Essentials of Wireless Mesh Networking

CAMBRIDGE WIRELESS ESSENTIALS SERIES

STEVE METHLEY

CAMBRIDGE www.cambridge.org/9780521876803

This page intentionally left blank

Essentials of Wireless Mesh Networking

Are you involved in implementing wireless mesh networks? As mesh networks move towards large-scale deployment, this highly practical book provides the information and insights you need. The technology is described, potential pitfalls in implementation are identified, clear hints and tips for success are provided, and real-world implementation examples are evaluated. Moreover, an introduction to wireless sensor networks (WSNs) is included. This is an invaluable resource for electrical and communications engineers, software engineers, technology and information strategists in equipment, content and service providers, and spectrum regulators. It is also a useful guide for graduate students in wireless communications and telecommunications.

STEVE METHLEY has over 20 years' experience in telecommunications and data communications innovation, having led teams in the laboratories of British Telecom, Hewlett–Packard and Toshiba. Currently a consultant based in Cambridge, UK, his work includes technology, regulation, business strategy, socio-economic analyses and futurology for a range of clients from start-ups to large global corporations.

The Cambridge Wireless Essentials Series

Series Editors
WILLIAM WEBB *Ofcom, UK*
SUDHIR DIXIT

A series of concise, practical guides for wireless industry professionals.

Martin Cave, Chris Doyle and William Webb *Essentials of Modern Spectrum Management*
Christopher Haslett *Essentials of Radio Wave Propagation*
Stephen Wood and Roberto Aiello *Essentials of UWB*
Christopher Cox *Essentials of UMTS*
Linda Doyle *Essentials of Cognitive Radio*
Steve Methley *Essentials of Wireless Mesh Networking*

Forthcoming
Albert Guiléni Fàbrigas *Essentials of Error Correction for Wireless Communications*

For further information on any of these titles, the series itself and ordering information, see www.cambridge.org/wirelessessentials.

Essentials of Wireless Mesh Networking

Steve Methley

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521876803

First published in print format 2009

ISBN-13 978-0-511-58070-3 eBook (NetLibrary)

ISBN-13 978-0-521-87680-3 Hardback

Contents

Preface

Wireless mesh networking is a hot and growing topic, still in its infancy in some ways, whilst already shown to be capable in others. From a military beginning, mesh networks moved to civilian use and are now being deployed worldwide as both local area networks (LANs) and metropolitan area networks (MANs). However, these deployments are still 'leading edge' and it is not yet clear what the most enduring applications of mesh will be – particularly as the market moves from early adopters towards widespread take up.

Some of the claims for what a mesh network may deliver have been very ambitious to say the least. In this book we investigate such claims versus the real qualities of mesh networks and identify the key time scales and drivers for the challenges involved with making meshes. Throughout the book we attempt to keep mathematics to a minimum. Where an equation is shown, it remains practical to follow the flow of the book without needing to understand the maths fully.

The book takes a very pragmatic but balanced approach to the issues. We are particularly interested in meshes with an external access capability, for example to the Internet. We supply a technical assessment of mesh and multi-hop networking, highlight the attractions, identify the pitfalls, provide clear and concise hints and tips for success – summarised inside the back cover – and finally evaluate some real-world examples of good mesh applications. These include wireless cities, community networking and vehicular ad hoc networks (VANETs). Wireless sensor networks (WSNs) are another important application of mesh techniques with their own unique challenges, and these receive their own chapter.

We conclude that, although some of the claims for what a mesh may deliver have been exaggerated, the real qualities of a mesh network when directed to a suitable application can still make mesh the best approach.

Acknowledgements

The author is very pleased to be able to thank past and present colleagues for many vigorous and entertaining technical discussions. Two former colleagues, Malcolm Crisp and James Newman, deserve special thanks for the many animated brainstorms we had together whilst deliberating the issues in Chapter 4. I was fortunate enough to be able to talk to many experts including Saleem Bhatti who expanded my appreciation of networking issues, especially for sensors. Other experts included Peter Massam who made many useful comments on modulation schemes, Peter Ramsdale who is now sadly missed, Frank Rowsell, Chris Davis, Nee Joo Teh, Stuart Walker, Ahmad Atefi and last but certainly not least William Webb, our book series editor, who provided many very helpful suggestions for the manuscript. Finally, I also belatedly want to thank Peter Cochrane who gave me so much encouragement in my first role all those years ago at BTRL.

Thanks are also due to my wife for her support at all times without exception and to our two sons for having such boundless energy which is a constant source of inspiration.

1 Mesh overview and terminology

The Internet is now firmly part of our everyday life. We perform many common tasks on-line, such as banking, grocery and gift shopping and the purchasing of travel or cinema tickets. Plus we get a growing portion of our entertainment from on-line sources: entertainment and social networking are two of the largest growth areas. We have seen the beginning of basic quality video from, for example, YouTube and the development of social networking sites such as MySpace and FaceBook, which have been enormously popular, especially amongst younger generations of consumers. If we are to continue in this trend of doing more on-line, our need for bandwidth will increase. And in future we might expect to generate appreciable content ourselves, for upload onto the Internet, as well as to continue to download content. But that is not all; our need for Internet availability and quality will also increase.

It would be very convenient if such future Internet access were also wireless, with the near ubiquitous service we are used to from cellular phones. However, building a new network to achieve this, or upgrading an existing network to support this, would mean installing or changing a great deal of infrastructure. What then if a method existed which promised improved Internet access with fewer requirements for new infrastructure? This is widely advertised as the domain of the mesh network.

This chapter begins with a top-level introduction to mesh networking, then looks at how meshes may fit into the larger telecommunications infrastructure, before moving on to classify and explain the basic properties of a mesh. Finally the chapter closes by bringing together the key issues for mesh networking and by linking these issues to the structure of the book. This includes four hypotheses of how meshes might be expected to bring networking benefits. The hypotheses are real, having been taken from a growing mass of mesh-centric literature at the time this book was written. Testing these hypotheses will form a useful basis for investigation and they will be revisited often as we progress through the book.

1.1 What is a mesh?

It is perhaps easiest to begin examining mesh networks by first taking a small step backwards, by reviewing how cellular and wireless local area networks (LANs) work, before highlighting the similarities and differences embodied in the mesh approach.

A cellular network, as its name suggests, consists of many cells of radio coverage, each having a base station near its centre which transmits radio signals over a wide area, for example several kilometres in diameter. The user's device is a small handheld unit of lesser complexity and capability than the base station. Where the coverage of one cell falls off, it is picked up by adjacent cells. In this manner, a large area is covered by a honeycomb of cells. Clearly the advantage is the promise of continuous coverage. This is a major advantage, but there are some downsides: a new network must be planned, plus, ideally, all cell sites should be rolled out simultaneously over the whole coverage area, but this means a large upfront cost for the network operator.

Despite best efforts, some black spots may occur where the user may be shielded from the radio signal by an obstruction. Whilst one way around this would be to install an additional small cell, too many of these would increase the cost of the network infrastructure and hence adversely affect the operator's business model. In reality a compromise is drawn whereby a typical network availability approaches 100%, but does not guarantee to please absolutely all the people, absolutely all the time.

The overarching method used to ensure good coverage is to choose a carrier frequency which propagates well across many terrain types and has good building penetration characteristics. This will jointly maximise coverage and capacity and minimise the number of cells.

It is no coincidence therefore that cellular systems worldwide centre around a very similar band of frequencies. Such spectrum is highly valued and already allocated and assigned in most countries. It is worth noting that cellular systems were designed primarily for voice communications, which are bursty in nature. The amount of spectrum they have is thus not ideally suited to modern multimedia communications, which may involve the streaming of delay sensitive data over a relatively long period, for

example to watch a video. The future evolution of cellular systems is therefore aimed towards better supporting multimedia applications.

Let us next look at the situation for wireless LANs. A wireless LAN, as its name suggests, is mostly about local area coverage. The objective is usually to cover a workplace or a home. Commonly, wireless LANs have a design objective to cover up to a 100 metre radius around a wireless access point on a wired LAN. The access point and base station concept are similar, but the distances under consideration are very different. One immediate benefit is that the type of spectrum required is different, specifically that the propagation characteristics need not be so good for wireless LANs. Higher frequencies are adequate to cover shorter distances, leading to typical wireless LANs operating at higher frequencies than typical cellular systems. Here the spectrum is in less demand. In fact many wireless LANs operate in spectrum which is licence exempt, meaning system costs are lower. But this is a double-edged sword; unlicensed wireless LANs do not have exclusive use of their spectrum as do cellular systems. This leads to the need to design the wireless LAN to be tolerant of interference, which has costs in terms of equipment and system operational efficiency. In the extreme, efficiency may drop to zero if congestion occurs, as it may do in any unlicensed system.

There are at least two more reasons why a wireless LAN may be simpler to implement than a cellular system. In general, there is no concept of cell to cell hand-over in a wireless LAN, nor does the wireless LAN offer much more than a raw data link between users; it is up to additional protocols such as IP and TCP to establish routes and transport the data reliably. It is worth noting that wireless LANs were designed primarily to transmit data rather than voice. However, a design constraint of transporting bursty data was assumed, which was realistic at the time. This means that whilst wireless LANs have good capacity for large amounts of data, they typically do not cope well with modern multimedia communications. In other words, whilst wireless LANs were designed primarily for cases where demand is higher at peak times, such peak times were not expected to be of very long duration. The future evolution of wireless LAN systems is now directed towards better support for multimedia applications, much like the evolution roadmap for cellular systems.

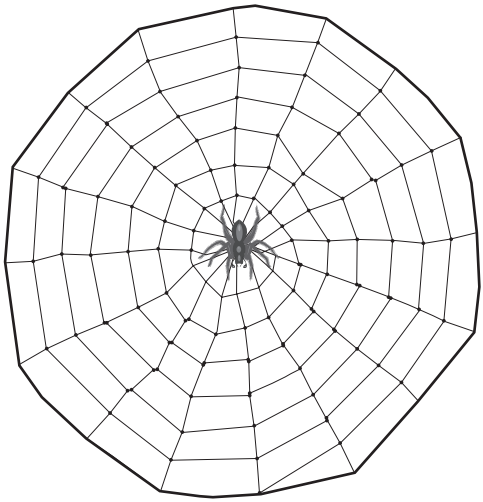


Figure 1.1 A spider's web, an example of a mesh.

Turning, at last, to the mesh network, once again we can infer its prime working attribute from its name. Meshes are all around us in one form or another – think of a spider's web, Figure 1.1, or the grid pattern of streets in a downtown area (and there is an application clue...). Imagine that at each material intersection there is a node. What these examples have in common, for our purposes, is two-fold:

1. there is no main node;
2. it is possible to reach any other node by traversing a number of intermediate nodes.

Immediately we can see that a mesh architecture is quite different from a cellular or wireless LAN architecture. All nodes are equal so there is no centralised control and, therefore, each node must participate in networking as well as be a source or sink of traffic. Rather than a single hop to a base, multi-hopping amongst nodes must be a common capability.

In fact this all brings the promise of great flexibility, particularly when we wish to create a new network, or expand an existing one. By way of example, consider that we wish to network five people in their house and garden, an increasingly common task. Let us try to do this firstly following the access point or cellular principle and secondly by the mesh principle. We will keep the example extremely simple.

Let us say that the range of all wireless units just exceeds the maximum dimension of the house. Then the situation for cell-based and mesh is equivalent; each approach can cover the house. Let us now say one person moves to the far end of the garden which is very long, much larger than the house dimensions. The cell based scheme cannot cope with this; the radios are simply out of range. The mesh based system can cope with one proviso, that a third person moves in between the house and the end of the garden. All units may now communicate as before. The unit at the end of the garden simply multi-hops, using the third person's unit as a relay. It is thus easy to see that multi-hopping can cope with distance, and we should bear in mind that this must mean it can also cope with clutter by hopping around obstructions.

Finally, let us say all the users and their wireless units move to a new house. Unless they take the base station with them, the cellular scheme clearly will not work. On the other hand, the mesh scheme works as before. This is the quality of meshes to work without infrastructure.

Whilst this example may be simplistic, it is also realistic. The principles remain the same when we look at more involved examples of where and how meshes might fit in with existing telecommunications systems, in the future.

1.2 The role of mesh in future networks

In order to understand the wider role of mesh networks it is necessary to place them in the context of an overall communications environment. Like many in this field, in the future we believe this is likely to consist of a wide range of different wireless communications systems, connected to a single core network based around IP packet switching.

We therefore see this future as an evolutionary integration of the cellular approach with the WLAN approach as depicted in Figure 1.2, where proprietary interfaces and protocols have largely been removed.

In the scientific literature, this is sometimes referred to as 'B3G', meaning 'beyond 3G'. Such terms are usually proffered by the existing cellular focus groups, but WLAN parties also have a similar vision – perhaps this should be called 'beyond WiFi'. Thus B3G and BWiFi are the same integrated vision and as such must surely delay or even remove

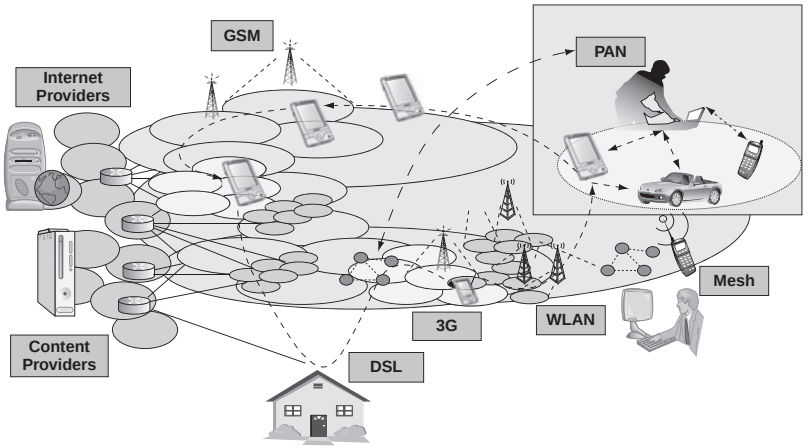


Figure 1.2 Future mobile integrated vision using an IP core.

the need for a wholly new and bespoke fourth generation (4G) network. In other words, we think 4G will look like Figure 1.2. Meshes are expected to provide a complementary access route into the core, alongside WLAN, 3G etc. In other words, mesh will be an additional access technology rather than a replacement. Whilst the foregoing may seem self-evident, there have been some who have thought mesh would be more of a unilateral revolution than a component of evolution.

It is worth looking at some of the typical types of networks presently used to link into the core of the network. The core is used to connect the users to the Internet via the Internet service providers (ISPs). Content providers may be particularly closely aligned with ISPs in some future business models. The core supports access at very many different data rates from different devices, which is a good match for the granularity of service offered by TCP/IP. Example technologies include the following.

- PAN – personal area network. This may for example use Bluetooth and may even be mesh-like in its structure. The speed is currently under 1 Mbps (although 100 Mbps is predicted) and the range is short, e.g. 10 metres. Access points are needed.
- GSM/3G – second and third generation cellular mobile. Data rates will initially be under 2 Mbps with the average around several hundred kbps.

The range is high, on the scale of several kilometres. It relies on a deployed infrastructure of base stations.

- DSL – digital subscriber line. This will typically be a point to point connection running over fixed copper pairs carrying Ethernet or ATM frames. The speed is currently around 8 Mbps, depending on distance and network loading (contention).
- WLAN – wireless local area network. This provides potentially the fastest access towards the core; up to 54 Mbps is available, with 11 Mbps being widespread today. Its range is around 100 metres. Some WLANs enable a choice of whether to use infrastructure or peer-to-peer, but the majority are using an access point infrastructure.
- Mesh – presently only very sparsely deployed in the 'early adopter' market. Its performance capability is to be investigated within this book. It is potentially high speed, with good coverage and with no need of infrastructure.

One key point from Figure 1.2 is that there are two ways at present to obtain relatively high speed wireless access to the core – via cellular and via WLAN. They have different performances in terms of quality, speed and range, but also critically in terms of cost. The cost benefits to the user from having a dual mode handset for attachment via a WLAN at close range and via cellular at longer range can be very attractive. However, such 'fixed-mobile convergence' presents a real challenge to operators who may well have to change their business model and adapt their technology. But mesh is relatively new to the vision and may enable radically fresh approaches to the challenge of high speed wireless access for the future.

1.3 How do meshes work?

Before we investigate mesh in detail we need to understand more closely what a mesh network actually is. This section discusses and defines the nomenclature and methodology used for mesh networking. It continues by clarifying the meanings of the terms mesh and ad hoc.

1.3.1 Forms of mesh

There are broadly three basic mesh type architectures, to which we shall refer throughout the discussions within this book. The three types are introduced here and described in detail in Section 3.2.2.

- Pure mesh – all traffic in a pure mesh is intra-mesh, i.e. the mesh is isolated. All traffic is relayed via a single node type, i.e. the user node.
- Hybrid mesh – a pure mesh with a hierarchy of node types, in order to improve efficiency via the introduction of backbone routes. In other words there is a further network overlaying the mesh, which exclusively takes longer distance traffic. In wireless meshes, such a hierarchy of routing layers may be created simply by using additional dedicated radio channels, or bands.
- Access mesh – a mesh with a hierarchy of node types, as above, but additionally where considerable traffic is extra-mesh. In other words, the overlay routing network also has gateways to other, external networks such as the Internet.

We shall see later that traffic flow and hence the most appropriate mesh architecture depends on whether the content to be accessed resides inside or outside the mesh. In other words, the type of mesh required in a given situation is driven by the needs of the user and the application. It is worth noting from the outset that most of the early published research was funded by the military to look at pure meshes and does not always translate well to public telecommunications requirements, which generally require an access mesh.

1.3.2 Planned versus ad hoc

A second distinction is concerned with the design rationale of the network. A planned network such as cellular has a predetermined maximum level of users and protected cells in which they may operate. The benefit of this is that interference is limited and it flows from this that guarantees of the quality of service delivery can be made. The downside is that infrastructure is needed; in other words the operator must first make a provision for everywhere the user requires a service.

An unplanned network allows ad hoc connections. 'Ad hoc' literally means 'for this purpose', i.e. a temporary arrangement as and when required. The benefit is that no infrastructure is needed and the users themselves may extend the area of coverage – although we shall see later that must involve a performance trade-off, if all else remains equal. The downside is that without planning, there is no control over interference effects from other users. Hence an application's performance is beyond the control of any one party, therefore absolute guarantees over the quality of the delivered service cannot be made.

1.3.3 Characteristics of an ad hoc pure mesh network

Table 1.1 briefly summarises the characteristics of a pure, ad hoc mesh network. These are the characteristics which will be examined in later chapters.

1.3.4 Characteristics of an access mesh

Moving up in complexity, a hybrid mesh is similar to the type of mesh described in Table 1.1, except that some nodes will additionally have independent network connections, enabling them to connect to an internal routing backbone. One logical step beyond this would be to add the ability for some nodes to be gateways to external networks. Of course this turns out to be exactly the configuration which is needed to provide Internet access to a group of users. For this reason, we refer to this mesh type as an access mesh. The access mesh is the most important mesh type for the rest of this book.

1.3.5 Meshing versus multi-hopping

We have already seen a simple example of home wireless communications, where multi-hopping was useful to increase coverage by increasing distance or hopping around obstacles. This begs the question of the difference between meshing and multi-hopping.

Table 1.1 *Characteristics of ad hoc mesh networks*

Ad hoc	Unplanned. Therefore the coverage and interference environment is uncontrolled, which is the exact opposite of the typical cellular case. This directly raises quality of service issues.
No separate infrastructure	All functionality is provided and carried out within the mesh. This includes any power control, routing, billing, management, security etc. There is no centralised equivalent to the base station or security and authentication centre of, for example, cellular networks. (Not true for an access mesh, see below.)
Mobility	Nodes are free to move and even disappear. The network interconnections may thus be very dynamic.
Wireless	In order to support mobility or avoid infrastructure, wireless operation is required. This could mean radio or optical, but this book concentrates on radio. Radio links are lower quality than wired links, packet loss in radio is 'normal', whereas on wired connections, loss is equated to congestion. Transport protocols (as developed for wired networks) may thus have the 'wrong' reaction when used on radio networks.
Relay	All nodes may be required to relay information for other nodes. This will lessen the bandwidth available to each node user.
Routing	All nodes will be required to participate in a routing protocol. This may be either proactively by maintaining up to date tables or reactively by creating routes on demand (which may also be cached in tables). Routing creates an overhead, which will depend on the protocol, the traffic and the mobility of the nodes.
Multi-hop	A corollary of relay and routing, multi-hop is an enabler of coverage, especially in the cluttered environment.
Inhomogeneity	Not all nodes need be equal, beyond the subset of capabilities needed for basic mesh operation. Some nodes may have additional network connections (external connections in the access mesh case, see below).

In fact the terms are not universally well defined and mesh is often used generically. Where it matters, in this book we will treat the difference as follows. Meshing may simply be thought of as multi-hopping with active route diversity. To use metaphors, a multi-hop network may be thought of as nodes within a tree and branch structure, whereas a mesh is more like nodes within a spider's web. In other words, in a mesh a given traffic flow may be split over two or more routes to the destination, whereas in a multi-hop network there is a single routing at any point in time. Nonetheless the multi-hop network preserves the ability to route around obstacles. This common, important property of hopping around obstructions will surface many times within this book and is perhaps the most enduring benefit offered by both mesh and multi-hop networks.

Looking a little deeper, it may be realised that the difference between meshing and multi-hopping can become quite blurred. If a multi-hop network may reconfigure another route to the destination, is that not a mesh in effect? One remaining difference which may matter becomes the fact that meshes may have two or more routes active simultaneously – but that brings problems along with benefits (problems may occur at the TCP level, which, in practical implementations, does not suffer reordered packets gladly).

However, a full mesh has redundant routing out to the network edge – multi-hopping tree and branch does not do this, neither does the mesh-star approach of ZigBee, which has a mesh core of full function nodes, whilst edge nodes are reduced function, not being redundantly routed. This is a cost-benefit trade-off; edge nodes may be designed to be less complex since they do not have to participate in routing. The drawbacks include the facts that redundancy and hence reliability is reduced, and the ability to extend the network beyond the edge, in an ad hoc manner, is lost.

Note that a fully meshed network is not necessarily a fully connected network, in the strict sense. The fully connected network has a single hop to all other nodes. A full mesh has single or multiple hops to all other nodes and redundant route potential at all nodes.

Where its detail is already familiar to the reader, it may helpful to recall that the Internet is a mesh. However, in the absence of load balancing or

link aggregation approaches, it often works simply as a per-packet or per-flow, dynamically reconfigured, multi-hop network.

Returning to the access mesh, this always tends towards multi-hopping since the presumption is often that the access node is the gateway for most traffic, creating a natural tree and branch structure.

1.4 Key mesh issues and the structure of this book

This chapter now moves on to list the key issues for mesh networking. The issues are briefly introduced here and each will be subsequently followed up in the succeeding chapters of this book. It is convenient to introduce the issues via a series of questions one may want to ask about meshes. In fact, the questions are real, having previously been gathered by interviewing networking industry professionals in 2005.

The first question is simply 'What are meshes good for?' The answers are best given by application examples, which include

- cellular or WLAN hot-spot multi-hopping,
- community networking,
- home and office indoor networking,
- micro base station backhaul,
- vehicular ad hoc networks (VANETs), and
- wireless sensor networks (WSNs).

These six application examples are described at the beginning of Chapter 2, with deeper consideration of the fundamental technology following in Chapter 3.

We found that the next questions typically asked about meshes included one or all of the following.

- Do meshes improve spectral efficiency ?
- Do meshes self-generate capacity ?
- Do directional antennas help a mesh?
- Can meshes improve the overall utilisation of spectrum?

These four questions come up so commonly, and so commonly together, that we shall spend quite some time looking at them. In fact these

questions are so core to understanding meshes that we shall elevate them all to the level of hypotheses in this book. We can then test them formally in Chapter 4, to determine whether they are true or false, by examining all the available evidence.

Next, having satisfied ourselves as to the validity or otherwise of the four hypotheses, we shall begin to look more closely at meshes in practical deployments. It transpires that one question a seasoned network operator would be right to ask is 'Are meshes more susceptible to interference?' Here the answer really is uncertain, since it depends on the detail of implementation. This is because if the mesh does encounter interference, it may re-route to a better path. But unfortunately this re-routing, especially if it occurs frequently, may upset the transport layer of the communications protocol, potentially leading to overall worse performance. We look at mesh susceptibility in Chapter 5.

The network operator may continue to ask 'Are there any unusual aspects of mesh networking?' We shall see the answer is indeed 'Yes'. For example, in a mobile mesh users may realise their battery is constantly being drained even if they are not actively using it. Looking back to the house coverage example earlier in this chapter, this would apply to the third person's node, which is relaying traffic to/from the end of the garden. This user might decide to turn off their node, with undesirable consequences for the whole mesh connectivity. A related point, but this time from the point of view of an operator, is the realisation that user connectivity is no longer under the operator's sole control. We shall see that there are some steps which can be taken to mitigate this effect. This book tackles the subject by first looking at mesh routing mechanisms in Section 5.3 as a foundation for quality of service issues in Chapter 6.

One question the reader may well have by now is 'Does this book offer any hints and tips for a good mesh deployment?' After all the main technical arguments have been developed, we include Chapter 7 which reviews and collects the common pitfalls to be avoided when deploying mesh networks. Inside the back cover of the book, we list a concise summary of hints and tips.

The six application examples given above are also revisited near the end of the book in Chapters 8 and 9, so that all the theory developed in the

book can be seen in context. Wireless sensor networks may use mesh but are quite distinct from telecommunication networks, so they receive their own place in Chapter 10.

However, to begin, we need to develop a better understanding of why meshes are perceived to be so attractive, which is the purpose of the next two chapters. Chapter 2 first takes an application level view, whilst Chapter 3 looks at the supporting technical fundamentals of meshes.

2 Attractive mesh attributes and applications

In the preceding chapter we showed that meshes were good for extending coverage beyond the existing network edge, without requiring additional infrastructure – a sufficient number of user nodes, in the right places, was all that was required. We also implied that this meant that obstacles to propagation such as buildings in the line-of-sight might be less of a problem, given that a mesh could hop around them in a way which cellular systems cannot. We even dropped a small application hint that the structure of a mesh can be quite similar to a grid of downtown city streets. In this chapter we look more closely at linking a number of useful application scenarios with the relevant attributes of a mesh.

We now propose that there are, at heart, only two worthwhile motivations for mesh networks. These are

1. coverage improvement,
2. lack of infrastructure.

All successful examples of meshing or multi-hopping known to us embody one or both of these core mesh attributes.

To support this conclusion, we now spend some time considering application scenarios. Overall, from a technology standpoint, we found it hard to envisage any new services which *only a mesh* could support, although vehicle ad hoc networks and wireless sensor networks are probably the closest – but even here a mesh is the best rather than the only solution. Rather it seems more likely that a mesh would contribute by delivering services in a new way. Six suitable application areas are identified below where mesh adoption is thought to be most likely. In hindsight, it is easy to see that all six applications are based on a mesh network's valuable attributes of coverage and/or reduced reliance on infrastructure.

Following these application examples, this chapter takes a closer look at the coverage attribute of meshes.

Table 2.1 *Our six example applications for mesh*

1	Cellular or WLAN hotspot multi-hopping
2	Community networking
3	Home and office indoor networking
4	Micro base station backhaul
5	Vehicular ad hoc networks (VANETs)
6	Wireless sensor networks (WSNs)

2.1 Example applications for mesh

Our mesh application examples are described below. Based on our definitions from Chapter 1, all are access meshes rather than pure meshes, since they all provide external network connections. All but one are user-side meshes; the remaining example is a network-side mesh of base stations. To clarify, the terms user-side and network-side are used to describe whether the user nodes or the access points form a mesh. A mesh of user nodes is perhaps what most people think of as a mesh, although a mesh of access points may well be attractive from an operator's perspective. They are not exclusive, for example there could be a hierarchy consisting of an access point mesh sitting above a user node mesh.

Our list of six applications is shown in Table 2.1. We now explain each of these in turn. Additional detail on selected applications will appear in Chapter 8.

2.1.1 Cellular or WLAN hotspot multi-hopping

In Figure 2.1, the base station is shown at the top of the large, central building. To the right is another building and to the left is foliage. These two obstructions would normally create shadowing of the base station signal, leading to poor signal strength. In order to improve the situation for the users at the extreme right and left of the figure, multi-hopping amongst all user nodes is enabled. Each path to the users at extreme right and left comprises three hops, which give a better signal strength than is normally available in the shadow of the base station.

A development of this approach, often used in 'wireless cities', is to site mesh nodes at street intersections of a downtown area. The grid of

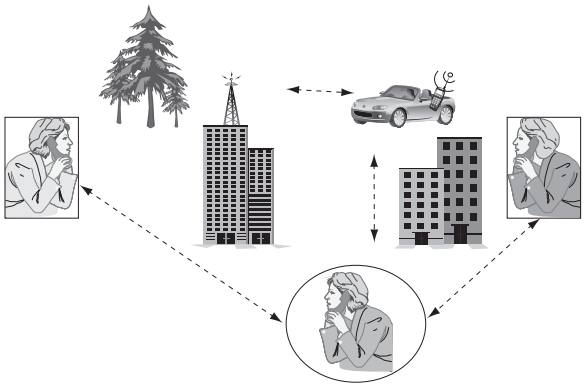


Figure 2.1 Cellular multi-hopping.

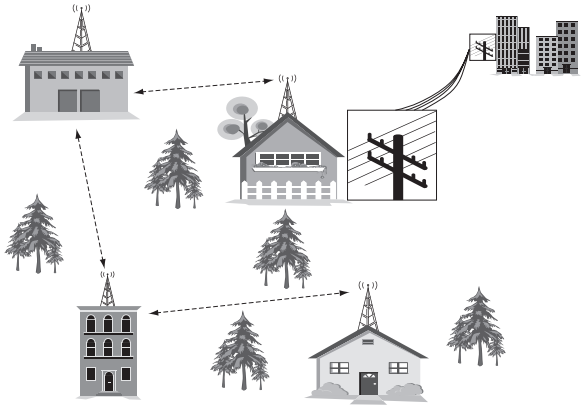


Figure 2.2 Community networks.

mesh links thus mimics the grid of streets. (This is the application hint we dropped earlier.) The benefit in each case is the improved coverage offered by the mesh.

2.1.2 Community networking

In Figure 2.2, the city at top right has broadband. But broadband is not distributed around the rural community by the city provider. If, however, one house does have a broadband feed, then this link may be shared out by a mesh network which may include all the other houses. The broadband link is shown as DSL over phone lines in the figure, but could equally well be satellite or any other substitute. Also, although the figure shows external antennas, we must bear in mind that ease of installation is likely to be key to this business model. Therefore, it is most likely that simple,

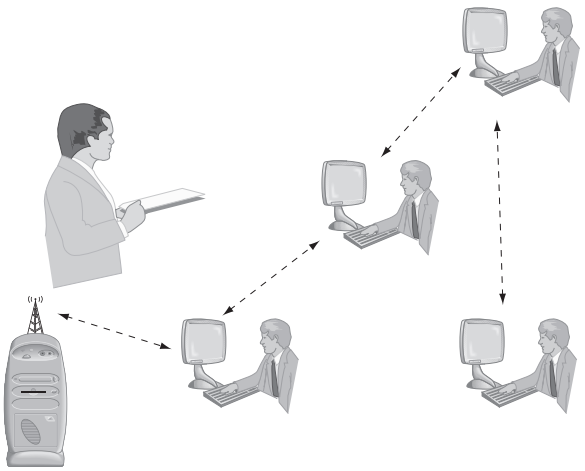


Figure 2.3 Home, office or lecture hall networks.

indoor omni-directional nodes will be preferred, in order to facilitate self-installation by the untrained broadband customer.

The network may be grown beyond its edge simply by adding more nodes. Connections to new houses within the mesh are also simply achieved by adding them as internal nodes. No infrastructure is required.

2.1.3 Home, office or college indoor networking

Quick set-up and tear-down of a service is also enabled by a wireless mesh. This is useful for temporary deployments of connectivity, which may be found in a lecture hall, for example. Figure 2.3 shows students who are connected to a mesh which includes their lecturer. As this group of students leave at the end of the lecture, their connections are no longer needed. However, the next group of students to enter the lecture hall may join up and create a new mesh with their own lecturer, for only as long as it is required.

Turning to wireless home and office networking, this is already available via current technology, but some applications would fit a mesh especially well. This is because there may be relatively small, closed user groups whose traffic flows could be purely peer to peer. In other words, here the mesh would be attractive both as a pure mesh for internal communications and as an access mesh for Internet connections. The difference is perhaps subtle, but where intense local traffic is expected, such as streaming HDTV,

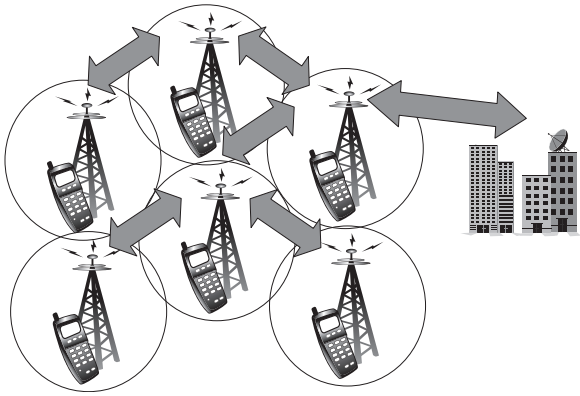


Figure 2.4 Backhaul of micro base stations.

the mesh option would restrict the traffic to the minimum number of user nodes, whereas the traditional WLAN system would have to route everything through the access point, potentially causing unnecessary congestion.

2.1.4 Micro base station backhaul

Figure 2.4 shows that a mesh may be applied to the network side, i.e. the access points, or in this case the micro base stations. In fact there is no reason why any base station from full size to femto-cells could not be meshed. On the whole though, with cellular schemes, the opportunity lies with new deployments, which in developed countries is likely to be newer, smaller cells. On the other hand, the micro base stations could equally well be WiFi access points, as used in wireless cities, which is a potentially larger market in developed countries.

The figure shows that only one pico cell site is connected to the upstream network; this connection is then shared out to all the other sites via a mesh. This is actually quite similar to the community networking example, but using meshing on the network side instead of the user side.

The advantage is that no infrastructure is necessary to roll out the new cells.

2.1.5 Vehicular ad hoc networks (VANETs)

Safety on our road systems could be improved. Many countries are looking to aid drivers to avoid known and common hazards. One way of doing this

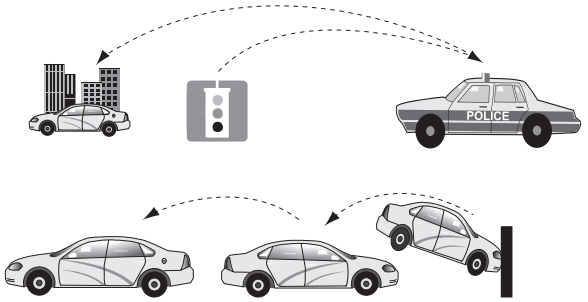


Figure 2.5 Examples of vehicular ad hoc networks.

is by enabling cars to communicate with each other and with the roadside infrastructure. At the top of Figure 2.5, we can see that a car has jumped a red light. This information is passed from the red light to a nearby enforcement vehicle, who in turn signals to the offending driver to pull over. In extreme cases, perhaps the offending vehicle could be stopped remotely.

At the bottom of Figure 2.5, we can see a second example. A car has hit a hazard in the road, maybe a fallen tree. This car has immediately signalled to the car behind, to warn of the hazard. This message is passed along the line of following cars. It may also be passed along the opposite carriageway in order to alert the nearest emergency services.

Both examples rely on ad hoc communications between cars and to the highway infrastructure. No installed communications infrastructure is required. The links are direct, ensuring good coverage even in urban shadowing situations.

2.1.6 Wireless sensor networks (WSNs)

There are many examples of where WSNs could be used. One such example is environmental monitoring, as shown in Figure 2.6. Sensors are scattered within industrial environments, office buildings or along the length of pipelines. These sensor networks self configure such that a fully meshed network is created by the relaying of messages over multiple hops. Each sensor network then connects to a gateway, which allows communication to the Internet. The sensor data can then be read from anywhere on the Internet. Action can be taken to circumvent developing environmental problems in industry, or simply to adjust the heating and lighting in an

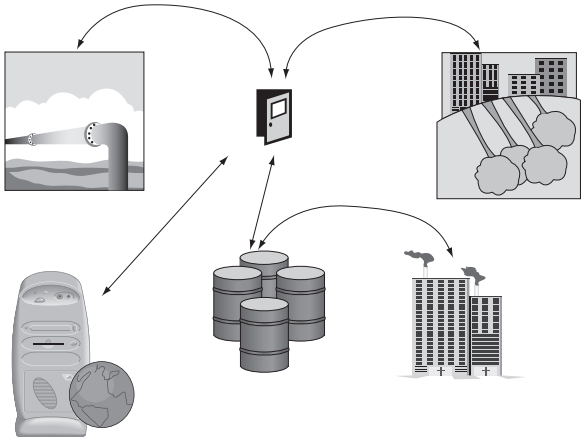


Figure 2.6 Environmental monitoring by WSNs and gateway.

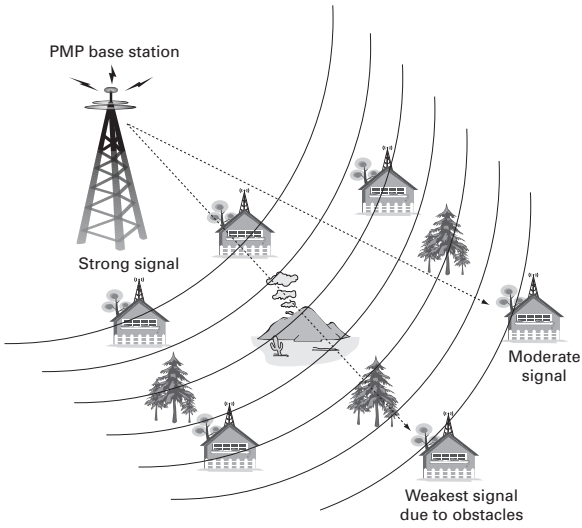


Figure 2.7 High power PMP approach in the cluttered environment.

office block, for best comfort and efficiency. Such sensor networks have no infrastructure by design. This brings a major cost saving.

2.2 The coverage attribute

The coverage attributes of a mesh are quite different to those of a point to multi-point (PMP) or cellular solution. The cellular approach uses high power to cope with the large range of path loss, as shown in Figure 2.7.

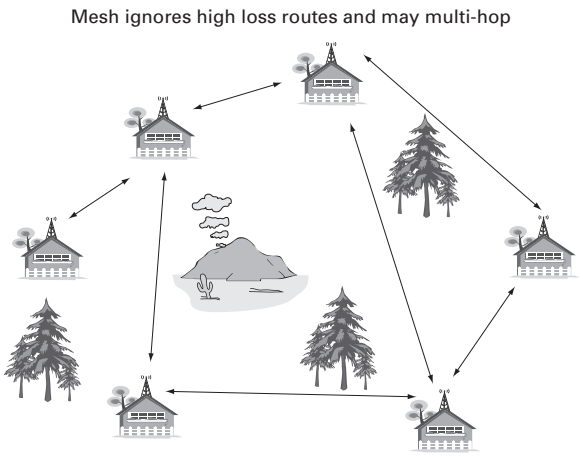


Figure 2.8 Low power mesh in the cluttered environment.

The strongest signal corresponds to a direct, short, line-of-sight link and the weakest signal corresponds to a long link which has partially obstructed line-of-sight. The base station is a relatively large structure, maximising any local terrain height advantage, such as a hill, which may be available. Where shadowing and obstructions occur, it is the combination of antenna height and signal strength which is used to overcome them. Whilst the design of cell sites is a well developed science with many sophisticated modelling tools available, the base station technique itself might be said to require a reasonable component of brute force.

In contrast, the mesh uses low power and 'skips' intelligently around obstacles, as shown in Figure 2.8. The success of the mesh thus does rely on having 'clever' software in the routing layers; this subject is covered in Chapter 6.

In the figure, some nodes are reachable only via two or more hops. The benefit is that high power transmitters are not needed since the distances are small and clear line-of-sight. The need for antenna elevation is also reduced for the same reason. Deploying a mesh can thereby use smaller, less conspicuous nodes than a cellular base station.

Let us now take a closer look at the path losses in the case of cellular and mesh. In each case we can consider a log-normal path loss following the well known formula:

$$\text{path loss}(d) = \text{path loss}(d_0) + 10n \log(d/d_0) + X_\sigma,$$

where d represents the transmitter–receiver spacing, n represents the severity of the distance-related loss and X_σ represents the variance of the normal distribution around this average, caused by a distribution of cluttered environment loss. The formula is not exact, but is usually derived from a curve-fit to a cloud of experimental measurements. But, at the moment, we are not interested in exact numbers, just that a path loss may be simply approximated, on average, to a distance-based loss and a clutter-based loss variance.

Now, referring back to Figures 2.7 and 2.8, it may be deduced that for the PMP example, a large loss variance is bad, since the system must be designed for the worst case, hence a high transmitter power is needed. But this high power will raise system costs and affect adjacent base stations negatively. This is an undesirable situation and might require the introduction of extra, lower power cells to avoid the problem, but these would come at a cost.

In complete contrast, for the mesh system, a large variance does not present a problem – if the node has the necessary software intelligence to pick the best link. This is because the best link may 'skip' around the obstacles over a distribution of wireless paths with much less loss and loss variance. In the mesh case, therefore, relatively low transmitter powers are needed to cope with the lower power budget, and a smaller interference footprint results. Note however that, firstly, some connections can be made only via multiple hops, which increases latency and that, secondly, a multiplicity of transmitters is now involved. We shall address these aspects later in the book.

In summary, we might reasonably expect a well-configured mesh to provide the greatest advantages compared to conventional networks in a cluttered environment.

To reinforce this point of view, let us look more closely at two extreme propagation environments, the open rural and the dense urban.

2.2.1 Rural environment with open propagation

This scenario is very simple; all links have low loss and low variance. There is neither advantage nor disadvantage in running a mesh system from a propagation point of view.

2.2.2 Urban environment with dense obstacles

This scenario is also very simple; the majority of cellular links would have a high average loss and a large loss variance. One extreme example would be the so-called Manhattan model (see the Appendix); this is a grid of US-style city streets. The potential for selective shadowing and thus loss variance is very high. The proposed advantage of a mesh here is the ability to hop around obstacles. As we have already shown, this has the potential to be more power efficient than the cellular case which would need to have sufficient power to cope with the worst case shadowing. Of course it should be mentioned that there is a middle ground – if they could be deployed, cellular pico cells would be an alternative step to offering benefits similar to what meshes could offer, but probably not at the same deployed cost point.

We can find evidence to support the coverage assertions discussed above in an experiment to compare unaided signal transmission with meshing around obstacles using a 'forwarder' node [1], see Figure 2.9. The aim was to test whether there is an advantage in throughput.

In Figure 2.9, the situation being modelled is as drawn in the diagram within the graph. The remote mobile terminal (RMT) desires to communicate with the access point (AP). A forwarding mobile terminal (FMT) may or may not be available; it is an experimental variable.

Communication from the AP to the RMT, which is sited around a corner, may be done in one of two ways:

- as a single hop, distance d, 'directly';
- as a two-hop, twice distance $d/\sqrt{2}$, 'hopping' around the corner via the FMT.

The experiment shows the results of each case, including the effect of adapting the modulation scheme.

In Figure 2.9, the potential gains to be had in all the two-hop cases are shown within the total shaded area of the graph which is made up of the gains from each two-hop case listed in the graph legend. The shaded area is large, indicating that the benefits from allowing a two-hop route are large. The small, white area enclosed between the axes and the shaded area is the single-hop performance, which is very much inferior, as might be expected.

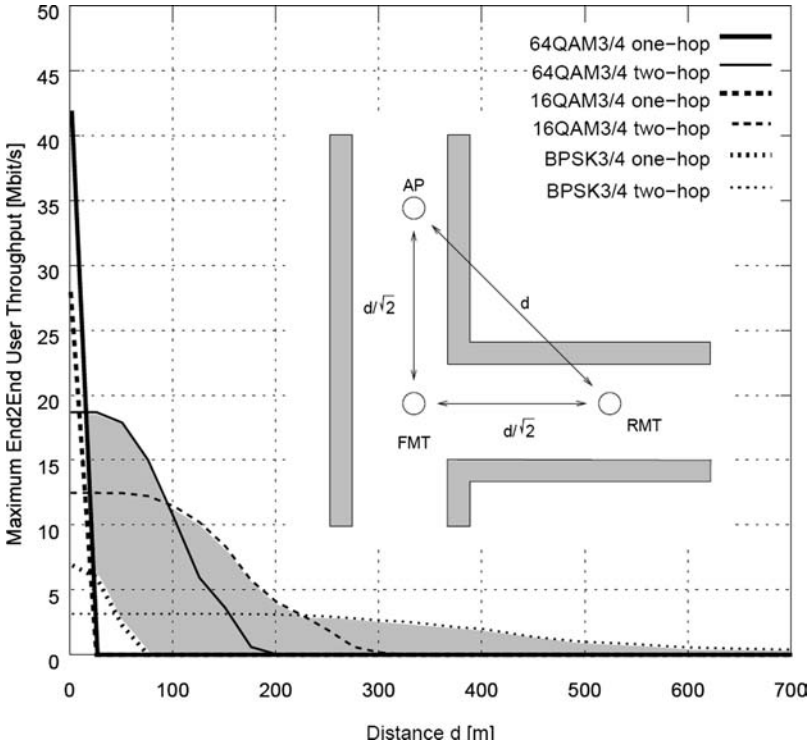


Figure 2.9 Use of a 'forwarder' to 'skip' around obstacles as in a mesh network, used with permission from Esseling *et al*. [1].

The results of Figure 2.9 related to extending the infrastructure mode of a particular short-range technology called HiperLAN/2 in the 5 GHz band. However, the implications of this simulation are of a general nature and thus may be applied to other hopping situations, such as a mesh network. In this book we employ the results of this and similar experiments as evidence that multi-hopping is a key enabler of coverage in a mesh network.

2.2.3 Extension to a mixed environment

Clearly in the real world, scenarios are in between the extremes outlined above. The best solution thus depends upon the detail of the situation. In general, however, meshes should show benefits where the problem faced

is one of coverage due to a cluttered or shadowed environment. Most often this would be expected to be a dense urban area.

It should not be assumed from the foregoing that the choice of mesh versus cellular is strictly one of 'either-or'. Recall that mesh can complement cellular by improving coverage within a cell, as already shown in Figure 2.1.

2.3 Summary

This chapter has shown that cellular and mesh systems operate on quite different principles. Although it undeniably works well, the cellular system is something of a brute force approach, relatively speaking, whilst the mesh is perhaps a more elegant method, which employs more intelligence in order to achieve coverage. We cited evidence that hopping around obstacles can improve system throughput.

For the sake of completeness and balance, we feel bound to point out that one tacit assumption we have made when discussing mesh in this chapter is that the mesh node density is high enough to maintain mesh connectivity. Ultimately, Chapter 6 will look at how this might be promoted and what happens when it is not the case.

But first we must complete building up our basic knowledge of mesh, hence Chapter 3 next looks at how the attractive features of mesh are fundamentally realised. It will begin to become clear that, like everything else, it is not possible to achieve something for nothing. In other words there are compromises to be understood and accepted, as we shall see once we are ready to delve into the detail which begins in Chapter 4.

Reference

1. Esseling N., Weiss E., Kramling A., Zirwas W., A multi hop concept for HiperLAN/2: capacity and interference, *Proc. European Wireless 2002*, Florence, Italy, February 2002, Vol. **1**, pp. 1–7.

3 Fundamentals of mesh technology

3.1 Overview

We listed the characteristics of an ad hoc mesh in Chapter 3, Table 1.1, and we built upon this to create the access mesh concept. But we have not so far attempted to offer any detailed explanation of the key mesh characteristics. The function of this chapter is to examine these fundamentals as a final foundation before Chapter 4, where we begin the detailed testing of the four key hypotheses of mesh performance which we introduced at the end of Chapter 1.

A logical way to address the fundamentals is to consider, in turn, each layer of a generic communications protocol stack as shown in Figure 3.1.

At the bottom of the stack is the physical layer, or PHY. This consists of the parts which directly concern the air interface, for example the antennas and transceiver electronics. By implication this also includes detail design elements, such as the choice of modulation scheme and transmit power.

But it does not include the method by which access to the air interface is determined – this is the job of the medium access control layer, or simply MAC. This, for example, will include schemes to allow multiple users to share the medium in some more or less fair fashion, such as the random collision avoidance approaches used in 802.11 or the structured time and frequency division multiplexing as used in GSM.

To enable nodes to find and communicate with each other, some sort of addressing scheme is required; this is contained in the routing layer. An example is the increasingly ubiquitous Internet protocol. This is so popular that its latest revision, IPv6, includes hugely expanded addressing space in response to a global demand which says that every conceivable device might need an IP address. IP itself is the addressing scheme, but there must also exist a routing protocol. On early Internet routers this was RIP (routing information protocol). Although RIP is still used, it has been augmented by more complex routing protocols as the Internet has

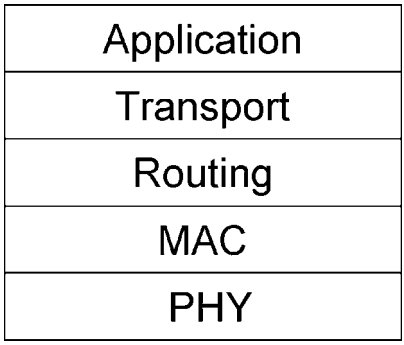


Figure 3.1 A generic protocol stack.

grown and grown. But, relative to the static Internet, a routing protocol for a mobile mesh may have to work very hard indeed, and this means different solutions may be required, as we shall see.

Next up the stack is the transport layer, which is responsible for arranging the way in which packets of data are sent over the link and what, if anything, is done when packets do not arrive at their destination. The most widely known example here is the transmission control protocol, TCP, which includes a congestion control mechanism and a packet delivery verification mechanism. TCP/IP for transport and addressing is expected to become a ubiquitous component in communications systems, as we outlined in Chapter 1. We will assume the TCP/IP stack throughout the book, but note that this does not necessarily mean we have to assume regular Internet routing protocols like RIP etc.

Finally, in our simplified stack, the application layer covers everything else all the way up to a user interface, such as a keyboard and screen. In other words, it supports whatever task the node user wishes to perform. In general, we are not interested in the precise detail of the application layer, but we are interested in the interaction of the applications with the transport protocol. For this reason we will consider transport and applications together.

Bringing all this together, an example stack would be an 802.11 radio (PHY) and 802.11 ad hoc MAC carrying TCP/IP to allow the transport of email application data.

This chapter begins at the physical layer by looking at the two ways a mesh may be made. Next it builds up, via a series of diagrams, how a pure mesh architecture may be developed into an access mesh architecture. This is guided by consideration of the expected application traffic flows. When considering the MAC, various approaches are introduced, noting that a MAC requiring centralised co-ordination is unlikely to be best suited to an ad hoc network. The mobility aspects of a mesh are key when considering routing. This immediately splits routing protocol approaches into two camps. Finally we consider the attributes of applications, for example real time, non-real time, specifically as they relate to transport protocols.

3.2 Physical layer

The most basic question to ask is how are meshes made. In fact there are two ways to accomplish this.

3.2.1 Physical versus logical meshes

A mesh may be made physically or logically. The distinction is important from an interference footprint point of view, which differs between the two cases. This in itself will become important when we begin to discuss mesh capacity in Chapter 4.

Physical meshes are those dictated by physical level constraints, for example a constraint caused by design such as directional antennas or perhaps a constraint of the signal path caused by terrain or medium, as we saw in Figure 2.8. An example is the wired Internet which is clearly a perfect physical mesh in that transmitting on one link does not interfere with any others. On the other hand, a logical mesh is configured above the physical layer, where there is not necessarily any physical constraint on a station's neighbours imposed by the system or environment. Omni-directional antennas in an open field could be connected as a logical mesh; of course this is done at the MAC level, since it cannot be done at the PHY level.

Figures 3.2 and 3.3 show a physical and a logical mesh respectively. Figure 3.2 illustrates antenna pointing directions, although the links themselves are bidirectional.

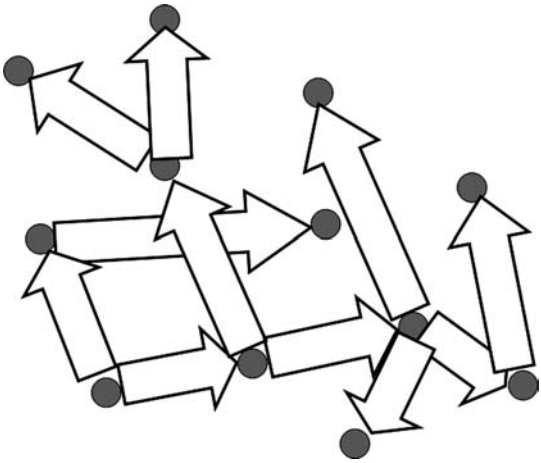


Figure 3.2 A physically created mesh.

Our six applications of interest shown in Table 2.1 of the previous chapter mostly involve either a degree of mobility or regular reconfigurability, which is much easier to achieve if omni-directional antennas are used and the mesh is therefore created logically. This is especially true of the newest and most exciting sectors of mesh networking – VANETs (vehicular ad hoc networks) and WSNs (wireless sensor networks). Discussions of capacity in Chapter 4 will thus assume omni-directionality as a basis, although the question of directional antenna possibilities will be included.

3.2.2 Intra-mesh and extra-mesh traffic flows

Historically, meshes were designed by the military to cope with situations where no infrastructure was available. No node could be allowed to be more important to any other, lest it be lost – whereupon system operation had to be unaffected. Therefore the concept of a centralised controller or server was seldom favoured. But our mesh applications are simply not like this; the traffic and applications are different, especially if our aim is

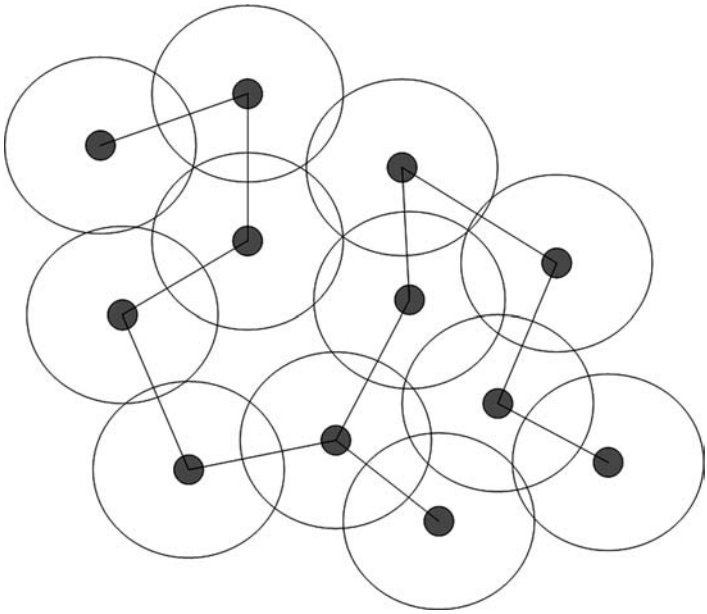


Figure 3.3 A logically created mesh.

to access the Internet, although we do wish to retain the element of ad hoc operation.

Candidate network architectures for mesh networks stem from considering traffic flows which depend on whether the content to be accessed is outside or inside the mesh network. Thus two fundamental traffic flows exist, defined here as 'extra-mesh' and 'intra-mesh'. Of course these are not exclusive – a real application will probably have a mixture of both.

First let us consider applications where there is no need to access external (or central) resources, for example team communication. For such applications traffic flow can be contained within the mesh with users communicating with each other either directly if range permits or by hopping via peers. In current markets such intra-mesh traffic tends to be limited to closed user groups such as

- file-sharing on-campus, intra-company, local community, etc., and
- emergency and disaster relief activities which cannot depend on infrastructure.

But these are not widely seen as majority applications. Because of this we conclude that the vast majority of traffic which is of commercial significance to users today would be classed as extra-mesh traffic, with an interface to the external networks including the Internet. It should be noted that even in the case of wholly intra-mesh data traffic, there is likely to be associated network management and/or billing traffic which must flow to a network management centre and so becomes extra-mesh. Thus, in practice, any mesh is very likely to require a hybrid architecture supporting both intra-mesh and extra-mesh traffic types.

Intra-mesh and extra-mesh traffic is next examined in more detail, with specific regard to traffic flow and potential traffic concentration and congestion.

Intra-mesh traffic architectures

In this case the sources and sinks of all traffic are within the mesh network – i.e. there is no requirement for connection to an external network, such as for connection to the Internet, control centre, etc. For such intra-mesh traffic the mesh may consist entirely of subscriber nodes, as illustrated in Figure 3.4. In this case, traffic concentrations will occur only where users are concentrated, for example around business centres, communities, retail centres, etc.

It therefore follows that the integrity and coverage of the network can be enhanced by the addition of fixed nodes added to assist with local traffic concentrations, as illustrated in Figure 3.5. The fixed nodes are new, additional nodes, not replacements of existing nodes.

These fixed nodes might be added for several reasons.

- To enhance connectivity or coverage when user nodes are sparsely distributed. This may be the case during early roll-out of the service when there are inadequate customer numbers to provide sufficient connectivity of the mesh. In this context they are often referred to as ‘seed nodes’.

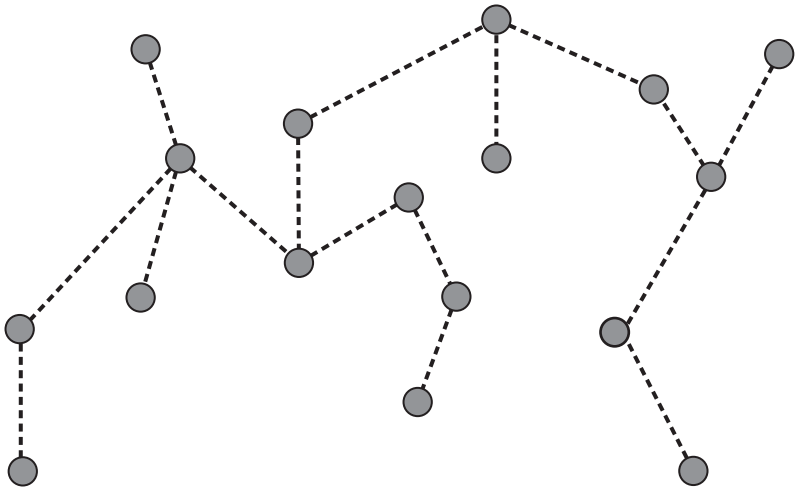


Figure 3.4 User nodes forming route connections.

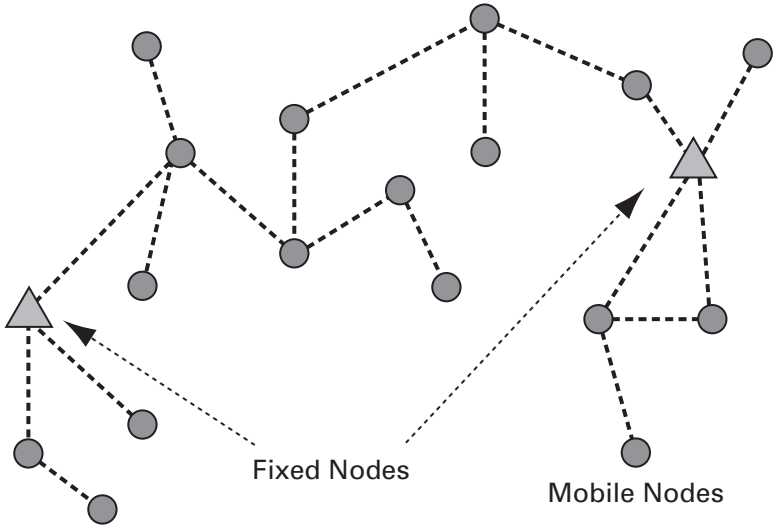


Figure 3.5 Adding fixed relay nodes within a network.

- To ensure a minimum degree of ongoing coverage and connectivity, independent of customer density. This may be required, for example, to address a time-varying lack of subscriber nodes, which arises as users commute in and out of a city.

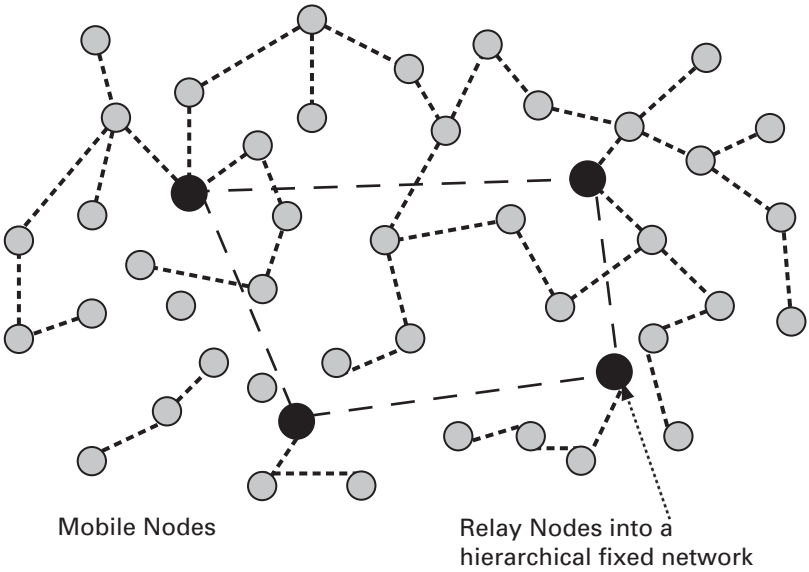


Figure 3.6 Mesh with backbone network.

- To enhance throughput in regions of high customer density.
- To enhance coverage by aiding routing around obstacles, such as in the urban environment.

Moving one step further, let us consider end-to-end traffic which flows over longer path lengths, which requires a large number of hops. To help here, a hierarchy of fixed relay nodes, each supporting a longer communication range, could be deployed. This forms a 'backbone network' within the mesh, Figure 3.6. The backbone can be wired or wireless. This architecture avoids a potential reduction in efficiency, or the increase in end-to-end delay, which is associated with many hops.

The backbone architecture would need to employ routing such that the chosen route would depend on the current state of each potential link in terms of some figure of merit which might include throughput, delay etc. Some commercial mesh offerings have one or more levels of backbone or backhaul using independent, dedicated radio channels or bands.

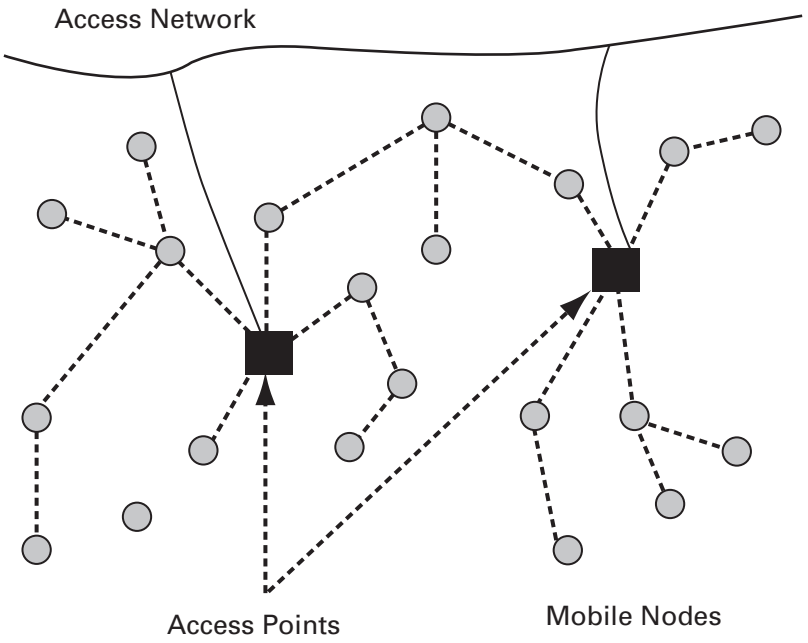


Figure 3.7 Extra-mesh traffic flow via access points – the 'access mesh'.

Having made this potential improvement via a backbone, note that the traffic flow may now no longer be evenly distributed throughout the mobile nodes. It may be concentrated around nodes in the vicinity of the relay nodes. This has significant implications for the required quantity, performance and battery endurance of the mobile nodes in proximity to the fixed relay nodes as well as for the relay nodes themselves. Thus the quality of service (QoS) becomes more dependent on the availability and behaviour of subscriber nodes in the vicinity of relay nodes.

Extra-mesh traffic flow

We turn now to the second type of traffic flow, extra-mesh. In this case traffic enters or leaves the mesh via one or more 'access points' which are connected to a public or private access network, Figure 3.7.

As with the backbone mesh network, the traffic flow is no longer so evenly distributed throughout the mobile nodes, but is concentrated around nodes in the vicinity of the access points. Again this has major

implications on achieving satisfactory quality of service when it is subject to the availability and behaviour of subscriber nodes in the vicinity of access points. However, it will be seen later in the book that the presence of the access points themselves has a positive impact on mesh scalability and quality of service.

In summary, an access mesh is the architecture we need for Internet access, and the traffic concentration effects that this brings are a consequence we must deal with.

3.3 Medium access control

Medium access techniques are required to allow a user to access the physical medium, which in our case is the air interface. It follows easily that the choice of MAC scheme and its related efficiency must thus be a determinant of system efficiency. In an ideal world, one might assume that each node's access to the channel would be fairly arbitrated. In some cases this is indeed so. In cases where it is not so perfect, a common reason is that the MAC function was not or was not able to be centralised. The alternative to centralisation is for each node to participate in a MAC function which is distributed amongst all nodes.

To take examples, in a cellular network there is centralised co-ordination of many of the PHY and MAC processes. However, in a randomly distributed network, such as a mobile mesh or a shared Ethernet, there is significantly less opportunity for co-ordination, so many of the desirable, efficient attributes of medium access schemes are not available. They are simply less easy to implement and manage in such a decentralised manner.

A simple test of whether a MAC is distributed or centralised is to ask whether the user node is ever in sole control of when it may transmit. If it can make this decision, then the MAC function is distributed, or decentralised.

3.3.1 MACs for fixed and planned applications

A centralised MAC is more appropriate where nodes are fixed and hence known by a central controller. This knowledge allows an entirely

deterministic MAC protocol to be used. Deterministic protocols are typically those which are 'slotted', meaning the MAC will give the node a slice of time, frequency, spreading code or other unique allocation of resource where it may transmit. The MAC examples just given are more commonly called TDMA, FDMA and CDMA (time, frequency and code division multiple access).

An interesting additional consideration is that quality of service is well able to be supported by a deterministic protocol; examples are cellular systems and IEEE 802.16d. The basic reason is that guarantees of service may be most easily made when there is a guarantee of a regular, unique communications channel. Unfortunately, centralised control and ad hoc mesh operation are largely exclusive concepts.

3.3.2 MACs for mobile and ad hoc applications

A distributed MAC is more appropriate where node location is not fixed. In a distributed MAC, each node participates in the MAC process. In this way it does not matter if any node disappears, either due to mobility or simply because it is turned off by the user. Distributed MACs can be operated under the principle of random access or, less likely, controlled access to the channel. Examples of controlled access, distributed MACs in the mainstream are not numerous, but may be said to include IEEE 802.16e, which provides traffic flow priorities at the MAC level. Sometimes the user node is allowed to decide when to transmit, although this is normally only to request a time slot to access the centralised MAC. This really makes 802.16 a hybrid MAC approach.

More common, and relevant to mesh, is the random access, distributed MAC, such as the well-known CSMA/CD (carrier sense multiple access/collision detection) of shared Ethernet or the CSMA/CA (collision avoidance) of 802.11, where collisions are (mostly) avoided by a listen-before-talk arrangement operated by the nodes themselves, rather than any assignment of priorities or other attempts at global co-ordination. Nonetheless, collisions may still occur in some circumstances, and the method is always less efficient than a non-random, deterministic access

scheme. The huge attraction is one of relative ease of implementation. As already mentioned, the prime example is 802.11.

3.4 Routing

First of all, let us remind ourselves what routing is; it is simply the function of knowing which path to take in order to deliver the data from one end to the other. It must include an addressing scheme and a routing protocol. To put it most simply, if the addressing scheme consisted of house addresses, then a knowledgeable postman would be the routing protocol.

Routing protocols are worked hard in a mesh, particularly if it is mobile. Imagine our postman's job if houses were regularly to disappear and re-appear somewhere else, or if they all moved constantly in different directions. This is the nature of the mobile routing protocol challenge. Routing protocols must be efficient and must either keep their knowledge proactively up to date, or be quickly reactive when routes are required. Which of theses two routing approaches is better depends on the behaviour of the mesh users (e.g. the number of hand-overs demanded) and the traffic (e.g. video versus simple file transfer) which must be carried. Each mesh node must participate in the routing and relay of other's traffic, which we describe next.

3.4.1 Every node is a router

In an ad hoc network environment, each node participates within the network not only as a possible source and sink of traffic, but also as a relay and a router, enabling the forwarding of traffic between nodes in the network. Thus, in an ad hoc mesh, each and every node needs to know what route traffic needs to follow in order to reach its destination. This is directly analogous to router behaviour in the Internet, although the big difference is that not all Internet nodes need be routers, as indeed most are not. As already illustrated, mobility adds volatility to the routing problem.

3.4.2 Every node is a relay

To be able to form a mesh it is necessary for nodes to act as relays. This is essential but has several consequences.

Firstly, by acting as a relay, a node takes on a workload over and above that needed to support its user's requirements. The accepted consensus from both academic and commercial published work is that nodes often need to be able to relay traffic of a volume a few times above that of their own service, i.e. they must handle not only their user generated traffic but potentially that of several other users as well. This implies that meshes will require more capable user nodes than otherwise equivalent cellular systems.

Secondly, a consequence of having a service level which depends on the performance of user nodes is an increased difficulty in maintenance and upgrade. For example with the introduction of EDGE onto GSM it was possible to upgrade base stations, and then allow subscribers to sign up to the new services as and when required, if at all. With a mesh system, users are dependent on the installed base of fellow subscribers and so new services cannot be provided unless all or at least a substantial fraction of existing users are persuaded to upgrade their units.

3.4.3 Proactive and reactive routing in ad hoc networks

Ad hoc routing protocols are used in environments where there is not necessarily a well-controlled infrastructure network but where there is a common routing policy. There are two main categories of protocols in ad hoc routing: proactive and reactive.

Proactive routing protocols are based on the 'normal' routing protocols used in wired networks, such as today's Internet. Algorithms based on distance vector or link state are common. Distance vector often uses the number of hops as a metric to minimise when selecting the best route, but can also go beyond this to consider more parameters, for example link bandwidth and delay. This is used to construct a route table, which is shared with other routers. Link state operation is more complex and requires each router to build its own map of the network.

Thus, in proactive protocols, there is an attempt to build locally, at each node, a picture of routes within the network before they are required for use. The routing tables are usually built periodically through the normal operation of the protocol exchanging routing update packets. In normal operation, this has the advantage that the routes are already pre-computed and so packet forwarding can take place as soon as a packet for a particular destination appears at a node. The drawback is that routes may be calculated and re-calculated (for example due to node mobility) when they are not actually required for data. This wastes bandwidth and, for mobile nodes, also wastes battery power via the sending and receiving of unnecessary routing updates.

Reactive routing takes an alternative approach by building routes only upon demand. It may also cache route information according to some short time-out or staleness policy. Cached routes can be used as required, but if a route is not known then it has to be 'discovered'. This has the advantage that routes are only evaluated when needed, although this approach adds latency to packet forwarding when routes are not already known.

Generally, the reactive routing approach is the one that has received most attention in the ad hoc networking community.

3.5 Transport and applications

We take these together because to consider a transport protocol without knowing what is to be transported would be to miss the point entirely. A simple example is that to transfer a large database file, large packets would be most efficient. If, however, the large file was actually a real-time video stream, then using small packets would make more sense. This is because if a real-time video packet is dropped, then it is usually ignored (the preceding frame is often frozen). There is no point in re-transmitting the packet as time has moved on and it is more acceptable to the user simply to drop the packet permanently. Clearly dropping a large packet is worse than dropping a small packet. This illustrates two points about transport protocols, namely packet size and re-transmission policy.

However, it is unlikely that transport protocols are going to be redesigned just to accommodate mesh networks, or any new network. IP, TCP and streaming protocols such as the various commercial streaming formats

and real-time streaming protocol (RTSP) are really quite established within the Internet. Plus, after all, the whole point of a layered communications stack is that higher layers can be made independent of lower layers.

We can safely assume then that the common transport protocols and applications, which were designed before meshing was envisaged, will simply continue. But anomalous effects may occur, in the sense that this can lead to unintended consequences. An example is that the normal behaviour of a mesh is to reconfigure when faced with poor links or nodes moving out of range, given a better link is available. This may happen very often in a mobile mesh. The problem lies when the transport protocol detects dropped packets during the period of the mesh reconfiguration. Dropped packets are seen by transport protocols as congestion and they reduce their throughput via back-off techniques. This works very well in the relatively static wired Internet, but is exactly the opposite to what is required in a wireless mobile mesh in this specific situation.

For completeness we should mention that any wireless link can cause similar TCP issues via dropped packets. The effect of mesh reconfiguration described above is in addition to the wireless effect.

Having shown that the type of application to be carried is pertinent, let us look at a common way to divide them into two distinct categories.

3.5.1 Elastic and inelastic applications

Applications may be classified as elastic or inelastic, depending on their tolerance to delay in transmission. Such applications need to be handled differently by a transport protocol, as we have already implied.

The majority of traffic in today's Internet is made up of elastic applications, that is applications which are tolerant to variations in the throughput but that typically require 100% packet reliability. Popular examples include email, web access and peer-to-peer file sharing.

But a growing portion of internet traffic is made up of non-elastic applications. These applications are intolerant to throughput variations, such as bandwidth, delay and delay variation. An obvious example is one way video, where delay variation is normally called jitter and leads to picture freeze and eventual break-up. In a two way video link, absolute

delay is also important, in order to maintain the spontaneity of normal conversation and reaction. On the other hand, these applications do not always demand 100% reliability of packet delivery.

Splitting applications into elastic and non-elastic types is very useful for many of the later discussions in this book.

3.6 Summary

In this chapter we saw that when a mesh is formed from nodes with omni-directional antennas, then the mesh must be formed as a logical mesh. Omni-directionality is a great enabler of the most attractive applications of mesh, but it affects how we look at mesh capacity in Chapter 4.

We saw that, just like any other communication node, mesh nodes need to participate in the MAC protocol. A big difference, however, is that every mesh node needs to be both a router and a relay of the traffic of others. These two additional requirements considerably increase the complexity and performance required of a mesh node. In the next chapter we shall see that the relay function in particular leads to a fundamental limit on the performance of mesh systems.

But, firstly, in the next chapter we drive into the detail of mesh operation by testing the first of our four hypotheses. As a reminder, these claim that a mesh will

- self-generate capacity,
- improve spectral efficiency,
- benefit from directional antennas, and
- improve the overall utilisation of spectrum.

Surprisingly, perhaps, we will find that to test these hypotheses fully we will need to use only PHY and MAC based arguments, albeit in some considerable detail. On the other hand, this may be seen as not too surprising, since all the claims are so closely connected to the lower levels of the protocol stack.

But, this is far from saying that higher layer effects are unimportant. On the contrary, we will show that specific higher layer techniques are essential to make a usable mesh, but we may safely leave discussion of this to subsequent chapters.

4 Mesh capacity, scalability and efficiency – hypothesis testing

Having covered the fundamentals of meshes, we now arrive at the point where we may begin to consider the big and often asked questions about mesh, four of which we consider together, via our list of hypotheses. As a reminder, these are that

1. meshes self-generate capacity,
2. meshes improve spectral efficiency,
3. directional antennas help a mesh, and
4. meshes improve the overall utilisation of spectrum.

We will examine them formally, via analysis of existing peer reviewed publications, followed by some more recent analysis and insight of our own [1, 2]. A key problem in assessing the published literature is that different assumptions are made in different published papers; a direct comparison is thus at risk of being inconsistent. We spend some time at the outset to ensure we avoid this issue.

We will bear in mind that we are predominantly interested in our six application examples of Chapter 2. This will set helpful bounds to our scope for testing the hypotheses.

When we look at Hypothesis 1 which is concerned with capacity, we form our initial viewpoint via a simple thought experiment, which looks at how we expect the capacity of a mesh might behave versus demand, relative to the known case of cellular. This is followed by a summary of four important peer reviewed research papers in the field, which concern system capacity. We contend that the important conclusions presented in these papers were never intended to be used by readers as evidence that a real-world mesh can self-generate capacity. To resolve this issue, we look at the capacity of a pure mesh by examining one of the four key papers in detail, being careful to note the caveats presented by the authors. Based on this we produce two types of insight into what does cause a limit to mesh

capacity, firstly based on mathematical analysis and secondly based on a more physical consideration of the situation. With this groundwork established, we return to reconsider all four key publications. We show why the results should not always be transposed to other situations as directly as might be first thought. Finally, with respect to the capacity hypothesis, we turn to other work, including simulations and measurements, which we use to support the view of mesh capacity presented in this book, both for a hybrid mesh (i.e. one with a backbone) and for an access mesh. Our overall conclusion is that the important figure of merit remains per-user throughput and that this factor always decreases as the number of mesh nodes increases, if all else remains equal. In other words, meshes can never self-generate capacity.

Our analysis of Hypothesis 1 occupies a large portion of this chapter.

When we look at Hypothesis 2, which concerns spectral efficiency, we begin by considering cellular to be our benchmark. We go on to show that in general there is no support for the notion that mesh will be more spectrally efficient than cellular. However, we note that in some cases mesh may be much better than cellular and that in other cases it may be much worse. That is to say that the performance of mesh is much more variable and we show that this depends on both the deployment environment and the actions of users themselves. In coming to these conclusions we take a close look at the advantages promised by multi-hopping, both in theory and from a more practical viewpoint.

When we look at Hypothesis 3, which concerns directional antennas, we do so in the knowledge that the analyses of the preceding hypotheses were conducted on the basis of employing omni-directional antennas. Because of this, we first look at the general benefits a directional antenna may give to any system. Then we look at the additional requirements imposed by a directional system, especially antenna control and steering. Finally we consider the problems of manufacture and performance, specifically for antennas which might be used in our six application examples of Chapter 2. This yields a very profound result and in fact quickly curtails our interest in directional antennas for our applications.

When we look at Hypothesis 4, which concerns utilisation of spectrum, we introduce the various different types of efficiency measure for a

wireless communications system. Even though mesh may not be the most spectrally efficient, it has the major attraction that it need not use the same sweet spot of spectrum which is so in demand by cellular. In fact, a mesh is more suited to using higher frequency, less precious spectrum.

4.1 Hypothesis 1 – Could customers self-generate capacity in a mesh?

Recall from Chapter 1 that this question was one often asked by many of the business and academic networking professionals to whom we spoke. There would be huge attractions to having 'self-generation of capacity' in a radio network. Notably, the network would be self-sustaining and it could avoid the so-called 'tragedy of the commons'. The commons ideology is often discussed in relation to radio spectrum, so we should explain it here. Such a tragedy relates to the days when common land was used for the grazing of livestock with free access for all. The real danger was that free access to such a finite resource could result in that resource being fully consumed, or compromised such that it lost its usefulness to all (the tragedy). What then, if each user were somehow to add grazing capacity as they joined the common? This would be 'self-generation of capacity', if it were possible.

As an aside and to be fair, it must be pointed out that not everyone believes in the tragedy of the commons. Conversations with UK mobile network operators suggest that they believe the commons policy worked well historically and they expect it to work well with radio spectrum with one key proviso – that there will always be access to premium spectrum and services for those who are willing to pay for it, if and when it is needed.

4.1.1 Starting with the answer

Fortunately, the self-generation of capacity argument can be examined independently of the commons argument. This is the approach we will take, but because the question of capacity is complex, we are going to state up-front what we have found the answer to be, before proving it. The

balance of this chapter thus works towards the following important conclusions.

1. For a pure mesh, subscribers cannot self-generate capacity at a rate which is sufficient to maintain their per-user throughput, as the network size and population increases.
2. The only viable way in which scalability can be achieved is by providing additional capacity either in the form of a secondary backbone, so forming a hybrid mesh, or in the form of access points, so forming an access mesh. In these two configurations scalability is possible and has characteristics broadly similar to those of a cellular network.

These conclusions are worded precisely and it will become clear why this needs to be so. We define and illustrate capacity and scalability next.

4.1.2 Capacity and scaling issues – a thought experiment

To set the scene further we are going to conduct a simple thought experiment. We know a great deal about cellular system capacity and how it behaves. What would this lead us to expect about mesh, based on the fundamentals we have covered so far? There is no proof here, just a first exploration of the situation.

In other words, we use this short section to draw a contrast between the known cases of cellular capacity and scalability and those which we speculate might apply to mesh. But first we should define our terms. By capacity, we mean the total bit rate which is present within the network. We will define throughput as the average per-user throughput, which each user can access. When we talk about scalability, we mean that if a mesh is scalable, then it will always perform equally well in terms of user throughput, independent of its growth.

Capacity growth in a standard cellular network is represented in Figure 4.1. In a single cell the capacity per unit area is defined by the capacity of the base station covering that area. Let us call this B bps. Then the mean traffic throughput per user in the cell is proportional to B/n, where n is the number of subscribers in the cell.

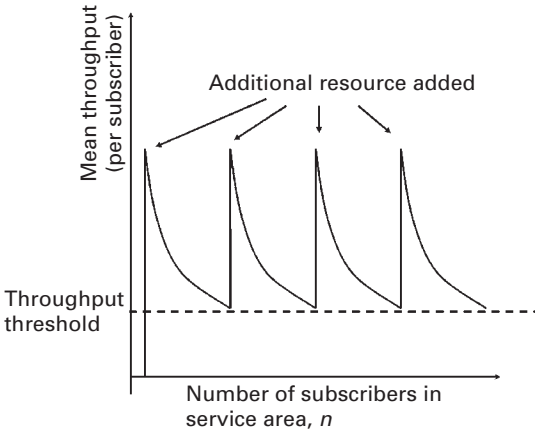


Figure 4.1 Throughput versus loading in a cellular network, or a mesh using fixed access points.

Capacity can be increased by deploying more base stations, and/or by sectoring the cells, for example by using the common three-sectored cell. Then the capacity becomes proportional to MB/n, where M is the number of cell-sectors.

Thus in terms of scalability a cellular network's per-user throughput is expected to decrease as $1/n$. At some point, an operator may decide to add more base stations when some chosen threshold of user throughput is approached. The effect of adding this extra resource is shown in Figure 4.1. In reality, adding extra resource to achieve this effect may not be possible as cells get smaller and obstructions get relatively more troublesome, but we will ignore this for our thought experiment.

Geographic coverage is extended simply by deploying more base stations over a wider area.

Having thought about the cellular case, let us now think about how Figure 4.1 might look for mesh. We will bear in mind what we said in Chapter 3 about intra-mesh and extra-mesh architectures, to suit pure meshes and access meshes respectively, and the ever-present need to relay the traffic of others.

We might expect a mesh network supporting extra-mesh traffic to have a very similar form to that shown in Figure 4.1, because the access points employed in this architecture would be parallels for the base stations of the cellular approach. However, at least two differences between the access mesh network and cellular might be expected, as follows:

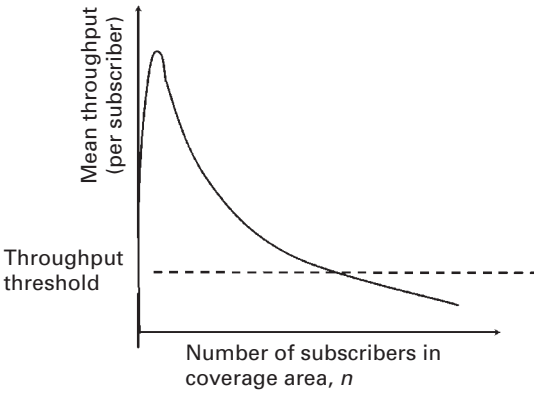


Figure 4.2 Throughput versus loading in a pure, isolated mesh network.

the rate of decay of per-node throughput may be steeper than $1/n$ because of the relaying process going on in the mesh, which must take up some capacity coverage, and

service availability in the mesh network will be dependent upon the size and distribution of the population of nodes, rather than being solely dependent on the base station.

Somewhat similar to cellular, geographic coverage is extended by deploying both more nodes and more access points over a wider area.

Moving on to the final thought experiment, for a pure mesh network supporting only intra-mesh traffic, the growth characteristics might be supposed to be different. We might expect per-node throughput to rise initially as the first few nodes are deployed and the network achieves its first degree of connectivity, see Figure 4.2. After this there is likely to be some form of throughput decay as demand from the node population increases. In some ways this is like the cellular and access mesh cases, but with additional base stations or access points never being added.

Geographic coverage is extended by deploying more nodes over a wider area, with sufficient density to ensure connectivity.

This section has consisted very much of supposing what might happen for a mesh, rather than proving it. But its purpose was not proof; next we begin the more rigorous evaluation. However, before we leave this section, consider finally what self-generation of capacity would look like on the graphs already shown – in fact it would be a straight, level line of fixed throughput per user, independent of the number of nodes added.

Table 4.1 *Capacity scaling approaches and stated assumptions*

Approach	Assumption
Mobile couriers, Grossglauser and Tse [3]	Unbounded delay
Spread spectrum, Shepard [4]	Unbounded spectrum
Infinite spectrum, Negi and Rajeswaren [5]	Unbounded spectrum
Clustered traffic only, Gupta and Kumar [6]	Strict localisation of traffic

4.1.3 Challenging self-generation of capacity models

Although this book concludes that meshes do not scale adequately without some form of additional infrastructure, we now highlight some concepts which we fear some observers may have misinterpreted as a contradiction of this.

Four published approaches to investigate capacity scaling behaviour are introduced briefly below and, whilst each presents a perfectly coherent argument based on its clearly stated assumptions, it will be shown in subsequent detail that these assumptions do not always translate well to practical applications. This is no real surprise at all, since careful reading of the publications shows that the authors themselves caution against results being used out of context. The four approaches, with their three key assumptions, are shown in Table 4.1.

Mobile couriers

This approach was taken as an input to an economics paper [7] which points to the benefits if a 'tragedy of the commons' could be avoided. It initially set the scene for an economic assessment by showing how the combination of two well-known papers on ad hoc networking could point to self-generation of capacity. The two well-known papers whose findings are combined are

- Gupta and Kumar's [6] identification that if traffic flows remain localised despite growth of the network then per-user throughput can stay constant as users are added, and
- Grossglauser and Tse's [3] assertion that mobility can be harnessed to achieve constant throughput as users are added.

In Grossglauser and Tse's approach the assumed long-range movement of nodes is used as the means of delivering traffic: traffic is transferred from source to a 'courier' (a wireless device that is in motion) and then from courier to recipient once the wireless device moves close by; thus all traffic is transferred via just two radio-hops. The key assumption is that a convenient courier node will be passing by at some timely moment. Clearly, the drawback is that there can be no guarantees on delivery or latency, and so the technique has very specific application. This method is discussed further in Section 4.1.7, where the localisation of traffic assumption is also covered.

Spread spectrum and infinite spectrum

We take these together as they are both based upon a similar key assumption.

Shepard [4] has a relatively 'out-of-the-box' approach in suggesting a mesh in which collisions are not the be-all and end-all for the MAC. He sees multiple transmissions as a signal-to-noise issue, rather than a requirement to back off and try again. He does this by using spread spectrum transmission, hence multiple transmissions simply raise the noise floor, as in a CDMA system. He develops a complete theory to enable meshes to scale to millions of nodes with an allegedly respectable throughput. The problem is that it is extremely spectrally hungry, due to the very large processing gain required. Other problems include the high computational effort and the distributed MAC complexity, which remains an open research issue. He also concludes that small hop count meshes are best, so that interference may be managed locally by his distributed algorithm.

Problems here concern the to-be-invented distributed MAC algorithm itself, but more importantly that an increase in spectrum must be traded for the necessary increase in signal-to-noise ratio. In any case the throughput of a large mesh is still only in the several kbps range in Shepard's own worked example with a realistic spectral bandwidth.

The idea of having a MAC which is more sophisticated than being simply collision based (CSMA/CA), like 802.11, clearly has merit. It is noted in Section 5.2 that such a MAC approach will result in better

efficiency of access to the available spectrum, albeit with an attendant need for increased network knowledge by the user nodes or by bespoke management nodes. We presently know of no reports of such a MAC.

A broadly similar approach with similar problems is that of using 'infinite' spectral bandwidth (bandwidth substantially greater than required to support the raw transmission rate used on links), for example something broadly akin to an extreme example of the ultra wide band (UWB) approach. This approach is discussed in Section 4.1.7.

Strict localisation of traffic [6]

Gupta and Kumar [6] identify that if traffic is confined to localised groups in a relatively large mesh, then scaling will be achieved. Clearly this is very application specific and it is arguable that this is no longer a large mesh, but is now a collection of smaller meshes. Further discussion appears in Section 4.1.7.

This paper contains much of the first analysis of mesh, having been published when pure mesh was first of interest to the military. We look at the key and valuable results for mesh from this paper in detail, below.

4.1.4 Analysis of pure mesh capacity limitations

Having introduced the seminal capacity papers, let us now return to analyse the pure mesh. We wish to look at the key results, with close consideration of the assumptions on which they are based. After an assessment of the scope of the analysis problem, we will make much use of the results from Gupta and Kumar [6].

When analysing a mesh, a full mathematical representation is complex because of the large number of parameters which must be specified, including amongst others

radio range per link, including as a function of bandwidth per link,
transmit power control procedures,
radio environment, including propagation law, clutter losses etc.,
signal-to-interference margins for different carrier modulation types and data rates,
type of multiple access scheme (MAC),

traffic models,
user mobility models,
routing algorithm, etc.

Because of this high level of complexity, there can be no simple full system model. Instead some parameters must be constrained in order to construct a tractable problem. In recognition of this, common problem constraints adopted in the literature include

- restricting nodes to be static only,
- fixing the link bandwidth,
- fixing the link range,
- assuming a regular distribution of nodes, and
- assuming an even distribution of traffic at a constant bit rate.

As a consequence of adopting constraints, published mesh system analyses generally examine just one or two aspects of system behaviour. Different constraints allow the influence of different system parameters to be investigated. To compare results between papers, therefore, it is necessary to be very specific about matching assumptions.

The alternative to mathematical modelling is system simulation. Again there is a considerable amount of activity in this field but the vast majority employs some form of IEEE 802.11 air interface and MAC. This is a choice based on convenience (especially easy access to information and equipment), rather than science, since it is generally agreed that the 802.11 series and associated MAC are sub-optimal for efficient mesh networking. For the same reason as cited above for analysis work, such simulations can serve only to indicate performance for a specific configuration.

Let us begin our analysis. As already implied, Gupta and Kumar [6] are the authors of much of the pivotal theoretical analysis of mesh networks which is taken as the reference work by many subsequent researchers. We will do the same. For the reasons above, like all other models, Gupta and Kumar's model is necessarily highly idealised and this leads to optimistic conclusions which the authors themselves caution against.

Firstly we need to take a look at some of the terms and constraints used in the analysis, of which there are necessarily quite a number.

The analysis includes concepts of arbitrary networks, random networks, a protocol model and a physical model. These definitions have been adopted quite widely and may be understood as follows.

The arbitrary network is defined as one in which the locations of the nodes, traffic sources and destinations, and their traffic demands, are arranged arbitrarily to achieve best performance, judged by some given metric.

The random network is one in which these parameters each have a statistical distribution. Therefore the random network is more relevant to the analysis of a real network, but still not ideal.

The protocol model assumes that the protocol is able to prevent neighbouring nodes transmitting at the same time on the same channel within a specified range, i.e. collisions are always assumed to be absent.

The physical model sets a minimum interference ratio for successful reception and allows collisions. Therefore the physical model is considered the more relevant to practical implementations.

The key constraints of the model are the following.

All nodes are static.

All nodes are uniformly, randomly distributed in a unified propagation environment.

All nodes transmit at the same power level.

All nodes have the same throughput bit rate.

Log-normal fading is not included.

All nodes employ omni-directional antennas.

All overheads for channel coding, routing and management traffic are ignored.

A single signal to interference ratio is set, where appropriate.

The work is specifically concerned with scalability, that is the behaviour of capacity as subscriber numbers are increased. This means that relative rather than absolute capacity is addressed. The key conclusions of the work may be summarised.

1. For a network of n identical randomly located nodes, each capable of relaying usable data at W bps over a fixed range, the upper bound for the throughput capacity of the total network is proportional to $W\sqrt{n}$. Thus the network's capacity increases in proportion to the square root of the node population. The phrase 'of order of' in this context means that, to a first-order approximation, the value tends towards this value as n approaches infinity.
2. This capacity is shared amongst the nodes such that the upper bound for the average throughput obtainable by each node for a randomly chosen destination is proportional to $W/\sqrt{(n \log n)}$ for the random network with physical model. In other words, the per-user throughput decreases with increasing node population.
3. The constants of proportionality in 1 and 2 above are functions of the signal-to-noise threshold required of the carrier modulation scheme and the rate of decay of RF signal power (i.e. the propagation law) such that for a high signal-to-noise threshold the capacity limits are reduced, whilst for a high propagation law the capacity limits are increased, as might be expected.

As we have said, the above results are frequently quoted by researchers. A specific issue often encountered is that it seems that whilst the first conclusion is noted, the second is overlooked. But capacity and throughput are quite different. The difference concerns the routing of the traffic of other nodes. Although the system capacity is going up, the per-user throughput is going down. The difference between the system capacity and the total per-user throughput is the traffic of others which each node must also carry. This is a component of system capacity, but it is not a component of per-user throughput, since the user cannot access it. With the benefit of hindsight, this is very similar to what our thought experiment was hinting at, at the start of this chapter.

Also we must remember that these are idealised theoretical upper bounds on performance. The paper includes a specific comment on this:

> *The results in this paper allow for a perfect scheduling algorithm which knows the locations of all nodes and all traffic demands, and which co-ordinates wireless transmissions temporally and spatially to avoid collisions which would otherwise result in lost packets. If such perfect node*

location information is not available, or traffic demands are not known, or if nodes move, then the capacity can only be reduced. [6]

Subsequently, other researchers have analysed less perfect scenarios and they have found a lesser performance. An example is the theoretical work by Arpacioglu and Zygmunt [8] in which it is concluded that the average per-user throughput obtainable for a randomly chosen destination has a faster rate of decay approximately proportional to $1/n$, rather than $1/n$ or $1/\sqrt{(n \log n)}$ (depending on model conditions) as predicted by Gupta and Kumar.

Looking at the detail shows that the primary reason for the more pessimistic performance is that in Gupta and Kumar's model the path loss is modelled as d^{γ}, but d is allowed to reduce towards zero as the density of nodes is increased. This leads to an anomalous decrease in path loss for $d < 1$. Arpacioglu and Zygmunt eliminate this by setting path loss as $(1+d)^{\gamma}$. This leads to the result that the capacity of their network does not increase monotonically with n, but rather there is an upper limit to the number of simultaneous transmissions that can be supported in a given area, which is independent of n. This would agree with what is often seen in practice.

But regardless of detailed differences in the order of proportionality with increasing node number, n, both these models agree that average per-user throughput diminishes towards zero as the number of nodes increases. Thus the mesh network does not scale indefinitely. To push the point firmly home, meshes do not self-generate capacity and both of the published works just examined show this.

4.1.5 Underlying causes of limited capacity – mathematical insights

Having established the case for mesh scalability behaviour, let us go further. It is interesting to consider what parameters, if any, might be changed to avoid this reduction in throughput. For this discussion let us consider a simplified view of a network model. Using Arpacioglu and Zygmunt's [8] model we can look at the influence of a range of parameters on the average throughput. Our question is whether any of these parameters could possibly be manipulated so as to reverse our conclusions.

The model states that the average throughput $\lambda(n)$ is proportional to (functions of)

$$\gamma,\ W,\ G/\beta,\ 1/L,\ 1/r,\ A,\ \text{and}\ 1/n$$

where γ is the propagation attenuation law, W is the channel transmission rate, G is the channel processing gain, β is the required signal-to-noise ratio (SNR), L is the mean end-to-end path length, r is the mean per-hop link length, A is the area covered by the network and n is the number of nodes.

This is very useful as it implies that, unless one or more of the parameters increases with n, then per-user throughput will be asymptotic to zero. Let us test each term in turn.

- W, the channel transmission rate, cannot grow arbitrarily large because of thermal noise constraints and limits on transmission power.
- G/β depends on the design of the communication system and increasing this generally makes it necessary to decrease the transmission rate W, so negating any potential advantage.
- Reducing the hop length r (e.g. by constraining transmit power) increases spatial re-use but at the expense of increased hop-count and hence increased relay traffic. However, it transpires that the preference is to reduce r to increase spatial re-use [6, 8]. But there is of course a limit here in that if r is too small then the network can become disconnected, i.e. minimum r is related to the inverse of node density (A/n).
- In random traffic flow models with uniform node density the mean end-to-end communication path length, L, is assumed to grow with coverage area A (L proportional to $\sqrt{A}$). This reduces capacity because of increased hop count. Thus, if one could conceive of services with more localised traffic (e.g. amongst localised communities) then A/L would increase more rapidly with increasing A. This would help to improve scalability. The idea of localised traffic crops up in many places, the problem being that many small isolated meshes are created with this approach, which limits the useful applications.
- The remaining parameter that might scale with n is the area A. Arpacioglu and Zygmunt ([8] Corollary 3) suggest that three factors

are required to achieve a non-zero throughput with increasing n: (i) the attenuation law γ needs to be greater than 3, (ii) the hop count H needs to be independent of n, and (iii) the area A needs to increase with n (i.e. the node *density* needs to be nearly constant or reducing with increasing A). However, (iii) requires that as the subscriber base increases those subscribers spread themselves out more thinly. It is not easy to see on what basis this might happen in any practical deployment.

- The propagation attenuation law γ strongly influences the above conclusions. A higher attenuation factor γ will permit higher throughput capacity [6, 8]. This means that meshes may well be more attractive at higher frequencies. We pick up this point when we look at spectrum utilisation in Hypothesis 4.

From the above list of options, one can see that there appears to be very little prospect of avoiding the asymptotic reduction in per-user throughput with increasing subscriber base.

4.1.6 Underlying causes of limited capacity – physical insights

It is instructive to identify the physical reasons why per-user throughput decreases with increasing population in a mesh supporting intra-mesh traffic. We cover quite a range of issues and provide a summary at the end of this section.

We aim to show that when considering average rates of traffic flow, this decreasing throughput is not due to blockage caused by the limited relay throughput of nodes. Instead it is due to having a limited spectrum for all nodes to share within the mesh (i.e. it is an interference effect) and the need to relay traffic through multiple hops. However, when considering specific traffic flows, the throughput may be constrained by the limited transport capability of nodes.

To set the scene consider the activity around a single node. The use of the air interface to communicate across a hop will impose an interference boundary around the transmitting node, recalling that our antennas are omni-directional. We say this is an interference boundary since, within this zone, the same frequencies cannot be used at the same time by any

other node. Other nodes wanting to communicate within this interference zone must choose other frequencies or other times to do so.

If the transmission rate of nodes is W bps then the maximum total throughput through this interference zone is of order W. Other traffic paths can pass through this zone, but the total throughput remains limited to W bps. If there are m other nodes in this zone then the zone's throughput can be fairly shared – providing a mean of W/m to each. If this zone is of area a, then we can consider having a maximum throughput of order W/a (bps per unit area) in this zone. Clearly, then, it is advantageous to keep this area, a, as small as possible. This confirms the conclusion of other researchers that short hop lengths and high propagation attenuation factor are conducive to high throughput capacity of the network.

Thus it can be seen that throughput is not limited by blocking at individual nodes but by the throughput capacity of the interference zones.

We can look for support for the conclusion that node relay throughput is not the dominant limiting factor, in a system simulation by Hekmat and Mieghem [7]. This paper addresses mesh throughput from the standpoint of signal-to-interference levels within a network, and from this it seeks to determine values for hop count, capacity and throughput per node. It uses only a regular lattice of nodes rather than a random distribution, but it does attempt to model some practical values for data rate and bandwidth, based on 802.11b.

The simulation compares the capacity or potential throughput at a node, which may not be fully used, with the throughput actually used, versus the number of nodes in the system. As the number of nodes increases, the potential throughput per node falls as we now know it must, whilst the throughput actually used is increasing. Figure 4.3 illustrates a 'saturation point' at which the achievable per-node throughput, which is limited by interference from other nodes, is equal to the required per-node throughput to support a given offered traffic level. This occurs at the point where the achievable capacity is approximately half that of the maximum potential throughput capacity of a node. We may conclude from this that at saturation, the node potential maximum throughput has not been reached, so the network saturation must have been caused by the effect of the interference zones.

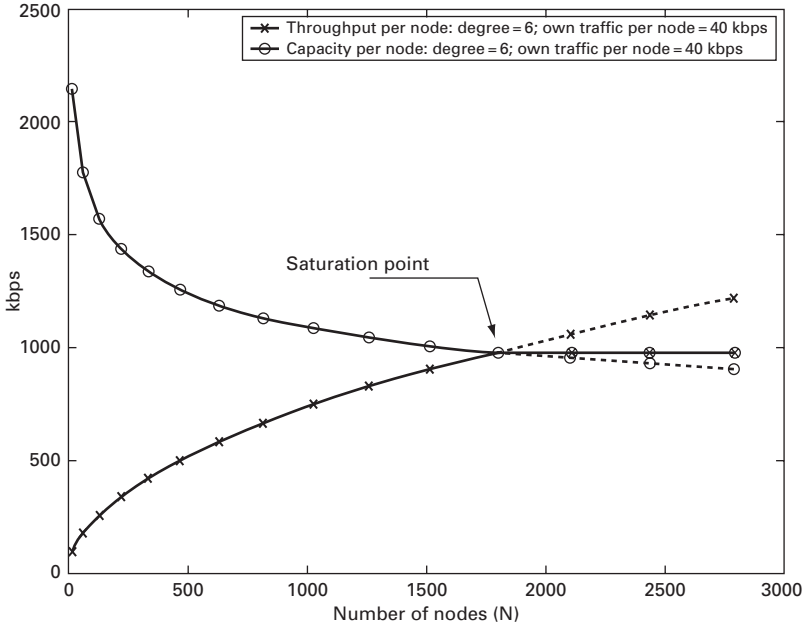


Figure 4.3 Comparison of available and required relay throughput per node (802.11 system with 22 MHz channel width, before CDMA de-spreading, and 2 Mbps relay rate), used with permission from Hekmat and Mieghem [9].

Traffic limitation

Let us continue to use our interference zone concept to look at blocking effects which may occur along a multi-hop route within a mesh. Figure 4.4 illustrates such a route, showing it as a sequence of transmit/receive boundaries.

In the figure the shaded circles with letters A, B and C represent the usage of different frequencies and/or transmit times to pass traffic along a route. We may refer to these frequency-time elements simply as the 'resource' we need to use. The smaller circles indicate the omni-directional boundary of the wanted signal on each link and the larger circles indicate the interference zones corresponding to each of these. For clarity only the interference boundary for resource A is shown. By way of explanation, within the wanted signal boundary a node will be able to receive the transmission successfully. Outside this region, but within the interference

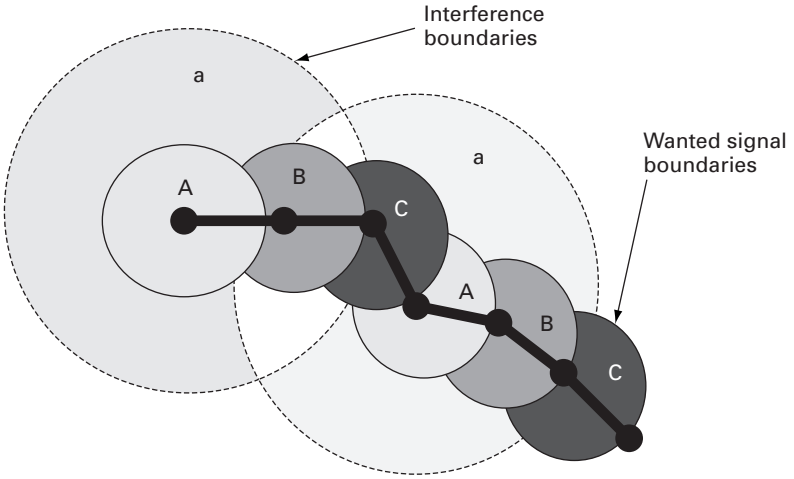


Figure 4.4 Spectrum resource re-use along a traffic route and associated interference zones around transmitters.

zone, the node cannot successfully receive the transmission, but it will see the transmission as a source of interference.

In our example, three resource elements are needed because the interference zone around a transmitter is assumed to extend to nominally twice the communication distance and so a minimum of three hop lengths separation is required between co-resource users. We derive a factor of two for the interference to wanted ranges according to the following rationale. Let the required SNR be β and let the propagation attenuation factor be γ. If we neglect log-normal fading, then the theoretical ratio of interference range versus wanted link length, Δ, is given by $10\gamma \log \Delta = \log \beta$. For example, if $\beta = 13$ dB and the propagation factor is 4 then the range-ratio is approximately equal to 2.

What we have considered actually represents a near-optimal situation for a 'string of pearls' route in which hops are nominally aligned and are of the same length. But this is ideal and in practice the resource utilisation is likely to be substantially higher. We can look to the work of Li *et al.* [10] who suggest a theoretical lower limit of four hop lengths between competing users, rather than three, although their simulations using a modified 802.11 MAC required up to seven.

In general, if the number of different frequency and time combinations needed per route is represented as b (where $b=3$ to 7, as discussed), then for a traffic rate of T bps a resource capacity of bT is used for each route.

We may now deduce that the total number of such traffic routes, m, which could cross within any one interference zone, area a, around an active node is limited by:

$$\sum_{i=1}^{m} bT_i \leq W.$$

To put this equation into words, the number of traffic routes which may cross any one interference zone may be calculated as the maximum number whose total required resource capacity does not exceed the bandwidth available in the interference zone.

Thus, for example, a system employing nodes with 20 Mbps relay throughput, supporting traffic rates up to 1 Mbps, and using an average $b=5$ spectrum resource elements per traffic route, could support only four such traffic routes passing through each interference zone around active nodes. Given that many diverse routes are expected within a mesh, this is quite a small number of path crossings which can be supported before route failures must occur.

Finding a way around the problem

There is more we can deduce from Figure 4.4. Notice that a single traffic route lays down a footprint of adjoining interference zones along its path and so extends the limited path crossing problem throughout its length. In effect it is the beginnings of cutting the mesh into separate pieces. There is literally a way around it, however, as shown in Figure 4.5.

The crossover 'bottleneck' caused by a specific interference zone could to some extent be reduced by diverse routing around it, but since the interference zone is large (a two-hop radius in our example) routing around it is likely to impose a considerable increase in route length. This will cause a degradation of performance, notably for latency. This issue is illustrated in Figure 4.5 for just three crossing routes.

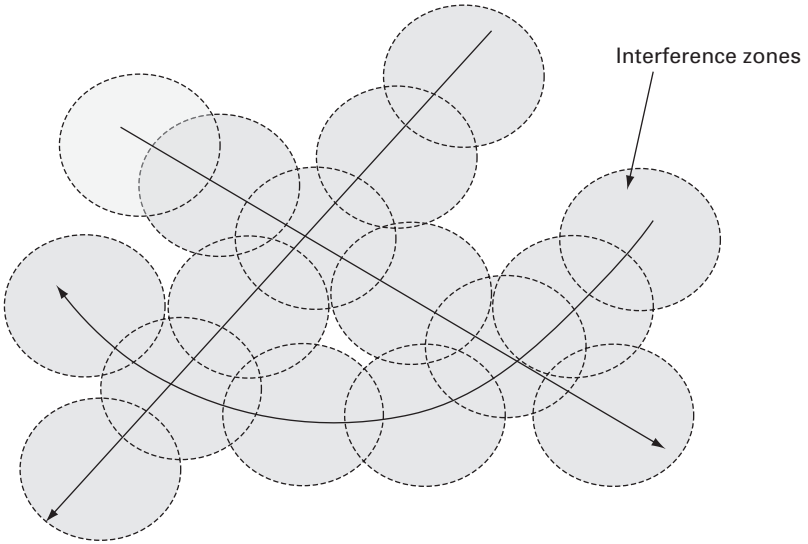


Figure 4.5 Routing three paths to avoid a three-route occupancy of a single interference zone.

Mesh cutting and partitions

Developing this analysis we can see yet another issue, which is that as the offered traffic rate, T, approaches a high proportion of the node relay rate, W, such that bT approaches W, then all the spectrum resource becomes fully employed on this one main route. Thus a single traffic route imposes an uncrossable boundary through its length and the mesh becomes partitioned as illustrated in Figure 4.6. In this case we have suffered the effect of creating two isolated meshes which cannot communicate while the main route traffic remains active.

Practical performance

All the above effects in combination will dictate the upper limit on the maximum traffic rate which can be supported by the mesh. Recall that the previous analysis [6] determined only the average throughput. In contrast, here we have given an insight into what limits the highest traffic rate, which could be offered to only a subset of the nodes.

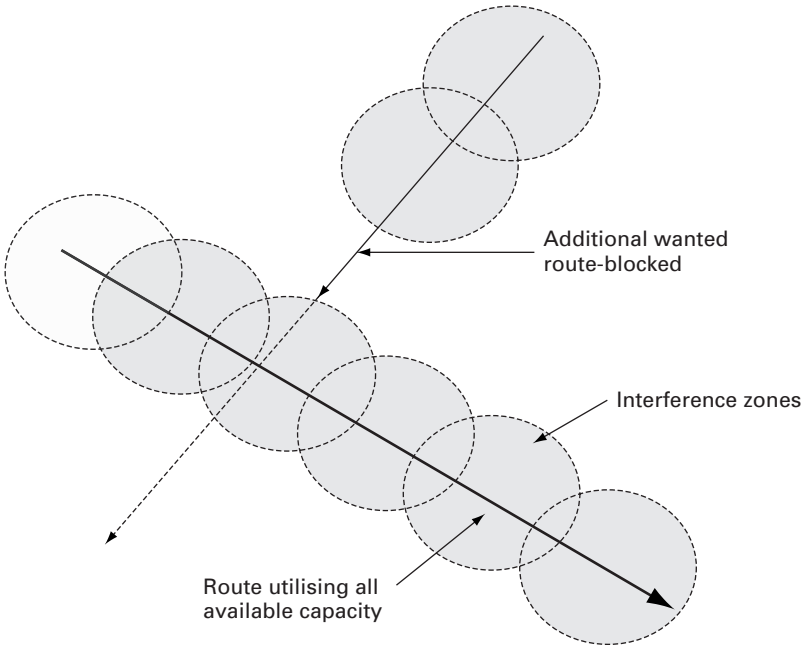


Figure 4.6 Single high throughput route causing partitioning of the network.

For any practical implementation the throughput bounds we have illustrated will not be achievable because of factors such as routing overheads, mobility and propagation environment.

To provide an indication of how this might translate to the real world, two practical examples from published literature are given.

- Gupta *et al*. [11] have found empirically that the scaling of 802.11 in particular is $1/n^{1.68}$. It is no great surprise to find this scales between $1/n$ and $1/n^2$, given the MAC is a collision avoidance MAC, ensuring only one hop is active at any one time.
- A commercial white paper [12] offers another view of scaling. The conclusion is a very substantial rate of degradation of per-user throughput which is proportional to $1/n^2$. However, this paper accepts that other researchers conclude that it is closer to $1/n$. The $1/n^2$ result includes the effect of routes not being the idealised 'string of pearls' illustrated in Figure 4.4. The authors have a three-radio solution which is said to solve the scaling issues, i.e. throughput scaling is unity,

independent of user numbers. But this is in fact a hybrid network comprising a fixed wireless-relay backbone network in additional spectrum.

Summary of our physical insights into mesh capacity limitations

We have covered quite a bit of ground with our physical insights. We have seen that interference imposes the limit on average throughput. We have also seen that in a given smaller section of mesh, maximum traffic rates may be above this average, but are limited by the presence of the paths and traffic levels of other routes. One way around this limitation is literally to go around the problem area, although this entails a longer path. If this is not possible, we have seen that it may be due to a high traffic route which has effectively chopped the mesh in two.

4.1.7 Further analysis of the myth of self-generation of capacity

Having completed our physical insights, it is time to return to the apparent mechanisms for overcoming the non-scalability of the pure mesh, which we identified at the beginning of this chapter in Table 4.1. We now examine the three assumptions used in the approaches in greater detail. These assumptions were

unbounded delay,
wide or limitless spectrum, and
that traffic would only ever exist in local pockets.

We take these assumptions in turn.

Unbounded delay: the 'mobile-courier' scenario

Grossglauser and Tse's paper [3] is often cited under the heading of 'mobile mesh'. However, this model specifically uses the mobility of nodes to act as intermediate couriers of data between source and destination. Datagrams are passed from source nodes to near neighbours and delivery occurs when the courier nodes encounter the target recipients. Under this idealised model the per-node throughput remains constant – independent of the

number of nodes, i.e. the mesh network is fully scalable in terms of capacity.

However, a clear consequence of this model is that the end-to-end packet delivery delay is related to the transit time of nodes moving throughout the area covered by the mesh. The average delivery time is of order of $2d/v$ where d is the diameter of the mesh network and v the average velocity of nodes within it. In a practical situation, of course, the courier nodes may never encounter the recipient, in which case traffic is lost. The authors conclude that this is clearly not acceptable for voice, or other real-time communications, and so direct the concept to non-critical store-and-forward messaging applications. Note also that the outward and return paths may well be completely uncorrelated in time and space, and so the concept of two-way communications is only loosely applicable.

Although of limited application in its basic form, others have commented that the technique might be enhanced to reduce the transport delay and increase the probability of message delivery by, for example, the following improvements.

1. The originating node passes its datagram to *all* of its in-range neighbours so that there is a number of nodes acting as couriers, so increasing the probability of intercepting the recipient within a given time. This is developed by Sharma and Mazumdar [13].
2. The courier nodes pass their datagrams on to neighbouring couriers at some event trigger such as a specified time interval, or the extent of their movement through the mesh. This spreads the data further throughout the network in a 'virus' fashion, again increasing the probability of intercepting the recipient within a specific time interval. This approach starts to bridge the gap between Grossglauser's basic single-courier approach and a fully connected real-time route through the mesh. This is developed by Sharma and Mazumdar [13].
3. For 2 above, we suggest that the courier process might be augmented by nodes retaining a database of all other nodes they have had contact with and so selecting couriers on the basis of those that have had recent contact with the recipient.

Each of the above enhancements, which are aimed to reduce end-to-end delay, will increase the traffic activity on the network and so will decrease its throughput somewhat. This illustrates a fundamental trade-off between throughput and delay.

The potential of using 'infinite' bandwidth

An interesting alternative scenario was suggested by Negi and Rajeswaren [5], in which the channel bandwidth, B, is very large and increases as the number of nodes, n, increases. These authors suggest that this approach could achieve a per-node throughput that is an increasing function of n, i.e. the network is fully scalable.

But it would be wrong to take this further and imply that this implies self-generation of capacity. This is because the transmission bandwidth must be increased independently as the number of nodes increases. In other words, for a fixed amount of spectrum the per-user throughput will remain a decreasing function with increasing number of nodes. Increasing spectrum clearly increases capacity, but this is no great differentiator as it does so whatever system we might apply it to.

For completeness, the specific constraints of Negi and Rajeswaren's model included the following assumptions.

- Link adaptation, such that the data rate over a link corresponded to the Shannon capacity limit. The Shannon capacity limit is equal to $B \log(1 + \text{SIR})$ where SIR is the signal-to-interference ratio and B is the bandwidth. This limit is unattainable in practice and so a real-world correction factor should be applied to the results.
- Transmission bandwidth and spreading gains which are sufficiently wide to permit all interference to be tolerated even when all nodes transmit simultaneously. This is a very convenient assumption, but hard to achieve in practice.
- Transmission bandwidth and spreading gains are increased with increasing node density. Once again convenient, but hard to translate into practice, not least since presently spectrum allocations are fixed, for example for cellular.
- Propagation law γ is greater than 2.
- Automatic transmit power control (APC) is used.

Pointing out that the assumptions are unrealisable should not be taken to detract from the contribution of the work. We have already been at pains to show that wide-ranging assumptions like these are necessary for an analysis to proceed at all. In fact, the work is included here because it may give useful clues to an intermediate solution. One potential clue is that the transmission bandwidth was set wide enough to render the transmission power from neighbouring nodes to below the thermal noise floor [5].

This may encourage us to look at wideband technologies more closely in future. These include direct sequence spread spectrum and fast frequency hopping, as used in later versions of 802.11 and Bluetooth. Both these spread the signal energy over a wider bandwidth such that relatively narrower-band interference is less likely to have a detrimental effect. Newer wideband technologies also include ultra wide band (UWB) transmission. This may be quite similar to what Negi and Rajeswaren had in mind, since the desired effect of UWB is to spread the signal energy out so extremely that the signal itself is below the prevailing noise floor. We summarise modulation schemes in Chapter 5.

The benefit of traffic localisation

The theoretical analysis of Gupta and Kumar [6] assumes a random distance between source and destination nodes. Thus path lengths range from nearest neighbour (one-hop) to the full diameter of the area covered (many hops). It follows easily that as the network size increases geographically and/or in terms of node density, the average number of hops per path must increase. We know from what we have covered so far that this is one of the primary factors which cause a reduction in capacity with increasing numbers of nodes.

It is clear, then, that if traffic flows were somehow localised amongst neighbouring nodes, regardless of the geographic size of the network, then the number of hops per path would not increase with mesh and size. In this case the mesh would scale better. That is to say that the more localised the traffic flows, the more capacity can be supported and the less this is affected by growth in population.

Accepting this for the moment, this suggests that we should try to identify applications whose traffic is predominantly amongst close

neighbours. However, whilst this might have some prospect for fixed mesh applications in which the geographic location of users is fixed and known, it has far less attractive prospects for mobile applications. The reason is simply that by definition node location is not fixed in a mobile application. Having said that, it is possible to think of some cases where mobile nodes may cluster, for example around city centre features such as shops and transport terminals. But this clustering would be temporary and unpredictable. Relying on traffic to be wholly localised all the time would seem to be no way to plan a real network.

When considering traffic flows and related architectures earlier in Chapter 3, we noted that a routing hierarchy would be helpful to avoid those routes which would otherwise consist of many hops. Given the localisation aspect just discussed, it is pertinent to revisit this next and collect our conclusions.

4.1.8 Hybrid mesh network capacity, examples and conclusions

As stated earlier, one of the key limitations to the efficient flow of intra-mesh traffic through a network is that, as geographic size and/or user-density increase, there arises a conflict. This conflict is between using small hops for good spectral efficiency (see Hypothesis 2), and using longer hops for reduced delay and better route connectivity. Clearly we cannot do both simultaneously.

We have already shown that a means of addressing this conflict is to add a fixed infrastructure in the form of an overlaid network configured as a fixed mesh network. As a reminder, this is shown in Figure 4.7.

In such a system the overlaid fixed mesh adds capacity and so the overall capacity and scalability are greatly enhanced. Some foundation work on this architecture has been presented by Liu *et al.* [14], with the following conclusions.

1. If the quantity of relay nodes, m, increases at a rate less than $\sqrt{n}$ then there is a substantial increase in the capacity of the network, but the rate of decay of per-user throughput with increasing node number, n, is not improved substantially. Capacity is improved but scalability is not.

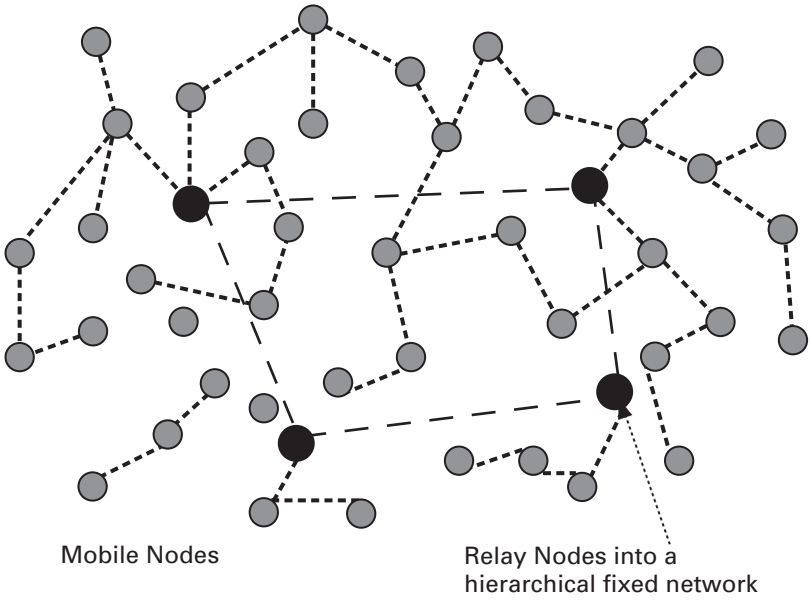


Figure 4.7 Hybrid network: intra-mesh traffic with infrastructure support.

2. If the quantity of relay nodes, m, increases at a rate greater than $\sqrt{n}$ then there is useful improvement in the scalability as well as capacity.
3. In the limit, when the quantity of relay nodes, m, increases at the same rate as the number of nodes, n, the throughput per user remains constant. The network is fully scalable.

It may be said with hindsight that the third conclusion is self-evident from the fact that in this case each relay node serves a constant number of user nodes. Therefore each node retains a constant share of the total transmission bandwidth available. At this point we have a good parallel to a cellular network where the network capacity scales directly with the number of base stations.

Boppana and Zheng [15] offer a simulation of such an architecture using 802.11 which illustrates the attainable improvement in capacity and per-node throughput with the addition of a hierarchical fixed network. In the simulation, adjacent pairs of fixed relay nodes were interconnected by a wired link. The same radio spectrum was used for peer-to-peer links between mobile nodes and for mobile-to-relay node links.

Table 4.2 *Capacity improvements in a hybrid network*

	Capacity gain (%), Boppana and Zheng [15]	Capacity gain (%), Liu *et al.* [14]
1000 nodes, no relays	reference level	reference level
1000 nodes, m=9 relays	57	66
1000 nodes, m=25 relays	190	384

For a simulated 1000-node network extending to 36km^2 in a fourth-order propagation environment, neglecting log-normal fading, the network throughput capacity is increased

by 63% by adding 9 fixed relay nodes with 12 interconnecting paths, and
by 190% by adding 25 relay nodes with 40 interconnecting paths.

These figures are of course specific to this particular model, but they do serve to indicate the potential benefits of a hybrid network. There are substantive differences between the working assumptions in Liu *et al.*'s theory and Boppana and Zheng's simulation and so direct comparison between the two can be misleading. Furthermore, the theoretical analysis is idealised and represents the limiting case for large networks. Nevertheless a comparison of the capacity gains is presented in Table 4.2.

Further benefits noted [15] were a reduced percentage of link breakages due to user mobility and reduced end-to-end delay due to lower hop count.

There is a potential down-side which may occur due to traffic bunching around the relay nodes into the hierarchical network. Clearly, in this respect, the network's performance is conditional on there being a co-operative balance between local and long-range traffic.

This hierarchical network could use radio, optical or wired links between relay nodes. Clearly, if the hierarchical network uses radio links then there is additional radio spectrum required for this network and so there is an impact on overall spectral efficiency. Whether or not the additional spectrum required is matched by the achieved gain in capacity,

such that overall spectral efficiency is not degraded, will depend on the scope for frequency re-use within the hierarchical network.

Conclusion on the hybrid mesh network

It is clear that adding a hierarchical infrastructure to a mesh network will enhance its performance substantially in terms of capacity, scalability, reliability and its ability to support long-distance traffic flows.

4.1.9 Access mesh network capacity, examples and conclusions

All of the above analysis relates to traffic that is routed between peer-to-peer users, either directly or via a fixed infrastructure mesh network. We need to build upon this since it has already been pointed out that there are very few user applications that would suit that model, particularly in the commercial and consumer sector. Far more relevant to today's applications for wireless communications are services which require access to a public network, such as the Internet. As discussed earlier in this book, there are also the potential requirements for network management and billing traffic which must flow to/from a management centre on an external network. This leads most real mesh deployments to be access meshes.

Traffic flows are centred around access points or gateways, as illustrated in Figure 4.8.

There is now a concentration of traffic around the access points and an associated higher burden of traffic flow through those user nodes which provide the final hop to the access points.

This traffic concentration substantially reduces the capacity of the underlying mesh network and the per-user throughput, compared to the intra-mesh traffic case. This arises because, in the absence of any sectoring of the access point coverage, all of the hops into the access point are contending to use the same resource. Hence the maximum combined throughput of all of these hops cannot be more than W, the node throughput.

This capacity is shared amongst the n nodes serviced by the access point, resulting in a scaling factor proportional to $1/n$. We might say the

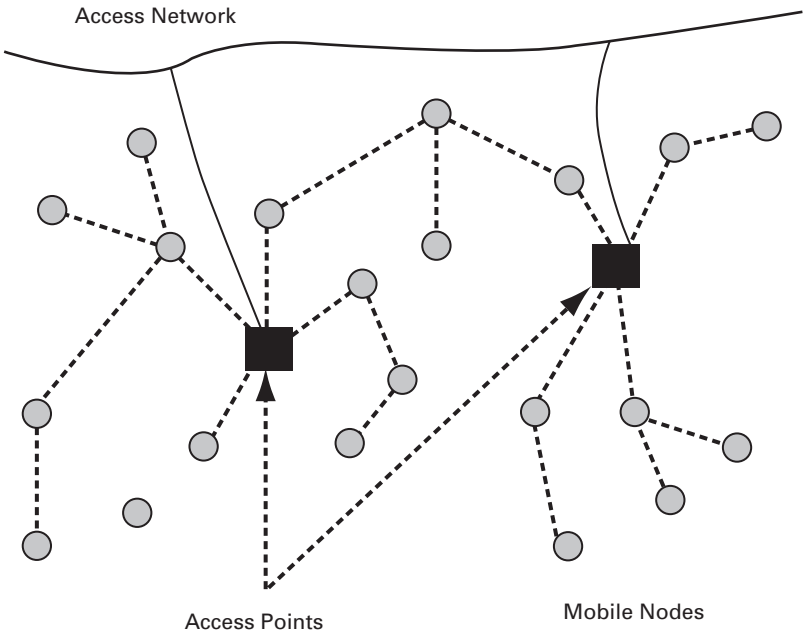


Figure 4.8 Mesh network with access points for extra-mesh traffic flow.

network has become closely akin to a cellular network, in which the capacity is set by the base station and this is shared amongst the cell's occupants.

However, in practice the interference zones of the first tier of links into the access point may also overlap most of the interference zones of the second tier of links into the nodes, and even possibly some of the third tier. For this reason, the maximum available network capacity may be practically limited to $W/2$ or $W/3$.

An example simulation of a small network with a single access point configuration is given by Jun and Sichitiu [16] This is presented in the context of a fixed mesh, but it serves to illustrate the traffic concentration issue. A network of 30 nodes using 802.11b at a raw transmission rate of 11 Mbps and an aggregate relay throughput of 5.1 Mbps (W=5.1 Mbps) is shown to achieve results which are asymptotic to a network capacity of about 1.5 Mbps. This is an average throughput per node of about 50 kbps, which implies a network capacity in the region of $W/3$. This is for a static

mesh with pre-determined, static, traffic routes. There would clearly be additional overheads for a mobile, dynamic network.

4.2 Conclusions – capacity

Mesh and ad hoc networks comprise user-nodes which relay traffic between themselves in order to achieve end-to-end traffic routes. As such, each new node joining the network adds capacity and connectivity. However, the raw statement that *as the number of nodes increases so does the capacity of the mesh* is disingenuous. It is true that network transport capacity increases with increasing node population, but in sharing this resource amongst the users there is a net decrease in the available throughput per node.

Our conclusion has been that meshes have no especially good properties with respect to scaling. In particular, as the node density and geographic size increase, the traffic rate available to any particular user decreases.

We have shown that this lack of scalability can only be overcome either by adding additional capacity in the form of a hierarchical or access network or by containing the end-to-end traffic flows to localised regions within the network. The latter approach would seem to be unrealistic.

It is principally the need to share the available mesh spectrum coupled with the need to relay other traffic which leads to a scaling of between $1/n$ and $1/n^2$ in per-user throughput.

Table 4.3 summarises the detailed findings with respect to mesh scaling.

As a final point, whilst the capacity of a cellular system can, within limits, be expanded post-deployment, mesh systems are less flexible. For a mesh, the network capacity is a function of the capability of the nodes. Network operators must therefore take a view at deployment time as to the long-term subscriber density and desired services and provide nodes that can support these. As performance depends on the lowest common denominator of the all users' equipment, it cannot be upgraded incrementally as it can for cellular. It follows therefore that nodes may have to be substantially over-dimensioned from the outset. This has clear capital cost implications.

Next we look at Hypothesis 2 which concerns efficiency.

Table 4.3 *Summary of detailed findings on capacity scaling*

Finding	Comments
1 Average per-user throughput decreases as the node population increases.	Estimated figures range from $1/\sqrt{n}$ to $1/n^2$, but the underlying trend is that throughput decreases asymptotically with increasing n and so the mesh network cannot scale indefinitely. This is due to mutual interference and to relaying traffic over multiple hops.
2 If traffic flows are localised in the network, then throughput can be independent of the total number of nodes, n.	This requires identification of specific user applications which do not demand ubiquitous flows across the whole area of the network.
3 Highest frequency re-use occurs when the hop length is minimised.	Short hop length increases vulnerability to link breakages due to mobility. There is thus a careful balance to be defined between these conflicting characteristics.
4 A higher attenuation propagation environment enables higher network capacity.	This is true to the extent that spatial re-use is increased in a high attenuation environment. However, performance is governed by the *density* of nodes: if this is too low for the prevailing environment then connectivity is poor and routing is unreliable.
5 The only way to improve substantially the scalability of a mesh network and its ability to carry long-haul traffic is to overlay a hierarchical fixed mesh network, resulting in a hybrid network.	Clearly this involves deploying a fixed, planned infrastructure and so is a departure from a stand-alone ad hoc network. The fixed network could employ wire/fibre interconnect, in which case overall spectral efficiency is not impacted, or it could employ wireless

Table 4.3 (*cont.*)

Finding	Comments
	interconnect. High spectral efficiency in the fixed network could be achieved by narrow-beam directional point-to-point links
6 The support of extra-mesh traffic requires connection to an access network.	Again this involves deploying a fixed, planned infrastructure and so is a departure from a stand-alone ad hoc network.

4.3 Hypothesis 2 – Are meshes more efficient?

Hypothesis 2 addresses the topic of network spectral efficiency. It brings practical issues together with theory. Once again we state our conclusions at the outset.

1. There seems little reason to believe that practical mobile meshes will be intrinsically more spectrally efficient than traditional cellular networks.
2. The mobility of users in the mesh can lead to a much bigger trade-off between efficiency and availability than is found for current cellular systems, since mesh users are also system routers.
3. Any preference for implementing a mesh should be based on benefits other than efficiency, such as coverage or lack of infrastructure.

Our evaluation begins with an introduction to the problem, followed by an outline of the fundamental supporting arguments for meshes offering higher spectral efficiency. We then describe the practicalities which may serve to negate this advantage.

4.3.1 Spectral efficiency

Here we assume omni-directional antennas; directional antennas form the basis of Hypothesis 3.

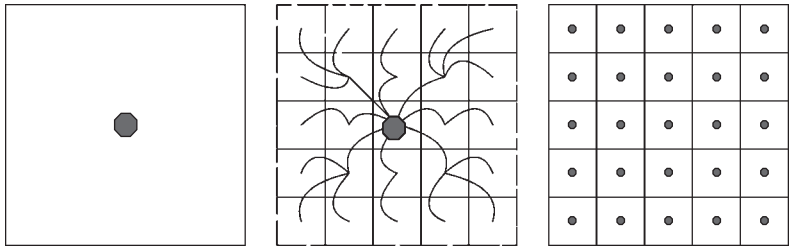


Figure 4.9 Examples of (a) cellular, (b) access mesh and (c) pico cellular networks.

As stated, a hypothesis *that mobile meshes are more spectrally efficient*, is essentially untestable without consideration of further information. For example, there are many possible deployment scenarios varying in range, numbers of subscribers, traffic flows etc. There are also several possible definitions of spectral efficiency; this book takes bps/Hz/km^2. Additional complexity arises from the fact that practical systems comprise not just a topology, but a whole range of components such as protocols, radios, antennas etc. Consideration of just one element may be misleading. For example, fixed mesh systems claim high spectral efficiency but this efficiency is known to arise principally from the use of highly directional antennas, rather than the use of the mesh architecture *per se*.

With so many variables the danger of comparison is that, by judicious choice of scenario and parameters, proponents of any given system may 'prove' the superior efficiency of their topology whether it be mesh or cellular. A similar conclusion was reached by earlier work [17] which also compared cellular and mesh architectures, although in a different context.

In order to provide a level and fair comparison, this section first addresses the question 'more spectrally efficient' *than what*? Figure 4.9 helps illustrate a simplified comparison of mesh with cellular, assuming a convenient, if practically unlikely, scalable propagation environment.

Figure 4.9(a) shows a single traditional cell. The cell is drawn as a square for simplicity. Figure 4.9(b) shows an access mesh approach with a snapshot of 25 mobile users distributed conveniently, if a little unrealistically, at the centres corresponding to the pico cell sites of Figure 4.9(c).

Figure 4.9(c) shows simple square pico cells, where 25 remote pico cell antennas are provided to serve the whole area. Figure 4.9(c) is included to show that adding more actual base stations is better than relying on the myth that each added mesh node 'effectively becomes a base station', see Section 7.1.1. The users are free to move over the whole area in all cases.

Although 25 users have been shown for simplicity in the mesh diagram, subsequent discussion assumes a 'significant' number of users, i.e. such that the user density is sufficient for the functioning of a mesh system. To simplify the discussion it is also assumed that all active users are allocated the same channel bandwidth. It should be emphasised that this is in terms of hertz, the unit of the underlying resource (spectrum), not bits per second as used by some other authors.

An alternative starting point would be to keep the delivered bit rate constant and vary the required channel bandwidth, but the argument remains the same. We already know that there are various possible access schemes such as TDMA, FDMA and CDMA, and each of these has its own benefits and drawbacks in different scenarios. However, for our purposes of considering spectral efficiency, each of these access schemes may be thought of simply as an equivalent division of the available spectral resource. This assumption makes them independent of our efficiency considerations.

To simplify the analysis, the premise is that the vast majority of commercially significant applications would generate extra-mesh traffic if implemented on a mesh. The analysis therefore focuses again on the access mesh in which all traffic flows to/from some central gateway, as shown.

If deploying a cellular system, Figure 4.9(a), the approach would be to locate a base station at some appropriate central position with enough power/sensitivity to meet the required objectives. If a fixed rate service is offered, for example a GSM voice call, then the worst-case users determine the link budgets, i.e. those furthest from the base station in terms of path loss. The data rate on the channel (or alternatively the width of the channel) is dictated by this worst-case link budget. For users nearer the base station the transmit power may be reduced. The spectral efficiency,

ignoring inter-cell effects, can then be simply calculated as the user bit rate per channel bandwidth per unit area of cell. Some cellular systems, for example 3G, offer a graded service to users whereby users nearer the base station are offered higher bit rates. This grading of service will serve to increase the spectral efficiency.

Now consider the introduction of hopping, in a mesh system, Figure 4.9(b). In principle the worst that can be done is the same as the cellular system, since it can default to a one-hop service to all users. In fact for users near the base station or access point it probably will use a single hop. For more distant users, however, it is possible to improve the delivered bit rate by transferring the data in a number of shorter hops, as we will show shortly in Section 4.3.3. Such gains, however, depend on user behaviour (distribution, mobility), quite unlike the predictability of a cellular system.

Note that the efficiency of the pico cell system, Figure 4.9(c), may never be approached by the mesh system, since the average path length within the mesh will always be longer, plus the pico cell air interface is not taxed by any backhaul or relay of other user's traffic. However, a fixed infrastructure is needed to backhaul traffic. Whilst the pico cell wins on spectral efficiency grounds, it is often discounted on financial grounds. It should be clear that this is a wholly different basis on which to argue and one which may be already changing as RF-on-fibre pico cells become available.

It is noteworthy that, even when effort is taken to compare like with like, a complication remains due to the fact that meshes in particular have a performance (including efficiency) which depends on instantaneous user-node distribution and mobility. However, this is not unlike the adaptive behaviour of 3G systems as mentioned above. Whilst a grading of service will serve to increase the spectral efficiency for some user distributions, it is not a guaranteed effect. The same is true for a mesh; a worst case ought to be used for mesh planning. The difference between worst case and best case in a mesh can be very high, since it is driven by user mobility, as already stated.

We now turn to the spectral efficiency of pure mesh and access mesh networks relative to cellular equivalents.

4.3.2 Comparative efficiency of pure mesh and cellular

To tackle the efficiency question, we must first recognise that none of our analyses within Hypothesis 1 on capacity quantified the absolute capacity and thereby spectral efficiency. They addressed only scalability, which concerns relative capacity. Thus the important question remains: What is the absolute spectral efficiency of a small mesh and how does this compare with a cellular network?

But direct comparison is not easy, not only because of the very large range of parameters involved and assumptions about functionality that must be made, but also because a cellular network can increase its capacity continually by deploying more base stations. Thus a first-stage comparison must be between a single base station cell and a pure mesh serving the same population over the same geographic area. Unfortunately most of the published work addressing absolute rather than relative measures of capacity makes use of IEEE 802.11 protocols rather than cellular networks. Nevertheless the Boppana and Zheng [15] simulation, which we used in Hypothesis 1, can be used to provide a first-order estimate of the absolute capacity of a pure mesh.

Let us consider the situation in which an isolated GSM cell provides services between users within its coverage area, thereby offering a parallel to intra-mesh traffic flow. For GSM telephony, of the order of 8 kbps in each direction is needed so a 2 Mbps capacity base station can support 125 duplex channels. GSM systems provision base stations on the basis of about 0.06 Erlang/user, so 2 Mbps should support of the order of 1600 users (leaving some headroom for peak loads). This figure is close enough for comparison by extrapolation from the Boppana and Zheng simulation of 1000 users to be worthwhile. Their simulation achieved a total usable capacity of about 500 kbps so that, using the capacity-scaling law deduced from their results (capacity $\propto n^{0.18}$), it would be anticipated that a mesh of 1600 users would be able to support an aggregate throughput of about $500 \times (1600/1000)^{0.18} = 544$ kbps which is equivalent to only 34 duplex channels.

In other words, even when providing an intra-mesh service, the mesh only achieves about 25% of the efficiency that a single-cell system

might be expected to achieve. And this excludes the effects of node mobility.

Furthermore, if the number of users increases either by increasing density within the same area or by increasing numbers over an extended area, then the mesh network capacity diminishes, whilst the cellular network can simply deploy more base station capacity. Thus the pure mesh network cannot possibly match the cellular equivalent.

The only way that the pure mesh could compete is by departing from purity by adding infrastructure in the form of an overlaid backhaul network to transport the longer range intra-mesh traffic, as we illustrated by the hybrid mesh with its hierarchical relaying architecture. This architecture effectively breaks the overall mesh down into a number of smaller, but still connected, meshes and so improves scalability and capacity. The same is true for the access mesh, and we have already noted that the access mesh is a good parallel for a cellular network, in terms of scalability. To understand whether the access mesh could offer improvements over cellular, for similar applications, we must understand the multi-hopping process.

4.3.3 Efficiency of multi-hopping

One of the traditionally used scenarios for suggesting that mesh operation into an access point might be more spectrally efficient than a base station cell is the concept that increased throughput can be achieved over a series of short hops rather than one long hop. It will be demonstrated that this is only true for an idealised single-path scenario, and is diminished by the dissimilar antenna gains of access points and mobiles, the overhead of relaying traffic from multiple users, the spectrum re-use contention within the mesh and finally, of course, the routing overheads to combat node mobility.

For the case of hopping between nodes of like type, consider node-to-node links in a mesh. If two hops of roughly equal length replace a single hop, as shown in Figure 4.10, then

- only half the time-bandwidth product of spectral resource is available for each hop, and this acts to reduce the delivered data rate by a factor of 2,

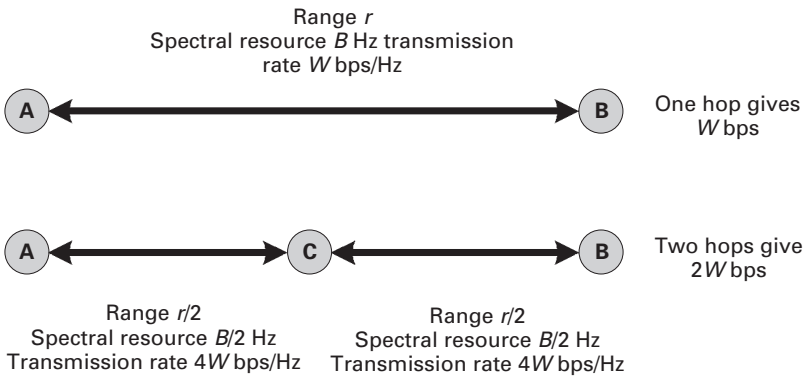


Figure 4.10 Two-hop versus one-hop rate improvement between mesh nodes.

- but as each hop is half the length of the original link, the link budget is improved. This improvement can be used to improve spectral efficiency either by increasing the transmission rate on each hop or by reducing the transmit power. For example, in a third-law propagation environment the link budget is improved by eight times (~9 dB); this would permit a four-fold increase in transmission rate by changing from QPSK to QAM64. Alternatively, with spread-spectrum the coding gain could be reduced to realise a similar increase in transmission rate.

Overall these two factors imply that twice as much data can be transferred using two shorter hops, i.e. spectral efficiency is doubled. But this only prevails when the path length is exactly halved. If instead there is asymmetry in the two-hop path lengths then the link budget gain in the longer hop will diminish and so the higher rate becomes unsupportable. This 'sweet spot' in the path length split is illustrated in the graph of link budget in Figure 4.11. The two-hop chain resultant has a peak at 0.5 (i.e. the centre), which tails off away from the centre in either direction, hence the term sweet spot. Overall, the two-hop chain is only as good as the weakest link, which must always be defined as the locus described by the two lower portions of the lines on the graph.

Note that, although there are now twice as many mesh transmitters using the same power, the interference situation is not necessarily degraded as each transmitter occupies only a fraction of the bandwidth.

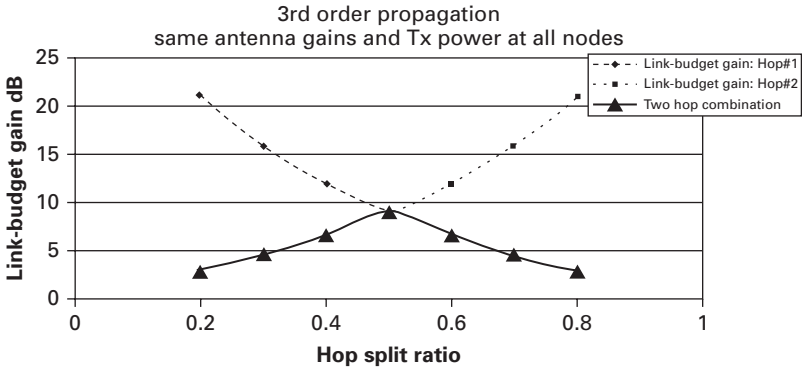


Figure 4.11 Two-hop link budget gain compared with the single-hop case.

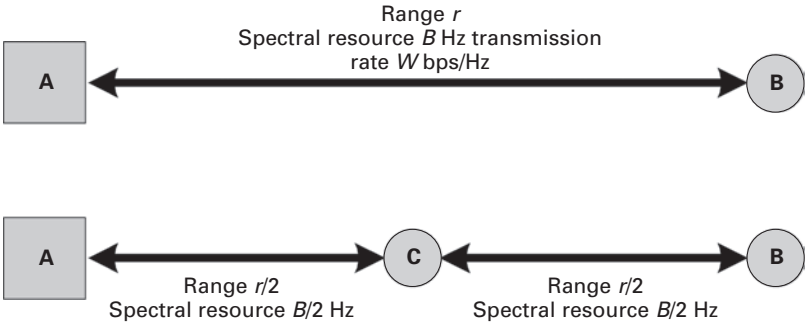


Figure 4.12 Two-hop versus one-hop into high gain access point.

The amount of interference is relatively unchanged, although its footprint is different.

The comparative performance is further eroded for the case of multi-hopping into a mesh access point or cellular base station as represented in Figure 4.12.

The hop(s) between mobiles lack the higher antenna gain and height of the link into the access point (item A in Figure 4.12). Due to this imbalance the 'sweet spot' does not occur at the 50:50 path-length split. The graph of Figure 4.13 illustrates this for the case when the access point antenna gain is just 13 dB above the mobile nodes' gain – the 'sweet spot' has moved to approximately 75:25 path-length ratio

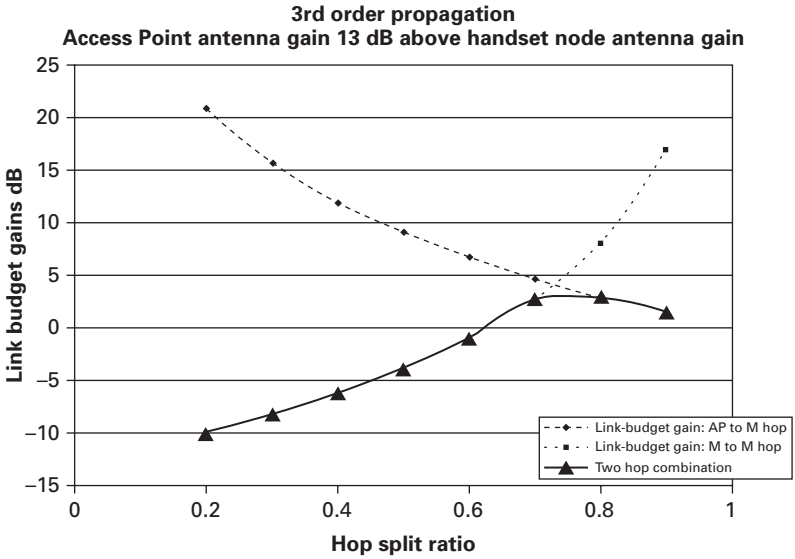


Figure 4.13 Two-hop versus one-hop link budgets with high antenna gain.

and the optimal link budgets on the two hops are only about 3–4 dB above the single-hop case. With this small link budget gain the transmission rate might be little more than doubled, but the spectral resource is still shared between the two nodes. Thus the best-case throughput rate of this specific two-hop route is the same as for the single-hop route.

Continuing the analysis to a three-hop scenario, each hop would now be allocated one-third of the spectral resource. The 'sweet spot' occurs at about 60:20:20 path-length split and at this point the link budgets are about 7 dB above the single-hop case. This link budget gain might just support a tripling of transmission rate and so again achieve about the same throughput rate as the single-hop case.

Extended route length

An implicit assumption in the above simplified analysis is that the total multi-hop path length is the same as the single-hop path length. In practice, of course this will not be the case; nodes will be unevenly distributed and routes may have to circumvent building and terrain clutter.

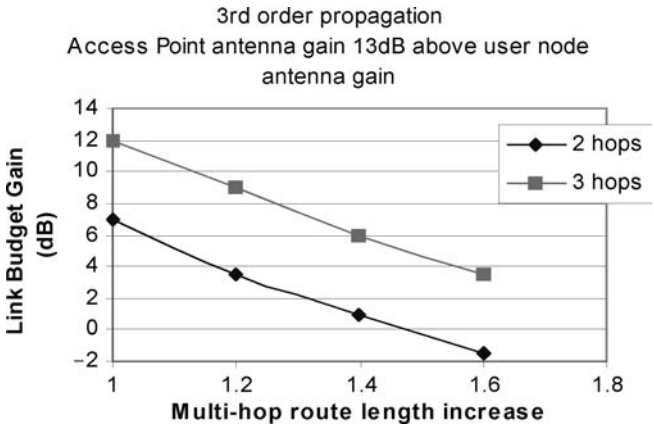


Figure 4.14 'Sweet spot' link budget gain versus extension in total route length.

The detrimental effect of increased route length is illustrated in the graph of Figure 4.14 which illustrates the reduction in link budget gain as the route length is increased.

The figure shows that, for example, for a two-hop system where the two-hop distance is no larger than the single hop, then about 7 dB advantage is to be had. If, instead, the two-hop route length is 50% longer than the single-hop path, then the two-hop advantage completely disappears.

On the basis of the above, simplified, analysis we may conclude that multi-hopping may rarely be much more spectrally efficient than single hopping, in practical situations.

We next look at how the propagation environment affects the multi-hopping situation.

Propagation law

The gains in link budget through multi-hopping increase for higher propagation law. This is illustrated in Figure 4.15 for idealised two-hop and three-hop paths.

Here we are showing that multi-hopping works better in a high order propagation environment. A little consideration reveals that this is the

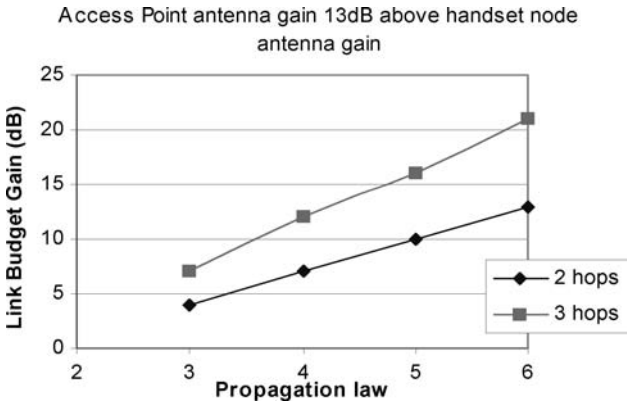


Figure 4.15 'Sweet spot' link budget gain versus propagation law.

basis on which multi-hopping can be used to circumvent high clutter losses on a single-hop path. This is in fact a primary argument in favour of mesh networking and is the basis for coverage improvement in our 'cellular multi-hopping' application example of Chapter 2.

We will pick up this point again when we cover Hypothesis 4 on spectrum utilisation. This is because one way to ensure a high order propagation environment is simply to work higher up the spectrum.

4.3.4 Practical mesh networking issues

The above discussion has addressed the fundamentals of multi-hopping in isolation from overall network issues. In a network, a number of practicalities will further erode spectral efficiency, these include

- traffic concentration,
- mobility and routing overheads, and
- efficiency/availability trade-off.

Traffic concentration

When we covered Hypothesis 1 on capacity, we highlighted that the resource contention around a mesh access point will erode capacity more than in the base station cellular case.

Mobility and routing overheads

As we saw when covering the fundamentals, the introduction of hopping in a mobile mesh brings a requirement for dynamic routing. Even leaving aside the need for route discovery, this adds an overhead which will diminish spectral efficiency. This is because, when measuring efficiency, only user bits should be counted, not bits used by the routing protocol. In other words, the protocol bits are an overhead. Looking ahead to Chapters 5 and 6 where we discuss routing and quality of service, we will see that routing overheads may be as high as 100% in a simple packet-based approach.

Efficiency/availability trade-off

In current cellular systems, 100% coverage availability is not guaranteed; instead outages of 2% to 5% are often used in planning. This is done to conserve base station power levels and hence spectrum efficiency. Further increases in availability and capacity may be achieved by in-filling cell coverage with pico cell sites but this is expensive, so is not always done. The resulting compromises set upper limits to cellular coverage, capacity and spectral efficiency.

The multi-hopping inherent within a mesh may increase coverage availability, but this service level enhancement is dependent on the statistics of the movements of users. The issue is that each user node is completely free to move – and may move to a position where coverage is lost. This must also affect any nodes that are downstream. With mesh systems the operator is not in control of all the network routing elements. The resultant compromise may well include a notably larger probability of mesh network outage, relative to current cellular systems.

4.4 Conclusions – omni-directional antennas

There seems little justification for a belief that practical mobile meshes will, as a general rule, provide a significantly more spectrally efficient means of carrying the commercially significant applications of today and the foreseeable future. This is for a number of key reasons.

- Realistic meshes will have a performance which will depend on user distribution and mobility behaviour. In some cases they could be comparable to current cellular systems. In the best case, where the environment allows, the efficiency could be very high on occasion, but the danger is that availability could dip to very low levels as users move. In other words, the mobility of users within the mesh can lead to a much bigger trade-off between efficiency and availability than for current cellular systems, since mesh users are also system routers.
- Gains predicted by multiple hopping may be reduced in the real world due to the diluting effect of unequal length hops in a path and the unequal antenna gains of user node and access points. Mobile to mobile link budgets are also likely to be far less than base station to mobile budgets for practical, height related reasons.

It seems reasonable to assume that, by creating small mesh islands through the use of fixed seed and routing nodes, dependence on user behaviour can be lessened. Several current commercial mesh deployments do this, for example the approach of mounting mesh routers regularly on street lighting in urban deployments. Of course, this is in support of hybrid and access mesh architectures, in preference to pure meshes.

4.5 Hypothesis 3 – Do directional antennas help a mesh?

This is the third of our real questions from people in the industry. We have hardly touched on directional antennas so far, except where we looked briefly at physical versus logical meshes in Chapter 3. There are at least four potential incentives for the use of directional antennas in any wireless network, each of which may have relevance to mobile mesh networks. In each case, listed below, the antenna gain is being exploited in slightly different ways.

1. High antenna gain usually means high directivity, which leads to narrow beams. This means that interference is less widespread and hence a greater spatial re-use of frequency may be achieved in a given area.

2. High antenna gain enables reduced transmitter power consumption for a given range and transmission rate. This can be significant, especially noting that the antenna gain appears at both ends of the link, and so the net effect on link budget is equivalent to twice the antenna gain.
3. High antenna gain permits a greater range for a given transmitter power input. This is not necessarily advantageous since it increases the interference footprint. However, it may be useful when constructing a limited number of backhaul or backbone links.
4. High antenna gain permits a higher transmission rate for a given transmitter power input.

Before we delve into any detailed efficiency calculations, we need also to recall that our interest is the six application examples of Chapter 2. These mostly require nodes to be portable or mobile, for example in vehicular ad hoc networks (VANETs). It is of great relevance to consider whether directional antennas are appropriate to such applications.

We begin by looking at what the steering and control requirements are for a directional antenna. We follow this by a consideration of the manufacturability of directional antennas for our preferred applications.

4.5.1 Antenna steering

There will be a finite overhead for antenna control, as the antenna must be steered, by some process, to match the direction of the wanted signal. The steering requirements below are independent of whether or not we are using a mesh.

In terms of performance demands, the required speed and accuracy of beam steering must increase as the beam width gets narrower. Hence, for a wide beam width antenna this overhead should be relatively lower. This is because in most environments the beam width will be sufficient to collect all the significant multi-path components, and so the steering dynamics can be relatively slow.

However, if we consider null steering, where the desire is to 'tune out' an interferer, the overhead is likely to be high because the null must align with the most significant multi-path component. Moreover, the orientation of this main source of interference can change rapidly in a

cluttered environment under Rayleigh fast-fading conditions. We explain null steering below.

Requirements for vertical directionality

We must remember that in practical deployments, users will be distributed in the vertical plane and not just the horizontal. As a consequence we would ideally want our antenna steering to cover the vertical axis. The alternative would be to ensure the beam width is sufficient to encompass the range of height differentials we might reasonably expect. Of course in reality this means that we will probably tolerate some outage due to vertical beam misalignments.

Vertical steering is a significant further complication to implementation and we know of no commercial examples of this. We must thus expect to suffer a penalty from this both due to outage as just described and due to a smaller antenna gain when looking off-axis.

Null steering

For completeness, we should explain null steering. The alternative to steering the antenna to maximise the signal is to steer the antenna to minimise the interferer. This can be accomplished by steering a null in the direction of the dominant interfering signal. In practice, the two approaches can be combined, by steering the antenna to maximise the signal to interference ratio.

Null steering is naturally limited to being a receiver function, for the following reasons.

- There is a reciprocity problem. In general one cannot reliably assume that the direction of a dominant interferer towards a node's receiver necessarily also gives us the best direction to avoid when transmitting.
- Even if we could determine the direction we should best transmit, it is not necessarily possible to steer a transmitter beam as accurately as a receiver. This is due to the higher powers involved in a transmitter and the consequently greater difficulty of finely adjusting the phase and amplitude.

A likely consequence is that a null steered receive antenna will be operated in simple beam steering mode during transmit. Thus, overall, it is unlikely that null steering will give any improvement over simple beam steering

4.5.2 Performance and manufacturability

Our aim in this section is to look at what sort of performance we might expect from relevant real-world directional antennas. We are considering directional antennas as a means of reducing mutual interference within the network, rather than extending range. This is the first in our earlier list of four reasons for using directional antennas. In other words we are more interested in what we may achieve for our system via directionality rather than via gain. Just as for beam steering, our discussion does not need to assume our system is a mesh.

Ideal antennas

Let us begin by considering an idealised antenna. This will have negligible side lobe responses and can therefore can be represented by the 'flat top' model. Here the antenna beam in the azimuthal (horizontal) plane is represented by an arc of a circle subtending an angle equal to the 3 dB beam width of the polar response.

This allows us to draw two simple pictures of the cases when two beams will interfere and when they will not, as illustrated in Figure 4.16. This requires no more than considering whether the beams cross at the position of the receiver; crossing elsewhere is not a problem.

For a network of randomly deployed nodes equipped with such antennas, the theoretical upper limit on the improvement of throughput capacity is as large as $4\pi^2/\alpha\beta$, where α and β are the beam widths of the transmit and receive antennas respectively [18]. Interestingly, this result can be confirmed geometrically by noting that the probability of a node falling within either beam is reduced by a factor $\alpha/2\pi$ and $\beta/2\pi$, compared to the omni-directional case. Thus the reduction in mutual interference is related to the product of these two probabilities.

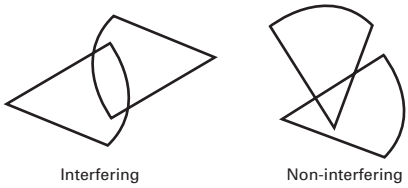


Figure 4.16 Interference model for directional antennas.

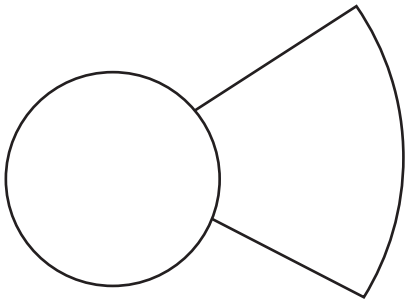


Figure 4.17 Radiation pattern for a composite antenna model.

By way of example, recall that we are primarily concerned with mobile applications. If we look forward to a result to be shown under Hypothesis 4, this means that we are likely to be working somewhere in the range 1–6 GHz, for propagation reasons. The relevant industry rule of thumb tells us that minimum achievable antenna beam width is likely to be in the region of 90°. For this beam width, the throughput equation above tells us that the improvement in system capacity is likely to be of the order of a factor of 16.

This is attractive. But let us next move on from the ideal, closer towards reality.

Real-world antennas

Any practical antenna has finite side lobe responses. Our aim now is to evaluate how these might affect the idealised gain we predicted above. Let us suppose our antenna has a uniform side lobe response outside the main beam, as illustrated in Figure 4.17.

The key effect of having a finite side lobe response is to extend the interference boundary around nodes [18], since we have increased the transmit footprint. We also know that the physical extent of this boundary is governed by the attenuation factor of the propagation environment. Let us try to put some numbers into our evaluation.

If an antenna has a mean side lobe level which is κ dB below the main beam, then we can show that in a propagation environment with propagation law γ, the differential coverage range, Δ_R, between main beam and side lobe is given by

$$\kappa = 10\gamma \log(1/\Delta_R), \quad \text{which implies} \quad 10\log(\Delta_R) = -\kappa/\gamma.$$

Looking once again for rules of thumb from the antenna industry, we find that for mobile or handheld products operating below approximately 6 GHz, the side lobe response is unlikely to be more than about 10–15 dB below the main beam. In other words we cannot build a better antenna than one which radiates unwanted power only 10 or 15 dB below the wanted power. We can calculate where the side lobes will cause an interference boundary as follows. If we take a likely figure for side lobe level of $\kappa = 13$ dB, then in a fourth-law propagation environment Δ_R is only 0.5. This means that in terms of distance from the node, the interference boundary for the side lobes is half that of the main beam, which is highly significant.

We next need to translate this into a system capacity improvement factor, so we can compare it to the factor of 16 which we found for an ideal antenna. We can do this via the same reference work as before [18], which indicates that the theoretical maximum capacity gain factor for an idealised random network is of the order of:

$$1/\left[(\Delta_R)^2 + \left\{1 - (\Delta_R)^2\right\} \alpha\beta/4\pi^2\right].$$

Let us put in the numbers again, using the same 90° main beam width, but now with −13 dB side lobes. This implies a capacity gain of only a factor of 3.3, versus the factor of 16 from earlier. This illustrates the detrimental

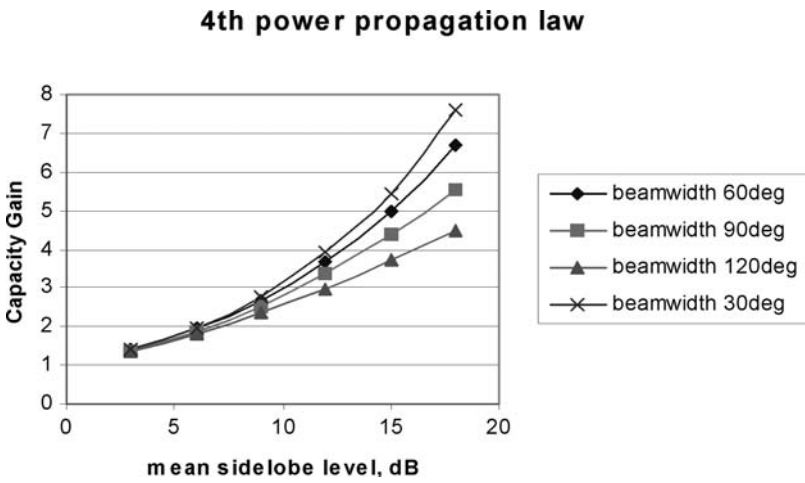


Figure 4.18 Theoretical capacity gain versus antenna performance.

effect of finite side lobe levels. In fact we can plot the influence of side lobe level as shown in Figure 4.18.

The figure illustrates that the capacity gain factor is a more sensitive function of side lobe level than it is of beam width. Furthermore, as the beam width is reduced the side lobe level dominates performance, thus indicating that there is little benefit in decreasing beam width without paying equal attention to reducing side lobe levels.

4.6 Conclusions – directional antennas

Our discussion has not had to touch on mesh at all to reach our conclusion. We have shown that there are two problem areas for directional antennas in wireless systems, which are particularly important from our point of view. One is due to needing to cope with the mobility of users and hence the introduction of the challenge of antenna steering. The other is due to the problem of actually making directional antennas which are small and efficient enough to suit handheld devices at typical useful mobile frequencies.

We therefore feel that we must discount directional antennas for mobile wireless systems, including mesh.

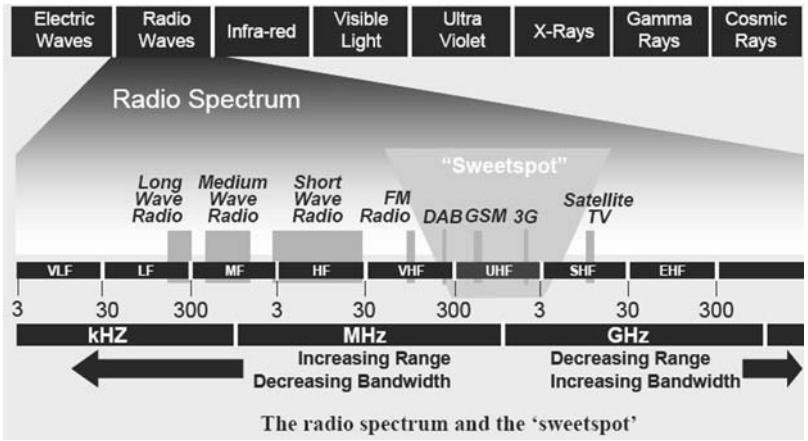


Figure 4.19 Spectrum 'sweet spot' used by kind permission of Ofcom.

4.7 Hypothesis 4 – Do meshes improve spectrum utilisation?

Our final question, in our group of four which are commonly asked of meshes, relates to spectrum utilisation. This is a different and wider issue than technical spectrum efficiency, where we have already shown that meshes do not generally offer higher spectrum efficiency than cellular.

Spectrum utilisation is an example of both functional and economic efficiencies, which look beyond basic bps/Hz/km^2 measures. Functional efficiency relates to how well an application fulfils its intended purpose, whilst economic efficiency relates to the cost of the solution, which may include the cost of acquiring spectrum.

To consider utilisation, therefore, we need to look at the spectrum space available and ask where a mesh might work well and also be cost efficient. We are able to draw on much experience with cellular systems and use these as a benchmark. Cellular systems are concentrated in a sweet spot of spectrum, which we discuss next.

4.7.1 Spectrum 'sweet spot'

The sweet spot, shown in Figure 4.19, is favoured by PMP and cellular systems worldwide. GSM and 3G are the examples shown in the figure.

Propagation in this region is good in open spaces, into buildings and through foliage. This is an attractive combination, hence the term sweet spot.

If we try to work at higher frequencies with cellular or PMP systems, we can run into reliability issues as propagation suffers in some environments. This represents a loss of functional efficiency. Although this problem could be overcome, in part, by working at higher transmit powers, this leads to problems with spectral efficiency and the increased system cost leads to a fall in economic efficiency.

However, recall that for mesh, from Section 2.2 onwards, we have shown that meshes do not rely on the same propagation behaviour as do cellular systems. Specifically, the potential of increased end-to-end throughput using multi-hopping is best realised when there is a high propagation loss. In particular, this is true when there is high clutter loss due to in-building penetration and the 'urban canyon' effect (see Section 4.3).

4.7.2 Use of less precious spectrum

So meshes are good users of high carrier frequencies based on their propagation behaviour. In fact there are two more drivers for mesh to work at higher frequencies.

Firstly, to achieve useful per-user throughput, the relaying capacity of mesh nodes needs to be high (e.g. the need for a high value of W in Section 4.2). Thus meshes need access to large allocations of bandwidth which are typically only available at higher frequencies.

Secondly, throughput latency remains a significant issue with any multi-hopping network. To some extent this can be reduced by using a wide bandwidth air interface whereby shorter frame structures in a TDMA architecture can be used. This again points to a wide bandwidth channel allocation and hence high carrier frequency operation for meshes.

4.8 Conclusions – utilisation

We have suggested that spectrum may be best utilised by targeting mesh systems to use 'less precious' spectrum, e.g. up to 6 GHz using short

line-of-sight (LoS) links, hence leaving the present cellular frequencies for best use where they are most needed, for longer-range cellular propagation environments.

4.9 Summary of hypothesis testing

The hypothesis testing resulted in the broad conclusions listed in Table 4.4, subject to the caveats and assumptions in the text of each discussion.

Table 4.4 *Summary of capacity constraints*

Capacity self-generation	FALSE	For a pure mesh. If a hybrid mesh with access points is considered, then with sufficient planned access points, the capacity may be made to scale, employing much the same principle as cellular.
Spectral efficiency	FALSE	A practical mesh does not intrinsically have a higher spectral efficiency than cellular. Meshes can, however, be very good at providing efficient coverage in certain situations.
Significant directional antenna benefits	FALSE	Directional antennas can out-perform omni-directional antennas, but antenna performance will be limited by physical size constraints at <6 GHz operation, and the complexity of the algorithm to control such antennas may be high. Overall, the realisable net benefit is not likely to be significant
Spectrum utilisation	TRUE	It is possible that the use of meshes will allow less precious spectrum to be utilised (e.g. up to 6 GHz), as a consequence of combating high clutter loss and high path loss with diverse routing and multiple short hops

Having judged the validity of our hypotheses, by considerations entirely at the PHY level, we may now move our considerations of mesh further up the protocol stack. This begins in Chapter 5, as we consider the influence of the MAC whilst discussing mesh susceptibility.

References

1. Methley S. *et al*., Efficient mobile mesh networking: attractions, myths and techno-economic roadmap to successful commercial innovation, *IEEE DySpan*, Baltimore, MD, 2005.
2. Methley S. *et al*., Efficient mobile mesh networking: testing scalability hypotheses, *IEE 3G and Beyond*, London, 2005.
3. Grossglauser M., Tse D., Mobility increases the capacity of ad hoc networks, *Proc. IEEE INFOCOM '01*, Anchorage, AL, April 2001.
4. Shepard T., A channel access scheme for large dense packet radio networks, *Proc. ACM SIGCOMM '96*, August 1996, Stanford University, CA.
5. Negi R., Rajeswaren A., Capacity of power constrained ad-hoc networks, *Proc. IEEE INFOCOM*, Hong Kong, China, March 2004.
6. Gupta P., Kumar P.R., The capacity of wireless networks, *IEEE Trans. Information Theory*, **46**, pp. 388–404, March 2000.
7. Jones S., Levine P., Rickman N., *The Economics of Radio Spectrum*, Ofcom Report, 2003.
8. Arpacioglu O., Zygmunt J.H., On the scalability and capacity of wireless networks with omnidirectional antennas, *Third International Symposium on Information Processing in Sensor Networks (IPSN 2004)*, Berkeley, CA, 27–28 April 2004.
9. Hekmat R., Mieghem P., Interference in wireless multi-hop ad-hoc networks and its effect on network capacity, *Wireless Networks*, **10** (4), pp. 389–399, July 2004.
10. Li J., Blake C., Couto D.D., Lee H., Morris R., Capacity of ad hoc wireless networks, *Proc. 7th ACM Int. Conf. Mobile Comput. Netw. (MobiCom '01)*, 2001, pp. 61–69.
11. Gupta P., Gray R., Kumar P.R., *An Experimental Scaling Law for Ad Hoc Networks*, Technical Report, University of Illinois at Urbana Champaign, available at http://decision.csl.uiuc.edu/#prkumar, accessed May 2001.

12. *High Performance Mesh for Dense Metro WiFi*, MeshDynamics White Paper, available at http://www.meshdynamics.com/Presentations.html, accessed 11 February 2005.
13. Sharma G., Mazumdar R.R., On achievable delay/capacity trade-offs in mobile ad hoc networks, *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Cambridge, UK, March 2004.
14. Liu B., Liu Z., Towsley D., On the capacity of hybrid wireless networks, *Proc. IEEE INFOCOM* 2003.
15. Boppana R.V., Zheng Z., Designing ad hoc networks with limited infrastructure support, *IEEE Consumer Communication and Networking Conference (CCNC)*, January 2005.
16. Jun J., Sichitiu M.L., The nominal capacity of wireless mesh networks, *IEEE Wireless Communications*, **10**, pp. 8–14, October 2003.
17. *Examination of Issues Related to Spectrum Efficiency of Point-to-Multipoint and Mesh Multimedia Wireless System Architectures Proposed for 40.5–43.5 GHz*, Radiocommunications Agency Report Ref. 1205/AE/MWS2/R/3, April 2000.
18. Yi S., Pei Y., Kalyanaraman S., On the capacity improvement of ad hoc wireless networks using directional antennas, *Proc. 4th ACM Int. Symp. on Mobile Ad Hoc Networking and Computing*, Annapolis, MD, 1–3 June 2003.

5 Mesh susceptibility

Continuing in the spirit developed in the previous chapter, rather than looking at meshes by pursuing a linear layer-by-layer exposition of the protocol stack as in Figure 3.1, we will continue to take a more pragmatically integrated view. This chapter and the next chapter therefore look at two key aspects of mesh systems, or indeed of any communications system; these are susceptibility to interference and quality of service. PHY, MAC, routing, transport and application behaviours along with their interactions are all relevant, although this chapter on susceptibility is more related to the lower layers and Chapter 6 on the quality of service is more related to the higher layers.

We begin by looking at interference and how the mesh may react to it. We do this by firstly classifying all the various forms which interference may take.

At the physical layer the effect of interference depends on the modulation and coding in use within the mesh. Of course this is true of any communications system, but we find an important distinction is that a mesh precludes the easy use of some common modulation approaches. The reason for this is the typical lack of any centralised control within a mesh, which precludes approaches demanding synchronisation of modulation across nodes. Examples include many versions of frequency hopping.

At the MAC, the effect of interference depends on the MAC scheme in use. Once again this is true for any communications system, yet again we find an important distinction is that a mesh precludes the easy use of many common MAC approaches. This includes the common slotted schemes of FDMA, TDMA and CDMA. The reason for this is the same as before: the typical lack of any centralised control within a mesh. Meshes thus usually employ random access MACs which are effectively co-operative schemes, having an element of 'politeness' built into the protocol. Such

politeness usually means listening before talking. Co-existence with meshes thus depends heavily on the politeness on the interfering protocol. It will become clear that if an interferer has no politeness built into its protocol (perhaps it is a deterministic time slotted protocol) then the mesh will simply allow the interferer to dominate. On the other hand the mesh may route around such an interferer.

Routing around an interference problem brings in the behaviour of routing protocols, which are designed to perform efficient re-routing, notably in response to the mobility expected of ad hoc meshes. However, any re-routing causes a temporary disruption in transmission and thus in turn brings into consideration the behaviour of the transport protocol. We look at basic routing and transport questions within this chapter, but such a discussion rapidly brings us into the major area of quality of service, which we discuss principally in Chapter 6, with particular regard to how it is supported in the presence of node mobility.

Finally in this chapter, we broaden our outlook to look at co-existence rather than simply susceptibility.

5.1 Interference types

Understanding the interference issues with any radio system is always important. Generally, as we have seen, meshes operate differently from conventional systems and so require a renewed level of investigation. Moreover the issue of interference itself is likely to grow in importance in the future as new attitudes to spectrum policy spread around the world. Notably many modern attitudes centre around a lightening of the regulatory approach and therefore an increased interest in the co-existence of different technologies and services. We will look at co-existence at the end of this chapter; at the moment our focus is susceptibility and interference.

There is a variety of types of interference which may affect a radio system. Figure 5.1 shows a simple breakdown of interferer categories which we will use in discussing the effect of interference on a mesh system. We will use the normal convention of referring to wanted signals and interferers in our descriptions.

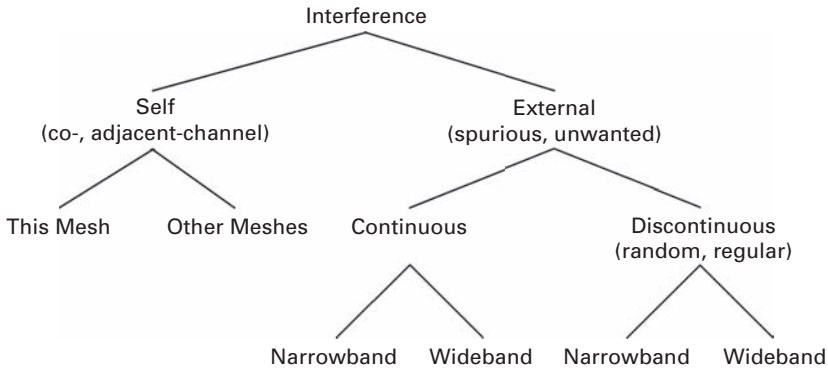


Figure 5.1 Types of interference.

Let us first make the distinction between self-interference and external interference. In the first case, the modulation, time and frequency characteristics of the interfering signal are the same as those of the wanted signal and this interference is commonly manifested as co- and adjacent-channel interference in practical situations.

In the second case, external interference refers to all other forms of interference and will therefore typically have modulation, time and frequency characteristics that are significantly different to those of the wanted signal. The interference commonly arises from unwanted or spurious transmission effects.

Self-interference may result from the nodes of a receiver's own mesh or from the nodes of another mesh which uses the same architecture and protocols, for example the case of two competing mesh service providers. In a single mesh case, the nodes could, at least in principle, exchange information which allows the interference to be mitigated, such as by reducing its power at source. But no other types of interference are controllable in this way and must be dealt with through the appropriate design of the radio system, its architecture and its protocols.

The time and frequency characteristics of external interference are of particular significance. A key time characteristic is the duration of the interference, specifically whether it appears to be continuous or discontinuous in nature. Discontinuous interference includes both random

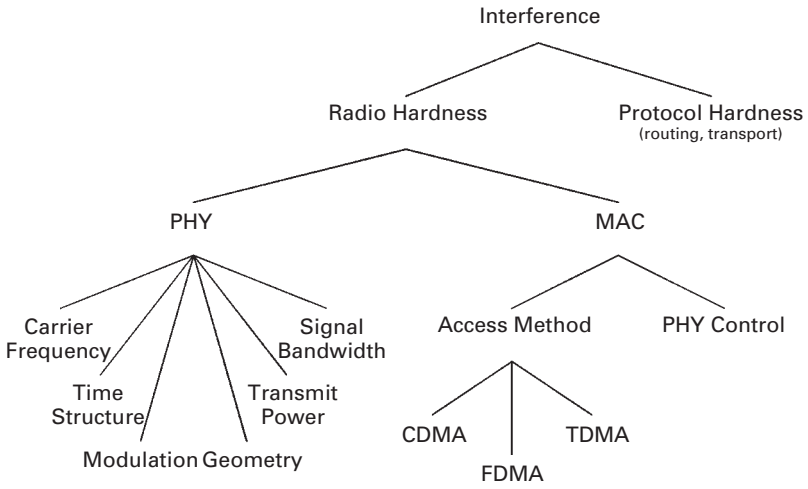


Figure 5.2 Elements affecting interference susceptibility.

interference, for example from a microwave oven, and regular interference, for example from a DECT cordless phone. A key frequency characteristic is the bandwidth occupied and in particular whether the bandwidth of the interference is small (narrowband) or large (wideband) compared to that of the wanted signal.

For completeness it is worth noting a distinction between real interference and virtual interference, although each is equally problematic to the system. By real interference, we mean any unwanted radio energy which appears in the frequency band occupied by the wanted signal. By virtual interference we mean radio energy which originates at other frequencies but which is converted to an in-band effect by receiver imperfections such as the practical limits of the performance of mixers and filters, such as intermodulation levels. In other words, if the interference effect appears in-band, we are not concerned with whether the original interferer was in-band or not.

Let us first look at susceptibility due to PHY and MAC issues.

5.2 Susceptibility to interference – PHY and MAC

Figure 5.2 shows the key elements of a radio network which affect its susceptibility to interference. The elements may be either dynamic or

static, meaning that they may or may not be able to change their characteristics in response to interference problems. The dynamic elements tend to be those in the PHY layer, for example a shift of carrier frequency or modulation level to avoid or reduce the effects of interference.

Even in carefully planned, cell-oriented, hierarchical networks where we have much experience, dynamic control of elements is complex. In mesh systems, stable control is likely to be even more difficult to achieve since control is typically decentralised. In the worst case, without central co-ordination there is a risk of an avalanche-like failure of significant areas of the mesh. For example, a system which relies on increasing node transmit power to overcome interference will, in doing so, also increase the self-interference affecting adjacent nodes, which will in turn then need to increase their power, and so on in escalation until failure of the system. In another example we might consider modulation level reduction as a way to increase the signal relative to interference. However, as the throughput of this link is reduced, additional routes must be found to share the given traffic load and so the effect multiplies through the mesh.

Hierarchical or centralised control algorithms can help in preventing these effects in a mesh system. But this brings with it at least two problems. Firstly, their presence introduces overhead in the form of control data that must traverse the mesh. Secondly, latency within the mesh may degrade the effectiveness of such algorithms, with the result that more localised control centres might be needed. This would curtail the attractive ad hoc feature of meshes.

Now, taking the higher level view, in essence all we need to ensure is that a node receives its wanted signal with enough energy to overcome an interference signal. There are a large number of methods that can be used to achieve this which fall into one or more of the following categories.

- Avoiding interference using time diversity, frequency diversity and spatial diversity, both individually and in combination. For example frequency hopping is a form of frequency diversity which can be used to avoid narrowband interference. It can also convert continuous interference into discontinuous interference so that time diversity can be additionally employed.

- Rejecting interference. Receivers can use antenna beam steering to reduce the energy from a few single-source interferers and phase continuous signals can allow multi-user detection techniques to be used to reject co-channel interference. (But antenna steering is not practical for mobiles, as we saw in Section 4.5.)
- Increasing transmission energy. A higher energy may be achieved by transmitting with a higher power spectral density, for a longer period of time or using a wider bandwidth.

In mesh systems, particularly mobile or ad hoc systems, the last of these is least preferable. This is because increasing the transmission energy of the signal means either increasing the self-interference effects on other users of the system, since transmit power is increased, or reducing the throughput of a link, due to transmit time being increased.

Let us begin by considering our options in the physical layer.

5.2.1 Physical layer

In Table 5.1 we list the key PHY layer elements in a radio system and their static/dynamic nature with respect to their most common practical implementations. For example, the carrier frequency is typically chosen and fixed at design time (static) and is not normally changeable during operation (dynamic), although limited exceptions clearly do exist, such as dynamic selection from a number of close, alternative in-band frequencies in response to measured interference levels. The table also describes the ways in which control of the element can counter interference, for example hopping is a technique to avoid interference, whereas wideband techniques aim to get more power into the transmission and hence overcome, rather than avoid, interference.

We will examine each line of the table in turn.

Carrier frequency: allocation

Ideally a frequency allocation should posses three attributes: it should be large in the sense the total bandwidth permits many channels, isolated in the sense that is separated in frequency from potential external interference, and dedicated solely to the radio network.

Table 5.1 *Key PHY layer elements with their static/dynamic nature as commonly implemented and their potential uses in mitigating interference effects*

	Static	Dynamic	Interference avoidance	Interference rejection	Transmit signal strengthening
Carrier frequency					
Allocation	✓		✓		
Sub-bands	✓		✓		
Hopping	✓	✓	✓		✓
Orthogonal frequency division multiplexing	✓	✓	✓		✓
Signal bandwidth					
Narrowband	✓		✓		
Wideband	✓				✓
Time structure					
Interleaving	✓		✓		
Throughput	✓	✓			✓
Modulation	✓	✓	✓	✓	
Transmit power		✓			✓
Geometry	✓	✓	✓	✓	✓

However, there are many other constraints, not least existing allocations and legislation which allocates particular spectrum to a particular use. Thus there is frequently little freedom available and the design of a radio system architecture must begin with the assumption that a particular allocation has been decided already. The allocation may be engineered to approach the ideal through appropriate selection of the remaining design parameters, such as those discussed below.

One promising technology which has been on the horizon for some time may change this however. The notion that regulators would allow cognitive radio schemes to operate in spectrum which the radios identify

as 'white space', or unoccupied, would enable dynamic spectrum access. The challenge at the moment is for the regulator to be satisfied that the technology can be employed without interference to others. One place it might be employed is in the UHF spectrum, in the bands used for TV. However, distinguishing a white space from a low power TV signal is technically especially challenging.

Carrier frequency: sub-bands

If the frequency allocation permits more than one channel, sub-bands may be defined which allow independent networks to co-exist with reduced self-interference. An example is the assignment of sub-bands within the GSM allocations to different service providers. Such a strategy works equally well for both mesh and cellular systems in reducing self-interference between networks. However, it requires a high degree of co-ordination and hence is only practical if the number of networks that must co-exist is small. Unfortunately this is also a typically inflexible example of a command and control approach to spectrum which is falling from favour.

Carrier frequency: hopping

Frequency hopping can be used to reduce the impact of self-interference and external interference when the bandwidth of the interference is less than that spanned by the frequency hopping. When used for this purpose it is one of the several techniques that are known as spread spectrum (frequency hopping spread spectrum, FHSS). In other words we are deliberately expanding our wanted signal spectrum to a range such that it is wider than the interferer, leading to an increased chance that we will 'miss' the interferer frequency on many of the hops.

There are two ways in which frequency hopping can be used.

- Through the use of orthogonal hopping sequences. Transmitters in such a radio network use hopping sequences which are guaranteed never to result in two devices using the same frequency, or frequencies close enough to result in adjacent-channel interference, at the same time. For example, the GSM system uses orthogonal hopping sequences within a cell. The base station defines one or more sets of

frequencies, each used by a group of mobiles. The mobiles use the frequencies in the same order but each has a different offset in the sequence. Clearly this requires centralised co-ordination, which may not be available in an ad hoc mesh.

- Through the use of non-orthogonal hopping sequences. The hopping sequences are designed so that the likelihood of co-channel or adjacent-channel interference is low. This allows error correction schemes to compensate for the effects of the self-interference.

The Bluetooth system is an example of this hopping approach. A pseudo-random sequence defined by the master in a piconet[1] uses up to 79 frequency channels. The sequence used by a master depends on its unique Bluetooth address and so independent piconets use different sequences.

A further refinement is that frequency hopping sequences do not need to be static. An example here was the introduction of adaptive frequency hopping in version 1.2 of the Bluetooth standard. This was specifically introduced to reduce the effects of interference between Bluetooth and 802.11 derivatives occupying the same unlicensed device band. The 79 channels used by the frequency hopping is a maximum figure. This number may be reduced during actual operation (i.e. dynamically), if measurements made by the receiver show that particular channels are subject to persistent interference or fading effects [1] and ought not be used.

Frequency hopping methods are attractive because they provide protection not only from interference effects but also from fading effects. This is because fading effects are frequency dependent. Hopping in frequency thus leads to a reduced chance of experiencing a deep fade over a given time.

But finally we must add a note to balance the attractions of hopping. In designing our system we should be mindful that hopping does have a number of implications.

[1] A piconet is the term for a networking cluster of Bluetooth devices.

- Some form of error correction or acknowledgement/repetition protocol is required. Frequency hopping alone does not protect against interference effects. It attempts effectively to shorten the duration of interference so that the performance of error correction and message acknowledgement/repetition protocols is improved. This is a key point.
- There must be guard periods in the on-air time structure which allow the frequency to be changed. In other words we must provide gaps where we allow the frequency to be in the process of change and stabilisation. Unfortunately this not only reduces the system throughput but also introduces phase discontinuities which may affect the subsequent demodulation performance.
- The channel characteristics at two different frequencies are unlikely to be the same. This is an advantage when frequency hopping is being used to mitigate fading effects as we have said, but it is potentially a drawback if the purpose is simply to avoid interference. Some form of process is typically required to allow the channel characteristics on a new frequency to be quickly estimated, which again constitutes a system overhead.
- Time synchronisation of the elements in the radio network is required so that frequency hops occur at the same time and the elements must agree on a sequence. This suggests the need for some form of master/slave structure to the network. This might be an ad hoc structuring such as occurs in Bluetooth piconets or a formal cellular structure. Nevertheless, the structure is required, and for mesh networks it is not clear how to impose such structure over the network. It might be done geographically or perhaps even by node traffic type. In any case we are returned to the recurrent theme of needing centralised co-ordination.

Carrier frequency: orthogonal frequency division multiplexing

OFDM is a multi-carrier technique in which data are transmitted over a number of sub-carriers. These sub-carriers are synchronised in time and phase and chosen to be orthogonal and so do not cause adjacent-channel interference to each other. The performance of each sub-carrier is therefore independent of the presence of the others.

The susceptibility of the method to interference is essentially dependent on the performance of the sub-carriers. However, since the original

data are distributed across the sub-carriers, the effective data rate on each sub-carrier is only a fraction of the original data rate. Thus the transmission of each bit is distributed across a greater period of time and the susceptibility to discontinuous interference is reduced, when compared to the same modulation scheme running at the original data rate. Further, if a continuous narrowband interferer is present, a relatively simple adaptation algorithm can completely suppress the use of the affected sub-carriers.

Signal bandwidth

Let us first define narrowband and wideband. Narrowband is used to describe signals with a bandwidth that is similar to or only slightly greater than the data rate. Wideband is used for signals with a bandwidth that is significantly greater than the data rate. We note one exception in that frequency or phase shift keying (FSK/PSK) modulation with a high modulation index, or high-order FSK, can result in a wide signal bandwidth, but with the signal energy concentrated at only a few discrete frequencies. Because of this concentration of energy, these modulation formats are often treated as narrowband.

Signal bandwidth: narrowband

Narrowband signals perform relatively poorly in the presence of interference of a similar power and bandwidth. However, the likelihood that a narrowband interfering signal will coincide with the wanted signal can be quite low and it is a relatively simple matter to use hopping techniques to avoid the interference.

Naturally, in the presence of wideband interference, narrowband hopping techniques do not work well, but nonetheless there remain conditions under which the use of narrowband signals can be advantageous. The fact that the power of the interference is spread across a wide band means that if the data rate is particularly low, the bandwidth of a receiver can be made small enough that most of the interferer's power is rejected. Note, however, the data rates required even only for speech are generally too great for this to work. In this case the only real remaining solution to overcome wideband interference is to increase the signal energy by increasing the transmit power.

So, in summary, as long as the system is subjected to only narrowband interference, the advantages and disadvantages to using a narrowband signal in a mesh system are similar to those for other systems. The relatively poor performance of narrowband signals with respect to narrowband interference is balanced by the relative ease with which hopping techniques can be used to avoid the interference. However, if wideband interference is likely to arise then the only real solutions are to increase the transmit power or reduce the throughput, if possible.

Signal bandwidth: wideband

In general, the fact that the signal energy is distributed over a wider bandwidth automatically means that the susceptibility to interference is relatively lower. Let us begin by looking at a number of types of wideband signal, each with their own characteristics.

Direct sequence spread spectrum At the transmitter, direct sequence spread spectrum (DSSS) achieves a wide bandwidth by replacing each data bit with a code sequence or 'chip' which has been modulated at a higher rate. The ratio of the chip rate to the data rate defines the processing gain.

There are various sequences which may be used as chips, each with different properties. Generally the choice of the type of sequence to be used is based on the multiple-access potential of DSSS (see CDMA in Section 5.2.2). The choice makes little difference in the susceptibility to interference.

At the receiver, the de-spreading process essentially converts a narrowband interferer into a noise-like signal at the input to the decision-making elements of a receiver [2]. An uncorrelated wideband interferer will also result in a noise-like signal. Thus the impact of both types of interferer is similar to that of white noise of the same power. The wanted signal is picked out strongly by correlation. This makes DSSS one of the best schemes for dealing with moderate levels of interference. This is true as long as we ensure the processing gain is reasonably high via our choice of chip rate relative to our data rate.

We do note, however, that at higher interference levels DSSS performance degrades quite quickly and the only recourse is to increase the

transmit power or reduce the throughput. Both these methods can cause problems in a mesh system, as already noted.

Frequency hopping spread spectrum We have already covered hopping schemes, but here we add more detail. In fact there are two classes of FHSS, fast and slow. In fast hopping, the rate at which the frequency is changed is higher than the data rate, and in slow hopping it is lower.

We have already seen that FHSS is highly resistant to narrowband interference. As the interference power increases, the performance degrades up to the point when the frequencies affected by the interference are completely blocked, but no further. Thus FHSS can handle high-power narrowband interference better than DSSS.

The key difference between fast and slow FHSS is that the blocking of slow FHSS frequencies results in the corruption of several data bits so that other schemes must be used to compensate for the effect (re-transmission, error correction coding etc.), resulting in a loss of throughput. With fast FHSS, blocked frequencies affect a fraction of a bit and so long as the number of blocked frequencies remains small the demodulation will be unaffected.

Time structure: interleaving

Interleaving is a diversity technique which can be used to counter discontinuous interference. The premise is that a non-zero error rate can be tolerated if the errors in the data are evenly distributed in time, rather than appearing as a long burst. Interleaving will cause long bursts of errors to be broken up as de-interleaving takes place at the receiver.

The interleaving technique is most efficient when the design takes account of the capabilities of the higher layer protocols, specifically how they perform with respect to error distribution in time. Usually, even distributions are handled best. The principal disadvantage of interleaving is that it introduces delay into the transmission process, due to the delay inherent in arranging the interleaving itself. In a multi-hop mesh system where delay is already an issue, this additional delay is likely to be unwelcome.

Time structure: throughput

The required information throughput is often the key element in a radio network's design. It is always considered relative to the level of interference to which the radio network is expected to be subject, for the following reason. By reducing the throughput the design can be made more resistant to interference because the energy which can be used to transmit the information bits increases as the bit rate is reduced. But often we do not know the exact conditions under which the radio system will operate and so the option to vary the throughput during normal operation is an attractive one. In this way we may dynamically maximise our throughput.

In the physical layer there are three groups of methods which can be used to vary the throughput dynamically.

- An adaptive modulation scheme, in which the on-air modulation rate changes, may be used. For example the sub-carriers of the IEEE 802.11a system are modulated using either binary or quadrature phase shift keying (BPSK/QPSK), 16-quadrature amplitude modulation (QAM) or 64-QAM.
- A repetition method may be used to introduce redundancy yet achieve a fixed on-air modulation rate, as follows. A system would be designed for a maximum throughput and as the throughput is reduced the information bits are repeated to maintain the same rate at the modulator. The method has the advantage of being simple to implement, but the spreading method is preferable since it brings additional advantages, including at the system level, as follows.
- Although the spreading method achieves the same aim of a fixed on-air modulation rate as the repetition method, the difference is that the information bits are not simply repeated, but are replaced by a sequence of modulation-bits, or chips. As we have already seen, the pattern of modulation-bits may be chosen to have different properties.
  - Pseudo-random sequences spread the bit energy across the channel bandwidth and form the basis of the direct sequence spread spectrum (DSSS) methods. By spreading the energy across the channel bandwidth the susceptibility to narrowband interference is reduced.

- Orthogonal sequences allow multiple information bits to be transmitted at the same time and in the same channel. These form the basis of the code division multiple access (CDMA) method.

For example, the 3GPP system uses orthogonal sequences to facilitate use of the CDMA method and scrambles the result to spread the bit energy across the channel bandwidth. We will look at CDMA in Section 5.2.2 when we consider multiple access techniques.

Varying the throughput to match the propagation conditions is a technique used in most cellular designs. The management of the throughput can be centralised in the base station and generally a point-to-point connection will include at most two radio hops so the negotiation of the throughput can be quite simple.

In meshes the management of throughput could be considerably more difficult. The variable number of RF hops and the possibility that this might change during a call means that the negotiation process could become very complex.

Modulation

The modulation scheme used in a system affects the susceptibility to both self-interference and external interference. The effects are a result of both the immediate characteristics of the modulation and the characteristics of the practical transmitter and receiver designs that must be used in the real world to realise each modulation format.

The differences between bandwidth efficient modulation methods can be subtle. The power spectral density of MSK, for example, has a wider central lobe than that of QPSK and BPSK but it has lower side lobes, which in principle should mean that it would offer better adjacent-channel performance. However, all these modulation schemes have most of their energy concentrated in the central lobe so that, in practice, additional channel filtering can be included in the transmitter without significantly degrading their performance.

The susceptibility of different modulation schemes to external interference is a similarly complex issue. Even studies which examine similar modulation schemes and similar interference effects can present different

results and conclusions [3, 4]. Differences arise from the assumptions made about the nature of the receiver designs.

On the whole it is not possible to identify one modulation scheme that is better than all others with respect to interference susceptibility, not least since other aspects of radio system design may have a relatively greater impact.

Transmit power

Increasing the power used to transmit a signal is the simplest method of overcoming interference. However, there are constraints which result from practical design considerations and from standardisation limits. The spectrum assigned to a particular use invariably has an associated maximum transmit power.

Further, increasing the power of the signal also increases the interference that it causes. Thus the self-interference of a system is increased, and in a mesh this may lead to instability as each node tries to increase its power to overcome the interference caused by its neighbours.

Geometry

The deployment geometry of the transmitters and receivers used in a system can be used to reduce the susceptibility to interference. Techniques such as antenna diversity, polarised antennas, directional antennas, beam steering and sectoring are all used in cellular systems and fixed systems.

However, such methods rely on the elements in a system having a relatively stable spatial position and orientation. In mobile meshes such stability cannot be assumed to exist.

5.2.2 Medium access control

Let us begin by saying that medium access control is synonymous with multiple access control and both are referred to by the acronym of MAC. The MAC challenge is that we have a shared medium which many users would like to use. Actually the users would like the medium all to themselves, but it is a limited resource to which none has exclusive access rights.

Controlled access to the medium, which in our case is the radio channel, needs to be

1. fair,
2. efficient, and
3. stable with respect to contention for resource, i.e. during excess demand.

On the whole there are two common approaches to satisfying the MAC challenge. In the first, there is an entity which is in control; the bandwidth share is organised centrally, for example as in a GSM system. New users need to be explicitly admitted by the system controller. In the second, no one node is in control, but each operates the same co-operative method of attempting to gain a share of the bandwidth. Access for any user is thus more random then deterministic. The best known example of a random access MAC is the original Ethernet based on a single, shared coaxial cable.

The first multiple access method is generally called either centralised, controlled or deterministic, whereas the second is generally called distributed, co-operative or random. We will use only the terms centralised and distributed to distinguish the two methods. In fact the other terms are not perfect synonyms, but the detailed differences need not concern our level of discussion.

Centralised medium access

Let us begin by looking at centralised access approaches. Many examples take a slotted approach, where each node is pre-allocated a slot of time, frequency or code space. In the following the basic multiple access techniques of frequency division multiple access (FDMA), time division multiple access (TDMA) and code division multiple access (CDMA) are described. Although they are described separately, most practical radio network designs use a combination of these.

Frequency division multiple access

FDMA allocates different carrier frequencies to different user channels. The allocation can be static or dynamic, with the latter combining well

with frequency hopping techniques for avoiding interference. Static allocations are generally avoided in mobile networks because a particular frequency may be subject to significant interference or poor propagation conditions. They are however frequently used in carefully planned point to point networks.

Dynamic allocation introduces the need to control access in time as well as in frequency, leading to an approach which combines FDMA with TDMA.

Time division multiple access

TDMA allocates different timeslots to different user channels. Nodes in the network are synchronised and the transmitter and receiver know when a transmission is due to take place. This requires a common time base which, in a mesh system, may be difficult to achieve without some form of common external reference. Alternatively it requires large guard periods to allow for imperfect synchronisation.

Code division multiple access

CDMA allocates different orthogonal spreading codes to different user channels, each sharing the same time and frequency space. In other words it is an access method which incorporates direct sequence spread spectrum (DSSS). Ideally the use of orthogonal spreading sequences allows signals that are superimposed in time and frequency to co-exist without interfering with each other. In practice there are some problems in achieving this.

There are known sets of sequences that are perfectly orthogonal as long as they are synchronised in time, and these can be used effectively to carry multiple data streams transmitted by the same user. However, if we have a multi-path environment then delayed versions of the transmitted signal are also received at the destination (unless highly directional antennas are used). Certain receiver designs such as Rake can take advantage of some of the delayed versions to increase the energy of the wanted signal. However, the signals which are no longer time aligned are no longer orthogonal and so act as interferers.

If there are multiple active transmitters, the signals at a receiver cannot all be synchronised. The signals therefore act as interference to each other, with signals from closer sources having a stronger effect. In the extreme case a signal from a nearby source can result in the signal from a distant source being completely hidden – the so-called near-far effect. To avoid this, cellular systems use closed-loop power control to ensure that the power of each signal at a receiver is the same.

An example of the use of CDMA is the 3GPP system. The W-CDMA variant uses orthogonal Hadamard codes so that the base station's transmissions to the mobiles in a cell are as orthogonal as possible given the propagation conditions. The mobiles use the same codes in their transmissions and are subject to strict closed-loop power control to try to ensure that their signals reach the base stations with the same received power. In fact the situation is further compounded since uplink synchronicity cannot be assured and further coding is also used with deliberately poor cross correlation properties, in order to enable separation.

Complex algorithms in the base station and in the network balance the power requirements of each mobile with the required throughput. They also balance the power used in neighbouring cells, as this is a major source of interference at a base station. As a guide, for a network of cells with three sectors, the interference power from adjacent cells can be 0.65 of the power of the wanted signal from the mobiles in a base station's own cell.

For mesh systems, particularly mobile ones, this self-interference is a serious problem. The centralised power control that exists in cellular systems would require significant flow of control data throughout a mesh and it is questionable whether it could do so fast enough to be effective.

Distributed medium access

In contrast to centralised systems, nodes participating in distributed medium access networks are not synchronised and hence each node transmission occurs at an uncontrolled, random time. This means that collisions may occur, and some form of re-transmission, error correction or avoidance scheme is therefore advisable.

The attraction of distributed methods is that they have simpler control algorithms. However, they have the disadvantages that the maximum throughput of a link is reduced and that the receiver must be active for longer periods of time. The maximum throughput is reduced as a consequence of dealing with collisions or due to the overhead of avoiding them.

The basic method is carrier sense multiple access (CSMA) in which the transmitter first checks that the channel is free. If the channel is occupied, the node will back off, in other words delay its next attempt at transmission until some random time later. This still leaves the possibility that two nodes will transmit at exactly the same time and collide. Refinements to the basic method include CSMA with collision avoidance (CSMA/CA) where a handshake is used to confirm transmission is about to commence and subsequently that no collision occurred during transmission. An example is IEEE 802.11, which is channelised by frequency to allow cells to be created where required, and within each channel CSMA/CA is used to control access to the medium.

In CSMA/CA, the MAC operates 'listen before talk', and it can thus be described as a 'polite' MAC. Politeness thus refers to a MAC's behaviour towards others. It is instructive to note that such a scheme often fails when operated in the same area as an impolite MAC. Examples of polite and impolite MACs are those of 802.11 and 802.16 respectively, as shown in Figure 5.3.

This politeness aspect of the MAC protocol has a large impact on susceptibility, as follows.

Because of the MAC behaviour, a system using a polite protocol will always cede control to a system which simply schedules transmissions regardless of any other band users. This is because the back-off algorithm inherent in a CSMA approach will continue to reduce packet size and increase wait time, thus reducing its access to the medium in response to interference. In contrast, the centralised MAC will continue to fill its regular resource slots (e.g. time slots) with automatic repeats of data which may have been lost in any collisions due to interference. This vicious circle has been studied for the case of 802.11 losing out to 802.16 and some steps have been suggested to overcome it, namely the inclusion

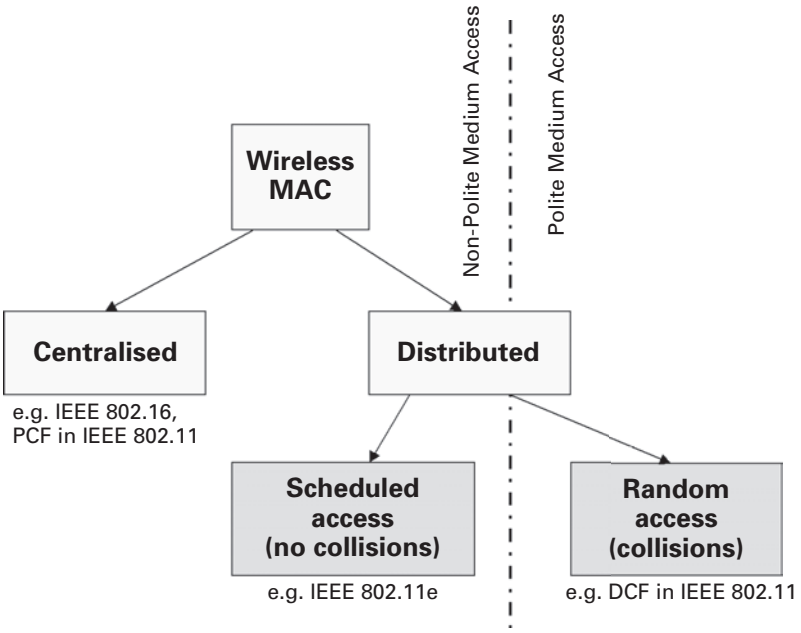


Figure 5.3 'Politeness' of MAC protocol for medium access.

of quiet periods in the 802.16 slotted MAC to allow the 802.11 CSMA MAC to transmit [5]. Very recent developments in 802.16 have focused on its operation in licence exempt bands. It is not presently clear whether this 802.16 development is meant to co-exist with only itself or with 'foreign' systems in the same band.

Finally, it should be clearly understood that politeness and throughput are competing aims. Maximising politeness means allowing time for sophisticated listening and handshaking processes, maximising throughput means maximising transmission time at the expense of everything else.

Distributed MAC inefficiencies

For completeness, we should mention there are several aspects regarding the efficiency of distributed MACs which are well known. They include the inefficiency of the CSMA approach in general, due to its back-off behaviour as we have already mentioned, plus the hidden-node and

exposed-node problems, which we have not needed to describe. The inefficiency and inequality of the specific 802.11 MAC is also known, where single-hop routes are always unfairly favoured over multiple hop routes. These effects, whilst important in themselves, are not mesh specific and we do not need to cover them here.

5.2.3 Conclusions

This section has reviewed PHY and MAC approaches. Those conclusions which are mesh specific are listed below. The major conclusions stem from the lack of centralised control within a mesh.

1. PHY schemes requiring synchronisation across nodes will not work. This includes many hopping schemes.
2. Common MAC schemes which are slotted such as FDMA/TDMA/CDMA will be hard if not impossible to realise.

Other conclusions include the following.

- The common cellular technique of managing throughput in response to propagation conditions is likely to be much more complex in a mesh where high hop counts occur.
- Current transmit power control schemes use centralised control, whereas in a mesh all control is traditionally expected to be distributed. Distributed power control on a wide scale has never been demonstrated to our knowledge.
- Self-interference on a mesh employing CDMA would be a serious problem, since it is doubtful whether fast power control for near-far effects could be implemented with a short enough response time over a wide area.
- Ad hoc systems preclude frequency planning by definition, therefore FDMA schemes are disadvantaged in a mesh.

The effect of MAC politeness, although not mesh specific, is so important that it deserves repeating: any co-located network using impolite protocols will always take all the contended resource away from a network using polite protocols. In other words, shouting loudest will work.

5.3 Dedicated mesh routing and transport approaches

We have already introduced transport and routing in Chapter 3. We noted that every mesh node is both a router and a relay. We also noted the two distinct traffic types: elastic traffic which is not delay sensitive and non-elastic traffic which does not tolerate delay.

The purpose of this section is to see how transport and routing applications fare in the presence of interference. First we look at the routing problem, which is driven by the mobility and volatility of users. This may stress the routing protocol, leading to increased routing overhead and potentially eventual connectivity failure.

We then look at what metrics are suitable for the assessment of routing protocols in an ad hoc environment. These are delivery success rates, delay and overhead. We use these metrics to examine a comparison of two common ad hoc protocols. The result is simply that no one protocol is always more suitable for all combinations of traffic and user load, when judged by all three of our metrics. This is a far less comfortable situation than we are used to with conventional networks.

When we look at transport we review how TCP and wireless systems do not match up well, often leading to inefficient throughput. This situation is likely to persevere since, whilst the required changes to TCP are known, there are so many legacy TCP implementations deployed that they are unlikely ever to be changed.

A better approach could be to exchange more information across the layers of the protocol stack. This would for example allow applications to be made aware of the medium over which they are communicating. Specific allowances could then be made. Whilst this would tackle wireless issues generally, it would not help with the mesh specific problem of relatively high delay, which also causes TCP inefficiency.

Finally we comment that much published investigation at the protocol level uses rather simpler traffic models than actually occur in practice.

5.3.1 Routing

In a manner very similar to the Internet, a mesh may be said to be self-healing, given appropriate network software. If a link drops due to failure

or, more likely, due to user mobility patterns, then another link will automatically be brought up, if one is available. Of course, the success of this method relies on the balance between system efficiency and the speed and range of user mobility. Unfortunately, such automatic re-routing can give rise to problems elsewhere, notably via packet re-ordering which upsets typical TCP stacks, and varying latency which leads to issues at the application layer. But let us start at the beginning by understanding how routing could work in an ad hoc mesh.

At the protocol level, route discovery and maintenance is a challenging task. The node connectivity map is very dynamic and must be refreshed regularly. This is driven by the mobility and the volatility of users;[2] it requires a large number of control packets to be sent. The issue raised by this is initially one of overhead and it is not uncommon for control packets to outnumber data packets under some conditions. Overhead, along with latency and packet loss rate, are normally used as the performance metrics when comparing routing protocols, as we shall see later. But routing packets are important system control packets and are expected to be delivered reliably. We should bear in mind that if they are not delivered, perhaps due to interference, then at best the routing overhead will rise but at worst the mesh may fail.

Routing and mobility

If PHY and MAC properties largely determine the scalability of meshes, as we proved in Chapter 4, then the behaviour of the mobile routing protocol largely determines the *degree of mobility* which may be handled. Efficient dynamic routing protocols are needed to cope with the frequently changing multi-hop network topology.

In fact, the degree of mobility which a routing protocol may handle and how this may be quantified is the subject of intense research at present. With this in mind, let us look at the current level of understanding.

An ad hoc routing protocol consists of three main functional components.

[2] After all, users may freely choose to turn off their nodes.

- Route discovery: this is the part of the protocol which is used to discover routes to destinations within the network.
- Data forwarding: this is how the data packets are organised to be forwarded by the protocol.
- Route maintenance: this is how the protocol deals with changes and faults within the network once routes have been established.

The design, implementation and interaction of these functional components within the ad hoc routing protocol greatly affects its performance. In assessing performance we need to consider several dimensions.

- Data forwarding performance. This might typically be measured by evaluating the 'load versus throughput' performance of the protocol under a range of network sizes (number of nodes), and various levels of background traffic and source traffic types (CBR, Poisson, on/off, etc.).
- Protocol overhead. This would be an assessment of what proportion of the available channel capacity is taken up by the operation of the routing protocol, in other words how efficient is the protocol.
- Delay and jitter. The effect of the protocol may be to introduce delay or delay variation; this will be important for inelastic traffic types.

This is quite a complex list of parameters by which to judge a routing protocol. Fortunately we can turn to published research in the field. This research, described next, compares dynamic source routing (DSR) with adaptive on-demand distance vector (AODV) routing. We do not need to understand these protocols in detail, but we do need to remember that in Chapter 3 we introduced proactive and reactive routing protocols. DSR and AODV are reactive protocols which build routes only upon request. Reactive protocols are favoured for ad hoc systems since they do not waste time calculating all routes, some of which may never be used before they disappear.

The quality metrics used to inform the comparison are essentially those we listed immediately above.

Routing overheads

A comprehensive overview paper is provided by Das *et al.* [6] which contrasts DSR and AODV, two popular reactive routing algorithms for

ad hoc networking. Examples from this paper are used to bring out the compromise between routing overhead, delay and packet loss. The computational capability a node would require for performing routing calculations are not included, but were not thought to be a major concern, relative to those already required by a node for modulation related signal processing.

Das *et al*. [6] tested the following metrics for AODV and DSR:

- packet delivery success rate,
- delay, and
- overhead.

The two main experimental variables were the number of users, indicated as 'sources' for all the graphs in Figure 5.4, and the mobility of users, represented by the 'pause time' on the horizontal axes; a user who pauses less is a more mobile user. The objective was to see whether DSR or AODV was overall a 'better' protocol. What was found was that there was no clear winner in all cases. Depending on which metric was more highly valued (from the list of three above), one or the other protocol was better – although this also depended on the number of users at the time, so making a choice impossible. Let us look at the details of the investigation.

Das *et al*. found a wide set of complex results depending on the parameters of the mobility model and which routing algorithm and metric were being tested. For example they found routing overheads ranged from 33% to 75%, see Figure 5.4 (left hand column). Since they are critical to the system, routing packets are always given preference over data, so any routing 'challenge' immediately pushes up the level of overhead.

The mobility model used was random waypoint (see the Appendix). A random, bounded mobile speed and a random, bounded pause were used before the next transmission in order to create different relative speeds between mobiles. Mobile speeds were 0–20 m/s and the area was 1500 m × 300 m for 50 nodes and 2200 × 600 m for 100 nodes. Pause times varied from zero to the full simulation time (900 or 500 seconds). A shorter pause was used to model higher mobility.

CBR traffic was used, always with 512 byte packets. Source–destination pairs were chosen at random and a variable number of pairs and transmission rates were used to load the network. Note that the model was

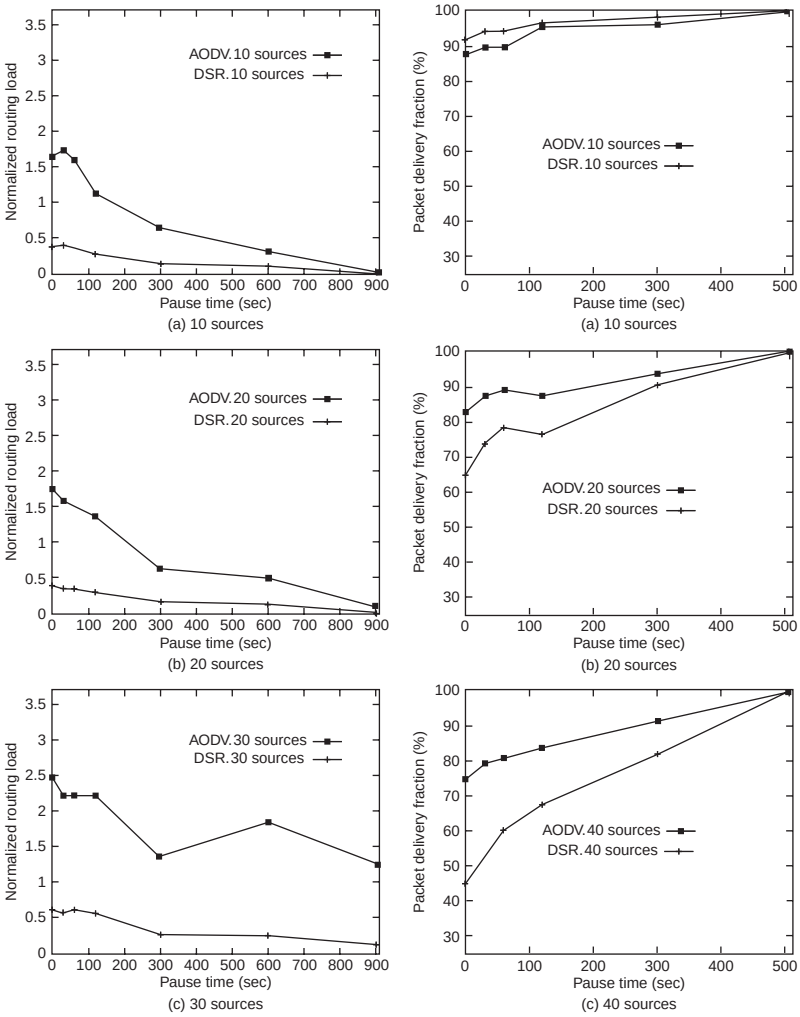


Figure 5.4 Left: normalised routing loads for DSR, AODV (0.5 = 33% overhead, 3 = 75%). Right: packet delivery success rate (100% corresponds to no dropped packets), used with permission from Das *et al*. [6].

never run with parameters which would break the connectivity of the mesh. Note also that these results are merely examples and are not presented as bounds for each metric, which could actually be much wider.

Figure 5.4 (right hand column) shows that at higher network loading (more traffic, more mobility) packet delivery success can be considerably

below 100%. Not shown are other graphs which report that packet delay is very dependent on load and protocol. Inelastic applications (e.g. video) may tolerate a few dropped packets well (due to both error control coding and human perception of video), but begin to fail with delay variation. Elastic applications like email require 100% delivery success but are not so sensitive to delay variation.

In summary, neither protocol was found to be better overall in terms of delivery, delay and overhead simultaneously, over all network loads, although DSR was better all round at lower loads and vice versa. This poses an obvious problem for the mesh system designer, who presently does not have the ability to satisfy all the requirements all of the time.

In summary we must conclude that operating a mesh is very sensitive to the traffic which is being carried and to the behaviour of the users. By the same token, the effect of lost packets due to interference is similarly sensitive. It is simply not possible to choose a routing protocol which is always better than all the others. With our present level of understanding all we can do is try to tailor our choice of protocol to our expectations of user traffic, mobility behaviour and interference, accepting that changes relative to our expectations may not be well handled by our system. It should be clear that this is a very different situation than we are used to with our more conventional communications systems.

5.3.2 Transport

Let us begin our examination of transport by looking at the familiar transmission control protocol and familiar elastic traffic, such as email and web browsing. Historically, TCP was designed to solve the congestion collapse problem in the Internet and is therefore extremely conservative when it encounters packet loss; it typically reduces to half the sending rate when it detects one lost packet. Since the Internet was, and still is, mostly a wired environment this assumption remains almost always correct.

In a wireless link, however, this is not the case. Losing a packet is much more common and does not usually imply congestion. Nonetheless, TCP will reduce the sending rate unnecessarily, causing lower network

efficiency and a degraded user experience. This is well understood and some steps can be taken to prevent this problem. For example some mobile devices are becoming available with modified versions of TCP that are much less conservative and more tolerant of packet loss, e.g. TCP Westwood. This is a solution when the full end-to-end link operates a modified TCP stack. However, it must be appreciated that no legacy devices are ever likely to receive new TCP stacks and there are numerous scenarios (e.g. home Internet access) where the end system is a legacy device which cannot be re-configured.

Even in those MAC implementations which include automatic repeat request (ARQ) to reduce packet loss, the delay caused by this may still cause TCP or other transmission protocols to time-out and thus deduce packet non-delivery. In other words the problem of TCP on a wireless link is here to stay.

A similar problem may arise from the increase in delay caused by a large number of hops in a mesh. Because the TCP data rate increases with each reception of an acknowledgement, indicating a good link, an increase in delay forces the TCP data rate to increase more slowly. This might be a problem, even for web browsing, since a small increase in the total delay of a page transfer may lead to a reduction in user satisfaction. It may also be a problem if high transfer rates are desirable, since TCP is well known to perform badly with high delay-bandwidth-product networks. In other words, if delay is present and the application is such that it requires high bandwidth transfer, TCP is proven to be inefficient. This is not foreseen to be a problem for any current application today but the near future may bring new applications with higher bandwidth requirements.

On another practical note, mesh networks where several routes can be used to maximise throughput can cause packet re-ordering since consecutive packets may follow differing paths. Although this does not violate the TCP/IP model it does generate inefficiencies since the majority of practical TCP implementations do not cope well with packet re-ordering. The potential success of mesh networks would thus partly depend upon a redesign of TCP stacks to overcome this issue – but such a change would be practically impossible to propagate into all the very many legacy TCP implementations already deployed, as we have said.

Let us now turn to consider inelastic applications and UDP.

Inelastic applications, typically interactive real-time ones like audio and video, have been gaining in popularity in the Internet and generating major concerns over its stability. The recent popularity of IP telephony, for example Skype, is forcing uncontrolled traffic into the network and some fear that the possibility of congestion collapse in the Internet has thus increased significantly. This is because these applications use UDP as a transport protocol, rather than TCP. UDP performs multiplexing through the UDP port but does not implement reliability and, more crucially, does not perform any congestion control. Users send at a given rate (in the case of Skype 10 kbps) and the network is simply expected to cope with this rate.

Although this remains a problem for a TCP/IP network in general, UDP is less affected by the behaviour of wireless mesh networks. Nevertheless, an increase in delay can easily destroy the quality of a phone call since anything over 150 ms begins to become problematic for the normal user.

Finally we note that a current research issue is the interface between applications and the lower layers. Newly proposed protocols like DCCP (datagram congestion control protocol) are trying to address this by creating different congestion control profiles which can be selected by the applications. This is increasingly important as the heterogeneity of applications in today's networks is increasing. In other words, the application would know over which type of network it was being operated, such as a wireless mesh network – and it could then instruct the lower layers to select an appropriate congestion profile. In this way wireless error rates could be less likely to be misinterpreted as congestion.

Traffic models and implications

It is appropriate to offer a note of caution with respect to findings offered for protocol testing. Many are based on models: network models, user mobility models and traffic models amongst others. The traffic model used will strongly affect findings, since the range of traffic types is great. We have already said that small file transfers can be very tolerant of delay, delay variation, end-to-end capacity and breaks in transmission, whereas full screen, real-time video places very stringent demands on the same three network parameters.

A key problem is that, despite the known diversity in user traffic types, many simulations simply use CBR (constant bit rate) traffic. CBR is representative of very little in the real world, but is convenient to handle computationally and analytically. Although simulation results for routing protocols, for example, will include items such as latency and packet loss, and thereby give some idea of how multimedia traffic might fare under the protocol, this may not be sufficient. We would suggest that a further technique is needed to stress the simulation based on the bursty nature of some traffic or the regular periods of very high speed traffic which might be expected in multimedia communications. Such an approach would stress the network as well as look for subsequent multimedia-friendly behaviour of the protocols.

Unfortunately, models of actual traffic loads appear to be largely absent from the present published literature.

5.4 Co-existence approaches

If susceptibility considerations are all about seeing other transmissions as interferers, then co-existence considerations are much more about being good neighbours. In other words, co-existence is a two-way co-operative approach to sharing the available resource with other users of the same spectrum. Of course, to be fair, much of the thinking in the previous sections about susceptibility assumed that there were no others who might wish to share the spectrum, so selfish efficiency was the primary aim, for example GSM. Such thinking was appropriate under the older command and control approach of dividing up spectrum for particular users. However, modern spectrum management is concerned with moving away from these old partitioning models towards a model where spectrum is seen more as a common resource. This leads directly to the need for spectrum users to think about co-existence. It is likely that when using shared spectrum in future, users will have to meet the normal power and other constraints, but may also have politeness constraints which promote co-existence.

Accepted co-existence tools like dynamic frequency selection (DFS) and transmit power control (TPC) are problematic for meshes as already

stated. On the other hand the use of polite protocols is perfectly realistic. In fact we have seen that meshes tend to use distributed protocols, which are based on random access and thus likely to be inherently polite.

It is the purpose of this section to look at what other techniques are available such that meshes may co-exist with other systems, i.e. without interference problems in either direction.

5.4.1 Knowledge based approaches

As we have said, co-existence is appropriate in shared spectrum, which is sometimes also called a spectrum commons, since it is spectrum which is commonly accessible by all. The guiding principle of knowledge based approaches to co-existence is very simple: it is easier to co-exist with an entity whose characteristics are known, than with one whose characteristics are unknown. Whilst this may seem self-evident, it is worth explicitly listing several drivers for moving to a knowledge based approach.

- In the spectrum commons, other users are likely to be present at times which suit them. This is quite the opposite to cellular system operating in licensed spectrum, where the interference environment is more predictable.
- The other users in a commons may use radio systems whose actual waveforms are unknown. This can make a large difference when trying to work alongside them.
- With no knowledge of the other systems, how is interference to be quantified and measured?

So, it is clearly preferable to have knowledge of other users, in order to simplify the task of co-existing with them. This will of course be true for all users. The knowledge required could be collected beforehand in a table, much like it is done for a fixed routing table in the Internet, or in theory it may be 'learnt' using some as yet to-be-developed technology.

The active approach of sensing whether anything else is using the spectrum is precisely what the field of cognitive radio is concerned with. A cognitive radio is one that can respond to its environment.

Much consideration of cognitive band sharing has occurred recently due to the freeing of spectrum within the UHF TV bands as many countries move towards more efficient digital TV transmission. Some spectrum will be freed entirely, but the more interesting case is spectrum which is free only in certain geographical areas (and perhaps times) but not in others. There is an incentive to detect this unused spectrum, usually referred to as 'white space', and to use it until such a time as the primary user requires it again. This has generated huge debate within the industry, especially as the white space detection process is technically very challenging, both in the laboratory and more so in real deployment environments (due to hidden node issues).

An FCC consultation document suggests three ways in which the avoidance of harmful interference to authorised users may be achieved.

1. Existing TV/radio stations are to transmit channel availability information directly to the unlicensed device – a control signal approach.
2. The unlicensed device is to determine its geographical location via for example GPS and then look up channel availability from a reference table – overseen by a professional installer (GPS indoors is a known issue).
3. The unlicensed device is to sense spectrum occupancy directly and hence available 'white space' (i.e. the smart or cognitive radio approach).

Although not mesh specific, this usefully lists the three options for planned co-existence as

- a control signal approach,
- a location table based approach, and
- spectrum sensing (cognitive radio).

However, the suggestions have not gone unchallenged; it has been pointed out by respondents that co-channel interference may be a problem over 75 square miles and that out of band interference will also cause concern. This is for the case of TV reception, which is quite prone to interference. The FFC Laboratory is to become involved making measurements with the ultimate aim of having co-existence products for sale by 2009.

The conclusion here is that the industry is moving towards band sharing and that the old command and control aspect of regulation is on the wane. Yet ways of enabling spectrum sharing are not yet mature.

5.4.2 Geographic spectrum planning

On a related note, whilst this book concludes that mobile mesh networks may not be spectrally efficient in terms of the traditional bps/Hz/km^2 metric, they offer certain benefits to spectrum planning. With regard to spectrum sharing on a regional basis, the interference footprint from a mobile mesh network is likely to have a sharp fall-off due to the low antenna height and transmit power of terminals, in comparison to an elevated cellular base station. Hence spectrum use is more contained spatially. This would facilitate re-use in adjoining areas.

5.5 Summary of susceptibility and co-existence issues

The lack of centralised control within an ad hoc mesh leads to problems at the PHY and MAC layers, where we find that conventional approaches are not so easy to implement. For example at the PHY layer hopping approaches are made difficult, and at the MAC layer slotted techniques such as FDMA, TDMA and CDMA are made difficult. Most meshes hence use a distributed MAC. This, being a polite approach, is a double edged sword; co-existence with similar polite systems is empowered, but the polite scheme will always loose out to those systems using slotted, deterministic MACs, such as FDMA, TDMA and CDMA.

Whilst mesh networks do have the capability to physically re-route under conditions of interference (for example), the drawbacks of doing so may be felt by the application via effects above the physical layer. Overall, the issues lie with the interaction of the communications protocols, which are constrained by the distributed quality of a mesh network, with the lossy nature of radio transmission.

Other approaches to co-existence apart from politeness include the set which is knowledge based: tables of all network transmitters, control signals from sensitive band users and cognitive approaches. The latter

appears not to be mature, although both effort in this area and regulator interest are high.

Finally we noted that due to the low antenna heights of typical meshes, compared to cellular base stations, the mesh interference footprint is likely to have a sharp fall-off at the edge. This should help any other systems which are planned to be adjacent.

References

1. Golmie N., Rebala O., Chevrollier N., Bluetooth adaptive frequency hopping and scheduling, *MILCOM 2003 – IEEE Military Communications Conf.*, no. 1, October 2003, pp. 1138–1142.
2. McCune E., DSSS vs. FHSS narrowband interference performance issues, *RF Design*, September 2000.
3. Martin W.L., Yan T.-Y., Gray A., Lee D.K., *CCSDS – SFCG, Efficient Modulation Methods Study at NASA/JPL, Phase 4: Interference Susceptibility*, NASA/JPL, 12 October 1999.
4. European Space Agency, *Interference Susceptibility of Selected Bandwidth-Efficient Modulation Schemes* (SFCG Action Item AI 18–17/2, Tasks 17/2–1 to 17/2–5), September, 1999, pp. 7–15.
5. Li L., Sarca O., Selea R., *A Solution/Scheme for Coexistence with 802.11 in the LE Bands*, submission IEEE C802.16a-02/74, IEEE 802.16 Broadband Wireless Access Working Group, 2 July 2002.
6. Das S., Perkins C., Royer E., Performance comparison of two on-demand routing protocols for ad hoc networks, *Proc. IEEE Conf. on Computer Communications (INFOCOM)*, Tel Aviv, Israel, March 2000.

6 Mesh services and quality of service

In this chapter we will show that a startling effect in meshes is that quality of service (QoS) is not under the operator's control but depends on mesh node behaviour. In a mobile mesh, this means that your QoS depends on your neighbours' behaviours at any point in time, potentially spanning a range all the way from having no discernable effect up to a complete loss of your service. There is nothing quite like this problem in the networks we commonly use today.

But we begin this chapter by looking at how QoS is defined and what QoS levels are required for the applications of today and into the future. Following this we look at whether there are any useful services which truly only a mesh could support. After considering node mobility and showing how node to node relative speed is the key parameter, we look at an example of how a mesh may break into disconnected pieces. This can occur before the full mesh capacity is approached. Finally we show that mesh quality of service is not entirely within the control of the network operator, but rather depends on user mobility and traffic, before showing how adding infrastructure can help improve the quality of service position. Mitigation techniques for QoS issues induced by normal user activity include the provision of extra network-owned nodes in order to regain some control, but this comes at a cost for the operator.

6.1 Quality of service and levels required

Over the next 10–20 years, the users' service requirement will become increasingly difficult to deliver as the mix of these services becomes increasingly biased towards real-time service types like video and voice over Internet protocol (VoIP). Quality of service in the business to business market is usually practically specified within a service level

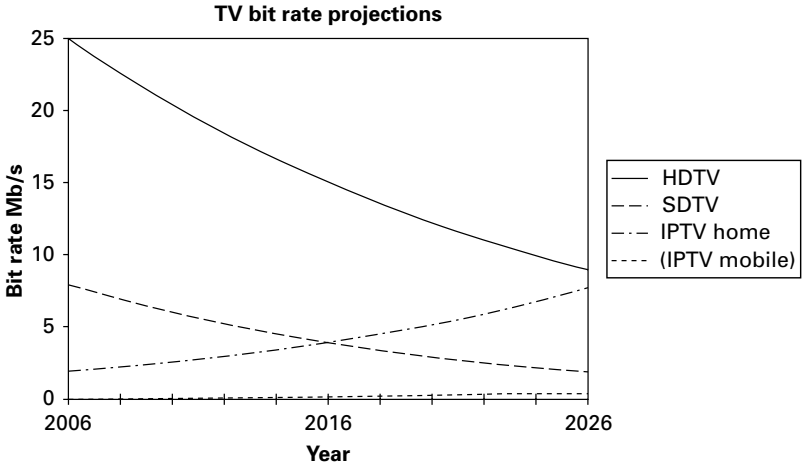


Figure 6.1 Predicted bandwidth requirements over time, used with permission from Ofcom.

agreement, although this is not yet the case in the consumer market. With or without a service level agreement, a user will require a quality of service covering all the same parameters. This traditionally includes at least the following:

- bandwidth,
- latency,
- packet loss, and
- availability.

Notable changes in the services expected to be required by users suggest that a move to more symmetrical bandwidth for uplink and downlink is going to be required, rather than the asymmetry inherent in ADSL, for example. The consumer applications driving this are home generation of content and multimedia file sharing. A further change in requirements is, as ever, an increase in the bandwidth required. This is also driven by multimedia.

Figure 6.1 shows that whilst the bandwidth required to deliver an SDTV or HDTV service may reduce over time, the requirement is still high. It also shows that the bandwidth required to deliver IPTV to a home

or portable device is likely to increase, due to improvements in screen quality and size plus the increasing quality demands of users. IPTV to the home is expected eventually to reach HDTV quality and this is a signal which, in future, may need to traverse a mesh to for example a laptop. IPTV to mobiles is a much smaller requirement and is shown for reference; the expectation here is that this will be over a limited capacity cellular channel to a relatively small screen.

Because of the increasingly real-time nature of the service mix, the QoS focus will naturally move to latency and latency variation. We can understand the importance of latency as follows.

Latency is a system parameter which is effectively fixed at design time by the chosen architecture; latency cannot be reduced in service, in the same way that bandwidth can simply be increased, for example by adding more parallel links and/or more spectrum. The next version of 3G, LTE, is undergoing a major redesign in order to reduce latency to the 20 ms level, since the present 3G architecture simply cannot achieve this. However, we should bear in mind that the end-to-end latency consists of much more than that within whatever access network is used. For example, to the 20 ms of 3G latency must be added the latency of the remainder of the total route, whose endpoint may be somewhere on the wired Internet, for example. Hence a holistic approach is demanded. Interestingly, the QoS of IEEE 802.16 comes at the expense of it never expecting to have to share the channel. This is the trade-off between politeness and efficiency which we saw in the previous chapter. Latency is also compromised when multi-hop systems are considered, like a chain of radio backhaul nodes or a mesh.

Turning now to packet loss, this typically means a short-term dropout, but it can have an amplified effect if the transport protocol reacts 'badly' to loss. As was discussed in the previous chapter, TCP's congestion control back-off is confused by high wireless error rates, for which it was never designed.

Finally, availability is coverage related in a radio system. Where coverage is poor, availability is also likely to be poor and may vary throughout the day. This is likely to be a growing problem as we are entering a world of 'always-on', which implies perfect availability.

6.2 Quality of service drivers

In order to establish the appropriate QoS requirements, we must first establish what application scenarios and user behaviour a mesh is likely to have to support. We already have our list of most likely application examples from Table 2.1. The potential applicability of mesh networking is next further tested via three key questions, with the emphasis on services and the role of QoS.

1. Would a mesh enable new services?
2. What degree of mobility could a mesh support?
3. Could a mesh guarantee a quality of service level?

This section concentrates mainly on access meshes, i.e. those with access to services beyond other user nodes in the mesh.

6.2.1 Would a mesh enable new services?

Like the rest of the Internet, meshes are expected to be required to carry a variety of traffic types. One way of answering the question of new services which could be delivered by mesh alone is to look for application types which are 'mesh-like' themselves. All meshes discussed in this book have an underlying attribute of relying heavily on peer-to-peer working for their basic operation. Comparison with Internet application traffic patterns quickly highlights peer-to-peer file sharing as being similar in operation to mesh peer-to-peer; indeed Chapter 4 showed that localised services would exploit the benefits of a mesh very well. If all peer-to-peer users were within the same mesh and if all sources of such traffic were also within the same file sharing mesh, then traffic would be localised within that mesh.

Example applications could be local file shares of local content, for example music files, video clips. Privately generated content would be freely distributable this way, although digital rights management (DRM) issues could quash plans for peer-to-peer file sharing with commercial content in the near future, depending on the licensing model. Another issue with peer-to-peer is the question of what the providers' charging and revenue model will be, since it may be that no traffic leaves the mesh.

At the very least some way of measuring the amount of each users' local traffic would be needed, or else a creatively marketed flat rate scheme would need to be offered to users. Finally, of course, there is no guarantee that peer-to-peer file sharers will all be within the same mesh. The closest potential application would seem to be home or office indoor networking, Section 8.1.4.

In summary, it is not clear that mobile meshes will enable any hitherto unattainable services which would create revenue for a telecommunications operator, although meshes may affect traffic patterns for existing services in the operator network. In other words, we can see no applications which a mesh alone could support, although they can deliver familiar services in new and potentially attractive ways. The case for meshes used within wireless sensor networks is quite distinct, and we shall look at this in Chapter 10.

We thus concentrate on providing already known service types over the six most likely application examples already identified in Chapter 2.

6.2.2 What degree of mobility could a mesh support?

The examination of capacity and efficiency in earlier chapters referred to predominantly physical factors. Those efficiency conclusions can never be diminished by consideration of higher level factors (transport, routing etc.), since these factors merely add yet more overhead and hence further reduce efficiency.

However, more detailed consideration of such higher level factors becomes key if a mesh is to be deployed realistically, so the cumulative, complicating effect on system performance may be assessed. This is the objective of the following discussion on mobility and, in Section 6.3.1, of quality of service. These two issues are quite complex and interlinked.

The effects of mobility are first put into context, followed by an examination of routing.

Effects of mobility

Here we are considering specifically mobile ad hoc networks using mesh approaches. Clearly the degree of mobility involved must strongly affect

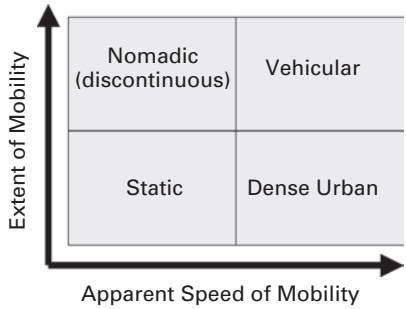


Figure 6.2 Application scenarios: matrix of range versus apparent speed of node mobility.

the results we observe, since mobility can vary from nomadic, for example a docking laptop on the one hand, to vehicular mobility at high speeds on the other hand. In summary,

- nomadic mobility is a limited form of mobility which eases system design, and
- truly ad hoc mobility is largely unfettered and is thus the hardest for routing to deal with.

When modelling or analysing mesh networks, the 'mobility model' assumption and the parameters used will be key. A high mobility node will be likely to pass by many other nodes and will thus cause many route changes in the mesh over the duration of the communication – and this will stress the routing algorithms most severely. This will lead to overheads via proportionally more control packets and fewer data packets being sent and hence to an overall poorer throughput being seen by all users. Mobility models themselves are expanded upon in the Appendix.

Figure 6.2 shows mobility scenarios mapped onto axes of extent and speed of mobility. Note that the x axis is labelled *apparent* speed of mobility – this is to convey the idea that the concept of speed is measured with respect to how many other nodes (which may also be mobile) are passed by the mobile node in question, rather than how fast it moves relative to a fixed point in space. The reason for this is that the number of nodes passed will have direct relevance for the routing algorithm, which

will have to discover new routes at a pace related to node speed measured in this way. This is why dense urban joins vehicular as a high speed scenario. In the dense urban case, even though absolute speed is low, there is a high node concentration: many other mobile nodes will be seen by the user's mobile node, meaning the relative speed is high. The routing protocol must work quickly to adapt to this, maybe even more quickly than in the vehicular case.

In contrast, nomadic computing is a slow, discontinuous process. Although a large range may be involved as a worker travels from fixed location to fixed location, the connectivity expected by the nomadic worker is discontinuous. When travelling, there is often no connectivity, nor any need for it.

Figure 6.2 is included here specifically to point out that the critical aspect of dealing with mobility is how quickly the system must adapt (e.g. routing) and this is not always obvious from first impressions of the generic application scenario. This might be borne in mind as the impact of mobility on throughput and routing protocol overhead is discussed later in this chapter.

The need for new mobility modelling processes

Little is really known about how nodes may move within a mesh, since few meshes have been deployed as yet, although one likely example is clustering around distributed hotspots. Thus we may imagine that nodes will move as a group around a hotspot. Where there is more than one hotspot, nodes will cluster generally, but individual nodes will move between the two (or more) hotspots. Although Camp *et al.* [1] (discussed in the Appendix) do cover group mobility models, no model exists for the above example. This is because prior models have assumed the military or emergency services scenarios, where nodes do not swap between hotspots (indeed there are no hotspots), rather they stay grouped in some form and the whole group moves as one, perhaps amongst other fixed membership groups.

A new mobility model for mesh networks is a ripe area for future research work. If real-world motion traces of node movements become available, they could be used directly or used to adapt models which could remain computationally attractive whilst still being representative. More

information on current modelling approaches can be found in Boukerche and Bononi's review paper [2].

The impact of mobility on throughput: volatility

One of the generally agreed conclusions from modelling work is that, for maximising network capacity and per-user throughput, it is more beneficial to use many short hops through the mesh (giving high frequency re-use) than it is to use longer hops (lower frequency re-use) to keep the peak relayed traffic level down at each node.

However, although this is a reasonably well accepted principle, there are limits to it. As the hop length reduces and hop count increases there is an increasing volatility in the routes through the mesh and hence an increasing overhead for route management and a re-transmission of lost packets. In other words, the mesh may eventually break up into pieces due to increased node mobility causing routing failures. This is termed 'partitioning'.

This volatility may be due to the following:

- potentially increased dependence on the accurate measurement of interference levels and increased susceptibility to short-term changes in these interference levels, as nodes move and/or the radio environment changes;
- an increased probability of link breakages as nodes move and the hop count increases;
- an increased susceptibility to errors in transmitter automatic power control (APC) if this is used to maintain a target signal-to-interference ratio on each link.

One interpretation of this is to say that the network becomes more 'brittle' as its capacity is increased by the spatial diversity techniques inherent in (mobile) mesh networking. Thus in practice it may be the network's stability and the integrity of its QoS which are likely to be real limits to its capacity. In other words, loading up the network may induce failures before capacity limits are reached.

Example of network 'brittleness'

One informative example of this 'brittleness' is given in a system simulation by Hsieh and Sivakumar [3] which explores the volatility of routes

through a mesh as nodes move. Although this is a specific 802.11-based simulation it does illustrate some interesting characteristics which may be expected to hold for mesh networks in general. This example is examined in detail in the remainder of this section.

The simulation sets all node transmissions to the same power level – regardless of the hop length between node pairs. Although this is not representative of a practical deployment it serves to illustrate the effect of power level (and hence link budget) on system performance.

For a random distribution of static nodes and routes, the simulation first increases the transmit power of all nodes until it determines a necessary minimum power level (defined as MIN) at which the routes are fully connected. Simulations are then run for increased power levels. Every node acts as a constant bit rate (CBR) traffic source.

The impact of node mobility is investigated by introducing the way-point mobility model (see Appendix), where the pause period is set to zero in order to simulate continuous motion. Movement speeds are 5, 10, 15 and 20 metres per second.

As a benchmark, the mean per-user throughput is first characterised for static users with a 100 node network and an offered traffic rate of 64 kbps per node, i.e. before the simulation of user motion.

This simulation showed that, as the transmit power is increased above the minimum required for connectivity, there was initially little change in the per-user throughput. This is because the network was being run below maximum capacity and so could tolerate the additional mutual interference introduced. As the transmit power was increased further the mutual interference started to dominate and throughput suffered.

Under these static conditions the variation in throughput amongst users was quite low.

Network partitions caused by mobility

Since initially the transmit power was set at a level sufficient just to achieve connectivity it is unsurprising that any motion started to cause disconnections. As the transmit power level was increased above the minimum the probability of partitioning decreased to zero again. When

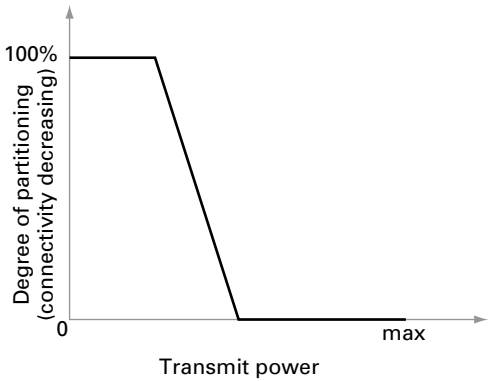


Figure 6.3 Occurrence of network partitions as a function of transmit power level.

it occurs, the effect of partitioning can be to lose key interconnections and thus global re-routing, i.e. routing to each and every node in the simulation, becomes impossible.

In summary, when user mobility was introduced, connections were lost and the network broke into isolated partitions. Increasing the transmit power from the minimum needed for static connectivity reversed the partitioning effect. In other words, the mesh was repaired and the percentage of the network which is partitioned fell back to zero. Figure 6.3 shows the general form of this effect, where 0% partitioning corresponds to a fully connected network.

Further conclusions from Hsieh and Sivakumar [3] included the following.

- In the mid-power region, where connectivity is good and per-node throughput is reasonable, there is a high probability of a need for re-routing.
- Near the minimum power setting there are fewer re-routes implemented, however there is a high level of network partitioning (in which state re-routing is not possible so end-to-end links are lost).
- At high power settings there are also fewer re-route occurrences, however there are now fewer active routes because capacity is very low (due to the increased level of mutual interference at these

higher transmit powers). Hence high power is generally not a good solution.

In summary, only at high transmit power levels is there a reduction in the percentage of link failures due to mobility. This is a direct consequence of the fact that the coverage area around each active node is so large that mobility does not cause links to break during the average period of a message transfer. However, at these high power levels the network capacity is very low due to the increased level of mutual interference.

This is clear illustration of a trade-off between resilience to mobility and network capacity.

6.3 Improving quality of service by adding network infrastructure

The example just given above was for a pure mesh with no additional infrastructure. This section now leads into examining how infrastructure might help.

It has become abundantly clear that there is a stark difference between mesh and cellular operation when considering user mobility and network performance. In a mesh, the network performance (e.g. availability/outage, service level) depends on the users themselves. It depends on their mobility, availability and the traffic carried. This makes it strictly impossible for an operator to offer a service level guarantee for a pure mesh, unless the dependence on users is mitigated somehow. Given that the problem occurs due to the uncertainty over node behaviour (i.e. the users), one solution must be to inject some certainty in the form of fixed nodes. How many to add depends on the projected user mobility and traffic levels. In a practical situation, it seems likely that a precautionary design margin will be required, since accurate prediction of user behaviour is likely to be limited.

This section begins by looking at whether QoS is at all realistic in a mesh.

A useful point to bear in mind whilst reading this section is that, in adding infrastructure to gain the ability to guarantee service levels, the opportunity should be taken to ensure that it is added in such a way

that the additional nodes also ensure scalability (in the same sense as cellular, via the incremental base station/access point concept), plus the infrastructure should be sufficient to 'seed' the network from day zero, when few real users may be available to provide mesh connectivity.

6.3.1 Could a mesh guarantee a quality of service?

Different applications will require different qualities of service and one very important attribute within a quality of service offering will be delay and its variation over time. We have already said that applications may be described as elastic or inelastic, meaning they may be tolerant or intolerant of delay (also called latency). Email is elastic as the time taken for delivery is not critical as long as it is within reasonable limits. Much of the LAN based ad hoc networking literature assumes elastic applications are the current norm and hence regular store and forward transport will be appropriate. This is unlikely to be true into the future of fixed mobile convergence, where new applications such as video calling or multimedia entertainment will certainly place inelastic demands on the network.

By way of an inelastic application example, whilst not specific to meshes, it is worthy of note that mobile displays are being called the 'fourth screen' concept (i.e. cinema, TV, PC, mobile). Fox Entertainment has its TV series '24' in 'mobisode' format already. This presently has a one minute run time, but is expected to be expanded. The fourth screen is specifically seen as a new entertainment medium which requires bespoke content creation. The famous Edinburgh Film Festival in the UK now has a 'films for mobiles' category.

Having confirmed the need for quality of service levels, an early question to be asked is whether quality of service is possible at all within an uncontrolled interference environment such as a mesh, since unlike cellular, mobile ad hoc mesh networks cannot be planned for a known interference environment *ab initio*, due to user movement. This will be at odds with the expectations of users, who will require guaranteed service for inelastic applications like video or anything else intolerant of delay variation or throughput variation.

By uncontrolled interference environment, here we primarily mean self-interference from other users, not interference from outside the mesh. Interference from outside is important of course, since this would also affect the quality of service; we looked at this in Chapter 5.

Perhaps because of this basic difficulty with self-interference, which is so different to the cellular case, many LAN biased research efforts have concentrated on the notion of best effort delivery, as used in the main on the Internet today. Such operation is often facilitated by a technique called 'store-and-forward', with the main objective being 100% delivery success (but this is in no way guaranteed, hence the need for the transmission control protocol, TCP). Latency is very much a secondary objective with a very low priority. Perhaps as evidence for this, some commercial latency figures might appear to be driven more by the desire of not overloading storage space queues in transmission equipment, rather than by the service level offered to the user. Conversely, from the cellular and telecoms world, service level agreements are much more the norm. These arrangements vary widely but generally incorporate guarantees of availability (uptime) as well as bandwidth, latency and error rate.

For whatever reason, the quality of service aspect of meshes does not always receive the attention it deserves. The move to being much more reliant on user behaviour for network performance is a situation starkly absent from present mobile networking. Its effects can be disruptive and it is examined next.

6.3.2 Dependence of QoS on user behaviour

Some published literature in this field is from the proponents of cellular extension by multi-hopping. This technique will be discussed more closely in Chapter 8. However, for the time being it is sufficient to appreciate that the technique involves extending the reach of a base station by allowing nodes to act as relays, such that links to the base station may be greater than the one hop of traditional cellular. In this way it is also hoped to increase coverage (by filling in black spots) as well as increase range. As such it is pertinent to multi-hop mesh networking.

Mobility versus connectivity

We may gain some insight into this via the work of Royer *et al*. [4] who noted the result of a well-known publication which stated that a node with the 'magic number' of six neighbours could be shown to possess the optimum trade-off between transmission power and self-interference for best throughput [5]. In other words, with more or less than six neighbours, throughput was shown to fall due to either increased relay hops, or bandwidth lost due to the bigger interference footprint of neighbouring nodes. This was derived for fixed networks and Royer *et al*.'s question was whether a similar optimum existed for mobile networks.

Their answer was 'no'. There is no single optimum power-efficiency trade-off for a range of node mobilities, although there is an optimum for each value of node mobility. Essentially, as the node mobility increases, Royer *et al*. found that the node density, in terms of connectivity, needed to rise: more mobile nodes need more neighbours. This could be ensured by increasing transmit power or by adding more nodes. The concept is quite easy to rationalise in principle: as nodes move more quickly, the likelihood of any link breakage increases along with their speed (already noted earlier in this chapter). In order to have a higher probability of remaining connected by some route, the number of neighbours within radio range must be increased by some method.

The clear issue for the mesh network planner is that network connectivity depends on parameters outside his control – the users. Only by adding some permanent seed or relay nodes can some degree of control be reasserted. This is next examined a little more closely and we note that what is really under discussion incorporates the notions of availability and quality of service.

Coverage, availability and quality of service

Firstly Royer *et al*. [4] showed that connectivity depended on user mobility. More recently Nilsson [6] showed that connectivity also depends on the traffic level within a multi-hop mesh.

Nilsson extended a similar approach to that of Royer *et al*., using a similar routing protocol in his model, plus a more modern protocol. In each case, he confirmed the dependence of packet delivery success on

user mobility and additionally noted that at low traffic loads the packet delivery rate could increase quickly with increasing transmit power (i.e. increasing connectivity). Conversely, as the traffic load increased, the rate of improvement in packet success rate slowed down. His conclusion was that both predicted traffic levels and node mobility/densities would need to be known at the planning stage, for viable network design.

Lungaro and Wallin [7] expanded on Lungaro's earlier work [8], where notably they added a caveat to his earlier enthusiasm for cellular multi-hopping. They noted that the principal drawback of the scheme was that, due to the 'partially uncontrolled infrastructure' (i.e. the user nodes), multi-hopping was not able to guarantee a quality of service. They proposed that a simple solution was the addition of nodes as relays, by the network operator. In fact they proposed a system with three node types: access node, relay node and user node. Their conclusion goes beyond Nilsson's; knowing traffic and user mobility is still not enough to guarantee quality of service in a mesh of user nodes. Infrastructure must be added. This is a key conclusion.

6.3.3 Directed QoS

Accepting this conclusion from Lungaro and Wallin, a contemporary paper from Sanzgiri and Belding-Royer [9] suggests a novel way of 'finding' the required quality of service within a mesh by directing the user to a new physical location within the mesh where the desired quality of service is known to be available. A protocol to achieve this was described. This is analogous to walking around in a marginal cellular coverage area until a stronger signal is found. It is not clear how acceptable this would be to users.

6.4 Quality of service summary

We began by defining QoS and looking at present and likely future drivers and requirements. Answering the three key questions within this section provided the following conclusions.

- The provision of current services via a mesh is the focus. Mesh itself does not enable any new services, although it may enable existing services to be delivered in an alternative way.
- The degree of mobility which a mesh can support should be measured by the number of other user nodes passed rather than by measuring speed relative to a fixed point. This metric begins to quantify the number of re-routes required of the protocol, which is a similar concept to cell hand-overs – both create an overhead for a similar, necessary reason.
- For large enough mobility, a mesh can break into several disconnected pieces, a process termed 'partitioning'. Before partitioning, an active mobile mesh may require as many routing packets as data packets for its operation. Further quantification of mesh mobility characteristics (as a planner would require) appears to be a large and open research subject well beyond the scope of this book.
- Quality of service within a mesh is user dependent, in terms of both mobility and traffic level. This effectively means that an operator cannot guarantee a quality of service level. There are two ways around this problem:
  1. add infrastructure, such as fixed nodes;
  2. direct the user to relocate physically to a 'better' part of the mesh, via a suitable protocol.

The key conclusion above is that the quality of service within a mesh is user dependent, in terms of both mobility and traffic level. Unless specific additional steps are taken to mitigate this, such as operator controlled nodes, an operator will be unable to provide service level guarantees. Even so such mitigation will need to be tuned to the actual mobility and traffic circumstances in each case, if they are known.

Beginning with Chapter 7 the book moves away from primarily theoretical and technical considerations towards practical deployment considerations and a closer examination of our six most likely application examples.

References

1. Camp T., Boleng J., Davies V., A survey of mobility models for ad hoc network research, *Wireless Communication and Mobile Computing*

(WCMC), Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, **2** (5), pp. 483–502, 2002.

2. Boukerche A., Bononi L., Simulation and modelling of wireless, mobile and ad hoc networks, in Basagni *et al*. (ed.), *Mobile Ad Hoc Networking*, IEEE Press, Wiley, 2004.
3. Hsieh H.-Y., Sivakumar R., Performance comparison of cellular and multi-hop wireless networks: a quantitative study, *Proc. ACM SIGMETRICS*, June 2001.
4. Royer E.M., Melliar-Smith P.M., Moser L.E., An analysis of the optimum node density for ad hoc mobile networks, *Proc. IEEE Commun.*, June 2001.
5. Kleinrock L., Sylvester J., Optimum transmission radii for packet radio networks or why six is a magic number, *Proc. IEEE National Telecommunications Conference*, Birmingham, AL, December 1978, pp. 4.3.1–4.3.5.
6. Nilsson A., Performance analysis of traffic load and node density in ad hoc networks, *Proc. European Wireless (EW2004)*, Barcelona, Spain, 24–27 February 2004.
7. Lungaro P., Wallin E., Coverage, capacity and QoS tradeoffs in hybrid multi-hop ad hoc cellular access systems, *Proc. Affordable Wireless Services and Infrastructure, 1st Annual Workshop*, Tammsvik, Bro, Sweden, 3–4 June 2003.
8. Lungaro P., *Coverage and Capacity in Hybrid Multi-Hop Ad Hoc Cellular*, Master of Science Thesis, TRITA-S3-RST-0315, Radio Dept. KTH, Stockholm, 2003.
9. Sanzgiri K., Belding-Royer E., Leveraging mobility to improve quality of service in mobile networks, *Proc. MobiQuitous* 2004, Boston, MA, August 2004.

7 Summary of potential mesh pitfalls to avoid

As we have seen from the previous chapters, there are numerous key considerations to bear in mind when planning to implement a mesh. Some of these key considerations, if not properly addressed, constitute potential pitfalls for the mesh system designer. The aim of this short chapter is to bring all such considerations together for easy reference, so the pitfalls may be avoided. This is particularly appropriate as not all pitfalls have familiar equivalents outside the world of mesh networking.

In summary, potential pitfalls already covered in the body of this book centred around

1. capacity,
2. infrastructure,
3. efficiency,
4. relay exhaustion,
5. initial roll-out,
6. upgradeability,
7. reliance of the system on user behaviour, and
8. ad hoc versus quality of service.

There are also two areas which we have covered implicitly, but now wish to highlight explicitly here:

9. security and trust, and
10. system economics.

Let us deal with these areas in turn.

7.1 Capacity

In Chapter 4 we noted that it was often rumoured that meshes self-generate capacity, as if this were a truism. The reasoning behind such a

claim was usually along the lines of 'each new user brings additional capacity to the mesh', or 'each new user effectively becomes a base station'. This book critically examined such statements and separated the reality from a something-for-nothing type of mythology. We outlined the difference between network capacity and the user throughput which is actually available, concluding that user throughput cannot grow as fast as the mesh grows. The simple reason is the relay requirement imposed on each node, due to the traffic of other nodes.

The pitfall to avoid is simply one of believing unrealistic performance claims.

7.2 Infrastructure

Following on from the findings on capacity, we also noted in Chapter 4 that for a mesh network to scale, such that performance is maintained as user numbers are increased, then infrastructure must be added. The purpose of this infrastructure is to separate local traffic from traffic which has a more distant destination, including external to the mesh. In this way the relay load, referred to above, may be limited.

The pitfall to avoid is trying to design a pure mesh where one is not strictly dictated by the application, since one with infrastructure would either be more capable, or would require less complex nodes for similar capability.

7.3 Efficiency

We noted two components of efficiency, firstly concerned with a single link in Chapter 4 and secondly with medium access control in Chapter 5. Whereas it is clear that there are efficiency advantages in dividing a single link into two equal hops, as is typical of a multi-hopping mesh, we questioned how often this could be relied upon to happen in a real deployment. We saw that if the split was not 50:50, then the advantage decreased, and if the hops introduced a kink or dog-leg in the path, then the extra path length quickly negated any advantage. Finally we noted that for the links from nodes to access points, the optimum path split depended on the power difference of the access point and the normal node.

In terms of MAC efficiency, the basic conclusion was that a mesh typically forces a decentralised approach. Most often this is achieved by a random access protocol with some back-off behaviour. This can never be as efficient as a centralised, scheduled MAC, such as TDMA.

The pitfall to avoid is to assume that meshes can be relied upon to bring increased efficiency relative to the more common point to multi-point approaches, such as cellular.

7.4 Relay exhaustion

Battery life is an issue for a number of reasons, to which we alluded briefly in Chapters 2 and 3. Although it is true that mesh nodes are transmitting at lower powers due to reduced link losses relative to typical cellular systems, their duty cycle of transmission is increased due to the relaying requirements. Such relaying requirements can be very demanding in specific circumstances.

In particular, nodes very close to an access point will suffer much more than nodes at the edge of a mesh. Such 'relay exhaustion' occurs when a node's battery power fails due to its over-use as a relay. Routing which could take into account the battery level of the nodes in the path appears to be well beyond any property of routing protocols under standards development for communications networks, although it is under consideration for wireless sensor networks, where trade-offs, for example against delay, may well be more acceptable, see Chapter 10. In any case, vulnerable nodes next to an access point could only benefit if traffic were routed to a completely different access point.

Additionally, this relay exhaustion problem can be exacerbated by the behaviour of users, which we cover below.

The pitfall to avoid is not designing mitigation for relay exhaustion scenarios.

7.5 Initial roll-out

When rolling out a new network, seed nodes may be needed. This situation has often been seen in practice with fixed meshes and is

beginning to be seen in reports of US 802.11-based mobile mesh deployments, which use dedicated relay nodes mounted on street corners or intersections. As we said in Chapter 6, the implication is that some infrastructure will be needed from day one, and probably for a longer time if predictable quality of service is desired, since this depends on user node density, mobility and traffic. These infrastructure nodes could be relatively stand-alone in that they may not all need data interconnections via a wired backbone, but they will need power supplied. As we note below, this kind of infrastructure may also alleviate problems created by the behaviour of users.

The pitfall to avoid is forgetting that connectivity will probably be at a minimum when the network is first rolled out in any new area.

7.6 Upgradeability

Upgrading a mesh for higher performance requires each and every node to be upgraded; it cannot be done piecemeal as it can for cellular.[1] In this respect the performance and capability of user terminals must be sufficiently well scoped to cover the service life of the network.

As we noted in Chapter 6, the total capacity of any mesh radio node must always exceed the capacity presented to the node user, in order to allow for the relay overhead which is intrinsic to mesh operation.

The pitfall to avoid is having to upgrade all nodes too early in the system life, due to having under-designed them at the outset.

7.7 Reliance on user behaviour

One of the interesting aspects of mesh is the co-operation aspect required of users; for the mesh to function it requires nodes to allow the use of their equipment to support other users, for example in the basic relay process. This becomes especially significant in the case of mobile applications where battery capacity is at a premium.

[1] In cellular, only the base station and chosen user nodes need be upgraded; in a mesh any node could be on the critical path at any time, so all nodes need to be upgraded.

We have stated above that there are circumstances where a user might find his battery exhausted by supporting other users without ever making a call on his own account. This could give rise to a strategy amongst users of not switching on their equipment until wanting to make a call, thus becoming 'selfish users'. Whilst it would initially appear to be rational to be a selfish user, in fact it would reduce the density of active nodes, thus compromising the integrity and performance of the mesh for all users. It would also impose an increased load on well-behaved users, speeding up the exhaustion of their batteries and thus perhaps ultimately speeding their conversion to becoming selfish users themselves. Techniques have been published for 'encouraging' proper behaviour, but these come with an overhead. Such schemes seek to reward users who leave their nodes on. The reward scheme method uses extra packets which monitor the end-to-end link in both directions; this takes up valuable resources which could otherwise be used for traffic [1].

In Chapter 6 we also looked at node mobility and brittleness. We showed that a mesh can be broken into several disconnected partitions by the mobility of users. Where users move beyond normal transmission range, link breakages occur. But, whilst these can be corrected by increasing transmission power, this increases interference and thus reduces capacity. Additionally, the traffic carried by nodes also affects mesh integrity by saturation of node capacity over certain crossing routes where capacity requirements are concentrated. Such vulnerabilities in mesh are often described as its brittleness.

The pitfall to avoid is forgetting that, unlike cellular, user behaviour can have dramatic effects in a mesh.

7.8 Ad hoc versus quality of service

Perhaps the largest attraction of meshes is that they can be entirely unplanned in pure form. To a service provider, the lure of a network which promises no planning phase must be high. But we have shown in Chapter 6 that ad hoc systems are poor supporters of quality of service.

The pitfall to avoid is missing the fact that there is a compromise between maintaining an ad hoc environment and facilitating quality of service.

7.9 Security and trust

Although we have covered it only implicitly, security is an important concern. Meshes will inherit all the security issues of radio systems in general and add some of their own. In particular, data will usually be transiting third party equipment not belonging to either the user or the operator, i.e. other user's nodes. It will therefore be more vulnerable to capture, delay and manipulation. Security concerns will also centre around user authentication. In a strictly ad hoc network there is no parallel to the present central authentication function of cellular networks which use AAA (access, authentication and authorisation) servers.

Security means both control over which nodes are allowed to join a mesh, and also integrity of the message (and its declared end points) during transit. This subject is strictly outside the scope of this book, but the following is included for consideration.

- Security issues arise when
  a new user wishes to join a trusted mesh, and
  user traffic must transit a third party device via hopping.
- Attacks can also be denial of service (DoS) via connection overload or via targeting battery life exhaustion in portables.

In summary, the challenge of security in an ad hoc environment is large and an open research issue. There are all the usual wireless problems and more, since the whole concept of ad hoc networking is contrary to the usual security approaches of centralised access control etc. Also related to the lack of central control are the network management and billing requirements, which will be more difficult. Ways around these problems include the introduction of 'special' nodes, such as the concept of trust centres as in ZigBee. We cover ZigBee in Chapter 10.

The pitfall to avoid is the very common one of not considering security early enough in the system design process.

7.10 Business case economics

We wish to add a final note on a non-technical aspect. Presumably the point of deploying a mesh system is to support or create a viable business.

The mesh must therefore have a sound business case, which usually encompasses identifying a customer need, a suitable technology and a sustainable revenue source.

With respect to the business case, in this book we have thus far identified our list of six likely mesh applications, thus identifying a market pull. We have also spent much time showing the capabilities of the technology. At this point we feel duty bound to make doubly clear that whilst these are necessary conditions for the success of a mesh deployment, they are not sufficient. In addition, of course, there must be a robust money making model. We will examine examples of this in Chapter 9.

The pitfall to avoid is one of being technology driven and forgetting the money making model.

7.11 Enduring attractions of mesh

The point in highlighting all the pitfalls is that they are only potential pitfalls, not necessary pitfalls. To see the benefits of mesh at their best, clearly we must not allow ourselves to fall into any of the pitfalls. Perhaps the most instructive aspect of the pitfalls we have described is that they are not already familiar to us from our experiences of non-mesh systems. Without specific preparation the danger is thus that we will not be ready for them.

To recap, if all the pitfalls are avoided, meshes do have some strongly attractive features, notably in the area of coverage extension or in-fill, where they offer complementary performance to that of cellular systems, as we first pointed out in Section 2.2.

The next chapter takes a closer look at our six most likely applications for mesh networking. Following that we will look at some successful mesh implementations in Chapter 9.

Reference

1. Salem N., Buttyan L., Hubaux J., Jakobsson M., A charging and rewarding scheme for packet forwarding, *ACM MobiHoc*, Annapolis, MD, 2003.

8 Appropriate telecommunications applications for mesh

To summarise once more, at this point in the book it has been shown that practical mobile meshes are not chosen primarily for spectral efficiency nor for any notion of self-generation of capacity. Meshes should be chosen because they have other benefits. Section 2.2 provided an introduction to how meshes offer coverage benefits, which is possibly their major attribute. In this chapter we revisit our six most likely applications which we have been considering throughout the book. These are

- cellular multi-hopping or WiFi hotspot extension,
- community networking,
- home and office indoor networking,
- micro base station backhaul,
- vehicle ad hoc networks (VANETs), and
- wireless sensor networks (WSNs).

The first five applications are considered in detail in this chapter, whilst wireless sensor networks receive their own treatment in Chapter 10, since they have some unique features. In this chapter, we also look at the barriers to mesh adoption and the time scales likely for them to be overcome.

For the following discussion we find it useful to group the applications into those which form a mesh on the user side and those which form a mesh on the network side, in other words those where the users' nodes themselves mesh together, versus those where only the backhaul forms a mesh. There is one case where the mesh can be for both users and network backhaul; this occurs in VANETs.

8.1 User side mesh applications

This includes

- cellular multi-hopping or WiFi hotspot extension,
- community networking, and
- home and office indoor networking.

We introduced these, with diagrams, in Chapter 2. We begin with a generalised consideration of how multi-hopping can generate an effective increase in cell coverage, i.e. the principle on which all three of these applications is based.

8.1.1 Cell boundary theory

An increase in usable cell radius can be achieved by multi-hopping. As this is so central to user side meshing, we explain and test the theory here, showing that an upper bound exists for cell extension. We believe the origin of this upper bound may not have been previously taken into account. For example, in the literature, Lungaro [1] presents the results of comprehensive traffic simulation modelling on a TDMA network and concludes that of the order of ×3 cell radius extension is achievable. We feel such estimates may be on the high side, for the following reason.

An aspect the above modelling paper does not consider is any limit which may be caused by high traffic loading on a large network. We will show that a major limit to range extension will be the finite throughput capacity of nodes in direct range of the base station. These must carry all the traffic for the whole extended cell coverage.

Considering this aspect allows us to derive an upper bound for cell extension.

Traffic to nodes outside the nominal range of the base station (or access point) is relayed via nodes within its range. Therefore, all of this outer traffic must pass through nodes in an area describing an annulus at the extremity of the base station's range. The width of this annulus must be the node to node range, as we show Figure 8.1.

To develop an upper bound, we make the assumption that the entire unused throughput of a node is consumed by acting as a relay. Then, from a consideration of the traffic concentration effect by taking a ratio of areas, the upper limit of range extension which can occur can be shown to be given by the expression:

$$E\frac{X^2R^2 - R^2}{R^2 - (R-r)^2} \leq (1-E)$$

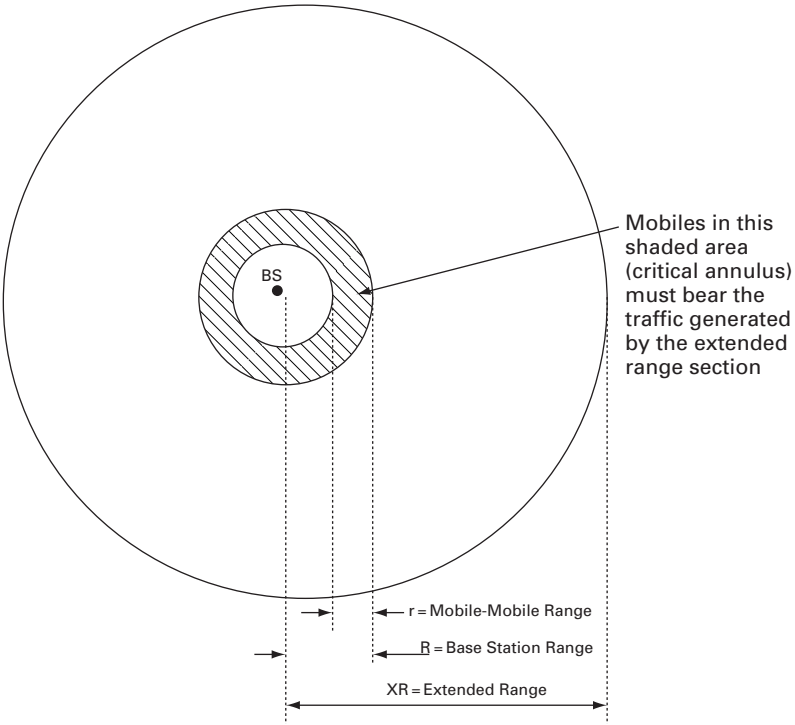


Figure 8.1 Cell boundary extension footprint.

where E is the offered traffic per user (Erlang), R is the mean base station to mobile communication range, r is the mean mobile to mobile communication range, and X is the cell boundary extension factor.

We can explain the origin of the above expression as follows. We can rephrase our condition of all unused throughput being used to relay extended traffic by saying (the traffic capacity in the outer ring) must be less than or equal to (the unused traffic capacity in the annulus). Plus, if the used capacity is E then the unused capacity must be $1-E$, under our conditions. Creating the above equation is then simply a matter of using geometry to find the appropriate areas; the numerator is the area of the outer ring and the denominator is the area of the annulus.

Going further, if we define the ratio of base station and node ranges as $Z = R/r$ then our expression becomes

$$E(X^2 - 1)\frac{Z^2}{2Z - 1} \leq (1 - E).$$

The important parameters are the resulting cell extension factor X, the ratio of the base station to node range Z, which we can choose, and the Erlang traffic loading which is controlled by the users. All other things being fixed by design, this implies that range extension scales inversely with the square root of the traffic loading.

In deriving the expressions above we have made simplifying assumptions, including perfect load balancing, even utilisation and uniform user density. By way of illustration, consider today's typical mobile-phone usage. We may assume a busy hour loading of about 0.04 Erlang. An approximate range for Z is 3 to 6 assuming 13–20 dB additional link margin base station–node versus node–node links and for a propagation law in the range of 3 to 4. Figure 8.2 illustrates this for a range of Z and E: there is a clear tendency towards a value of approximately 2.

Our interpretation of this is that diminishing returns exist for attempts to increase the cell extension factor by increasing the ratio of base station range to node range. In the figure we see that a cell extension factor of about 2 seems to be an asymptotic limit. In this case, designing base

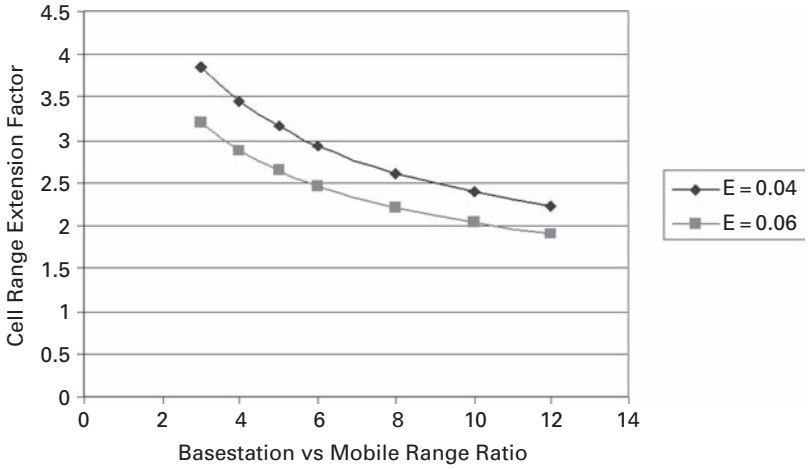


Figure 8.2 Cellular range extension through multi-hopping (upper limits).

station ranges more than 10 times node ranges would become an increasingly wasted effort.

Having described the common cell extension principle, we next look at our three user side mesh applications in turn.

8.1.2 Cellular multi-hop or WLAN hotspot extension

This is the prospect of multi-hopping between mobile, personal digital assistants (PDAs), laptops or other terminals within some future cellular or WLAN network. Within the cellular industry this proposed architecture is called 'cellular with multi-hopping'. In the data communications industry an implementation is currently being discussed within IEEE 802.16j. The concept is one of using the relaying function either

1. to extend the range of higher bandwidth services, or
2. to increase coverage availability within a cell.

Each of these offers the potential benefit of reduced infrastructure costs via fewer base sites. An additional bonus, given the discussion of Section 4.7, must also surely be the potential to establish services at much higher frequencies than presently used, in spectrum which is currently less highly valued. This is of interest since contiguous cellular or hotspot coverage with the present model is highly unlikely at higher frequencies, because of reduced link budget and higher clutter losses. But multi-hopping via relay nodes could potentially offer access to this new, more cost-effective spectrum.

For example, in the case of 3G, with its higher transmission rate capability, the poor availability of high data rate services over a cell's coverage area is an acknowledged weakness. Operators are unlikely to invest in the necessary additional infrastructure to remedy this over their entire service areas, but would welcome a more cost-effective solution.

WLAN hotspot extension is very similar to multi-hop cellular in its basic aim. However, there is a potential reduction in service expectation, which may make WiFi hotspot extension more easily deployable than multi-hop cellular. This is a perceived lesser requirement for the prime

quality of service parameters of delay and delay variation (recall that we showed QoS issues in Chapter 6). But this is an assumption based on the present traffic types for WLAN having a greater proportion of applications which place only elastic demands on the network, for example email or web browsing. However, we must be cautious, since such an assumption may not hold well into the future, as cell phones become more like laptops and laptops begin to be used for VoIP.

Points which mitigate against the take-up of cellular or WLAN multihopping are security on the one hand and billing on the other. Both are problems, as we pointed out in Chapter 7, since user data transit nodes owned by other users. This introduces a security question and a question of how to recompense or otherwise encourage those users who are part of a relay chain to maintain their participation.

Despite a level of interest, we know of no actual deployments of this application.

8.1.3 Community networking

Where a remote community has no broadband connection, installing just one connection and sharing it can be the most effective solution. The user nodes are meshed together without infrastructure and one mesh node is connected to the broadband backhaul.

There are several examples of shared ADSL services in community networks. Remote communities can use the mesh to share a single expensive internet link, like a satellite or leased line, among enough users to make the service affordable. A T1 or satellite connection is often out of reach of individual small businesses and personal users, but if there is enough local interest then the combined purchasing power can tip the balance and help to provide excellent value within the local community.

Nevertheless we do not see this as an area of major growth, since most developed populations are well served by broadband, outside niche rural areas. It has however gained a foothold and could grow rapidly in less developed countries. It is for this reason that we include it as a good example of mesh usage in Chapter 9.

8.1.4 Home and office indoor networking

This is essentially similar to WLAN hotspot extension with respect to the requirements for mobility, lack of infrastructure, and the need for high bandwidth. But additionally the naturally closed user group transmissions would play to the best attributes of a mesh, as we showed in Section 6.2.1 which discusses potential new mesh based services.

But, whilst there is activity in IEEE 802.11 aimed at meshing with an indoor, ad hoc focus, it is unfortunately the case that the trend in home networking is not in this direction. Rather the trend is towards higher data rates in order to support real-time multimedia services. For example 802.11n is targeting over 100Mbps with low latency LANs. Similarly 802.15.3 is targeting high data rates in the PAN space.

We must conclude that mesh does not fit this application well due to latency issues, limited upgradeability paths and the problem of relay bandwidth. We discussed these pitfalls in Chapter 7.

8.1.5 Conclusion on user side meshing

As we have seen, the suitability of this technique depends heavily on the application. We can generalise by pointing out the pros and cons as follows.

Looking back through the chapters of this book we can see that promised benefits of cellular multi-hopping include the following.

- A reduction of the link-budget planning margin for log-normal fading which traditionally burdens cellular networks.
- A range extension of around ×2 seems achievable in normal circumstances. This represents a significant reduction in base station density. The extent to which the cell boundary can be extended by multi-hopping depends on the ratio of base station to node antenna gains and heights and on the throughput capacity of the relay nodes.

But the technique is not without drawbacks which include:

- quality of service (QoS) issues such that the operator cannot guarantee service levels; this may be tackled by adding fixed infrastructure, but that conflicts with the aim of cellular multi-hop, which is cost reduction via infrastructure reduction;

- an unavoidably increased latency as a result of multi-hopping, which will also affect QoS.

Drawing on the above, the overall conclusion is that the multi-hop technique is useful where QoS and coverage are more important than operating without infrastructure. They would best be implemented by including fixed relay nodes, which have access to a good power supply.

8.2 Network side or backhaul mesh applications

The operator of any wireless access network with many cells must have a requirement to link the cell sites back to an operating centre and then, potentially, the wider wired network. Such a process is usually called backhaul. This need is present whether we consider a cellular operator or a WiFi hotspot operator.

Just to be clear, we are not concerned with the users forming meshes in this case, we are interested in the interconnections between cell sites forming a mesh. This gives rise to the term network side meshing or backhaul meshing.

8.2.1 Micro base station backhaul

We have referred to micro base station backhaul in particular simply because there is currently a lot of activity with smaller cell sites; they are variously called microcells, pico cells and hotspots. The smaller cell sites are driven by the market pull of users wanting more bandwidth. Next generation cellular and WLAN solutions are expected to converge into 4G as we saw in Chapter 1 – and both will be expected to serve higher bandwidth applications.

Even today, 3G networks are being used for higher bandwidth, and their cell coverage areas are necessarily smaller. Smaller cells means more cells deployed more densely and clearly this means more backhaul. Given also that many high bandwidth deployments are to satisfy demand in cities, then backhaul can become a significant capital cost of the deployment. In fact cites are very expensive places to install traditional means of backhaul, such as copper or fibre links. Digging up a city street entails a

staggering cost per mile, often into five figures. The other alternative of a microwave link around 10, 20 or 30 GHz is already used for backhaul, but this does not suit the dense urban landscape.

So the attraction of mesh for backhaul is that its coverage attribute makes it attractive for the urban environment and its lack of a need for traditional infrastructure makes it cheaper to install in urban environments, although we should not forget that even mesh nodes need power supplied. Wireless cities are currently a big driver of interest in mesh backhaul and we will look at some examples of this in the next chapter.

8.3 Joint user and network side mesh applications

User side and network side meshing are not exclusive, as we see in the case of vehicular ad hoc networks.

8.3.1 Vehicular ad hoc networks (VANETs)

Various national transport agencies are most interested in moving to more intelligent transport systems (ITS) in the near future. There are at least three drivers for this, which are improved safety, congestion avoidance and better environmental performance.

The safety aspect of ITS is creating an opportunity for mesh systems. The benefits of improved safety systems are easy to accept; fewer deaths on the roads could be facilitated by the deployment of systems which help drivers to make the right decisions. Many examples exist, the two most commonly cited being electronic brake lights and stop sign warnings. Electronic brake lights (EBL) involve a signal being sent from a car in front to warn the following driver that the first car has applied its brakes. If the brakes were applied very heavily, then the system could go as far as applying the brakes of the second car automatically. Clearly this requires communication from one car to the next – and only a small delay can be tolerated. Stop sign warning is appropriate where a vehicle is approaching a stop sign in the distance. The system could warn the driver that the lights are red before the driver can see them clearly. The system could also advise the driver that at the current speed, stopping by the sign would not

be possible unless swift action were taken. Clearly this requires a communication between the car and the roadside infrastructure, in this case the traffic light.

This safety aspect of ITS has prompted the creation and testing of vehicular ad hoc networks or VANETs, which are wireless networks which interconnect the transport infrastructure and vehicles. These are usually split into car-to-car (C2C) schemes and roadside-to-vehicle (R2V) schemes. The car-to-car schemes usually demand the lowest latency. Trials have been ongoing for some time in Japan, and the USA and EU are catching up. The level of activity and momentum can be realised from the fact that spectrum for these services has already been allocated in Japan and the USA, and the EU has a harmonised spectrum allocation at the final draft stage.

VANETs are interesting since they may be combination meshes, in that both user side (C2C) and network side (R2V) meshing can be appropriate.

We expect a large growth in VANET activity commercially and we also expect that many mesh companies who have been legacy communications focused will now also focus on these new transportation opportunities. We therefore look at real-world VANET examples in the next chapter.

8.4 Time scales

When looking generally at time scales for mass mesh adoption, naturally we must consider technology and regulatory viewpoints. But, interestingly, there are also human factor aspects, due to the unfamiliar way in which meshes operate, such as the selfish user effects and user expectations of service level variation, covered earlier in the book. It is potentially such 'softer', user-focused issues which could have hard to predict effects on the overall time scales for mesh adoption. In terms of regulation, the move away from the command and control method of spectrum management is expected to encourage innovation, but some unexpected regulatory obstacles may need clearing, since much legislation predates mesh networking concepts.

In general it is not the maturity of electronic hardware which is limiting mobile mesh networking. The elements of transport, routing, medium

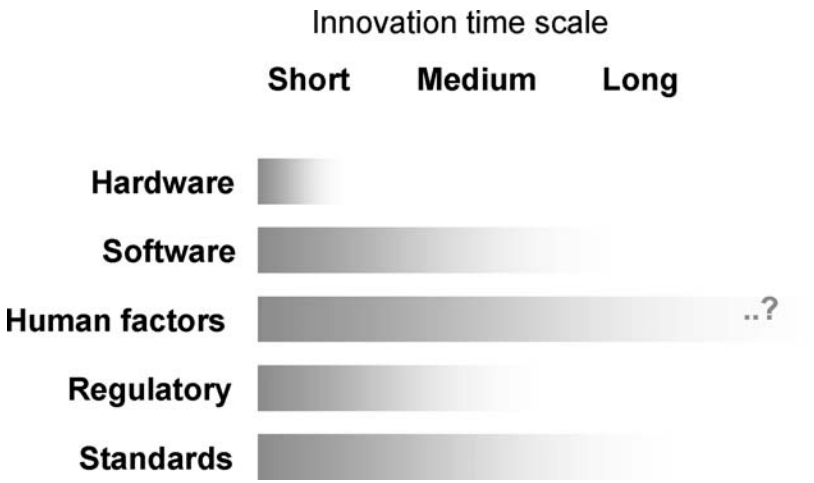


Figure 8.3 Time scales for mesh adoption (prediction).

access and cross-layer protocol co-operation, plus modelling software are notably less mature.

Critically it may be necessary to find means to modify user behaviour if this acts as a deterrent to successful mesh network operation. Such a task is very difficult to quantify.

Figure 8.3 summarises the relative time scales for the various aspects of innovation.

Reference

1. Lungaro P., *Coverage and Capacity in Hybrid Multi-Hop Ad Hoc Cellular*, Master of Science Thesis, TRITA-S3-RST-0315, Radio Dept. KTH, Stockholm, 2003.

9 Successful mesh implementations

Here we examine two of the most common real-world mesh deployments: firstly wireless cities and secondly community Internet. We show how their reasons for success align with the content presented in earlier chapters in this book. Interestingly, wireless city deployments are targeted at urban areas which already have wired Internet connectivity but where the addition of mobility is valued, whilst in contrast community Internet is targeted at those places where the wired Internet is sparse and connectivity can be added most easily by using wireless to serve fixed locations.

Thirdly, we also show a rising application of mesh networking – vehicular ad hoc networks (VANETs). These systems are targeted at improving road safety and have had spectrum allocated in many countries, and enjoyed success in industrial trials. We expect VANETs to experience particularly strong future growth.

9.1 Wireless cities

Several wireless cities are now up and running which provide easy Internet access on the move. In the UK, London and Bristol were early examples, whilst in the USA there is New York, Portland, OR and a rapidly growing number of others. The aim in each case is to enable easy mobile connection to the Internet. This can serve the general public, business users and the city authorities, who may use it for operational purposes, including for public services such as law enforcement.

The wireless nodes are deployed at street level and each includes a normal WiFi access point, so that users may connect with their existing WiFi enabled devices, such as laptops and a growing number of converged cellular-WiFi mobile handsets and PDAs. The challenge for the network operators has been to provide backhaul for all these access

points. The solution for this was to arrange that the access points are meshed together wirelessly. This creates a network side or backhaul mesh. This is an appropriate solution, given the environment in which the backhaul must take place – the urban area is one of intersecting streets in between tall buildings, in other words an environment with many obstacles. We showed this exact example application in Sections 2.1.4 and 8.2.1, where we generically termed it micro base station backhaul, as it may apply equally well to cellular and WLAN deployments. The principle of using the mesh to improve coverage in cluttered situations, such as the urban area, was shown in Section 2.2. The wireless city is a perfect example of this, employing mesh in a lower power, intelligent, multi-hopping approach in preference to the higher power cellular approach.

The mesh approach is suitable for another reason, which is cost of installation. In an urban centre, the prospect of laying cables is quite daunting. There are already many cables laid in ducts and there are very many other services also buried in an underground maze. This leads to the cost of digging up city streets being very high, in fact the cost per mile can easily run into five figures. In summary, the use of mesh will remove the need to dig up streets and the low power radio nodes used are adequate if deployed at street corners.

An example of a mesh node being installed in Toronto in 2007 is shown in Figure 9.1 [1]. By installing a number of nodes across a city, an operator can ensure coverage in a chosen area, as we show in Figure 9.2, for a rather simplified city.

If the continued success of the wireless city can be called into question, it is over the business case [1]. As an example of the price charged to users, we may take the example of Minneapolis, MN, which appears to be fairly typical. Here, $20 per month will buy a usage level of 3 Mbps, whilst $29 is needed for 6 Mbps. A city visitor may pay $5.95 for a single day's access [1, 2]. But in all wireless cities, there is competition to deliver Internet access, which comes from cellular at broadly similar prices, or from hotspots in cafes, which is often free. Secondly, workers in wireless

Figure 9.1 Mesh node installation in a wireless city, used by kind permission of Toronto Hydro Telecom, Inc.

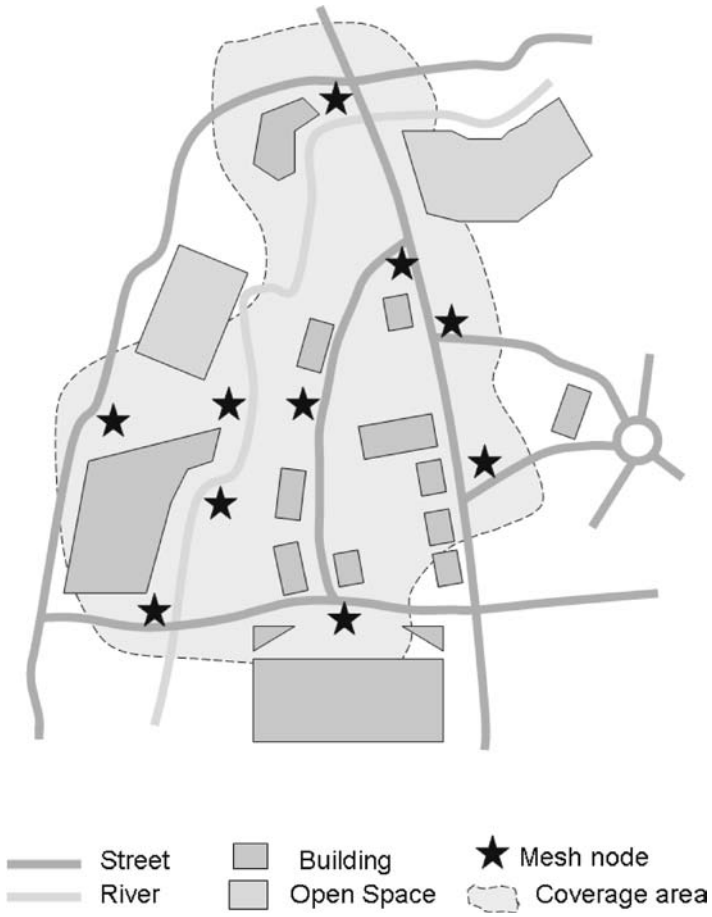


Figure 9.2 Map of a wireless city.

cities often have access to the Internet at their desks, so what is the incentive to pay for a wireless connection? Minneapolis thinks it may have the answer, which lies in using the network for a combination of public and private uses, but where the emphasis lies in public use [2]. In other words a city may use and pay for wireless access for operational and public safety reasons. This gives the operator a guaranteed income irrespective of any less certain private use of the network.

We expect this public–private partnership model will grow, meaning that wireless cities will not fall into the last of our pitfalls of Chapter 7, the money making model. We also expect that city WiFi will provide strong competition to cellular alternatives in the area of speed. With a greater density of higher capacity nodes, dense WiFi may better accommodate the traffic of the future, which is likely to be more bandwidth hungry. At the time of writing the number of wireless cities continues to increase and the trend is to move towards a public–private partnership to ensure the money making model.

In summary, at the beginning of this book, it was suggested that all successful mesh implementations should use either or both mesh properties of (a) hopping around obstacles and (b) requiring no infrastructure. It is easy to see that wireless cities employ both properties to solve their problem.

9.2 Community Internet

In community Internet, unlike wireless cities, it is not a greater convenience in attaching to the Internet which is sought – it is the need to connect at all. As we showed in Sections 2.1.2 and 8.1.3, it may not be cost effective for a remote community to connect to the Internet as the density of homes may be too low for an operator to amortise the cost of the necessary backhaul and local distribution. This often occurs in two broad cases – specifically in rural communities within the well developed areas of the world and generally within the less well developed areas of the world. Interest in the latter appears highest.

The community Internet solution proposes that mesh equipment can be used to share a single Internet feed within community networks. Such systems are appearing because traditional cost and service inhibitors become controlled by a ‘community’. The potential capacity of the radio network is so much greater than the limit at the Internet gateway access point, that it is not an issue for network design, so users are free to install as they wish without too much pressure for optimisation. The cost of equipment is also low as it can be based on a modified WiFi installation.

Figure 9.3 Community Internet node installation, used with permission of Locust World.

Figure 9.3 shows an installation using Locust World equipment in Bolivia. In this case the WiFi hardware is installed in an external box directly with a high gain antenna, in order to achieve the best radio performance. The installation does not need to be to professional standards, which is one of the great attractions. On the other hand, we suspect the relative difficulty of the installation becomes tolerable because there is no viable alternative and hence this situation would not translate to well developed countries.

Figure 9.4 shows a map view of several community wireless network clouds which each have one feed to the Internet. The cost of distribution is low and the cost of backhaul is shared by each community. The community is free to set the coverage, quality and reliability of the wireless distribution network at a level which suits them, and it is here that the majority of savings can be made.

Lack of infrastructure and ease of set-up are the prime mesh properties taken advantage of by community Internet schemes. Deployment opportunities are highest where no suitable alternative exists or is affordable.

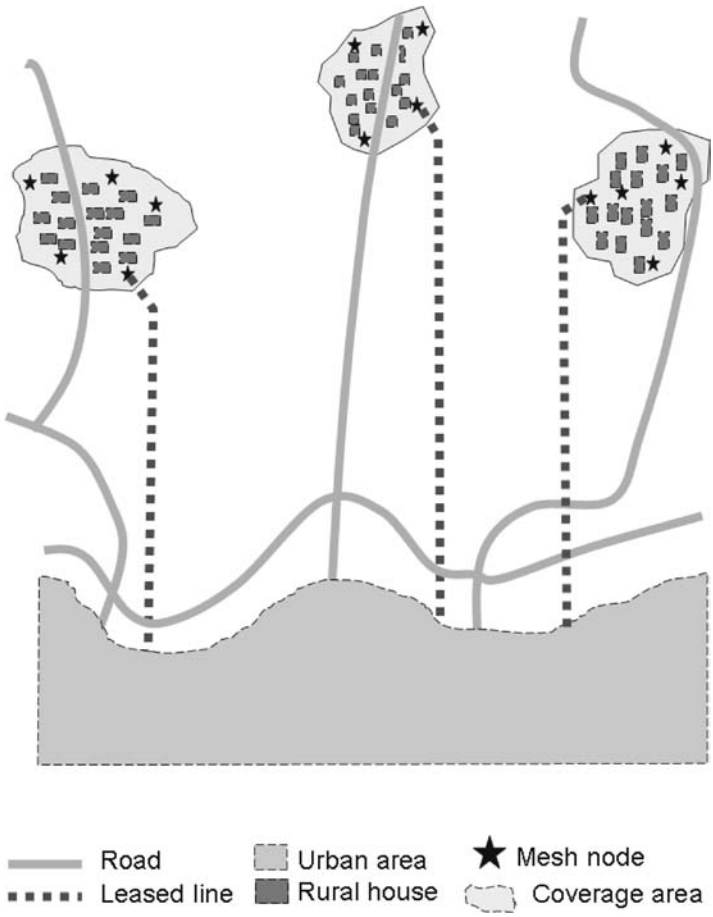


Figure 9.4 Map of a community Internet.

9.3 Vehicular ad hoc network (VANET) applications

The desire to introduce intelligent transport systems (ITS) is high in many countries today, including the USA, Japan and Europe. The initial driver for effort in this area was safety, which has now been joined by the economy (due to the economic cost of congestion) and the environment. Current work in the USA includes the DoT's 10 year vision (to 2012), which aims for a 15% reduction in road fatalities [3].

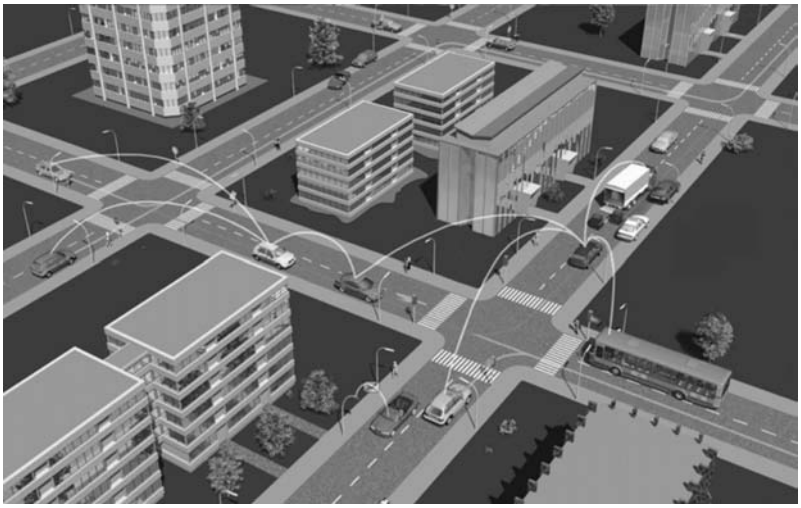


Figure 9.5 Artist's impression of vehicle safety communications using short range wireless, used by kind permission of the Car to Car Consortium.

Similarly, the EU's eSafety initiative aims to reduce road deaths by a half in the same time frame [4]. Increased safety will also lead to public spending benefits, for example in the emergency services and at hospitals.

In the future we expect there will be provision to warn individual vehicle drivers of any potential adverse interaction of their real-time driving behaviour with any upcoming road infrastructure feature or danger. Examples of this would be warnings of red traffic lights, or a dangerous bend which is being approached too fast. Such communication schemes using short range wireless might look as shown in Figure 9.5, where the meshing and multi-hopping aspect is clearly apparent.

We expect the main service offerings will be assistance for drivers in situations which are already known, by prior experience, to be prone to accident and injury. In the USA, the Vehicle Safety Consortium of car manufacturers found that the four applications which would have most impact, most quickly, on improving safety statistics were as follows.

- Traffic signal violation warnings, where drivers are warned that they have failed or are probably about to fail to obey a traffic signal. An example is a driver approaching a red light at a speed which is faster than that ideally required to stop by the stop line position.
- Pre-crash sensing, where if a vehicle senses a crash is unavoidable, then internal vehicle safety features can be activated and the driver warned, in a timely manner. This might include pre-tensioning of seatbelts.
- Emergency electronic brake lights (EBL), where vehicles behind one which is under hard braking automatically receive a warning.
- Curve speed warning, where the driver is simply made aware of the safe speed for the curve about to be entered. This may be very much less than the general speed limit on the road, and may be weather dependent.

There are very many other wireless vehicle safety examples currently under discussion and testing worldwide.

Whilst some safety applications impose critical demands on latency and other quality of service parameters, other safety related applications have less onerous requirements. Because of this it may be appropriate to use separate spectrum for these two classes of safety applications. For example, whilst electronic brake lights are highly latency critical, simply passing a warning message of an accident ten minutes ahead is relatively less critical of latency. This is exactly the view which is being taken within the EU today.

The preventative safety applications discussed above would be expected to operate in the emerging EU ITS band at 5.9 GHz, which is harmonised with similar bands in Japan and the USA. These applications are not safety-of-life critical,[1] but they are safety related and do require low latency and good reliability of delivery. The band allows for car-to-car (C2C) and roadside-to-vehicle (R2V) communications.

It is generally accepted that cellular services, whilst adequate for post crash information sharing, do not have the very low latencies or

[1] This phrase has a very special meaning which would not permit the presently proposed way of allocating the spectrum at 5.9 GHz to proceed. Specifically, the sharing of spectrum with other services would be a problem.

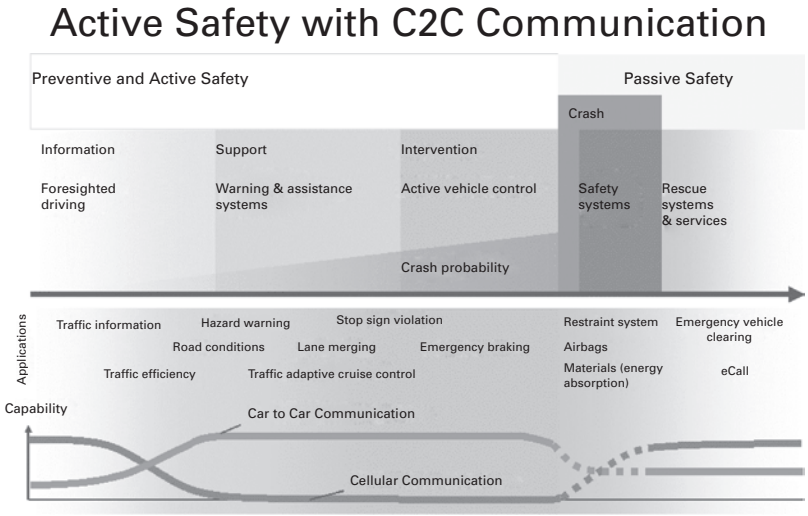


Figure 9.6 Contrasting cellular with C2C communications for safety applications, used by kind permission of the Car to Car Consortium.

availability needed for demanding safety related applications. We could take electronic brake lights as an example. If a cellular data channel was already open, then perhaps the latency would be adequate, but the latency involved in setting up the call (which itself may or may not succeed) most certainly would be inadequate for the application. Figure 9.6 clarifies this point of view and additionally shows that cellular would be suitable for other services which are less latency critical such as traveller information and road network management. In summary we can see that C2C communications is needed where latency is critical in achieving safety.

The attribute of mesh which is most of use in this application is that of being easy to set up and tear down, and of course that no infrastructure is needed. The routing of links in a mesh, which has no need of centralised co-ordination, is also a great enabler for this application. Low latency is not normally associated with mesh operation since end-to-end latency can be poor over a path of many hops, but here it is the low latency in connection set-up, i.e. routing, which is of prime importance, often over a single hop.

We expect that many mesh networking companies will come to see the synergy with applications within the road safety area, although at the time of writing we know of only two companies who have begun to connect with this opportunity.

9.4 Summary

We have shown three real-world applications of mesh, whose success may be correlated with the points made in the preceding chapters of this book. One further application area which we think will enjoy success is sensor networking. This is based on many mesh techniques, but the application makes it dissimilar enough from conventional mesh networking that wireless sensor networks deserve their own chapter.

References

1. Conti J.P., Metro Wi-Fi, *IET Engineering and Technology Magazine*, March 2008.
2. *Minneapolis Nets Industry*, BelAir Networks Case Study, available at www.belair.com, March 2008.
3. *National ITS Program Plan: A Ten-Year Vision*, US DoT, January 2002.
4. *Final Report of the eSafety Working Group on Road Safety*, European Commission, November 2002.

10 Wireless sensor networks (WSNs) as mesh networks

This book has so far focused on meshes for telecommunications, however another use of multi-hop networking is the wireless sensor network. In fact it is potentially beginning to look like WSNs might outstrip telecommunications as a use for multi-hop and mesh technology. We have kept this chapter separate as WSNs have some unique properties, but we find that many aspects of mesh discussed earlier in the book apply to WSNs in much the same ways. In terms of applications, at the time of writing, smart buildings (advanced control of lighting and HVAC, heating ventilation and air-conditioning) and logistics look like the top two likely WSN applications, in terms of earliest uptake.

Let us begin with an introduction to wireless sensor networks. We take quite a broad overview before concentrating on the networking aspects of WSNs.

The role of a wireless sensor network is essentially that of a monitor. Broadly speaking, what is being monitored can usually be placed within one of three groups:

- area monitoring – i.e. monitoring *somewhere*; examples include the environment or area alarms (intrusion etc.);
- entity monitoring – i.e. monitoring *something*; examples include a civil structure (bridge, building etc.) or a human body;
- area–entity interaction monitoring – i.e. monitoring *something*, *somewhere, in context*; examples include vehicles on the road, asset tracking or the flow of a manufacturing process.

As to why a sensor *network* is important, it is most simply understood by realising that, often, individual sensors themselves are limited in their ability to monitor a given situation. Specifically, a single sensor is not likely to embody sufficient scope to sense a complete phenomenon alone, nor is the system reliability likely to be very good, since the sensor presents a single point of failure. The method of communicating the resulting information to a base unit may also bring a system challenge.

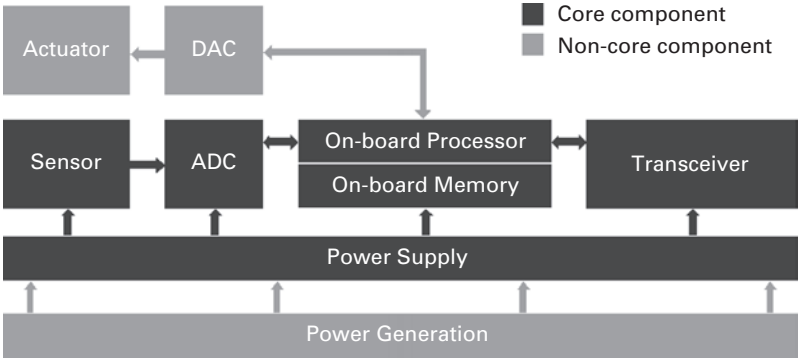


Figure 10.1 Block diagram of a basic WSN mote.

The strength of a sensor network comes from the fact that even though the individual nodes are quite limited, the whole array becomes very powerful when networked. Thus sensor networks are likely to be large in scale, in the sense that they have many nodes and they are likely to be self-configuring, to bring reliability. Also, the nodes themselves are likely to be cheap, such that many nodes may be economically deployed.

In this chapter, we will introduce

- sensing technologies and interfacing,
- power sources and harvesting,
- differences between mesh, WSN and RFID,
- key WSN standards efforts, and
- the question of structure in WSNs.

We will conclude by drawing parallels between mesh and sensor networks, but first let us look at what makes up a wireless sensor node.

10.1 Introduction

A practical wireless sensor node must consist of at least the following, Figure 10.1:

- a sensor, e.g. a microelectromechanical system (MEMS) accelerometer, or a light sensor,
- a signal converter, e.g. an analogue to digital converter,

- a processor and memory, of minimal power requirements,
- a wireless network interface, e.g. radio or optical, and
- a power supply, or a method of harvesting power, e.g. from vibration or light.

The term 'mote' is commonly used to describe a node, either with or without its relevant sensors. It is interesting to note that the dictionary definition of a mote is a 'speck of dust', or similar, and this describes the role of an individual sensor node very well. Each is relatively small, like a speck of dust, but there are lots of specks of dust in the network.

A central theme of motes is that they need to be low power or self-powered, so that they may last for many years in situ, without suffering the cost and inconvenience of maintenance. This theme of low power drives much of our commentary in this chapter.

Let us begin by considering the transducer, which is the very 'front-end' of the sensor node, the part which monitors a property of the phenomenon to be sensed and converts it into some electronic form.

10.2 WSN sensors

One way sensors may be grouped is by their main operating principle as follows:

- physical sensors,
- electromagnetic, optical and acoustic sensors,
- thermal sensors, and
- chemical and biological sensors.

We will not cover all the sensors here, but we will point out that there have been many advances in sensors in recent years

One example is the recent breakthrough in integrating sensors onto silicon, which is the watershed enabler for integrated devices. Examples are MEMS accelerometers and gyroscopes. These fall under the descriptor of integrated MEMS, or iMEMS and are worth a special mention.

Having a MEMS device close to its associated electronics brings operational benefits and reduces the cost, particularly if the sensor and electronics are on the same silicon process. An example device which

includes piezo and capacitance principles and which has integrated electronics is the nano-gyro, or iMEMS gyro.

The importance of this device is not that it is a new concept, since gyros have been around for very many years. Its importance lies in the fact that this device is now affordable to enable many more applications. For example, rather than rely on GPS alone to help navigate vehicles, an iMEMS gyro can provide inertial navigation. This has the potential to be cheaper than GPS and to work where GPS cannot, where the view of GPS satellites is blocked. It also could not be jammed in the way that GPS can.

10.3 WSN power sources

Strict power consumption requirements arise from the need for the sensor node to be independent and capable of running unattended for a long time, perhaps for several years. Wireless sensor nodes lose their attraction if costly maintenance visits have to be made, for example to replace batteries. Variables in the design include

- the choice of battery or power harvesting schemes, and
- low power electronic design approaches.

Here, we look at power and power harvesting options and we address power saving in Section 10.10.3.

Power sources are usually split into primary and secondary cells, where the difference is that primary cells cannot be recharged, by design, whereas secondary cells have a need for regular charging. Key parameters of primary and secondary cells include capacity, temperature range, current drain level and self-discharge characteristics. Alternatively, there is also a storage method based on capacitors, called supercapacitors. Whilst these are strictly not power cells, they are useful energy stores. A potential future energy source is fuel cells, if they could be made small and safe enough for more applications.

10.3.1 Energy scavenging/harvesting

Where secondary cells are used, the charging source may be scavenged or harvested from the cell's working environment. The most well-known

example of this is probably the harvesting of solar energy to charge a battery. However, many other energy modes are available for harvesting and scavenging, such as vibration, wind power and thermal energy. For flow and mechanical driven harvesting a battery may not be required where the harvested movement is constant, such as in a pipeline.

Taking the best known as an example, solar systems need some degree of installation to ensure optimum orientation, particularly at higher latitudes. Currently the best silicon solar cells offer efficiencies in the mid 20% range, with low cost commercial cells being significantly lower than this.

Most recently, success has been found with vibration harvesting, including working with the 100 Hz 'hum' found in industrial applications or flow from an oil pipeline.

10.4 Wireless sensor technologies and applications

In this section we look at what else remains to be provided at the physical level in order that we might proceed to use sensors easily in applications, i.e. what are the outstanding technical enablers. This introduces transducer electronic data sheets (TEDS).

10.4.1 Sensor interfacing and calibration

There is a large gap between, on the one hand, producing a sensing element and, on the other hand, successfully deploying that element in a network. In fact, the two major problems are

- firstly the lack of a standard sensor interface, and
- secondly that networking knowledge lies outside the core competencies of of the majority of sensor manufacturers.

The knowledge that this was the case and that it was holding up the industry was what drove the creation of the IEEE 1451 standard for smart sensors.

IEEE 1451 specifies electronic data sheets and wired/wireless communications for sensor networks. Electronic data sheets are contained within the smart sensors and define the sensor's networking interface, its sensor

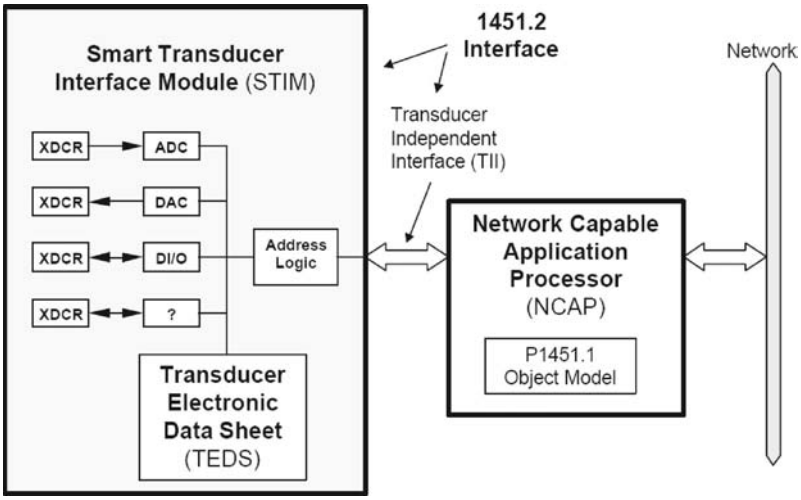


Figure 10.2 IEEE 1451 functional diagram.

type and its calibration data. This is shown in Figure 10.2 where sensors (transducers or XDCRs) with analogue or digital inputs and outputs (I/O) – and even a sensor whose output format is not known to the user – are linked to their specific TEDS. In this way each sensor has a known interface and stores its own specifications, meaning that the user need no longer know the lowest level details of the sensor and network implementation. It is easy to see why sensors conforming to IEEE 1451 are referred to as smart sensors. The interface to the sensor module (STIM) is a digital interface defined by IEEE 1451. Later sections of IEEE 1451 (1451.3/4/5) provide for the interface to the wider network to employ various forms of wired or wireless link.

The wireless parts of IEEE 1451 (1451.5) simply reference the use of existing wireless networks, presently including 802.11, 802.15.4/ZigBee and Bluetooth. One of the strengths of IEEE 1451 is that the sensor network may be grown by enlarging the host network (Ethernet, 801.22 etc.) rather than having to extend bespoke sensor links.

In summary, the usefulness of IEEE 1451 is that it creates defined sensor interfaces where presently none exist, it includes sensor data in an embedded manner and it connects easily to existing wired and wireless

Table 10.1 *Differentiating RFID (not active RFID), wireless mesh and wireless sensor networks – today*

RFID	Mesh network node	Wireless sensor node
Insignificant price for simplest RFID	Expensive	Low cost
No power source necessary (option)	Good battery, rechargeable	Restricted energy resources
Not a network, needs a reader	Multi-functional	One-trick pony
None or limited processing	Can run 'big' protocols like TCP/IP	Can run only lightweight processes
Powered RFID has medium range	Radio range can be large	Radio range is small
Embeddable	Handheld or larger	Tiny
None or nomadic	Full mobility	None or nomadic

local area networks. An IEEE 1451 wireless sensor network clearly possesses infrastructure, a point which we shall pick up later.

10.5 Differentiating RFID, mesh and sensor networks

It is useful to be clear on the differences between WSNs and mesh networks. RFID is also of interest mainly for its anticipated future progress, which may well take it into WSN territory.

Let us begin by comparing, as directly as possible, the various attributes of RFID,[1] mesh and WSN network nodes as they exist today. This is shown in Table 10.1. Examples of each type of network are as follows.

10.5.1 RFID

RFID is used for asset tracking and car immobilisers or remote keyless entry. RFID really stands alone in the table in that it is intended for the

[1] We have chosen to differentiate traditional RFID (today's generation, simple, polled) from active RFID (next generation, capable, always on).

lowest cost, even literally throw-away applications and is not operated as a network. Having said that, in fact there are four possible classes of RFID, with the lower classes being overwhelmingly more common today, and so forming the basis of comparison in the table. The four classes are as follows.

- Passive RFID. This has no built-in power source. It relies on back-scatter and only one device can be read at once in the field of a scanner. However, it is often said that many of these devices can be read 'simultaneously' by a single scanner – in fact many devices are read sequentially using a back-off algorithm. This is the cheapest tag, used for item ID and theft prevention.
- Semi-passive RFID. This has a built-in power source for use in processing and by other peripherals, perhaps sensors, but the power source is not used for the transceiver and so does not boost range. It is also a backscatter mode tag. This type of tag is used in road tolling schemes, for example the new bridge over the River Thames at Dartford.
- Semi-active RFID. This has a power source which is available to the transceiver as well as everything else. However, the node is expected to sleep for much of the time, i.e. it has a low duty cycle. The tag is capable of initiating communication, making it quite different to the passive tags.
- Active RFID. This is powered and the transceiver can be always on, which can make it somewhat similar to a WSN node. To avoid doubt note that we have excluded active RFID from Table 10.1 on the grounds of maintaining simplicity for this comparison. Active RFID is really the next generation of RFID and much more powerful than the RFID most of industry is familiar with today. Active RFID can be very capable and could become indistinguishable from a WSN in an application. We thus expect active RFID and WSN applications to converge.

10.5.2 Mesh networks

We have already shown mesh applications which today include municipal wireless roll-outs and, in the near future, we expect vehicle ad hoc networks (VANETs) to be a large application.

Full mobility is clearly a huge differentiator of mesh networks from RFID or WSNs. Hand in hand with this goes a larger radio range. This larger range, together with the levels of performance expected by the applications, imply a good battery will be required. And if needed this battery can be recharged on a daily basis, since the application environments support this. Range and power source are thus very different for mesh networks, when compared to both WSN and RFID.

Earlier in the book, we noted that additional network infrastructure could be used to improve scalability and quality of service for mesh networks. We will show that this conclusion carries over to WSNs.

10.5.3 Wireless sensor networks

Logistics, environmental and industrial monitoring and smart buildings are the most commonly cited applications for WSNs.

Big differentiators of WSN from mesh are the limited data capability and the associated power savings. WSNs may require only a few bits per second per day, on average. WSNs do not carry real-time streaming services, nor are they used where latency is critical. In other words, WSNs are not video or even voice capable, although people have tried VoIP over 802.15.4 with limited success. The biggest source of power saving of WSNs is their low duty cycle, less than 1%.

We will look more closely at requirements for WSNs in Section 10.8, but we next compare WSNs to mesh networks to establish the major differences and similarities at the highest level.

10.5.4 Comparisons between mesh and sensor networks

Putting RFID completely aside, let us now list the differences and similarities between mesh and sensor networks. These are as follows.

- Having cheaper WSN nodes probably means having less reliable nodes than mesh networks.
- WSNs contain more nodes at a higher density, and the radio range is lower.

- WSN traffic is lower complexity, certainly not real time, and bit rates may be only several bits per day.
- WSN traffic is very application specific, and node design may follow this, making nodes inflexible.
- WSNs can also include actuators as well as sensors.

An additional useful insight is that whilst the routing challenge for wireless mesh networks is coping with mobility, the WSN routing challenge is coping with limited energy.

There are also similarities between mesh and sensor networks.

- Both are self organising networks.
- Both, in principle, do not need infrastructure.
- Both, in practice, benefit from infrastructure. The mesh benefits from access points to improve QoS as we have shown in earlier chapters, and the WSN from gateways/routers to improve power consumption for edge nodes, as we shall show in this chapter.
- Both, in practice, benefit from clustering to cheat scalability issues.
- Both have security challenges.
- Both have privacy challenges.
- Both introduce a dependence of the network on node behaviour.

It is useful to remember that both are co-operative networking approaches, by which we mean that medium access control is decentralised and there is an element of fair contention for resources involved for each node, as discussed in Chapter 5. This means that such systems ought to co-exist as well as possible with each other, but that other systems operating centralised medium access may dominate. In other words, mesh and sensor networks typically run polite protocols and may suffer when co-located with systems operating impolite protocols.

10.6 Differentiating 802.15.x, ZigBee and 6LoWPAN

There are a great many acronyms for ad hoc and de facto standards in use in the field of wireless sensor networks.[2] At the top level, we can clarify

[2] We remind the reader of the list of abbreviations.

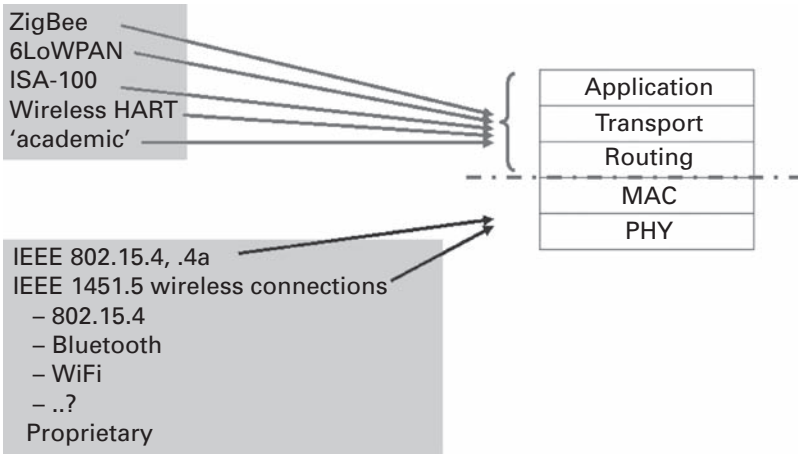


Figure 10.3 Standards and where they fit.

the situation somewhat by reference to Figure 10.3, which splits the standards into those which apply to the physical layer/MAC and those which apply to the networking/application layers. This is particularly useful in showing the difference between ZigBee and 802.15.4, which are often used together in an application. The figure makes clear that other pairings are possible.

Note that Figure 10.3 presents a rather simplified stack, but is sufficient to show the simple layer-level split we have outlined. The term 'academic' refers to university research networks which often use 802.15.4 as a base for investigations of the higher levels, such as routing algorithms.

Let us now examine 802.15.4/ZigBee and 6LoWPAN.

10.6.1 IEEE 802.15.4 and ZigBee

In the future, many WSN implementations are expected to use 802.15.4 for the radio functionality. Very many of these are expected to run ZigBee on top of 802.15.4, whilst others will run application specific networking, including the two industrial sensor standards Wireless HART and ISA-100. Those which are not 802.15.4 based may use a variety of proprietary

radio protocols. As an example of the majority interest in 802.15.4, 90% of WSNs in smart buildings are expected to be 802.15.4 based by 2011.[3]

The various radio standards can be compared with the aid of Figure 10.4. As is typical, ZigBee has been added to 802.15.4, WiMAX to 802.16 and Bluetooth to 802.15.1. Given that we have already shown the layer-level differences in Figure 10.3, it is nonetheless useful to add such labels, as otherwise the 802 numbering can become confusing.

The IEEE 802 standards referred to in the figure include the following.

802.11, commonly used in both business and consumer applications for wireless LANs with a range of up to 100 m and raw data rates around 11–54 Mbps in the 2.4 GHz band. A variant also exists for the 5 GHz band.[4] 802.11 is a WLAN, a wireless local area network.

802.16, recently added to the IMT-2000 family as the sixth terrestrial radio interface alongside 3G etc. 802.16 is a WMAN, a wireless metropolitan area network.

802.20, a standard plagued by delay, it is intended to offer high rates to mobile users. Like 802.16, 802.20 is a WMAN.

802.22, intended to serve a regional area to provide broadband wireless access as a WWAN, a wireless wide area network. This is only broadband in the same sense as ADSL, since contention levels are similarly high.

Finally, the figure includes 802.15, which is a series of WPAN, wireless personal area network, standards. There are three interesting components of 802.15. These are 802.15.1 which is the radio layer used by Bluetooth, 802.15.3 which is for high rate PANs (e.g. over 20 Mbps) and 802.15.4 which has the following features:[5]

- data rates of 250 kbps (2.4 GHz), 40 kbps (915 MHz), and 20 kbps (868 MHz),
- 16 channels in the 2.4 GHz ISM band, 10 channels in the 915 MHz band and one channel in the 868 MHz band,

[3] Source: OnWorld.

[4] This is 802.11a/b for 2.4 GHz and 802.11a for 5 GHz. However, 802.11n (2.4 GHz, up to 200 Mbps) has yet to be finalised.

[5] Note that number 2 in this sequence, 802.15.2, was a co-existence working group looking at operating 802.11 with 802.15.1 at 2.4 GHz.

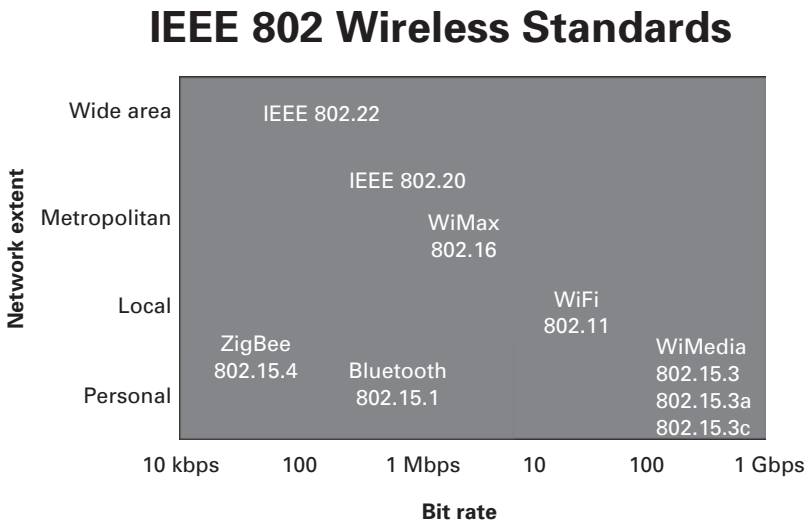


Figure 10.4 Radio (PHY/MAC) comparison, used with permission from Gleichmann.

- two addressing modes, 16-bit short and 64-bit IEEE addressing,
- support for critical latency devices such as joysticks (the beaconing option),
- CSMA/CA channel access (the non-beaconing option),
- automatic network establishment by a dedicated co-ordinator,
- fully hand-shaking protocol for transfer reliability, and
- power management to ensure low power consumption.

Of all the 802 wireless standards shown in Figure 10.4, only 802.15.4 is targeted at very low power and very long battery life. It is specifically targeted at sensor networks, interactive toys, smart badges, remote controls and home automation, operating in international licence exempt device bands. We also look at 802.15.4 more closely when we examine structured WSNs in Section 10.10.2.

It is interesting to note that 802.15.4a also exists as a task group whose aim is to provide high precision location services using UWB. Apart from location sensing this would also be good for WBANs, wireless body area networks, which typically require a single gateway radio, with flexible data rates to cover the widely different BAN applications from cardiac pulse rate to streamed entertainment.

ZigBee

In the words of the ZigBee marketing slogan, ZigBee provides 'Wireless Control that Simply Works ™'. Its aim is, like Bluetooth, to identify the common applications and to make them particularly easy to implement and to ensure interoperability between compliant devices. Over 200 companies have joined together to define the upper layers shown in Figure 10.3 and to add security. ZigBee provides various profiles in the following groups:

- a general group including simple on/off and RSSI applications,
- an HVAC group (heating, ventilation and air conditioning),
- a lighting group,
- a security group, and
- a measurement and sensors group.

The security aspect of ZigBee relies on it being a structured network (see Section 10.8), hence a trust centre will always be available. Security is built upon access control lists and data encryption at various layers.

10.6.2 6LoWPAN

If, like ZigBee, 6LoWPAN had a slogan, it would probably be 'IP Everywhere'. Historically, IP was thought to be too heavyweight for low power radios like 802.15.4 due to the amount of management traffic as well as typical payload sizes, which are much larger in TCP/IP than the 802.15.4 MAC would expect/accept.

Hence the desire to run IP over sensor networks was neglected for a while since it was thought impractical. But, by using 6LoWPAN, this is now being achieved, and without the large penalty originally expected. There is no doubt that a penalty is paid, however, but it need not be prohibitive for certain applications, where a suitable performance trade-off can be made.

It is the IETF, who produce the major Internet standards, who are also aiming to produce the 6LoWPAN standard.[6] It is intended for IP on low

[6] 6LoWPAN is an IETF Working Group whose output is intended for the IETF Standards Track.

power devices, including sensor nodes. 6LoWPAN comes later to the market, but competes directly with the ZigBee approach by offering a more open standard and direct access to nodes via IP, rather than through a protocol translating gateway. In other words, a user anywhere on the Internet may address any 6LoWPAN sensor node as freely as any other Internet device. Of course, not all applications require this, but where they do it is an appealing prospect for the end user. Note, however, that the '6' in '6LoWPAN' designates that 6LoWPAN is based on the use of the IPv6 packet header format and IPv6 addresses, rather than IPv4. Considering differences in the size of the maximum transmission unit (MTU) between IPv6 (1280 bytes) and 802.15.4 (as low as 81 bytes when using link-layer security), the use of header compression to reduce overhead in the 802.15.4 frame, and possible requirement to interwork with IPv4 applications, means that there will still be requirements for gateways between the 6LoWPAN/802.15.4 realm and the Internet to more than just bridge packets from the 802.15.4 radio realm. In other words, some basic network layer capability will be required in the gateways, rather than being simply a MAC layer bridge.

For balance, it should be pointed out that for really low resource sensors, 6LoWPAN still remains too heavyweight in memory footprint and computational power (e.g. as needed for header compression). For larger wireless sensors, however, it may be a good match, to allow existing applications or control protocols for wired, legacy IP devices (alarms, lighting control) to be ported to WSNs. In other words the trade-off is the possible flexibility offered by running IP (even if it is IPv6) versus not having quite the lowest power consumption. The main industrial proponents of 6LoWPAN are currently start-up companies, who offer the so-called 'Internet enabled sensor'.

10.6.3 Summary

In summary, the compromise encountered when choosing sensor network protocols is one of flexibility versus power consumption. Bespoke implementations are the most efficient but least flexible – and probably the least attractive to users, since they are not standards driven. ZigBee sits in the

middle, whilst 6LoWPAN offers the greatest potential for seamless integration with the IP world, but pays somewhat in terms of power consumption and the ongoing need for gateway devices, albeit relatively simple ones. No one approach is a clear technical winner and it seems that the choice must be dictated by the needs of the application. In the market, however, it is not always the most elegant technical option which wins. Often the most reliable and standardised solution will succeed, as long as the technical solution is fit for purpose.

10.7 A suggested taxonomy of WSNs: structure and equality

As there is a great deal of information in the literature concerning WSNs, it is instructive to form a basic taxonomy in order to avoid confusion. We have found, as we did with mesh networks, that the presence of structure, such as a wired infrastructure or a node hierarchy, makes a great deal of difference to network performance. Network design and performance are both considerably different when the network has structure compared to when it does not.

On a related but distinct note, a similar case can be made for node equality. When all nodes are equal, the design of a network proceeds differently to when some nodes have unequal performance. This inequality can be for better or worse, for example, a node with powerful processing ability versus a node which can exist only on scavenged power. There are clearly advantages to having nodes with unequal capability, namely only such nodes need be specialised, leaving others to be less complex and thus less power hungry.

Therefore our basic taxonomy is based on the concepts of

- network structure and
- node equality.

We justify this approach by detailed examination next.

10.8 System architecture in sensor networks

In this section we will review WSN requirements from a system design point of view. We follow this by looking at how the Internet is classically

organised using the TCP/IP suite and draw out some comparisons. Taking these two aspects together allows us to examine the relative applicability and suitability first of unstructured WSNs and then of structured WSNs. This eventually leads also into a discussion of node equality, where we use 802.15.4 as a specific example of a structured network with different classes of nodes. The ideas of node equality and network structure are related, but the idea of node equality is broader than network structure. There are many other ways in which nodes could be made unequal, such as offering translating gateways, acting as security centres or simply being more powerful in computing terms, which often also implies access to a capable power supply.

10.8.1 WSN system requirements

We have already observed that traditionally WSNs are driven by the need to be low power. This means they can be battery powered and placed anywhere which is convenient for the application. Low power design can go only so far, however, and another technique is needed to make the system itself low power. This is to operate the system with a low duty cycle. As might be expected, nothing saves power quite as well as regularly turning the nodes off for long periods of time (look ahead to Figure 10.10 for confirmation). The downside of doing this is that latency must suffer an increase.

Latency affects the transmission of data, but it also affects network management, where node discovery[7] becomes more of a problem, since the nodes are not always available to be found. Ways around this include having special network devices which store and forward discovery information and indeed data when normal nodes are sleeping. However, then we have immediately created a new class of node, which goes against the principles of having an unstructured network where all nodes are equal. On the other hand, much research has been published on how to cope with WSN requirements whilst still having only one class of node, which may be scattered at will, with no planning whatsoever.

[7] When a new node joins, or more likely rejoins, a network, it has to be 'discovered' by the network.

In this section we look at unstructured networks first, followed by structured networks. Looking ahead to our conclusions, it is fair to say that we cannot see that many large-scale applications will really require an unstructured network, and that structured networks bring many advantages with them. One probable reason for such a high level of research into unstructured networks may be military funding,[8] where, like mesh, mobile wireless sensors are also of great importance. It is notable that industry and standards interest favour the structured network, where mobility is not of greatest concern, for example for building monitoring.

Finally we should say that security, including authentication issues, is much easier when a network has structure and one node may be regarded as a permanent trust centre. However, let us begin, as we have said, by reviewing how things are presently done in the IP world of the Internet.

10.8.2 Classic IP address-based routing and transport – review

We are going to go into quite some detail in this section, to show how Internet based routing and transport works in practice, so that we may contrast it with how several sensor network approaches work. Much of the way the Internet works is based on it having evolved over a period of time. This means that it does contain a few idiosyncrasies and imperfections, which we would probably avoid if we were to design it again with all the knowledge we now have. The point is that we are fortunate to have (almost) such a blank sheet of paper for sensor networks, which is an opportunity perhaps to think a little differently.

The classic Internet approach to addressing is to use a network layer address which is tied to a particular physical interface. For example, a 32-bit IP(v4) address, normally written in the form *a.b.c.d* could be assigned to an interface on a device. If we think about a home PC this means that the network card 'owns' the IP address, say address 192.168.0.2. If we have two network cards (laptops commonly do, where one is WiFi), then the PC has two IP addresses, perhaps adding

[8] Initial research into mesh networks also focused on absence of structure due to military funding of the research.

address 192.168.0.10. The IP address thus does not in fact identify the PC, but as we shall see later, many applications assume that it does. This is for historical reasons during the development of the Internet, which we need not delve into here. This leads to some entanglement of the layers of the protocol stack,[9] which can be unfortunate in certain circumstances.

Further entanglement of the supposedly independent layers of the protocol stack occurs because although the address is a network layer identifier it is tied to a physical interface. In other words the IP address has an element of identification, but also an element of topology or location, due to the hierarchical organisation within IP addresses. With our earlier example, the first part of the address 192.168.x.x is a particular network prefix, which usually implies something about topological location.[10] This has implications for routing, especially in the face of mobility, which we discuss later.

Looking deeper, the network layer provides only for node–node packet delivery. But a finer grained de-multiplexing of packets is required, since we expect many different types of packets to traverse the node–node link, and we wish to keep these packets separate so they may go to the correct destination within each node. This means another layer of addressing is used on top of the network layer. For example, in TCP, a 16-bit unsigned number is used to identify, within the scope of a given IP address, a given connection endpoint. This means a TCP endpoint consists of a pair of numbers: the IP address plus the 16-bit number, called the port number. Together the IP address and port number comprise a TCP socket (a connection endpoint), and two sockets (remote and local) define a TCP connection.

Note also that TCP provides other functions, such as reliability and ordered delivery, which may or may not be required for a given sensor network application. Finally, the application itself may have some information it wishes to use in identifying or routing the data at the application level, and so there is yet another level of naming/addressing at the application level. In many real cases, this last level of addressing will

[9] We showed a simple stack in Figure 10.3.

[10] Actually 192.168.x.x is a non-routable address, but it does show the principle of containing more than simply identification within the IP address.

often overload information[11] such as IP addresses and port numbers, not necessarily because it is technically beneficial to do so, but because the IP address bits and port number happen to be conveniently available for use. This leads to the idea of 'well-known' port numbers, such as port 80 for http and so on.

However, strictly speaking, there is actually no requirement for the application[12] to be aware of the network and/or transport layer address bits. An important observation is that this last feature can be exploited by sensor network protocols, as we discuss later.

Before we do, however, it is worth noting the limitations of the current Internet oriented addressing model. Our discussion uses IPv4 addresses for convenience, but most of the same arguments apply to IPv6. Let us collect together the observations we have just made on IP addresses in a shorter format.

1. An address of p bits is assigned to a physical interface (layer 1/2).
2. The address consists of a *network part* (also known as the *network prefix* or routing prefix) and forms the n most significant bits of the p address bits. The routing prefix provides an addressing aggregation opportunity to help with hierarchical routing.
3. The remaining $m = p - n$ bits form the *host part* of the address.
4. Only the n bits of the routing prefix are used in routing – the remaining m bits are only used when the packet reaches the edge network for which it is to be delivered to a node.
5. All p bits of the address are used for the transport layer state in each transport layer session, thus entangling identification with topological location.
6. The p bits of the address have a dual role:
   a. as a **locator**, providing a topological name that is used as information to aid in routing, and
   b. as an **identifier**, providing a unique routable name for each interface, but also as part of the state information in transport layer and sometimes application layer names.

[11] That is, use the same bits for an expanded purpose.
[12] Unless it is a network management or control application.

The location-identification entanglement problem relates to IP-like addressing in general. Let us begin to put this into the context of WSNs. Where a sensor network consists of a few tens of nodes, it may be possible, with a suitable routing algorithm, simply to use flat addressing and so none of the preceding discussion is an issue. However, where we move to thousands of nodes, and/or hierarchical structure and so hierarchical addressing in sensor networks, for example tiered networks, then the use of flat addresses is no longer viable and these issues will become significant for WSNs.

However, at the application layer for a sensor network, there may not be a need to know anything about the exact nodes from where the data arrives. Such abstraction of content from address is a key idea for some approaches to WSNs, which we discuss in the Section 10.9.

For example, imagine a sensor network that has been dynamically deployed to monitor earth tremors in a given geographical area. The sensors may record and report in application packets the timestamp, GPS location and seismic activity reading. In this case, the seismic *monitoring application* is not at all interested in the network address of the sensor node – all the information it needs is in the data described. However, the address would be useful for the routing protocol, and the address plus the GPS location would probably be of interest to a network management application. So, it may be that the control and management plane applications have different addressing needs to the user plane applications within the same network.

This leads us naturally into a discussion of unstructured wireless networks, i.e. those with no hierarchy.

10.9 Unstructured WSNs

In contrast to structured approaches, unstructured networks are homogeneous with respect to node type and thus have no physical hierarchy. Another way of saying this is that all nodes are equal, physically and architecturally.

In an unstructured WSN, the sensors have no mechanism for out-of-band communication or control – all communication is through their

single wireless communication interface. We can compare this directly to the familiar case of the pure mesh which we first introduced in Chapters 2 and 3. In fact we can re-use Figure 3.4 from Chapter 3, to illustrate the unstructured WSN.

Once deployed, unless there is a very carefully managed deployment with careful pre-configuration of the sensor nodes, the sensor nodes must perform all of the following tasks:

- locate/discover other sensor nodes in their network;
- discover a route back to the gateways or sinks, the points in the network at which the collected information must be presented, perhaps for onwards transmission beyond the sensor network;
- forward relevant data towards the gateways or sinks using the other sensor nodes as relays;
- maintain/update routes to the gateways or sinks in the case of node failure, nodes and/or gateway/sink mobility, and/or due to other policy requirements, for example network load sharing, conservation of power across sensor nodes through diverse routing, etc.

Such networks and related technical issues are often the focus of academic and military research. Typically, the approach taken is to use ad hoc networking, with the additional constraint of resource limitation. This additional limitation could take the form of any combination of limits on network capacity, CPU power, memory, battery life, etc. In this way the WSN challenge becomes greater than the ad hoc networking challenge.

Unstructured WSNs are often thought to have the attribute of lowest power operation, usually because they have been designed to be so application specific that all unnecessary functionality is simply not included. This may not always be true, however. If we contrast this with a structured WSN, whilst the gateway node(s) may be mains powered, the actual sensor nodes can be very low power, since they may not have the same level of responsibility for providing functions such as routing, network time synchronisation, localisation and data filtering. The main disadvantage with structured networks is, of course, that mobility and the ad hoc aspect are potentially lost. In other words, low power alone is not necessarily a driver for an unstructured approach, and it may be that other

factors, such as the ability to be mobile, are important enough that the overall system design is adjusted in order that other constraints are given a lower priority.

On the other hand, it transpires that many potential applications of unstructured WSNs are not in fact mobile, but the convenience of an unstructured approach – not having to deploy and maintain a 'backbone' for the sensor network nodes to gateways/sinks – is simply preferable. Clearly the balance of costs and benefits will vary from application to application.

In either case, we have to deal with basic functions of communication such as addressing, routing, discovery, data transfer and route maintenance, including robustness to failure of nodes as well as change of topology through mobility.

We next look at how routing may be organised, specifically for the case of the unstructured WSN, via three methods: data-centric routing, geographic routing and other methods including energy aware routing. All are quite different to the classical IP addressing which we reviewed earlier.

10.9.1 WSN approaches – data-centric routing

Following on from the argument presented at the end of Section 10.8.2, a specific application based on a sensor network may be more concerned with the nature of the data from the sensor field rather than the addressing or routing information. This is why there has been a paradigm proposed for routing in sensor networks, which takes a data-centric approach to routing and data dissemination across the sensor nodes, rather than an address based approach. One of the most cited works in this area describes directed diffusion [1], and this gives a very good description of the principle of a data-centric approach.

The basic principle arises from considering the motivation of the user of a sensor network, who is interested in data and so would like answers to questions such as, 'How many events of type A occurred in area X?' Such a query would be sent to the sensor network and we say that sensors have been *tasked* with collecting data to answer the query. The nodes then

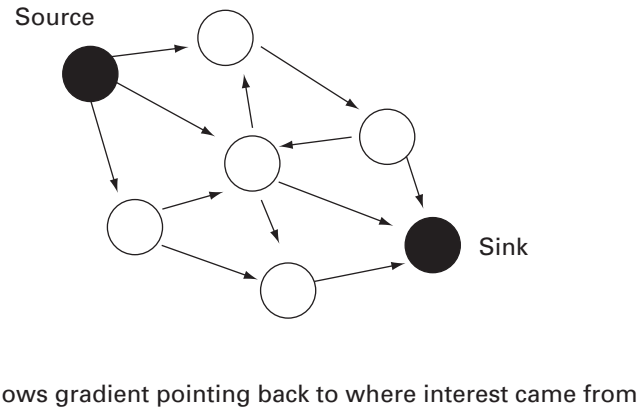


Figure 10.5 Formation of gradients from interests.

co-operate to 'answer the question' and report the result to the user. In order to enable this in a scalable, robust and energy efficient manner, the paper proposes the use of routing using attribute-value pairs to name data that are generated by sensor nodes. Note that the data are named and not the nodes themselves. By advertising, or *diffusing* (*sending interests* for), the named data to its neighbours, the data collected by nodes are drawn towards the node that generated the name. Intermediate nodes can use the name to initiate caching, perform data aggregation, forward historic/stored results, or forward new results matching that name towards the source of the name.

In summary, the general principle is that the knowledge of the data required by sensors, as advertised in their interests, enables their neighbours to forward data appropriately. The choice of the sink (the node that generated the interest) is arbitrary. The sink periodically broadcasts its interest(s) and its neighbours maintain an interests cache. If the interest is not already present in the cache, the receiving node records the interest in the cache and the node from where it has come, and then sends the interest message on to its neighbours. In this way, a 'gradient' is formed pointing back towards the source, indicating a path of data flow towards the sink that broadcasted the interest (see Figure 10.5).

But data-centric routing is far from the only option, as we see in the following two sections.

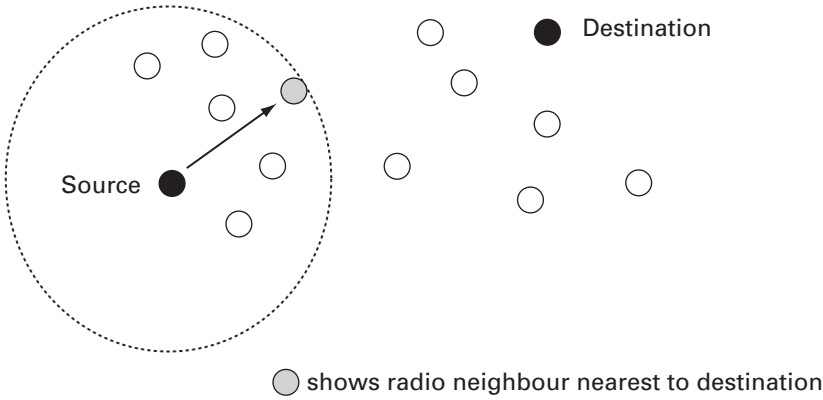


Figure 10.6 The principle of geographic routing.

10.9.2 WSN approaches – geographic routing

In a real-world deployment of a sensor network, it is clear from our earlier discussion that location, i.e. geography, is important for a number of applications. We note that even for the example of the data-centric approach above, sensor nodes and events have some location information that is important to the application. So, it seems natural to consider location or geographic information as candidates for routing for sensor networks.

Again, we take as our example one of the most-cited works in the field, which in this case is the one of greedy perimeter stateless routing (GPSR) [2]. The basic principle of GPSR reflects the basic principle of location or geographic routing – simply that we should always forward packets to nodes that get us closer (geographically) to our destination. In Figure 10.6, we see the source and its intended destination. In the greedy approach, the source always uses the node that is within radio range and closest to the destination. In this case, it is the shaded node. This process of selecting the forwarding node continues until the packet reaches its destination. Various refinements may be added to the method to overcome obstacles such as local maxima.

The principal advantage of geographic routing is that a node needs information about only its immediate neighbours, as forwarding decisions are made on location information of the destination and the neighbours.

However, geographic routing in this way requires the presence of an authoritative and secure source of location information, and a mechanism (e.g. a secure server) that will provide mapping between locations and node addresses in order that packets can be transmitted on to neighbouring nodes.

10.9.3 WSN approaches – other routing mechanisms

Data-centric and geographic approaches are not the only mechanisms under consideration for routing in unstructured WSNs. There is a large amount of literature relating to energy aware or energy efficient routing. This is designed to take into account the consumption of valuable battery power which results from packet forwarding. Various strategies can be used for energy conservation.

- Energy conservation for the network as a whole by load distribution, for example multi-path routing, to avoid draining batteries on a single path. This may be implemented by discovering energy rich nodes and using those first.
- Controlling transmission power so that only enough energy is used for the transmission to reach known neighbours.
- Optimisation of sleep/power-down/idle states of a sensor node in order to conserve energy. This option is already used to some degree in most WSNs.

Of course, the importance of energy conservation will vary depending on the application. Indeed, the various approaches (data-centric, geographic, energy aware) try to optimise for a specific network scenario and so they may not be appropriate for the general case. Moving to the general case is complicated, for a number of reasons.

- The number of parameters in a mobile network and their range of possible values, coupled with possible different traffic models, different failure regimes, node densities, radio ranges, MAC protocols, and different mobility models, for example, means that the problem space is very complex.
- The complexity of the scenario means that networks are relatively rarely built, and much of the evaluation work is conducted under

simulation. It is not clear that the research community has complete agreement on simulation scenarios and how simulations should be conducted [3].

In summary, we are of the view that energy aware routing is not yet suitable for the implementation stage.

Let us next move on to consider WSNs where structure and organisation is allowed.

10.10 Structured WSNs

To illustrate the structured WSN we may once again borrow from our mesh figures in Chapter 3. The structured WSN may have internal structure as shown in Figure 3.6, or more likely it will have both structure and a way to reach outside the WSN, analogous to the access mesh of Figure 3.7.

It is interesting to note the similarity between the structure shown in Figure 3.7 and the structure enabled by the IEEE 1451 approach for WSNs which we showed in Section 10.4.1. Legacy sensor applications which are moving to wireless clearly expect to be able to use existing infrastructure for backhaul.

To support this observation, and in contrast to unstructured approaches, structured WSNs are often the focus of industry and standards activity. Here, mobility needs are low, typically zero, but flexibility and interoperability needs are high. For widespread industrial deployment, the wireless sensor network must be seen to be reliable and easy to work with and the creation of a standard is often the best way forward in this case.

10.10.1 WSN approaches – hierarchical

Before we move into detail, let us review some general features of the hierarchical approach, unspecific to WSNs. In fact, there are several reasons why hierarchical networks have been preferred, historically:

- hierarchical routing is efficiently achieved,
- super nodes can take the processing load off regular nodes,

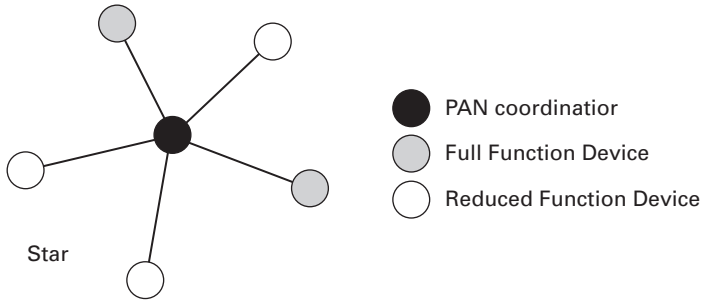


Figure 10.7 802.15.4 star configuration.

- super nodes can guide traffic off the network most quickly (this reduces hop count and improves throughput), and
- security also drives a hierarchical approach – via trust centres.

As we discuss structure in this section, we will also come across the idea that not all nodes need be equal. Hierarchy is connected, but not equivalent, to the idea that not all nodes need be equal, and that there are diverse benefits of inequality, such as bigger processors for routing information and storage, gateways to the outside world including the Internet and access to bigger power supplies.

Let us examine structured approaches by a convenient and relevant example.

10.10.2 Structured versus unstructured

We can look at structured versus unstructured approaches by taking the example of 802.15.4, since this includes options to create either type of network. In reality IEEE 802.15.4 does not specify the configurations, but does supply a PHY and MAC which are capable of being configured thus, typically by ZigBee.

The first of three examples is the star network, shown in Figure 10.7. This is the simplest configuration, which also provides the lowest latency, by disallowing multi-hopping.

Two types of device are shown in the figure and are defined as follows. The full function device (FFD) is the more capable device and can fulfil

any network function; the reduced function device (RFD) is much simpler (cheaper) and can exist only as an end device. Put slightly differently, the FFD can talk to any other device, whilst the RFD can talk only to the FFD, which is its direct parent in the hierarchy. In the star or any other configuration, the most child nodes an FFD can support is 254 – this compares very favourably to Bluetooth's limit of only 7 slaves.

In any 802.15.4 personal area network (PAN), one of the FFDs must elect itself as the PAN co-ordinator. This then has many responsibilities. For example, it must select a free channel at start-up, it must handle address allocation, it must organise beacons where they are used, it must act as the bridge to other networks, if there is a need for one, it must act as the trust centre to co-ordinate security, including cryptographic key distribution, where this is used. Usually the PAN co-ordinator is mains powered. FFDs which are not the PAN co-ordinator will act as routers in those PANs which are more complex than the star. Already we can see a link between node equality and system hierarchy, which goes beyond routing, i.e. the pull in this case for some nodes to be mains powered.

Multiple instances of stars are completely independent of each other. A useful comparison is that the star is no more than the access point architecture of 802.11 and it could be expanded by adding wired infrastructure, just as with 802.11. However, 802.15.4 can be more capable than this in the wireless domain alone, as our second of three examples shows.

The most flexible and complex approach is the mesh, or peer-to-peer network as it is termed in 802.15.4, as shown in Figure 10.8. In contrast to all other network options the mesh, as shown, is an unstructured network, i.e. it is has a flat hierarchy. Whilst it would be possible to terminate the mesh with RFDs at the edge, this would curtail the possibility for mesh expansion beyond these devices, since they are incapable of acting as routers. As one of the main attractions for mesh is its ad hoc expansion, this would be an unusual step, unless a specific application clearly demanded it. Hence we shall assume that the mesh is most attractive when populated in the main by FFDs to realise its full performance.

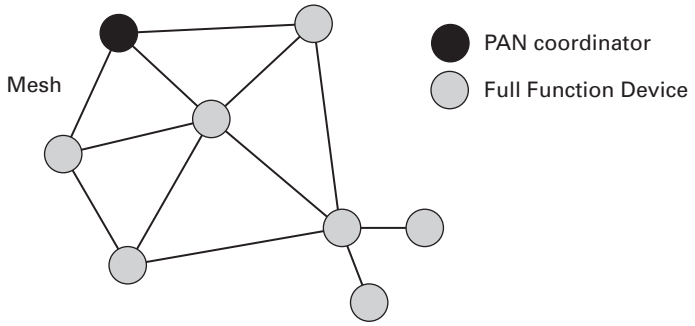


Figure 10.8 802.15.4 mesh (peer-to-peer) configuration.

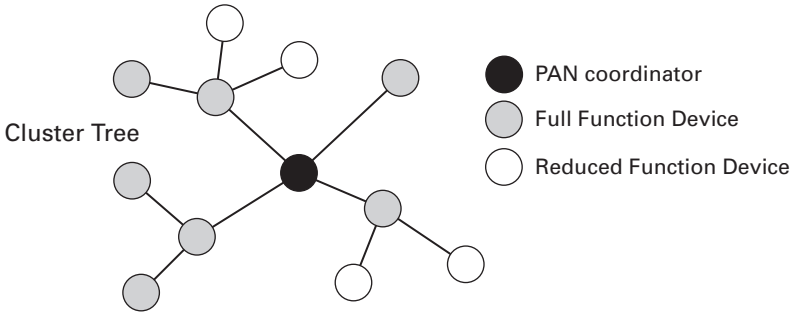


Figure 10.9 802.15.4 cluster tree configuration.

The principle advantages of the mesh are

- good and flexible coverage which can be extended simply by adding more nodes without any planning required, and
- redundancy due to the potential for multipath routing.

Since the mesh has no hierarchy, any node can communicate with any other node which is in range. This makes medium access most conveniently implemented by a contention based scheme. 802.15.4 offers the option of CSMA/CA. There is also a slotted version of CSMA/CA, supported by beacons generated from the PAN co-ordinator, but this fits better with stars and the last of our three examples, the cluster tree, shown in Figure 10.9. The reason for this is that propagating beacon information, which is clearly time sensitive, will become increasingly difficult across a larger mesh.

The cluster tree clearly has structure, but it represents a compromise between the simple star and the complex mesh, since it is quite flexible, but has simplified routing. It is also lower cost than the mesh since the use of RFDs as end nodes is expected. Clearly this does involve an element of planning the deployment. Usually, cluster heads are mains powered and end nodes are battery powered.

Routing in the star and the cluster tree works a little like the hierarchical approach of IP, as we showed earlier in this chapter. The node addresses similarly contain both identification and location information. A node in the cluster tree determines routing information directly from the destination node address. In contrast, in the mesh, AODV is used, which is a more complicated protocol.

ZigBee/802.15.4 configuration

Common to all three network architectures is the need to configure the system. One choice to make is the network architecture itself, and the other key choice is whether to use beacons or not.

With respect to architecture, there are two key parameters which can be set to influence how the network will automatically build its topology. These are the maximum network depth and the maximum number of routers (devices which are FFDs, but excluding the PAN co-ordinator). For example, to create a star, the network depth can be set at 1 and the number of routers can be set at 0.

In architectures other than the mesh, which typically does not use beacons, the choice of whether to use beacons or not comes down to the application needs. If the application is driven by timeliness of communication, such as a wireless mouse requiring regular communication, then it is most convenient to use beacons. If the application is event driven, such as a monitoring application, then non-beaconing may be more appropriate. Note that the use of beacons can enable longer sleep modes for suitable periodic applications and so conserve power better.

The foregoing has illustrated that adopting network structure can dictate that node inequality, i.e. diversity, should also exist, but that node diversity can go far beyond the network level. We examine node inequality as distinct from hierarchy next.

10.10.3 All nodes equal versus unequal

Apart from an inequality in the wireless routing functionality as discussed above, nodes may also be deliberately chosen to be unequal in other ways. This includes, for example, their access to power, their processing ability and their extra network connections, for example to the wired network.

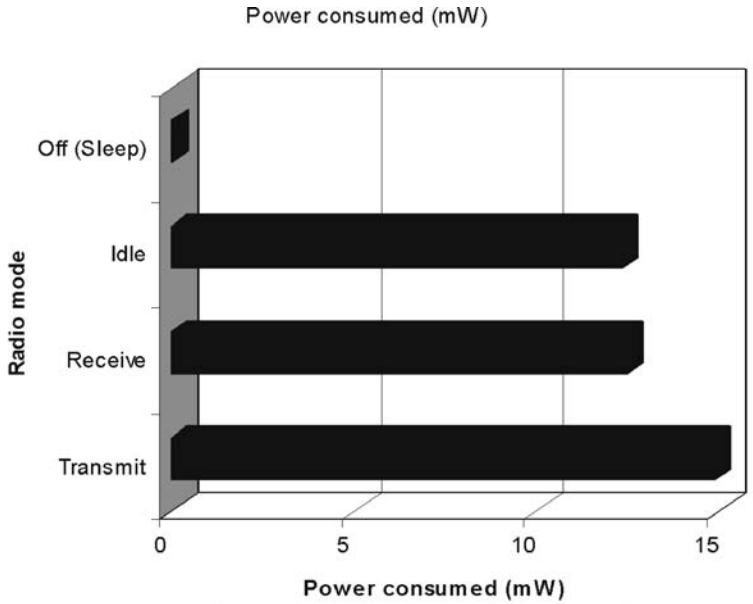
As we have stressed, WSNs are power constrained. However, if we allow inequality of nodes, then we may distribute the power constraint unevenly. This can really help with those nodes which are the most power constrained. If, for example, we have a star network then the nodes at the edge may be battery powered or may be harvesting energy, whilst the central node could be mains powered. This is exactly what happens in a WSN light switch application. The edge nodes are the light switches. In some cases these are power harvesting from the push action on the switch itself. Such nodes do not need to transmit regularly, only when operated.

However, simply reducing the transmit time is only one of the available possibilities for power saving. We could design the system so as to make the node idle[13] most of the time rather than actively receiving. But better than this, as shown in Figure 10.10, is actually to switch the transceiver/node off when not in use, i.e. put the circuits into a sleep mode. To have nodes mostly sleeping is the aim of many power constrained WSNs since, as the figure shows, active or idle transceiver circuitry is the largest power drain within a node, including the sensors and the processor. The power consumption data in Figure 10.10 were gathered from measurements by Shurgers *et al.* [4].

Clearly we have then sacrificed latency for power savings. This is because if a node sleeps we must either wait for a node to wake up if it does so periodically, or we must cause it to wake up by some external action. We have also sacrificed network flexibility since the edge nodes are not fully functional routers and cannot be used to extend the network.

This is directly related both to the choice of system design in 802.15.4 of end nodes versus co-ordinators (reduced functionality, and thus reduced power consumption) and to the choice of whether 802.15.4 uses regular beaconing or a random access mode (which allows the nodes to sleep for longer).

[13] Idle means powered up and ready to transmit or receive data.



	Transmit	Receive	Idle	Off (Sleep)
Power consumed (mW)	14.88	12.5	12.35	0.016

Figure 10.10 Relative power consumption of different radio modes.

Finally, for balance, it must be said that not all WSNs have power consumption at the very top of their list of requirements. Emergency services use WSNs for vital signs monitoring of active staff and environment monitoring. Here battery life is important, but equally important is network flexibility and low latency. Such WSNs are designed for low power, but do not typically enter sleep modes. The personnel monitoring units need to be fully functional routers such that any node may extend the network, but they may have reduced processing power compared to the base unit.

10.11 External routing and transport options

We have assumed so far that there is always a gateway device providing access to the sensor network. Information is injected into the network through a gateway and the gateway may return the result. However, in

principle, it is possible to access any node in the sensor network directly. But, typically gateways are used because:

1. the radio transceivers in the sensors have limited range so remote access to the sensor field requires a gateway which can bridge the technologies between the client system and the sensor network,
2. the communication protocols being used in the sensor network are typically not based on the Internet protocol suite, as these are typically thought to be too heavyweight (in terms of both code base and use of bits in transmission) for the resource limited environments of the sensor network realm. As we have said, even 6LoWPAN using IPv6 requires some gateway functionality, albeit relatively simple.

Finally we would like to point out that, just as for mesh networks, we should not fall into the pitfall of considering security too late in the system design. In fact, with the possibility of external access to the sensor network, system security and access control becomes a major issue. For example, consider a road traffic control system that uses information from sensors in order to control traffic flow within an urban area. Should that network be subverted by an attacker (for whatever purpose) the consequences could range from a large degree of inconvenience to motorists to potentially dangerous conditions on the highways and roads.

10.12 WSN summary

In this chapter we have introduced WSNs and shown that they have many similarities to mesh networks, but with even greater constraints, notably in terms of power consumption. Both WMNs and WSNs offer ease of deployment and good coverage properties – in the case of mesh this allows easy avoidance of RF shadowing in dense environments, and in the case of sensor networks this allows a sufficient density and arrangement of sensors to be deployed on site without constraint.

Similarly to mesh we have seen that the presence of structure in the network can markedly influence its performance. We can draw some clear parallels. For mesh we have shown in earlier chapters that adding infrastructure helps avoid scalability and quality of service problems. In the

WSN case, structure also helps scalability, in that by reducing the relay load for many nodes within the network, such as edge nodes, scalability is possible without power consumption problems, i.e. the low duty cycle of WSNs can be preserved. In a WSN this requires that some nodes are more capable than others and these nodes are typically not battery powered. Such nodes would be co-ordinators in 802.15.4 networks and would have mains power, external network connections and serve as trust centres for security applications.

We take the view that network structure and node diversity are enablers of flexible yet practical systems. Moreover, these can be efficiently standardised, which in turn enables adoption by industry. In other words, we continue to question the need for truly unstructured WSNs in many cases. Finally we note that major applications foreseen for WSNs at the moment include legacy wired applications evolving to wireless, rather than a revolutionary new killer WSN application.

References

1. Intanagonwiwat C., Govindan R., Estrin D., Directed diffusion: a scalable and robust communication paradigm for sensor networks, *Proc. Sixth Annual Int. Conf. on Mobile Computing and Networking (MobiCOM '00)*, Boston, MA, August 2000.
2. Karp B., Kung H.T., GPSR: greedy perimeter stateless routing for wireless networks, *Proc. Sixth Annual Int. Conf. on Mobile Computing and Networking (MobiCOM '00)*, Boston, MA, August 2000.
3. Kurwoksi S., Camp T., Colagrosso M., MANET simulation studies: the Incredibles, *ACM SIGMOBILE Mobile Computing and Communications Review*, **9** (4), pp. 50–61, October 2005.
4. Shurgers C., Tsiatsis V., Ganeriwal S., Srivastava M., Optimizing sensor networks in the energy-latency-density design space, *IEEE Trans. Mobile Computing*, **1** (1), January–March 2002.

Abbreviations

3G	third generation mobile, meant generically
6LoWPAN	IETF effort focused on running IPv6 over 802.15.4
AAA	access, authentication and authorisation (server)
ADSL	asymmetric digital subscriber line
AODV	adaptive on-demand distance vector, a favoured reactive ad hoc routing protocol
APC	automatic power control
ARQ	automatic repeat request
BER	bit error rate
BPSK	binary phase shift keying
CBR	constant bit rate
CDMA	code division multiple access
CPE	customer premises equipment
CSMA/CA	carrier sense multiple access/collision avoidance
CSMA/CD	carrier sense multiple access/collision detection
DCCP	datagram congestion control protocol
DFS	dynamic frequency selection
DNS	domain name server
DoS	denial of service
DRM	digital rights management
DSL	digital subscriber line
DSR	dynamic source routing (routing protocol)
DSSS	direct sequence spread spectrum
DTI	Department of Trade and Industry (UK)
EDGE	enhanced data rates for GSM evolution
ERP	effective radiated power
FDD	frequency division duplex
FDMA	frequency division multiple access

FFD	full function device
FHSS	frequency hopping spread spectrum
FQDN	fully qualified domain name
FSK	frequency shift keying
FWA	fixed wireless access
GPS	global positioning system
GPSR	greedy perimeter stateless routing
GSM	second generation mobile, meant generically
HART	sensor networking standard, now includes wireless
HVAC	heating, ventilation and air conditioning
IEEE	Institute of Electrical and Electronic Engineers (publish IEEE 802.x etc.)
IEEE 1451	wireless networking for sensors and TEDS standard
IEEE 802.11	'WiFi' standard (PHY and MAC layers)
IEEE 802.15.4	low power network standard (PHY and MAC layers)
IETF	Internet Engineering Task Force (responsible for TCP/IP, mobileIP etc.)
IP	Internet protocol
IPv6	expansion of current Internet protocol (IPv4) to handle more address space plus other improvements
ISA-100	sensor networking standard, now includes wireless
ISP	Internet service provider
ITS	intelligent transport system
LAN	local area network
LoS	line of sight (for RF path)
MAC	medium access control
MAN	metropolitan area network
MANET	mobile ad hoc networking (also MANet, a specific IETF group)
MEMS	microelectromechanical system
MIMO	multiple in, multiple out (an RF diversity technique)
MTU	maximum transmission unit
MUD	multi-user detection (CDMA systems)
NCAP	network capable application processor

OFDM	orthogonal frequency division multiplexing
OOB	out of band (signalling which does not need to share the data channel)
PAN	personal area network
PDA	personal digital assistant
PHY	physical layer
PMP	point to multi-point
PSK	phase shift keying
QAM	quadrature amplitude modulation
QoS	quality of service
QPSK	quadrature phase shift keying
RADIUS	remote access dial in user service
RF	radio frequency
RFD	reduced function device
RFID	radio frequency identification
RIP	routing information protocol
RTS/CTS	ready/clear to send (handshaking protocol)
Rx	receive, receiver
SIR	signal to interference ratio
STIM	smart transducer interface module
TCP/IP	transmission control protocol/Internet protocol
TDD	time division duplex
TDMA	time division multiple access
TEDS	transducer electronic data sheets, see IEEE 1451
TPC	transmit power control
Tx	transmit, transmitter
UDP	user datagram protocol
UMTS	universal mobile telecommunication service
UWB	ultra wide band
VANET	vehicular ad hoc network
VBR	variable bit rate
VoIP	voice over Internet protocol
WBAN	wireless body area network
WLAN	wireless local area network

WMAN	wireless metropolitan area network
WMN	wireless mesh network
WPAN	wireless personal area network
WSN	wireless sensor network
WWAN	wireless wide area network
ZigBee	low power network standard (networking and application layers)

Selected definitions

access mesh	a development of a pure, isolated mesh to include connection to other networks such as the internet via access points. Access meshes are expected to be the dominant form of mesh deployed for public use. The density of access points has a large effect on mesh properties.
access points	see **access mesh**
allocation	relates to allocating spectrum for a service type and is usually internationally agreed, cf. assignment
arbitrary network	as defined by Gupta and Kumar (see Chapter 4), see also **random network**
assignment	relates to assigning a pre-existing allocation to an operator and is usually nationally agreed, cf. allocation
hop	a point to point connection, with no intermediate nodes
link	a point to point connection
network capacity	the average total traffic rate (bps) that can circulate within the network.
node transport capability	the average rate at which data can be passed through a node (from Rx to Tx)
path	equivalent to **route**
per-user average throughput	the average data rate that can be sourced/sinked from/to the user of a node
per-user traffic throughput	the data rate and associated Erlang loading that can be sourced/sinked from/to the user of a node

pure mesh	an isolated mesh
random network	as defined by Gupta and Kumar (see Chapter 4), see also **arbitrary network**
raw transmission rate	the raw on-air bit-rate (bps) over a radio **link**
relay nodes	mesh internal infrastructure connection points
route	a point to point connection description via intermediate network nodes
seed node	a node or nodes deployed on day one to ensure connectivity by guaranteeing a minimum density of nodes
spectrum management models	three main spectrum management models: command and control (licensed e.g. cellular commons, unlicensed e.g. WLAN market); secondary trading; flexible, 'change of use' allowed, probably via the regulator.
transmission bandwidth	the spectrum bandwidth (Hz) required to accommodate the **raw transmission rate**

Appendix: Mobility models

The aim of using a mobility model is to reflect as accurately as practicable the real conditions themselves. One way to do this is to use motion traces, which are logs of real-life node movements over a representative period of time. There are not many such logs available for use even with established cellular schemes, and none are known to this author which cover mesh environments. The focus then must move to synthetic models. Such a model will deal with a number of nodes and may include parameters such as speed and direction of movement, the ability to pause at some locations and a bound to the model area. The models available are mostly fairly simple to implement, since they are intended for use in simulators where a tractable run time is expected. It is probably the case that present models err on the side of simplicity at the expense of realism. On the other hand, moving too close to the actual environment requires a very specific model – which may then not be adequately representative of all environments. The choice of model is thus a subject which needs to be understood, in order to interpret specific protocol and other simulation results for wider contexts.

Camp *et al*. [1] review 12 different mobility models which have been applied to mesh simulations at various points in the published literature. Their work is an often quoted indication that the choice of model alone can strongly affect the results when testing the exact same routing protocol. For the purposes of this book three models are noted as being appropriate.

The random waypoint mobility model

This model has a base of randomly distributed destinations to which any node may move. The node will move with a random speed and will pause for a random amount of time at each point, before moving on again. The

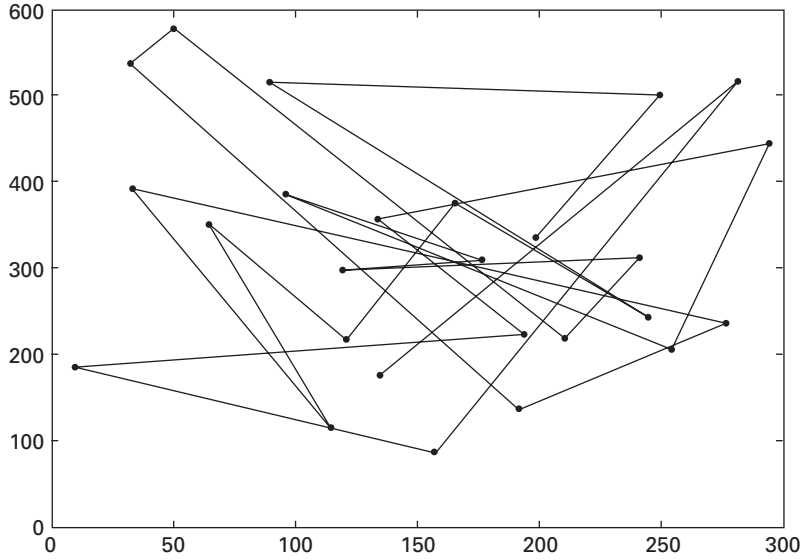


Figure A1 Random waypoint model, used with permission from Camp *et al*. [1].

model input parameters are the number of nodes, the bounds on both speed and pause time and the physical boundaries of the model area. Figure A1 shows the pattern of a single node in this model. The axes are in units of distance.

It is a matter of some current debate whether this model tends to concentrate the nodes at the centre of the model over time. If it does, the effect is thought to be quite light, but for the purposes of this book, this behaviour is interesting as it is reminiscent of the clustering effect which users may exhibit within a mesh, depending on the scenario. This fact may be referred to when simulation results using this model are interpreted. Like all the models, the pattern in Figure A1 can be scaled by changing parameters, even to the point of having quite a static distribution.

The random direction mobility model

Any tendency to concentrate nodes at the centre of the model area is avoided in this model, by design. Nodes travel all the way to the

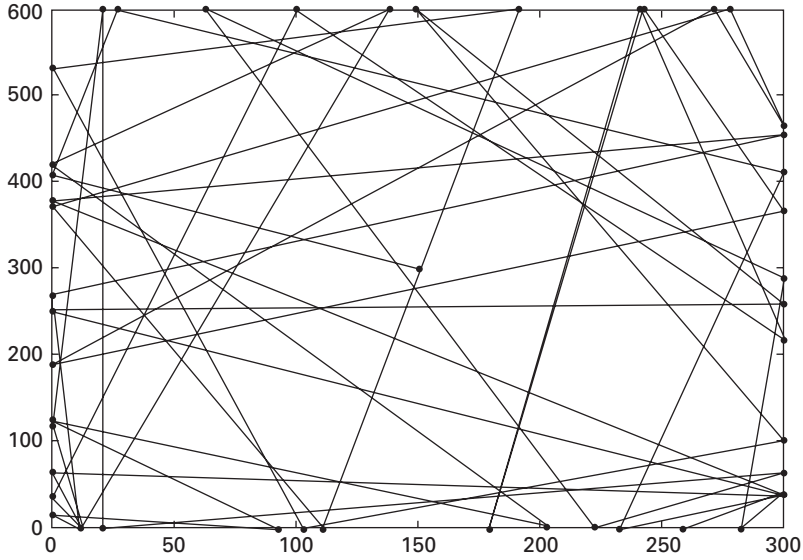


Figure A2 The random direction mobility model, used with permission from Camp *et al*. [1].

boundaries of the model before setting off again in another direction, with another speed. At this level it is often called the billiard ball model. However, pauses may be added, so that nodes spend time at the boundary before moving off again, in order to mimic the pauses seen in real-world user behaviour, see Figure A2. Of course not all users will really pause at the boundaries of a given area. Despite this, the model with pauses is useful in that it has been found to create a multiple hop situation more often than not, precisely due to pausing at the boundaries (maximising inter-node distance). This makes the model useful for modelling multi-hop networks. It is also a tractable model to use in a simulation environment.

The city section or Manhattan mobility model

The approach here is very direct. A dense urban environment with a regular grid of streets is assumed, hence the term Manhattan. Nodes are constrained to move within the streets and may be made to pause at street

corners to improve coverage around the corners. The inputs to the model are the street length, width and spacing. Such a model was used in the assessment of the 3GPP and GSM coverage quality.

Despite the appeal of a direct analogue of the real situation, this model has not seen much application in the study of mesh networks. It may be surmised that this is due to the very specific nature of the model, in other words how are the street dimensions to be picked? If a typical case is taken, how useful is that in reality, where a spread occurs? It would seem that that such models are worthy of further investigation.

Reference

1. Camp T., Boleng J., Davies V., A survey of mobility models for ad hoc network research, *Wireless Communication and Mobile Computing (WCMC),* Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, **2** (5), pp. 483–502, 2002.

About the author

Steve Methley has over 20 years' experience in telecommunications and data communications innovation, having led teams in the laboratories of British Telecom, Hewlett–Packard and Toshiba.

Work at BT Labs included the earliest long haul wavelength division multiplexing, used in live traffic field trials in BT's national optical network and BT's first demonstration of wireless last mile using fibre for antenna remoting. Several detailed aspects of this work were patented. Later physical layer work at HP Labs resulted in contributions being made to the ATM Forum, and IEEE 802.3 100 Mbps and Gigabit Ethernet committees for data communications, and led to participation in the ADSL Forum for last mile broadband access. Hewlett–Packard products resulting from Steve's work include networking hardware and test equipment.

Subsequently, Steve's interest moved higher up the protocol stack and he became the first research manager for Toshiba's new Telecoms Research Laboratory in Europe, responsible for setting the agenda for networks and protocols research for future wireless systems.

Most recently, as a consultant based in Cambridge, UK, work has been detailed and diverse, including radio and optical physical layers, transport protocols, algorithm modelling, business strategy, socio-economic analyses, regulation and futurology for a range of clients from start-up to international household-name companies.

Steve was awarded his PhD from Imperial College and his MBA from the Open University. He has a number of patents and published papers and is currently a Visiting Fellow at the University of Essex. He continues to be an active IEEE Standards Association contributor and an independent expert for the European Commission's Research Framework.

He can be contacted at steve.methley@ieee.org.

Index

Mesh hints and tips

This section offers 14 hints and tips, in the area of practical deployment. All hints and tips are drawn from material already presented. They are summarised in the table.

Table of hints and tips

Potential problem area	Hints and tips
Scalability	Use relay nodes or access points to split the original mesh into smaller meshes
Initial connectivity and QoS	Use seed nodes
Ongoing QoS	Deploy operator controlled nodes to control hop count and signal quality
Network stability	Deploy operator controlled nodes to reduce the extent of the users' impact on the network
Inelastic traffic demands	Minimise hop count and minimise re-routing
Security and trust aspects of ad hoc networks	Consider creating trust centres
Selfish user behaviour	Consider reward schemes, improved batteries or power harvesting
Upgradeability of handsets	This cannot be done piecemeal, and must include any extra performance from the outset
Mobility requirements	Plan performance via considering re-routing load (hand-overs per time), rather than node movement speed
Achievable antenna directionality	Not viable for typical handhelds, i.e. size constrained

Table of hints and tips (cont.)

Potential problem area	Hints and tips
Working outside the spectrum sweet spot	Can take advantage of higher frequencies and shorter hops
Meshing on user side and/or network side	Both are possible and are not mutually exclusive
Whether mesh or multi-hop is best	Multi-hop is less complex, but the distinction can blur depending on the application
Battery life effects in handhelds versus user position	Consider providing fixed operator controlled nodes near access points