions of the accelerated particles (e.g., Ramaty et al., 1990). The particles are probably accelerated in the coronal regions of the loops. But they can produce nonthermal radiation both in the coronal and subcoronal regions. The ambient gas density and the magnetic field are highly variable throughout the loops; typical values are $B_\perp = 200$ Gauss and $n_H = 10^{10}$ cm^{-3} in the corona and $B_\perp = 1000$ Gauss and $n_H = 10^{17}$ cm^{-3} in the photosphere at the foot points of the loops. Here $B_\perp = \sqrt{2/3}B$ is the perpendicular component of the magnetic field, appropriate for an isotropic accelerated electron distribution; n_H is the total (neutral plus ionized) hydrogen density. We also consider the ambient He, with a He to H ratio by number of 0.1. For the variation of the magnetic field with density we adopt the simple relationship,

$$B_\perp(Gauss) = 20\left[n_H(\ cm^{-3})\right]^{0.1} \qquad (2.1)$$

which satisfies the above boundary conditions.

High energy electrons interacting with ambient gas and magnetic fields lose energy as a result of Coulomb interactions, bremsstrahlung and synchrotron radiation. The expressions for these energy loss rates are well known (e.g., Skibo, 1993). In Figure 2.1 we plot the quantity $\tau_l = E/\frac{dE}{dt}$ for various ambient densities and magnetic fields given by Equation (2.1); here E is the energy of the electron, $\frac{dE}{dt}$ is the energy loss rate, and τ_l provides a measure of the electron lifetime against the energy losses. All the solid curves are for losses in a neutral medium. At the density of 10^{10} cm^{-3}, which is typical of the corona, we also show τ_l in an ionized gas (dashed curve). We see that at low energies, where Coulomb losses dominate, τ_l in the

ionized case is smaller by about a factor of 2 than in the neutral case. At the high energies, however, synchrotron losses dominate and these do not depend on the state of ionization of the ambient medium. Because of the weak dependence of the magnetic field on density, as the ambient density increases, the transition from Coulomb and bremsstrahlung dominated losses to synchrotron dominated losses, manifested by the turnover of τ_l, occurs at increasingly higher energies. Thus, while in the corona ($n = 10^{10}\ cm^{-3}$) synchrotron losses will dominate already at a few MeV, near the photosphere synchrotron losses are negligible practically at all energies.

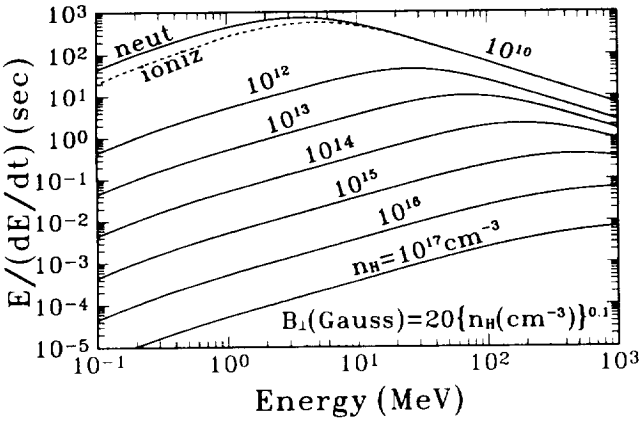


Figure 2.1: Lifetime of accelerated electrons against Coulomb, bremsstrahlung and synchrotron energy losses in the solar atmosphere; for each density the magnetic field is computed from the displayed equation.

We now consider a simplified transport model referred to as the trap and precipitation model. In this model the electrons, accelerated in the coronal region of a loop, are trapped there for a time τ, and subsequently precipitate into the subcoronal regions where they lose their energy. The trapping time τ depends on the magnetic field structure and the energy density in plasma turbulence. We do not consider these issues here but refer the reader to the detailed discussion given in Ramaty and Mandzhavidze (1994). We see from Figure 2.1 that for a coronal density of $10^{10}cm^{-3}$, if τ is less than about 10 s, electrons over the entire range from 0.1 to 1000 MeV will escape from the coronal trap without significant energy loss; in the energy range from 1 to 10 MeV, electrons can be trapped for as long as several hundreds of seconds without losing much energy.

A region from which the electrons escape without losing their energy is referred to as thin target region. Assuming that the coronal trap is a thin target, the differential electron number $N_c(E,t)$ in the region, measured in electrons MeV^{-1}, is given by

$$N_c(E,t) = \int_{-\infty}^{t} dt' Q(E,t') e^{-\int_{t'}^{t} \frac{dt''}{\tau(E,t'')}} \qquad (2.2)$$

where $Q(E,t)$ is the differential electron source function in the corona measured in electrons $MeV^{-1}\ s^{-1}$. This source function is the product of the acceleration, i.e., its energy and time dependence depending on the acceleration mechanism and the physical conditions under which it operates. But independent of the variability of the source function Q, Equation (2.2) ensures that N_c will not vary on time scales shorter than τ.

We refer to a region in which the electrons remain trapped for a time long enough to lose all of their energy as a thick target. The subcoronal regions, particularly those at the higher densities, are thick targets. As the electrons are assumed to be accelerated in the coronal region, the source function for the subcoronal regions is given by the electrons escaping from the coronal region, i.e.,

$$Q_{sc}(E,t) = \left[\tau(E,t)\right]^{-1} N_c(E,t) \qquad (2.3)$$

If the energy loss time in the subcoronal region is shorter that the characteristic time scale of the variations of $Q_{sc}(E,t)$, we can write

$$N_{sc}(E,t) = \left[\frac{dE}{dt}(E)\right]^{-1} \int_E^{\infty} dE' Q_{sc}(E',t) \qquad (2.4)$$

where $N_{sc}(E,t)$ is the time and energy dependent electron number in the subcoronal region. By substituting Equation (2.3) into Equation (2.4), and by assuming that τ is independent of energy, we obtain

$$N_{sc}(E,t) = \left[\frac{dE}{dt}(E)\right]^{-1} \tau(t)^{-1} \int_E^{\infty} dE' N_c(E',t) \qquad (2.5)$$

Equation (2.5) relates the coronal and subcoronal electron numbers. As mentioned, this equation is valid if N_c/τ remains essentially constant over time scales of the order of the energy loss time in the thick target region, τ_l. As can be seen from Figure 2.1, τ_l depends on the electron energy and the ambient density. For example, for a density of $10^{14}cm^{-3}$, in the 1–10 MeV range, τ_l is about 0.1 s. As N_c will not vary on time scales shorter than τ (Equation 2.2), for this density Equation (2.5) will be valid if the trapping time in the corona is longer than a few tenths of seconds.

Before considering in the next two sections the radiations produced by the electrons, we mention here that the concepts of thin and thick target refer to the electrons and not to the photons that they produce. In our subsequent calculations we shall assume that both the bremsstrahlung and the gyrosynchrotron radiation are optically thin. For the bremsstrahlung this assumption

is valid unless the flare is located at or behind the limb of the Sun. This is not so for the gyrosynchrotron radiation which can be strongly absorbed even if the flare is located near the center of the solar disk. However, by limiting our discussion to the very high frequency of 80 GHz, we can safely assume that the gyrosynchrotron emission is also optically thin.

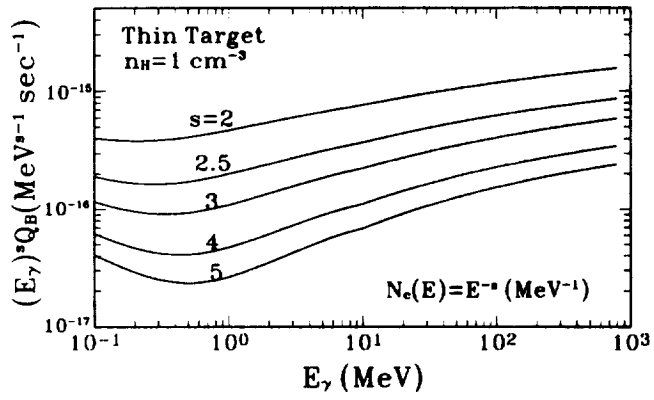


Figure 2.2: Thin target bremsstrahlung spectra; the differential number of accelerated electrons is displayed in the figure.

2.3 BREMSSTRAHLUNG

We calculate the thin target bremsstrahlung produced in the coronal region and the thick target bremsstrahlung produced in the subcoronal region. We consider only bremsstrahlung produced by electrons having an isotropic momentum distribution. The reader is referred to Dermer and Ramaty (1986) and Miller and Ramaty (1989) for anisotropic bremsstrahlung calculations. The thin target coronal bremsstrahlung is given by

$$Q_B(E_\gamma,t) = n_H \int_{E_\gamma}^{\infty} dE v(E) N_c(E,t) \frac{d\sigma}{dE_\gamma}(E,E_\gamma) \tag{2.6}$$

where $v(E)$ is electron velocity, N_c is the differential electron number in the corona, and $\frac{d\sigma}{dE_\gamma}(E,E_\gamma)$ is an effective angle integrated differential bremsstrahlung cross section. This cross section is the combination of the non-screened electron-proton (Koch and Motz, 1959) and electron-electron (Haug, 1975) cross sections, and it is adjusted for the presence of He (see Skibo, 1993 for more details). The differential bremsstrahlung production rate Q_B is measured in photons $MeV^{-1}\ s^{-1}$.

The thick target subcoronal bremsstrahlung is given by an expression analogous to Equation (2.6), with N_c replaced by N_{sc} given by Equation (2.5). By making the substitution we obtain

$$Q_B(E_\gamma,t) = [\tau(t)]^{-1} \int_{E_\gamma}^{\infty} dE v(E) \frac{d\sigma}{dE_\gamma}(E,E_\gamma) \times \left[\frac{1}{n_H}\frac{dE}{dt}(E)\right]^{-1} \int_E^{\infty} dE' N_c(E',t) \tag{2.7}$$

Thin target bremsstrahlung spectra for $N_c = E^{-s}$ and $n_H = 1cm^{-3}$ are shown in Figure 2.2 for various values of the spectral index s. Thick target bremsstrahlung spectra for $N_c/\tau = E^{-4}$ and one thin target spectrum for $n_H = 10^{10}cm^{-3}$ and $N_c = 10E^{-4}$ are shown in Figure 2.3. The thick target spectra are independent of ambient density as long as the synchrotron energy losses are negligible. At energies where the Coulomb losses dominate, the thick target spectrum is flatter than the thin target spectrum because of the flattening of the electron spectrum in the thick target due to these energy losses. The ratio of the thin to thick target spectra are proportional to τ. As can be seen, for the coronal density of $10^{10}\ cm^{-3}$ and $\tau = 10\ s$, the thin target bremsstrahlung is negligible in comparison to the thick target bremsstrahlung. At 1 MeV these two bremsstrahlung components will become comparable only for a trapping time of about a few hundreds seconds.

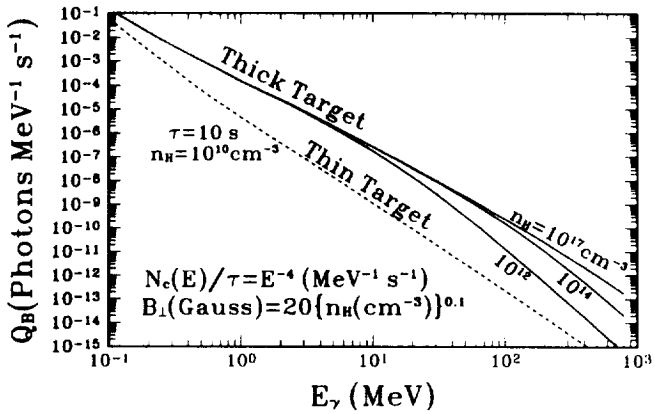


Figure 2.3: Solid curves: thick target bremsstrahlung spectra for various ambient densities and an electron spectrum with power law spectral index $s = 4$; dashed curve: thin target spectrum for a trapping time of 10 s and ambient density $10^{10}\ cm^{-3}$.

2.4 GYROSYNCHROTRON EMISSION

A detailed treatment of gyrosynchrotron emission and absorption was presented by Ramaty (1969). Here one is limited to considerations of optically thin gyrosynchrotron emission and apply our results to millimeter

data at 80 GHz where the emission is clearly optically thin.

In analogy to Equations (2.6) and (2.7) we have that

$$F_{GS}(\nu,t) = R^{-1}\int_0^\infty dE N_c(E,t)\frac{dG}{d\nu}(\nu,\theta,E,B) \quad (2.8)$$

and

$$F_{GS}(\nu,t) = R^{-1}[\tau(t)]^{-1}\int_0^\infty dE\frac{dG}{d\nu}(\nu,\theta,E,B)\times$$
$$\left[\frac{dE}{dt}(E)\right]^{-1}\int_E^\infty dE' N_c(E',t) \quad (2.9)$$

for the thin and thick target cases, respectively. Here $F_{GS}(\nu,t)$ is the gyrosynchrotron flux density at Earth, measured in $erg\ cm^{-2}\ s^{-1}\ Hz^{-1}$, ν is frequency in Hz, R is the Sun–Earth distance, θ is angle between the direction of observation and the magnetic field assumed to be uniform, and $\frac{dG}{d\nu}(\nu,\theta,E,B)$, measured in $erg\ \mathrm{sr}^{-1}\ s^{-1}\ Hz^{-1}$, is the angle-dependent gyrosynchrotron emission from an isotropic electron distribution normalized to 1 electron; $\frac{dG}{d\nu}$ can be calculated by evaluating a series of Bessel functions as described in detail in Ramaty (1969).

We show results for the thin and thick target cases in Figure 2.4, where the coronal electron number N_c and trapping time τ are the same as those used in the bremsstrahlung calculations. We assume that the angular distribution of the electrons, both in the thin and thick target cases, is isotropic. This simplifying assumption is probably valid for the coronal emission, but could lead to the overestimation of the thick target emission because in the subcoronal regions the electron distribution could be downward peaked toward the photosphere, a geometry that would produce less gyrosynchrotron emission in the direction of the Earth than would an isotropic distribution. Gyrosynchrotron radiation from anisotropic electrons is calculated in Ramaty (1969). The calculations of Figure 2.4 do not take into account the variety of absorption and suppression effects that cut off the gyrosynchrotron emission at low frequencies. We do not consider these effects here. However, as already mentioned, at 80 GHz where we shall compare the calculations with data, the solar flare gyrosynchrotron source is most likely optically thin, allowing us to carry out this comparison.

We further note in Figure 2.4 that, unlike the thick target bremsstrahlung, which over a broad energy range is independent of the ambient density, the thick target gyrosynchrotron emission decreases with increasing density. This is caused by the much slower increase with depth of the magnetic field than that of ambient density (see Equation 2.1). Consequently, the deeper the interaction site the more likely it is that the electrons will lose their energy by Coulomb interactions and bremsstrahlung rather than by gyrosynchrotron radiation. The production site of bremsstrahlung at MeV energies in the solar atmosphere can currently only be determined by calculations (because imaging observations are not available). The calculations are of course model dependent (i.e., Miller and Ramaty, 1989), but they do show that the bulk of the bremsstrahlung is probably produced at densities of $10^{14}\ cm^{-3}$ or greater. From Figure 2.4 we then see that the thick target gyrosynchrotron radiation produced in the subcoronal regions is essentially negligible in comparison with the thin target gyrosynchrotron radiation produced in the coronal region for a trapping time in the corona of the order of 10 s. In the next section we shall compare gamma-ray continuum and millimeter wave observations assuming thick target bremsstrahlung and thin target gyrosynchrotron emission.

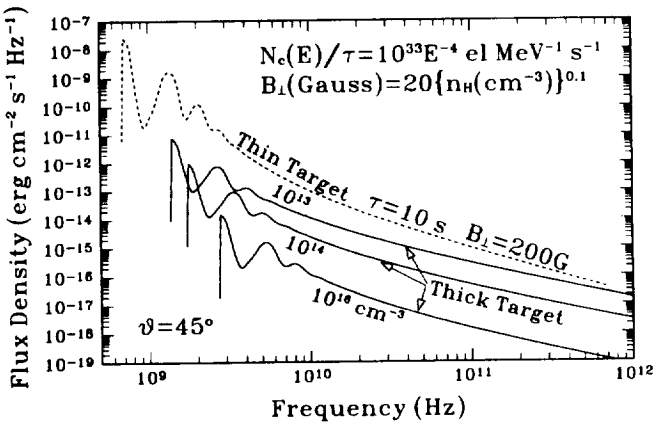


Figure 2.4: Thin and thick target gyrosynchrotron spectra for the same parameters as in Figure 2.3; the magnetic field is assumed to be uniform and the accelerated electrons are isotropic; θ is the angle between the magnetic field and the direction of observation; all absorption effects are ignored.

2.5 COMPARISON WITH DATA

We consider the 4 June 1991 flare from which both gamma-rays and 80 GHz emission were observed. A detailed description of these and other related data is given in Ramaty et al., 1994. It was shown in that paper that during an impulsive peak lasting for about a few tens of seconds, the bulk of the gamma- rays observed with the Charged Particle Detector (CPD) of the BATSE instrument on the Compton Gamma-Ray Observatory and the 80 GHz emission were produced by electrons of similar energies, typically around a few MeV. In the top panel of Figure 2.5 we show the time profiles of the 80 GHz flux, $F_{GS}(80\ GHz)$, and the greater than 1 MeV bremsstrahlung flux $\Phi_B(E_\gamma > 1\ MeV)$. The 80 GHz

flux, observed at the Nobeyama Radio Observatory in Japan, is measured in solar flux units (1SFU = 10^{-19} $erg\ cm^{-2}\ s^{-1}\ Hz^{-1}$); the gamma-ray flux is measured in photons $cm^{-2}\ s^{-1}$. In terms of Equations (2.6) or (2.7),

$$\Phi_B(E_\gamma > 1\ MeV\ ,t) = \left[4\pi R^2\right]^{-1} \int_{1\ MeV}^{\infty} dE_\gamma Q_B(E_\gamma, t) \tag{2.10}$$

We now show that the observed gamma-ray and millimeter fluxes are consistent with the theoretical consideration of the previous sections. We also suggest interpretations for the observed relative displacement of time profiles seen in the top panel of Figure 2.5. The middle panel of the Figure 2.5 shows the ratio of the two emissions. We see that the displacement translates into a dip in the ratio, with a minimum achieved at the time of maximum of the bremsstrahlung flux. We assume that the 80 GHz emission is thin target gyrosynchrotron emission produced in the coronal region of a loop or loops, and that the gamma-rays are thick target bremsstrahlung produced in the subcoronal regions. We thus calculate F_{GS}/Φ_B using Equations (2.7), (2.8), and (2.10). This ratio depends on the magnetic field in the coronal region, on the spectrum of the accelerated electrons, and on the trapping time τ. From Equations (2.7) and (2.8) we see that the ratio depends linearly on τ. However, the dependence on B and s is more complicated.

In the lower panel of Figure 2.5 we plot F_{GS}/Φ_B calculated from Equations (2.7), (2.8), and (2.10) for $\tau = 15$ s. We have fixed this trapping time so that the calculated ratio as a function of time is about equal to the observed ratio as a function of spectral index for $B = 250$ Gauss (this is a total B, not $B_\perp$). By varying s from about 4.5 to 3 we see that F_{GS}/Φ_B decreases from about 800 to 300 which could reproduce the observed decrease during the rise of the emissions from about 3:41UT to 3:41:15UT. The hardening of the electron spectrum during the rising portions of the emissions is expected, as efficient particle acceleration must occur during this period. The choice of $\tau = 15$ $s$is not arbitrary, as the trapping time cannot be much longer than the characteristic time scale of the variations of both the 80 GHz and gamma-ray fluxes, which are on the order of 15 s. It thus appears that the decline of the emissions after their maximum is controlled by the trapping and not the acceleration. In this case the spectrum should remain unchanged orharden with time. Therefore, it is unlikely that the increase of F_{GS}/Φ_B after its minimum at 3:41:15 UT is due to the softening of the electron spectrum. Ramaty et al., 1994 suggested that the entire variation of F_{GS}/Φ_B is due to a variation in the trapping time τ. The physical process that could lead to such a variation is the damping of plasma turbulence in the loop caused by the acceleration of the particles (Ramaty et al., 1994; Ramaty and Mandzhavidze, 1994). In principle, the variation could also be the consequence of changes in B. It is thus possible that the observed displacement of the bremsstrahlung and gyrosynchrotron time profiles is caused by a combination of all three of these effects. It is important, however, to remember that the data overall can be understood with very reasonable values for the parameters, i.e., a coronal magnetic field of about 250 Gauss, a coronal trapping time of about 15 s and an electron spectral index hardening during the rise of the emissions.

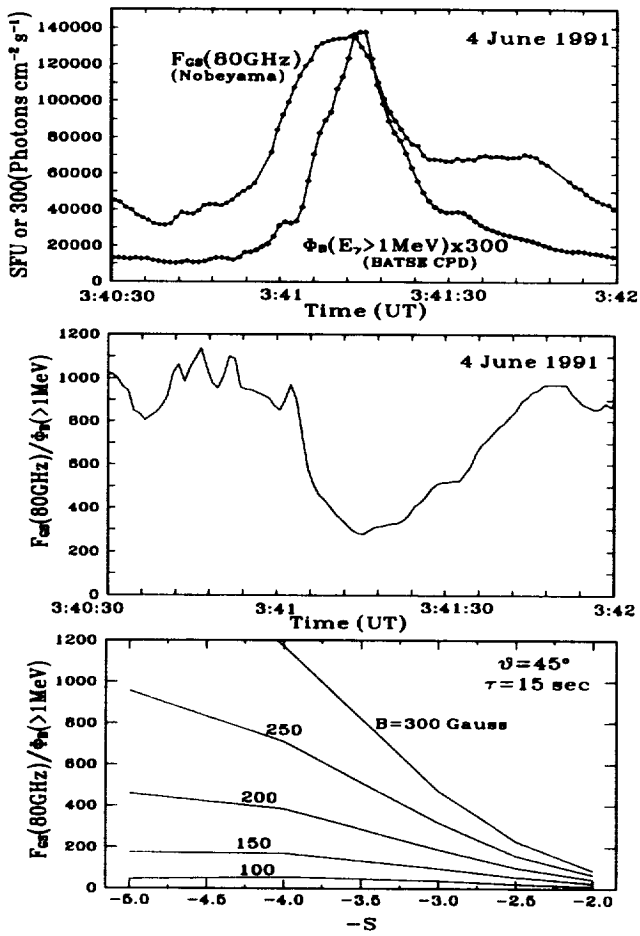


Figure 2.5: Top panel: gamma-ray and millimeter wave time profiles near the impulsive peak of the 4 June 1991 flare; middle panel: the ratio of the millimeter and gamma-ray fluxes showing that the displacement seen in the top panel corresponds to a dip in the ratio with a minimum at the time of maximum of the gamma-ray flux; bottom panel: calculations of the millimeter to gamma-ray flux ratio showing that for a given magnetic field and trapping time, the ratio decreases as the spectrum hardens.

We conclude our discussion with a calculation of the total number of accelerated electrons required to produce the observed gamma-ray and millimeter emissions during the impulsive peak of the 4 June flare. We mention here

that gamma- rays from this flare were observed for much longer periods of time than indicated in Figure 2.5 (Murphy et al., 1993), so that the calculated number should be considered as a lower limit.

For simplicity we assume that all the electrons are accelerated instantaneously at the peak of the emissions (at about 3:41:12 UT), that the source function Q in Equation (2.2) is a power law with spectral index s equal to either 3 or 4, and that τ is a constant equal to 15 s. Thus

$$Q(E,t) = Q_o(E)\delta(t) = AE^{-s}\delta(t) \qquad (2.11)$$

where $\int_E^\infty Q_o(E)dE$ is the total number of accelerated electrons with energies greater than E, and $t = 0$ at the peak of the emissions. By substituting into Equation (2.2) we obtain

$$N_c(E,t) = AE^{-s}e^{-\frac{t}{\tau}} \qquad (2.12)$$

Then by substituting into Equation 7, combining the result with Equation 10, evaluating the integrals numerically, and comparing the result with the observed peak bremsstrahlung flux of about 460 photons $cm^{-2}\ s^{-1}$ (top panel Figure 2.5), we obtain

$$Q_o(E) \simeq 2.2 \times 10^{34} E^{-3} \ \ electrons \quad MeV^{-1} \qquad (2.13)$$

for $s = 3$, and

$$Q_o(E) \simeq 2.3 \times 10^{35} E^{-3} \ \ electrons \quad MeV^{-1} \qquad (2.14)$$

for $s = 4$.

In Figure 2.6 we compare the total number of electrons of energies greater than 1 MeV with similar data for other flares (Ramaty et al., 1993). We plot these as a function of the electron spectral index, which, for each of these flares was obtained from the observed gamma-ray spectrum. As such spectral observations are not available for the 4 June flare, we plotted the results for the two assumed spectral indexes, $s = 3$ and $s = 4$. We see that the 4 June flare was quite unusual. It either had a very hard spectrum or a very large number of accelerated electrons. Both of these possibilities explain the finding that the gamma-ray and millimeter fluxes from this flare were the largest ever observed.

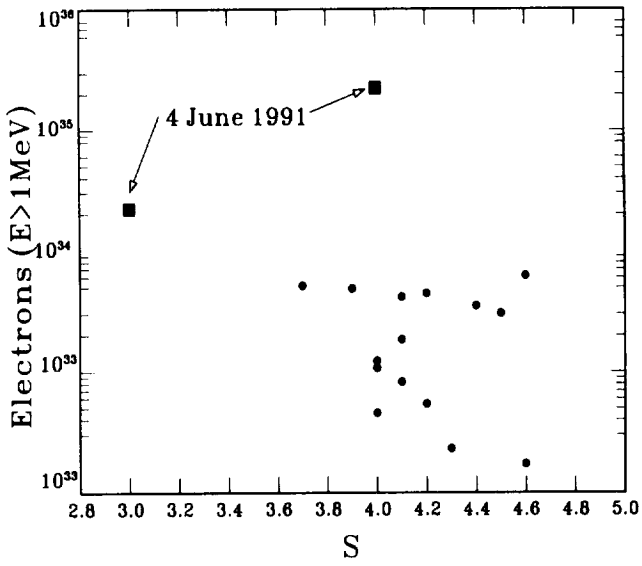


Figure 2.6: Derived total numbers of electrons with energies above 1 MeV showing the exceptional nature of the 4 June 1991 flare.

2.6 REFERENCES

Benka, S.G., and Holman, G.D. 1994, ApJ, 435, 469.

Dulk, G.A. 1985, Ann. Rev. Astron. Astrophys., 23, 169.

Ellison, D.C., and Ramaty, R. 1985, ApJ, 298, 400.

Dennis, B.R. 1985, Solar Physics, 100, 465.

Dermer, C.D., and Ramaty, R. 1986, ApJ, 301, 962.

Enome, S., and Hirayama, T. 1994, Proceedings of the Kofu Symposium, NRO Report No. 360.

Forman, M.A., Ramaty, R., and Zweibel, E.G. 1986, in Physics of the Sun, P. A. Sturrock (ed.), D.Reidel, p. 251.

Haug, E. 1975, Zs. Naturforsch., 30a, 1099.

Koch, H.W., and Motz, J.W. 1959, Rev. Mod. Phys., 31, 920.

Mandzhavidze, N., and Ramaty, R. 1993, Nuclear Phys. B, 33, 141.

Miller, J.A., and Ramaty, R. 1989, ApJ, 344, 973.

Miller, J.A., and Steinacker, J. 1992, ApJ, 399, 284.

Murphy, R.J., et al., 1993, in "Proceedings 23rd International Cosmic Ray Conference" (Calgary), 3,99.

Ramaty, R. 1969, ApJ, 158, 753.

Ramaty, R., and Mandzhavidze, N. 1994, in "High Energy Solar Phenomena" eds. J. M. Ryan and W. T. Vestrand, (AIP: New York), 26.

Ramaty, R., Mandzhavidze, N., Kozlovsky, B., and Skibo, J.G. 1993, Adv. Space Res. 13, (9)275.

Ramaty, R., Miller, J.A., Hua, X-M., and Lingenfelter, R.E. 1990, ApJ Suppl., 73, 199.

Ramaty, R., and Murphy, R.J. 1987, Space Science Revs., 45, 213.

Ramaty, R., Schwartz, R.A., Enome, S., and Nakajima, H. 1994, ApJ, 436, 941.

Rieger, E. 1989, Solar Physics, 121, 323.

Skibo, J.G. 1993, Diffuse Galactic Positron Annihilation Radiation and the Underlying Continuum, PhD Dissertation, Univ. of Maryland.

Chapter 3

GAMMA RAYS AS A TRACER OF COSMIC RAYS AND MATTER IN OUR GALAXY AND OTHER GALAXIES

CARL FICHTEL
Laboratory for High Energy Astrophyics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

As the energetic electrons and nuclei, called cosmic rays, travel through the Galaxy, they interact with matter and photons to produce high-energy gamma-rays. These high-energy gamma-rays are very penetrating and can pass through even the central part of the Galaxy with very little attenuation. They can therefore be studied in detail to obtain a picture of the distribution of the galactic cosmic rays as they interact. A study of the Magellanic Clouds has permitted the determination that cosmic rays are galactic and not metagalactic, and therefore that galactic dynamic balance should exist. This fact, combined with a knowledge of the cosmic rays, matter, and photon distributions, permits a prediction of the galactic gamma-ray energy spectrum and the gamma-ray distribution with relatively few adjustable parameters. A comparison of the observed and predicted spatial distributions shows that there is very good agreement between the two. Further, the coupling scale of the cosmic rays and the matter can be determined as well as the normalization for the molecular hydrogen density distribution.

3.1 INTRODUCTION

In thinking of our Galaxy, one is often accustomed to considering the Milky Way and the stars that can be seen with the eye, or with an optical telescope. There are many other important components besides the visible stars which play a major role in the Galaxy. These components include the interstellar gas, the magnetic fields, photons of greater and smaller frequency, and the cosmic rays. The latter, which are misnamed from a time in the past when their nature was unknown, are in fact charged particles, some of quite high energy. There is a tug-of-war occuring in the Galaxy between the cosmic rays, the gas particles, and the galactic magnetic fields, all of which are trying to expand the Galaxy, and the gravitational attraction of the matter which is trying to pull the whole together.

This latter statement has implicit in it the belief that the cosmic rays are galactic and not from beyond our Galaxy. In this chapter, it will be seen that, although this assumption had seemed reasonable for some time, recent high-energy gamma-ray measurements have now shown it to be correct with a high degree of certainty. Thus, the description of the cosmic rays must be consistent with their galactic nature and the galactic forces that come into play. On a broad galactic scale, the balance described above must be taken into account, although on a smaller scale there are believed to be local perturbations, due particularly to supernovae explosions.

It is worth remembering at this point that, although we live in our Galaxy and some measurements can be made very well, in some respects we know less about it than other galaxies. This situation results from the fact that the interstellar matter prevents us from seeing very far into the Galaxy in either the optical band or the X-ray band. Also, it is not possible to step back and take a picture of it. There is no beautiful picture of our Galaxy, such as the one of the Great Galaxy in Andromeda. For our Galaxy, we see what is appropriately called the Milky Way. Thus, the overall structure must be deduced from line-of-sight measurements in the plane, mostly from the radio band, and the deconvolution is not without problems.

The primary goal of this chapter, as the title implies, is to consider what can be learned about our Galaxy from high-energy gamma-rays coming both from our Galaxy and from the neighboring Magellanic Clouds. It is desirable to begin by describing the nature of the principal components of our Galaxy that are involved in the production of high-energy gamma-rays. It will be seen that, as is often the case in astrophysics, it is necessary to use information obtained from other wavelengths, particularly the radio band, as well as measurements of the cosmic-ray particles. The first section is followed by a description of the relevant interaction processes, to see how the gamma-rays that are of interest are produced and to have the quantitative basis for calculations. The next section provides a brief description of the manner in which high-energy gamma-rays are detected. This is followed by a discussion of galactic dynamic balance, wherein the expansive pressures in the Galaxy are balanced by gravitational attraction. The subsequent two sections describe the gamma-ray observations in our Galaxy and their interpretation in terms of our current understanding of the processes involved and the manner in which the high-energy gamma-ray observations of the Magellanic Clouds assisted in the study of the cosmic rays in galaxies.

The high-energy interaction processes that you have been studying in your physics courses have application beyond the high-energy physics laboratories, and a knowledge of these processes is very necessary for an understanding of the astrophysical phenomena being considered here.

3.2 COSMIC RAYS AND THE INTERSTELLAR MATTER, PHOTONS, AND MAGNETIC FIELDS

Galactic high-energy ($\geq 20\ MeV$) gamma-rays, as will be shown in the following sections, are primarily produced by cosmic rays interacting with interstellar matter and to a lesser extent with interstellar photons. Hence, it is these three constituents of our Galaxy whose properties will be summarized here.

3.2.1 Cosmic Rays

Cosmic rays are bare nuclei and electrons, some of very great energy. Cosmic rays were discovered by Victor Hess in 1912, but their nature at that time was unknown other than that they came from beyond the Earth. Now we know many details of their properties, which will be summarized here, but it is fair to say that even today, over 80 years later, the origin of the bulk of the radiation, or to be more specific the manner and location of the primary acceleration of the particles, is primarily a matter of speculation. The energy density of the cosmic rays in the Galaxy is about equal to that of starlight, the kinetic motion of interstellar matter, and the magnetic fields; so they are not a trivial phenomena. (At a more personal level cosmic rays bombard the top of the atmosphere and some of their secondary products proceed all of the way down to sea level. There are a few secondary cosmic-ray particles, mostly muons, passing through each of us every second.)

Figures 3.1 and 3.2 show the composition and energy spectra of some of the major components of the

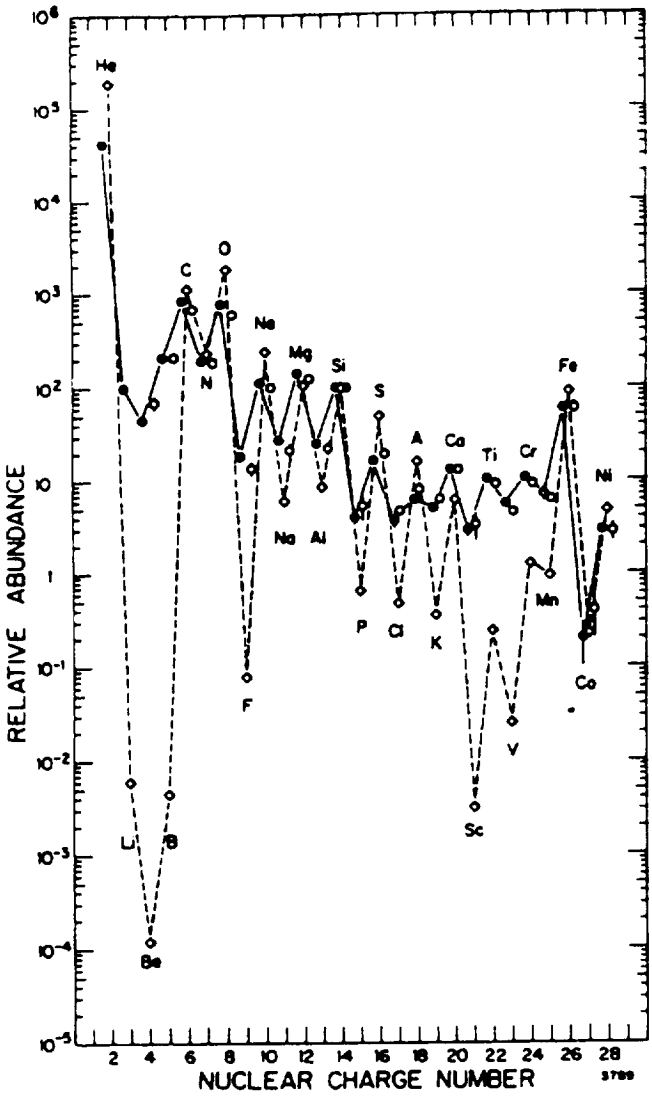


Figure 3.1: The cosmic ray nuclear abundances (He-Ni) measured at 1 AU compared to the solar system abundances, all relative to Silicon (Si $\equiv$ 100). The solid circles represent low-energy data, 70-280 MeV/n; the open circles represent a compilation of high-energy measurements 1000-2000 MeV/n; the diamonds are solar system abundances. This figure is reproduced with permission from "Annual Review of Nuclear and Particle Science," Simpson, (1983). The data are referenced therein.

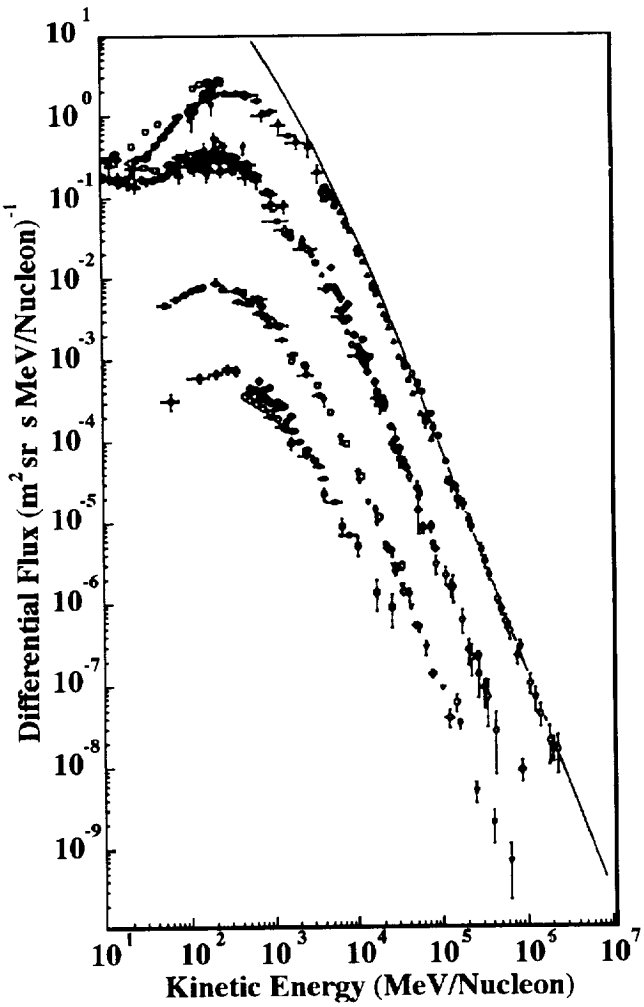


Figure 3.2: Energy spectra of cosmic ray nuclei measured at 1 AU based on many satellite and balloon-borne experiments near the solar minimum modulation. Differential energy spectra for the elements (from top) hydrogen, helium, carbon, and iron. The solid curve shows the hydrogen spectrum extrapolated to interstellar space by unfolding the effects of solar modulation. The "turn-up" of the helium flux below ~60 MeV/nucleon is caused by the additional flux of the anomalous ^{4}He component. The figure is from Simpson (1983), and is reproduced with permission from "Annual Review of Nuclear and Particle Science," Vol. 33c, 1983 by Annual Reviews Inc. The data are referenced therein.

cosmic ray nuclei. Hydrogen nuclei are the dominant constituent, helium nuclei are next, and all of the rest amount to only about one percent of the total. The composition is similar to that of the solar system, except for the large excess of Li, Be, B, and the nuclei just below the iron group. These excesses are generally believed to be secondaries, produced in the interaction of the cosmic rays with the instellar matter. Cosmic-ray electrons have an energy density much smaller than that of the protons, but are of importance for producing high-energy gamma-rays, as will be discussed later.

In the figure showing the energy spectra, it is seen that the cosmic-ray component spectra are very similar over the range shown. Beyond the energy range shown, there is no information on the composition of individual nuclei, but from ground level secondary air shower mea-

surements it is known that the spectra continue with the same slope to about 10^{15} eV. At this point the spectra begin to steepen and the composition begins to change to there being a greater abundance of heavier nuclei. These characteristics are generally attributed to the facts that:

(1) Protons with rigidities above about 10^{15} eV/c for a proton cannot be held in the Galaxy.

(2) For the same rigidity, heavier nuclei have a higher total energy (e.g., Fichtel and Linsley, 1986).

(It also is possible that sources in our Galaxy cannot accelerate particles above this point.) At about 10^{18} eV/c, the spectrum begins to become less steep. This change suggests the possibility of an extragalactic component (Fichtel, 1963).

An analysis of the secondary cosmic rays mentioned above, suggests that the cosmic rays have typically passed through about 10 to 15 $g\ cm^{-2}$, although there is a distribution in path lengths, and that the higher energy (10^{11} eV) cosmic rays have passed through somewhat less material on the average than the lower-energy (10^{9} eV) cosmic rays. Some secondaries, such as Be7, are unstable and act as clocks giving an estimate of the age of the cosmic rays, now thought to be (1 to 2) $\times 10^{7}$ years.

Combining the lifetime of the nuclei, the velocity, and the average amount of matter traversed, shows that the average density of matter seen by the nuclear cosmic rays is (0.2 to 0.3) cm^{-3}. Hence, on the basis of the discussion in the last paragraph, they do not spend most of their time in the thin matter disk where the density is about 1 cm^3, but rather they have a broader distribution relative to the galactic plane.

The nonthermal continuum radio emission, which is generally attributed to the synchrotron radiation from cosmic-ray electrons interacting with the galactic magnetic fields (e.g., Ginzburg and Syrovatskii, 1964, 1965), provides information about the distribution of high-energy cosmic-ray electrons. Baldwin (1967, 1976) estimates the equivalent disk thickness for synchrotron emission to be about 750 pc, and some analyses have suggested that it is even larger. Significant nonthermal emission is even seen as high as 2 kpc above the plane. If it is assumed that both the electron density and magnetic field density have the same distribution on the average and that it is Gaussian, then, since the synchrotron radiation is proportional to the product of the two, the scale height of each individually is $\sqrt{2} \times 0.75$ or just over 1 kpc. It seems reasonable to assume that cosmic-ray electrons and protons and the magnetic fields, at least in our local region of the Galaxy and probably elsewhere, all have a scale height of this order.

3.2.2 The Interstellar Matter

The matter that is relevant here is the diffuse interstellar matter with which the cosmic rays interact. It is in three forms: atomic, molecular, and ionized. The first two are the most important. The instellar atomic hydrogen can be measured directly by observing the 21 cm hyperfine structure line, and the distance distribution can be estimated from the line shift due to the assumed motion of the Galaxy. The molecular hydrogen can only be estimated indirectly through measurement of other lines, principally the CO lines, and assuming a constant ratio of molecular hydrogen to CO, or at least one that varies only very slowly with position. A more accurate determination of this ratio has been possible in part through the use of the high-energy gamma-ray measurements. There are some problems in detail, especially in the general direction of the galactic center, but a reasonable understanding seems to exist. The ionized hydrogen is estimated from dispersion measurements, particularly of pulsars of known distances. The matter in our Galaxy appears to have an approximate spiral pattern, but some details are not well-defined. The molecular hydrogen is concentrated more towards the center of the Galaxy relative to the atomic hydrogen.

3.2.3 Photons and Fields

The relevant photons are the stellar ones of all wavelengths, but principally visible and near visible, and the blackbody radiation. They are important for the discussion here, because cosmic-ray electrons interact with them and produce gamma-rays. It is interesting to note that once again the energy density of these photons is about 1 eV in the plane.

The magnetic fields are important, not because of their leading to gamma-rays (The gamma-ray production from cosmic-ray interactions with them is negligible.), but because they control the motions of the cosmic rays and also because they are one of the expansive pressures in the Galaxy.

3.3 INTERACTION PROCESSES PRODUCING HIGH-ENERGY GAMMA-RAYS IN OUR GALAXY

There are three important processes for the production of diffuse high energy gamma-rays in the Galaxy. They are nucleon-nucleon interactions, bremsstrahlung, and the Compton interaction. The diffuse radiation dominates over the contribution from point sources; this feature will be seen quantitatively in Section 3.6 of this chapter. Synchrotron radiation from cosmic rays spiraling around galactic magnetic fields can be shown to be negligible. These interaction processes are illustrated in Figure 3.3. Each of the important processes will be considered now.

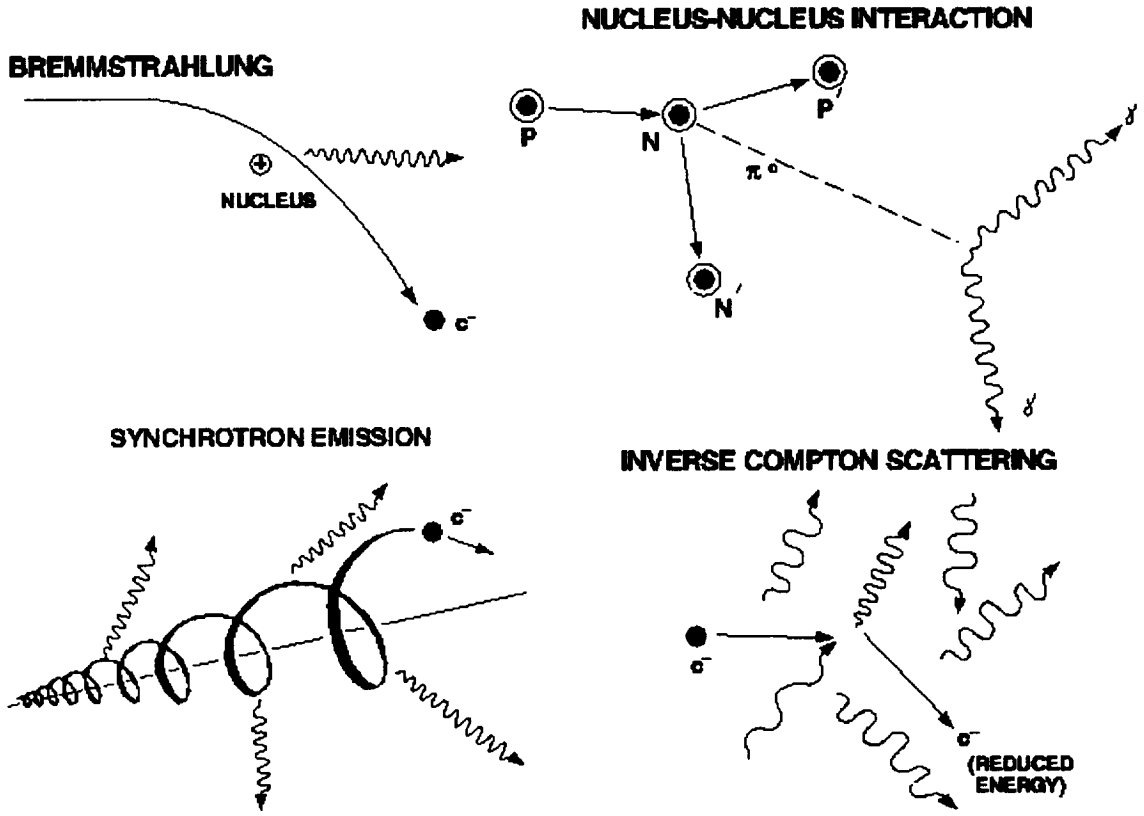


Figure 3.3: Illustrations of high-energy, gamma-ray-producing interactions

Much of the following material in this section is from Fichtel and Trombka (1981).

3.3.1 Interactions of Cosmic-Ray Nuclei with Interstellar Matter

High-energy physics has provided a wealth of information on nucleus, nucleus interactions–enough information to permit a good, but long and tedious calculation of the expected gamma-ray emission. The primary contribution comes from the proton-proton interaction with additions for the proton-helium, helium-protron, helium-helium, etc. interactions. The cosmic-ray composition is well known and the interstellar medium composition, although less well known, is believed to be quite adequately known for this purpose. A very brief summary of the interaction process now follows.

Of the many mesons produced in the interactions of the cosmic rays with interstellar matter, the most commonly produced are mesons, which can be either charged or neutral. The π° mesons decay in about $10^{16}s$ into two gamma-rays each with about 68 MeV in the rest frame. Many of the other mesons and hyperons also decay into π mesons, e.g., $K^\pm \rightarrow \pi^\pm + \pi^\circ$, $K^\circ \rightarrow 2\pi$, $K^\circ \rightarrow 3\pi$, $K^\circ \rightarrow \pi^+ + \pi^- + \pi^\circ$, and $\Lambda \rightarrow \eta + \pi^\circ$. The detailed calculations leading to the predicted intensity and energy spectrum of the gamma-rays (based on the average numbers of mesons formed in an interaction, their angular distribution, and the resulting energy spectrum) are quite lengthy. This is due to the need to study the many different products, to take into account the different cosmic-ray species and interstellar nuclei in the correct proportions, follow their decay, integrate over all angles, and then integrate over the cosmic-ray energy spectrum. These calculations have, however, been performed and refined by many scientists (e.g., Caravello and Gould, 1971; Stecker, 1971; Morris, 1984; Dermer, 1986). The results of these calculations are compared to existing gamma-ray data in Section 3.6.

It is possible to understand the general shape of the predicted energy spectrum because it is dominated by π° decay. The energy spectrum of the gamma-rays resulting from π° decay is quite different from that resulting from most other astrophysical processes, such as bremsstrahlung, synchrotron radiation, and the inverse Compton effect. The origin of the unique spectrum is the decay of the π° into two massless particles of equal energy in the rest frame. The probability distribution of the gamma-rays in the laboratory system can be shown (e.g., Fichtel and Kniffen, 1974) to be a constant from

$E_\gamma^* [(1-\beta)/(1-\beta)]^{1/2}$ to $E_\gamma^* [(1+\beta)/(1-\beta)]^{1/2}$ and zero outside this range, where E_γ^* is the energy of the π^0 decay gamma-ray in the rest frame. Since this range always includes E_γ^* and since the integral is performed over a distribution of $\beta's$, it is seen at once that the distribution would have a peak at E_γ^* and be symmetrical when plotted against in E_γ , as seen in Figure 3.4.

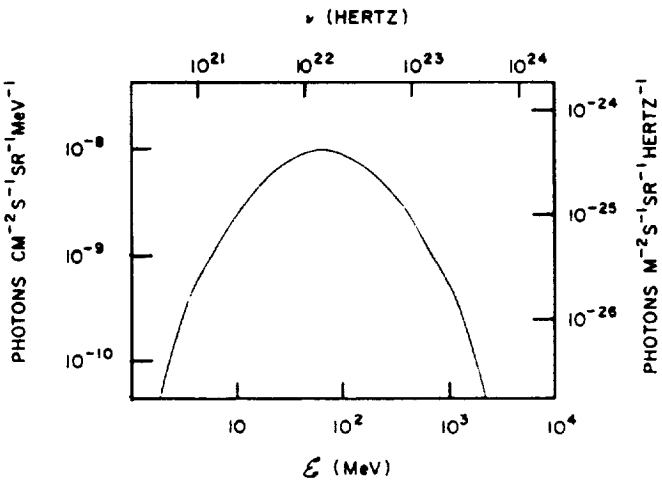


Figure 3.4: Typical differential spectrum expected from the decay of neutral pions produced in cosmic-ray interactions with interstellar gas (adopted from Stecker, 1971). The curve is symmetric about E*, since the abscissa is chosen to be log E. The figure is from Fichtel and Trombka (1981).

3.3.2 Cosmic-Ray Electron Interactions with Matter

As cosmic-ray electrons pass through the interstellar medium they interact with the matter and produce photons over a wide range of energies. In the case of the energetic electrons, the resulting high-energy photon spectrum reflects that of the electrons, at least qualitatively, in the sense that it is a monotonically decreasing function with increasing energy. The largest gamma-ray intensity from electrons probably comes from bremsstrahlung. The calculation of the radiation from this mechanism in the region below 10^2 MeV is rather uncertain because the interstellar cosmic-ray electron spectrum is not well known at low energies, where the electron spectrum observed near the Earth has undergone strong solar modulation which suppresses the lower energies. The degree of the suppression is a function of the solar cycle. This problem did not arise in the case of cosmic-ray nucleon matter interactions because there the higher-energy cosmic rays are most significant in the production of gamma-rays, and their spectrum and general composition are well known and not affected markedly by solar modulation. The gamma-ray production from cosmic-ray electron, matter interactions can be calculated using the bremsstrahlung cross-section formulas of Koch and Motz (1959). The calculations in general are very complex. However, to a good approximation, if the parent electrons differential intensity has the form

$$j_e = \frac{dj_e}{dE_e} = KE_e^{-a} \tag{3.1}$$

then the gamma-ray source function is given by

$$q(E_\gamma) = \sum_S \frac{\rho_s}{X_o} \left(\frac{4\pi}{E_\gamma}\right) \int_{E_\gamma}^{\infty} j_e dE_e \tag{3.2}$$

where ρ_s is the mass density of the particular nuclear species and Xo is the radiation length. The sum is over all target species. Assuming the interstellar composition to be 10 percent helium and 1 percent heavy nuclei, the result is

$$q(E_\gamma) \cong 4.7 \times 10^{-25} n_H \left(\frac{KE_\gamma^{-a}}{a-1}\right) \frac{photons}{cm^3\ s\ \ MeV} \tag{3.3}$$

where n_H is the number of protons per unit volume in both atomic and molecular form. This last equation includes the effect of the nuclei heavier than the proton as long as they are in the assumed proportions. Notice that the spectral index of the gamma-ray differential source function is the same as that of the differential electron spectrum.

3.3.3 Compton Interactions

Cosmic-ray electrons also interact with starlight photons, for which both the optical and infrared ranges are important, and with the 2.7° microwave blackbody radiation to produce Compton gamma-rays. The source function for these interactions is much smaller in the galactic plane in the vicinity of the solar system. The total contribution to the galactic gamma-radiation, however, is significant, because the cosmic-ray and the stellar photon scale heights above the galactic plane are much greater than those of the matter. Of course, the blackbody photon density is uniform. Hence, the integral intensity along a line of sight is closer to that of the bremsstrahlung than the source functions would imply.

The calculations associated with the production of Compton gamma-rays are quite complex; however, they have been performed in some detail for the cases of astrophysical interest by Ginzburg and Syrovatskii (1964). A power law in energy for the electron spectrum of the form given by Equation (3.1) will produce a power law gamma-ray source function of the form

$$q(E_\gamma) \propto E_\gamma^{-(a+1)/2} \tag{3.4}$$

Whereas, in the case of bremsstrahlung, the astrophysical gamma-rays come predominantly from electrons of energies similar to those of the gamma-rays, or only a factor of several higher in energy, for the Compton case, the energy of the parent electron is given approximately by the equation

$$E_\gamma \cong \left(\frac{E_e}{mc^2}\right)^2 E_p \qquad (3.5)$$

where E_p is the photon energy before the electron interaction. Therefore, if the interaction is with starlight where the typical photon energy is a few eV, in order for the gamma-ray to be about a hundred MeV, E_e must be several GeV. For the 3° blackbody radiation, the typical photon energy is about 8×10^{-4} eV; therefore, the parent electrons must be in the range of 2×10^5 MeV to produce 100 MeV gamma-rays. In the energy range above a GeV, there is no serious uncertainty in the electron spectrum due to a lack of knowledge of the solar modulation. However, in the few GeV range, the electron energy spectrum changes shape, so the calculation must be performed carefully. At energies as high as a few times 10^5 MeV, the electron energy spectrum is less well known, and an increased uncertainty in the calculated gamma-ray intensity is introduced.

3.3.4 The Sum of the Three Processes

When the energy spectra for the three processes just discussed are combined, the bremsstrahlung and nucleon-nucleon radiation dominate. This feature is generally true regardless of direction, although at higher latitudes the Compton radiation becomes somewhat more important. Because of the comparable intensities of the bremsstrahlung and nucleon-nucleon emission, the latter leads to a spectral feature better described as a change in slope than a "bump," as will be seen in Section 3.6.

3.4 THE DETECTION OF HIGH-ENEGY GAMMA-RAYS

The instruments used to detect high-energy gamma-rays bear no resemblance to optical telescopes; hence, it is worth having some understanding of how they work so that one can better appreciate the observations, including what can be achieved and the limitations to the measurements. The knowledge of the detection techniques also gives a better understanding of why the field of high-energy gamma-ray astronomy developed the way it did and why it is truly a product of the space age.

In the region of the electromagnetic spectrum defined by high-energy gamma-ray astronomy, the pair production process dominates. There is no possibility of using reflection or defraction techniques. One must detect the secondary electron and positron formed in the interaction, and, if one is to be able to obtain good information on the direction of the arrival direction of the high-energy gamma-ray, the directions of these two charged particles must be determined before there is excessive electron scattering in the material in which the gamma-ray has converted. It is also, of course, desirable to have an estimate of the energy of the gamma-ray by measuring that of the electron and positron. Finally, the high-energy gamma-ray must be detected in a high background of charged particles, whose number is typically ten thousand times that of the high-energy gamma-rays that are to be detected and measured.

The properties outlined in the last paragraph have led to the development of high-energy gamma-ray telescopes which have as their central detector system an automated pictorial device consisting of a series of thin plates of high nuclear charge. The high nuclear charge is chosen so that for a given probability for interaction of the gamma-ray, the energy loss of the secondary electrons is kept to a minimum. The location of the electron tracks is made between each of these thin plates so that the directions may be reasonably well determined before there is large scattering. Of the instruments flown on satellites, the spark chamber has been the detector of choice thus far, because of its high reliability, rugged possible design, and reasonable position accuracy. Other detectors of greater position accuracy may replace it in the future.

The spark chambers have been triggered by a directional counter telescope, either a Cerenkov system or a time-of-flight system combined with an anticoincidence shield. This latter is essential to the success of the telescope because of the high charged-particle background. These charged particles must be detected so that they and any interactions they cause may be rejected. This is true not only because the early and present telescopes could not tolerate the high charged-particle rate, but also because, even more critically, there are rare interactions that could masquerade as gamma- rays. Early experiments demonstrated that this was a very real problem that could completely negate the validity of a result. Hence, experimental groups took great care to be certain that the anticoincidence system was of very high quality. The one used with The Energetic Gamma-Ray Experiment Telescope (EGRET) flown on the Compton Gamma-Ray Observatory (CGRO), for example, was known to be good to at least one part in ten million. A schematic diagram of EGRET showing how a gamma-ray event is recorded is shown in Figure 3.5, and a more detailed illustration of the telescope is given in Figure 3.6. For a discussion of high-energy telescopes in general, their variations, and the long history which led to the current approach, see Fichtel and Trombka (1981).

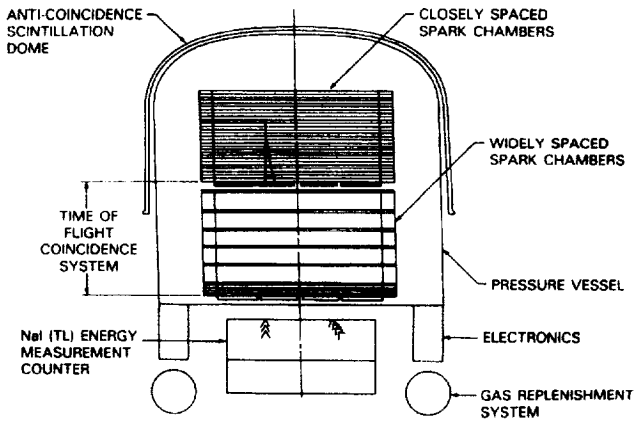


Figure 3.5: Schematic Diagram of EGRET

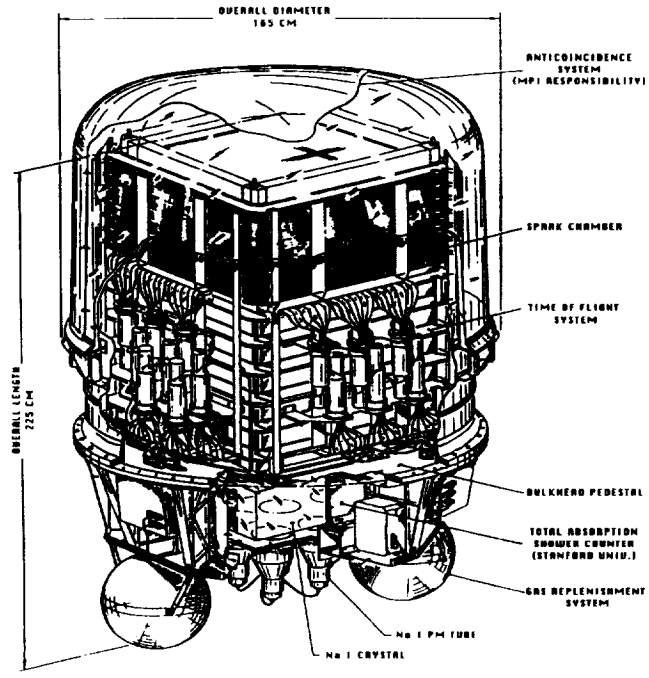


Figure 3.6: Detailed Illustration of the EGRET

In the data analysis, an automatic program searches for the characteristic inverted "V" or "Y" and rejects events that do not satisfy rigid criteria established to identify the high-energy gamma-rays, which start in the chamber wall or above the top deck. The program also identifies any event that might be ambiguous, so that it may be viewed on a computer display terminal and analyzed further. The current computer program has been found to be extremely reliable.

3.5 GALACTIC DYNAMIC BALANCE

Several fundamental theoretical considerations place constraints on the cosmic-ray distribution. It is now generally accepted that the cosmic rays and magnetic fields are primarily galactic and not universal. The recent important support for this belief from high-energy gamma-ray observations will be described in Section 3.7 of this chapter. If these fields and cosmic rays are galactic, they can only be constrained to the galactic disk by the gravitational attraction of the matter (Biermann and Davis, 1960; Parker, 1966, 1969, and 1977). The local energy density of the cosmic rays ($\approx 1\ eV/cm^3$) is about the same as the estimated energy density of the magnetic field and that of the kinetic motion of matter. Together, the total expansive pressure of these three effects is estimated to be approximately equal to the maximum that the gravitational attraction can hold in equilibrium. Assuming the solar system is not at an unusual position in the Galaxy, these features suggest that the cosmic ray density throughout the Galaxy may generally be as large as could be contained under near-equilibrium conditions.

Further theoretical support is given to this concept by the calculated slow diffusion rate of cosmic rays in the magnetic fields of the Galaxy and the small cosmic-ray anisotropy. These considerations then lead to the hypothesis that the energy density of the cosmic ray is larger where the matter density is larger on a coarse scale such as that of the galactic arms. On a smaller scale, the pressures of the cosmic-ray gas and magnetic fields cause the cosmic-ray gas and field system to expand between the large clouds through which the magnetic fields thread. As noted earlier, the radio continuum measurements and cosmic-ray results support the picture of the cosmic rays having a large scale height relative to the matter and spending only a relatively small amount of time passing through the dense cloud region. This large scale height is important in determining the Compton contribution to the diffuse emission, as will be seen later.

The picture that seems to emerge is that cosmic rays, at least below 10^{16} or $10^{17}\ eV$/nucleon, are bound to the magnetic lines of force and the field lines are closed (or else the cosmic rays would escape too quickly). The cosmic rays are constrained and not free to escape individually. Thus, their escape is the result of cosmic-ray group pressure inflating the magnetic field lines and pushing outward from the Galaxy and diffusion.

The fact that the cosmic-ray density is approximately as large as can be contained suggests a plentiful source. What is this source? At present most astrophysicists in the field feel that cosmic rays come from supernovae and flare stars, but may receive subsequent acceleration. If supernovae were to supply the total cosmic-ray energy, each supernova would have to produce between

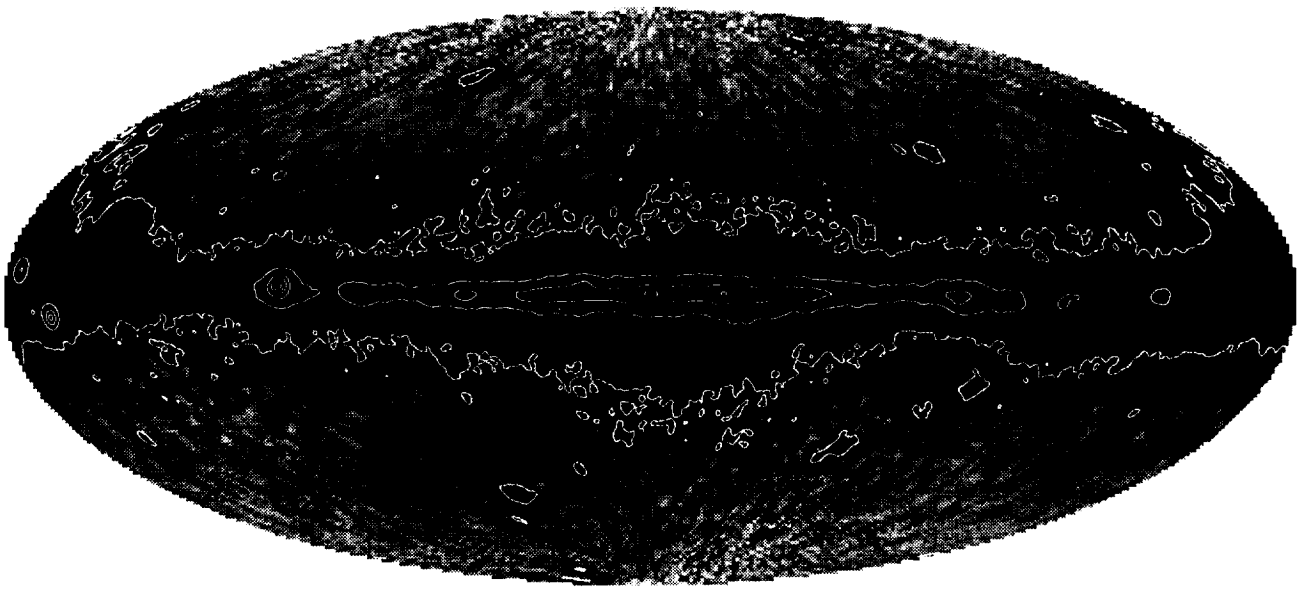


Figure 3.7: EGRET Gamma-Ray All-Sky Survey above 100 MeV (Courtesy of Carl Fichtel and the EGRET Instrument Science Team)

10^{49} *and* 10^{50} *erg* of cosmic rays, assuming a supernova rate of 1 every 30 years. This seems not a prohibitive amount of energy for a supernova to give to energetic particles, provided a reasonable mechanism can be found and verified. Even if supernovae are not the primary initial source, they may supply an important part of the total energy of the cosmic rays by shock wave acceleration of existing cosmic rays. (See Blandford and Ostriker, 1980.)

3.6 THE HIGH-ENERGY GAMMA-RAY SKY AND THE GALACTIC DIFFUSE RADIATION

The most intense celestial high-energy gamma-radiation observed is that from the galactic plane. This feature was observed first by the pioneering counter telescope flown on OSO-3 (Kraushaar et al., 1972), and the major features of this galactic radiation were defined by measurements made with the SAS-2 satellite launched in 1972 (Kniffen et al., 1973; Fichtel et al., 1975; Hartman et al., 1979), and the COS-B satellite launched in 1975 (e.g., Mayer-Hasselwander et al., 1980). Te results from these satellites revealed the galactic plane gamma-radiation is strongly correlated with galactic structural features, especially when the known strong gamma-ray sources are subtracted from the total radiation. These important results now have been followed by the observations of (EGRET) flown on the Compton Gamma-Ray Observatory. Because EGRET is far more sensitive than the gamma-ray telescopes on the earlier missions and has improved angular resolution, the EGRET results are now the ones to be considered and compared to theoretical expectations.

As Figure 3.7 shows, the gamma-ray sky looks quite different from the optical sky, even when the detector resolution is considered. The galactic diffuse radiation dominates strongly, as shown even more clearly in Figure 3.8, where the strong point sources have been removed. To the best of our knowledge, this gamma-radiation is primarily diffuse. Not surprisingly, in view of the earlier discussion in this chapter, it is more similar to the radio sky, particularly the 21 *cm* sky. The radiation is particularly intense and more narrow in galactic latitude in a quadrant centered about the galactic center and least intense in the general direction of the anticenter, where it is also broader in galactic latitude. How well does the expected radiation, based on calculations, match what is observed? It matches very well, as it will now be considered.

The number of gamma-rays per unit area and time

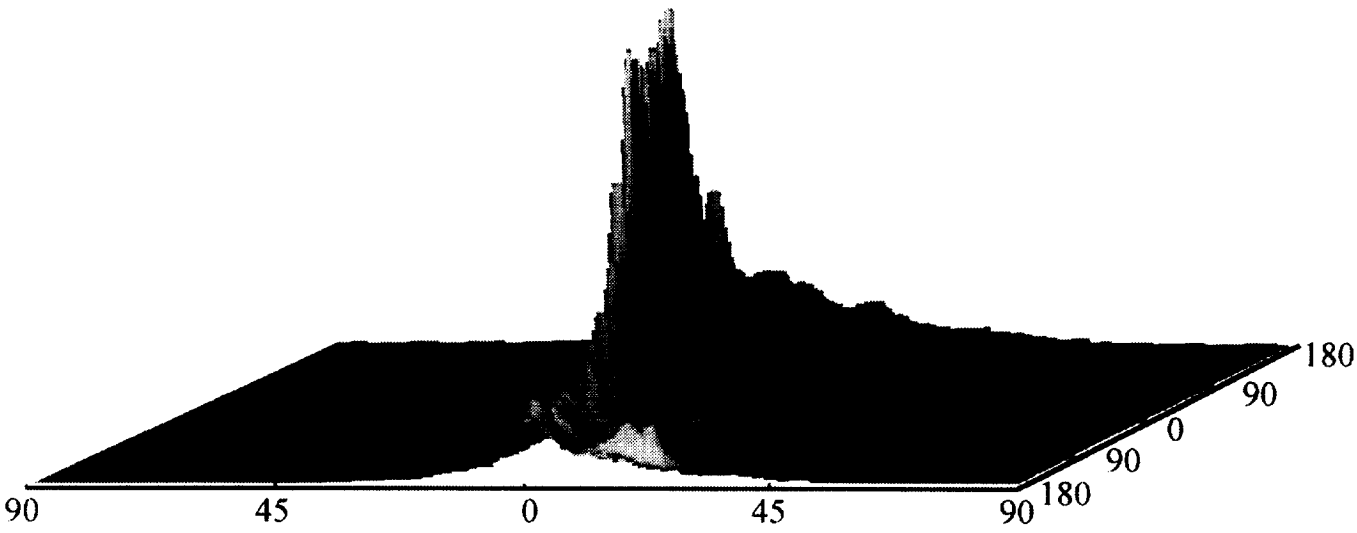


Figure 3.8: Three-dimensional portrayal of the high-energy, gamma-ray intensity (E > 100 MeV) measured by EGRET. (Courtesy of Carl Fichtel and the EGRET Instrument Science Team)

within dE falling on a surface a distance r from a volume element dV with a source strength q is

$$\frac{1}{4\pi r^2} q\left(E_\gamma, r\right) dV dE \tag{3.6}$$

Hence, the differential intensity is

$$j = \frac{dJ}{dE} = \frac{1}{4\pi} \int_0^\infty q\left(E_\gamma, r\right) dr \tag{3.7}$$

where j is the number of photons per unit time, area, solid angle, and energy. The "q" values have already been described earlier in this chapter.

There are several considerations still to be addressed. First of all, for some time, there had been a debate on whether cosmic rays are galactic or metagalactic. This is a very important consideration because, if they are metagalactic, they are likely to be uniform in intensity throughout the Galaxy, whereas if they are galactic, then the discussion of Section 3.5 applies. Although there were many reasons for thinking that they were probably galactic, it was the high-energy gamma-ray measurements of the Magellanic Clouds, and particularly the small one, that provided the definitive evidence showing that cosmic rays are galactic. These observations will be discussed in Section 3.7.

With the clear elimination of the metagalactic alternative for cosmic rays, the dynamic balance concept (described in Section 3.5) is appropriate on a broad scale, whose dimension must be determined. The coupling scale is then an adjustable parameter to be determined on the basis of the comparison of the observations to the predicted distribution. In fact, only the coupling scale and the normalizing factor for molecular hydrogen are left as adjustable parameters, if indeed the cosmic-ray density is taken as being as large as it can be based on dynamic balance, as supported by the local value.

There are two primary tests that may be applied to see if there is good agreement between theory and experiment. These are the energy spectrum and the spatial distribution. Turning first to the energy spectrum, it has been found through an analysis of the EGRET results (Hunter, et al., 1996) that the spectrum did not change with position in the Galaxy within uncertainties. This is consistent with the electron-to-proton ratio being essentially the same throughout the Galaxy. With regard to the spectrum itself, the best spectrum is obtained in the general direction of the galactic center, because the intensity is larger and more photons are collected. This spectrum is shown in Figure 3.9. Notice that one does see the effect on the spectrum caused by the neutral pion decays resulting from the nucleon-nucleon interactions. Generally, there is excellent agreement between the predicted spectrum and the observed one in both magnitude and shape. The exception is the deviation in the GeV region. This deviation increases from about 1 to 5 GeV and then remains fairly constant. The cause of this is not known. The two most likely explanations are the following. One is that for some reason, the cosmic-ray spectrum seen at the Earth is different at higher energies than that in the rest of the Galaxy. The second is that the theoretical calculations of the gamma-ray production may not be adequate because of the approximations made in the case of the nucleon-nucleon interactions. Other possibilities are considered in the paper of Hunter et al. (1996)

For the second set of tests, there are the spatial distributions. Figure 3.10 compares the longitudinal distribution of the predicted galactic high-energy gamma-ray diffuse radiation to that observed by EGRET, using the

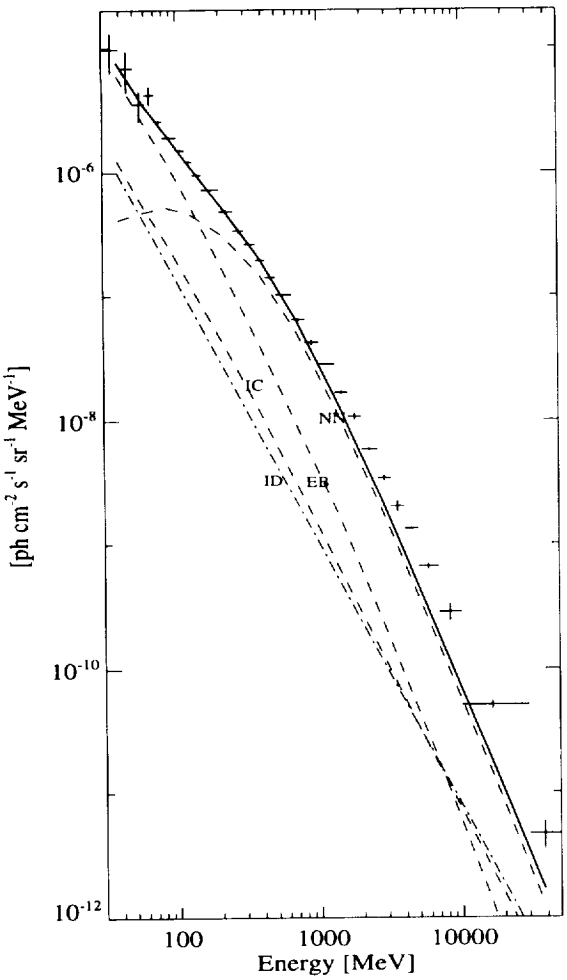


Figure 3.9: The average, diffuse gamma-ray spectrum of the galactic center region, $300° < l < 60°, -10° < b < 10°$, by EGRET telescope. Expected contributions from the nucleon-nucleon interactions, bremsstrahlung, the inverse Compton interactions, and the sum of the three. The point sources detected over 5σ significance are removed. (The figure is from Hunter et al., 1996).

results of Hunter et al., 1996. The theoretical model in Hunter et al. is based on that of Bertsch et al. (1993). In this figure, only the radiation from the five strong high-energy gamma-ray pulsars has been subtracted; there are other known sources. As can be seen, the agreement is reasonable with the excesses generally seen associated with point sources or with localized excesses. Figure 3.11 shows a distribution in latitude for one longitude range. Again, the agreement is reasonable as it is for other longitudes. More details can be found in the paper of Hunter et al. (1996), including distributions in latitude at many longitudes and distributions as a function of energy. The radiation from local galactic clouds is generally consistent with the cosmic rays at the local level as would be expected (e.g., Hunter et al., 1994, Digel et al., 1995, and Hunter et al., 1996).

Thus, there is generally good agreement between the observed diffuse radiation and the expected high-energy gamma-radiation based on the concept of cosmic-ray interactions with matter and with photons and on the concept of dynamic balance. It is not clear whether there is any positive evidence for perturbations by supernovae. Excesses have been seen in association with supernovae (Esposito et al., 1994), but they are consistent with the resolution of the EGRET detector and may be the radiation from a yet undiscovered neutron star pulsar.

Finally, there is a matter of the two parameters that were determined in the process of the analysis of the diffuse gamma-ray distribution. The first is the value of X, the conversion factor to proceed from the CO data to the estimate of the molecular hydrogen normalization discussed in Section 3.2. It was found to be $(1.56 \pm 0.05) \times 10^{20}$ H-atoms cm^{-2} (K $km\ s^{-1})^{-1}$, an improved estimate over earlier ones, but consistent with other estimates. The value of the cosmic-ray, matter coupling scale was found to be (1.75 ± 0.20) kpc, consistent with the expectations of radio observations, galactic structure, and theoretical considerations. See Hunter et al., 1996 for further details and Fichtel and Trombka, 1996, for a fuller general discussion.

3.7 THE LARGE AND SMALL MAGELLANIC CLOUDS

These two galaxies are the closest ones to our own. Although they are relatively small, they are important for astrophysical studies because they are close and therefore can be studied more fully. One contribution that high-energy gamma-ray measurements of the Magellanic Clouds have made is the determination that cosmic rays are galactic. This conclusion can be reached in a very straightforward manner. If the cosmic rays pervade the metagalaxy, or some larger region, they must create high-energy gamma-rays through their interaction with the interstellar matter and photons in the Magellanic Clouds. The high-energy gamma-radiation that is expected in this case can be determined. It was also calculated for the case of galactic dynamic balance, as noted by Ginzburg and Ptuskin (1976).

These calculations have been made by Fichtel et al. (1991) and Sreekumar and Fichtel (1991) using recent radio information for the Large and Small Magellanic Clouds. The values are listed in Table I along with the experimentally observed values by EGRET (Sreekumar et al., 1992, and Sreekumar et al., 1993). It is seen clearly that the Small Magellanic Cloud high-energy gamma-ray emission is inconsistent with a metagalactic cosmic-ray concept. Since the Magellanic Clouds are the nearest

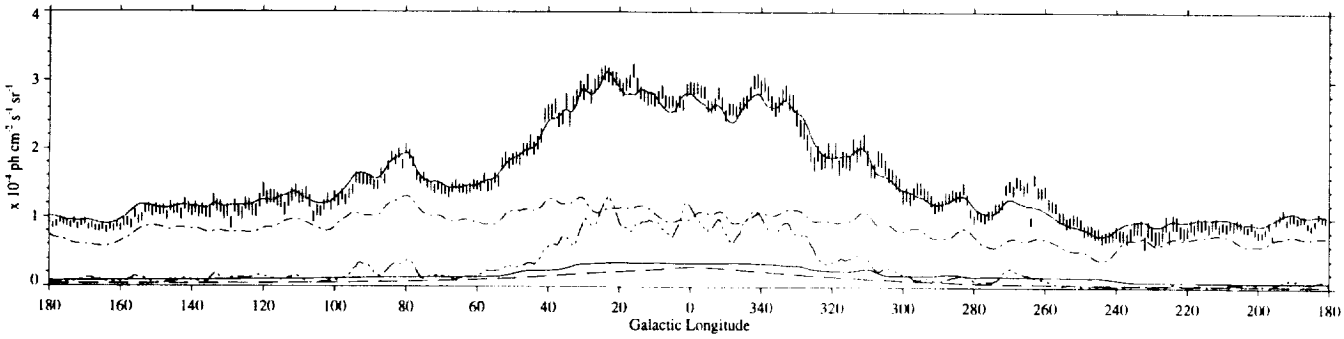


Figure 3.10: Predicted and observed high-energy, gamma-ray galactic diffuse radiation within 10° of the galactic plane. Hunter et al. (1996).

Table 3.1: Predicted and Observed Values for the Magellanic Clouds High Energy Gamma-Ray Emission

LARGE MAGELLANIC CLOUD		
Predicted Intensities[1]		
		Photons (E > 100 MeV) cm^{-2} s^{-1}
	Quasi-Stable Equilibrium	(2.0 ± 0.4) x 10^{-7}
	Universal Cosmic Rays	(2.1 ± 0.4) x 10^{-7}
Observed[2]		(1.9 ± 0.4) x 10^{-7}
SMALL MAGELLLANIC CLOUD		
Predicted Intensities[3]		
	Universal Cosmic Rays	(2.4 ± 0.4) x 10^{-7}
	Quasi-Stable Equilibrium	(1.2 ± 0.3) x 10^{-7}
	Disrupted	(0.15 *to* 0.3) x 10^{-7}
Observed[4]		< 0.5 x 10^{-7}

1. Fichtel et al. (1991)
3. Sreekumar and Fichtel (1991)
2. Sreekumar et al. (1992)
4. Sreekumar et al. (1993)

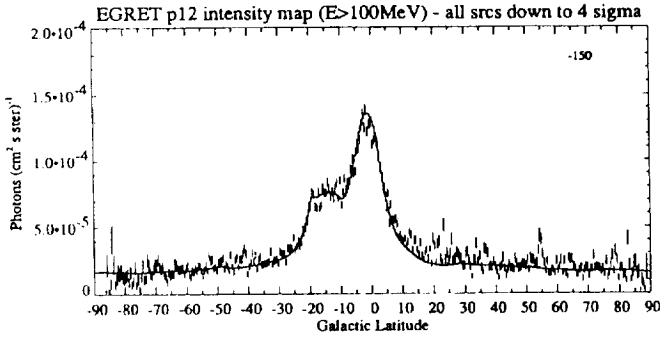


Figure 3.11: Comparison of the predicted and observed high-energy, gamma-ray diffuse radiation of a function of galactic latitude for $l = 210°$. The figure is from Sreekumar et al. (1995). Note the molecular cloud between $b = -10°$ and $b = -20°$.

galaxies to our own, the only reasonable conclusion is that cosmic rays are galactic. This important conclusion was anticipated on the basis of several less direct measurements of other phenomena, but the observation described here is the definitive one.

Having established that cosmic rays are galactic and not metagalactic, the value shown in Table I for the Large Magellanic Cloud, on the other hand, shows that the cosmic rays in the Large Magellanic Cloud are in quasi-stable equilibrium or at least close to it.

From Table 3.1, it is seen that not only is the high-energy gamma-radiation from the Small Magellanic Cloud not consistent with the metagalactic concept, but also it is not consistent with quasi-stable equilibrium. This finding is consistent with the belief that the Small Magellanic Cloud is in a state of irreversible disintegration in agreement with the tidal interaction model of Murai and Fujimoto, (1980) and the experimental work of Mathewson, Ford, and Visvanathan, (1986 and 1988). The specific predicted level shown in the table was calculated by Sreekumar and Fichtel (1991) based on the synchrotron data and on the assumption that the Galaxy was disrupted. Thus, the conclusion is that the Small Magellanic Cloud is in a disrupted state presumably as a result of an encounter with the Large Magellanic Cloud in the past and that it is not in quasi-stable equilibrium.

3.8 SUMMARY

The high-energy gamma-ray sky is dominated by diffuse radiation from the galactic plane. The distribution has some similarity to the 21 *cm* radio sky, but is even more enhanced in the central quadrant of the galactic plane. The majority of the high-energy gamma-radiation is explained well by the interaction of the galactic cosmic rays with the interstellar matter and photons. The radiation is specifically consistent with the assumption that the cosmic rays are galactic and vary with position in the Galaxy. The concept that the cosmic rays are galactic is supported by a number of observations, but particularly by the recent measurement of the high-energy gamma-ray emission from the Small Magellanic Cloud, which shows that the bulk of the cosmic rays are galactic, and hence play an important role in galactic dynamic balance. The cosmic-ray density seems to be at the local value in nearby galactic molecular clouds–that is, there appears to be nothing special about these clouds. The low value of the high-energy gamma-radiation coming from the Small Magellanic Cloud is consistent with that Galaxy being in a state of disruption, consistent with other observations. The Large Magellanic Cloud, however, seems to be in quasi-stable equilibrium, or at least close to it.

3.9 REFERENCES

Baldwin, J.E., 1967, in "Radio Astronomy and the Galactic System" IAU Symp., 31, 337.

Baldwin, J.E., 1976 in "The Structure and Content of the Galaxy and Galactic Gamma Rays" NASA CP-002 (Washington, DC) 206.

Bertsch, D.L., Dame, T.M., Fichtel, C.E., Hunter, S.D., Sreekumar, P., Stacy, J.G., and Thaddeus, P., 1993, ApJ.

Biermann, L., and Davis, L.Jr., 1960, Zs, F., ApJ, 51. 19.

Blandford, R.D., and Ostriker, J.P., 1980, ApJ, 237 , 793.

Cavallo, G., and Gould, R.J., 1971, Nuovo Cimento, 2B 77.

Dermer, C.D., 1986, A and A, 157, 223.

Digel, S.W., Hunter, S.D., and Mukherjee, R., 1995, ApJ, 441, 270.

Esposito, J., Hunter, S., Kanbach, G., and Sreekumar, P., 1994, AAS Meeting, Minn. MN.

Fichtel, C.E., and Kniffen, D.A., 1974, in "High Energy Particles and Quanta in Astrophysics" ed. McDonald and Fichtel, 428.

Fichtel, C.E., Hartman, R.C., Kniffen, D.A., Thompson, D.J., Bignami, G.F., Ögelman, H.B., Özel, M., and Tümer T., 1975, ApJ, 198, 163.

Fichtel, C.E., and Trombka, J., 1981, NASA SP-453, Chapter 12.

Fichtel, C.E., Özel, M., Stone, R., and Sreekumar, P., 1991, ApJ, 374, 134.

Fichtel, C.E., and Trombka, J.I., 1996, in "Gamma Ray Astrophysics, New Insight into the Universe" Second Edition, to be published.

Ginzburg, V.L., and Syrovatskii, S.I., 1964, in "The Origin of Cosmic Rays" (Oxford: Pergamon Press, 1964).

Ginzburg, V.L., 1972, Nature, 239, 8.

Ginzburg, V.L., and Ptuskin, V.S., 1976, Rev. Mod. Phys. 48, 161.
Hartman, R.C., Kniffen, D.A., Thompson, D.J., and Özel, M.E., 1979, ApJ, 230, 597.
Hunter, S., et al., 1996, submitted to ApJ.
Hunter, S.D., de Geus E.J., Digel, S.W., Kanbach, G., 1994, ApJ, to be published.
Kniffen, D.A., Hartman, R.C., Thompson, D.J., and Fichtel, C.E., 1973, ApJ, Letters, 186, L105.
Kraushaar, W.L., Clark, G.W., Garmire, G.P., Borken, R., Higbie, P., Leong, C., Thoros, T., 1972, ApJ, 177, 341.
Mathewson, D.S., Ford, V.L., and Visvanathan, N., 1986, ApJ, 301, 664.
Mathewson, D.S., Ford, V.L., and Visvanathan, N., 1988, ApJ, 333, 617.
Mayer-Hasselwander, H. et al., 1980, in "Annals of the New York Academy of Sciences" Proceedings of the 19th Texas Symposium, eds. J. Eklers, J.J. Perry, and M. Walker.
Morris, D.J., 1984, J. Geophys. Res. 89, 10685.
Murai, T., and Fujimoto, M., 1980, Publ. Astron. Soc. Japan, 32, 581.
Parker, E.N., 1966, ApJ 145, 811.
Parker, E.N., 1969, Space Sci. Rev. 9, 654.
Parker, E.N., 1977, in "The Structure and Content of the Galaxy and Galactic Gamma Rays" Space Sci. Rev., NASA CP 002, 283-300 (U.S. GPO, Wash, DC).
Simpson, J.A., 1983, in "Elemental and Isotopic Composition of the Galactic Cosmic Rays" Ann. Rev. Nucl. Part. Sci., 33, 323.
Sreekumar, P., and Fichtel, C.E., 1991, Astron. Astrophys., 251, 447.
Sreekumar, P., et al., 1992, ApJ, 400, L67.
Sreekumar, P., et al., 1993, Phys. Rev. Letters 70, 127.
Sreekumar, P., et al., 1995, submitted to ApJ.
Stecker, F.W., 1971, in "Cosmic Gamma Rays" (Baltimore, MD: Mono Book Corp., 1971).

Chapter 4

THE ROLE OF X-RAY OBSERVATIONS IN UNDERSTANDING SUPERNOVAE AND SUPERNOVA REMNANTS

R. PETRE
Laboratory for High Energy Astrophyics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

Supernova (SN) explosions are the most energetic events observed in our Galaxy. The 10^{51} ergs of thermal energy they release is one of the primary energy sources for the interstellar medium. They are responsible for dispersion of all atomic species heavier than He. They also leave behind the exotica of the stellar menagerie, neutron stars and black holes. The explosive release of such a large amount of thermal energy leads to the heating via shock waves of both the supernova ejecta and the surrounding interstellar medium to X-ray emitting temperatures for tens of thousand of years after the explosion. X-ray imaging and spectroscopic studies of supernovae and their remnants provide insight into the nature of the progenitor, the symmetry of the explosion, the degree of mixing of ejecta, the structure of the ambient interstellar medium, and the presence of particle acceleration by a pulsar or via shock acceleration. We present an up to date discussion of X-ray observations, placed in the context of the fundamental issues that can be addressed.

4.1 SUPERNOVAE AND SUPERNOVA REMNANTS IN THE GRAND SCHEME

One of nature's most dramatic events is a supernova, the explosive self-destruction of a star. A supernova event marks the end of the life of one star, but stimulates the birth of the subsequent generation. These events play a number of critical roles in galactic evolution. First, supernovae are the only means by which most atoms heavier than He are dispersed into interstellar space. It is no understatement to say that we are supernova remnants. Second, many atomic species are created exclusively in the explosive nucleosynthesis that takes place during a supernova event. Third, the energetic shock waves produced by supernovae trigger the collapse of interstellar clouds into new stars. And finally, and most interesting for high energy astrophysicists, supernovae give rise to the exotica of the stellar menagerie, namely neutron stars and stellar mass black holes.

4.2 WHAT CAN X-RAY OBSERVATIONS TELL US?

A supernova explosion is extremely violent: a total energy of 10^{53} ergs is produced via the gravitational collapse of the stellar core, mostly in the form of neutrinos. The small percentage of this that becomes kinetic energy ($\sim 10^{51}$ ergs) is still an enormous amount of energy, sufficient to heat hundreds of solar masses of material to millions of degrees, temperatures at which the primary emission band is the X-ray. For most of their lives, supernova remnants emit copiously in the X-ray band. By contrast, most supernova events themselves emit X-rays only briefly, if at all.

While X-ray astronomy as a whole is a relatively young field, the observation of supernovae is younger still. X-ray observations of supernovae can in principle reveal some important information. For instance, it can reveal the mass, composition, and spatial distibution of the matter in the circumstellar medium. Since this medium consists primarily of material blown off from the star as a wind during its blue supergiant phase, following the evolution of the X-ray emission from a supernova can provide us a history of the stellar wind evolution. Additionally, X-ray observations could reveal the presence of a central compact object well before the shell of ejected material becomes transparent at visible wavelengths. Finally, observation of line emission from Fe and Ni (whose K shell transitions fall in the X-ray band around 7 keV) can reveal the amount of these metals expelled by the explosion. As will be discussed below, X-ray observations of supernovae have not yet fulfilled their potential, largely because of the small number of events that have been observed using detectors capable of making these kinds of measurements.

On the other hand, X-ray observations of supernova remnants (SNR) is nearly as old as the field of X-ray astronomy. The Crab Nebula, the remnant of a supernova that was observed in 1054, is one of the strongest X-ray emitters in the sky, and was the first X-ray source identified with a known object. While the evolutionary timescale of SNR is sufficiently long to prevent us from detecting (in ~ 35 y) any evolution, they are ideal targets for both imaging and spectoscopic observations. X-ray imaging observations can reveal a number of important properties, including the symmetry of the inital explosion, the spatial distribution of the hot gas, the underlying structure of the medium into which the remnant is expanding, the clumpiness of the supernova ejecta, the details of the interaction between strong shock waves and small and large obstacles (interstellar clouds), and the presence or absence of a stellar remnant (a neutron star). X-ray spectroscopic studies reveal complementary information, including the atomic composition of the ejecta (in young remnants) and the interstellar medium (in old remnants), the amount of thermal energy liberated in the explosion, the mass of the progenitor star, and the effects of a pulsar. Most recently, it has been possible to combine these two approaches and perform spatially resolved X-ray spectroscopy. This technique has yielded a number of important discoveries, such as the fact that in some supernova remnants the ejecta are still stratified as they were in the outer layer of the progenitor star. Most recently, this technique has been used to prove the 40 year old conjecture that supernova remnant shocks are the site of acceleration of high energy cosmic rays.

4.3 SUPERNOVAE

There are two basic classes of supernovae, as determined via optical spectroscopy. There are also two basic physical mechanisms for producing supernovae. Interestingly, there is not a one-to-one match between these two schemes.

Optically, supernovae are classified by the presence or absence of certain lines in their optical spectra. Remnants are classified as Type I if no hydrogen lines are detected. These kinds of remnants are usually (but not always) found in elliptical galaxies, or sites dominated by an old stellar population (Population II stars). Type II remnants, on the other hand, do show H lines, primarily from the Balmer series. These objects are found mainly in the spiral arms of galaxies, usually near H II regions. Hence they are associated with sites of massive star formation.

A closer look at the Type I remnants reveals a bifur-

cation into three subclasses, Types Ia, Ib, and Ic. Type Ia supernovae are distinguished observationally from Types Ib and Ic by the presence of silicon absorption lines in their optical spectra at maximum light. Similarly, Types Ib and Ic are distinguished by the presence (Type Ib) or absence (Type Ic) of helium lines. A more physical distinction is that Types Ib and Ic are found in similar locations to Type II events, and are the only Type I events that can have strong radio emission (often detected from Type II events). The presence of radio emission means that the supernova shock is encountering an envelope of circumstellar material. Thus the Type Ib and Ic supernovae are thought to be physically more related to Type II events than Type Ia events, and are massive stars that have lost their hydrogen envelopes prior to exploding.

The fact that some Type I events are more similar to Type II events than they are to other Type I's shows how misleading the optical classification scheme is. This is underlined by the fact that some supernovae change their identities. For instance, the spectrum of SN 1993j, in the nearby spiral galaxy M81, changed from a that of a Type II to a Type Ib (and the supernova is now classified as a Type Ic). The rest of this chapter will draw the distinction between supernovae and their remnants based on the explosion mechanism.

4.4 CREATION OF SUPERNOVAE

The two basic mechanisms that produce supernovae are the collapse of the core of a massive star, and the disruption of a solar mass star.

Type II, Ib, and Ic supernovae are almost certainly due to the collapse of the Fe rich core of a massive star ($\geq$9 $M_\odot$). Much of the strongest observational evidence for this comes from the recent SN 1987a, in the Large Magellanic Cloud. As a massive star evolves, its inner layer undergoes successively higher stages of nuclear burning. At each stage, the energy yield per atom becomes lower. Once an atomic nucleus is transmuted to iron, there is no additional energy that can be liberated via fusion, and it becomes inert. After a sufficient mass of Fe has been produced at the core, about 1.5 $M_\odot$, the central temperature, determined by the gravitational compression, rises above $3\times10^9 K$, and the nuclei begin to photodisintegrate. The pressure of the outer core can no longer counteract the gravitational force of the outer layers, and the star collapses. This collapse releases an enormous amount of gravitational energy, 10^{53} ergs. This is mainly released in the form of neutrinos produced in the photodissociation of the core atoms. About one percent of this, 10^{51} ergs, is converted into bulk motion, which drives off the stellar envelope. It should be noted that the details of how the neutrino luminosity is transferred into bulk motion is not well understood. One outcome of the core collapse is that very little Fe survives; most is swallowed in the core or destroyed (only about 0.07$M_\odot$ of Fe was observed in the envelope of SN 1987a). The core can become a neutron star or a black hole, or it could be completely disrupted.

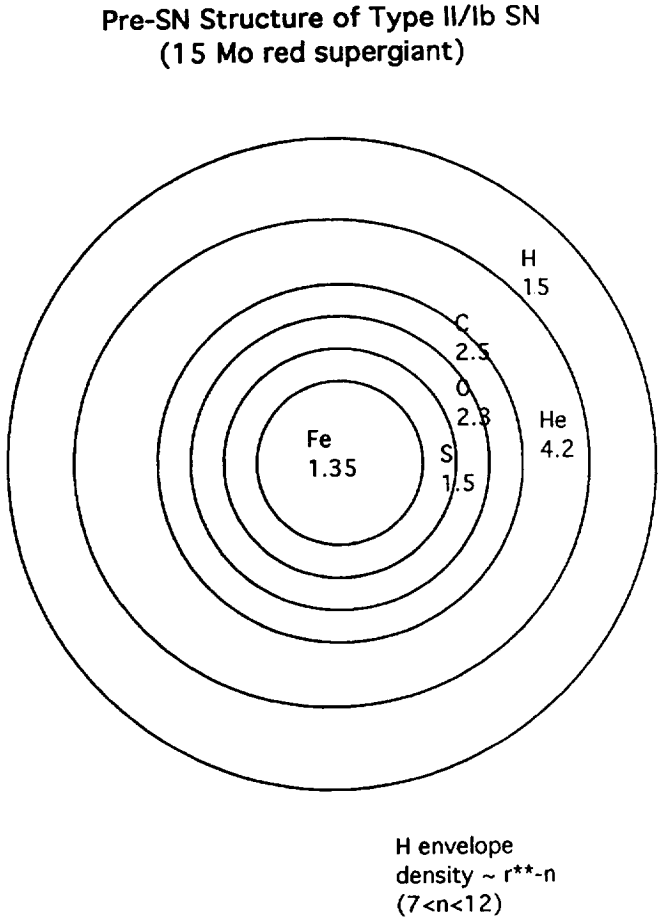


Figure 4.1: Internal structure of a massive star prior to gravitational collapse and creation of a Type II supernova. Primary fuel in each shell is indicated, along with shell mass in solar masses.

The mechanism that produces a Type Ia explosion is still not well understood, although many of the basic attributes of the event are. It is currently thought that these events are due to either detonation or deflagration of a carbon-oxygen white dwarf star. (The difference between deflagration and detonation is whether the velocity of the nuclear burning front through the stellar interior is supersonic (detonation) or subsonic (deflagration)). This occurs when a white dwarf accreting mass from a binary companion approaches a critical core mass of 1.38$M_\odot$. At this point, the energy produced by the nuclear burning of carbon exceeds the neutrino losses, the star experiences a nuclear runaway, and it explodes. Only a small stellar

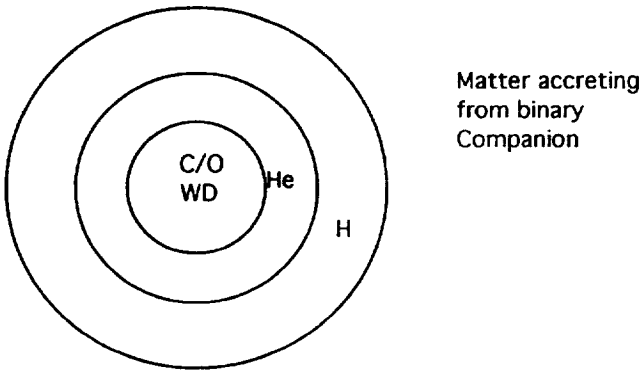


Figure 4.2: Internal structure of a carbon-oxygen white dwarf prior to explosion as Type Ia supernova.

remnant or no remnant at all is left behind. Deflagration models predict that half of the stellar mass is converted to Fe. The remainder of the mass becomes so-called intermediate a-burning products, such as Mg, Si, and S. In the detonation models, most of the mass is Fe. The late time light curves of these supernovae are powered by the decay of ^{56}Ni produced during the explosion.

4.5 X-RAY EMISSION MECHANISMS

There are four mechanisms by which X-rays might be produced in supernovae: impulsive shock heating, downscattering of gamma-rays from radioactive decay, ejecta-wind interaction, and leakage of synchrotron radiation from an underlying pulsar.

Breakout: The breakout of the supernova shock through the surface of the progenitor star in a Type II event could produce a burst of prompt X-ray emission lasting for ~1,000 s (Falk, 1978; Chevalier and Klein, 1989). A radiative precursor to the breakout might occur when the shock approaches the radius at which the optical depth to X-rays is ~25. The X-ray luminosity of this precursor could in principle reach 10^{45} erg s^{-1} (a luminosity for a bright X-ray quasar) with a temperature of ~2×10^5 K. This would correspond to very soft X-rays, less than 1 keV. As the shock breaks through the stellar surface, it could produce 100 keV bremsstrahlung radiation from a thin layer of shock-heated surface material at temperatures of 10^8-$10^9 K$. This mechanism has not been observed in the X-ray band.

Radioactive decay: Radioactive ^{56}Ni is produced abundantly in the nuclear statistical equilibrium attained during the explosion as a result of the high temperature and density. The ^{56}Ni is quickly frozen out on a dynamical timescale of $\leq$1 s. It decays by the following path

$$^{56}\mathrm{Ni} \xrightarrow[6.1\mathrm{d}]{} {}^{56}\mathrm{Co} \xrightarrow[78.5\mathrm{d}]{} {}^{56}\mathrm{Fe}$$

Significant gamma-radiation accompanies each reaction, as decay products are produced in excited nuclear states. The key lines are

$$^{56}\mathrm{Ni} \rightarrow {}^{56}\mathrm{Co}$$
$$0.158 \text{ MeV}$$
$$^{56}\mathrm{Co} \rightarrow {}^{56}\mathrm{Fe}$$
$$0.847 \text{ MeV}, 1.238 \text{ MeV}, 0.511 \text{ MeV}$$

Deflagration/detonation supernovae produce $\geq$ 0.5 $M_\odot$ of ^{56}Ni per event, core collapse supernovae produce $\leq$ 0.1 $M_\odot$. (Observational proof of the production of ^{56}Ni came from detection of the gamma ray lines from SN 1987A). X-rays are produced by multiple Compton downscattering of the gamma rays. Models of this downscattering predict a low energy cutoff of the X-rays at $E_{crit} \sim 30\ \zeta^{1/4}\ keV$, (where ζ represents the metallicity of the gas) but not the shape of the spectrum. It is unclear whether the high energy X-rays observed from SN 1987A starting ~6 months after the explosion were produced by this mechanism.

Ejecta/wind interaction: When the high velocity stellar ejecta encounter resistance in the form of circumstellar matter from the progenitor wind, shock waves are produced. One shock wave propagates rapidly through the relatively low density wind material, while a "reverse shock" propagates into the ejecta. This reverse shock heats the ejecta to X-ray emitting temperatures (Chevalier and Fransson, 1994). The denser the wind, the stronger the reverse shock, an the higher the luminosity of the X-rays. The luminosity arising from this mechanism is estimated to be

$$L \sim 1.6\times10^{39}\left(\frac{\dot{M}}{10^{-5}M_\odot y^{-1}}\right)\left(\frac{w_1\ r_{15}^{-1}\ T_9^{1/2}}{10\ km\ s^{-1}}\right) erg\ s^{-1}$$

where $\dot{M}$ represents the presupernova mass loss rate, w_1 represents the progenitor wind velocity, and T_9 the post shock temperature in units of $10^9 K$. For most supernovae, the values of $\dot{M}$, w_1, and T_9 are expected to be less than unity. However, the observed properties of more than one of the supernovae detected in X-rays suggest this is not always the case.

Another idea is that strong X-ray emission could be produced via much the same mechanism if the stellar wind material has dense clumps in it (Chugai, 1993). As the ejecta will be overabundant in Fe and intermediate α-burning elements compared with the wind, careful X-ray

spectroscopic measurements should be able to discriminate between these two models.

Underlying pulsar: Pulsars, like that in the Crab Nebula, are prodigious emitters of synchrotron radiation. Some of this radiation would "leak out" of the expanding supernova shell, but not until the later stages (years after the explosion), when the shell becomes optically thin to X-rays.

4.6 X-RAY OBSERVATIONS OF SUPERNOVAE

It is only in recent years that X-ray instruments have had the sensitivity, and the good fortune, required to detect X-rays from supernovae. There are now several detections of supernovae at varying ages, from 5 days to 16 years. Curiously, most of these would not be classified as "typical" supernovae. Below are summarized the detections and some of the more interesting non-detections.

4.6.1 SN 1980K

This was a Type II event in the nearby starburst galaxy NGC 6946. It was the first supernova detected in the X-ray band, observed using the imaging detectors on the *Einstein* Observatory (Canizares, Kriss, and Feigelson, 1982). It was detected at 35 days after optical maximum, but not at 82 days. Insufficient information was obtained to determine the emission mechanism, but the discoverers speculated that inverse Compton scattering of optical photons produced the X-rays, but could not rule out other mechanisms. Based on subsequent observations of other supernovae, it is more likely they were observing thermal emission from a SN shock/stellar wind interaction.

4.6.2 SN 1987A

This is the closest known SN in $\sim$ 300 years, and the first for which modern instrumentation facilitated detailed study. It was an unusual Type II event, with a blue supergiant progenitor (as opposed to a red supergiant, thought to be the progenitor of a more typical Type II). X-rays were first detected 130 d after optical maximum. The broad, hard X-ray spectrum (6-30 keV, with a possible Fe line; Dotani et al., 1987, Sunyaev et al., 1987) indicated that it arose from multiple components. The X-ray emission decayed over the course of $\sim$1 year. The X-ray emission above $\sim$15 keV is thought to arise from Compton downscattered γ-ray line radiation; the origin of the lower energy emission, especially the Fe line, has never been fully explained (see Sutherland, 1990 for a summary of the early observations and the attempts to interpret them). Recent ROSAT observations have detected SN 1987A weakly (Gorenstein, Hughes, and Tucker, 1994; Beuermann et al., 1994). It is thought that this emission arises from the interaction between a reverse shock and the expanding ejecta.

4.6.3 SN 1986J

This is another unusual supernova, in the spiral galaxy NGC 891. It was first discovered as a luminous radio source. Its high radio luminosity indicates the presence of a dense circumstellar medium. This is confirmed by the X-ray detection of the SN, at a luminosity of 2×10^{40} erg s^{-1} (Bregman and Pildis, 1992). The X-ray spectrum is thermal, with indications of line emission, which suggests that the X-ray emission arises from shocked ejecta (Houck and Bregman, 1996).

4.6.4 SN 1978K

This unusual supernova, in the nearby spiral galaxy NGC 1313, was the first supernova to be identified as such on the basis of its X-ray emission (Ryder et al., 1993). It was observed, but not detected, 18 months after its explosion in mid-1978. Thirteen years later, it was detected an X-ray luminosity of 1×10^{40}erg s^{-1}, at least a factor of five brighter than the early upper limit. Since the first detection in 1991, its luminosity and spectrum have not changed (Schlegel et al., 1996). The luminous X-rays, combined with a high radio luminosity, suggest that, like SN 1986J, this object has a dense circumstellar medium, in this case as large as 80 $M_\odot$. Curiously, the X-ray spectrum shows no evidence for X-ray lines, and might be consistent with synchrotron emission (Petre et al., 1994).

4.6.5 SN 1993J

Located in the nearby spiral galaxy M81, this unusual supernova evolved from a Type II into a Type Ib. This metamorphosis can be explained if the progenitor star had a very thin hydrogen envelope. Contrary to all theoretical models existing at the time, X-rays were observed from this SN 5 days after optical maximum, and they are still detectable after over 2 years (Zimmermann et al., 1994). In its first $\sim$8 months, the temperature of the X-ray emission declined from $\sim$10 keV (with a strong, broad Fe K line) to $\sim$1 keV. The X-ray evolution of SN 1993J has provided the first opportunity for us to study the wind structure in a red supergiant.

4.6.6 SN 1988Z

This is the most distant, and with a luminosity of $\sim1\times10^{41}$ erg s^{-1}, the most luminous supernova detected in X-rays (Fabian and Terlevich, 1996). It is another member of the class of supernovae with evidence for massive circumstellar envelopes. If the luminosity of the

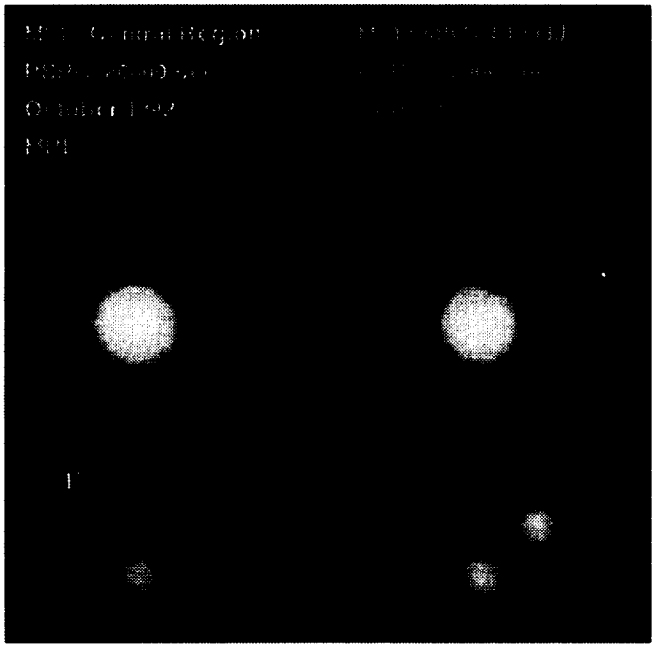


Figure 4.3: ROSAT images of M81, before (left) and after (right) the discovery of SN 1993J.

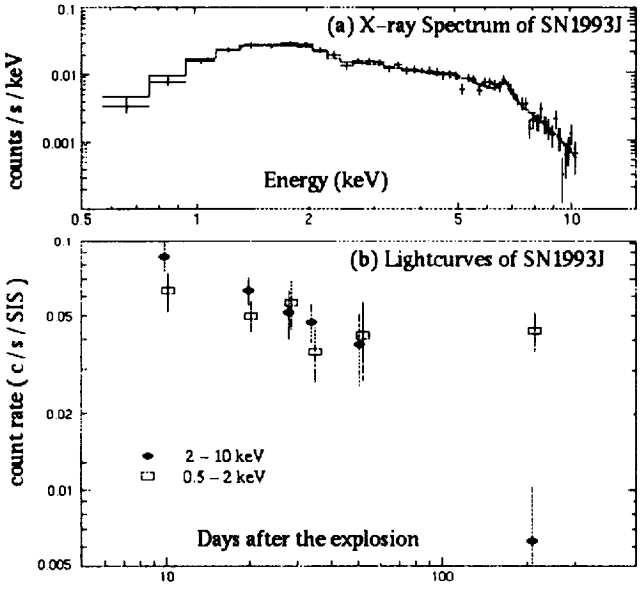


Figure 4.4: ASCA spectrum of SN 1993J, 9 days after optical maximum. The temperature is higher than 10 keV, and a broad Fe emission line is marked by the arrow (from Kohmura et al., 1994.)

supernova does not change dramatically, it will radiate away all its kinetic energy in approximately 300 y (compared with $\sim 10^5$ y for a supernova remnant in a more typical environment.

4.6.7 SN 1992A

This is an important non-detection (Schlegel and Petre, 1993). SN 1992A was an archetypical Type Ia supernova, in the elliptical galaxy NGC 1380. The X-ray upper limit of $3\text{-}5 \times 10^{38}$ erg s^{-1} is more than an order of magnitude lower than the previous best upper limit on emission from Type Ia's. This limit constrains the mass loss rate of the progenitor to less than $10^{-6} M_\odot$ y^{-1}.

4.6.8 Summary

It is clear from the above that the study of X-ray emission from supernovae is still in its infancy. Virtually all the detected supernovae are anomalous in one way or another, and there is still a substantial element of luck involved in finding X-ray supernovae. Systematic studies of groups of supernovae aimed at determining X-ray flux have only just begun. A new generation of X-ray observatories with unprecedented sensitivity (NASA's AXAF, ESA's XMM, Japan's Astro-E) will provide the clearest view of X-ray emission from supernovae. And should by chance a superova explode in our Galaxy in the next few years, we are extremely well equipped to learn all we can from whatever X-rays it emits.

4.7 SUPERNOVA REMNANTS

Supernova remnants are one of the primary sources (along with the winds from massive stars) of heating the interstellar medium (ISM). With a typical total energy content of 10^{51} ergs, they can also power the galactic halo. SNR are the primary means of transport of heavy elements, and the interaction between their shock fronts and interstellar clouds catalyze star formation.

Observationally, SNR are complicated objects, due to the influence of both heredity (complexities in the SN explosion) and environment (the multiphase ISM into which they expand). We'll start with an idealized view of SNR in order to establish a perspective on the observations, and then discuss the complications that have the most important ramifications for X-ray observations. We'll also ignore the issue of pulsars and synchrotron nebulae, which are the subject of another lecture.

4.8 SNR EVOLUTION

The lifecycle of a SNR can be broken up into the main stages, free expansion, the adiabatic phase, and the radiative phase.

In the initial stage of expansion, the ejecta expand unimpeded, carried along behind the supernova blast wave. Energy from the blast wave is transferred to the ISM (in ways we do not fully comprehend), and a pair of shock fronts is formed. An outwardly moving shock front

marks the periphery of the the remnant, which sweeps up and accelerates the outer presupernova circumstellar medium. A so-called "reverse" shock (in the ejecta frame of reference) propagates into the ring of ejecta. The transition to the next evolutionary stage occurs when the forward shock begins decelerating as a consequence of sweeping up several ejecta masses of interstellar and circumstellar material. The exact age when this occurs depends upon the initial shock velocity and the density of ISM, but it is thought to happen when the forward shock has swept up from the ISM 1-10 times the ejecta mass.

The second expansion stage is referred to as adiabatic because radiative losses are small compared with the total energy, meaning that the energy of the remnant is (nearly) conserved. On the other hand, the gas moving across the shock suffers a large entropy increase. During this phase a strong "collisionless" shock forms, and sweeps up the interstellar medium. Kinetic temperatures of $\sim 10^7 K$ are typical of this phase. The dynamics of a supernova remnant propagating through a uniform ISM can be described by a self-similar blast wave model, first formulated independently by Sedov (1959) and Taylor (1950). In this model, the dynamics depend on only two parameters, the age t, and E/n, the ratio between the explosion energy E and the ambient density n.

$$\begin{aligned} R_s &= 13.7\, t_4^{2/5} \left(\frac{E_o}{n_o}\right)^{1/5} pc \\ V_s &= 535\, t_4^{-3/5} \left(\frac{E_o}{n_o}\right)^{1/5} km\ s^{-1} = R_s^{-3/5} \\ T_s &= 3.2 \times 10^6\, t_4^{-6/5} \left(\frac{E_o}{n_o}\right)^{2/5} K \\ n_s &= 4\, n_o \end{aligned} \tag{4.1}$$

The density of the shocked material is low, at most only a few particles per cubic centimeter. Since the only mechanism for ionization is electron-ion collisions and these are rare at such low densities, atoms retain their ionization state for a long time.

The onset of the radiative stage takes place when the radiative cooling timescale becomes comparable with the expansion timescale. This happens approximately at

$$t^{(d)} = 2.9 \left(\frac{E_o}{n_o}\right)^{2/11} (10^{22} L n_o)^{5/11} \times 10^4\ yr \tag{4.2}$$

An abrupt and strong change in the character of the shock occurs at this time. Since the radiative cooling is now efficient, the shocked material cools rapidly. This produces a drop in the pressure immediately behind the shock. This in turn causes more rapid deceleration of the shock, as well as an increase of the velocity of the material further inside. As a result, most of the remnant mass compresses into a thin dense shell as it cools and collects behind the decelerated shock. The shock becomes isothermal, with its temperature driven by the that at which the dominant atomic transitions occur. The expansion is driven by the momentum and the pressure of the remnant interior. Meanwhile, the interior of the remnant is too hot to cool radiatively, so it cools adiabatically as the shell expands. The radius R_s and velocity v_s of the shell are characterized by the following expressions

$$\begin{aligned} R_s &= 21 \left(\tfrac{E_o}{n_o}\right)^{5/22} (10^{22} L n_o)^{9/44} \left(\tfrac{t}{t^{(d)}}\right)^{1/4} pc \\ V_s &= 280 \left(\tfrac{E_o}{n_o}\right)^{5/22} (10^{22} L n_o)^{9/44} \left(\tfrac{t}{t^{(d)}}\right)^{-3/4} km\ s^{-1} \end{aligned} \tag{4.3}$$

Ultimately (after about 10^6 y) when the shell velocity decreases to 10-20 km s^{-1}, its internal pressure becomes comparable with that of the ISM. At that point, further expansion is impeded, and the SNR disappears into the ISM.

The idealization presented here is naive. The ISM is full of structures on all size scales. These include small ($\sim$1 pc), isolated clouds, large, dense molecular and HI clouds, walls of swept up material left behind by older SNR, and cavities produced by SNR and OB associations. Oftentimes it is these structures, and not the ideal, that the X-ray observations confront.

4.9 X-RAY PRODUCTION IN SNR

As most of the SNR observed in the X-ray band manifest at least some of the properties of the adiabatic stage (and the models we use to model the emission are based assume the adiabatic stage), the discussion of X-ray production here will focus on that stage. As SNR in this stage have shock velocities corresponding to 1×10^6 - $5 \times 10^7 K$, this is the stage at which the X-ray luminosity peaks. (Indeed, once the remnant reaches the radiative stage, the dominant coolant is the Lyα line of hydrogen so the temperature is more characteristically $\sim 10^4 K$.) In this stage, X-rays are one of the three prime emission bands. The others are the radio whose luminosity is often comparable, and the infrared which sometimes dominates. Unlike the other bands, however, the X-ray observations provide more fundamental information about the remnant, its composition and its dynamics. This is because the X-ray emission arises primarily as line emission from interstellar and ejecta gas heated directly by the forward and reverse shocks, while the radio arises as synchrotron emission from electrons along the shock front, and the infrared from dust whose heating is less

closely coupled to the remnant dynamics. The X-ray luminosity of a SNR, L_x, can be described by

$$L_x = \int dV \int dE n e^2 j(E,T)\ erg\ s^{-1}$$

where j(E,T) is the emissivity of the plasma and n_e is the electron density. In the soft X-ray band, if we assume a constant density, then the integral $\int$dE j(E,T) is fairly constant when integrated over a fairly broad band. So the overall luminosity, and for imaging studies the surface brightness, is dependent on ne^2.

Various models for j(E,T) have been used historically. The most simplistic view is that the gas is fully ionized and dominated by the thermal bremsstrahlung from the hot electrons. This model worked fine in the early days of X-ray astronomy, until spectrometers were developed with sufficient spectral resolution to demonstrate that the emission is not dominated by a featureless continuum, but by X-ray emission lines.

The gas in supernova remnants is low density, typically 0.1-10 cm^{-3}. It is therefore optically thin to its own radiation, a so-called coronal plasma. In such a gas, the only mechanism for ionizing or exciting atoms is collisions with electrons. If one assumes the gas is in ionization equilibrium, then the population of a given ionic state depends only upon the atomic physics of collisional excitation and radiative recombination. Uncertainties that creep into the comparison of this (or any plasma model) with a measured spectrum arise from uncertainties in atomic physics parameters such as collision strengths and in astrophysical parameters such as the "cosmic" abundances. Incorrect values of these affect the line strength predictions, and thus lead to faulty conclusions about the plasma.

A key fact about the atomic physics missing from the coronal plasma models is that the actual timescale for heating of electrons and the ionization of atoms in a low density medium can be longer than the age of the supernova remnant. For instance, any particle crossing a shock boundary is given a boost in velocity. Since the kinetic temperature is proportional to the square of the velocity, the electron gas is cooler than the ion gas by the ratio of the masses. If electron-ion Coulomb scattering is the only mechanism by which the temperatures equilibrate, then the electron equilibration time is

$$T_{eq} = 2T^{3/2} n_o^{-1}$$

or 8,000 y for $T=10^7 K$, and n_o=1 cm^{-3}. This is longer than the age of most remnants we observe. In the case of electron heating, there are other mechanisms which might come into play to equilibrate the electrons faster. For instance, microturbulence due to plasma instabilities might cause a rapid electron-ion energy exchange. Measurements of the solar wind bow shock against the earth

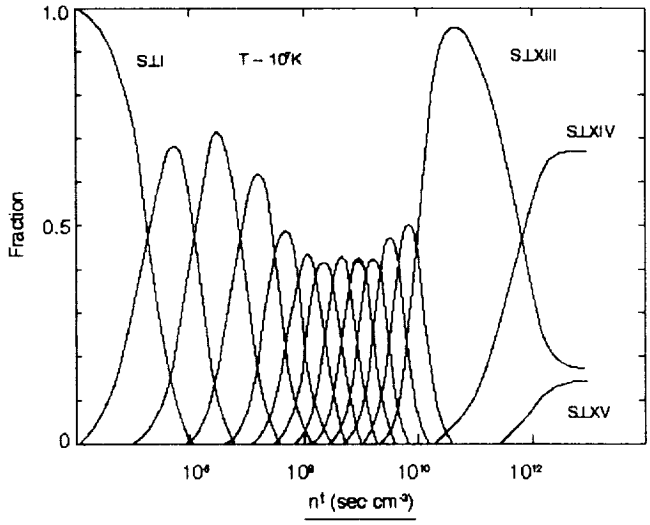


Figure 4.5: Ionization stage of Si as a function of nt, the plasma density times the time since the atom encountered the shock. Note that the atom enters its He-like state, that most often observed in X-rays, only after a long time.

shows a more rapid electron heating than would be expected from Coulomb processes alone, though electron-ion equilibration is never attained. Our current spectroscopic capabilities in the X-ray band are insufficent for addressing the issue of electron-ion equilibration.

More importantly than this, the collisions responsible for ionization are so infrequent that gas takes long time to reach ionization equilibrium. Thus the X-ray emitting gas in SNR is always underionized compared to its kinetic temperature, and always "ionizing up." The ionization history of a Si atom is shown in the figure below. In practical terms, this means that comparison of an X-ray spectrum with equilibrium models will give two wrong results. First, since the lines dominating the emission are at a lower ionization state than expected from the kinetic temperature, the models will yield too low a temperature. Also, since the strong lines are stronger than they would be in equilibrium, the model would yield too high an abundance. So to correctly interpret spectral information, need to compare with model that correctly includes time dependence of ionization.

A number of such non-equilibrium ionization (NEI) models exist (e.g., Hamilton, Sarazin, and Chevalier, 1983; Hughes and Helfand, 1985). In general, they also depend upon two parameters: a temperature, either the shock temperature, the electron temperature; and a parameter characterizing the ionization conditions. Two such parameters are nt, the product of the ambient density and the age of the explosion, and η=n^2E, where n is the ambient density and E is the initial kinetic energy of the explosion.

4.10 HIGHLIGHTS OF X-RAY SUPERNOVA REMNANT OBSERVATIONS

The study of the X-ray emission from supernova remnants is continually producing new and interesting results. Below are a few examples indicating the kinds of things we can learn by studying SNR images and spectra.

4.10.1 Cas A

Cas A is the youngest known galactic SNR, with an age of about 315 y, thought to be the remnant of a Type II event. It is also the brightest radio source in the sky and the brightest thermal X-ray emitting SNR (the Crab is brighter, but all of its emission is from nonthermal synchrotron processes). Its X-ray spectrum is quite dramatic, with strong lines from Mg through Fe (the lines from lighter elements are obscured due to line-of-sight absorption by interstellar gas). In Cas A, we observe several dramatic effects. First, in the integrated spectrum, the lines are broader than would be expected from thermal effects. If we measure the line centroids from various parts of the remnant, we find that they are narrow and shifted in energy in a systematic way; when blended into the integrated spectrum, the form the observed broadening. This is interpreted as a Doppler shift, with maximum velocity of $\sim \pm 2{,}000$ km s^{-1}. The distribution of shifts can be understood if the line emission is confined to an inclined ring on the sky, half of which is approaching and half receding (Markert et al., 1983; Holt et al., 1994). Also, a comparison between the distribution of emission from lines and from high energy continuum is qualitatively different. The continuum image resembles the radio image, suggesting that the continuum emission is related to the nonthermal processes responsible for the radio emission (Holt et al., 1994).

4.10.2 Typing the LMC remnants

For many remnants it is possible to use X-ray data to determine the nature of the progenitor; i.e., whether the supernova explosion was caused by core collapse of a massive star or deflagration/detonation of a white dwarf. It is possible to do this because the two types of explosions produce very different nucleosynthesis yields, which are reflected in their X-ray spectra. Type Ia supernovae produce relatively large amounts of Mg, Si, S and Fe, and virtually no O and Ne (e.g., Nomoto et al., 1986). Core collapse explosions, on the other hand, produce an overabundance of O and Ne compared with Si, S and Fe (e.g., Thielemann et al., 1994). The qualitatively different resulting X-ray spectra are shown in Figure 4.1.

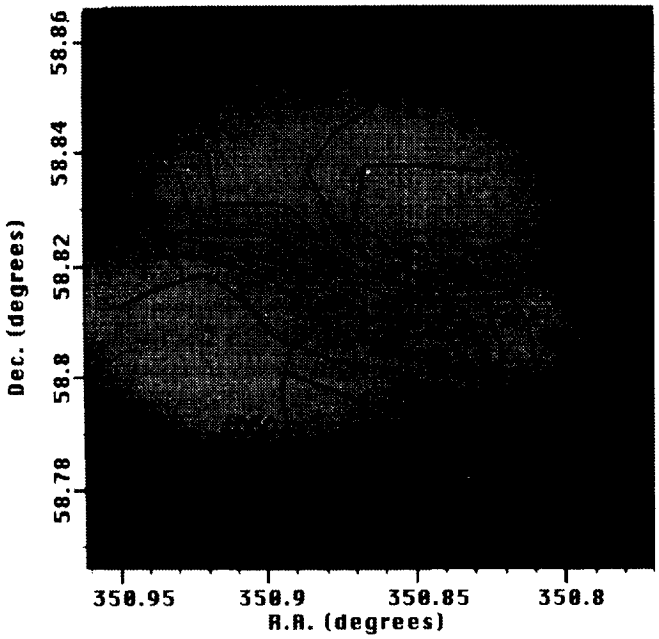


Figure 4.6: X-ray "Doppler" image of Cas A from ASCA spectra. The northwestern half of the remnant is receding, while the southeastern half is approaching. This suggests that the supernova ejecta are confined to an inclined ring. Figure from Holt et al., 1994.

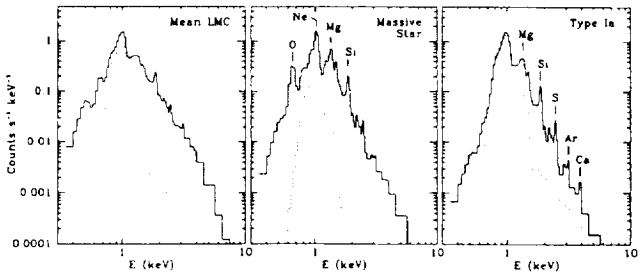


Figure 4.7: Simulated ASCA SIS spectra of shock heated gas with metal abundances typical of the Large Magellanic Cloud (left), LMC abundances enhanced by ejecta from a core collapse supernova (center), and LMC abundances enhanced by ejecta from a Type Ia supernova (right). The qualitative differences in the line ratios allow a straightforward determination of progenitor type for young SNR. Figure from Hughes et al. (1995).

The Magellanic Clouds are among the most convenient locations for studying the composition of supernova remnants, for a number of reasons. The Clouds contain quite a number of high luminosity remnants; they are sufficiently nearby that SNR there have fluxes suitable for observation by existing spectrometers; the small amount of intervening material in the direction towards

the Clouds means that, unlike most galactic remnants, the spectra of Cloud remnants are not subject to the severe absorption of low energy X-rays; and the metal abundances in the Cloud ISM are considerably less than solar, making differences due to enrichment by supernovae easier to detect in principle.

The most definitive work to date for LMC remnants (or any other remnants) using the X-ray spectral signature to determine the explosion type is that of Hughes et al. (1995). Their comparison of two remnants whose optical emission is dominated by H-α from nonradiative shocks, and one with a peculiar X-ray morphology, shows that all are products of Type Ia explosions.

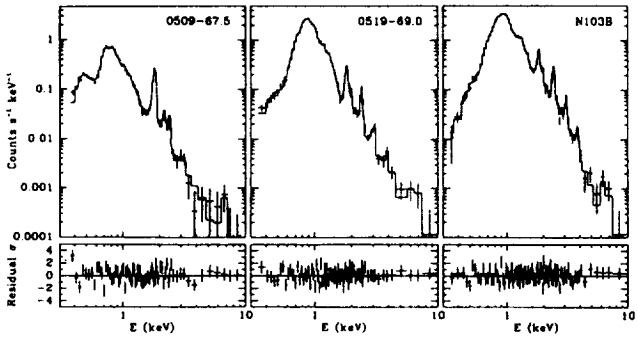


Figure 4.8: ASCA spectra three LMC supernova remnants. The line ratios compare favorably with those expected for a Type Ia remnant. Figure from Hughes et al. (1995).

In Figure 4.2 we show the X-ray spectrum of the brightest remnant in the SMC, E0102-72 (Hayashi et al., 1994). Its nearly symmetrical, limb-brightened X-ray morphology suggests a symmetric explosion into a relatively uniform interstellar medium (Hughes, 1994). Its small angular diameter (40 arc seconds) makes possible a comparison between its integrated spectrum and those produced by NEI models (Hamilton, Sarazin, and Chevalier, 1983; Hughes and Helfand, 1985). However, we find that even in this apparently symmetrical object no single component NEI model provides a satisfactory representation of the spectrum. To achieve an acceptable fit, at least three different (T, nt) components are required, one for each prominent set of lines. This unanticipated spectral complexity appears consistently in the supernova remnants observed by ASCA and constrains our ability to make more detailed comparisons with the predictions of nucleosynthesis models. In E0102-72 it indicates that each element is encountering different shock conditions, and thus the ejecta are stratified.

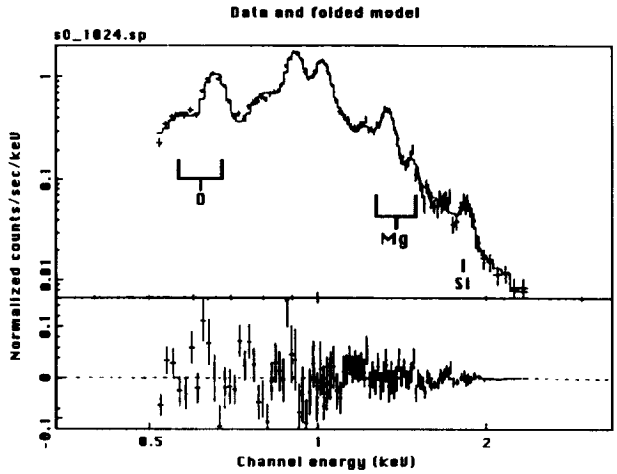


Figure 4.9: ASCA spectrum of SNR E0102-72 in the Small Magellanic Cloud. Unlike those of the LMC remnants in the figure above, this spectrum is dominated by O and Ne lines, consistent with a core collapse progenitor. Figure from Hayashi et al. (1994).

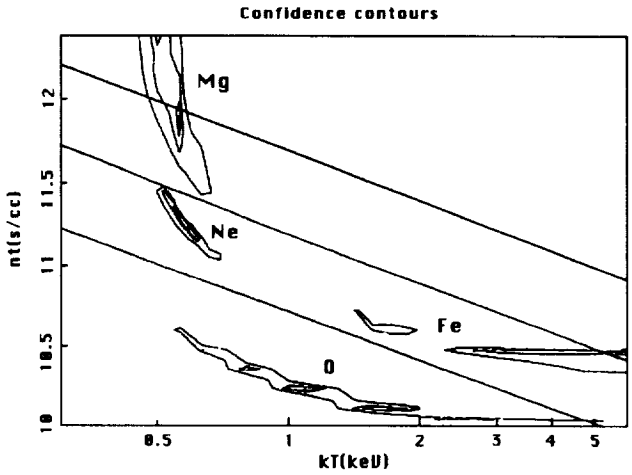


Figure 4.10: Confidence ranges for temperature and ionization parameter derived from the line ratios in the ASCA spectrum of E0102-72. The disjoint confidence ranges indicates each element is encountering different ionization conditions, and thus argues for stratification within the shell of ejecta. Figure from Hayashi et al. (1994).

4.10.3 Tycho

Tycho (SN 1572) is the archetype of Type Ia events. Its X-ray spectrum is one of the most closely studied of that of any SNR, and probably the best understood and the most consistent with conventional wisdom. The integrated X-ray spectrum can be understood in the context of a model with a forward shock front propagating

into a uniform ISM, and a reverse shock propagating into uniform density ejecta consisting of pure heavy elements, stratified into layers (Hamilton et al., 1986). The total ejecta mass of 1.4 $M_\odot$, about $1M_\odot$ of which is either Si group or Fe group metals, is clearly consistent with that expected from a Type Ia progenitor. Subsequent observations have confirmed the general predictions of this model. A BBXRT observation showed that the ionization conditions experienced by the Fe are different from those of the other metals, suggesting that the Fe filling the interior has lower density than the outer ejecta layers (Petre et al., 1993). A comparison between the Einstein and ROSAT HRI images (Vancura et al., 1995) and narrow band imaging using ASCA data (Hwang and Gotthelf, 1996) suggests the presence of minor temperature and abundance variations within the remnant. While the presence of the abundance variations indicate that at most incomplete mixing has taken place between the ejecta layers, modeling of the radial profiles of the ASCA narrow band images requires that some mixing must have occurred.

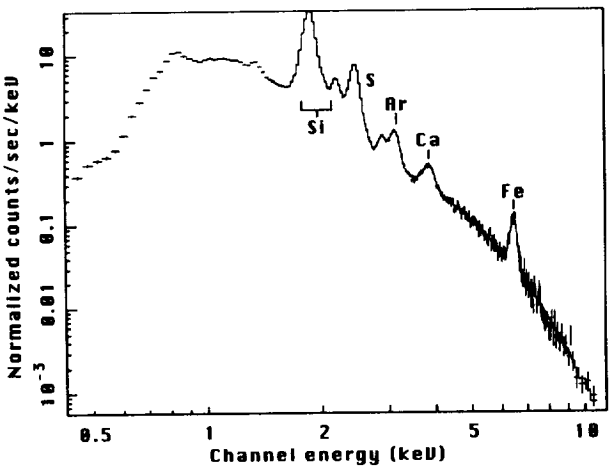


Figure 4.11: ASCA spectrum of Tycho. Spectrum is dominated by lines from Mg, Si, S, and Fe, as expected from a Type Ia remnant whose spectrum is dominated by emission from ejecta. Figure from Hwang and Gotthelf (1996).

4.10.4 SN1006

One of the most exciting ASCA results to date involves SN1006. Koyama et al. (1995) demonstrated that the dominant, featureless spectral component that has long posed a mystery, is isolated to two segments of the limb, and arises as synchrotron radiation from electrons accelerated within the shock to energies as high as a few hundred TeV. As the physical processes responsible for accelerating relativistic particles to these high energies do not

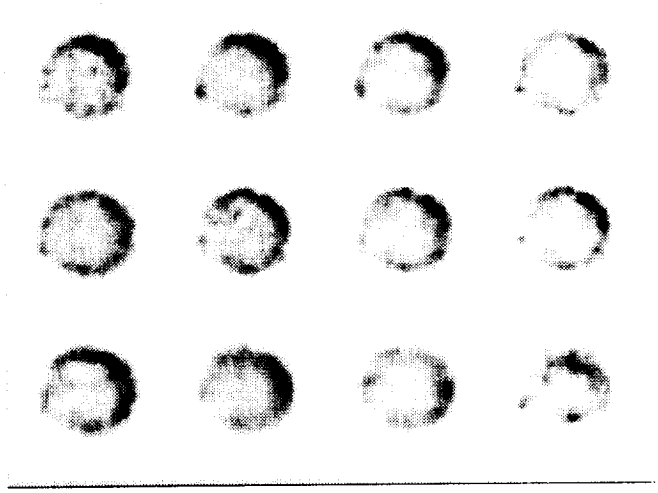


Figure 4.12: Narrow spectral band X-ray maps of Tycho from ASCA. The lack of significant differences suggests well mixed ejecta and uniform dynamics (heating and expansion. Figure from Hwang and Gotthelf (1996).

differentiate between electrons and ions, they infer that the nonthermal X-ray spectrum in the rims of SN1006 is the first strong evidence for the acceleration of cosmic rays by supernova shocks. The ASCA data showed that everywhere else in the remnant, the X-ray emission has a thermal spectrum, with strong line emission typical of other supernova remnants. The strengths of the lines from the intermediate α-burning elements, magnesium, silicon, and sulfur, are significantly enhanced, requiring abundances well in excess of solar. This is the first X-ray evidence that SN1006 was a Type Ia supernova, as inferred from its historical light curve (Stephenson et al., 1977).

4.10.5 W49B

W49B has one of the more spectacular X-ray spectra among supernova remnants, with intense lines from silicon through iron. Narrow band images constructed using ASCA rays by supernova shocks. The data show that while the Si and S line emission arise in a shell, the Fe emission is centrally concentrated (Fujimoto et al., 1995). This is clear evidence for metal stratification, similar to that predicted by the Hamilton et al. (1986) model of Tycho, and thus indirect evidence for a Type Ia progenitor. The X-ray emission suffers significant line-of-sight absorption below $\sim$1.3 keV due to high column density, however, rendering us unable to observe the strength of the O and Ne lines, and the corresponding high optical extinction makes impossible a search for optical line emission from O-rich ejecta. Thus the nature of the W49B progenitor remains undetermined.

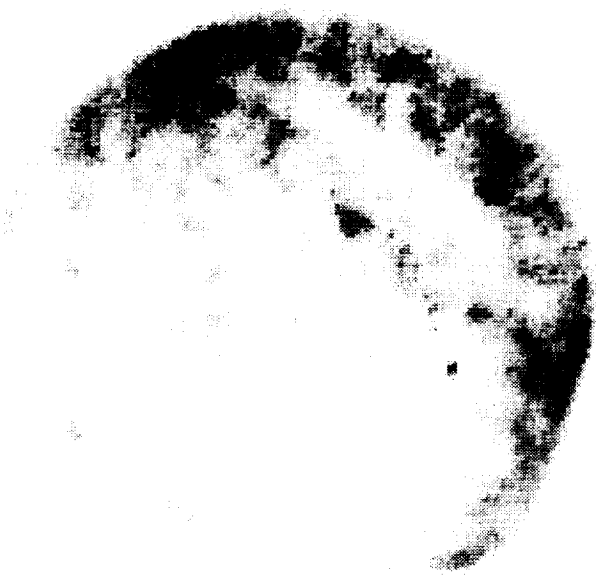


Figure 4.13: ROSAT PSPC image of SN1006 (Willingale et al., 1996). Brightness of image represents X-ray surface brightness; color represents spectral "hardness," an indication of effective temperature. The two bright rims are substantially harder than the remainder of the remnant.

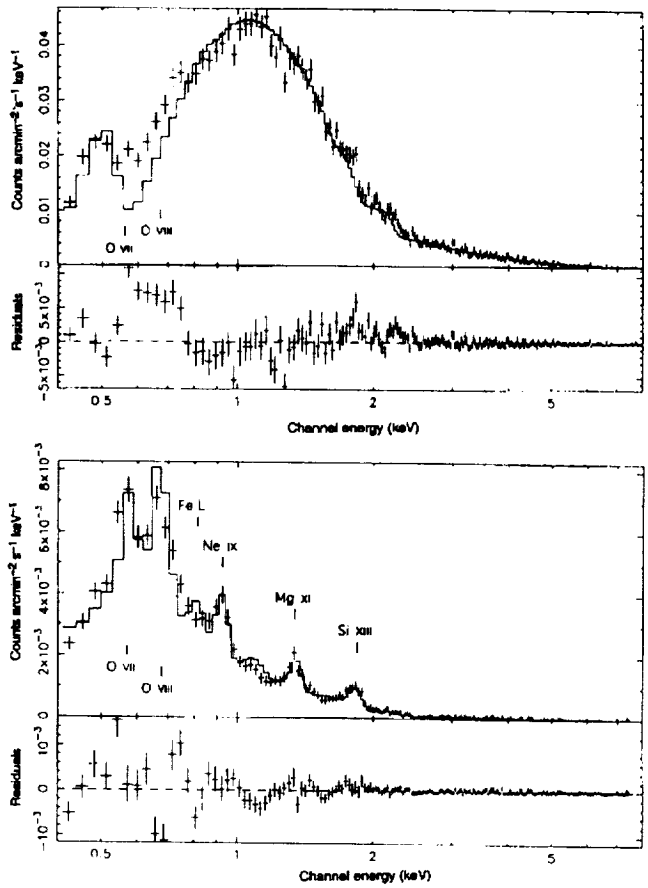


Figure 4.14: ASCA spectra of rim (top) and interior (bottom) of SN1006. The interior is dominated by line emission as expected from shock-heated gas. The featureless rim spectrum is indicative of a nonthermal emission process, most likely synchrotron emission from very high energy electrons. Figure from Koyama et al. (1996)

4.10.6 Cygnus Loop

Most of the SNR discussed above are relatively young, with their X-ray emission dominated by shock-heated ejecta. We can also learn a great deal studying the X-ray emission from more evolved remnants. In these remnants, the mass of swept up interstellar material far exceeds that of the ejecta, and so the shock heated ISM material dominates the emission in all bands. As these SNR are older, their shocks have also decelerated significantly, so that the temperature is typically a few million degrees, as opposed the the tens of millions of degrees in the younger remnants. Thus the X-ray spectrum is dominated by line emission at lower energies. In particular, the lines associated with helium-like and hydrogen-like Fe around 7 keV are absent. Instead, copious emission from the Fe L band is produced, along with lines from the He-like and H-like ions of C, O, Ne and Mg. In these SNR it is possible to use X-ray images to learn about the structure of the ISM by virtue of the enhanced X-ray emission produced when a SNR shock front encounters a higher density region in the ISM.

The many X-ray observations of the ~20,000 y old Cygnus Loop, one of the nearest (~700 pc) and best studied SNR, provide an example about what we can learn about evolved SNR and their interaction with the ISM. High quality spectral observations using the *Einstein* Crystal Spectrometer and ASCA show that the X-ray emitting plasma, even in this old SNR, has not attained ionization equilibrium (Vedder et al., 1986; Miyata et al., 1994). The reason for this is the X-ray emission is dominated by recently shocked gas, not by gas that has had sufficient time to equilibrate. ASCA observations show that the X-ray emitting gas is densest along the rim, and that the temperature increases toward the interior, as would be expected from Sedov-Taylor dynamics (Miyata et al., 1994). Imaging studies using ROSAT show that along the rim, interactions of the shock front with interstellar clouds lead to enhanced X-ray emission as the shock progagates into the cloud, as well as from regions behind clouds where reflected shocks propagate (Graham et al., 1995; Levenson et al., 1996). Thus the Cygnus Loop, through its X-ray emission has become a laboratory for the study of strong shocks.

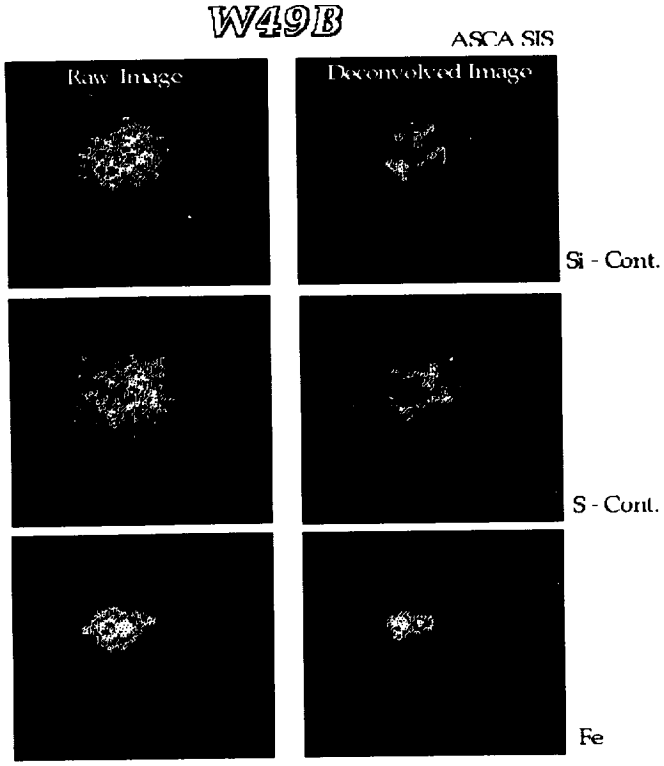


Figure 4.15: ASCA images of W49B in the light of the Si (1.86 keV), S (2.45 keV) and Fe (6.7) keV lines. Note that the Si and S are distributed in a shell, and the Fe is centrally filled. This provides direct evidence for stratification of SN ejecta. (from Fugimoto et al., 1995)

4.11 CONCLUSION

The discussion above was meant to be only an overview of the role X-ray observations play in helping us to understand these important cosmic events. As improved instrumentation becomes available, it will be possible to examine in greater detail the interactions between supernova shocks and the circumstellar and interstellar medium into which they propagate, learn more about explosive nucleosynthesis from more accurate measurements of abundant metals and the first X-ray detections of lines from less abundant metals, and shed light on other fundamental phenomena like cosmic ray acceleration.

4.12 REFERENCES

Aschenbach, B. 1993, in "UV and X-Ray Spectroscopy of Laboratory and Astrophysical Plasmas" eds. E. Silver, and S. Kahn, (New York: Springer) p. 434.

Bregman, J.N., and Pildis, R. 1992, ApJ 398, L107.

Canizares, C.L., Kriss, G.A., and Feigelson, E.M. 1982, ApJ 253, L17.

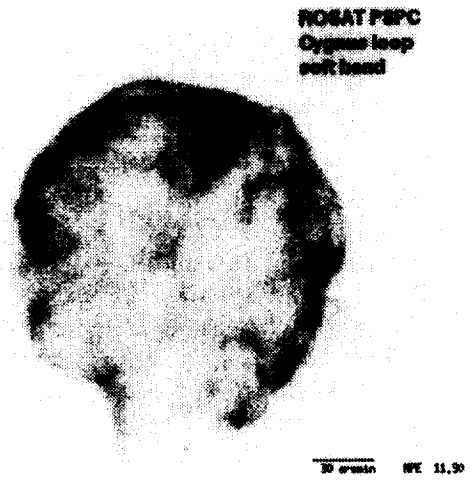


Figure 4.16: ROSAT PSPC image of the Cygnus Loop (Aschenbach, 1993).

Figure 4.17: ROSAT HRI image of a portion of the western rim of the Cygnus Loop (Levenson et al., 1993).

Chevalier, R.A., and Fransson, C. 1994, ApJ 420, 268.

Chevalier, R.A., and Klein, R.I. 1979, ApJ 234, 597.

Chugai, N.N. 1993, ApJ 414, L101.

Dotani, T., et al., 1987, Nature 330, 230.

Fabian, A.C., and Terlevich, R. 1996, MNRAS 280, L5.

Falk, S.W. 1978, ApJ 226, L133.

Fujimoto, R., et al., 1995, PASJ 47, L31.

Hughes, J.P., and Helfand, D.J. 1985, ApJ 291, 544.

A and A, 151, 52.

Gorenstein, P., Hughes, J.P., and Tucker, W.H. 1994,

ApJ.
Graham, J.R., Levenson, N.A., Hester, J.J., Raymond, J.C., and Petre, R. 1995, ApJ 444, 787.
Hamilton, A.J.S., Sarazin, C.L., and Chevalier, R.A. 1983, ApJS 51, 115.
Hamilton, A.J.S., Sarazin, C.L., and Szymkowiak, A.E. 1986, ApJ 300, 713.
Hayashi, I., Koyama, K., Ozaki, M., Miyata, E., Tsunemi, H., Hughes, J.P., and Petre, R. 1994, PASJ 46, L121.
Holt, S.S., Gotthelf, E.V., Tsunemi, H., and Negoro, H. 1994, PASJ 46, L15.1
Houck, J., and Bregman, J. 1996, in preparation.
Hughes, J.P., and Helfand, D.J. 1985, ApJ 291, 544.
Hughes, J.P. 1994, in "The Soft X-Ray Cosmos," eds. E.M. Schlegel, and R. Petre, (New York: AIP), p. 144.
Hughes, J.P., et al., 1995, ApJ 444, L81.
Hwang, U., and Gotthelf, E.V. 1996, ApJ, in press.
Kohmura Y., et al., 1994, PASJ 46, L157.
Koyama, K., Petre, R., Gotthelf, E.V., Hwang, U., Matsuura, M., Ozaki, M., and Holt, S.S. 1995, Nature 378, 255.
Levenson, N.A., Graham, J.R., Hester, J.J., and Petre, R. 1996, ApJ 468, 323.
Markert, T.H., Canizares, C.R., Clark, G.W., and Winkler, P.F. 1983, ApJ 268, 134.
Miyata, E., Tsunemi, H., Pisarski, R., and Kissel, S.E. 1994, PASJ 46, L101.
Nomoto, K., Thielemann, F.-K., and Yokoi, K. 1986, ApJ 286, 644.
Petre, R., et al., 1993, in "UV and X-Ray Spectroscopy of Laboratory and Astrophysical Plasmas" eds. E. Silver, and S. Kahn, (Cambridge: Cambridge University Press) p. 424.
Petre, R., Okada, K., Mihara, T., Makishima, K., and Colbert, E.J.M. 1994, PASJ 46, L115.
Ryder, S.L., et al., 1993, ApJ 416, 167.
Schlegel, E.M., and Petre, R. 1993, ApJ 412, L29.
Schlegel, E.M., Petre, R., and Colbert, E.J.M. 1996, ApJ 456, 187.
Sedov, L.I., 1959, in "Similarity and Dimensional Methods in Mechanics" (Ney York: Academic Press).
Stephenson, F.G., Clark, D., and Crawford, D. 1977, MNRAS 180, 567.
Sunyaev, R., et al., 1987, Nature 330, 227.
Taylor, G. 1950, Proc. Roy. Soc. Lon. A201, 159.
Thielemann, F.-K., Nomoto, K., and Hashimoto, M. 1994, in "Supernovae, Les Houches Session LIV" eds. S. Bludman, R. Mochkovitch, and J. Zinn-Justin, (North Holland: Elsevier) p. 629.
Vancura, O., Gorenstein, P., and Hughes, J.P. 1995, ApJ 441, 680.
Vedder,P.W., Conizares, C.R., Morkert, T.H., Prodhan, A.K., 1986, ApJ, 307, 269.
Willingale, R., West, R.G., Pye, J.P., and Stewart, G.C., 1996, MNRAS 278, 749.
Zimmermann, H.-U., et al., 1994, Nature 327, 621.

Chapter 5

PARTICLE ACCELERATION IN ASTROPHYSICAL SHOCKS

FRANK C. JONES
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD, 20771, USA

ABSTRACT

Collisionless shocks in astrophysical plasmas are believed to be a primary source of energetic particles in the solar system, the galaxy, and possibly in intergalactic space. We review the development of this idea from its earliest expressions to its latest manefestations. We will also discuss the various theoretical approaches to the process and discuss how well the current ideas and theories fit the observations.

5.1 INTRODUCTION

One of the most striking aspects of the tenuous plasmas common in astrophysics is that they often contain particle populations which are not in thermodynamic equilibrium. Nonthermal populations, often with power law distributions [1], are observed either directly or by inference (ie., from radiative properties) in most low density environments. The most exceptional example of this aspect is the galactic cosmic ray distribution which extends from MeV energies to above 10^{20} eV! The highest energy cosmic rays are, by far, the most energetic particles known in the universe, and energies this high have not occurred elsewhere since the temperature in the early universe dropped below $\sim 10^{24}$ K (about 10^{-28} s after the big bang). In a low density environment, particles "scatter collisionlessly" against magnetic turbulence rather than against other particles, and the scattering can be nearly elastic in the plasma frame. Particles are forced to equilibriate with the collective body of particles held together by the nearly frozen-in magnetic field rather than with each other, and individual particle energies can become extremely large. Nonthermal particle distributions can develop and persist for long times.

Tenuous astrophysical plasmas differ from laboratory plasmas in another important way; in most environments where energetic populations are observed, typical sound speeds are considerably less than easily obtainable bulk flow velocities, and shocks are expected to develop. In fact, *shocks are associated with most energetic particle populations seen in astrophysics.* The exceptions include pulsar magnetospheres, the aurorae, and sites where magnetic reconnection occurs (although shocks may be important in reconnection as well). The dominance of collective field-particle processes means that, unlike terrestrial shocks, the dissipation mechanism in astrophysical shocks is almost always collisionless. In the heliosphere, collisionless shocks are directly observable with spacecraft, and they have received considerable attention. In every case where direct observation has been made, shocks are seen to accelerate particles, often to power law distributions. Observations of heliospheric shocks, along with a great deal of theoretical and simulation work, also show that field-particle interactions control the shock dissipation and structure. Furthermore, initial results from large-scale plasma simulations suggest that not only is particle acceleration likely to accompany collisionless shocks, but also that the injection and acceleration of ambient particles may be even *essential* for dissipation in quasiparallel shocks and in high Mach number quasiperpendicular shocks. The plasma physics of shock dissipation and particle acceleration seem to be intimately related. [2]

[1] A distribution in which the density in phase space is proportional to the momentum or energy raised to a negative power, e.g., $dN/dp \propto p^{-\alpha}$ where α is the spectral index.

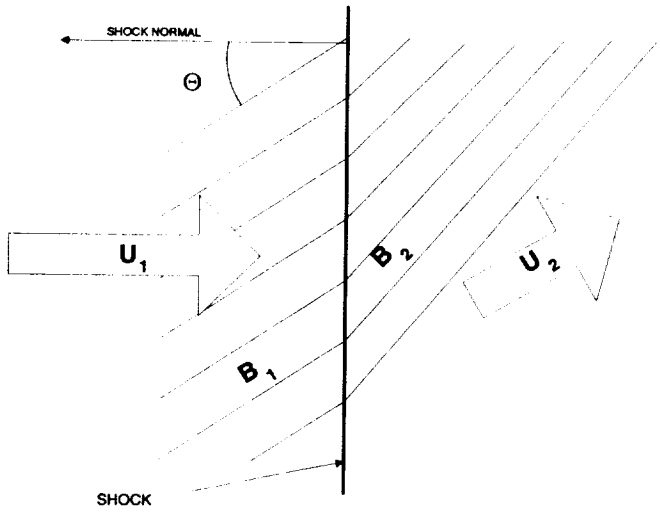


Figure 5.1: Shock geometry showing shock normal, velocity, and magnetic field vectors.

The promise of shock acceleration as a "universal" acceleration mechanism has prompted a great deal of work in the past decade or so, and we refer readers to several reviews for a full theoretical and observational background, including source references (Toptyghin, 1980; Axford, 1981; Drury, 1983; Völk, 1984; Forman and Webb, 1985).

5.2 HISTORY

5.2.1 Early Ideas

Darwin (1949)

Perhaps the earliest suggestion that shock waves were responsible for accelerating cosmic rays came from Sir Charles Darwin (1949). Starting with a discussion of the pitting of marine propellers by cavitation in the water, Darwin noted that hydrodynamics could lead to regimes where a few particles could be accelerated to extreme energy. He then applied this idea to turbulence in the solar atmosphere and suggested that cosmic rays might be accelerated by this process. He pointed out that the solar surface manifested such phenomena as flares and prominences and that "from some of them will spurt out spray, which occasionally will have very high velocities indeed." Darwin suggested that even if this process could not produce the cosmic rays directly, it could act as the

[2] The term quasiparallel means shocks with normals within approximately 45° of the magnetic field direction; quasiperpendicular means shocks with normals greater than 45° to the magnetic field direction; oblique refers to any shock which is not strictly parallel or perpendicular. See figure 5.1.

injector to Fermi's (1949, 1954) statistical acceleration mechanism.

This notion of a hydrodynamic wave steepening until it breaks and spews out a froth of energetic particles clearly prefigures the ideas set out some eleven years later by Colgate and Johnson (1960). It should be noted that in this paper, Darwin pointed out the phenomenon of "equipartition of velocity rather than energy" that is characteristic of such shock processes. This concept is, today, an important ingredient of current shock acceleration ideas as it leads to a preferred injection of heavy nuclei into the accelerator.

Parker (1958)

Parker (1958) noted that the turbulent gas motions of the galactic disk and halo were approximately Mach one when the pressure of the cosmic rays was included in computing the speed of sound. The Mach number is the ratio of the fluid flow speed to the local sound speed. He proposed that this was no accident, rather it was an example of a self-regulating system: if the cosmic ray energy density were too low, the motions would steepen into shock waves which would accelerate more cosmic rays. On the other hand, when the cosmic ray energy density is high enough to make the motions Mach one, the shocks fade and there is no more acceleration. This self-regulation was named the "Mach One Effect."

In Parker's work, however, the shocks were considered to be the moving irregularities that produced second-order Fermi acceleration. Later work has changed that picture somewhat, as we shall see.

Colgate and Johnson (1960)

Colgate and Johnson proposed that cosmic rays are the last fraction of the atmosphere of a star that has undergone a supernova explosion. The shock traveling upward through an ever decreasing density would achieve relativistic velocities and impart high energies to the last particles that were shocked. In their model, the energy spectrum depended on the density versus radius function of the exploding star. This result, derived from the equations of radiative-hydrodynamic stability, produced a differential spectrum that was an inverse power law. The spectral index that they derived was equal to -3, somewhat steeper than observed but, considering the complexity of the derivation, it was, perhaps, closer to observation than one might have imagined.

Hoyle (1960)

Hoyle considered the case in which a considerable portion of the momentum and energy of a gas through which a shock is passing is carried by relativistic particles. He pointed out that in this situation the ratio of specific heats, γ, can approach a value of 4/3 and thus the compression ratio can approach 7 rather than 5/3 and 4, respectively, for a nonrelativistic gas. The increased compression can result in a large increase in particle energy. Hoyle gives an example of the cosmic-ray gas in which a 23 fold increase in particle energy results for a shock speed of $10^8\ cms^{-1}$. This appears to be essentially the "cosmic ray dominated" shock of the two-fluid approach to shock acceleration discussed in Section 5.5.

Schatzman (1963)

In an analytical calculation, Schatzman considered particles traversing a shock and was able to deduce that a power law energy spectrum would result. He found that on one traversal of the gyro-orbit through the shock front the magnetic moment was conserved approximately and when scattering was added, he obtained a power law spectrum, but without the simple relationship between the exponent and shock parameters obtained in later work.

Fisk (1971)

Fisk was first to note that a shock propagating through a diffusive medium would accelerate particles by scattering the particles with converging scattering centers. Since he applied this approach to a particular class of events observed at travelling interplanetary shocks, the generality of the method was not noticed. Also, Fisk did not obtain the result, now well known, that the spectrum of the accelerated particles depends on the shock's compression ratio, r, as $N \propto p^{-\sigma}$ where N is the particle density in space and in scalar momentum, p, and $\sigma = (r+2)/(r-1)$ is the spectral index.

5.2.2 "Modern Era" of Shock Acceleration

The important breakthrough in understanding shock acceleration occurred in 1977-78 when a series of independent papers showed how a power law momentum spectrum of accelerated particles resulted from very general properties of a plasma shock traveling through a medium in which energetic particles were diffusing (Axford, Leer, and Skadron, 1977; Krymsky, 1977; Bell, 1978a,b; and Blandford and Ostriker, 1978). These papers showed that the steady-state power law spectrum that resulted was

- independent of the injection spectrum, provided the injection spectrum was steeper than the resultant spectrum [to be shown],
- independent of the details of the scattering process as long as the distribution function of the accelerated particles could be considered isotropic to first order, and
- independent of the shock geometry if the diffusion scale was smaller than any curvature of the shock front but larger than the shock thickness.

The wide appeal of this process is due to the fact that, with the simplest assumptions, the particle spectrum depends only on the compression ratio of the shock. Furthermore, most astrophysical shocks, since they are strong, have compression ratios that are constrained to a rather narrow range of values near $r = 4$ (assuming $\gamma = 5/3$). For shocks with Mach numbers greater than about 4, $3.4 < r < 4$ and $2 < \sigma < 2.3$ as we see in Figure 5.2.

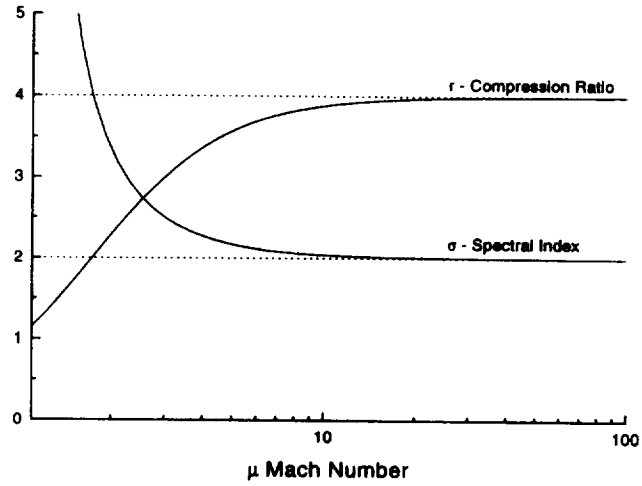


Figure 5.2: Compression ratio, r, and spectral index, σ, vs. Mach no.

5.3 DIFFUSION THEORY APPROACH

Axford, Leer, and Skadron (1977), Krymsky (1977), and Blandford and Ostriker (1978) applied the diffusion-convection equation (a differential equation that describes the diffusion of particles through a flowing medium) to show that a shock, propagating through a region in which energetic particles are diffusing, would produce a superthermal population of particles with a power law momentum distribution. The derivation is quite straightforward in the one-dimensional case and involves only the simplest notions. If we picture the shock as an infinite, plane discontinuity in a flowing plasma, the plasma flows in from $x = -\infty$ and out to $x = +\infty$ with a discontinuous transition in flow speed from a supersonic upstream speed, u_1, to a subsonic downstream speed, u_2, at $x = 0$. If we assume that the distribution function in space and scalar momentum of the accelerated particles, $f(x,p)$, is isotropic to first order [i.e., $f(x,p)$ is the same in all reference frames to first order in u/v, where u is the plasma flow velocity and v and p are the individual particle velocity and momentum measured in the local plasma frame], then the steady-state Boltzmann equation describing the transport of particles with $v \gg u$ in space and momentum can be written in the form of the following diffusion-convection equation

$$\frac{\partial}{\partial x}\left[uf(x,p) - \kappa\frac{\partial f(x,p)}{\partial x}\right] = \frac{1}{3}\left(\frac{\partial u}{\partial x}\right)\frac{\partial}{\partial p}[pf(x,p)] \tag{5.1}$$

where $\kappa = \kappa_{\parallel}\cos^2\theta_{Bn} + \kappa_{\perp}\sin^2\theta_{Bn}$ is the diffusion coefficient in the direction normal to the shock, where θ_{Bn} is the angle between the shock normal and the mean magnetic field, and where $\kappa_{\parallel}$ and $\kappa_{\perp}$ are the diffusion coefficients parallel and perpendicular to the magnetic field, respectively.

As long as scattering is strong enough to insure that the assumptions [3] of the diffusion-convection equation are valid, this equation holds for all θ_{Bn} and contains shock-drift acceleration. Forman and Webb (1985) give a good discussion of how one derives the diffusion coefficient, κ, from a description of turbulent magnetic fields employing a perturbation technique known as quasilinear theory.

Integrating from $x = -\infty$ to $x = +\infty$ and employing the boundary conditions $f(x = -\infty, p) = f_1(p)$ $f(x \geq 0, p) = f_2(p)$ and $\partial u/\partial x = (u_2 - u_1)\delta(x)$ we obtain

$$u_2 f_2(p) - u_1 f_1(p) = \frac{1}{3}(u_2 - u_1)\frac{\partial}{\partial p}[pf_2(p)] \tag{5.2}$$

Using the definition of the compression ratio $r = u_1/u_2 > 1$, we may express this differential equation in p for the downstream distribution function in the form

$$p\frac{\partial f_2(p)}{\partial p} + \left(\frac{r+2}{r-1}\right)f_2(p) = \frac{3r}{r-1}f_1(p) \tag{5.3}$$

the solution to this equation is

$$f_2(p) = \left(\frac{3r}{r-1}\right)p^{-\sigma}\int_{p_0}^{p} dp' p'^{\sigma-1} f_1(p') + Bp^{-\sigma} \tag{5.4}$$

where $\sigma = (r+2)/(r-1)$ is the spectral index, B is an arbitrary constant of integration that multiplies the homogeneous term, the distribution $f_1(p)$ is the far upstream spectrum of ambient particles which convects into the shock and is accelerated, and p_0 is large enough so that the assumption $v \gg u$ holds.

If, in Equation (5.4), $f_1(p)$ is taken to be a δ-function distribution at momentum $p(E_i)$, where E is kinetic energy, normalized at an upstream number density n_1, we may write the downstream spectrum as a differential flux in particles/(cm^2-s-$steradian$-E) given by

$$\frac{dJ}{dE} = \frac{1}{4\pi}\frac{n_1 c}{m_0 c^2}\frac{3r}{(r-1)}\frac{1}{\sqrt{\varepsilon_i(\varepsilon_i+2)}}\left[\frac{\varepsilon(\varepsilon+2)}{\varepsilon_i(\varepsilon_i+2)}\right]^{-\sigma/2} \tag{5.5}$$

[3] The particle distribution is very nearly isotropic and the flow speeds are enough slower than that of the particles that this approximate isotropy holds in all fluid rest frames.

where, $\varepsilon = E/(m_0c^2)$, and m_0 is the rest mass of the particles.

This equation is properly normalized and fully relativistic. As a power law in kinetic energy

$$\left(\frac{dJ}{dE}\right)_{\substack{NR\\UR}} \propto E^{-\sigma_{\substack{NR\\UR}}} \tag{5.6}$$

the limiting nonrelativistic and ultrarelativistic spectral indexes for the flux are $\sigma_{NR} = \sigma/2$ and $\sigma_{UR} = \sigma$, respectively. When measured versus energy, the power law differential flux steepens above $E \sim m_0c^2$, and the spectral index doubles.

No other source of particles is included in the above equations other than $f_1(p)$. In actual shocks with no particle creation (such as electron-positron creation), the homogeneous term should be zero. The downstream spectrum, $f_2(p)$, is produced solely from the ambient upstream spectrum, $f_1(p)$. It has been speculated by Drury (1983) that the homogeneous term might represent the injection into the acceleration process of particles from the thermal plasma making up the shock. However, this does not seem possible because the homogeneous term is a pure power law with no cutoff, at low or high energy and hence cannot be normalized. If one were to impose a cutoff at some convenient energy, the function would no longer be a homogeneous solution. Furthermore, results from Monte Carlo calculations, in which the diffusion of charged particles in a fluid with a shock is simulated on a computer, strongly indicate that the injection of thermal particles should be treated exactly like the injection of any other seed particles. This is shown in Figure 5.3, which shows spectra produced by a Monte Carlo simulation. The smooth curves are plots of Equation (5.5), and they are shown to match quite well, in shape and normalization, the Monte Carlo results at energies high enough so that $v \gg u$ holds. Below about 1 MeV this condition is not well satisfied, and the slopes of the Monte Carlo results are a bit too steep. In the Monte Carlo model, injection comes solely from upstream thermal particles convected into the shock [i.e., $f_1(p)$] when some fraction of the shock-heated plasma scatters back across the shock and is accelerated. It can be seen that the diffusion theory and Monte Carlo results agree surprisingly well even when the injected particles (the seed population, f_1) have thermal energies that are considerably less than that of their flow energy. (The flow speed of 500 $km\ s^{-1}$ corresponds to a kinetic energy of 1.3 keV.) It is clear from Figure 5.3 that thermal ions that make up the shocked plasma are quite capable of serving as the seed population and that there is no need for injection from the homogeneous term.

In the proceeding discussion we have assumed that the injected particles had a single (low) energy and were accelerated to a power law spectrum. We can see, moreover, that the same result prevails even if the injected

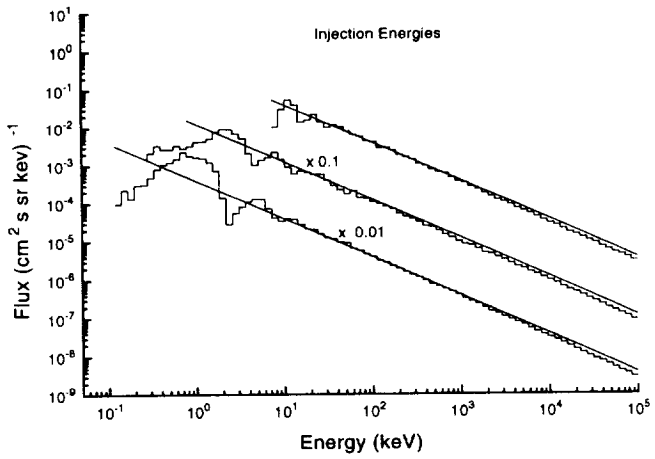


Figure 5.3: Particle flux versus energy for nonrelativistic shock velocities. The upper six curves were calculated with $u_1 = 500\ km\ s^{-1}$, $r = 4$, and $\alpha = 1$. The lower two curves were calculated with $u_1 = 100\ km\ s^{-1}$. The smooth curves are the test particle predictions, while the histograms are the Monte Carlo results. All spectra here and elsewhere are calculated in the reference frame of the shock, at 1a downstream position. Particle fluxes are normalized to 1 incoming particle/(cm^2-s). Injection energies are shown by arrows. Curves with numerical factors have been displaced by those factors for clarity. (Monte Carlo results from M. Baring private communication.)

particles already have a power law distribution, provided that the initial spectrum is steeper than the spectrum that we have previously calculated. Assume that $f_1(p)$ is a power law of the form $f_1(p) \propto p^{-\sigma+\delta}$ for $p > p_0$ and zero for $p < p_0$, where δ may be any value, then the spectrum given by Equation (5.4) is proportional to

$$p^{-\sigma}\left(\frac{p^\delta - p_0^\delta}{\delta}\right) \tag{5.7}$$

If $f_1(p)$ is steeper than $p^{-\sigma}$ (i.e., $\delta < 0$), at momenta well above the lowest injection momentum ($p \gg p_0$), the second term in the parentheses will dominate the first term and the shock will produce an accelerated spectrum proportional to $p^{-\sigma}$. If, on the other hand, $f_1(p)$ is flatter than $p^{-\sigma}$ (i.e., $\delta > 0$), the first term will dominate and the accelerated spectrum will retain the flatter slope of the injected spectrum with the injected particles shifted up in energy. Since most ambient distributions are expected to be thermal or steeper than $p^{-\sigma}$, strong shocks naturally produce power law distributions with spectral indexes near 2.

The most remarkable property of this result is that the power law index depends only on the compression ratio, all details of the scattering process and θ_{Bn} (contained in κ) drop out as long as the assumptions used to

derive the equations, namely that the accelerated particles have speeds sufficiently large that they can appear isotropic in both the upstream and downstream fluid rest frames (i.e., $v \gg u$), continue to hold. While it is clear that these assumptions must breakdown at some point, many plasma shocks in astrophysical settings seem to approach the ideal case over a wide dynamic range. The simplicity and lack of free parameters, coupled with the fact that shock acceleration has been observed taking place at the Earth's bow shock and interplanetary shocks, are the major reasons why shock acceleration has received such a great deal of attention.

5.4 INDIVIDUAL PARTICLE APPROACH OF BELL

A different approach was taken by Bell (1978a,b) who considered the average energy gain experienced by a particle that traversed a discontinuity in the flow of a fluid to which the particle was bound. From the general transport Equation (5.1) it is straightforward to see that the rate of change of momentum of a particle is given by

$$\dot{p} = -\frac{1}{3}p\left(\frac{\partial u}{\partial x}\right) \tag{5.8}$$

Integrating this equation through the shock from upstream (region 1) to downstream (region 2) we obtain

$$\int_1^2 \dot{p}dt = -\int_1^2 \frac{1}{3}p\left(\frac{\partial u}{\partial x}\right)\left(\frac{dx}{v_x}\right) = \frac{1}{3}\left(\frac{p}{v\mu}\right)(u_1 - u_2) \tag{5.9}$$

where μ is the cosine of the angle that the particle's velocity makes with the shock normal.

The rate at which particles cross the shock is proportional to μ (because particles approaching the shock straight on are more likely, in a given increment of time, to hit and cross it than those approaching from an oblique angle). Therefore, the average momentum change of a particle crossing the shock is given by multiplying the right hand side of Equation (5.9) by μ and integrating over the 2π solid angle (flux averaging). This proceedure gives

$$\langle \delta p \rangle = \frac{2}{3}\left(\frac{p}{v}\right)(u_1 - u_2) \tag{5.10}$$

For particles traversing the shock in the opposite direction, u_1 and u_2 are interchanged but the sign of v_x is also reversed so the equations remain unchanged and a particles energy is *incremented* on each traversal regardless of the direction of the traversal. It should be noted that this same expression is obtained by arguments involving frame transformations (Drury, 1983).

We now can see that if a particle with initial momentum p_0 has traversed the shock N times its average momentum will be

$$\langle p(N) \rangle = \prod_{i=1}^{N}\left[1 + \frac{2}{3}(u_1 - u_2)/v_i\right] p_0 \tag{5.11}$$

In order to obtain a distribution of particles, however, we must include the probability that a particular particle will cross the shock N times. The probability that a particle that has traversed the shock from upstream to downstream will return to the shock can be calculated by means of a simple argument first given by Bell (1978a) and generalized to relativistic shocks by Peacock (1981).

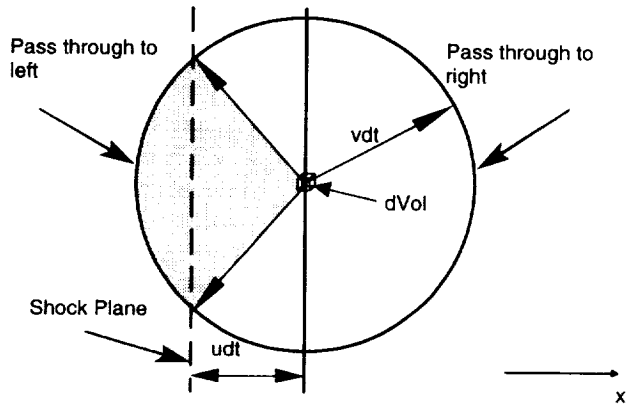


Figure 5.4: Return Probability from velocity space. The circle shown is the projection on the $x - z$ plane of a sphere with radius vdt, a portion of which protrudes through the shock plane.

From Figure 5.4 we see that the probability that a particle that crosses the shock from left to right will return to the shock and cross it from right to left is given simply by the ratio of the respective fluxes. If the particles are isotropic in the downstream fluid frame, then the probability of return is given by

$$\mathcal{P}(\text{return}) = \frac{\frac{1}{2}(u_2 - v)^2}{\frac{1}{2}(u_2 + v)^2} = \left(\frac{1 - u_2/v}{1 + u_2/v}\right)^2 \tag{5.12}$$

The probability that a particle that has traversed the shock from upstream to downstream will return to the shock *at least* $N/2$ times and thus cross the shock *at least* N times is given by

$$\mathcal{P}(N) = \prod_{i=1}^{N/2}\left(\frac{1 - u_2/v_i}{1 + u_2/v_i}\right)^2 = \left[\prod_{i=1}^{N/2}\left(\frac{1 - u_2/v_i}{1 + u_2/v_i}\right)\right]^2 \tag{5.13}$$

Taking the logarithm of Equation (5.11) we obtain

$$\ln\left[\frac{\langle p \rangle (N)}{p_0}\right] = \frac{4}{3}(u_1 - u_2)\sum_{i=1}^{N/2}\frac{1}{v_i} \tag{5.14}$$

and taking the logarithm of Equation (5.13) we obtain

$$\ln[\mathcal{P}(N)] \approx -4u_2 \sum_{i=1}^{N/2} \frac{1}{v_i} \tag{5.15}$$

Combining these two equations gives

$$\ln[\mathcal{P}(p)] = -\frac{3u_2}{u_1 - u_2} \ln\left(\frac{p}{p_0}\right) \tag{5.16}$$

or

$$\mathcal{P}(p) = \left(\frac{p}{p_0}\right)^{-3u_2/(u_1-u_2)} \tag{5.17}$$

Equation (5.11) represents the probability that a particle will cross and recross the shock enough times to achieve a momentum of p or higher. If the upstream number density is N_0 particles per unit volume, the resulting differential spectrum is given by

$$\begin{aligned} f(p) &= -n_1 \left(\frac{u_1}{u_2}\right)\left(\frac{\partial \mathcal{P}}{\partial p}\right) \\ &= \left(\frac{n_1}{p_0}\right)\left(\frac{3u_1}{u_1 - u_2}\right)\left(\frac{p}{p_0}\right)^{-(u_1+2u_2)/(u_1-u_2)} \\ &= \left(\frac{n_1}{p_0}\right)\left(\frac{3r}{r-1}\right)\left(\frac{p}{p_0}\right)^{-\sigma} \end{aligned} \tag{5.18}$$

where the extra factor of $r = u_1/u_2$ in the first equality accounts for the simple compression of the upstream density by the shock.

If we set $f_1(p')$ equal to $n_1\delta(p' - p_0)$ in Equation (5.4) we obtain the above equation showing that the two approaches are equivalent.

5.4.1 Problems and Limitations

Four aspects of this approach to shock acceleration limit its validity

- it assumes isotropy and $\frac{u}{v} \ll 1$,
- it assumes shock thickness $\ll \frac{\kappa}{u}$,
- it assumes shock curvature radius $\gg \frac{\kappa}{u}$,
- it assumes presence of turbulence.

It is often stated that these four conditions are necessary for shock acceleration to occur at all. This is a myth! These limitations are only a problem for the mathematics of the simple theory. Anisotropy of the particles in the acceleration process has been treated by Kirk and Schneider (1987a,b, 1989). The finite thickness of the shock, which affects the spectral shape, has been treated by Drury, Axford, and Summers (1982). The effects of the finite size of the shock, which imposes an upper limit to the energy that may be achieved, has been treated by Eichler (1981) and Lee (1982,1983). It can be seen that there are means of carrying the simple theory into more complicated regimes. The problem is that the resulting equations get complicated very quickly, and it is never completely clear just how good the approximation is. In Section 5.5 we will see that certain nonlinear effects lead to quite different methods for studying shock acceleration. Bell (1978 a,b) employed a self-consistent treatment of the generation of waves upstream from the shock. These waves were generated by the streaming of the particles reflected from the shock front and moving in an upstream direction. Lee (1982, 1983) developed a similar theory of the Earth's bow shock and interplanetary shocks traveling in the solar wind. Both authors were able to show that particles initially reflected from the shock front would generate, via a plasma streaming instability, enough wave turbulence to scatter the particles. This scattering is required to produce the particle diffusion back and forth across the shock front and allow diffusive shock acceleration to occur.

5.5 HYDRODYNAMIC EFFECTS OF THE ACCELERATED PARTICLES: TWO FLUID THEORIES

If a plasma shock is at all efficient in converting some of its flow energy into energetic particles, the pressure exerted by these particles must play some role in defining the structure of the shock itself. We may express the pressure of the energetic particles, P_c, as due to a "second" fluid that carries pressure and energy but no mass. Including the effects of this fluid with the basic plasma that supports the shock, yields the "Two Fluid" model of shock acceleration (Drury and Völk, 1981). The energetic particle pressure is given by

$$P_c = \frac{4\pi}{3} \int_0^\infty p^3 v f(p) dp \tag{5.19}$$

where $f(p)$ is the distribution function of the energetic particles. If we multiply Equation (5.1) by $p^3 v$ and integrate over momentum, we obtain

$$\frac{\partial P_c}{\partial t} + \frac{\partial}{\partial x}(uP_c) + (\gamma_c - 1)P_c \frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\overline{\kappa}\frac{\partial P_c}{\partial x} \tag{5.20}$$

where, in order to make the equation look more hydrodynamic, we have converted the factor $1/3$ on the right hand side of Equation (5.1) to the expression $\gamma_c - 1$ where γ_c $(= 4/3)$ is the ratio of specific heats for a gas of relativistic particles.

We may express the continuity of the flux of mass, momentum, and energy through the shock with the following differential equations

$$\frac{\partial(\rho u)}{\partial x} = 0 \quad \text{mass flux} \tag{5.21}$$

$$\frac{\partial}{\partial x}\left(\rho u^2 + P_g + P_c\right) = 0 \quad \text{momentum flux} \qquad (5.22)$$

$$\frac{\partial}{\partial x}\left(\frac{1}{2}\rho u^3 + \frac{\gamma_g}{\gamma_g - 1} u P_g + \frac{\gamma_c}{\gamma_c - 1} u P_c - \frac{\overline{\kappa}}{\gamma_c - 1}\frac{\partial P_c}{\partial x}\right) = 0 \quad \text{energy flux}$$

where P_g is the pressure of the "first" fluid or plasma and γ_g is the ratio of specific heats of the thermal particle gas.

Equations of this type are usually used to determine the jump conditions af various fluid properties across a fluid shock transition, the Rankine-Hugoniot conditions. We have simply extended the usual set of equations to include the dynamic effects of the energetic particles. The difficulty with this set of equations is that it is not complete; there are three equations and four unknown quantities; ρ, u, P_g, and P_c. These equations do not supply enough information to integrate a set of starting values through the shock transition. In principle, one could employ Equation (5.20) as the fourth equation, but a simpler solution is effected by employing two supplemental assumptions.

The first assumption is that the energetic particles do not transfer any heat to the gas, they only push it around. We may, therefore, add to the above equations one that describes the continuity of the entropy flux of the gas

$$\frac{\partial(P_g u^{\gamma_g})}{\partial x} = 0 \qquad (5.23)$$

This equation may not be used to deduce continuity of the entropy flux across a shock transition in the thermal plasma. In the gas shock, dissipative processes will produce an increase in entropy and continuity will not be preserved across the discontinuity.

We must find some other consideration to reduce the number of independent variables. The additional assumption used in the two-fluid approach is that the presence of a gas shock has no "local energetic" effect on the energetic particle population, and, thus, its pressure and energy flux are continuous across the shock. This implies that, although the thermal background gas is undergoing a strongly dissipative transition in which flow energy is being converted to heat, the energetic particles do not participate in this highly localized event. (The conservation of entropy and the continuity of the energetic particles across the thermal gas shock can, in fact, be derived from the three conservation equations plus Equation (5.20) However, it is equally true that one may take entropy conservation and continuity on an equal footing with the conservation equations and derive Equation (5.20) from them.)

With this set of equations one may solve for the shock structure as modified by the pressure of the energetic particles. Dissipative effects in the thermal gas still can produce a discontinuity in the gas but there is also a long range slowing down of the gas upstream of the shock due to the pressure gradient of the energetic particles.

Actually solving the above equations can be a rather complex procedure requiring several piecewise processes using different equations in different regions. For a more complete look at such solutions one should refer to the review article by Drury (1983), where it is pointed out that three following basic types of solutions occur, depending on the input parameters

- Weak Shock Solution - as described above - regular shock in the gas but with upstream slowing,
- Cosmic Ray (energetic particle) Dominated Solution - no shock in the thermal gas,
- Efficient Shock Solution - no energetic particles upstream, but with such particles downstream - even with no thermal shock!

5.5.1 Advantages of Two Fluid Approach

The basic diffusion equation or test particle approach to shock acceleration is very difficult to extend into the nonlinear regime where the accelerated particles influence the shock structure. The Two Fluid approach handles this difficulty very nicely by turning the entire problem into a purely hydrodynamic one. One unusual solution, the cosmic ray dominated solution, is one that would never occur with the other approaches. While the appearance of this solution is somewhat startling, there appears to be no reason such cases could not exist under the right conditions.

5.5.2 Problems with the Two Fluid Approach

No number conservation law for the energetic particles

The lack of a conservation law for the number of energetic particles leads to the Efficient Shock Solution with these particles being "created" out of nothing. Since the energetic particle fluid has nothing but pressure or energy density, this property can be obtained easily from the low energy fluid or plasma. This technique, therefore, does not address the problem of energetic particle injection in real shocks. For this reason it is believed that this solution is probably not a physical one.

No spectral information

Since the particles are treated entirely as fluids, no information about the distribution function of either the plasma or the energetic particles can be obtained. Since

the particle spectrum is usually what is desired in astrophysics, this omission must be considered serious.

5.6 NUMERICAL SIMULATIONS

Although a discussion of this topic is beyond the scope of this lecture, many of the problems that one encounters with the techniques discussed so far may be overcome by modeling the shock acceleration process on a computer. Although, this technique does not come under the strict heading of the "theory" of shock acceleration, many interesting characteristics of the process have been revealed by these studies. It is now understood that plasma shocks can often act as their own "injectors," picking up particles from the background plasma and subsequently accelerating them to high energies. Also, the "universal" spectrum (a strict power law in momentum) is not produced whenever the pressure of the accelerated particles is strong enough to broaden the shock transition. A more detailed discussion of these and other results and an extensive reference list can be found in the review article by Jones and Ellison (1991).

5.7 REFERENCES

Axford, W.I., 1981, Proc. 17th Int. Cosmic Ray Conf., Paris, 12, 155.

Axford, Leer, and Skadron, 1977, Proc. 15th Int. Cosmic Ray Conf., Plovdiv, 11, 132.

Bell, A.R., 1978a, MNRAS, 182, 147.

Bell, A.R., 1978b, MNRAS, 182, 443.

Baring, M., 1994, private communication.

Blandford, R.D., and Ostriker, J.P., 1978, ApJ, 331, L29.

Colgate, S.A., and Johnson, M.H., 1960, Phys. Rev. Lett., 5, 235.

Darwin, C., 1949, Nature, 164, 1112.

Drury, L.O'C., 1983, Rep. Prog. Phys., 46, 973A.

Drury, L.O'C., Axford, W.I., and Summers, D., 1982, MNRAS, 223,353.

Drury, L.O'C., and Völk, H.J., 1981, ApJ, 248, 344.

Eichler, D., 1981, ApJ, 224, 711A.
ApJ, 360, 702.

Fisk, L., 1971, J. Geophys. Res., 76, 1662.

Fermi, E., 1949, Phys. Rev. Letters, 75, 1169.

Fermi, E., 1954, ApJ, 119, 1.

Forman, M.A., and Webb, G.M., 1985, in "Collisionless Shocks in the Heliosphere: A Tutorial Review" eds. R.G. Stone, and B.T. Tsuritani, AGU Monograph, Vol 34, AGU, Washington, DC, P. 91.

Hoyle, F., 1960, MNRAS, 120, 338.

Jones, F.C., and Ellison, D.C., 1991, Sp. Sci. Rev., 58, 259.

Kirk, J.G., and Schneider, P., 1987a, ApJ, 315, 425.

Kirk, J.G., and Schneider, P., 1987b, ApJ, 322, 256.

Kirk, J.G., and Schneider, P., 1989, å, 217, 344.

Krymsky, G.F., 1977, Dokl. Akad. Nauk SSSR, 243, 1306.

Lee, M.A., 1982, J. Geophys. Res., 87, 5063.

Lee, M.A., 1983, J. Geophys. Res., 88, 6109.

Parker, E.N., 1958, Phys. Rev., 109, 1328.

Peacock, J.A., 1981, MNRAS, 196, 135.

Schatzman, E., 1963, Ann. Astrophys., 26, 234.

Toptyghin, I.N., 1980, Space Sci. Rev., 26, 157.

Völk, 1984, in "High Energy Astrophysics" ed. Tran Than Van, Proc. 19th Recontre de Moriond, Editions Frontières, Gif-Sur-Yvette, p. 281.

Chapter 6

511 keV ANNIHILATION RADIATION FROM THE CENTER OF THE MILKY-WAY

MARVIN LEVENTHAL
Department of Astronomy
University of Maryland
College Park, Maryland 20742

ABSTRACT

Annihilation of electrons and positrons (anti-electrons) can produce 511 keV gamma-ray line radiation. Such radiation has been observed coming from our own galactic center. This chapter describes those observations and discusses the pros and cons of various models for the source of the radiation. A favored interpretation of the data set is in terms of a disk plus spheroid model. It is suggested that the 1E 1740.7-2942 source, a "microquasar" like system located near the galactic Center can account for most of the observations within a few degrees of the center of our Milky Way galaxy.

6.1 INTRODUCTION

For the past 20 years more than a dozen balloon and satellite observations have been reporting 511 keV electron/positron annihilation line radiation from the general direction of our own galactic Center (GC). The history of these observations and their implications have been reviewed by a number of workers (Tueller et al., 1993; Skibo et al., 1992; Von Ballmoos 1991; Lingenfelter and Ramaty, 1989). The data set has now grown large, complicated, and confusing. No general consensus yet exists as to what it all means. Quite the contrary, a certain amount of controversy surrounds these observations. The most fundamental questions remain unresolved such as: What is the origin of the positrons?; What is their spatial distribution on the sky?, Does a variable point source near the GC really exist?, etc. What we will attempt to do is first present a balanced overview of the observational situation emphasizing some recent measurements while pointing out the areas of controversy and then suggest a possible interpretation, which on the surface at least appears to tie together many of the conflicting results. This will naturally lead us to a "microquasar" near but not at the GC. In the end the reader must decide for himself what the truth is.

We will be discussing galactic "positron" annihilation but what is really meant by that most of the time is galactic "positronium" annihilation. In dilute astrophysical gases and plasmas when the temperature is below $\sim 10^6$ K the positronium annihilation mode usually dominates (Guessoum et al., 1991). (Positrons annihilating in earthly laboratories usually do so from the "free" state. An electron and a positron wave function will briefly overlap and the mass energy will reappear as two anti-parallel 511 keV gamma-ray line photons.) Positronium is the hydrogen like atom of an electron bound to a positron. Its annihilation modes are indicated in Figure 6.1. Depending on the spin alignments, one either gets the familiar two anti-parallel 511 keV photons or the 3-photon orthopositronium continuum. The orthopositronium continuum extends from 511 keV to 0 keV having its maximum value at 511 keV. In real detectors gamma-ray are counted one a time, their energy measured and a spectrum accumulated. Real detectors all have a finite energy resolution which determines the minimum width which a line like feature can have. A pure positronium spectrum folded through detectors of degrading energy resolution is shown in Figure 6.2[6]. The peak in the distribution eventually shifts to energies below 511 keV as more and more of the continuum is included in the line in an asymmetric fashion. This is purely an instrumental effect due to the finite detector resolution. Practically what this means is that observers are usually dealing with a low energy tail on the annihilation line. In the most general case one would expect some combination of positronium and

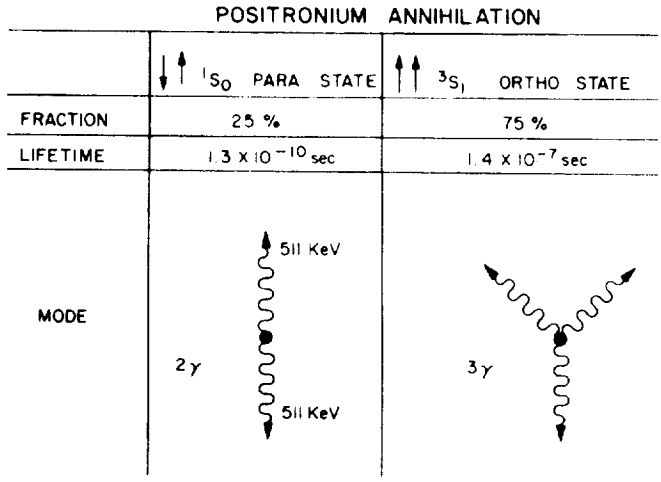


Figure 6.1: The annihilation modes of positronium.

"free" annihilation which always gives 511 keV photons. Detailed calculations and laboratory simulation experiments indicate that in HI regions (neutral atomic hydrogen) and in cool molecular clouds a positronium fraction f of $\sim$ 90% is expected (Guessoum et al., 1991; Brown and Leventhal, 1987). Here f is defined as the fraction of incident positrons annihilating from the bound positronium state i.e. (positronium/(positronium + free)).

6.2 OBSERVATIONS

Until quite recently all of the "telescopes" were of a rather primitive nature. Since mirrors don't work at these energies the "telescopes" were simply arrays of detectors and collimators. As a modern example the Gamma-Ray Imaging Spectrometer (GRIS) balloon "telescope" is shown in Figure 6.3. It is built around an array of 7 of the largest available high purity Ge detectors (Tueller et al., 1994). The detectors are operated as electrical diodes and cooled to cryogenic temperatures to reduce the leakage current. The gamma-ray energy generates a large number of electron-hole pairs causing a current to flow through the detector. The charge is integrated and proportional to energy. A resolution of $\sim$ 2 keV is achieved at 511 keV. Surrounding the Ge detectors is an active anti-coincidence shield made from gamma-ray sensitive NaI. NaI is a light scintillator (the gamma-ray energy is converted to light within the transparent crystal) with an energy resolution of $\sim$ 40 keV at 511 keV. The amplitude of the light flash is measured with photo multipliers attached to the crystal and is proportional to gamma-ray energy. The shield is doing 3 things:

1. It defines the field-of view in this case 18° FWHM.
2. It suppresses less than full energy deposition events in the Ge detectors. (There is a significant probability that a fraction of the incident gamma-ray energy will be

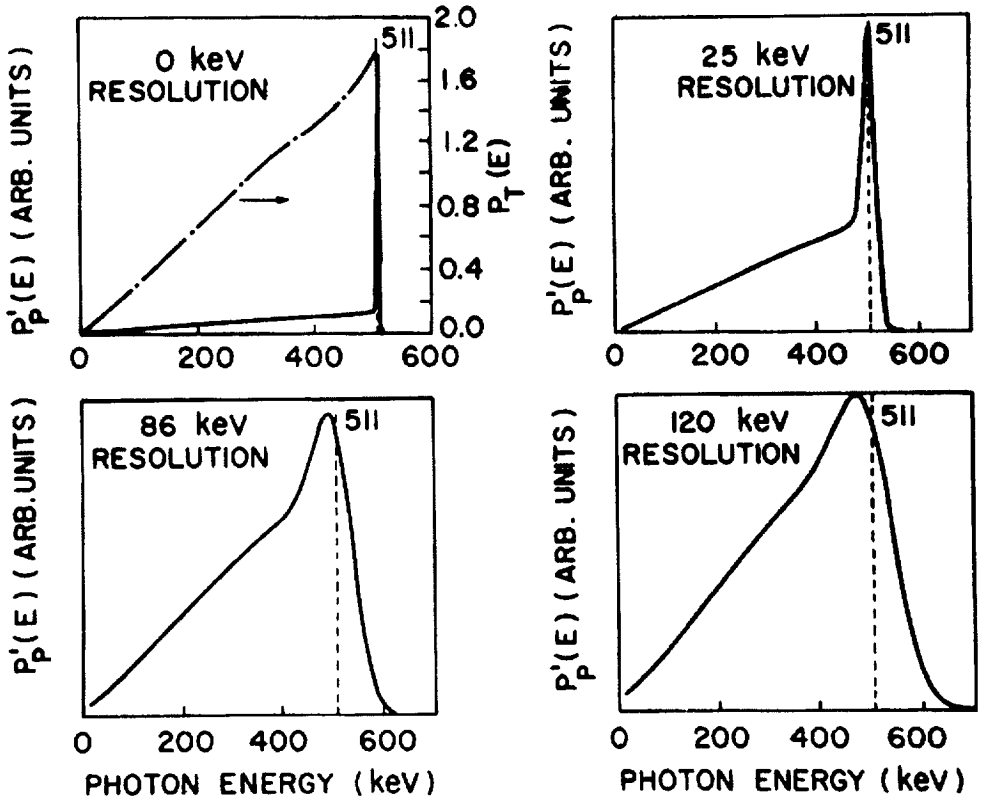


Figure 6.2: A pure positronium spectrum convoluted with degrading resolution functions. The broken line is the 3-photon ortho-positronium spectrum shown by itself.

Compton scattered out of the detector.)

3. It suppresses a large gamma-ray background generated by cosmic-ray interactions.

Energetic cosmic-ray particles, mostly energetic protons, are continually striking the atmosphere and the telescope and creating secondary gamma-rays. GC annihilation line spectra achieved in May and October of 1988 with GRIS are shown in Figure 6.4. Data were accumulated in alternating 20 minute target background segments allowing the instrumental background to be determined and subtracted before plotting in Figure 6.4. Line flux, line width, and f values were determined by doing least-squares curve fitting of the data to a function consisting of a single Gaussian, a positronium continuum of adjustable amplitude and an inverse power law continuum. The sum of the squares of the deviations between the observed data points and the theoretical model is constructed (chi-squared) and minimized by varying the adjustable parameters (Smith et al., 1995). The October balloon flight was long enough to allow a pointing at a target in the galactic plane but removed by 25° from the GC. The off-center signal has almost completely disappeared confirming the fact that GRIS is studying a source that is largely localized in the GC region. GRIS observed the GC again in April and May of 1992 (Leventhal et al., 1993). These results are shown in Figure 6.5. The line fluxes appear to be intermediate between the 1988 values. All of the GRIS results are tabulated in Table 6.1. A fundamental question to ask about the GRIS data set is whether or not statistically significant changes in line flux and line width are being observed with the implied presence of a variable point source. (Astronomers often make a light travel time argument which implies that an astronomical source observed to vary in its emission on a time scale t must be able to communicate with itself on this same time scale and hence is less than ct in extent.) This question can be answered in a quantitative way by assuming that GRIS is looking at statistical fluctuations in the emission from a steady source. Weighted mean quantities and chi-squared per degree of freedom (the number of data points minus the number of variable parameters) values for this simple model are indicated in Table 6.1. The probability of getting chi-squared values this large or larger by chance are 0.18 and 0.10 for the lines fluxes and line widths, respectively (Smith et al., 1995). Hence the evidence for source variability within the GRIS data set alone is suggestive but not compelling. We will see shortly that within the larger historical data set the case for variability is much stronger.

Some of the telescopes involved in this saga have

central detectors made from NaI. The most sensitive of these is the Oriented Scintillation Spectrometer Experiment, (OSSE) currently flying on the Compton Gamma-Ray Observatory (Johnson et al., 1993), launched in April 1991 and shown in Figure 6.6. The field-of view of this instrument is $\sim 4° \times 11°$. As of April of 1992 OSSE had viewed the GC region for seven two-week observing periods. The last period ended 3 days before the first GRIS 1992 balloon flight. A typical OSSE GC spectrum is shown in Figure 6.7 clearly indicating the annihilation line and the positronium continuum (Purcell et al., 1993). For its first two years of operation the OSSE data suggest that they are looking at a steady source within a few degrees of the GC with flux $\sim 2-3\times10^{-4}$ photons $cm^{-2}s^{-1}$, be it a point source, a diffuse source or some combination (Purcell et al., 1993). This flux value is about a factor of 4 less than that reported by the near simultaneous GRIS balloon flight and seems to be independent of the orientation of the OSSE collimator. OSSE data obtained since then continues to support this conclusion but has not yet been published. One obvious conclusion to reach from this comparison is that the GRIS data cannot be understood in terms of emission from a single point source near the GC. Additional information on the spatial extent of the distribution has been obtained by OSSE pointing in

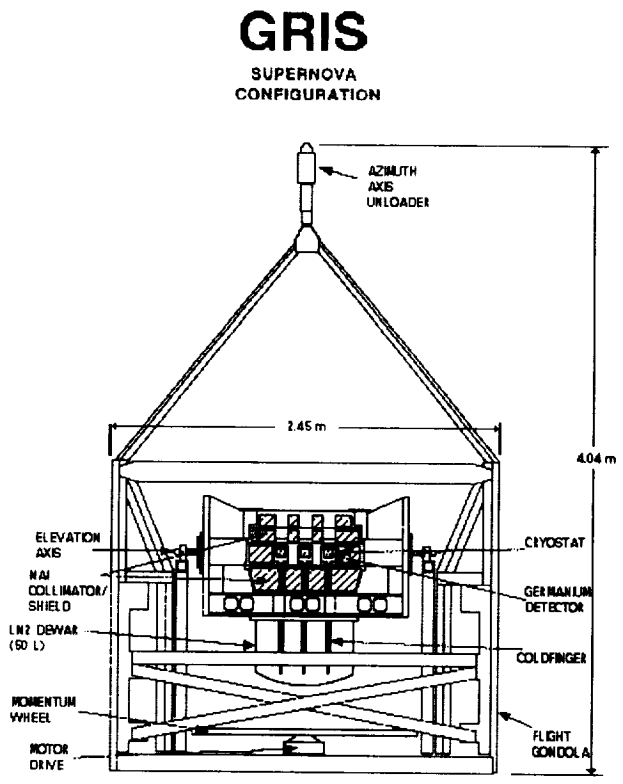


Figure 6.3: A schematic diagram of the GRIS balloon instrument (Tueller et al., 1994).

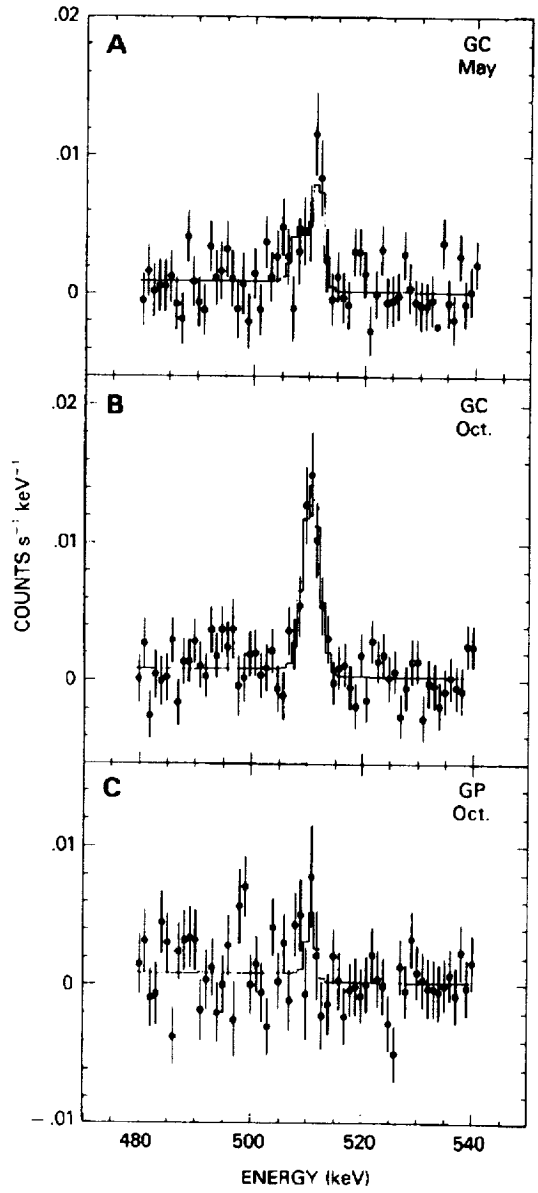


Figure 6.4: Spectra and model fits in the vicinity of the 511 keV line for the 1988 GRIS observations (Gehrels et al., 1991). The galactic plane observation was made in the direction l = 335°, b = 0°.

the galactic plane away from the GC as indicated in Figure 6.8. The innermost dotted curve is a model fit to a point source at the GC and the outermost solid curve is a model fit to a disk plus spheroid model. Various model fits to the data will be discussed below.

The complete history of these measurements through April 1992 is shown in Figure 6.9. The vertical axis is the point source equivalent flux which would have produced the observed 511 keV signal. Each obser-

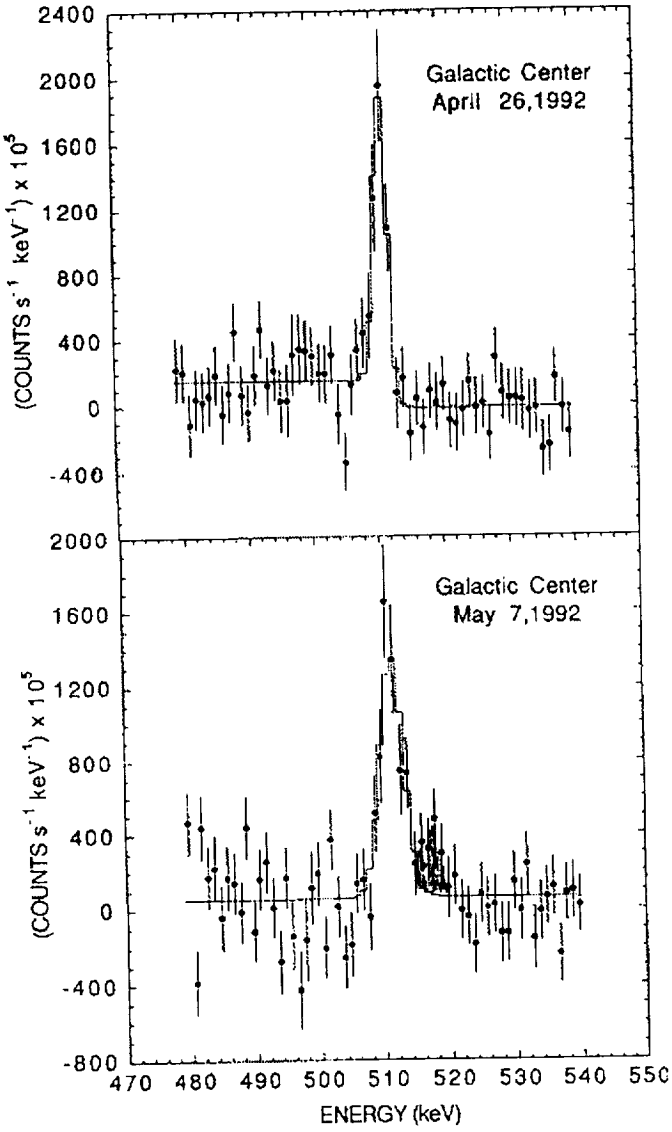
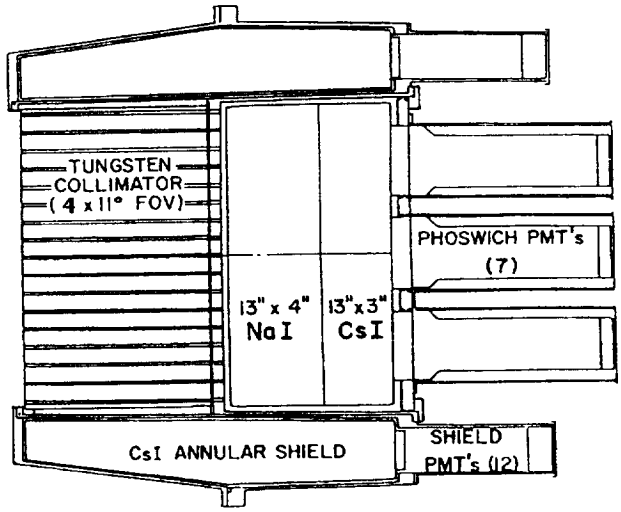


Figure 6.5: Spectra and model fits of the 511 keV line for the 1992 GRIS observations (Leventhal et al., 1993).



OSSE Detector

Figure 6.6: A schematic diagram of 1 of 4 identical OSSE modules.

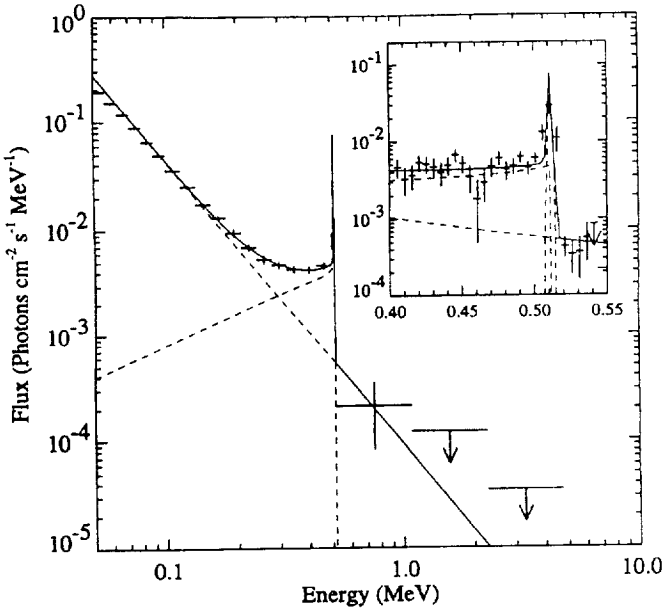


Figure 6.7: The OSSE GC spectrum accumulated over the period 13-24 July 1991 (Purcell et al., 1993).

vation is plotted with +/- 1 sigma (1 standard deviation) statistical error bars only. The FWHM field-of view of each instrument is indicated. Short horizontal bars indicate experiments having central Ge detectors. Possible systematic errors in the various experiments are thought to be small and are ignored in the discussion to follow. Many workers studying this data set have reached two general conclusions. The first is that instruments with larger fields-of view seem to be detecting larger signals. The second is that the narrower field-of view Ge detectors seem to be detecting a variable signal which perhaps was high in the late 1970s, went low in the early 1980s and reappeared in the late 1980s. These conclusions were formalized into a two-component model for the source by Lingenfelter and Ramaty (1989). They suggested that the source contained a truly diffuse component extended over the galactic disk probably due to the radioactive byproducts of galactic supernovae and novae (such as ^{56}Co, ^{44}Ti, ^{26}Al etc.) and a compact, variable source located near the GC, probably an accreting stellar mass black hole producing pairs by gamma-gamma

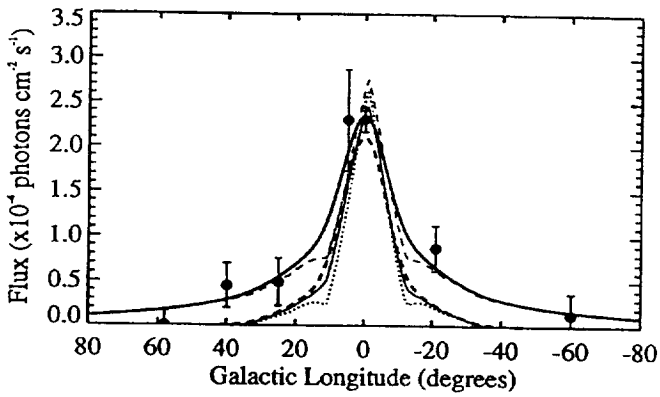


Figure 6.8: The fitted 511 keV line flux for the galactic plane observations. The curves represent the expected OSSE responses for several galactic distributions fitted to the data (Purcell et al., 1993).

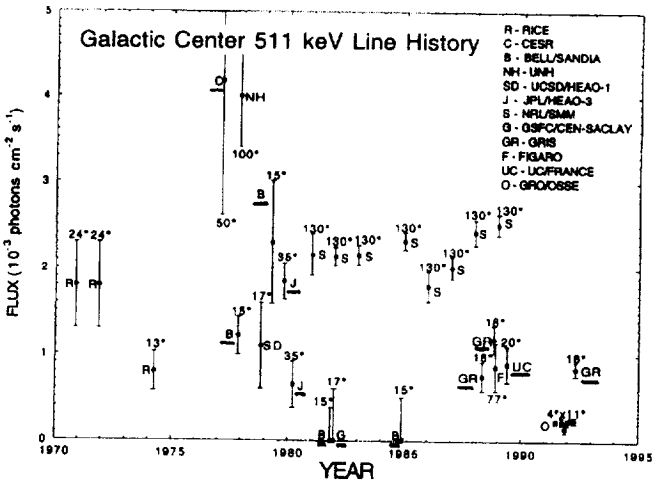


Figure 6.9: A history of the GC 511 keV line observations.

interactions in a relativistic plasma near the inner edge of an accretion disk. Gas falling into a black hole will, in general, carry some angular momentum with it and first form a hot disk outside of the event horizon. It is generally believed that about 10% of the mass energy of accreted material is radiated away before disappearing into the black hole. At temperatures above about 10^8 K electron positron pairs are generated spontaneously by the scattering of energetic photons off of each other in the hot plasma. A number of other sources of galactic disk positrons have been considered in the past such as radio pulsars and secondary cosmic-ray production but have ultimately been discarded because they do not produce sufficient quantities. Recently Ozernoy has suggested that star burst activity in the GC might contribute to the central component.

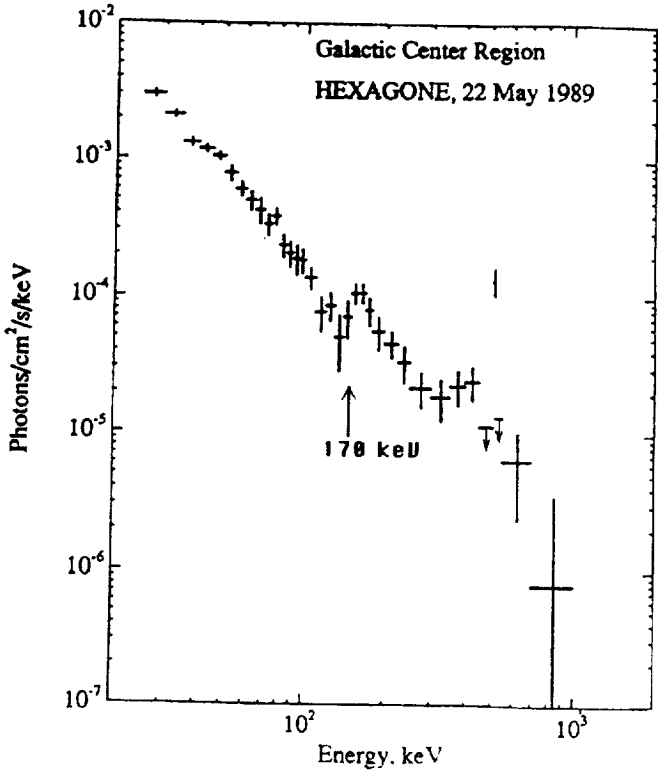


Figure 6.10: The Hexagone 1989 GC spectrum showing a possible 170 keV backscattered feature (Lingenfelter and Hua, 1991).

6.3 DISCUSSION

At the heart of the controversy surrounding this data set is the need for a variable point source of narrow 511 keV emission near the GC. The NaI gamma-ray spectrometer on the Solar Maximum Mission, SMM, monitored the GC source during January and February for most on the decade of the 1980s (Share et al., 1990). While the scatter in their results is more than one would expect statistically, they concluded that their data set is consistent with but does not require the presence of a variable point source with time-averaged flux $\sim 4 \times 10^{-4}$ photons $cm^{-2}s^{-1}$. This conclusion has been reinforced by the OSSE results which suggest steady emission for most of 1991-93. However, the earlier Ge results (particularly the Bell/Sandia and HEAO-3 results) seem inconsistent with steady emission. An additional complication is a controversial re-analysis of the HEAO-3 results which diminishes but does not eliminate the variability in the HEAO-3 results (Mahoney et al., 1994). Indirect evidence supporting the presence of a variable point sources can also be found in the occasional detection of 170 keV features in the GC spectra (see Figure 6.10). These have been interpreted in terms of backscattered 511 keV line photons from the inner edge on an accretion disk sur-

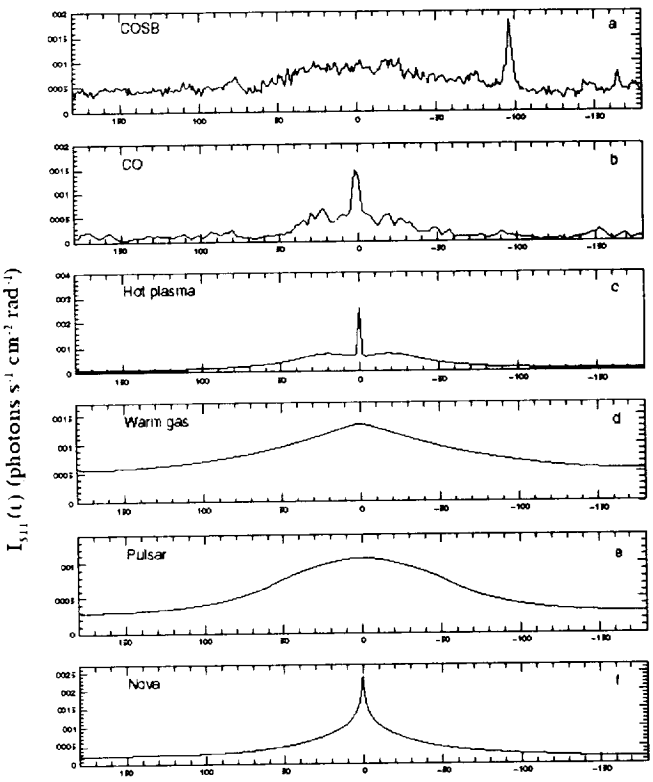


Figure 6.11: The intensity distributions for the six assumed models integrated over latitude (Smith et al., 1993).

rounding a compact object (Lingenfelter and Hua, 1991; Smith et al., 1993). However no such features have been detected in the OSSE data set (Smith et al., 1995).

The OSSE team (Purcell et al., 1993) and a second group based at Goddard Space Flight Center (Skibo et al., 1992) have independently attempted to resolve this controversy by doing model fits to the complete data set. Both groups agree that no single component model based on previously obtained galactic distributions at other wavelengths, which conceivably might be related to the positron distribution, fits the data. The problems are that the 511 keV line distribution is too strongly peaked in the GC direction and the variability. The various distributions tried by the Goddard group are shown in Figure 6.11. The nova (thought to be explosions on the surface of accreting white dwarf stars in binary systems) distribution is taken from Higdon and Fowler (1989) and gave the best of the unsuccessful fits. This distribution is also believed to be similar to the galactic Type 1a supernova (thought to be disruptive explosions of accreting white dwarf stars in binary systems) distribution. An escape fraction of $\sim 1\%$ for the ^{56}Co positrons produced in galactic Type 1a supernovae could account for much of the observed distribution. The analytical form of the

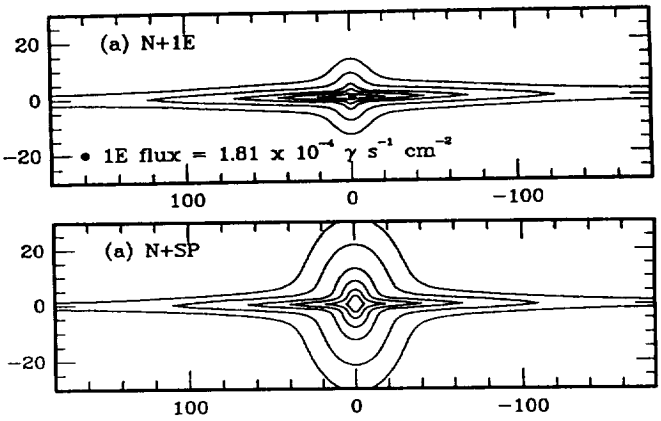


Figure 6.12: Fitted contour maps of the 511 keV line intensity in galactic longitude and latitude with an added point source or an enhanced spheroid. The intensities along the contours are 0.5, 1, 2.5, 5, 10 and 20 where the outer contours represent the 0.5×10^{-3}ph $cm^{-2}s^{-1}$ intensity (Ramaty et al., 1994).

distribution is

$$\epsilon(\rho, z) = N_d \exp[-44.5z^2 - 0.297(\rho - R_\odot)]$$

for the disk, and by

$$\epsilon(R) = N_{sph} \exp\left[10.093\left(1 - \left(\frac{R}{R_\odot}\right)^{1/4}\right)\right] \times \left[1.25\left(\frac{R}{R_\odot}\right)^{-3/4}\right] \quad R \le 0.03R_\odot$$

$$\times \left[\left(\frac{R}{R_\odot}\right)^{-7/8}\left[1 - 0.08669\left(\frac{R}{R_\odot}\right)^{-1/4}\right]\right] \quad R \ge 0.03R_\odot$$

for the spheroid

This indicates its disk plus spheroid ($\sim 8°$ FWHM) nature. A more successful but still not completely satisfactory (2 sigma) fit to the complete data set is obtained either by significantly enhancing the intensity of the spheroid component as preferred by the OSSE team or by adding a GC point source as preferred by the Goddard group. Contour maps of the fitted 511 keV line intensity in galactic longitude and latitude are shown in Figure 6.12. The total number of counts required in the spheroidal component is about 4 times that in the disk component , the total number of galactic annihilations per second is $\sim 10^{43}$ and the flux from the central radian of the galaxy is $\sim 2 \times 10^{-3}$ photons $cm^{-2}s^{-1}$. Successful models can also be obtained by adding a point source or a spheroid to a number of the other one-component distributions. A basic conclusion to reach about the galactic positron distribution is that it has a bright central component (FWHM $< 5 - 10°$) which has remained steady over the last few years.

The question of the existence of time variable narrow 511 *keV* line emission from the GC remains controversial with the debate turning on how to include the ever increasing number of OSSE observations in the fitting. The inclusion of the large number of OSSE measurements which show little variability on an equal footing with the historical measurements seems to statistically reduce the need for a variable source to about the 2 sigma level. However those favoring an historically variable source argue that the OSSE results are just indicating that the source has been steady in the recent past.

6.4 THE GREAT ANNIHILATOR

A dramatic development in this saga took place on October 13-14, 1990 when the French SIGMA experiment (an imaging NaI "telescope") on the Russian GRANT spacecraft detected an intense 1 day long burst of annihilation radiation from the X-ray source 1E1740.7-2942, a stellar mass black hole candidate located in the galactic plane at an angular distance of 0.9° from the GC (Bouchet et al., 1991; Sunyaev et al., 1991). In Figure 6.13 we show this transient feature superposed upon the continuum seen often from this highly variable source. A broad annihilation line feature with flux $\sim 1.3 \times 10^{-2}$ photons $cm^{-2}s^{-1}$ centered at $\sim$ 480 *keV* with a FWHM of $\sim$ 240 *keV* is apparent. This broad feature is clearly different from the historically observed 511 *keV* galactic line discussed above, which is much narrower, much less intense and centered exactly on 511 *keV*. The continuum shown in Figure 6.13 is thought to come directly from the hot plasma in an accretion disk and is best fit by the Comptonized disk model of Sunyaev and Titarchuk (1980) with an electron temperature of 30-60 *keV* and an optical depth 1-2. This spectrum is similar in shape and intrinsic luminosity with that seen often from the classic black hole candidate Cyg X-1. It should be pointed out however that a number of other workers viewing the 1E1740.7-2942 source at other times have not been able to detect transient line features near 511 *keV*(Harris et al., 1994; Jung et al., 1995).

Since the detection of the broad annihilation feature from the 1E1740.7-2942 source it has been the leading candidate for the variable, narrow 511 *keV* GC feature. It was quickly pointed out that this source is aligned with a dense (n $\sim 10^5$ cm^{-3}) giant molecular cloud (Bally and Leventhal, 1991). A contour map of the ^{13}CO emission from the GC region showing gas in the velocity range -200 to -100 $km\ s^{-1}$is shown in Figure 6.14. This cloud is dense enough to slow down and annihilate energetic positrons on a time scale of about 1 year and cold enough to yield a narrow 511 *keV* line. Hence Ramaty et al., 1992, suggested that positrons injected impulsively into the cloud from the 1E source could account for the historically variable results. In this model the thermally broadened features arises from positrons stopping promptly in the inner region of a hot (T $\sim 10^8$ K) and dense accretion disk surrounding the black hole. The apparent redshift requires the annihilation region to be closer than $\sim$ 10 Schwarzschild radii from a stellar mass black hole. Roughly equal numbers of positrons annihilating in the disk and escaping into the cloud coupled with a duty cycle of a few percent for the broad line state are required to make such a model work. It is interesting to note that a simple Bondi-Hoyle accretion calculation indicates that the system could be powered by accretion directly from the molecular cloud rather than from a stellar companion. Dust obscuration in the galactic plane (50 mag in the visible) severely hampers the search for an optical companion. Nevertheless no companion has been found to 17th magnitude K band suggesting that the companion is not a giant star as in the case of Cyg X-1[37].

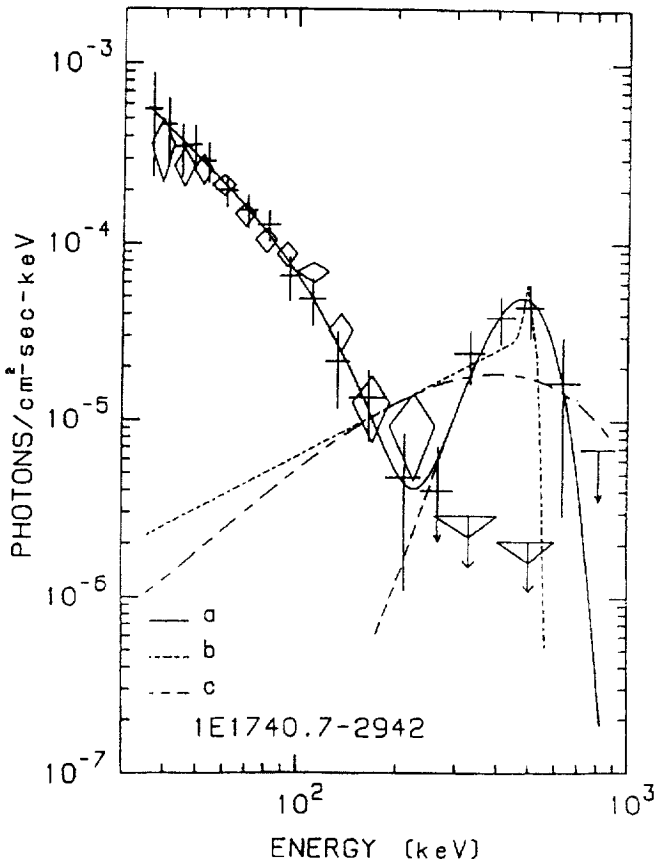


Figure 6.13: The energy spectrum of 1E1740.7-2942 as obtained by SIGMA on 1990 October 13 (crosses); the spectrum derived from 1990 March/April (diamonds) is shown for comparison (Bouchet et al., 1991).

A recent series of remarkable Very Large Array (VLA) of radio telescopes in Socorro New Mexico observations, by Mirabel et al., 1992 and his colleagues, have

established the existence of a highly variable, nonthermal radio counterpart of the 1E source which in many respects exhibits "microquasar" like properties. (Quasars, QSOs, are the most luminous objects in the known universe. They are believed to be the nuclei of young active galaxies, AGN's and often exhibit variability and jet like morphology emanating from an unresolved central core. They are widely believed to be massive, accreting black hole engines. See Table 6.2.) The central radio source was unambiguously identified by demonstrating correlated X-ray and radio variability (by about a factor of4). The radio spectrum was similar to that of Cyg X-1, following a power law with variable index $\sim$ - 0.8. In fact the best argument that the 1E source is an accreting stellar mass black hole system is the similarity of its X-ray and radio spectrum with that of Cyg X-1. Perhaps the most surprising discovery is the detection of twin radio jets and lobes emanating from a central radio source. See Figure 6.15. Mirabel and Rodriguez (1994) have determined that the particles in the jets from the October 1990 flare are moving at $\sim$ 0.9c and will reach the outer lobes at the tip of the jets by the end of 1994. The jets which

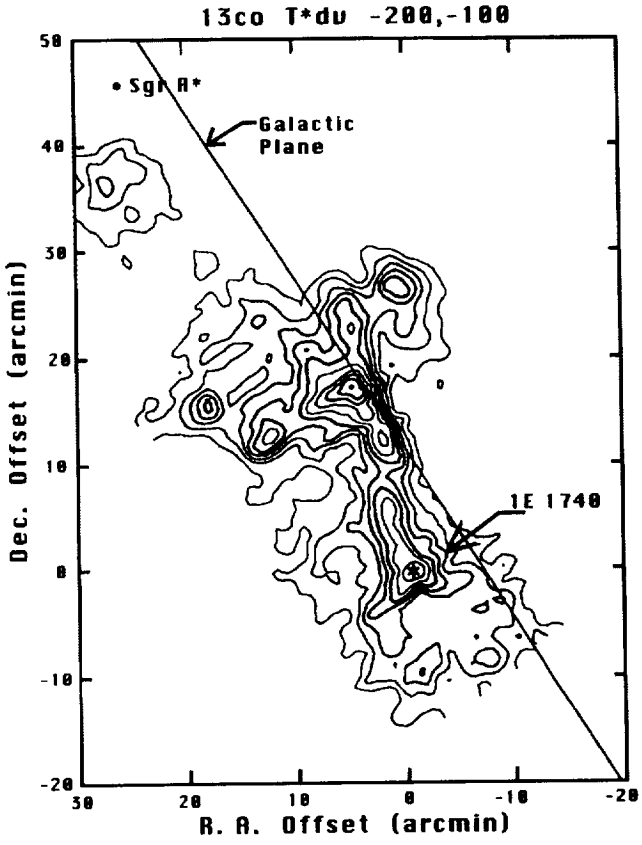


Figure 6.14: A contour map of ^{13}CO emission around the position of 1E1740.7-294223. The position of 1E1740.7-2942 is indicated by a star.

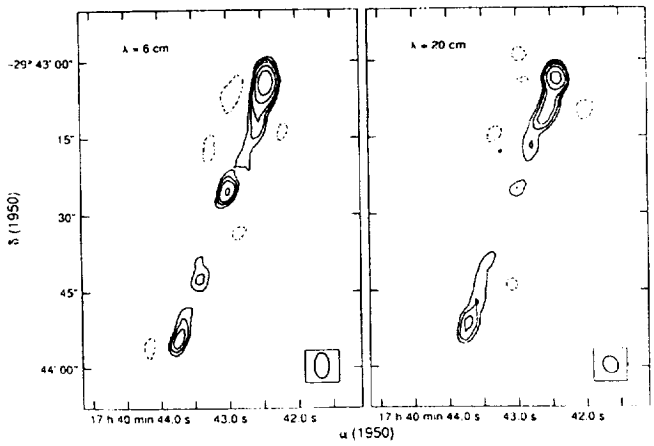


Figure 6.15: The radio jet morphology of the 1E source. The thin solid lines are the 6 *cm* VLA contour map (Mirabel et al., 1992) overlayed on the HCO+ contours of the molecular cloud along the line of sight.

are about 0.5 *arcmin* in extent or about 1 *pc* at the GC distance are probably due to synchrotron emission from e^+e^- pairs streaming out at high velocities from the high energy source. Similar radio morphology has also been detected from a second black hole candidate GRS1758-258 removed about 5° from the GC although no annihilation radiation has yet been detected from it. Chen, Gehrels, and Leventhal (1994) have concluded that the 1E black hole is likely to have a stellar companion with a mass less than 9 M_o and that the accretion is mainly from the ISM. They also suggest that the bipolar jets are composed of a combination of e^+e^- pairs and ordinary matter.

6.5 CONCLUSIONS

It is tempting to try and tie together the entire history of variable annihilation radiation from the GC center with the model indicated in Figure 6.16. Here we have a stellar mass black hole slowly moving through and accreting material from a giant molecular cloud near the GC. Instabilities in the accretion process temporarily produce a relativistic pair plasma. Some of the positrons annihilate promptly within a few hundred kilometers of the black hole producing the broad line observed by SIGMA. Some fraction of the e^+e^- pairs will be accelerated by radiation pressure and collimated perhaps by magnetic fields or an accretion disk. These pairs streaming away at relativistic velocities will produce the aligned radio jets observed with the VLA. After spending several years traversing the jets the positrons will eventually slow down and annihilate in high density clumps in the molecular cloud on a time scale of 1 year giving rise to the variable narrow line seen historically (Misra and Melia, 1993).

Table 6.1: GRIS galactic Center Region 511 keV Line Results

Flight	Flux (photons $cm^{-2}s^{-1}$)	FWHM (keV)	Centroid (keV)
1988 May 1	$(7.5 \pm 1.7) \times 10^{-4}$	≤ 3.6	511.46 ± 0.38
1988 Oct 29	$(11.8 \pm 1.6) \times 10^{-4}$	2.9 ± 0.6	510.97 ± 0.26
1992 April 26	$(7.7 \pm 1.2) \times 10^{-4}$	1.3 ± 0.7	511.34 ± 0.18
1992 May 7	$(8.9 \pm 1.1) \times 10^{-4}$	3.6 ± 1.0	511.37 ± 0.30
Weighted Mean	$(8.8 \pm 0.7) \times 10^{-4}$	2.5 ± 0.4	
Chi-Squared	1.65 (3 D. O. F.)	2.29 (2 D. O. F.)	
Probability	0.18	0.10	

Positrons reaching dilute regions of the molecular cloud or other dilute components of the ISM will slow down and annihilate on a much longer time scale giving rise to a quasi-steady source which is perhaps what the OSSE experiment is observing. A crucial piece of observational data remains to be achieved before this exciting model can be taken seriously, namely a clear association of the 1E source with a narrow 511 keV line. SIGMA searches for such a line have only yielded upper limits of a few $\times 10^{-4}$ photons cm^{-2} s^{-1} so far. However some new support for such a model has recently been produced by Tueller, Purcell, and Leventhal (private communication). There have only been two OSSE scans of the GC taken with the collimator perpendicular to the galactic plane. This scanning mode should give the highest spatial resolution for determining the galactic longitude centroid of the 511 keV distribution. The result is shown in Figure

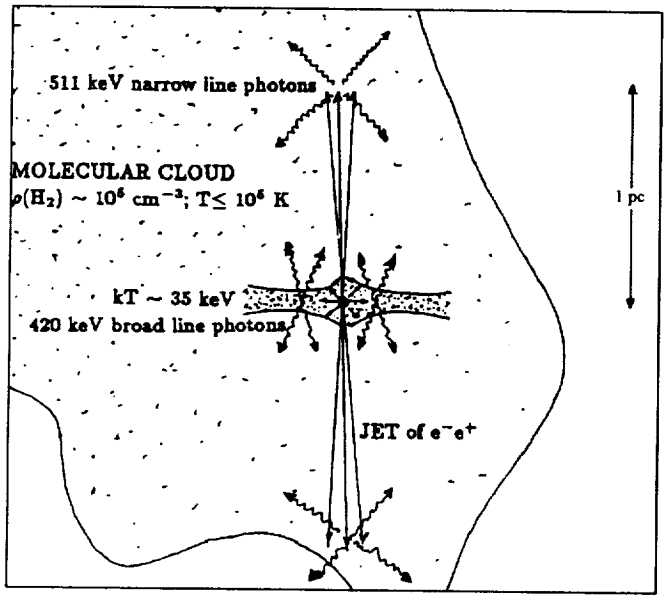


Figure 6.16: Schematic diagram for the "Great Annihilator" as a microquasar source of GC positrons (Mirabel et al., 1992).

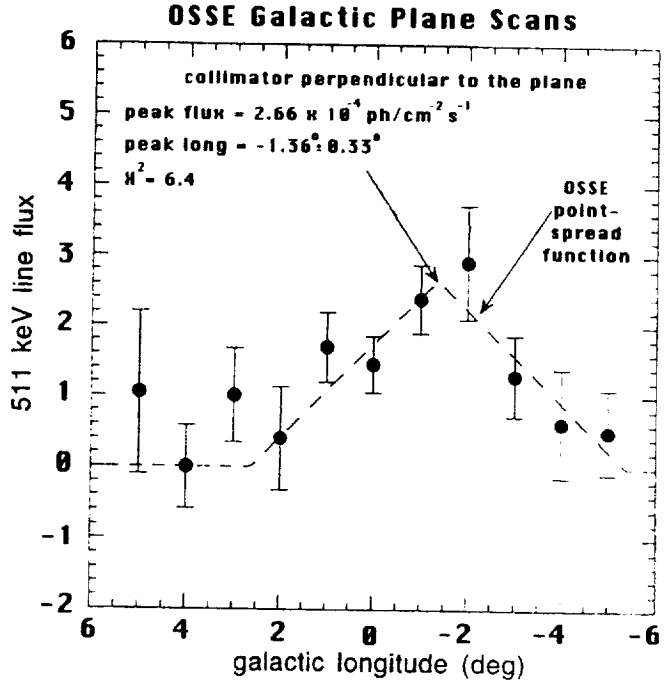


Figure 6.17: OSSE scans of the GC 511 keV line along the plane with the long dimension of the collimator perpendicular to the plane. The peak emission is 4.1 sigma off of the GC. The data are fit with a 4° FWHM, triangular response.

6.17. The error bars indicate the statistical uncertainty only which is believed to be the dominant uncertainty. Clearly the peak of the distribution is not at the GC but is shifted off to negative longitude by an amount consistent with the location of the 1E source. A bright GC spheroid component might well be due to one or a small number of black hole systems periodically injecting positrons into the interstellar medium. In this regard it is interesting to point out that a second transient annihilation line source has been detected in the archival HEAO A4 data set by Briggs (1991). This source is removed in latitude by about 14° from the GC. The idea that accret-

Table 6.2: Quasar and 1E Properties

Property	Quasar	1E1740.7-2942
$M_b h$	$\sim 10^{6-8}$ M_o	$\sim 10^{1-2}$ M_o
Size	$\sim 10^6_{ly}$	$\sim 10_{ly}$
L_{x-ray}	$\sim 10^{44}$ erg s^{-1}	$\sim 10^{37}$ erg s^{-1}
L_{radio}	$\sim 10^{44}$ erg s^{-1}	$\sim 10^{30}$ erg s^{-1}

ing black hole systems can generate bursts of positrons is also substantiated by the detection of a transient annihilation line from Nova Musca (Goldwurm et al., 1992). It seems more than coincidental that all of the essential ingredients of an AGN or a QSO seem to be present in the 1E source i.e. a black hole, an accretion disk and collimated radio jets although all are greatly reduced in scale. Table 6.2 compares some of the fundamental properties of a quasar and the 1E source. Perhaps nature has indeed provided us with a microquasar right here in the Milky-Way.

6.6 REFERENCES

Bally, J., and Leventhal, M., 1991, Nature, 353, 234.

Bevington, P.R., 1969, in "Data Reduction and Error Analysis for the Physical Sciences" McGraw-Hill.

Bouchet, L., et al., 1991, ApJ, 383, L45.

Briggs, M.S., 1991, Ph.D. thesis, Univ. California, San Diego.

Brown, B.L., and Leventhal, M., 1987, ApJ, 319, 637.

Chen, W., Gehrels, N., and Leventhal, M., 1994, ApJ, 426, 586.

Gehrels, N., Barthelmy, S.D., Teegarden, B.J., Tueller, J., Leventhal, M., and MacCallum, C.J., 1991, ApJ, 375, L13.

Goldwurm, A., et al., 1992, ApJ, 389, L79.

Guessoum, N., Ramaty, R., and Lingenfelter, R.E., 1991, ApJ, 378, 170.

Higdon, J.C., and Fowler, W.A., 1989, ApJ, 339, 956.

Johnson W.N., et al., 1993, ApJS, 86, 693.

Leventhal, M., Barthelmy, S.D., Gehrels, N., Teegarden, B.J., Tueller, J., and Bartlett, L.M., 1993, ApJ, 405, L25.

Lingenfelter, R.E., and Hua, X.M., 1991, ApJ, 381,426.

Lingenfelter, R.E., and Ramaty R., 1989, ApJ, 343, 686.

Mahoney, W.A., Ling, J.C., and Wheaton, W.A., 1994, ApJS, in press.

Mirabel, I.F., Morris, M., Wink, J., Paul, J., and Cordier, B., 1991, A and A, 251, L43.

Mirabel, I.F., Rodriguez, L.F., Cordier, B., Paul, J., and Lebrun, F., 1992, Nature, 358, 215.

Mirabel, I.F., and Rodriguez, L.F., 1994, in "The Second Compton Symposium" eds. C.E. Fichtel, N. Gehrels, and J.P. Norris (New York: AIP), 413.

Misra, R., and Melia, M., 1993, ApJ, 419, L25.

Purcell, W.R., et al., 1993, ApJ, 413, L85.

Purcell, W.R., Grabelsky, D.A., Ulmer, M.P., Johnson, W.N., Kinzer, R.L., Kurfess, J.D., Strickman, M.S., and Jung, G.V., 1993, in "Compton Gamma-Ray Observatory" eds. M. Friedlander, N. Gehrels and D.J. Macomb (New York: AIP), 107.

Ramaty, R., Leventhal, M., Chan, K.W., and Lingenfelter, R.E., 1992, ApJ, 392, L63.

Ramaty, R., Skibo, J.G., and Lingenfelter, R.E., 1994, ApJS, in press.

Share, G.H., Leising, M.D., Messina, D.C., and Purcell, W.R., 1990, ApJ, 358, L45.

Skibo, J.G., Ramaty, R., and Leventhal, M., 1992, ApJ, 397, 135.

Smith, D., et al., 1995, ApJ, in press.

Smith, D.E. et al., 1993, ApJ, 414, 165.

Sunyaev, R., et al., 1991, ApJ, 383, L49.

Sunyaev, R., and Titarchuk, L.G., 1980, A and A, 86, 121.

Tueller, J., 1993, in "Compton Gamma-Ray Observatory" eds. M. Friedlander, N. Gehrels, and D.J. Macomb (New York: AIP), 97.

Tueller, J., Barthelmy, S.D., Bartlett, L.M., Gehrels, N., Leventhal, M., Palmer, D.M., and Teegarden, B.J., 1994, ApJS, in press.

von Ballmoos, P. 1991, ApJ, 380,98.

Chapter 7

PULSARS: PERSPECTIVES FROM X-RAYS AND GAMMA-RAYS

ALICE K. HARDING
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

With the increased sensitivity of gamma-ray detectors on the Compton Gamma-Ray Observatory (CGRO) the number of presently known gamma-ray pulsars has grown. The new detections are beginning to provide clues to the origin of the high-energy radiation in the form of emerging patterns and correlations among observed quantities such as gamma-ray efficiency and spectral index vs. age. But there are still many questions about the location of the emission and its relation to the radio, optical and X-ray pulses. This chapter will describe some of the basic physics of pulsars and their high-energy emission.

7.1 INTRODUCTION

Since their serendipitous discovery in 1967 (Hewish et al., 1968) during radio observations of sources of interplanetary scintillation, more than 500 radio pulsars have been detected. Presently, seven have been discovered to pulse at gamma-ray energies and ten have pulsed emission at X-ray energies. The Crab and Vela pulsars, the two youngest, have been known gamma-ray pulsars for nearly two decades. Both show pulsed emission up to several GeV, and the Crab has unpulsed emission up to at least several TeV. Increased sensitivity of detectors on the Compton Gamma-Ray Observatory (CGRO) has now tripled the number of gamma-ray pulsars. These detections provides new clues to the origin of the high-energy radiation, which is a powerful source of information on the acceleration of particles in the pulsar magnetosphere and thus on the origin of emission at all other wavelengths.

This lecture is intended to be a summary of and an introduction to the high-energy emission from pulsars. This emission is the most direct way to study the production of the high energy particles by a pulsar. Gamma rays are the "smoking gun" that tell us that these particles must exist. In addition, the pulsed-radiation efficiency reaches a maximum at the highest energies, which is an incredible 30% of the rotational energy loss in PSR1055-52 (Thompson, Chapter 4). Therefore, high-energy emission is also an important part of the pulsar energy budget.

A brief summary of the basic model for pulsars based on a rotating, magnetized neutron star will be given in the next section, followed by a discussion of the high-energy emission, including the physics of the radiation processes thought to be operating and the current acceleration and emission models. A number of reference books on pulsars give a good background on the observational aspects of radio pulsars (Manchester and Taylor, 1977; Lyne et al., 1990) and on the theoretical aspects of pulsar magnetospheres (Michel, 1991). However, the models for high energy emission and acceleration are still being developed and compared to the data, so we do not yet have a clear picture of how pulsars radiate.

7.2 THE ROTATING NEUTRON STAR MODEL

The periods of observed pulsars lie in the range 1.5 ms - 4.3 s and are remarkably stable to 1 part in 10^7. Such stability requires a compact object with high inertia, either a neutron star or white dwarf. Stellar pulsation is ruled out because the expected periods would be too short ($\sim ms$) for neutron stars and too slow for white dwarfs. Orbital motion in a binary system is also ruled out for white dwarfs, because the period would be too long, and for neutron stars because gravitational radiation would cause the orbit to decay too quickly. Finally, rotation of a white dwarf can be ruled out because a 10 ms period would cause break up of the star due to centrifugal forces at the surface. Thus, by process of elimination, it is believed that pulsars are rotating neutron stars. This model has been adopted as the explanation for pulsar periods and has held up for over two decades to numerous tests and predictions. The most fundamental prediction (Gold, 1968; Gunn and Ostriker, 1969) is that pulsar periods should undergo a secular increase as a result of the torque from magnetic dipole radiation. This dipole radiation causes a rotational energy loss (Shapiro and Teukolsky, 1983)

$$\dot{E} = -\frac{B_o^2 \sin\alpha \Omega^4 R^6}{6c^3} = I\Omega\dot{\Omega} \qquad (7.1)$$

where B_o, Ω, R and I are the surface magnetic field strength, rotation frequency, radius, and moment of inertia of the neutron star and α is the angle between the spin axis and magnetic dipole axis Figure 7.1. All pulsars exhibit period increases at rates consistent with that expected for magnetic dipole radiation.

Expressing the change in pulsar rotation frequency in terms of a "braking index" n

$$\dot{\Omega} = -K\Omega^n \qquad (7.2)$$

one can differentiate this equation to obtain an expression for n

$$n = \frac{\ddot{\Omega}\Omega}{(\dot{\Omega})^2} \qquad (7.3)$$

that involves only directly measurable quantities. The braking indices measured for several pulsars, including the Crab (n = 2.515, Growth, 1975) and PSR1509-58 (n = 2.83, Manchester et al., 1985), are near but slightly below the braking index $n = 3$ for magnetic dipole radiation from Equation (7.1). The fact that n is observed to be less than 3 may be due to the distortion of the field by plasma near the light cylinder, c/Ω, where the corotation velocity reaches the speed of light. Integration of Equation (7.2) allows us to derive a "characteristic age" for spindown of the pulsar

$$\tau = \frac{-\Omega}{(n-1)\dot{\Omega}}\left[1 - \left(\frac{\Omega}{\Omega_i}\right)^{n-1}\right] \qquad (7.4)$$

If we assume that the initial spin frequency of the pulsar, $\Omega_i \gg \Omega$, and $n = 3$ for dipole spin-down, one obtains the commonly used formula, $\tau \simeq P/2\dot{P}$. This is an upper limit to the true age of a pulsar because there are other mechanisms, such as gravitational radiation, which can contribute to the rotational energy loss.

The identification of the rotational energy loss of a pulsar with magnetic dipole radiation in Equation

(7.1) also allows an indirect measure of the surface field strength

$$B_o \simeq \left(\frac{3Ic^3P\dot{P}}{2\pi^2R^6}\right)^{1/2} \tag{7.5}$$

where $P = 2\pi/\Omega$ is the pulsar period. Observed values of P and $\dot{P}$ for most of the pulsars gives B_o between 10^{11} and 10^{13} G, for assumed values of I ($\sim 10^{45}\ g\ cm^2$) and R ($\sim 10^6\ cm$) consistent with neutron star equations of state.

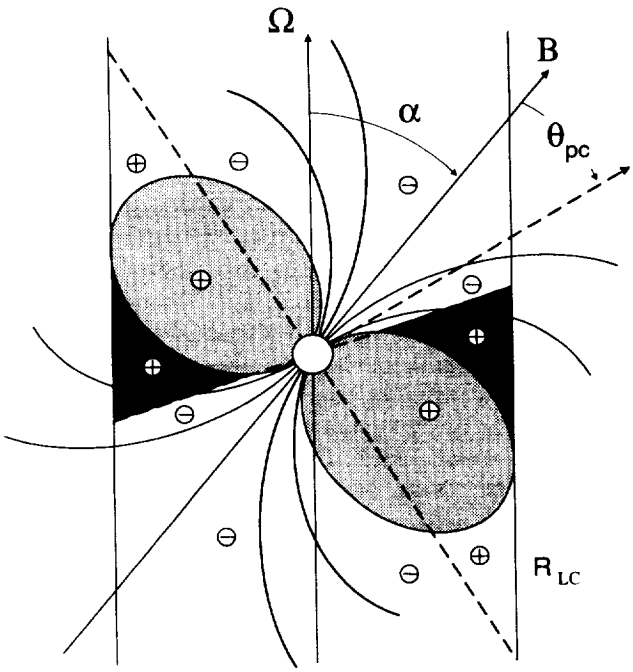


Figure 7.1: Schematic illustration of pulsar magnetosphere: dashed lines are the null charge surface ($\mathbf{\Omega} \cdot \mathbf{B} = 0$), the + and - symbols indicate the sign of the corotation charge; light shaded areas are the closed field regions; dark shaded areas are the active outer gaps. $R_{LC} = c/\Omega$ is the light cylinder.

The concept of a charge-filled neutron star magnetosphere was born of the observation by Goldreich and Julian (1969) that if one assumed that a rotating magnetized star were surrounded by a vacuum, the derived component of the induced electric field parallel to the magnetic field at the surface

$$E_{\parallel} \simeq \frac{\mathbf{v} \times \mathbf{B}}{c} = \frac{\Omega R B_o}{c} \tag{7.6}$$

would produce a force on a charged particle much larger than the gravitational force. Thus, charges of one sign would be pulled from the surface to short out the $E_{\parallel}$ and make $\mathbf{E} \cdot \mathbf{B} = 0$, so that an observer not rotating with the star will see only a component of $\mathbf{E}$ perpendicular to $\mathbf{B}$. The resulting $\mathbf{E} \times \mathbf{B}$ drift velocity of the charges causes the plasma to corotate with the neutron star. The charge density which maintains corotation of both the plasma and the magnetic field is obtained by solving $\nabla \cdot \mathbf{E} = 4\pi\rho_e$ with the condition $\mathbf{E} \cdot \mathbf{B} = 0$ and is known as the corotation or Goldreich-Julian charge density

$$\rho_e = -\frac{\mathbf{\Omega} \cdot \mathbf{B}}{2\pi c} \tag{7.7}$$

Pulling primarily one sign of charge from both polar caps would charge up the neutron star, an undesirable result, unless the charge were being simultaneously replaced. While it is believed that pulsars must maintain current closure, how this happens is one of the major unsolved problems in pulsar physics. (See Michel, 1991.)

The plasma can maintain corotation only out to the light cylinder, $R_{LC} = \Omega/c$, beyond which the field lines are distorted by the plasma into an open toroidal configuration. This open field plus plasma forms a wind that flows away from the pulsar at a relativistic speed, and carries a Poynting flux nearly equal to that carried by an electromagnetic dipole wave. The open dipole field lines, those which intersect the light cylinder, map out a polar cap region at the neutron star surface about the magnetic axis which has an opening angle

$$\sin\theta_{pc} = \left(\frac{R\Omega}{c}\right)^{1/2} \tag{7.8}$$

Particle acceleration and radiation will occur in the open field region and will thus define the geometry of the pulsed emission.

7.3 HIGH ENERGY EMISSION

Models for pulsar gamma-ray emission divide into two classes: the so-called polar cap and outer gap models. Gamma-rays can originate from any location in the pulsar magnetosphere where particles are accelerated to high energies. The polar cap model (Sturrock, 1971; Daugherty and Harding, 1982) assumes that acceleration takes place near the neutron star surface at the magnetic poles, a picture generally supported by studies of radio emission properties. The outer gap model (Cheng, Ho, and Ruderman, 1986) assumes that particles are accelerated in the outer magnetosphere, in vacuum gaps that develop along null charge surfaces. In both cases, electromagnetic cascades develop, creating large numbers of electron-positron pairs which radiate most of the observed gamma-rays. The pairs produced in these cascades may be responsible for radiating the observed emission at other wavelengths, such as X-ray, optical, and radio. It is believed that electron-positron pairs may be *required* for production of the radio emission by a coherent radiation process, possibly coherent curvature radiation by particle bunches.

7.3.1 Acceleration

The potential drop across the open field lines, found by integrating Equation (7.6) from the magnetic pole to the edge of the polar cap Equation (7.8), provides an upper limit on pulsar acceleration

$$V_{pc} = \left(\frac{2\pi}{cP}\right)^2 B_o R^3 = 10^{13}\ eV \quad B_{12} P^{-2} \qquad (7.9)$$

where $B_{12} \equiv B_o/10^{12}$ G. In the standard (Goldreich-Julian) model for a pulsar magnetosphere, the charges can only undergo drift perpendicular to the magnetic field to enforce corotation. The magnetosphere is force-free (i.e., no acceleration of particles) because the magnetic field lines are equipotentials and there is no parallel electric field. In the frame corotating with the plasma there is no electric field at all. How then does particle acceleration take place to generate the observed emission? Acceleration and, therefore, gamma-ray emission can take place *only* in parallel electric fields caused by local *departures* from force-free conditions. In polar cap models, these departures from the corotation charge occur above the magnetic poles; in the outer gap models they occur in the outer magnetosphere. These possibilities are not mutually exclusive; acceleration could be taking place in both regions of the magnetosphere in the same pulsar.

There have been several suggested ways of causing acceleration due to an $E_\parallel$ near or above the polar cap. One early model by Ruderman and Sutherland (1975) proposed that it would be easier to pull electrons than ions from the neutron star surface, due to the higher ion work function, causing a depletion of charge and resulting potential drop near the surface of one pole. More recent calculations of the ion work function (Flowers et al., 1977) give lower values than that assumed by Ruderman and Sutherland, so that the surface vacuum gaps may not grow as easily. However, Arons and Scharlemann (1979) realized that even with a free supply of both signs of charge, the constraint that the particles move along curved field lines will cause the supply of charge to fall short of the corotation charge along those magnetic field lines curving toward the rotation axis (favorably curved), allowing parallel electric fields to grow. Acceleration occurs only over half of the polar cap in the original picture, with acceleration to the highest energies occurring in "slot gaps" or in a narrow range of magnetic colatitudes where pair production is minimal. However, the general relativistic effect of inertial frame dragging near the neutron star surface produces an additional $E_\parallel$ (Muslimov and Tsygan, 1992) even along unfavorably curved field lines, which could substantially change any polar cap acceleration model.

In these models, growth of accelerating potential is always self-limited, ultimately by electron-positron pair production in the strong magnetic field (see next section), since pairs can short out $E_\parallel$. In the Arons-Scharlemann model, the strong variation of pair production optical depth with magnetic colatitude across the polar cap leads to the formation of "slot gaps," where the pair production optical depth is minimal and where acceleration to the highest energies occurs (Arons, 1983). The resulting maximum polar cap potential available for acceleration in the slot gap gives approximately

$$\Phi_{\rm PC} \simeq 1.7 \times 10^{12}\,{\rm V}\, P^{-3/8} \qquad P > 0.02\, B_{12}^{8/21} \quad (7.10)$$

Due to the dependence of both the $E_\parallel$ and the pair production optical depth on P and B, the potential drop along **B** is, at most, weakly dependent on these parameters.

Vacuum gaps may grow in the outer magnetosphere along null charge surfaces, which separate regions of the opposite sign of the corotation charge. Along unfavorably curved open field lines threading this surface (see Figure 7.1) charges of one sign will escape through the light cylinder and cannot be replaced because charge coming from the polar cap below has the opposite sign. Particles can be accelerated by the resulting $E_\parallel$ in the gap to energies limited by curvature radiation reaction (Cheng, Ho, and Ruderman, 1986). The maximum gap potential (Ho, 1989) depends on several free parameters that include the gap height a, the length L and the radius of curvature of the field lines ρ_c. If one takes $a = 0.5\,\rho_c$ and $\rho_c = L = c/\Omega$, then the maximum gap acceleration voltage is

$$\Phi_{\rm OG} \simeq 8 \times 10^{11}\,{\rm V}\, B_{12}\, P^{-2} \qquad (7.11)$$

and the curvature radiation reaction limited primary energy is

$$\gamma \simeq 5 \times 10^7\, B_{12}^{1/4}\, P^{-1/4} \qquad (7.12)$$

Thus, the outer gaps are capable of achieving somewhat higher particle acceleration energies than the polar cap.

7.3.2 Gamma-Ray Emission in Polar Cap Models

If the accelerated primary particles reach energies of $\gamma \gtrsim 10^6$, then curvature radiation losses

$$\frac{d\gamma}{dt} = \frac{2}{3}\frac{e^2}{mc^3}\left(\frac{c}{\rho_c}\right)^2 \gamma^4 \qquad (7.13)$$

will dominate, as the particles move along field lines with radius of curvature, ρ_c. The curvature radiation spectrum of the primary accelerated particles is a power law that extends to the critical energy

$$E_c = \frac{3}{2}\frac{c}{\rho_c}\gamma^3 \qquad (7.14)$$

above which it cuts off exponentially. If $\gamma \lesssim 10^6$, inverse Compton scattering of X-rays, either nonthermal cascade emission or thermal emission from a hot polar cap, will be more important. More specifically, the energy loss for curvature radiation of an electron moving parallel to a magnetic field with radius of curvature $\rho_6 \equiv \rho_c/10^7\ cm$ will exceed the energy loss due to non-magnetic inverse Compton scattering of blackbody radiation at temperature $T_6 \equiv T/10^6\ K$, which is valid for $\gamma \gtrsim 10^4$ in fields around 10^{12} G (Dermer, 1990), when

$$\gamma_{\rm curv} \gtrsim 2 \times 10^6\, \rho_7\, T_6^2 \tag{7.15}$$

At $50 \lesssim \gamma \lesssim 10^4$, resonant Compton scattering will be important because the soft photons at temperatures near $10^6\ K$ will be blueshifted into the cyclotron resonance in the electron rest frame, greatly enhancing the scattering cross section and thus the energy loss rate (Daugherty and Harding, 1989; Dermer, 1989).

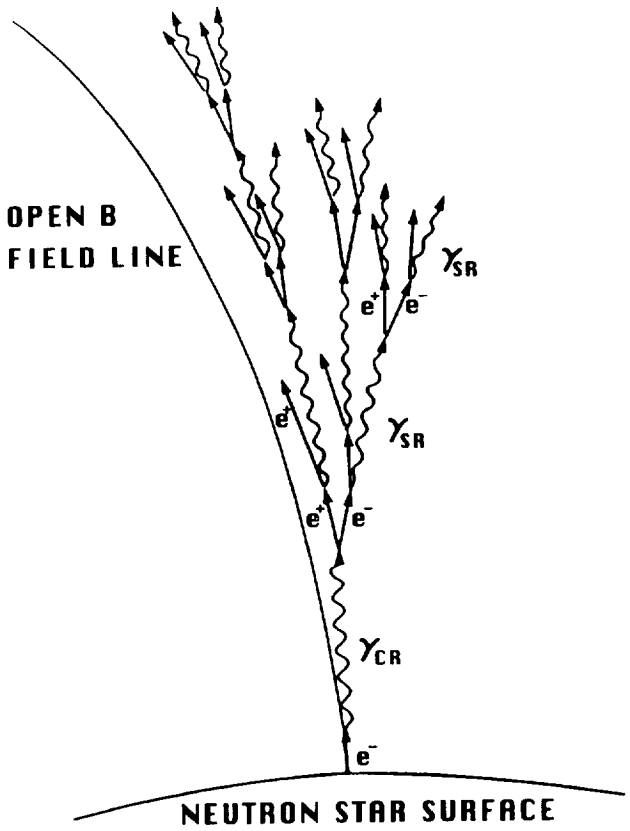


Figure 7.2: Schematic illustration of pulsar polar cap cascade initiated by curvature radiation of primary accelerated electron.

The accelerated particles will radiate gamma-rays by curvature radiation, by inverse Compton scattering, or by a combination of both. These photons then pair produce by the process, $\gamma \rightarrow e^+\, e^-$, in the strong magnetic field. This process, known as magnetic or one-photon pair production, occurs only in very high magnetic fields when the product of the photon energy E_γ, angle to the field $\sin\theta$ and field strength B satisfy both the approximate condition (Erber, 1966; Daugherty and Harding, 1983)

$$\frac{E_\gamma \sin\theta}{mc^2}\,\frac{B}{B_{cr}} \gtrsim 0.1 \tag{7.16}$$

where $B_{cr} \equiv 4.413 \times 10^{13}$ G is the critical field strength for quantum effects to become important, and the threshold condition, $E_\gamma > 2mc^2/\sin\theta$. The pairs are produced in excited Landau states that decay by emission of synchrotron photons, many of which will also produce pairs. The cascade will continue reprocessing and thus softening the primary emission spectrum until the photons escape. Figure 7.2 shows an illustration of a polar cap pair cascade initiated by curvature radiation.

These pair cascades are the basis for models of the high energy emission from the polar cap. The predicted spectrum is made up of photons that are produced in the cascade and escape from the magnetosphere. One important prediction of the polar cap models is a cutoff of the emission around a few GeV due to magnetic pair production. In fact, such a sharp turnover is seen in the spectra of most of the observed pulsars and may be an indication that the gamma-radiation originates in the strong field region. The gamma-ray emission pattern from each polar cap will be a hollow cone centered on the magnetic pole. In both curvature radiation and inverse Compton initiated pair cascades, the emission peaks at the polar cap rim where the field line radius of curvature is smallest. Thus, an observer viewing emission from a single pole may see a pulse profile with a single broad peak or two narrower peaks (Daugherty and Harding, 1994; Sturner and Dermer, 1994), depending on the viewing angle to the magnetic axis. Observers viewing emission from both poles (which requires a large obliquity angle α) will see two narrow peaks always separated by 180°. Pulse profiles like those seen in the Crab and Vela pulsars (see Thompson, Chapter 8), where the two peaks are separated in pulse phase by 140° must be produced by single-pole emission when the obliquity angle α is comparable to the polar cap angle θ_{pc} Equation (7.8).

7.3.3 Gamma-Ray Emission in the Outer Gap Model

Unlike the case in polar cap models, the magnetic fields in the outer gap are too weak for one-photon pair production to play a significant role in the cascade. Instead, the curvature photons undergo photon-photon pair production with thermal X-rays from the polar caps (as in the case of Vela or Geminga) or nonthermal X-rays produced by the cascade pairs (as in the Crab). In the Crab pulsar, the gamma-ray emission is produced by synchrotron

self-Compton radiation from the pairs, with the lower energy part of the spectrum dominated by the synchrotron component and the higher energy spectrum dominated by inverse Compton. In Vela-like pulsars with periods around 0.1 s, the gaps are at larger radii, the gap magnetic fields are weaker, and the pairs do not radiate much synchrotron emission. The spectra of these pulsars will be primarily curvature radiation.

The gamma-ray pulse profiles from the outer gaps are considerably more complicated than those from the polar caps. Detailed pulse profiles from cascade calculations in the outer gap model have not been computed, but the profiles based on a simple outer gap geometry have been studied (Chiang and Romani, 1992; Romani and Yadigaroglu, 1995). Most of the observed radiation comes from the two outer gaps closest to the star, as shown by the shaded regions in Figure 7.1. The emission from gap accelerated particles is beamed along the magnetic field lines bordering the closed field regions and is thus radiated in a fan beam having a larger solid angle than polar cap emission. Because electron-positron pairs are accelerated in opposite directions in the gap, an observer, depending on orientation, may see one or both beams of radiation from one or both gaps. Thus, up to four peaks in the pulse profile may be visible. Two peaks with phase separation of 140° is fairly easy to generate.

7.3.4 Energetics and Efficiency

There are presently six known gamma-ray pulsars detected by CGRO, but one of these, PSR1509-58, is detected only at energies below about 4 MeV. Another of these, Geminga, identified as a pulsar by ROSAT (Halpern and Holt, 1992) at the location of the mysterious COS-B point source of that name and confirmed by EGRET (Bertsch et al., 1992), is radio-quiet. The five pulsars detected above 100 MeV are listed in Table 7.1 with their maximum luminosity as derived from the observed flux by assuming an emission solid angle of 1 steradian. Also shown is the gamma-ray efficiency, $\eta_\gamma(>100\ MeV) \equiv L_\gamma(>100\ MeV)/\dot{E}$ or the fraction of the dipole spin-down power, $\dot{E}$, radiated in pulsed gamma-rays. Unfortunately, our lack of knowledge of the emission solid angle introduces an uncertainty of possibly an order of magnitude or more in the actual gamma-ray luminosity as well as in other quantities derived from it, such as in the gamma-ray efficiency. Despite this uncertainty, there is a striking trend among the observed gamma-ray pulsars that the efficiency increases with apparent age, $\tau = P/2\dot{P}$. This trend is actually expected in polar cap models because the gamma-ray luminosity should be proportional to the current of primary particles from the polar caps which flow out in the open field line region,

$$\dot{N}_p \simeq 2\rho_e \pi R_p^2 c/e = \frac{\Omega^2 R^3 B_o}{ec} = 2.7\times10^{30}\,\mathrm{s}^{-1}\,B_{12}P^{-2} \qquad (7.17)$$

where ρ_e is the corotation charge density, $R_p = R(\Omega R/c)^{1/2}$ is the polar cap radius. Values of $\dot{N}_p$ for the known gamma-ray pulsars are listed in Table 7.1. The total power available for pulsed gamma-rays is $L_{\rm tot} \simeq \dot{N}_p E_p$, where E_p is the energy to which primary particles are accelerated. Since $\dot{E} \propto B^2P^{-4}$ Equation (7.1), then $\eta_\gamma \propto \tau^{3/2}\dot{P}$, if E_p is assumed to be only weakly dependent on B and P and $B \propto (P\dot{P})^{1/2}$ Equation (7.5). This dependence is somewhat modified by magnetic pair production attenuation of the high energy portion of the gamma-ray spectrum and the predicted efficiency (Harding, 1981) is $\eta_\gamma \propto \tau^{1.8}\dot{P}^{1.3}$. Figure 7.3 shows that the five high-energy gamma-ray pulsars do fit this relation reasonably well, although there is some uncertainty in the distance to Geminga, as well as possible dispersion in the emission solid angle. (A solid angle of 1 sr has been assumed for all of the pulsars.)

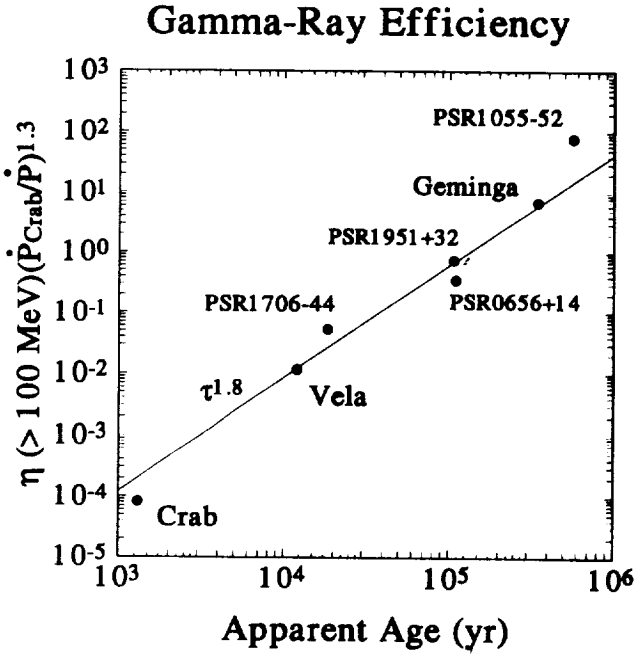


Figure 7.3: Efficiency vs. apparent age for known gamma-ray pulsars.

Requirements on the energy to which the primary particles must be accelerated come from comparing $L_{\rm tot} \simeq \dot{N}_p E_p$ to the observed gamma-ray luminosities. For an emission solid angle of 4π, primary energies must be at least several hundred TeV, which is much larger than the estimated polar cap potential drops but which is

Table 7.1: Observed Gamma-Ray Pulsars

PSR	P (s)	$\dot{E}$ (erg/s)	L^a_{max} (erg/s)	η_γ	N_p (s^{-1})
Crab	0.033	4.8×10^{38}	3.9×10^{34}	8.2×10^{-5}	1×10^{34}
Vela	0.089	7.2×10^{36}	1.7×10^{34}	2.3×10^{-3}	1.4×10^{33}
1706-44	0.102	3.6×10^{36}	2.6×10^{34}	7.3×10^{-3}	8×10^{32}
1951+32	0.039	3.7×10^{36}	1.4×10^{34}	4×10^{3}	1.7×10^{33}
Geminga[b]	0.237	3.5×10^{34}	2.2×10^{33}	6.3×10^{-2}	8×10^{31}
1055-52	0.197	3.1×10^{34}	9.2×10^{33}	0.30	7.3×10^{31}

[a]From observed flux $> 100\ MeV$(Thompson et al., 1994), assuming 1 sr solid angle

[b]Distance of 100 pc is assumed

achievable in outer gap models. However, polar cap models would predict a higher degree of beaming and only tens of TeV are required if the solid angle is 1 *steradian*.

7.4 POLAR CAP HEATING AND THERMAL X-RAY EMISSION

In models where high-energy gamma-rays produce pair cascades, some fraction of these particles will be accelerated toward the stellar surface and will heat the polar cap. Thus, this process will add to any other heating or cooling that may be present to produce thermal X-rays. The contribution from polar cap heating has been estimated for both polar cap and outer gap accelerators and can be compared to observed X-ray fluxes and can also be extrapolated to make predictions for fluxes from millisecond pulsars.

In the polar cap model where electrons are accelerated along favorably curved magnetic field lines, the break down of $E_\parallel$ occurs in a thin "pair formation front," above which $\mathbf{E} \cdot \mathbf{B} = \mathbf{0}$. The fraction of the total number of positrons produced by the accelerated primary electrons that are produced in the pair formation front and which can be accelerated back toward the neutron star surface is small. Consequently, the energy going into polar cap heating in this model is smaller than the energy radiated directly in nonthermal emission. Arons (1981a) estimates that the X-ray luminosity from polar cap heating is

$$L_x^{pc} \simeq 2 \times 10^{26}\, P^{-27/8}\, B_{12}\, erg\ s^{-1} \quad P > P_{1D}$$
$$L_x^{pc} \simeq 1 \times 10^{29}\, P^{-9/5}\, B_{12}^{2/5}\, erg\ s^{-1} \quad P < P_{1D} \quad (7.18)$$

where $P_{1D} = .03\ B_{12}^{8/21}$ is the period below which a one-dimensional model of the acceleration zone may be used. Arons (1981b) finds that positron trapping may be important in pulsars with weaker magnetic fields and longer periods, which would significantly increase the estimated luminosities in Equation (7.18).

Electrons and positrons produced in cascades in the outer gaps are accelerated in opposite directions along open field lines that are connected to the polar caps. Because the outer gap acceleration is radiation reaction-limited, half of all of the particles produced in the gaps are accelerated toward the star and will heat the polar caps. Therefore, the X-ray luminosity due to polar cap heating can be comparable to the gamma-ray luminosity. An estimate of the X-ray luminosity from polar heating in the outer gap model is given by (CHR; Halpern and Ruderman, 1993)

$$L_x^{og} \simeq E(R)\, \dot{N}_g \simeq 2 \times 10^{31}\, P^{-5/3}\, B_{12}\, erg\ s^{-1} \quad (7.19)$$

where $\dot{N}_g$ is the maximum pair flux directed toward the polar cap from an outer gap accelerator (basically the Goldreich-Julian current) and $E(R_o)$ is the energy per particle at the polar cap after curvature radiation losses. $E(R_o)$ depends on r_{min}, the radius at the inner edge of the gap which decreases with increasing α. An average value of $r_{min} = 0.5\, R_{LC}$ was assumed to obtain Equation (7.19). However, the polar cap heating luminosity may be somewhat smaller than the estimate given in Equation (7.19) since it also will depend on the pitch angles of the particles as they come out of the accelerator.

Table 7.2 compares the predicted luminosities for the polar cap and outer cap models to the observed X-ray luminosities of those pulsars with pulsed thermal emission. The general trend is for the predicted outer gap luminosities to be higher than observed and for the predicted polar cap luminosities to be too low. However, the polar cap (as well as the outer gap) luminosity is higher than observed for the single millisecond pulsar which has detected X-ray emission, J4037-4715. In general, X-ray luminosity due to sources other than polar cap heating

Table 7.2: Predicted X-rays from Polar Cap Heating

	X-RAY LUMINOSITY ($erg\ s^{-1}$)		
	Observed[a]	Predicted	
PSR		Polar Cap[b]	Outer Gap[c]
0833-45	5×10^{31}	4×10^{30}	4×10^{33}
0656+14	6×10^{32}	3×10^{28}	4×10^{32}
Geminga[d]	8×10^{30}	6×10^{28}	3×10^{32}
1055-52	4×10^{33}	7×10^{28}	3×10^{32}
1929+10[e]	1×10^{29}	2×10^{28}	1×10^{32}
J0437-4715	$1 - 3 \times 10^{30}$	7×10^{31}	9×10^{31}

[a] from Ogelman (1994), and references therein
[b] Arons (1981a) and Equation (7.18)
[c] Halpern and Ruderman (1993) and Equation (7.19)
[d] assuming a distance of 100 *pc*
[e] Yancopoulos and Helfand (1994)

will be present. Thermal emission from initial cooling of the neutron star and emission due to internal heating from friction between the neutron star superfluid and crust (Shibazaki and Lamb, 1989; Umeda et al., 1993) both decrease with increasing age. The observed X-ray emission from several of the younger pulsars (Vela and 0656+14) could be due primarily to initial cooling (Ogelman, 1994). The observed *ratio* of X-ray to gamma-ray luminosity of Geminga is explained very well by Arons' (1981a) polar cap heating model (Harding et al., 1993), but to produce the X-ray luminosity from polar cap heating alone requires its distance to be around 40 *pc*. Fits to spectra of magnetized atmospheres (Meyer et al., 1994) (which deviate from blackbody spectra) are compatible with Geminga being at such distances, much closer than the distance limits derived from blackbody fits (Halpern and Ruderman, 1993). The luminosity observed from J0437-4715, whose age is 8×10^8 *yr*, is well above that predicted for both initial cooling or for internal frictional heating and so must be due to polar cap heating and/or magnetospheric emission.

7.5 SUMMARY

It was hoped that with the improved sensitivity of CGRO and the detection of more gamma-ray pulsars we would be able to identify the site(s) of particle acceleration in pulsar magnetospheres. So far, this has not been possible; primarily, because attempting to explain the very diverse gamma-ray and multiwavelength behavior of the observed pulsars is difficult with either model. In addition, several features of the observed high energy emission can be explained in either model.

Both models are capable of producing almost any phase separation of the pulses, including 140° for the Crab and Vela, as well as interpulse emission. However, in the outer gap model one would not expect to see a 180° phase separation by chance as in the case of Geminga. It is also possible to observe as many as four pulses from the outer gaps for some observer orientations, so it may be surprising that all the pulsars observed so far show no more than two.

Both models can predict high energy spectral turnovers, although the location of the turnover in the outer gap spectra can be moved up or down more freely by changing free parameters. The outer gap model spectrum for the Crab pulsar predicts a dip between the inverse Compton and synchrotron portions of the gamma-ray spectrum, which has not been observed. However, synchrotron radiation from the outer gap can account for the observed Crab optical emission, whereas synchrotron radiation from the inner magnetosphere where the cyclotron energy is well above optical, cannot. The outer gap predicts an order of magnitude higher ratio of X-ray to gamma-ray luminosity due to polar cap heating than is observed in Geminga (Halpern and Ruderman, 1993). On the other hand, the polar cap model of Arons (Arons, 1981) predicts $L_x/L\gamma \simeq 10^{-3}$ for Geminga (Harding, Ozernoy, and Usov, 1993), about what is observed.

There is convincing evidence that radio emission in pulsars occurs near the polar cap region. For example, the sweep of polarization position angle through the pulse profile displays the characteristic "S" shape expected for an observer seeing emission from near the magnetic pole. Detailed studies of radio pulse morphology and polarimetry (Lyne and Manchester, 1988; Rankin 1990) find that the pulse widths exhibit the dependence on period expected in the polar cap model with remarkable regularity. It is not so clear that the gamma-ray emission observed from some of these pulsars also comes from the polar cap cascade. However, both peaks of the optical pulse profile of the Crab pulsar show a polarization position angle sweep of nearly 180°. One might argue that the gamma-ray pulses are produced near the magnetic pole since they are coincident in phase with the optical pulses. In other pulsars like Vela and PSR1706-44, though, the radio pulses are not coincident with the gamma-ray pulses.

In the present situation, we need both better observations of high-energy emission from pulsars, e.g., detection of more gamma-ray pulsars and better measurements of the spectra and pulse profiles of the known pulsars, and more detailed theoretical predictions, e.g., calculations of phase-resolved spectra and statistics on numbers and profile types expected. It is becoming clearer that the emission geometry may be the deciding test for the models and the most accurate probe of the radiation physics and, ultimately, the acceleration process.

7.6 REFERENCES

Alpar, M.A., Cheng, A., Ruderman, M.A., and Shaham, J., 1982, Nature, 300, 728.

Arons, J., 1981a, ApJ, 248, 1099.

Arons, J., 1981b, in IAU Symposium No. 125, "The Origin and Evolution of Neutron Stars" eds. D.J. Helfand, and J.H. Huang, Reidel: Dordrecht, p. 207.

Arons, J., 1983, ApJ, 266, 215.

Arons, J., and Scharlemann, E.T., 1979, ApJ, 231, 854.

Bertsch, D.L., et al., 1992, Nature 357, 306

Chen, K., and Ruderman, M.A., 1993, ApJ, 402, 264.

Cheng, K.S., Ho, C., and Ruderman, M.A., 1986, ApJ, 300, 500.

Chiang, J., and Romani, R.W., 1992, ApJ, 400, 629.

Dermer, C.D., 1989, ApJ, 347, L13.

Dermer, C.D., 1990, ApJ, 360, 197.

Dermer, C.D., and Sturner, S.J., 1994, ApJ, 420, L75.

Daugherty, J.K., and Harding, A.K., 1982, ApJ, 252, 337.

Daugherty, J.K., and Harding, A.K., 1983, ApJ, 273, 761.

Daugherty, J.K., and Harding, A.K., 1989, ApJ, 336, 861.

Daugherty, J.K., and Harding, A.K., 1994, ApJ, 429, 325.

Erber, T., 1966, Rev. Mod. Phys., 38, 626.

Flowers, E.G., Lee, J.F., and Ruderman, M.A., 1977, ApJ, 215, 291.

Foster, R.S., Fischer, J., Johnson, K.J., and Grove, J.E., 1994, preprint.

Gold, T., 1968, Nature, 218, 731.

Goldreich, P., and Julian, W.H., 1969, ApJ , 157, 869.

Growth, E.J., 1975, ApJ, 200, 278.

Gunn, J.E., and Ostriker, J.P., 1969, Nature, 221, 454.

Halpern, J.P., and Holt, S.S., 1992. Nature, 357, 222.

Halpern, J.P., and Ruderman, M.A., 1993, ApJ, 415, 286.

Harding, A.K., 1981, ApJ, 245, 267.

Harding, A.K., 1993, Ann. NY Acad. Sci., 688, 573.

Harding, A.K., Ozernoy, L.M., and Usov, V.V., 1993, MNRAS, 265, 921.

Hewish, A.S., Bell, J.S., Pilkington, J.D.H., Scott, P.F., and Collins, R.A., 1968, Nature, 217, 709.

Ho, C., 1989, ApJ, 342, 396.

Kanbach, G., et al., 1994, A and A, 289, 855.

Lyne, A.G., and Graham-Smith, F., 1990, in "Pulsar Astronomy" (Cambridge Univ. press).

Lyne, A.G., and Manchester, R.N., 1988, MNRAS, 234, 477.

Manchester, R.N., Durdin, J.M., and Newton, L.M., 1985, Nature, 313, 374.

Manchester, R.N., and Taylor, J.H., 1977, in "Pulsars" (San Francisco: Freeman).

Meyer, R.D., Pavlov, G.G., and Meszaros, P., 1994, ApJ, 433, 265.

Michel, F.C., 1991, in "Theory of Neutron Star Magnetospheres" (Univ. of Chicago Press).

Muslimov, A.G., and Tsygan, A.I., 1992, MNRAS, 255, 61.

Ogelman, H., 1994, in "Procedings Kemer NATO ASI: Lives of Neutron Stars" eds. J. Van Paradijs, and A. Alpar, Kluwer, in press.

Rankin, J.M., 1990, ApJ, 352, 247.

Romani, R.W., and Yadigaroglu, I.-A., 1995, ApJ, 438, 314.

Ruderman, M.A., and Sutherland, P.G., 1975, ApJ, 196, 51.

Shapiro, S.L., and Teukolsky, S.A., 1983, in "Black Holes, White Dwarfs and Neutron Stars" (New York: J. Wiley and Sons).

Shibazaki, N., and Lamb, F.K., 1989, ApJ, 346, 808.

Sturner, S.J., and Dermer, C.D., 1994, ApJ, 420, L79.

Sturrock, P.A., 1971, ApJ, 164, 529.

Thompson, D.J., et al., 1995, ApJ, 436, 229.

Thompson, D.J., 1996, in "AGN and Pulsars: High Energy Gamma-Ray Beams" Chapter 9, of this book.

Umeda, H., Shibazaki, N., Nomoto, K., and Tsuruta, S., 1993, ApJ, 408, 186.

Yancopoulos, S., Hamilton, T.T., and Helfand, D.J., 1994, ApJ, 429, 832.

Chapter 8

X-RAY EMISSION AS A TRACER OF MATTER

KEITH JAHODA
Laboratory for High Energy Astrophyics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

X-rays can be used to trace the distribution of matter within the local interstellar medium, the galaxy, and the universe at large. This chapter reviews the characteristic, and energy dependent, interaction path lengths for X-rays and provide several examples of how the X-ray surface brightness constrains the distribution of hot gas responsible for the soft X-ray background and how more energetic X-rays trace the distribution of matter in the nearby universe.

8.1 INTRODUCTION

The purpose of this chapter is to provide several examples of how X-rays can be used to trace the distribution of matter. The scales that are accessible are a function of both the energy band observed, the solid angle covered, and the resolution of the experiment. The discussion is restricted to measurements of X-ray surface brightness (rather than catalogs of sources). The abbreviation SXRB for Soft X-Ray Background refers to surface brightness measurements in the band below 0.28 keV and the abbreviation CXB for Cosmic X-ray Background refers to surface brightness measurements in the band above 2 keV.

This chapter is divided into several sections. Section 8.2 provides an operational definition of an X-ray and examines the scales which are accessible to different energies. Section 8.3 presents maps of the X-ray background in three energy ranges in order to emphasize how different the "X-ray sky" appears. Section 8.4 gives examples of several topics that are related largely as distinct methods which nonetheless all use X-ray emission to trace the distribution of matter.

8.2 DEFINITION OF AN X-RAY

What is an X-ray? I adopt a practical definition related to the interaction between photons and matter: if photoelectric absorption is the primary interaction, then that photon is an X-ray. This definition has the disadvantage of depending on the type of absorbing matter, but in general the condition is met for photons between 0.05 and 100 keV. Above about 100 keV, Compton scattering becomes the dominant interaction. Figure 8.1, reproduced from Evans (1955), shows the energy at which photoelectric and Compton interactions have equal cross-section, as a function of z. Figure 8.1 illustrates the energies at which photoelectric and Compton cross sections are equal, but does not display the magnitude of the cross section, or give any indication of the range. Figure 8.2a, reproduced from Morrison and McCammon (1983), shows the absorption cross section per hydrogen atom, under the assumption that the relative abundances of the elements are as measured in the local interstellar medium. Note that the vertical scale is multiplied by E^3, removing the strong energy dependence. There are sharp jumps at the K- and L- absorption edges of elements that are abundant in the interstellar medium. From such data, we can estimate the unit optical depth, measured in atoms, as a function of energy. Figure 8.2b gives the X-ray range as a function of energy, under the assumption that the path traverses regions with the mean interstellar density of 1 atom cm^{-3}. At 1 keV, the mean free path is < 1 kpc while at 2 keV the range is 8 kpc, about the distance to the center of the galaxy. At energies substantially below 1 keV it is, therefore, not possible to see out of our own galaxy, while at energies above 2 keV, the galaxy is largely transparent. The opacity of the galaxy, which varies substantially over the X-ray band, controls the range over which X-rays are a useful probe of the distribution of matter.

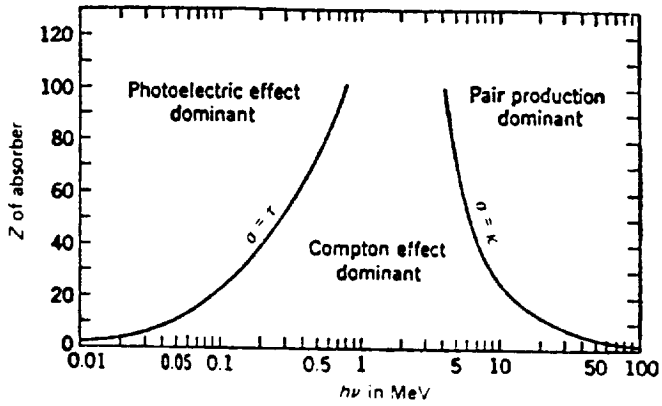


Figure 8.1: Dominant interaction processes vs. absorber Z and energy.

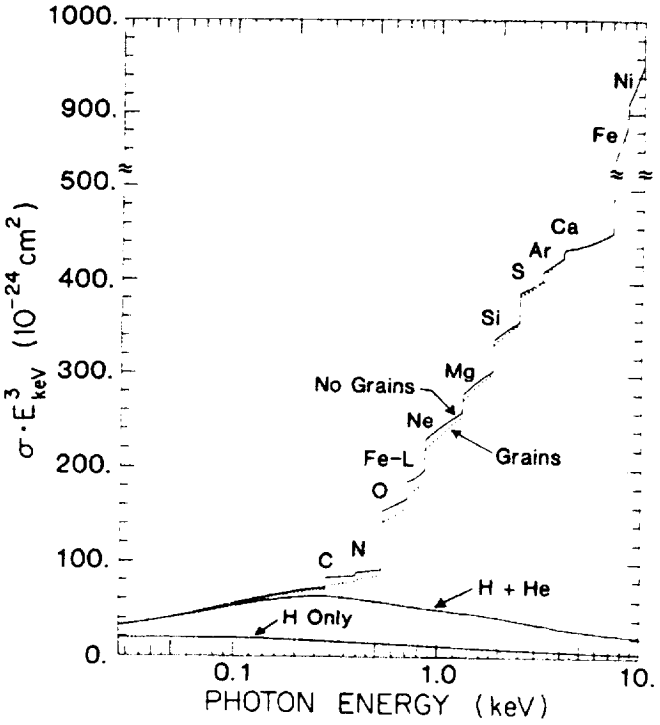


Figure 8.2a: Interstellar cross section vs. energy.

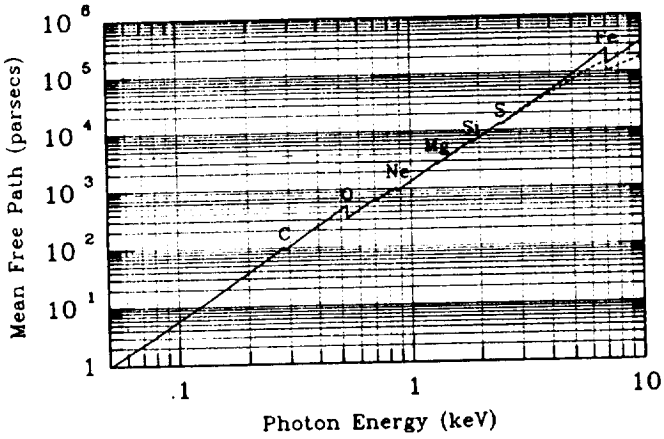


Figure 8.2b: Interstellar range vs. energy.

8.3 VIEWS OF THE SKY

8.3.1 The Soft X-Ray Background

The observed two-dimensional structure of the SXRB has been used to infer details of the three-dimensional organization of the local interstellar medium.

Figure 8.3 (McCammon and Sanders, 1990) shows the soft X-ray background in a band that covers 0.1-0.28 *keV*. It has been observed for a long time that the SXRB is anticorrelated with the neutral HI in the galaxy (McCammon et al., 1983 present the HI data). The simplest explanation is that we are observing an extragalactic source of the SXRB and absorption by intervening and local material is responsible for the anti-correlation

This explanation remained popular for two decades despite several difficulties. At a minimum this explanation is incomplete; the large column densities in the plane of the galaxy, and the nonzero X-ray flux observed at low latitudes, require that at least some X-rays are produced locally. It is simply not possible to observe an extragalactic source at low latitude in this band.

Second, at high latitude, the variation of X-ray surface brightness is not as strong as predicted by the variation in measurements of the HI column density. Although the HI is used as a tracer of all matter, and the number of helium, and heavier, atoms must be estimated as a function of the number of hydrogen atoms (and similarly for any molecules, including H_2), the band below 0.28 *keV* is particularly simple. Only Hydrogen and Helium are significant absorbers. Since the Helium number density is well known to be $\sim 1/10$ of the Hydrogen number density (from nucleo-synthesis arguments), the predicted absorption cross section is simple to calculate. Because the apparent variation of X-ray brightness is not as strong as even the absorption accounted for by H and He, ignoring the possible presence of H_2 or other atomic or molecular species can only make the problem worse. Before the launch of the ROSAT satellite, there was limited evidence that any individual feature observed in the HI maps of the sky cast a shadow in the SXRB. (But see the section on shadows in the SXRB below.)

8.3.2 Intermediate Energies

Figure 8.4, also from McCammon and Sanders (1990), shows the X-ray sky in the 0.5-1.0 *keV* band. The sky looks quite different from the lower energy band. Immediately obvious is a large-scale enhancement at the center of the map (which is coincident with radio loop I (Berkhuijsen et al., 1971)). Outside of this enhancement, the map is surprisingly uniform, particularly as a function of galactic latitude. Although the mean free path in this band is expected to be ~ 1 *kpc* for typical galactic densities, there is no hint of the galactic plane, either as a source of X-ray emission or as casting a shadow on an extragalactic source. Though puzzling, this result seems to require that any galactic emission (which must exist as stars emit in this band (Schmitt and Snowden, 1990)), which should be enhanced at low latitude where the path length through the galactic disk is large must be balanced, nearly exactly, by absorption of extragalactic flux by neutral material in the plane.

In this band, we turn around the argument used below 0.28 *keV*: the presence of emission at high latitudes requires that at least some emission is extragalactic, as there is no plausible source from stars or hot gas that can produce the entire observed flux.

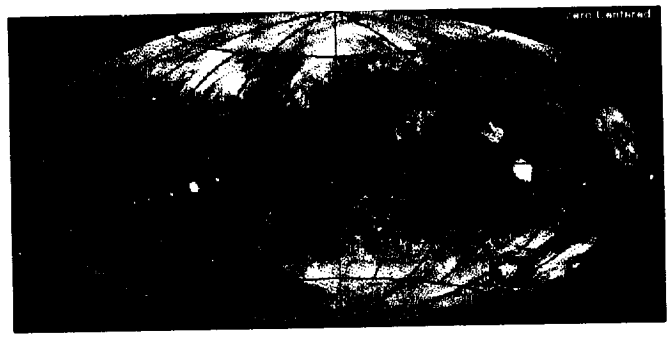


Figure 8.3: The X-ray Sky in the 0.1-0.28 *keV* band. The sky is variable and largely anticorrelated with neutral hydrogen column density.

8.3.3 High Energies

Figure 8.5 shows the X-ray background observed by the HEAO-1 A2 experiment (Boldt, 1987; Allen et al., 1994). The effective band for this image is 2-20 *keV*, and the sky is different again. At these energies, the galaxy is virtually transparent to X-rays. The striking features of this map are the uniformity of emission at high lati-

tudes, the presence of a ridge of emission along the galactic plane, and a handful of bright individually unresolved sources. (The sources are a mixture of galactic binaries, supernova remnants, and stars and extragalactic sources; galaxies, active galaxies, and clusters (Piccinotti et al., 1982; Wood et al., 1984)). The uniformity of the high latitude emission argues that the sky is dominated by an extragalactic source (Shafer, 1983: Boldt, 1987), which is now believed to be a combination of individually unresolved point sources (Fabian and Barcons, 1992). The ridge of emission along the galactic plane is believed to be galactic in origin, and while much is known about many classes of sources which must be partial contributors, the complete picture is still unknown.

8.4 SEVERAL TOPICS

8.4.1 Shadows in the SXRB

The X-ray band below 0.28 keV, first observed in the late 1960s (Bowyer et al., 1968; Bunner et al., 1969), has long been known to be anticorrelated with the column density of neutral hydrogen as inferred from measurements of 21 cm emission. The anticorrelation remained striking even after the whole sky was mapped by several experiments (McCammon et al., 1983; Marshall and Clark, 1984; Garmire et al., 1992) and was a primary motivation of the absorption model in which the bulk of the SXRB arises as an extragalactic or coronal source which is absorbed by the neutral galactic material.

Figure 8.4: The X-ray sky in the 0.5-1.0 keV band. Outside of the Loop I enhancement there is very little structure associated with galactic latitude.

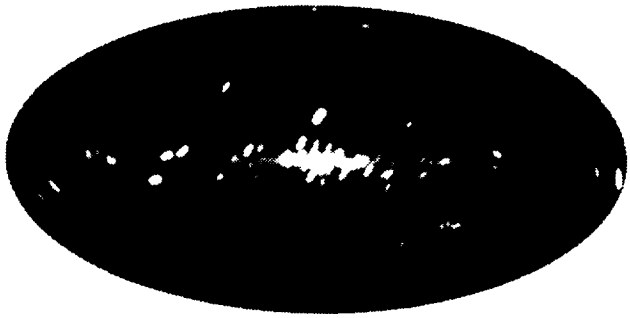


Figure 8.5: The X-ray sky in the 2-20 keV band. The sky is largely uniform, with emission from the galaxy and several point sources detectable.

There were several difficulties with this simple picture: 1) The SXRB flux does not fall to zero in the galactic plane, as would be expected if an extragalactic source were responsible; 2) The variation in intensity with increasing column density is not as strong as predicted from photo-electric cross sections; 3) There was no clear evidence of any shadows cast by galactic objects with substantial column density.

Additional features were added to the absorption model to overcome these drawbacks. The hypothesis that there is also a local source component, arising in the hot gas left by a recent supernova, allows a nonzero flux at low latitude. The predicted absorption cross section of neutral material can be reduced under the assumption of clumping of the absorbing medium, which results in a screening of some of the absorbing material. And the lack of detailed shadows can be hidden by variations in both the local and distant components which are uncorrelated with the galactic absorber. Searches for the required clumping, however, via detailed mapping of the 21 cm brightness distribution (Jahoda et al., 1985; 1986) failed to detect the required clumpiness and the requirements that variations in emission conspire to hide variations in absorption seemed contrived.

By the launch of the ROSAT satellite in June of 1990, most researchers in this field believed it was likely that greater than 90% of the soft flux originated from the near side of most of the absorbing material (see review by McCammon and Sanders, 1990). Nonetheless, ROSAT provided an unprecedented ability to produce detailed images of fields 2 degrees in diameter (or larger using survey data and the mosaicing techniques of Snowden and Petre (1994)).

Early in the ROSAT mission two independent investigations of the Draco region provided evidence of shadows in the SXRB cast by neutral material which was known to be 600 pc distant and >400 pc above the plane of the galaxy (Burrows and Mendenhall, 1991; Snowden et al., 1991). Figure 8.6 is from Burrows and Mendenhall. The gray scale image shows SXRB brightness while the contours represent the brightness at 100 μm, measured by the IRAS satellite, and a good tracer of neutral material. The 100 μm emission itself arises from the glow of interstellar dust. It is immediately apparent that at this scale, and in this direction, there is a shadow which cuts out 50% or more of the SXRB. This leads to the conclusion that we are indeed observing, on either side of the dust finger, emission from beyond the neutral material.

It is unlikely that the distant emission represents a

truly cosmological background as the extragalactic emission must be patchy. This requirement comes from assuming that the derived distant flux is typical and putting that flux behind the galactic absorbing layer in all directions. This results, however, in a prediction of too much SXRB flux (i.e. more than is observed) in many directions.

Shadowing observations have now been made in numerous (~30) directions. In each direction where the distance to the absorbing material can be estimated and where the depth of a shadow can be measured (including a null result) the fraction of the emission arising in front of and behind the cloud can be measured. In this way a three-dimensional map of the important emission regions can be constructed. Though the wide field of view

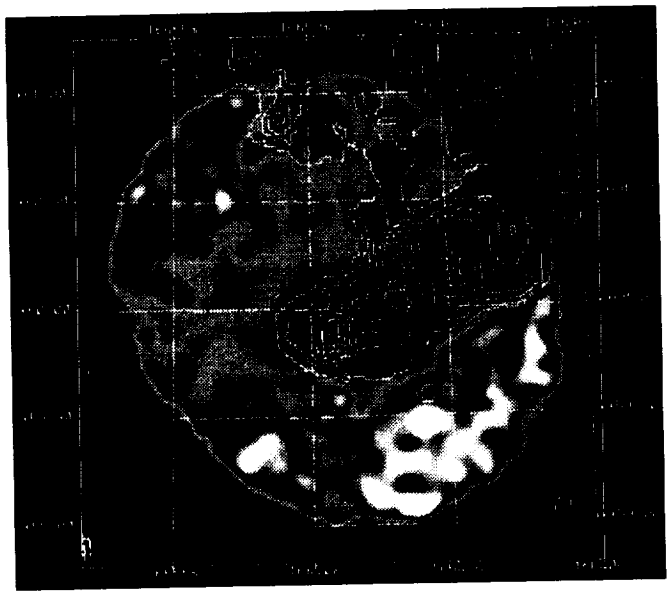


Figure 8.6: The X-ray shadow in Draco. The contours represent 100 μm surface brightness while the gray scale gives SXRB brightness.

instrument aboard ROSAT (the Position Sensitive Proportional Counter) is no longer usable, most of the suitable shadowing targets (i.e. substantial changes in optical depth occurring over a spatial scale smaller that the 2° diameter field of view) have been observed. While the simplest model for the origin of the SXRB, the absorption model, remains too contrived, it is now known that there is a distant contribution in at least some directions.

The study of the location of the source of the SXRB emission is most nearly parallel to the vernacular use of the verb "X-ray"; a shadow or an absence of expected flux is used to infer absorption and thus locate the emission beyond the shadowing object. This is possible because an optical depth at these energies is comparable to the column density of dense objects which happen to reside nearby in the galaxy. Although shadowing is physically possible at higher energies, nature does not provide thick enough targets.

8.4.2 CXB as a Tracer of Matter

Having demonstrated that the galaxy is transparent to radiation with energy >2 keV, and concluding from the basic isotropy of the observed surface brightness that the radiation is primarily extragalactic, deviations from isotropy can be used to study the distribution of matter in the local universe. There are several categories of possible anisotropy: bumps, microscopic, and macroscopic. Bumps are individual features; microscopic anisotropies are statistical variations; and macroscopic anisotropies (which could include bumps) are large angular scale, perhaps entire sky, anisotropies. Examples of each follow.

Bumps

Figure 8.5 above shows that the most prominent feature in the >2 keV sky, after the large scale isotropy, is the ridge of emission associated with the galaxy. Ignoring this feature, if possible, one can attempt to ask what is the largest anisotropy not obviously related to some feature of the galaxy? On a more limited scale one can search for emission that is correlated with some other feature of the known universe. In the late 1980s, the discovery of the Great Attractor, an overdensity believed to make a significant contribution to the peculiar velocity of the Local Group of Galaxies, motivated such a search. The search was based on the assumptions that (a) light, and in particular X-ray light, is a tracer of matter, and (b) that the Great Attractor is not a point mass but spread out over a substantial solid angle in the sky.

Figure 8.7 shows the X-ray sky in the vicinity of the Great Attractor, again using data from the A2 experiment. The coordinates are ecliptic, the easiest for describing the shape of the beam, and the dashed contour outlines the redshift survey of Dressler (1988) which showed a peak near 4000 $km\ s^{-1}$. The area is centered on the nominal direction of the Great Attractor identified by Lynden- Bell et al. (1988) at l = 307, b = 9 (though note that the precise center is not well determined: Faber and Burstein (1989), using the same galaxy velocity data, derive a position of l = 309, b = 18). Although numerous point sources are visible in Figure 8.7, and all of the sources with 2 or more closed contours are identified, very few have a redshift that allows an identification as part of the Great Attractor (Jahoda and Mushotzky, 1989). This object, which has an inferred mass equal to a supercluster, could not be identified by the usual means, as it has no clusters of galaxies.

Jahoda and Mushotzky (1989) searched for a diffuse enhancement by subtracting from each pixel within the Great Attractor region the pixel exactly 180 degrees

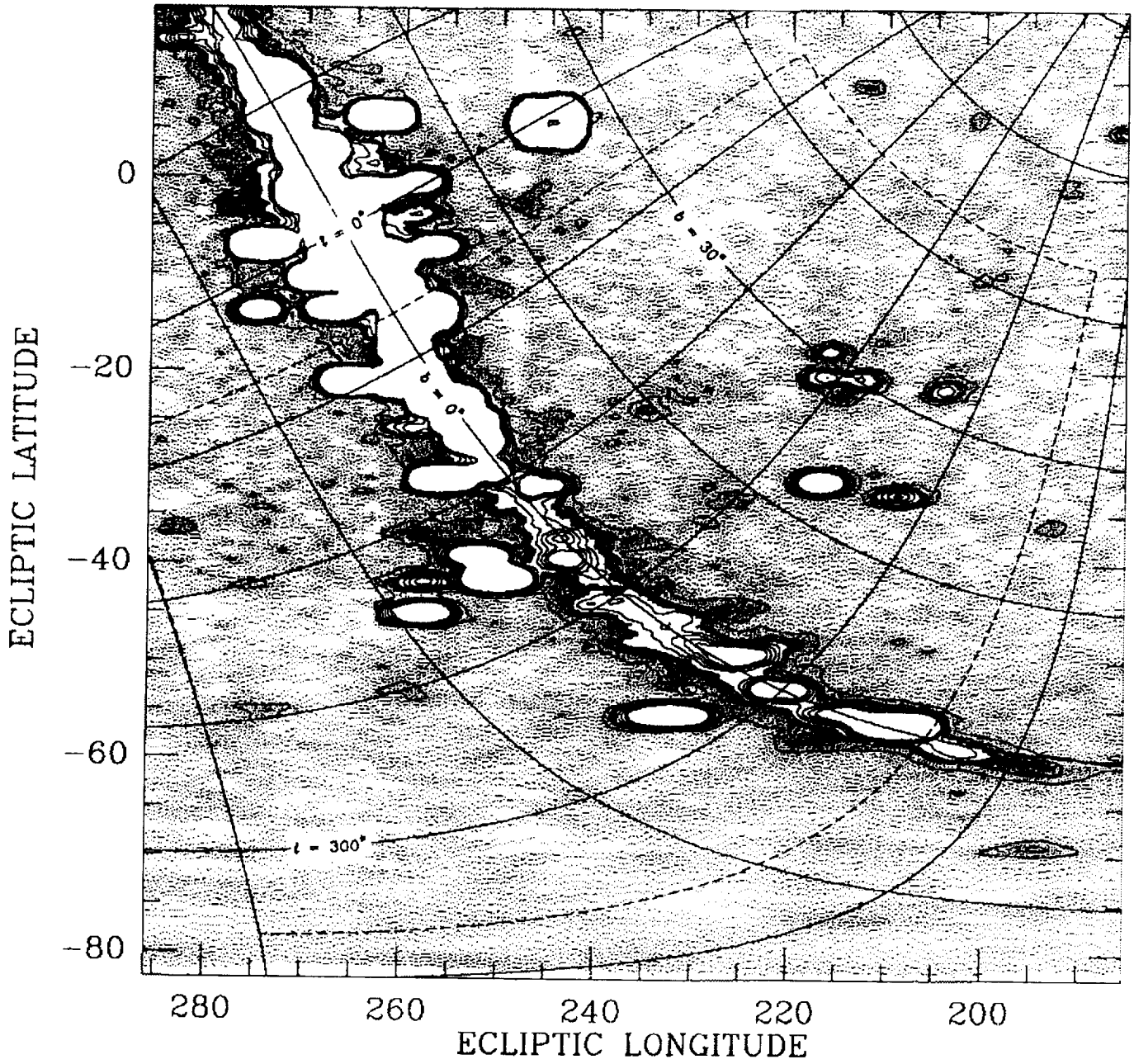


Figure 8.7: X-ray surface brightness in the Great Attractor region. The dashed contour outlines the redshift survey of Dressler (1988) and is approximately the boundary of the Great Attractor.

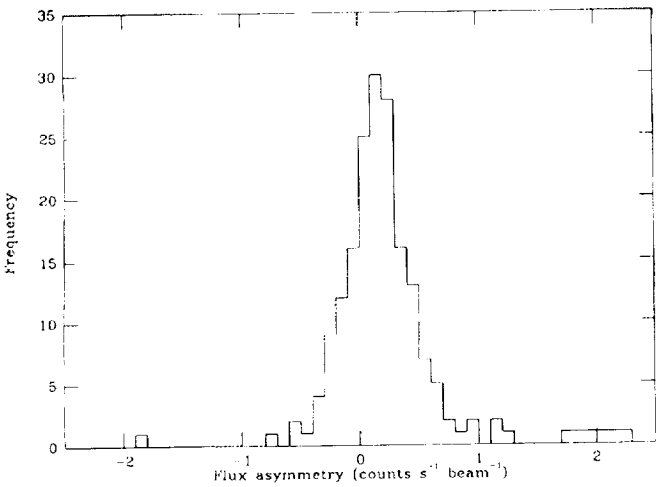


Figure 8.8: Flux assymetry in the 400 square degrees surrounding the Great Attractor.

away. For a completely uniform sky, this would produce a distribution with 0 mean, and width governed by counting statistics. For the actual sky, which contains point sources, the distribution has large outliers at positive and negative fluxes (corresponding to the sources) and a decidedly nonzero mean as shown in Figure 8.8, reproduced from Jahoda and Mushotzky (1989). This evidence alone is not enough to convince one that the enhancement is not galactic. However, this enhancement is the largest anywhere on the sky when the same technique is repeated; in particular it is larger than the offsets obtained from other points that are at similarly located (reflections through l = 0, b = 0 or both) with respect to the likely symmetry axes of the galaxy.

Under that assumption that the excess flux is associated with the Great Attractor, Jahoda and Mushotzky derive that 25% of the mass of the Great Attractor could exist in the form of hot baryons in equilibrium with the gravitational potential. This seems consistent with expectations based on the fraction of cluster mass which appears as baryons.

Microscopic Anisotropies

Figure 8.9 shows the 2-20 keV map, although with a greatly enhanced contrast. This emphasizes that while the sky is often described as "isotropic," there remain small fluctuations that represent cosmic (rather than measurement) variations. The study of the distribution and magnitude of these fluctuations, believed to arise from sources that are not individually resolved, can yield insights into the source population.

There are many statistical measures that might give insight into the underlying distribution of sources of the XRB such as the autocorrelation function (Danese et al., 1993) or fluctuation analysis (Shafer, 1983; Hayashida, 1989). This chapter illustrates only one: correlations between the CXB and another population of objects.

The hypotheses here are quite simple. Since it is believed (a) that the CXB results from the combined emission from many individually unresolved sources, and (b) since objects such as galaxies are known to both emit X-rays themselves and trace the underlying matter distribution, it should be true that a complete survey of galaxies is correlated with the CXB, either because the galaxies themselves are X-ray emitters or because the galaxies are distributed like X-ray emitters. Turner and Geller (1980) give an elegant description of the problem although they do not detect a correlation.

More recent efforts have detected correlations between the CXBbrightness distribution and optically selected galaxies (Jahoda et al., 1991; 1992; Carrera et al., 1995) IRAS selected galaxies (Lahav et al., 1993; Miyaji

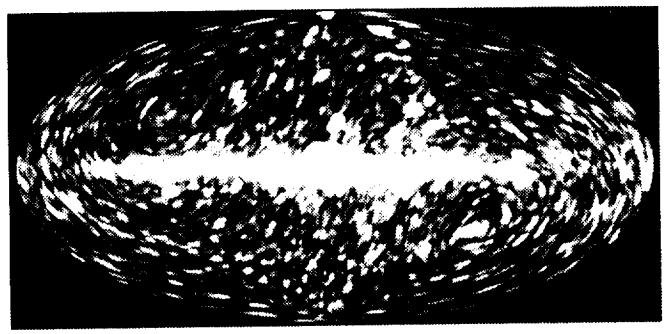


Figure 8.9: Same data as Figure 8.5, but with contrast enhanced to emphasize cosmic fluctuations.

et al., 1994) and at greater distances, the faint blue galaxies (Treyer et al., 1995).

In principle this is a simple test based on a correlation statistic

$$W(\theta) = \frac{\langle \delta N / \delta I \rangle_\theta}{\langle N \rangle \langle I \rangle}$$

where $\langle N \rangle$ is the average number of galaxies per square degree, $\langle I \rangle$ is the average X-ray surface brightness, and the numerator is the sum of the product of deviations from the average summed over all pairs of cells separated by θ degrees. If there is no correlation between the two data sets then W(θ) is expected to be 0. Evaluating and interpreting, W(θ) has many subtleties, which are discussed in the papers referenced above.

An immediate difficulty is how to perform a fair comparison between a surface brightness survey and a galaxy catalog which is just a collection of locations where a galaxy is known to exist. The solution is to smear the galaxy catalog by "observing" it with an experiment with the same point spread function as the X-ray survey (Jahoda et al., 1991). Although there is unavoidable loss of information in the galaxy catalog, this allows a binning on the sky which maintains, by design, the same

cosmic correlation between adjacent cells, due to the finite beam, in both surveys. Jahoda et al. (1991; 1992) evaluated the correlation in cells of several square degrees at 0° lag: W(0). Although the galaxy surveys are shallow (Jahoda et al. used subsets of the UGC and ESO catalogs with effective depths <100 Mpc), and the CXB is believed to originate at truly cosmological distances (Boldt, 1987), one can interpret the value of W(0) in terms of an X-ray volume emissivity in the local universe which is sampled by the galaxy catalogs. This is extrapolated to predict the contribution to the CXB expected from a nonevolving population of X-ray emitters correlated with present epoch galaxies but integrated back over the redshift range where the CXB originates. Jahoda et al. (1992) concluded that a large fraction, greater than half, of the CXB could be produced by such sources.

The analysis of Jahoda et al. however, ignored two possible, and indeed certainly present, effects of source clustering. The first effect is a self-correlation. If some fraction of the galaxies are themselves also X-ray emitters, then some cells in the analysis will be enhanced in both galaxy surface density and X-ray surface brightness. Failure to account for this causes an overestimate of the X-ray volume emissivity and consequently of the fraction of the CXB arising from a nonevolving population. A second effect is the potential clustering of sources. Since it is well known that the positions of galaxies are correlated (c.f. Lahav et al., 1990), it is reasonable to assume that the positions of X-ray emitters correlated with galaxies may also be clustered. This too increases the measured value of W(0) (Lahav et al., 1993; Miyaji et al., 1994; Miyaji, 1994). Corrected for clustering, W(0), evaluated with both optical and IR galaxy catalogs, allows about 20% of the CXB to be produced by nonevolving populations. Given that strong evolution is observed in source counts (Maccacaro et al., 1991) this may be a less surprising result.

Carrera et al. (1995) have extended the formalism in order to evaluate, simultaneously, the values measured for W(θ) for $0 \leq \theta \leq 5°$, where the upper limit is large enough so that the intrinsic signal due to source clustering has vanished. Although some assumptions about the underlying parametrization of the correlation function must be made, these results are largely consistent with the 0 lag results that account for clustering (Lahav et al., 1993; Miyaji et al., 1994). Treyer and Lahav (1995) have recently extended the formalism to allow the correlation of faint blue galaxies with the CXB; because these galaxies cover a wide range of redshift, it is necessary to allow for a parametrization of the X-ray volume emissivity as a function of redshift. Although this study was done in the context of the SXRB observed with the imaging ROSAT experiment, which allows small cell sizes that can sensibly be correlated with the limited solid angles over which there are deep optical surveys, the formalism will allow an extension to data from ASCA or Astro-E which will be more directly relevant to studies of the harder CXB and comparable to the series of studies outlined in this section.

Macroscopic Anisotropies

There may also be anisotropies that are truly all sky. Indeed, one predicts a dipole effect due to our peculiar motion with respect to the inferred source of the bulk of the CXB. Though other large-scale features are plausible, this section is limited to a discussion of the possible existence and detection of an all sky dipole.

The microwave background has long been known to have a dipole with amplitude of about 1 part in 1,000 (Smoot et al., 1992) of the mean temperature. This is interpreted as the signature of motion with respect to the distant origin of this radiation. The net motion is measured to be $\sim$365 $km\ s^{-1}$ in the direction of l = 265°, b = 48°. This motion is the sum of several terms: the peculiar motion of the sun relative to the local standard of rest, galactic rotation, the motion of the galaxy with respect to the Local Group of galaxies, and peculiar motion of the Local Group. Under the assumption that the CXB arises from a distant source (beyond whatever mass anisotropy causes the motion of the Local Group) one predicts that there will be a dipole anisotropy of about $\pm$0.5% (Boldt, 1987). This effect, known as a Compton-Getting effect (Compton and Getting, 1935), results as a combination of a Doppler shifted input spectrum observed with a fixed instrumental energy band combined with special relativistic aberrations, which concentrate the solid angle in the direction of motion

Observation of such a dipole, however, is complicated by the anisotropies previously discussed. The galactic plane is clearly large compared to 0.5%, and even at high latitudes there is a galactic contribution of $\sim$1% of the total flux (Iwan et al., 1992). The peak enhancement associated with the Great Attractor is 4% (Jahoda and Mushotzky, 1989) and there are possible enhancements associated with the Super Galactic Plane (Mushotzky and Jahoda, 1992; Jahoda, 1993). The galactic enhancements are pure noise in this search. The extragalactic enhancements, however, to the extent that X-ray light traces matter (the central assumption of this section) are expected to have some relationship to the matter responsible (through gravitational acceleration) for the peculiar motion of the Local Group. The direction of motion of the Local Group (towards l = 277, b = 30) is not far from the net direction of the solar system relative to the microwave background. The search for a true all sky dipole is, therefore, complicated by the facts (a) that some regions of the sky are masked by the relatively bright galactic emission, and that a galactic correction may be necessary everywhere; (b) the presence of point sources of

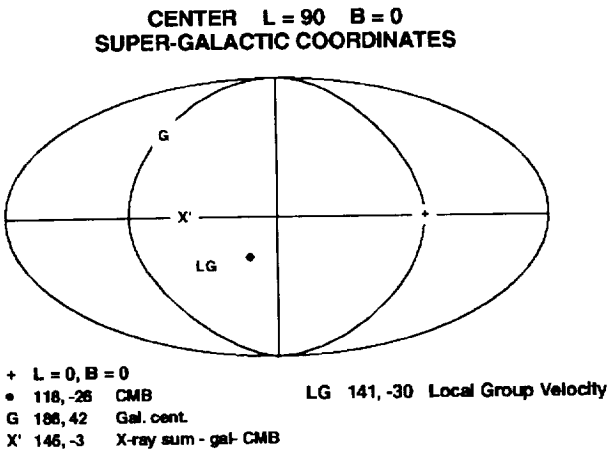


Figure 8.10: Interesting directions in super galactic coordinates.

emission such as nearby galaxies; and (c) the emission, potentially patchy, which traces the mass distribution in the nearby universe.

A good observation of the motion induced dipole would be an important confirmation of the interpretation that the microwave dipole is due to motion (though not in doubt, this important result rests on a single observable feature of the universe), and therefore several attempts have been made to look for the dipole despite all of the foreground noise described above. Jahoda (1993) performs a vector sum of the high latitude X-ray emission, ignoring directions with bright point sources and their antipodes as well as the galactic plane. He corrects for a galactic contribution using the model of Iwan et al. (1982) and a predicted Compton Getting effect. The results, shown in super galactic coordinates in Figure 8.10, show that the corrected sum lies on the line between the galactic center and the direction of motion of the Local Group. The + marks the super galactic center; the map is centered on (90,0); GC marks the galactic center; LG is the direction of motion of the Local Group; CG is the direction of the predicted Comption Getting effect; X' is the direction of the inferred vector sum of the X-ray sky. The dominant uncertainties are likely to be related to the excluded regions and the galactic model. Further attempts at a reconstruction with more complicated models of the foreground emission are underway.

8.5 CONCLUSIONS

The X-ray sky has several different appearances, depending on the energy band selected and the method of analyzing the data. At the lowest energies where absorption is important, cosmic X-rays can be used to "X-ray" the interstellar medium and yield clues as to the organization of hot (emitting) and cold (absorbing) regions. At higher energies, where absorption is negligible, the emission can be used to give clues about the distribution of matter in the nearby universe through searching for enhancements in the CXB, either related to large-scale or statistical features that are known or suspected from other surveys.

8.6 REFERENCES

Allen, J., Jahoda, K., and Whitlock, L., 1994, Legacy, 5, 27.

Bowyer, C.S., Field, G.B., and Mack, J.E., 1968, Nature, 217,32.

Boldt, E., 1987, Physics Reports, 146(4), 215.

Bunner, A.N., et al., 1969, Nature, 223, 1222.

Berkhuijsen, E.M., Haslam, C.G.T., and Salter, C.J., 1971, Astronomy and Astrophysics, 14, 252.

Burrows, D.N., and Mendenhall, J.A., 1991, Nature, 351, 629.

Carrera, F., et al., 1995, MNRAS, 275, 22.

Compton, A.H., and Getting, I.A., 1935, Physical Review, 47, 8127.

Danese, L., et al., 1993, ApJ, 412, 56.

Dressler, A., 1988, ApJ, 329, 519.

Evans, R.D., 1955, in "The Atomic Nucleus" McGraw-Hill.

Faber, S.M., and Burstein, D., 1989, in "Large Scale Motions in the Universe" eds. V.C. Rubin, and G. Coyne (Princeton: Princeton Univ. Press).

Fabian, A.C., and Barcons, X., 1992, Annual Reviews of Astronomy and Astrophysics, 30, 429.

Garmire, G., et al., 1992, ApJ, 399, 694.

Hayashida, K., 1989, Ph.D. thesis, U. Tokyo.

Iwan, D., et al., 1982, ApJ, 260, 111.

Jahoda, K., 1993, in "Advances in Space Research" 13(12), 231.

Jahoda, K., et al., 1992, ApJ, Lett. 399, L107.

Jahoda, K., et al., 1991, ApJ, Lett. 378, L37.

Jahoda, K., et al., 1985, ApJ, 290, 229.

Jahoda, K., McCammon, D., and Lockman, F.J., 1986, ApJ, Letters, 311, L57.

Jahoda, K., and Mushotzky, R.F., 1989, ApJ, 346, 638.

Lahav, O., Nemiroff, R.J., and Piran, T., 1990, ApJ, 350, 119.

Lahav, O., et al., 1993, Nature, 364, 693.

Lynden-Bell, D., et al., 1988, ApJ, 326, 19.

Maccacaro, T., et al. ApJ, 374, 117.

Marshall, F.J., and Clark, G.W., 1984, ApJ, 287, 633.

McCammon, D., et al., 1983, ApJ, 269, 107.

Miyaji, T., et al., 1994, ApJ, 434, 424.

Miyaji, T., 1994, Ph.D. thesis, U. MD.

Morrison, R., and McCammon, D., 1983, ApJ, 270, 119.

Mushotzky, R.F., and Jahoda, K., 1992, in "The X-ray Background" eds. X. Barons, and A.C. Fabian,

proc. of a workshop in Laredo, Spain, September 1991 (Cambridge University Press: Cambridge)
Piccinotti, G., et al., 1982, ApJ, 253, 485.
Schmitt, J.H.M.M., and Snowden, S.L., 1990, ApJ, 361, 207.
Smoot, G.F., et al., 1992, ApJ, Letters, 396, L1.
Snowden, S.L., and Petre, R., 1994, ApJ, Letters, 436, L123.
Snowden, S.L., et al., 1991, Science, 252, 1529.
Shafer, R.A., 1983, Ph.D. thesis, U.MD.
Treyer, M.A., and Lahav, O., 1995, MNRAS, submitted.
Turner, E.L., and Geller, M.J., 1980, ApJ, 236, 1.
Wood, K.S., et al., 1984, ApJ, Supplement, 56, 507.

Chapter 9

AGN AND PULSARS: HIGH ENERGY GAMMA-RAY BEAMS

D. J. THOMPSON
Laboratory for High Energy Astrophysics
NASA Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

Nonthermal, high-energy physical processes that produce energetic gamma-radiation indicate particle acceleration, which can occur in beams in some astrophysical sites. Gamma-ray studies of pulsars and blazars help understand both the particle acceleration and the beaming process.

9.1 INTRODUCTION

In astrophysics, the goal is to understand the physical nature of whatever is being studied. In short, "What is it?" and "How does it work?". Meeting this goal in high-energy astrophysics is often complicated by unfamiliar physical processes, nontraditional detectors, and challenging observational conditions. This chapter addresses two classes of objects that involve beams of gamma-rays: pulsars and active galactic nuclei (AGN). Because the emphasis is on what can be learned directly from gamma-rays, the first part of the discussion is a brief review of gamma-rays in astrophysics.

A schematic diagram of the electromagnetic spectrum Figure 9.1 shows gamma-rays occupying a broad band at the high-energy end. Although the transition from X-rays to gamma-rays is somewhat arbitrary, most photons with energies above 10^5 electron volts ($100\ keV$) are considered gamma-rays. In other terms, gamma-rays have frequencies above $2.4 \times 10^{19}\ Hz$ and wavelengths shorter than 1.2×10^{-11} meters. Gamma-rays, because of their high energies, are important to many astrophysical studies involving large energy transfers.

9.1.1 Gamma-Ray Production

Although the lowest energy gamma-rays can be produced in thermal processes, most sources of gamma-rays involve nonthermal processes such as radioactive decay, annihilation radiation, nuclear collisions, and electromagnetic interactions. The first three of these are discussed in more detail in other chapters. Electromagnetic processes, which appear to be the most important ones for the topics of this chapter, are summarized below in terms of gamma-ray production.

Synchrotron Radiation

For an electron energy E_e (in MeV) in a magnetic field B (in $gauss$), the peak of the resulting synchrotron spectrum is at frequency ν (in Hz)

$$\nu \simeq 5 \times 10^{26} B E_e^2 \tag{9.1}$$

In order to produce 100 MeV gamma-rays ($\nu = 2.4 \times 10^{22}\ Hz$), several combinations of parameters are possible, as shown in Table 9.1.

Inverse Compton Scattering

For an electron with energy E_e to scatter a photon with initial energy E_i up to a final energy of E_f requires

$$E_f \simeq ((E_e/mc^2)^2 E_i \tag{9.2}$$

where the electron rest mass energy $mc^2 = 0.511\ MeV$. Some possible parameters to produce a $100\ MeV$ gamma-ray by inverse Compton scattering are given in Table 9.2.

Table 9.1: Gamma-Ray Production by Synchrotron Radiation

B ($gauss$)	E_e(eV)	Site
10^{-6}	7×10^{16}	Space
1	7×10^{13}	AGN
10^{12}	7×10^{7}	Pulsar

Table 9.2: Gamma-Ray Production by Inverse Compton Scattering

Target Photon	E_i (eV)	E_e (GeV)
Microwave	0.001	160
Optical	1	5
X-ray	1000	0.16

For both synchrotron radiation and inverse Compton scattering, the gamma-ray energy spectrum starting with a differential electron spectrum

$$J(E_e)\ E_e^{-n} electrons/(\ cm^{\ 2}\ s\ \ MeV\ ) \tag{9.3}$$

has the approximate form

$$J(E)\ E^{-(n+1)/2} photons/(\ cm^{\ 2}\ s\ \ MeV\ ) \tag{9.4}$$

9.1.2 Gamma-Ray Detection

Gamma-rays interact with matter in three principal ways which can be used to construct gamma-ray detectors

1. Photoelectric Effect – the gamma-ray energy is transferred to an atomic electron. The photoelectric effect dominates at low gamma-ray energies, below about $0.5\ MeV$.

2. Compton Scattering – a gamma-ray scatters a low energy electron to higher energy. This process dominates at intermediate energies, between about 0.5 MeV and 10 MeV.

3. Pair production – a gamma-ray converts into an electron and its antiparticle, a positron. Pair production is only possible above 1 MeV and is the dominant process at higher energies. Conservation of energy and momentum prevents spontaneous pair production. Some possible mediators of this effect are matter (an atomic nucleus), a magnetic field, or another photon.

What all these gamma-ray interaction processes have in common is that a significant fraction of the gamma-ray energy is transferred to one or more charged particles. There are immediate implications for observational gamma-ray astrophysics

1. Gamma-rays cannot be reflected or refracted. There are no mirrors or lenses for gamma-rays.

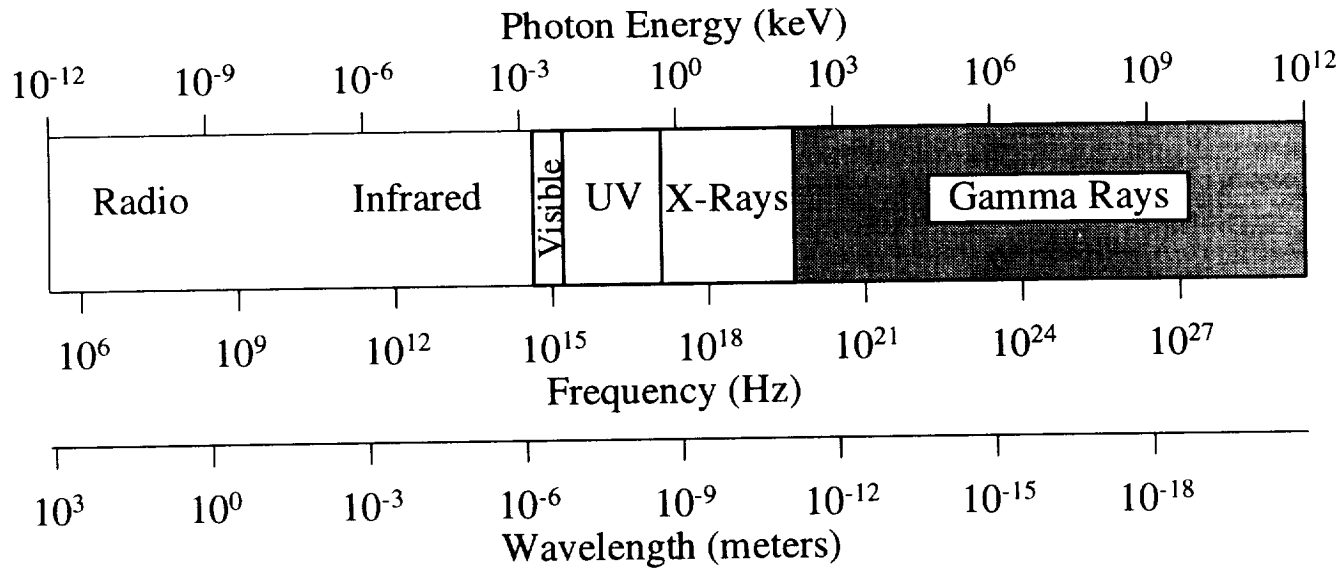


Figure 9.1: The Electromagnetic Spectrum

2. Gamma-ray telescopes are in reality charged particle detectors.

3. No single type of detector can operate successfully over the entire gamma-ray part of the spectrum.

The most comprehensive study of the gamma-ray sky to date is being carried out by NASA's Compton Gamma-Ray Observatory. A space mission is necessary, because the earth's atmosphere is opaque to gamma-rays. The Compton Observatory, a 16,000 kg satellite which was launched by the Space Shuttle Atlantis on April 5, 1991, carries four gamma-ray telescopes, shown schematically in Figure 9.2.

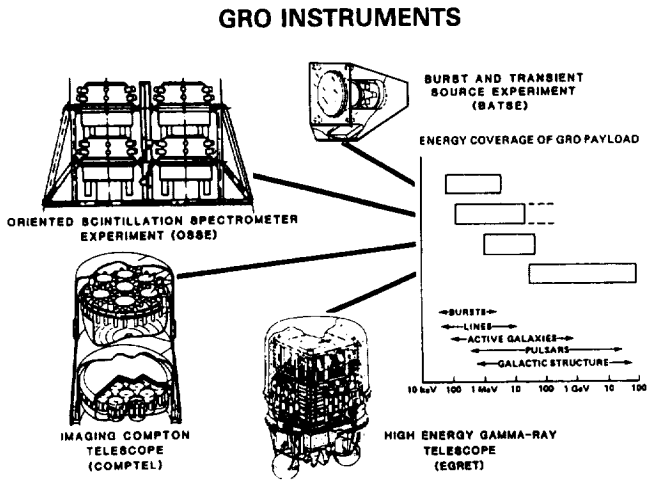


Figure 9.2: The four gamma-ray telescopes on NASA's Compton Gamma-Ray Observatory. The combined energy range spans nearly six orders of magnitude.

9.2 GAMMA-RAY PULSARS

First discovered by radio astronomers in 1967 (Hewish et al., 1968), pulsars are now known with high confidence to be spinning, magnetized neutron stars which were formed in supernovae. The collapsed star has a mass about 1.4 times that of the Sun and a radius of about 15 km. At the implied density (10^{14} g/cm^2) neutrons are more stable than independent protons and electrons; hence much of the star is composed of neutrons.

A pulsar is characterized by its spin period P and its rate of slowing $dP/dt = \dot{P}$. Based on these observable quantities and reasonable physical models, other parameters can be approximated for pulsars (see the chapter by Harding in this volume for some details)

Age (Timing age):

$$t = P/2\dot{P}$$

Rate of energy loss:

$$dE/dt = \dot{E} = 4\pi^2 \; I \; \dot{P} \; P^{-3} erg \; s^{-1}$$

where I = moment of inertia $\simeq 10^{45}$ gcm^2

Surface magnetic field:

$$B \simeq 3.2 \times 10^{19} (P \; \dot{P})^{1/2} \; gauss$$

In addition to these derived quantities, the distance can be estimated from the dispersion of the radio signal in the interstellar medium. A summary of data for 558

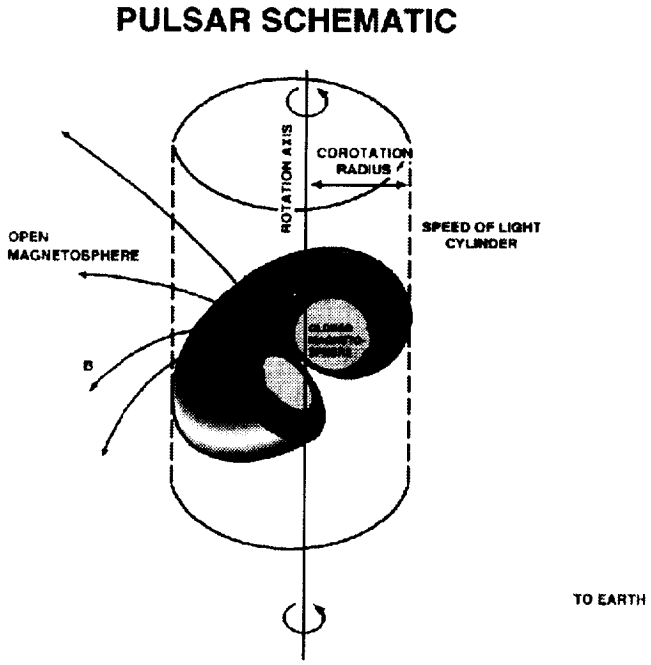


Figure 9.3: Pulsar schematic. In general, the angle between the rotation axis and the magnetic dipole axis of the neutron star is not known.

known pulsars is given by Taylor, Manchester, and Lyne (1993). Figure 9.3 shows a schematic diagram of a pulsar in one possible configuration.

Despite the wealth of knowledge about pulsars, some critical questions remain open

1. How do pulsars evolve? Does the magnetic field decay? What is the origin of millisecond pulsars?

2. What is the beaming geometry, the relationship of the magnetic field to the spin axis?

3. How is the rotational energy converted into observable radiation?

With the strong magnetic field expected at a neutron star (on the order of 10^{12} *gauss*), coupled with the rapid rotation, large electric fields can be produced which can in principle accelerate charged particles to high energies (up to 10^{16} eV). As described in Section 9.1., gamma-rays can be produced by such charged particles interacting with magnetic fields and with lower-energy photons. Gamma-rays may, therefore, help answer some of the questions about these high-energy objects.

Seven pulsars have now been seen at gamma-ray energies. The shapes of the pulses (light curves) at gamma-ray and other wavelengths is shown in Figure 9.4.

It is clear from this figure that not all pulsars are alike, either within one wavelength band or from one band to another. The beams of radiation appear to have a complicated geometry. The region above the dipole magnetic field poles is the likely location for particle acceleration and gamma-ray production. Whether this region is close to the poles or near the speed of light cylinder (at a distance $r = cP/2\pi$ beyond which the magnetosphere would be rotating faster than the speed of light) remains an open question. The commonality of double pulses, coupled with the absence of many 180° separations, argues against orthogonal rotator models in which both poles are sources of gamma-rays. The gamma-ray light curves are consistent with single-pole, hollow-cone models of emission (see the chapter by Harding).

A second observable property of a pulsar is its energy spectrum. The energy spectrum of the Crab pulsar across much of the electromagnetic spectrum is shown in Figure 9.5. This type of presentation multiplies the energy flux by frequency, so that the plotted quantity represents the observed power per decade of frequency. It is clear from this figure that the vast majority of the observed power from the Crab pulsar is in the X-ray and gamma-ray bands of the spectrum. This pattern holds for all seven gamma-ray pulsars, a fact which emphasizes the importance of studying the high-energy radiation.

Another feature of pulsar energy spectra is visible in Figure 9.5: the gamma-ray spectrum is fairly well represented by a straight line (power law). In terms of a photon (rather than energy) spectrum

$$J(E_\gamma)\ E_\gamma^{-n} photons/(\ cm^{\,2}\ s\ \ MeV\ ) \qquad (9.5)$$

the spectral index n for the Crab is 2.15 ± 0.04 (Nolan et al., 1993). For the six pulsars which are seen above 100 MeV (PSRB1509-58 is seen only up to 1 MeV), the spectral index shows a trend as a function of pulsar age, illustrated in Figure 9.6. The implication is that older pulsars (at least those which have been detected) preferentially produce higher energy gamma-rays.

The observed energy spectra can be integrated to give an observed gamma-ray energy flux F_γ. Using the pulsar distance D obtained from radio or optical observations, the high-energy (X-ray and gamma-ray) luminosity can be determined from

$$L_\gamma = \Omega F_\gamma D^2 \qquad (9.6)$$

where Ω is the solid angle into which the gamma-rays are beamed. As discussed above, the beaming geometry is not known. Ω must lie between 0 and 4π, and different pulsar models predict it to be between 0.1 and nearly 4π. An intermediate choice of 1 *steradian* provides a result which is easily scaled to any particular model prediction. The results are shown in Table 3. The efficiency is the ratio of the observed gamma-ray energy flux to the total available energy from the rotational spin-down of the pulsar, $\dot{E}$.

The older pulsars show a significantly higher efficiency for converting rotational energy into gamma-rays, with PSRB1055-52 reaching over 15% under these assumptions. The high efficiency calculated for B1055-52

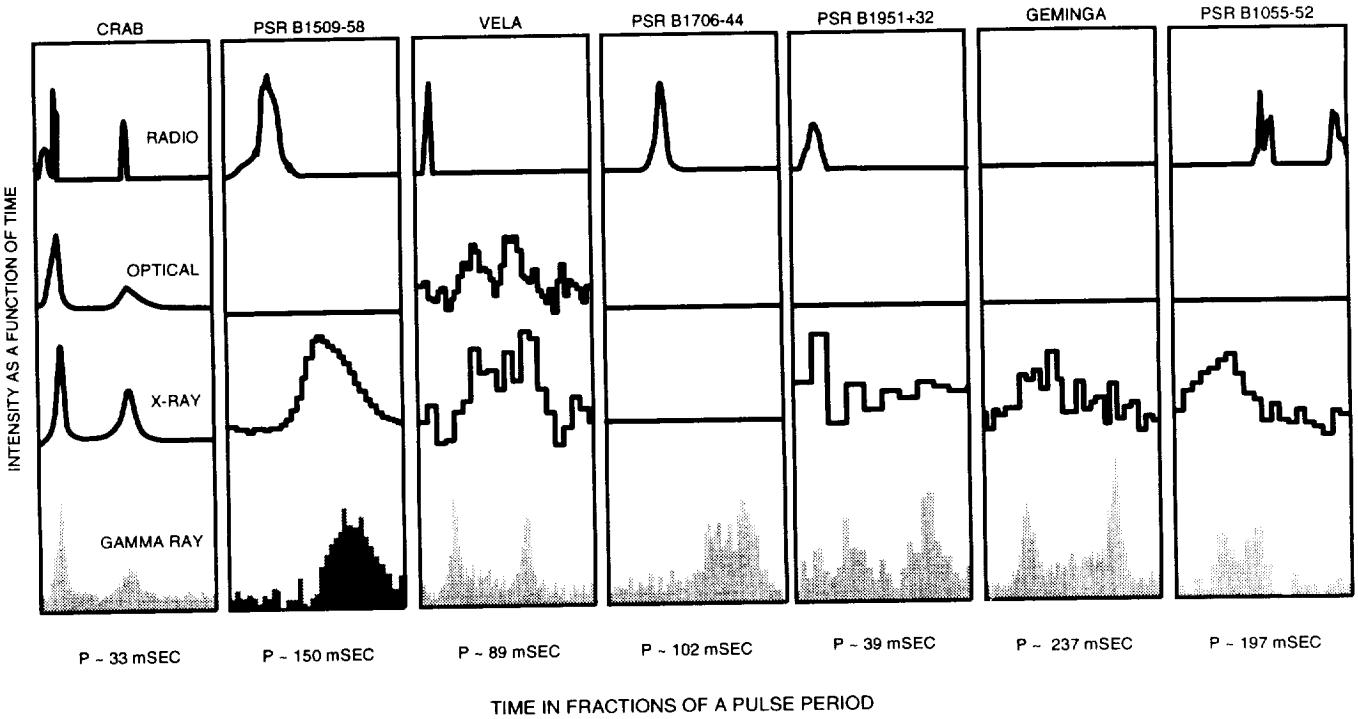


Figure 9.4: Multiwavelength light curves for the known gamma-ray pulsars. Crab: Manchester, 1971 (radio), Groth, 1975 (optical), Harnden and Seward, 1984 (X-ray), Nolan et al., 1993 (gamma-ray). B1509−58: Ulmer et al., 1993 (radio and gamma-ray), Kawai et al., 1991 (X-ray). Vela: Kanbach et al., 1994 (radio and gamma-ray), Wallace et al., 1977 (optical), Ögelman, 1994 (X-ray). PSRB1706−44: Johnston et al., 1992 (radio), Thompson et al., 1996 (gamma-ray). PSRB1951+32: Kulkarni et al., 1988 (radio), Safi-Harb et al., 1995 (X-Ray), Ramanamurthy et al., 1995 (gamma-ray). Geminga: Bertsch et al., 1992 (gamma-ray), Halpern and Ruderman, 1993, and Halpern, 1995 (X-ray). PSRB1055−52: Fierro et al., 1993 (radio and gamma-ray), Ögelman and Finley, 1993 (X-ray).

shows that assuming a much larger solid angle would produce a nonphysical result of more than 100% efficiency.

A variety of models have been developed to interpret these results on gamma-ray pulsars. Several of these are summarized in the chapter by Harding. At present, none of these models is completely successful in explaining the observed properties of these high-energy pulsars. In particular, all models predict that some additional gamma-ray pulsars should have been seen. Upper limits (Nel et al., 1994) to gamma-ray emission from other radio pulsars suggest that the true physical conditions around these rotating neutron stars are more complicated than the first generation of models.

In summary, the gamma-ray observations of pulsars show

1. Pulsars must accelerate electrons to high energies, at least into the *GeV* range.

2. Gamma-rays can be an important part of the energy loss of pulsars.

3. The beaming geometry of pulsars is complicated.

4. Models for gamma-ray pulsars do not yet explain all the observable features.

9.3 ACTIVE GALACTIC NUCLEI

The term Active Galactic Nuclei (AGN) encompasses a wide variety of objects, including quasars, BL Lacertae objects, Seyfert galaxies, and radio galaxies. All these are characterized by unusually luminous central regions compared to normal galaxies like the Milky Way. Many properties of AGN are discussed in the chapter by Kazanas. Although not accepted by all astrophysicists, the "Standard Model" of an AGN is a supermassive (10^6 to 10^9 solar masses) black hole surrounded by an accretion disk formed by material that is spiraling into the center. Often a "jet" of material and radiation is thought to emerge perpendicular to the accretion disk, along the rotation axis of the black hole. Figure 9.7 is a sketch of this concept of an AGN.

Many questions about AGN are important topics in current astrophysics. Among these are

1. What is the origin and evolution of AGN?

2. How does the accretion disk form and evolve?

3. What produces "jets," and how do they evolve?

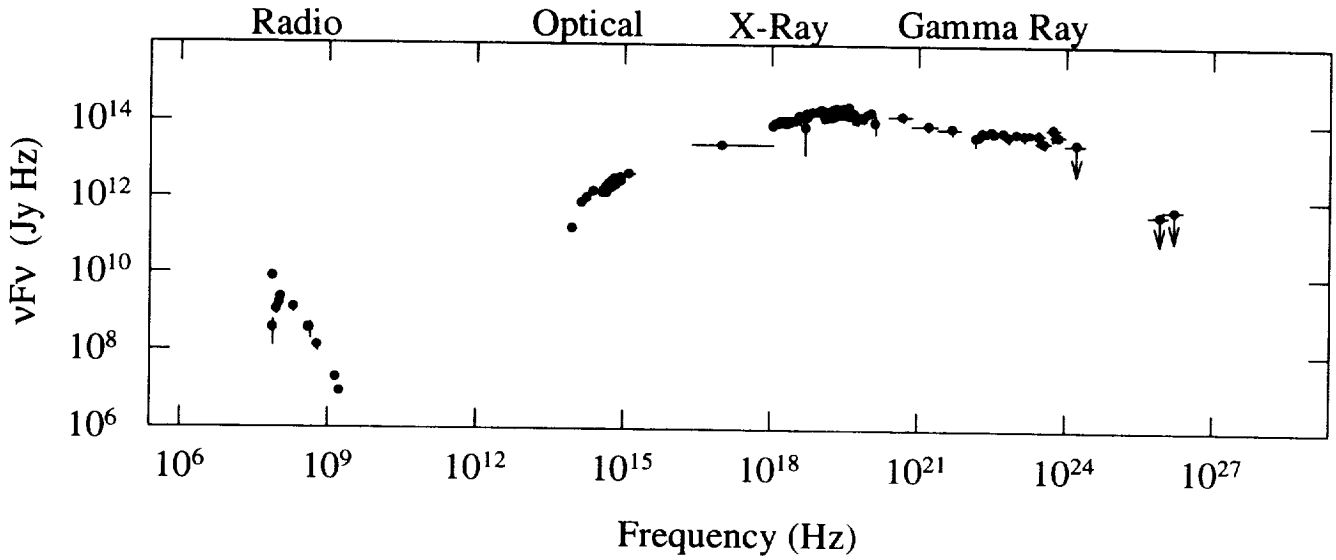


Figure 9.5: Multifrequency energy spectrum of the Crab Pulsar. Radio: Taylor, Lyne, and Manchester, 1993; IR: Middleditch, Pennypacker, and Burns, 1983; Optical: Oke, 1969; X-ray: Pravdo and Serlemitsos, 1981; Gamma-ray: Ulmer et al., 1993; Bennett, et al., 1993; Nolan et al., 1993; TeV: Vacanti et al., 1991, Akerlof et al., 1989.

Table 9.3: Gamma-Ray Pulsar Characteristics

Pulsar	Period(s)	Age(yr)	Luminosity ($erg\ s^{-1}$)	Efficiency
Crab	0.033	1300	4.5×10^{35}	0.001
B1509-58	0.150	1550	1.9×10^{35}	0.010
Vela	0.089	11,000	2.3×10^{34}	0.003
B1706-44	0.102	17,000	3.7×10^{35}	0.011
B1951+32	0.040	110,000	1.4×10^{34}	0.004
Geminga	0.237	340,000	2.6×10^{33}	0.080
B1055-52	0.197	530,000	5.0×10^{33}	0.150

4. How is gravitational energy converted to observable radiation?

Gamma-radiation has been found to be important in two classes of active galatic nuclei

1. Low energy gamma-rays are seen from a number of Seyfert galaxies, such as NGC 4151 (Maisack et al., 1993). Based on their energy spectra, these gamma-rays are thought to be radiation from the hot gas in the accretion disk. Beaming is unlikely for such sources.

2. High-energy gamma-rays are seen from a number of quasars and BL Lacertae objects. These are likely to be beamed and are the subject of the remainder of this discussion.

In its survey of the high-energy gamma-ray sky, EGRET on the Compton Observatory found evidence that more than 50 Active Galactic Nuclei are powerful sources of gamma-radiation (von Montigny et al., 1995; Thompson et al., 1996). None of these are Seyfert galaxies. Many are Optically Violent Variable (OVV) quasars, some are BL Lacertae objects, and several show evidence in radio of having apparently superluminal components. All were known as radio-bright objects with flat radio spectra in the 2-5 GHz range.

Figure 9.8 is a multi-wavelength energy spectrum for one of these gamma-ray AGN. Others are similar. The distinguishing characteristic is the extremely high power per decade seen in gamma-rays, exceeding that of other parts of the spectrum by an order of magnitude or more. If the radiation is assumed to be isotropic at the source, then the gamma-ray luminosity $L_\gamma = 4\pi F_\gamma D^2$ for a gamma-ray flux F_γ and a distance D. For some of the observed AGN, the implied gamma-ray luminosity is $10^{48}\ erg\ s^{-1}$.

A second observable feature of the gamma-radiation from these AGN is time variability, as shown in Figure 9.9. Most of the AGN seen at high energies show some variability. In this case, the gamma-ray flux changes significantly on a time scale of 1-2 days.

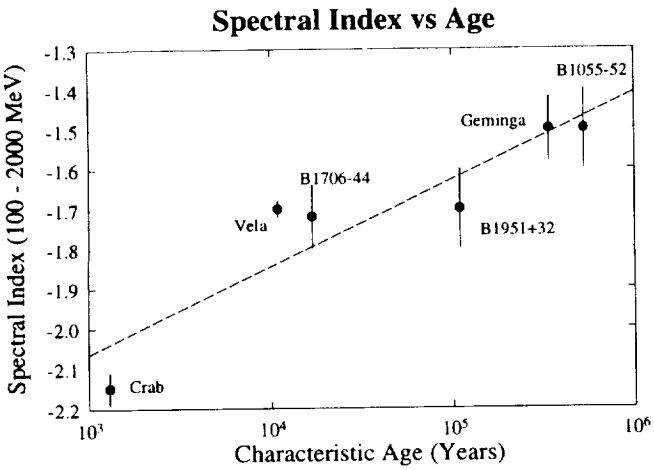


Figure 9.6: Gamma-ray spectral index as a function of pulsar age.

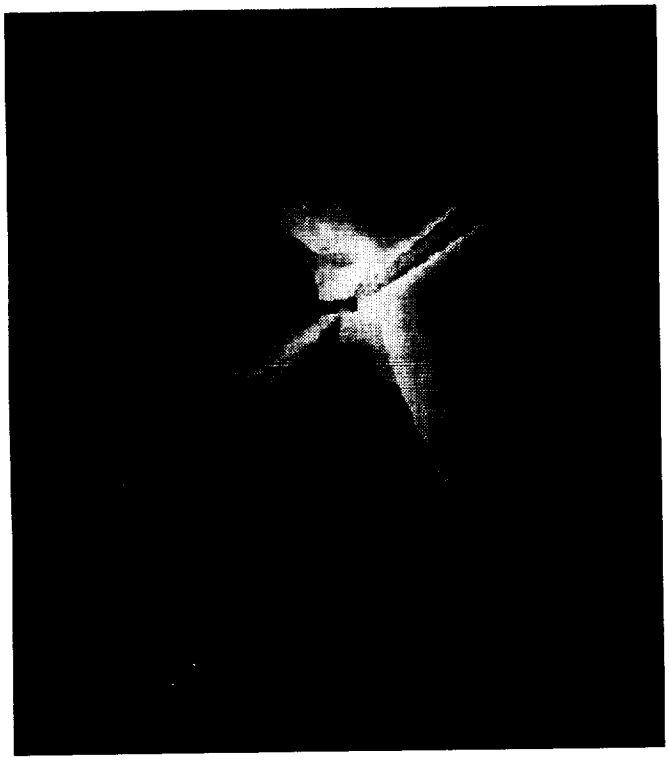


Figure 9.7: Artist rendition of an active galactic nucleus.

Observed variability of a source on a time scale ΔT sets a limit on the size of the emitting region. Causality implies that the region size R cannot be larger than light can travel in ΔT,

$$R \leq c\Delta T/(1+z) \tag{9.7}$$

where z is the redshift of the object. The $(1+z)$ is a Doppler shift due to the motion of the object away from us. For time variability with a scale of 1 day (as seen in Figure 9.9), the emitting region size is on the order of 10^{15} cm, about twice the size of the orbit of Pluto. This relatively small size might suggest that the gamma-rays are produced near the central black hole of the AGN. Such a region has a high density of X-ray and other photons, however, and gamma-rays can be absorbed by photon-photon pair production $\gamma + \gamma \rightarrow e^+ + e^-$.

The cross section for this process peaks at about

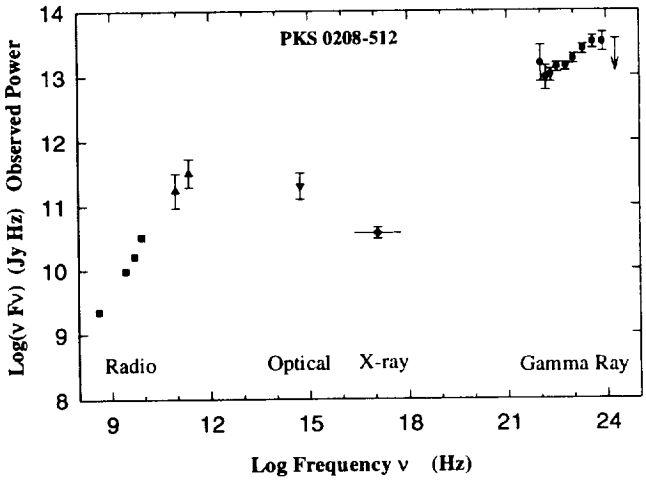


Figure 9.8: Multiwavelength (νF_ν) energy spectrum of PKS 0208−512 (Bertsch et al., 1992). On this type of plot, νF_ν is the observed power per decade of frequency.

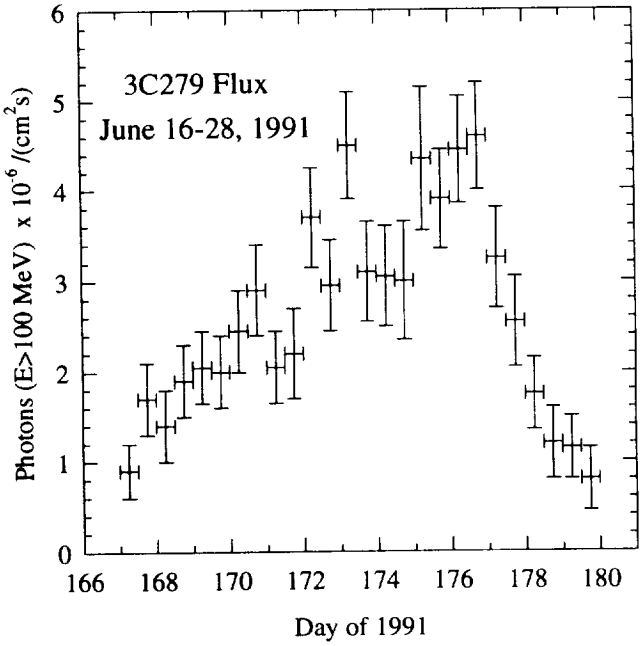


Figure 9.9: Gamma-ray time variability of quasar 3C279 in June 1991 (Kniffen et al., 1993).

twice the threshold, so that

$$E_\gamma E_t \sim 2(m_e c^2)^2 \simeq 0.5 \ \ MeV \qquad (9.8)$$

where E_γ is the gamma-ray energy and E_t is the target photon energy. For 1000 MeV gamma-rays, the principal target photons are X-rays with energy about 0.5 keV.

The optical depth for gamma-rays due to this process is large when the dimensionless compactness parameter l is greater than about 10.

$$l = (L_X/R)(\sigma_T/m_e c^3) \simeq (L_X/\Delta T)/10^{39} \qquad (9.9)$$

where L_X is the X-ray luminosity in $erg\ s^{-1}$, R is the size of the emitting region, σ_T is the Thomson cross section ($7 \times 10^{-25}\ cm^2$), and ΔT is the variability time scale in seconds. For several of the AGN seen by EGRET, l is greater than 10. If the gamma-rays were produced isotropically, they should not be able to escape to be seen.

The solution to this dilemma is that the gamma-rays probably originate in the jet. The class of AGN seen by EGRET is often termed "blazars", and the model for these objects is one in which the jet is directed nearly along the line of sight. For a jet with flow velocity v_p,

$$\beta_p = v_p/c \text{ and } \Gamma_p = (1-\beta_p^2)^{-1/2}$$

The Doppler factor δ, for an angle θ between the jet and the line of sight, is (for $\beta = 1$)

$$\delta = \Gamma_p^{-1}(1 - cos\theta)^{-1} \qquad (9.10)$$

with typical values of 3 to 10.

Beaming changes many of the calculated properties of the AGN

1. The size implied by the time variation is increased from R to δR.

2. The intrinsic luminosity L_{int} is related to the observed luminosity L_{obs} by

$$L_{int} = \delta^{-n} L_{obs} \qquad (9.11)$$

where n is between 3 and 4 (depending on the assumed shape of the emitting region). This beaming changes the calculated gamma-ray luminosity by several orders of magnitude, although some are still found in the 10^{46} $erg\ s^{-1}$ range.

3. For beamed radiation, the optical depth is reduced to $\delta^{-(4+2a)}$ of the isotropic value, where a is the power law energy index of the absorbing X-rays, typically 0.5 - 1.0. The observed GeV gamma-rays have lower energy before the Doppler shift by the beam, so that the absorbing X-rays lie in the multiple keV range (and are, therefore, less numerous). Even the highest energy gamma-rays observed by EGRET can escape without significant attenuation in this case.

Models for gamma-ray blazars all require acceleration of particles to very high energies in the jet. Three general model types are

1. Synchrotron Self Compton. Within the jet, high-energy electrons radiate in the radio to optical bands by synchrotron radiation. Some of these synchrotron photons can then be upscattered by the same electron population to gamma-ray energies.

2. External Compton. Photons from outside the jet, either thermal radiation from the accretion disk itself or photons scattered into the jet, can be inverse Compton scattered to gamma-ray energies by the high-energy electrons in the jet.

3. Proton Neutron Cascade. High-energy protons or neutrons start a cascade of secondary particles within the jet, including electrons and photons.

In all these models, high-energy particles are assumed. Some possible origins of these energetic particles are

1. The rotation of the black hole and the accretion disk sets up large electric fields, which accelerate particles to extremely high energies. In effect, this is analogous to the acceleration in a pulsar, although on a much larger scale.

2. Particles are accelerated to high energies in the jet by shocks traveling along the jet.

At present, the data do not provide definitive answers about the models or the acceleration processes. In addition to the study of the energy spectrum at all wavelengths, a key test for models is the relationship of time variability in different parts of the electromagnetic spectrum to each other. For example, in Synchrotron Self Compton models, the same electrons produce UV by synchrotron and gamma-rays by Compton scattering. Variations at these wavelengths should be simultaneous, with variations in the IR and hard X-ray lagging. Models with external Compton scattering of photons offer at least two possibilities. If the electrons are being accelerated in the jet, then the higher energy variation should come later than the lower energies. If the electrons are injected into the jet with high energy, then the high energy variation may start first (assuming they are far enough out to avoid gamma gamma pair production).

The gamma-ray observations of blazars can be summarized

1. For some blazars, gamma-rays are the dominant observable form of radiation.

2. The time variability of the gamma-radiation implies emission in a jet, probably directed toward us.

3. Various models for these jets have been developed, although there is no consensus on either the particle acceleration mechanism or the gamma-ray production process.

4. Multiwavelength studies of spectra and time variability should help resolve at least some of the open questions about these objects.

9.4 CONCLUSION

Pulsars and active galactic nuclei that are gamma-ray sources are sites in which particles are accelerated to high energies. Beaming, as evidenced by periodicity in the pulsars and time variability in the AGN, is a characteristic of both object classes. The gamma-rays help study both the beaming and the underlying particle acceleration.

9.5 REFERENCES

Akerlof, C., et al., 1989, in "Gamma Ray Observatory Workshop" Proceedings, NASA/GSFC, ed. W.N. Johnson, 4-49.

Bennett, K., et al., 1993, private communication.

Bertsch, D.L., et al., 1992, Nature 357, 306.

Fierro, J.M., et al., 1993, ApJ 413, L27.

Groth, E.J., 1975, ApJS, 293, 431.

Halpern, J.P., 1995, private communication.

Halpern, J.P., and Ruderman, M., 1993, ApJ 415, 286.

Harnden, F.R., Jr. and Seward, F.D., 1984, ApJ 283, 279.

Hewish, A., Bell, S.J., Pilkington, J.D.H., Scott, P.F., and Collins, R.A., 1968, Nature 217, 709.

Johnston, S., et al., 1992, MNRAS, 255, 401.

Kanbach, G., et al., 1994, A and A, 289, 855.

Kawai, N., et al., 1991, ApJ 383, L65.

Kniffen, D.A., et al., 1993, ApJ 411, 133.

Kulkarni, S.R., et al., 1988, Nature, 331, 50.

Maisack, M., et al., 1993, ApJ, 407, L61.

Manchester, R.N., 1971, ApJ 163, L61.

Middleditch, J., Pennypacker, C., and Burns, M.S., 1983, ApJ 273, 261.

Montigny, C. von, et al., 1995, ApJ, 440, 525.

Nel, H.I., et al., 1996, ApJ, in press.

Nolan, P.L., et al., 1993, ApJ 409, 697.

Ögelman, H.B., and Finley, J.P., 1993, ApJ 413, L31.

Ögelman, H.B., 1994, in "Lives of Neutron Stars" NATO ASI, Kemer, Turkey, in press.

Oke, J.B., 1969, ApJ 156, L49.

Pravdo, S.H., and Serlemitsos, P.J., 1981, ApJ 246, 484.

Ramanamurthy, P.V., et al., 1995, ApJ, 447, L109.

Safi-Harb, S., Ögelman, H.B., and Finley, J.P., 1995, ApJ, 439, 722.

Taylor, J.H., Manchester, R.N., and Lyne, A.G., 1993, ApJS 88, 529.

Thompson, D.J., et al., 1996, ApJ, in press.

Ulmer, M.P., et al., 1993, ApJ 413, 738.

Vacanti, G., et al., 1991, ApJ 342, 379.

Wallace, P.T., et al., 1977, Nature 266, 692.

Chapter 10

AN INTRODUCTION TO THE PHYSICS OF AGN

DEMOSTHENES KAZANAS
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center, Code 665
Greenbelt, MD 20771

ABSTRACT

Orders of magnitude of the relevant physical parameters and the basic processes and physics associated with Active Galactic Nuclei are given. The two most broad and distinct AGN classes (the radio loud and radio quiet) are discussed. The characteristics and physical processes believed to be responsible for the emission in these two classes are discussed with particular emphasis on the recent observations of high-energy γ-rays from radio loud AGN.

10.1 INTRODUCTION

Active Galactic Nuclei (AGN) are, as the term indicates, the central regions of certain galaxies at which there is apparent activity, manifest by the emission of radiation, which can rival or even surpass the power output of the entire galaxy by as much as a thousand fold. What makes AGN exceptional is the fact that all this power is emitted from a region which is quite small by galactic standards, and whose size is of the order of that of our solar system.

AGN come under a host of names depending on their total luminosity, morphology, and frequency band in which they were first discovered or studied. The most familiar name to the nonexpert is that of "quasar," denoting a class of objects, controversial at the time of their discovery, precisely because of their prodigious power requirements, if they indeed were at the distances indicated by their redshifts. We now know that quasars are only the brightest members in the AGN class, which came to include other objects such as Seyfert galaxies, and radio galaxies. The difference between a quasar and a Seyfert is their absolute luminosity and their distance to the observer: Because quasars are very far away and much brighter than the host galaxy, it is almost impossible to detect the associated (but much fainter) galaxy, thus giving the impression that quasars represent "naked" sources of energy output.

To appreciate the extreme conditions associated with AGN, the following scaling analogy is instructive. If the electric power output of a major metropolitan area (e.g. Washington or Boston) represents the light distribution in a galaxy, then a Seyfert nucleus would represent the same power output from the volume smaller than the head of a pin, while a quasar in the same scale would emit as much power as the entire United States from the volume of a basketball. It becomes thus apparent that AGN physics deal with extremes of power and power density, higher than those which could be attributed to nuclear reactions. It was precisely these power requirements that genererated controversy with the name of quasars and AGN in general. The absolute values of the power associated with AGN are $\lesssim 10^{12}\ L_\odot = 10^{44}\ erg\ s^{-1}$ for a Seyfert and $10^2 - 10^3$ times higher for a quasar.

The requirement for such tremendous power outputs leaves only one choice for the source that powers AGN: gravity. It was proposed that, by radiatively releasing the energy available from the accretion of matter into a deep gravitational potential of a black hole, one could achieve efficiencies approaching mc^2. In contrast, nuclear reaction efficiencies cannot exceed $10^{-3}mc^2$. Therefore, it has been accepted, and it is generally agreed within the community, that AGN are powered by matter accretion onto a black hole. Although the source of energy is agreed upon by almost everyone, that is where the agreement, for the most part, stops. Despite twenty years of study of these objects, the particular details of the accretion process and the energy release remain elusive. Their study still yields surprises and an in depth understanding of the physics involved is still lacking.

10.2 THE PHYSICS OF AGN

This section presents an incomplete summary of the physics of AGN. Its purpose is to help a nonspecialist develop a feeling of the order of magnitude of the physical quantities involved and the observational significance of the various processes associated with the emission of high-energy radiation in photons and neutrinos.

10.2.1 The Physical Quantities

The assumption that AGN are powered by accretion of plasma onto a black hole does provide estimates of the order of magnitudes of the various quantities associated with the accretion in terms of the mass of the black hole M and the accretion rate $\dot{M}$.

1. Length

The size of the black hole horizon sets the natural unit of length in the problem. The length of the Schwarzschild radius in cm is given in terms of the black hole mass by

$$R_S = 3 \times 10^{14} M_9 \quad cm \tag{10.1}$$

where M_9 is the mass of the black hole in units of $10^9\ M_\odot$. With this unit of len-gth one can scale the radial distance r by defining the dimensionless radial coordinate $x = r/R_S$.

2. Velocity and Time Scales

The characteristic velocities and time scales usually quoted in association with AGN variability are the free-fall velocity $v_{ff} \simeq (2GM/r)^{1/2} = c/x^{1/2}$ and the free-fall time scale $t_{ff} \simeq r/v_{ff} \simeq 10^4 x^{3/2} M_9$. These expressions are usually employed when trying to estimate the size of the AGN "central engine" region and the mass of the putative black hole, responsible for powering the AGN. It should be cautioned that in the most popular models, namely those involving accretion disks and viscous dissipation of energy, the associated time scales are much longer than the above estimates (See below, item 3).

3. Accretion Rate

The unit of accretion rate in AGN used (arbitrarily) is one solar mass per year because the observed AGN luminosities imply accretion rates which range from a few orders of magnitude smaller to roughly 100 times larger

than this value. One solar mass per year, therefore, is roughly the geometric mean of the extreme values of this quantity associated with AGN. The luminosity corresponding to the unit of accretion rate, at an efficiency of energy conversion into radiation equal to mc^2, is 5×10^{46} $erg\ s^{-1}$. It is certain that the efficiency is less than this, and it is loosely assumed that it is $\simeq 0.1mc^2$, so that $\dot{M} = 1$ (i.e., one solar mass per year) provides roughly $10^{46}\ erg\ s^{-1}$. A more natural unit for the accretion rate will be discussed later on in this section.

4. Particle Density

The density of matter in the vicinity of the accreting black hole apparently depends on both the accretion rate, $\dot{M}$, the mass, M_9 (i.e., the mass in units of 10^9 solar masses), and the dimensionless distance, x, from the black hole. Unfortunately, it also depends on the speed with which matter is accreted onto the black hole, a quantity not very accurately determined. A lower limit on the density can nonetheless be obtained by assuming the highest possible accretion speed, which in the case of accretion is given by the free-fall velocity, i.e., by assuming a largely spherically symmetric free-fall of the accreting matter. Under this assumption, the density of the accreting matter is

$$n(x) \simeq 10^9 x^{-3/2} \frac{\dot{M}}{M_9^2} \quad cm^{-3} \qquad (10.2)$$

Different assumptions about the mode of accretion can lead to vastly different estimates for the values of the particle density in the vicinity of the black hole. In particular, in the popular accretion disk models, the accretion of matter, (and therefore the rate of energy release per unit mass) takes place over (largely unknown) viscous time scales leading to much different values and functional dependences for the particle density. (Shapiro and Teukolsky, 1983).

5. Luminosity

The fact that AGN are powered by gravity-induced accretion provides a natural scale for the luminosity of these objects, the so called Eddington luminosity L_{Edd}, given the requirement that the radiation pressure does not overcome the force of gravity. This yields

$$\frac{GMm_p}{r^2} = \frac{L}{4\pi r^2 c}\sigma_T \qquad (10.3)$$

or

$$\begin{aligned} L_{Edd} &= 4\pi GMm_p c\sigma_T \simeq 1.3 \times 10^{38} M_0 \\ &= 1.3 \times 10^{47} M_9\ erg\ s^{-1} \end{aligned} \qquad (10.4)$$

Because both forces, for spherically symmetric geometry, drop-off like r^{-2}, the condition of balance between radiation pressure and gravity does not depend on the radius but only on the mass of the accreting object. The existence of this characteristic luminosity then allows the definition of an equivalent accretion rate, the so called Eddington accretion rate, through the relation

$$\dot{m}_{Edd} \equiv L_{Edd}/c^2 \simeq 2M_9 \quad M_\odot\ \mathrm{yr}^{-1} \qquad (10.5)$$

It should be pointed out that the above expression implicitly assumes that the accreting material is fully ionized and therefore the coupling of radiation with the infalling matter involves only the scattering of radiation with the ambient electrons, that is scattering at the Thomson cross section σ_T ($\sigma_T = 8\pi r_e^2/3 = 6.65 \times 10^{-25}\ cm^2$; $r_e = e^2/m_e c^2$ is the classical electron radius). Were the matter non-fully ionized, one would have to use, instead of σ_T in Equation (10.3), the relevant atomic cross sections which are larger by several orders of magnitude. Steady state accretion would then be impossible even for luminosities well below the Eddington values. In principle, the ionization structure of the accreting gas has to be solved in conjunction with the dynamical effects of radiation pressure, which will not be attempted here.

6. Energy Density

Estimates of the particle energy density suffer from the same uncertainties as those of number density. In the case of quasispherical accretion, in which the density can be estimated rather accurately and in which the velocity dispersion of the accreting matter is of the same order as the free-fall velocity, the energy density, similar in magnitude to the ram pressure, is

$$\begin{aligned} P &\simeq \rho v_{ff}^2 \simeq n m_p c^2 (v_{ff}/c)^2 \\ &\simeq 1.5 \times 10^6 x^{-5/2} \dot{M} M_9^{-2} \quad erg \quad cm^{-3} \end{aligned} \qquad (10.5)$$

In the case of a thin accretion disk one has to bear in mind that the ambient density $n(x)$ can be much higher and that the velocity dispersion can be much less than the azimuthal velocity $v_\phi \simeq v_{ff}$.

7. Magnetic Field

This is the most poorly determined parameter associated with AGN. Yet, its knowledge is important because, in many cases, it is believed that the continuum emission is due to synchrotron radiation. The magnitude of this parameter is simply determined from equipartition arguments, that is by equating the energy density associated with a magnetic field B, $U_{mag} = B^2/8\pi$, wity a characteristic energy density. As "characteristic energy density" it is typically used the photon energy density $L/4\pi r^2 c$ (mainly because it can be easily estimated from the observed luminosity and from the variability of a source which provides an estimate of r). An upper limit to the mean magnetic field value in quasi-spherically

symmetric accretion can be obtained by using as a characteristic energy density the ram pressure. These considerations yield values for the magnetic field which scale inversely with the mass of the black hole, $B \sim 1 - 100\, M_9^{-1}$ *gauss*.

10.2.2 The Physical Processes

This section deals with several of the physical processes associated with AGN. The list is by no means complete, reflecting mainly the author's personal prejudices and the limited of the available space. Some references for futher reading for the interested are provided. The processes discussed include dynamics, radiation processes, the opacities of high-energy radiation, and the effects of relativistic beaming on the emission from AGN.

1. The Dynamics of Accretion

The notion that the prodigious energy output of AGN is due to accretion onto a compact object (most likely a massive black hole) dates back to the late sixties. The mode of accretion in these earliest models was considered to be spherical accretion. The dynamics for this process are sufficiently simple to be easily reproduced here. They are governed by the mass, momentum, and energy conservation of a fluid in the presence of a gravitational field, assumed to be that of a point mass. These are the mass Equation (10.6) and momentum Equation (10.7) conservation equations below (given in steady state), supplemented by an equation of state for the gas (assumed to be a polytrope i.e., $P = K\rho^{\Gamma}$, with K, Γ constants)

$$\nabla\cdot(\rho\mathbf{u}) = \frac{1}{r^2}\frac{d}{dr}(r^2\rho u) = 0 \qquad (10.6)$$

$$u\frac{du}{dr} = -\frac{1}{\rho}\frac{dP}{dr} - \frac{GM}{r^2} \qquad (10.7)$$

These equations can be integrated to yield the familiar mass conservation and Bernoulli integrals

$$4\pi\rho r^2 u = \dot{M} = \text{constant}$$

$$\frac{1}{2}u^2 + \frac{1}{\Gamma-1}a^2 - \frac{GM}{r} = \frac{1}{\Gamma-1}a_\infty^2 = \text{constant}$$

where we have used the boundary conditions at infinity to determine the constant in Eq (10.7), and $a^2 = dP/d\rho$ is the speed of sound of the accreting gas.

These equations govern the rather idealized problem of matter accretion onto a point mass embedded into an infinite "sea" of gas with sound speed a_∞. It is apparent from these equations that when the gravitational potential dominates the dynamics, the pressure terms can be neglected and the gas *free-falls* with velocity $u \propto (2GM/r)^{1/2}$. This law for the velocity, when combined with the mass conservation constraint gives the expression for the density profile in spherical accretion given in Equation (10.2). On the other hand, if the pressure term is dominant (i.e., at distances at which the speed of sound is much larger than that of the local escape velocity), the effects of the presence of the point mass are irrelevant, and the density is almost constant. This condition, in conjunction with the mass conservation integral, indicates that in this case, $u \propto r^{-2}$.

The above equations can be used in conjunction with the polytropic equation for the gas to determine the qualitative behavior of the accreting flow. In particular, one would like to estimate whether gas can, in fact be in free-fall. This latter condition is realized only if the local gas pressure $P(r)$ is much less than the ram pressure of the accreting gas $\rho u^2 \propto r^{-5/2}$. If the gas is isothermal ($\Gamma = 1$) there is no doubt that free-fall can be achieved for sufficiently small distances. However, for a general adiabatic flow ($\Gamma > 1$) this condition is not guaranteed, since, the increase in gas pressure due to the compression induced by its infall could, in principle, surpass the gain in kinetic energy due to the infall.

By solving the the mass conservation equation, $\rho \propto 1/r^2 u(r)$, for the density ρ and substituting it into the polytropic equation of state, one can obtain the following expression for the pressure of the accreting as a function of the radius r (assuming the free-fall law for the velocity $u(r) \propto r^{-1/2}$),

$$P_g \propto r^{-3\Gamma/2} < P_{ram} \propto r^{-5/2}$$

This last inequality then translates to a condition for the adiabatic index $\Gamma < 5/3$. Therefore, in order that a gas be in free-fall in the gravitational potential of a black hole it should be less "stiff" than a monatomic gas. For $\Gamma = 5/3$, the gas is almost in hydrostatic equilibrium in the sense that its pressure gradient matches the force of gravity. The speed of sound increases as the radius r decreases ($a^2 \propto 1/r$), and the accretion flow never becomes supersonic (except close to the horizon where the acceleration of gravity becomes stronger than its Newtonian value).

In reality, the accreting gas is presumably fully ionized due to the compressional heating. The subsequent, line, or inverse Compton cooling, can decrease the adiabatic index sufficiently to allow its free-fall inflow. However, it was realized (Shapiro, 1974) that spherical accretion, coupled with free-free emission by the compressed and heated gas is an inefficient means of radiation production. Because of the very low radiative efficiency of free-free emission, most of the available thermal energy is advected into the black hole and never radiated away.

The problem of low radiative efficiency prompted the consideration of disk rather than spherical accretion to provide the required radiative efficiency. In the case of disk accretion, matter is considered to be in Keplerian orbit, slowly sinking toward the black hole by dissipat-

ing and radiating away its azimuthal kinetic energy. The difference between disk and spherical accretion is that, in disk accretion, matter can accrete only by losing (radiatively) its azimuthal kinetic energy, thus guaranteeing high efficiency. No advection of thermal energy into the black hole occurs in this case.

The major problem with modeling disk accretion is that the viscosity responsible for the dissipation of the azimouthal kinetic energy is completely unknown, a fact which makes quantitative predictions of this model difficult (Shapiro and Teukolsky, 1983). Nonetheless, accretion disk models have been used extensively to model the emission of AGN, especially in the optical-UV part of the spectrum, as discussed later. These models assume that the disk emits like a black body (Malkan, 1983), which requires the viscosity of the disk be sufficiently small to allow it to be optically thick to free-free absorption. This is not a severe restriction except for the fact that it is not consistent with the observed optical-UV variability. There is little doubt that, at least near the black hole, the dynamics of accretion are dominated by the rotation, since, even a small amount of angular momentum would dominate the dynamics at small distances. (Remember that a fluid element with angular momentum L "sees" an effective potential "$L^2/2r^2$".) The issue then is what are the parameters of the resulting disk, i.e., density and temperature, which would determine the character of the radiation we observe.

An alternative class of models, based mainly on the spherical mode of accretion is that of Kazanas and Ellison (1986a). These models use a standing shock surrounding the black hole as a means for randomizing the accretion kinetic energy and slowing down the infall, thereby increasing the radiative efficiency. Similar models are those of Mészáros and Ostriker (1983) and Babul et al. (1989). The problem with these models is that formation of a shock in spherical accretion requires a boundary condition on the horizon other than simple free-fall. The model of Kazanas and Ellison (1986a) addresses this issue by providing the required back-pressure for shock formation with a sufficiently high concentration of relativistic protons which are not pulled by gravity and do not fall into the black hole.

These relativistic protons are thought to be produced by the Fermi acceleration mechanism at the same shock whose existence they guarantee by their presence. Protons are necessary for such a model, as relativistic electrons suffer losses on time scales much shorter than the free-fall time and cannot provide the required back-pressure to ensure the formation of such a shock.

The interesting feature about this model is that once a mechanism can convert a sizable fraction of the accretion power into relativistic protons, the subsequent formation of the shock is almost inevitable. In addition, the dissipation of energy within this model is understood very well. The dissipation time scale is the proton-proton collision time scale $t_{pp}^{-1} = n(r)c\sigma_{pp}$, ($\sigma_{pp} \simeq 3 \cdot 10^{-26}\, cm^{-2}$ is the strong interaction cross section) and it is *computable* from first principles once the density of the ambient gas is given. The position of the shock is then set at $r_s \simeq v_{ff} t_{pp}$. The dynamics of accretion in this case depend on the ratio of the dissipation time scale to the free-fall time scale $t_{ff}/t_p \propto \dot{M}/M_9 \sim \dot{M}/\dot{M}_{Edd}$. One should not that this ratio does not depend on M and $\dot{M}$ separately! Kazanas and Ellison (1986a) indicate that AGN cover a rather limited range in this parameter, despite the wide range of their luminosities.

2. Radiative Losses

The energy distribution across the AGN spectrum suggests the presence of relativistic particles within the AGN central engine. The origin of these relativistic particles is not certain. However, given that velocities approaching that of light are present, shocks could be the mechanism for producing nonthermal particle populations (F. C. Jones, this lecture series).

The main energy loss process for relativistic electrons in the AGN environment is that of inverse Compton scattering, in which a photon of energy ϵ increases its energy by scattering with an electron of higher energy, considered to be relativistic of Lorentz factor γ. If θ is the angle between the electron and the photon, in the electron rest frame the photon energy will be $\epsilon' = \gamma\epsilon(1 - \beta cos\theta)$. If this energy is much smaller than the electron rest mass, then the scattering takes place in the Thomson regime, and the effects of electron recoil and the quantum structure of the electron can be ignored. The electron in this approximation behaves like a billiard ball in its interaction with the photon. The collision will be elastic in the electron rest frame. Retransforming the final photon energy back to the lab frame, we obtain $\epsilon_f \simeq 2\gamma^2\epsilon(1 - \beta cos\theta)$, i.e., the final energy is $\simeq 4\gamma^2$ times larger than the original. To calculate the total rate of energy loss by the electrons one has to integrate the rate of interactions $\sigma_T c(1 - \beta cos\theta)$ times the energy gained per interaction over all angles and over all photon energies weighting each photon with their spectral distribution $n(\epsilon)d\epsilon$. The assumption that the target photon distribution is isotropic gives the value 4/3 for the integral of $(1 - \beta cos\theta)^2$, over all angles and yields for the total electron loss

$$\dot{E} = \frac{4}{3}\int \gamma^2 \sigma_T c n(\epsilon)\epsilon d\epsilon \quad \text{or} \quad \dot{E} = \frac{4}{3}\gamma^2 \sigma_T c \rho_{rad}$$

where ρ_{rad} is the total energy density of the radiation field present. The latter quantity is given in terms of the total luminosity L and the radius of the source r by $\rho_{rad} = L/\pi R^2 c$ and given that $\dot{E} = m_e c^2 \gamma$, the electron

Compton loss time scale is

$$t_C = \frac{\gamma}{\dot{\gamma}} \simeq \frac{m_e c^2 \pi}{\gamma \sigma_T} \frac{R^2}{L}$$

Compare the expression above to the light crossing time through the source $t_{lc} = R/c$

$$\frac{t_{lc}}{t_C} \simeq \frac{L}{R} \frac{\sigma_T}{m_e c^3} \gamma$$

This ratio depends only on the combination L/R of the parameters of the source which is called *compactness* and it can be expressed in terms of the natural units associated with accretion

$$\frac{t_{lc}}{t_C} = \frac{2\pi}{3} \left(\frac{m_p}{m_e}\right) \frac{L}{L_{Edd}} \left(\frac{3R_S}{R}\right) \gamma \qquad (10.8)$$

The lifetime of electrons (even those with $\gamma \simeq 1$), in a source emitting close to its Eddington rate, is much shorter than the light crossing time through the source. This condition guarantees that the pressure of relativistic electrons cannot have any dynamical effects in AGN and that there is need for an efficient accelerator of electrons if these particles play any role in the formation of AGN spectra.

The above estimates hold for any process which operates at the Thomson cross section rate: the particular value of the cross section guarantees that the radiative losses will be faster than any dynamical change in the system. Such a process is the production of e^+e^- pairs in photon-photon collisions. This process is possible for photons with energies $E_\gamma > m_e c^2$ interacting with photons of lower energy ϵ fulfilling the threshold condition $E_\gamma \epsilon \gtrsim (m_e c^2)^2$. The opacity to this process is effectively the ratio of the photon escape time to the photon interaction time scale to produce a e^+e^- pair. Because the γ-ray photons interact with photons of lower energy to produce a pair, the expression for the opacity depends on the spectral distribution of lower energy photons. If the energy spectrum of the lower energy photons is $\propto \epsilon^{-\alpha}$ the latter expression reads

$$\tau_{\gamma\gamma}(E_\gamma) \simeq \frac{2\pi}{3} \left(\frac{m_p}{m_e}\right) \frac{L}{L_{Edd}} \left(\frac{3R_S}{R}\right) \left(\frac{E_\gamma}{m_e c^2}\right)^\alpha \qquad (10.9)$$

One should note the emergence, in the above arguments, of a compactness, namely the combination $\pi m_e c^3/\sigma_T$, which depends only on natural constants, namely σ_T, m_e, c. The value of this quantity in cgs units is $\pi m_e c^3/\sigma_T \simeq 10^{29}\ erg\ s^{-1}\ cm^{-1}$. This value (which would make the opacity of MeV γ-rays approximately one) is equivalent to having a solar luminosity of high-energy radiation being emitted from a room of size of a few tens of meters; nonetheless, this value is still much smaller than the characteristic compactness associated with accretion problems i.e., $L_{Edd}/R_S \simeq 3 \times 10^{32}\ erg\ s^{-1}\ cm$, implying that γ-rays cannot escape to the observer if created at regions close to the black hole.

3. Relativistic Beaming

The term "relativistic beaming" is used to indicate the perceived amplification of the radiation emitted by plasma moving with a bulk Lorentz factor $\Gamma \gg 1$, if viewed at a sufficiently small angle θ with respect to its direction of motion (typically $\theta \lesssim 1/\Gamma$). Light travel time effects cause any intrinsic variability of the emission to be shortened by one power of the Doppler factor, $\delta \equiv [\Gamma(1 - \beta cos\theta)]^{-1}$. At the same time, conservation of the photon phase space implies both an increase of the observed radiation flux density by $\delta^{3+\alpha}$ (where the intrinsic emission has an intensity of the form $F_\nu \propto \nu^{-\alpha}$) and also an increase in the observed bolometric luminosity by δ^4 (Rybiki and Lightman, 1979). Thus, emission from plasma in bulk relativistic motion could appear extremely luminous and yet vary on time scales short compared to the light travel time across the emitting source.

If a "blob" of plasma moves relativistically at an angle θ with respect to the observer's line of sight, its apparent transverse velocity will be $v_{app} = c\beta\ sin\theta/(1-\beta cos\theta)$. The factor in the denominator comes about because the "blob" chases the light which itself emits and, therefore, wavefronts emitted at points separated by a distance L come to the observer separated only by a distance $L(1 - \beta cos\theta)$. The above expression for v_{app} has a maximum value $v_{app,m} = \Gamma c$ for $cos\theta = \beta$ or $\theta \simeq 1/\Gamma$, indicating that such "blobs" will appear moving superluminally on the plane of the sky.

While the observation of radio "blobs" moving superluminally has unequivocally established the presence of bulk relativistic motion, its effects on the spectra of these sources, though potentially extremely important, are far from being firmly established in a quantitative manner, mainly because the emitted radiation depends on a large number of parameters, each of which can significantly affect the resulting spectrum.

10.3 AGN PHENOMENOLGY

Armed with the knowledge of the physical quantities and the processes relevant to the study of AGN physics, we now can discuss some associated phenomenology and see how these processes are applied in the observational context.

One feature that distinguishes AGN from other galaxies, is their spectra: Figure 3.1 shows the spectra of a number of AGN in a νF_ν *vs.* ν diagram, in which the luminosity per decade is plotted as a function of frequency. The striking features of this diagram are its extent in frequency (it covers roughly 12 decades for certain objects extending from radio to X-rays and gamma-rays) and the fact that the energy per decade seems to be roughly constant across the electromagnetic spectrum.

These spectra are radically different from those of

normal galaxies, that (representing the ensemble of the spectra of the resident stars) span a much smaller range in frequency, from IR to UV. It is precisely these two properties that have prompted several authors (Jones, O'Dell and Stein, 1974) to suggest that nonthermal processes are responsible for the emission of radiation in AGN. These authors have suggested that the observed radiation is the result of synchrotron emission from a population of relativistic electrons (spanning the radio to UV frequency range), coupled with inverse Compton scattering of the synchrotron radiation by the same relativistic electrons (to produce observed the X-ray and gamma-ray radiation). This model, referred to as "synchrotron-self-Compton " (or SSC), has been the bedrock of the nonthermal models of AGN emission. Interestingly, the observed spectral energy distribution in AGN (of roughly equal energy per decade) is that expected by the SSC process, if the radiating relativistic electron population was the result of acceleration in a strong shock. The form of the spectral energy distribution in AGN has been a motivating force in the development of such models (Protheroe and Kazanas, 1983; Kazanas and Ellison, 1986a).

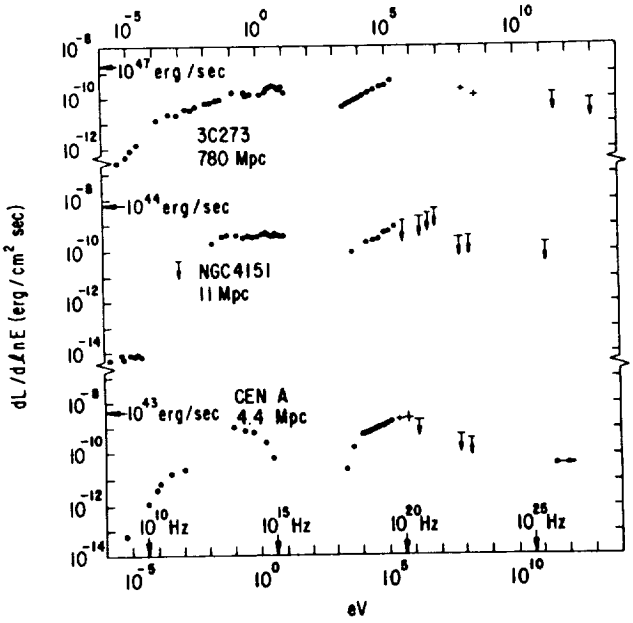


Figure 10.1: Energy distribution in AGN across the electromagnetic spectrum (from Ramaty and Lingenfelter, 1982).

The broad spectral distribution of the AGN emission does not necessarily guarantee the nonthermal character of the radiation, despite the comments given above. Reprocessing of high frequency radiation by matter distributed over a large range of radii could stretch the emission into the IR and far IR part of the spectrum. More careful examination of the spectral distribution in the optical-UV range uncovered a hardening of the spectrum towards high frequencies. This feature, ter-med "the Blue Bump," has been interpreted and modeled as the spectrum of an accretion disk emitting thermal radiation (Malkan, 1983; Sun and Malkan, 1989). It was suggested (Sanders et al., 1989) that the emission in the infrared, rather than being nonthermal, is due to reprocessing of the Blue Bump photons by dust present in a region of size $\sim$ 1 - 100 parsec ($\sim 10^{18} - 10^{20}$ cm) surrounding the AGN "central engine." In fact, the absence of AGN variability in the IR spectra of Seyfert galaxies (Edelson and Malkan, 1987) supports this conclusion and suggests that, at least for this class of objects, the IR emission, thought to be due to synchrotron radiation in the context of the SSC models, is probably due to thermal emission due to dust.

Subsequent observations have made clear the distinction of AGN in two major classes: the radio quiet and the radio loud. While quite similar in their spectral distribution from IR to UV and X-rays, the spectral radio properties of these two classes are distinctly different: Besides the more powerful radio emission associated with the radio loud compared to radio quiet AGN, their distinction can, in fact, be made on the basis of the radio emission in relation to their bolometric luminosity alone: The 1 GHz luminosity of the radio quiet AGN comprises only $\sim 10^{-6}$ of their bolometric luminosity while that of the radio loud is a much larger fraction of it, $\sim 10^{-3}$. In addition, the emission in radio loud AGN is continuous between the radio and the IR-O-UV bands (see Figure 10.2), indicating that a single emission mechanism operates across this entire frequency range; this is suspected to be synchrotron radiation, based on the observed polarization and variability of the radio - UV spectra of radio loud AGN (however, some fraction of it might also be due to reprocessing from dust as is thought to be the case in radio quiet AGN). By contrast, the emission in radio quiet AGN essentially cuts off at frequencies $\lesssim 10^{12}$ Hz. The observed radio emission in the GHz band from this AGN class, besides being much a smaller fraction of the bolometric emission, it is also extended, indicating processes at work far from the AGN central engine.

The distinction of AGN in radio loud and radio quiet has been made more clear with the recent observations by the EGRET instrument aboard the Compton Gamma-Ray Observatory, which detected roughly 55 AGN at energies in excess of 100 MeV, all belonging to the general class of radio loud AGN and, in particular, to its subcategory of blazars (see Figure 10.2). Members of the blazar class exhibit high polarization and large amplitude variability from optical to radio frequencies). Most important, the γ-ray luminosity, in most objects, outshines that in any other band, sometimes by more than a factor of 30. The discovery of high-energy gamma-rays

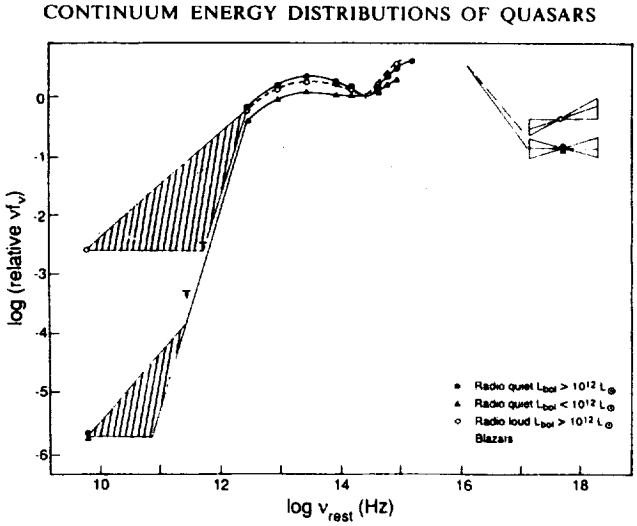


Figure 10.2: Schematic continuum energy distribution in quasars (Sanders et al., 1989). Note the difference in the fraction of the total luminosity emitted in the radio $\simeq 10^{10}$ Hz for radio loud and radio quiet AGN.

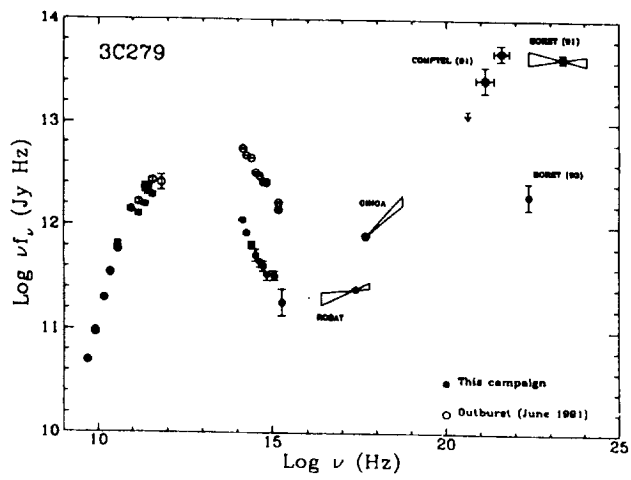


Figure 10.3: The energy distribution from radio to γ-rays in the radio loud quasar *3C 279* (from Marachi et al., 1994). The EGRET results are not simultaneous with those at other frequencies. The X-ray flux during the EGRET observations was closer to that of the June 1987 observation.

from this particular class of AGN confirmed the suspicion that nonthermal processes (i.e., SSC mechanism) are at work at least in this class of AGN and that observations of radio emission imply the presence of highly relativistic electrons in regions of high photon densities, naturally accounting for the observed high-energy gamma-rays.

The level of high-energy γ-ray emission exceeded the the most optimistic estimates based on the SSC mechanism. Indeed, the observed γ-ray flux implies, under conditions of isotropic emission from a stationary source, γ-ray luminosity in excess of 10^{48} $erg\ s^{-1}$. Such prodigious power outputs are exceptional even by AGN standards and pose severe problems with their fueling. Of particular interest is the fact that the EGRET observations cover the energy range in which where the gamma-gamma opacity should be very large ($\sim 10^3$) if the implied luminosity is emitted by a source of size consistent with the observed gamma-ray variability. In such a case, the large opacity should manifest itself as a sharp break in the spectra of the EGRET sources. Figure 3.3 shows the spectrum of such a source namely that of the quasar *3C 279*. As it is apparent, no obvious breaks are present in its gamma-ray spectrum, thus precluding the presence of significant gamma-gamma absorption.

The absence of gamma-ray absorption has been considered as a model independent evidence for the presence of beaming in this class of objects: the absence of gamma-ray absorption implies that the instrinsic luminosity is much smaller than that implied by multiplying the observed flux by $4\pi D^2$, where D is the luminosity distance to the source. Because relativistic beaming boosts the apparent luminosity by δ^4, values of $\delta \simeq 10$ could render the γ-ray observations compatible with the gamma-ray absorption constraints.

The presence of radiation of such high-energy from a relativistically moving fluid puts several severe restrictions on the dynamics and on the associated particle acceleration. Although bulk relativistic motion in AGN has been taken as a given (as the natural explanation of the apparent superluminal motion of radio "blobs"), models which could account in a simple fashion for the acceleration of an electron (actually e^+e^- for neutrality) plasma to bulk Lorentz factors $\Gamma \simeq 10$ simply are not readily available. The difficulties are compounded by the additional demand that the individual electron Lorentz factors, in the fluid rest frame, be in excess of 10^4. Models that can fulfill these constraints do exist but it is not apparent that any of them can address the issue of beamed AGN emission with a sufficiently small number of parameters to provide a coherent model for the emission of radio loud AGN across the electromagnetic spectrum. While most models assume the acceleration of relativistic electrons *in situ* by shocks within a given relativistic outflow, the recent observation of *Mkn 421* in the *TeV* range with the atmospheric Cerenkov imaging technique (Punch et al., 1992) (the luminosity in this energy band is similar to that in the O-UV and EGRET range), presents difficulties for this mechanism, too. The energy loss rates for electrons of these energies would be so fast that it is generally assumed that the particles responsible for the *TeV* radiation are protons. This fact suggests that this may indeed be the case for all the other EGRET sources,

a point which we will not pursue further at present.

The issue of detailed AGN classification, despite intense study over three decades, is far from complete. A great leap forward has been made in the past several years with polarization observations of radio quiet AGN which indicated that Seyfert 1 galaxies (which exhibit prominent broad lines and strong X-ray emission) and the Seyfert 2 ones (of strong narrow lines and weak X-ray emission) are manifestations of the same entity (a point-like source surrounded by an optically thick torus) viewed from different angles with respect to the torus axis. Similar arguments have also been made for the interpretation of subclasses within the more general class of radio loud AGN. This subject is a very interesting and fertile one. For more information and developments the reader can consult the recent review by Antonucci, 1993; Osterbrock, 1991.

The most interesting fact about AGN classification, in my opinion, is that such schemes can exit at all. The extent of AGN emission, both in frequency and in radius, suggests that modeling this emission can easily involve 10-20 parameters. Clearly, neither all these parameters are important nor the associated parameter space is not covered uniformly, because, in such a case it would be extremely hard to discern any distinct AGN classes. To see why this is indeed the case, one has only to bear in mind that the entire classification of stars in main sequence, red giants and white dwarfs (plus the stellar evolution) can be achieved to a large extent with only one parameter, namely, their mass. In light of this argument, the fact that any AGN classes can exist at all, means, at least to me, that AGN are much more constrained than the freedom of our models seems to imply. Understanding the physics of AGN, therefore, boils down to discovering the additional pieces of physics that limit the apparent freedom of our current models. In this direction, I think, we have got quite a way to go.

10.4 REFERENCES

Antonucci, R., 1993, Ann. Rev. As. Ap., 31, 473.

Babul, A., et al., 1989, ApJ, 347, 59.

Edelson, R.A., and Malkan, M.A., 1987, ApJ, 323, 516.

Hartman, R.C., et al., 1992, ApJ (Lett.), 385, L1.

Jones, T.W., O'Dell, S.L., and Stein, W.A., 1974, ApJ, 188, 353.

Kazanas, D., and Ellison, D.C., 1986a, ApJ, 304, 178.

Malkan, M.A., 1983, ApJ, 268, 582.

Maraschi, L., et al., 1994, ApJ, 435,L91

Meszáros, P.M., and Ostriker J.P., 1983, ApJ (Lett.), 273, L59.

Osterbrock, D., 1991, Rep. Prog. Phys., 54, 579.

Protheroe, R.J., and Kazanas, D., 1983, ApJ, 265, 620.

Punch, M., et al., 1992, Nature, 358, 477.

Ramaty, R., and Lingenfelter, R.E., 1982, Ann. Rev. Nuc. Part. Sci., 32, 235.

Rybiki, G.B., and Lightman, A.P., 1979, "Radiative Processes in Astrophysics" Wiley.

Sanders, D.B., et al., 1989, ApJ, 347, 29.

Shapiro, S.L., 1974, ApJ, ApJ, 240, 246.

Shapiro, S.L., and Teukolsky, S.A., 1983, in "White Dwarfs, Neutron Stars, Black Holes, The Physics of Compact Objects" Wiley.

Sun, W.S., and Malkan, M., 1989, ApJ, 346, 68.

Chapter 11

EXTRAGALACTIC GAMMA-RAY BACKGROUND

NEIL GEHRELS AND CYNTHIA CHEUNG
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

We first provide a historical overview on the observations and theoretical interpretations of the extragalactic gamma-ray background that were obtained before the launch of the Compton Gamma Ray Observatory. We then summarize the recent observations by Compton that may help to explain the origin of the extragalactic gamma-ray background.

11.1 INTRODUCTION

The question of whether cosmic gamma rays can be detected among primary cosmic rays has been investigated since the earliest balloon-borne experiments (Hulsizer and Rossi, 1949). At that time, the only measurable gamma-ray fluxes identified as extraterrestrial in origin were from solar flares. The rest were attributed to the Earth's radiation belts. It was not until the launch of deep space probes that carried instruments far from the Earth's high radiation environment did scientists measure an interstellar flux of gamma rays. An extraterrestrial and non-solar component of gamma-rays was first positively detected by the Ranger 3 Moon probe (Arnold et al., 1962). Historically, the measurement of the diffuse gamma-ray background (GRB) was beset by great experimental difficulties. The high charged particle intensities in the balloon and near-Earth space environment induce an intense radioactive background in gamma-ray instruments. Most early instruments also had large fields-of-view and poor spatial resolution, which added complexities to the extraction of the diffuse GRB. The Orbiting Solar Observatory-3 (OSO-3) (Kraushaar et al., 1972) was the first satellite to carry a directional detector and the analyses showed that the detected cosmic gamma rays could be separated into two components: a highly anisotropic galactic component, and an isotropic extragalactic component with a softer energy spectrum. Subsequently, data were obtained by many balloon- and satellite-borne instruments in the last 30 years (Table 11.1). Each experiment attempted to achieve greater sensitivity by increasing the detector size and/or using better background-reduction techniques.

Many science students have become familiar with the infrared background radiation that is the relic of the Big Bang due to the well-known measurements made by the Cosmic Background Explorer (COBE). But many are not aware that an extragalactic component of the diffuse background is detected in all wavelength regimes of the electromagnetic spectrum, including X-rays and gamma rays. Their origins are not that well understood (see for example, articles in Calzetti et al., 1995). Recent ROSAT observations (Hasinger et al., 1993) seem to have resolved the puzzle concerning the diffuse X-ray background which was discovered by the first X-ray satellite Uhuru. The low-energy X-ray background is likely to be the integrated emission of active galactic nuclei (AGN) (Zdziarski et al., 1995; Chen et al., 1996). However, in the gamma-ray regime, five years after the launch of the Compton Gamma Ray Observatory (Compton), there is still no general consensus on the origin of the extragalactic GRB, although the new data are beginning to shed some light on the problem.

The extragalactic GRB is of particular interest to cosmology because of the transparency of the Universe to gamma rays back to redshifts of 100 or more. A determination of the isotropy of the GRB can be a sensitive test of its cosmological origin and a detailed study of the complete GRB spectrum and its spatial fluctuations may provide constraints to different cosmological models. There are several possible origins for the extragalactic GRB: it is truly diffuse, it is the integrated emission of various distant unresolved gamma-ray sources, or it may be a combination of diffuse and point sources and each component may predominate in different portions of the gamma-ray band. This is according to our understanding of the radiative processes that produce gamma-rays at these possible sites and how their relative importance vary according to the energy range (Stecker et al., 1993). From below 10 MeV to 100 MeV, the most likely gamma-ray production mechanisms are electron-positron annihilation, bremsstrahlung between cosmic ray electrons and interstellar nuclei, and inverse Compton interactions between cosmic ray particles and lower- energy photons. Above 100 MeV, the dominant process is likely πo decay from nucleon interactions.

Though some of the earlier experimental results gave only upper limits, the general form and magnitude of the GRB spectrum had been measured with reasonable accuracy before the launch of Compton (Figure 11.1). The low-energy portion (10 keV to 60 keV) was characterized by a bremsstrahlung spectral shape that could be approximated by a power-law segment of energy index ~0.4. It transitioned to a power law of index ~1.6 above 60 keV, flattened to an index of ~0.7 around 1 MeV, then steepened again above several MeV to an index of ~1.7 that extended to energies well above 100 MeV. When plotted in intensity per logarithmic energy interval EIE , the GRB spectrum exhibited two peaks, one at ~30 keV due to thermal bremsstrahlung and the other at a few MeV (called "the MeV bump") (Figure 11.2). There were uncertainties concerning the measurements in the 10 to 30 MeV range because of the extraordinary high instrumental and atmospheric background. Recent results by the Imaging Compton Telescope (COMPTEL) instrument on Compton have revised downwards the measurements in this range and called into question the existence of the MeV bump (Kappadath et al., 1995). Better data have also been obtained for the higher energies by the Energetic Gamma Ray Experiment Telescope (EGRET) instrument (Sreekumar and Kniffen, 1996). The updated GRB spectrum (Zdziarski, 1996) is given in Figure 11.3.

In this chapter we summarize the state of our knowledge on the extragalactic GRB. We present the observations and their interpretations in two parts: (1) for energies from 0.1 to 10 MeV, and (2) for energies greater than 10 MeV. We then present the recent results from Compton observations and discuss their relevance to the GRB.

Table 11.1: Observations of the Diffuse Gamma-Ray Background (adapted from Gruber (1992))

Energy Band	Detector	Platform	Reference
0.1-3 MeV	scintillator	Ranger 3	Arnold et al. (1962)
1-6 MeV	scintillator	ERS 18	Vette et al. (1969)
1-5 MeV	scintillator	balloon	Vedrenne et al. (1971)
0.1-8.5 MeV	scintillator	balloon	Daniel et al. (1972)
0.03-4.1 MeV	scintillator	Kosmos	Mazets et al. (1975)
0.1-4 MeV	scintillator	balloon	Fukuda et al. (1975)
0.3-10 MeV	scintillator	Apollo	Trombka et al. (1975)
2-10 MeV	Compton telescope	balloon	White et al. (1977)
0.3-6 MeV	scintillator	balloon	Mandrou et al. (1979)
3-60 keV	proportional counter	HEAO-1 A2	Marshall et al. (1980)
1.1-10 MeV	scintillator	balloon	Schönfelder et al. (1980)
0.013-4.0 MeV	scintillator	HEAO-1 A4	Gruber et al. (1985)
0.45-10 MeV	Ge spectrometer	HEAO-3	Wheaton et al. (1995, 1996)
> 50 MeV	scintillator/Cerenkov	OSO-3	Kraushaar et al. (1972)
30-50 MeV	spark chamber	balloon	Pinkau et al. (1973)
35-200 MeV	spark chamber	SAS-2	Fichtel et al. (1978)
4-100 MeV	spark chamber	balloon	Lavine et al. (1982)

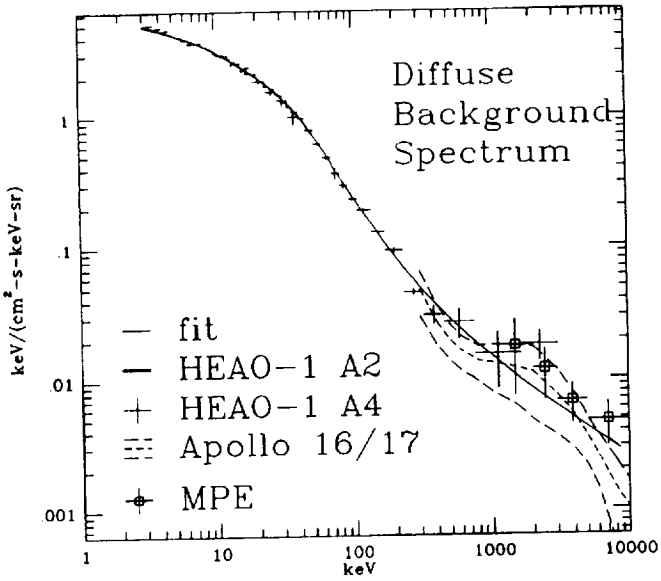


Figure 11.1: The pre-Compton energy spectrum of the diffuse extragalactic background (Gruber, 1992).

11.2 LOW-ENERGY MEASUREMENTS (0.1 – 10 MeV)

11.2.1 Spectral Characteristics

The first measurement of the GRB was obtained by a CsI scintillation detector onboard the Ranger 3 Moon probe in 1962 in the energy range of 0.1 to 3 MeV (Arnold et al., 1962). The lunar missions Apollo 15, 16 and 17 in the early 1970s carried identical NaI scintillation spectrometers and obtained data in the 300 keV to 10 MeV energy range (Trombka et al., 1977). The measured spectrum was well-fitted by an electron bremsstrahlung model after subtracting the unwanted background components caused by charged particle activation of the detector and the spacecraft. The data points between 100 keV and 10 MeV were supplemented by measurements from balloon experiments (Vedrenne et al., 1971; Daniel et al., 1972; Fukuda et al., 1975). The results generally agreed within experimental uncertainties and connected smoothly to data between 30 MeV to 200 MeV from the Small Astronomical Satellite–2 (SAS-2), the first high-energy gamma-ray satellite that was launched in 1972 (Fichtel et al., 1975) (Figure 11.2). There was some indication of a flatter slope in the 1 to 5 MeV region (Figure 11.1). In the EIE plot (Figure 11.2), this translated into a peak, called the MeV bump.

The High Energy Astronomy Observatory - 1 (HEAO-1) was launched in 1977 with one of its primary objectives to measure the spectrum of the high-energy cosmic diffuse background. Two of the experiments, the Goddard Space Flight Center's Cosmic X-Ray Experiment (A-2) and the Hard X-Ray and Low-energy Gamma-Ray Experiments (A-4) by the University of California, San Diego and the Massachusetts Institute of Technology, together spanned the X-ray to low-energy gamma ray regime from 3 keV to 4.0 MeV. They used different detection techniques, yet obtained two spectra

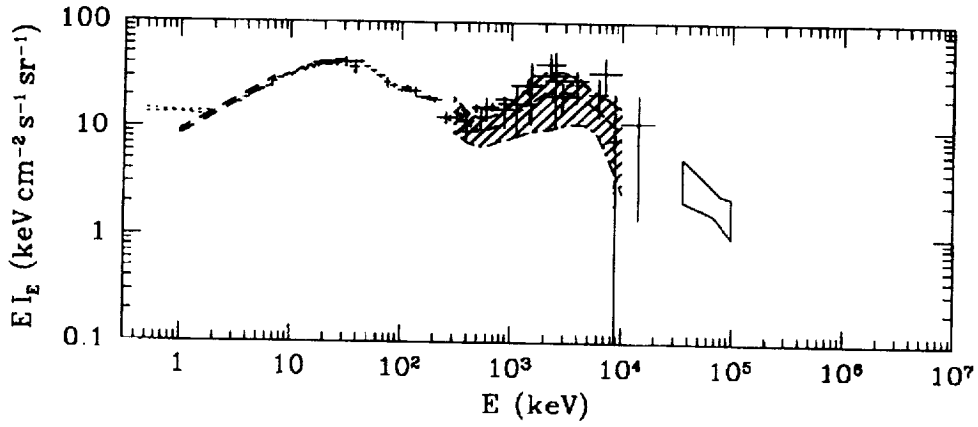


Figure 11.2: The extragalactic background spectrum in intensity per logarithmic energy interval (adapted from Zycki et al., 1993). Circles correspond to the data points compiled by Gruber (1992) in Figure 11.1. The elongated 0.5-2 keV error contour gives the ROSAT results (Hasinger, 1992). The dashed region is the systematic error contour for the Apollo results (Trombka et al., 1977), with the middle dot-dashed curve the best estimate of the spectrum. The 35-100 MeV error contour is the SAS-2 result from Fichtel et al. (1978).

that joined smoothly and fitted remarkably well to a single thermal bremsstrahlung model characterized by kT $\sim$ 40 keV in the transitional energy range of 60 to 100 keV (Marshall et al., 1980, Gruber et al., 1985). Besides measuring the GRB, the A-4 experiment also detected about 50 X-ray sources in the Galactic plane (Levine et al., 1984), the diffuse Galactic ridge that extends approximately $\pm 50°$ in longitude and $\pm 5°$ in latitude (e.g., Gehrels and Tueller, 1993 and references therein), and around 20 active galactic nuclei (AGN). The AGN have simple power-law spectra with energy indices clustered around the value of $\sim$0.7 (Mushotzky, 1982, 1984; Rothschild et al., 1983).

There was also evidence in the HEAO-1 A-4 data of spectral flattening around a few MeV. Data from balloon-borne Compton telescopes in the energy range of 1 to 20 MeV (Schönfelder et al., 1980; White et al., 1977) were consistent with the existence of the MeV bump. Beyond 5 MeV, the balloon data showed that GRB spectrum steepens to a power law of energy index $\sim$2 (Schönfelder et al., 1980), in excellent agreement with the extrapolation of SAS–2 measurements (Fichtel et al., 1978) towards lower energies.

Gruber (1992) has performed a best fit using all the available spectral data published prior to 1990 (Figure 11.1). He found that below 60 keV, the diffuse background energy flux was fitted well by a bremsstrahlung spectrum. But above 60 keV, he needed to add to the bremsstrahlung extension a power law component with an index of $\sim$0.7, which characterizes the energy spectra of AGN measured by HEAO-1 A-4 experiment. His empirical functional fit to the energy flux between 3 keV and 6 MeV (in units of $keV\ /\ cm^2 - s - keV - sr$) is:

$$7.877E^{-0.29}\exp\left(\frac{-E}{41.13\ keV}\right) \quad 3\ keV < E < 60\ keV$$

$$1652E^{-2.00} + (1.75 \times E - 0.70)\quad 60\ keV < E < 6\ MeV$$

Recent detailed analysis of A-4 Medium Energy Detector (MED) data by Kinzer et al. (1996) confirmed the spectral fit between 100 to 400 keV. However, the analysis also indicated that above 400 keV, scintillation detectors may not be able to discriminate between the internal instrumental background and true cosmic flux. This calls into question the accuracy of the data points above $\sim$500 keV and the existence of the MeV bump.

Data from other kinds of gamma-ray detectors can provide an independent measurement of the GRB. A germanium High Resolution Gamma-Ray Spectrometer was first flown on HEAO-3 in 1979 (Mahoney et al., 1981). A preliminary GRB spectrum that resulted from using a new data analysis approach (Wheaton et al., 1996) matched Gruber's fit from 50 keV to 150 keV.

11.2.2 Spatial Characteristics

Comprehensive studies have been performed on the spatial structure of the low-energy GRB using HEAO-1 data (Boldt, 1987). The HEAO-1 detectors had nominal fields- of-view of $3° \times 3°, 3° \times 1.5°$, and $3° \times 6°$. Shafer (1983), using the A-2 High Energy Detector (HED) data from 2.5 to 60 keV, derived a dipole amplitude of dI/I_0 = (0.510.2)%, with a peculiar velocity ν = 475 $\pm$ 165 $km\ s^{-1}$ and apex at l = 282°, b = 30°. Gruber (1992), using the A-4 Medium Energy Detector (MED) data from 95 to 165 keV, found a similar dipole anisotropy with $dI/I_0 = (2.2 \pm 0.7)\%$, ν = 1450 $\pm$ 440 $km\ s^{-1}$ and apex at

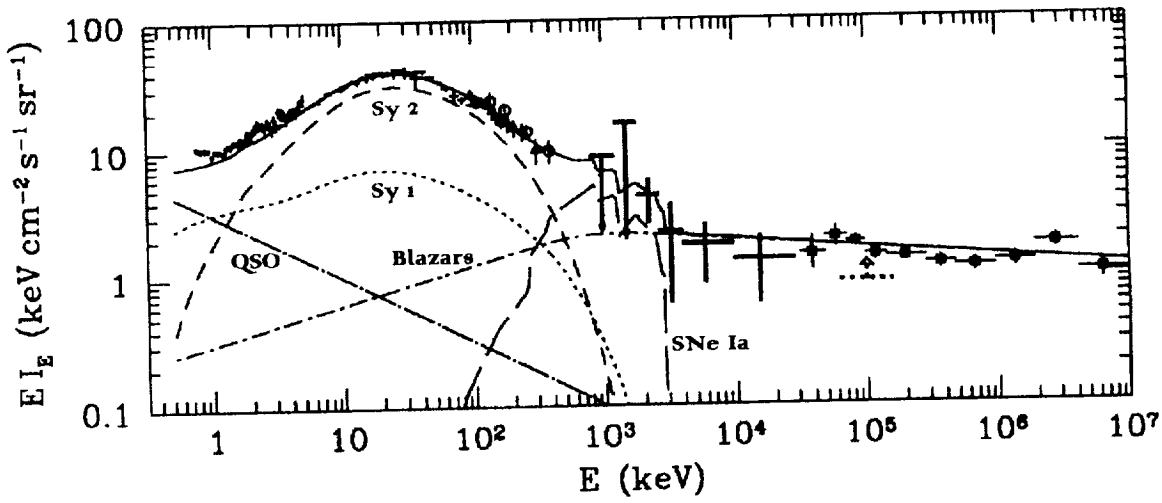


Figure 11.3: The updated extragalactic background spectrum in intensity per logarithmic energy interval (from Zdziarski, 1996). The crosses marked by filled circles below 5 keV are the ASCA results from Gendreau et al. (1995). Crosses in the 3–100 keV correspond to the data points from HEAO-1 A-2 and A-4 LED compiled by Gruber (1992) in Figure 11.1. The data marked by open triangles and open circles are from HEAO-1 A-4 MED (Kinzer et al., 1996), and Nagoya balloon (Fukada et al., 1975) measurements. The bold crosses in the 0.8–30 MeV range are from COMPTEL (Kappadath et al., 1995), and the filled squares in the > 30 MeV range are from EGRET (Kniffen et al., 1996).

$l = 304°$, b = 26°. These results are consistent, within the experimental uncertainties, with each other and also with the recent microwave result from COBE (Smoot et al., 1992, Kogut et al., 1993), which gives a peculiar velocity $\nu = 627 \pm 22$ $km\ s^{-1}$, and apex at $l = 264°$, b = 48°. Gruber noted that the large peculiar velocity measured in A-4 could be due to a local enhancement of the diffuse background.

11.3 INTERPRETATION OF LOW-ENERGY GRB SPECTRUM

The empirical bremsstrahlung shape of the low-energy GRB spectrum prompted early models to attribute its origin to thermal bremsstrahlung by a hot diffuse intergalactic gas at a temperature of $\sim 4 \times 10^8$ K (Marshall et al., 1980; Daly, 1988; Brown and Stecker, 1979; Olive and Silk, 1985). These models were essentially ruled out after the recent observations by COBE (Mather et al., 1990, 1994). The lack of significant deviations from a blackbody spectrum in the microwave background implies a much smaller Sunyaev-Zel'dovich (1980) y-parameter than predicted. The y-parameter is an indicator of the temperature of the intergalactic gas. So the COBE results exclude the possibility that the Universe was filled with hot intergalactic gas in the past (Terasawa, 1991; Rogers and Field, 1991).

Other models attempted to explain the low-energy GRB spectrum by the integrated flux of extragalactic sources. These were confronted by the well-known spectral paradox: AGN such as Seyferts and quasars which are the most likely contributors produce continuum spectra that are markedly different from that of the GRB (Mushotzky, 1984). The average AGN spectrum is characterized by a power law of energy index a ~ 0.7, whereas the GRB spectrum has an index of a ~ 0.4 below ~ 60 keV. To calculate properly the contribution of cosmological sources to the GRB, one needs to take in account the evolutionary effects of the AGN luminosity function. Avni (1978) found that the contribution from AGN depends on the form and amount of density evolution, on the deceleration parameter, and on the AGN formation epoch. The observed integrated flux is given by

$$I_N(E) = \frac{1}{4\pi} \int_0^{Z_F} dz \frac{dV(z)}{dz} \frac{(1+z) B_N(E(1+z), z)}{4\pi d_L^2(z)}$$

$$\frac{dV(z)}{dz} = \frac{4\pi c}{H_0} \frac{1}{(1+z)^3} \frac{1}{\sqrt{1+2q_0 z}} d_L^2(z)$$

where I_N is the flux in $keV\ cm^{-2} s^{-1}$- keV, B_N(E,Z) is the emissivity in $erg/(\text{s-}Mpc^3\text{-}keV)$, and Z_F is the epoch of first formation. V(z) is the comoving volume to redshift z, $d_L(z)$ is the luminosity-distance to redshift z (Weinberg, 1972), and q_0 is the deceleration parameter. For a pure density evolution where the source density $\rho(z)$ can be parametrized by $(1+z)^k$, the emissivity is

given by $B_N(E(1+z),z) = (1+z)^{k+\alpha} B_N(E, z=0)$, and α is the energy spectral index.

Since the AGN luminosity function in the gamma-ray regime is not well-known, work to estimate the AGN contributions to the GRB has to adopt extrapolations of X-ray or radio luminosity functions in the computation. Assumptions about the emissivity $B_N(E,Z)$ of a class of AGN also have to be made based upon the few measurements or existing upper limits of 'typical' AGN of that class. Rothschild et al. (1983) and Bassani et al. (1985) integrated the X-ray luminosity function of the HEAO-1 AGN derived by Piccinotti et al. (1982) to obtain the relative contributions of different classes of AGN. The results are shown in Figure 11.4. They found that Seyferts are likely to be the most important contributors to the GRB. Radio galaxies, quasars, field galaxies and BL Lac objects contribute only a small percentage. The Seyfert contribution increases from around 20% at energies $\sim 40\ keV$ till at $\sim 160\ keV$, they can account for all of the observed background. In fact, the observed spectral shape and intensity of GRB are used to set limits on the luminosity cutoff and spectrum break energy of AGN (Rothschild et al., 1983; Zycki et al., 1993) to prevent an overproduction of diffuse background above 100 keV. Indeed, recent AGN observations by Compton and GRANAT are indicating that Seyferts do typically have low-energy spectral breaks (see Section 11.6).

After subtracting the Seyfert contribution, a significant residual GRB remains below 100 keV that needs to be explained. Various authors have postulated the existence of different AGN populations in earlier epochs to explain the residual spectrum: for example, Eddington-limited thermal-type sources (Boldt, 1987), AGN precursors with high compactness (Leiter and Boldt 1992), and AGN with harder power-law spectra (Gruber, 1992). The recent ROSAT deep survey detected an unexpectedly high number of faint QSOs at $1 < z < 2$ (Shanks et al., 1991; Hasinger et al., 1993; Georgantopoulos et al., 1993). When the new QSO X-ray luminosity function (Boyle et al., 1993) with its derived evolution is integrated to a maximum redshift of $z_{max} \sim 4$, it is found that QSOs can account for 30% to 90% of the diffuse background at $\sim 2\ keV$. The respective contributions of different classes of AGN: Seyfert 1s, strongly-absorbed radio-quiet Seyfert 2s, and QSOs to the low-energy GRB have also been re-computed by Madau et al. (1993), Comastri et al. (1995), Zdziarski et al. (1995), and Chen et al. (1996) using the Boyle et al..'s (1993) observed X-ray luminosity function. They are plotted in Figure 11.3. It can be seen that the low-energy GRB up to the 100's keV range is well-explained by the integrated emission of various types of AGN.

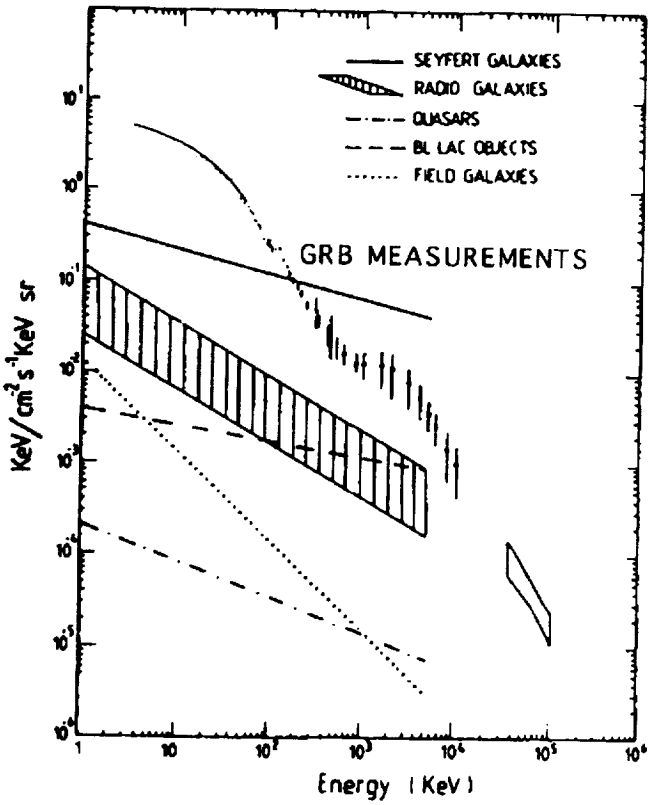


Figure 11.4: Contributions of the different classes of AGN to the GRB (Bassani et al., 1985)

11.4 HIGH-ENERGY MEASUREMENTS OF GRB (>10 MeV)

The first all-sky survey of cosmic gamma rays of energies above 50 MeV was carried out by OSO-3 in 1967-1968. The directional scintillator/Cerenkov detector recorded 621 "sky events" in 16 months of operations (Figure 11.5). The observations can be attributed to three components: (1) a Galactic component concentrated along the Galactic plane and well-correlated with the column density as deduced from 21-cm radio measurements; (2) a Galactic center component; and (3) an isotropic, extragalactic component with a steep power-law spectrum. Subsequent measurements made using an entirely different technique - spark chamber detectors - confirmed the general picture. These included the SAS-2 (Fichtel et al., 1975) mission and several balloon flights (Table 11.1). SAS-2 measured a diffuse GRB component with a very steep differential power law of energy index a ~ 1.7 (photon index = a+1 ~ 2.7) between 35 MeV and 200 MeV (Fichtel et al., 1978) (Figure 11.6). The extrapolated intensity of this component to lower energies agreed well with measurements at 10 MeV. Above several hundred

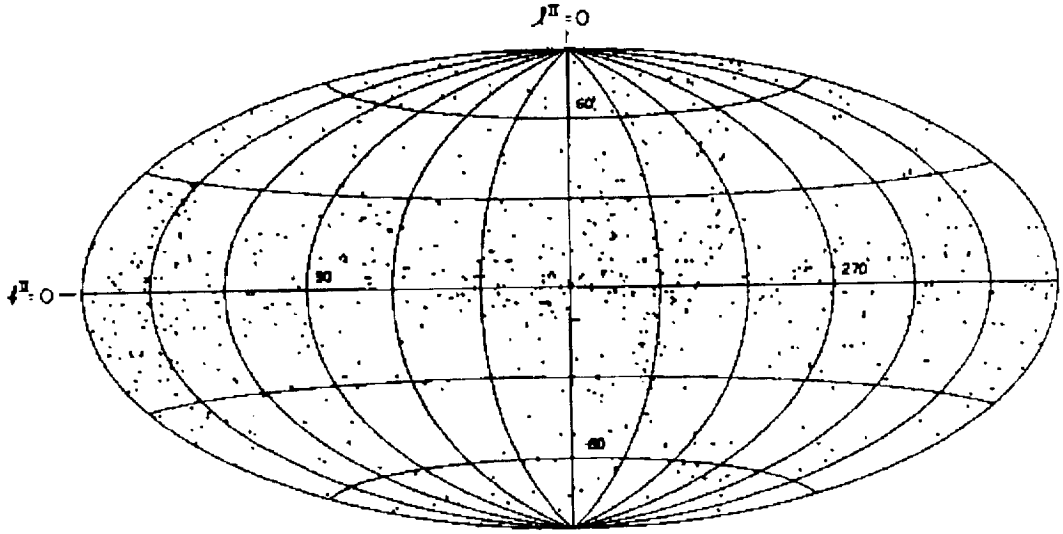


Figure 11.5: OSO-3 sky events plotted in galactic coordinates (Kraushaar et al., 1972)

MeV, the flux of the extragalactic GRB falls below the galactic high latitude background and the determination of its value is highly dependent on data analysis techniques.

11.5 INTERPRETATION OF HIGH-ENERGY GRB SPECTRUM

Before the launch of Compton, only one extragalactic source, 3C 273, had been detected above 10 MeV (Bignami et al., 1979). So theorists tended to explain the origin of the high-energy GRB spectrum by diffuse mechanisms rather than by a superposition of unresolved discrete extragalactic sources. This has now been changed by Compton observations of blazars by EGRET and the GRB measurements by COMPTEL (see Section 11.6), though the diffuse models can still explain some portions of the high-energy GRB spectrum. Below we describe four theoretical models that postulate a cosmological origin for the 1–100 MeV GRB.

11.5.1 GRB from Proton-Antiproton Annihilation

Stecker et al. (1971) proposed a diffuse emission model in which the GRB arises in a baryon symmetric Big Bang cosmology from matter-antimatter annihilations. In a grand-unified-theory model with spontaneous CP (charge parity) violation (Stecker, 1985), it is possible for the Universe to have evolved very large regions of pure matter and pure antimatter containing masses the size of galaxy clusters or superclusters in its early history. These regions are essentially the fossils of the vanished CP domains. In a baryon symmetric cosmology, annihilations occur at the boundaries between these regions at all redshifts to produce an extragalactic GRB. Puget (1973) has computed the annihilation rate as a function of redshift. The annihilations produce πo-mesons with gamma-ray producing decay modes. (πo decay is also the principal mechanism for producing the galactic high-energy diffuse gamma radiation.) Figure 11.7 shows a typical rest-frame spectrum produced by proton-antiproton annihilation with πo decay (Stecker, 1971), with maximum intensity at $m_\pi c^2/2 \sim 70\ MeV$. The spectrum is nearly flat near the maximum, with a minimum energy of $\sim$5 MeV and a maximum cutoff at $\sim$1 GeV. To arrive at the predicted GRB spectrum observed at the current epoch, it is necessary to solve a cosmological photon transport equation and take into account pair production and Compton scattering at high redshifts which may cause energy loss (Stecker et al., 1971). The resultant spectrum (Stecker, 1989) matches the observed steep slope of the extragalactic GRB above $\sim$1 MeV very well. Furthermore, it can account for the observed MeV bump (Figure 11.8). Absorption due to Compton scattering and pair production causes the spectrum to bend over below $\sim$1 MeV.

11.5.2 GRB from Primordial Black Holes

Page and Hawking (1976) postulated a population of primordial black holes (PBH) created in the early Universe that could explain the GRB. The PBHs cannot be created in the present epoch since the necessary compressional forces do not exist. They undergo quantum mechanical decay by radiating gravitons, neutrinos, elec-

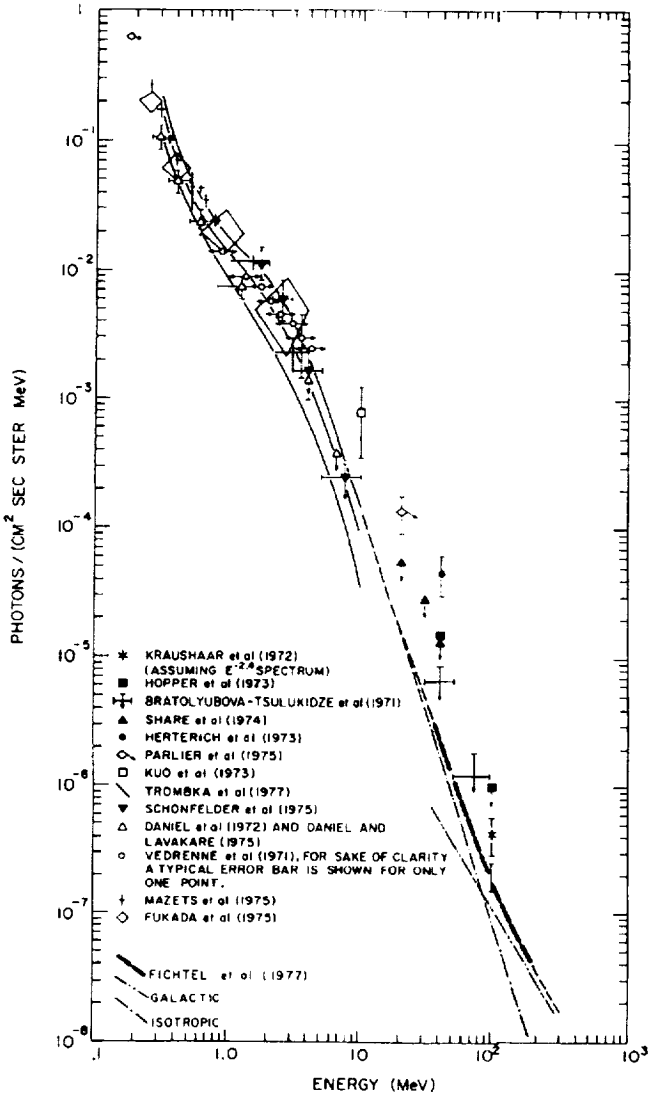


Figure 11.6: Differential photon spectrum of the SAS-2 diffuse gamma-ray background (Fichtel et al., 1978)

trons, positrons, and gamma rays. PBHs with initial masses less than a critical mass of 5×10^4 g would have completely evaporated by now. PBHs of slightly greater initial mass would be radiating energy at the rate of 2.5×10^{17} erg s^{-1}. As they reach the end of their life, they evaporate by ejecting all remaining rest mass in a very short time. The particles emitted in this final release would decay rapidly (t << 1 sec), giving a short burst of gamma rays between 100 MeV and 1 GeV. The EGRET instrument onboard Compton may be able to detect PBH bursts with this particular timing signature, but has not seen any such events to date. A uniform distribution of PBHs of initial mass $\geq 5\times10^4$ g, when integrated over the cosmological time scale, would give a GRB photon spectrum of power law index ~3 above 120 MeV. This matches the observed GRB spectral slope in

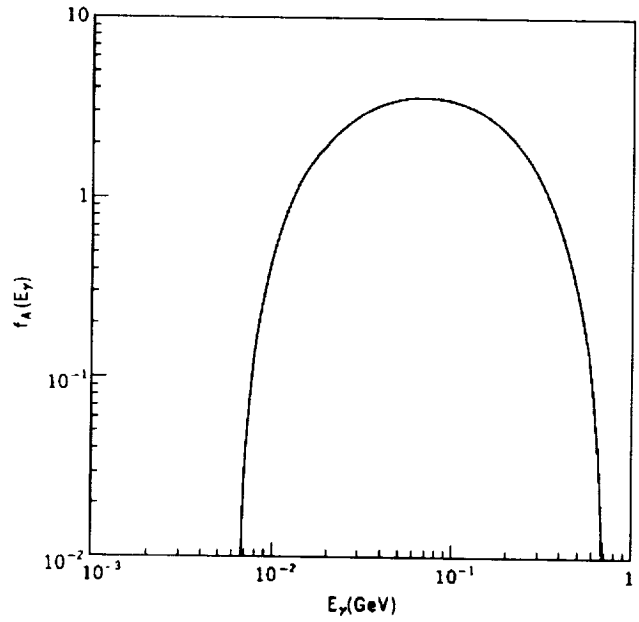


Figure 11.7: Gamma-ray spectrum from proton-antiproton annihilation at rest (Stecker, 1971)

the 100 MeV range rather well. Below 120 MeV, the spectrum may flatten depending on the space density of PBHs.

11.5.3 GRB from Deuterium Formation

Leventhal (1973) suggested that deuterium production by the two-body recombination process (p+n → D+2.22 MeV γ ray) in an environment transparent to the emitted gamma rays could give rise to the observed 1 MeV bump in the GRB. He noted that the abundance ratio n(D)/n(H) might be an order of magnitude higher than those of other radioactive elements that produce gamma-ray lines and could more easily account for the excess. For example, the deuterium 2.22 MeV line has been clearly detected in solar flares (Chupp et al., 1973). The line can also be produced in the outer envelope of a supernova by spallation reaction (Hoyle and Fowler, 1973). (To date, Compton has not detected the 2.22 MeV D line from supernova remnants.) The integrated contribution to the GRB over cosmological history from this deuterium line is expected to take the form of an edge at the rest energy of 2.2 MeV and a continuous distribution towards redshifted energies as one looks at earlier epochs (Clayton and Silk, 1969). The spectral shape fits the MeV bump well.

11.5.4 GRB from Type Ia SN

The cumulative gamma-ray spectrum of Type Ia supernovae (SN Ia's) during the history of the Universe are

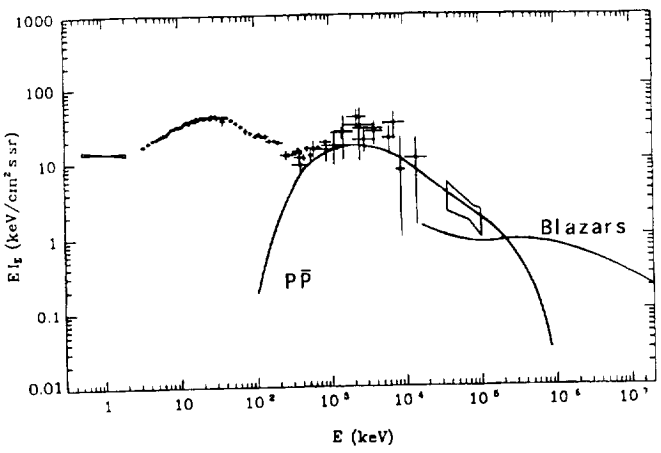


Figure 11.8: Contribution to the extragalactic background by proton-antiproton annihilations in a baryon-symmetric universe (Stecker et al., 1971 and Zdziarski et al., 1993).

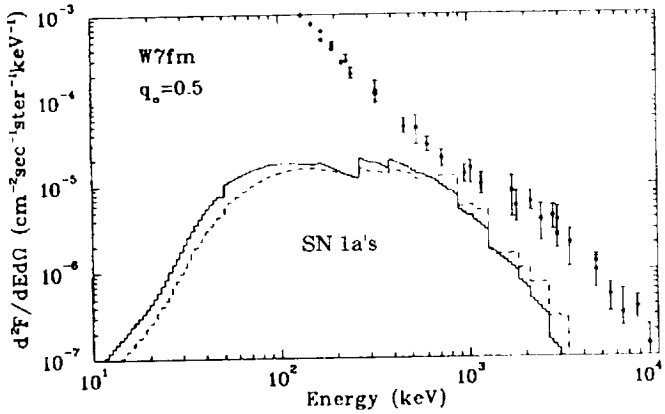


Figure 11.9: The differential flux of cosmic Type Ia supernovae in the Einstein-de Sitter ($q_o = 0.5$) universe. The data points are measurements of the GRB. (The, et al., 1993)

expected to contribute significantly to the GRB. This idea was first advanced by Clayton and Silk (1969), followed by a more detailed discussion by Clayton and Ward (1975) and a recently improved calculation by The et al. (1993). SN Ia's emit strong gamma-ray lines at 847, 1238, 2599 and 3250 keV from the radioactive decay of $56_{Ni} \rightarrow 56_{Co} \rightarrow 56_{Fe}$ produced in the explosion. This has been confirmed by the observations of SN 1987a (e.g., Leising and Share 1990). Integrating over the cosmological history of nucleosynthesis smears the lines and they become edges near the rest energies in the GRB spectrum. In particular, the SN Ia's may contribute a significant fraction of the GRB at $\sim$1 MeV (Figure 11.9). Zdziarski (1996) indeed found that the SN Ia contribution was necessary to explain the discrepancy between the latest COMPTEL measurements of the GRB and theoretical model predictions due to AGN. If high resolution spectral measurements can be made of the GRB between 0.1 and 1.0 MeV, the model of The et al. (1993) may be used to derive the history of SN Ia nucleosynthesis in the Universe. Such high resolution spectral measurements of the GRB have never been made.

11.6 COMPTON RESULTS AND IMPLICATIONS FOR GRB

11.6.1 Compton Observations of AGN

Before the launch of Compton, only four AGN had been detected at gamma-ray energies: 3C 273, Cen A, NGC 4151 and MCG 8-11-11 (see, e.g., Bassani et al., 1985). 3C 273 was the only quasar that has been detected above 10 MeV. It was during one of the early Compton pointings which included 3C 273 in the field-of-view that another quasar, 3C 279, was first detected by the EGRET instrument from 30 MeV to over 5 GeV (Hartman et al., 1992). The photon flux intensity of 3C 279 was very high, comparable to that of the Crab or Geminga above 100 MeV. The differential photon spectrum is well represented by a single power law of photon index $\sim$2 over the entire observed energy range. The lack of detection by COS-B in 1976 implies that 3C 279 is highly variable. Later observations by EGRET confirmed that 3C 279 was initially detected in a flaring state and it is variable on the time scale of days with no significant change in spectral index (Kniffen et al., 1993). Subsequently, many more AGN have been detected by EGRET, COMPTEL and OSSE.

Table 11.2 (from Dermer and Gehrels, 1995) lists the AGN detected by Compton in Phase I (1991 April to 1992 November) and Phase II (1992 November to 1993 September), together with their characteristics (see e.g., Thompson et al., 1995; Fichtel, 1993; Bertsch et al., 1993; Schönfelder 1994; McNaron-Brown et al., 1995; and references therein). These are supplemented by AGN detections by other high energy missions and the ground-based Whipple Observatory. The 20– 165 keV column refers to HEAO-1, Exosat, and Ginga observations, the 50–150 keV column to OSSE and SIGMA/Granat (a French-Russian mission) observations, the 1–30 MeV column to COMPTEL observations, the E > 100 MeV column to EGRET observations, and the E > 0.5 TeV column to Whipple Observatory observations.

All the AGN detected in high-energy gamma rays are core-dominated, flat spectrum radio quasars or radio-loud BL Lac objects (or "blazars"). Blazars generally exhibit strong variability, significant optical polarization

and superluminal flow. The rapid variability indicates that the radiation is likely to be beamed in a relativistic jet, which also produces the radio emission. The BL Lac object Mrk 421 has also been detected in TeV gamma rays by the Whipple Observatory ground-based Cerenkov telescope (Punch et al., 1992). The data are consistent with a single power law of photon index $\sim$2 from the 100 MeV to the TeV energy range, though it must be kept in mind that the Whipple observations were not simultaneous with the EGRET observations (Lin et al., 1992). Note that no Seyfert galaxies have been detected by EGRET in the 100 MeV range.

In the intermediate-energy range (1 to 30 MeV), the COMPTEL instrument team (Bloemen et al., 1995; Blom et al., 1995) has reported evidence of MeV blazers that have not been detected by EGRET. These are compact, flat-spectrum radio-loud AGN, with a spectral break around a few MeV, where the radiated power strongly peaks.

In the low-energy range of 50 keV to 10 MeV, the Oriented Scintillation Spectrometer Experiment (OSSE) instrument has observed 26 AGN and reported 24 detections, with 12 detected at a significance of 4s or higher (Cameron et al., 1993; Johnson et al., 1994; McNaron-Brown et al., 1995). Of the 24 OSSE-detected AGN, 12 are Seyfert 1 galaxies; 5 are Seyfert 2 galaxies (including NGC 4151); 5 are QSOs; and 2 are BL Lac objects. The spectrum of NGC 4151 obtained in July 1991 by OSSE is steeper than most previous observations (Maisak et al., 1993). Such a steep spectrum has been observed once before by the Granat instruments SIGMA and ART-P (Jourdain et al., 1992; Apal'kov et al., 1992). The data can be fitted by a broken power law or a thermal Comptonization spectrum (Sunyaev and Titarchuk, 1980). There is evidence that the spectrum steepens at times of source brightening (Perola et al., 1986; Yaqoob et al., 1989). It is possible that Seyfert galaxies in general have variable spectral states correlated with luminosity (Paciesas et al., 1993). These results together with observations from Ginga (e.g., Williams et al., 1992) and ROSAT (e.g., Turner et al., 1993) at lower energies indicate that AGN have more complex spectral characteristics than can be represented by a simple "canonical" power-law spectrum that resulted from HEAO-1 (Rothschild et al., 1983). For example, Ginga observations have revealed an additional flat component in the $>$ 10 keV portion of Seyfert 1 spectrum (e.g., Pounds et al., 1990; Matsuoka et al., 1990). This has been attributed to reflection or Comptonization of the incident power-law spectrum by an accretion disk.

Johnson et al. (1994) have combined the OSSE observations of Seyferts to produce an average Seyfert spectrum (Figure 11.10), excluding the most intense sources like NGC 4151 so as not to bias the result. The average Seyfert spectrum clearly exhibits a break around 50 keV. Below 50 keV, the HEAO-1 "canonical" power law with photon index 1.7 characterizes the spectrum well. Above 50 keV, the spectrum is much softer. The spectrum can be described by a simple exponential with e-folding energy of 46$\pm$5 keV, but thermal Comptonization models provide the best parameterization. These are plotted in Figure 11.10 for comparison. This new result helps to solve the potential excessive Seyfert emission problem mentioned in Section 11.3, and brings the composite AGN spectrum more in line with the observed GRB spectrum.

A picture is emerging that the gamma-ray emitting AGN can be divided into two classes (Dermer and Gehrels, 1995) with different emission mechanisms: the comparatively weak Seyferts and radio galaxies, in which we observe quasi-isotropic gamma rays with inferred isotropic luminosities in the range $\sim 10^{41}$ - 10^{44} $erg\ s^{-1}$; and the radio-loud blazars, in which we observe beamed emission from collimated relativistic jets, and which are highly variable and have luminosities in the range $\sim 10^{45}$ - 10^{49} $erg\ s^{-1}$.

11.6.2 Compton Observations of GRB

Careful background analyses are beginning to yield preliminary results from COMPTEL and EGRET on the GRB spectrum. Both of these instruments have been designed to maximize their use for obtaining the extragalactic diffuse background. The preliminary COMPTEL results (Kappadath et al., 1995) indicate that the GRB flux in the 2-10 MeV range is 5 to 10 times lower than those measured by Trombka et al. (1977) and balloon experiments. The 10-30 MeV flux is consistent with other measurements. The existence of the MeV bump is called into question because of better background reduction technique and the identification of additional contributions by long-lived radioactive nuclei to the internal background. The MeV bump, if present, is restricted for COMPTEL to energies of less than $\sim$2 MeV.

The preliminary EGRET result for the GRB above 100 MeV is consistent with the previous measurement by SAS-2. Much work remains to be done to make sure that there is no contamination from the galactic diffuse emission. Both the COMPTEL and EGRET results have been included in the updated GRB plot in Figure 11.3.

11.6.3 Implications for GRB

The Compton observations firmly identify AGN as gamma-ray emitters that can contribute to the GRB. In the low-energy regime, several groups have proposed schemes to explain the GRB spectrum based on the "reprocessed" AGN model (Zdziarski et al., 1993; Madau et al., 1993; Rogers and Field, 1991; Terasawa, 1991; Mereghetti, 1990). The intrinsic AGN spectrum can be

Table 11.2: Active Galactic Nuclei Detected at $> 20keV$ (Dermer and Gehrels 1995)

Source	Type	z	20–165 *keV*	50–150 *keV*	1–30 *MeV*	>100 *MeV*	>0.5 *TeV*
NGC 4151	Sy 1.5–2	0.0033	Y: 3.4–10	Y: 15.–31	N	N: <0.05	N
NGC 7582	Sy 2	0.0049		Y: 2.6–5.1	N		
NGC 6814	Sy 1	0.0053	Y: 2.6	Y: 3.2	N		
MCG–6–30–15	Sy 1	0.0078		Y: 4.0–5.0	N	N: <0.1	
MCG–5–23–16	Sy 2	0.0082	Y: 3.3	Y: ~4.0			
NGC 4388	Sy 2	0.0087		Y: 6.4			
NGC 3783	Sy 1	0.0092	Y: 0.8	Y(2σ):~4	N	N: <0.12	
IC 4329A	Sy 1	0.0157		Y: 4.2–5.9	N	N	
NGC 5548	Sy 1	0.017	Y: 0.6	Y(3σ):~4	N	N	
MCG+8–11–11	Sy 1	0.0205	Y: 8.0	Y: 2.2–4.0	N	N: <0.11	
Mrk 279	Sy 1	0.031	Y: 1.0	Y: 2.6		N: <0.07	
Mrk 509	Sy 1	0.0355	Y: 1.1	Y(3σ):~4	N	N: <0.08	
ESO 141–G55	Sy 1	0.037	Y: 0.6	Y(3σ):~3	N	N: <0.15	
Cen A	RG, Sub,Sy 2	0.0008	Y: 28.	Y: 25.–63.	Y	N: <0.15	Y(?)
NGC 1275	RG, Sub,Sy 2	0.0183	Y: 3.3	N: <2.4		N: <0.09	N
3C 120	Sy 1,RG,SL	0.0334	Y: 1.8	Y(3σ):~3	N	N: <0.09	
3C 111	Sy 1,RG,SL?	0.049		Y: 2.1-3.1	N	N: <0.09	N
3C 390.3	Sy 1,RG,SL?	0.0569	Y: 3.0	Y: <1.5-4.1		N: <0.07	
Mrk421(1101+384)	FS,BL	0.031	Y	N: <0.9		Y:0.14	Y: 1.5
PKS 0521–365	FS,HP	0.055				($4\ \sigma$): ~0.2	
PKS 2005–489	FS,BL,OV	0.071				Y(4σ): ~0.18	
1219+285	FS,BL,HP,OV	0.102				Y: 0.17	
PKS 2155–304	BL	0.117	Y(?)	Y: <2.0–4.9		N	
3C273(1226+023)	FS,SL,HP,OV	0.158	Y	Y: 3.6–19.7	Y	Y:<0.13–0.21	N
0829+046	FS,BL,OV	0.18				Y(4σ): ~0.14	
PKS 1510–089	FS,HP,OV?	0.361				Y: 0.18–0.23	
0954+658	FS,BL,HP	0.368				Y: 0.21	
4C21.35(1222+216)	FS	0.435				Y: 0.17	
3C279(1253–055)	FS,SL,HP,OV	0.538		Y: <2.4–6.2	Y	Y: 0.9–2.7	N
1E 1227+0224	QSO(near 3C273)	0.57		Y(SIGMA)			
4C29.45(1156+295)	FS,SL,HP,OV	0.729				Y: 0.63	
PKS 0454–463	FS	0.858				Y: <0.16–0.29	
3C454.3(2251+158)	FS,SL,HP,OV	0.859		Y: <2.2–4.5	Y(?)	Y: 0.8–1.35	N
PKS 0537–441	FS,BL,HP	0.894				Y: 0.17–0.32	
PKS 0420–014	FS,HP,OV	0.92				Y: <0.14–0.45	
PKS 0235+164	FS,BL,SL?,HP	0.94				Y: <0.3–0.82	
PKS 1933–400	FS	0.966				Y(4σ): ~0.88	
PKS 0208–512	FS,HP	1.003				Y: 0.4–1.1	
PKS 0454-234	FS,HP	1.009				Y(4σ): ~0.14	
CTA 102 (2230+114)	FS,SL?,HP	1.037		Y: <1.6–4.0	Y(?)	Y: 0.25–0.46	
PKS 1229–021	FS	1.045				Y(4σ): ~0.12	
PKS 1741–038	FS,HP	1.054				Y: 0.2–0.34	
PKS 2356+196	FS	1.066				Y(4σ): ~0.29	

Table 11.2: Continued

Source	Type	z	20–165 *keV*	50–150 *keV*	1–30 *MeV*	>100 *MeV*	>0.5 *TeV*
PKS 0506–612	FS	1.093			Y(?)	Y(4σ): $\sim$0.06	
PKS 0446+112	FS	1.207				Y: <0.16–1.04	
PKS 1313–333	FS	1.21				Y: 0.1–1.3	
4C28.07 (0234+285)	FS,SL,HP	1.213				Y: 0.16	
4C10.45 (1606+106)	FS	1.23				Y: 0.27–0.53	
4C51.37 (1739+522)	FS	1.38				Y: 0.36	
1611+343	FS,OV	1.40				Y: 0.33	
0804+499	FS,HP	1.43				Y: 0.29	
PKS 2052–474	FS	1.489				Y: 0.26	
PKS 1406–076	FS	1.494				Y: <0.10–0.41	
4C38.41(1633+382)	FS,SL,OV	1.81				Y: 1.0	N
0827+243	FS,OV	2.046				Y: 0.21	
PKS 0528+134	FS,OV	2.06		Y: 2.0–4.5	Y	Y: <0.5–1.6	N
4C71.0 (0836+710)	FS,SL,HP	2.17				Y: 0.14–0.34	N
PKS 0458-020	FS,OV	2.286				Y(4σ): $\sim$0.28	
4C15.05(0202+149)	FS,HP	?		N: <8.7		Y: 0.26	
0716+714	FS,BL,SL?,HP	?		N: <1.4		Y: 0.2–0.5	N
4C15.54(1604+159)	BL,HP	?				Y(4σ): $\sim$0.17	
PKS 1622–253	FS	?				Y: 0.19–0.47	
2022-077	FS	?				Y: <0.15–0.63	
PKS 2209+236	FS	?				Y(4σ): $\sim$0.15	
H1517+656	BL	?		Y: 4.9			

Type: Sy = Seyfert, RG = radio galaxy, Sub =subluminal, SL = superluminal, FS = flat spectrum radio galaxy, BL = BL Lac, HP = high optical polarization, OV = optically violent variable (OVV).

Note 1. - With some exceptions, the columns refer to detections by instruments as follows: 20-165 *keV*: HEAO-1, Exosat, and Ginga; 50-150 *keV*: OSSE and SIGMA; 1-30 MeV: COMPTEL; >100 *MeV*: EGRET; >0.5 *TeV*: Whipple.

Note 2. - A "Y" indicates a detection of significance greater than 5.0σ for the >100 *MeV* column and greater than 3.5σ for the other columns. The significance level for marginal detections is given in parentheses, where, for example, *3sigma* means between $2.5\sigma \leq$ significance $< 3.5\sigma$.

Note 3. - Numbers in detection columns are fluxes where available. Units and references are: 20–165 *keV*: 10^3 ph $cm^{-2}s^{-1}$(Rothschild et al., 1983, from luminosities in their Table 11.2; NGC 4151 from Baity et al., 1984; MCG+8-11-11 from Ubertini et al., 1984; Cen A from Baity et al., 1981); 50-150 *keV*: 10^4 ph $cm^{-2}s^{-1}$(Johnson et al., 1994; 0.1 *MeV* times numbers in their Table 11.1); >100 MeV: 10^6 ph $cm^{-2}s^{-1}$(Fichtel et al., 1994; Lin et al., 1993; von Montigny et al., 1994): > 0.5 *TeV*: 10^{11} ph $cm^{-2}s^{-1}$(Punch et al., 1992).

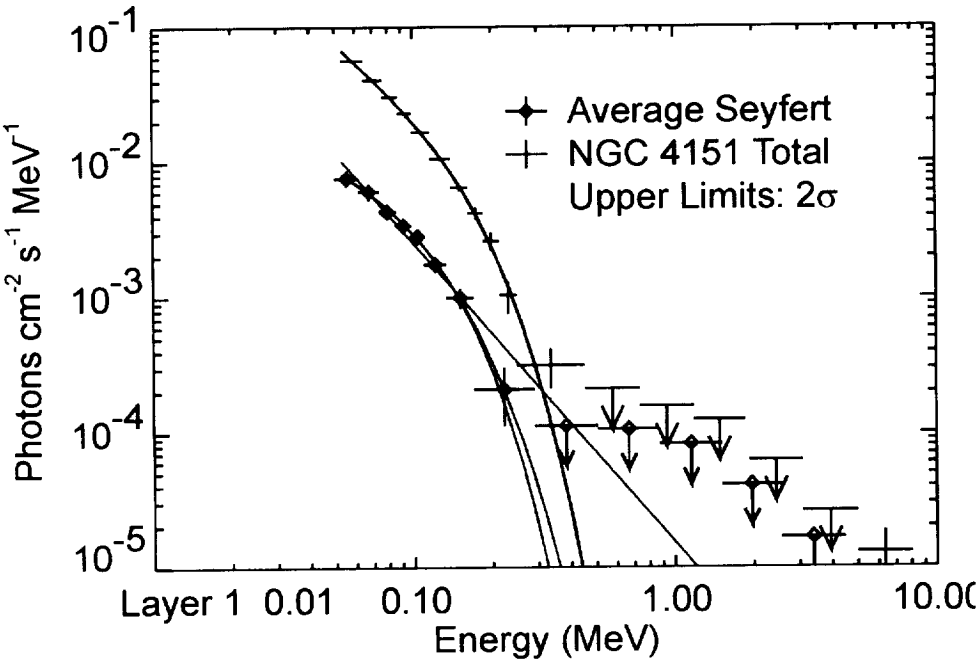


Figure 11.10: Average OSSE Seyfert spectrum. Best fit thermal Comptonization, power law and exponential are shown. The OSSE NGC 4151 spectrum from June 1991 is shown for comparison with thermal Comptonization fit (Johnson et al., 1994).

generated either through thermal Comptonization of low-energy photons by hot plasma or by non- thermal pair cascade process in a compact plasma (Svensson, 1987). Both mechanisms produce a spectral break at ~80 keV. Zdziarski et al. (1993), using a Seyfert 1 thermal Comptonization model and the X-ray luminosity function derived by Boyle et al. (1993), integrated the contributions over redshift and obtained a good fit to the GRB spectrum from 2 to 100 keV, including the 30 keV bump (Figure 11.11). The dominant contribution comes from the AGN at the highest redshift. However, the model did not seem satisfactory at the time because it could not explain the MeV bump as the Seyfert spectrum cuts off above 80 keV. Now the MeV bump problem seems to have disappeared, the AGN models (e.g., Zdziarski, 1996) can very well explain a good part of GRB spectrum, though Zdziarski (1996) still found that AGN contributions alone cannot account for all emission in the few MeV range. He obtained a good fit with the preliminary COMPTEL measurements by adding the SN Ia contribution (The et al., 1993). If there is indeed a class of MeV Blazers as indicated by the COMPTEL observations, then they could also contribute significantly towards the MeV range GRB emission.

In the high-energy regime, Padovani et al. (1993), Dermer and Schlickeiser (1992), and Stecker et al. (1993) estimated the contribution of blazars to the GRB. Since the gamma-ray luminosity function of blazars has not yet been determined, they used the radio luminosity evolutionary function by Dunlop and Peacock (1990) instead. This is based on a significant (possibly linear) correlation between the 100 MeV gamma-ray luminosity and the 5-GHz radio luminosity seen in the gamma-ray blazars (Figure 11.12). The computed background falls an order of magnitude below that observed by SAS-2 in the 100 MeV range (Stecker et al., 1993) and so fails to be a major contributor to the GRB in this energy range. Moreover, the average blazar spectrum is a power law of photon index ~2, which is much harder than the observed GRB (Figure 11.6). However, if the hard spectrum continues unattenuated to higher energies, blazars could be important contributors to the GRB in the multi-GeV energy range. Recently, Stecker and Salamon (1996) have recomputed the contribution of blazars to the GRB based on the second EGRET catalog (Thompson et al., 1995), taking account of their high variability and the apparent hardening of their spectra when they are in a flare state. They found that the entire extragalactic GRB can be explained by the contributions from unresolved blazars.

It may well be that the GRB is the sum of a number of different components with different origins. Each component's contribution varies as a function of energy. We need to keep in mind that when we study the gamma-ray background, we are looking at an energy span of over six decades. A particular physical mechanism probably dominates in a specific energy range, for example:

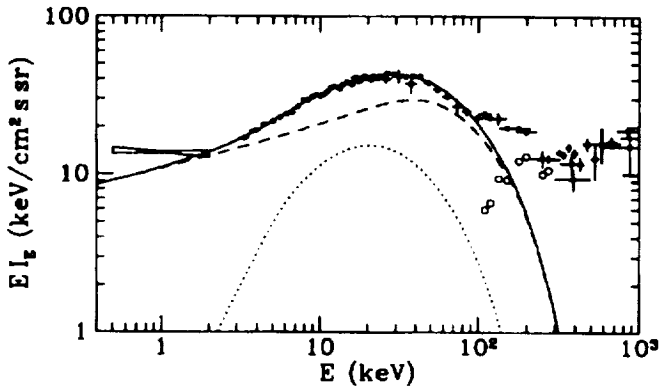


Figure 11.11: Contribution to the GRB by thermal AGN. The solid and dot-dashed curves are Comptonization models without and with absorption. The dashed curve gives the thermal component, and the dotted curve the reflection component. (Zdziarski et al., 1993)

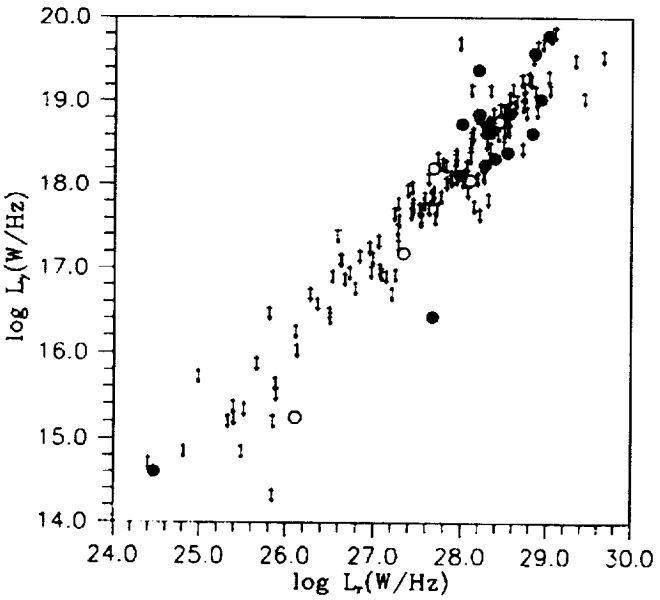


Figure 11.12: Gamma-ray luminosity vs. radio luminosity plot for blazars in the EGRET survey. The filled circles represent solid detections, the hollow circles represent marginal detections, and the crossed arrows represent the upper limits for those radio sources which EGRET did not detect. (Stecker et al., 1993)

the radio-quiet AGN at < 1 MeV; the diffuse matter-antimatter annihilations at $\sim$1 to 100's MeV; the MeV blazars and/or the SN Ia's at $\sim$1 MeV; the primordial black holes at 100's MeV; and blazars at 100's GeV.

11.7 FUTURE OBSERVATIONS

In some energy bands there will be significant new data on the diffuse gamma-ray background from current or planned missions, while in other bands there are no missions. Prospects are particularly bright in the 30 MeV - 30 GeV band where EGRET will provide a definitive observation. The task is not easy though, requiring full understanding of the dominant galactic diffuse emission. OSSE and COMPTEL also have diffuse background objectives, but both have high internal backgrounds that must be modeled and subtracted. This is a generic problem in the 100 keV - 30 MeV range.

There are several possible future approaches to measuring the diffuse background in the MeV range. One idea is to make a differential measurement in which identical and exchangeable detectors with different fields-of-view observe the sky. The internal background level should be the same in both detectors, so that subtracting their rates should give the diffuse flux in the differential field-of-view. Another idea is to have a detector with very little surrounding passive material located on a long extendible boom from a light spacecraft flown outside the Earth's magnetosphere. The small background contribution from the spacecraft could be determined by moving the detector in and out on the boom (like with the Apollo measurements (Trombka et al., 1977). The diffuse signal would be maximized by having an almost unobstructed view of the entire sky. The internal background in the detector itself from beta decays could be minimized by using background-reduction techniques such as pulse-shape discrimination (Roth et al., 1984), detector segmentation (Gehrels, 1985) or detector isotope enrichment (Gehrels, 1990). A third idea is to have an aperture with a changeable field-of-view. If this is done carefully with active anticoincidence shield and aperture elements, it should be possible to have the internal background be almost constant as the aperture changes, so that the only the diffuse flux is modulated.

The history of SN Ia nucleosynthesis in the Universe can be derived if high resolution spectral measurements can be made of the GRB between 0.1 and 1.0 MeV (The et al., 1993). The current generation of germanium gamma-ray spectrometers has sufficient spectral resolution to measure such edges in the GRB spectrum. For example the GRIS balloon instrument has successfully resolved the 56Co lines of SN 1987a (Tueller et al., 1989). Observations of the GRB by GRIS have been carried out in 1995 and analysis is in progress. Better spatial resolution in future instruments would allow determination of the GRB fine-scale anisotropy. This is essential to distinguish between the baryon-symmetric cosmological origin and the AGN origin of the GRB (Gao et al., 1990a,b; Cline and Gao, 1990), as these models predict different intensity fluctuation patterns in the GRB.

The great hope with these future observations will be to accurately determine the spectrum of the diffuse background and thereby learn about its origin. Questions we hope to answer include: Is the MeV bump real? What is the contribution of Seyfert AGN to the low-energy gamma-ray background? What is the contribution of blazars AGN to the high-energy gamma-ray background? Is there a residual true cosmological background?

11.8 REFERENCES

Apal'kov, Y., et al., 1992, in "Frontiers of X-Ray Astronomy" eds. Y. Tanaka and K. Koyama (Tokyo: Universal Academy Press), 251.

Arnold, J.R., Mertzger, A.E., Anderson, E.C., and Van Dilla, M.A., 1962, J. Geophys. Res., 67, No. 12, 4878.

Avni, Y., 1978, Astr. Ap., 63, L13.

Baity, W.A., et al., 1981, ApJ, 244, 429.

Baity, W.A., Mushotzky, R.F., Worrall, D.M., Rothschild, R.E., Tennant, A.F., and Primini, F.A., 1984, ApJ, 279, 555.

Bassani, L., Dean, A.J., Di Cocco, G., and Perotti, F., 1985, in "Active Galactic Nuclei, ed. J.E. Dyson" (Manchester: Manchester University Press), 252.

Bertsch, D.L., et al., 1993, ApJ, 405, L21.

Bignami, G.F., Fichtel, C.E., Hartman, R.C., and Thompson, D.J., 1979, ApJ, 232, 649.

Bloemen, H., et al., 1995, Astr. Ap., 293, L1.

Blom, J.J., et al., 1995, Astr. Ap., 298, L33.

Boldt, E., 1987, Phys. Reports, 146, No. 4, 215.

Boyle, B.J., Griffiths, R.E., Shanks, T., Stewart, G.C., and Georganopoulos, I., 1993, MNRAS, 260, 49.

Brown, R.W. and Stecker, F.W., 1979, Phys. Rev. Letters, 43, 315.

Calzetti, D., Livio, M., and Madau, P., 1995, in "Extragalactic Background Radiation" STScI Conf. Proc. (Cambridge University Press).

Cameron, R.A., et al., 1993, in "Proc. Compton Gamma Ray Observatory Symposium" eds. M. Friedlander, N. Gehrels, and D. Macomb (AIP: New York), 478.

Chen, L.W., Fabian, A.C., and Gendreau, K.C., 1996, MNRAS, in press.

Chupp, E.L., et al., 1973, Nature, 241, 333.

Clayton, D.D. and Silk, J., 1969, ApJ, 158, L43.

Clayton, D.D. and Ward, R.A., 1975, ApJ, 198, 241.

Cline, D.B. and Gao, Y.-T., 1990, ApJ, 348, 33.

Comastri, et al., 1995, Astr. Ap., 296, 1.

Daly, R.A., 1988, MNRAS, 232, 853.

Daniel, R.R., Joseph, G., and Lavakare, O., 1972, Astrophys. Space Sci., 18, 462.

Dermer, C.D. and Gehrels, N., 1995, ApJ, 447, 103.

Dermer, C.D. and Schlickeiser, R., 1992, Science, 257, 1642.

Dunlop, J.S. and Peacock, J.A., 1990, MNRAS, 247, 19.

Fichtel, C.E., 1993, Ann. NY Acad. Sci. (Texas/PASCOS, ed. C.W. Akerlof and M.A. Srednicki), 688, 136.

Fichtel, C.E., et al., 1994, Ap.J. Suppl., 94, 551.

Fichtel, C.E., Hartman, R.C., Kniffen, D.A., Thompson, D.J., Bignami, G.F., Ögelman, H., Özel, M.F., and Tümer, T., 1975, ApJ, 198, 163.

Fichtel, C.E., Simpson, G.A., and Thompson, D.J., 1978, ApJ, 222, 833.

Fukuda, Y., Hayakawa, S., Kasahara, I., Makino, F., and Tanaka, Y., 1975, Nature, 254, 398.

Gao, Y.T., Cline, D.B., and Stecker, F.W., 1990a, ApJ, 357, L1.

Gao, Y.T., Stecker, F.W., Gleiser, M., and Cline, D.B., 1990b, ApJ, 361, L37.

Gehrels, N., 1985, Nucl. Inst. Meth., A239, 324.

Gehrels, N., 1990, Nucl. Inst. Meth., A292, 505.

Gehrels, N. and Tueller, J., 1993, ApJ, 407, 597.

Gendreau, K.C., et al., 1995, PASJ, 47, L5.

Georgantopoulos, I., et al., 1993, MNRAS, 262, 619.

Gruber, D.E., 1992, in "X-Ray Background" eds. X. Barcons and A.C. Fabian (Cambridge: Cambridge Univ. Press), 44 (G92).

Gruber, D.E., Matteson, J.L., and Jung, G.V., 1985, Proc., 19th Intl. Cosmic Ray Conf., ed. F.C. Jones, OG-1, 349.

Hartman, R.C., et al., 1992, ApJ, 385, L1.

Hasinger, G., Burg, R., Giacconi, R., Hartner, G., Schmidt, M., Trümpler, J., and Zamorani, G., 1993 A and A, 275, 1.

Hoyle, F. and Fowler, A., 1973, Nature, 241, 384.

Hulsizer, R. and Rossi, B., 1949, Phys. Rev., 76, 164.

Johnson,W.N., et al., 1994, in "The Second Compton Symposium" eds. C.E. Ficthel, N. Gehrels, and J.P. Norris (New York: AIP), 515.

Jourdain, E., et al., 1992, Astr. Ap., 256, L38.

Kappadath, S.C., et al., 1995, in "Proc. 24th Int. Cosmic Ray Conf." Rome, 2, 230.

Kinzer, R.L., Jung, C.V., Gruber, D.E., Matteson, J.L., and Perterson, L.E., 1996, preprint.

Kogut, A., et al., 1993, ApJ, 419, 1.

Kniffen, D.A., et al., 1993, ApJ, 411, 133.

Kniffen, D.A., et al., 1996, Astr. Ap., in press.

Kraushaar, W.L., Clark, G.W., Garmire, G.P., Borker, R., Higbie, P., Leong, V., and Thorsos, T., 1972, ApJ, 177, 341.

Lavigne, J.M., et al., 1982, ApJ, 261, 720.

Leising, M.D. and Share, G.H., 1990, ApJ, 357, 638.

Leiter, D. and Boldt, E., 1992, in "Proc. Compton Observatory Science Workshop" (NASA Conf. Pub. 3137), 359.

Leventhal, M., 1973, Nature, 246, 136.

Levine, A.M., et al., 1984, Ap.J. Suppl., 54, 581.

Lin, Y.C., et al., 1992, ApJ, 401, L61.

Lin, Y.C., et al., 1993, ApJ, 416, L53.
Madau, P., Ghisellini, G., and Fabian, A.C., 1993, ApJ, 410, L7.
Mahoney, W.A., Ling, J.C., and Jacobson, A.S., 1981, JGR, 86, 11098.
Maisack, M., et al., 1993, ApJ, 407, L61.
Mandrou, P., Vedrenne, G., and Niel, M., 1979, ApJ, 230, 97.
Marshall, E.F., et al., 1980, ApJ, 235, 4.
Mather, J.C., et al., 1990, ApJ, 354, L37.
Mather, J.C., et al., 1994, ApJ, 420, 439.
Matsuoka, M., et al., 1990, ApJ, 361, 440.
Mazets, E.P., Golenetskii, S.V., Il'inskii, V.N., Gur'yan, Y.A., and Kharitonova, T.V., 1975, Astrophys. Space Sci., 33, 347.
McNaron-Brown, K., et al., 1995, ApJ, 451, 575.
Mereghetti, S., 1990, ApJ, 354, 58.
Mushotzky, R.F., 1982, ApJ, 256, 92.
Mushotzky, R.F., 1984, Astrophys. Space. Sci., 3, Nos. 10-12, 157.
Olive, K.A. and Silk, J., 1985, Phys. Rev. Letters, 55, 2362.
Paciesas, W.S., et al., 1993, in "Proc. Compton Gamma Ray Observatory Symposium" ed. M. Friedlander, N. Gehrels, and D. Macomb (AIP: New York), 473.
Padovani, P., Ghisellini, G., Fabian, A.C., and Celotti, A., 1993, MNRAS, 260, L21.
Page, D.N. and Hawking, S.W., 1976, ApJ, 206, 1.
Perola, G.C., et al., 1986, ApJ, 306, 508.
Piccinotti, G., Mushotzky, R.F., Boldt, E.A., Holt, S.S., Marshall, F.E., Serlemitsos, P.J., and Shafer, R.A., 1982, ApJ, 253, 485.
Pinkau, K., 1973, in "Gamma Ray Astrophysics" eds. F.W. Stecker and J.I. Trombka, NASA SP-339 (US Govt. Printing Office, Washington, DC), 133.
Pounds, K.A., et al., 1990, Nature, 344, 132.
Puget, J.L., 1973, in "Gamma Ray Astrophysics" eds. F.W. Stecker and J.I. Trombka, NASA SP-339 (US Govt. Printing Office, Washington, DC), 367.
Punch, M., et al., 1992, Nature, 358, 477.
Rogers, R.D. and Field, G.B., 1991, ApJ, 366, 22.
Rogers, R.D. and Field, G.B., 1991, ApJ, 370, L57.
Roth, J., Primbsch, J. H., and Lin, R.P., 1984, IEEE Trans. Nucl. Sci., NS-31, 367.
Rothschild, R.E., Mushotzky, R.F., Baity, W.A., Gruber, D.E., Matteson, J.L., and Peterson, L.E., 1983, ApJ, 269, 423.
Schönfelder, V., Graml, F., and Penningsfeld, F.-P., 1980, ApJ, 240, 350.
Schönfelder, V., 1994, Ap.J. Suppl., 92, 593.
Shafer 1983, Ph.D. Dissertation, University of Maryland, NASA TM85029.
Shanks, T., et al., 1991, Nature, 353,315.
Smoot, G.F., et al., 1992, ApJ, 396, L1.
Sreekumar, P. and Kniffen, D.A., 1996, in "U Symp. 168, Examining the Big Bang and Diffuse Background Radiation" eds. M. Kafatos and Y. Kondo (The Hague: Kluwer), 279.
Stecker, F.W., 1971, in "Cosmic Gamma Rays" NASA SP-249 (US Govt. Printing Office, Washington, DC).
Stecker, F.W., 1985, in "Nuclear Physics" B252, 25.
Stecker, F.W., 1989, in "Proc. of Gamma Ray Observatory Science Workshop" ed. W.N. Johnson, 4–73.
Stecker, F.W., Morgan, D.L., and Bredekamp, J., 1971, Phys. Rev. Letters, 27, 1467.
Stecker, F.W., Salamon, M.H., and Malkan, M.A., 1993, ApJ, 410, L71.
Stecker, F.W. and Salamon, M.H., 1996, ApJ, 464, 600.
Svensson, R., 1987, MNRAS, 227, 403.
Sunyaev, R.A. and Titarchuk, L.G., 1980, Astr. Ap., 86, 121.
Sunyaev, R.A. and Zel'dovich, Ya.B., 1980, Ann. Rev. Astr. Astrophys., 18, 537.
Terasawa, N., 1991, ApJ, 378, L11.
The, L.-S., Leising, M., and Clayton, D.D., 1993, ApJ, 403, 32.
Thompson, D.J., et al., 1995, Ap.J. Suppl., 101, 259.
Trombka, J.I., Dyer, C.S., Evans, L.G., Bielefeld, M.J., Seltzer, S.M., and Metzger, A.M., 1977, ApJ, 212, 925.
Tueller, J., Barthelmy, S.D., Gehrels, N., Teegarden, B.J., Leventhal, M., and MacCallum, C.J., 1989, ApJ, 351, L41.
Turner, T.J., George, I.M., and Mushotzky, R.F., 1993, ApJ, 412, 72.
Ubertini, P., Bazzano, A., LaPadula, C., Polcaro, V.F., and Manchanda, R.K., 1984, ApJ, 284, 54.
Vedrenne, G., Albernhe, F., Martin, I., and Talon, R., 1971, Astr. Ap., 15, 50.
Vette, J.I., Gruber, D.E., Matteson, J.L., and Peterson, L.E., 1969, in "IAU Symp. No. 37" ed. L. Gratton (Dordrecht: Reidel), 335.
von Montigny, C., et al., 1994, ApJ, 440, 525.
Weinberg, S., 1972, in "Gravitation and Cosmology" (New York: Wiley).
Wheaton, W., et al., 1995, ApJ, 438, 322.
Wheaton, W., et al., 1996, preprint.
White, R.S., Dayton, B., Moon, S.H., Ryan, J.M., Wilson, R.B., and Zych, A.D., 1977, ApJ, 218, 920.
Williams, O.R., et al., 1992, ApJ, 389, 157.
Yaqoob, T., et al., 1989, MNRAS, 236, 153.
Zdziarski, A.A., 1996, MNRAS, 281, L9.
Zdziarski, A.A., Zycki, P.T., and Krolik, J.H., 1993, ApJ, 414, L81.
Zdziarski, A.A., Johnson, N.W., Done, C., Smith, D., and McNaron-Brown, K., 1995, ApJ, 438, L63.

Chapter 12

THE X-RAY SKY AS A MEASURE OF HISTORY

ELIHU BOLDT
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

The high latitude X-ray sky is dominated by an isotropic cosmic background that provides a powerful demonstration of the cosmological principle – that we are not in a special place. The observed flux represents the emission of all the X-ray sources in previous epochs of the universe. Analysis of this background within the context of accretion driven active galactic nuclei indicates that most black hole galactic nuclei must be dormant at present but suggests that they were X-ray bright during a brief epoch signaling an early phase of galaxy formation.

12.1 INTRODUCTION

There are only two bands of the electromagnetic spectrum where the sky is dominated by a clearly isotropic extragalactic background. The most celebrated is the microwave band (Mather et al., 1990) in which resides the 2.7° K black body relic radiation cooled from an early stage of the expanding universe when its energy equivalent mass density exceeded that of ordinary matter. The other is the X-ray band (Giacconi et al., 1962) where we find a well-defined component associated with the subsequent epochs in which subrelativistic matter dominates. This all-sky Cosmic X-ray Background (CXB) has a characteristic photon energy (E) in the sense that the spectrum of energy flux $[I(erg\ s^{-1}cm^{-2}\text{sterad}^{-1})]$, when expressed as $dI/d(logE)$, peaks at E = 29 keV, corresponding to an equivalent Boltzmann temperature of about 3×10^8 K (Boldt, 1987). Since the Planck spectrum of the 2.7° K background is undistorted the possible Compton scattering with intervening ionized matter must be minimal. Hence, there is not an extensive hot intergalactic plasma contributing to the CXB; the X-rays must then arise mainly from discrete sources

The volume X-ray emissivity [i.e., the power $(erg\ s^{-1})$ radiated in X-rays per unit volume] in the present epoch universe is mainly due to galaxies with clearly active nuclei (Miyaji et al., 1994) even though such objects represent less than 1% of all galaxies. The concept of what we mean by the present epoch is challenged when we consider that the typical distance between such neighboring "local" active galaxies corresponds to a light travel time of several million years, not short compared to the time scales that could be associated with the growth of a central black hole in the nucleus, due to the matter accretion process thought to power such sources (e.g., see Rees, 1984). Clearly, space and time are inexorably intertwined for the CXB. As we shall exhibit, its flux provides us with a revealing measure of *all* the accreted matter responsible for the growth of supermassive black hole galactic nuclei, integrated over the entire visible history of the baryonic universe.

Recent broadband X-ray observations of active galaxies (Johnson et al., 1994) with the Compton Gamma-Ray Observatory indicate that such sources also have a characteristic photon energy, but its value is about three times that characteristic of the CXB. This shows that the major contribution of such sources to the observed CXB must be highly redshifted from substantially earlier epochs. Observations of soft X-rays with the ROSAT focusing telescope have in fact directly traced X-ray bright active galaxies back to the earliest epochs already revealed by quasars detected optically at the very highest redshifts, corresponding to a look-back time that is about 90% the age of the universe (Warren, Hewitt, and Osmer, 1994).

12.2 EXTRAGALACTIC SCALES

Distances in extragalactic astronomy are commonly expressed in megaparsecs (1 Mpc= $3.09\times10^{24}cm$, the distance traveled by light in 3.26 million years). Although the core diameter of a rich cluster of galaxies is about 1 Mpc, the chance is only one in a million that a cluster centroid is actually to be found within a randomly located sphere of this diameter (e.g., see Bahcall, 1988). The number density of active galaxies is greater. In particular, data from the HEAO-2 Einstein Observatory and ROSAT have been used by Boyle et al. (1993) to determine that, for X-ray luminosities $\geq 10^{42}h^{-2}erg\ s^{-1}$ (0.3 - 3.5 keV), the number density of AGN (active galactic nuclei) is $n = 1.9\times10^{-5}h^3\ \ Mpc^{-3}$, where $h \equiv H_o/(50\ km\ s^{-1}/\ Mpc)$ exhibits the dependence on the Hubble constant (H_o), discussed later. This number density is comparable to that of Seyfert 1 galaxies at $M_B < -17.5$ (Huchra and Burg, 1992). According to the unified model for Seyfert galaxies (Antonucci and Miller, 1985; Awaki, 1991) Seyfert 1 AGN are the same as Seyfert 2 but viewed from a direction that avoids the obscuring matter intrinsic to such objects. Although Seyfert 2 galaxies are likely to dominate the AGN population their X-ray spectra are highly absorbed, making these sources relatively inaccessible within the soft X-ray spectral band considered by Boyle et al. (1993). From a representative vantage point in the present epoch, a survey depth of $\sim 20h^{-1}Mpc$ would be needed to encounter one Seyfert 1 galaxy, on the average, within the corresponding horizon in look-back time of $\sim 60h^{-1}$ million years.

In general, the radiation reaching us from sources having a velocity of recession due to the expansion of the universe, is Doppler-shifted in wavelength (λ), manifested as a redshift. This redshifted wavelength depends on source recession velocity (v) as follows

$$\frac{\lambda}{\lambda_s} = \left[\frac{(c+v)}{(c-v)}\right]^{1/2} \approx 1+(v/c) \quad for\ v \ll c \qquad (12.1)$$

where c is the velocity of light and λ_s is the wavelength emitted at the source. By defining the redshift as

$$z \equiv \frac{(\lambda - \lambda_s)}{\lambda_s}$$

it follows that

$$\frac{\lambda}{\lambda_s} = \frac{E_s}{E} = (1+z) \qquad (12.2)$$

We note further that $z \approx (v/c)$ for $v \ll c$. Hubble's law states that, for $v \ll c$, the distance (D) to an object receding with the expansion velocity (v) of the universe is $D = v/H_o$. Assuming that the observed redshift arises from the Doppler effect, we note that this

implies $D = (c/H_o)z$, for $z \ll 1$. [For $z \geq 1$ the distance used for determining luminosity is a more complicated function of z that depends on Ω, the ratio of overall gravitational mass density to the critical closure value of an Einstein-deSitter model (e.g., see Weinberg, 1972)]. The Hubble distance that characterizes the cosmological scale is $(c/H_o) = 5996h^{-1}Mpc$. The Hubble time $H_o^{-1} = 19.6 \times 10^9 h^{-1}$ years is a measure of the age since the "big bang" marking the origin of the expansion; for $\Omega = 1$ this age is $t_H = (2/3)H_o^{-1} = 13 \times 10^9 h^{-1}$ years.

12.3 BLACK HOLE MASS GROWTH

An active galactic nucleus is generally regarded to be powered by the accretion of matter induced by the gravity field of a massive black hole at the core (Rees, 1984). The resulting bolometric luminosity (L_{bol}) is related to the rate of mass infall as

$$L_{bol} = \epsilon\, c^2 \left(\frac{dM}{dt}\right) \qquad (12.3)$$

where ϵ is the efficiency for the conversion of accreted mass energy into radiation. In typical cases this luminosity is limited by the ambient mass fuel supply available for accretion. There is, however, a limiting luminosity determined by the mass of the black hole itself. In particular, the total luminosity can not exceed the Eddington limit where outward radiation pressure would cancel out the black hole's gravitational pull on infalling matter. For accreting matter where there is one electronic charge per atomic mass unit, this occurs at

$$L_{Edd} = 1.26 \times 10^{38}\ \frac{M_{BH}}{M_\odot}\ erg\ s^{-1} \qquad (12.4)$$

where $M_{BH}/M_\odot$ is the black hole mass in units of solar mass; for an incisive discussion see Begelman (1978).

The growth of black hole mass (M_{BH}) arising from accretion is given by

$$c^2 \frac{dM_{BH}}{dt} = \left(\frac{L_{bol}}{\epsilon} - L_{bol}\right) \qquad (12.5)$$

In terms of the Eddington ratio ($\Lambda \equiv L_{bol}/L_{Edd} \leq 1$) the expression for mass growth is

$$\left[\frac{dM_{BH}}{dt}\right] \tau_{Edd} = \Lambda(\epsilon^{-1} - 1)M_{BH} \qquad (12.6)$$

where

$$\tau_{Edd} \equiv \frac{Mc^2}{L_{Edd}} = 4.5 \times 10^8\ years \qquad (12.7)$$

In general, the effective e-folding time (τ) for black hole mass growth is then

$$\tau = \frac{\tau_{Edd}}{[(\Lambda(\epsilon^{-1} - 1)]} \qquad (12.8)$$

As emphasized by Begelman (1978) the luminosity from a steadily accreting compact neutron star or white dwarf is always of order (dM/dt)(GM/r), where r is the radius of the stellar surface. Black holes, of course, do not have a hard surface. Instead, the maximum value for the radiation efficiency (ϵ) for an accretion disk around a black hole is determined by the binding energy per unit mass (ϵc^2) of the most tightly bound stable orbit. For a non-spinning Schwarzschild hole this occurs at $r = 6GM_{BH}/c^2$, where $\epsilon = 0.057$. For an extreme Kerr spinning hole (i.e., characterized by a specific angular momentum GM_{BH}/c) the most tightly bound orbit co-rotates with the spin at $r = GM_{BH}/c^2$ where $\epsilon = 0.42$. Although the matter swallowed by the hole from an accretion disk would by itself eventually spin the hole up to the theoretical Kerr limit, radiation from the disk swallowed by the hole produces a counter torque which prevents spin-up beyond $0.998GM_{BH}/c$ (Thorne, 1974); this then limits $\epsilon \approx 0.3$ for the physically more realistic limiting case of a "canonical Kerr" hole. The spin-up time from a Schwarzschild to a canonical Kerr hole is comparable to the effective e-folding time (τ) for mass growth, as given by Equation (12.8). In this connection, it is interesting to note that, for $\epsilon = 0.057 - 0.3$

$$\Lambda\tau = (3 - 19) \times 10^7\ years$$

a span of possible characteristic growth times comparable to the look-back horizon time needed to encounter a neighboring Seyfert 1 galaxy; see Section 12.2.

12.4 SURFACE BRIGHTNESS

The surface brightness (I) of a black body of temperature $T°\ K$ is given by

$$I(erg\ s^{-1}\ cm^{-2}\ sterad^{-1}) = \left(\frac{\sigma}{\pi}\right) T^4 \qquad (12.9)$$

where σ is the Stefan-Boltzmann constant. With the understanding that the temperature is *as observed* this law is universal in Friedmann cosmology, in which a redshifted black body is also a black body. Since the mean photon energy $\langle E \rangle$ is proportional to T, [i.e., $\langle E \rangle = 2.68\ kT$], and gets redshifted as (1+z) we note that Equation (12.9) implies

$$I_{observed} = (1 + z)^{-4} I_{source} \qquad (12.10)$$

As stressed by Sandage and Perelmuter (1990) this result for surface brightness is completely general for an expanding universe, independent of both Ω and H_o. In terms of volume emissivity $q(erg\ s^{-1}cm^{-3})$, the differential contribution to the surface brightness, *observed at close range*, may be expressed as

$$\delta I = q\ c\ \frac{\delta t}{4\pi} \qquad (12.11)$$

where (δt) is the associated differential interval in light travel time. From Equations (12.10-12.11) we derive the general expression for I observed now,

$$4\pi\ I = c \int (1+z)^{-4}\ q\ \delta t \qquad (12.12)$$

obtained by integration over the finite interval (Δt) from any earlier cosmic time of interest to any later time $t \leq t_H$, the Hubble age of the universe. Considering discrete sources of number density (n) and mean luminosity (L) at cosmic time (t), we substitute $q = nL$ in Equation (12.12) to obtain the expression for the corresponding surface brightness of the sky

$$4\pi I = c \int (1+z)^{-4} nL dt \qquad (12.13)$$

arising from all such sources within the Δt starting at $t_{initial}$. (The maximum look-back light travel time is then $t_H - t_{initial}$.) In terms of the comoving density $\boldsymbol{n} \equiv n(1+z)^{-3}$, which is invariant with redshift for conserved objects, we obtain the simple and powerful result that

$$4\pi I = c \int (1+z)^{-1} \boldsymbol{n} L dt \qquad (12.14)$$

Defining the comoving volume emissivity as $\boldsymbol{q} \equiv \boldsymbol{n} L$, then yields the succinct and useful expression

$$4\pi I = c \int (1+z)^{-1} \boldsymbol{q} dt \qquad (12.15)$$

12.5 ACCRETION HISTORY

Taking $\boldsymbol{q} = \boldsymbol{n} L_{bol}$ as the comoving volume emissivity arising from the *bolometric* luminosity of the associated black hole galactic nuclei, Equation (12.5) implies that the comoving mass density of accreted matter (ρ_A) increases with cosmic time at the rate

$$\frac{d\rho_A}{dt} = c^{-2}(\epsilon^{-1} - 1)\boldsymbol{q} \qquad (12.16)$$

which integrates to

$$\rho_A = c^{-2} \int (\epsilon^{-1} - 1) \boldsymbol{q} dt \qquad (12.17)$$

As pointed out by Soltan (1982), Equation (12.17) is remarkably similar in form to Equation (12.15), both being integrals over the radiative history. However, if we take the sky surface brightness (I) as that of the CXB, we must take into account that the appropriate $\boldsymbol{q}$ for x-radiation in Equation (12.15) is then a fraction $(1/\eta)$ of the bolometric $\boldsymbol{q}$ in Equation (12.17), since $\eta \equiv L_{bol}/L_x \geq 1$. From Equations (12.15, 12.17) we obtain that the present-epoch mass density of black hole growth arising from accretion (ρ_A) is proportional to the total CXB omnidirectional flux $(4\pi I)$, viz

$$\rho_A = c^{-3}[\langle \eta(\epsilon^{-1} - 1)(1+z) \rangle](4\pi I) \qquad (12.18)$$

where the proportionality constant depends on η, ϵ *and* z via the indicated average over the relevant history of accretion.

12.6 COSMIC X-RAY BACKGROUND

The nearest resolved AGN X-ray sources ($at\ z \leq 0.02$) exhibit a highly anisotropic full-sky distribution indicative of the pronounced underlying large-scale structure of the present-epoch space sampled (Miyaji and Boldt, 1990). Evidently then, a region of this size is still much too small a portion of the universe for invoking the key *cosmological principle* of spatial homogeneity and isotropy (Weinberg, 1972). These foreground sources ($at\ z < 0.02$), however, make no more than a one-percent contribution to the total flux of extragalactic X-rays (Miyaji et al., 1994); the bulk of the total flux comes from the X-ray sources at much larger redshifts that constitute most of the CXB. This all-sky CXB is in fact remarkably isotropic (Boldt, 1987), thereby demonstrating that the matter-dominated regime of the universe that it traces is indeed consistent with the cosmological principle. In effect, this validates our use of light travel look-back time as the appropriate "third dimension" coordinate for tracking X-ray source evolution in cosmic time. The principal component of the CXB has a spectrum which may be integrated in closed form. This spectrum characteristic of the CXB (Boldt, 1992), based on observations over the 3-50 keV band with HEAO-1 A2 proportional counters (Marshall et al., 1980), is well described by the simple function

$$I_E \equiv \frac{dI}{dE} = 7.8 E^{-0.29} \exp\left(\frac{-E}{E_o}\right)\ cm^{-2} s^{-1} sr^{-1} \qquad (12.19)$$

where $E_o = 40\ keV$.

Results from HEAO-1 A4 scintillators indicate that this spectrum (as given by Equation 12.19) provides a good fit to the data up to about 100 keV (Gruber, 1992) while recent CCD data from the ASCA orbiting X-ray observatory (Gendreau et al., 1995) indicate that it also remains valid all the way down to 1 keV, below which a softer component associated with our galaxy becomes evident. Expressed as EI_E this spectrum exhibits a well-defined peak at E = 28.4 keV. Considering a three-component, six-parameter fit to the broadband composite HEAO-1 (A2 + A4) spectrum (Gruber, 1992) yields a peak in EI_E that occurs at E = 29.2 keV with a flux that is within 3% of the simple fit. The total omnidirectional

energy flux associated with integrating Equation (12.19) over all energies is

$$4\pi I = 2.8 \times 10^{-6} erg\ s^{-1}\ cm^{-2} \qquad (12.20)$$

Considering the more complete broadband background observed with HEAO-1 (Gruber, 1992) increases this flux by only $\sim$ 20% (for energies up to 0.5 MeV). Hence the dominant background component having the spectrum given by Equation (12.19) provides an especially good measure of the CXB flux. Under the currently favored hypothesis that essentially all the CXB arises from sources associated with black hole galactic nuclei powered by accretion (Daly, 1991) this flux is particularly appropriate in evaluating Equation (12.18) to obtain the total present-epoch mass density of black hole growth resulting from the entire history of accretion.

Using the flux given by Equation (12.20) in Equation (12.18), under the assumption that η and ϵ are invariant with epoch (i.e., independent of z), yields

$$\rho_A \left(\frac{M_\odot}{Mpc^3} \right) = 1.6 \times 10^3 \eta (\epsilon^{-1} - 1)(1 + \langle z \rangle) \qquad (12.21)$$

Evaluating Equation (12.21) for $\epsilon < 0.3$ *and* $\eta \geq 2$ gives the lower limit

$$\rho_A > 7.5 \times 10^3 (1 + \langle z \rangle) \frac{M_\odot}{Mpc^3} \qquad (12.22)$$

Since the Seyfert galaxies observed with the Compton GRO exhibit spectra that peak in EL_E at $E \approx 90$ keV (Johnson et al., 1994), their main contribution to the CXB, which peaks in EI_E at 29 $keV$is expected to be coming from redshifts at $z \approx 2$. Indeed, the X-ray luminosity evolution for AGN obtained with ROSAT by Boyle et al. (1993), for the case $\Omega = 1$, increases strongly with redshift only up to $z \approx 2$ and is then relatively flat up to z = 3. On the other hand, the evolution of comoving density for the optically brightest quasars peaks at z = 3.3 (Warren, Hewitt, and Osmer, 1994). Hence, it is reasonable to anticipate that $\langle z \rangle \geq 2$ for the CXB; using this in Equation (12.22) yields

$$\rho_A > 22 \times 10^3 \frac{M_\odot}{Mpc^3} \qquad (12.23)$$

Assuming that the AGN at $> 10^{42} erg\ s^{-1}$ (0.3-3.5 keV) in the sample used by Boyle et al. (1993) are mainly Seyfert 1 galaxies and that the number of Seyfert 2 galaxies is 2.3 times greater (Huchra and Burg, 1992) we can employ the mass function for Seyfert 1 galaxies (Padovani, Burg, and Edelson, 1990) to infer that the corresponding total present-epoch mass density of all Seyfert AGN nuclei is

$$\rho_{AGN} \approx 2 \times 10^3 h \frac{M_\odot}{Mpc^3} \qquad (12.24)$$

with a mean black hole mass

$$\langle M_{AGN} \rangle \approx 3 \times 10^7 h^{-2} M_\odot \qquad (12.25)$$

From Equations (12.23-12.24) we conclude that

$$\rho_A > 10 h^{-1} \rho_{AGN} \qquad (12.26)$$

12.7 BLACK HOLE POPULATION

Stellar-kinematic evidence for dark central masses in several galaxies (Kormendy and Richstone, 1992) indicates that the total density of supermassive Black Hole Galactic Nuclei (BGN) in the present epoch (ρ_{BGN}) is consistent with the lower limit for ρ_A from Equation (12.23) , much larger than ρ_{AGN}, viz

$$\rho_{BGN} \approx 14 \times 10^4 h\ (M_\odot / Mpc^3) \qquad (12.27)$$

and hence dominated by dormant objects, whereby

$$\rho_{BGN} \approx 70\ \rho_{AGN} \qquad (12.28)$$

With $n_{AGN} \approx 6 \times 10^{-5} h^3\ Mpc^{-3}$, we obtain that

$$\frac{\rho_{BGN}}{n_{AGN}} \approx 2 \times 10^9 h^{-2} M_\odot \qquad (12.29)$$

a mass much larger than the average value cited in Equation (12.25) for present-epoch Seyfert nuclei. Even if we consider only the *accreted* mass density (ρ_A) directly inferred from the CXB, clearly a lower limit for ρ_{BGN}, the same trend emerges, viz

$$\frac{\rho_A}{n_{AGN}} > 4 \times 10^8 h^{-3} M_\odot \qquad (12.30)$$

This indicates that the duty cycle for mass accretion (and hence being on as an X-ray bright AGN) has been small during the overall lifetime of an average individual black hole galactic nucleus (see also Padovani, Burg, and Edelson, 1990). It implies that the history of accretion responsible for the CXB was distributed over a substantially larger underlying population of BGN. Taking $\langle M_{BGN} \rangle = \langle M_{AGN} \rangle$, in the spirit of generalizing the unified AGN model, we conclude that the current AGN number density n_{AGN}) is probably only a few percent that of the underlying population n_{BGN}) and that

$$\frac{\rho_{BGN}}{n_{BGN}} = \langle M_{AGN} \rangle \qquad (12.31)$$

Since the comoving number density of Seyfert galaxies (n_{AGN}) appears to be invariant with respect to redshift for the X-ray sample considered by Boyle et al. (1993), at least up to z = 3, the constant comoving number density (n_{BGN}) expected for the much larger underlying population would imply that the probability for being "on"

is essentially the same throughout the evolution already traced (but see Small and Blandford, 1992). It has been speculated, however, that during the epoch of galaxy formation, when there was ample fuel for accretion, there could well have been a brief phase at $z > 3$ when all these nuclei were on as strong Eddington limited X-ray sources (Leiter and Boldt, 1993; Boldt and Leiter, 1994, 1995).

12.8 A COMPLETE HISTORY?

Boyle et al. (1993) have used ROSAT and HEAO-2 Einstein Observatory data on AGN observed up to z = 3 to determine the X-ray luminosity evolution of these objects. Since the luminosity determined for each AGN depends on the luminosity distance to the source as well as the observed flux the actual numerical values derived are not independent of the assumed cosmological model, as characterized by Ω, especially for $z \geq 1$. For $\Omega = 0$ they find that the luminosity function (dn /dL) is a power-law that breaks at a luminosity that increases with redshift, reaching $3.9 \times 10^{45} erg\ s^{-1}$ (0.3-3.5 keV) at $z = 3$. For $\Omega = 1$ the break luminosity reaches $1.2 \times 10^{45} erg\ s^{-1}$ at $z = 2$ and remains essentially invariant to z = 3. The corresponding broadband X-ray luminosities relevant to the CXB are appreciably larger. Using the appropriate luminosity functions, the associated average broadband X-ray luminosities may be obtained by assuming AGN spectra of the form compatible with that of the CXB. The resulting average broadband X-ray luminosities can be used to evaluate the corresponding mean mass values for black hole galactic nuclei at z = 3 (Leiter and Boldt, 1996). This yields

for $\Omega = 0$

$$\langle M_{AGN} \rangle = 15 \times 10^6 \left(\frac{\eta}{\Lambda}\right) M_\odot \qquad (12.32)$$

for $\Omega = 1$

$$\langle M_{AGN} \rangle = 7 \times 10^6 \left(\frac{\eta}{\Lambda}\right) M_\odot \qquad (12.33)$$

where we expect $\eta/\Lambda > 2$.
Since the comoving density of the brightest quasars decreases sharply at $z \geq 3.3$ (Warren, Hewitt, and Osmer, 1994), the mass values given by Equations (12.32-12.33) are taken to characterize the supermassive black holes that already exist at the presumed onset of the AGN phenomenon at $z \approx 3$.

Where did these massive black holes already present at $z \approx 3$ come from? Haehnelt and Rees (1993) consider that, in the course of galaxy generation, the baryonic component of the central part of a newly forming dark matter halo collapses into a self-gravitating structure. They suggest that relatively rapid black hole creation could then occur by protogalactic core collapse via a process of "supply driven" accretion whose initial rate vastly exceeds the Eddington value ($dM/dt > M/\tau_{Edd}$). In this situation radiation is initially trapped (i.e., does not escape). However, in their model, most of the actual mass buildup occurs during the final phase of core collapse when the corresponding Eddington limit for accretion has grown to be comparable to the externally enforced rate and during which time essentially all of the total radiation that eventually escapes actually does emerge. Efficient production of X-radiation during this brief final phase would yield a significant CXB component. For Eddington limited accretion the characteristic time for black hole mass growth and spin-up could be as short as 3×10^7 years, as shown in Section 12.3. During this phase these objects would qualify as viable candidates for the Precursor Active Galaxies (PAG) proposed as the sources to be associated with the earliest epoch of CXB generation (Leiter and Boldt, 1982, 1993; Boldt and Leiter, 1987; Zdziarski, 1988).

Another possibility involves the pre-galactic / post-decoupling formation of seed black holes having a Jeans mass ($10^{5-6} M_\odot$) at $z \approx 10^{2-3}$ (Carr, 1980; Carr and Rees, 1984; Gnedin and Ostriker, 1992). A subsequent brief period (e.g.,at $z \approx 3-4$) of black hole exponential growth to supermassive ($> 10^7 M_\odot$) galactic nuclei via Eddington- limited accretion could then also account for a PAG component of the CXB. Cosmological creation of full-grown supermassive black holes at $z = 10^{2-3}$, such as considered by Loeb (1993), would not yield a significant X-ray signal directly. Such objects could yield appreciable X-radiation only after the onset of the AGN phenomenon at much lower redshifts during the presumed epoch of galaxy formation.

As the total CXB intensity (I_{CXB}) is a direct measure of the *entire* history of accretion responsible for producing the supermassive black hole galactic nuclei (BGN) that exist in the present epoch (at z = 0), we likewise expect that there is a PAG portion (I_{PAG}) of this background that signals the early history of accretion responsible for producing the BGN already present at the onset of the AGN phenomenon, here taken to be at z = 3. Assuming the comoving density n_{BGN} and the efficiency ϵ as invariant, and that the mean BGN mass at z = 3 already greatly exceeds that of possible earlier seed masses, we use Equation (12.18) to construct the ratio $[\rho(z=0)/\rho(z=3)]$ that provides the relation

$$\frac{I_{PAG}\,\langle \eta(1+z)\rangle_{PAG}\ \ \langle M_{AGN}\rangle_0}{I_{CXB}\,\langle \eta(1+z)\rangle_{CXB}\ \ \langle M_{AGN}\rangle_3} = \qquad (12.34)$$

where $\langle M_{AGN}\rangle_0$ is the mean AGN (BGN) mass at z = 0 (see Equation 12.25), $\langle M_{AGN}\rangle_3$ is the mean mass at z = 3 (see Equations 12.32-12.33), and the indicated averages $\langle \eta(1+z)\rangle$ are over *all* z for the CXB and *only* $z \geq 3$ for the PAG portion. Taking $\eta/\Lambda > 2$ for the AGN at z = 3, $\eta_{PAG} \leq \eta_{AGN}, \langle z_{AGN}\rangle \approx 2$ *and* $6 > \langle z_{PAG}\rangle \geq$

3 *yields*

$$\frac{I_{PAG}}{I_{CXB}} > 0.2 \qquad (12.35)$$

These PAG are not only probable major contributors to the unresolved CXB (as suggested by Equation 12.35) but should be so numerous as to render the surface brightness of such a residual background much smoother on small angular scales than expected on the basis of AGN alone (Leiter and Boldt, 1996); this is due to the fundamental aspect of a picture in which $n_{PAG} = n_{BGN} \gg n_{AGN}$. Preliminary evidence obtained from ROSAT already suggests that there may indeed be an especially large number of faint unresolved sources of the residual CXB (Hasinger et al., 1993; Barcons et al., 1994).

Acknowledgement

Much of the material presented here is based on studies of the CXB carried out in collaboration with Darryl Leiter. His determination for obtaining a *complete* physically compelling picture of this phenomenon is greatly appreciated.

12.9 REFERENCES

Antonucci, R., and Miller, J., 1985, ApJ, 297, 621.
Awaki, H., 1991, PhD thesis, Nagoya University.
Bahcall, N., 1988, Ann. Rev. Astron. Astrophys., 26, 631.
Barcons, X., et al., 1994, MNRAS, 268, 833.
Begelman, M., 1978, MNRAS, 184, 53.
Boldt, E., 1987, Phys. Rept., 146, 215.
Boldt, E., 1992, in "The X-ray Background" eds. X. Barcons, and A. Fabian, (Cambridge: Cambridge Univ. Press), 115.
Boldt, E., and Leiter, D., 1994, in "The First Stromlo Symposium: The Physics of Active Galaxies" 1993, eds. G.V. Bicknell, M.A. Dopita, and P.J. Quinn, ASP Conf. Ser., 54, 131.
Boldt, E., and Leiter, D., 1987, ApJ, 322, L1.
Boldt, E., and Leiter, D., 1995, Nucl. Phys. B (Proc. Suppl.) 38, 440.
Boyle, B., Griffiths, R., Shanks, T., Stewart, G., and Georgantopoulos, I., 1993, MNRAS, 260, 49.
Carr, B., 1980, Nature 284, 326.
Carr, B., and Rees M., 1984, MNRAS 206, 315.
Daly, R., 1991, ApJ, 379, 37.
Gendreau, K., et al., 1995, PASJ, 47, L5.
Giacconi, R., Gursky, H., Paolini, F., and Rossi, B., 1962, Phys. Rev. Lett., 9, 439.
Gnedin, N., and Ostriker, J., 1992, ApJ, 400, 1.
Gruber, D., 1992, in "The X-ray Background" eds. X. Barcons, and A. Fabian, (Cambridge: Cambridge Univ. Press), 229.
Haehnelt, M., and Rees, M., 1993, MNRAS, 263, 168.
Hasinger, G., et al., 1993, Astron. Astrophys. 271, 1.
Huchra, J., and Burg, R., 1992, ApJ., 393, 90.
Johnson, N., et al., 1994, in "Second Compton Symposium" Proceedings, (1993), eds. C. Fichtel, N. Gehrels, and J. Norris, AIP Conf. Proc. 304, 515.
Kormendy, J., and Richstone, D., 1992, ApJ, 393, 559.
Leiter, D., and Boldt, E., 1982, ApJ, 260,1.
Leiter, D., and Boldt, E., 1993, ASP Conf. Ser., 51, 473.
Leiter, D., and Boldt, E., 1996, in preparation.
Loeb, A., 1993, ApJ 403, 542.
Marshall, F., et al., 1980, ApJ, 235, 4.
Mather, J., et al., 1990, ApJ, 354, L37.
Miyaji, T., and Boldt, E., 1990, ApJ, 353, L3.
Miyaji, T., Lahav, O., Jahoda,.K., and Boldt, E., 1994, ApJ, 434, 424.
Padovani, P., Burg, R., and Edelson, R., 1990, ApJ, 353, 438.
Rees, M., 1984, Ann. Rev. Astron. Astrophys., 22, 471.
Sandage, A., and Perelmuter, J-M., 1990, ApJ, 361, 1.
Small, T., and Blandford, R., 1992, MNRAS, 259, 725.
Soltan, A., 1982, MNRAS, 200, 115.
Thorne, K., 1974, ApJ, 191, 507.
Warren, S., Hewitt, P., and Osmer, P., 1994, ApJ, 421, 412.
Weinberg, S., 1972, in "Gravitation and Cosmology" New York: Wiley.
Zdziarski, A., 1988, MNRAS, 233, 739.

Chapter 13

INVESTIGATING DARK MATTER IN THE UNIVERSE WITH X-RAY OBSERVATIONS

MICHAEL LOEWENSTEIN
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center, Code 662
Greenbelt, MD 20771

ABSTRACT

An overview of the motivation and methods of the X-ray astronomy approach to the dark matter problem is presented. The basic equations and techniques of solution are given, some recent results are reviewed, and the historical and cosmological contexts are briefly discussed.

13.1 INTRODUCTION

Between 75 and 99 percent of the mass in the universe is invisible, detectable only through its gravitational influence on luminous matter. This fact is of great interest to particle physicists and cosmologists trying to understand the most fundamental questions about the very early universe and the subsequent formation and evolution of the galaxies, clusters of galaxies, and clusters of clusters of galaxies that make up the observable universe today.

Current dark matter research efforts can be thought of as an attempt to seek answers to the following three major questions:

1. What is the dark matter made of?
2. How much dark matter is there, and is there enough to close the universe?
3. Where is the dark matter? (How is it distributed in the universe and within individual objects?)

Because hot, X-ray emitting gas in elliptical galaxies and clusters of galaxies traces the dark matter in these systems one can use X-ray observations to directly address questions (2) and (3). And because different kinds of dark matter are expected to be distributed differently question (1) can, in principle, be indirectly addressed as well.

13.2 HISTORICAL CONTEXT: THE MISSING MASS PROBLEM

If one measures mass and luminosity in solar units then the mass-to-light ratio (M/L) for observed populations of stars is generally within the range 1-5, depending on the age, metallicity, and distribution of stellar masses. When dynamical estimates of the mass divided by the observed integrated light greatly exceed this value in some system there, historically, was said to be a 'missing mass' problem (even though the light, rather than the mass, is what is actually missing). Discrepancies between the observed light and the inferred mass have been noted at least as far back as the early 1930s with studies by Oort (1932) of the dynamics of stars in the local galaxy and by Zwicky (1933) of the dynamics of galaxies in the Coma Cluster, with other important milestones including the extensive work on flat rotation curves in spiral galaxies by Rubin, Ford, and Thonnard (1978) and the estimate of the mass needed to bind the hot gas in the elliptical galaxy M87 by Mathews (1978).

For more detailed historical background see one of the reviews by Trimble (1987, 1995) or the popular text by Parker (1989).

13.3 HYDROSTATIC EQUILIBRIUM

13.3.1 When is Hydrostatic Equilibrium a Good Approximation?

A point mass in a circular orbit at a distance r from the center of a galaxy has a speed (the circular speed)

$$v_{circ} = \left(\frac{GM(<r)}{r}\right)^{\frac{1}{2}} \tag{13.1}$$

where $M(<r)$ is the mass within the sphere enclosed by r. Crudely, this is how the mass distribution in spiral galaxies is estimated from the measured speeds of gas clouds or stars.

Similarly, the equation of motion (or, equivalently, an expression of the conservation of momentum) for a gas in an external gravitational field is

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} = -\frac{1}{\rho}\frac{\partial P}{\partial r} - \frac{v_{circ}^2}{r} \tag{13.2}$$

I have assumed that spherical symmetry holds and will continue to do so; however, everything can be generalized to two or three dimensions. In Equation (13.2), r is the radial coordinate, u the gas flow speed, ρ the mass density of the gas, and P the gas pressure. The circular velocity is defined as above with the mass referring to the total mass enclosed within r. The first term on the right-hand side indicates that forces are induced if there is a gradient in the pressure. From the ideal gas law $P = \rho kT/\mu m_p = \rho c_s^2$, where k is Boltzmann's constant, T the gas temperature, μm_p the mean mass per particle, m_p the proton mass, and $c_s \equiv (kT/\mu m_p)^{1/2}$ the isothermal sound speed. I have assumed that all nonthermal/nongravitational forces (e.g., from magnetic fields) are negligible.

Equation (13.2) simplifies considerably (to the equation of hydrostatic equilibrium) if one ignores the two terms on the left-hand side. Under what conditions is this justified? A necessary (but not always sufficient) condition for neglecting the time-dependent first term is that there has been sufficient time for any initial inhomogeneities to be erased. Since disturbances propagate at the sound speed, a spherical gas distribution will have settled down out to a radius equal to the the sound speed times the age of the system (τ), the sound crossing radius

$$r_{sc} = c_s\tau \approx 10\left(\frac{T}{10^8\ K}\right)^{\frac{1}{2}}\left(\frac{\tau}{10^{10}\ yr}\right)\ Mpc \tag{13.3}$$

where temperature and age are in units appropriate for clusters of galaxies. This expression can also be derived by comparing the first and third terms in Equation (13.2) and using dimensional analysis.

If one compares the orders of magnitude of the second and pressure gradient terms in Equation (13.2) it is easy to show that the former is negligible if the flow speed is much less than the sound speed ($u << c_s$). Another way of expressing this condition is in terms of the mass flux, $\dot{M} \equiv 4\pi r^2 \rho u$,

$$u << c_s \rightarrow \dot{M} << 2\ 10^5 \left(\frac{\rho}{10^{-27}\ g\ \ cm^{-3}}\right) \left(\frac{r}{1\ \ Mpc}\right)^2 \left(\frac{T}{10^8\ K}\right)^{\frac{1}{2}} M_\odot\ \ yr^{-1} \qquad (13.4)$$

where again quantities have been scaled to physical units appropriate for the intracluster medium. I apply this equation, as well as Equation (13.3), to intracluster and elliptical galaxy interstellar media in subsequent sections.

13.3.2 Hydrostatic Equilibrium and Characteristic Temperatures

In the previous subsection it has been shown that if a system has been undisturbed for a length of time longer than the time for a sound wave to cross the system, and if any gas flows are subsonic ($u << c_s$), that hydrostatic equilibrium is a good approximation. One can then write Equation (13.2) as

$$\frac{v_{circ}^2}{r} = -\frac{1}{\rho}\frac{dP}{dr} = -\frac{1}{\rho}\frac{d\left(\rho c_s^2\right)}{dr} \qquad (13.5)$$

that expresses the balance between (inward) gravitational and (outward) pressure forces.

The following similar equation holds for the stars orbiting in a galaxy, or for galaxies orbiting in a cluster of galaxies (Binney and Tremaine, 1987):

$$\frac{v_{circ}^2}{r} = -\frac{1}{\rho_*}\frac{d(\rho_* v_r^2)}{dr} - \frac{2\beta v_r^2}{r} \qquad (13.6)$$

This equation, the spherical Jeans equation, essentially treats a system of self-gravitating particles as a collisionless gas with mass density (for stars in galaxies or galaxies in clusters) ρ_* and pressure $\rho_* v_r^2$, where v_r^2 is the square of the radial velocity dispersion at r. Unlike the case for a system of collisional gas particles, the velocities in a collisionless gas may not be isotropic; thus the extra term in Equation (13.6). The anisotropy of the velocity distribution at r is characterized by

$$\beta \equiv 1 - \frac{v_t^2}{v_r^2} \qquad (13.7)$$

where v_t^2 is the square of the tangential velocity dispersion. If the velocity dispersion is isotropic ($v_r = v_t$) then $\beta = 0$. In general, there is no *a priori* way to estimate β, which lends the application of Equation (13.6) a built-in uncertainty lacking in the application of Equation (13.5).

Systems in hydrostatic equilibrium have a characteristic 'gravitational' temperature that can be derived from Equation (13.5) under the assumption that the gas is isothermal, i.e., the temperature and sound speed are constant. In that case

$$v_{circ}^2 = -c_s^2 \frac{dlog\rho}{dlogr} \qquad (13.8)$$

Since the density slope is generally on the order of unity; equating the sound speed and circular speeds defines the temperature

$$T_{grav} \equiv \frac{G\mu m_p}{k}\frac{M}{r} \approx 0.3 \left(\frac{M}{M_\odot}\right)\left(\frac{r}{1\ \ pc}\right)^{-1}\ K \qquad (13.9)$$

For galaxies, where $M \sim 10^{12}$ $M_\odot$ and $r \sim 3\ 10^4$ pc, $T_{grav} \sim 10^7$ K, while for clusters $M \sim 10^{15}$ $M_\odot$ and $r \sim 3\ 10^6$ pc so that $T_{grav} \sim 10^8$ K. These temperatures correspond to X-ray energies, which is why X-ray observations of hot gas in hydrostatic equilibrium is a primary method for deriving the gravitational potential in these systems.

13.4 APPLICATION TO REAL X-RAY EMITTING GRAVITATIONALLY BOUND PLASMAS

Using the definitions for circular velocity and thermal pressure Equation (13.5) can be rewritten as

$$M(<r) = -\frac{kTr}{G\mu m_p}\left(\frac{dlog\rho}{dlogr} + \frac{dlogT}{dlogr}\right) \qquad (13.10)$$

So, in principle, all one need do is measure the temperature, and density and temperature gradients at every radius of interest to derive the mass distribution. In practice one can only measure the three-dimensional density and temperature distributions projected onto the plane of the sky and with limited precision and angular resolution.

13.4.1 Isothermal Plasmas

Since there is considerable simplification if one assumes isothermality it is instructive to initially consider the constant temperature case. Observationally, isothermality seems to be not too bad an approximation over a substantial range in radius for many clusters and elliptical galaxies. Under this assumption the temperature gradient term in Equation (13.10) vanishes, and the mass simply depends on the values of the (constant) temperature and the density gradient.

Now, the gas density is not directly observed. Instead the X-ray surface brightness (Σ_x), which is the volume emissivity (ϵ_x) integrated along the line of sight (z), is measured.

$$\Sigma_x(R) = 2\int_0^\infty \epsilon_x dz \tag{13.11}$$

where the factor of two is a result of mirror symmetry about the midplane, and R (as distinguished from r) is the distance from the center of the system in projection so that $z^2 = r^2 - R^2$ and

$$\Sigma_x(R) = 2\int_R^\infty \frac{\epsilon_x r dr}{(r^2 - R^2)^{1/2}} \tag{13.12}$$

This integral is of the Abel form (see Binney and Tremaine, 1987) and can be inverted to yield

$$\epsilon_x = -\frac{1}{\pi}\int_r^\infty \frac{d\Sigma}{dR}\frac{dR}{(R^2 - r^2)^{1/2}} \tag{13.13}$$

For hot, diffuse plasmas in elliptical galaxies and clusters of galaxies emission mechanisms include free-free (thermal bremmstrahlung), bound-free (recombination), and bound-bound (collisionally excited line emission) processes (Sarazin, 1986). A common property of all these is that they depend on collisions between ions and electrons and, therefore, the total emissivity is proportional to the square of the gas density and can be written as

$$\epsilon_x = \rho^2 \Lambda(T) \tag{13.14}$$

The function Λ subsumes the temperature dependence of the emissivity derived from detailed consideration of the appropriate atomic physics (and can be generalized to include the instrumental efficiency if the above surface brightness is actually observed with a detector).

The temperature itself is derived from *spectral fitting*. For diffuse plasmas of the sort observed in elliptical galaxies and clusters of galaxies the shape of the energy spectrum depends only on the temperature and elemental abundances. Therefore, by varying these parameters until a model that reproduces the observed spectrum is obtained one can measure the temperature (and metallicity) of the hot gas.

Once the temperature and surface brightness distribution have been measured it is possible to derive the density profile from Equations (13.13) and (13.14) and thus the total mass profile from Equation (13.10). Often the surface brightness is well-fit by a function of the form

$$\Sigma_x = \Sigma_o \left(1 + \frac{R^2}{a^2}\right)^{-\alpha} \tag{13.15}$$

from which it follows from Equations (13.13) and (13.14) that

$$\rho = \rho_o \left(1 + \frac{r^2}{a^2}\right)^{-(2\alpha+1)/4} \tag{13.16}$$

where

$$\rho_o = \left(\frac{\Sigma_o \Gamma(\alpha + 1/2)}{\sqrt{\pi}\Lambda(T) a \Gamma(\alpha)}\right)^{1/2} \tag{13.17}$$

and Γ is the gamma function. Therefore, fitting the surface brightness profile for the core radius a, the normalization Σ_o, and the slope α determines the density profile and, therefore, the total gravitational mass profile as

$$\begin{aligned} M(<r) &= \frac{kT(\alpha + 1/2)}{G\mu m_p}\frac{r^3}{r^2 + a^2} \\ &\approx 3\ 10^{14}(\alpha + 1/2)\left(\frac{T}{10^8\ K}\right) \\ &\quad \left(\frac{r}{1\ Mpc}\right)\left(\frac{r^2}{r^2 + a^2}\right) M_\odot \end{aligned} \tag{13.18}$$

13.4.2 Relaxing the Isothermal Assumption

At present one can obtain very accurate X-ray surface brightness profiles with good spatial resolution for elliptical galaxies and clusters of galaxies using the imaging instruments on the ROSAT satellite. Because $\Lambda(T)$ (Equation 13.14) is a weak function of temperature accurate density profiles can be obtained even when the gas is not isothermal. With the recently (1993) launched Advanced Satellite for Cosmology and Astrophysics (ASCA) very accurate spatially resolved X-ray spectroscopy from which precise temperature profiles can be derived (see section 13.7) is now possible. However, the spatial resolution of ASCA is relatively crude; therefore, Equation (13.10) cannot be directly applied.

Instead, recast Equation (13.5) into yet another form

$$\frac{kT}{\mu m_p} = \frac{p_\infty}{\rho} - \frac{1}{\rho}\int_\infty^r \frac{GM(<r)\rho}{r^2} dr \tag{13.19}$$

where p_∞ is a pressure boundary condition. The total mass ($M(<r)$) can be broken up into an observed component (consisting of the hot gas observed in X-rays plus all the stars in the case of galaxies or all the stars in all the galaxies in the case of clusters) that is directly measured plus a dark matter component that one hopes to place constraints on. The dark matter component is assumed to have some functional form, and the parameters of this function are then varied until the temperature calculated using Equation (13.19) matches the observed temperature profile. One choice, motivated by recent numerical simulations of cluster formation, is

$$M_{dark} = 4\pi\rho_{dark}a_{dark}^3\left(log\frac{r+a_{dark}}{a_{dark}} - \frac{r}{r+a_{dark}}\right) \quad (13.20)$$

where the dark matter normalization, ρ_{dark}, and scale-length, a_{dark}, are the parameters to be fit for.

I now turn to some applications of the methods just described.

13.5 DARK MATTER IN ELLIPTICAL GALAXIES

It is now well-established from the study of rotation curves that spiral galaxies have extended dark halos that account for $\sim 90\%$ of their total masses (e.g., Persic and Salucci, 1990). For elliptical galaxies the situation has been much less clear. There are, generally, no extended HI disks from which to derive rotation curves. One can measure the stellar density and velocity dispersion profiles and apply Equation (13.6); however, there are two substantial difficulties. Firstly, the degree of anisotropy of the velocity dispersion (Equation 13.7) is not known - some uncertain assumptions about stellar orbits must invariably be made. Secondly, deriving stellar velocity dispersions in elliptical galaxies involves precise measurements of weak absorption lines and becomes increasingly difficult at radii beyond 3-10 *kpc* where the optical surface brightness falls to very low levels. The optical data is consistent with little or no dark matter inside a few *kpc*; however, in spiral galaxies dark matter becomes more important the further out one looks and this could very well be the case in ellipticals as well.

This is where observations of X-ray emitting gas become crucial. Many elliptical galaxies, including virtually all of the most luminous ones, have very extended distributions of hot gas. Advantages of X-ray observations of hot gas over optical observations of stars include freedom from uncertainties due to velocity anisotropy and the feasibility of measuring the density and temperature out to tens of *kpc* where one might expect dark matter to dominate the gravitational potential. For a review of X-ray properties of elliptical galaxies see Fabbiano (1989).

13.5.1 The Origin of the Gas in Ellipticals and the Hydrostatic Assumption

By measuring the X-ray surface brightness profile one can derive the gas density profile as explained in the previous section and integrate this to derive the total gas mass. For a typical bright elliptical galaxy ($L \sim 10^{11}$ $L_\odot$) $M_{gas} \sim 10^9$-10^{10} $M_\odot$. The gas originates as stellar mass loss from red giant winds and planetary nebulae that collide and are shock-heated to approximately the gravitational temperature of $\sim 10^7$ K (Equation 13.9). Additional heating may result from supernova explosions, and some galaxies in cluster environments may accrete intracluster gas. The amount of mass return, $\dot{M}_*$, is on the order of 1 $M_\odot$ yr^{-1} per 10^{11} $L_\odot$, so that the total gas masses are easily explained as the accumulated mass loss over 10^{10} yr. Actually, because the mass-losing stellar population has evolved, $\dot{M}_*$ was larger in the past and some gas must have either been converted into stars or escaped the galaxy.

Hydrostatic equilibrium is generally a good approximation in elliptical galaxies. The sound crossing radius (Equation 13.3) is much larger than than the size of a galaxy, so there is adequate time for the gas to have settled down. The evolution of the source of the hot gas ($\dot{M}_*$) means that a steady state is never exactly attained, but the evolution is slow enough that the time-dependent term in Equation (13.2) can generally be neglected. In this quasi-steady state the flow of gas through the galaxy will be on the order of the mass input rate, $\dot{M}_*$. Plugging in the observed values ($\rho \sim 10^{-26}$ g cm^{-3} and $T \sim 10^7$ K at $r \sim 10$ kpc) into Equation (13.4), one finds that the gas flow must be highly subsonic, and that the hydrostatic assumption is justified.

13.5.2 Dark Matter Estimates in NGC 4636 and Other Giant Ellipticals

One of the best studied elliptical galaxies in X-rays is NGC 4636. Figure 13.1 shows the temperature profile from Mushotzky et al., 1994. Superposed on the observed temperature is the best-fit model obtained using Equation (13.19). The mass-to-light ratio required to bind hot gas with the observed density and temperature profiles is shown in Figure 13.2. The ratio starts to significantly deviate from the stellar value (~ 11) inside 10 kpc (close to the radius that encloses half of the light in NGC 4636), and dark matter dominates beyond 20 *kpc* where M/L_B monotonically increases outwards. Nearly isothermal gas distributions extending well beyond the optical galaxy have been found in a number of giant elliptical galaxies and absolutely require the presence of dark matter if they are bound.

Table 13.1 shows a 1994 compilation of various X-ray mass estimates for elliptical galaxies, the radii within which these estimates were acquired, and the satellite used to obtain the temperature profile. The mass-to-light ratios range from 25 to 150, all well above the stellar value of $\sim$ 5-10. The total mass mostly ranges from 10^{12} to $10^{13}M_\odot$; however, larger masses are inferred for galaxies with very extended gas halos - these may be more appropriately categorized with intragroup media (see next section).

Table 13.1: Dark Matter in Elliptical Galaxies

galaxy (NGC)	mission	R (kpc)	$M(<R)$ ($M_\odot$)	M/L_B ($M_\odot/L_\odot$)	$M(<R)/R$ ($10^{10}\ M_\odot/kpc$)
1399	BBXRT	38	$\sim 2\ 10^{12}$	~ 40	~ 5
1399	PSPC	125	$4.3-8.1\ 10^{12}$	$38-75$	$3-6$
1399	ASCA	750	$\sim 5.0\ 10^{13}$	~ 400	~ 6
720	PSPC	22-44	$0.41-1.4\ 10^{12}$	$35-120$	$2-3$
1332	PSPC	44	$0.38-1.7\ 10^{12}$	$32-143$	$1-4$
507	PSPC	460	$1.8-2.6\ 10^{13}$	$81-113$	$4-5$
499	PSPC	140	$0.82-1.1\ 10^{13}$	$76-104$	$6-8$
4261	PSPC	620	$1.5-2.6\ 10^{13}$	$115-200$	$2.5-4$
4472	PSPC	100	$5.8-6.9\ 10^{12}$	$78-93$	$6-7$
4472	ASCA	80	$3.8-4.6\ 10^{12}$	~ 45	~ 5
4406	ASCA	88	$2.8-3.6\ 10^{12}$	~ 70	~ 4
1404	ASCA	13	$\sim 6\ 10^{11}$	~ 40	~ 5
4374	ASCA	15	$\sim 1\ 10^{12}$	~ 24	~ 7
4636	ASCA	100	$4.4-5.2\ 10^{12}$	$126-153$	$4-5$

13.6 DARK MATTER IN GALAXY GROUPS

Roughly half of all the galaxies in the universe (including the Milky Way) reside in loose groups that consist of 3-30 galaxies within a $\sim 1\ Mpc$ region. Detection of very faint, diffuse gas in the space between galaxies in groups (the intragroup media) has become possible only relatively recently using ROSAT. Forty-eight groups have been analyzed by Mulchaey et al., 1996, and intragroup gas with a temperature of about $10^7\ K$ has been detected in about half of them. The centroid of the extended emission is

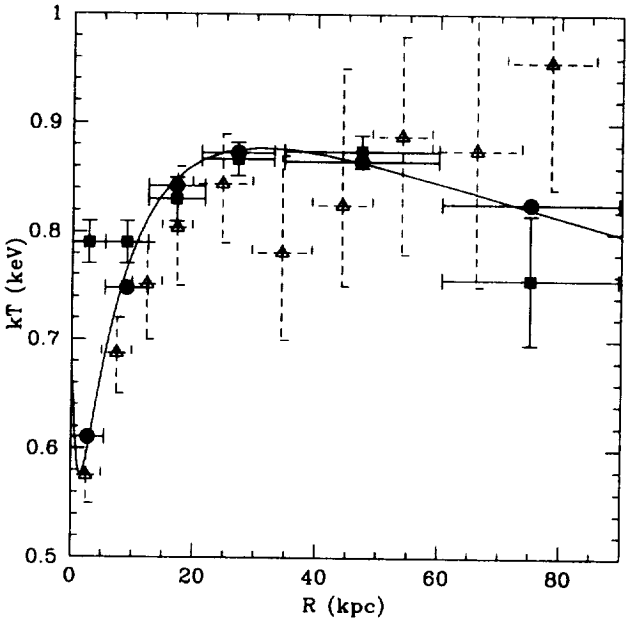


Figure 13.1: Filled boxes and open trianges with error bars show the observed temperature profiles from ASCA and ROSAT, respectively. The filled disks connected by the solid line show the temperatures derived using Equation (13.19) and the mass model shown in Figure 13.2 that represent the best-fit to the data.

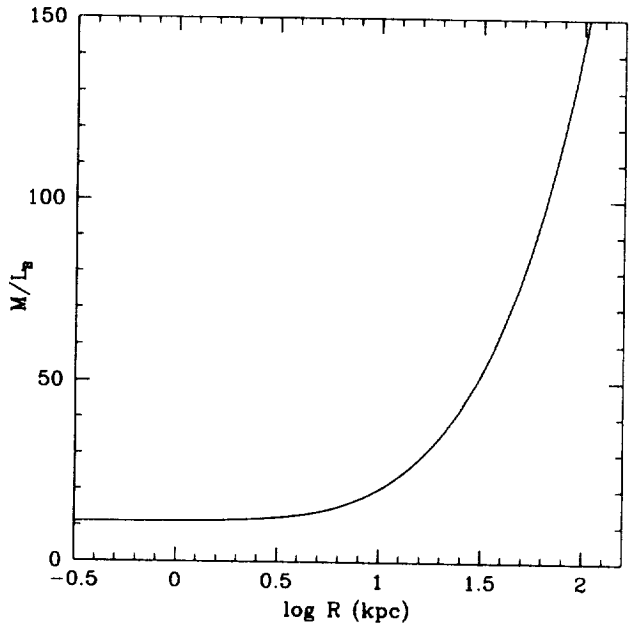


Figure 13.2: The mass-to-light ratio, M/L_B, that provides the best fit to the observed temperature profiles shown in Figure 13.1.

not necessarily coincident with the brightest galaxy in the group. The temperature gradient is generally not well determined (and when it is, is consistent with being flat), so isothermality is assumed. Equations (13.15)-(13.18) are then used to derive the total gravitating mass. Typically these groups are found to have masses of 1-4 10^{13} $M_\odot$ within radii of 200-600 kpc. When all the mass in all the groups in the universe is summed a mass on the order of that needed to close the universe is obtained.

13.7 DARK MATTER IN RICH CLUSTERS OF GALAXIES

Rich clusters of galaxies are the largest bound entities in the universe and are much more rare than groups. Each cluster can contain hundreds or even in excess of a thousand galaxies and spans several Mpc. X-ray emission from clusters was discovered 30 years ago and identified as hot intracluster gas ten years later. (See Sarazin, 1986 for a comprehensive review.) Soon after, Equation (13.10) was applied to estimate the total mass in clusters and it was immediately clear that dark matter was required to bind the hot gas (Mathews, 1978). The total amount of hot gas exceeds that in stars by factors of two to five, and the temperature of the hot gas ranges from 2-15 10^7 K. The hot gas is mostly primordial material left over from the galaxy formation epoch; however, early star formation activity and supernova-driven ejection in galaxies has enriched the intracluster gas to, typically, one-third solar abundances.

From Equation (13.3) it is clear that clusters should be relaxed out to $\sim$ 10 Mpc and thus in a steady state. However, it should be noted that many clusters are still forming at the present epoch, and in these systems the steady state assumption breaks down. Clusters form by the aggregation of smaller units (subclusters), and evidence of substructure is seen in many clusters. Equation (13.5) and its variants cannot be applied to such systems, or can only be applied by 'removing', with due caution, regions of substructure. Somewhere between one- and two-thirds of nearby systems are relaxed. As with galaxies and groups, the gas density profile can be obtained from the X-ray surface brightness profile (Equations 13.13-13.17) and integrated to give the total gas mass. M_{gas} ranges from $\sim$ 10^{13} to several times 10^{14} $M_\odot$. If the gas were flowing supersonically the mass flux would exceed the expression given in Equation (13.4) and the cluster would be emptied of gas in $M_{gas}/\dot{M} \sim 10^8$-$10^9$ yr. Since this is much less than the age of clusters and hot gas is observed in all rich clusters the gas must be, by-and-large, subsonic and one is justified in treating the intracluster media in relaxed clusters as hydrostatic.

13.7.1 The β Model

A commonly used and fairly accurate function used to characterize the density of galaxies in clusters is the King (1962) function,

$$\rho_{gal} = \rho_{gal,o}\left(1+\frac{r^2}{a^2}\right)^{-3/2} \tag{13.21}$$

This is an analytic approximation for a distribution of galaxies with isotropic orbits and a Gaussian radial velocity distribution with a constant velocity dispersion. Now, assume that both the gas and galaxies are in hydrostatic equilibrium in the same potential. One can then equate the right-hand sides of Equations (13.5) and (13.6). Moreover, assume that the galaxy orbits are isotropic - $\beta = 0$ in Equation (13.6) - and that both the gas and galaxy distributions are isothermal - v_r^2 and c_s^2 are independent of r in Equations (13.5) and (13.6). One can then easily derive the relationship

$$\frac{dlog\rho_{gas}}{dlog\rho_{gal}} = \frac{v_r^2}{c_s^2} \tag{13.22}$$

or

$$\rho_{gas} \propto \rho_{gal}^{\beta} \tag{13.23}$$

or

$$\rho_{gas} = \rho_{gas,o}\left(1+\frac{r^2}{a^2}\right)^{-3\beta/2} \tag{13.24}$$

where (utilizing the definition of the isothermal sound speed from section 13.3)

$$\beta \equiv \frac{v_r^2}{c_s^2} = \frac{\mu m_p v_r^2}{kT} \tag{13.25}$$

(Unfortunately, the same Greek letter, β, is used to denote this galaxy-to-gas energy ratio that is used for the orbital anisotropy parameter defined in Equation 13.7.) The β parameter is related to the surface brightness slope α (Equation 13.15) by $\alpha = 3\beta - 1/2$. The β model form for the gas density (Equation 13.24) is often utilized as a convenient parameterization, even in cases where the gas and/or the galaxies are not isothermal.

A simple way to parameterize a nonisothermal gas distribution is to assume a polytropic relation between the pressure and density (Cowie, Mushotzky, and Henriksen, 1987),

$$P \propto \rho^{\gamma}, T \propto \rho^{\gamma-1} \tag{13.26}$$

and therefore

$$T = T_o\left(1+\frac{r^2}{a^2}\right)^{-3\beta(\gamma-1)/2} \tag{13.27}$$

where γ is called the polytropic index. Equation (13.10) can now be used to generalize Equation (13.18) to

$$M(<r) \approx 10^{15}\beta\gamma \left(\frac{T_o}{10^8\ K}\right)\left(\frac{r}{1\ Mpc}\right) \times \frac{r^2/a^2}{(1+r^2/a^2)^{1+3\beta(\gamma-1)/2}} M_\odot. \quad (13.28)$$

13.7.2 Recent Determinations of Masses of Rich Clusters of Galaxies

An important uncertainty in all mass determinations using X-ray observations prior to 1990 or so was the lack of availibility of accurate temperature profiles. This has been remedied with the launch of ASCA. Figure 13.3 shows the ASCA spectrum of the central region of the Virgo cluster.

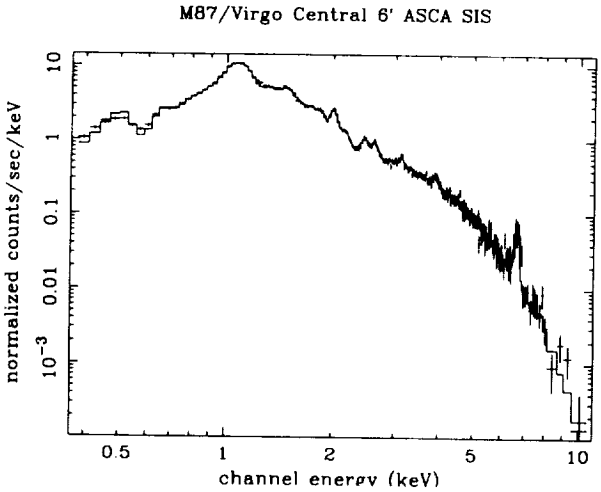


Figure 13.3: ASCA X-ray spectrum (error bars) and best-fit spectral model (histogram) for the central 6′ ($\sim$ 25 *kpc*) of the Virgo cluster (courtesy of U. Hwang). The best-fit model consists primarily of hot gas with a temperature of $\sim 2 \times 10^7$ *K* and elemental abundances about one-half solar. The features are emission lines from O, Mg, Si, S, Ca, Ar, and Fe (as the hot gas increases in temperature there are fewer of these features as more atoms become fully ionized). Because of the good energy resolution and signal-to-noise ratio the temperature is determined to a precision of a few percent.

The spectral resolution $E/\Delta E \sim$ 10-30; therefore, the shape of the spectrum is very well determined and the temperature can be derived with errors of only a few percent. The ASCA detectors are imaging spectrometers, and temperatures in 3-7 concentric annuli can be obtained out to a radius of $\sim$ 20′ (or the radius where the cluster ends, whichever comes first). This angular scale (assuming a Hubble constant $H_o = 50\ km\ s^{-1}\ Mpc^{-1}$) corresponds to $\sim$ 80 *kpc* for the nearest rich cluster (the Virgo cluster at a distance of $\sim$ 15 *Mpc*), and $\sim$ 1.6 *Mpc* for a typical well-studied rich cluster at a redshift of 0.05 (a distance of $\sim$ 300 *Mpc*). When combined with the gas density profiles derived from the ROSAT surface brightness profiles these temperatures have been recently utilized to constrain the dark matter distribution in clusters via Equation (13.19). The total mass within the radius where the temperature is measured can now be determined to accuracies of 10-30%; at larger radii one must extrapolate and the uncertainties grow with radius. At the maximum radius where hot intracluster gas is detected – typically 2-3 *Mpc* – the total mass is generally determined to within a factor of two.

Table 13.2 on the following page shows the total mass, in units of 10^{14} $M_\odot$, evaluated at the indicated radius (R_{max} measured in *Mpc*) for five clusters of galaxies with X-ray surface brightness profiles measured using ROSAT and temperature profiles measured using ASCA. Also shown are the mass breakdowns, in terms of fractions of the total mass at R_{max}, into stars (in galaxies), gas, and dark matter. The sum of the stellar and gas mass fractions is called the baryon fraction. This quantity generally slowly rises with radius (the dark matter is more concentrated than the intracluster gas). The baryon fraction is 20-25% for the largest (i.e., hottest, most X-ray luminous, most massive) clusters such as Abell 496, Abell 2199, and Abell 1795 and seems to show a wider range of variation from 10-25% for smaller clusters. Table 13.2 also provides a rough idea of the global range for the total mass and mass breakdown in the three types of systems – giant elliptical galaxies, groups of galaxies, and rich clusters of galaxies – discussed in this review. Interestingly, the range of baryon fractions is quite similar from giant elliptical galaxy through group all the way up to rich cluster scales although, on average, there may be a tendency for more massive systems to have larger baryon fractions. The form that the baryons take changes dramatically with scale: most of the baryonic matter in rich clusters is intracluster gas, while most of the baryons in galaxies are in stars. In groups the relative proportions of gas and stars vary considerably; however, on average the two components are comparable.

13.8 DIGRESSION: HUBBLE CONSTANT DEPENDENCES

It should be noted that the quantities in Table 13.2 are calculated assuming a Hubble constant $H_o = 50\ km\ s^{-1}\ Mpc^{-1}$ and scale in various ways with H_o. Directly observed quantities such as the flux and temperature are, of course, independent of H_o. All length-scales (including

Table 13.2: Total Mass and Mass Breakdown

	R_{max} (Mpc)	M_{total} $(10^{14}M_{\odot})$	gas/total	stars/total	dark/total
Abell 496	0.95	2.9-3.2	0.19-0.21	0.019-0.021	0.77-0.79
Abell 2199	0.86	2.4-2.7	0.17-0.19	0.045-0.050	0.76-0.78
Abell 1795	0.92	3.8-3.9	0.20-0.22	0.028-0.032	0.75-0.76
Abell 1060	0.38	0.68-0.76	0.069-0.077	0.053-0.059	0.86-0.88
AWM 7	0.61	1.1-1.4	0.15-0.19	0.056-0.074	0.73-0.79
clusters	1-5	3-30	0.1-0.3	0.02-0.05	0.7-0.9
groups	0.3-1	0.3-1	0.01-0.15	0.03-0.15	0.7-0.9
e-galaxies	0.03-0.2	0.01-0.2	0.01-0.08	0.05-0.3	0.7-0.9

R_{max} in Table 13.2) scale as H_o^{-1}, as does (by Equation 13.9 or 13.18) the total mass. Optically determined stellar masses also scale as H_o^{-1}, and since luminosities (in any wavelength band) scale as H_o^{-2}, (optical or total) mass-to-light ratios scale as H_o. The X-ray luminosity is the integral of Equation (13.14) over the emitting volume, from which it follows that the gas density scales as $H_o^{1/2}$, the gas mass as $H_o^{-5/2}$, and the baryon fraction (when X-ray emitting gas is the dominant form of baryonic matter) as $H_o^{-3/2}$. When comparing results from different investigators it is necessary to confirm that the quantities of interest are normalized to the same Hubble constant.

13.9 COSMOLOGICAL IMPLICATIONS

I would like to conclude with a brief discussion of how the results summarized at the end of the previous section fit in (or don't fit in!) with the standard view of the origin and early evolution of the universe. We can divide this view into two parts: cosmology – the origin, geometry, and composition of the universe as a whole, and cosmogony – the origin and evolution of structure in the universe.

13.9.1 Dark Matter and Cosmology

The standard Friedmann-Robertson-Walker cosmology assumes only that the universe is homogeneous and isotropic, that Einstein's general theory of relativity is correct, and that energy is locally conserved (e.g., Peebles, 1993). Model universes constructed under these assumptions expand from an initial singularity (the Big Bang), the evolution of the expansion depending on the total average mass-energy density in the universe. If the average density exceeds the critical density,

$$\rho_{crit} = \frac{3H_o^2}{8\pi G} = 4.7\ 10^{-30} h_{50}^2\ g\ \ cm^{-3} \qquad (13.29)$$

the universe will eventually turn around and recollapse; if the average density is less than ρ_{crit} it will expand forever. The Hubble constant has been expressed in units of 50 $km\ s^{-1}\ Mpc^{-1}$ (h_{50}). The symbol Ω is used to express density in units of ρ_{crit}. An inflationary model of the universe with $\Omega = 1$ exactly is now theoretically favored since it resolves a number of cosmological paradoxes (e.g., the 'flatness problem') and is a natural consequence of some of the more popular grand unified theories in particle physics.

One of the pillars of the Big Bang model (along with the expansion of the universe and the microwave background radiation) is its ability to predict the abundances of light elements (^{2}H, ^{3}He, ^{4}He, ^{7}Li) that are mostly or exclusively synthesized in the primeval fireball during the first three minutes after the universe began. Since the elements are created by reactions among neutrons, protons, and light nuclei their abundances depend on the total baryon density of the universe and are correctly reproduced only if

$$0.03 < \Omega_{baryon,BBN} < 0.1 \qquad (13.30)$$

Standard inflationary big-bang cosmology then predicts that 90-97% of the universe is nonbaryonic, and that clusters of galaxies should have baryon fractions (f_{baryon}) of 3-10% instead of 10-30% as is observed (section 13.7, Table 13.2). Assuming the X-ray results are correct, one (or more) of the following must be true:

1. Clusters of galaxies (or at least those regions we have observed) are not representative of the universe as a whole.
2. The theory of Big Bang nucleosynthesis is incorrect and $\Omega_{baryon,BBN} \sim 0.2$.
3. The inflationary model with $\Omega = 1$ is incorrect, and

$$\Omega = \frac{\Omega_{baryon,BBN}}{f_{baryon}} = 0.15 - 0.5 \qquad (13.31)$$

Since there is no theory that explains why clusters of galaxies should have more baryons than average and we have no indication that f_{baryon} is decreasing with radius in clusters, alternative (1) seems unlikely. Therefore it seems that the standard cosmological model is inconsistent with the observed universe. This paradox is sometimes referred to as the 'baryon catastrophe'.

13.9.2 Dark Matter and Cosmogony

The standard model of the formation of structure in the universe is the hierarchical clustering model where galaxies form as gas cools and falls into dark matter halos, and clusters form by the accretion and merging of galaxies and groups of galaxies. The evolution is controlled by gravitational interactions. Numerical simulations predict dark matter halos of the form given by Equation (13.20). The X-ray data are indeed consistent with this functional form for the dark matter with scale lengths, a_{dark}, consistent with those predicted by the simulations. However there is a fundamental discrepancy. Clusters of galaxies in the simulations reflect the average baryon fraction in the universe with little dispersion; however, as we have seen, real clusters show variations of factors of 2-5. This might imply that some process in addition to gravity has a hand in determining the formation and state of structures on scales of galaxies and larger.

13.10 SUMMARY

In this contribution I have presented an overview of the investigation of dark matter in the universe using X-ray observations, data analysis, and modeling. The presentation has been, to a large extent, pedagogical with explanations of the fundamental dimensions of the dark matter problem, the basic equations needed to address the problem, and the bases for the techniques of solution. I have also included some of the latest results on the dark matter content in elliptical galaxies, groups of galaxies, and rich clusters of galaxies. Finally, these new results have raised great challenges to standard cosmological and cosmogenic scenarios, and I have tried to offer a glimpse of some of the outstanding issues and puzzles that are directing current research and motivating new lines of inquiry regarding the most fundamental of astrophysical questions.

13.11 REFERENCES

Binney, J., and Tremaine, S., 1987, in "Galactic Dynamics" (Princeton: Princeton Univ. Press).

Cowie, L.L., Henriksen, M.J., and Mushotzky, R.F., 1987, ApJ, 317, 593.

Fabbiano, G., 1989, Ann. Rev. As. Ap., 27, 87.

King, I.R., 1962, AJ, 67, 471.

Mathews, W.G., 1978, ApJ, 219, 413.

Mulchaey, J.S., Davis, D.S., Mushotzky, R.F., and Burstein, D., 1996, ApJ, 456, 80.

Mushotzky, R.F., Loewenstein, M., Awaki, H., Makishima, K., Matsushita, K., and Matsumoto, H., 1994, ApJ, 436, L79.

Oort, J., 1932, Bull. Astron. Inst. Neth, 6, 249.

Parker, B., 1989, in "Invisible Matter and the Fate of the Universe" (New York: Plenum Press).

Peebles, P.J.E., 1993, in "Principles of Physical Cosmology" (Princeton: Princeton Univ. Press).

Persic, M., and Salucci, P., 1990, MNRAS, 245, 577.

Rubin, V.C., Ford, W.K. Jr., and Thonnard, N., 1978, ApJ, 225, L107.

Sarazin, C.L., 1986, Rev. Mod. Phys., 58, 1.

Trimble, V., 1987, Ann. Rev. As. Ap., 25, 425.

Trimble, V., 1995, in "Dark Matter" eds. Holt, S.S., and Bennett, C.L. (New York: AIP), p. 57.

Zwicky, F., 1933, Helv. Phys. Acta, 6, 110.

Chapter 14

THE MYSTERY OF GAMMA RAY BURSTS: UNDECIPHERED RADIATIONS FROM UNKNOWN ORIGINS

THOMAS CLINE
Laboratory for High Energy Astrophyics
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

The history of an unsolved astrophysical puzzle is reviewed here. Through the 1970s to the present time, explanations of gamma-ray bursts have ranged from a minor variation in neutron-star X-ray burst phenomenology to evidence for an unusually energetic and new process in Nature. Some of the latest observations are interpreted as giving confirmation for the cosmological extreme, and other GRB data seem to present greater confusion. This topic has become more fascinating than ever.

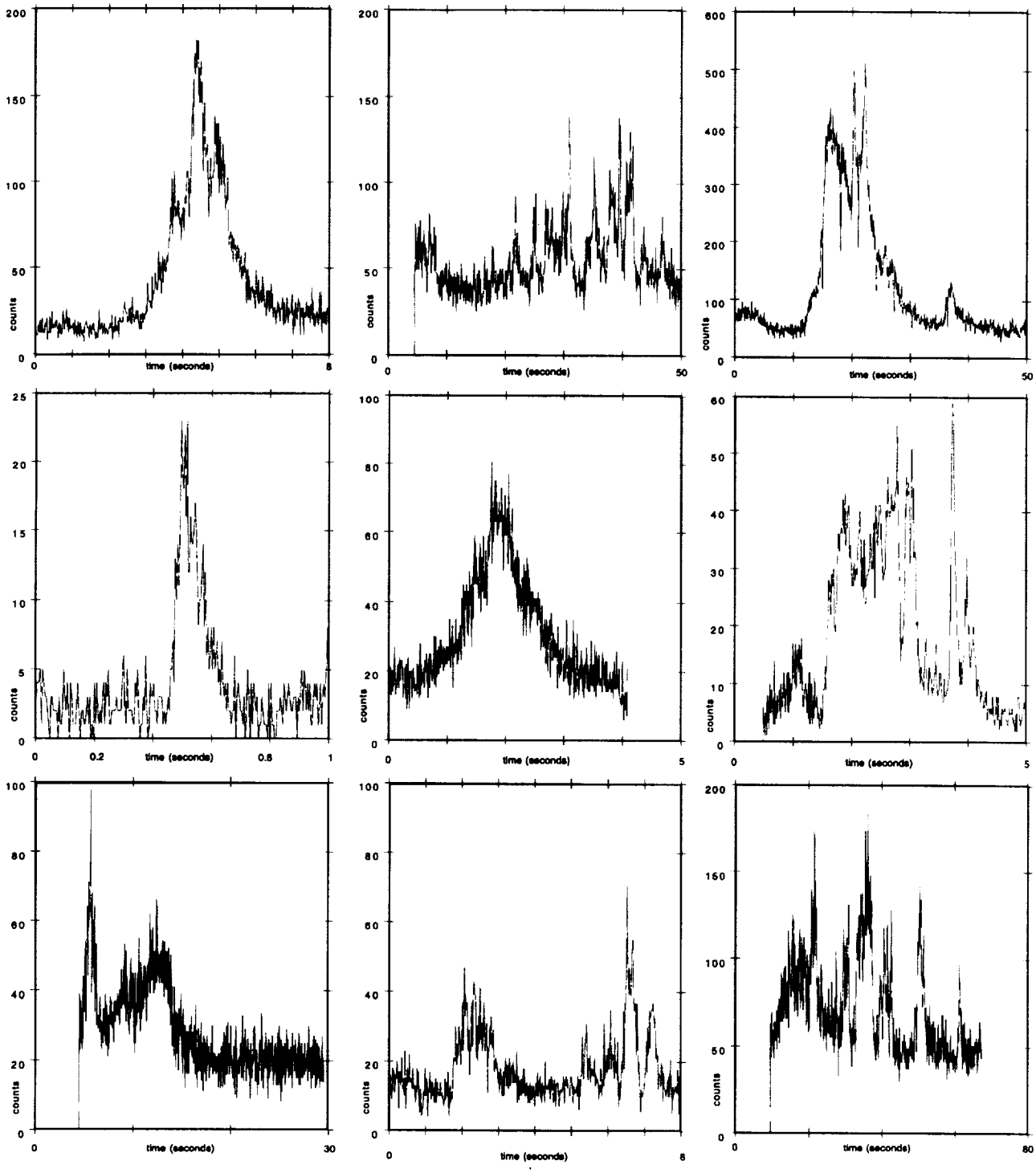


Figure 14.1: Typical gamma-ray bursts (GRB). From the beginning, GRB intensity vs. time plots have appeared to be random in all regards. Some appear to fluctuate to the measurement limits, whereas other do not. Other hints of nonrandom pattern can be inferred or, at least, imagined. (If a high-statistics GRB profile could be obtained at 0.1-ms or finer binning, one could make it into an audio frequency recording!)

14.1 INTRODUCTION

The gamma-ray burst (GRB) phenomenon, for the third time in twenty some years, has become a preoccupation of many astrophysicists and a favorite subject of science popularizers. The peculiar space-age drama of its discovery in 1973 - not the result of scientific research but an accidental byproduct of the cold war - inspired source ideas varying from concepts in 'new physics' to fantasies of space warfare. The early observations of GRBs could set few limitations, so modelling was free to range from the conservative and presumably sensible models involving neutron stars or stellar flares to the exotica of antimatter annihilation, the explosion of black holes into gamma-ray visibility, or the collapse of stars at cosmological distances. Thus, with source distances permitted all the way from the solar system to the edge of the universe, the resulting uncertainty in energy, going as the square of the distance, clearly set a world's record for scientific ignorance of an observed natural phenomenon.

Following the few years' time needed to develop some necessarily primitive burst detection instrumentation that could be adapted to the space vehicles then available, a rash of new clues renewed interest in this unsolved mystery. Meanwhile, X-ray bursts had been discovered and were quite amenable to rapid phenomenological definition and to galactic neutron-star source identifications. Despite the observational uncertainties and the awkward or contradictory data regarding GRBs (some of which appeared to be consistent with neutron-star sources), and despite the continuing lack of confirmed source identification, opinions seemed to force a consensus, almost by default, that nearby neutron stars must produce gamma-ray bursts as well.

After another period of languish from the lack of further new results (and from the common assumption that any loose ends to the neutron-star model would soon be tied up), the present and third period of excitement centers on the results from the new missions of the 1990s. The Compton Gamma-Ray Observatory, ASCA, and Ulysses measurements force the conclusions that the sources of the primary burst population still cannot be identified and that the GRB phenomenology is murkier than ever. Rather than solving the gamma-ray burst mystery with the galactic disk source pattern required of the interstellar neutron-star paradigm, the results from the very sensitive burst experiment on ComptonGRO show a continued and amazingly precise fit to isotropy. However, the CGRO results also exhibit too many bright (or too few weak) events to be compatible with an extended and uniform source region, given its calibrated detection capabilities. This result implies a source vs. distance relationship such as a limited source volume or even a source region cutoff. This combination of this result with isotropy would indicate a source region of unknown shape but necessarily spherically centered around us, clearly incompatible with any fit to the galactic disk.

Dramatic results from these various missions also confirm that a subclass of transients comes from distant supernova remnants, i.e., at considerably greater distances than the typical X-ray burst sources. These repeaters, intermediate in energy above the X-ray burst and below the usual GRB regimes, are rare, with only three having been found in our galaxy and in its two satellite (Magellanic Cloud) galaxies, combined. They were, in fact, responsible for some of the neutron-star clues formerly attributed to all gamma-ray bursts. Whether or not there is an astrophysical kinship between these repeaters and the apparently nonrepeating, primary GRB population is, however, a completely open issue. Further, quite unexpected new gamma-ray burst properties keep turning up, more curious and difficult to model than ever. Thus, with the mystery ever deepening and with a source volume appropriate to a Ptolemaic celestial picture (rendering a cosmological model as likely a bet as any), it's deja vu, all over again.

The appeal of this subject attracts a considerable amount of scientific talent. Experimenters and theorists have devoted great efforts to the scrutiny of the observations, suggesting over 100 different source ideas. New experiments have been developed, not only to study GRBs directly, but to search for associated effects in the X-ray, radio and optical domains as well. This considerable industry goes with the claim so often printed (arrogant, naive, or biased as it may be) that gamma-ray bursts present one of the major unresolved problems in astrophysics. Despite the number of published GRB reviews, one more, from the perspective of someone involved nearly from the start, may yet be justified here: tailored to this audience. Given the time delays inherent in the promotion, development and execution of new research, it is quite possible that the definitive GRB clue could be found by someone now in the student generation.

14.2 DISCOVERY

The first chapter in any GRB review begins with the discovery, published after more than a dozen of the inexplicable increases of the extraterrestrial $>100\text{-}keV$ photon count rate had been detected over several years' time. This observation, made outside the Earth's atmosphere and magnetosphere, was not accomplished with NASA or foreign scientific instrumentation, but was an accidental finding with the 'Vela' spacecraft system, designed at Los Alamos for the detection of nuclear explosions beyond the atmosphere, to monitor adherence to the 1963 test-ban treaty. The publication was delayed until after

several years of the consistent accumulation of data had provided confidence in the effect, with all attempts to explain it away exhausted. The events occurred rarely and at random, and really did seem random in other regards as well, with irregular intensity vs. time profiles differing in shape and in duration. Although the electronics employed counted the photons in geometrically lengthening time bins (no doubt, to fit the gamma-ray flashes expected from nuclear devices), specific timing marks or fiducials could be established at each spacecraft, with uncertainties short compared with the roughly 1-second light travel times that separated the various Vela spacecraft from each other in their orbits, about halfway out to the Moon. It was thus possible to measure the arrival direction of each burst as, one presumes, must have been intended in the case of nuclear weapons tests. The source pattern found in this manner was adequate to demonstrate no relationship to the locations of the Earth, Moon, or Sun.

There was also no apparent ecliptic or galactic plane correlation. Further, no evidence for event repetition from a single source was found, indicating that numerous source objects must independently exist. The maximum GRB counting rates were similar to those seen from solar flares, indicating many orders of magnitude greater total emission energy, to say the least, assuming that the events originated somewhere well outside the solar system. With no relationship found between the events and known or candidate supernovae (the only astronomical sources at that time having been suggested as capable of some sort of gamma-ray transients), and since this discovery predated that of X-ray bursts, this novel experience took the high-energy astrophysics community completely by surprise.

14.3 CONFIRMATION

The year before the discovery publication, we had begun a low-energy gamma-ray astronomy effort at Goddard, with a search for the gamma-ray transients then predicted from distant supernovae. We were building a balloon experiment for this purpose and also had been plowing through spacecraft data records to look for evidence of anomalous rate increases, requesting simultaneous data from other missions in order to eliminate spurious effects. Only weeks before the discovery publication, several large increases were found in the background rates from our IMP-6 solar flare X-ray detector, too ephemeral to have been noticed immediately. If the Velas had never been flown, these data and other observations that were soon to come could have established the existence of a curious new effect, although without the Vela directional picture, at first. Our events all turned out to be on the Vela list, giving immediate confirmation to the GRB phenomenon, and also providing the first, primitive burst energy spectra. These were consistent with maximum strength in the several-hundred keV region rather than with, e.g., a high-energy extension of a typical X-ray spectrum. Thus, 'gamma-ray' had been an appropriate term, although, of course, processes other than nuclear transitions (the original definition of gamma-rays) could be involved.

The GRB source process still remains, in fact, undefined.

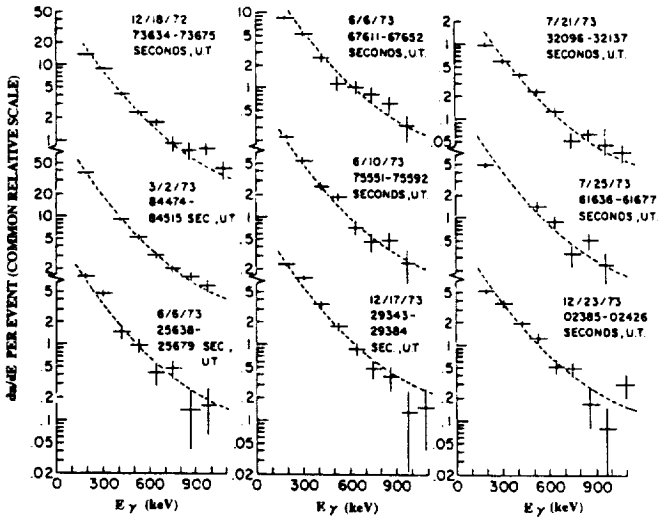


Figure 14.2: Early gamma-ray burst total-event spectra in the 0.1 to 1.0 MeV range. These appeared to form a population separate from, e.g., a high-energy extension of X-ray bursts. However, despite a roughly uniform appearance of the time-averaged spectra in this energy band, each event spectrally varies with time, and in both the lower and the higher energies, the average event spectra vary rather dramatically.

The potential of this new specialty injected real excitement into gamma-ray astronomy. Despite an earlier prediction in 1958 that astrophysical information might be transmitted by nuclear gamma-rays in the few-MeV band and by positron-electron annihilation radiation, the new field of X-ray astronomy, instead, had matured in the intervening 15 years with more generous rewards. Also, higher energy (>100-MeV) gamma-rays had been found from the galactic plane, but more as a neutral component of the cosmic radiation than as signals from individual astronomical objects. The development of "gamma-ray" (roughly 25-keV to 25-MeV) astronomy had seemed to require such elaborate instrumentation as to be quite difficult. Yet, here in GRBs were huge counting rates from unsophisticated detectors that didn't even have collimation or shielding to enhance the signal- to-noise ratios! And, regardless of their origins, enormous energies could be inferred, guaranteeing some

new astrophysical relevance. Further, the curiosity value of this discovery was enhanced by the surprisingly brief, several-second, durations of these events. The fluctuation times were fractional-second and were possibly more rapid than the detectors could sense, certainly shorter than almost anything then known in traditional astronomy. The thrill of all that, while prompting many imaginative ideas and speculations, left researchers unable to immediately respond with productive new experiments: space experiments generally took more than a few years to get through the mission schedules. And, with a rate of several big bursts per year, a rocket or even a balloon flight had a poor chance of being aloft during one such event. Weaker events should presumably be more numerous, but it was not verified until recently that no arbitrarily sensitive balloon-borne instrument could accumulate more than several per week.

Only one of our half-dozen IMP events turned up to be in the data records from other spacecraft: a bright increase out of a list of hundreds of count-rate fluctuations logged by OSO-7 in near-Earth orbit. Its energy spectrum, agreeing with the IMP spectrum but extending it at both ends, providing the second confirmation of the burst effect. However, as the only GRB detection thus far using a collimator, it was of even greater interest. The source definition of tens of square degrees was no improvement in resolution, but its location verified their method with agreement. Also, at high galactic-latitude, the source direction provided no new key to the mystery but complicated it with a location far from the region of typical X-ray emitters along the disk, enhancing the feeling that a truly new scientific phenomenon was ripe for exploration.

After a year or two, the total number of GRBs found with the IMPs had exceeded the initial Vela total by several times. The timing resolution of the solar flare detectors used was too coarse to record the spectral evolutions of the bursts, but the total-event GRB spectra were found to all cluster around the same curve in the 0.1-MeV to 1.0-MeV interval. This result could tell nothing of the details within a burst, of course, but it did confirm that the GRB population formed a class qualitatively distinct from other known transients.

14.4 CHALLENGE

The second chapter in this history concerns the early generation of experiments and projects that were undertaken to research the GRB topic. Our supernova search was, of course, continued in order to look for smaller-sized GRBs that might occur more frequently. A balloon flight in 1974 found no new events, and even the successful 1975 attempt to use two gondolas (simultaneously flown several midwestern states apart in order to better reduce background count-rate fluctuations) also came up with no new bursts. The balloon-borne measurements of many groups suggested a deficiency in smaller events but were statistically inadequate to quantify this, vaguely previewing the ComptonGRO result of two decades later that the all-sky rate of bursts is barely more than two per day.

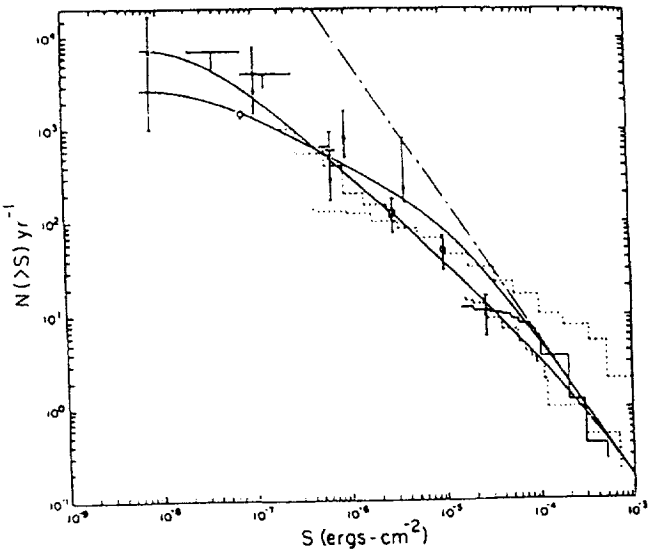


Figure 14.3: Early gamma-ray burst size spectra. GRB detection rates, then obtained circumstantially, were statistically weak and often difficult to quantify in a fully objective manner. The apparent change of slope towards the smaller sizes was ambiguously interpreted either as instrumental – promising a higher event detection rate when superior detectors could be used – or as a genuine absence of weaker events.

Meanwhile, we also promoted the opportunities to modify space hardware then under construction and to use any capability available on missions in progress, for burst detection. NASA responded positively to our plea, permitting some piggyback instruments and certain modifications, also enabling, e.g., the Los Alamos group to place a GRB monitoring experiment on a mission to Venus. These primitive devices ultimately formed an interplanetary network, designed to make possible precise GRB source localization by the timing of each burst arrival at the widely separated locations. European researchers also enthusiastically started building new GRB instrumentation, with French detectors getting rides on several Soviet satellites and two Venus probes. By the late 1970s, four deep space probes had augmented the near-Earth complement of over a half dozen more, creating the first interplanetary gamma-ray burst network. This was accomplished one and a half decades before the launch of ComptonGRO, the first mission to carry a full-scale, major experiment designed for the study of gamma-ray bursts.

14.5 NETWORK

The first instrument to be intentionally put into space for GRB studies was our initial contribution to the interplanetary network, a 1-kilogram package piggybacked on a previously designed cosmic-ray experiment on the Helios-2, then under construction. Later, we were able to add more advanced burst capabilities on another, relatively nearby mission, as well. Helios-2, launched in early 1976 into a solar orbit that carried it up to twice the Sun's distance from us, made possible source precision superior to the previously obtained several-degree variety – but with a catch. For the first two and a half years when no other burst detector existed at great distance from the Earth, timing comparisons could employ only this single baseline to near-Earth missions, such as the surviving Vela spacecraft. The resulting analyses gave celestial source fields in the shape of ring segments, as thin as one arc minute but from several to several tens of degrees in length. Although these source loci were too large and too irregular in shape to be useful to astronomers or to identify single stars or galaxies from catalogs, the location of any object of interest in the high-energy astrophysics category could be meaningfully compared for agreement or rejection. It resulted that the Helios-2 annuli did not include any of the known objects such as pulsars, quasars or cosmic X-ray sources that could have been candidates for identification. This result, of course, showed that the source objects were 'dim' over many wavelengths, and heightened the GRB mystery.

The interplanetary network was fully completed in 1978 when Helios-2 was given company in deep space with the burst monitor from Los Alamos on Pioneer-Venus Orbiter and with two Franco-Soviet missions sent to and beyond Venus. The near-Earth collection included those Velas that still operated, Franco-Soviet Earth orbiters, and ISEE-3, with several of our detectors at five light-seconds distance in an orbit maintained at the 'first Lagrangian point.' Circumstantial GRB detection could sometimes also occur with other near-Earth missions as well, including even the Einstein X-ray observatory. For about the first year, PVO and the two Veneras were all on the way to or in the vicinity of Venus. Helios-2 thus optimized the long-baseline array with a separately directed baseline. After the Veneras cruised past Venus into solar orbits, all five elements of this array became mutually well separated for a year or so, until the components began failing in the early 1980s. This network finally made possible the first relatively precise GRB source field locations. Naturally, the critical element or potential Achilles' heel in this novel arangement was the interspacecraft timing: great efforts were made to expose inappropriate assumptions and to eliminate errors of all kinds. For all the near-Earth spacecraft, at least, the dispersion was found to be less than the burst fluctuation

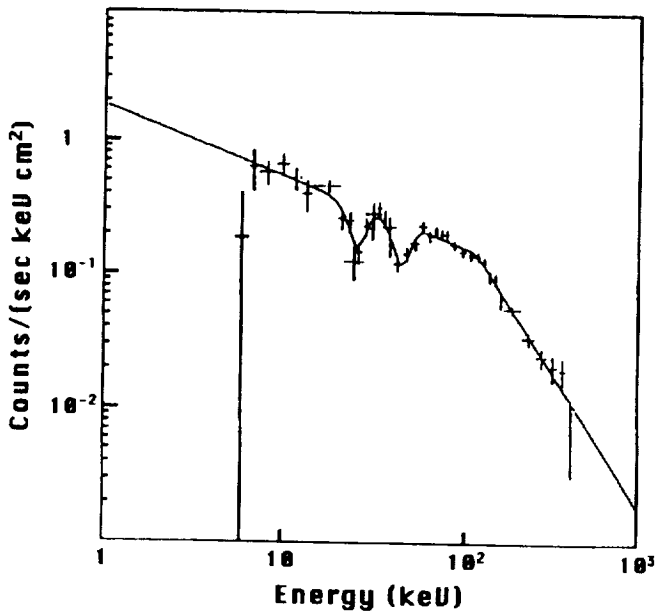


Figure 14.4: A GRB energy spectrum with multiple variations. This was taken as confirmation that simple variations in GRB spectra could be identified as cyclotron resonance features, supporting the high magnetic field neutron star burst origin model. However, no features, single or plural, have yet been observed with two instruments simultaneously.

times. In one event with a rapid onset, it was shown to be at the 2 *ms* instrumental limit. For the distant space probes, the only way to verify the absolute timing was to send artificial burst commands, since there were no naturally occurring calibrations with all the network detectors too insensitive to known X-ray sources. (A new kind of emitter the "bursting pulsar", could provide confirmation of the time calibration for the Ulysses-based interplanetary network in 1996, however.)

The network scheme determined source regions with greatest accuracy for those events having the highest intensities and the shortest temporal fluctuations. These fields or "error boxes" varied in size down to less than one square arc minute, small enough that narrow-field optical scrutiny could be carried out, particularly at high galactic latitudes where the stellar background is minimal. With the exception of the March 5, 1979 event, discussed below, the regions found contained no known X-ray emitters, neutron stars, pulsars, and other exotic objects. Also, the source areas were generally empty of stars, or they contained foreground stars within random expectations. Further deep optical searches of these source fields to even fainter limits, such as with Kitt Peak Observatory, gave no further clues. Thus, GRB sources (and their possible companion objects as well) had eluded all methods of identification available.

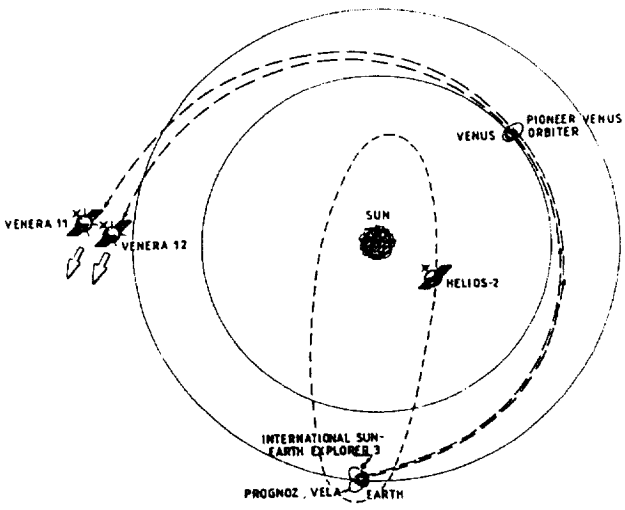


Figure 14.5: The first interplanetary network. This artist's schematic illustrates the placement in the 1970s of four GRB detectors into deep space, along with other missions near the Earth. The mutual separations of up to 1000 light-seconds made possible the timing that produced about ten arc-minute-sized GRB source fields, and several tens with less precision, during the full-complement lifetime of under two years.

14.6 PHENOMENOLOGY

14.6.1 Number

Progress in defining the phenomenology of GRBs proceeded only slowly at first. Although several hundred events had been accumulated by all the instruments in operation over the decades of the 1970s and 80s, only those bursts detected with any one sensor in a given configuration could provide an internally consistent sample for study. Each all-sky survey consisted of a limited sample of events for various circumstantial reasons, and each was consistent with an isotropic source pattern for its population. However, due to the varying instrumental threshold levels in energy or in counting rate, these various samples could not be reliably mixed together for quantitative comparisons. Of particular interest was the burst size distribution: the number of events observed to have a given size or greater when plotted as a function of intensity or some other measure of 'size' (the so-called N > S vs. S plots). This function should, in principle, provide source region information. In particular, a three-dimensional volume of sources uniformly distributed throughout a region of space of infinite extent or extent beyond visibility limits would exhibit a power law of index -1.5. An index nearer -1.0 would result for a population of events coming from a two- dimensional source volume such as the galactic disk (and, of course, in that case an anisotropy should be observed). An obvious complication to any quantitative deduction from this is the lack of knowledge regarding the possible source luminosity function – the distribution of sizes and types of GRBs to be seen at a calibrated distance. Also, possible variations with distance, such as a cosmological red shift, or an ageing factor could be relevant. With this ignorance, the luminosity function could lie anywhere between one imaginary extreme, with all bursts having similar maximum brightnesses but situated at a great range of distances from us, to another, with most sources clustered in similar apogees but having greater intrinsic size variations. However, given that the distribution of brightnesses could range well beyond the visible, and that the temporal fluctuations could exceed the limits of detector resolution – and with the event diversity and the chaotic appearance of most bursts – the assignment of a meaningful 'size' to any event could well be a subjective act.

Nevertheless, the GRB size spectrum was researched over the years, always with a great deal of uncertainty but generally with finding a persistent flattening, or deficiency in small events. This could have been interpreted as evidence for a population transition to the galactic disk, except that the expected disk anisotropy was never seen. This fact provided the basic problem that had been the subject of great concern, with compromises in those years ranging from the suggestion of a galactic halo source region, with a rough isotropy consistent with the limited measurements, to the redefinitions of 'size' (using, for example, peak intensity rather than total intensity). The default view that burst sources are surely close by could neither be ruled out nor proven.

The issue of the true number of burst sources is another question of obvious relevance. Do sources repeat and so is the vast number of events detected over the years due to the poor directional resolution of a considerably smaller number of objects that emit repeatedly? Or, could the emission be so tightly beamed that the number of sources must be much greater than can be seen here? The fact that the well-determined error boxes never overlap could be due simply to their scarcity in number, if the repetition time constant is long enough. In fact, there has always been relatively poor evidence that the primary GRB sources do not repeat (except for the subclass that, by definition, does – see below).

14.6.2 Spectra

The temporal and spectral properties of GRBs have been studied for any possible clues regarding the source process. The brevity of the temporal fluctuations requires that the source regions be small. Estimates of the photon densities in the emission volume, along with self-absorption considerations, seemed at first to support the claim that the distances to the sources must be less than

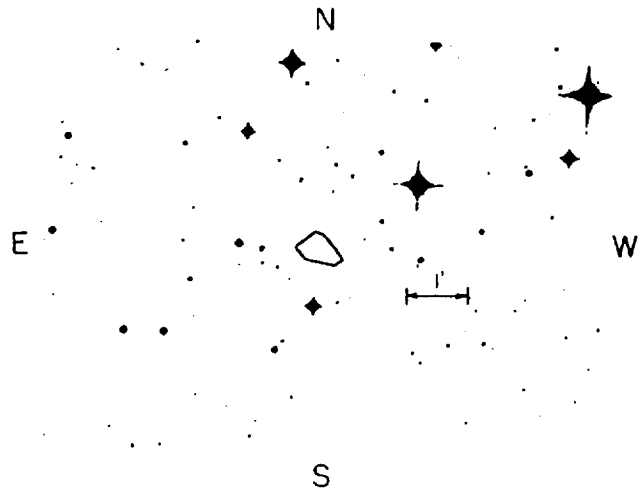


Figure 14.6: One of several high-precision GRB source fields. With the one exception described below, no candidate source objects existed in the catalogs of known astronomical high-energy sources or could be found with optical surveys of moderate power. Deep optical scrutiny, on the other hand, reveals too many faint objects in all fields of this size.

several *kpcs*. This interpretation is no longer popular. Studies of the continuum energy spectra have remained unable to lock onto specific emission properties, since determination of the input photon spectrum from the observed counting rate spectrum (due to a curious 'obliging' nature of the fitting procedure) is compatible with almost any model. However, certain spectral features were claimed to be found in the 40-60 *keV* region and also at about 400 *keV*. These detections had specific implications. The more energetic feature is consistent with the 20-percent gravitational redshift expected of the 511-*keV* positron- electron annihilation line at the surface of a neutron star, and could also be fit to an annihilation gamma-ray coherence effect. The low-energy features were claimed to be cyclotron resonance phenomena, providing possible evidence for burst emission within the strong magnetic field of a neutron star.

A little later, a Ginga spacecraft X-ray monitor gave apparent confirmation to the 'cyclotron' model, with multiple features that looked even more like what one would hope to see. However, up to the present, no other confirmation has been obtained. In particular, the ComptonGRO Burst and Transient Source Experiment data bank shows no convincing spectral features in over one thousand burst events. Earlier, those questioning the interpretation of the features as 'cyclotron' and hence neutron-star effects suggested the possibility of a false instrumental summation of time-varying continua with differing low-energy cutoffs. Now, after two more decades: since data showing these features have in fact never been found with two different instruments for the same event, it is becoming reasonable to suspect that the features were count-rate fluctuations, or were otherwise misleading, all along. In those days, however, their appearances convinced many theorists that GRBs most surely come from nearby neutron stars.

14.7 OPTICAL ASSOCIATIONS

One totally unexpected discovery in the early 1980s was that of an optical transient coincident with a precisely determined GRB source field from the interplanetary network. It was anything but coincident in time, however, having been found in a diligent search through hundreds of old stellar photographic plates,.stored at Harvard. Taken in the year 1928 (!), it was in one of a series of several exposures photographed during the same night, fortuitously providing 'before' and 'after' comparisons. Both the character of the image in the emulsion and the fact of its rounded shape (vs. the elliptical star images made by motion of the camera) were consistent with a brief duration relative to the 45-minute exposure. Although disputed by some as a possible fluke, this discovery was received by most with great enthusiasm, since the possibility of arc-second precision, accessible with optical tools, might finally make possible the identification of GRB source objects. Of course, the 50-year separation between the optical transient and the burst might defeat any present-day optical search due to source proper motion, if the source was close enough. Soon after, hope was fueled when additional candidates turned up in other searches. However, some of these seemed suspicious, and one was found to be only in close proximity to a GRB source, suggesting that any relationship between GRBs and these optical flashes may have been entirely accidental. Deep searches of these several-arc-seconds sized regions have so far been disappointingly marginal and inconclusive.

It was realized that optical transient observations obtained in real time, rather than recovered from decades-old archives, could not only make possible the localization of GRB sources using the precision of optical astronomy, but also do that as they occur (assuming that GRBs actually are accompanied, at the same time, by visible optical flashes). Both ground-level and spacecraft experiments have been attempted, and a variety are now in various phases of development, to exploit this hoped-for effect. The first sky survey used a wide-angle CCD array to sense and record optical transient effects with moderately good precision. It was also used to orient a mirror in real time for a second, narrower field telescope to localize the transient to even better precision. Set up in the desert at Kitt Peak Observatory,

this promising combination operated at low duty cycle for several years, but was never successful, due to the unexpectedly high rate of man-made and other optical transient backgrounds. A reorienting device, called the 'Sunflower' for obvious reasons, was also orbited by the Soviets recently, but failed to operate successfully. A new gamma-ray/X-ray/optical-transient mission will be flown soon, perhaps by early 1997. It should both establish meaningful upper limits to the rate of optical flashes accompanying GRBs, and provide relatively accurate GRB source fields, with moderate delay, for deeper scrutiny with optical telescopes, even if no flashes are seen.

14.8 THE MARCH 5, 1979 EVENT

One burst observed with the interplanetary network was an immediate curiosity. It seemed to be so unusual that this author was convinced that it was a 'once in a lifetime' experience. All the instruments then in operation logged it as the most intense to date (of course, one of them had to be) – but its shape was unique. The single brief spike with a long tail presented a meaningful form, in contrast to all the other noisy-looking events. The risetime caught our attention: in the first and only event to activate the fastest (4-ms) Helios-2 trigger mode, the onset was certainly shorter than 2 ms and consistent with instantaneity. Also, an 8-second periodicity of the decay (first noticed in the Soviet data) was the only evidence for periodicity in any gamma-ray transient. This effect was consistent with what was expected from a rotating neutron star.

The interplanetary network found the source location to be the first to actually agree with the celestial position of a 'candidate' astronomical object, with an initial two-arc-minute fit to N49, an X-ray-emitting supernova remnant at a distance of 55 *kpc* (about 5 times as far from us as the galactic center), and within the Large Magellanic Cloud, a galaxy that is satellite to our own. Our final analysis gave a much narrower source field only arc-seconds from (but not quite fitting onto) the center of N49. This was, and still is, the most precise directional measurement in gamma-ray astronomy. Although the probability of chance fit to any of the known supernova remnants, pulsars, and neutron stars was under one in a million, and despite the fact that any such SNR could contain an 8-second rotating neutron star, N49 was not generally accepted as a source identification, and would have to wait for well over a decade for confirmation.

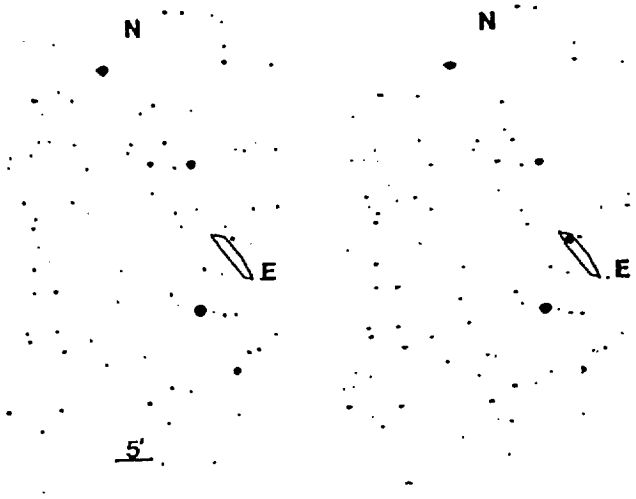


Figure 14.7: Evidence for an optical flash from a GRB source. The fit of this 1928 optical transient (found in a search through sky photos in astronomical archives) onto a 1978 GRB source field set into motion an industry of researching at other wavelengths possibly associated with GRBs.

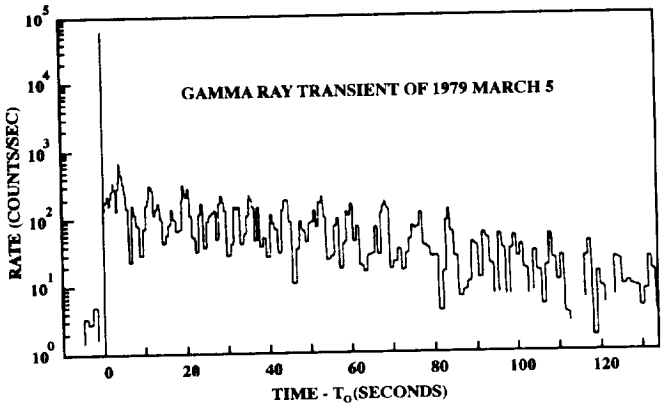


Figure 14.8: The March 5, 1979 event time profile. This transient rose in less than a few milliseconds to an intensity four orders of magnitude over the all-sky background, in its >0.1 MeV region. Its single peak was followed by an eight-second oscillation for many minutes. Other anomalous features include the facts that it appeared to sire a series of less intense events for months afterwards and that, for the first time, its source direction fitted that of a candidate source object.

Another special property of the March 5th event was the series of several much weaker transients that followed it over a few months. Their source directions were both mutually consistent and consistent with the March 5th source, to about one degree acuracy. Possible only with Soviet detectors due to their greater sensitivity from the greater size, this result complicated the GRB picture. Another series of small events had been also found with the Soviet detectors in early 1979, but without an obvious parent burst. It was then questioned whether repetitive events might be a common occurrence,

thus diluting the singularity of the March 5th event, and whether such phenomena might even be typical, but too weak to be observed. It was years before the detection of only one additional series (and of later flare-ups of the earlier series) had established that the so-called "soft gamma-repeaters" were quite rare: three series were totalled over a monitoring period that included one and a half thousand GRB events.

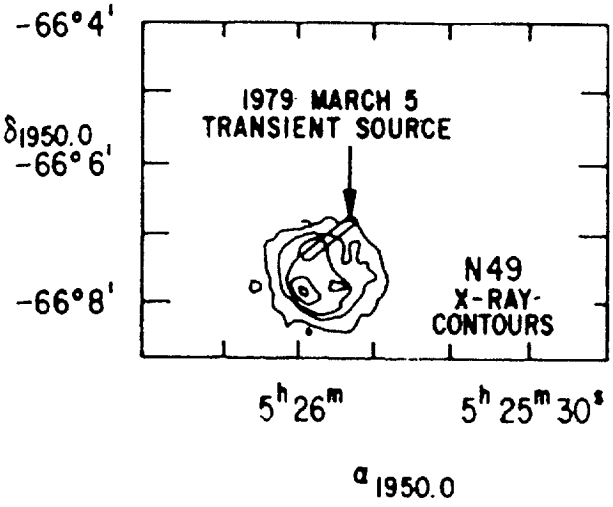


Figure 14.9: The precise March 5, 1979 event source location. The association with the supernova remnant N49 in the neighboring LMC galaxy had to wait for a dozen years for confirmation, when an SGR was found to be produced from a similarly distant supernova remnant. This new result would argue for some snr/neutron-star source model for all GRB and SGR phenomena, if the GRB source isotropy problem were overcome, or it would revive the contention that SGRs and GRBs may be separate phenomena.

The N49 measurement produced considerable controversy as a candidate source identification for the March 5th event, with its distance, with all other GRB source locations fitting onto blank space, and with general source isotropy. The question of 'How could the brightest event come from the farthest possible source distance?' was based both on the premise that there was only one category of gamma-ray bursts (economy of assumptions) and on the premise that GRBs must originate nearby (economy of energetics). Our speculation that this event was anomalous, if not unique, was entirely consistent with a possible origin at N49 in the LMC, since this might little bearing on the source pattern of most GRBs. However, one discovery seemed unlikely enough, let alone the probability that the study of one new phenomenon (GRBs) would prompt the discovery of a second one (SGRs). The 1996 picture, however, turns the energy issue around: GRBs can now be respectably discussed as possibly coming from sources orders of magnitude more distant than the LMC. Also, the fit of SGR sources at supernova remnants has been confirmed, but how the intense March 5, 1979 event relates to its sequel SGRs, and how - or if - SGRs relate to typical GRBs, are still mysteries.

The apparently singular properties of the March 5 event presented a complex turn of events. For example, the fact that this was a very bright burst with a periodicity that existed only in a very weak tail might be taken as evidence that such a low-level ringing simply could not be discerned in other GRBs, or that the March 5th event could be the only big burst from a rotating neutron star. If it was a separate kind of event, then inferences regarding its nature and origin would be inappropriate to the larger population of bursts that did not share its properties. Also, the weak SGR events contrasted the other GRBs not only by the fact of their repetition, but by their softer energy spectrum, in the 25-keV region between the X-ray bursts at a few keV and GRBs at several hundred keV. Also, SGRs (and March 5) had no cyclotron resonance features or spectral evolution, then claimed for many typical GRBs. And, in additon, SGR time histories are always brief and simgle-peaked, unlike the chaotic 'white noise' varieties of GRBs. The SGR phenomenology, however, received little attention in those days, being regarded as a minor variation on the GRB issue.

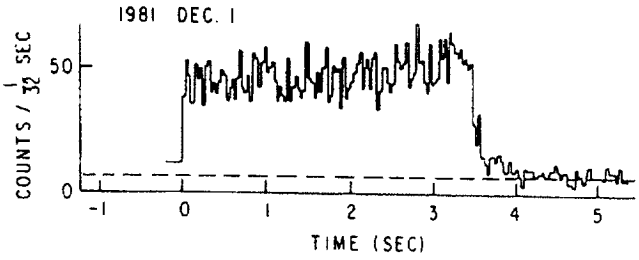


Figure 14.10: Time history of a long-duration SGR. Soft gamma repeaters differ from GRBs with their simple and uncomplicated time histories, their narrow luminosity range and their absence of spectral features (in addition to the softer spectra and fact of repetition, as labelled), but the randomicity of their repetition patterns and of their durations are not unlike the chaos typical of so many of the GRB characteristics.

One somewhat unexpected turn of events in the mid-1980s (before the launches of the missions that changed everything: Compton, Ulysses, and ASC'A) was the discovery of another soft gamma-repeater. This was the third time that weaker, softer events had been found with GRB instrumentation and was not that dramatic, in itself. However, what struck this writer was the fact that its source direction, a few degrees from the galactic center, taken together with those of the other two, sug-

gested a meaningful pattern. Naturally, three points are too few to demonstrate any mathematical arrangement, but these were all consistent with high-density regions in our galaxy and in the neighboring LMC. Also, the typical intensities of the three series were consistent with the same idea, since the March 5, 1979 sequel rates were weaker than the other two, roughly by the square of the ratio of the distance of N49 to that of the galactic center. Several years had to elapse before the fact of this pattern was fully confirmed.

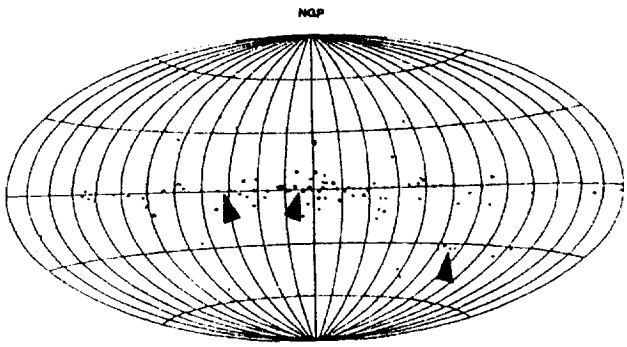


Figure 14.11: The late-1980s sky map of the three SGRs (arrows), with the X-ray binaries. This pattern was seen as three random points in the sky (if the N49 location was taken to be a meaningless coincidence) or it suggested (with statistically negligible evidence) that all SGRs might ultimately be found in the galactic disk and in the LMC (if the N49 association was taken as a real possibility). All three SGRs have since been located in the galactic disk and LMC, and there are only three.

14.9 REVOLUTION

The third chapter in the GRB story features the major leaps forward from the Compton Gamma-Ray Observatory (most, but not all, from its Burst and Transient Source Experiment) and from other experiments in association with it. The GRO was launched in April, 1991 into Earth orbit and remains fully operating to the present, in late 1996. The BATSE uses eight instruments facing, as it were, from the corners of a cube, so as to provide all-sky coverage with isotropic sensitivity. It was initially promoted as a burst monitor, to service the GRO gamma-ray astronomy experiments with real-time burst data (due to the fact that GRBs did not have much scientific priority in 1979 to the early 1980s, when GRO was designed. Only after a GRO high-resolution spectrometer was cancelled was it possible for BATSE to acquire full experiment status). The BATSE principal investigator credits us with inspiring his design and enabling his successful proposal, based largely on a large 'gamma-dome' monitor we had proposed for the early Space Shuttle, inheriting a certain amount of its style and content. However, our gamma-dome was designed to detect an anisotropy in a week, if the GRB size spectrum extends to the small events, which BATSE showed that it does not. We credit and congratulate the BATSE PI in turn with executing the strategy that accomplished the definition of the characteristics of the GRB population: the precise source pattern isotropy and the limitation on the size distribution.

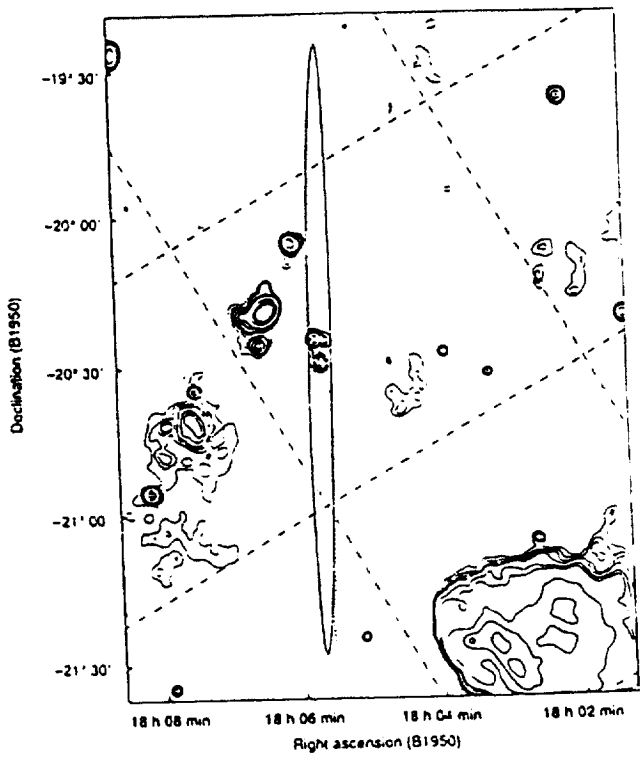


Figure 14.12: The network-determined third SGR source field, with a radio supernova remnant unknown when the SGR was found in the early 1980s. When viewed years later with an X-ray telescope, this SNR produced an outburst in real time, with the low-energy attenuation consistent with absorption from a source located on the other side of the galaxy. This result confirmed N49 as the March 5, 1979 event and earlier SGR source, thereby liberating GRBs from a necessary connection to that event, to a range of possibilities including it but conceivably wider than ever.

The GRB isotropy measurement of the CGRO BATSE is by far the most precisely determined yet. However, by itself, that could be seen as almost no surprise. The Leningrad team had found an isotropic distribution over a dozen years earlier with an event rate of about 75 bursts per year. Since the BATSE rate is about 600 per year, the effective source volume sampled (assuming

an unbounded, homogeneous, three-dimensional model) would be eight times as much, i.e., with only double the radius. Thus, if the GRB source distance limit was unknown before, to have learned that it is twice as much would not be a major advance.

However, the other BATSE result, the so-called 'V/Vmax' analysis, shows that this cannot be the case. This procedure effectively normalizes the measurement of burst sizes, by comparing the size of each burst to the minimum size an event of that particular character could have to be detected by the instrument in its given configuration. Thus, the ratio of measured to minimum count rate is translated to a ratio of source volume to maximum source volume, per event. The distribution of V/Vmax values from all the observed events would be flat with an average value of 0.5 if the detector is sampling only the nearby members of a large, unbounded space, filled smoothly with emitters. On the other hand, a transition to a disk-shaped source volume, or to some other limiting shape, would give other results. The BATSE V/Vmax, given the perfect isotropy, shows that the source volume must have some structure, such as a spherical edge or a nesting of concentric round shells. The implication is that burst sources may be confined to a round 'Oort cloud' around the solar system, or to an immense halo around the galaxy (large enough such that our displacement from the galactic center is unnoticed), or to a round interstellar cluster or a broken piece of the spiral arm accidentally centered on us, or to the distant universe itself. For the first time, a possible cosmological origin could finally be taken seriously.

Beyond the precise isotropy of source directions and the first statistically valid V/Vmax index, other observations of the Compton mission have done much to advance – and confound – the study of gamma-ray bursts. For example, regarding the issue of spectroscopy, BATSE failed to find any evidence either for the 400- keV (redshifted annihilation) bump or for any cyclotron resonance features above 50 keV, that had previously been cited as the basic evidence (from the primary population of GRBs, not from SGRs or the March 5th event) supporting a neutron-star origin model. The burst phenomenology coming out of the Compton mission covers a richly diverse range, with the biennial workshop in 1993 producing a 150-manuscript text from BATSE alone. And, other experiments on Compton have made startling GRB discoveries as well; e.g., that of very high-energy photons detected with the spark-chamber spectrometer (nearly 20 GeV!) coming from the direction of GRB events, but trailing the bursts by as much as 1.5 hours! The theorists who try to model emission concepts find such data outrageous, indeed.

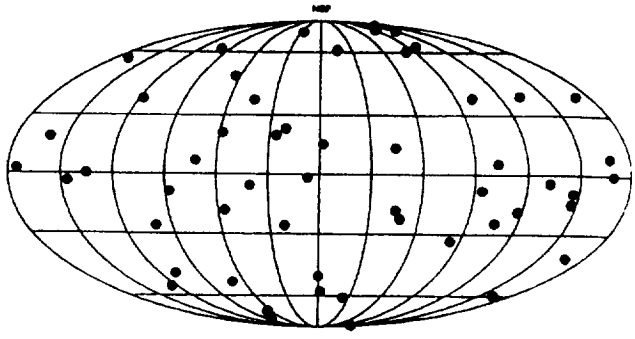


Figure 14.13: The sky map of GRB sources, as of the early 1980s. It was long assumed that the apparently isotropic celestial distribution would soon yield, with more sensitive monitoring, to a source pattern having an obvious relationship to the structure of our galaxy. Instead, the Compton GRO has found an incredibly isotropic pattern (next figure).

14.10 BREAKTHROUGH

An incredibly unlikely occurrence of significance to the GRB puzzle recently took place, somewhat as the grand finale of a drama combining profound research efforts with serendipity – just pure luck. A soft gamma-repeater was identified with a pointed, narrow-field X-ray telescope in real time. First, the fact that SGRs would be detected with the BATSE instrument was one surprise; also, the fact that an apparently new SGR would be rather quickly determined to be an obvious reactivation of one of the three (and only three) known SGRs from the 1970s and '80s was another surprise. Meanwhile however, radio astronomers, openminded as to the possible fit of the March 5th event to the SNR N49, and curious as to the origin of other SGRs, had deduced that a very likely source candidate for one of the SGRs found nearly a decade earlier with the interplanetary network (and localized to a thin crescent-shaped source field several degrees from the galactic center) could also fit a supernova remnant. This one, optically unseen through the galactic disk material, was not known in the radio band at the time of the detection of the SGR in the earlier 1980s. The successful promotion of this remnant as a possible source candidate resulted in the X-ray astronomers orienting their ASCA instrument towards that direction. They were rewarded with a single SGR transient being detected in real time. This was localized (with x- rays) exactly at the radio remnant, in the center of the extended, earlier source field. Further, the cut-off in the X-ray spectrum below 1 keV was an excellent fit to the absorption expected for a source at the distance estimated for the radio snr, about 18 kpc. This is across the galactic center, within the disk, but on the other side! The fact that this is one-third of the distance to the LMC

provided convincing evidence that the N49 fit could be finally a confirmed source identification and no longer taken as unlikely and ironic accident.

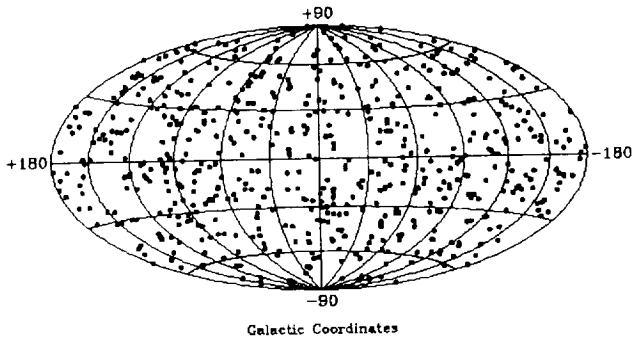


Figure 14.14: The sky map of gamma-ray burst source directions, from several years' monitoring with the Compton Gamma Ray Observatory. The centers of each location are shown without the uncertainty radii, which vary from about three degrees minimum size. (This sky survey must be normalized with the relative viewing pattern to quantify the isotropy.)

The Ulysses mission, launched in 1990, was the starship of a new interplanetary network of burst instruments. It was the survivor from spacecraft twins, the 'Out of the Eliptic' mission, designed to stereoscopically sample the Sun's north- and south-polar vicinities at up to 5-AU distances. Our proposal for a pair of burst detectors had been selected, enabling the European half to continue when the NASA half was cancelled. This new network, however, was compromised with a series of misfortunes. The Ulysses instrument, at a new record's distance of 5 AU, the 15-year-old Venus Orbiter, and BATSE in Earth orbit, completed an interplanetary array until PVO finally entered the Venusian atmosphere in late 1992. It so happened that Jupiter, towards which Ulysses was initially cruising (in order to anticipate the gravitational encounter, to boost it out of the ecliptic) was in approximately the same direction from the Earth as the planet Venus, so that many of the early Ulysses-PVO-GRO triangles collapsed, with thin but long source fields as a result. The next unfortunate circumstance was the demise of the Mars Observer (MO) as it reached the planet, in 1993. We had made sure that the MO gamma-ray spectrometer included a burst triggering and timing capability, but the next unfortunate circumstance was that Mars Observer was then replaced with a smaller Mars mission having no spectrometer, and thus no GRB trigger. Instead, however, our next host mission will the the Russian Mars-96, when it is to be flown later this year. Thus, the interplanetary network of the 1990s provided thus far only a similarly small sample of a dozen or so well determined GRB source fields. Study of these has not yet yielded any new source object clues, other than one possible ROSAT X-ray source association.

However, Ulysses observations were only recently used in a very clever way to determine the location of the third and last SGR source. This analysis was prompted by the notice from BATSE that another repeater series had appeared to start up again, after years of no new data, from the approximate direction of the one SGR that remained unlocalized, due to its initial coarse directional definition. The discovery, using only the Ulysses-Compton baseline in the absence of a third network vertex, incorporated the repetitions of the bursts to provide intersecting arcs to localize the common source direction. Since the events were too low in intensity to trigger Ulysses, the clever part was getting the Ulysses data out of the background by scanning with the timing geometry in reverse. A distant galactic supernova remnant was found as a candidate source, entirely consistent with the identifications of N49 and the ASCA snr. The soft gamma-repeater picture now seems self-consistent: the three and only repeating series of transients seen with all GRB instrumentation over 20 years' time, can be associated with supernova remnants at the distances ranging from within the disk, to beyond its center to the other side of the galaxy, and to the LMC.

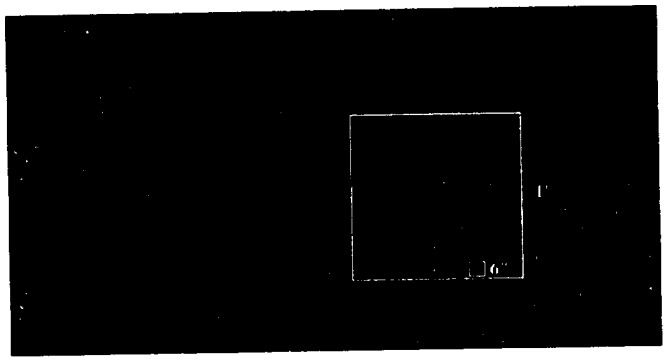


Figure 14.15: The celestial background of the typical precise GRB source field. Given that the distant galaxies necessarily populate a one-arc-minute field, it is clear that if a future instrument could determine GRB source locations to one-arc-second accuracy, only a statistical study of the coincidences or associations with objects could provide evidence for a claimed identification of the GRB source population.

14.11 CONCLUSION

The source identifications of SGRs and of the March 5, 1979 event may be history, but these issues are replaced by new questions: What are the source mechanisms for the weak SGR events? Why is the March 5 event so much more intense than its SGR sequels? The neutron-star

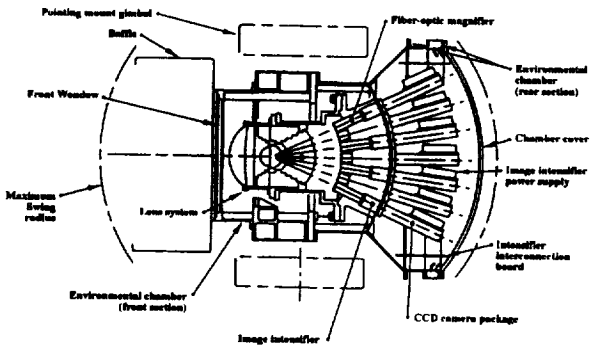


Figure 14.16: This wide angle optical telescope was originally developed for the Strategic Defense Initiative to pinpoint enemy rocket launches. It is one of a variety of instruments that are automatically notified with GRB source locations, in order to reorient and search for other emissions that may accompany the bursts.

origin explanation promoted earlier for the GRBs now finds some satisfaction in the SGRs, with N49 and the other two remnants surely containing neutron stars. But can 'classical' gamma-ray bursts be from neutron stars, as well? Is the gamma-ray burst mystery really back to where it might have been if SGRs had never 'existed', i.e., never been detected? One idea, that GRBs may be another manifestation of neutron star activity and may yet be found to be in an enormous halo, is compatible with the up-to-LMC distances known for the SGRs. The push towards a cosmological origin is winning the polls, however. Recent conferences and workshops have given warm receptions both to theorists' ideas such as a binary red-hole/neutron-star collapse and to data analysts' findings of the compatibility of both the red-shift and time dilation effects with the BATSE data bank, giving total plausibility to this preference.

New experiments based on the hope of finding transients in other wavelength regimes associated with gamma-ray bursts are underway. First, optical and radio studies are possible with a new burst alert system, the 'BACODINE' or BATSE COordinate DIstribution NEtwork. This system automatically calculates the burst direction from BATSE data and sends that information directly from Goddard to observatories in real time, as the burst takes place. Only if optical transients take place entirely before, or within the first few seconds of a GRB, will they be missed by this system. At the present time, about thirty optical and radio instruments are connected to it. The curious advantage of radio astronomy is that, if the bursts indeed come from cosmological distances, the expected radio events should be delayed in the intervening medium, making their detection easier. The High Energy Transient Experiment (HETE), an imaginative, new mission designed as a GRB alert system giving source fields small enough for telescopic view without scanning, was tragically lost at launch on November 4, 1996. HETE also carried a CCD array to directly search for GRB-associated optical transients that might pinpoint the sources of gamma ray bursts. Hardly two weeks later, the launch of the Russian Mars-96 mission also failed. This additional disaster destroyed any chances of resurrecting a full-scale interplanetary gamma-ray burst network until after the turn of the century.

As of the time of this writing, no optical transients have yet been seen by any instruments that are connected to the BACODINE system. Of course, the telescopes in use are somewhat limited in their sensitivity, relative to to the highly focussed telescopes that would require narrower fields of view than can come from BATSE. Perhaps another generation of systems may be further required, before the appropriate optical magnitude is reached, for success. Or, prehaps the supposed connection between GRBs and optical transients could turn out to be as misleading as the earlier assumption that a galactic disk anisotropy would soon be found.

Next-generation spacecraft missions for gamma-ray burst studies are also in the design stages at the time of writing. Some use new technologies, such as CdZnTe detectors, that can make precise GRB localizations directly, without the intermediate optical stage (that has yet to be shown to be viable). However, none of these has been selected for flight by NASA or by any foreign space agency. At present, the design for any experiment promoted to find the 'solution to the gamma-ray burst msystery' may have to be so highly focused towards the anticipated form of the answer that the experiment itself would have little 'survey instrument' or other value to the high-energy astrophysics community - beyond the existential gamble itself.

14.12 REFERENCES

eds. Lingenfelter, R. E., Hudson, H., and Worrall, D. 1981, in "Gamma Ray Transients and Related Astrophysical Phenomena" AIP Proc. 77.

eds. Woosley, S. 1983, in "High Energy Transients in Astrophysics" AIP Proc. 115.

eds. Liang, E., and V. Petrosian 1984, in "Gamma Ray Bursts" AIP Proc. 141.

eds. Paciesas, W., and G. Fishman 1991, "Gamma Ray Bursts" AIP Proc. 265.

eds. Fishman, G., Brainerd, J., and Hurley, K. 1993, in "Gamma Ray Bursts: 2nd Workshop" AIP Proc. 307.

eds. Rothschild, R., and Lingenfelter, R. 1995, in "High Velocity Neutron Stars and Gamma Ray Bursts" AIP Proc. 366.

ed. Maran, S. 1992, in "The Astronomy and Astrophysics Encyclopedia" (Van Nostrand Reinhold, publ.).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1996	Reference Publication

4. TITLE AND SUBTITLE
Current Perspectives in High Energy Astrophysics

5. FUNDING NUMBERS
Code 660

6. AUTHOR(S)
Jonathan Ormes, Editor

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS (ES)
Goddard Space Flight Center
Greenbelt, Maryland 20771

8. PEFORMING ORGANIZATION REPORT NUMBER
96B00115

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS (ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
NASA RP-1391

11. SUPPLEMENTARY NOTES
Ormes: Goddard Space Flight Center, Greenbelt, Maryland 20771

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category 90

Availability: NASA CASI (301) 621-0390.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

High energy astrophysics is a space-age discipline that has taken a quantum leap forward in the 1990s. The observables are photons and particles that are unable to penetrate the atmosphere and can only be observed from space or very high altitude balloons. The lectures presented as chapters of this book are based on the results from the Compton Gamma-Ray Observatory (CGRO) and Advanced Satellite for Cosmology and Astrophysics (ASCA) missions to which the Laboratory for High Energy Astrophysics at NASA's Goddard Space Flight Center made significant hardware contributions. These missions study emissions from very hot plasmas, nuclear processes, and high energy particle interactions in space.

Results to be discussed include gamma-ray beaming from active galactic nuclei (AGN), gamma-ray emission from pulsars, radioactive elements in the interstellar medium, X-ray emission from clusters of galaxies, and the progress being made to unravel the gamma-ray burst mystery. The recently launched X-ray Timing Explorer (XTE) and prospects for upcoming Astro-E and Advanced X-ray Astronomy Satellite (AXAF) missions are also discussed.

14. SUBJECT TERMS
Active galactic nuclei, AGN, Bremsstrahlung, gamma ray, X-ray

15. NUMBER OF PAGES
188

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL