Viral radio

A Lippman and D P Reed

This paper defines a domain of study, some experiments and a research agenda to explore a topic we term viral radio. The premise is that we can make energy- and spectrum-efficient radio communications systems that scale (almost) without bound. We do this by treating the RF signals in a given space as a distributed optimisation process whereby each radio uses the presence of other radios to assist and co-operate in the delivery of messages. Any relaying that occurs is done in the RF domain; we thus eliminate delays normally associated with multi-hop ad hoc networks. Further, we embed the routing decision in the RF processing and view it as a matter of 'flux-propagation' rather than path definition — data is delivered from a source transmitter to the ultimate recipient with some RF amplification provided by any radios that are in the propagation path. Our goal is to develop a simple radio networking architecture organised on an end-to-end design basis. We expect that we can build scalable and efficient real-time telecommunications and broadcast systems that rely on no central radiator or suite of cell towers.

1. Introduction

1.1 Viral and traditional radio

Historically, radio has been viewed as a restricted resource to be used when no other signalling method is available. This is due to many factors, including the perceived lack of security of a radio communications system and the inherent detectability of radiation. Perhaps most important, modulation techniques and radio designs for consumer systems have been optimised for an inexpensive receiver (or in the case of two-way systems, both an inexpensive receiver and a simple modulator.) As a result, the use of the RF spectrum has been so inefficient that a myth of scarcity has evolved.

Associated with this is a regimen for allocation and use that entails leaving most of the spectrum 'dark' with only a few permitted radiators in a region. When open communications are allowed at all, such as with Citizen's Band or the Family Radio Service (FRS) in the USA, the cacophony that results as the service succeeds reinforces support for limiting access to 'real' or 'important' systems. The extent to which the scarce spectrum is used in practice has been vividly demonstrated by informal measurements taken in urban areas over the course of a day (Fig 1). It is remarkably quiescent, considering its economic value.

Exclusive spectrum registration is not an artefact of the physics of radio; it is the combination of engineering limitations of the 1920s and the interests of existing stakeholders. A television broadcaster, for example, once given a licence, has little incentive to invest in a more efficient system that might allow newcomers on-the-air or require a new receiver for its customers. Yet the popularity of digital spectrum-sharing systems, such as cellular telephony and IEEE802.11 indicates a public interest in increased use of radio.

Further, the grassroots nature of WiFi data networks, Citizen's Band radio, and the Family Radio Service support the more general notion that radio systems that are evolvable by the users are economically and socially valuable. The forum for innovation is open to large segments of the user community, the use architecture is open to change, and the low cost of entry promotes new ideas. (This is the more general theme of 'viral innovation' that argues that such end-to-end systems are inherently more responsive than ones with an expensive or inflexible central architecture.)

> the forum for innovation is open to large segments of the user community

The above noted spectrum sharing also uses a fundamentally different design philosophy that is predicated more on statistical presence of desired communications than a full-time reservation. They mimic the design of the original Ethernet in many regards — the medium assumed the occurrence of collisions that would impede individual packets, but the overall communication integrity was sufficient. Yet, even with the Ethernet, when the system was scaled and used more intensively, designers looked to other options. The single cable first used was subdivided by routers and bridges. In the

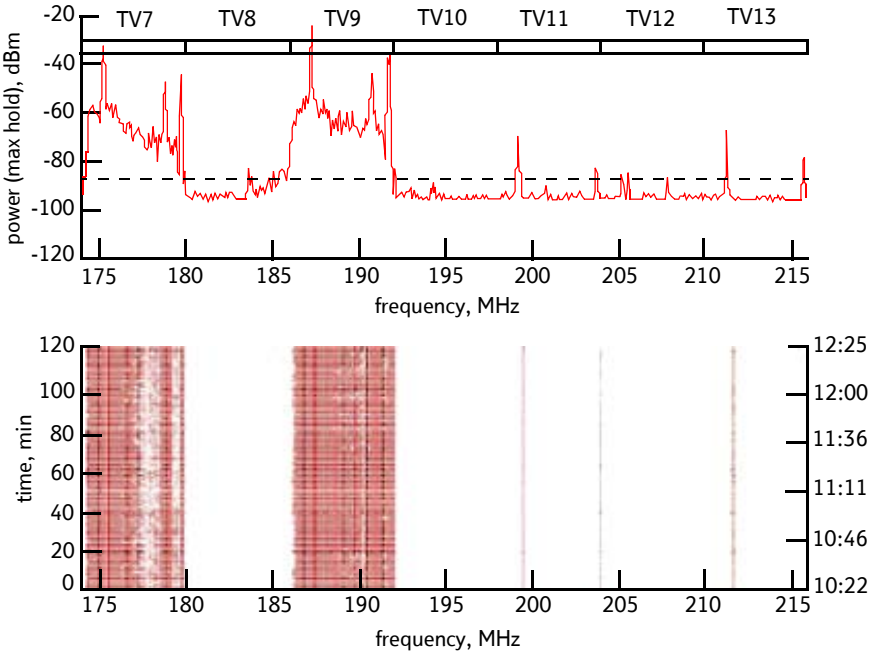

Fig 1 Spectrum use in Dupont Circle, Washington, DC [courtesy New America Foundation].

case of statistical radio communications systems in use today, the same end is accomplished by other means, such as idiosyncratic protocols that 'listen before talking'. Sometimes, the point of fragility has not yet been reached[1]. In all cases, the systems technically do not scale — ultimately, their capacity is fixed by design and divided among the users.

We present a design methodology for radio systems that avoids these pitfalls. It is optimised for three parameters:

- scalability — the system must allow virtually unlimited use and access,
- incremental growth — the system should be deployable without first constructing a backbone,
- value-conservation — ideally, each additional element should contribute to the capacity of the system as a whole.

viral systems are both relatively infrastructure-free and also inherently open to innovation

Taken together, we call such a system 'viral' [1]. In its most general form, the implications of a viral system are that it is both relatively infrastructure-free (and thus can gain grassroots adoption), and also inherently flexible and open to innovation in that there need be no large-scale deployed core system on which it is based. There are two main motivations for this approach. Its modularity provides a flexible basis for innovation in both use and system evolution. Our goal is a network design that promotes development, can be widely used for personal and embedded applications, and works both locally and when densely deployed. Also, this work implicitly takes a view of spectrum use that is based on co-operation among elements and a more global estimation of spectrum capacity. We invert the historical notion that derives from simple radio design. Instead, co-operative radios that are practical, with today's processing capabilities and design methods, extend the notion of spectrum capacity, interference, and allocation.

Technically, the basic issues are building a radio system:

- where the capacity increases with the number of elements,
- where co-operation among the elements optimises the distribution of information.

We see no limit to the potential growth of radio networking, and our goal is that radio should become the default communications medium, with wires reserved for the special occasions when one cannot distribute power any other way.

2. Sharing an electromagnetic field for communications

Metaphors have shaped our thinking about radio communications for many years. In most cases these metaphors have been elaborated into a mathematical foundation, in some cases a deep mathematical foundation. But these metaphors were invented to simplify thinking, and as such they are merely approximations to physical reality. A deep mathematical foundation without a correspondingly sound physical foundation can produce misleading conclusions. For example, we often talk about the 'range' of a radio transmitter or a transmission, as if the wave originating at the transmitter could be thought of as 'detectable' over all

[1] Technically, the cited systems are scarce in that they are resource inefficient. Even 802.11 will saturate with enough use. However, in their early phases, before these pitfalls became evident, they were seemingly capacious.

distances up to a certain point, and then past that point, would be 'undetectable'. It is quite easy to see that this metaphor is almost never correct in the real physics of the world. Most listeners to FM signals in their cars are familiar with the phenomenon — when reception gets marginal at a stoplight, you merely need to move the car a few feet to improve the signal strength, and thereby dramatically improve reception. Figure 2 shows how signal energy at a receiver can vary dramatically as the receiver's position varies in real propagation environments. Note that the change due to variation is more significant than the average decline with distance.

As engineers, we know why this happens in great detail. Electromagnetic waves energy propagate in a richly structured environment, where phenomena such as reflection, refraction, multipath, and attenuation (these are merely descriptive metaphors, also, of course) create localised fading[2]. In the environments most humans inhabit (indoor and urban ones), propagation is dominated by these effects. Thus the notion of 'range' as a way to model propagation of an individual transmission makes little sense as a practical model of the reality that must be addressed by systems designers who would build scalable radio networks.

For point-to-point links, it has been sufficient to model propagation statistically. In other words, to get a reliable link, the design problem is to cope with the likelihood that the received signal strength exceeds the level required for effective communication. One treats the received signal strength as a random function that has a distribution that is known once one knows the distance between the two end-points. This model implies that distance is a dominant parameter, and that the other effects are random and uncontrollable, i.e. the received signal $R(t)$ can be expressed as:

$$R(t) = \frac{H(t) \times S(t)}{\alpha A(d)} + N(t)$$

where $S(t)$ is the transmitted signal, $H(t)$ is an impulse response that characterises the physical path between source and destination, $A(d)$ is the attenuation as a function of distance, α is a random variable that characterises the statistical variability of attenuation, and $N(t)$ is a random noise input measured at the receiver. Clearly for common situations as shown in Fig 2, separating out a distance-based attenuation is not very helpful, since α is the dominant source of variation.

However, it is not necessary to view propagation as a random process characterised by statistical parameters, since it is physically deterministic. An alternative way to think about propagation is to view it as a measurable property of the system environment that can be exploited by the system by design.

[2] Fading is a radio communications engineering term that is used to refer to local gain or attenuation variations in a received signal due to propagation effects. In physics, fading would be called (constructive or destructive) interference, since it results from interactions from a wave (or photon) with itself. However, 'interference' in radio communications refers to the difficulty of separating independent signal or noise waves from a linear superposition at a particular antenna sensor. In this paper we use the radio engineering terminology.

What we do not know, however, is how to exploit the true nature of radio propagation. We have no examples of system designs that can scale in propagation environments that do not behave as simply as a collection of dipole antennas in free-space. A whole line of theoretical research focused on scalability of wireless network communications is grounded in models based on the naïve notion that the single key parameter of radios is 'range', which is controlled by transmit power. For example, Gupta and Kumar model this assumption using Voronoi diagrams to describe coverage regions of distinct transceivers as a partitioning of space, in order to prove (within the limits of their model) that planar networks of packet repeaters have a system-wide transport capacity that scales as $N^{1/2}$ where N is the number of nodes [2].

Similarly, designers of wireless LANs make assumptions about 'range' of this simple sort in their designs. For example, the 802.11b/g/a standards assume that all stations in a network can receive all other stations, even in the '*ad hoc*' mode where they use a distributed co-ordination function (DCF). This is usually described by saying that all stations are within a certain distance of each other, relying on the notion of range.

It is difficult to create analytical models that characterise typical indoor and urban environment propagation with accuracy. Attempts to create models that represent a concept like 'range' posit that measured signal strength is a monotonically decreasing function of distance in such environments. For example, a well-known rule-of-thumb used by engineers for indoor signal propagation is that received signal strength declines according to a 'path loss exponent', that is $A(d) \approx d^q$ where q is between 3 and 4 [3][3]. Such an empirical model used to design individual radio links should be applied very carefully to networks, since it was created to model the minimum received signal strength that can be assumed as a function of distance. The high degree of variation in signal strength due to the physical environment is not represented in such models.

it is not necessary to predict these propagation effects to exploit them adaptively

As a design rule for estimating a worst-case propagation bound for a link, a power law model is adequate, when used to

[3] Such an empirical model calls for an explanation via physics, not just a curve fit to some data. Conservation of energy requires that the average energy flux over an enclosing sphere at a distance d from a source is $1/d^2$, when energy is not absorbed (accumulated in some energy absorbing material). There may be considerable variation over the sphere's surface due to propagation effects, but the average should remain constant. For long distance radio links, the curvature of the earth and energy absorption by the ground may result in a uniform deviation from $1/d^2$ near the ground over a certain span of distance, but this clearly does not apply to microwave signals from omnidirectional antennas with no objects in their Fresnel zone. Similarly, off-axis reception with directional antennas can decline via a function other than $1/d^2$, but again, propagation effects create a high degree of variation. Our investigation of the propagation literature suggests that any single-exponent characterisation of path loss may not have a firm scientific footing.

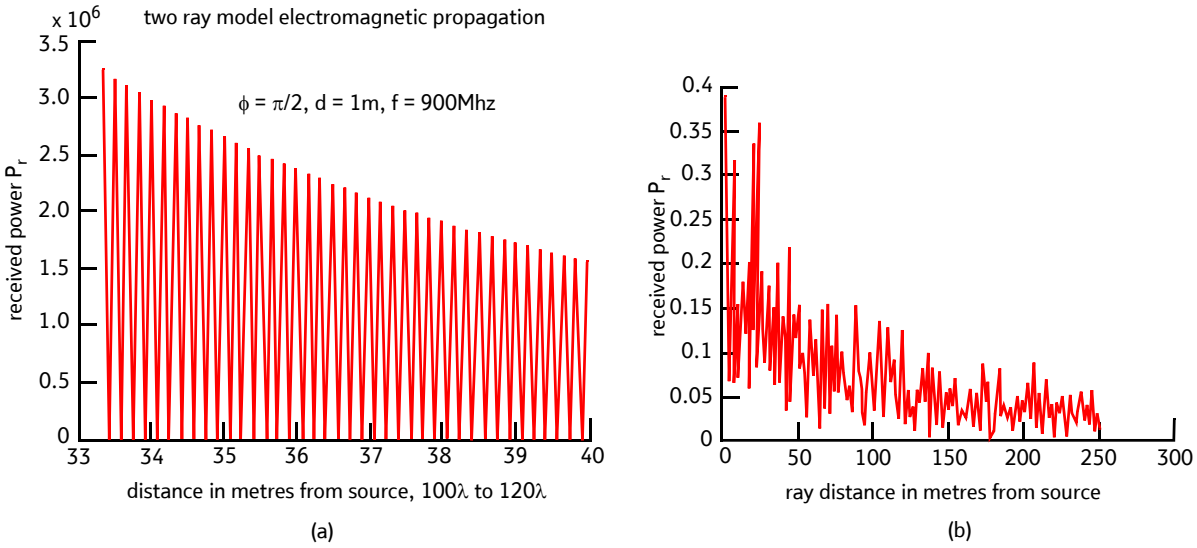

Fig 2 Calculated examples of signal power versus range — (a) an example of a signal source and one 0 dB reflector such as a wall, where signal strength is plotted as a function of distance from the source along an axis parallel to the reflector surface; (b) a collection of 0 dB reflectors arranged randomly, where signal strength is plotted as a function of distance on a ray from the source.

predict the minimum signal power needed to achieve a particular link capacity against a constant noise background. But a worst-case propagation bound for a desired signal is a 'best case' bound for that signal as interference. As the 'worst case' for an interfering signal is the strongest value it can attain, assuming the worst-case propagation for interfering signals in complex environments requires consideration of local signal levels that far exceed the inverse-square law of distance predicted for free space.

As we will show later, it is not necessary to predict these propagation effects to exploit them adaptively. In particular, we can select the relay paths used between intermediate relay nodes, thereby turning the high degree of both fading and interference variation to our advantage. When a relay is in a deep fade from one source, it perforce has a greater signal-to-noise ratio for the channel between it and another source. Indeed, with enough radios, there is a strong likelihood of a multihop path between any source and destination that need not encounter energy from other communicators [4].

In a network of many nodes co-operating in communications in a shared electromagnetic field, especially in an indoor or urban environment, links cannot be treated in isolation using simple worst-case assumptions. Transmitting at a higher-than-necessary power on one link often has an adverse impact on the capacity of other links in the system. An efficient use of the realisable capacity among a network of radios would adjust the emitted power from each transmitter to transmit the largest number of bits possible while at the same time minimising its reduction of capacity effects on other currently active links.

The traffic capacity that can be achieved in such a system involves complex trade-offs. A common means of describing such a capacity is the achievable rate region (ARR) [5], which represents the information delivery rate for each end-to-end channel as a distinct axis in a space that has as many dimensions as there are potential channels. A point in such a space represents a particular combination of rates on all channels. The set of points that represents combinations that can be achieved by a physical system is the achievable rate region. For simple systems, the achievable rate region may be a convex *N*-dimensional polyhedron. However, in complex physical environments the ARR may be topologically quite complex.

The achievable rate region is useful because it illustrates that the same radio network can handle a diverse combination of end-to-end loads by shifting its operation from one point in the achievable rate region to another. A system with a large achievable rate region may be more valuable than a system that is statically limited to one combination of end-to-end rates, even if the total throughput is maximised at one point. The value of a network architecture that can adapt to wide load variability is much greater than a system that has pre-assigned fixed rates to each end-to-end link. Consequently, the area of the ARR itself, not just its maxima, creates part of a system's economic value, when it allows adaptation to unpredicted loads[4].

Such a picture still leaves out key elements of the trade-offs in real wireless communications. First of all, the combinations of power levels, modulation schemes, and access protocols used by transmitters to achieve the rates in the different portions of space is not represented — so factors like battery life and computational complexity may be hard to extract from that representation. Secondly, the achievable rate region's shape is crucially dependent on the physics of the propagation environment, as noted above. As small perturbations in the

[4] Such contingent economic value may be characterised precisely as a form of real option value [6].

[5] The vector and matrix elements are signals, i.e. functions of time, so in the matrix operations and other equations that follow, multiplication of signals means convolution, and addition means ordinary point-by-point addition.

physical layout of transmitters and receivers can result in large changes in the achievable rate region, the representation does not capture this sensitivity at all.

The use of achievable rate region as a term is also misleading in itself. The word 'achievable' implies a basic or fundamental limit. But it is clear that any particular ARR is quite narrow in its applicability to the particular physical conditions and system design assumptions, and does not arise from physical law.

2.1 Propagation space

Rather than focusing on 'range', a better approach is to use the network itself to measure and share information about propagation. A complete representation of the propagation space might be in the form of an impulse response matrix, which has an entry $P_{i,j}(t)$ that is the response measured at station *j* for an impulse at station *i*. The response to any given combination of transmissions could then be calculated by multiplying[5] the vector *S* of waveforms that represents the transmitted signals by the *P* matrix, which would then give a vector *R* of received signals that would be measured at all receivers:

$$R = S \times P$$

Thus the *P* matrix is a compact representation of propagation. (It may not be practical to compute all of the information in *P* in a real system, but any real system may estimate aspects of *P*.)

2.2 Noise

Noise arises in many places in radio systems. One key observation, though, is that noise is highly localised and localisable. A 'pulse of noise' does not arise simultaneously in many places, nor is noise uncorrelated from place to place. Noise takes the form of waves that propagate just like signals. One way to include them in the system, then, would be just to add rows to the propagation matrix, viewing noise as a collection of independent sources whose effects propagate just as signals do. This model is useful when considering the scaling properties of the system, because as the number of radios in a system grows, the role of noise may grow less significant. In particular, as the scale of the system grows, correlation of noise between different receivers grows, and the architecture of the system can take advantage of this correlation.

2.3 Interference

Each radio in the common electromagnetic field senses the effect of all disturbances in the field. Consequently there is potential for an information-destructive trade-off between multiple independent users of the field. However, there is an equal and largely unexploited potential for radio nodes to constructively assist in delivering information, both when they are idle and when they are transmitting, which arises from relaying, joint coding, and other co-operative activities. Thus interference in the sense of interaction need not be bad or avoided — instead it should be a part of the normal operation of the system. The crucial measure for interference is whether it reduces the capacity of the system in terms of actual end-to-end messages delivered over a period of time. The structure of propagation space and the offered load of messages to be delivered jointly determine whether the end-to-end message capacity increases or decreases when a particular transmitter's relative amplitude is increased or when a particular message's mapping into waveforms is changed.

2.4 Orthogonality grows with the number of stations

The problem of 'coding' is best considered in propagation space. Space-time coding is based on the idea that in diffusive environments, portions of the propagation matrix are of full rank, and therefore invertible without loss of information. In such a case, the information capacity of propagation space is equivalent to a set of independent channels equal to the rank of the propagation matrix. This has been shown to be realisable in architectures like BLAST [7].

The ability to create many independent channels in the same space seemed surprising when BLAST was invented, but it should not have been. It results naturally from a simple insight — the waveforms emitted from any random set of spatially distinct antennas are linearly independent. That is, there is no set of coefficients by which one can multiply those waveforms to make them sum to zero over the entire field. This is easy to show mathematically for free-space EM waves, and empirically seems to be true for EM waves in real 3-D situations. There are special cases where the physics does not provide linear independence, for example a set of transmitters spaced within a one-dimensional waveguide.

2.5 Incremental scalability versus optimality at a fixed scale

An important lesson from computational complexity theory (a branch of computer science that analyses the cost of algorithms) is that an algorithm that scales well with the size of the problem is often quite different from an algorithm that is optimal at a particular scale. Viral networking seeks network architectures that scale well as the network of radios in an area gets denser and denser. There is no reason to believe that viral architectures will be the optimal network at any particular scale or for any particular physical environment. The crucial issue is one of sufficiency, not optimality — does the viral network work 'well enough' at a particular scale of deployment, and can radio nodes be added to it without requiring a whole new approach as the scale increases?

interference in the sense of interaction need not be bad or avoided

A static partition of the electromagnetic field (by frequency, time, direction, etc) may optimise the sharing of the network at any particular scale, and for any particular set of communications demand. But optimum partitionings are often quite brittle — adding a single new node, or small changes in demand may require complete reorganisation if one is to obtain a new optimum. Given that we expect networks to grow and change, our interest is not in obtaining theoretical optimality in any particular physical situation, but there is a major benefit to developing architectures that improve in efficiency as they scale incrementally.

3. A new 'hourglass model' of radio networking

Traditionally, the job of radio spectrum management is to subdivide the electromagnetic field into disjoint pieces that can be allocated to different applications. Division according to frequency, geography, angle of arrival, polarisation, modulation technique, and time may create such pieces, for example. This partition into pieces is then used to create links, or virtual wires, that can then optionally be assembled into networks.

An alternative view with very different implications involves thinking about all radio communications with a new layered model in the form of an 'hourglass model' of Internet architecture [8] akin to the depiction of the Internet composed by the Computer Science and Technology Board of the National Research Council in 1994 (Fig 3) [9] that focuses on the end-to-end message delivery function as the fixed central concern.

This hourglass model defines the common function of all radio systems abstractly as delivering messages from a set of sources to a set of destinations. Messages are strings of bits, and the sources and destinations are the ultimate producers and consumers of those messages. The source-destination messaging function is the 'neck of the hourglass' because all applications layered on top of that are implemented in terms of messages.

there is a major benefit to developing architectures that improve in efficiency as they scale incrementally

Below the neck of the hourglass are radios and radio networks; these can then co-operate in their use of the electromagnetic field to optimise the end-to-end delivery of messages, without the applications being aware of the network.

The difference between this 'hourglass model' and the traditional radio network architecture is illustrated by the composition rule used to build systems. The traditional radio network first creates point-to-point links that are virtual wires. Then the links are interconnected into a graph where the edges are links and the nodes are switches. The new mode builds a larger radio network by co-ordinating the interactions between smaller networks. The component networks in a radio network have inherent interactions — the RF energy of each subnetwork impinges on the others. The co-ordination of interaction exploits useful interactions and mitigates the effect of other interactions. There is no need to create 'links'.

In other words, rather than try to simulate the isolation inherent in wire-based links, the new model makes a virtue of the lack of isolation. Through co-operation the individual radios gain access to many more degrees of freedom or vibration modes in the electromagnetic field that all radios share.

3.1 Co-operative radio precedents

Only in the past twenty years have radio systems for widespread use been a fertile area for research. Before that, the equipment for a sophisticated system was beyond the reach of most researchers (perhaps with the exception of defence-oriented imaging such as radar and some secure communications systems), and consumer application was not envisioned. Beginning in the 1980s, when digital processing became fast enough and accessible, there was a renewal of interest. An early example is the problem of 'ghost elimination' in television. Urban propagation results in severe multipath distortion, and manufacturers attempted to solve the problem. Philips, for example, used baseband processing of the video signal to deconvolve the delayed replicas of the television signal and thus reduce the echo visibility. By doing this, two problems are solved — firstly the distorting signal is minimised, but, more importantly, the second, or reflected signal, by shifting it into coincidence with the original, adds energy to the net signal and thus can improve picture quality. We call this multipath gain.

Radio systems with inherent multipath immunity have been devised. The notion behind frequency domain multiplexing (FDM), for example, is that by using a signalling interval that is long compared to the potential multipath delay, successive copies of a signal would not result in inter-symbol interference. Such a signalling rate implies a low data rate, but a correspondingly narrow channel. Therefore, a multiplicity of such channels are concatenated by frequency to achieve the desirable rate. When the symbol rate is chosen so that each channel is orthogonal, there is then no adjacent channel interference. The orthogonality is realised in practice by using an inverse DFT to create the broadband signal [10].

Guard intervals between successive bits help make the system robust in the face of multipath interference. Frequency-dependent channel response within the band of interest is generally accounted for by using forward error correction redundancy in the channel (COFDM). A by-product of this multipath immunity that concerns us here is that each signal reflector between the transmitter and the receiver is, in effect, a secondary relay for the data. The impact of this is that one can create a wide area single frequency network (SFN) by substituting additional transmitters for the reflectors. Each transmitter radiates a synchronised version of the original signal, and each receiver anywhere in the reception area receives energy from one or more transmitters and any reflecting elements in the path. Each transmitter is equivalent to a multipath source but it adds power to the transmission. In practice, the separate transmitters for this type of broadcast are wired together or synchronised by backhaul on a separate channel. SFN thus provides multi-transmitter gain. Schreiber demonstrated this in 1995 [11].

Foschini et al [12, 13] showed that diverse paths of a signal can be exploited at the receiver to provide better transmission efficiency. V-Blast demonstrates that by using multiple antennas, multipath can result in better reception. V-Blast is primarily a fixed demonstration. Laneman and Wornell [14, 15] showed how multiple antenna diversity can be used to improve efficiency in mobile *ad hoc* communications, for example by using addition relay nodes in a space to improve

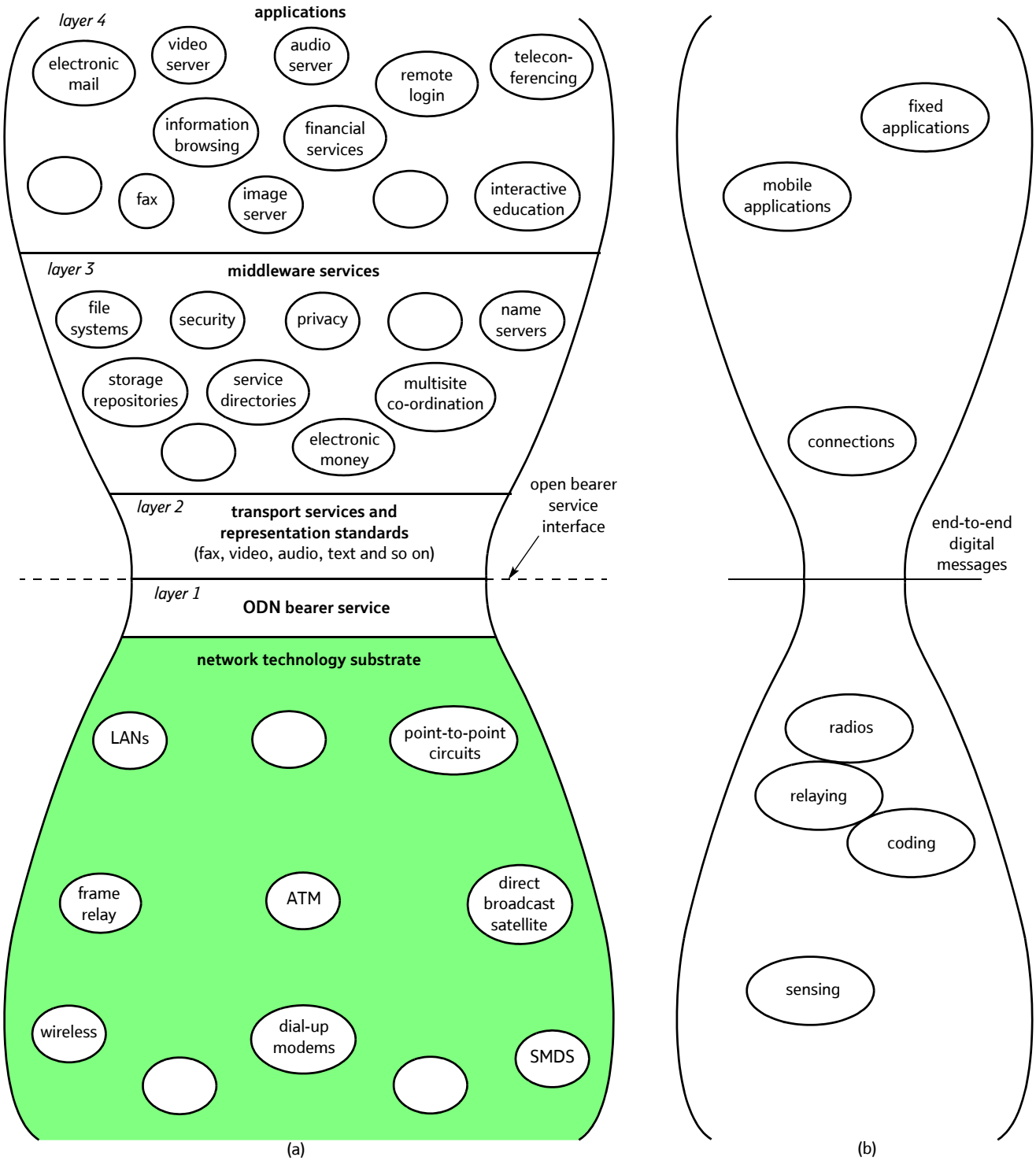

Fig 3 (a) Internet 'hourglass model' and (b) our radio networking 'hourglass model'.

signalling; Zheng [16] formalised the gain in terms of using the additional channel capacity thus gained for either better energy conservation or for additional channel capacity.

Gupta and Kumar [2], and Gastpar and Vetterli [17], have explored the efficiency limits possible in such schemes, although the general area is not fully understood and the ultimate limits of capacity are unknown both in theory and in practice.

A key idea that we bring into our concept is based on the idea of the 'end-to-end' principle, which suggests that function be moved from low architectural layers to higher ones, and from central control points to the 'edges' or clients of the network [18, 19].

In the radio network case, we treat the distributed network as an adaptive shared radio channel, rather than trying to simulate wires. The network elements are simple remotely

controlled radio repeaters, much like the 'bent pipe' geosynchronous satellite designs that amplify and reflect signals back to earth.

4. Co-operative network architectures

To date, the approaches to radio described above have been used primarily to demonstrate the theory or for land-based communications systems. They provide a theoretical and practical basis for radio systems that scale through their efficiency, either by power conservation at each transmitter, or by allowing dynamic channel estimation. Here we extend that work to construct a scalable communications network. Our goal, as stated earlier, is a network where additional nodes create additional network capacity through co-operation, where delay is minimised by making routing decisions based on the RF environment (rather than the packet destination alone), and where distributed discovery of propagation parameters contribute to the overall system efficiency. We note that although the primary thrust of our work is in the domain of wireless networks, the same principles of co-operation also apply in wired networks. We have built a wired implementation of a co-operative distribution network for audio and video information and a protocol to support it [20]. In a wired system, locality of information distribution is the source of the economy.

Large-scale, mobile wireless networks where multiple relay hops are utilised for a communication path benefit most from such a scheme. Indeed, Shepard [21] did early work to assemble an *ad hoc* set of radios into a scalable network based on power conservation. However, he did not dwell on the issues of delay. In fact, if we allow arbitrary delay, we could account for the vagaries of most network traffic patterns quite simply with sufficient buffering at all nodes in the system. (Even Internet routers avoid this; it is pessimistically scalable.)

In a multihop scenario, delay increases rapidly when each intermediate note must analyse the packet to make a routing decision. Since each digital relay must accept a full frame, decode, and retransmit the frame, the delay is equal to the time it takes to fully accept the transmitted packet plus the processing time to decode, re-encode and retransmit. We can minimise lag by making link bandwidth arbitrarily large; however, this is not practical in a radio communications system with limited or restricted bandwidth.

Of course, in a real system, each potential intermediate node still has to make some decision about whether it will re-transmit the information or else the space will become overloaded with spurious broadcasts. It is sufficient for the work done here that when the motion of the nodes is slow compared to the packet rate, an intermediate node can make the relay decision once and maintain state for the remainder of the transmissions. It is also true that the decision to continue to relay can be made in parallel with analogue retransmission and expire when the conditions so warrant.

An example of the network architecture is shown as in Fig 4(b). The distinction from '*ad hoc* networking' shown in Fig 4(a) is that the signal at the receiver is the sum of the relay and original signals and the receiver treats the original signal as information-bearing, rather than as part of the noise[6].

4.1 Analogue relaying

Analogue relays are a new way to change the structure of propagation space. For example, a sequence of k relays between a particular source and destination can dramatically reduce the power/bit needed to transmit a message from a source to its destination — to get the same end-to-end SNR one can use less than $1/k^2$ energy per hop, or less than $1/k$ of the energy in total [22]. At the same time, even without directional antennas, the signal is distributed much more along the end-to-end path. Note that the ultimate receiver receives a signal that contains 'echoes' of itself, essentially a form of artificial multipath distortion. This is an important observation since delivery techniques that emulate radio as virtual wires assume that a given link is a point-to-point communication that propagates no further than the destination node. Given a coding technique that exploits multipath distortion to achieve gain such as COFDM for example[7], the multipath distortion can be deconvolved from the signal to achieve improved capacity with reduced energy release. Since analogue relays look like a generalised form of 'multipath' distortion, they can be easily represented in our P matrix as modifications to the impulse response between every pair of stations. Alternatively, channel pairing can be used so that any intermediate relay can receive on one channel

[6] For the sake of illustration, we make the critical simplification that masks the essence of the work: that the propagation is related to distance. We show the design in this manner to make the architecture clear. Since we argue that propagation is a function of more than distance, we use the notion of 'propagation proximity' as a proxy for physical distance, but we illustrate it physically.

[7] We note that OFDM is a blockwise encoding. Therefore the end-to-end delay will never be less than the encoding time of one block. However, block formation is a prerequisite for packet communi-cations and for most audio and video compression schemes.

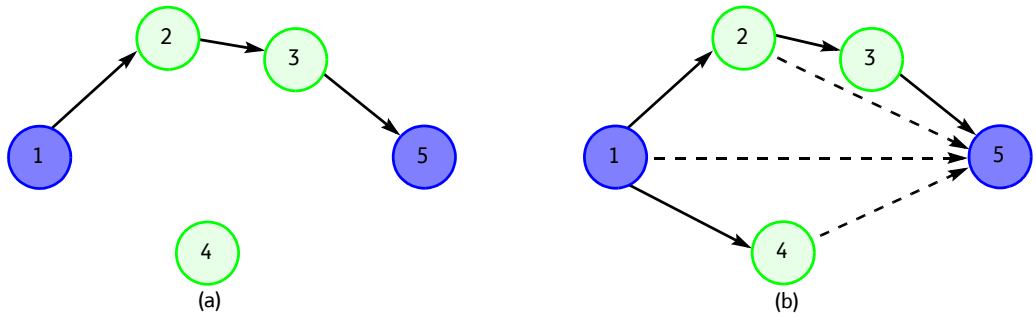

Fig 4 *Ad hoc* routing (a) and RF relaying (b). RF relaying makes use of all signal energy incident at the ultimate receiver to carry information.

and simultaneously transmit on a second one. Other relays can choose their transmit and receive pair and the ultimate recipient (who does no re-transmission) processes the sum of signals on both channels. An example of such a network architecture is shown in Fig 5.

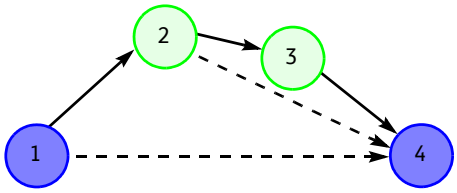

Fig 5 Channel pairing example.

The ability to use idle stations as analogue relays allows the network as a whole to 'distort' propagation space to create more information-efficient signal propagation to enhance communications along certain paths, while limiting interaction between communications along other paths. As noted earlier, the distortion has only a small effect on latency — because like multipath, it changes the delay spread of a signal — which does not scale in a way that depends on the packet size or message size.

Relays must be realised in real physical form. The biggest problem to solve is to add energy to the incoming signal without responding to the relay's own transmitted energy. One simple analogue relay we have explored has the relay listen to incoming signals for a period of time equal to half of a 'symbol period', and then repeat that signal for an equal period of time. This function reinforces the information content of the signal at the ultimate receiver in a predictable way, while guaranteeing that the relay does not suffer from feedback. Essentially it convolves the signal with a delayed copy of the signal, behaving much like a multipath interferer whose impact remains within the symbol period. As is shown by Bletsas and Lippman [23], this analogue relay increases the ultimate information capacity of the link dramatically without adding significant delay. In the same noise environment, more bits can be delivered with the same radiated energy, and the energy radiated becomes more concentrated near the relays. There are lots of ways a simple relay can reinforce an incoming signal by adding energy to the field that avoids self-feedback at the relay — the analysis is not crucially dependent on the technique used.

As the density of radios grows, the ability to distort propagation space grows in proportion to that density. For example, intermediate analogue relays can reduce or eliminate the effects of shadow fading. This kind of distortion involves relays that may not be arranged on a linear path, but instead on multiple paths (such as the two relays on each side of a large intermediate obstacle between the source and destination in Fig 6).

Maximising the beneficial distortion of propagation space, and minimising the negative impact of energy on stations transmitting simultaneously between other source-destination pairs, involves adapting the relays dynamically to demand. The primary parameter of each relay is its gain. If we assume that the ultimate destination can deconvolve any number of

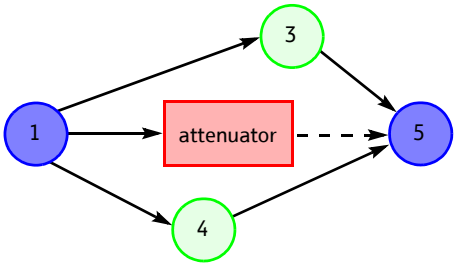

Fig 6 Relaying reducing shadow fading.

relays, increasing the gain on certain intermediate relays (or the amplitude of the original source) increases the achievable end-to-end bit rate on that particular path. At the same time, there will be a negative impact on the achievable rate on some concurrent end-to-end transmissions. This arises because the gain applied to non-signal data exceeds the noise level that allows the desired signalling rate.

This suggests that there are one or more optimal settings for the gains on intermediate relay nodes that achieve a desired combination of end-to-end signalling rates using minimal transmission energy/bit.

4.2 Dynamic propagation

Our approach is to develop a distributed control algorithm that dynamically adjusts the gains on the individual relays in response to 'congestion'. The trade-off between flows is very much like the end-to-end TCP congestion-control problem, except that instead of buffer overflow, overloads occur when you cannot separate the signals that have been mixed. Ultimately, only the target-recipient can determine when some competing signal is getting too much gain — their error rate goes up. The key insight here is that the target recipient (or the source) can respond to this by signalling back to the intermediate nodes to cut their gain.

4.3 Routing by flux propagation

The goal of the routing algorithm is that an intermediate node that is in between a source and a destination in propagation space will amplify and re-transmit an incident signal, but only when such energy addition does not increase the error rate of other communications in the region.

The goal of the routing algorithm is to move the system operation within a large achievable rate region that is available by altering the relative gains of potential relay nodes (i.e. nodes that have available capacity to act as a relay). Note that analogue relays can combine in any number of ways to increase capacity between a source and a destination — they can form a long 'bucket-brigade' (Fig 7(a)) that carries the signal like a wire, or a set of relays laid out like a lens or reflector around the source or destination can spread the signal out in such a way that it converges on the destination (Fig 7(b)). The former is useful in free space to allow many concurrent 'parallel paths' while minimising cross-interactions, and the latter is useful to provide ways to distribute information around obstacles that create shadow fading.

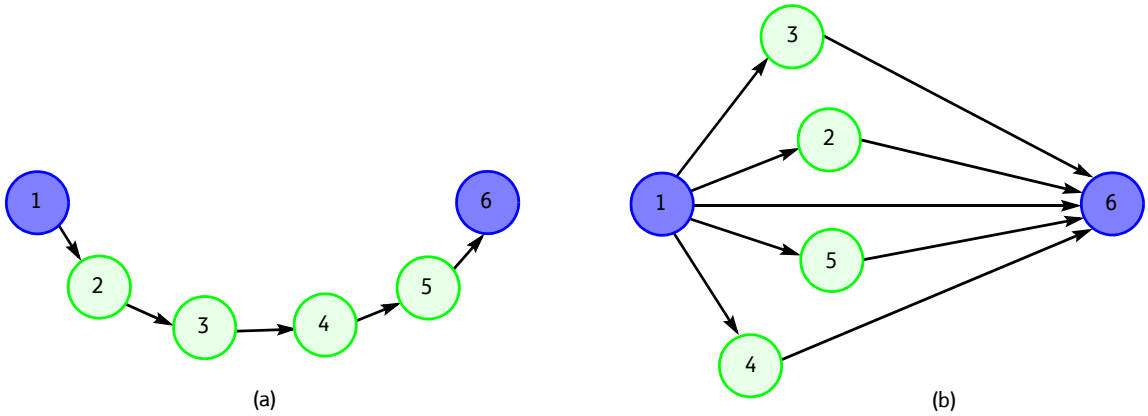

Fig 7 Bucket-brigade (a) or convergent spreading (b) repeaters.

The fundamental routing algorithm that we want to investigate works by locally adapting to satisfy the offered demand with the lowest cost. We assume that the network carries messages from sources to sinks where flow-rate control is shared between source and sink, just as in TCP. FCC rules and coexistence with other systems suggests that a primary goal should be to keep the local energy density in any part of the network below some bound. Additionally, one may desire to enforce some kind of 'fair sharing' among competing users of the network, or develop a 'charging scheme' that makes 'hogs' pay for excess use of the shared resources in the network. Note that the shared resources include elements of varying value and cost, such as battery power and power from the grid.

Unlike a wired network, which is built out of fixed links, the links in our wireless network can shift capacity dynamically by warping propagation space, using intermediate nodes as repeaters. So the nodes of the network can respond to the offered demand by increasing or decreasing the gains on intermediate repeaters.

A radio in any particular region can detect that their region is underutilised by sharing information with neighbours about the energy flux profile in their region. If they are also on a path between (in propagation space) an active source-destination pair, they can improve the efficiency of that communication by acting as a repeater, which allows adjacent nodes on the path to lower their amplification. Such routing decisions depend on the current messaging activity, as suggested in Fig 8.

We assume that the source and destination of a flow are able to communicate with all of the repeaters involved in delivering a flow by an 'in-band' channel that is part of the flow that allows the end-points to address the intermediate repeaters (for example asking them to increase or decrease tto add itself as a repeater by joining the end-to-end flow temporarily and signalling to the destination its willingness to participate. The

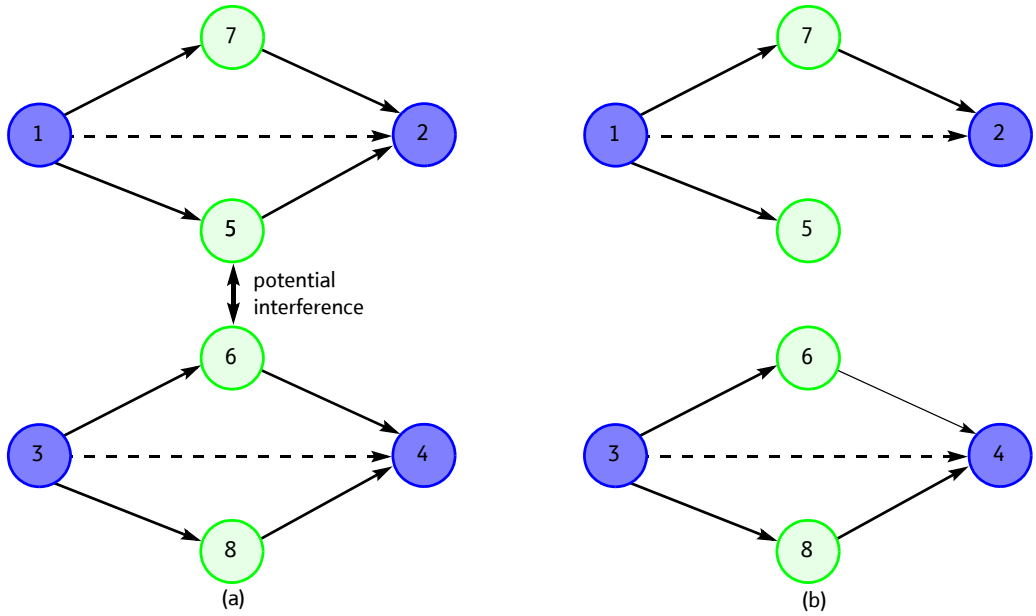

Fig 8 A complex flux routing decision. If node 5 and node 6 both choose to act as relays, the total end-to-end information capacity achievable by both links will be reduced by their interaction. Choosing not to repeat at node 5 improves the joint capacity of the two links. Note that, though all nodes affect each other, we have omitted the 'weaker' paths for clarity.

4 Valenti M C and Correal N: 'Exploiting macrodiversity in dense multihop networks and relay channels', Proc IEEE Wireless Communication and Networking Conference, New Orleans, LA (March 2003) — http://citeseer.ist.psu.edu/565750.html

5 Cover T and Thomas J: 'Elements of information theory', New York, Wiley, pp 389 (1991).

6 Trigeorgis L: 'Real Options: Managerial Flexibility and Strategy in Resource Allocation', MIT Press (1996).

7 Foschini G J and Gans M J: 'On limits of wireless communications in a fading environment when using multiple antennas', Wireless Personal Commun, 6, pp 311—335 (March 1998).

8 Clark D D: 'The design philosophy of the DARPA Internet protocols', Proc SIGCOMM 88, ACM CCR, 18, No 4, pp 106—111 (August 1988) (reprinted in ACM Computer Communications Review, 25, No 1, pp 102—111 (January 1995)).

9 'Realizing the information future: The Internet and Beyond', Computer Science and Technology Board (1994).

10 Weinstein S B and Ebert P M: 'Data transmission by frequency-division multiplexing using the discrete Fourier transform', IEEE Transactions on Communication Technology, COM-19, No 5, (October 1971).

11 Schreiber W F: 'Advanced television systems for terrestrial broadcasting: some problems and some proposed solutions', Proceedings of the IEEE, 83, No 6 (June 1995).

12 Golden G D, Foschini C J, Valenzuela R A and Wolniansky P W: 'Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture', Electronics Letters, 35, No 1 (January 1999).

13 Wolniansky P W, Foschini G, Golden G and Valenzuela R A: 'V-BLAST: an architecture for realising very high data rates over the rich-scattering wireless channel', Proc URSI International Symposium on Signals, Systems, and Electronics, IEEE, New York, NY, pp 295—300 (1998).

14 Laneman J N and Wornell G W: 'Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks', IEEE Trans Inform Theory, 59, No 10, pp 2415—2525 (October 2003) — http://www.nd.edu/~jnl/pubs/it2003.pdf

15 Laneman J N and Wornell G W: 'Distributed spatial diversity techniques for improving mobile *ad hoc* network performance', Internal Digital Signal Processing Group Paper, Massachusetts Institute of Technology (2000) — http://www.nd.edu/~jnl/pubs/atirp2000.pdf

16 Zheng L and Tse D N C: 'Optimal diversity-multiplexing tradeoff in multiple antenna channels', Proc of the 39th Allerton Conference on Communication, Control and Computing, Monticello, IL (October 2001).

17 Gastpar M and Vetterli M: 'On the capacity of wireless networks: the relay case', Proc INFOCOM, New York, NY (June 2002).

18 Reed D P, Saltzer J H and Clark D D: 'Commentaries on active networking and end-to-end arguments', IEEE Network, 12, No 3, pp 66—71 (May/June 1998).

19 Saltzer J H, Reed D P and Clark D D: 'End-to-end arguments in systems design', ACM Transactions in Computer Systems, 2, No 4, pp 277—288 (November 1984).

20 Vyzovitis D: 'An active protocol architecture for collaborative media', SM Thesis, Massachusetts Institute of Technology (June 2002).

21 Shephard T J: 'A channel access scheme for large dense packet radio networks', Proc ACM SIGCOMM'96, San Francisco (1996).

22 Srinivas A and Modiano E: 'Minimum energy disjoint path routing in wireless *ad hoc* networks', MobiCom'03, San Diego, CA (September 2003) — http://lids.mit.edu/~modiano/papers/T9.pdf

23 Bletsas M and Lippman A: 'Efficient collaborative (viral) communication in OFDM-based WLANs', IEEE/ITS Int Sym on Advanced Radio Technologies (ISART 2003), Institute of Standards and Technology, Boulder, Colorado (March 2003).

24 Floyd S and Jacobson V: 'Random early detection gateways for congestion avoidance', IEEE/ACM Transactions on Networking, 1, No 4, pp 397—413 (August 1993).

Andrew Lippman's work at the Media Lab has ranged from wearable computers to global digital television. Currently, he heads the Lab's Viral Communications program and co-directs MIT's inter-disciplinary Communications Futures program. He also directs the Digital Life consortium, which works to create a networked world where communication becomes fully embedded in our daily lives. He has written both technical and lay articles about our digital future and given over 250 presentations on the future of information and its commercial and social impact. He received both his BS and MS in electrical engineering from MIT, and his PhD from the EPFL in Lausanne, Switzerland.

David P Reed's research focuses on designing systems that manage, communicate, and manipulate information shared among people. He is best known for co-developing the Internet design principle known as the 'end-to-end argument', and 'Reed's Law', which describes the economics of group formation in networks. An adjunct professor at the Media Lab, he has been instrumental in developing the Lab's Viral Communications programme. He has worked as a consultant, as well as for Interval, Lotus, and Software Arts, and has been a faculty member in MIT's Department of Electrical Engineering and Computer Science (EECS), working in the Laboratory for Computer Science (LCS). He earned his BS, MS, EE, and PhD degrees in EECS, while conducting research at LCS, and its predecessor, Project MAC.