Flexible Spectrum Use Rights
Tutorial - ISART 2005

Robert J. Matheson,
National Telecommunications and Information Administration
Institute for Telecommunication Sciences[1]
Matheson@its.bldrdoc.gov
303-497-3293, fax: 303-497-3680

Abstract: Although "command and control" spectrum management techniques have provided licenses for many specific services since the early days of radio, such licensing may not easily permit new technologies and new services. This paper describes the necessary principles of flexible use spectrum rights, which may allow a wide variety of spectrum uses in a single general-purpose band. Based on the electrospace description of the radio spectrum, these principles allow general aggregation or division of licensed electrospace regions via secondary markets, providing rules for how regulatory limits change under aggregation or division. These flexible-use principles limit transmitter behaviors that tend to create a more difficult operating environment for receivers, while making receivers responsible for handling any remaining interference. Flexible-use principles could provide a basis for real-world flexible-use frequency bands.

1. Introduction

The process of spectrum management determines what users and services can be provided at certain frequencies. In the past (and most of the present also) spectrum managers have assigned radio licenses that tightly prescribe exactly what frequency, bandwidth, modulation, transmitter power, geographical location, services, and type of user can be active on a specific frequency. Since a central authority assigned a specific use at each frequency, following a pre-engineered formula for that particular service, radio users could be squeezed together in a band just tightly enough for maximum efficiency, but not too tightly to cause interference. This method of spectrum management – called "command and control" – seemed to be the best way to simultaneously ensure high efficiency and freedom from interference.

The command and control technique has worked for many years, but it also has some major problems. The most obvious problem is in smoothly accommodating new services or new technologies, for which no pre-engineered formulas and associated frequency bands have been established yet. How can a new service, "A," be offered, when there are no rules and frequencies at which this new service can operate? Equally frustrating, other services, "B" and "C," may have less demand than anticipated, and their corresponding earmarked frequency bands will lie fallow for lack of users.

Keeping in mind the lack of application flexibility as the major problem of command and control regulations, one might ask if there could be a more flexible way to license radio spectrum for a broad range of technologies and applications. This paper describes a set of rules and principles – called "flexible use spectrum rights"– that would allow the spectrum to be used for a wide range of user-selected services, as well as to be freely traded, aggregated, and divided via a secondary market. Throughout the remainder of this paper, the specific term "electrospace" will be used in place of "spectrum" to eliminate confusion with the more traditional use of the word "spectrum," which usually refers to only the "frequency" dimension of the electrospace. The electrospace will be described in more detail in the next section.

The proposed flexible-use spectrum rights environment includes the following major features:

1. All signals must remain within their respective licensed electrospace region.
2. There are scalable limits on transmitter power or field intensity at ground level.
3. Receivers are unregulated, without any guarantee of freedom from interference.
4. Aggregation or division along any electrospace dimension can occur via secondary markets.

Definitions:
1. A licensed region of electrospace is a hyper-space volume described by dimensions of frequency, location, time, and angle-of-arrival.
2. A "signal" is defined to be present wherever spectral energy flux density is larger than X Watts per MHz per square-meter ($W/MHz/m^2$).

[1] The views, opinions and/or findings contained in this report do not represent an official position of the National Telecommunications and Information Administration or of the U.S. Department of Commerce.

Although the flexible-use environment described here may be especially suitable for rapidly changing radio services requiring substantial infrastructure investment, it might have some disadvantages compared to traditional environments. In particular, since a flexible-use band might have a wide variety of possible services, a receiver might need to be capable of rejecting interference from a wide variety of different types of signals. This might require a more expensive receiver design, compared to receivers in traditional bands, which typically would need to reject interference from a much smaller variety of signal types.

It remains likely that the traditional regulatory environments will continue to be especially useful for some services. In particular, low-power non-licensed (Part 15) devices and spectrum for large government (especially defense) systems do not seem likely to benefit from flexible-use principles. Traditional command-and-control regulations will surely remain for many years, while low-power and "commons" environments will probably grow. Fortunately, there seem to be few impediments to establishing different regulatory environments in different frequency bands, so we could possibly select multiple sets of regulatory features, as needed to best serve various types of users and services. Therefore, this description of flexible-use spectrum regulations should be understood as applying only to (currently non-existent) flexible-use frequency bands, without prejudicing in any way the regulatory practices that apply to other bands.

2. The Electrospace as Property

This section shows how the electrospace reasonably fits into the category of "property," so that a set of normal property transactions – including the unrestricted ability to buy or sell via a secondary market – can be established for the electrospace. It must be understood that the issues we will discuss here concern only the rights and obligations of the current holder of electrospace property, not whether the user holds the electrospace property permanently or temporarily. We describe the rights and obligations of the current property holder to use the electrospace – no matter whether the holder has a long-term lease (e.g., based on a 10-year license from the FCC), a short-term rental agreement (e.g., a temporary 1-week rental from the primary license-holder), or permanent ownership (via a permanent title transfer from the FCC or a future secondary market).

2.1 A Description of the Electrospace

The electrospace is a formalized description of the radio signal environment, as it might be seen by a hypothetical ideal measurement receiver [1]. It applies to all types of radio systems and all regulatory environments. It can also be used to describe ways in which the radio environment can be shared among multiple radio systems. It is particularly useful in flexible-use environments, where it provides a straightforward basis for unambiguously describing licensed electrospace regions, as well as for aggregating and dividing electrospace regions.

The electrospace describes radio signals, which means that it describes the domain of transmitters and transmission paths. Any description of real receiver characteristics is totally separate from the electrospace description. Although any real radio system must consider all system components – i.e., the electrospace and the receiver – it is appropriate to divide these two components for regulatory purposes. The crucial regulatory difference between the two components is that the electrospace describes the ability of radio signals to cause interference to other users – which involves an externalized cost that must be regulated. The receiver domain, however, includes only components that do not cause interference to others. Therefore, the receiver domain has no associated externalized costs that need to be controlled by regulations; it can operate completely free of regulation.

An "ideal" electrospace model is based on the radio universe as seen from the viewpoint of an ideal receiver. In Section 3.3, various assumptions about the limitations of real receivers to reject unwanted signals will be used to modify electrospace management rules, giving interference rights that are more appropriate and efficient for real-world use.

The electrospace describes the radio field strength at a given electrospace "location" that is defined by the 7 electrospace dimensions. These 7 dimensions are all independent of each other, which means that the electrospace can be considered to be a 7-dimensional hyperspace. A "location" in the electrospace can be described by assigning specific values to several independent variables. It should be noted that different investigators have sometimes included other sets of variables in the electrospace. The set shown in Table 1 is a useful starting point, and probably no great harm is done by including or omitting some marginal variables such as polarization and modulation.

The physical location of a test point or hypothetical receiver is defined by the three spatial dimensions. The field strength characteristics at that location are described by the remaining variables, including the frequency, time of occurrence, and angle-of-arrival. In a frequency band whose licensing is based on the electrospace, a numerical limit will typically be established, such that field strengths in excess of X are considered to be signals, which are not permitted outside of the user's licensed regions of the electrospace. An electrospace "region" consists of all points within a described 7-dimension hyperspace volume. An electrospace region is typically used to denote the hyperspace volume defined by an electrospace license (e.g., the licensed electrospace region) or the hyperspace volume occupied by a signal (e.g., the region where field strength is greater than X).

Table 1 - Electrospace Dimensions

Quantity	Units	# of dimensions
Frequency	kHz, MHz, or GHz	1
Time	seconds, hours, or years	1
Spatial location (geography)	latitude, longitude, altitude	3
Angle-of-arrival	azimuth, elevation angle	2

One characteristic of the electrospace is that an ideal receiver can theoretically separate any radio signals that differ by at least one of their seven electrospace dimensions. For example, two co-located radio receivers could function without interference if the signals were at different frequencies, or if the signals occurred at different times, or if the signals came from different directions. Radio signals using the same frequency, operating time, and angle-of-arrival could be separated without interference if the receivers were present at different locations.

2.2 Comments on Electrospace Dimensions

Frequency. The frequency dimension of the electrospace has the standard meanings of the word, namely a description of the frequency or range of frequencies (bandwidth) at which field strength is being characterized. Frequencies can be divided over a wide range of increments, typically matching the channelization of particular services.

Time. The time dimension can be subdivided over a wide range of increments. Useful time divisions might include the several-year-duration of a licence, an agreement to allow a particular user to transmit regularly during the midnight-to-5 AM time block (when bandwidth would be inexpensively available to update computer files for the following day), or a one-time use during a 4-hour special events broadcast. On a much smaller time scale, a user could use a particular time slot on a TDMA system, to transmit during a 2.5-ms time slot that would be available once every 20 ms, or transmit data during the vertical blanking interval of an NTSC television signal 30 times every second.

Spatial location. The spatial dimensions represent a physical (geographical) location. They can be problematic, because there is no practical way to confine radio signals within a desired region. In typical hilly terrain, there are many distant locations that have higher signal amplitudes than many closer locations. Therefore, although one might easily select an arbitrary spatial region, the selected region might be extremely inconvenient to use efficiently. In order to prevent excessive signal levels (larger than "X") outside the boundaries of the selected spatial region, it might be necessary to greatly diminish signal amplitudes at many otherwise-useful locations within the spatial boundaries. Transmitter power, details of the terrain, and the use of directional transmitting antennas are operative in establishing the spatial boundaries of the electrospace associated with a given transmitter.

Angle-of-arrival. This factor describes the angle-of-arrival or direction of radio signals at a given location, including the possible effect of multipath components scattered from many objects in many different directions from the receiver location. Note that this factor is not created by physical antenna pointing angles. The pointing direction of transmitting antennas primarily affects the spatial dimensions of the occupied electrospace, i.e., the geographical areas where signals are larger than X. No aspect of physical receivers – including the pointing angle of receiving antennas – ever has any effect on the electrospace. Therefore, neither transmitting nor receiving antennas influence the angle-of-arrival factor. On the other hand, receivers that exploit the angle-of-arrival dimension will often employ directional antennas. Recently developed "multiple-input, multiple-output" (MIMO) technology exploits multipath reflections coming from different directions, handled by multiple transmitting and receiving antennas and mathematically processed to generate independent transmission channels. MIMO technology can be considered to be a generalization of the angle-of-arrival dimension of the electrospace.

3. Flexible-use Rights in the Electrospace

3.1 Use of the Electrospace to Divide the Radio Environment

The electrospace model can be directly applied to a flexible-use, market-based frequency management environment, since the model describes a way that the use of the radio spectrum can be unambiguously divided (shared) among multiple users. The only significant regulatory principle is that a licensee has the right to radiate a signal within a licensed electrospace region. Outside the licensed region, signals must be kept below a specified very low spectral power flux density limit, X, in Watts/m^2/MHz. X includes power coming from all directions, though in many cases the great majority of total power will arrive from the direction of the respective transmitter location. Note that X is a power that is proportional to bandwidth. There are no restrictions on type of service, transmitter power, bandwidth, modulation, antenna height, number of sites, etc., as long as the signal is kept lower than X at all points (including all dimensions of time, frequency, location, and direction-of-arrival) outside the licensed electrospace boundaries.

Ideal Electrospace Rules
(assumes ideal receivers)

1. Transmit without any restrictions inside your licensed electrospace region.

2. Keep your signals below X outside licensed electrospace region.

As described in greater detail in Section 4, an electrospace region is permitted unlimited aggregation or subdivision along all of its dimensions. This allows electrospace to be freely repackaged and resold as a market-based commodity, redistributing spectrum without requiring prior approval by a regulator. A 1-MHz bandwidth could be subdivided into 40 channels of 25 kHz each or augmented with 4 MHz of additional adjacent frequencies to make a single 5-MHz bandwidth. A given channel could be subdivided into TDMA time slots of 10 ms occurring once every second and rented to a hundred separate transmitters. A statewide geographic coverage area could be divided into much smaller geographical cells and rented to short-range neighborhood wireless ISPs. Multiple fixed transmitters could be allowed to radiate signals into a common receiver location, if the transmitters are arranged to provide signals that have different angles-of-arrival.

Although the electrospace model is critically based on a specified spectral power flux density limit, X W/m^2/MHz, which cannot be exceeded outside the licensed region, it is not obvious what numeric value to choose for X. Presumably X will be chosen so that systems licensed outside the region will not receive interference from the signal. However, the minimum level of interfering signal for various types of systems varies over a wide range – perhaps 50-60 dB – depending on the system. Since all types of systems are assumed to operate in a flexible-use band, which type of system should X protect? One answer is that the selection of a specific value for X might be done to make that band particularly suitable or unsuitable for various types of services; multiple bands could use different values of X to efficiently accommodate various services.

3.2 Practical Limitations on the Electrospace

Although the electrospace model is conceptually powerful and potentially very useful, there are a few important problems with its application to the real world. One major problem, non-ideal receivers, will be discussed in the next section. Other problems are discussed in this section.

The division of the electrospace along any selected dimensions – while theoretically possible – may or may not produce a useful division in the real world. Arbitrary spatial regions, for example, may not match easily achievable propagation/coverage areas. A more useful spatial division technique may be to use propagation models to determine easily achievable coverage areas and divide the electrospace regions in a corresponding way. The angle-of-arrival dimensions may be compromised by unintended scattering from the terrain or by lack of sufficiently-narrow-beamwidth receiving antenna performance (especially at lower frequencies). Division into very narrow time slots may produce systems that are difficult to synchronize properly. Division into very narrow frequency slots may produce unreasonable requirements for frequency stability and Doppler shift.

The spatial dimensions pose some other problems. The field strength at a particular location is often the vector sum of many multipath signals. These multiple signals can occasionally add up to a field strength that is larger than the average field strength in the general vicinity. Therefore, it may be desirable that the field strength limit contain a statistical parameter, which would allow the occasional presence of signals above the limit. However, the inclusion of a statistical limit might make it much more difficult to show that a user had violated the electrospace limits, since a single instance of excess field strength might not be sufficient proof of a violation.

One obvious application of spatial coordinates is to describe licensed regions using some imaginary lines drawn on the ground – e.g., lines described by latitudes and longitudes, a circle centered on a designated location, geographical boundaries, political boundaries, etc. For many applications, radio signals will be attenuated by buildings, terrain, and the earth's curvature, which all tend to give the greatest attenuation at ground level. Raising a receiving antenna farther above ground will usually increase the received signal level. Therefore, a transmitted signal that is below X at ground level will often increase greatly at higher elevations above ground.

Many radio systems have receiving antennas located on tall buildings, towers, or mountaintops. Therefore, a simple electrospace boundary at ground level may describe only part of the real world; the success of a given application may depend on a much more complex understanding of how field strength changes with all three of the spatial electrospace dimensions – possibly with much more complex 3-D descriptions of the associated electrospace regions.

The frequency dimension can also cause problems. Although a transmitter can radiate any amount of power inside the licensed frequency range, the signal strength outside the licensed band must be less than X. Presumably this condition must be met at all locations – even very close to the transmitting antenna, where the in-band field strength is very high. To meet the "X" condition near a transmitter may involve a very high signal level inside the licensed region to drop below X immediately outside the licensed region (bandwidth) – requiring a very rapid decrease in signal strength over a small change in frequency. Therefore, the out-of-region absolute limit, X, may need to be supplemented by establishing an optional "relative-dB" emission mask that provides a legal "safe harbor." The relative-dB emission mask would allow higher out-of-region levels in locations (e.g., near transmitters) where the field strength is very high. However, in areas where the in-region field strength was already low, a relative-dB emission mask would require out-of-region levels lower than X. Therefore, although the relative-dB emission mask could replace the absolute X criterion at any location. In practice the relative-dB emission would be invoked only near strong transmitters.

3.3 Receiver Regulatory Theory

The most serious limitation on the practical application of the electrospace model to flexible-use spectrum management is that the electrospace model assumes that all receivers are "ideal." In this context, "ideal" means that the receiver has infinite rejection of unwanted frequencies (i.e., signal power at frequencies outside of the nominal receiver bandpass), infinite dynamic range (strong out-of-band signals will not cause intermodulation products or gain compression), and directional receiving antennas that have infinite rejection of signals coming from unwanted directions. At the end of Section 2.2, we stated that an ideal receiver could theoretically separate any two signals that were different in at least one of their electrospace coordinates. Some electrospace dimensions are easier to separate than others; even a simple non-ideal receiver could separate signals that were present in substantially different locations, frequencies, or times. However, an ideal receiver could even separate signals occurring at the same location, frequency, and time – as long as the signals came from different angles-of-arrival (separated through the use of directional receiving antennas).

Two non-identical signals (i.e., two signals with different electrospace coordinates) can always be separated and received without interference if the receiver is good enough, including receiving antennas as part of the receiver. This means that all interference is always caused because the receiver is not good enough. There is no theoretical line separating cases where interference is caused by poor receiver performance from cases where interference is caused by an "actual" interfering signal. Although an inadequate receiver is *always* the cause of interference, the required "adequate" receiver might be extraordinarily complex and expensive, and it might not actually be achievable with today's technology. For example, an adequate receiver might require an elaborate adaptive antenna array to null out unwanted signals, while generating a high-gain receive beam in the direction of the desired signal. Such technology would be quite difficult today for even large fixed base stations; it would surely be completely impossible today for handheld portable radios. But tomorrow ... who knows?

If all receivers were ideal receivers, we would only need to worry about foreign signals that illegally intruded within our licensed electrospace region to appear at the frequency of our desired signal – so-called "in-band" interfering signals. In-band interference is controlled chiefly through the electrospace parameter "X," which sets a limit on the level of signal that can be present outside its licensed electrospace region. If all receivers were ideal receivers, the electrospace rules that control "X" would be all that is needed to control interference. Unfortunately, none of the receivers that are available to users at reasonable prices are ideal receivers. Even worse, the most popular and rapidly growing class of receivers – handheld, multi-band cellphones – are especially non-ideal, with performance constrained by small size, low cost, and limited battery power. An important characteristic of real (i.e., non-ideal) receivers is that they can generate interference even when no unwanted signal is actually present at the tuned receiver frequency. Strong signals at close-in frequencies or very strong signals at frequencies further away from the tuned frequency can also cause receiver distortions that are seen as interference; this is known as "out-of-band" interference.

Fortunately, real radio systems do not usually require ideal receivers for satisfactory operation. Instead, they merely require "good-enough" receivers. A "good-enough" receiver is a receiver whose performance is at least good enough to achieve the desired system performance in the actual radio signal environment. The required level of performance for a receiver that is good enough to reject unwanted signals without experiencing interference will vary greatly, depending on the specific characteristics of the electrospace environment in which the receiver is operating.

The overall regulatory strategy is to supplement the electrospace rules to produce a more benign signal environment that allows the successful operation of less-expensive "good-enough" receivers. Hopefully, this will allow an improvement of the overall cost/benefits that can be achieved from operating radio systems. The various supplemental rules that are selected for each band (some of which are described in the following paragraphs) should be selected on a principle of maximizing overall benefits – balancing the benefits from less-expensive "good-enough" receivers with the disadvantages of adding some restrictions on transmitter characteristics. Presumably, different rules could be selected for different bands, since this will differentially maximize the benefits for various types of systems that could be built in each band. Under these supplemented rules, the important principles that regulate interference are now:

Practical Electrospace Rules
(assumes non-ideal receivers)

1. Transmit within power restrictions inside your licensed electrospace region.

2. Keep your signals below X outside licensed electrospace region.

Transmitters must still follow the electrospace rules, including the supplemental rules that make a more benign environment for receivers. As before, receivers are not regulated in any manner. They are allowed to be as-good-as or as-poor-as their owners permit. There is no implied protection against interference, except that there is the expectation that the radio environment will probably allow the use of cheaper receivers. The aforementioned limitations on transmitter power will be discussed in more detail in section 3.4, and the way that these limitations scale when electrospace regions are aggregated or divided will be discussed in section 4.

A major advantage of these principles (compared to the current command-and-control rules, which tend to try to guarantee interference-free performance) is that there is much less legal ambiguity about who is responsible for fixing interference situations. Assuming that transmitters obey the supplemented electrospace rules, the receiver owner is completely responsible for solving his own interference problems. There is never any assumption that a transmitter operating within these rules has any further obligations to prevent interference to any receiver. An exception to this general rule would apply to tightly-grouped transmitters and receivers, where site managers could have the authority to adjust radio systems to reduce interference. Note that this situation is a well-known "exceptional" case in traditional frequency management, also.

In an interference situation, the receiver owner has several basic options to deal with the problem:

a. Show that a specific transmitter is violating one of the applicable supplemented electrospace rules, and require that the offending transmitter change its operation to become compliant.

b. Improve his own system, as needed, to eliminate the interference. Depending on the exact cause of the interference, the changes might involve improving the performance of the victim receiver, increasing desired transmitter power (if permitted under electrospace rules), adding better error correction, etc.

c. Figure out how to tolerate the interference. This might involve changing operating procedures, restricting the operation to areas where interference is not a problem, ignoring the issue, issuing the customer a partial refund, etc.

d. Negotiate with the interferer. This is a strictly voluntary negotiation for both parties. After investigation of the possible alternatives in (a) – (c), it might turn out that an adjustment of the interfering transmitter would be the best way to solve the problem. If so, negotiations between the parties might result in an appropriate mutually voluntary business arrangement that could become a legal attachment to the respective electrospace licenses.

The opportunity to select an appropriate receiver remains completely with the receiver owner. The supplemented electrospace rules help to define the statistics of the expected unwanted signal environment in which a receiver must operate. The receiver owner has complete flexibility to select whatever level of receiver performance that he has judged to be adequate to accomplish the mission of his radio system. One would expect that there will be a wide variation of performance requirements among the population of operating radio systems, and the radio system owner has much better knowledge of the specific economic and operational requirements of his own mission than any federal regulator. Moreover, the radio system owner is more highly motivated than anyone else to make a correct decision about how to get that required performance. Finally, the owner's selection of receiver performance does not cause any additional interference to any other radio system. Therefore, this set of decisions can be left completely in the hands of the radio system owner.

3.4 A Limit on Power or Maximum Field Strength

The original electrospace rules control external signals at the receiver operating frequency (in-band interference) by requiring that they are always less than X outside of their own electrospace region. The supplemented electrospace rules are needed to control the presence of (legal) strong signals at frequencies outside the receiver tuned frequency (out-of-band interference). Strong transmitters that cause high signal levels within the relatively wideband first RF stages in a practical receiver are usually the major cause of out-of-band interference. Therefore, supplementing the electrospace rules by placing a limit on transmitter power (or EIRP) is one obvious approach to controlling the occurrence of strong signals in the radio environment and reducing out-of-band interference for practical receivers. Note that limiting transmitter power would not be expected to eliminate all out-of-band interference. However, out-of-band interference would tend to be limited to a much smaller set of circumstances where the victim receiver is located very close to a transmitter tuned to a nearby frequency. Therefore, the use of supplemented electrospace rules will tend to allow interference-free operation in more locations using cheaper receivers.

It should be noted, however, that transmitter power is not solely responsible for causing out-of-band interference in receivers; additional factors must also be present. Specifically, the direct cause of out-of-band interference to receivers is when receivers are subject to high-field-strength, out-of-band signals. The cause of the high-field-strength signal interference to receivers is the result of a combination of three factors:

1. High transmitter power,
2. Transmitter vertical antenna patterns that produce high-field-strength signals on the ground near the transmitter, and
3. The presence of susceptible receivers in the high field strength areas.

Presumably, the out-of-band interference could also be prevented if any suitable combination of these three factors could be arranged, including controlling transmitter power, controlling the transmitter antenna patterns underneath/nearby the transmitting antenna, or placing transmitting antennas in locations where receivers will only rarely be found in the nearby high field strength locations. Therefore, instead of controlling the interference only by limiting transmitter power, it would provide more user flexibility to also allow the control of interference by controlling transmitter antenna patterns, and/or by carefully separating transmitter sites from high concentrations of susceptible receivers.

Thus, a more effective supplementary rule to protect receivers might include a limit on signal field strength at ground level instead of a limit on maximum transmitter power. This alternative rule would state that field strength at ground level must be less than E_{max}, where E_{max} corresponds to a maximum watts/m^2. Note that this limit is not bandwidth-dependent, since the total power at the receiver input is usually what causes the problems, and the receiver front-end circuits will tend to be much wider bandwidth than most transmitters. This limitation must be met only in areas where there is a likelihood that susceptible receivers will normally be found there. In some circumstances, it might be necessary to similarly protect additional not-at-ground-level outdoor locations where people are often found (e.g., elevated walkways, rooftop cafes on nearby buildings, etc.).

This maximum-field-strength rule would allow much more flexibility in building a wide variety of radio systems, and it would protect receivers better.[2] Although this supplemented electrospace rule would not limit maximum transmitter power, it would still ensure that receivers are protected from the high-level fields that can cause interference. A higher power transmitter will still need to stay below a fixed maximum field strength at ground level. Part of the "cost" of using a higher transmitter power is that the field strength at ground level will need to be suppressed relatively more, so that it still meets the E_{max} field strength limit. In an economic sense, this rule would tend to ensure that the higher cost of using a more powerful transmitter is borne entirely by the transmitter owner, instead of being partly externalized to unrelated receiver owners.

In summary, two basic methods could be used to control out-of-band interference problems. A transmitter power limit (or EIRP, or equivalent) is the simpler rule to apply. This "EIRP"rule indirectly tends to control the high-field-strength locations that can cause interference in receivers. The alternative version of the rule directly establishes a maximum field strength, E_{max}, and leaves the details up to the transmitter owner. Although the "E_{max}" rule is more complex in application, it provides better protection to receivers and allows more freedom in designing transmitter systems. There is no reason to require that a single rule would need to be applied to all flexible-use bands. One rule could be applied to one band, the other rule to another band.

[2] Note that the interference to public safety LMR in the 800-MHz band was caused by allowing apparently reasonable changes in antenna locations, even without allowing any changes in transmitter power. This is an example where E_{max} limits would have allowed greater flexibility in system architecture, while simultaneously providing better protection from interference.

There are some possible refinements to either version of this supplementary rule. Since the possibility of overload is actually caused by the total power into a receiver, the E_{max} limit should be determined by the total equivalent power from all the fields from various transmitters in a given location. Thus, the rule will probably require an obvious adjustment in areas where multiple transmitters produce high field strengths. However, since most receiving antennas do not operate efficiently over a very wide frequency band, the total power counted at a location would include only transmitters within fairly close frequencies. The actual algorithm for computing the weighting of field strengths with frequency might change according to the typical receiver front-end or antenna technologies used in nearby frequency bands.

It will also probably prove useful to adjust the values of X (maximum signal leakage outside of licenced electrospace) and E_{max} (maximum field strength) in various flexible-use bands to preferentially optimize their use for various types of service. Special consideration should be given to nearby bands that contain large numbers of receivers that might be particularly susceptible to strong signals, such as portable (cellular, PCS) transceivers. For operational simplicity, it may also be useful to include "safe harbor" rules, so that transmitters with EIRP below a certain power limit would be automatically assumed to meet the E_{max} field strength rules.

Other limitations on flexible use might also be beneficial in certain frequency bands. For example, one large class of radio systems (including most LMR and cellular/PCS services) will benefit from frequency bands that are engineered into duplex band architectures, where base station receiver frequencies are systematically separated from base station transmitter frequencies. Therefore, although "maximum-flexibility-of-use" remains a key principle, some bands will benefit from a requirement that specific sets of frequencies can be used only for base station transmitters or only for mobile transmitters. Similar generic restrictions may prove useful for other flexible-use bands intended to efficiently support other types of services, though additional examples of such restrictions are not yet obvious.

The actual values of X will need to be determined according to the performance of receivers that operate within or nearby the various flexible-use bands. It seems reasonable to expect that the selection of different combinations of values for these parameters will create bands that have different "sweet spots" for systems of different bandwidths and services. It would seem useful to eventually allocate a variety of flexible-use bands having different "sweet spots," which would be expected to differentially attract a varied mix of applications in each band.

Note that the value of the parameter E_{max}, is totally determined by current (and past) practical receiver technology; there is nothing theoretically binding about these values. If a future change in receiver technology causes the performance of receivers to change substantially, this numerical value should also be expected to change (presumably after sufficient discussion and rule-making). Over the years, receiver performance has occasionally changed dramatically. The development of the "superhet" receiver created a major improvement in receiver performance, including much better receiver selectivity. The recent development of the receiver-on-a chip technologies have surely made receivers much smaller and cheaper, but not necessarily much better. Major changes in receiver performance may result from much smarter receivers (that figure out how to move to a better frequency or a better modulation), from receivers using digital RF or IF processing (where certain types of receiver distortions can be recognized and processed away), from room-temperature superconductors (producing very-narrow-band, very-high-Q, RF filters that could reject many of the signals that would cause out-of-band interference in today's receivers), or from adaptive antenna technology (that nulls out many strong unwanted signals).

Possibly none or possibly all of these receiver changes will actually occur in the next few decades. Since there is a substantial possibility of change, however, it might be useful to figure out how to easily change the values of the operational parameters that regulate the use of flexible-use bands. This would allow the band "sweet spots" to track the changes in markets and technologies.

4. Freedom to Aggregate or Divide

An important feature of the flexible-use regulatory environment is freedom to aggregate or divide an electrospace region along any or all of the 7 electrospace dimensions, presumably according to a secondary market and without the permission of a regulator. If this freedom is permitted, it will be necessary to define how the rules (including the values for X and E_{max}) can be made to scale in a reasonable manner for the resulting new electrospace regions.

4.1 General Principles

The applicable principle here is that aggregation or division of an electrospace region should not expose electrospace neighbors to any greater threat of interference after a "transaction" than existed before the transaction. All allowable transactions must meet this general principle.

Only electrospace regions regulated by identical sets of rules can be aggregated. Whenever electrospace regions are combined, any original regional borders that are now interior to the new region can be ignored. No limits that were associated with these "interior" borders need to be obeyed anymore. When a single party owns multiple adjacent electrospace regions, the owner can decide whether to consider the two regions as a single region or as multiple independent regions. In most cases, the common owner simply chooses not to enforce (against himself) the rules associated with excessive signal levels (signals greater than X) leaking across interior borders.

Whenever electrospace regions are divided, the new borders associated with the new regions must now meet all of the conditions associated with the borders of the original electrospace regions. All external boundaries maintain the same set of rules as before the transaction. No sets of internal changes can be construed to change the rules or values associated with external regions.

Any set of electrospace regions can be joined together. Similarly, a given electrospace region can be sub-divided into multiple new electrospace regions – essentially without any constraints or limits. However, the mere ability to identify and create a new electrospace region should not be understood to imply that the resulting electrospace region will necessarily be useful for any specific job. This limitation is particularly important when dividing or combining along the geographical dimensions, where natural terrain and buildings will tend to set limits – instead of being set by any arbitrary latitude/longitude boundaries.

4.2 Rules for Scaling X

The limit, X, for the amount of signal allowed outside of a licensed electrospace region is scaled in terms of $W/MHz/m^2$. When geographical areas are added or subtracted from a region, the change merely affects the geographical position of the boundaries outside of which the signal must be suppressed below X. Similar effects are applied to changes in the time and angle-of-arrival boundaries. When the frequency boundary is changed, the bandwidth of the signal leaking across geographical boundaries will presumably change with the bandwidth of the primary signal, but the value of X at any particular frequency will remain the same.

The only complication of aggregating or combining electrospace regions comes from a very basic understanding of what constitutes a "signal." In particular, each independent signal source is allowed to leak a very small amount of power (up to "X") at any or all electrospace locations outside the licensed region. If a single owner had ten base stations within a geographical region, each using the same frequency, those ten base stations would be part of the same electrospace license, and cumulatively they could not leak more than X signal at any given frequency outside of the licensed electrospace region. If the single owner divided his electrospace region geographically into ten regions (one base station in each region), each of the base stations would be part of a different electrospace region and could presumably leak X outside of its own electrospace region. Potentially, this could represent a cumulative leakage of 10X from the ten independent stations.

Similarly, a region with a 10-MHz bandwidth might be filled with a single 10-MHz bandwidth signal that leaked X power at various frequencies outside the licensed frequency region. The single owner might divide the electrospace region into ten frequency regions, each having a 1-MHz bandwidth and containing a 1-MHz portion of the previously described 10-MHz signal. The new owner of the ten 1-MHz regions could claim that each 1-MHz region could individually radiate X energy to various other frequencies outside the ten regions, producing as much as 10X cumulative energy at any frequency outside the licensed regions (assuming that one could show that each 1-MHz region independently produced X energy at a certain outside frequency).

In each of these cases, the owner of a single region could apparently get permission to leak more signal outside his electrospace region by simply claiming that a single electrospace region (and signal) had been divided into multiple electrospace regions (and signals) – each "signal" with a separate allowance for X. Therefore, it may be necessary to understand that an independent "signal" must actually be independent of other signals in order to qualify for a separate right to radiate X outside the region. A COFDM signal could not be arbitrarily split into a thousand multiple carriers and corresponding electrospace regions, if they shared error-correction mechanisms and data between the various carriers. The entire COFDM signal would have to be treated as a single signal, entitled to leak no more than X into neighboring regions.

The exact rules by which a single signal is defined may be a little tricky to define exactly, since many independent signals might be used in a network of various load-sharing paths using different signals. Some other independent signals are very closely coordinated (e.g., synchronized spreading codes in multiple CDMA signals at a single base station). Are simulcast transmitters one signal or many signals? One could imagine future systems where multiple independent signals are combined to be amplified by a single broadband transmitter power amplifier and radiated from a single antenna, or where a single signal is split among multiple transmitting antennas in an adaptive array (after each antenna feed signal is adjusted for gain and phase, amplified, and maybe more?).

In summary, X does not need to scale in any way under aggregation or division of electrospace regions. The arbitrary division of a signal into separate pieces to acquire a separate allowance for X for each divided portion is not permitted.

4.3 Rules for Scaling EIRP and E_{max}

This section describes two possible rules for scaling the power of transmitters under aggregation or division – one for the transmitter power/EIRP case and one for the E_{max} case. It will be noted that these two cases develop somewhat different sets of rules.

EIRP/transmitter power model. In the case of an electrospace model that includes a maximum transmitter power (or EIRP) = Y, the actual definition is in terms of Y = Watts/MHz. If a wider bandwidth is divided to give two smaller bandwidths, each of the smaller bandwidths will have a maximum transmitter power proportional to the relative bandwidths of the new regions. Moreover, the maximum transmitter power of the original bandwidth will be equal to the sum of the maximum power of the two smaller bandwidths.

An additional refinement could be added to the above rule, based on typical receiver performance. The probability of interference from a strong unwanted signal is affected by the total power of the unwanted signal and also by how close the unwanted signal is to the frequency of the desired signal. There is more chance of interference if the strong unwanted signal is close to the frequency of the desired signal. Therefore, a further rule could be proposed, which states that the radiated energy should be spread out evenly across the licensed bandwidth, instead of being allowed to be concentrated at the edges of the licensed bandwidth. Otherwise, the entire extra transmitter power allowed by aggregating more bandwidth could be placed in a CW signal at one extreme edge of the bandwidth, creating a stronger signal immediately next to the frequency range used by an electrospace neighbor. Therefore, the proposed rule states that the cumulative radiated power measured from the edge of the licensed bandwidth to any point inside the licensed bandwidth cannot be more than twice the total power that would result if the allowable average power/EIRP were totaled over that same frequency range. The factor of two allows a wide range of modulations to be used without any derating of total transmitter power.

This scaling rule for EIRP is a very natural rule for scaling spectrum use rights, since the total transmitter power does not change when a given transmitter is divided into smaller bandwidths. However, such a rule is not totally effective in preventing interference to receivers. As a transmitter becomes wider in bandwidth (presumably, by aggregating additional frequencies), its maximum power can increase (proportional to bandwidth), until the transmitter is possibly powerful enough to cause out-of-band interference in a nearby victim receiver. This is a problem that is quite similar to the current FCC rules concerning the situation where the number of separate transmitters at a given base station increases until a certain total power threshold is crossed. Eventually, there is sufficient total power radiated from the base station transmitters that they will cause out-of-band interference in nearby receivers (actually, the FCC maximum-power rules are intended to prevent health dangers to people). At that point, the combination of transmitters becomes equally responsible to control the problem, with any transmitter that contributes more than a certain percentage (e.g.,10%) of the total power being responsible to decrease transmitter power as required to cause the total power to drop below a certain threshold.

Similarly, a useful set of rules for the EIRP model would be to scale maximum transmitter power proportional to bandwidth under normal conditions of division and aggregation. However, once a certain maximum power threshold had been crossed, power would be limited on an absolute basis.

In terms of scaling power along other electrospace axes, the dimensions of time, space, or angle-of-arrival do not cause any difference in transmitter power scaling. Of course, extending the geographical area of a region may allow more powerful transmitters to be employed, simply because the new regional boundaries are further from the transmitter site, permitting more transmitter power without violating the leakage of signals above X outside the new boundary.

Although the installation of additional transmitter sites within a region can increase the total power radiated at a given frequency, this will generally not increase the risk of interference to other users. The probability of interference from out-of-band signals is primarily related to the presence of *strong* unwanted signals, not by the total area over which a weaker unwanted signal is available. Therefore, there is no reason to limit the total power radiated by multiple sites, as long as the total power radiated by a single site is controlled.

E_{max} model. An alternative model for scaling transmitter power under aggregation and division is used to control the maximum field strength = E_{max} = V/m at ground level where receivers will be present. Since the occurrence of out-of-band interference is mostly related to the total amount of signal within the very-wide-bandwidth electronic circuits at the receiver front end, it should be assumed that all of the energy from any transmitter at any nearby frequency will be available to cause out-of-band interference. Therefore, the E_{max} limit does not scale with transmitter bandwidth, but remains tied to a certain maximum field strength. Presumably, the cumulative power from multiple transmitters should be included within this limit, with some rules for requiring compliance by any group of multiple transmitters that cumulatively violates the field strength limit. No other constraints are imposed to prevent out-of-band interference from other transmitters.

5. Summary

The preceding sections have described a possible spectrum management approach to flexible-use spectrum rights. This set of concepts could provide a market-based method that would give great flexibility in the use of spectrum, while controlling most interference. In the few circumstances where interference might result, the model includes clear rules for assigning responsibility to mitigate the interference.

This flexible-use model is believed to provide a highly flexible environment in which new or modified services can be rapidly provided by following a very small number of rules. It should be noted that there are still some areas of ambiguity, but some of these can be resolved with "safe harbor" practices or more detailed rules. Also note that many of the ambiguities refer to situations that are also substantial ambiguities under current command-and-control spectrum management practices. The flexible-use rules do not somehow make the radio world simpler than it now is, and even under flexible-use rules it will be necessary to occasionally make complex and difficult technical trade-offs. However, unlike with the current command and control management, the spectrum owner would be directly authorized to immediately make and implement these decisions, instead of waiting for an expensive and problematic federal regulatory process.

It is likely that a flexible-use environment will also have some disadvantages. What appears as "freedom" to one licensee might appear as a "lack of needed guidance and prescribed practices" to another licensee. A higher degree of technical expertise might be required to put a new system in a flexible-use band. The possible lack of expertise in flexible-use system design might lead to higher levels of interference in a flexible-use band. The lack of narrow standards in the band might mean that a new system would have to be designed to withstand interference from a wide variety of possible interferers. This might require a more expensive system design than would be necessary in a more traditional band (where only one type of interferer would usually be present).

All of this suggests that a traditional single-service frequency band might remain the most suitable band for radio systems that comfortably fit there. However, it is also expected that technological obsolescence may fairly rapidly create situations where new applications no longer fit the existing band allocations. A major question will then be whether additional uses can be painlessly grafted into existing band allocations, using some of the principles of flexible-use rights described here, or whether some other conversion technique will be more useful.

Some of the remaining questions about possible flexible-use bands include:

1. What specific values for X, Y, E_{max} should be chosen for a specific flexible-use band? Which sets of values would best match specific technologies or services?
2. What are appropriate characteristics for relative-dB safe-harbor emission masks?
3. What is a usable definition of a single signal? (to prevent multiple-X emission limits)
4. What is the best way to describe the geographical limits – especially re the vertical dimensions?
5. Are there any other holes in the model?
6. Would this model be too difficult to administer or enforce? Who would be responsible for enforcement?

6. References

1. R. J. Matheson, "The Electrospace Model as a Frequency Management Tool." Addendum to the Proceedings of the 2003 ISART Conference, March 4-7, 2003. J. W Allen and T.X Brown, editors. NTIA Special Publication SP-03-401, March 2003.